AFRL-RI-RS-TR-2023-202

LEARN: LABEL-EFFICIENT ACTIVE RESILIENT NETWORKS

UNIVERSITY OF CALIFORNIA

NOVEMBER 2023

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2023-202 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
MILVIO FRANCO
Work Unit Manager

/ S /
MICHELE GRIECO, signed for
JULIE BRICHECK
Chief, Information Systems Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
NOVEMBER 2023	FINAL TECHNICAL REPORT	SEPTEMBER 2019	MAY 2023

4. TITLE AND SUBTITLE

LEARN: LABEL-EFFICIENT ACTIVE RESILIENT NETWORKS

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8750-19-1-0504	N/A	61101E
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
		R2TZ

6. AUTHOR(S)

Trevor Darrell

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
The Regents of the University of California 1608 4th ST STE 220 Berkeley CA 94710-1749	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL/RISB 525 Brooks Road Rome NY 13441-4505	DARPA/I20 675 N. Randolph St. Arlington VA 22203-2114	AFRL/RI & DARPA	AFRL-RI-RS-TR-2023-202

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The LEARN LwLL team conducted fundamental research on Zero-shot Image Classification, Few-Shot Image Classification, Few-Shot Object Detection, Few-Shot Video Classification, Few-Shot Machine Translation, and on Data-Efficient Fine-Tuning for Image Classification. Significant results were found across many research areas leading to improved performance on established benchmarks.

15. SUBJECT TERMS

Object Detection, Video Classification, Few-Shot Machine Translation, Data-Efficient Fine-Tuning for Image, algorithm classification

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	109

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
MILVIO FRANCO	N/A

PREVIOUS EDITION IS OBSOLETE.

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Table of Contents

1 Summary **1**

2 Introduction **1**

3 Methods, Assumptions, and Procedures **7**
3.1 Zero-Shot Image Classification 7
3.2 Few-Shot Image Classification 9
3.3 Few-Shot Object Detection 15
3.4 Few-Shot Video Classification 17
3.5 Few-Shot Machine Translation 21
3.6 Data-Efficient Fine-Tuning for Image Classification 23

4 Results **25**
4.1 Zero-Shot Image Classification 25
4.2 Few-Shot Image Classification 27
4.3 Few-Shot Object Detection 33
4.4 Few-Shot Video Classification 38
4.5 Few-Shot Machine Translation 40
4.6 Data-Efficient Fine-Tuning for Image Classification 44

5 Conclusions **49**
5.1 Zero-Shot Image Classification 49
5.2 Few-Shot Image Classification 49
5.3 Few-Shot Object Detection 50
5.4 Few-Shot Video Classification 50
5.5 Few-Shot Machine Translation 51
5.6 Data-Efficient Fine-Tuning for Image Classification 51
5.7 Final Observations 52

A Appendix: Full List of Team Publications **62**

List of Figures

1 A conceptual comparison of previous multi-label recognition (MLR) methods and our approach . . . 7
2 DualCoOp learns a pair of positive and negative prompts to quickly adapt powerful pretrained vision-text encoders to the MLR task. . . . 7
3 Comparisons to previous works on unsupervised domain adaptation. . . . 8
4 We propose Selective Entropy Optimization via Committee Consistency (SENTRY) for unsupervised DA. . . . 8
5 Overview of PACMAC. . . . 9
6 Active Domain Adaptation. . . . 10
7 We propose CLustering Uncertainty-weighted Embeddings (CLUE), a novel label acquisition strategy for Active DA. . . . 11
8 Pipeline of the proposed entropy minimization method. . . . 12
9 An illustration of our proposed open-set soft-consistency loss used to enhance outlier detection. . . . 12
10 Comparisons to previous works on semi-supervised learning. . . . 13
11 Few-shot class incremental learning . . . 14
12 We modernize a standard ConvNet (ResNet) towards the design of a hierarchical vision Transformer (Swin), without introducing any attention-based modules. . . . 15
13 **MaskCut** can discover multiple object masks in an image without supervision. . . . 16
14 VideoCutLER is a simple unsupervised multi-instance segmentation method for video. 16
15 ODISE Overview and Training Pipeline . . . 17
16 Predictions of STIN and I3D models. . . . 18
17 The Spatial-Temporal Interaction Network (STIN). . . . 18
18 Cross-video cycle-consistency for image representation learning. . . . 19
19 Illustration of our Temporal Contrastive Learning TCL Framework. . . . 20
20 Each curve represents a bag and points on the curve represent instances in the bag. . . . 21
21 Our EM-MIL model architecture . . . 21
22 Zip it model framework. . . . 23
23 ZipIt! merges models layer-wise by exploiting redundancy in their features. . . . 24
24 Multi-Session FSCIL accuracy (%) results on *mini*ImageNet. . . . 32
25 Zero-shot unsupervised object detection and instance segmentation using our CutLER model, which is trained without human supervision. . . . 35
26 Fine-tuning on MS-COCO with various annotation ratios. . . . 35
27 We fine-tune VideoCutLER for semi-supervised video instance segmentation on the YouTubeVIS-2019 dataset . . . 36
28 Qualitative Visualization on COCO (first 2 rows) and ADE20K (last row) validation and test sets. . . . 37
29 Applying our read-out network to prefixes of the syntactic tag sequence. . . . 47
30 Varying Partial Zip. . . . 48
31 Model Scale. . . . 48

List of Tables

1 Multi-label Recognition on MS-COCO and VOC2007 with partial labels. 25
2 Zero-Shot Multi-Label Recognition on MS-COCO[1] 26
3 Zero-Shot Multi-label Recognition on NUS-WIDE [2]. 26
4 Per-class average accuracies on DomainNet. 27
5 Accuracies on label-shifted and standard DA benchmarks. 27
6 Target test set accuracy on OfficeHome across MAE [3] and DINO [4] pretraining. . . . 28
7 Target test set accuracy on DomainNet across MAE [3] and DINO [4] pretraining. . . . 28
8 Accuracies on target test set for 4 DomainNet shifts of increasing difficulty spanning 5 domains: Real (R), Clipart (C), Sketch (S), Painting (P) and Quickdraw (Q). 29
9 Accuracy on the DomainNet dataset (%) for one-shot and three-shot settings on 4 domains. 30
10 Results on Office-Home and Office dataset (%). The value is the accuracy averaged over all adaptation scenarios. Performance on each setting is summarized in supplementary material. 30
11 Error rates (%) with standard deviation for CIFAR10, CIFAR100 on 3 different folds. . 31
12 AUROC of Table 11. 31
13 Evaluation of outlier detection on outliers unseen in unlabeled training data (AUROC). 32
14 Multi-Session FSCIL weighted average of accuracy (%) results on *mini*ImageNet using an identical setup to fig. 24 with memory distinction. 33
15 Classification accuracy on ImageNet-1K. 34
16 Zero-shot unsupervised multi-instance video segmentation on YouTubeVIS-2019 and YouTubeVIS-2021. 36
17 Open-vocabulary semantic segmentation performance. 37
18 Compositional action recognition over 174 categories. 38
19 Few-shot compositional action recognition on base categories and few-shot novel categories. 38
20 Video action recognition accuracy comparison with other unsupervised representation methods on UCF101. 39
21 Performance Comparison in Mini-Something-V2. 39
22 Performance Comparison on Jester and Kinetics-400. 40
23 Our EM-MIL detection results on THUMOS14 in percentage. 41
24 F1 on the WSJ test set for parsers using different syntactic representations and pre-trained encoders. 41
25 BLEU scores for previous supervised and unsupervised results from [1][5], [2][6] and our models. 43
26 Comparison of effect of using unsupervised MT training ($BT + AE$) with supervised data. 43
27 Comparison of effect of using unsupervised MT training (BT + AE) with comparable data. 43
28 Comparison of effect of using code switch (MLM_{cs}) with supervised data. 43
29 Comparison of effect of using **code switch** (MLM_{cs}) with **comparable** data. 43
30 Results for action sequence generation with Interscript [7] and topic-based summarization by [8]. 44
31 Results on VTAB-1k with ViT backbone (Part 1). 44
32 Results on VTAB-1k with ViT backbone (Part 2). 44
33 CIFAR Results. 45
34 ImageNet-1k (200+200) Results. 45
35 Multi-Dataset Results. 46

Approved for Public Release; Distribution Unlimited.

1 Summary

The LEARN LwLL team conducted fundamental research on Zero-shot Image Classification, Few-Shot Image Classification, Few-Shot Object Detection, Few-Shot Video Classification, Few-Shot Machine Translation, and on Data-Efficient Fine-Tuning for Image Classification. Significant results were found across many research areas leading to improved performance on established benchmarks. The following sections summarize key results obtained in each area. An appendix contains a complete list of publications supported across the team by this program.

2 Introduction

Zero-Shot Image Classification. Solving multi-label recognition (MLR) for images in the low-label regime is a challenging task with many real-world applications. Recent work learns an alignment between textual and visual spaces to compensate for insufficient image labels, but loses accuracy because of the limited amount of available MLR annotations. In this work, we utilize the strong alignment of textual and visual features pretrained with millions of auxiliary image-text pairs and propose Dual Context Optimization (DualCoOp) as a unified framework for partial-label MLR and zero-shot MLR. DualCoOp encodes positive and negative contexts with class names as part of the linguistic input (i.e. prompts). Since DualCoOp only introduces a very light learnable overhead upon the pretrained vision-language framework, it can quickly adapt to multi-label recognition tasks that have limited annotations and even unseen classes. Experiments on standard multi-label recognition benchmarks across two challenging low-label settings demonstrate the advantages of our approach over state-of-the-art methods.

Many existing approaches for unsupervised domain adaptation (UDA) focus on adapting under only data distribution shift and offer limited success under additional cross-domain label distribution shift. Recent work based on self-training using target pseudolabels has shown promise, but on challenging shifts pseudolabels may be highly unreliable and using them for self-training may cause error accumulation and domain misalignment. We propose Selective Entropy Optimization via Committee Consistency (`SENTRY`), a UDA algorithm that judges the reliability of a target instance based on its predictive consistency under a committee of random image transformations. Our algorithm then selectively minimizes predictive entropy to increase confidence on highly consistent target instances, while maximizing predictive entropy to reduce confidence on highly inconsistent ones. In combination with pseudolabel-based approximate target class balancing, our approach leads to significant improvements over the state-of-the-art on 27/31 domain shifts from standard UDA benchmarks as well as benchmarks designed to stress-test adaptation under label distribution shift.

Visual domain adaptation (DA) seeks to transfer trained models to unseen, unlabeled domains across distribution shift, but approaches typically focus on adapting convolutional neural network architectures initialized with supervised ImageNet representations. In this work, we shift focus to adapting modern architectures for object recognition – the increasingly popular Vision Transformer (ViT) – initialized with modern pretraining based on self-supervised learning (SSL). Inspired by the design of recent SSL approaches based on learning from partial image inputs generated via masking or cropping – either by learning to predict the missing pixels, or learning representational invariances to such augmentations – we propose `PACMAC`, a two-stage adaptation algorithm for self-supervised ViTs. `PACMAC` first performs in-domain SSL on pooled source and target data to learn task-discriminative features, and then probes the model's predictive consistency across a set of partial target inputs generated via a novel attention-conditioned masking strategy, to identify reliable candidates for self-training. Our simple approach leads to consistent performance gains over competing methods that use ViTs and self-supervised initializations on standard object recognition benchmarks.

Few-Shot Image Classification. Generalizing deep neural networks to new target domains is critical to their real-world utility. In practice, it may be feasible to get some target data labeled, but to be cost-effective it is desirable to select a maximally-informative subset via active learning (AL). We study the problem of AL under a domain shift, called Active Domain Adaptation (Active DA). We demonstrate how existing AL approaches based solely on model uncertainty or diversity sampling are less effective for Active DA. We propose Clustering Uncertainty-weighted Embeddings (`CLUE`), a novel label acquisition strategy for Active DA that performs uncertainty-weighted clustering to identify target instances for labeling that are both uncertain under the model and diverse in feature space.

CLUE consistently outperforms competing label acquisition strategies for Active DA and AL across learning settings on 6 diverse domain shifts for image classification.

Contemporary domain adaptation methods are very effective at aligning feature distributions of source and target domains without any target supervision. However, we show that these techniques perform poorly when even a few labeled examples are available in the target domain. To address this semi-supervised domain adaptation (SSDA) setting, we propose a novel Minimax Entropy (MME) approach that adversarially optimizes an adaptive few-shot model. Our base model consists of a feature encoding network, followed by a classification layer that computes the features' similarity to estimated prototypes (representatives of each class). Adaptation is achieved by alternately maximizing the conditional entropy of unlabeled target data with respect to the classifier and minimizing it with respect to the feature encoder. We empirically demonstrate the superiority of our method over many baselines, including conventional feature alignment and few-shot methods, setting a new state of the art for SSDA.

Semi-supervised learning (SSL) is an effective means to leverage unlabeled data to improve a model's performance. Typical SSL methods like FixMatch assume that labeled and unlabeled data share the same label space. However, in practice, unlabeled data can contain categories unseen in the labeled set, *i.e.*, outliers, which can significantly harm the performance of SSL algorithms. To address this problem, we propose a novel Open-set Semi-Supervised Learning (OSSL) approach called *OpenMatch*. Learning representations of inliers while rejecting outliers is essential for the success of OSSL. To this end, OpenMatch unifies FixMatch with novelty detection based on one-vs-all (OVA) classifiers. The OVA-classifier outputs the confidence score of a sample being an inlier, providing a threshold to detect outliers. Another key contribution is an open-set soft-consistency regularization loss, which enhances the smoothness of the OVA-classifier with respect to input transformations and greatly improves outlier detection. OpenMatch achieves state-of-the-art performance on three datasets, and even outperforms a fully supervised model in detecting outliers unseen in unlabeled data on CIFAR10.

Few-shot class incremental learning—the problem of updating a trained classifier to discriminate among an expanded set of classes with limited labeled data—is a key challenge for machine learning systems deployed in non-stationary environments. Existing approaches to the problem rely on complex model architectures and training procedures that are difficult to tune and re-use. In this paper, we present an extremely simple approach that enables the use of ordinary logistic regression classifiers for few-shot incremental learning. The key to this approach is a new family of *subspace regularization* schemes that encourage weight vectors for new classes to lie close to the subspace spanned by the weights of existing classes. When combined with pretrained convolutional feature extractors, logistic regression models trained with subspace regularization outperform specialized, state-of-the-art approaches to few-shot incremental image classification by up to 23% on the *mini*ImageNet dataset. Because of its simplicity, subspace regularization can be straightforwardly configured to incorporate additional background information about the new classes (including class names and descriptions specified in natural language); this offers additional control over the trade-off between existing and new classes. Our results show that simple geometric regularization of class representations offers an effective tool for continual learning.

The "Roaring 20s" of visual recognition began with the introduction of Vision Transformers (ViTs), which quickly superseded ConvNets as the state-of-the-art image classification model. A vanilla ViT, on the other hand, faces difficulties when applied to general computer vision tasks such as object detection and semantic segmentation. It is the hierarchical Transformers (e.g., Swin Transformers) that reintroduced several ConvNet priors, making Transformers practically viable as a generic vision backbone and demonstrating remarkable performance on a wide variety of vision tasks. However, the effectiveness of such hybrid approaches is still largely credited to the intrinsic superiority of Transformers, rather than the inherent inductive biases of convolutions. In this work, we reexamine the design spaces and test the limits of what a pure ConvNet can achieve. We gradually "modernize" a standard ResNet toward the design of a vision Transformer, and discover several key components that contribute to the performance difference along the way. The outcome of this exploration is a family of pure ConvNet models dubbed ConvNeXt. Constructed entirely from standard ConvNet modules, ConvNeXts compete favorably with Transformers in terms of accuracy and scalability, achieving 87.8% ImageNet top-1 accuracy and outperforming Swin Transformers on COCO detection and ADE20K segmentation, while maintaining the simplicity and efficiency of standard ConvNets.

Few-Shot Object Detection. We propose **Cut**-and-**LE**a**R**n (CutLER), a simple approach for

training unsupervised object detection and segmentation models. We leverage the property of self-supervised models to 'discover' objects without supervision and amplify it to train a state-of-the-art localization model without any human labels. CutLER first uses our proposed MaskCut approach to generate coarse masks for multiple objects in an image, and then learns a detector on these masks using our robust loss function. We further improve performance by self-training the model on its predictions. Compared to prior work, CutLER is simpler, compatible with different detection architectures, and detects multiple objects. CutLER is also a zero-shot unsupervised detector and improves detection performance AP_{50} by over 2.7× on 11 benchmarks across domains like video frames, paintings, sketches, etc. With finetuning, CutLER serves as a low-shot detector surpassing MoCo-v2 by 7.3% AP^{box} and 6.6% AP^{mask} on COCO when training with 5% labels. We have also demonstrated that CutLER can significantly improve the performance of label-efficient learning on datasets from various domains, including the natural image dataset MSCOCO and remote sensing images xView.

Existing approaches to unsupervised video instance segmentation typically rely on motion estimates and experience difficulties tracking small or divergent motions. We present VideoCutLER, a simple method for unsupervised multi-instance video segmentation without using motion-based learning signals like optical flow or training on natural videos. Our key insight is that using high-quality pseudo masks and a simple video synthesis method for model training is surprisingly sufficient to enable the resulting video model to effectively segment and track multiple instances across video frames. We show the first competitive unsupervised learning results on the challenging YouTubeVIS-2019 benchmark, achieving 50.3% AP_{50}^{video}, surpassing the previous state-of-the-art by over 10× (50.3% vs. 4.8%). VideoCutLER can also serve as a strong pretrained model for supervised video instance segmentation tasks, exceeding DINO by 15.9% on YouTubeVIS-2019 in terms of AP^{video}.

We present ODISE: Open-vocabulary DIffusion-based panoptic SEgmentation, which unifies pretrained text-image diffusion and discriminative models to perform open-vocabulary panoptic segmentation. Text-to-image diffusion models have the remarkable ability to generate high-quality images with diverse open-vocabulary language descriptions. This demonstrates that their internal representation space is highly correlated with open concepts in the real world. Text-image discriminative models like CLIP, on the other hand, are good at classifying images into open-vocabulary labels. We leverage the frozen internal representations of both these models to perform panoptic segmentation of any category in the wild. Our approach outperforms the previous state of the art by significant margins on both open-vocabulary panoptic and semantic segmentation tasks. In particular, with COCO training only, our method achieves 23.4 PQ and 30.0 mIoU on the ADE20K dataset, with 8.3 PQ and 7.9 mIoU absolute improvement over the previous state of the art.

Few-Shot Video Classification. Human action is naturally compositional: humans can easily recognize and perform actions with objects that are different from those used in training demonstrations. In this paper, we study the compositionality of action by looking into the dynamics of subject-object interactions. We propose a novel model which can explicitly reason about the geometric relations between constituent objects and an agent performing an action. To train our model, we collect dense object box annotations on the Something-Something dataset. We propose a novel compositional action recognition task where the training combinations of verbs and nouns do not overlap with the test set. The novel aspects of our model are applicable to activities with prominent object interaction dynamics and to objects which can be tracked using state-of-the-art approaches; for activities without clearly defined spatial object-agent interactions, we rely on baseline scene-level spatio-temporal representations. We show the effectiveness of our approach not only on the proposed compositional action recognition task, but also in a few-shot compositional setting which requires the model to generalize across both object appearance and action category.

Recent works have advanced the performance of self-supervised representation learning by a large margin. The core among these methods is intra-image invariance learning. Two different transformations of one image instance are considered as a positive sample pair, where various tasks are designed to learn invariant representations by comparing the pair. Analogically, for video data, representations of frames from the same video are trained to be closer than frames from other videos, i.e. intra-video invariance. However, cross-video relation has barely been explored for visual representation learning. Unlike intra-video invariance, ground-truth labels of cross-video relation is usually unavailable without human labors. In this paper, we propose a novel contrastive learning method which explores the cross-video relation by using cycle-consistency for general image representation learning. This allows to collect positive sample pairs across different video instances, which we hypothesize will lead to

Approved for Public Release; Distribution Unlimited.

higher-level semantics. We validate our method by transferring our image representation to multiple downstream tasks including visual object tracking, image classification, and action recognition. We show significant improvement over state-of-the-art contrastive learning methods.

Learning to recognize actions from only a handful of labeled videos is a challenging problem due to the scarcity of tediously collected activity labels. We approach this problem by learning a two-pathway temporal contrastive model using unlabeled videos at two different speeds leveraging the fact that changing video speed does not change an action. Specifically, we propose to maximize the similarity between encoded representations of the same video at two different speeds as well as minimize the similarity between different videos played at different speeds. This way we use the rich supervisory information in terms of 'time' that is present in otherwise unsupervised pool of videos. With this simple yet effective strategy of manipulating video playback rates, we considerably outperform video extensions of sophisticated state-of-the-art semi-supervised image recognition methods across multiple diverse benchmark datasets and network architectures. Interestingly, our proposed approach benefits from out-of-domain unlabeled videos showing generalization and robustness. We also perform rigorous ablations and analysis to validate our approach.

Weakly-supervised action localization problem requires training a model to localize the action segments in the video given only video level action label. It can be solved under the Multiple Instance Learning (MIL) framework, where a bag (video) contains multiple instances (action segments). Since only the bag's label is known, the main challenge is to assign which key instances within the bag trigger the bag's label. Most previous models use an attention-based approach. These models use attention to generate bag's representation from instances and then train it via bag's classification. In this work, we explicitly model the key instances assignment as a hidden variable and adopt an Expectation-Maximization (EM) framework. We derive two pseudo-label generation schemes to model the E and M process and iteratively optimize the likelihood lower bound. We also show that previous attention-based models implicitly violate the MIL assumptions that instances in negative bags should be uniformly negative. In comparison, Our EM-MIL approach more accurately models these assumptions. Our model achieves state-of-the-art performance on two standard benchmarks, THUMOS14 and ActivityNet1.2, and shows the superiority of detecting relative complete action boundary in videos containing multiple actions.

Few-Shot Machine Translation. We present an incremental syntactic representation that consists of assigning a single discrete label to each word in a sentence, where the label is predicted using strictly incremental processing of a prefix of the sentence, and the sequence of labels for a sentence fully determines a parse tree. Our goal is to induce a syntactic representation that commits to syntactic choices *only as they are incrementally revealed by the input*, in contrast with standard representations that must make output choices such as attachments speculatively and later throw out conflicting analyses. Our learned representations achieve 93.72 F1 on the Penn Treebank with as few as 5 bits per word, and at 8 bits per word they achieve 94.97 F1, which is comparable with other state of the art parsing models when using the same pre-trained embeddings. We also provide an analysis of the representations learned by our system, investigating properties such as the interpretable syntactic features captured by the system and mechanisms for deferred resolution of syntactic ambiguities.

We conduct an empirical study of neural machine translation (NMT) for truly low-resource languages, and present a training curriculum fit for cases when both parallel training data and compute resource are lacking, reflecting the reality of most of the world's languages and the researchers working on these languages. Previously, unsupervised NMT, which employs back-translation (BT) and auto-encoding (AE) tasks has been shown barren for low-resource languages. We demonstrate that leveraging comparable data and code-switching as weak supervision, combined with pre-training with BT and AE objectives, result in remarkable improvements for low-resource languages even when using only modest compute resources. The training curriculum proposed in this work achieves BLEU scores that improve over supervised NMT trained on the same backbone architecture by +12.3 and +13.5 BLEU for Gujarati→English and vice versa, +11.2 BLEU and +9.6 for Kazakh→English and vice versa, and +7.6 and +3.7 BLEU for Somali→English and vice versa, showcasing the potential of weakly-supervised NMT for low-resource languages. We also observe that adding more time and increasing batch size in training can further improve performance, which underscores the importance of reporting compute resource usage in MT research. Through insights we obtain in our experiments, we provide recommendation for training effective NMT under low data- and compute-setting[1].

[1]Our training scripts are in the supplementary materials

Approved for Public Release; Distribution Unlimited.

Despite their unprecedented success, even the largest language models make mistakes. Similar to how humans learn and improve using feedback, previous work proposed providing language models with *natural language feedback* to guide them in repairing their outputs. Because human-generated critiques are expensive to obtain, researchers have devised learned critique generators in lieu of human critics while assuming one can train downstream models to utilize generated feedback. However, this approach does not apply to black-box or limited access models such as ChatGPT, as they cannot be fine-tuned. Moreover, in the era of large general-purpose language agents, fine-tuning is neither computationally nor spatially efficient. In this work, we introduce RL4F (Reinforcement Learning *for* Feedback), a multi-agent collaborative framework where the critique generator is trained to maximize end-task performance of GPT-3, a fixed model 200 times its size. RL4F produces critiques that help GPT-3 revise its outputs. We study three datasets for action planning, summarization and alphabetization and show improvements (~5% on average) in multiple text similarity metrics over strong baselines across all three tasks.

Data-Efficient Fine-Tuning for Image Classification. For a task (*e.g.*, satellite image classification) with limited training data, the most common approach is to borrow a pretrained large model and finetune it. Current methods for personalizing such large models (*e.g.*, fully finetuning, LoRA, Adapter) usually modifies the weights of feedforward network. Our insight is, once pretrained on large datasets, the feedforward backbone already contains all the information one may need, and instead of tweaking the feedforward backbone which may lose information learned during pretraining, we only need a way to "probe" the information required for downstread tasks while leaving the feedforward backbone untouched. In this work, we propose **To**p-Down **T**uning (**ToT**) which adds a top-down path to the pretrained network, and only finetunes the top-down path while keeping the feedforward backbone frozen when transfering to downstread tasks. We show superior performance compared to other tuning methods such as fully finetuning, VPT, LoRA, etc.

Typical deep visual recognition models are capable of performing the one task they were trained on. In this paper, we tackle the extremely difficult problem of combining completely distinct models each solving a separate task into one multi-task model **without any additional training**. Prior work in model merging permutes one model to the space of the other then adds them together. While this works for models trained on the same task, we find that this fails to account for the differences in models trained on disjoint tasks. Thus, we introduce "ZipIt!", a general method for merging two arbitrary models of the same architecture that incorporates two simple strategies. First, in order to account for features that aren't shared between models, we expand the model merging problem to additionally allow for merging features *within* each model by defining a general "zip" operation. Second, we add support for *partially zipping* the models up until a specified layer, naturally creating a multi-head model. We find that these two changes combined account for a staggering 20-60% improvement over prior work, making the merging of models trained on disjoint tasks feasible. We summarize relevant related work to our effort below for each major area of effort.

Zero-Shot Image Classification: DualCoOp: Fast Adaptation to Multi-Label Recognition with Limited Annotations [9]: `DualCoOp` is an algorithm for multi-label image recognition under the low-label regime. This reduces the effort of data annotation and enables fast adaptation to novel data domains.

SENTRY: Selective Entropy Optimization via Committee Consistency for Unsupervised Domain Adaptation [10]: `SENTRY` is an algorithm for adapting a model even without any new annotations in a target domain. Advances in unsupervised adaptation can enable learning from fewer labeled examples overall by only requiring labels in the initial source domain.

Adapting Self-Supervised Vision Transformers by Probing Attention-Conditioned Masking Consistency [11]: `PACMAC` is an algorithm for adapting a visual transformer based model even without any new annotations in a target domain. Advances in unsupervised adaptation can enable learning from fewer labeled examples overall by only requiring labels in the initial source domain.

Few-Shot Image Classification: Active Domain Adaptation via Clustering Uncertainty-weighted Embeddings [12]: `CLUE` is an algorithm for selecting the most informative samples in a target domain to label. This directly reduces the number of labeled samples needed by actively selecting the subset that can enable fastest performance gains. Semi-Supervised Domain Adaptation via Minimax Entropy [13]: MME helps to improve performance for semi-supervised domain adaptation settings where a few target labels are available. OpenMatch: Open-set Consistency Regularization for Semi-supervised Learning with Outliers [14]: OpenMatch is a method for open-set semi-supervised learning, where

samples of novel categories are present in unlabeled data. This helps to reduce the effort for data annotation and improve models' robustness to outliers in training data.

Subspace Regularization for Few-Shot Class Incremental Learning [15]: The methods proposed in this work enables augmenting the set of objects that a learned classifier can correctly classify by using only a handful of new images and without training from scratch.

A ConvNet for the 2020s [16]: ConvNeXt is a visual recognition architecture that could be used to learn the representations needed for few-shot classifications and other downstream few-shot computer vision tasks. It was also later shown in follow-up research to be competitive in zero-shot image classification on CLIP-style image-text paired training [17].

Few-Shot Object Detection: Cut and Learn for Unsupervised Object Detection and Instance Segmentation [18]: **CutLER** is a versatile algorithm for unsupervised object detection and instance segmentation that can be applied to a wide range of domains, including natural images, sketches, traffic images, paintings, and more. Furthermore, the model trained on ImageNet can serve as a robust pretrained model for both label-efficient and fully-supervised object detection and instance segmentation tasks. VideoCutLER: Surprisingly Simple Unsupervised Video Instance Segmentation [19]: Our approach, VideoCutLER, does not require labels, and does not rely on motion-based learning signals like optical flow. In fact, VideoCutLER does not need real videos for training as we synthesize videos using natural images from the ImageNet-1K dataset. Despite being simpler, VideoCutLER outperforms models that use additional learning signals or video data, achieving 10× their performance on benchmarks like YouTubeVIS. Additionally, VideoCutLER serves as a strong self-supervised pretrained model for supervised video instance segmentation. We hope that our approach enables both a wide range of applications in video recognition, as well as its simplicity enables easy future research.

Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models [20]: `ODISE` is an algorithm for open-vocabulary recognition, object detection, and segmentation, that can directly generalize to unseen categories in a zero-shot manner. This can serve as a pre-trained model for downstream few-shot recognition tasks.

Few-Shot Video Classification: Something-Else: Compositional Action Recognition with Spatial-Temporal Interaction Networks [21]: `Something-Else` provides a new model and new dataset for few-shot video classification. Contrastive Learning of Image Representations with Cross-Video Cycle-Consistency [22]: This approach provides a new self-supervised contrastive learning approach for better video representations for video classification. This can provide pre-trained representations for few-shot video classification.

Semi-Supervised Action Recognition with Temporal Contrastive Learning [23]: `TCL` is a method for the semi-supervised video recognition task. This helps to reduce the effort for data annotation and exploit unlabeled videos to improve generalization and robustness

Weakly-Supervised Action Localization with Expectation-Maximization Multi-Instance Learning [24]: `EM-MIL` is a method for weakly-supervised action localization, using only video-level action labels to learn activity recognition and localization. These weakly-supervised models can significantly reduce the annotation needed, and have the additional benefit that existing search engines could be leveraged to generate annotations.

Few-Shot Machine Translation:

Low-Resource Machine Translation Training Curriculum Fit for Low-Resource Languages [25]: We dissect the components that go into machine translation training based on their impact on the translation performance under a truly low-resource setting.

Generating Natural Language Feedback with Reinforcement Learning for Repairing Model Outputs [26]: We propose a feedback mechanism to automatically improve model outputs without requiring human feedback. The approach is generic to apply generation tasks beyond those studied in the paper.

Data-Efficient Fine-Tuning for Image Classification:

ZipIt! Merging Models from Different Tasks *without Training*: ZipIt! designs an algorithm for merging models together without retraining. It allows on the fly creation of new models with expanded knowledge. While each model may be trained with labeled data you do not need access to all data sources at once so data can be accumulated and learned from efficiently at disjoint sites and then merged into a single model at a later point without requiring more data to be collected.

3 Methods, Assumptions, and Procedures

3.1 Zero-Shot Image Classification

3.1.1 Method: DualCoOp: Fast Adaptation to Multi-Label Recognition with Limited Annotations [9]

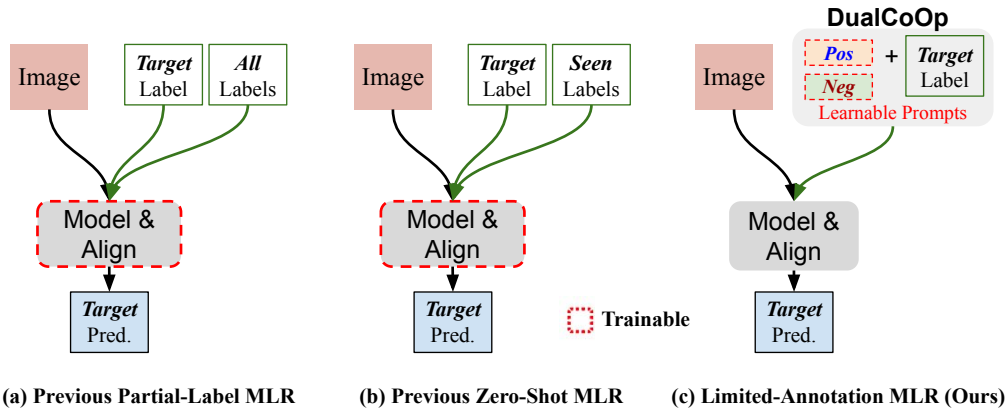


Figure 1: **A conceptual comparison of previous multi-label recognition (MLR) methods and our approach**. In Partial-Label MLR (a) and Zero-Shot MLR (b), previous works learn to model and align the visual and textual inputs as well as explore the correlation between the target label with all/seen labels depending on the limited semantic annotations available on the dataset, which leads to sub-optimal performance and complex model architectures. In contrast, we propose a unified framework to tackle both limited-annotation tasks (c). We rely on the modeling and alignment of visual and textual inputs contained a large-scale pretrained vision-language model, and only learn a pair of positive and negative prompts to this model.

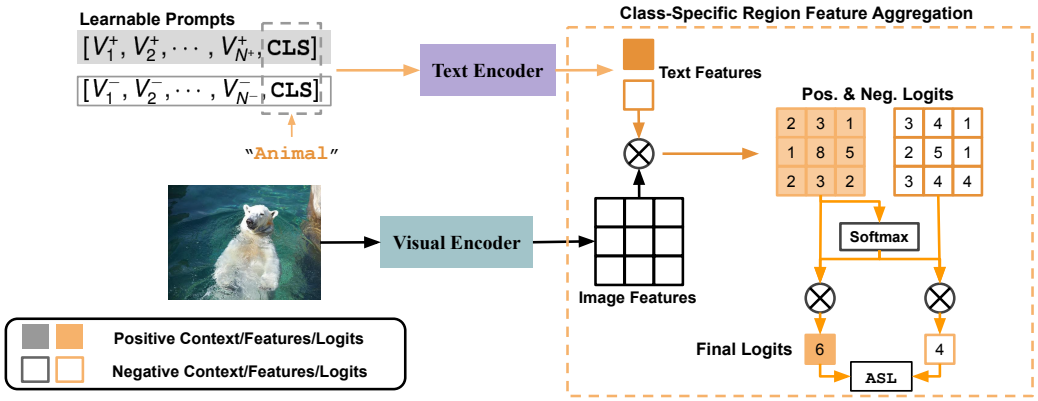


Figure 2: `DualCoOp` learns a pair of positive and negative prompts to quickly adapt powerful pretrained vision-text encoders to the MLR task. For each class, two prompts generate two contrastive (positive and negative) textual embeddings as the input to the text encoder. Furthermore, we propose *Class-Specific Region Feature Aggregation* to first project each region's feature to the textual space and then aggregate the spatial logits by the magnitude of class-specific semantic responses. During inference, we compare the positive and negative logits to make a prediction for each class.

Prompt learning provides a convenient way to transfer pretrained vision-language models to other tasks. It designs additional templated or learnable prompt tokens for textual input to "inform" the model about downstream tasks and avoids finetuning the entire model which can be inefficient and data-hungry. By doing so, recent works like CoOp [27] have demonstrated CLIP's remarkable generalisation to various zero-shot image tasks [28, 29]. However, these methods mainly focus on matching each image with a single label, hence they are not able to handle the multi-label setting. To adapt the knowledge learned in CLIP to multi-label image recognition, we propose the DualCoOp framework. As shown

in Fig. 1 (c) and Fig. 2, DualCoOp learns a pair of differentiable prompts to provide positive and negative contexts for the target class. Instead of using hand-crafted thresholding to determine positive labels [30], the dual prompts naturally result in a positive and a negative classifier, so the existence of the target class in the image can be easily decided by comparing their scores. Unlike prior models, shown in Fig. 1 (a)(b), we avoid fine-tuning the full vision-language model and only learn the prompts, which are much smaller compared to the entire model. Therefore, our simple framework achieves much higher efficiency when adapting to different datasets. Additionally, we modify the attention mechanism of CLIP to better model spatial information in images, improving its ability to recognize multiple objects in MLR. With these design choices, we achieve a unified framework for addressing the general challenges of multi-label recognition with limited annotations.

