REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
21-11-2019	Final Report	1-Aug-2015 - 30-Apr-2019

4. TITLE AND SUBTITLE
Final Report: Topic 5.2.1: Action Co-Discovery as a Cross-Reconstruction Problem

5a. CONTRACT NUMBER
W911NF-15-1-0354

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Michigan - Ann Arbor
3003 South State Street

Ann Arbor, MI 48109 -1274

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
67429-CS.9

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Jason Corso
UU	UU	UU	UU		19b. TELEPHONE NUMBER 734-647-8833

RPPR Final Report
as of 23-Jan-2020

Agency Code:

Proposal Number: 67429CS **Agreement Number: W911NF-15-1-0354**
INVESTIGATOR(S):

Name: Jason Corso
Email: jjcorso@umich.edu
Phone Number: 7346478833
Principal: Y

Organization: **University of Michigan - Ann Arbor**
Address: 3003 South State Street, Ann Arbor, MI 481091274
Country: USA
DUNS Number: 073133571 EIN: 386006309
Report Date: 31-Jul-2019 Date Received: 21-Nov-2019
Final Report for Period Beginning 01-Aug-2015 and Ending 30-Apr-2019
Title: Topic 5.2.1: Action Co-Discovery as a Cross-Reconstruction Problem
Begin Performance Period: 01-Aug-2015 **End Performance Period:** 30-Apr-2019
Report Term: 0-Other
Submitted By: Michele Feldkamp Email: careymrz@umich.edu
Phone: (734) 647-1813
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 0 **STEM Participants:** 0

Major Goals: 1 Objective

This project seeks a new method for automatically discovering actions that are common to a given set of videos. These common actions, called coactions, are
represented either as a set of frames or in more detail as a set of space-time segmentations (Figure 1). The project seeks a formulation that does not require strong
assumptions on the content or quality of the video signals and yet is computationally efficient.

The proposed formulation is based on the idea that each video's role in a certain coaction is measured by how well that video can be used to reconstruct the other videos also participating in the coaction. The proposed methodology explicitly does not incorporate features from the given video into the basis that is representing it to avoid the basis being overwhelmed by the background rather than the action itself. Hence, it is called a cross-reconstruction problem by the investigators. Neither does this novel formulation require a common or joint representation over all videos to
represent the coaction, which is the de facto approach for co-detection and co-segmentation methods; in particular, in video, it is not clear if the necessary underlying action invariants exist in sufficient descriptiveness to actually specify such a joint model. The novel formulation does not require any such assumption.
Ultimately, the main objective is to formulate, solve and study the new cross-reconstruction problem.

(See PDF for Figure 1)

Motivating CONOP Consider a forward operating base (FOB) in a highly dangerous location that sees dozens or hundreds of vehicles and people in its vicinity in any given day. The FOB is enabled with dozens of security cameras that are constantly on and acquiring video. FOB security team members are under constant watch for suspicious behavior. Some such behaviors have been noticed, such as peculiar walking paths in the neighboring hills, or vehicles making U-turns. However, there are hours and hours of video; too much to watch manually. Detectors for vehicle and people can be used to cull some video, but there still remains too much. The vehicles and people are hard to distinguish, so direct visual matching and reidentification is not plausible.

Only behaviors evolving over space and time are distinguishable. Currently, the security team would need to manually evaluate all or most of the video, which would take days or weeks to evaluate, potentially long after the threat has actually been realized.

With action co-discovery, the security team would instead be able to extract a small set of query videos and automatically detect commonly occurring across the videos. The result of the process would be dozens of common actions to analyze rather than thousands of videos to analyze. Example common actions in this CONOP could be a suspicious U-Turn in the same place near the gate, by different vehicles; or a meandering walk stopping at specific locations peering into the FOB, made by the same or different people.

Accomplishments: 2 Activities and Findings
2.1 Overall Approach
The problem at hand is the discovery of commonly occurring actions across multiple unlabeled videos. For example, in two surveillance videos, a van makes a three-point turn near the building. The na¨ıve system would implement current video analysis methods, such as action detection, specific object detection and tracking, and so on. However, these methods are not accurate enough to be later used in cross-view, crossrate common action discovery, which is the project focus. To overcome these limitations of existing state of the art methods, the project proposes to formulate a cross-reconstruction problem, which does not require any high-performing extraction and matching method. Instead, the cross-reconstruction problem localizes commonly occurring sub-signals (e.g., groups of frames, or groups of pixels/segments across frames), by identifying those sub-signals that can be used as a basis to reconstruct the sub-signals of other videos. Intuitively, for a sub-signal, or a set of frames, to be a common action, it must be able to serve as a representation capable of matching to a sub-signal, or set of frames, in the other videos.

