AFRL-RI-RS-TR-2023-182

TOWARDS LESS LABELS BY ACTIVE LEARNING, EXPLOITING UNLABELED DATA AND LEARNED AUGMENTATION (TOLEDA)

NEDERLANDSE ORGANISATIE VOOR TOEGEPAST-NATUURWETENSCHAPPELIJK ONDERZOEK TNO

OCTOBER 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-2023-182 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
NANCY A. SEPULVEDA
Work Unit Manager

/ S /
MATTHEW J. KOCHAN
Technical Advisor
Intelligence 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
OCTOBER 2023	FINAL TECHNICAL REPORT	AUGUST 2019	MAY 2023

4. TITLE AND SUBTITLE

TOWARDS LESS LABELS BY ACTIVE LEARNING, EXPLOITING UNLABELED DATA AND LEARNED AUGMENTATION (TOLEDA)

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8750-19-C-0514	N/A	DoD, DARPA

5d. PROJECT NUMBER	5e. TASK NUMBER	5f. WORK UNIT NUMBER
		R2U0

6. AUTHOR(S)

Klamer Schutte, Gertjan Burghouts, Wyke Pereboom, Maarten Kruithof

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

TNO
Oude Waalsdorperweg 63
2597 AK The Hague, Netherlands

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Air Force Research Laboratory/RIEA 525 Brooks Road Rome NY 13441-4505	AFRL/ RI	AFRL-RI-RS-TR-2023-182

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

Within the DARPA Learning with Less Labels (LwLL) Program, TNO has executed the TOLEDA: Towards Less Labels by Active Learning, Exploiting Unlabeled Data and Learned Augmentation. This report provides the findings of that research. The goal of the LwLL program was to reduce the number of labeled samples by a factor of 1000 (Phase I) to 1000000 (phase II). Our research aimed at these goals, i.e. using very few or no labels. They have developed a method to find clusters and use the cluster centers as initial labels. Due to the nature of clustering, such labels are representative and diverse. At the same time these clusters provide a good set of pseudo labels. This leads to a strong image classification system, needing only very few labeled samples (down to 1 labeled sample per class) to obtain fairly good results.

Another innovation is the use of semantic construction. Here they exploit a network pretrained on a widely available, large-scale dataset, in combination with a text embedding of the class labels to construct a new classifier. This provides a zero-shot capability to image classification and object detection. In addition, they have explored use of upcoming externally trained language-vision foundation models such as CLIP and GLIP. These provided very powerful capabilities as compared to the technologies developed within LwLL.

15. SUBJECT TERMS

Image classification, object detections, active learning, zero-shot learning

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	26

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

PREVIOUS EDITION IS OBSOLETE.

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

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 SUMMARY

Within the DARPA Learning with Less Labels (LwLL) Program, TNO has executed the TOLEDA: Towards Less Labels by Active Learning, Exploiting Unlabeled Data and Learned Augmentation. This report provides the findings of that research.

The goal of the LwLL program was to reduce the number of labeled samples by a factor of 1000 (Phase I) to 1000000 (Phase II). Our research aimed at these goals, i.e. using very few or no labels.

We have developed a method to find clusters and use the cluster centers as initial labels. Due to the nature of clustering, such labels are representative and diverse. At the same time these clusters provide a good set of pseudo labels. This leads to a strong image classification system, needing only very few labeled samples (down to 1 labeled sample per class) to obtain fairly good results.

Another innovation is the use of semantic construction. Here we exploit a network pretrained on a widely available, large-scale dataset, in combination with a text embedding of the class labels to construct a new classifier. This provides a zero-shot capability to image classification and object detection.

In addition, we have explored use of upcoming externally trained language-vision foundation models such as CLIP and GLIP. These provided very powerful capabilities as compared to the technologies developed within LwLL.

2 INTRODUCTION

The DARPA Learning with Less Labels program considers how to learn high performance models with very small amounts of labeled data samples. This report describes the results of efforts of the TNO TOLEDA team performing in that program.

2.1 TOLEDA Approach

In the original TOLEDA proposal, we stressed the use of the cognitive functions: *Adaptation, Association, Exploitation, Exploration* and *Judgement*. In phase I we have learned that there is great benefit in using *clustering*; that it is beneficial to have a different *embedding* for the initial *clustering* and the final *deep learning model*; that an initial clustering can yield quite beneficial *pseudo labels*; and that using a *transport matrix* does not properly scale to domain transfer cases with no one-to-one match between source and target domains. Yet we still see the same cognitive functions used in our updated approach.

We *Explore* and *Exploit* the target domain in the *initial clustering* module. We *Exploit* the target data using our *data augmentation*. We *Explore* and *Exploit* the source domain with the source selection module and through the use of source models in the *deep learning model*. We *Adapt* the source domains to the target domain by enforcing a common *deep learning model*. We *Associate* source domains and target domain with the *initial clustering,* based upon the *initial SimCLR embedding*. *Judgement* is applied in the *target model*.

2.1.1 Key Innovations

In our *deep learning model,* we force both source and target domain data to be classified by the same classifiers. This approach is inspired by the self-ensembling approach [2] incorporating a

student-teacher network [1], which seems to surpass other domain transfer and semi-supervised learning approaches in recent benchmarks. We investigate other Self-Supervised Learning (SSL) approaches as alternatives to the self-ensembling approach.

The key innovations of the project are:

- The use of initial clustering, where the cluster centers are expected to be representative and diverse, and thus the perfect samples to be queried.
- The use of the class identity of the cluster center samples as likely class identity for all cluster members, to be used as pseudo labels.
- The use of an embedding obtained by SimCLR [3] as input for the initial clustering.
- The use of multiple runs of UMAP[4]-based dimension reduction and clustering in a super-clustering process to obtain a more robust initial clustering.
- The use of a SWIN [5] transformer as a base network.
- The use of samples of source datasets within the deep learning model as domain transfer.
- The use of such source samples to train a model to generate pseudo labels.
- The use of semantic construction to construct image classifiers based on label-space text embedding distances, including using this to generate pseudo labels.
- The use of a combined loss function, combining a labeled-sample term, a pseudo label term, an SSL term, and a domain transfer term.

While we recognize that each single innovation might not reach the LwLL goal of reducing the need for labeled data, we found that the combination of innovations pursued here will lead to the LwLL goal.

Within the LwLL program we applied these key innovations for the image classification, object detection and video classification tasks. In addition to the standard program evaluations, we have applied these key innovations to semantic segmentation.

3 METHODS, ASSUMPTIONS, AND PROCEDURES

3.1 Image Classification – Phase I System

For image classification we achieved an improvement over state-of-the-art for semi-supervised learning techniques that use only a very few labeled samples. Without a per-class preselection, our technique 1) selects these few samples to annotate, and 2) it provides high quality pseudo labels, that are commonly used in semi-supervised learning to train the image classification model.

We apply clustering on a reduced embedding that is pretrained using contrastive learning on the ImageNet1k dataset. Contrastive frameworks learn representations by maximizing agreement between differently augmented views of the same image. Such a model results in an embedding where similar images are clustered together. The model maps the images in a high-dimensional space, which we reduce with a dimension reduction technique to improve clustering with k-means.

To increase the robustness of the clustering result, we repeat the clustering multiple times on randomly chosen subsets of the training data and combine these clusterings in so-called superclusters. Besides robustness, this also provides a consistency measure. This measure is used to select a mix of random and consistently clustered samples, for which we use the assigned cluster as pseudo labels.

The samples closest to the cluster centers will be diverse and representative samples and therefore very suitable for annotation. With these samples, we assign a label to each cluster and are used for pseudo labeling.

With the few selected labels and pseudo labels we train a CNN. We selected a ResNet 50 pretrained on the ImageNet1k dataset. For training we optimize a combined loss, where we weigh a cross-entropy loss on the pseudo labels and on the few true labels.