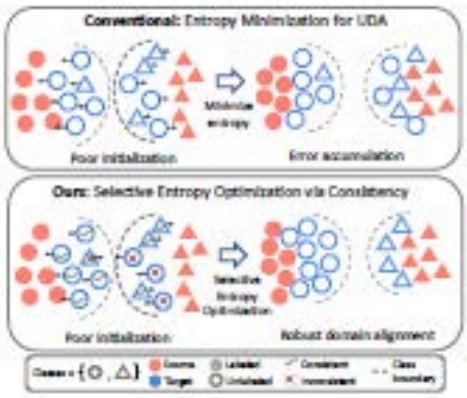


Figure 3: Comparisons to previous works on unsupervised domain adaptation. **Top**: Conventional entropy-minimization based approaches for unsupervised domain adaptation (UDA) operate by increasing model confidence on unlabeled target instances. Under strong distribution shifts, some instances may initially be strongly misaligned and entropy-minimization can lead to error accumulation. **Bottom**: We propose Selective Entropy Optimization via Committee Consistency (SENTRY), a UDA algorithm that first uses consistency under a set of random image transformations to select target instances for alignment, and then optimizes conditional entropy on selected instances to induce alignment.

3.1.2 Method: SENTRY: Selective Entropy Optimization via Committee Consistency for Unsupervised Domain Adaptation [10]

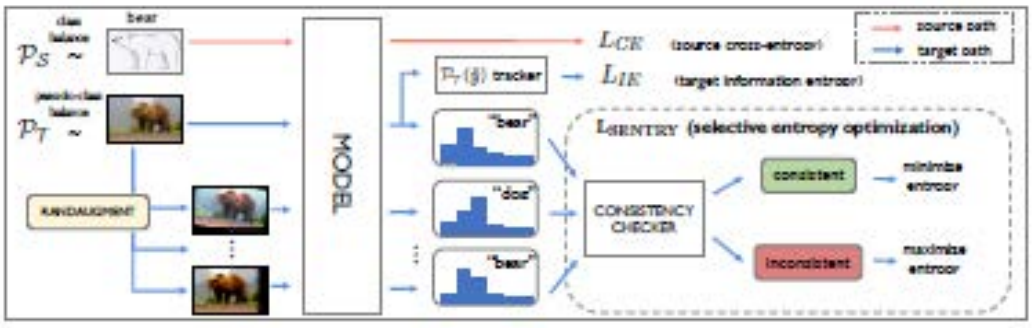


Figure 4: We propose Selective Entropy Optimization via Committee Consistency (SENTRY) for unsupervised DA. For each target instance, we generate a committee of random, label-preserving image transformations. A consistency checker then computes the consistency between model predictions for the original and augmented versions. The algorithm then minimizes predictive entropy (increasing model confidence) on highly consistent target instances, and maximizes predictive entropy (reducing model confidence) on highly inconsistent ones.

In this work, we propose SENTRY, a *selective* entropy optimization algorithm for UDA that addresses

these challenges. First, rather than using model confidence to select unlabeled target instances for self-training, we propose using the *consistency* of model predictions over a committee of random, label-preserving image transformations. Such consistency checks have been found to be a reliable way to detect model errors [31]. Second, to overcome the challenge of error accumulation under a poor initialization, we propose using our selection strategy to optimize a novel *selective* entropy optimization objective. Specifically, we consider a highly consistent target instance as likely correctly aligned, and increase model confidence by minimizing predictive entropy for one of its transformed consistent versions. Similarly, we consider an instance with high predictive inconsistency over transformations as likely misaligned, and reduce model confidence by maximizing predictive entropy for one of its transformed inconsistent versions. See Figure 4 (bottom).

3.1.3 Method: Adapting Self-Supervised Vision Transformers by Probing Attention-Conditioned Masking Consistency [11]

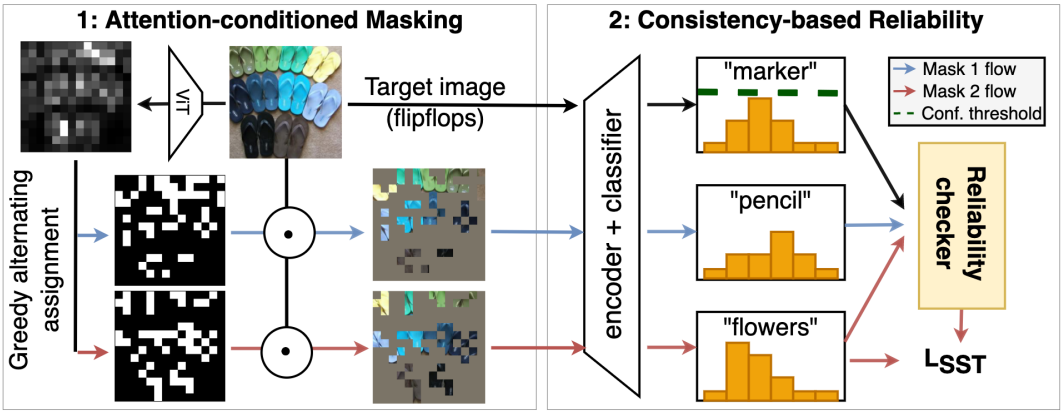


Figure 5: Overview of PACMAC. **Left.** First, the model's attention over a given target image is used to generate a set of disjoint masks that only keep highly attended patches of the input image via a greedy assignment strategy. **Right.** Next, the model's predictive consistency across the original and masked images is used as a self-supervised reliability measure to select target instances for self-training.

we observe that several recent state-of-the-art SSL methods for ViTs focus on learning from partial inputs [32, 33, 3, 34, 4] generated via masking or cropping strategies. In the context of vision transformers, such methods learn to either reconstruct the missing information [3, 35] or to be invariant to such cropping [4]. We find that the *predictive consistency* of a model initialized with such representations across partial inputs generated via masking acts as an effective self-supervised reliability measure on the unlabeled target domain. Further, rather than masking inputs randomly, we leverage the ViT attention mechanism to perform *attention-conditioned masking*: we generate a set of disjoint masks that only keep highly attended patches from the input image. We then probe the reliability of a given target instance by measuring the model's predictive consistency across such masked images and the original image, and mark target instances with high consistency as reliable. We call our selection strategy Probing Attention-Conditioned Masking Consistency (PACMAC).

We further mark instances with high confidence as reliable, and then selectively self-train the model against the current prediction on target instances identified as reliable via either scheme. This leads to an easy to implement selective self-training adaptation algorithm that outperforms competing methods on standard benchmarks.

3.2 Few-Shot Image Classification

3.2.1 Method: Active Domain Adaptation via Clustering Uncertainty-weighted Embeddings [12]

We propose a novel label acquisition strategy for Active DA that combines uncertainty and diversity sampling in a principled manner without the need for complex gradient or domain discriminator-based

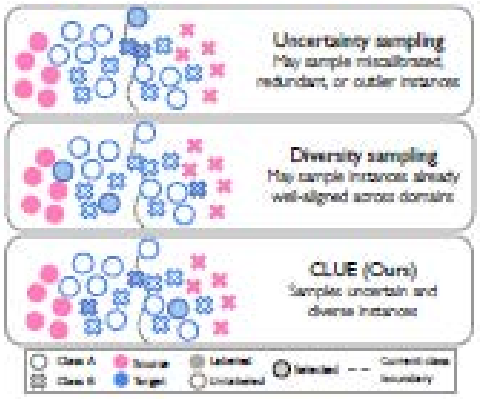


Figure 6: Active Domain Adaptation. The goal of **Active Domain Adaptation** [36] (Active DA) is to adapt a source model to an unlabeled target domain by acquiring labels for selected target instances via an oracle. Existing active learning (AL) methods based solely on uncertainty [37, 38, 39] or diversity-sampling [40, 41] are less effective for Active DA (Row 1, 2). We propose `CLUE`, an AL method designed for Active DA that selects instances that are both uncertain (thus informative to the model) and diverse in feature space (thus minimizing redundancy, Row 3), and leads to more cost-effective adaptation than competing AL and Active DA methods.

diversity measures. Our approach, Clustering Uncertainty-weighted Embeddings (`CLUE`), identifies informative and representative target instances from dense regions of the feature space. To do so, `CLUE` clusters deep embeddings of target instances *weighted* by the corresponding uncertainty of the target model. Our weighting scheme effectively increases the density of instances proportional to their uncertainty. To construct non-redundant batches, `CLUE` then selects nearest neighbors to the inferred cluster centroids for labeling. Our algorithm then leverages the acquired target labels and, optionally, the labeled source and unlabeled target data, to update the model, consistently leading to more cost-effective domain alignment than competing (and frequently more complex) alternatives.

3.2.2 Method: Semi-Supervised Domain Adaptation via Minimax Entropy [13]

In this work, We propose a novel approach for semi-supervised domain adaptation called Minimax Entropy (MME). We exploit a cosine similarity-based classifier architecture. The classifier (top layer) predicts a K-way class probability vector by computing cosine similarity between K class-specific weight vectors and the output of a feature extractor (lower layers), followed by a softmax. Each class weight vector is an estimated "prototype" that can be regarded as a representative point of that class. While this approach outperformed more advanced methods in few-shot learning and we confirmed its effectiveness in our setting, as we show below it is still quite limited. In particular, it does not leverage unlabeled data in the target domain. Our key idea is to minimize the distance between the class prototypes and neighboring unlabeled target samples, thereby extracting discriminative features. The problem is how to estimate domain-invariant prototypes without many labeled target examples. The prototypes are dominated by the source domain, as shown in the leftmost side of Fig. 8 (bottom), as the vast majority of labeled examples come from the source. To estimate domain-invariant prototypes, we move weight vectors toward the target feature distribution. Entropy on target examples represents the similarity between the estimated prototypes and target features. A uniform output distribution with high entropy indicates that the examples are similar to all prototype weight vectors. Therefore, we move the weight vectors towards target by maximizing the entropy of unlabeled target examples in the first adversarial step. Second, we update the feature extractor to minimize the entropy of the unlabeled examples, to make them better clustered around the prototypes. This process is formulated as a minimax game between the weight vectors and the feature extractor and applied over the unlabeled target examples.

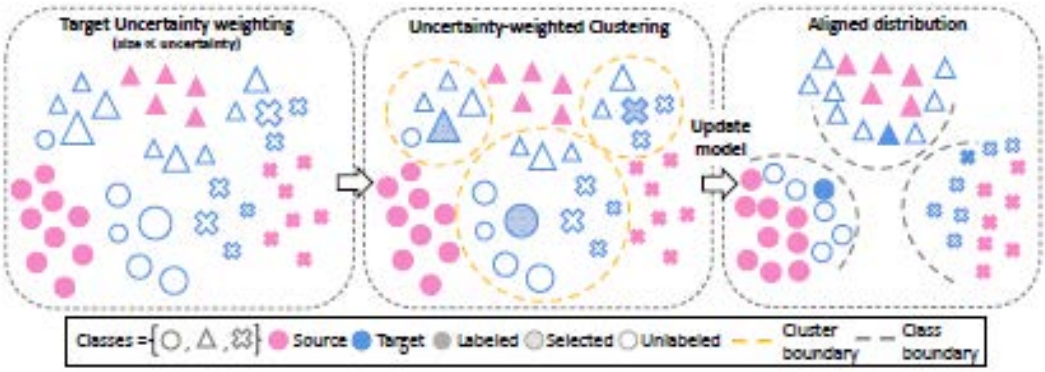


Figure 7: We propose CLustering Uncertainty-weighted Embeddings (CLUE), a novel label acquisition strategy for Active DA. It identifies a diverse set of target instances that are informative and representative. First, deep embeddings of target instances are reweighted based on model entropy to emphasize uncertain regions of feature space (*left*). Next, to select diverse instances, these uncertainty-weighted embeddings are clustered, and the instance closest to each cluster centroid is acquired for labeling (*middle*). Finally, the acquired target labels (and optionally, the labeled source and unlabeled target data) are used to update the model, leading to well-classified target data (*right*).

3.2.3 Method: OpenMatch: Open-set Consistency Regularization for Semi-supervised Learning with Outliers [14]

In this work, we propose a new framework, *OpenMatch*, to address the above drawbacks of OSSL. First, we propose to utilize a One-Vs-All (OVA) network [42] that can learn a threshold to distinguish outliers from inliers. A separate OVA-classifier is trained for each class, and a sample is labeled an outlier if all of the classifiers determine it to be one. Thus, this technique allows us to identify outliers in an unsupervised way. We call this an *outlier detector*. Second, we propose a novel *open-set soft-consistency loss* to learn more effective representations for detecting outliers. We first transform an unlabeled input in two ways and obtain two logits from the outlier detector. Then, we minimize the distance between the two logits to encourage consistency (See Fig. 9 and 10). The main difference between SSL and OSSL is that unlabeled outliers do not have any neighboring labeled samples, which makes it risky to perform hard-labeling such as pseudo-labeling. The outlier detector outputs a distance from inliers given an input, and enhancing the smoothness of this function allows us to improve its ability to find outliers. Empirically, this objective provides significant improvements in detecting outliers. Finally, to correctly classify inliers, we propose to apply FixMatch [43] to unlabeled samples that are considered to be inliers by the outlier detector.

3.2.4 Method: Subspace Regularization for Few-Shot Class Incremental Learning [15]

Standard approaches to classification in machine learning assume a fixed training dataset and a fixed set of class labels. But for many real-world classification problems, these assumptions are unrealistic. Classifiers must sometimes be updated on-the-fly to recognize new concepts (e.g. new skills in personal assistants or new road signs in self-driving vehicles), while training data is sometimes unavailable for reuse (e.g. due to privacy regulations, [44, 45]; or storage and retraining costs, [46]). Development of models that support **few-shot class-incremental learning** (FSCIL), in which classifiers' label sets can be easily extended with small numbers of new examples and no retraining, is a key challenge for machine learning systems deployed in the real world [47].

As a concrete example, consider the classification problem depicted in Figure 11. A model, initially trained on a large set of examples from several **base classes** (*snorkel*, *arctic fox*, *meerkat*; Figure 11a), must subsequently be updated to additionally recognize two **novel classes** (*white wolf* and *poncho*; Figure 11b), and ultimately distinguish among all five classes (Figure 11c). Training a model to recognize the base classes is straightforward: for example, we can jointly optimize the parameters of a feature extractor (perhaps a convolutional network parameterized by θ) and a linear classification

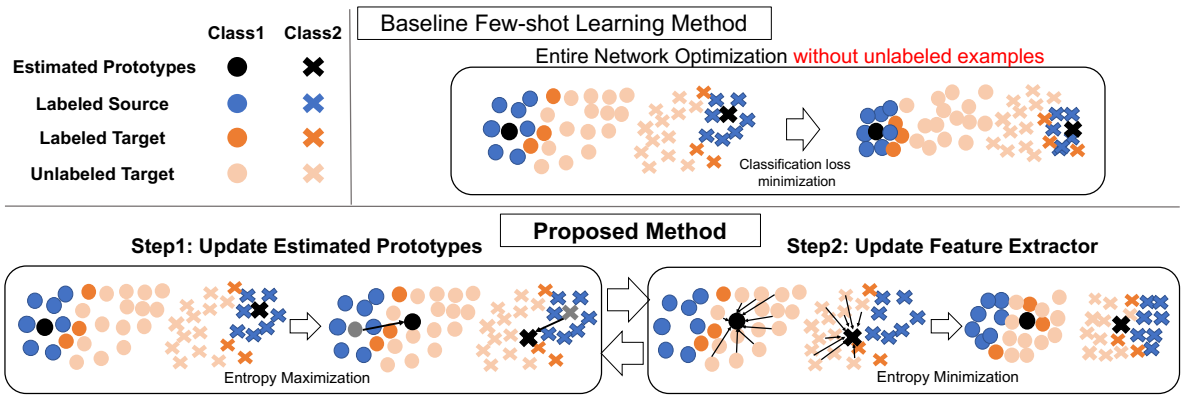


Figure 8: Pipeline of the proposed entropy minimization method. Top: baseline few-shot learning method, which estimates class prototypes by weight vectors, yet does not consider unlabeled data. Bottom: our model extracts discriminative and domain-invariant features using unlabeled data through a domain-invariant prototype estimation. Step 1: we update the estimated prototypes in the classifier to maximize the entropy on the unlabeled target domain. Step 2: we minimize the entropy with respect to the feature extractor to cluster features around the estimated prototype.

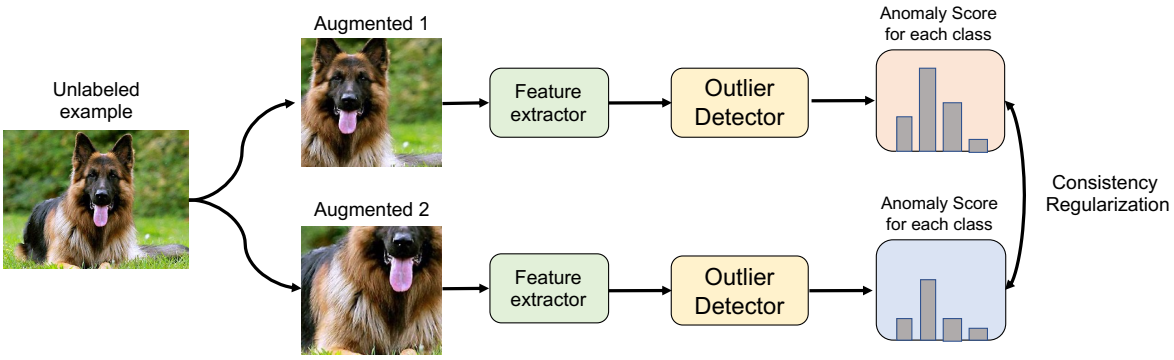


Figure 9: An illustration of our proposed open-set soft-consistency loss used to enhance outlier detection. Two differently augmented inputs are fed into the network to obtain the predictions of the outlier detector. The detector consists of one-vs-all classifiers and is able to detect outliers in an unsupervised way. The consistency loss is computed in a soft manner, *i.e.*, without sharpening logits.

layer (η) to maximize the regularized likelihood of (image, label) pairs from the dataset in Figure 11a:

$$\mathcal{L}(\theta, \eta) = \frac{1}{n} \sum_{(x,y)} \log \frac{\exp(\eta_y^\top f_\theta(x))}{\sum_{y'} \exp(\eta_{y'}^\top f_\theta(x))} + \alpha \left(\|\eta\|^2 + \|\theta\|^2 \right) \tag{1}$$

But how can this model be *updated* to additionally recognize the classes in Figure 11b, with only a few examples of each new class and no access to the original training data?

Naïvely continuing to optimize Equation 1 on (x, y) pairs drawn from the new dataset will cause several problems. In the absence of any positive examples of those classes, performance on *base classes* will suffer due to **catastrophic forgetting** [48], while performance on *novel classes* will likely be poor as a result of **overfitting** [49].

As a consequence, most past work on FSCIL has focused on alternative approaches that use non-standard prediction architectures e.g. [50] or optimize non-likelihood objectives e.g. [51, 52]. This divergence between approaches to standard and incremental classification has its own costs—state-of-the-art approaches to FSCIL are complicated, requiring nested optimizers, complex data structures, and numerous hyperparameters. When improved representation learning and optimization techniques are developed for standard classification problems, it is often unclear to how to apply these to the incremental setting.

Approved for Public Release; Distribution Unlimited.

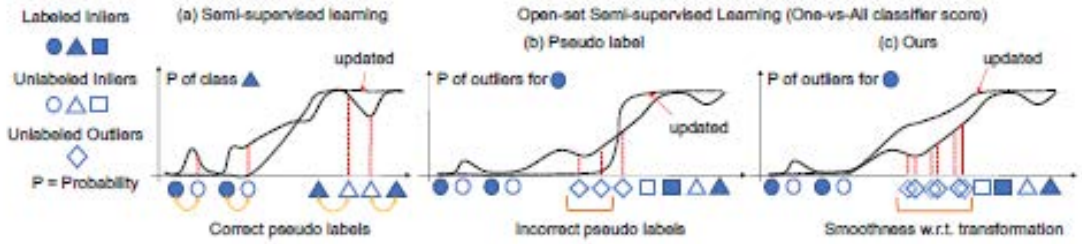


Figure 10: Comparisons to previous works on semi-supervised learning. **(a) SSL with pseudo-labels.** Hard labels such as pseudo-labels are often used in SSL to propagate training signals from labeled samples to neighboring unlabeled ones. **(b) OSSL with pseudo-labels.** However, in OSSL, the outlier unlabeled samples do not have any labeled neighbors. Therefore, the pseudo-labels can be highly unreliable. **(c) OSSL with soft consistency (ours).** To address this, we propose to ensure smoothness with respect to data augmentation by minimizing a soft consistency loss. This separates the outliers from inliers and avoids using incorrect pseudo-labels.

We turn the standard approach to classification into a surprisingly effective tool for FSCIL. Specifically, we show that both catastrophic forgetting and overfitting can be reduced by introducing an additional **subspace regularizer** (related to one studied by [53] and [54]) that encourages novel η to lie close to the subspace spanned by the base classes. On its own, the proposed subspace regularizer produces ordinary linear classifiers that achieve state-of-the-art results on FSCIL, improving over existing work in multiple tasks and datasets.

Because of its simplicity, this regularization approach can be easily configured to incorporate additional information about relationships between base and novel classes. Using language data as a source of background knowledge about classes, we describe a variation of our approach, which we term **semantic subspace regularization**, that pulls weight vectors toward particular convex combinations of base classes that capture their semantic similarity to existing classes. Semantic subspace regularization results in a comparable performance to subspace regularization on average; it better preserves performance on existing classes at the expense of higher accuracy on newer classes. These results suggest that FSCIL and related problems may not require specialized machinery to solve, and that simple regularization approaches can solve the problems that result from limited access to training data for both base and novel classes.

3.2.5 Method: A ConvNet for the 2020s [16]

In this work, we provide a trajectory going from a ResNet to a ConvNet that bears a resemblance to Transformers. We consider two model sizes in terms of FLOPs, one is the ResNet-50 / Swin-T regime with FLOPs around 4.5×10^9 and the other being ResNet-200 / Swin-B regime which has FLOPs around 15.0×10^9. For simplicity, we will present the results with the ResNet-50 / Swin-T complexity models.

At a high level, our explorations are directed to investigate and follow different levels of designs from a Swin Transformer while maintaining the network's simplicity as a standard ConvNet. The roadmap of our exploration is as follows. Our starting point is a ResNet-50 model. We first train it with similar training techniques used to train vision Transformers and obtain much improved results compared to the original ResNet-50. This will be our baseline. We then study a series of design decisions which we summarized as 1) macro design, 2) ResNeXt, 3) inverted bottleneck, 4) large kernel size, and 5) various layer-wise micro designs. In Figure 12, we show the procedure and the results we are able to achieve with each step of the "network modernization". Since network complexity is closely correlated with the final performance, the FLOPs are roughly controlled over the course of the exploration, though at intermediate steps the FLOPs might be higher or lower than the reference models. All models are trained and evaluated on ImageNet-1K.

After the modifications, the end model can outperform the Swin Transformer for ImageNet-1K classification in this compute regime. It is also worth noting that all design choices discussed in this work are adapted from vision Transformers. Our ConvNeXt model has approximately the same FLOPs, #params., throughput, and memory use as the Swin Transformer, but does not require specialized

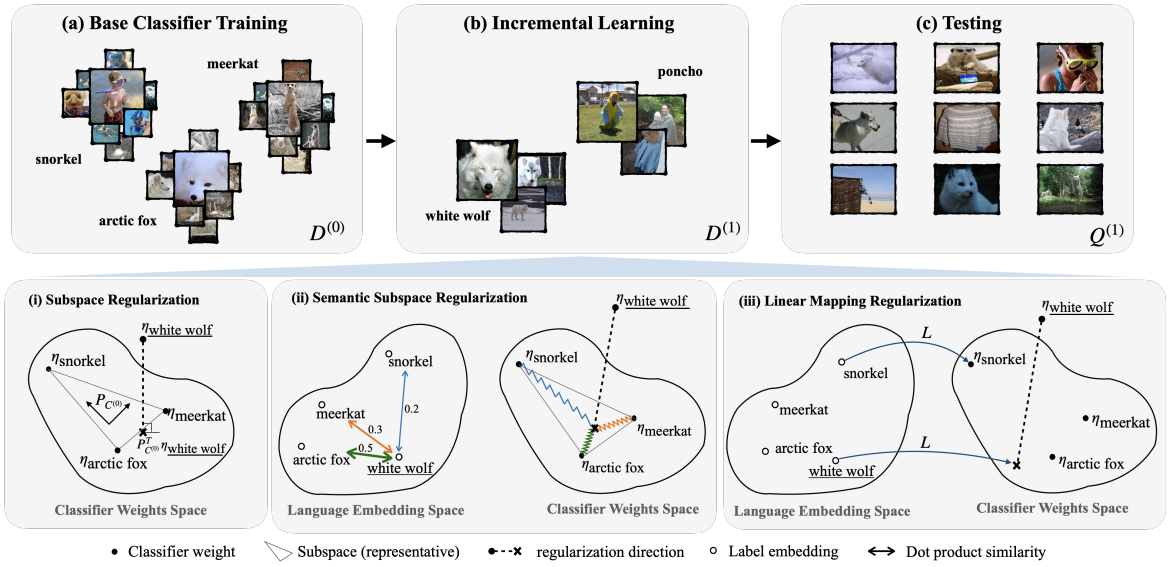


Figure 11: Few-shot class incremental learning. (a) A *base classifier* is trained on a large dataset ($D^{(0)}$). (b) This classifier is extended to also discriminate among a set of new classes with a small number of labeled examples ($D^{(1)}$). (c) Models are evaluated on a test set that includes all seen classes ($Q^{(1)}$). This paper focuses on extremely simple, regularization-based approaches to FSCIL, with and without side information from natural language: (i) We regularize novel classifier weights toward the shortest direction to the subspace spanned by base classifier weights. (ii) We regularize novel classifiers pulling them toward the weighted average of base classifiers where weights are calculated using label/description similarity between novel and base class names or one-sentence descriptions. (iii) We learn a linear mapping L between word labels and classifier weights of the base classes. Later, we project the novel label *white wolf* and *regularize* the novel classifier weight $\eta_{\text{white wolf}}$ towards the projection.

modules such as shifted window attention or relative position biases. Vision Transformers' scaling behavior is what truly distinguishes them. We scale up our ConvNeXt models both in terms of data and model size, and evaluate them on a diverse set of visual recognition tasks.

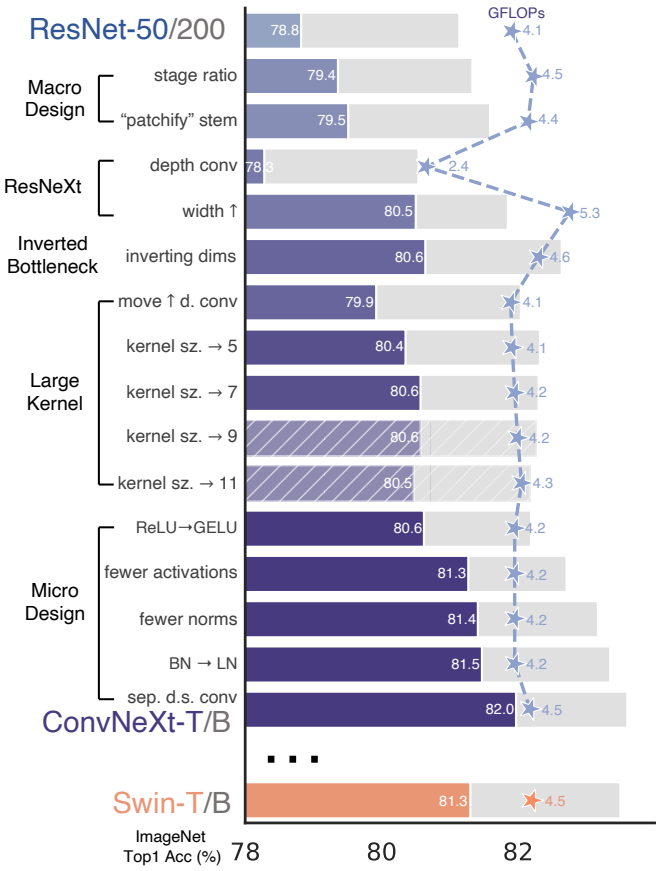


Figure 12: We modernize a standard ConvNet (ResNet) towards the design of a hierarchical vision Transformer (Swin), without introducing any attention-based modules. The foreground bars are model accuracies in the ResNet-50/Swin-T FLOP regime; results for the ResNet-200/Swin-B regime are shown with the gray bars. A hatched bar means the modification is not adopted. Many Transformer architectural choices can be incorporated in a ConvNet, and they lead to increasingly better performance. In the end, our pure ConvNet model, named ConvNeXt, can outperform the Swin Transformer.

3.3 Few-Shot Object Detection

3.3.1 Method: Cut and Learn for Unsupervised Object Detection and Instance Segmentation [18]

We tackle the problem of unsupervised object detection and segmentation with a simple cut-and-learn pipeline. The resulting model can be used as a strong pretrained model for label-efficient object detection and instance segmentation. Our method builds upon insights from recent work [4, 55], showing that self-supervised representations can discover objects. While these methods often find a single object per image, we propose a simple approach that can discover multiple objects and significantly improves segmentation and detection performance. *First*, we propose MaskCut that generates multiple binary masks per image using self-supervised features from DINO [4]. The overview of our MaskCut pipeline is illustrated in Fig. 13. *Second*, we show a dynamic loss dropping strategy, called DropLoss, that can learn a detector from MaskCut's initial masks while encouraging the model to explore objects missed by MaskCut; *Third*, we further improve the performance of our method through multiple rounds of self-training.

Approved for Public Release; Distribution Unlimited.

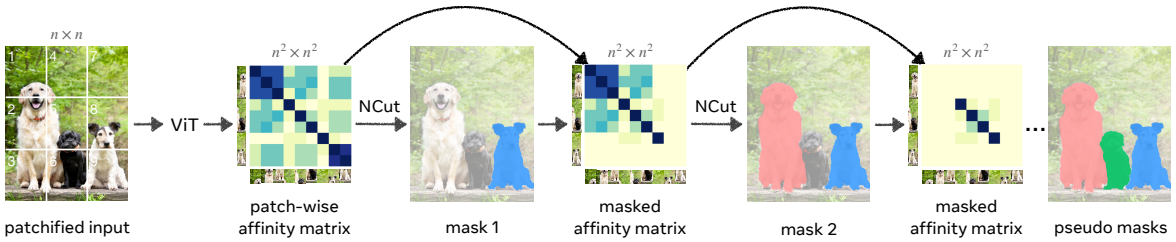


Figure 13: **MaskCut** can discover multiple object masks in an image without supervision. We build upon [4, 55] and create a patch-wise similarity matrix for the image using a self-supervised DINO [4] model's features. We apply Normalized Cuts [56] to this matrix and obtain a single foreground object mask of the image. We then mask out the affinity matrix values using the foreground mask and repeat the process, which allows MaskCut to discover multiple object masks in a single image. In this pipeline illustration, we set $n=3$.

3.3.2 Method: VideoCutLER: Surprisingly Simple Unsupervised Video Instance Segmentation [19]

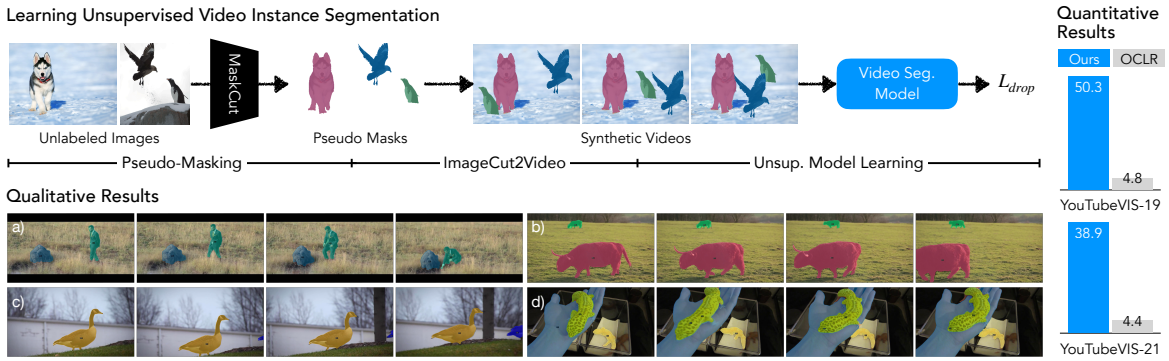


Figure 14: VideoCutLER is a simple unsupervised multi-instance segmentation method for video. We show the first competitive unsupervised results on the challenging YouTubeVIS benchmark. Moreover, unlike most prior approaches, ours does not train on natural videos and does not rely on motion estimates such as optical flow. ***Row 1:*** We propose **VideoCutLER**, a simple cut-synthesis-and-learn pipeline that involves three main steps. Firstly, we generate pseudo-masks for multiple objects in an image using MaskCut [18]. Then, we convert a random pair of images in the minibatch into a video with corresponding pseudo mask trajectories using ImageCut2Video. Finally, we train an unsupervised video instance segmentation model using these mask trajectories. ***Row 2:*** Despite being trained only on unlabeled images, at inference time VideoCutLER can be directly applied to unseen videos and can segment and track multiple instances across time, even for small objects, objects that are absent in specific frames, and instances with high overlap. ***Column 2:*** Our method surpasses the previous SOTA method OCLR [57] by a factor of 10 in terms of class-agnostic AP_{50}^{video}.

Our method, **VideoCutLER**, is an unsupervised **Video** instance segmentation model that employs a **Cut**-syn**t**hesis-and-**LEaR**n pipeline. First, given unlabeled images, we extract pseudo-masks for multiple objects in an image using MaskCut [18], leveraging a self-supervised DINO [4] and a spectral clustering method Normalized Cuts [56]. Second, given unlabeled images and their pseudo-masks in a minibatch, we propose Image2Video, a surprisingly simple video synthesis scheme that generates a video from those with corresponding pseudo mask trajectories. Finally, those mask trajectories are used to train a video instance segmentation model, aiming to perform object segmentation with temporal consistency across video frames. We utilize VideoMask2Former [58] as our video instance segmentation model, which operates by attending to the 3D spatiotemporal features of our synthetic videos and generating 3D volume predictions of pseudo-mask trajectories using shared queries across frames. The shared queries across frames enable the model to segment and track object instances based on their appearance (feature) similarities.

3.3.3 Method: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models [20]

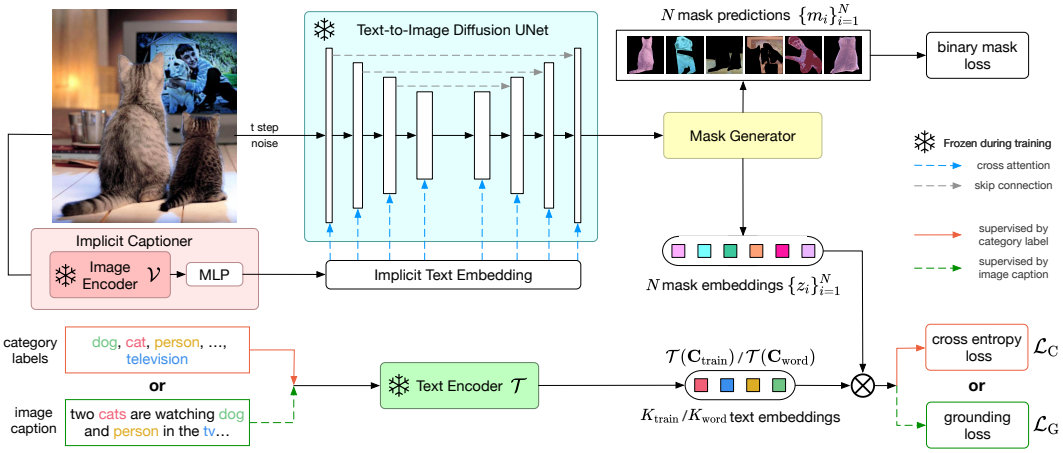


Figure 15: ODISE Overview and Training Pipeline. We first encode the input image into an implicit text embedding with an implicit captioner (image encoder $\mathcal{V}$ and MLP). With the image and its implicit text embedding as input, we extract their diffusion features from a frozen text-to-image diffusion UNet. With the UNet's features, a mask generator predicts class-agnostic binary masks and their associated mask embedding features. We perform a dot product between the mask embedding features and the text embeddings of training category names (red box) or the nouns of the image's caption (green box) to categorize them. The similarity matrix for mask classification is supervised by either a cross entropy loss with ground truth category labels (red solid path), or via a grounding loss with the paired image captions (green dash path).