The primary proposed methodological direction for investigating the cross-reconstruction problem is joint sparse coding across multiple videos; the approach will jointly learn the bases for cross-reconstruction while segmenting which elements (frames, flow, trajectories or segments) of each video are actually part of the common action. The team will apply common action co-discovery both on synthetic data and real data, emphasizing plausible abstractions of Army-relevant scenarios. They will further generalize the method from co-discovery of single coaction to multiple coactions in a given video, and will scale the method to hundreds of videos being concurrently processed.

Core elements of the current approach are the automatic bottom-up segmentation of video sequences into semantically labeled (automatically) segments that can then be processed by the joint sparse coding (and other cross-reconstruction-based methodologies), the use of optical flow as the core feature space on which to compute the sparse coding-based representation, and independent cross-reconstruction from a single video to a second video. Finally, the current approach emphasizes both temporal coaction discovery and spatiotemporal coaction discovery, but at two different levels of development. In temporal coaction discovery, the project is evaluating the core premise of cross-reconstruction with sparse coding. In spatiotemporal coaction discovery, the team is in the processing of improving the quality of general semantic spatiotemporal segmentation of video to be later used in the cross-reconstruction of spatiotemporal coactions.

2.2 Major Research and Education Activities
The project involved numerous activities.

Weakly Supervised Coaction Discovery and Segmentation We developed grouping process models for coaction discovery using a weakly supervised, multi-task ranking model. This met our original problem statement in the proposal and has led to marked success in the problem space. The work here was published in a CVPR 2017 paper.

Sparse Decomposition in Deep Networks Our modeling paradigm for this project is sparse coding and dictionary learning. However, it is well understand that the underlying optimization problems in this paradigm do not scale well to large data sets. Based on recent findings relating to the continuity of these underlying optimization problems, we have developed a mechanism for embedding them into deep networks. We investigated whether introducing sparse decomposition operations into deep neural networks would improve their transparency and data dependence. Clearing a series of technical hurdles, we discovered that our predictions were optimistic, and failed to find suitable evidence that sparse decomposition improved network performance, transparency, or data dependence in a significant way. This was released as an ArXiv technical report in 2016.

Learning to detect procedures in long videos We considered an alternative approach to discovering common actions across different videos this past year. This new approach seeks multiple videos of the same procedure, such as fixing a bicycle tire or cooking a certain dish. In this work, we developed a deep network-based method,

called Procedure Networks or ProcNets, that is capable of learning to automatically segment the video into clips containing meaningful content pertaining to the underlying procedure in the video. The work here was published in a AAAI 2018 paper.

2.3 Major Findings

Each major finding is demarcated by a horizontal rule, and a large, boldface heading.

Weakly Supervised Coaction Discovery and Segmentation
Fine-grained activity understanding in videos has attracted considerable recent attention with a shift from action classification to detailed actor and action understanding that provides compelling results for perceptual needs of cutting-edge autonomous systems. However, current methods for detailed understanding of actor and action have significant limitations: they require large amounts of finely labeled data, and they fail to capture any internal relationship among actors and actions. To address these issues, our focus for the previous year was the investigation of a robust multi-task ranking model for weakly-supervised actor-action segmentation where only video-level tags are given for training samples. Our model is able to share useful information among different actors and actions while learning a ranking matrix to select representative supervoxels for actors and actions respectively. Final segmentation results are generated by a conditional random field that considers various ranking scores for video parts.

The overview of proposed weakly supervised actor-action segmentation framework is shown in Figure 2. We first segment videos into supervoxels using the graph-based hierarchical supervoxel method (GBH) [GKHE10]. Meanwhile, we generate action tubes as the minimum bounding rectangles around supervoxels. We extract

(See PDF for Visual and Figure 2)

features at different GBH hierarchy levels to describe supervoxels and action tubes. It is our assumption that information contained in supervoxel segments in adult-running videos should be correlated with supervoxel segments in adult-walking videos as they share same actor adult. Similarily, the correlation of action tubes among fine-grained actions in a same general action, e.g. cat-walking and dog-walking, should be larger than the correlation among non-relevant action pairs. Therefore, three different kinds of potentials (action, actor, actor-action) are computed via our robust multi-task ranking model by considering information sharing among different groups of actors and actions. Finally, we devise a CRF model for actor-action segmentation. Example resuts of our new weakly supervised model applied to the Actor-Ation dataset are shown in Figure 3. We observe that our approach can generate better visual qualitative results than other approaches. However, our method fails in some cases, such as cat-jumping. This is probably because there are several cats jumping simultaneously and motion is significant in the video.