Figure 1 shows a diagram of our Phase I image classification approach.

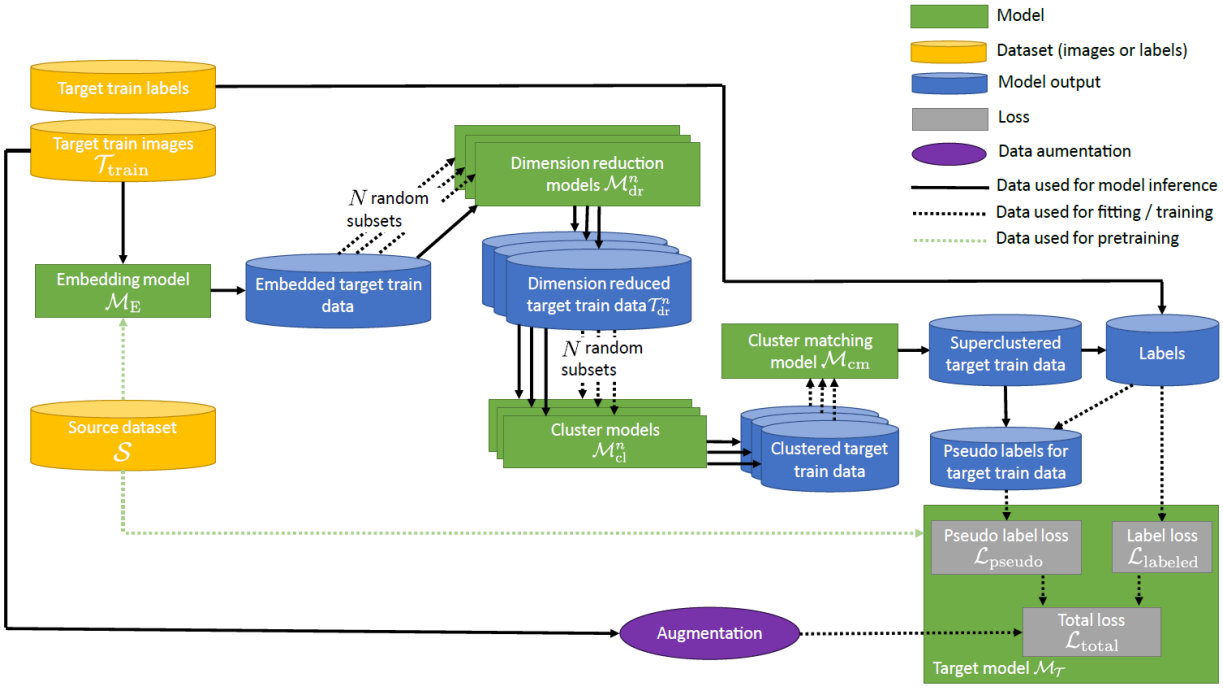

Figure 1 Diagram of our Phase 1 image classification method. The legend in the upper-right corner explains all.

3.2 Image Classification – Phase II System

The Phase II image classification system is an extension of the Phase I system. The sub-sections below will discuss pseudo labels by domain transfer, pseudo labels by semantic construction, combination of pseudo labels, SWIN transformer, and using ImageNet-22k as base dataset.

3.2.1 Pseudo Labels by Domain Transfer

For domain transfer we use overlapping source samples. As source datasets we use datasets on the whitelist as well as the base dataset in case of adapt case. Within the source datasets, we select these classes whose names map to names of the classes in the target dataset, and as overlapping source samples we take all samples from these source classes. Note that some translation of the text labels for these classes has to be performed, including translating the wordnet identifiers of ImageNet to their normal text equivalents.

For the Phase I system we had an extra loss term in the loss function which was trained such that the prediction on a overlapping source sample matched its class ID.

For the Phase II system, we instead trained a classifier on the overlapping source samples and used that classifier to predict source domain pseudo labels using the unlabeled target data samples.

3.2.2 Pseudo labels by Semantic Construction

The basic idea behind semantic construction is explained in Figure 2. This proves an effective way to create a zero-shot classifier. In the Phase II system, we used this zero-shot classifier to generate pseudo labels, by running this classifier on the unlabeled target data samples.

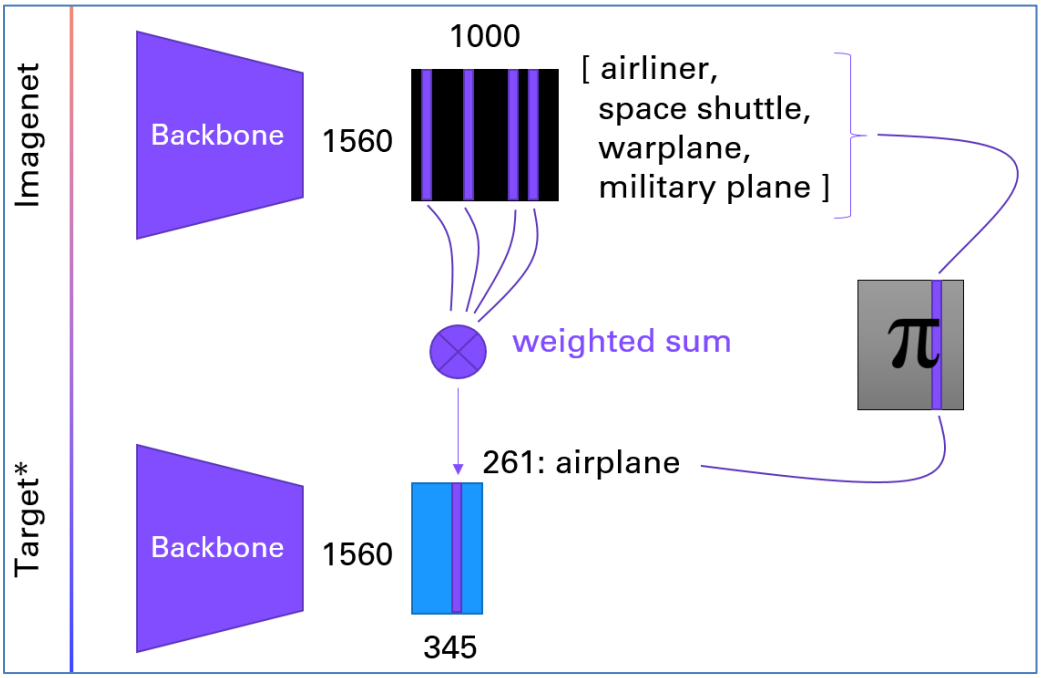

Figure 2 Semantic construction. The target class name "airplane" is matched in a text embedding space to the class names of a source dataset; subsequently the weights and biases in the final classifier layer are linearly combined according to the textual match.

3.2.3 Combination of Pseudo Labels

In the Phase II system, we have three different sources of pseudo labels: those of domain transfer, those of semantic construction, and those of clustering (as was used in the Phase I system). The combination of pseudo labels is critical to performance of the final system. In the design of the combination, we have tried to use the best pseudo labels of each pseudo label source, and to select pseudo labels in order accordingly to expected quality of the source.

For the domain transfer pseudo labels and the semantic construction pseudo labels, the best pseudo labels are selected using the softmax output as provided by the classifier generating these pseudo labels. For the clustering pseudo labels, the best pseudo labels are selected by their consistency over multiple clusters within a supercluster.

We have combined pseudo labels according to the following strategy:

1. Select the best pseudo labels of the domain transfer.

2. For those classes which are under-represented in that set of pseudo labels, add the best pseudo labels from semantic construction. Note to include all classes for which no direct match is present in the source data.
3. Subsequently, add the best clustering pseudo labels for classes underrepresented in the set of pseudo labels.

3.2.4 SWIN Transformer

In the Phase II system we have used a SWIN-B network in favor of the ResNet-50 used in the Phase I system, as it proved to have superior performance.