The field of open-vocabulary recognition in computer vision has attracted attention for its potential to replicate human-like understanding of categories and spatial relations in images. Current approaches mainly use text-image discriminative models, but these models often struggle with spatial relations. On the other hand, text-to-image diffusion models have revolutionized image synthesis, suggesting a possible correlation between internal representations and semantic concepts. We propose Open-vocabulary DIffusion-based panoptic SEgmentation (ODISE), a model that leverages both text-image diffusion and discriminative models for state-of-the-art panoptic segmentation. ODISE inputs an image and caption into a pre-trained text-to-image diffusion model and extracts internal features for mask generation. A mask classification module then associates these features with text embeddings to categorize the masks. ODISE outperforms existing baselines across numerous open-vocabulary recognition tasks, establishing a new state of the art in the field.

3.4 Few-Shot Video Classification

3.4.1 Method: Something-Else: Compositional Action Recognition with Spatial-Temporal Interaction Networks [21]

We investigate the capability of video modeling to generalize actions across different combinations of verbs and nouns, focusing on actions represented by changes in geometric arrangements between agents and objects. We introduce the Something-Else task, which extends the Something-Something dataset and requires models to recognize actions performed with unseen objects. Current state-of-the-art action recognition models struggle in this setting, as they do not fully capture the compositionality of action and objects. To address this, we propose a model based on a sparse and semantically-rich object graph learned for each action and introduce the Spatial-Temporal Interaction Network (STIN), as shown in Figure 17, to reason on candidate sparse graphs found from the detection and tracking results.

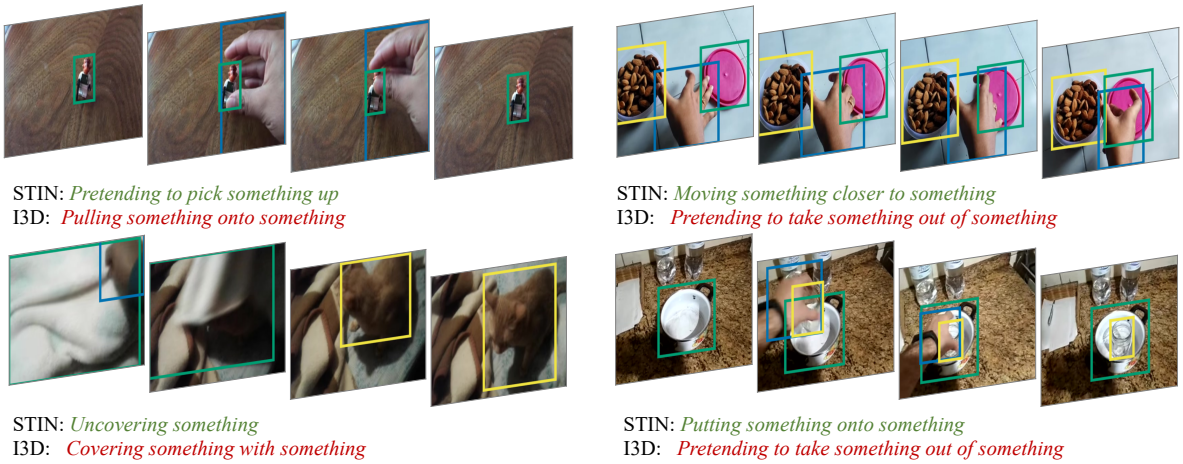


Figure 16: Predictions of STIN and I3D models. Correct predictions are in green, incorrect in red. STIN can keep tracking the relations between subjects and objects change over time in complicated actions.

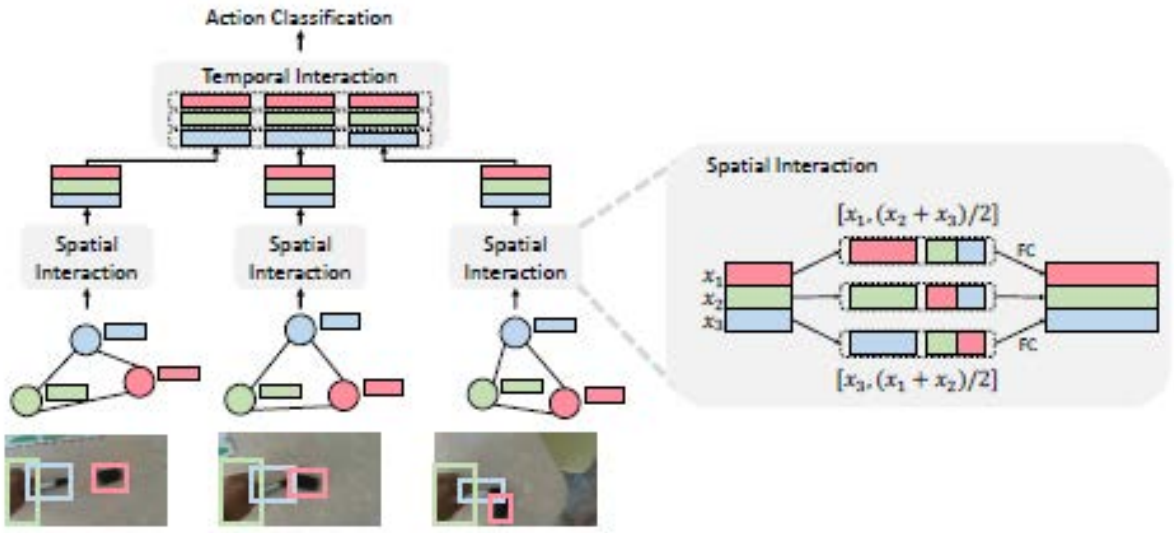


Figure 17: The Spatial-Temporal Interaction Network (STIN).

The STIN model first performs spatial interaction reasoning on the locations and shapes of objects and subjects in each frame, followed by temporal interaction reasoning over the boxes along the same tracklet. Finally, the trajectories for the agent and objects are composed together to understand the action. We introduce two new compositional tasks for testing model generalizability and dense object bounding box annotations in videos. We find that even with low dimensional coordinate inputs, the model shows comparable results and significantly improves appearance-based models in few-shot settings. This research presents a substantial performance gain over appearance-based models on compositional action recognition.

3.4.2 Method: Contrastive Learning of Image Representations with Cross-Video Cycle-Consistency [22]

Contrastive learning has recently gained popularity, outperforming supervised pre-training in various visual recognition tasks. Its success relies on constructing different views and transformations of the same instance, making deep representation invariant to view changes. However, artificial augmentation of single instances has shown limitations in multiple applications. Researchers have turned to videos as a source for obtaining positive pairs of training images, using temporal augmentation to learn viewpoint and deformation invariance. We propose performing contrastive learning with positive image pairs sampled across different videos, as shown in Figure 18, potentially capturing higher-level semantics

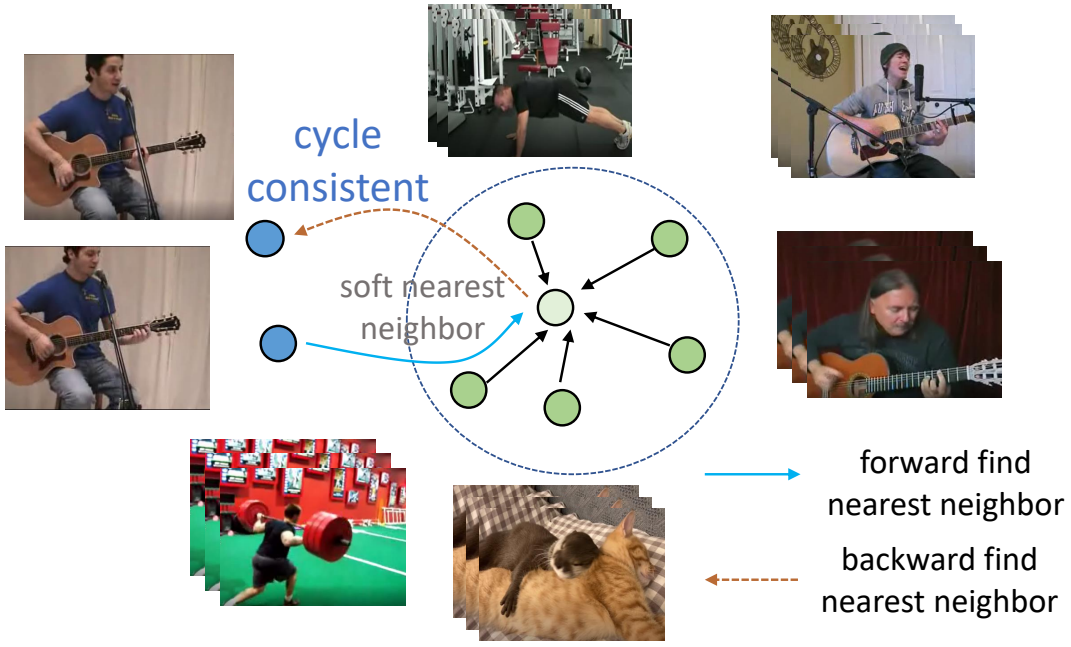


Figure 18: Cross-video cycle-consistency for image representation learning. Starting from one frame in a video, we find its soft nearest neighbor from other videos as a forward step, then the cycle-consistency is achieved when the soft nearest neighbor finds its closest frame within the same video as the start frame in a backward step.

and categorical information beyond low-level intra-instance invariances. We introduce Cycle-Consistent Contrastive Learning, where we explain one frame by composing frames from other videos that are similar to it and then compare the composed frames for contrastive learning.

We test our self-supervised representation learning on the unlabeled video dataset Random Related Video Views (R2V2) and transfer the learned representation to various downstream tasks such as visual object tracking, image classification, and action recognition. Our goal is to use the temporal signal to learn a general image-level representation for multiple applications beyond video-level recognition tasks. We report significant improvements over multiple state-of-the-art approaches and conduct extensive ablation studies of different components and design choices. Our contributions include a novel cross-video cycle-consistent contrastive learning objective that explores cross-video relations and enforces image representations from the same category closer without explicitly generating pseudo labels, achieving significant improvement in multiple downstream tasks.

3.4.3 Method: Semi-Supervised Action Recognition with Temporal Contrastive Learning [23]

Supervised action recognition has benefited a lot by using differently paced versions of the same video during training [59]. Motivated by the success of using slow and fast versions of videos for supervised action recognition as well as by the success of the contrastive learning frameworks [60], we propose *Temporal Contrastive Learning* (TCL) as illustrated in Fig. 19 for semi-supervised action recognition in videos where consistent features representing both slow and fast versions of the same videos are learned. Starting with a model trained with limited labeled data, we present a two-pathway model that processes unlabeled videos at two different speeds and finds their representations. Though played at two different speeds, the videos share the same semantics. Thus, similarity between these representations are maximized. Likewise, the similarity between the representations of different videos are minimized. We achieve this by minimizing a modified NT-Xent contrastive loss [32] between these videos with different playback rates. While minimizing a contrastive loss helps to produce better visual representations by learning to be invariant to different views of the data, it ignores information shared among samples of same action class as the loss treats each video individually. To this end, we propose a new perspective of contrastive loss between neighborhoods. Neighborhoods are compact groups of

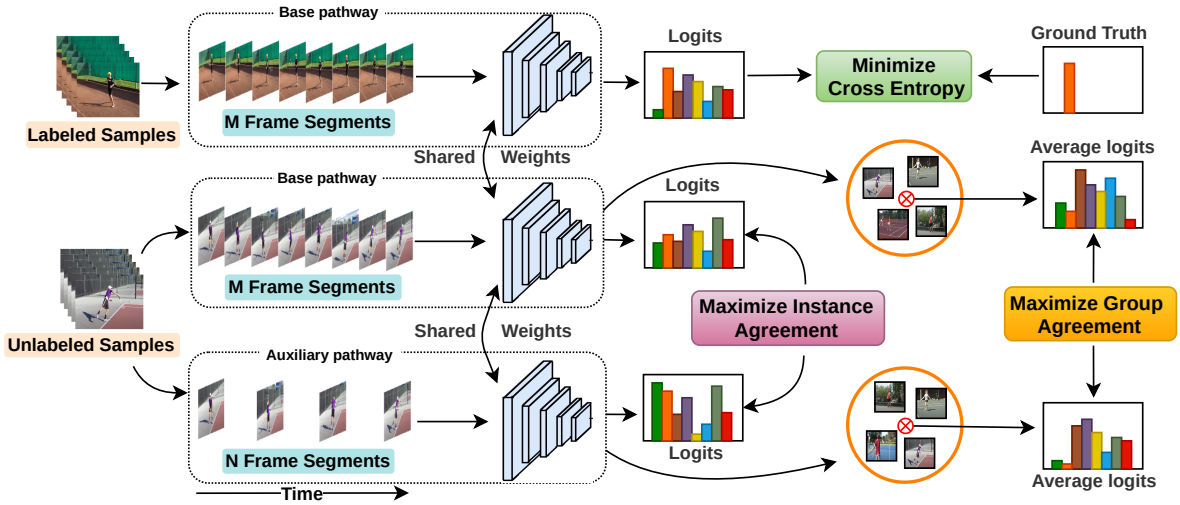


Figure 19: Illustration of our Temporal Contrastive Learning TCL Framework. Our approach consists of *base* and *auxiliary* pathways that share the same weights. Base pathway accepts video frames sampled at a higher rate while the auxiliary pathway takes in frames at a lower framerate. At first, the base network is trained using limited labeled data. Subsequently, the auxiliary pathway comes into picture for the unlabeled samples by encouraging video representations to match in both pathways in absence of labels. This is done by maximizing agreement between the outputs of the two pathways for a video while minimizing the same for different videos. In addition, originally unlabeled videos with high semantic similarity are grouped by pseudo-labels assigned to them. To exploit the high consistency and compactness of group members, the average representations of groups with same pseudo-label in different pathways are made similar while that between the varying groups are made maximally different. Two separate contrastive losses are used for this purpose. Given a video at test time, only the base network is used to recognize the action.

unlabeled videos with high class consistency. In absence of ground-truth labels, groups are formed by clustering videos with same pseudo-labels and are represented by averaging the representations of the constituent videos. Contrastive objective between groups formed off the two paths explores the underlying class concept that traditional NT-Xent loss among individual video instances does not take into account.

3.4.4 Method: Weakly-Supervised Action Localization with Expectation-Maximization Multi-Instance Learning [24]

In this work, we focus on weakly-supervised action localization paradigm, using only video-level action labels to learn activity recognition and localization. We formulate this task as a Multiple Instance Learning (MIL) problem with each video as a bag, and the clips of the video as the instances inside the bag. Most previous works [61] used attention-based approaches to model the key instance assignment process. They used attention weights to combine instance-level classification to produce the bag's classification. The learned attention weights imply the contribution of each instance to the bag's label, and thus can be used to localize the positive instances (action clips) [62, 63]. Models of this variety tend to produce incomplete action proposals as discussed in [64, 65], and implicitly violate the MIL assumptions. They use attention weights in negative bags, while the MIL assumption states that instances in negative bags are uniformly negative. We adopt a different method using the Expectation–Maximization framework to solve the interlocking steps of key instance assignment and action concept classification, explicitly model key instance assignment as a hidden variable and optimize for this as our target. To formulate our learning objective, we derive two pseudo-label generating schemes to model the E and M process respectively. We show that our alternating update process optimizes a lower bound of the MIL-objective. Our method can better model the data generating procedure of both positive and negative bags, and achieves better temporal localization performance. See Figure 20 and Figure 21.

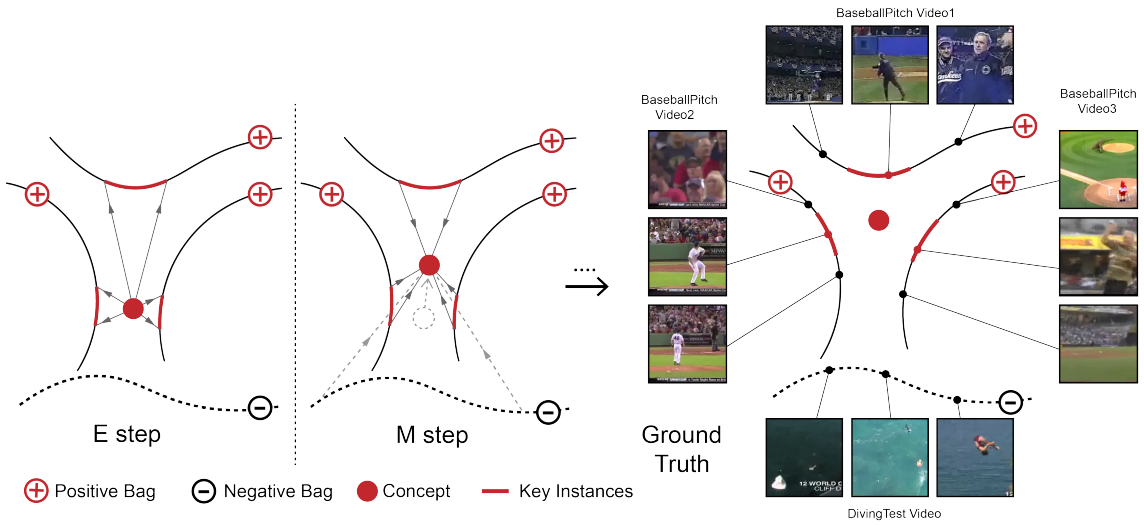


Figure 20: Each curve represents a bag and points on the curve represent instances in the bag. We aim to find a concept point such that each positive bag contains some key instances close to it while all instances in the negative bags are far from it. In E step we use the current concept to pick key instances for each positive bag. In M step we use key instances and negative bags to update the concept.

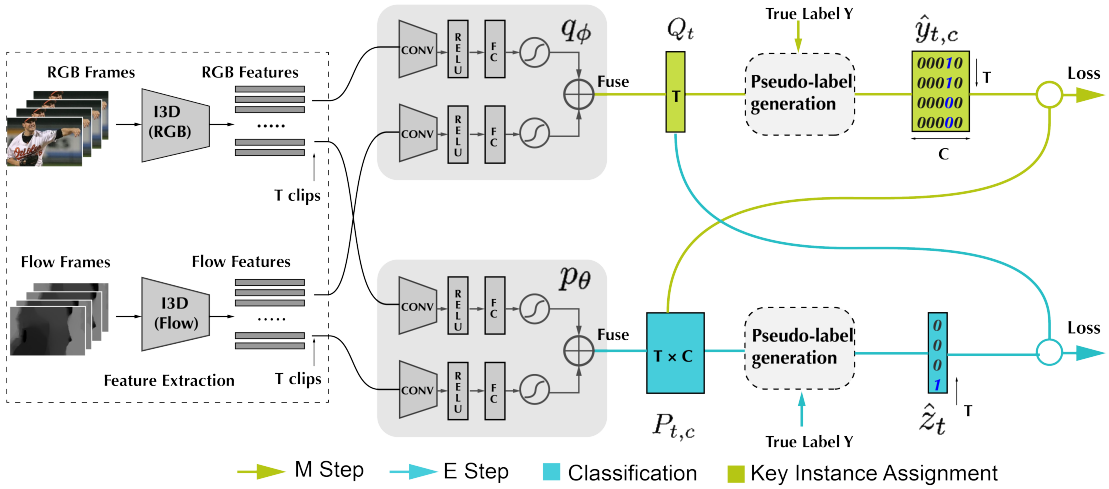


Figure 21: Our EM-MIL model architecture. It was built on fixed two-stream I3D features and alternates between updating the key-instance assignment branch q_ϕ (E Step) and the classification branch p_θ (M Step). We use the classification score and key instance assignment result to generate pseudo-labels for each other, and alternate freezing one branch to train the other.

3.5 Few-Shot Machine Translation

3.5.1 Method: Learned Incremental Representations for Parsing [66]

Our goal in this work is to design a representation for parsing that is maximally *speculation free*. In other words, it should record commitments to syntactic choices only as they are incrementally revealed by the input. We additionally want our representation to relate to constituency trees in a similar way to how transition-based actions relate to them: that is, through a deterministic transformation function.

A sequence of shift-reduce or attach-juxtapose actions is not identical to a parse tree, but it can be mapped to a tree using a deterministic automaton that interprets the discrete actions as operations on a tree fragment or stack of tree fragments. A sequence of supertags is likewise not the same as a tree, and mapping it to a tree requires further processing in the form of finding and applying an appropriate series of combinators. These mappings are non-trivial, especially in the case of CCG, so we should not expect our mapping to be trivial either — our only requirement is that it be deterministic and operate entirely from our representation, having already discarded the raw text in the sentence. Finally, we would like our representation to take the familiar form of a sequence of discrete symbols.

We propose to arrive at such a representation through *end-to-end learning*, rather than manual construction. The model can then make its own decisions about when syntactic decisions take place, how to handle cases of ambiguity, and how to represent belief states within the learned system itself. This system will learn to encode linguistic and structural features to allow effective incremental parsing. Our end-to-end approach is a model that proceeds in two stages. The first stage maps from individual words to syntactic decisions, which are represented as discrete tokens drawn from a small, bounded vocabulary. The second component of our system is a read-out network that takes a sequence of discrete tags as input and produces a conventional parse tree as output. Both stages are trained jointly in an end-to-end manner. Crucially, we do not a priori assign meaning to the discrete tokens (e.g. actions like *shift*, or supertags like in CCG); we only specify the total number of symbols available to the model to control the complexity of the representation. Unlike a speculative system, our representation can be used by finding the single highest-scoring tag at each position in a sentence, and then converting the resulting sequence of tags into a tree.

3.5.2 Method: Low-Resource Machine Translation Training Curriculum Fit for Low-Resource Languages [25]

In this work, our contribution is in providing a training recommendation for low-resource NMT. We show through extensive experiments that our training recommendation that combines (1) code-switching LM pre-training, followed by (2) training with the unsupervised NMT objective, and optionally (3) training on comparable data if the data is larger than the available parallel data is effective for truly low-resource MT and that together they outperform their individual scores.

Among the low-resource languages in MT for which there are test corpora to evaluate MT models on, we focus our attention to three: Gujarati (gu), Somali (so), the *scraping-by* languages—according to [67]—for which *some* monolingual data are available, and Kazakh (kk), the *rising-star* language, for which *more* monolingual data is available. These are distant languages that are diverse in terms of script, morphological complexity and word order from English (en) and for which unsupervised NMT has either never been explored (for so) or has been shown to perform poorly i.e., 0.6 and 0.8 BLEU for en→gu and en→kk [68].

For the source of comparable data in this work, we use Wikipedia, which has over 300 languages, both high and low-resource, allowing our approaches here to scale to other languages. We mine comparable sentences from Gujarati, Somali, and Kazakh Wikipedia articles that are linked to their corresponding English articles. We use a publicly available dictionary—Panlex—which is available in 5,700 languages) to mine sentences with lexical overlap from the linked Wikipedia articles. We also use the dictionary to translate monolingual data in the languages word-by-word when applicable in order to obtain code-switched sentences that we use to augment data for pre-training LMs. Although we have strived to use resources that are publicly available for many languages (Panlex, Wikipedia and news), we acknowledge as a limitation to our work that these resources may still not apply to many of the world's languages.

Further, unless otherwise indicated, we assume limited access to compute resource and restrict our compute expenses by the average monthly income for the language speaking region[2]. Our assumption of limited access to compute resource is realistic and driven by our personal experiences and other previous works that have observed how indeterminate access to a large number of GPUs is unrealistic beyond large industrial/academic environments [69]. By using limited compute resources, we also hope to shed light on the true state of low-resource NMT and the researchers working on it when both data *and* access to compute resource are lacking. We observe in our experiments that without

[2]For each language, we use the average monthly income estimated in https://www.numbeo.com/cost-of-living, and calculate how many hours of 1 GPU (32GB) training time we can get for 1 month of income using the AWS EC2 rate https://calculator.aws. This comes down to 40, 60, and 72 hours of training for gu, so, and kk respectively

this constraint, even a modest increase in batch size (from 2k to 4k tokens per batch) and training time (training till convergence using validation BLEU as a stopping criterion) can already significantly improve NMT performance. Since previous works have shown that increasing compute resources comes with a high environmental cost [70, 71], we echo their suggestions to factor in the compute resources when reporting results.

3.5.3 Method: RL4F: Generating Natural Language Feedback with Reinforcement Learning for Repairing Model Outputs [26]

Correcting model outputs is a pressing challenge in natural language generation [72], emerging across many use-cases such as style transfer [73], grammatical or factual error correction [74], debiasing and detoxification, to name a few. Unlike humans who can understand natural language feedback and improve using the information, most of the previous work relied on sequence tagging [72], retraining from scratch or parameter editing [75] to repair model predictions.

We focus on a multi-agent collaborative framework of language models where one model's job is to criticize another as the latter performs a task of interest, such as summarization. This setting comprises a task model, denoted by LM_{task}, which learns the mapping from an input x (e.g. passage) to a ground truth output y (e.g. summary); and a critiquing model $\text{LM}_{\text{critique}}$ which provides natural language critiques for LM_{task}'s outputs $\hat{y} \sim \text{LM}_{\text{task}}(x)$. The framework can additionally involve a separate model (say $\text{LM}_{\text{refine}}$) for repairing model outputs conditioned on critiques. We follow past work [76], and merge LM_{task} and $\text{LM}_{\text{refine}}$ into a single model. Hence, in addition to predicting y given x, LM_{task} is also tasked to improve its initial output conditioned on a critique $\hat{c}$ sampled from $\text{LM}_{\text{critique}}(x, \hat{y})$.

We introduce RL4F (Reinforcement Learning for Feedback Generation), a cascade [77] of two language models for automatic critique generation and refinement. RL4F is trained to maximize target task performance of LM_{task} by learning to provide critiques for its outputs via $\text{LM}_{\text{critique}}$. RL4F advances retrieval-based methods with learned critique generation [78]. Moreover, unlike previous work which teaches LM_{task} to read a crowd-sourced set of critiques [76, 8], RL4F learns the particular set of critiques that will steer LM_{task} into improving its predictions without requiring any updates to LM_{task} parameters. Treating LM_{task} as fixed is especially important in era of limited-access large language models which are costly, if not impossible, to fine-tune.

3.6 Data-Efficient Fine-Tuning for Image Classification

3.6.1 Method: A Top-Down Approach to Personalizing Large Models

For a pretrained model (*e.g.*, a ViT-B pretrained on ImageNet), we first add a top-down path on it (as described in our previous paper [79]), then only tune the top-down path when transfering to a downstream tasks.

3.6.2 Method: ZipIt! Merging Models from Different Tasks *without Training*

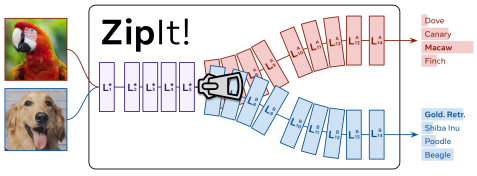


Figure 22: [Zip it model framework. **Top**: ZipIt! merges models trained on completely separate tasks *without any additional training* by identifying their shared features. Depending on the architecture and task, ZipIt! can nearly match their ensemble performance.

In this work, we treat model merging as jointly combining the checkpoints (i.e., collection of weights) of two models into a single checkpoint that can perform all the tasks of its constituents. We accomplish this by merging each layer of one model with the corresponding layer in the other, *while modifying both* (in contrast to permutation-based merging, which only permutes one of the models).

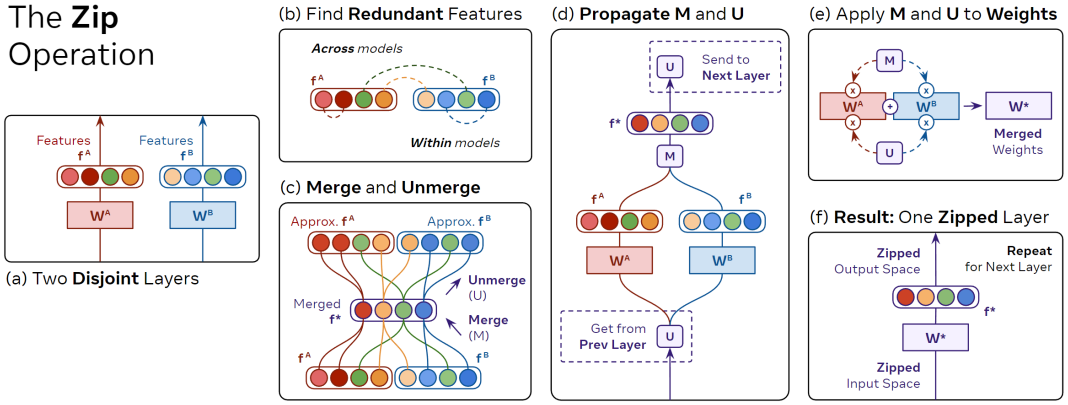


Figure 23: ZipIt! merges models layer-wise by exploiting redundancy in their features. (a) Starting with completely disjoint layers with weights W^A and W^B from models trained on different tasks, (b) we match redundant features by comparing their activations f^A and f^B. (c) We use this matching to produce a merge matrix M to combine f^A and f^B into a single shared feature space f^* and a corresponding unmerge matrix U that undoes this operation. (d) In order to align the input space of the next layer, we propagate U forward along network and at the same time receive a U matrix from the previous layer. (e) Once we have both an M for the output, and a U for the input, we can "zip" the layers together. (f) The result is a single layer with a shared input and output space, and we can now repeat from (a) on the next layer.

For instance, if layer $L_i \in \mathcal{L}$ is a linear layer (where $\mathcal{L}$ denotes a model), it has parameters $W_i \in \mathbb{R}^{n_i \times m_i}, b_i \in \mathbb{R}^{n_i}$ and takes input $x \in \mathbb{R}^{m_i}$, with an output feature vector $f_i \in \mathbb{R}^{n_i}$ where

$$f_i = L_i(x) = W_i x + b_i \tag{2}$$

Then our goal is to take $L_i^A \in \mathcal{L}^A$ from model A and $L_i^B \in \mathcal{L}^B$ from model B and merge them into a layer L_i^* that combines their feature spaces such that information from both f_i^A and f_i^B is retained in f_i^*. Note that we consider activation and normalization to be distinct layers in the network as they are implemented in practice, rather than one combined unit.

In order to construct the combined features f_i^*, we assume that there are some redundant features in f_i^A and f_i^B. That is, some elements of the two feature vectors are *highly correlated* over a sample of data. In this work, we consider correlations between features *within* the same model and *across* the two models. If two features are highly correlated, then we can average them without losing much information. Thus, if we can find a good pairing for elements of f_i^A and f_i^B, we can construct a merged feature f_i^* that contains an efficiently compressed representation of f_i^A and f_i^B by averaging each pair of features. In practice, we define a merge matrix M_i that maps highly correlated features from f_i^A and f_i^B together.

However, when computing matches in the following layer, we now have the problem that the next layers, L_{i+1}^A, L_{i+1}^B, are not compatible with this merged representation. Instead, we need to *undo* the merge operation before passing the features to the next layer. We denote this operation by an "unmerge" matrix U_i such that: $U_i f_i^* \approx (f_i^A, f_i^B)$. Using M and U we can merge arbitrary segments of distinct models together, either completely, or up to a point — where-after we have multiple heads for the last layers of f_i^A and f_i^B. We denote the latter as "partial zips,"

4 Results

4.1 Zero-Shot Image Classification

4.1.1 Experimental Results: DualCoOp: Fast Adaptation to Multi-Label Recognition with Limited Annotations [9]

Multi-Label Recognition with Partial Label. We conduct experiments on MS-COCO [1] and VOC2007 [80] to evaluate multi-label recognition with partial labels. Table 1 shows the comparison of mAP between DualCoOp and all baselines optimized with 10% to 90% of labels. For the two most recent works (SST [81] and SARB [82]), we further substitute the ImageNet pretrained weights [83] with the CLIP pretrained weights [29] when initializing of their visual encoders, which results in SST* and SARB* in Table 1. Since we learn class-specific prompts, DualCoOp on MS-COCO adopts more learnable parameters than VOC2007. Our proposed DualCoOp achieves the best performance across all proportions of labels available during the training with the smallest learnable overhead (1.3M vs. 29.6M in SARB* on MS-COCO and 0.3M vs. 29.6M in SARB* on VOC2007). Notably, DualCoOp yields a great improvement over the second-best method, 3.2% on MS-COCO and 6.8% on VOC2007, especially when only providing 10% of labels during the training. This indicates that DualCoOp can quickly adapt to the multi-label recognition task with a few labels.

Table 1: **Multi-label Recognition on MS-COCO and VOC2007 with partial labels.** * indicates previous models using weights pretrained by CLIP [29]

Methods	#P	10%	20%	30%	40%	50%	60%	70%	80%	90%	Avg.
MS-COCO [1]											
SSGRL [84]	64.7M	62.5	70.5	73.2	74.5	76.3	76.5	77.1	77.9	78.4	74.1
GCN-ML [85]	44.9M	63.8	70.9	72.8	74.0	76.7	77.1	77.3	78.3	78.6	74.4
KGGR [86]	$\geq$ 25M	66.6	71.4	73.8	76.7	77.5	77.9	78.4	78.7	79.1	75.6
Curriculum labeling [87]	$\geq$ 38M	26.7	31.8	51.5	65.4	70.0	71.9	74.0	77.4	78.0	60.7
Patial BCE [87]	$\geq$ 38M	61.6	70.5	74.1	76.3	77.2	77.7	78.2	78.4	78.5	74.7
SST [81]	33.5M	68.1	73.5	75.9	77.3	78.1	78.9	79.2	79.6	79.9	76.7
SST*	33.5M	69.1	78.5	79.3	79.9	80.1	80.5	81.1	80.7	80.7	78.9
SARB [82]	29.6M	71.2	75.0	77.1	78.3	78.9	79.6	79.8	80.5	80.5	77.9
SARB*	29.6M	75.5	78.5	79.0	79.5	80.4	80.2	80.8	80.6	80.8	79.4
DualCoOp (ours)	**1.3M**	**78.7**	**80.9**	**81.7**	**82.0**	**82.5**	**82.7**	**82.8**	**83.0**	**83.1**	**81.9**
PASCAL VOC 2007 [80]											
SSGRL [84]	66.6M	77.7	87.6	89.9	90.7	91.4	91.8	91.9	92.2	92.2	89.5
GCN-ML [85]	44.9M	74.5	87.4	89.7	90.7	91.0	91.3	91.5	91.8	92.0	88.9
KGGR [86]	$\geq$ 25M	81.3	88.1	89.9	90.4	91.2	91.3	91.5	91.6	91.8	89.7
Curriculum labeling [87]	$\geq$ 38M	44.7	76.8	88.6	90.2	90.7	91.1	91.6	91.7	91.9	84.1
Patial BCE [87]	$\geq$ 38M	80.7	88.4	89.9	90.7	91.2	91.8	92.3	92.4	92.5	90.0
SST [81]	32.4M	81.5	89.0	90.3	91.0	91.6	92.0	92.5	92.6	92.7	90.4
SARB [82]	29.6M	83.5	88.6	90.7	91.4	91.9	92.2	92.6	92.8	92.9	90.7
DualCoOp (ours)	**0.3M**	**90.3**	**92.2**	**92.8**	**93.3**	**93.6**	**93.9**	**94.0**	**94.1**	**94.2**	**93.2**

Zero-shot Multi-Label Recognition. We conduct experiments on MS-COCO [1] and NUS-WIDE [2] to perform zero-shot multi-label recognition. Table 2-3 shows the comparison between DualCoOp and all SOTA methods of zero-shot learning and generalized zero-shot learning on MS-COCO and NUS-WIDE datasets. DualCoOp achieves the best F1 score in all cases with a very light learnable overhead (0.02M) and improves the performance of zero-shot learning (unseen labels) with a significant margin: F1 score improves by 12.5 @Top-3 on MS-COCO, and by 10.8 @Top-3 and 10.9 @Top-5 on NUS-WIDE. This shows the power of exploiting the pretrained alignment of textual and visual spaces in CLIP via DualCoOp to solve multi-label recognition.

4.1.2 Experimental Results: SENTRY: Selective Entropy Optimization via Committee Consistency for Unsupervised Domain Adaptation [10]

Results on label-shifted DA benchmarks. We present results on 12 shifts from DomainNet (Table 4) and 6 shifts from OfficeHome RS→UT (Table 5). On DomainNet, `SENTRY` outperforms the next

Table 2: **Zero-Shot Multi-Label Recognition on MS-COCO[1]**. DualCoOp achieves the best F1 score in both ZSL and GZSL settings.

Methods	#P	ZSL			GZSL		
		P	**R**	**F1**	**P**	**R**	**F1**
CONSE [88]	-	11.4	28.3	16.2	23.8	28.8	26.1
Fast0Tag [89]	0.61M	24.7	61.4	25.3	38.5	46.5	42.1
Deep0Tag [90]	$\geq$ 23M	26.5	65.9	37.8	43.2	52.2	47.3
SDL (M=2) [91]	30.6M	26.3	65.3	37.5	**59.0**	60.8	59.9
DualCoOp (ours)	**0.02M**	**35.3**	**87.6**	**50.3**	58.4	**68.1**	**62.9**

Table 3: **Zero-Shot Multi-label Recognition on NUS-WIDE [2].** DualCoOp achieves the best F1 score over all SOTA methods at Top-3/Top-5 predictions in both ZSL and GZSL settings.