(See PDF for Visual and Figure 3)

Sparse Decomposition in Deep Networks CNNs are powerful and flexible, and scale naturally to problems of high dimensionality and large scale, but are highly dependent on the acquisition of large supervised datasets. Dictionary learning and sparse decomposition are successful alternatives for problems of low dimensionality and in general, do not require the same volume of training data to perform, but suffer from poor scalability to higher dimensions. Motivated by analytical work demonstrating the differentiable nature of one sparse decomposition scheme, we propose a ..(See PDF for remaining Text and Figures)

Training Opportunities: Continued support of Graduate Student, Parker Koch.
This grant partially support Dr. Chenliang Xu while he was a PhD student with Professor Corso. Dr. Xu is now an Assistant Professor at the University of Rochester in Computer Science

RPPR Final Report
as of 23-Jan-2020

Results Dissemination: We summarize all of the publications that were in part or full supported by this award; this is an all-inclusive list from the project's start.

[1] X. Sun, R. Szeto, and J. J. Corso. A Temporally-Aware Interpolation Network for Video Frame Inpainting. In Proceedings of Asian Conference on Computer Vision (ACCV), 2018.

[2] L. Zhou, C. Xu, and J. J. Corso. Towards automatic learning of procedures from web instructional videos. In Proceedings of AAAI Conference on Artificial Intelligence, 2018.

[3] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong. End-to-end dense video captioning with masked transformer. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[4] Y. Yan, C. Xu, D. Cai, and J. J. Corso. Weakly supervised actor-action segmentation via robust multitask ranking. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[5] P. Koch and J. J. Corso. Sparse factorization layers for neural networks with limited supervision. CoRR, abs/1612.04468, 2016.

[6] C. Xu and J. J. Corso. Actor-action semantic segmentation with grouping-process models. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[7] W. Chen and J. J. Corso. Action detection by implicit intentional motion clustering. In Proceedings of IEEE International Conference on Computer Vision, 2015.

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: YouCook2. As part of our efforts in creating large video data sets, we have released the YouCook2 dataset, which contains more than 2000 videos of procedures. It is available online at http://youcook2.eecs.umich.edu/.

ProcNets. We have released the source code for the method in our AAAI 2018 paper. It reproduces our experimental results and is available at https://github.com/LuoweiZhou/ProcNets-YouCook2.

Grouping Process Model. The software underlying the grouping process model for enhanced semantic segmentation in video, which is a possible mid-layer representation to underly the coaction discovery problem, has been released on the web. http://web.eecs.umich.edu/~jjcorso/r/a2d/

PARTICIPANTS:

Participant Type: Graduate Student (research assistant)
Participant: Parker Alexander Koch
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: PD/PI
Participant: Jason J Corso
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N

Other Collaborators:

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IEEE Computer Vision Pattern and Recogntion 2016
Date Received: 31-Aug-2017 Conference Date: 26-Jun-2016 Date Published: 31-Dec-2015
Conference Location: Las Vegas, NV
Paper Title: Actor-Action Semantic Segmentation with Grouping Process Models
Authors: Chenlian Xu, Jason Corso
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IEEE Conference on Computer Vision and Pattern Recognition
Date Received: 16-Oct-2018 Conference Date: 24-Jul-2017 Date Published: 24-Jul-2017
Conference Location: Honolulu, Hawaii
Paper Title: Weakly supervised actor-action segmentation via robust multi-task ranking
Authors: Yan Yan, Chenliang Xu, Dawen Cai, Jason J. Corso
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IEEE Conference on Computer Vision and Pattern Recognition
Date Received: 16-Oct-2018 Conference Date: 21-Jul-2017 Date Published: 03-Apr-2018
Conference Location: Honolulu, Hawaii
Paper Title: End-to-end dense video captioning with masked transformer
Authors: L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: AAAI Conference on Artifical Intelligence, 2018
Date Received: 16-Oct-2018 Conference Date: 02-Feb-2018 Date Published: 12-Nov-2017
Conference Location: New Orleans, Louisiana
Paper Title: Towards Automatic Learning of Procedures from Web Instructional Videos
Authors: Luowei Zhou, Chenliang Xu, Jason J. Corso
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: ACCV 2018, Asian Conference on Computer Vision
Date Received: 21-Nov-2019 Conference Date: 02-Dec-2018 Date Published: 03-Nov-2018
Conference Location: Perth, Australia
Paper Title: A Temporally-Aware Interpolation Network for Video Frame Inpainting
Authors: Ximeng Sun, Ryan Szeto, Jason J. Corso
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 23-Jan-2020