3.2.5 Using ImageNet-22k as Pretraining Dataset

For the Phase II system we have used ImageNet-22k instead of ImageNet-1K as pretraining dataset, assuming this would provide better generalization to new classes.

In addition, we have used ImageNet-22k pretrained models to generate the embeddings as used for clustering as well as for semantic construction. This is especially important for semantic construction, as ImageNet-22k provides a richer vocabulary and thus better matches.

As no SimCLRv2 model pretrained on ImageNet-22k was available, we have used a BIT-M [6] model instead to generate the embeddings.

4 RESULTS AND DISCUSSION

4.1 Results on LwLL Development Tasks

In this section, we will describe the results on our Phase-I and Phase-II system on the development tasks as defined in the LwLL program.

Figure 3 provides the results for the Domain_net-real and Domain_net-sketch tasks. Domain net [7] is a dataset aimed for Unsupervised Domain Adaptation. In addition, the types of classes are quite similar to the Imagenet [8] dataset. It can be seen that both the Phase I and Phase II systems outperform the JPL baseline. For the Phase I system, the UDA case, with no labeled data, performed better than the few-labels per class cases. This was likely due to more reliance on the domain transfer in the UDA case. This behavior is corrected in the Phase II system, where the few labels per class cases outperform the UDA case. Also, the Phase II system clearly outperforms the Phase I system.

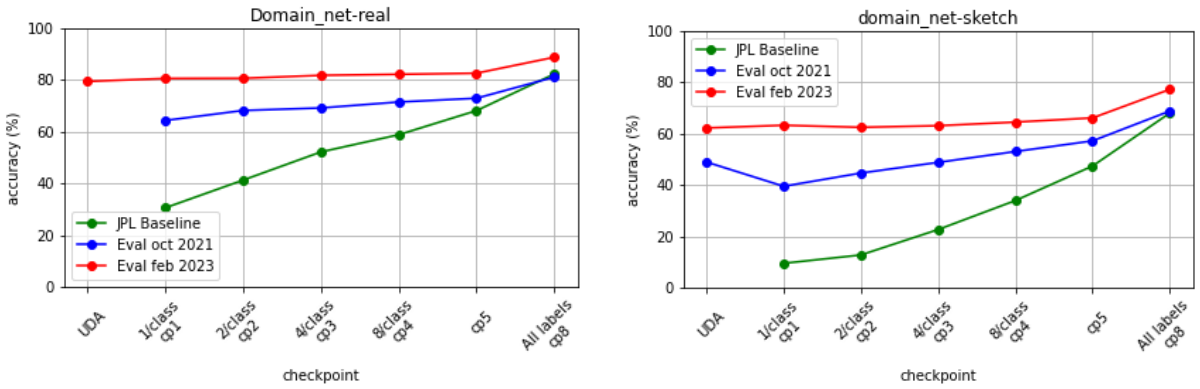

Figure 3 Results on the Domain_net-real and Domain_net-sketch datasets for the Phase I (Oct 2021) and Phase II (Feb 2023) systems

Figure 4 provides the results for the CIFAR100 and MNIST datasets. For CIFAR100 the results are comparable to the Domain_net results presented above. However, it can be seen that for more that 2 labels class, the performance slightly drops; most likely this is the case due to the method of combining the different sets of pseudo labels, which was optimized for the few-label case. Performance for the Phase-II is probably slightly lower as this system was not optimized for small images, where the Phase-I system had a dedicated 32x32 Wide Resnet [9] pretrained network. For MNIST we see a slightly lower performance for the Phase-II compared to the Phase-I system; we expect that is due to the rather worse match between the semantics of MNIST (digit numbers) and the classes of ImageNet as used in the semantic construction. This also would explain the result of the UDA case for MNIST; here the result is just above random, indicating that the semantic knowledge used here does not make very much sense.

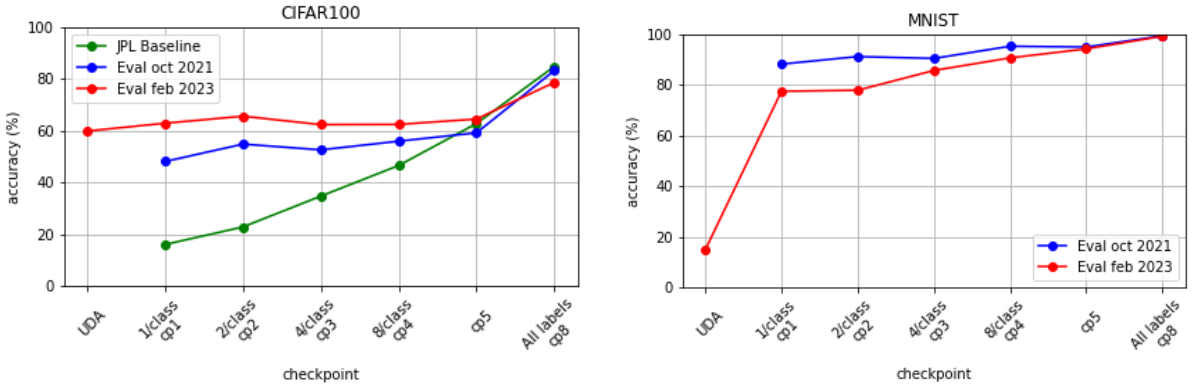

Figure 4 Results on the CIFAR100 and MNIST datasets for the Phase I (Oct 2021) and Phase II (Feb 2023) systems

Figure 5 provides results on the MARS_surface_images dataset. Here we can see that for the Phase I system, our performance is slightly worse than the JPL baseline – obviously this dataset is not consistent with our assumptions. Most likely the similarity to ImageNet, containing only Earth-

found objects, is too small for these images from Mars. In the evolution from Phase I to Phase II, we strived to improve where we already were good; in that light it is not surprising to see the Phase II system fail on this dataset. However, we find it surprising that with more and more labels the performance of the Phase II system does not increase. We expect this to relate to our system not converging to a decent solution, for which a different parameterization of the hyper parameters of the system might be required.

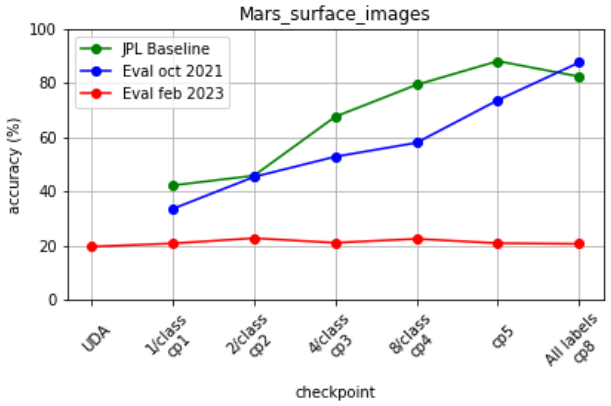

Figure 5: Results on Mars_surface_images for the Phase I (Oct 2021) and Phase II (Feb 2023) systems.

4.2 Results on LwLL Program Evaluations

Figure 6 shows the results of the Phase I system as provided for the October 2021 LwLL program evaluation. To save computation time, we did not provide results for #5 and #6. For 5 out of 6 tasks we performed best for the lowest checkpoints.