Methods	#P	Top-3			Top-5			mAP
		P	**R**	**F1**	**P**	**R**	**F1**	
Zero-Shot Learning (ZSL)								
CONSE [88]	-	17.5	28.0	21.6	13.9	37.0	20.2	9.4
LabelEM [92]	-	15.6	25.0	19.2	13.4	35.7	19.5	7.1
Fast0Tag [89]	0.61M	22.6	36.2	27.8	18.2	48.4	26.4	15.1
One Attention per Label [93]	$\geq$ 12.8M	20.9	33.5	25.8	16.2	43.2	23.6	10.4
LESA (M=10) [94]	$\geq$ 0.45M	25.7	41.1	31.6	19.7	52.5	28.7	19.4
BiAM [95]	3.8M	–	–	33.1	–	–	30.7	26.3
SDL (M=7) [91]	33.6M	24.2	41.3	30.5	18.8	53.4	27.8	25.9
DualCoOp (ours)	**0.02M**	**37.3**	**46.2**	**41.3**	**28.7**	**59.3**	**38.7**	**43.6**
Generalized Zero-Shot Learning (GZSL)								
CONSE [88]	-	11.5	5.1	7.0	9.6	7.1	8.1	2.1
LabelEM [92]	-	15.5	6.8	9.5	13.4	9.8	11.3	2.2
Fast0Tag [89]	0.61M	18.8	8.3	11.5	15.9	11.7	13.5	3.7
One Attention per Label [93]	$\geq$ 12.8M	17.9	7.9	10.9	15.6	11.5	13.2	3.7
LESA (M=10) [94]	$\geq$ 0.45M	23.6	10.4	14.4	19.8	14.6	16.8	5.6
BiAM [95]	3.8M	–	–	16.1	–	–	19.0	9.3
SDL (M=7) [91]	33.6M	27.7	**13.9**	18.5	23.0	**19.3**	21.0	**12.1**
DualCoOp (ours)	**0.02M**	**31.9**	**13.9**	**19.4**	**26.2**	19.1	**22.1**	12.0

best performing method InstaPBM [96] on every shift, and by 3.55% mean accuracy averaged across shifts. On OfficeHome RS-UT, `SENTRY` outperforms the next best performing method MDD+I.A [97] on 5 out of 6 shifts, and on average by 3.58% mean accuracy. Our method also significantly outperforms F-DANN [98] (by 11.87% and 11.37%) and COAL [99] (by 5.50% and 6.85%), which are both UDA strategies designed for adaptation under LDS.

Results on standard DA benchmarks. Table 5 presents results on 2 standard DA benchmarks: OfficeHome and VisDA 2017. As seen, `SENTRY` improves mean accuracy over the next best method by 2.7% averaged over 12 shifts (full table in supp.) on OfficeHome, and by 0.4% on VisDA.

4.1.3 Experimental Results: Adapting Self-Supervised Vision Transformers by Probing Attention-Conditioned Masking Consistency [11]

Tables 6, and 7, present results. We observe:

▷ **Several existing DA methods underperform on top of SSL representations.** On both OfficeHome and DomainNet, we observe existing DA methods (CDAN [110], MCC [111], TVT [112]) to frequently underperform even the source model on average, despite hyperparameter tuning. The fact that many existing DA methods struggle with such initializations suggests the need for specialized solutions, particularly as SSL pretraining becomes more common. SENTRY [10], Shen *et al.* [113], and `PACMAC` however offer consistent improvements.

▷ `PACMAC` **outperforms prior work across benchmarks and initializations.** On OfficeHome (12 shift average), `PACMAC` improves over the next-best method by 1.3% (MAE init.) and 0.2% (DINO). On DomainNet, we observe gains of 1.1% (MAE) and 0.4% (DINO), and 1.7% (MAE) on VisDA.

Table 4: Per-class average accuracies on DomainNet. Bold and underscore denote the best and second-best performing methods respectively.

Method	**R → C**	**R → P**	**R → S**	**C → R**	**C → P**	**C → S**	**P → R**	**P → C**	**P → S**	**S → R**	**S → C**	**S → P**	AVG
source	65.75	68.84	59.15	77.71	60.60	57.87	84.45	62.35	65.07	77.10	63.00	59.72	66.80
BBSE [100]	55.38	63.62	47.44	64.58	42.18	42.36	81.55	49.04	54.10	68.54	48.19	46.07	55.25
PADA [101]	65.91	67.13	58.43	74.69	53.09	52.86	79.84	59.33	57.87	76.52	66.97	61.08	64.48
MCD [102]	61.97	69.33	56.26	79.78	56.61	53.66	83.38	58.31	60.98	81.74	56.27	66.78	65.42
DAN [103]	64.36	70.65	58.44	79.44	56.78	60.05	84.56	61.62	62.21	79.69	65.01	62.04	67.07
F-DANN [98]	66.15	71.80	61.53	81.85	60.06	61.22	84.46	66.81	62.84	81.38	69.62	66.50	69.52
UAN [104]	71.10	68.90	67.10	83.15	63.30	64.66	83.95	65.35	67.06	82.22	70.64	68.09	72.05
JAN [105]	65.57	73.58	67.61	85.02	64.96	67.17	87.06	67.92	66.10	84.54	72.77	67.51	72.48
ETN [106]	69.22	72.14	63.63	86.54	65.33	63.34	85.04	65.69	68.78	84.93	72.17	68.99	73.99
BSP [107]	67.29	73.47	69.31	86.50	67.52	70.90	86.83	70.33	68.75	84.34	72.40	71.47	74.09
DANN [108]	63.37	73.56	72.63	86.47	65.73	70.58	86.94	73.19	70.15	85.73	75.16	70.04	74.46
COAL [99]	73.85	75.37	70.50	89.63	69.98	71.29	89.81	68.01	70.49	87.97	73.21	70.53	75.89
InstaPBM [96]	80.10	75.87	70.84	89.67	70.21	72.76	89.60	74.41	72.19	87.00	79.66	71.75	77.84
SENTRY (Ours)	**83.89**	**76.72**	**74.43**	**90.61**	**76.02**	**79.47**	**90.27**	**82.91**	**75.60**	**90.41**	**82.40**	**73.98**	**81.39**

Table 5: Accuracies on label-shifted and standard DA benchmarks.

(a) OfficeHome RS-UT (12 shift avg)

Method	acc (%)
source	54.39
BSP [107]	40.97
PADA [101]	42.66
BBSE [100]	44.33
MCD [102]	45.94
DAN [103]	52.90
F-DANN [98]	53.88
JAN [105]	54.24
DANN [108]	56.91
MDD [109]	55.44
COAL [99]	58.40
InstaPBM [96]	58.90
MDD+I.A [97]	61.67
SENTRY (Ours)	**65.25**

(b) OfficeHome (12 shift avg)

Method	acc (%)
Source	46.1
DAN [103]	56.3
DANN [108]	57.6
JAN [105]	58.3
CDAN [110]	65.8
BSP [107]	66.3
MDD [109]	68.1
InstaPBM [96]	69.2
MDD+I.A [97]	69.5
SENTRY (Ours)	**72.2**

(c) VisDA2017

Method	acc (%)
Source	41.0
JAN [103]	61.6
MCD [102]	69.8
CDAN [110]	70.0
MDD [109]	74.6
MDD+I.A [97]	75.8
InstaPBM [96]	76.3
SENTRY (Ours)	**76.7**

4.2 Few-Shot Image Classification

4.2.1 Experimental Results: Active Domain Adaptation via Clustering Uncertainty-weighted Embeddings [12]

Tables 8 demonstrates our results on DomainNet. We make the following observations:

▷ **Uncertainty and diversity sampling are less effective for Active DA, frequently underperforming even random sampling.** Approaches solely based on uncertainty (e.g. `margin` [38]) work well on relatively easier shifts (R→C with `MME`, SVHN→MNIST), but overall we find uncertainty-based (`margin`, `entropy`), and diversity-based (`coreset`) approaches generalize poorly to challenging shifts (e.g. S→P, C→Q), frequently underperforming even random sampling! On the other hand, hybrid approaches (`CLUE` and `BADGE`) that combine uncertainty and diversity are versatile across shift difficulties.

▷ `CLUE` **outperforms prior AL methods in the Active DA setting.** Across learning strategies, shifts, benchmarks, and most rounds, `CLUE` consistently performs best. Averaged over 4 DomainNet shifts, `CLUE` outperforms `margin`-based uncertainty sampling and `coreset`-based diversity sampling at $B = 2k$ by 1.4% and 3% when finetuning, and 1.3% and 2.3% when adapting via `MME` (Tab. 8).

Approved for Public Release; Distribution Unlimited.

Table 6: Target test set accuracy on OfficeHome across MAE [3] and DINO [4] pretraining.

IN1K Init.	Method	A → C	A → P	A → R	C → A	C → P	C → R	P → A	P → C	P → R	R → A	R → C	R → P	AVG
MAE [3]	source	46.4	57.6	71.0	51.1	60.0	62.6	51.4	46.9	70.5	66.3	52.2	77.2	59.4
	CDAN [110]	45.3	58.8	69.1	51.6	60.7	61.5	53.4	45.5	72.4	67.7	49.9	78.0	59.5
	MCC [111]	43.9	61.2	70.7	52.8	59.9	62.8	51.1	40.3	70.9	66.2	48.3	76.3	58.7
	Shen *et al.** [113]	57.1	63.6	71.9	57.9	65.6	67.1	55.5	56.7	71.2	69.0	62.6	79.4	64.8
	SENTRY [10]	54.8	65.6	74.4	56.5	65.8	69.8	57.6	54.9	75.5	68.9	60.0	81.6	65.5
	PACMAC (Ours)	58.9	68.2	74.1	60.6	67.1	67.2	57.3	59.2	74.4	68.6	63.9	82.7	**66.8**
DINO [4]	source	53.1	65.0	75.2	62.0	66.2	70.4	60.8	50.5	77.0	72.8	53.9	81.2	65.7
	CDAN [110]	49.0	70.0	76.4	60.0	67.3	71.2	64.7	47.0	79.9	75.1	56.4	81.8	66.5
	MCC [111]	44.4	74.2	79.6	61.9	67.6	72.4	63.0	40.1	79.2	73.3	47.1	82.8	65.5
	TVT [112]	48.3	65.7	73.6	60.6	68.8	64.6	57.1	44.1	75.4	71.0	53.7	77.2	63.3
	Shen *et al.** [113]	53.1	69.4	76.7	62.6	68.9	71.4	62.2	51.8	76.0	73.5	56.3	81.8	67.0
	SENTRY [10]	59.5	72.0	76.8	66.1	71.1	73.4	63.7	56.2	77.8	72.4	63.0	81.9	69.5
	PACMAC (Ours)	54.9	74.7	79.3	65.7	74.0	74.5	63.3	55.8	79.2	73.1	58.4	83.9	**69.7**

Table 7: Target test set accuracy on DomainNet across MAE [3] and DINO [4] pretraining.

IN1K Init.	Method	R → C	R → P	R → S	C → R	C → P	C → S	P → R	P → C	P → S	S → R	S → C	S → P	AVG
MAE [3]	source	71.0	77.6	62.9	73.7	61.5	63.3	82.4	63.1	66.1	76.6	71.9	69.6	70.1
	CDAN [110]	72.2	74.5	59.3	80.6	57.3	59.2	78.5	57.4	61.2	81.4	73.2	69.4	68.7
	Shen *et al.** [113]	83.0	80.2	77.7	84.2	74.6	74.7	85.0	77.5	76.5	83.7	80.7	76.2	79.5
	SENTRY [10]	84.2	82.8	76.4	86.9	77.1	74.1	86.9	76.2	73.3	88.8	81.6	77.6	80.5
	PACMAC (Ours)	86.0	81.9	78.8	86.0	74.8	76.3	87.4	84.0	77.5	85.2	83.1	78.3	**81.6**
DINO [4]	source	75.7	82.8	68.2	81.9	73.9	71.1	82.6	70.6	69.5	80.8	76.9	78.1	76.0
	CDAN [110]	81.1	84.2	77.2	84.8	76.2	72.5	84.6	69.0	69.1	84.9	79.8	80.1	78.6
	TVT [112]	70.4	79.7	64.2	80.2	68.1	65.2	81.6	61.9	65.6	80.3	71.4	74.1	71.9
	Shen *et al.** [113]	76.6	81.9	76.8	82.1	74.9	74.3	84.7	69.7	74.3	83.0	77.9	80.1	78.0
	SENTRY [10]	81.8	80.9	73.4	89.1	79.4	75.8	86.5	75.6	71.6	88.3	81.9	82.4	80.6
	PACMAC (Ours)	80.7	82.9	82.0	85.7	78.8	78.3	87.3	75.5	75.2	84.7	79.6	82.0	**81.0**

Similarly, CLUE outperforms the next best-performing method (BADGE) by 0.7% at $B = 2k$ when finetuning and 0.8% with MME (4 shift average).

4.2.2 Experimental Results: Semi-Supervised Domain Adaptation via Minimax Entropy [13]

Overview. The main results on the DomainNet dataset are shown in Table 9. First, our method outperformed other baselines inlcuding S+T [116], DANN [117], ADR [118], CDAN [110], ENT [119] for all adaptation scenarios and all three networks except for one case. On average, our method outperformed S+T with 9.5% and 8.9% in ResNet one-shot and three-shot setting respectively. The results on Office-Home and Office are summarized in Table 10, where MME also outperforms all baselines. Due to the limited space, we show the results averaged on all adaptation scenarios.

Comparison with UDA Methods. Generally, baseline UDA methods need strong base networks such as VGG or ResNet to perform better than S+T. Interestingly, these methods cannot improve the performance in some cases. The superiority of MME over existing UDA methods is supported by Tables 9 and 10. Since CDAN uses entropy minimization and ENT significantly hurts the performance for AlexNet and VGG, CDAN does not consistently improve the performance for AlexNet and VGG.

Comparison with Entropy Minimization. ENT does not improve performance in some cases because it does not account for the domain gap. Comparing results on one-shot and three-shot, entropy minimization gains performance with the help of labeled examples. As we have more labeled target examples, the estimation of prototypes will be more accurate without any adaptation. In case of ResNet, entropy minimization often improves accuracy. There are two potential reasons. First, ResNet pretrained on ImageNet has a more discriminative representation than other networks. Therefore, given a few labeled target examples, the model can extract more discriminative features, which contributes to the performance gain in entropy minimization. Second, ResNet has batch-normalization (BN) layers. It is reported that BN has the effect of aligning feature distributions [120]. Hence, entropy minimization was done on aligned feature representations, which improved the performance. When there is a large domain gap such as C to S, S to P, and R to S in Table 9, BN is not enough to handle the

Table 8: Accuracies on target test set for 4 DomainNet shifts of increasing difficulty spanning 5 domains: Real (R), Clipart (C), Sketch (S), Painting (P) and Quickdraw (Q). Accuracies on target test set for 4 DomainNet shifts of increasing difficulty spanning 5 domains: Real (R), Clipart (C), Sketch (S), Painting (P) and Quickdraw (Q). We perform 10 rounds of Active DA with $B = 500$ and report results at 3 intermediate rounds (full plots in supplementary), as well as the 4-shift average (AVG). We compare `CLUE` against state-of-the art methods for AL (`entropy` [39], `margin` [38], `coreset` [41], `BADGE` [114]) and Active DA (`AADA`), spanning different AL paradigms: uncertainty sampling (U), diversity sampling (D), and hybrid (H) combinations of the two. We use multiple learning strategies: finetuning (`ft`), `MME` [13] (state-of-the-art semi-supervised DA method), and semi-supervised DA via `DANN` [108]. Best performance is in bold, gray rows are our method.

DA method	AL method	AL Type	**R** → **C** (easy)			**C** → **S** (moderate)			**S** → **P** (hard)			**C** → **Q** (very hard)			AVG		
			1k	2k	5k	1k	2k	5k	1k	2k	5k	1k	2k	5k	1k	2k	5k
`ft` from source	`uniform`	-	51.5	55.3	60.6	42.1	44.4	47.0	41.1	43.8	47.2	23.3	28.1	35.3	39.5	42.9	47.5
	`entropy` [39]	U	48.1	52.1	58.6	41.1	42.7	45.7	41.2	43.8	47.2	21.9	26.4	34.0	38.1	41.3	46.4
	`margin` [38]	U	51.0	54.8	60.7	42.3	44.3	47.0	41.4	44.0	47.1	23.6	28.4	35.8	39.6	42.9	47.7
	`coreset` [41]	D	50.0	54.0	59.6	41.2	42.8	44.9	40.1	42.2	45.4	22.4	26.0	32.4	38.4	41.3	45.6
	`BADGE` [114]	H	52.4	56.1	61.7	42.8	45.2	48.1	41.7	44.9	47.9	23.1	28.2	**35.5**	39.8	43.6	48.3
	`CLUE` (Ours)	H	**52.9**	**57.1**	**62.0**	**43.3**	**45.8**	**48.6**	**42.4**	**45.3**	**48.3**	**24.3**	**28.8**	**35.5**	**40.7**	**44.3**	**48.6**
`MME` [13] from source	`uniform`	-	55.2	59.3	63.5	45.7	47.8	49.7	42.9	45.3	47.8	24.5	30.3	38.1	42.1	45.7	49.8
	`entropy` [39]	U	53.8	58.6	64.4	44.2	45.7	48.5	41.6	43.9	47.2	21.9	25.7	32.8	40.4	43.5	48.2
	`margin` [38]	U	55.6	**60.7**	**65.7**	46.0	48.1	50.8	42.2	44.8	48.2	23.1	28.3	36.6	41.7	45.5	50.3
	`coreset` [41]	D	54.3	59.1	64.6	45.1	46.7	48.9	42.4	44.2	47.1	23.9	27.8	34.3	41.4	44.5	48.7
	`BADGE` [114]	H	56.2	60.6	**65.7**	45.8	48.2	50.7	43.1	45.7	48.7	24.3	29.6	38.3	42.4	46.0	50.9
	`CLUE` (Ours)	H	**56.3**	**60.7**	65.3	**46.8**	**49.0**	**51.4**	**43.7**	**46.5**	**49.4**	**25.6**	**31.1**	**38.9**	**43.1**	**46.8**	**51.3**
`DANN` [108] from source	`AADA` [115]	H	53.2	57.4	62.8	44.8	46.5	49.2	41.3	43.5	46.1	21.9	25.8	32.4	40.3	43.3	47.6
	`CLUE` (Ours)	H	**54.6**	**58.9**	**63.8**	**45.3**	**47.9**	**50.8**	**43.2**	**45.5**	**48.3**	**24.4**	**29.2**	**35.4**	**41.9**	**45.4**	**49.6**

domain gap. Therefore, our proposed method performs much better than entropy minimization in such cases. We show an analysis of BN in our supplemental material, revealing its effectiveness for entropy minimization.

4.2.3 Experimental Results: OpenMatch: Open-set Consistency Regularization for Semi-supervised Learning with Outliers [14]

We evaluate the efficacy of OpenMatch on several SSL image classification benchmarks. Specifically, we perform experiments with varying amounts of labeled data and varying numbers of known/unknown classes on CIFAR10/100 [121] and ImageNet [122].

Evaluation on CIFAR10/100. We compare OpenMatch with several baselines on the standard benchmark datasets of SSL, CIFAR10, and CIFAR100. Following [123], we use a randomly initialized Wide ResNet-28-2 [124] with 1.5M parameters in these experiments. For CIFAR10, we split the classes into known and unknown classes by defining animal classes as known and others as unknown, which results in 6 known classes and 4 unknown classes. For CIFAR100, we first split the classes by their super-classes provided by the dataset so that known and unknown classes do not share super-classes. Then, we split the super-classes into known and unknown. To evaluate performance with different numbers of outliers, we run experiments on two settings: 80 known classes (20 unknown classes) and 55 known classes (45 unknown classes).

Tables 11 and 12 describe the error rate on inliers and AUROC values respectively. Note that the number of labeled samples *per class* is shown in each column. With respect to the error rate, OpenMatch achieves state-of-the-art performance in all cases, and the gain over the baselines is remarkable in many cases. In particular, we see about a 10% improvement in CIFAR10 error at 50 labels. In addition, the improvements in AUROC are also clear. MTC [123] does not improve the performance very much compared to the labeled-only model on CIFAR100 because their method of identifying outliers, *i.e.*, Otsu's thresholding [125], is not very robust to the number of outliers as discussed in their paper. On the other hand, OpenMatch improves the error rate by more than 3% and AUROC by more than 6% in all cases. Simply applying FixMatch [43] also does not always improve the error compared to the labeled-only model. This is because FixMatch confuses inliers and outliers and assigns pseudo-labels to outliers, as the significant decrease in AUROC indicates.

Evaluation on ImageNet. We evaluate OpenMatch on ImageNet to see its behavior on more complex and challenging datasets. Since training on the entire ImageNet is computationally expensive, we choose ImageNet-30 [126], which is a subset of ImageNet containing 30 classes. The subset is useful for evaluation on unseen outliers, because the classes do not include several super-classes such as a bird class. We pick the first 20 classes, in alphabetical order, as known classes, and the remaining 10

Table 9: Accuracy on the DomainNet dataset (%) for one-shot and three-shot settings on 4 domains, R: Real, C: Clipart, P: Clipart, S: Sketch. Our MME method outperformed other baselines for all adaptation scenarios and for all three networks, except for only one case where it performs similarly to ENT.

Net	Method	R to C		R to P		P to C		C to S		S to P		R to S		P to R		MEAN	
		1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot
AlexNet	S+T	43.3	47.1	42.4	45.0	40.1	44.9	33.6	36.4	35.7	38.4	29.1	33.3	55.8	58.7	40.0	43.4
	DANN	43.3	46.1	41.6	43.8	39.1	41.0	35.9	36.5	36.9	38.9	32.5	33.4	53.6	57.3	40.4	42.4
	ADR	43.1	46.2	41.4	44.4	39.3	43.6	32.8	36.4	33.1	38.9	29.1	32.4	55.9	57.3	39.2	42.7
	CDAN	46.3	46.8	45.7	45.0	38.3	42.3	27.5	29.5	30.2	33.7	28.8	31.3	56.7	58.7	39.1	41.0
	ENT	37.0	45.5	35.6	42.6	26.8	40.4	18.9	31.1	15.1	29.6	18.0	29.6	52.2	60.0	29.1	39.8
	MME	**48.9**	**55.6**	**48.0**	**49.0**	**46.7**	**51.7**	**36.3**	**39.4**	**39.4**	**43.0**	**33.3**	**37.9**	**56.8**	**60.7**	**44.2**	**48.2**
VGG	S+T	49.0	52.3	55.4	56.7	47.7	51.0	43.9	48.5	50.8	55.1	37.9	45.0	69.0	71.7	50.5	54.3
	DANN	43.9	56.8	42.0	57.5	37.3	49.2	46.7	48.2	51.9	55.6	30.2	45.6	65.8	70.1	45.4	54.7
	ADR	48.3	50.2	54.6	56.1	47.3	51.5	44.0	49.0	50.7	53.5	38.6	44.7	67.6	70.9	50.2	53.7
	CDAN	57.8	58.1	57.8	59.1	51.0	57.4	42.5	47.2	51.2	54.5	42.6	49.3	71.7	74.6	53.5	57.2
	ENT	39.6	50.3	43.9	54.6	26.4	47.4	27.0	41.9	29.1	51.0	19.3	39.7	68.2	72.5	36.2	51.1
	MME	**60.6**	**64.1**	**63.3**	**63.5**	**57.0**	**60.7**	**50.9**	**55.4**	**60.5**	**60.9**	**50.2**	**54.8**	**72.2**	**75.3**	**59.2**	**62.1**
ResNet	S+T	55.6	60.0	60.6	62.2	56.8	59.4	50.8	55.0	56.0	59.5	46.3	50.1	71.8	73.9	56.9	60.0
	DANN	58.2	59.8	61.4	62.8	56.3	59.6	52.8	55.4	57.4	59.9	52.2	54.9	70.3	72.2	58.4	60.7
	ADR	57.1	60.7	61.3	61.9	57.0	60.7	51.0	54.4	56.0	59.9	49.0	51.1	72.0	74.2	57.6	60.4
	CDAN	65.0	69.0	64.9	67.3	63.7	68.4	53.1	57.8	63.4	65.3	54.5	59.0	73.2	78.5	62.5	66.5
	ENT	65.2	71.0	65.9	69.2	65.4	71.1	54.6	60.0	59.7	62.1	52.1	61.1	75.0	**78.6**	62.6	67.6
	MME	**70.0**	**72.2**	**67.7**	**69.7**	**69.0**	**71.7**	**56.3**	**61.8**	**64.8**	**66.8**	**61.0**	**61.9**	**76.1**	78.5	**66.4**	**68.9**

Table 10: Results on Office-Home and Office dataset (%). The value is the accuracy averaged over all adaptation scenarios. Performance on each setting is summarized in supplementary material.

Net	Method	Office-Home		Office	
		1-shot	3-shot	1-shot	3-shot
AlexNet	S+T	44.1	50.0	50.2	61.8
	DANN	45.1	50.3	55.8	64.8
	ADR	44.5	49.5	50.6	61.3
	CDAN	41.2	46.2	49.4	60.8
	ENT	38.8	50.9	48.1	65.1
	MME	**49.2**	**55.2**	**56.5**	**67.6**

Net	Method	Office-Home		Office	
		1-shot	3-shot	1-shot	3-shot
VGG	S+T	57.4	62.9	68.7	73.3
	DANN	60.0	63.9	69.8	75.0
	ADR	57.4	63.0	69.4	73.7
	CDAN	55.8	61.8	65.9	72.9
	ENT	51.6	64.8	70.6	75.3
	MME	**62.7**	**67.6**	**73.4**	**77.0**

classes as unknown. We utilize the ResNet-18 [83] network architecture. Hyper-parameters such as a batch-size, learning rate, data augmentation are the same as ones used in CIFAR10 and CIFAR100. As shown in the rightmost columns of Table 11 and 12, OpenMatch achieves the best performance in both accuracy and AUROC. Unlike the results on CIFAR, FixMatch improves the performance over the Labeled Only model. This is probably because outliers are easier to detect, as the AUROC score of the Labeled Only model already shows a higher score than other scenarios. Therefore, even without any outlier detection objectives, FixMatch can correctly select pseudo-inliers with confidence thresholding.

Novelty Detection. We have seen that OpenMatch can learn to detect outliers from unlabeled data. In this experiment, we evaluate how well our OSSL model can separate inliers from outliers unseen in the unlabeled data. As seen in Table 13, OpenMatch outperforms OSSL baselines by 5.8% on average and even the supervised model by 3.4% on average on CIFAR10. On ImageNet-30, OpenMatch outperforms OSSL baselines by more than 4.1% on average. This result verifies that OpenMatch is robust to various out-of-distribution data by virtue of being exposed to unlabeled data containing outliers.

4.2.4 Experimental Results: Subspace Regularization for Few-Shot Class Incremental Learning [15]

Figure 24 and Table 14 show the results of multi-session experiments with and without memory. Session 0 indicates base class accuracy after feature extractor training. We compare subspace and language-guided regularization (linear mapping and semantic subspace reg.) to simple fine-tuning (a surprisingly strong baseline) and joint-training. We also compare our results to three recent benchmarks: [50], [127] and [128].

Approved for Public Release; Distribution Unlimited.

Table 11: Error rates (%) with standard deviation for CIFAR10, CIFAR100 on 3 different folds. Lower is better. For ImageNet, we use the same fold and report averaged results of three runs. Note that the number of labeled samples *per each class* is shown in each column.

Dataset	CIFAR10			CIFAR100		CIFAR100		ImageNet-30
No. of Known / Unknown	6 / 4			55 / 45		80 / 20		20 / 10
No. of labeled samples	50	100	400	50	100	50	100	10 %
Labeled Only	35.7±1.1	30.5±0.7	20.0±0.3	37.0±0.8	27.3±0.5	43.6±0.5	34.7±0.4	20.9±1.0
FixMatch [43]	43.2±1.2	29.8±0.6	16.3±0.5	35.4±0.7	27.3±0.8	41.2±0.7	34.1±0.4	12.9±0.4
MTC [123]	20.3±0.9	13.7±0.9	9.0±0.5	33.5±1.2	27.9±0.5	40.1±0.8	33.6±0.3	13.6±0.7
OpenMatch	**10.4±0.9**	**7.1±0.5**	**5.9±0.5**	**27.7±0.4**	**24.1±0.6**	**33.4±0.2**	**29.5±0.3**	**10.4±1.0**

Table 12: AUROC of Table 11. Higher is better. Note that the number of labeled samples *per each class* is shown in each column.

Dataset	CIFAR10			CIFAR100		CIFAR100		ImageNet-30
No. of Known / Unknown	6 / 4			55 / 45		80 / 20		20 / 10
No. of labeled samples	50	100	400	50	100	50	100	10 %
Labeled Only	63.9±0.5	64.7±0.5	76.8±0.4	76.6±0.9	79.9±0.9	70.3±0.5	73.9±0.9	80.3±1.0
FixMatch [43]	56.1±0.6	60.4±0.4	71.8±0.4	72.0±1.3	75.8±1.2	64.3±1.0	66.1±0.5	88.6±0.5
MTC [123]	96.6±0.6	98.2±0.3	98.9±0.1	81.2±3.4	80.7±4.6	79.4±2.5	73.2±3.5	93.8±0.8
OpenMatch	**99.3±0.3**	**99.7±0.2**	**99.3±0.2**	**87.0±1.1**	**86.5±2.1**	**86.2±0.6**	**86.8±1.4**	**96.4±0.7**

When models are evaluated on combined base and novel accuracy, subspace regularization outperforms previous approaches (by 23% (-M) and 10% (+M) at session 8); when semantic information about labels is available, linear mapping and semantic subspace regularization outperform [128] (Table 14). Evaluating only base sample accuracies (Figure 24b), semantic subspace reg. outperforms others; compared to regularization based approaches fine-tuning is subject to catastrophic forgetting. R_{old} is still useful in regulating forgetting. The method of [127] follows a similar trajectory to our regularizers, but at a much lower accuracy (Figure 24a). In Figure 24c, a high degree of forgetting in base classes with fine-tuning allows higher accuracy for novel classes—though not enough to improve average performance (Figure 24a). By contrast, subspace regularizers provide a good balance between *plasticity* and *stability* [129]. In Figure 14, storing as few as a single example per an old class substantially helps to reduce forgetting. Results from linear mapping and semantic subspace regularization are close, with semantic subspace regularization performing roughly 1% better on average. The two approaches offer different trade-offs between base and novel accuracies: the latter is more competitive for base classes and vice-versa. A similar argument can be made when comparing the subspace regularizer with the semantic counterpart where the latter is more efficient in base classes than the former. Last but not least, keeping as little as one example per old class in the memory along with our subspace regularizers is sufficient to surpass joint-training—a procedure which is extremely inefficient in terms of both space and computation.

4.2.5 Results: A ConvNet for the 2020s [16]

We construct different ConvNeXt variants, ConvNeXt-T/S/B/L, to be of similar complexities to Swin-T/S/B/L. In addition, we build a larger ConvNeXt-XL to further test the scalability of ConvNeXt. The variants only differ in the number of channels C, and the number of blocks B in each stage. Following both ResNets and Swin Transformers, the number of channels doubles at each new stage.

Training on ImageNet-1K. We train ConvNeXts for 300 epochs using AdamW [130] with a learning rate of 4e-3. There is a 20-epoch linear warmup and a cosine decaying schedule afterward. We use a batch size of 4096 and a weight decay of 0.05. For data augmentations, we adopt common schemes including Mixup [131], Cutmix [132], RandAugment [133], and Random Erasing [134]. We regularize the networks with Stochastic Depth [135] and Label Smoothing. Layer Scale of initial value 1e-6 is applied. We use Exponential Moving Average (EMA) as we find it alleviates larger models' overfitting.

Approved for Public Release; Distribution Unlimited.

Table 13: Evaluation of outlier detection on outliers unseen in unlabeled training data (AUROC). Higher is better. Supervised models use the same batch size, learning rate as OpenMatch, but are trained with fully labeled inliers.

		Unseen Out-liers						
Method	ImageNet-30	LSUN	DTD	CUB	Flowers	Caltech	Dogs	MEAN
Labeled Only	80.3±0.5	85.9±1.4	75.4±1.0	77.9±0.8	69.0±1.5	78.7±0.8	84.8±1.0	78.6±1.1
FixMatch [43]	88.6±0.5	85.7±0.1	83.1±2.5	81.0±4.8	**81.9±1.1**	83.1±3.4	86.4±3.2	83.0±1.9
MTC [123]	93.8±0.8	78.0±1.0	59.5±1.5	72.2±0.9	76.4 ±2.1	80.9±0.9	78.0±0.8	74.2±1.2
OpenMatch	**96.3±0.7**	**89.9±1.9**	**84.4±0.5**	**87.7±1.0**	80.8±1.9	**87.7±0.9**	**92.1±0.4**	**87.1±1.1**
Supervised	92.8±0.8	94.4±0.5	92.7±0.4	91.5±0.9	88.2±1.0	89.9±0.5	92.3±0.8	91.3±0.7

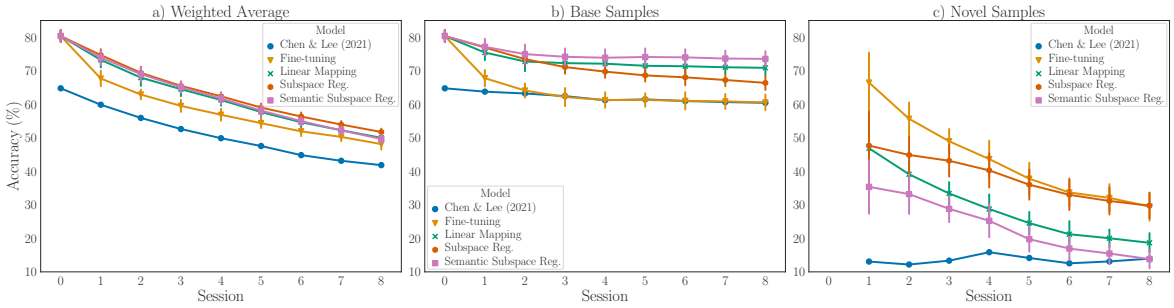


Figure 24: Multi-Session FSCIL accuracy (%) results on *mini*ImageNet. In the first session *0*, there are a total of 60 classes (*base*). Every session following the first one introduces 5 *novel* classes with 5 labeled samples from each. Each session provides accuracy over all classes that were seen thus far. Weighted average is the weighted combination of novel and base accuracies with respect to the number of classes in each category. Error bars are standard deviation (95% CI). In accordance with [127] we preserve only one sample per class from previous classes and append them to the support set during fine-tuning (+M variant). Regularization based approaches i.e. *subspace regularization*, *semantic subspace regularization* and *linear mapping* consistently outperform previous benchmark and *fine-tuning* in weighted average (a).

Pre-training on ImageNet-22K. We pre-train ConvNeXts on ImageNet-22K for 90 epochs with a warmup of 5 epochs. We do not use EMA. Other settings follow ImageNet-1K. **Fine-tuning on ImageNet-1K.** We fine-tune ImageNet-22K pre-trained models on ImageNet-1K for 30 epochs. We use AdamW, a learning rate of 5e-5, cosine learning rate schedule, layer-wise learning rate decay, no warmup, a batch size of 512, and weight decay of 1e-8. The default pre-training, fine-tuning, and testing resolution is 224^2. Additionally, we fine-tune at a larger resolution of 384^2, for both ImageNet-22K and ImageNet-1K pre-trained models. Compared with ViTs/Swin Transformers, ConvNeXts are simpler to fine-tune at different resolutions, as the network is fully-convolutional and there is no need to adjust the input patch size or interpolate absolute/relative position biases.

4.2.6 Results

ImageNet-1K. Table 15 (upper) shows the result comparison with two recent Transformer variants, DeiT [139] and Swin Transformers [142], as well as two ConvNets from architecture search - RegNets [136], EfficientNets [137] and EfficientNetsV2 [138]. ConvNeXt competes favorably with two strong ConvNet baselines (RegNet [136] and EfficientNet [137]) in terms of the accuracy-computation trade-off, as well as the inference throughputs. ConvNeXt also outperforms Swin Transformer of similar complexities *across the board*, sometimes with a substantial margin (e.g. 0.8% for ConvNeXt-T). Without specialized modules such as shifted windows or relative position bias, ConvNeXts also enjoy improved throughput compared to Swin Transformers.

A highlight from the results is ConvNeXt-B at 384^2: it outperforms Swin-B by 0.6% (85.1% vs. 84.5%), but with 12.5% higher inference throughput (95.7 vs. 85.1 image/s). We note that

Approved for Public Release; Distribution Unlimited.