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IEEE ICCV 2015, International Conference on Computer Vision
Date Received: 21-Nov-2019 Conference Date: 13-Dec-2015 Date Published: 13-Dec-2015
Conference Location: Santiago, Chile
Paper Title: Action Detection by Implicit Intentional Motion Clustering
Authors: Wei Chen, Jason J. Corso
Acknowledged Federal Support: **Y**

Jason J. Corso
University of Michigan
Grant No.: W911NF1510354

FINAL PROGRESS REPORT ATTACHMENT
Action Co-Discovery As A Cross-Reconstruction Problem
University of Michigan

This document contains material for the cumulative project period. The material presented is fully or partially supported by this grant.

1 Objective

This project seeks a new method for automatically discovering actions that are common to a given set of videos. These common actions, called coactions, are represented either as a set of frames or in more detail as a set of space-time segmentations (Figure 1). The project seeks a formulation that does not require strong assumptions on the content or quality of the video signals and yet is computationally efficient.

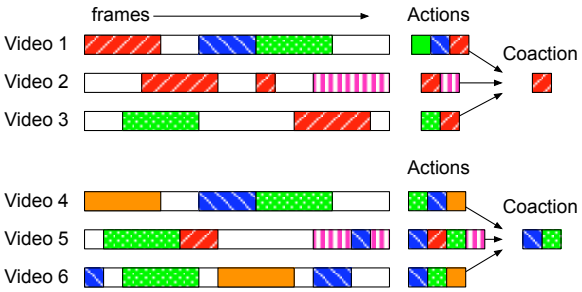

Figure 1: Two illustrative examples of a temporal coaction set over three videos each.

The proposed formulation is based on the idea that each video's role in a certain coaction is measured by how well that video can be used to reconstruct the other videos also participating in the coaction. The proposed methodology explicitly does not incorporate features from the given video into the basis that is representing it to avoid the basis being overwhelmed by the background rather than the action itself. Hence, it is called a cross-reconstruction problem by the investigators. Neither does this novel formulation require a common or joint representation over all videos to represent the coaction, which is the de facto approach for co-detection and co-segmentation methods; in particular, in video, it is not clear if the necessary underlying action invariants exist in sufficient descriptiveness to actually specify such a joint model. The novel formulation does not require any such assumption. Ultimately, the main objective is to formulate, solve and study the new cross-reconstruction problem.

Motivating CONOP Consider a forward operating base (FOB) in a highly dangerous location that sees dozens or hundreds of vehicles and people in its vicinity in any given day. The FOB is enabled with dozens of security cameras that are constantly on and acquiring video. FOB security team members are under constant watch for suspicious behavior. Some such behaviors have been noticed, such as peculiar walking paths in the neighboring hills, or vehicles making U-turns. However, there are hours and hours of video; too much to watch manually.

Detectors for vehicle and people can be used to cull some video, but there still remains too much. The vehicles and people are hard to distinguish, so direct visual matching and reidentification is not plausible.

Only behaviors evolving over space and time are distinguishable. Currently, the security team would need to manually evaluate all or most of the video, which would take days or weeks to evaluate, potentially long after the threat has actually been realized.

With action co-discovery, the security team would instead be able to extract a small set of query videos and automatically detect commonly occurring across the videos. The result of the process would be dozens of common actions to analyze rather than thousands of videos to analyze. Example common actions in this CONOP could be a suspicious U-Turn in the same place near the gate, by different vehicles; or a meandering walk stopping at specific locations peering into the FOB, made by the same or different people.