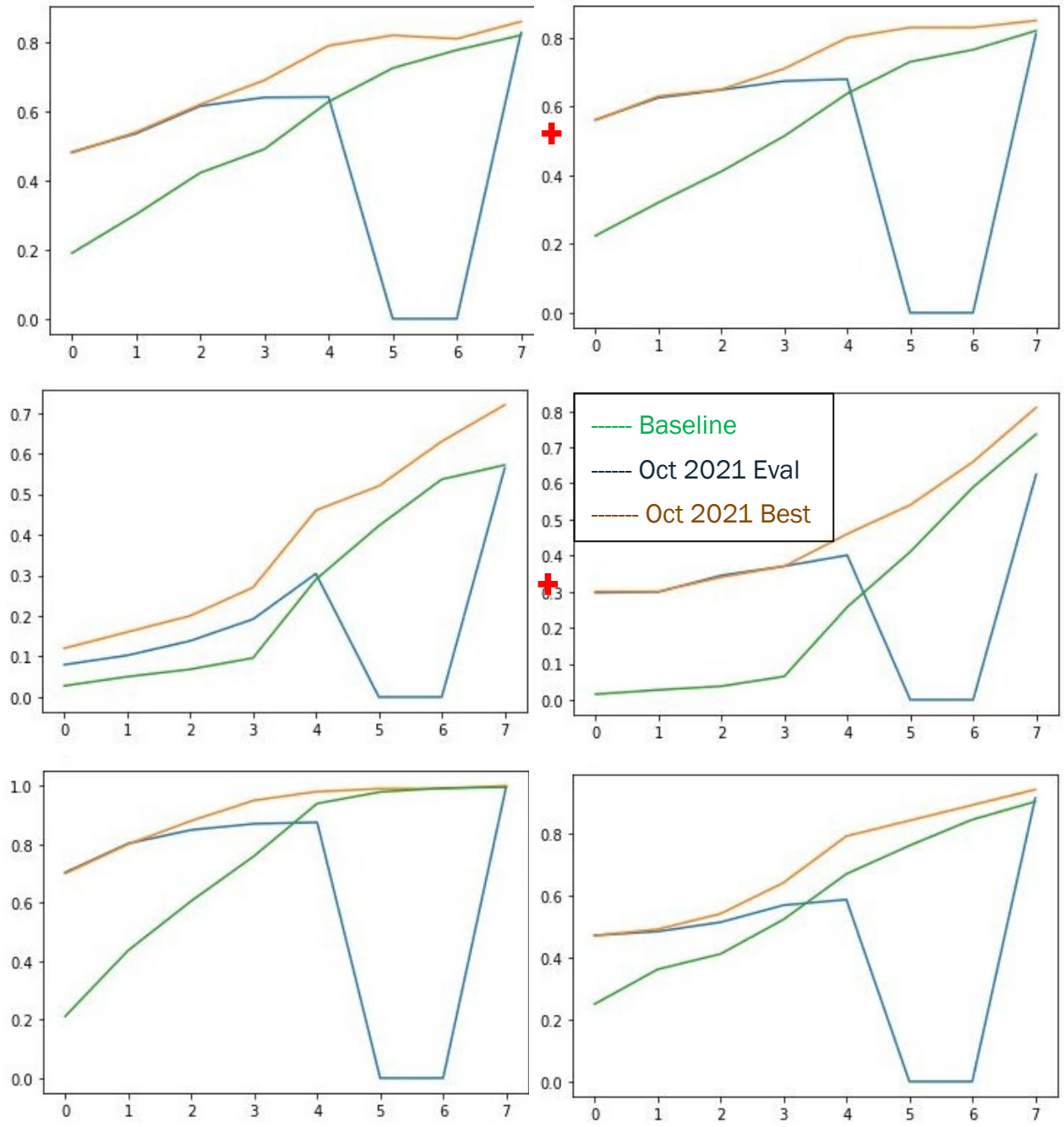

Figure 6 Results of the Phase I (October 2021) LwLL program evaluation. Top to bottom shows IC1, IC2 and IC3. Left shows the base task, and right the adaptation task. The red crosses indicate the UDA results.

Figure 7 the results of the Phase II system in the final LwLL program evaluation. For convenience, the results of the Phase I system are included in these graphs. Note that the checkpoint numbers are different between Figure 6 and Figure 7 (counting from 0 versus counting from 1.)

It came to us as a big disappointment that in general our performance decreased from our Phase I to our Phase II system. While we did see significant improvements on the development data, we did not see that on the program evaluation. Part of this might be due to a DDP style multi-GPU training used in the program evaluation, driven by our switch to a SWIN-B base network, and the program requirement to finish a task within a week. Another problem might be that the text labels belonging to the classes in IC tasks do not match the textual form we expect. Since the program evaluations are blind evaluations, we cannot confirm what is the reason for this unexpected drop in performance.

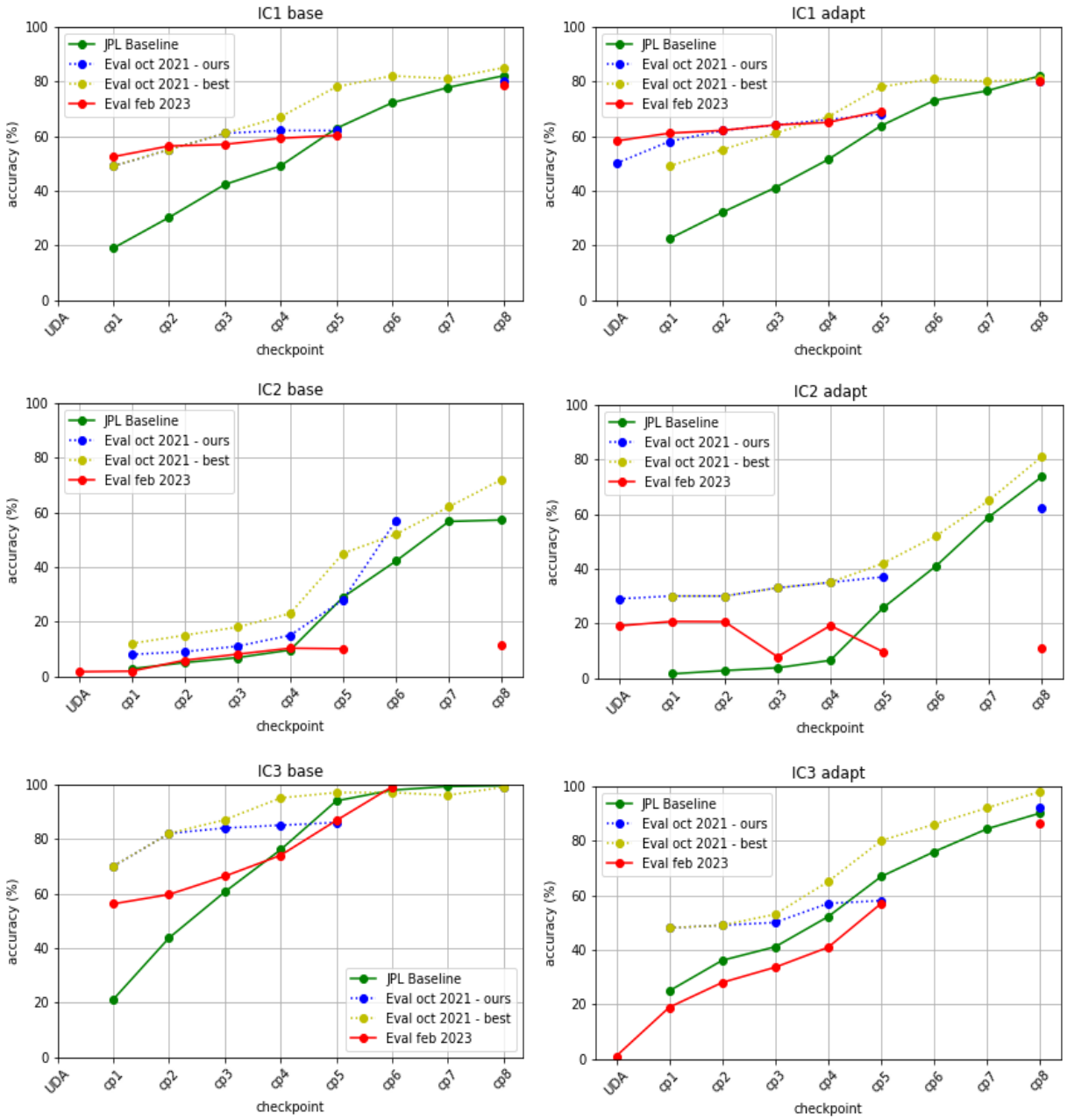

Figure 7 Results of the Phase II system on the final LwLL program evaluation Feb 2023.