Table 14: Multi-Session FSCIL weighted average of accuracy (%) results on *mini*ImageNet using an identical setup to fig. 24 with memory distinction. We report the average results over 10 random splits of the data for incremental sessions 1, 2 and 8. $\pm M$ indicates 1 sample per class is kept (or not) in the memory to further regularize forgetting. Our Fine-tuning baseline is already superior to previous results. In both memory settings, our regularizers substantially outperform respective benchmarks for all 1-8 sessions. *Results are only estimates from the plot in the respective work. Bold indicates the highest.

Session	1		2		8	
Model	$-M$	$+M$	$-M$	$+M$	$-M$	$+M$
[50]	50.1		45.2		24.4	
[127]		59.9		55.9		41.8
Fine-tuning	61.8	67.7	49.9	62.9	26.5	48.1
Subspace Reg.	**74.0**	**74.7**	**68.9**	**69.4**	**50.6**	**51.7**
+language						
[128]*	58.0		53.0		39.0	
Linear Mapping	72.6	73.2	67.1	68.0	46.9	**50.0**
Semantic Subspace Reg.	**73.8**	**73.9**	**68.4**	**69.0**	**47.6**	49.7
Joint-training		73.1		67.5		47.3

the FLOPs/throughput advantage of ConvNeXt-B over Swin-B becomes larger when the resolution increases from 224^2 to 384^2. Additionally, we observe an improved result of 85.5% when further scaling to ConvNeXt-L.

ImageNet-22K. We present results with models fine-tuned from ImageNet-22K pre-training at Table 15 (lower). These experiments are important since a widely held view is that vision Transformers have fewer inductive biases thus can perform better than ConvNets when pre-trained on a larger scale. Our results demonstrate that properly designed ConvNets are *not* inferior to vision Transformers when pre-trained with large dataset — ConvNeXts still perform on par or better than similarly-sized Swin Transformers, with slightly higher throughput. Additionally, our ConvNeXt-XL model achieves an accuracy of 87.8% — a decent improvement over ConvNeXt-L at 384^2, demonstrating that ConvNeXts are scalable architectures.

On ImageNet-1K, EfficientNetV2-L, a searched architecture equipped with advanced modules (such as Squeeze-and-Excitation [143]) and progressive training procedure achieves top performance. However, with ImageNet-22K pre-training, ConvNeXt is able to outperform EfficientNetV2, further demonstrating the importance of large-scale training.

4.3 Few-Shot Object Detection

4.3.1 Experimental Results: Cut and Learn for Unsupervised Object Detection and Instance Segmentation [18]

Zero-shot Unsupervised Object Detection. We conduct extensive experiments on eleven different datasets, covering various object categories, image styles, video frames, resolutions, camera angles, etc., to verify the effectiveness of CutLER as a universal unsupervised object detection and segmentation method. CutLER is trained solely using images from ImageNet and evaluated in a zero-shot manner on all downstream datasets without finetuning on any labels or data.

We evaluate CutLER on a variety of datasets and report the detection performance using AP^{box}_{50} and AR^{box}_{100} metrics in Fig. 25. CutLER uses a *smaller* model size and *less* training data than prior work. Compared to the previous SOTA approach, FreeSOLO [144] with a backbone of ResNet101, CutLER, with the smaller ResNet50 backbone, significantly outperforms it in each of these benchmarks spanning various image distributions, more than doubling performance on 10 of them. Also note that, FreeSOLO requires FreeMask pre-training using approximately 1.3M ImageNet images and model fine-tuning using additional data in test benchmarks.

We observe that on different domains, e.g. watercolor or frames from videos (UVO dataset), CutLER improves performance by over 4× and 2×, respectively. Fig. 25 shows some qualitative examples of CutLER's predictions.

Label-Efficient and Fully-Supervised Learning We now evaluate CutLER as a pretraining method

Table 15: Classification accuracy on ImageNet-1K. Similar to Transformers, ConvNeXt also shows promising scaling behavior with higher-capacity models and a larger (pre-training) dataset.

model	image size	#param.	FLOPs	throughput (image / s)	IN-1K top-1 acc.
ImageNet-1K trained models					
• RegNetY-16G [136]	224^2	84M	16.0G	334.7	82.9
• EffNet-B7 [137]	600^2	66M	37.0G	55.1	84.3
• EffNetV2-L [138]	480^2	120M	53.0G	83.7	85.7
○ DeiT-S [139]	224^2	22M	4.6G	978.5	79.8
○ DeiT-B [139]	224^2	87M	17.6G	302.1	81.8
○ Swin-T	224^2	28M	4.5G	757.9	81.3
• ConvNeXt-T	224^2	29M	4.5G	774.7	**82.1**
○ Swin-S	224^2	50M	8.7G	436.7	83.0
• ConvNeXt-S	224^2	50M	8.7G	447.1	**83.1**
○ Swin-B	224^2	88M	15.4G	286.6	83.5
• ConvNeXt-B	224^2	89M	15.4G	292.1	**83.8**
○ Swin-B	384^2	88M	47.1G	85.1	84.5
• ConvNeXt-B	384^2	89M	45.0G	95.7	**85.1**
• ConvNeXt-L	224^2	198M	34.4G	146.8	**84.3**
• ConvNeXt-L	384^2	198M	101.0G	50.4	**85.5**
ImageNet-22K pre-trained models					
• R-101x3 [140]	384^2	388M	204.6G	-	84.4
• R-152x4 [140]	480^2	937M	840.5G	-	85.4
• EffNetV2-L [138]	480^2	120M	53.0G	83.7	86.8
• EffNetV2-XL [138]	480^2	208M	94.0G	56.5	87.3
○ ViT-B/16 [141]	384^2	87M	55.5G	93.1	85.4
○ ViT-L/16 [141]	384^2	305M	191.1G	28.5	86.8
• ConvNeXt-T	224^2	29M	4.5G	774.7	**82.9**
• ConvNeXt-T	384^2	29M	13.1G	282.8	**84.1**
• ConvNeXt-S	224^2	50M	8.7G	447.1	**84.6**
• ConvNeXt-S	384^2	50M	25.5G	163.5	**85.8**
○ Swin-B	224^2	88M	15.4G	286.6	85.2
• ConvNeXt-B	224^2	89M	15.4G	292.1	**85.8**
○ Swin-B	384^2	88M	47.0G	85.1	86.4
• ConvNeXt-B	384^2	89M	45.1G	95.7	**86.8**
○ Swin-L	224^2	197M	34.5G	145.0	86.3
• ConvNeXt-L	224^2	198M	34.4G	146.8	**86.6**
○ Swin-L	384^2	197M	103.9G	46.0	87.3
• ConvNeXt-L	384^2	198M	101.0G	50.4	**87.5**
• ConvNeXt-XL	224^2	350M	60.9G	89.3	**87.0**
• ConvNeXt-XL	384^2	350M	179.0G	30.2	**87.8**

for training object detection and instance segmentation models. While CutLER can discover objects without any supervision, finetuning it on a target dataset aligns the model output to the same set of objects labeled in the dataset. Fig. 26 shows the results of fine-tuning the detector on different subsets of COCO. When tested with low-shot settings, e.g., 2% and 5% labeled data, our approach achieves 5.4% and 7.3% higher AP^{box} than the MoCo-v2 baseline, respectively. Even when training with full annotations, MaskCut still consistently gives more than 2% improvements, outperforming MoCo-v2 for both object detection and segmentation. More impressively, MaskCut outperforms prior SOTA methods - FreeSOLO [144] and DETReg [145] despite using an older detection architecture.

4.3.2 Experimental Results: VideoCutLER: Surprisingly Simple Unsupervised Video Instance Segmentation [19]

Evaluating unsupervised video instance segmentation poses two main challenges. Firstly, as unsupervised learning methods train the model without semantic classes, the class-aware video segmentation setup cannot be used directly for an evaluation. As a result, following previous works, we evaluate video instance segmentation results in a class-agnostic manner. Secondly, video instance

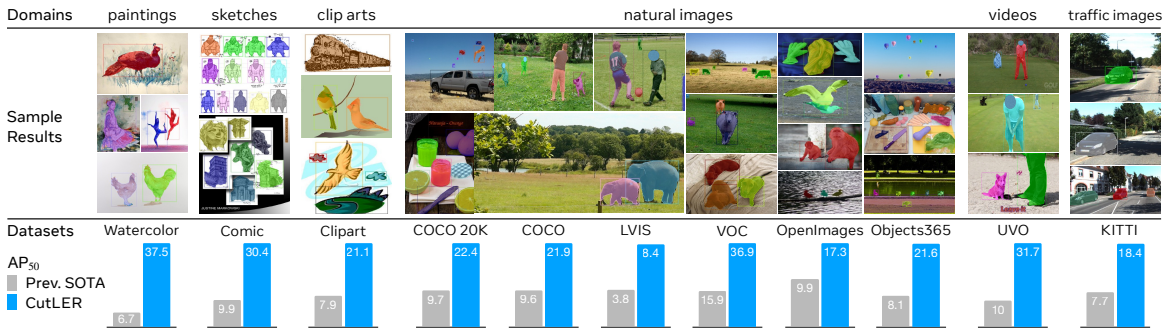


Figure 25: Zero-shot unsupervised object detection and instance segmentation using our CutLER model, which is trained without human supervision. We evaluate the model using the standard detection AP_{50}^{box}. CutLER gives a strong performance on a variety of benchmarks spanning diverse image domains - video frames, paintings, clip arts, complex scenes, etc. Compared to the previous state-of-the-art method, FreeSOLO [144] with a backbone of ResNet101, CutLER with a backbone of ResNet50 provides strong gains on all benchmarks, increasing performance by more than 2× on 10 of the 11 benchmarks. We evaluate [144] with its official code and checkpoint.

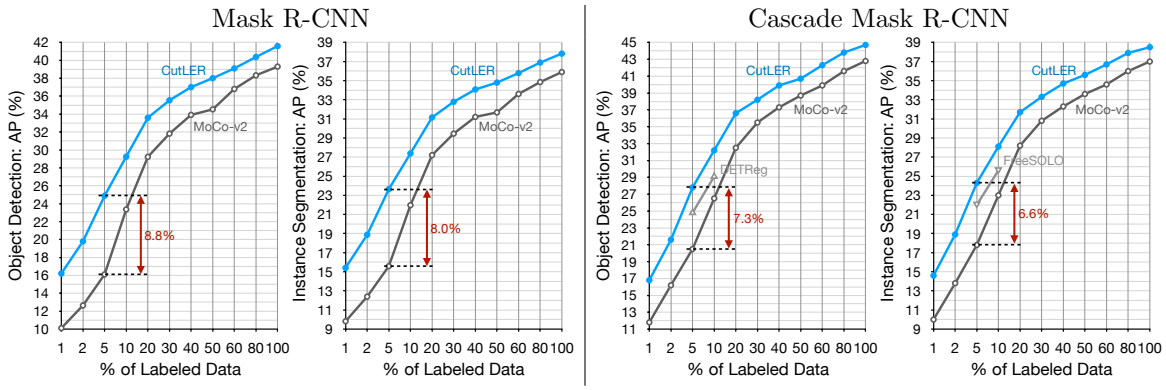


Figure 26: Fine-tuning on MS-COCO with various annotation ratios. We report results using Mask R-CNN and Cascade Mask R-CNN with a backbone of ResNet-50 as the detector.

segmentation datasets often annotate only a subset of the objects in the video, which makes Average Recall (AR) a valuable metric that does not penalize models for detecting novel objects not labeled in the dataset [18]. Therefore, we report both AR and AP for YouTubeVIS. Regarding DAVIS, we use the official unsupervised learning metrics $\mathcal{J}$, $\mathcal{F}$, and $\mathcal{J}\&\mathcal{F}$. All these metrics assess the performance of unsupervised video instance segmentation in a class-agnostic manner.

Detailed comparisons on YouTubeVIS. Table. 16 presents a summary of the results for unsupervised zero-shot video instance segmentation on the YouTubeVIS-2019 datasets. We compare our method's results with the previous state-of-the-art methods OCLR [57] and motion grouping [146]. We reproduce their results using their official code and checkpoints to ensure fairness.

Although OCLR is trained on synthetic videos, it relies on the off-the-shelf optical flow estimator RAFT [147] to compute optical flows for RGB sequences. It is worth noting that RAFT is pretrained on a combination of synthetic videos [148, 149] and human-annotated videos such as KITTI-2015 [150] and HD1K [151]. Our approach, VideoCutLER, despite not using any optical flow estimations like many previous works on unsupervised video segmentation, achieves over 10× higher AP_{50} and 18× higher AP than OCLR on YouTubeVIS-2019. Additionally, we achieve over 30% higher recall. Furthermore, unlike the previous state-of-the-art method OCLR, which exhibits poor performance in segmenting small objects (with 0.0% AP_S), our approach significantly outperforms it. Similar performance gains can be observed on YouTubeVIS-2021.

Label-Efficient and Fully-Supervised Learning Most prior approaches on self-supervised representation learning [33, 32, 152, 4] are limited to providing initializations only for the model backbones, with the remaining layers, such as Mask2Former's decoders, being randomly initialized. In contrast,

Table 16: Zero-shot unsupervised multi-instance video segmentation on YouTubeVIS-2019 and YouTubeVIS-2021. Zero-shot unsupervised multi-instance video segmentation on YouTubeVIS-2019 and YouTubeVIS-2021. We report the instance segmentation metrics (AP and AR) and training settings. *: reproduced MotionGroup [146] and OCLR [57] results with the official code and checkpoints. †: While OCLR is trained on synthetic data, the optical flow estimator it employs (RAFT [147]) is pretrained on both synthetic data [148, 149] and human-annotated data, such as KITTI-2015 [150] and HD1K [151]. Sup and flow denote human supervision and optical flow information, respectively. We evaluate results on YouTubeVIS's `train` splits in a class-agnostic manner.

Methods	Training settings			YouTubeVIS-2019						
	flow	videos	sup.	AP_{50}	AP_{75}	AP	AP_S	AP_M	AP_L	AR_{100}
MotionGroup*	✓	✓	✗	1.3	0.1	0.3	0.2	0.3	0.5	1.7
OCLR*	✓	✓	✗†	4.8	0.4	1.3	0.0	1.2	5.5	11.0
VideoCutLER	✗	✗	✗	50.3	23.2	25.1	5.3	20.6	36.7	41.2
vs. SOTA				**+45.5**	**+22.8**	**+23.8**	**+5.3**	**+19.4**	**+31.2**	**+30.2**

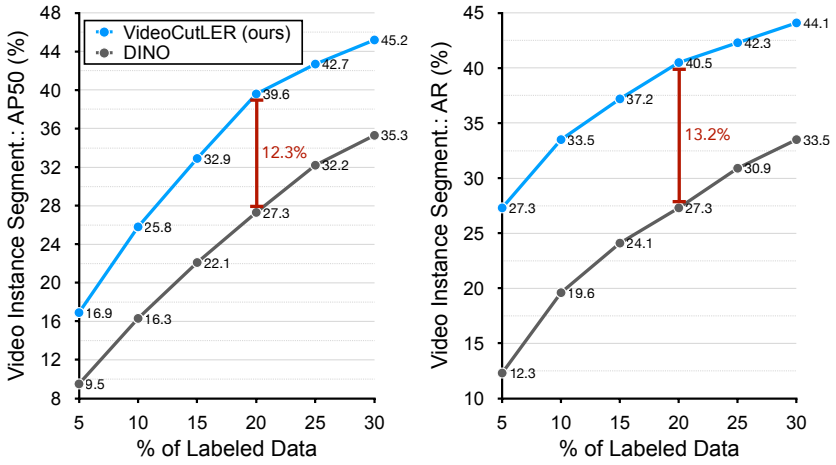


Figure 27: We fine-tune VideoCutLER for semi-supervised video instance segmentation on the YouTubeVIS-2019 dataset using different percentages of labeled training data. We evaluate the performance of our method by reporting the average precision and recall on the validation set of YouTubeVIS-2019. To establish a strong baseline, we use the self-supervised DINO [4] model and initialize the weights of VideoMask2Former with DINO. To ensure a fair comparison, both baselines and VideoCutLER are trained using the same schedule and recipe.

VideoCutLER takes a more comprehensive approach that allows all model weights to be pretrained, resulting in a stronger pretrained model better suited for supervised learning. As a result, as shown in Fig. 27, our method outperforms these prior works significantly, offering a strong pretrained model for fully-/semi-supervised learning tasks.

In Fig. 27, it can be observed that VideoCutLER consistently outperforms the strong baseline method DINO [4] across all label-efficient learning settings with varying proportions of labeled YouTubeVIS-2019 videos. The most significant performance gains are observed when 20% labeled data is provided, where VideoCutLER exceeds DINO by over 12% in video instance segmentation precision AP_{50} and 13.2% in terms of video instance segmentation recall.

4.3.3 Experimental Results: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models [20]

We show a comparison of ODISE to previous work on open-vocabulary semantic segmentation in Table 17. We evaluate mIoU on 5 semantic segmentation datasets: (a) A-150 with 150 common classes and (b) A-847 with all the 847 classes of ADE20K, (c) PC-59 with 59 common classes and (d) PC-459 with full 459 classes of Pascal Context, and (e) the classic Pascal VOC dataset with 20

Figure 28: Qualitative Visualization on COCO (first 2 rows) and ADE20K (last row) validation and test sets. To demonstrate open-vocabulary recognition capability, we merge category names of LVIS, COCO and ADE20K together and perform open-vocabulary inference with $\sim$1.5k classes directly. "Bamboo", "swimming pool", "conveyer belt", "chandelier", "booth", "stool", "column", "pool table", "bannister", *etc.*, are novel categories from LVIS/ADE20K that are not annotated in COCO. ODISE shows plausible open-vocabulary panoptic results. The supplement contains more visual results.

foreground classes and 1 background class (PAS-21). For a fair comparison to prior work, we train ODISE with either category or image caption labels. ODISE outperforms the existing state-of-the-art methods on open-vocabulary semantic segmentation by a large margin: by 7.6 mIoU on A-150, 4.7 mIoU on A-847, 4.8 mIoU on PC-459 with caption supervision; and by 6.2 mIoU on A-150, 4.5 mIoU on PC-459 with category label supervision, versus the next best method. Notably, it achieves this despite using supervision from panoptic mask annotations, which is noted to be suboptimal for semantic segmentation. The qualitative results can be found in Fig. 28.

Table 17: Open-vocabulary semantic segmentation performance.

Method	Training Dataset	Supervision			mIoU					
		label	mask	caption	A-847	PC-459	A-150	PC-59	PAS-21	COCO
SPNet	Pascal VOC	✓	✓		-	-	-	24.3	18.3	-
ZS3Net	Pascal VOC	✓	✓		-	-	-	19.4	38.3	-
LSeg	Pascal VOC	✓	✓		-	-	-	-	47.4	-
SimBaseline	COCO	✓	✓		-	-	15.3	-	74.5	-
ZegFormer	COCO	✓	✓		-	-	16.4	-	73.3	-
LSeg+	COCO	✓	✓		3.8	7.8	18.0	46.5	-	55.1
MaskCLIP	COCO	✓	✓		8.2	10.0	23.7	45.9	-	-
ODISE (Ours)	COCO	✓	✓		**11.1**	**14.5**	**29.9**	**57.3**	**84.6**	**65.2**
GroupViT	GCC+YFCC			✓	4.3	4.9	10.6	25.9	50.7	21.1
OpenSeg	COCO		✓	✓	6.3	9.0	21.1	42.1	-	36.1
ODISE (Ours)	COCO		✓	✓	**11.0**	**13.8**	**28.7**	**55.3**	**82.7**	**52.4**

Table 18: Compositional action recognition over 174 categories.

(a) Compositional action recognition with **ground-truths**.

model	split	top-1	top-5
STIN	Shuffled	54.0	79.6
STIN	Comp.	47.1	75.2
STIN + OIE	Comp.	51.3	79.3
STIN + OIE + NL	Comp.	51.4	79.3
I3D	Shuffled	61.7	83.5
I3D	Comp.	46.8	72.2
STRG	Comp.	49.2	74.9
I3D + STIN + OIE + NL	Comp.	54.6	79.4
I3D, STIN + OIE + NL	Comp.	58.1	83.2

(b) Compositional action recognition with **detections**.

model	split	top-1	top-5
STIN	Comp.	37.1	62.6
STIN + OIE	Comp.	40.0	66.2
STIN + OIE + NL	Comp.	39.8	66.0
I3D	Comp.	46.8	72.2
STRG	Comp.	49.2	74.9
I3D + STIN + OIE + NL	Comp.	49.3	73.2
I3D, STIN + OIE + NL	Comp.	51.8	77.2
STRG, STIN + OIE + NL	Comp.	54.9	80.4

Table 19: We show results with (a) ground-truth bounding boxes and (b) object detection boxes.

(a) Few-shot compositional setting with **ground-truths**.

model	*base* top-1	*base* top-5	*few-shot* 5-shot	*few-shot* 10-shot
STIN	65.7	89.1	24.5	30.3
STIN + OIE	69.5	91.4	25.8	32.9
STIN + OIE + NL	70.2	91.4	27.7	33.5
I3D	73.6	92.2	21.8	26.7
STRG	75.4	92.7	24.8	29.9
I3D + STIN + OIE + NL	80.6	95.2	28.1	33.6
I3D, STIN + OIE + NL	81.1	96.0	34.0	40.6

(b) Few-shot compositional setting with **detections**.

model	*base* top-1	*base* top-5	*few-shot* 5-shot	*few-shot* 10-shot
STIN	54.0	78.9	14.2	19.0
STIN + OIE	58.2	82.6	16.3	20.8
STIN + OIE + NL	58.2	82.6	17.7	20.7
I3D	73.6	92.2	21.8	26.7
STRG	75.4	92.7	24.8	29.9
I3D + STIN + OIE + NL	76.8	93.3	23.7	27.0
I3D, STIN + OIE + NL	76.1	92.7	27.3	32.6
STRG, STIN + OIE + NL	78.1	94.5	29.1	34.6

4.4 Few-Shot Video Classification

4.4.1 Experimental Results: Something-Else: Compositional Action Recognition with Spatial-Temporal Interaction Networks [21]

Compositional Action Recognition. In our experiments in Table 18, we first experiment with the Spatial-Temporal Interaction Network (STIN) model using ground-truth object bounding boxes and observe a significant performance drop in the I3D baseline from a shuffled setting to a compositional setting. The STIN model outperforms I3D in the compositional split, and applying the Object Identity Embedding (OIE) further improves the model's performance. Combining the STIN model with the I3D baseline yields a substantial improvement over the appearance-only model. We then build our model on object bounding boxes obtained via object detection and tracking, observing that OIE still enhances the STIN model's performance. When combining the STIN model with I3D, we notice improvements in both joint learning and model ensemble approaches. The better results achieved with ensemble combining might be due to the differing pace of convergence between the two models during training, leading to optimization difficulties for joint training.

Few-shot Compositional Action Recognition. In the few-shot compositional action recognition task as shown in Table 19, we train our model on 88 base categories and finetune it on 86 novel categories. We evaluate the model on over 50k videos to assess its generalizability. Instead of using the n-way, k-shot setting, we directly finetune our model with all the novel categories. Our full model (STIN+OIE+NL) outperforms the I3D model by almost 6% in both 5-shot and 10-shot learning settings, demonstrating that our model generalizes better than I3D, which tends to overfit to object appearance. Individually and cooperatively, the OIE and non-local block improve few-shot performance. When combined with I3D through model ensemble, we achieve 12.2% improvement on 5-shot and 13.9% on 10-shot settings. Using object detection boxes, our model generalizes better in the compositional setting. Combining our model with the I3D model, joint learning yields a 1.9% improvement and model ensemble yields a 5.5% improvement in the 5-shot setting, with similar improvement in the 10-shot setting. By replacing the I3D base network with STRG, our method (STRG, STIN + OIE + NL) still provides substantial improvement over STRG in both 5-shot and 10-shot settings.

Approved for Public Release; Distribution Unlimited.

Table 20: Video action recognition accuracy comparison with other unsupervised representation methods on **UCF101**. We compare with methods using RGB modality. We mainly compare with models that has **similar parameters** as ours. K-400: Kinetics-400.

Method	Backbone(#Param)	Dataset	UCF101
3D-RotNet [153]	3D-ResNet18-full (33.6M)	K-400	62.9
3D-ST-Puzzle [154]	3D-ResNet18-full (33.6M)	K-400	63.9
SpeedNet [155]	I3D (12.1M)	K-400	66.7
DPC [156]	3D-ResNet18 (14.2M)	K-400	68.2
DPC [156]	3D-ResNet34	K-400	75.7
Video-Pace [157]	R(2+1)D (33.3M)	K-400	77.1
CBT [158]	S3D	K-400	79.5
MemDPC [159]	R-2D3D (32.4M)	K-400	78.1
Temporal-ssl [160]	R(2+1)D (33.3M)	K-400	81.6
VTHCL [161]	3D-ResNet50 (31.7M)	K-400	82.1
Ours	ResNet-18 (**11.69**M)	R2V2	76.8
Ours	ResNet-50 (**25.56**M)	K-400	81.6
Ours	ResNet-50 (**25.56**M)	R2V2	**82.1**

4.4.2 Experimental Results: Contrastive Learning of Image Representations with Cross-Video Cycle-Consistency [22]

The learned feature representation is evaluated on video action recognition using the UCF101 dataset, which has 13,320 videos from 101 action categories (Table 20). A ResNet-18 backbone is used, with video representation obtained by averaging the frame representation from the video. The results show that our method achieves 76.8% top-1 accuracy on UCF101, surpassing the previous best result of 68.2%. Our method is not directly comparable to methods using large 3D ConvNets, as we use 2D ConvNets with fewer parameters and less FLOPs. However, our representation can solve multiple downstream tasks, while the previous 3D ConvNets are designed only for action recognition. Using ResNet-50 as the backbone, our method achieves 82.1%, surpassing other methods with larger 3D ConvNets models. Additionally, we provide a final model using our method to train with Kinetics-400 and ResNet-50, validating the effectiveness and transferability of our approach across datasets.

4.4.3 Experimental Results: Semi-Supervised Action Recognition with Temporal Contrastive Learning [23]

Tables 21 and 22 compare performance on different datasets including Mini-Something-V2 [162], Jester [163], and Kinetics-400 [164], in terms of average top-1 clip accuracy and standard deviation over 3 random trials.

Table 21: **Performance Comparison in Mini-Something-V2.** Numbers show average Top-1 accuracy values with standard deviations over 3 random trials for different percentages of labeled data. TCL significantly outperforms all the compared methods in both cases.

	ResNet-18			ResNet-50		
Approach	1%	5%	10%	1%	5%	10%
Supervised (8f)	5.98±0.68	17.26±1.17	24.67±0.68	5.69±0.51	16.68±0.25	25.92±0.53
Pseudo-Label [165] (ICMLW'13)	6.46±0.32	18.76±0.77	25.67±0.45	6.66±0.89	18.77±1.18	28.85±0.91
Mean Teacher [166] (NeurIPS'17)	7.33±1.13	20.23±1.59	30.15±0.42	6.82±0.18	21.80±1.54	32.12±2.37
S4L [167] (ICCV'19)	7.18±0.97	18.58±1.05	26.04±1.89	6.87±1.29	17.73±0.26	27.84±0.75
MixMatch [168] (NeurIPS'19)	7.45±1.01	18.63±0.99	25.78±1.01	6.48±0.83	17.77±0.12	27.03±1.66
FixMatch [43] (NeurIPS'20)	6.04±0.44	21.67±0.18	33.38±1.58	6.54±0.71	25.34±2.03	37.44±1.31
TCL (Ours)	7.79±0.57	29.81±0.77	38.61±0.91	7.54±0.32	27.22±1.86	40.70±0.42
TCL w/ Finetuning	8.65±0.76	30.55±1.36	40.06±1.14	8.56±0.31	28.84±1.22	41.68±0.56
TCL w/ Pretraining & Finetuning	9.91±1.84	30.97±0.07	41.55±0.47	9.19±0.43	29.85±1.76	41.33±1.07

Mini-Something-V2. Table 21 shows the performance comparison with both ResNet-18 (left half) and ResNet-50 (right half) backbones on Mini-Something-V2. TCL outperforms the video extensions of all the semi-supervised image-domain baselines for all three percentages of labeled training data. The improvement is especially prominent for low capacity model (ResNet-18) and low data (only

Approved for Public Release; Distribution Unlimited.

Table 22: **Performance Comparison on Jester and Kinetics-400**. Numbers show the top-1 accuracy values using ResNet-18 on both datasets. Our approach achieves the best performance across different percentages of labeled data.

	Jester			Kinetics-400	
Approach	1%	5%	10%	1%	5%
Supervised (8f)	52.55±4.36	85.22±0.61	90.45±0.33	6.17±0.32	20.50±0.23
Pseudo-Label [165] (ICMLW'13)	57.99±3.70	87.47±0.64	90.96±0.48	6.32±0.19	20.81±0.86
Mean Teacher [166] (NeurIPS'17)	56.68±1.46	88.80±0.44	92.07±0.03	6.80±0.42	22.98±0.43
S4L [167] (ICCV'19)	64.98±2.70	87.23±0.15	90.81±0.32	6.32±0.38	23.33±0.89
MixMatch [168] (NeurIPS'19)	58.46±3.26	89.09±0.21	92.06±0.46	6.97±0.48	21.89±0.22
FixMatch [43] (NeurIPS'20)	61.50±0.77	90.20±0.35	92.62±0.60	6.38±0.38	25.65±0.28
TCL (Ours)	75.21±4.48	93.29±0.24	94.64±0.21	7.69±0.21	30.28±0.13
TCL w/ Finetuning	77.25±4.02	93.53±0.15	94.74±0.25	8.45±0.25	31.50±0.23
TCL w/ Pretraining & Finetuning	82.55±1.94	93.73±0.25	94.93±0.02	11.56±0.22	31.91±0.46

1% and 5% data with labels) regime. Notably, our approach outperforms the most recent approach, FixMatch by 1.75% while training with only 1% labeled data. The improvement is **8.14%** for the case when 5% data is labeled. These improvements clearly show that our approach is able to leverage the temporal information more effectively compared to FixMatch that focuses on only spatial image augmentations.

Jester. Our approach also surpasses the performance of existing semi-supervised approaches in Jester as shown in Table 22 (left). In particular, TCL achieves 10.23% absolute improvement compared to S4L (the next best) in very low labeled-data regime (1% only). Adding finetuning and self-supervised pretraining further increases this difference to 17.57%. Furthermore, TCL with pretraining and finetuning achieves a top-1 accuracy of 94.93% using 10% labeled data which is only 0.32% lower than the fully supervised baseline trained using all the labels (95.25%).

Kinetics-400. Table 22 (right) summarizes the results on Kinetics-400, which is one of the widely used action recognition datasets consisting of 240K videos across 400 classes. TCL outperforms FixMatch by a margin of 1.31% and 4.63% on 1% and 5% scenarios respectively, showing the superiority of our approach on large scale datasets. The top-1 accuracy achieved using TCL with finetuning and pretraining is almost twice better than the supervised approach when only 1% of the labeled data is used. The results also show that off-the-shelf extensions of sophisticated state-of-the-art semi-supervised image classification methods offer little benefit to action classification on videos.

4.4.4 Experimental Results: Weakly-Supervised Action Localization with Expectation-Maximization Multi-Instance Learning [24]

Results on THUMOS14 Dataset and ActivityNet1.2 Dataset. We compare our model's results on the THUMOS14 dataset with state-of-the-art results in Table 23. Compared to the best result among the four recent models [181, 183, 184, 182], we get 3% significant improvement at mAP@0.5. Our model also shows more significant improvement at high threshold metrics tIoU=0.6 and tIoU=0.7, which implies our action proposals are more complete. Our model also outperforms previously published models on ActivityNet1.2 dataset and gets the value of 37.4% in mAP@0.5.

4.5 Few-Shot Machine Translation

4.5.1 Experimental Results: Learned Incremental Representations for Parsing [66]

In Table 24, we present an F1 score comparison that highlights the behavior of different syntactic representations with different choices of encoder. When directly predicting either per-span label probabilities following the span classification approach of [185], or actions in the attach-juxtapose transition system, failing to include bidirectional layers on top of a unidirectional GPT-2 incurs a strong accuracy penalty. This is despite the fact that both systems can discard speculative attachment decisions. In the case of the chart parser with representations that consist of label probabilities for each span, adding an additional word can cause a switch to a new analysis by way of the CKY decoding procedure. In the case of the attach-juxtapose parser, the same can be achieved via the use of beam search.

Approved for Public Release; Distribution Unlimited.

Table 23: Our EM-MIL detection results on THUMOS14 in percentage mAP at different tIoU thresholds α are reported. The top half shows fully-supervised methods while the bottom half shows weakly-supervised ones including ours.

Supervision	Models	α 0.1	0.2	0.3	0.4	0.5	0.6	0.7
Fully-Supervised	CDC [169]	-	-	40.1	29.4	23.3	13.1	7.9
	Dai et al. [170]	-	-	-	33.3	25.6	15.9	9.0
	R-C3D [171]	54.5	51.5	44.8	35.6	28.9	-	-
	Gao et al. [172]	-	-	50.1	41.3	31.0	19.1	9.9
	SSN [173]	66.0	59.4	51.9	41.0	29.8	19.6	10.7
	TAL-Net [174]	59.8	57.1	53.2	48.5	42.8	33.8	20.8
	Alwassel et al. [175]	-	-	51.8	42.4	30.8	20.2	11.1
	BSN [176]	-	-	53.5	45.0	36.9	28.4	20.0
Weakly-Supervised	Sun et al. [177]	-	-	8.5	-	4.4	-	-
	Hide [178]	36.4	27.8	19.5	12.7	6.8	-	-
	UntrimmedNet [63]	44.4	37.7	28.2	21.1	13.7	-	-
	STPN [62]	52.0	44.7	35.5	25.8	16.9	9.9	4.3
	Autoloc [179]	-	-	35.8	29.0	21.2	13.4	5.8
	W-TALC [180]	55.2	49.6	40.1	31.1	22.8	-	7.6
	Liu et al. [181]	-	-	37.0	30.9	23.9	13.9	7.1
	Yu et al. [182]	-	-	39.5	-	24.5	-	7.1
	3C-Net [183]	59.1	53.5	44.2	34.1	26.6	-	8.1
	Nguyen et al. [184]	**64.2**	**59.5**	**49.1**	**38.4**	27.5	17.3	8.6
	EM-MIL (ours)	59.1	52.7	45.5	36.8	**30.5**	**22.7**	**16.4**

Table 24: F1 on the WSJ test set for parsers using different syntactic representations and pre-trained encoders. Our representation results in an F1 score that is close to the maximum achievable with the same pre-trained GPT-2-medium model. †Results based on integrating GPT-2 with publicly available code for prior methods, with and without bidirectional layers on top of GPT-2.

	Encoder Type		
	Bi ($\leftrightarrow$)		Uni ($\rightarrow$)
Representation	BERT	GPT-2	GPT-2
Span Classification	95.59	**95.10**†	93.95†
Attach-Juxtapose	**95.79**	94.53†	87.66†
Learned	95.55	–	**94.97**
(This work)			

Nevertheless, incrementally predicting either of these representations fails to leverage the full power of the pre-trained encoder.

The choice of GPT-2 rather than a stronger bidirectional model has a large effect on the performance on the Penn Treebank. To give a more accurate comparison with other models, Table 24 also shows F1 scores for models based on BERT, with the recognition that no model with such a deeply bidirectional encoder can truly be referred to as incremental. Our approach of inducing learned representations with vector quantization also performs well in this setting, validating the method.

Even higher scores are achievable by using stronger pre-trained models, different forms of bidirectional processing, and additional supervision in the form of dependency trees; [186] combine all of these elements to achieve 96.38 F1. However, many of these techniques are either orthogonal to our work, or they cannot be borrowed into an incremental setting due to their focus on deeply bidirectional neural processing.

Approved for Public Release; Distribution Unlimited.

4.5.2 Discussions on Incremental Learning

Having achieved high F1 scores, we must next demonstrate that our representation is, in fact, incremental. An incremental representation has meaningful syntactic information in each of its prefixes, and we can probe this by running our read-out network after each word in the sentence, as shown in Figure 29. The resulting trees involve mostly local changes from word to word, which shows that important information is not being deferred to the very end of a sentence.

It should be noted that our read-out network was *never trained* on anything but complete sentences. Applying it to fragments will produce individual trees that may not be representative of the ambiguity present in the underlying representation. For example, after the word "on" the read-out network outputs a prepositional phrase that initially appears to attach to the verb. Depending on the label chosen for the next word, however, the final attachment can be to either the verb or the noun phrase.