2 Activities and Findings

2.1 Overall Approach

The problem at hand is the discovery of commonly occurring actions across multiple unlabeled videos. For example, in two surveillance videos, a van makes a three-point turn near the building. The naïve system would implement current video analysis methods, such as action detection, specific object detection and tracking, and so on. However, these methods are not accurate enough to be later used in cross-view, cross-rate common action discovery, which is the project focus. To overcome these limitations of existing state of the art methods, the project proposes to formulate a cross-reconstruction problem, which does not require any high-performing extraction and matching method. Instead, the cross-reconstruction problem localizes commonly occurring sub-signals (e.g., groups of frames, or groups of pixels/segments across frames), by identifying those sub-signals that can be used as a basis to reconstruct the sub-signals of other videos. Intuitively, for a sub-signal, or a set of frames, to be a common action, it must be able to serve as a representation capable of matching to a sub-signal, or set of frames, in the other videos.

The primary proposed methodological direction for investigating the cross-reconstruction problem is joint sparse coding across multiple videos; the approach will jointly learn the bases for cross-reconstruction while segmenting which elements (frames, flow, trajectories or segments) of each video are actually part of the common action. The team will apply common action co-discovery both on synthetic data and real data, emphasizing plausible abstractions of Army-relevant scenarios. They will further generalize the method from co-discovery of single coaction to multiple coactions in a given video, and will scale the method to hundreds of videos being concurrently processed.

Core elements of the current approach are the automatic bottom-up segmentation of video sequences into semantically labeled (automatically) segments that can then be processed by the joint sparse coding (and other cross-reconstruction-based methodologies), the use of optical flow as the core feature space on which to compute the sparse coding-based representation, and independent cross-reconstruction from a single video to a second video. Finally, the current approach emphasizes both temporal coaction discovery and spatiotemporal coaction discovery, but at two different levels of development. In temporal coaction discovery, the project is evaluating the core premise of cross-reconstruction with sparse coding. In spatiotemporal coaction discovery, the team is in the processing of improving the quality of general semantic spatiotemporal segmentation of video to be later used in the cross-reconstruction of spatiotemporal coactions.

2.2 Major Research and Education Activities

The project involved numerous activities.

Weakly Supervised Coaction Discovery and Segmentation We developed grouping process models for coaction discovery using a weakly supervised, multi-task ranking model. This met our original problem statement in the proposal and has led to marked success in the problem space. The work here was published in a CVPR 2017 paper.

Sparse Decomposition in Deep Networks Our modeling paradigm for this project is sparse coding and dictionary learning. However, it is well understand that the underlying optimization problems in this paradigm do not scale well to large data sets. Based on recent findings relating to the continuity of these underlying optimization problems, we have developed a mechanism for embedding them into deep networks. We investigated whether introducing sparse decomposition operations into deep neural networks would improve their transparency and data dependence. Clearing a series of technical hurdles, we discovered that our predictions were optimistic, and failed to find suitable evidence that sparse decomposition improved network performance, transparency, or data dependence in a significant way. This was released as an ArXiv technical report in 2016.

Learning to detect procedures in long videos We considered an alternative approach to discovering common actions across different videos this past year. This new approach seeks multiple videos of the same *procedure*, such as fixing a bicycle tire or cooking a certain dish. In this work, we developed a deep network-based method, called Procedure Networks or ProcNets, that is capable of learning to automatically segment the video into clips containing meaningful content pertaining to the underlying procedure in the video. The work here was published in a AAAI 2018 paper.

2.3 Major Findings

Each major finding is demarcated by a horizontal rule, and a large, boldface heading.

Weakly Supervised Coaction Discovery and Segmentation

Fine-grained activity understanding in videos has attracted considerable recent attention with a shift from action classification to detailed actor and action understanding that provides compelling results for perceptual needs of cutting-edge autonomous systems. However, current methods for detailed understanding of actor and action have significant limitations: they require large amounts of finely labeled data, and they fail to capture any internal relationship among actors and actions. To address these issues, our focus for the previous year was the investigation of a robust multi-task ranking model for weakly-supervised actor-action segmentation where only video-level tags are given for training samples. Our model is able to share useful information among different actors and actions while learning a ranking matrix to select representative supervoxels for actors and actions respectively. Final segmentation results are generated by a conditional random field that considers various ranking scores for video parts.