4.3 Results on Other Datasets

Table 1 provides results on a k-shot transfer case form ImageNet1K to Domain_net-real and CUB200 [10]. These results are taken from [11]. In this table, Naïve Transfer is a method where a network is only trained on the k labelled samples. SimPLE is the method described in [12]. CLaP is the Phase I system, although here no domain transfer has been applied. In line with the results shown in Figure 3, the performance of our system outperforms other methods for Domain_net-real. Accuracy reported in Table 1 is lower than the accuracy for the Phase I system in Figure 3, as their domain transfer was used – which for the Domain_net datasets is quite valuable. In addition, it shows the Phase I system outperforms other methods for the CUB200 dataset.

Table 1: Classification accuracy with standard deviation for k-shot transfer from ImageNet1K

	DomainNet real		
Method	$k = 1$	$k = 2$	$k = 4$
Naive transfer	26.21 ± 0.04	37.25 ± 0.06	47.41 ± 0.09
SimPLE	33.26 ± 0.72	49.80 ± 0.24	62.95 ± 0.84
CLaP	56.98 ± 1.00	62.93 ± 0.66	65.95 ± 0.40
	CUB200		
Method	$k = 1$	$k = 2$	$k = 4$
Naive transfer	13.25 ± 0.14	18.69 ± 0.10	25.88 ± 0.06
SimPLE	16.56 ± 0.45	24.72 ± 0.73	49.82 ± 1.06
CLaP	36.00 ± 0.84	43.63 ± 1.03	53.45 ± 0.54

Table 2 shows the results of CLaP – the Phase I system without domain transfer – to the BSCD-FSL benchmark. These results are taken from [11]. We apply CLaP on the BSCD-FSL benchmark introduced in [13] to show the added value of CLaP when domains differ. There are four datasets in the benchmark, all of which come from very different domains: ChestX (diagnosing chest X-rays), ISIC2018 (identifying melanoma from images of skin lesions), EuroSAT (predicting land-use from satellite images) and CropDiseases (recognizing plant diseases in leaf images), with a decreasing similarity to ImageNet1k. These four datasets are interesting because they each differ more from the source dataset with ChestX being most different and CropDisease least different but still challenging. We evaluate five-way k-shot classification tasks (five classes and k examples per class) with k = 1, 5 and report the test accuracy mean and standard deviation over 200 few-shot experiments per task. We compare our results to the techniques reported in [14], [15], [16]. [14] include naive transfer which trains a CNN to classify the base dataset and uses the resulting representation to learn a linear classifier on the target dataset.

To measure the effect of the different parts of our method, we performed an ablation study on the datasets of the BSCD-FSL benchmark. For this ablation study, we performed five configurations of our method:

1) only super-clustering using random labels per class,
2) super-clustering using the cluster centers as class labels,

3) CLaP with random labels per class and training only the last layer,
4) CLaP with training the last layer on the given labels only without pseudo-labels, and
5) CLaP with only the last layer trainable.

This experiment will show the effect of training a network on the pseudo-labels (4 and 5), of using the cluster centers as labeled samples (1, 2, 3, and 4), and see the effect of fully training the whole network instead of only the last layer (5).

Results: The results of naive transfer, STARTUP [14], Dynamic Distillation [15], SSL [16] and the ablation study of our method applied to datasets of the BSCD-FSL benchmark are shown in Table 2. For different datasets, other configurations of CLaP performs best. For datasets that are more similar to ImageNet1k (EuroSAT and CropDisease), the full CLaP method performs best. For the other two datasets, CLaP without pseudo labels performs better than the full CLaP method. This is because the clustering accuracy on ChestX and ISIC is low due to the difference between these datasets and ImageNet1k used to train the embedding network. Thus, the accuracy pseudo-labels for these datasets will be low. Training a network using these low-quality pseudo-labels leads to a decrease in performance.

Table 2: Classification accuracies over 200 5-way few-shot transfer experiments from ImageNet1K to four domain of the BSCD-FSL benchmark using ResNet18

	ChestX		**ISIC**	
Methods	$k = 1$	$k = 5$	$k = 1$	$k = 5$
Transfer with random labels per class*	21.97 ± 0.39	25.85 ± 0.41	30.27 ± 0.51	43.88 ± 0.56
STARTUP	23.03 ± 0.42	27.24 ± 0.46	31.69 ± 0.59	46.02 ± 0.59
Dynamic Distillation	23.38 ± 0.43	28.31 ± 0.46	34.66 ± 0.58	49.36 ± 0.59
SSL (best)	22.75 ± 0.41	28.80 ± 0.49	36.69 ± 0.66	49.26 ± 0.64
Super-clustering with random labels per class	18.25 ± 8.56	22.11 ± 4.88	27.64 ± 17.64	31.80 ± 11.18
Super-clustering with cluster center labels	34.41 ± 7.46	29.50 ± 3.59	60.36 ± 14.70	58.88 ± 16.44
CLaP with random labels per class*	21.06 ± 7.72	23.97 ± 4.54	28.51 ± 15.93	37.90 ± 10.77
CLaP without pseudo-labels*	33.10 ± 8.48	32.48 ± 5.77	59.79 ± 13.79	62.05 ± 14.80
CLaP*	32.21 ± 6.69	28.29 ± 4.21	56.22 ± 10.11	59.10 ± 13.27
CLaP	32.37 ± 6.63	27.38 ± 3.71	55.26 ± 11.75	57.27 ± 13.15
	EuroSAT		**CropDisease**	
Methods	$k = 1$	$k = 5$	$k = 1$	$k = 5$
Transfer with random labels per class*	66.08 ± 0.81	85.58 ± 0.48	74.17 ± 0.82	92.46 ± 0.42
STARTUP	73.83 ± 0.77	89.70 ± 0.41	85.10 ± 0.74	96.06 ± 0.33
Dynamic Distillation	73.14 ± 0.84	89.07 ± 0.47	82.14 ± 0.78	95.54 ± 0.38
SSL (best)	84.30 ± 0.73	94.12 ± 0.32	91.00 ± 0.76	97.46 ± 0.34
Super-clustering with random labels per class	64.11 ± 17.58	75.66 ± 10.46	76.23 ± 18.96	91.88 ± 7.89
Super-clustering with cluster center labels	86.28 ± 8.41	89.93 ± 5.38	91.82 ± 8.35	97.69 ± 3.52
CLaP with random labels per class*	61.20 ± 12.90	77.27 ± 7.05	76.86 ± 17.00	92.05 ± 6.66
CLaP without pseudo-labels*	61.88 ± 9.75	71.18 ± 8.55	79.20 ± 9.18	86.63 ± 6.15
CLaP*	74.96 ± 7.45	81.06 ± 6.11	89.78 ± 7.71	95.78 ± 3.53
CLaP	86.90 ± 8.35	91.43 ± 4.71	92.43 ± 8.03	98.70 ± 2.11

4.4 Discussion

In the results above, we have presented a promising approach to image classification with less labeled samples. We see areas where the Phase II system clearly outperforms the Phase I system, as well as the other way around. We expect that actual performance of system using less labeled samples is dependent on the inductive biases used by a system, and the match of such biases against an actual problem task. This is true especially for the blind evaluation, such as performed in the LwLL program evaluations. It is – during development and submission time of such a system – unknown in how far inductive biases are true. Yet in the few labels domain, as in the core of the LwLL program, we need an extensive generalization capability of the system – which in our opinion is rooted in inductive biases.