Nevertheless, this approach allows us to probe the degree to which the representation encodes syntactic decisions immediately, versus deferring them to some later point. For each span in the final tree, we can walk backwards through the partial readouts to find the furthest point when the span still appears in a readout; we call this the point in time that a span is *finalized*. In Figure 29, the noun phrase "The Council" is finalized after the word "Council," and the verb phrase is finalized after the word "approved." For the purposes of identifying whether a span is the same across two different trees, we assume that a span is uniquely identified by its label, its starting position, and the position of the last word in the leftmost child of the span (or the position of the single word in the span, if it only covers one word). The last of these we also refer to as the *split point* of the span.

4.5.3 Experimental Results: Low-Resource Machine Translation Training Curriculum Fit for Low-Resource Languages [25]

Time-Constrained Setting In Table 25, we explore BLEU scores for several NMT training configurations. First, we provide performances for time-constrained baselines for supervised MT (i.e., $BT + MT_s$) and unsupervised MT, which is LM pre-training of XLM with the MLM objective, followed by back-translation BT and auto-encoding AE objective i.e., $MLM + (BT + AE)$. This results in poor BLEU scores similar to what has been reported in previous works [5]. For unsupervised MT, among the three languages, we can see that Somali, which is typologically closer to English than Gujarati or Kazakh in terms of lower syntactic distance, higher character and token overlap, and more shared typological features with English has the highest unsupervised MT performances among the three languages. This reiterates previous findings that unsupervised MT works better for typologically closer languages.

Next, while adhering to the time constraints specified in, we explore starting from pre-trained LM with the MLM objective followed by MT training with parallel data (i.e., $MLM + (BT + MT_s)$). Our experiments show that training a NMT model starting from pre-trained LMs (even bilingual LMs) can already improve performance over training from scratch (i.e., $BT + MT_s$). In addition to starting with pre-trained LMs, if we also pre-train our LMs with unsupervised MT objective before training with MT objective (i.e., $MLM + (BT{+}AE) + (BT{+}MT_s)$), we can further improve performance over just starting from pre-trained LMs (also shown in Tables 26 and 27).

We also observe that pre-training LM with augmented data obtained with code-switching i.e., MLM_{cs} before training with unsupervised MT objective (i.e., $MLM_{cs} + (BT{+}AE) + (BT{+}MT_s)$) improves performance further over just pre-training LM with monolingual data (i.e., $MLM + (BT{+}AE) + (BT{+}MT_s)$) – also shown in Tables 28 and 29.

Further, we also observe in Table 25 that training with comparable data only improves performance when the size of the mined comparable data is bigger than the available parallel data. In our case, only Gujarati has comparable data that is larger than its available parallel data. Training on this comparable data (i.e., $MLM_{cs} + (BT{+}AE) + (BT{+}MT_c)$) significantly improves performance further for Gujarati. To show significance, a paired samples t-test was performed to compare BLEU scores of Gujarati models trained with parallel sentences and comparable sentences. There was a significant difference in BLEU scores between models trained with parallel sentences (M = 6.2, SD = 3.9) and models trained with comparable sentences (M = 9.2, SD = 4.5); T-statistic = 15.42, $p = 4.43 \times 10^{-8}$.

Overall, under time-constrained setting, we observe that when combined together: code-switching LM pre-training (MLM_{cs}), pre-training with unsupervised NMT objective ($BT{+}AE$), and training on larger comparable data ($BT{+}MT_c$), they outperform individual employments of these objectives

suggesting that pre-training on the related tasks of *MLM* and the unsupervised objectives *BT* and *AE* are pivotal to *MT* training.

Table 25: BLEU scores for previous supervised and unsupervised results from [1][5], [2][6] and our models. Test and validation sets are from WMT19 for Gujarati and Kazakh and from [187] for Somali. *MLM*, *AE*, *BT* and *MT* stand for *MLM*, Auto-Encoding loss, Back Translation loss and Machine Translation loss, respectively. MT_s and MT_c utilize human-labeled parallel data and mined comparable data, respectively. MLM_{cs} utilizes both code-switched and original forms of monolingual data. Best results overall are **bolded** while best results in each section are underlined. All our models here use 1 (32GB) GPU. Parentheses in training objectives refer to two simultaneous losses while the ones separated with "+" are used successively.

Name	Supervision	en-gu	gu-en	en-kk	kk-en	en-so	so-en
Baseline (Time-Constrained)							
$(BT + MT_s)$	Supervised	3.7	1.2	1.9	3.1	20.1	23.1
$MLM + (BT + AE)$	Unsupervised	1.7	1.2	1.0	1.3	8.1	7.4
Ours (Time-Constrained)							
$MLM + (BT + MT_s)$	Supervised	6.8	3.0	5.6	8.4	23.1	29.4
$MLM + (BT + AE) + (BT + MT_s)$	Semi-supervised	11.6	6.9	7.2	10.7	23.5	29.2
$MLM_{cs} + (BT + AE) + (BT + MT_s)$	Semi-supervised, Code-switching	13.1	7.7	7.9	10.7	23.6	29.3
$MLM_{cs} + (BT + AE) + (BT + MT_c)$	Weakly-supervised, Code-switching	15.0	11.8	5.5	7.5	14.7	13.9
Ours (Best of Time-Constrained, ↑ batch size: 4k tokens per batch)							
$MLM_{cs} + (BT + AE) + (BT + MT)$	Semi/weakly-supervised, Code-switching	16.3	12.8	9.7	12.7	23.7	29.7
Ours (Best of Time-Constrained, ↑ batch size: 4k tokens per batch, ↑ time: MT trained to convergence)							
$MLM_{cs} + (BT + AE) + (BT + MT)$	Semi/weakly-supervised, Code-switching	**17.3**	**13.6**	**11.5**	**14.2**	**23.8**	**30.7**
Large Multilingual Models (MT trained to convergence)							
mBART25[1]	Supervised	0.1	0.3	2.5	7.4	-	-
XLM-R[2] + (BT+MT)	Supervised	13.2	7.8	6.0	8.4	22.5	27.5

Table 26: Comparison of effect of using **unsupervised MT training** $(BT + AE)$ with **supervised** data.

UMT	en-gu	gu-en	en-kk	kk-en	en-so	so-en
N	6.8	3.0	5.6	8.4	23.1	29.4
Y	11.6	6.9	7.2	10.7	23.5	29.2

Table 27: Comparison of effect of using **unsupervised MT training** $(BT + AE)$ with **comparable** data.

UMT	en-gu	gu-en	en-kk	kk-en	en-so	so-en
N	6.8	3.0	5.6	8.4	23.1	29.4
Y	14.2	10.4	5.2	6.9	13.8	13.7

Table 28: Comparison of effect of using **code switch** (MLM_{cs}) with **supervised** data.

CS	en-gu	gu-en	en-kk	kk-en	en-so	so-en
N	11.6	6.9	7.2	10.7	23.5	29.2
Y	13.1	7.7	7.9	10.7	23.6	29.3

Table 29: Comparison of effect of using **code switch** (MLM_{cs}) with **comparable** data.

CS	en-gu	gu-en	en-kk	kk-en	en-so	so-en
N	14.2	10.4	5.2	6.9	13.8	13.7
Y	15.0	11.8	5.5	7.5	14.7	13.9

4.5.4 Experimental Results: RL4F: Generating Natural Language Feedback with Reinforcement Learning for Repairing Model Outputs [26]

Our main results for Interscript and topic-based summarization are provided in Table 30. Given the free-from nature of the outputs, we evaluate planning and summarization tasks using text similarity

Table 30: Results for action sequence generation with Interscript [7] and topic-based summarization by [8]. We evaluate the performance of different sources for natural language critiques in steering LM_{task} to improve its predictions. Best scores in each column are made bold. We compare our method, RL4F, to three strong baselines and human-generated critiques. Self-Refinement prompts GPT-3 to self-repair its answer. MemPrompt uses memory to store human-generated critiques to previous outputs [78]. ROUGE and BERTScore are out of 100 while BLEURT can be negative or positive and should be used in comparing different methods.

	Interscript			Topic-Based Summarization		
Source of Critiques	**BLEURT ↑**	**BERTScore↑**	**R1/R2/RL↑**	**BLEURT↑**	**BERTScore↑**	**R1/R2/RL↑**
Self-Refinement	-1.07	86.97	15.8 / 0.9 / 15.5	0.09	93.1	54.3/46.0/50.9
SUPERVISED	-1.02	86.99	19.4 / 0.5 / 18.5	0.06	92.9	53.2/46.4/50.7
MemPrompt	-1.18	**87.45**	16.9 /**1.9** / 16.7	0.09	91.9	48.8 /40.4/45.6
RL4F (Ours)	**-0.92**	87.23	**22.1** / 0.9 / **21.3**	**0.10**	**93.6**	**55.1/48.2/52.6**
With gold feedback	*-0.69*	*89.56*	*40.7 / 6.8 / 39.1*	*0.22*	*94.2*	*58.3/50.3/55.8*

metrics to capture semantic and lexical similarities. We utilize learned metrics such as BLEURT and BERTScore [188] along with ROUGE [189]. We compare the performance achieved by using different sources of critiques to that of human-written critiques. Across all metrics, RL4F yields one of the closest outcomes to human-written critiques.

4.6 Data-Efficient Fine-Tuning for Image Classification

4.6.1 Experimental Results: A Top-Down Approach to Personalizing Large Models

We tested on VTAB which contains 19 tasks, each with 1000 training images. Preliminary results are shown in Table 31-32.

Table 31: Results on VTAB-1k with ViT backbone (Part 1).

	#Params	CIFAR100	Caltech	EuroSAT	Flowers	CLEVR-distance	CLEVR-count	DTD	Pet
Finetune	87M	48.0	78.4	95.4	76.0	53.7	62.8	57.3	87.9
VPT	0.9M	65.5	90.6	94.9	88.7	52.8	53.9	68.1	**89.1**
ToT	0.7M	**74.4**	**92.1**	**96.2**	**93.1**	**57.0**	**66.2**	**68.5**	89.0

Table 32: Results on VTAB-1k with ViT backbone (Part 2).

	#Params	Camelyon	SUN	Dmlab	SVHN	NORB-azi	NORB-ele	dSprites-pos	dSprites-ori
Finetune	87M	**86.7**	18.9	34.4	86.2	15.5	37.9	67.2	44.6
VPT	0.9M	83.0	40.7	**45.1**	83.3	26.2	29.5	68.0	-
ToT	0.7M	83.2	**42.6**	44.3	**87.7**	**28.6**	**42.9**	**86.6**	**48.0**

4.6.2 Experimental Results: ZipIt! Merging Models from Different Tasks *without Training*

We devise two types of experiments to benchmark disjoint task model merging: (1) Merging models trained on disjoint category splits of the same dataset (i.e., *same dataset and different label sets*), and (2) merging models trained on completely different datasets (i.e., *different datasets and label sets*).

Results on disjoint classes of same dataset. Table 33 and Table 34 present results when using ZipIt! to merge ResNet [190] models of varying sizes on the CIFAR100 [191] and ImageNet1k [122] datasets. Notably, ZipIt! significantly outperforms prior work, with partial zipping—denoted by fractional subscripts which refer to the number of layers merged between models—nearly matching the ensemble accuracy!

Results on disjoint datasets. Notably, ZipIt! can also be used to merge models trained on *completely different datasets*, each with *different label sets*. Table 35 showcase ZipIt!'s propensity to capture information across disjoint models into *a single one* while maintaining strong performance across all datasets.

Partial zipping: trading speed for accuracy. In Fig. 30, we plot the average per task accuracy by the number of layers zipped in ResNet-20×8 for CIFAR-100 (50+50) and ResNet-50 for ImageNet-1k (200+200). Overall, we find partial zipping to be a simple yet effective technique to add capacity back

Approved for Public Release; Distribution Unlimited.

Table 33: ZipIt! vs. baselines on combining a model trained on half the classes (Task A) with one trained on the other half (Task B) *without extra training*. We report both joint (10/100-way) and per-task (5/50-way) accuracy. All models are trained to predict among a set of labels encoded by a frozen off-the-shelf CLIP[29] model. ZipIt! *significantly* outperforms its baseline and closes in on the upper bound (ensemble accuracy).

(a) **CIFAR-10 (5+5).** Using ResNet-20 (4× width).

Method	FLOPs (G)	Joint Acc (%)	Per-Task (%) Task A	Task B	Avg
Model A	0.68	48.2±1.0	97.0±0.6	45.1±8.6	71.0±4.4
Model B	0.68	48.4±3.8	49.1±9.3	96.1±1.1	72.6±4.9
Git Re-Basin [192]	0.68	46.2±0.8	76.8±8.9	82.7±5.1	79.8±6.5
ZipIt!$_{20/20}$	0.68	**79.1**±1.1	**92.9**±1.0	**91.2**±1.4	**92.1**±1.0
Ensemble	1.37	87.4±2.6	97.0±0.6	96.1±1.1	96.6±0.4
ZipIt!$_{13/20}$	0.91	**83.8**±3.1	**95.1**±0.5	**94.1**±1.8	**94.6**±0.6

(b) **CIFAR-100 (50+50).** Using ResNet-20 (8× width).

Method	FLOPs (G)	Joint Acc (%)	Per-Task (%) Task A	Task B	Avg
Model A	2.72	41.6±0.3	82.9±0.7	24.8±0.4	53.9±0.5
Model B	2.72	41.6±0.2	25.1±1.2	82.8±0.2	54.0±0.6
Git Re-Basin [192]	2.72	40.9±0.2	57.3±1.5	56.7±0.7	57.0±0.8
ZipIt!$_{20/20}$	2.72	**54.9**±0.8	**68.2**±0.6	**67.9**±0.6	**68.0**±0.4
Ensemble	5.45	73.5±0.4	82.9±0.7	82.8±0.2	82.8±0.4
ZipIt!$_{13/20}$	3.63	**70.2**±0.4	**80.3**±0.8	**80.1**±0.8	**80.2**±0.6

Table 34: ImageNet-1k (200+200) Results. Merging ResNet-50 models trained from scratch on disjoint 200 category subsets (Task A and B) of ImageNet-1k. All models are trained to predict label indices as in standard classification. Prior work performs poorly, but ZipIt! makes this task feasible.

Method	FLOPs (G)	Joint Acc (%)	Per-Task (%) Task A	Task B	Avg
Model A	4.11	37.2±2.0	74.3±4.0	0.5±0.1	37.4±2.0
Model B	4.11	35.3±1.6	0.5±0.1	70.5±3.2	35.5±1.6
Git Re-Basin [192]	4.11	3.1±1.2	5.3±2.6	5.7±2.4	5.5±1.7
ZipIt!$_{50/50}$	4.11	**8.6**±4.7	**12.4**±5.9	**14.7**±7.8	**13.5**±6.6
Ensemble	8.22	63.3±4.9	74.3±4.0	70.5±3.2	72.4±2.5
ZipIt!$_{22/50}$	6.39	55.8±4.1	65.9±2.5	64.1±3.0	65.0±2.3
ZipIt!$_{10/50}$	7.43	**60.9**±4.1	**70.7**±3.0	**69.0**±2.9	**69.9**±1.9

to the merged model. For CIFAR-100, we can obtain near ensemble accuracies at a 1.5× speed-up. Similarly on a difficult setting like ImageNet, partial zipping is *necessary* to obtain any reasonable accuracy.

Model Scale. In Fig. 31, we test the effect of model scale directly by evaluating joint accuracy on our CIFAR-100 (50+50) setting with ResNet-20 models of increasing width. Here, we explicitly see that when the width of the models are too small for the task (e.g., < 4), ZipIt! and prior work perform identically (though both much better than Git Re-Basin). However, when the scale increases, ZipIt! trends toward the ensemble upper bound of 75%, while other work plateau at around 45%. This indicates that our method uses the extra capacity of these models effectively, much better than prior work.

Approved for Public Release; Distribution Unlimited.

Table 35: Merging pairs of differently initialized ResNet-50 models trained on *completely different datasets*: Stanford Dogs [193] (SD), Oxford Pets [194] (OP), CUB200 [195] (CUB), and NABirds [196] (NAB). We report average per-task accuracy over all pairs (2-way merging) and per-task accuracy for each head (4-way merging). [192] doesn't support models with different outputs.

	FLOPs	Per-Task (%)				
Method	(G)	SD	OP	CUB	NAB	Avg
Merging Pairs						
ZipIt!$_{49/50}$	4.11	**51.2**	**67.7**	**40.6**	**15.6**	**43.8**
Ensemble	8.22	72.7	83.2	71.0	77.2	76.0
ZipIt!$_{37/50}$	4.92	56.8	73.8	54.6	37.9	55.8
ZipIt!$_{22/50}$	6.39	**65.3**	**79.7**	**64.8**	**61.2**	**67.7**
Merging All 4						
ZipIt!$_{49/50}$	4.12	32.1	**55.3**	**14.7**	**6.9**	**27.3**
Ensemble	16.44	72.7	83.2	71.0	77.2	76.0
ZipIt!$_{37/50}$	6.5	39.9	66.4	44.3	24.6	43.8
ZipIt!$_{22/50}$	11.0	**58.2**	**78.5**	**58.6**	**55.1**	**62.6**

Approved for Public Release; Distribution Unlimited.

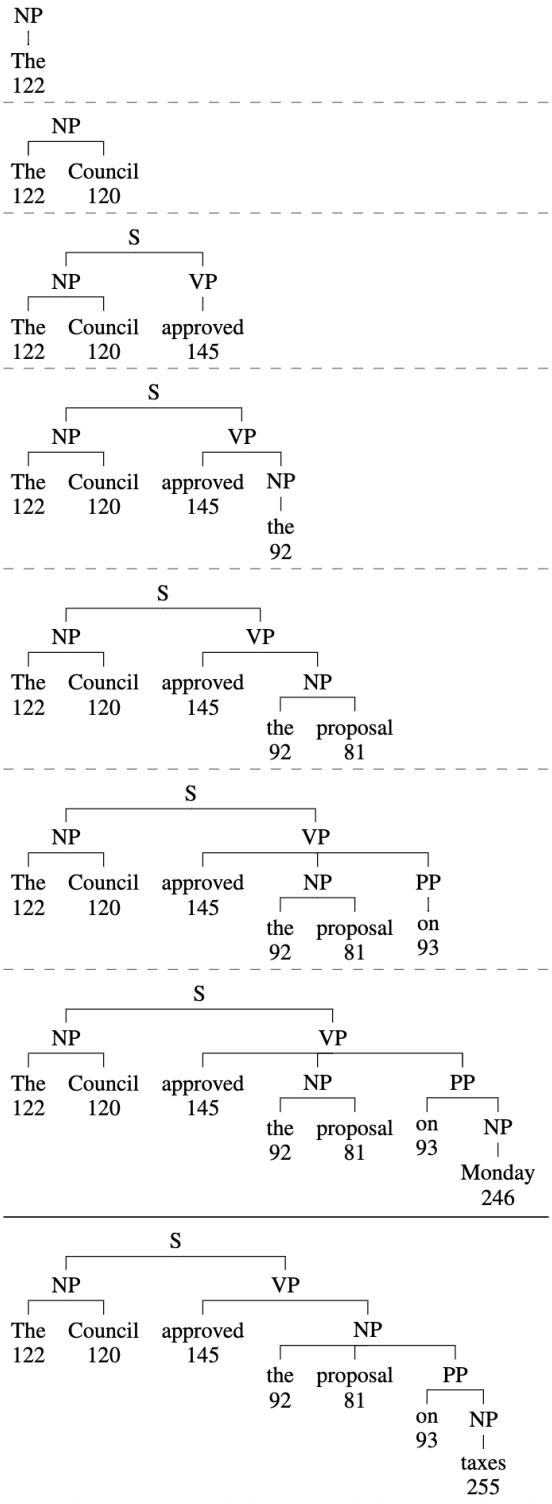
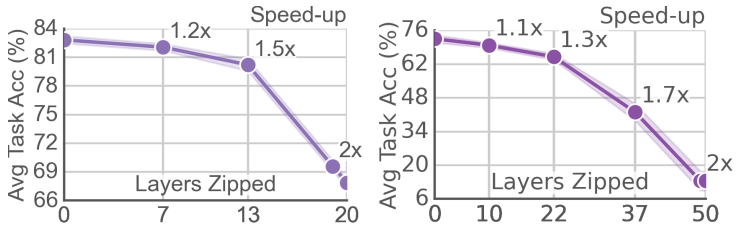


Figure 29: [Applying our read-out network to prefixes of the syntactic tag sequence. It demonstrates that syntactic decisions are committed to incrementally, and are not all deferred to the end of the sentence. Nevertheless, the same tag sequence may decode to different trees, depending on a future tag. Numbers below each word denote the discrete tags assigned by our model.



(a) **CIFAR-100 50+50.** (b) **ImageNet-1k 200+200.**

Figure 30: Varying Partial Zip. By leaving some layers unzipped, we can recover a significant amount of performance while still merging most of the model.

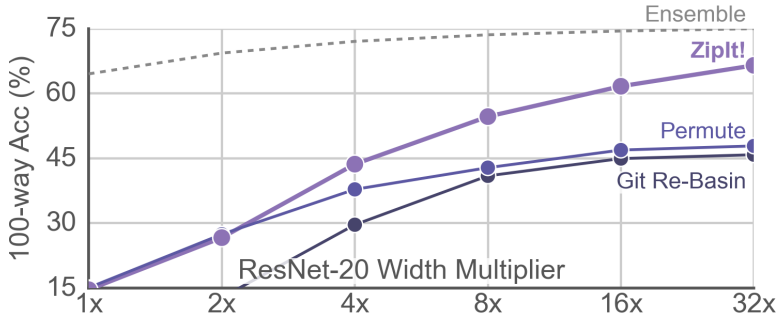


Figure 31: Model Scale. As we increase the width of the ResNet-20 models used for the CIFAR-100 (50+50) setting, ZipIt! makes effective use of that extra capacity, quickly approaching ensemble accuracy. Other methods only slightly benefit from the extra scale.

5 Conclusions

5.1 Zero-Shot Image Classification

We proposed a unified framework, DualCoOp, for two types of multi-label recognition with limited annotations. It utilizes the powerful vision-language pretraining from a large-scale dataset. By introducing a lightweight learnable overhead, it can quickly adapt to solve multi-label recognition after receiving a small amount of labels. In DualCoOp, we learned a pair of positive and negative prompts followed by the target class name as the linguistic input. Furthermore, to better aggregate visual region features for each class, we reformulate the original visual attention in the pretraining model as a class-specific region feature aggregation. We conducted extensive experiments for both partial-label MLR and Zero-Shot MLR across MS-COCO, VOC2007, and NUS-WIDE datasets showing the efficacy of our proposed approach over state-of-the-art methods.

We developed SENTRY, an algorithm for unsupervised domain adaptation (UDA) under simultaneous data and label distribution shift. Unlike prior work that suffers from error accumulation arising from unconstrained self-training, SENTRY first judges the reliability of a target instance based on its predictive consistency under a committee of random image transforms, and then selectively minimizes entropy (increasing confidence) on consistent instances, while maximizing entropy (reducing confidence) on inconsistent ones. We have shown that SENTRY significantly improves upon the stateof-the-art across 27/31 shifts from several UDA benchmarks

Despite its simplicity, PACMAC requires i) additional in-domain pretraining and ii) additional forward passes over masked target images (but no additional backpropagation), which increases its computational overhead. Further, PACMAC's effectiveness at identifying reliable instances varies across categories (see appx.). In our paper we only experimented with ViTs, and the generality of our approach across architectures is not established. Finally, our approach still lags behind methods that use supervised initializations, suggesting the need to rethink SSL with cross-domain transfer in mind.

5.2 Few-Shot Image Classification

CLUE addressed active domain adaptation, where the goal is to select target instances for labeling so as to generalize a trained source model to a new target domain. We show how existing active learning strategies based solely on uncertainty or diversity sampling are not effective for Active DA. We developed CLUE, a novel label acquisition strategy for active sampling under a domain shift, that performs uncertaintyweighted clustering to select diverse, informative target instances for labeling from dense regions of the feature space. We demonstrate CLUE's effectiveness over competing active learning and Active DA methods across learning settings and domain shifts, and comprehensively analyze its behavior.

We proposed a novel Minimax Entropy (MME) approach that adversarially optimizes an adaptive few-shot model for semi-supervised domain adaptation (SSDA). Our model consists of a feature encoding network, followed by a classification layer that computes the features' similarity to a set of estimated prototypes (representatives of each class). Adaptation is achieved by alternately maximizing the conditional entropy of unlabeled target data with respect to the classifier and minimizing it with respect to the feature encoder. We empirically demonstrated the superiority of our method over many baselines, including conventional feature alignment and few-shot methods, setting a new state of the art for SSDA.

We proposed a method for open-set semi-supervised learning (OSSL), where samples of novel categories are present in unlabeled data. To approach the task, we propose a novel framework, OpenMatch, which unifies one-vs-all classifiers and FixMatch. Our proposed objective, open-set soft consistency loss, is shown to be effective to detect outliers from unlabeled data, which allows FixMatch to work well in the OSSL setting. In addition, OpenMatch sometimes detects outliers unseen in unlabeled data better than a supervised model. We believe that our framework for OSSL will make label-efficient techniques more practical.

We have described a family of regularization-based approaches to few-shot class-incremental learning, drawing connections between incremental learning and the general multi-task and zero-shot learning literature. The proposed regularizers are extremely simple—they involve only one extra hyperparameter, require no additional training steps or model parameters, and are easy to understand and implement. Despite this simplicity, our approach enables ordinary classification architectures to

achieve state-of-the-art results on the doubly challenging few-shot incremental image classification across multiple datasets and problem formulations.

In the 2020s, vision Transformers, particularly hierarchical ones such as Swin Transformers, began to overtake ConvNets as the favored choice for generic vision backbones. The widely held belief is that vision Transformers are more accurate, efficient, and scalable than ConvNets. We developed ConvNeXts, a pure ConvNet model that can compete favorably with state-of-the-art hierarchical vision Transformers across multiple computer vision benchmarks, while retaining the simplicity and efficiency of standard ConvNets. In some ways, our observations are surprising while our ConvNeXt model itself is not completely new — many design choices have all been examined separately over the last decade, but not collectively. We hope that the new results in this study will challenge several widely held views and prompt people to rethink the importance of convolution in computer vision.

5.3 Few-Shot Object Detection

Object localization is a fundamental task in computer vision. We have shown that a simple yet effective cut-and-learn approach can achieve extraordinary performance on challenging object detection and instance segmentation tasks without needing to train with human annotations. As a zero-shot unsupervised detector, CutLER, trained solely on ImageNet, outperforms the detection performance of previous works by over 2.7x on 11 benchmarks across various domains.

We presented a simple unsupervised approach to segment multiple instances in a video. Our approach, VideoCutLER, does not require labels, and does not rely on motionbased learning signals like optical flow. In fact, VideoCutLER does not need real videos for training as we synthesize videos using natural images from the ImageNet-1K dataset. Despite being simpler, VideoCutLER outperforms models that use additional learning signals or video data, achieving 10× their performance on benchmarks like YouTubeVIS. Additionally, VideoCutLER serves as a strong selfsupervised pretrained model for supervised video instance segmentation. We hope that our approach enables both a wide range of applications in video recognition, as well as its simplicity enables easy future research.

We took the first step in leveraging the frozen internal representation of large-scale text-to-image diffusion models for downstream recognition tasks. ODISE shows the great potential of text-to-image generation models in openvocabulary segmentation tasks and establishes a new state of the art. This work demonstrates that text-to-image diffusion models are not only capable of generating plausible image but also of learning rich semantic representations. It opens up a new direction for how to effectively leverage the internal representation of text-to-image models for other tasks as well in the future.

5.4 Few-Shot Video Classification

Motivated by the appearance bias in current activity recognition models, we developed a new model for action recognition based on sparse semantically grounded subject-object graph representations. We validated our approach on novel compositional and few shot settings in the SomethingElse dataset; our model is trained with new constituent object grounding annotations Our STIN approach models the interaction dynamics of objects composed in an action and outperforms all baselines.

CycleContrast proposed a cross-video cycle-consistent contrastive learning objective to perform self-supervised learning for image representations. The proposed method could learn to explore cross-video relations to make not only the representations of images within the same video closer, but also representations from different videos with similar visual structures close, without using ground truth class labels or generating pseudo labels. We performed self-supervised representation learning on unlabeled R2V2 video dataset, and show that the learned image representation transfer well to multiple downstream tasks including visual tracking, image classification and action recognition. Extensive ablation studies are conducted and validate the effectiveness of our proposed method. We hope our approach can open up an opportunity to utilize cross-instance paired data for learning general image representations.

We developed a novel temporal contrastive learning framework for semi-supervised action recognition by maximizing the similarity between encoded representations of the same unlabeled video at two different speeds as well as minimizing the similarity between different unlabeled videos run at different speeds. We employed contrastive loss between different video instances including groups of videos with similar actions to explore high-level action semantics within the neighborhood of different videos

depicting different instances of the same action. We demonstrated the effectiveness of our approach on four standard benchmark datasets, significantly outperforming several competing methods.

We proposed a EM-MIL framework with pseudo labeling and alternating training for weakly-supervised action detection in video. Our EM-MIL framework was motivated by traditional MIL literature which is under-explored in deep learning settings. By allowing us to explicitly model latent variables, this framework improves our control over the learning objective of the instance-level MIL, which leads to state of the art performance. While this work uses a relatively simple pseudo-labeling scheme to implement the EM method, more sophisticated EM methods can be designed, e.g. explicitly parameterize the latent distribution for instances and directly optimize the instance likelihood in E and M steps. Incorporating the video's temporal structure is also a promising direction for further performance improvement.

5.5 Few-Shot Machine Translation

We developed an approach to inducing syntactic representations that associate each token in the input with a discrete symbol from an arbitrarily-sized vocabulary, where the representations can be predicted incrementally in a strictly append-only manner. Our models achieve high F1 on the WSJ test set despite a steep information bottleneck limiting the information that can be associated with each token. The token-level tags produced by our model encode relevant syntactic information suitable for the given bit rate, while the locations of these tags serve to concretely define the location at which syntactic decisions can be committed to in a speculation-free manner. These systems can serve to improve our understanding of incremental parsing and sequential decision making, and the underlying computational methods may be useful in the analysis of other incremental contexts.

In this work we explored a wide range of techniques from the NMT toolkit for low-resource languages and propose a training curriculum that is effective under low-data and low-compute settings. Among those, effective utilization of comparable data with the correct succession of training objectives results in substantial gains. Despite the fact that the three languages we examine here are low-resource, we demonstrated that a simple bitext mining technique yielded quality comparable corpora. While using lexical translations for parallel corpora extraction has a long history in MT, it is under-explored for NMT of low-resource languages. We also observed that improvement with using comparable data depends on the size of the comparable data—for Gujarati, this results in significant improvement (since the mined comparable data is more than 3 times the size of the supervised data) while for Somali supervised results are still better (since the supervised data is 25x the size of the mined comparable data). Kazakh BLEU percentages are lower compared to those of Somali and Gujarati. We believe this is due to the morphologically more complex structure of Kazakh. Yet, the introduction of comparable data substantially increases the weakly-supervised BLEU scores for Kazakh from unsupervised, by 4.6 and 6.2 points, even without additional compute power.

We have described a collaborative framework involving two language models where one model, the critique generator, is trained to improve the performance of the task model. We train the former via policy gradient learning while treating the task model as a black-box. We show that RL4F leads to superior final performance across three domains compared to other strong baselines without resulting as the critiques remain fluent and natural. Future work might focus on generalizing the critique generator into a mixture of experts allowing humans and other models to contribute to critiqueing procedure.

5.6 Data-Efficient Fine-Tuning for Image Classification

For a task (*e.g.*, satellite image classification) with limited training data, the most common approach is to borrow a pretrained large model and finetune it. Current methods for personalizing such large models (*e.g.*, fully finetuning, LoRA, Adapter) usually modifies the weights of feedforward network. Our insight is, once pretrained on large datasets, the feedforward backbone already contains all the information one may need, and instead of tweaking the feedforward backbone which may lose information learned during pretraining, we only need a way to "probe" the information required for downstread tasks while leaving the feedforward backbone untouched.

In this work, we proposed **To**p-Down **T**uning (**ToT**) which adds a top-down path to the pretrained network, and only finetunes the top-down path while keeping the feedforward backbone frozen when

transfering to downstread tasks. We have shown superior performance compared to other tuning methods such as fully finetuning, VPT, LoRA, etc.

We tackled the extremely difficult task of merging two or more models trained on completely disjoint tasks without additional training. We have shown experimentally how prior work falls short in this setting and posit that this is due to not merging features within models as well as merging every layer in the model at once. We then introduced ZipIt!, a generalized framework for merging models that deals with these issues and find it to significantly outperform both prior work [2] and our own strong baseline on a number of difficult model merging settings. We then analyzed the behavior of our method and find that at smaller model capacities, it performs similarly to permutation-based methods, but can perform much better as the model capacity increases. We hope ZipIt! can serve as a strong starting point for practical applications of merging models trained on different tasks.

5.7 Final Observations

Overall, we assess that significant progress was made leveraging improved base models, including ConvNeXT and CLIP. The relative improvement made by active learning methods, which were expected at the start of the program to be large, turned out to be less impactful than was hoped. The precise setting in which developed domain adaptation methods were successful was often more narrow than anticipated, and the semantics of domain adaptation often differed across different practitioners, so performance in the wild was only mildly positive. Fully unsupervised methods also offered striking performance capability that aligned well with the original program goals. "Zero-shot" methods were not originally considered to be in scope for the program, but in the end offered some of the most striking capabilities. Looking to the future, the avent of large scale multimodal LLMs is transforming the level of capability of such models and increasingly performant capabilities are anticipated.

References

[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[2] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *ICIVR*, 2009.

[3] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2022.

[4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

[5] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual Denoising Pre-training for Neural Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 11 2020.

[6] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 8440–8451, Association for Computational Linguistics, July 2020.

[7] N. Tandon, A. Madaan, P. Clark, K. Sakaguchi, and Y. Yang, "Interscript: A dataset for interactive learning of scripts through error feedback," *arXiv preprint arXiv:2112.07867*, 2021.

[8] W. Saunders, C. Yeh, J. Wu, S. Bills, L. Ouyang, J. Ward, and J. Leike, "Self-critiquing models for assisting human evaluators," *arXiv preprint arXiv:2206.05802*, 2022.

[9] X. Sun, P. Hu, and K. Saenko, "Dualcoop: Fast adaptation to multi-label recognition with limited annotations," in *NeurIPS*, 2022.

[10] V. Prabhu, S. Khare, D. Kartik, and J. Hoffman, "Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8558–8567, 2021.

[11] V. U. Prabhu, S. Yenamandra, A. Singh, and J. Hoffman, "Adapting self-supervised vision transformers by probing attention-conditioned masking consistency," in *Advances in Neural Information Processing Systems*.

[12] V. Prabhu, A. Chandrasekaran, K. Saenko, and J. Hoffman, "Active domain adaptation via clustering uncertainty-weighted embeddings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8505–8514, 2021.

[13] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8050–8058, 2019.

[14] K. Saito, D. Kim, and K. Saenko, "Openmatch: Open-set consistency regularization for semi-supervised learning with outliers," *NeurIPS*, 2021.

[15] A. F. Akyürek, E. Akyürek, D. T. Wijaya, and J. Andreas, "Subspace regularization for few-shot class incremental learning," in *International Conference on Learning Representations*, 2022.

[16] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[17] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," July 2021. If you use this software, please cite it as below.

[18] X. Wang, R. Girdhar, S. X. Yu, and I. Misra, "Cut and learn for unsupervised object detection and instance segmentation," *arXiv preprint arXiv:2301.11320*, 2023.

[19] X. Wang, I. Misra, Z. Zeng, and T. Darrell, "Videocutler: Surprisingly simple unsupervised video instance segmentation," *under review*, 2023.

[20] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," *arXiv preprint arXiv:2303.04803*, 2023.

[21] J. Materzynska, T. Xiao, R. Herzig, H. Xu, X. Wang, and T. Darrell, "Something-else: Compositional action recognition with spatial-temporal interaction networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1049–1059, 2020.

[22] H. Wu and X. Wang, "Contrastive learning of image representations with cross-video cycle-consistency," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10149–10159, 2021.

[23] A. Singh, O. Chakraborty, A. Varshney, R. Panda, R. Feris, K. Saenko, and A. Das, "Semi-supervised action recognition with temporal contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10389–10399, 2021.

[24] Z. Luo, D. Guillory, B. Shi, W. Ke, F. Wan, T. Darrell, and H. Xu, "Weakly-supervised action localization with expectation-maximization multi-instance learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pp. 729–745, Springer, 2020.

[25] G. Kuwanto, A. F. Akyürek, I. C. Tourni, S. Li, A. G. Jones, and D. Wijaya, "Low-resource machine translation training curriculum fit for low-resource languages," *arXiv preprint arXiv:2103.13272*, 2021.