The overview of proposed weakly supervised actor-action segmentation framework is shown in Figure 2. We first segment videos into supervoxels using the graph-based hierarchical supervoxel method (GBH) [GKHE10]. Meanwhile, we generate action tubes as the minimum bounding rectangles around supervoxels. We extract

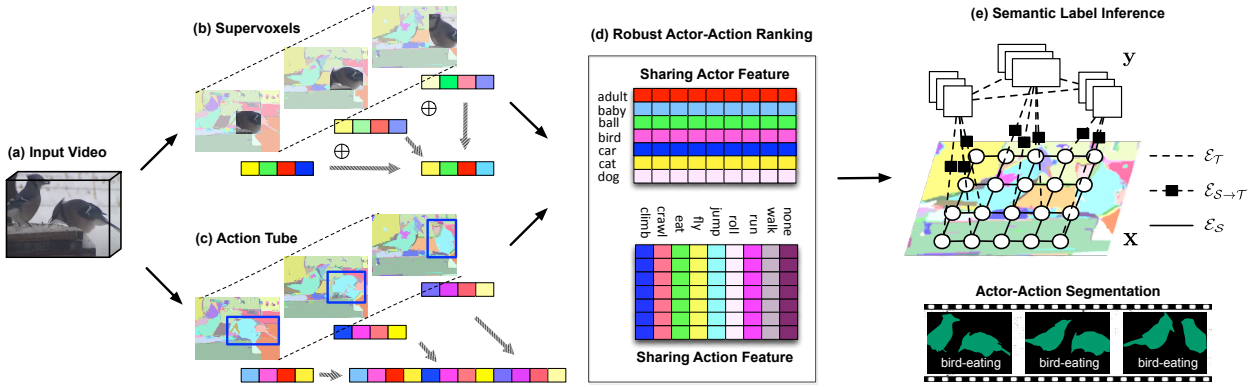

Figure 2: Overview of our proposed weakly supervised actor-action segmentation framework. (a) Input videos from the A2D dataset. (b) Supervoxel generation and feature extraction. (c) Action tube generation and feature extraction. (d) Sharing features among different actors and actions. (e) Semantic label inference for actor-action segmentation. Figure is best viewed in color and under zoom.

features at different GBH hierarchy levels to describe supervoxels and action tubes. It is our assumption that information contained in supervoxel segments in *adult-running* videos should be correlated with supervoxel segments in *adult-walking* videos as they share same actor *adult*. Similarily, the correlation of action tubes among fine-grained actions in a same general action, e.g. *cat-walking* and *dog-walking*, should be larger than the correlation among non-relevant action pairs. Therefore, three different kinds of potentials (action, actor, actor-action) are computed via our robust multi-task ranking model by considering information sharing among different groups of actors and actions. Finally, we devise a CRF model for actor-action segmentation. Example resuts of our new weakly supervised model applied to the Actor-Ation dataset are shown in Figure 3. We observe that our approach can generate better visual qualitative results than other approaches. However, our method fails in some cases, such as *cat-jumping*. This is probably because there are several cats jumping simutaneously and motion is significant in the video.

- Y. Yan, C. Xu, D. Cai, and **J. J. Corso**. Weakly supervised actor-action segmentation via robust multi-task ranking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Figure 3: Qualitative results shown in sampled frames for several video sequences from the A2D dataset. Columns from left to right are input video, ground-truth, our method, GPM [X**Corso**16], WSS [TZY16], RSVM [Joa06], DM-MTL [JRSR10] and AHRF [LRKT14] respectively. Our method is able to generate correct actor-action segmentation expect for *cat-jumping* and *adult-running* in these examples.

Sparse Decomposition in Deep Networks

CNNs are powerful and flexible, and scale naturally to problems of high dimensionality and large scale, but are highly dependent on the acquisition of large supervised datasets. Dictionary learning and sparse decomposition are successful alternatives for problems of low dimensionality and in general, do not require the same volume of training data to perform, but suffer from poor scalability to higher dimensions. Motivated by analytical work demonstrating the differentiable nature of one sparse decomposition scheme, we propose a marriage of the two models by creating neural network layers that perform this sparse decomposition.

The elastic-net formulation of sparse coding seeks a sparse $\alpha^* \in \mathbb{R}^k$ to represent an input $x \in \mathbb{R}^n$ as the coefficients in a linear combination of the columns of a dictionary matrix $D \in \mathbb{R}^{n \times k}$, i.e. α^* should have $x \approx D\alpha^*$. This α^* is found via the minimization:

$$\alpha^* \equiv \arg\min_{\alpha} \quad \frac{1}{2}\|x - D\alpha\|_2^2 + \lambda\|\alpha\|_1 + \frac{\lambda_2}{2}\|\alpha\|_2^2 \tag{1}$$

This factorization of x into D and α^* can be viewed as an operation $\alpha^*(x, D)$, and Mairal et al. [MBP12] prove that this operation is continuous and differentiable with respect to both x and D almost everywhere under mild assumptions on the distribution of x. This allows for the training of D for an external (supervised) loss function, in contrast to traditional dictionary learning, which minimizes the above (unsupervised) loss function over D. Most importantly, it allows the two loss functions to be combined for semisupervised learning, taking advantage of unlabeled training data, which is often much easier to obtain than labeled data.