Our Phase II system consisted of the components: embedding; clustering; pseudo labels by clustering; pseudo labels by semantic construction; pseudo labels by domain transfer; combination of pseudo labels; network backbone using SWIN-B. Below we will discuss these components, based on all our experiments.

Embedding: In the Phase-II system we have embedded the data using BIT-M network pretrained on ImageNet-22K. We expect that pretraining on ImageNet-22K is a strong choice, especially when the domain of ImageNet-22K spans the target domain. It is unclear whether a BIT-M network is a proper choice compared to the SSL SimCLRv2 network. For future work, one might explore the use of clustering a SSL network pretrained on the target data.

Clustering: The results shown in Section 4.3 show the effectiveness of the clustering approach. We expect its sample and pseudo label selection to be extremely powerful especially for a small amount of labeled samples. Note that for larger amount of labeled samples, we no longer expect to find very deterministic substructures in the embedded data, and thus less value of using a clustering mechanism. While using a clustering mechanism to find which samples to label is, from the operator standpoint, much easier than having to provide balanced seed labels, it has the disadvantage that we might provide unbalanced amounts of labeled samples, something which is hard to handle for the backend networks.

Pseudo labels by clustering: these pseudo labels seem quite effective. Note that in our system, we switch off the use of pseudo labels by clustering when we expect that the clustering has failed, as we observe by the amount of classes found when clustering the data

Pseudo labels by semantic construction: this seems an effective method to obtain pseudo labels as long as the pretraining data for the embedding (we used ImageNet22K) properly spans the target domain.

Pseudo labels by domain transfer: this is a very effective method to obtain pseudo labels when we have a large overlap in classes between source and target AND the target data has a similar appearance to the source data.

Combination of pseudo labels: for the Domain_net data sets we have seen that our combination strategy is effective. However, when the relative performance of the pseudo labels by clustering / semantic construction / domain transfer is different compared to what we did observe for Domain_net we might need a different strategy to combine these pseudo labels.

Network backbone using SWIN-B: it seems that the SWIN-B network typically outperforms the ResNet-50 (or other ResNets) we have used before. However, it comes with a larger computational cost, which makes the comparison rather unfair.

5 OBJECT DETECTION

For object detection, we aimed for a combination of two SOTA methods. Decoupled Faster R-CNN (DeFCRN) [17] is a SOTA method for few-shot object detection and therefore we are integrating it in our LwLL pipeline. It is based on Faster R-CNN [18] with an additional decoupling of localization and classification. We have confirmed that the decoupling makes sense on a 1 label / class setting on VOC splits: it outperforms Faster R-CNN. We validated it also on other datasets such as DOTA and LVIS. We selected these datasets because they are imbalanced and contain small objects and under different viewpoints.

For initialization, we use the semantic construction in the same way as for image classification, except now for two heads resp. for classification and localization. Our hypothesis is that it should help with really low label budgets.

Our approach is detailed in this figure:

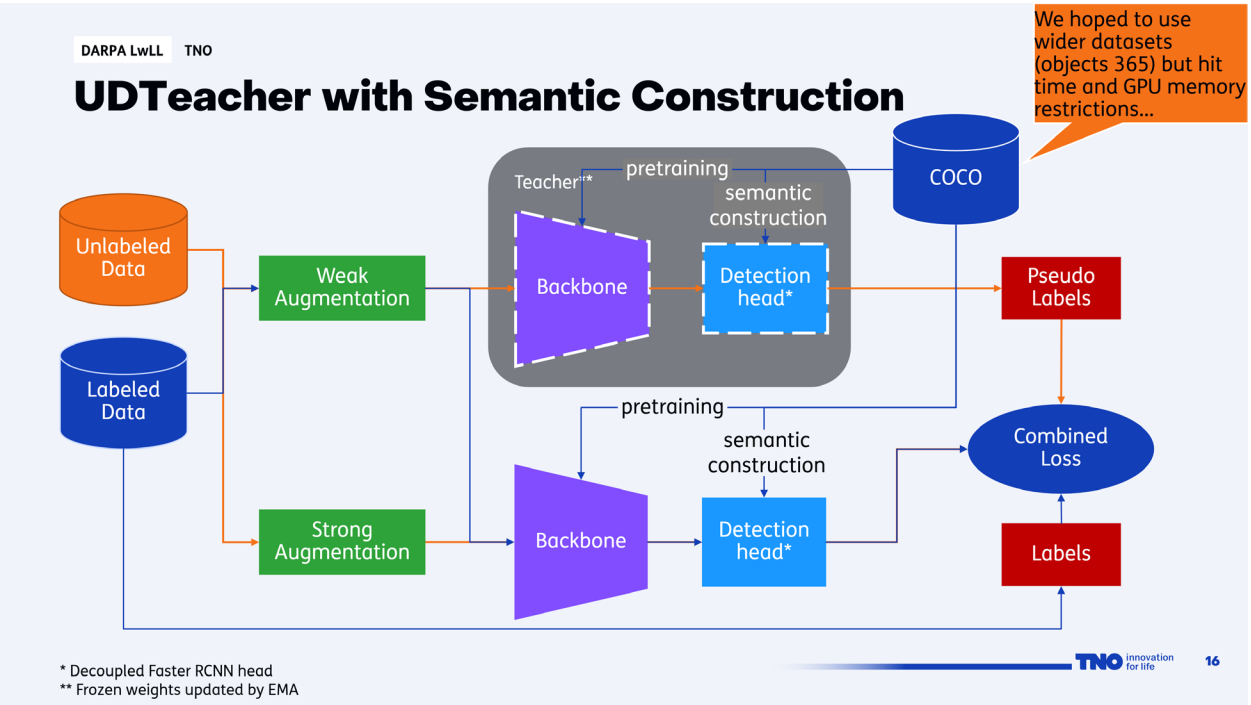

Figure 8: Overview of our Object Detection system

In the results (Figure 9), we see that our strategy provides a performance gain if the target domain is similar in the type of images and classes, compared to the source domain (COCO). The results:

- Outperforms UBTeacher [19] on LVIS and open-images-small
- Outperforms Decoupled on LVIS
- Semantic construction improves results on open-images-small

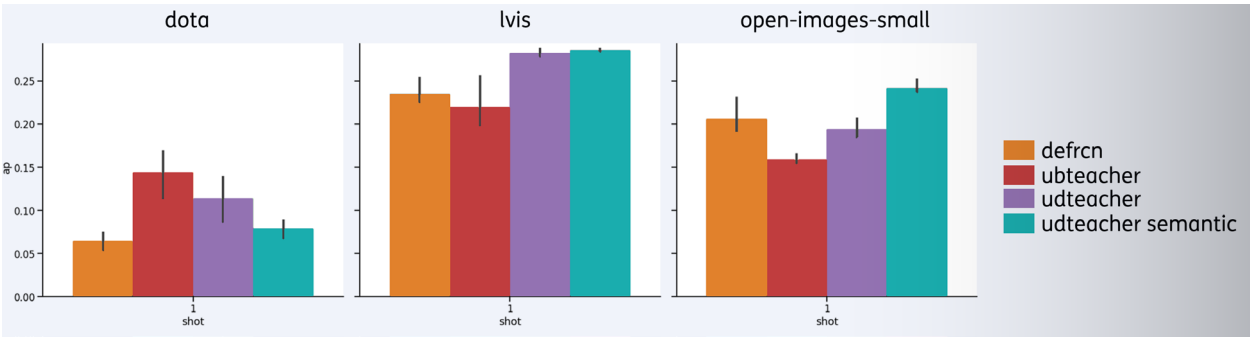

Figure 9: In-house evaluation results of our Object Detection system

Unfortunately, the results of Eval-3 of our object detection system were disappointing (consistently scoring lower than the baseline). In year 1 and year 2, we took small steps for incremental improvement for few shot object detection. In year 3 we took a leap: combining two SOTA methods for few shot object detection. Unfortunately, this high risk turned out into disappointment for the year 3 evaluation.