[26] A. F. Akyurek, E. Akyurek, A. Kalyan, P. Clark, D. Wijaya, and N. Tandon, "Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs," in *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*, (Online), Association for Computational Linguistics, July 2023.

[27] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *arXiv preprint arXiv:2109.01134*, 2021.

[28] T. Huang, J. Chu, and F. Wei, "Unsupervised prompt learning for vision-language models," *arXiv preprint arXiv:2204.03649*, 2022.

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[30] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in *ICCV*, 2021.

[31] Y. Bahat, M. Irani, and G. Shakhnarovich, "Natural and adversarial error detection using invariance to image transformations," *arXiv preprint arXiv:1902.00236*, 2019.

[32] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[33] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

[34] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.

[35] M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat, and N. Ballas, "Masked siamese networks for label-efficient learning," in *European Conference on Computer Vision*, pp. 456–473, Springer, 2022.

[36] P. Rai, A. Saha, H. Daumé III, and S. Venkatasubramanian, "Domain adaptation meets active learning," in *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pp. 27–32, Association for Computational Linguistics, 2010.

[37] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1183–1192, JMLR. org, 2017.

[38] D. Roth and K. Small, "Margin-based active learning for structured output spaces," in *European Conference on Machine Learning*, pp. 413–424, Springer, 2006.

[39] D. Wang and Y. Shang, "A new active labeling method for deep learning," in *2014 International joint conference on neural networks (IJCNN)*, pp. 112–119, IEEE, 2014.

[40] D. Gissin and S. Shalev-Shwartz, "Discriminative active learning," *arXiv preprint arXiv:1907.06347*, 2019.

[41] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *International Conference on Learning Representations*, 2018.

[42] K. Saito and K. Saenko, "Ovanet: One-vs-all network for universal domain adaptation," *arXiv preprint arXiv:2104.03344*, 2021.

[43] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv preprint arXiv:2001.07685*, 2020.

[44] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, and N. Diaz Rodriguez, "Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges," *Information Fusion*, vol. 58, 12 2019.

[45] P. McClure, C. Y. Zheng, J. Kaczmarzyk, J. Rogers-Lee, S. Ghosh, D. Nielson, P. Bandettini, and F. Pereira, "Distributed weight consolidation: A brain segmentation case study," in *NeurIPS*, 2018.

[46] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021.

[47] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. van de Weijer, "Class-incremental learning: survey and performance evaluation," *arXiv preprint arXiv:2010.15277*, 2020.

[48] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgetting in gradient-based neural networks," *arXiv preprint arXiv:1312.6211*, 2013.

[49] D. Anderson and K. Burnham, "Model selection and multi-model inference," *Second. NY: Springer-Verlag*, vol. 63, no. 2020, p. 10, 2004.

[50] X. Tao, X. Hong, X. Chang, S. Dong, X. Wei, and Y. Gong, "Few-shot class-incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12183–12192, 2020.

[51] S. W. Yoon, D.-Y. Kim, J. Seo, and J. Moon, "Xtarnet: Learning to extract task-adaptive representation for incremental few-shot learning," in *International Conference on Machine Learning*, pp. 10852–10860, PMLR, 2020.

[52] M. Ren, R. Liao, E. Fetaya, and R. Zemel, "Incremental few-shot learning with attention attractor networks," in *Advances in Neural Information Processing Systems*, 2019.

[53] A. Agarwal, S. Gerber, and H. Daume, "Learning multiple tasks using manifold regularization," in *Advances in Neural Information Processing Systems* (J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, eds.), vol. 23, Curran Associates, Inc., 2010.

[54] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[55] Y. Wang, X. Shen, Y. Yuan, Y. Du, M. Li, S. X. Hu, J. L. Crowley, and D. Vaufreydaz, "Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut," *arXiv preprint arXiv:2209.00383*, 2022.

[56] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[57] J. Xie, W. Xie, and A. Zisserman, "Segmenting moving objects via an object-centric layered representation," in *Advances in Neural Information Processing Systems*, 2022.

[58] B. Cheng, A. Choudhuri, I. Misra, A. Kirillov, R. Girdhar, and A. G. Schwing, "Mask2former for video instance segmentation," *arXiv preprint arXiv:2112.10764*, 2021.

[59] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *ICCV*, 2019.

[60] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, and Y. Cui, "Spatiotemporal contrastive video representation learning," *arXiv preprint arXiv:2008.03800*, 2020.

[61] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *arXiv preprint arXiv:1802.04712*, 2018.

[62] P. Nguyen, T. Liu, G. Prasad, and B. Han, "Weakly supervised action localization by sparse temporal pooling network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6752–6761, 2018.

[63] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 4325–4334, 2017.

[64] D. Liu, T. Jiang, and Y. Wang, "Completeness modeling and context separation for weakly supervised temporal action localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1298–1307, 2019.

[65] Y. Yuan, Y. Lyu, X. Shen, I. W. Tsang, and D.-Y. Yeung, "Marginalized average attentional network for weakly-supervised learning," *arXiv preprint arXiv:1905.08586*, 2019.

[66] N. Kitaev, T. Lu, and D. Klein, "Learned incremental representations for parsing," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3086–3095, 2022.

[67] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 6282–6293, Association for Computational Linguistics, July 2020.

[68] Y. Kim, M. Graça, and H. Ney, "When and why is unsupervised neural machine translation useless?," in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, (Lisboa, Portugal), pp. 35–44, European Association for Machine Translation, Nov. 2020.

[69] N. Ahmed and M. Wahed, "The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research," *arXiv preprint arXiv:2010.15581*, 2020.

[70] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 3645–3650, Association for Computational Linguistics, July 2019.

[71] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, (New York, NY, USA), p. 610–623, Association for Computing Machinery, 2021.

[72] M. Reid and G. Neubig, "Learning to model editing processes," *arXiv preprint arXiv:2205.12374*, 2022.

[73] E. Malmi, Y. Dong, J. Mallinson, A. Chuklin, J. Adamek, D. Mirylenka, F. Stahlberg, S. Krause, S. Kumar, and A. Severyn, "Text generation with text-editing models," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, (Seattle, United States), pp. 1–7, Association for Computational Linguistics, July 2022.

[74] E. Mitchell, C. Lin, A. Bosselut, C. D. Manning, and C. Finn, "Memory-based model editing at scale," in *International Conference on Machine Learning*, pp. 15817–15831, PMLR, 2022.

[75] E. Mitchell, C. Lin, A. Bosselut, C. Finn, and C. D. Manning, "Fast model editing at scale," in *International Conference on Learning Representations*, 2022.

[76] T. Schick, J. Dwivedi-Yu, Z. Jiang, F. Petroni, P. Lewis, G. Izacard, Q. You, C. Nalmpantis, E. Grave, and S. Riedel, "Peer: A collaborative language model," *arXiv preprint arXiv:2208.11663*, 2022.

[77] D. Dohan, W. Xu, A. Lewkowycz, J. Austin, D. Bieber, R. G. Lopes, Y. Wu, H. Michalewski, R. A. Saurous, J. Sohl-Dickstein, *et al.*, "Language model cascades," *arXiv preprint arXiv:2207.10342*, 2022.

[78] A. Madaan, N. Tandon, P. Clark, and Y. Yang, "Memory-assisted prompt editing to improve gpt-3 after deployment," *arXiv preprint arXiv:2201.06009*, 2022.

[79] B. Shi, T. Darrell, and X. Wang, "Top-down visual attention from analysis by synthesis," *arXiv preprint arXiv:2303.13043*, 2023.

[80] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[81] T. Chen, T. Pu, H. Wu, Y. Xie, and L. Lin, "Structured semantic transfer for multi-label recognition with partial labels," 2022.

[82] T. Pu, T. Chen, H. Wu, and L. Lin, "Semantic-aware representation blending for multi-label image recognition with partial labels," in *AAAI*, 2022.

[83] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[84] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *ICCV*, 2019.

[85] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *CVPR*, 2019.

[86] T. Chen, L. Lin, X. Hui, R. Chen, and H. Wu, "Knowledge-guided multi-label few-shot learning for general image recognition," *IEEE Trans. on PAMI*, 2020.

[87] T. Durand, N. Mehrasa, and G. Mori, "Learning a deep convnet for multi-label classification with partial labels," in *CVPR*, 2019.

[88] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," *arXiv preprint arXiv:1312.5650*, 2013.

[89] Y. Zhang, B. Gong, and M. Shah, "Fast zero-shot image tagging," in *CVPR*, 2016.

[90] S. Rahman and S. Khan, "Deep multiple instance learning for zero-shot image tagging," in *ACCV*, 2018.

[91] A. Ben-Cohen, N. Zamir, E. Ben-Baruch, I. Friedman, and L. Zelnik-Manor, "Semantic diversity learning for zero-shot multi-label classification," in *ICCV*, 2021.

[92] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE Trans. on PAMI*, vol. 38, no. 7, pp. 1425–1438, 2015.

[93] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear Attention Networks," in *NIPS*, 2018.

[94] D. Huynh and E. Elhamifar, "A shared multi-attention framework for multi-label zero-shot learning," in *CVPR*, 2020.

[95] S. Narayan, A. Gupta, S. Khan, F. S. Khan, L. Shao, and M. Shah, "Discriminative region-based multi-label zero-shot learning," in *ICCV*, 2021.

[96] B. Li, Y. Wang, T. Che, S. Zhang, S. Zhao, P. Xu, W. Zhou, Y. Bengio, and K. Keutzer, "Rethinking distributional matching based domain adaptation," *arXiv preprint arXiv:2006.13352*, 2020.

[97] X. Jiang, Q. Lao, S. Matwin, and M. Havaei, "Implicit class-conditioned domain alignment for unsupervised domain adaptation," in *International Conference on Machine Learning*, 2020.

[98] Y. Wu, E. Winston, D. Kaushik, and Z. Lipton, "Domain adaptation with asymmetrically-relaxed distribution alignment," in *International Conference on Machine Learning*, pp. 6872–6881, 2019.

[99] S. Tan, X. Peng, and K. Saenko, "Class-imbalanced domain adaptation: An empirical odyssey," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2020.

[100] Z. Lipton, Y.-X. Wang, and A. Smola, "Detecting and correcting for label shift with black box predictors," in *International Conference on Machine Learning*, pp. 3122–3130, 2018.

[101] Z. Cao, L. Ma, M. Long, and J. Wang, "Partial adversarial domain adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 135–150, 2018.

[102] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3723–3732, 2018.

[103] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*, pp. 97–105, PMLR, 2015.

[104] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Universal domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2720–2729, 2019.

[105] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *International conference on machine learning*, pp. 2208–2217, PMLR, 2017.

[106] Z. Cao, K. You, M. Long, J. Wang, and Q. Yang, "Learning to transfer examples for partial domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2985–2994, 2019.

[107] X. Chen, S. Wang, M. Long, and J. Wang, "Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation," in *International Conference on Machine Learning*, pp. 1081–1090, 2019.

[108] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*, pp. 1180–1189, 2015.

[109] Y. Zhang, T. Liu, M. Long, and M. Jordan, "Bridging theory and algorithm for domain adaptation," in *International Conference on Machine Learning*, pp. 7404–7413, 2019.

[110] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Advances in Neural Information Processing Systems*, pp. 1640–1650, 2018.

[111] Y. Jin, X. Wang, M. Long, and J. Wang, "Minimum class confusion for versatile domain adaptation," in *European Conference on Computer Vision*, pp. 464–480, Springer, 2020.

[112] J. Yang, J. Liu, N. Xu, and J. Huang, "Tvt: Transferable vision transformer for unsupervised domain adaptation," *arXiv preprint arXiv:2108.05988*, 2021.

[113] K. Shen, R. Jones, A. Kumar, S. M. Xie, J. Z. HaoChen, T. Ma, and P. Liang, "Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation," 2022.

[114] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds.," in *International Conference on Learning Representations*, 2020.

[115] J.-C. Su, Y.-H. Tsai, K. Sohn, B. Liu, S. Maji, and M. Chandraker, "Active adversarial domain adaptation," in *The IEEE Winter Conference on Applications of Computer Vision*, pp. 739–748, 2020.

[116] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," *arXiv preprint arXiv:1904.04232*, 2019.

[117] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015.

[118] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Adversarial dropout regularization," *arXiv preprint arXiv:1711.01575*, 2017.

[119] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *NIPS*, 2004.

[120] F. M. Cariucci, L. Porzi, B. Caputo, E. Ricci, and S. R. Bulò, "Autodial: Automatic domain alignment layers," in *ICCV*, 2017.

[121] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[122] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[123] Q. Yu, D. Ikami, G. Irie, and K. Aizawa, "Multi-task curriculum framework for open-set semi-supervised learning," in *European Conference on Computer Vision*, pp. 438–454, Springer, 2020.

[124] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[125] D. Liu and J. Yu, "Otsu method and k-means," in *2009 Ninth International Conference on Hybrid Intelligent Systems*, vol. 1, pp. 344–349, IEEE, 2009.

[126] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[127] K. Chen and C.-G. Lee, "Incremental few-shot learning via vector quantization in deep embedded space," in *International Conference on Learning Representations*, 2021.

[128] A. Cheraghian, S. Rahman, P. Fang, S. K. Roy, L. Petersson, and M. Harandi, "Semantic-aware knowledge distillation for few-shot class-incremental learning," *arXiv preprint arXiv:2103.04059*, 2021.

[129] M. Mermillod, A. Bugaiska, and P. Bonin, "The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects," *Frontiers in psychology*, vol. 4, p. 504, 2013.

[130] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[131] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.

[132] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *ICCV*, 2019.

[133] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *CVPR Workshops*, 2020.

[134] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *AAAI*, 2020.

[135] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *ECCV*, 2016.

[136] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *CVPR*, 2020.

[137] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019.

[138] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *ICML*, 2021.

[139] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *arXiv:2012.12877*, 2020.

[140] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big Transfer (BiT): General visual representation learning," in *ECCV*, 2020.

[141] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *arXiv preprint arXiv:2106.10270*, 2021.

[142] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021.

[143] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[144] X. Wang, Z. Yu, S. De Mello, J. Kautz, A. Anandkumar, C. Shen, and J. M. Alvarez, "Freesolo: Learning to segment objects without annotations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14176–14186, 2022.

[145] A. Bar, X. Wang, V. Kantorov, C. J. Reed, R. Herzig, G. Chechik, A. Rohrbach, T. Darrell, and A. Globerson, "Detreg: Unsupervised pretraining with region priors for object detection," in *CVPR*, 2022.

[146] C. Yang, H. Lamdouar, E. Lu, A. Zisserman, and W. Xie, "Self-supervised video object segmentation by motion grouping," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7177–7188, 2021.

[147] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 402–419, Springer, 2020.

[148] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2758–2766, 2015.

[149] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *European Conf. on Computer Vision (ECCV)* (A. Fitzgibbon et al. (Eds.), ed.), Part IV, LNCS 7577, pp. 611–625, Springer-Verlag, Oct. 2012.

[150] M. Keuper, B. Andres, and T. Brox, "Motion trajectory segmentation via minimum cost multicuts," in *Proceedings of the IEEE international conference on computer vision*, pp. 3271–3279, 2015.

[151] D. Kondermann, R. Nair, K. Honauer, K. Krispin, J. Andrulis, A. Brock, B. Gussefeld, M. Rahimimoghaddam, S. Hofmann, C. Brenner, *et al.*, "The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 19–28, 2016.

[152] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *NeurIPS*, 2020.

[153] L. Jing and Y. Tian, "Self-supervised spatiotemporal feature learning by video geometric transformations," *arXiv preprint arXiv:1811.11387*, vol. 2, no. 7, p. 8, 2018.

[154] D. Kim, D. Cho, and I. S. Kweon, "Self-supervised video representation learning with space-time cubic puzzles," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8545–8552, 2019.

[155] S. Benaim, A. Ephrat, O. Lang, I. Mosseri, W. T. Freeman, M. Rubinstein, M. Irani, and T. Dekel, "Speednet: Learning the speediness in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9922–9931, 2020.

[156] T. Han, W. Xie, and A. Zisserman, "Video representation learning by dense predictive coding," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[157] J. Wang, J. Jiao, and Y.-H. Liu, "Self-supervised video representation learning by pace prediction," in *European Conference on Computer Vision*, pp. 504–521, Springer, 2020.

[158] C. Sun, F. Baradel, K. Murphy, and C. Schmid, "Learning video representations using contrastive bidirectional transformer," *arXiv preprint arXiv:1906.05743*, 2019.

[159] T. Han, W. Xie, and A. Zisserman, "Memory-augmented dense predictive coding for video representation learning," *arXiv preprint arXiv:2008.01065*, 2020.

[160] S. Jenni, G. Meishvili, and P. Favaro, "Video representation learning by recognizing temporal transformations," *arXiv preprint arXiv:2007.10730*, 2020.

[161] C. Yang, Y. Xu, B. Dai, and B. Zhou, "Video representation learning with visual tempo consistency," *arXiv preprint arXiv:2006.15489*, 2020.

[162] C.-F. R. Chen, R. Panda, K. Ramakrishnan, R. Feris, J. Cohn, A. Oliva, and Q. Fan, "Deep analysis of cnn-based spatio-temporal representations for action recognition," in *CVPR*, 2021.

[163] J. Materzynska, G. Berger, I. Bax, and R. Memisevic, "The jester dataset: A large-scale video dataset of human gestures," in *ICCVW*, 2019.

[164] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[165] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICMLW*, 2013.

[166] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *NeurIPS*, 2017.

[167] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *ICCV*, 2019.

[168] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *NIPS*, 2019.

[169] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, "Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5734–5743, 2017.

[170] X. Dai, B. Singh, G. Zhang, L. S. Davis, and Y. Qiu Chen, "Temporal context network for activity localization in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5793–5802, 2017.

[171] H. Xu, A. Das, and K. Saenko, "R-c3d: Region convolutional 3d network for temporal activity detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 5783–5792, 2017.

[172] J. Gao, Z. Yang, and R. Nevatia, "Cascaded boundary regression for temporal action detection," *arXiv preprint arXiv:1705.01180*, 2017.

[173] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2914–2923, 2017.

[174] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster r-cnn architecture for temporal action localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1130–1139, 2018.

[175] H. Alwassel, F. Caba Heilbron, and B. Ghanem, "Action search: Spotting actions in videos and its application to temporal action localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 251–266, 2018.

[176] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "Bsn: Boundary sensitive network for temporal action proposal generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

[177] C. Sun, S. Shetty, R. Sukthankar, and R. Nevatia, "Temporal localization of fine-grained actions in videos by domain transfer from web images," in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 371–380, ACM, 2015.

[178] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3544–3553, IEEE, 2017.

[179] Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang, "Autoloc: Weakly-supervised temporal action localization in untrimmed videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 154–171, 2018.

[180] S. Paul, S. Roy, and A. K. Roy-Chowdhury, "W-talc: Weakly-supervised temporal activity localization and classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 563–579, 2018.

[181] Z. Liu, L. Wang, Q. Zhang, Z. Gao, Z. Niu, N. Zheng, and G. Hua, "Weakly supervised temporal action localization through contrast based evaluation networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3899–3908, 2019.

[182] T. Yu, Z. Ren, Y. Li, E. Yan, N. Xu, and J. Yuan, "Temporal structure mining for weakly supervised action detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5522–5531, 2019.

[183] S. Narayan, H. Cholakkal, F. S. Khan, and L. Shao, "3c-net: Category count and center loss for weakly-supervised action localization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8679–8687, 2019.

[184] P. X. Nguyen, D. Ramanan, and C. C. Fowlkes, "Weakly-supervised action localization with background modeling," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5502–5511, 2019.

[185] M. Stern, J. Andreas, and D. Klein, "A minimal span-based neural constituency parser," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Vancouver, Canada), pp. 818–827, Association for Computational Linguistics, July 2017.

[186] K. Mrini, F. Dernoncourt, Q. H. Tran, T. Bui, W. Chang, and N. Nakashole, "Rethinking self-attention: Towards interpretability in neural parsing," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, (Online), pp. 731–742, Association for Computational Linguistics, Nov. 2020.

[187] J. Tracey, S. Strassel, A. Bies, Z. Song, M. Arrigo, K. Griffitt, D. Delgado, D. Graff, S. Kulick, J. Mott, and N. Kuster, "Corpus building for low resource languages in the DARPA LORELEI program," in *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, (Dublin, Ireland), pp. 48–55, European Association for Machine Translation, Aug. 2019.

[188] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.

[189] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, pp. 74–81, 2004.

[190] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016.

[191] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[192] S. K. Ainsworth, J. Hayase, and S. Srinivasa, "Git re-basin: Merging models modulo permutation symmetries," *arXiv:2209.04836*, 2022.

[193] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, 2011.

[194] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *CVPR*, 2012.

[195] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD Birds 200," Tech. Rep. CNS-TR-2010-001, California Institute of Technology, 2010.

[196] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *CVPR*, 2015.

A Appendix: Full List of Team Publications

Full List of Earlier Team Publications (2019-2021)

Title Indicate the title of the publication.	Variational Adversarial Active Learning
Author(s) Indicate the authors of the publication.	Samarth Sinha*, Sayna Ebrahimi*, Trevor Darrell
Publication Date Indicate the date of publication.	October, 2019
Publication Venue Indicate the journal, conference, or magazine name.	Proceedings of the IEEE International Conference on Computer Vision (ICCV)
Keywords Enter at least three keywords for the publication.	Semi-supervised active learning, Adversarial learning, Variational auto-encoders
URL Enter the URL associated with this publication (if any.)	http://openaccess.thecvf.com/content_ICCV_2019/html/Sinha_Variational_Adversarial_Active_Learning_ICCV_2019_paper.html
Comments Enter any relevant comments about this publication.	First two authors contributed equally, listed alphabetically.

Title Indicate the title of the publication.	Uncertainty-Guided Continual Learning with Bayesian Neural Networks
Author(s) Indicate the authors of the publication.	Sayna Ebrahimi, Mohamed Elhoseinyi, Trevor Darrell, Marcus Rohrbach
Publication Date Indicate the date of publication.	February, 2020
Publication Venue Indicate the journal,	International Conference on Learning Representations (ICLR)

conference, or magazine name.	
Keywords Enter at least three keywords for the publication.	Continual learning, Bayesian neural networks, Predictive uncertainty
URL Enter the URL associated with this publication (if any.)	https://openreview.net/pdf?id=HklUCCVKDB
Comments Enter any relevant comments about this publication.	Code: https://github.com/SaynaEbrahimi/UCB

Title Indicate the title of the publication.	Few-Shot Learning via Embedding Adaptation with Set-to-Set Functions
Author(s) Indicate the authors of the publication.	Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, Fei Sha
Publication Date Indicate the date of publication.	March, 2020
Publication Venue Indicate the journal, conference, or magazine name.	Conference on Computer Vision and Pattern Recognition (CVPR)
Keywords Enter at least three keywords for the publication.	Few-Shot Learning, Embedding Adaptation
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/1812.03664
Comments Enter any relevant comments about this publication.	Code: https://github.com/Sha-Lab/FEAT

Title Indicate the title of the publication.	Few-shot Object Detection via Feature Reweighting
Author(s) Indicate the authors of the publication.	Bingyi Kang*, Zhuang Liu*, Xin Wang, Fisher Yu, Jiashi Feng, Trevor Darrell
Publication Date Indicate the date of publication.	October, 2019
Publication Venue Indicate the journal, conference, or magazine name.	Proceedings of the IEEE International Conference on Computer Vision (ICCV)
Keywords Enter at least three keywords for the publication.	Few-shot Learning, Object Detection
URL Enter the URL associated with this publication (if any.)	http://openaccess.thecvf.com/content_ICCV_2019/papers/Kang_Few-Shot_Object_Detection_via_Feature_Reweighting_ICCV_2019_paper.pdf
Comments Enter any relevant comments about this publication.	Code: https://github.com/bingykang/Fewshot_Detection

Title Indicate the title of the publication.	Semi-Supervised Domain Adaptation via Minimax Entropy
Author(s) Indicate the authors of the publication.	Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, Kate Saenko
Publication Date Indicate the date of publication.	October, 2019
Publication Venue Indicate the journal, conference, or magazine name.	Proceedings of the IEEE International Conference on Computer Vision (ICCV)
Keywords Enter at least three keywords for the publication.	Semi-Supervised learning, Domain adaptation, Minimax entropy

URL Enter the URL associated with this publication (if any.)	https://openaccess.thecvf.com/content_ICCV_2019/papers/Saito_Semi-Supervised_Domain_Adaptation_via_Minimax_Entropy_ICCV_2019_paper.pdf
Comments Enter any relevant comments about this publication.	Code: https://github.com/VisionLearningGroup/SSDA_MME

Title Indicate the title of the publication.	DIPNet: Dynamic Identity Propagation Network for Video Object Segmentation
Author(s) Indicate the authors of the publication.	Ping Hu, Jun Liu, Gang Wang, Vitaly Ablavsky, Kate Saenko, Stan Sclaroff
Publication Date Indicate the date of publication.	March, 2020
Publication Venue Indicate the journal, conference, or magazine name.	The IEEE Winter Conference on Applications of Computer Vision (WACV)
Keywords Enter at least three keywords for the publication.	Semi-Supervised learning, Dynamic identity propagation, Object segmentation
URL Enter the URL associated with this publication (if any.)	https://openaccess.thecvf.com/content_WACV_2020/html/Hu_DIPNet_Dynamic_Identity_Propagation_Network_for_Video_Object_Segmentation_WACV_2020_paper.html
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Something-Else: Compositional Action Recognition with Spatial-Temporal Interaction Networks
Author(s) Indicate the authors of the publication.	Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu†, Xiaolong Wang†, Trevor Darrell†

Publication Date Indicate the date of publication.	March, 2020
Publication Venue Indicate the journal, conference, or magazine name.	Conference on Computer Vision and Pattern Recognition (CVPR)
Keywords Enter at least three keywords for the publication.	Compositional action recognition, Spatial-Temporal interaction, few-shot learning
URL Enter the URL associated with this publication (if any.)	https://openaccess.thecvf.com/content_CVPR_2020/html/Materzynska_Something-Else_Compositional_Action_Recognition_With_Spatial-Temporal_Interaction_Networks_CVPR_2020_paper.html
Comments Enter any relevant comments about this publication.	Last three authors contributed equally. Code/Dataset: https://github.com/joaanna/something_else

Title Indicate the title of the publication.	Multi-Label and Multilingual News Framing Analysis
Author(s) Indicate the authors of the publication.	Feyza Akyurek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, Derry Wijaya
Publication Date Indicate the date of publication.	April, 2020 (In submission)
Publication Venue Indicate the journal, conference, or magazine name.	The 58th Annual Meeting of the Association for Computational Linguistics (ACL)
Keywords Enter at least three keywords for the publication.	Multi-Label, Multilingual, News Framing Analysis
URL Enter the URL associated with this publication (if any.)	https://derrywijaya.github.io/Multi_label_and_Multilingual_News_Framing_Analysis.pdf
Comments Enter any relevant	Code: https://github.com/feyzaakyurek/newsframing

comments about this publication.	

Title Indicate the title of the publication.	A New Meta-Baseline for Few-Shot Learning
Author(s) Indicate the authors of the publication.	Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell
Publication Date Indicate the date of publication.	March, 2020
Publication Venue Indicate the journal, conference, or magazine name.	arXiv
Keywords Enter at least three keywords for the publication.	Meta learning, few shot learning, Meta-Baseline
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2003.04390v1
Comments Enter any relevant comments about this publication.	Code: https://github.com/cyvius96/few-shot-meta-baseline

Title Indicate the title of the publication.	Revisiting Few-shot Activity Detection with Class Similarity Control
Author(s) Indicate the authors of the publication.	Huijuan Xu, Ximeng Sun, Eric Tzeng, Abir Das, Kate Saenko, Trevor Darrell
Publication Date Indicate the date of publication.	March, 2020
Publication Venue Indicate the journal, conference, or magazine name.	arXiv

Keywords Enter at least three keywords for the publication.	Few-shot Activity detection
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2004.00137
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	CURL: Contrastive Unsupervised Representations for Reinforcement Learning
Author(s) Indicate the authors of the publication.	Aravind Srinivas, Michael Laskin, Pieter Abbeel
Publication Date Indicate the date of publication.	July 2020
Publication Venue Indicate the journal, conference, or magazine name.	Proceedings of the 37th International Conference on Machine Learning (ICML 2020)
Keywords Enter at least three keywords for the publication.	
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/pdf/2004.04136.pdf
Comments Enter any relevant comments about this publication.	https://www. github.com/MishaLaskin/curl.

Title Indicate the title of the publication.	Reinforcement Learning with Augmented Data
Author(s) Indicate the authors of the publication.	Michael Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, Aravind Srinivas
Publication Date Indicate the date of publication.	April 2020
Publication Venue Indicate the journal, conference, or magazine name.	arXiv (under review at NeurIPS 2020)
Keywords Enter at least three keywords for the publication.	
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/pdf/2004.14990.pdf
Comments Enter any relevant comments about this publication.	Code: https://github.com/MishaLaskin/rad

Title Indicate the title of the publication.	Explainable Deep Classification Models for Domain Generalization
Author(s) Indicate the authors of the publication.	Andrea Zunino*, Sarah Adel Bargal*, Riccardo Volpi, Mehrnoosh Sameki , Jianming Zhang , Stan Sclaroff , Vittorio Murino, Kate Saenko
Publication Date Indicate the date of publication.	arXiv
Publication Venue Indicate the journal, conference, or magazine name.	

Keywords Enter at least three keywords for the publication.	Explainability, Deep Models, Domain Generalization
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/pdf/2003.06498.pdf
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Online Variable-Shot Meta-Learning
Author(s) Indicate the authors of the publication.	Tianhe Yu*, Xinyang Geng*, Chelsea Finn, Sergey Levine
Publication Date Indicate the date of publication.	N/A (under review)
Publication Venue Indicate the journal, conference, or magazine name.	N/A (under review)
Keywords Enter at least three keywords for the publication.	Meta-learning, online learning, few-shot image recognition
URL Enter the URL associated with this publication (if any.)	Coming soon
Comments Enter any relevant comments about this publication.	Proposes incremental online learning, where new data points are revealed one at a time in an online process for each task.

Title Indicate the title of the publication.	Adaptive Risk Minimization: A Meta-Learning Approach for Tackling Group Shift

Author(s) Indicate the authors of the publication.	Marvin Zhang*, Henrik Marklund*, Abhishek Gupta, Sergey Levine, Chelsea Finn
Publication Date Indicate the date of publication.	N/A (under review)
Publication Venue Indicate the journal, conference, or magazine name.	N/A (under review)
Keywords Enter at least three keywords for the publication.	Meta-learning, distributional robustness, distributional shift
URL Enter the URL associated with this publication (if any.)	Coming soon
Comments Enter any relevant comments about this publication.	Proposes ARM, a framework for handling distributional shift by adapting to new test-time distributions without any labels at test time. Provides an adaptive alternative to previously proposed DRO methods.

Title Indicate the title of the publication.	Amortized Conditional Normalized Maximum Likelihood
Author(s) Indicate the authors of the publication.	Aurick Zhou, Sergey Levine
Publication Date Indicate the date of publication.	N/A (under review)
Publication Venue Indicate the journal, conference, or magazine name.	N/A (under review)
Keywords Enter at least three keywords for the publication.	Uncertainty estimation, neural network calibration, out-of-distribution detection
URL Enter the URL associated	Coming soon

with this publication (if any.)	
Comments Enter any relevant comments about this publication.	Proposes ACNML, a method for tractably approximating the NML distribution with deep networks by using approximate parameter posteriors. Demonstrates the benefit of this approach for uncertainty estimation and detecting distributional shift.

Title Indicate the title of the publication.	Learning to Balance Specificity and Invariance for In and Out of Domain Generalization
Author(s) Indicate the authors of the publication.	Prithvijit Chattopadhyay, Yogesh Balaji, Judy Hoffman
Publication Date Indicate the date of publication.	August 2020
Publication Venue Indicate the journal, conference, or magazine name.	European Conference in Computer Vision (ECCV)
Keywords Enter at least three keywords for the publication.	Domain generalization, domain adaptation, representation learning
URL Enter the URL associated with this publication (if any.)	Coming soon
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Universal Domain Adaptation through Self-supervision
Author(s) Indicate the authors of the publication.	Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Kate Saenko

Publication Date Indicate the date of publication.	Under review
Publication Venue Indicate the journal, conference, or magazine name.	Under review
Keywords Enter at least three keywords for the publication.	Universal Domain Adaptation, Self-supervised clustering
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2002.07953
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	CRaWL: Cyclic Random Walk Loss for Self-Supervised Domain Adaptation
Author(s) Indicate the authors of the publication.	Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Kate Saenko
Publication Date Indicate the date of publication.	Under review
Publication Venue Indicate the journal, conference, or magazine name.	Under review
Keywords Enter at least three keywords for the publication.	Universal Domain Adaptation, Self-supervised clustering
URL Enter the URL associated with this publication (if any.)	
Comments Enter any relevant	

comments about this publication.	