The sparsity of the output α^* is also noteworthy, as prior work has been devoted to studying activation sparsity in neural networks, induced by, e.g., linear rectification [BC$^+$08]. Furthermore, work exploring this joint space of learning that is both sparse and deep is currently flourishing, with many novel techniques being developed [GBB11] [TMSV16].

We introduce the sparse factorization (SF) layer for neural networks, which takes input x and parameters D and produces α^*. Parameter-for-parameter, this layer can directly replace the traditional linear (or fully-connected) layer, which we have done in our comparisons.

Dictionary learning in computer vision has found success mainly when used on image patches, rather than whole images. This inspires another sparse factorizing layer analogous to the convolutional layer. Much like convolutional layers perform a linear operation locally over patches of an input image, our convolutional sparse factorization (CSF) layer performs the SF operation over image patches and is a parameter-for-parameter replacement for the convolutional layer.

We obtain preliminary results by comparing three networks on the task of handwritten digit classification: a baseline traditional four-layer CNN, the baseline network with the first linear layer replaced by an SF layer, and the baseline network with the first convolutional layer replaced by a CSF layer. We use the MNIST dataset [LBBH98], as well as three of its variants: MNIST-rot, where each image is rotated randomly; MNIST-rand, where each image has a white noise background; and MNIST-img, where each image has an image patch for a background.

Trained on all 60,000 training samples, the modified networks demonstrate performance generally on par with the baseline network. When training data becomes very scarce, however, the modified networks show improved performance.

The bottleneck with the new layers is computation; sparse factorization being a more complex operation than matrix multiplication, it is relatively slow, and current work is primarily devoted to investigating more

Table 1: Classification accuracy scores on the digit classification datasets, in percent. Boldface indicates the best score for each dataset.

	LeNet	CSF	SF
MNIST	99.14	**99.20**	98.31
MNIST-rot	**91.84**	91.76	78.63
MNIST-rand	95.11	**95.34**	93.33
MNIST-img	91.84	**93.42**	89.42