After Eval-3, we have explored another route for object detection, i.e. language-vision models: GLIP v2 [20]. It has potential for military vehicles based on a textual description:

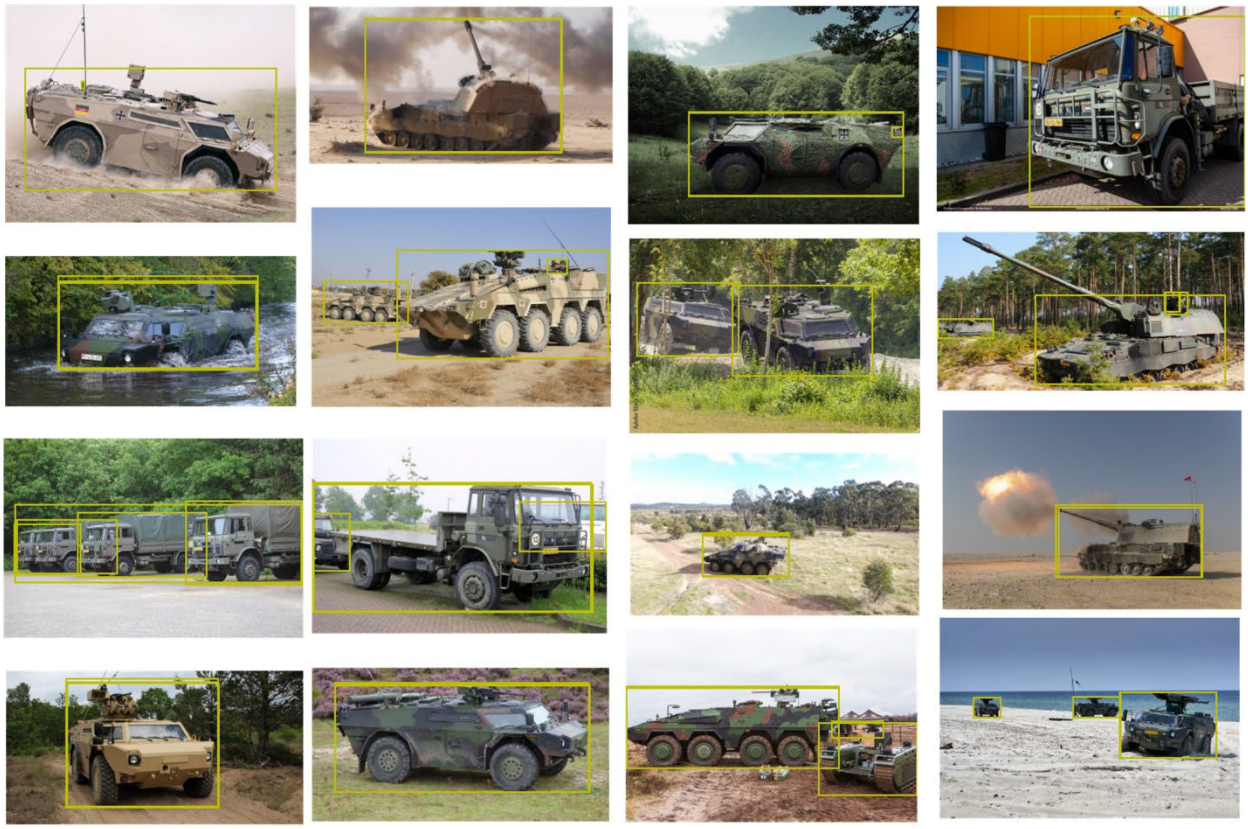

Figure 10: Zero-shot object detection results using GLIP v2.

6 TRANSITION

6.1 Low-shot Image Retrieval

There was a transition opportunity about Low-shot Image Retrieval. TNO achieved very good results. Results on the task from the LwLL API:

*Table 3: Low-shot Image Retrieval results **

Session_id	CHK_1	CHK_2	CHK_3	CHK_4	CHK_5	CHK_6	CHK_7	CHK_8
pQtJDQ49jFadSsFUBgVO	0.84	0.93	0.96	0.96	0.95	0.95	0.94	0.95

** metrics are taken from LwLL API: precision @ 50 (also reported in DARPA spreadsheet).*

For comparison, CLIP [21] zero-shot: 0.86. Baseline at CHK_1: 0.61 (ours 0.84), CHK_2: 0.80 (ours 0.93). At higher checkpoints we may improve by combining with semantic construction or finetune CLIP (CoCoOp/VPT).
Our method was to leverage both the language and the visual information via fusion on top of CLIP:

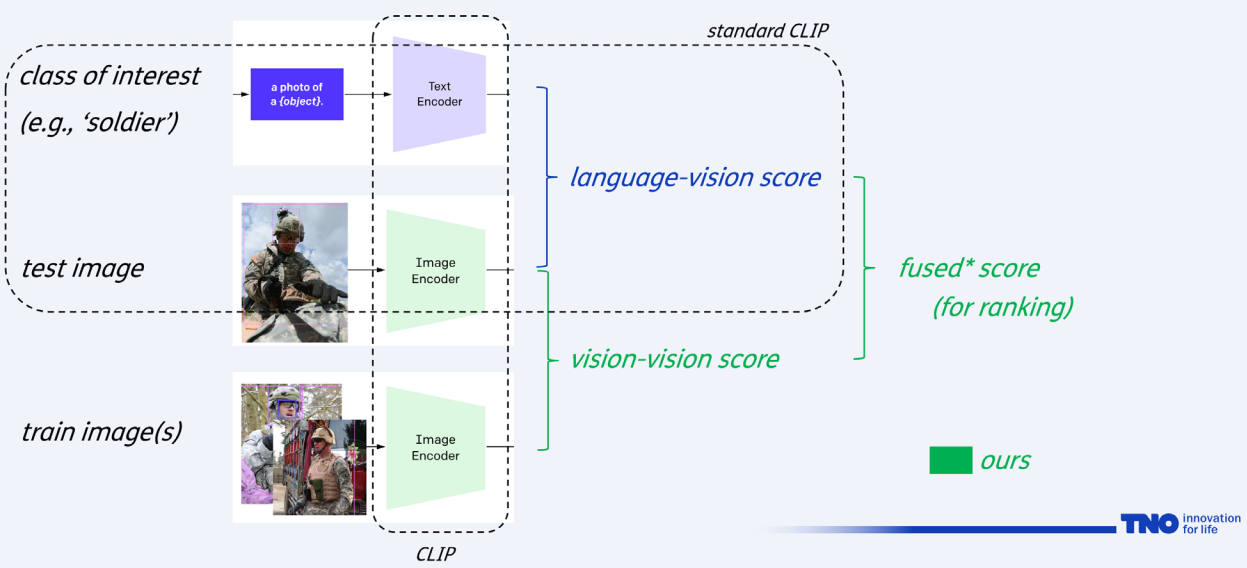

Figure 11: approach for Low-shot Image Retrieval

It is an interesting use-case with various applications for military analysts (confirmed by our MOD). We are very much interested to continue this with DARPA's client.

6.2 CARVE Proposal to IWTSD

We have prepared the CARVE proposal for the Irregular Warfare Technical Support Directorate (IWTSD). This proposal was using the image clustering by content methods as we did develop in

LwLL. A quadchart was submitted April 2021, a whitepaper was submitted in June 2021, a full proposal July 2021, and an Statement of Work was agreed in August 2021. However, no contract was awarded due to lack of funds.

6.3 Other Transition Efforts:

- Delivered a proof-of-concept regarding point 1, for image analysts of our (Netherlands) MOD. They believe that such developments are key to handle the huge amounts of images to be analyzed.
- Continuation of research on low-label regime in European Defence Fund project Faradai – including transition to European MODs and Defence Industry in that consortium.