Title Indicate the title of the publication.	Cross-domain Self-supervised Learning for Domain Adaptation with Few Source Labels (CDS)
Author(s) Indicate the authors of the publication.	Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A. Plummer, Stan Sclaroff, Kate Saenko
Publication Date Indicate the date of publication.	Under review
Publication Venue Indicate the journal, conference, or magazine name.	Under review
Keywords Enter at least three keywords for the publication.	Domain adaptation, Self-supervised clustering, Representation Learning
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2003.08264
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Auxiliary Task Reweighting for Minimum-data Learning
Author(s) Indicate the authors of the publication.	Baifeng Shi, Judy Hoffman, Kate Saenko, Trevor Darrell and Huijuan Xu
Publication Date Indicate the date of publication.	Dec 6, 2020
Publication Venue Indicate the journal,	Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS 2020)

conference, or magazine name.	
Keywords Enter at least three keywords for the publication.	Learning using privileged information, Multi-task learning, Task reweighting
URL Enter the URL associated with this publication (if any.)	Project webpage: https://sites.google.com/view/auxiliary-task-reweighting
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Minimax Active Learning
Author(s) Indicate the authors of the publication.	Sayna Ebrahimi, William Gan, Kamyar Salahi, Trevor Darrell
Publication Date Indicate the date of publication.	Dec 18, 2020
Publication Venue Indicate the journal, conference, or magazine name.	Submitted to CVPR
Keywords Enter at least three keywords for the publication.	Active learning, semi-supervised learning, object detection, semantic segmentation
URL Enter the URL associated with this publication (if any.)	Project webpage: https://people.eecs.berkeley.edu/~sayna/mal.html
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	A Review of Single-Source Deep Unsupervised Visual Domain Adaptation
Author(s) Indicate the authors of the publication.	Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E Gonzalez, Alberto L Sangiovanni-Vincentelli, Sanjit A Seshia, Kurt Keutzer
Publication Date Indicate the date of publication.	Oct, 2020
Publication Venue Indicate the journal, conference, or magazine name.	IEEE Transactions on Neural Networks and Learning Systems
Keywords Enter at least three keywords for the publication.	Single-Source, Visual Domain Adaptation, Review
URL Enter the URL associated with this publication (if any.)	https://ieeexplore.ieee.org/abstract/document/9238468
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Domain-agnostic Contrastive Learning with Informative Prediction for Unsupervised Domain Adaptation
Author(s) Indicate the authors of the publication.	Bo Li, Yezhen Wang, Tong Che, Shanghang Zhang, Sicheng Zhao, Xiangyu Yue, Pengfei Xu, Trevor Darrell, Kurt Keutzer
Publication Date Indicate the date of publication.	Nov, 2020
Publication Venue Indicate the journal, conference, or magazine name.	Submitted to CVPR

Keywords Enter at least three keywords for the publication.	Domain adaptation, Contrastive learning, Informative prediction
URL Enter the URL associated with this publication (if any.)	Will be released after accepted
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Prototypical Cross-domain Self-supervised Learning for Few-shot Unsupervised Domain Adaptation
Author(s) Indicate the authors of the publication.	Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, Alberto L Sangiovanni-Vincentelli
Publication Date Indicate the date of publication.	Nov. 2020
Publication Venue Indicate the journal, conference, or magazine name.	Submitted to CVPR
Keywords Enter at least three keywords for the publication.	Domain adaptation, Contrastive learning, Few-shot
URL Enter the URL associated with this publication (if any.)	Will be released after accepted
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Learning Invariant Representations and Risks for Semi-supervised Domain Adaptation
Author(s) Indicate the authors of the publication.	Bo Li, Yezhen Wang, Shanghang Zhang, Dongsheng Li, Kurt Keutzer, Trevor Darrell, Han Zhao
Publication Date Indicate the date of publication.	Nov. 2020
Publication Venue Indicate the journal, conference, or magazine name.	Submitted to CVPR
Keywords Enter at least three keywords for the publication.	Domain adaptation, Semi-supervised, Risk minimization
URL Enter the URL associated with this publication (if any.)	Will be released after accepted
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Region Similarity Representation Learning
Author(s) Indicate the authors of the publication.	Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, Trevor Darrell
Publication Date Indicate the date of publication.	24 Mar 2021
Publication Venue Indicate the journal, conference, or magazine name.	arXiv
Keywords Enter at least three	Self-supervised, object detection, transfer learning

keywords for the publication.	
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2103.12902
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Learning Continuous Image Representation with Local Implicit Image Function
Author(s) Indicate the authors of the publication.	Yinbo Chen, Sifei Liu, Xiaolong Wang.
Publication Date Indicate the date of publication.	1 Mar 2021
Publication Venue Indicate the journal, conference, or magazine name.	CVPR
Keywords Enter at least three keywords for the publication.	Self-supervised, image representation
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2012.09161
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Semi-Supervised 3D Hand-Object Poses Estimation with Interactions in Time
Author(s) Indicate the authors of the publication.	Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, Xiaolong Wang
Publication Date Indicate the date of publication.	1 Mar 2021
Publication Venue Indicate the journal, conference, or magazine name.	CVPR
Keywords Enter at least three keywords for the publication.	Semi-supervised learning, video, 3D pose estimation
URL Enter the URL associated with this publication (if any.)	https://drive.google.com/file/d/1pO9KDB0Ww-ScGg6hwIZqPk2AIAsx-fWJ/view?usp=sharing
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Rethinking Self-supervised Correspondence Learning: A Video Frame-level Similarity Perspective.
Author(s) Indicate the authors of the publication.	Jiarui Xu, Xiaolong Wang.
Publication Date Indicate the date of publication.	31 Mar 2021
Publication Venue Indicate the journal, conference, or magazine name.	arXiv
Keywords Enter at least three keywords for the publication.	Self-supervised, video, correspondence, tracking

URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2103.17263
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Temporal Action Detection with Multi-level Supervision
Author(s) Indicate the authors of the publication.	Baifeng Shi, Qi Dai, Judy Hoffman, Kate Saenko, Trevor Darrell, Huijuan Xu
Publication Date Indicate the date of publication.	18 Feb 2021
Publication Venue Indicate the journal, conference, or magazine name.	arXiv
Keywords Enter at least three keywords for the publication.	Semi-supervised, temporal action detection
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/pdf/2011.11893.pdf
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	AdaFuse: Adaptive Temporal Fusion Network for Efficient Action Recognition
Author(s) Indicate the authors of the publication.	Yue Meng, Rameswar Panda, Chung-Ching Lin, Prasanna Sattigeri, Leonid Karlinsky, Kate Saenko, Aude Oliva, Rogerio Feris

Publication Date Indicate the date of publication.	02/10/2021
Publication Venue Indicate the journal, conference, or magazine name.	ICLR 2021
Keywords Enter at least three keywords for the publication.	Action recognition, temporal fusion network, Adaptive temporal fusion
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/pdf/2102.05775.pdf
Comments Enter any relevant comments about this publication.	Website: https://mengyuest.github.io/AdaFuse/ Code: https://mengyuest.github.io/AdaFuse/

Title Indicate the title of the publication.	Semi-Supervised Action Recognition with Temporal Contrastive Learning
Author(s) Indicate the authors of the publication.	Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, Abir Das
Publication Date Indicate the date of publication.	02/4/2021
Publication Venue Indicate the journal, conference, or magazine name.	CVPR 2021
Keywords Enter at least three keywords for the publication.	Action recognition, Contrastive Learning, Semi-supervised Learning
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/pdf/2102.02751.pdf
Comments Enter any relevant comments about this publication.	Website: https://cvir.github.io/TCL/

Title Indicate the title of the publication.	Surprisingly Simple Semi-Supervised Domain Adaptation with Pretraining and Consistency
Author(s) Indicate the authors of the publication.	Samarth Mishra, Kate Saenko, Venkatesh Saligrama
Publication Date Indicate the date of publication.	01/29/2021
Publication Venue Indicate the journal, conference, or magazine name.	arXiv
Keywords Enter at least three keywords for the publication.	Domain adaptation, Semi-supervised Learning, Consistency regularization
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/pdf/2101.12727.pdf
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	OpenMatch: Open-set Consistency Regularization for Semi-supervised Learning with Outliers
Author(s) Indicate the authors of the publication.	Kuniaki Saito, Donghyun Kim, Kate Saenko
Publication Date Indicate the date of publication.	28 May 2021
Publication Venue Indicate the journal, conference, or magazine name.	arXiv
Keywords Enter at least three keywords for the publication.	Semi-supervised, consistency regularization
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2105.14148.pdf

Comments Enter any relevant comments about this publication.	Code: https://github.com/VisionLearningGroup/OP_Match

Title Indicate the title of the publication.	Fighting Gradients with Gradients: Dynamic Defenses against Adversarial Attacks
Author(s) Indicate the authors of the publication.	Dequan Wang, An Ju, Evan Shelhamer, David Wagner, Trevor Darrell
Publication Date Indicate the date of publication.	2021/5/18
Publication Venue Indicate the journal, conference, or magazine name.	arXiv
Keywords Enter at least three keywords for the publication.	Adversarial attacks, domain adaptation
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2105.08714
Comments Enter any relevant comments about this publication.	Code: https://github.com/DequanWang/dent

Title Indicate the title of the publication.	DETReg: Unsupervised Pretraining with Region Priors for Object Detection
Author(s) Indicate the authors of the publication.	Amir Bar, Xin Wang, Vadim Kantorov, Colorado J Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, Amir Globerson
Publication Date Indicate the date of publication.	2021/6/8
Publication Venue Indicate the journal, conference, or magazine name.	arXiv
Keywords Enter at least three keywords for the publication.	Self-supervised learning, object detection, Transformers
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2106.04550

Comments Enter any relevant comments about this publication.	Code: https://github.com/amirbar/DETReg

Title Indicate the title of the publication.	Semi-Supervised Action Recognition with Temporal Contrastive Learning
Author(s) Indicate the authors of the publication.	Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, Abir Das
Publication Date Indicate the date of publication.	June 2021
Publication Venue Indicate the journal, conference, or magazine name.	CVPR 2021
Keywords Enter at least three keywords for the publication.	Video classification, self supervised learning, semi supervised learning.
URL Enter the URL associated with this publication (if any.)	https://openaccess.thecvf.com/content/CVPR2021/html/Singh_Semi-Supervised_Action_Recognition_With_Temporal_Contrastive_Learning_CVPR_2021_paper.html
Comments Enter any relevant comments about this publication.	Code: https://github.com/CVIR/TCL

Title Indicate the title of the publication.	Robust Object Detection via Instance-Level Temporal Cycle Confusion
Author(s) Indicate the authors of the publication.	Xin Wang, Thomas E Huang, Benlin Liu, Fisher Yu, Xiaolong Wang, Joseph E Gonzalez, Trevor Darrell
Publication Date Indicate the date of publication.	2021/4/16
Publication Venue Indicate the journal, conference, or magazine name.	ICCV 2021
Keywords Enter at least three keywords for the publication.	Object detection, cycle consistency, cycle confusion

URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2104.08381
Comments Enter any relevant comments about this publication.	https://github.com/xinw1012/cycle-confusion

Title Indicate the title of the publication.	Rethinking Self-supervised Correspondence Learning: A Video Frame-level Similarity Perspective
Author(s) Indicate the authors of the publication.	Jiarui Xu, Xiaolong Wang
Publication Date Indicate the date of publication.	04/06/2021
Publication Venue Indicate the journal, conference, or magazine name.	ICCV 2021
Keywords Enter at least three keywords for the publication.	Self-supervised, space-time correspondence, video classification
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2103.17263
Comments Enter any relevant comments about this publication.	https://jerryxu.net/VFS/

Title Indicate the title of the publication.	Contrastive Learning of Image Representations with Cross-Video Cycle-Consistency
Author(s) Indicate the authors of the publication.	Haiping Wu, Xiaolong Wang
Publication Date Indicate the date of publication.	05/13/2021
Publication Venue Indicate the journal, conference, or magazine name.	ICCV 2021
Keywords Enter at least three keywords for the publication.	Contrastive learning, video classification, cycle-consistency

URL Enter the URL associated with this publication (if any.)	https://arxiv.org/pdf/2105.06463.pdf
Comments Enter any relevant comments about this publication.	https://happywu.github.io/cycle_contrast_video/

Title Indicate the title of the publication.	OVANet: One-vs-All Network for Universal Domain Adaptation
Author(s) Indicate the authors of the publication.	Kuniaki Saito, Kate Saenko
Publication Date Indicate the date of publication.	2021/4/7
Publication Venue Indicate the journal, conference, or magazine name.	ICCV 2021
Keywords Enter at least three keywords for the publication.	one-vs-all classifier, universal domain adaptation
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2104.03344
Comments Enter any relevant comments about this publication.	https://github.com/VisionLearningGroup/OVANet

Title Indicate the title of the publication.	Object-Region Video Transformers
Author(s) Indicate the authors of the publication.	Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, Amir Globerson
Publication Date Indicate the date of publication.	-
Publication Venue Indicate the journal, conference, or magazine name.	-
Keywords Enter at least three keywords for the publication.	Video transformers - object-centric action recognition

URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2110.06915
Comments Enter any relevant comments about this publication.	https://roeiherz.github.io/ORViT/

Title Indicate the title of the publication.	VisDA-2021 Competition Universal Domain Adaptation to Improve Performance on Out-of-Distribution Data
Author(s) Indicate the authors of the publication.	Dina Bashkirova, Dan Hendrycks, Donghyun Kim, Samarth Mishra, Kate Saenko, Kuniaki Saito, Piotr Teterwak, Ben Usman
Publication Date Indicate the date of publication.	23 Jul 2021
Publication Venue Indicate the journal, conference, or magazine name.	NeurIPS Dataset Track
Keywords Enter at least three keywords for the publication.	Domain adaptation - outlier detection
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/pdf/2107.11011.pdf
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Contrast and Mix: Temporal Contrastive Video Domain Adaptation with Background Mixing
Author(s) Indicate the authors of the publication.	Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, Abir Das
Publication Date Indicate the date of publication.	
Publication Venue Indicate the journal, conference, or magazine name.	NeurIPS 2021

Keywords Enter at least three keywords for the publication.	Contrastive learning - video domain adaptation
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/pdf/2107.11011.pdf
Comments Enter any relevant comments about this publication.	https://cvir.github.io/projects/comix

Title Indicate the title of the publication.	OVANet: One-vs-All Network for Universal Domain Adaptation
Author(s) Indicate the authors of the publication.	Kuniaki Saito, Kate Saenko
Publication Date Indicate the date of publication.	Oct. 11
Publication Venue Indicate the journal, conference, or magazine name.	ICCV 2021
Keywords Enter at least three keywords for the publication.	Universal Domain Adaptation, One-vs-All network, Outlier-detection
URL Enter the URL associated with this publication (if any.)	https://openaccess.thecvf.com/content/ICCV2021/papers/Saito_OVANet_One-vs-All_Network_for_Universal_Domain_Adaptation_ICCV_2021_paper.pdf
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density
Author(s) Indicate the authors of the publication.	Kuniaki Saito, Donghyun Kim, Piotr Teterwak, Stan Sclaroff, Trevor Darrell, Kate Saenko
Publication Date Indicate the date of publication.	Oct. 11
Publication Venue Indicate the journal, conference, or magazine name.	ICCV2021

Keywords Enter at least three keywords for the publication.	Domain adaptation, unsupervised validation, unsupervised clustering
URL Enter the URL associated with this publication (if any.)	https://openaccess.thecvf.com/content/ICCV2021/papers/Saito_Tune_It_the_Right_Way_Unsupervised_Validation_of_Domain_Adaptation_ICCV_2021_paper.pdf
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	OpenMatch: Open-set Consistency Regularization for Semi-supervised Learning with Outliers
Author(s) Indicate the authors of the publication.	Kuniaki Saito, Donghyun Kim, Kate Saenko
Publication Date Indicate the date of publication.	Dec.6
Publication Venue Indicate the journal, conference, or magazine name.	NeurIPS 2021
Keywords Enter at least three keywords for the publication.	Semi-supervised learning, Out-lier detection, Consistency Regularization
URL Enter the URL associated with this publication (if any.)	https://proceedings.neurips.cc/paper/2021/file/da11e8cd1811acb79ccf0fd62cd58f86-Paper.pdf
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Learning to Detect Every Thing in an Open World
Author(s) Indicate the authors of the publication.	Kuniaki Saito, Ping Hu, Trevor Darrell, Kate Saenko
Publication Date Indicate the date of publication.	Dec.3
Publication Venue Indicate the journal, conference, or magazine name.	Arxivz

Keywords Enter at least three keywords for the publication.	Object detection, open-world instance segmentation, copy-paste data augmentation
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/pdf/2112.01698.pdf
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	CDS: Cross-Domain Self-Supervised Pre-Training.
Author(s) Indicate the authors of the publication.	Kim, D., Saito, K., Oh, T. H., Plummer, B. A., Sclaroff, S., & Saenko, K.
Publication Date Indicate the date of publication.	Oct 2021
Publication Venue Indicate the journal, conference, or magazine name.	In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9123-9132).
Keywords Enter at least three keywords for the publication.	Pre-training; self-supervised learning; domain adaptation
URL Enter the URL associated with this publication (if any.)	https://openaccess.thecvf.com/content/ICCV2021/papers/Kim_CDS_Cross-Domain_Self-Supervised_Pre-Training_ICCV_2021_paper.pdf
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Selective Entropy Optimization via Committee Consistency for Unsupervised Domain Adaptation
Author(s) Indicate the authors of the publication.	Viraj Prabhu, Shivam Khare, Deeksha Karthik, Judy Hoffman
Publication Date Indicate the date of publication.	Oct 2021
Publication Venue Indicate the journal, conference, or magazine name.	In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Keywords Enter at least three keywords for the publication.	domain adaptation; object recognition; long-tailed distribution
URL Enter the URL associated with this publication (if any.)	https://openaccess.thecvf.com/content/ICCV2021/papers/Prabhu_SENTRY_Selective_Entropy_Optimization_via_Committee_Consistency_for_Unsupervised_Domain_ICCV_2021_paper.pdf
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Rethinking Self-supervised Correspondence Learning: A Video Frame-level Similarity Perspective
Author(s) Indicate the authors of the publication.	Jiarui Xu, Xiaolong Wang.
Publication Date Indicate the date of publication.	Oct 2021
Publication Venue Indicate the journal, conference, or magazine name.	ICCV 2021
Keywords Enter at least three keywords for the publication.	Self-supervised Learning, contrastive learning, videos
URL Enter the URL associated with this publication (if any.)	https://openaccess.thecvf.com/content/ICCV2021/papers/Xu_Rethinking_Self-Supervised_Correspondence_Learning_A_Video_Frame-Level_Similarity_Perspective_ICCV_2021_paper.pdf
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	A-SDF: Learning Disentangled Signed Distance Functions for Articulated Shape Representation
Author(s) Indicate the authors of the publication.	Jiteng Mu, Weichao Qiu, Adam Kortylewski, Alan Yuille, Nuno Vasconcelos, Xiaolong Wang.

Publication Date Indicate the date of publication.	Oct 2021
Publication Venue Indicate the journal, conference, or magazine name.	ICCV 2021
Keywords Enter at least three keywords for the publication.	3D vision, domain adaptation, implicit function
URL Enter the URL associated with this publication (if any.)	https://openaccess.thecvf.com/content/ICCV2021/papers/Mu_A-SDF_Learning_Disentangled_Signed_Distance_Functions_for_Articulated_Shape_Representation_ICCV_2021_paper.pdf
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Meta-Baseline: Exploring Simple Meta-Learning for Few-Shot Learning
Author(s) Indicate the authors of the publication.	Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, Xiaolong Wang.
Publication Date Indicate the date of publication.	Oct 2021
Publication Venue Indicate the journal, conference, or magazine name.	ICCV 2021
Keywords Enter at least three keywords for the publication.	Few-shot learning, meta-learning, image classification
URL Enter the URL associated with this publication (if any.)	https://openaccess.thecvf.com/content/ICCV2021/papers/Chen_Meta-Baseline_Exploring_Simple_Meta-Learning_for_Few-Shot_Learning_ICCV_2021_paper.pdf
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Video Autoencoder: self-supervised disentanglement of static 3D structure and motion
Author(s) Indicate the authors of the publication.	Zihang Lai, Sifei Liu, Alexei A. Efros, Xiaolong Wang.
Publication Date Indicate the date of publication.	Oct 2021
Publication Venue Indicate the journal, conference, or magazine name.	ICCV 2021
Keywords Enter at least three keywords for the publication.	Self-supervised learning, 3D vision, videos
URL Enter the URL associated with this publication (if any.)	https://openaccess.thecvf.com/content/ICCV2021/papers/Lai_Video_Autoencoder_Self-Supervised_Disentanglement_of_Static_3D_Structure_and_Motion_ICCV_2021_paper.pdf
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	“Wikily” Supervised Neural Translation Tailored to Cross-Lingual Tasks
Author(s) Indicate the authors of the publication.	Mohammad Sadegh Rasooli, Chris Callison-Burch, Derry Tanti Wijaya
Publication Date Indicate the date of publication.	Nov 2021
Publication Venue Indicate the journal, conference, or magazine name.	EMNLP 2021
Keywords Enter at least three keywords for the publication.	Machine Translation, Low resource setting, Crosslingual task
URL Enter the URL associated with this publication (if any.)	https://aclanthology.org/2021.emnlp-main.124/

Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Subspace Regularizers for Few-Shot Class Incremental Learning
Author(s) Indicate the authors of the publication.	Afra Feyza Akyürek, Ekin Akyürek, Derry Wijaya, Jacob Andreas
Publication Date Indicate the date of publication.	Oct 2021
Publication Venue Indicate the journal, conference, or magazine name.	Arxiv
Keywords Enter at least three keywords for the publication.	Few-shot learning, image classification
URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2110.07059
Comments Enter any relevant comments about this publication.	

Title Indicate the title of the publication.	Low-Resource Machine Translation Training Curriculum Fit for Low-Resource Languages
Author(s) Indicate the authors of the publication.	Garry Kuwanto, Akyürek, Afra Feyza, Isidora Chara Tourni, Siyang Li, Derry Wijaya
Publication Date Indicate the date of publication.	Nov 2021
Publication Venue Indicate the journal, conference, or magazine name.	Arxiv
Keywords Enter at least three keywords for the publication.	Machine Translation, low resource languages, code-mixing

URL Enter the URL associated with this publication (if any.)	https://arxiv.org/abs/2103.13272
Comments Enter any relevant comments about this publication.	

Full List of Most Recent Team Publications (2022-2023)

Title: A Broad Study of Pre-training for Domain Generalization and Adaptation
Author(s): Donghyun Kim, Kaihong Wang, Stan Sclaroff, Kate Saenko
Publication date: 2022/3/22
Publication venue (e.g. journal, conference or magazine name): arXiv
Publication keywords: pre-training, domain adaptation, domain generalization
URL: https://arxiv.org/abs/2203.11819
Abstract: Deep models must learn robust and transferable representations in order to perform well on new domains. While domain transfer methods (e.g., domain adaptation, domain generalization) have been proposed to learn transferable representations across domains, they are typically applied to ResNet backbones pre-trained on ImageNet. Thus, existing works pay little attention to the effects of pre-training on domain transfer tasks. In this paper, we provide a broad study and in-depth analysis of pre-training for domain adaptation and generalization, namely: network architectures, size, pre-training loss, and datasets. We observe that simply using a state-of-the-art backbone outperforms existing state-of-the-art domain adaptation baselines and set new baselines on Office-Home and DomainNet improving by 10.7\% and 5.5\%. We hope that this work can provide more insights for future domain transfer research.

Title: Better Quality Estimation for Low Resource Corpus Mining
Author(s): Muhammad Yusuf Kocyigit, Jiho Lee, Derry Wijaya
Publication date: 2022/3/15
Publication venue (e.g. journal, conference or magazine name): ACL Findings 2022
Publication keywords: low resource, parallel corpus mining, quality estimation, machine translation
URL: https://arxiv.org/abs/2203.08259
Abstract: Quality Estimation (QE) models have the potential to change how we evaluate and maybe even train machine translation models. However, these models still lack the robustness to achieve general adoption. We show that State-of-the-art QE models, when tested in a Parallel Corpus Mining (PCM) setting, perform unexpectedly bad due to a lack of robustness to out-of-domain examples. We propose a combination of multitask training, data augmentation and contrastive learning to achieve better and more robust QE performance. We show that our method improves QE performance significantly in the MLQE challenge and the robustness of QE models when tested in the Parallel Corpus Mining setup. We increase the accuracy in PCM by more than 0.80, making it on par with state-of-the-art PCM methods that use millions of sentence pairs to train their models. In comparison, we use a thousand times less data, 7K parallel sentences in total, and propose a novel low resource PCM method.

Title: Extending the WILDS Benchmark for Unsupervised Adaptation
Author(s): Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori Hashimoto, Sergey Levine, Chelsea Finn, Percy Liang
Publication date: April 2022
Publication venue (e.g. journal, conference or magazine name): ICLR

Publication keywords: distribution shifts, adaptation, unlabeled data
URL: https://arxiv.org/abs/2112.05090
Abstract: Machine learning systems deployed in the wild are often trained on a source distribution but deployed on a different target distribution. Unlabeled data can be a powerful point of leverage for mitigating these distribution shifts, as it is frequently much more available than labeled data and can often be obtained from distributions beyond the source distribution as well. However, existing distribution shift benchmarks with unlabeled data do not reflect the breadth of scenarios that arise in real-world applications. In this work, we present the WILDS 2.0 update, which extends 8 of the 10 datasets in the WILDS benchmark of distribution shifts to include curated unlabeled data that would be realistically obtainable in deployment. These datasets span a wide range of applications (from histology to wildlife conservation), tasks (classification, regression, and detection), and modalities (photos, satellite images, microscope slides, text, molecular graphs). The update maintains consistency with the original WILDS benchmark by using identical labeled training, validation, and test sets, as well as identical evaluation metrics. We systematically benchmark state-of-the-art methods that use unlabeled data, including domain-invariant, self-training, and self-supervised methods, and show that their success on WILDS is limited. To facilitate method development, we provide an open-source package that automates data loading and contains the model architectures and methods used in this paper. Code and leaderboards are available at https://wilds.stanford.edu.
Comments:

Title: Knowledge Based Template Machine Translation In Low-Resource Setting
Author(s): Zilu Tang, Derry Wijaya
Publication date: 2022/9/8
Publication venue (e.g. journal, conference or magazine name): arXiv preprint
Publication keywords: knowledge base, template machine translation
URL: https://arxiv.org/pdf/2209.03554.pdf
Abstract: Incorporating tagging into neural machine translation (NMT) systems has shown promising results in helping translate rare words such as named entities (NE). However, translating NE in low-resource setting remains a challenge. In this work, we investigate the effect of using tags and NE hypernyms from knowledge graphs (KGs) in parallel corpus in different levels of resource conditions. We find the tag-and-copy mechanism (tag the NEs in the source sentence and copy them to the target sentence) improves translation in high-resource settings only. Introducing copying also results in polarizing effects in translating different parts-of-speech (POS). Interestingly, we find that copy accuracy for hypernyms is consistently higher than that of entities. As a way of avoiding "hard" copying and utilizing hypernym in bootstrapping rare entities, we introduced a "soft" tagging mechanism and found consistent improvement in high and low-resource settings.
Comments:

Title: AugCSE: Contrastive Sentence Embedding with Diverse Augmentations
Author(s): Zilu Tang, Muhammed Yusuf Kocyigit, Derry Wijaya
Publication date: forthcoming
Publication venue (e.g. journal, conference or magazine name): AACL 2022
Publication keywords: data augmentation, sentence embedding, contrastive learning
URL: forthcoming
Abstract: Data augmentation techniques have been proven useful in many applications in NLP fields. Most augmentations are task-specific, and cannot be used as a general-purpose tool. In our work, we present AugCSE, a unified framework to utilize diverse sets of data augmentations to achieve a better, general purpose, sentence embedding model. Building upon the latest sentence embedding models, our

approach uses a simple antagonistic discriminator that differentiates the augmentation types. With the fine tuning objective borrowed from domain adaptation, we show that diverse augmentations, which often lead to conflicting contrastive signals, can be tamed to produce a better and more robust sentence representation. Our methods (code and data can be found at https://github.com/PootieT/AugCSE) achieve state-of-the-art results on downstream transfer tasks and perform competitively on semantic textual similarity tasks, using only unsupervised data.
Comments:

Title: Adapting Self-Supervised Vision Transformers by Probing Attention-Conditioned Masking Consistency
Author(s): Viraj Prabhu*, Sriram Yenamandra*, Aaditya Singh, Judy Hoffman.
Publication date: December 2022
Publication venue (e.g. journal, conference or magazine name): Neural Information Processing Systems (NeurIPS)
Publication keywords: domain adaptation, object recognition
URL: https://github.com/virajprabhu/PACMAC
Abstract: Visual domain adaptation (DA) seeks to transfer trained models to unseen, unlabeled domains across distribution shift, but approaches typically focus on adapting convolutional neural network architectures initialized with supervised ImageNet representations. In this work, we shift focus to adapting modern architectures for object recognition -- the increasingly popular Vision Transformer (ViT) -- and modern pretraining based on self-supervised learning (SSL). Inspired by the design of recent SSL approaches based on learning from partial image inputs generated via masking or cropping -- either by learning to predict the missing pixels, or learning representational invariances to such augmentations -- we propose PACMAC, a simple two-stage adaptation algorithm for self-supervised ViTs. PACMAC first performs in-domain SSL on pooled source and target data to learn task-discriminative features, and then probes the model's predictive consistency across a set of partial target inputs generated via a novel attention-conditioned masking strategy, to identify reliable candidates for self-training. Our simple approach leads to consistent performance gains over competing methods that use ViTs and self-supervised initializations on standard object recognition benchmarks.
Comments: Investigating integration into LEARN system

Title: AugCSE: Contrastive Sentence Embedding with Diverse Augmentation
Author(s): Zilu Tang, Muhammed Yusuf Kocyigit, Derry Tanti Wijaya.
Publication date: November 2022
Publication venue (e.g. journal, conference or magazine name): Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL 2022) and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)
Publication keywords: sentence embedding, data augmentation, contrastive learning, antagonistic discriminator
URL: https://github.com/PootieT/AugCSE
Abstract: Data augmentation techniques have been proven useful in many applications in NLP fields. Most augmentations are task-specific, and cannot be used as a general-purpose tool. In our work, we present AugCSE, a unified framework to utilize diverse sets of data augmentations to achieve a better, general purpose, sentence embedding model. Building upon the latest sentence embedding models, our approach uses a simple antagonistic discriminator that differentiates the augmentation types. With the fine-tuning objective borrowed from domain adaptation, we show that diverse augmentations, which often lead to conflicting contrastive signals, can be tamed to produce a better and more robust sentence

representation. Our methods achieve state-of-the-art results on downstream transfer tasks and perform competitively on semantic textual similarity tasks, using only unsupervised data

Title: Pic2Word: Mapping Pictures to Words for Zero-shot Composed Image Retrieval
Author(s): Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li , Chen-Yu Lee , Kate Saenko , Tomas Pfister
Publication date: 2023, June
Publication venue (e.g. journal, conference or magazine name): CVPR
Publication keywords: Zero-shot, Image Retrieval
URL: https://arxiv.org/pdf/2302.03084.pdf
Abstract: In Composed Image Retrieval (CIR), a user combines a query image with text to describe their intended target. Existing methods rely on supervised learning of CIR models using labeled triplets consisting of the query image, text specification, and the target image. Labeling such triplets is expensive and hinders broad applicability of CIR. In this work, we propose to study an important task, Zero-Shot Composed Image Retrieval (ZS-CIR), whose goal is to build a CIR model without requiring labeled triplets for training. To this end, we propose a novel method, called Pic2Word, that requires only weakly labeled image-caption pairs and unlabeled image datasets to train. Unlike existing supervised CIR models, our model trained on weakly labeled or unlabeled datasets shows strong generalization across diverse ZS-CIR tasks, e.g., attribute editing, object composition, and domain conversion. Our approach outperforms several supervised CIR methods on the common CIR benchmark, CIRR and Fashion-IQ.

Title: Prefix Conditioning Unifies Language and Label Supervision
Author(s): Kuniaki Saito · Kihyuk Sohn · Xiang Zhang · Chun-Liang Li · Chen-Yu Lee · Kate Saenko · Tomas Pfister
Publication date: 2023, June
Publication venue (e.g. journal, conference or magazine name): CVPR
Publication keywords: Vision-language contrastive learning
URL: https://arxiv.org/pdf/2206.01125.pdf
Abstract: Vision-language contrastive learning suggests a new learning paradigm by leveraging a large amount of image-caption-pair data. The caption supervision excels at providing wide coverage in vocabulary that enables strong zero-shot image recognition performance. On the other hand, label supervision offers to learn more targeted visual representations that are label-oriented and can cover rare categories. To gain the complementary advantages of both kinds of supervision for contrastive image-caption pre-training, recent works have proposed to convert class labels into a sentence with pre-defined templates called prompts. However, a naive unification of the real caption and the prompt sentences could lead to a complication in learning, as the distribution shift in text may not be handled properly in the language encoder. In this work, we propose a simple yet effective approach to unify these two types of supervision using prefix tokens that inform a language encoder of the type of the input sentence (e.g., caption or prompt) at training time. Our method is generic and can be easily integrated into existing VL pre-training objectives such as CLIP or UniCL. In experiments, we show that this simple technique dramatically improves the performance in zero-shot image recognition accuracy of the pre-trained model.

Title: Mind the Backbone: Minimizing Backbone Distortion for Robust Object Detection
Author(s): Kuniaki Saito, Donghyun Kim, Piotr Teterwak, Rogerio Feris, Kate Saenko
Publication date: 2023, March
Publication venue (e.g. journal, conference or magazine name):Arxiv
Publication keywords: Object Detection

URL: https://arxiv.org/abs/2303.14744
Abstract: Building object detectors that are robust to domain shifts is critical for real-world applications. Prior approaches fine-tune a pre-trained backbone and risk overfitting it to in-distribution (ID) data and distorting features useful for out-of-distribution (OOD) generalization. We propose to use Relative Gradient Norm (RGN) as a way to measure the vulnerability of a backbone to feature distortion, and show that high RGN is indeed correlated with lower OOD performance. Our analysis of RGN yields interesting findings: some backbones lose OOD robustness during fine-tuning, but others gain robustness because their architecture prevents the parameters from changing too much from the initial model. Given these findings, we present recipes to boost OOD robustness for both types of backbones. Specifically, we investigate regularization and architectural choices for minimizing gradient updates so as to prevent the tuned backbone from losing generalizable features. Our proposed techniques complement each other and show substantial improvements over baselines on diverse architectures and datasets.

Title: Selective Self-training by Probing Predictive Consistency for Unsupervised Domain Adaptation
Author(s): Viraj Prabhu, Shivam Khare, Deeksha Kartik, Judy Hoffman
Publication date: In submission
Publication venue (e.g. journal, conference or magazine name): IJCV
Publication keywords: Domain Adaptation
Abstract: Many existing approaches for unsupervised domain adaptation (UDA) focus on adapting under only data distribution shift and offer limited success under additional cross-domain label distribution shift. Recent work based on self-training using target pseudolabels has shown promise, but on challenging shifts pseudolabels may be highly unreliable and using them for self-training may cause error accumulation and domain misalignment. We propose Augmentation Consistency guided Self-training (Augco), a UDA algorithm that probes the reliability of a target instance based on its predictive consistency under random image transformations. Our algorithm then selectively increases model confidence on consistent target instances, while reducing it on inconsistent ones. AugCo achieves significant improvements over the state-of-the-art on 27/31 domain shifts for classification adaptation.Further, Augco is directly extensible to other practical settings -- to *source-free* UDA, wherein source data is no longer available due to memory or compute restrictions, and to adapting *semantic segmentation* models, achieving strong performance in each. Our code is available at https://github.com/virajprabhu/AUGCO.

Title: Bridging the Sim2Real gap with CARE: Supervised Detection Adaptation with Conditional Alignment and Reweighting
Author(s): Viraj Prabhu, David Acuna, Andrew Liao, Rafid Mahmood, Marc T. Law, Judy Hoffman, Sanja Fidler, James Lucas
Publication date: In submission
Publication venue (e.g. journal, conference or magazine name): ICML
Publication keywords: Domain Adaptation
Abstract: Sim2Real domain adaptation (DA) research focuses on the constrained setting of adapting from a labeled synthetic source domain to an unlabeled or sparsely labeled real target domain. However, for high-stakes applications (e.g., autonomous driving), it is common to have a considerable amount of human-labeled real data in addition to plentiful auto-labeled source data (e.g., from a driving simulator). We study this setting of *supervised* Sim2Real DA applied to 2D object detection. We propose Domain Translation via Conditional Alignment and Reweighting (CARE), a novel algorithm that systematically combines synthetic and real data by exploiting target labels to close the Sim2Real appearance and content gaps. We present an analytical justification of our algorithm and demonstrate strong gains over competing methods on standard benchmarks.

Title: VideoCutLER: Surprisingly Simple Video Instance Segmentation

Author(s): Xudong Wang, Ishan Misra, Ziyun Zeng, Rohit Girdhar, Trevor Darrell
Publication date: In submission
Publication venue (e.g. journal, conference or magazine name): ICCV
Publication keywords: Unsupervised Learning
Abstract: Existing approaches to unsupervised video instance segmentation typically rely on optical flow, and experience difficulties tracking small or divergent motions. We present VideoCutLER, a frustratingly simple method for unsupervised multi-instance video segmentation without using optical flow or training on video data. Our key insight is that using high-quality pseudo masks and a simple video synthesis method for model training is sufficient to enable the resulting video model to effectively segment and track multiple instances across video frames. We show the first competitive unsupervised learning results on the challenging YouTubeVIS-2019 benchmark, achieving 50.3% APvideo, surpassing the previous SOTA by over 10 times (50.0% vs. 4.8%). VideoCutLER is a strong pretrained model for supervised video instance segmentation, exceeding DINO by 18.7% on YouTubeVIS-2019.

Title: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models
Author(s): Jiarui Xu, Sifei Liu*, Arash Vahdat*, Wonmin Byeon, Xiaolong Wang, Shalini De Mello
Publication date: Accepted by CVPR 2023
Publication venue (e.g. journal, conference or magazine name): CVPR 2023
Publication keywords: Open-vocabulary Panoptic Segmentation
Abstract: We present ODISE: Open-vocabulary DIffusion-based panoptic SEgmentation, which unifies pre-trained text-image diffusion and discriminative models to perform open-vocabulary panoptic segmentation. Text-to-image diffusion models have the remarkable ability to generate high-quality images with diverse open-vocabulary language descriptions. This demonstrates that their internal representation space is highly correlated with open concepts in the real world. Text-image discriminative models like CLIP, on the other hand, are good at classifying images into open-vocabulary labels. We leverage the frozen internal representations of both these models to perform panoptic segmentation of any category in the wild. Our approach outperforms the previous state of the art by significant margins on both open-vocabulary panoptic and semantic segmentation tasks. In particular, with COCO training only, our method achieves 23.4 PQ and 30.0 mIoU on the ADE20K dataset, with 8.3 PQ and 7.9 mIoU absolute improvement over the previous state of the art.

List of Acronyms

- MLR: Multi-Label Recognition
- DualCoOp: Dual Context Optimization
- UDA: Unsupervised Domain Adaptation
- UDA: Unsupervised Domain Adaptation
- SENTRY: Selective Entropy Optimization via Committee Consistency
- DA: Domain Adaptation
- ViT: Vision Transformer
- SSL: Self-Supervised Learning
- AL: Active Learning
- Active DA: Active Domain Adaptation
- OSSL: Open-set Semi-Supervised Learning
- SSDA: semi-supervised domain adaptation
- CutLER: Cut-and-LEaRn
- ODISE: Open-vocabulary DIffusion-based panoptic SEgmentation
- MIL: Multiple Instance Learning
- EM: Expectation-Maximization
- NMT: neural machine translation
- ToT: Top-Down Tuning
- VPT: Visual Prompt Tuning
- LoRA: Low-Rank Adaptation of Large Language Models
- MLR: Multi-Label Recognition
- PACMAC: Probing Attention Conditioned Masking Consistency
- CLUE: Clustering Uncertainty-weighted Embeddings
- MME: Minimax-Entropy
- OVA: One-Vs-All
- VideoCutLER: Video Cut-synthesis-and-LEaRn
- SOTA: State-of-the-art