7 CONCLUSIONS AND REFLECTION

Our conclusions and reflection on the LwLL program:

7.1 External Developments

It is hard to perform well on datasets when nothing is known beforehand. This is very different from our other research with known datasets. LwLL is a relevant setup that forces to be generic without overfitting. Yet some metadata about viewpoint, size distribution of objects, type of domain gap, will be helpful.

Good classification models, pretrained on incredibly large datasets, appeared (SimCLR, DINO, SWIN, CLIP, etc.). Using these as starting point for zero- and few-shot modelling is often more effective than clever tricks in the learning itself. E.g. CLIP is often hard to beat with zero labels and sometimes even with few labels. The same happened for object detection with the arrival of models such as GLIP v2.

7.2 Scientific Insights

Selecting the right images to be labelled is an important matter.

We did not find active learning methods to be very helpful beyond random.

However, clustering showed to be effective, providing diverse and representative first labels – improving performance over using seed labels. Even for object detection, while the object does not cover the whole image. The quality heavily depends on the quality of the embeddings.

Pseudo-labels are effective for learning with few labels. Clustering is a simple yet effective way to obtain them. Complementary sources are helpful to obtain good pseudo labels: semantic construction, domain transfer.

Domain transfer is very powerful when few target labels are available. Effective domain transfer can be obtained using student-teacher and pseudo label approaches. We expect this to include use of simulated data.

For zero/few-shot learning, it is helpful to transfer prior knowledge from seen classes to unseen classes via learning of a compositional object-attribute graph or semantic construction.

7.3 Goals and Future

Domain transfer is very powerful when few target labels are available – meeting the program goal for some data sets. Effective domain transfer can be obtained using student-teacher; pseudo label approaches; and by exploiting huge models like CLIP.

We expect that for problems not currently meeting the program goals there might be solutions involving simulation. However, an unsolved problem there is how to properly address the domain gap between real and simulated data.

Another problem in using data with few labeled samples is how to assure that a given model indeed meets performance in operational conditions – where those conditions might even deviate from the few labeled samples available.

8 REFERENCES

[1] Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, Antti Tarvainen, Harri Valpola, 2017

[2] Self-ensembling for visual domain adaptation, Geoffrey French, Michal Mackiewicz, Mark Fisher, ICLR 2018

[3] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029, 2020

[4] Leland McInnes. Umap: Uniform manifold approximation and projection for dimension reduction, 2018

[5] Z. Liu and Y. Lin and Y. Cao and H. Hu and Y. Wei and Z. Zhang and S. Lin and B. Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, 2021 IEEE/CVF International Conference on Computer Vision (ICCV)

[6] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, Neil Houlsby, Big Transfer (BiT): General Visual Representation Learning. ECCV 2020.

[7] Peng, Xingchao, Bai, Qinxun, Xia, Xide, Huang, Zijun, Saenko, Kate, Wang, Bo , Moment matching for multi-source domain adaptation, Proceedings of the IEEE International Conference on Computer Vision , 2019

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV), Vol. 115, pp. 211-252, 2015

[9] Sergey Zagoruyko and Nikos Komodakis, Wide Residual Networks, BMVC 2016

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

[11] Wyke Huizinga, Maarten Kruithof, Gertjan Burghouts and Klamer Schutte, Efficient Transfer by Robust Label Selection and Learning with Pseudo-Labels, accepted for publication ICIP 2023

[12] Z. Hu, Z. Yang, X. Hu, and R. Nevaita, “SimPLE: Similar Pseudo Label Exploitation for Semi-Supervised Classification,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021

[13] Y. Guo, N. C. Codella, L. Karlinsky, J. V. Codella, J. R. Smith, K. Saenko, T. Rosing, and R. Feris, “A broader study of cross-domain few-shot learning,” in Proceedings of the European Conference on Computer Vision (ECCV) 2020. ECCV, 2020.

[14] C. P. Phoo and B. Hariharan, “Self-training for few-shot transfer across extreme task differences,” in International Conference on Learning Representations, 2021

[15] A. Islam, C.-F. Chen, R. Panda, L. Karlinsky, R. Feris, and R. J. Radke,“Dynamic distillation network for cross-domain few-shot recognition with unlabeled data,” in 35th Conference on Neural Information Processing Systems, NeurIPS 2021

[16] J. H. Oh, S. Kim, N. Ho, J.-H. Kim, H. Song, and S. Yun, “Understanding cross-domain few-shot learning based on domain similarity and few-shot difficulty,” in 36th Conference on Neural Information Processing Systems, NeurIPS, 2022

[17] Limeng Qiao Yuxuan Zhao Zhiyuan Li Xi Qiu* Jianan Wu Chi Zhang, DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection, ICCV, 2021

[18] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NeurIPS, 2015

[19] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, Peter Vajda , Unbiased Teacher for Semi-Supervised Object Detection, ICLR, 2021

[20] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, Jianfeng Gao , GLIPv2: Unifying Localization and Vision-Language Understanding, NeurIPS, 2022

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, Learning Transferable Visual Models From Natural Language Supervision, ICML, 2021

APPENDIX A – PUBLICATIONS AND PRESENTATIONS

List the dates, times, title, event and speakers of any presentations made under this effort and the title author and publication information for any publication made under this effort.

Netherlands AI Coalition 19 May 2021, Deep Dive Learning with Less Labels, Klamer Schutte

NeurIPS, 2021: Independent Prototype Propagation for Zero-Shot Compositionality

ICIAP 2021, Cluster Centers Provide Good First Labels for Object Detection

NeurIPS 2022, Maximum Class Separation as Inductive Bias in One Matrix

ICRA 2023, Improved Zero-Shot Object Localization using Contextualized Prompts and Objects in Context

CVPR 2023: Self-Guided Diffusion Models, Vincent Tao Hu, David W. Zhang, Yuki M. Asano, Gertjan J. Burghouts, Cees G. M. Snoek

ICIP 2023: Efficient Transfer by Robust Label Selection and Learning with Pseudo-Labels, Wyke Huizinga, Maarten Kruithof, Gertjan Burghouts and Klamer Schutte

Artificial Intelligence for Security and Defence Applications, SPIE Vol 12742 (5 Sep 2023): Aerial image segmentation with minimal annotation effort using clustering of pre-trained embeddings, Maarten C. Kruithof, Klamer Schutte, Wyke Huizinga, Gertjan Burghouts, Sabina van Rooij

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

BIT-M	Big Transfer (trained on ImageNet-21k)
BSCD-FSL	Broader Study of Cross-Domain Few-Shot Learning
CLaP	Clustering, Label selection and Pseudo-labels
CLIP	Contrastive Language-Image Pretraining
CIFAR	Canadian Institute for Advanced Research
CNN	Convolutional Neural Network
COCO	Common Objects in Context
CPU	Central Processing Unit
CUB200	Caltech UCSD Birds-200-2011
DARPA	Defense Advanced Research Projects Agency
DeFCRN	Decoupled Faster R-CNN
DDP	Distributed Data Parallel
DOD	Department of Defense
DOTA	Dataset for Object deTection in Aerial Images
GLIP	Grounded Language-Image Pre-training
IC	Image Classification
ISIC	International Skin Imaging Collaboration
JPL	Jet Propulsion Laboratory at NASA
LVIS	Large Vocabulary Instance Sepmentation
LwLL	Learning with Less Labels
MNIST	Modified National Institute of Standards and Technology database
MOD	Ministry of Defence
R-CNN	Region-Based Convolutional Neural Network
SOTA	State of the Art
SSL	Self-Supervised Learning
SWIN	Shifted Window
TNO	Netherlands Organisation for Applied Scientific Research
TOLEDA.	Towards Less Labels by Active Learning, Exploiting Unlabeled Data and Learned Augmentation
UDA	Unsupervised Domain Adaptation
UMAP	Uniform Manifold Approximation and Projection
VOC	Visual Object Class