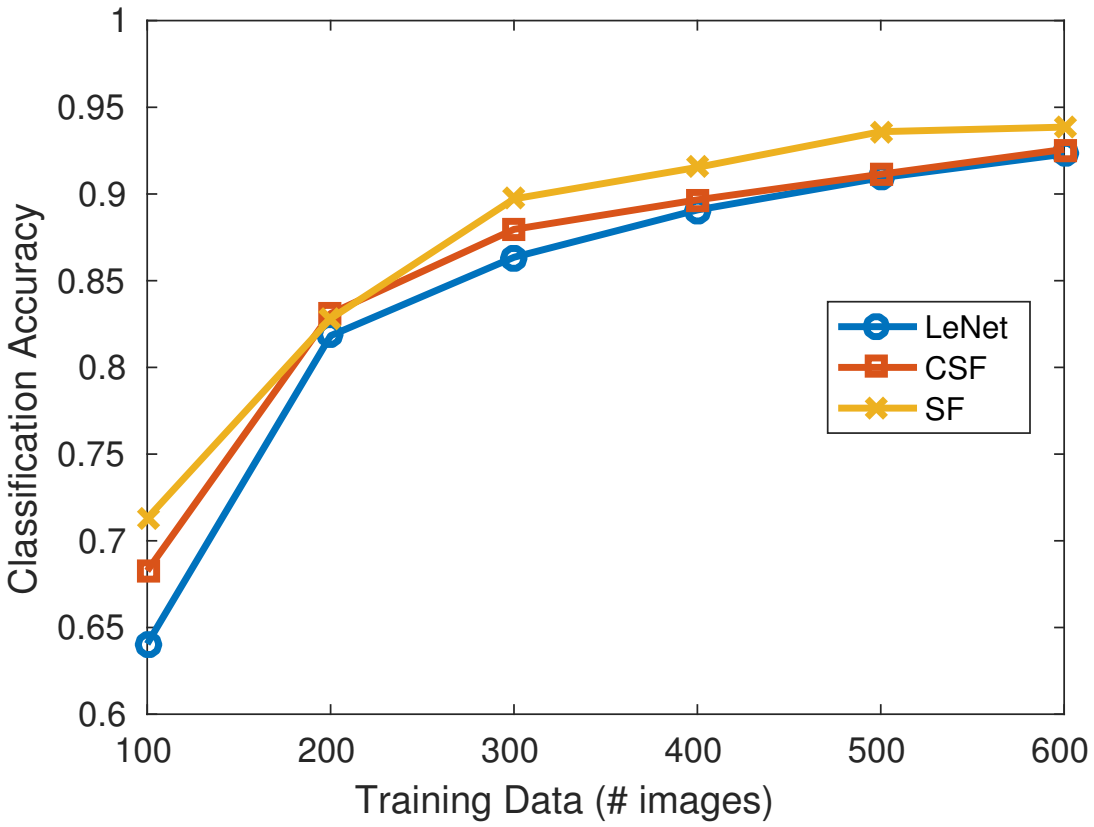

Figure 4: Classification accuracy of the baseline LeNet, CSF, and SF networks for limited amounts of training images, averaged over several trials.

efficient implementations of this operation. There is also much to be learned about the convergence properties and acceptable hyperparameter ranges for networks with sparse factorization layers, as conventional guidelines (e.g. Xavier initialization, common learning rates) seem less effective.

Analysis and Discussion Mairal et al. show how a shallow network can be trained with this method for a variety of tasks. However, the extension of their methods to deeper networks has proven to be highly nontrivial and resource-intensive. A handful of reasons for this are:

- The calculations required to find α^* and its gradient are much more computationally complex than the corresponding matrix multiplications for a linear layer. Our most efficient implementation, though imperfect, sometimes performed a factor of 40 times slower than the linear layer it replaced. In general, the deeper the layer is in the network, the worse this factor scales.
- Sparse coding requires the dictionary to be sufficiently uncorrelated and normalized. Though seemingly simple, maintaining such properties while minimizing a loss function is the subject of active research, and techniques have yet to be developed to sufficiently solve the problem for deep dictionary learning.
- The gradient derived by Mairal et al. for sparse coding behaves in a deep network much differently than that of a linear layer, and traditional learning schemes that coax deep networks to converge on a local minima completely fail to do so with sparse factorization layers. Without a more analytical and nuanced understanding of how this minimization environment interacts with such gradients and schemes, we may never know how to design around this gradient computation in general.
- Mairal et al. emphasized the necessity of pretraining the dictionary and the network in a multi-stage procedure as important to the convergence of their shallow network, and it seems to us that this is also true for deep networks. However, with more layers, this pretraining procedure quickly becomes untenable, and can take longer than the training itself. This is highly undesirable for networks looking to train on large amounts of data, as deep networks may be, and may not work at all if the network has multiple sparse factorization layers, a consideration not addressed in Mairal et al.
- Perhaps most disappointingly, the encouraging preliminary results we saw do not scale, and in some cases, couldn't be replicated on new deep learning frameworks. It is not apparent to us that continuing to endeavor to fix the above problems would yield further positive results.

In sum, these issues demonstrate that though analytically very interesting, deep sparse coding and dictionary learning is less practically fruitful than initially anticipated.

- P. Koch and **J. J. Corso**. Sparse factorization layers for neural networks with limited supervision. *CoRR*, abs/1612.04468, 2016.

Learning to Detect Procedures in Long Videos

The potential for machines to learn by observing humans/agents performing procedures involving objects and actions is rich. Current research on automatic video procedure learning heavily relies on action labels or video subtitles, even during the inference phase, which makes them infeasible in real-world scenarios. This leads to our question: can the human-consensus structure of a procedure be learned from a large set of long, unconstrained videos (e.g., instructional videos from YouTube) with only visual evidence? To answer this question, we introduce the problem of *procedure segmentation*—to segment a video procedure into category-independent procedure segments.

We define *procedure* as the sequence of necessary steps comprising such a complex task, and define each individual step as a *procedure segment*, or simply *segment* for convenience. For example, there are 8 segments in the making BLT sandwiches video shown in Fig. 5. We represent these segments by their start and end temporal boundaries in a given video. Note that one procedure segment could contain multiple actions, but it should be conceptually compact, i.e., described with a single sentence. The number of procedure segments and their locations reflect human consensus on how the procedure is structured. We show in this work how machines can learn a consensus on proposing semantically-meaningful segments for a procedure.

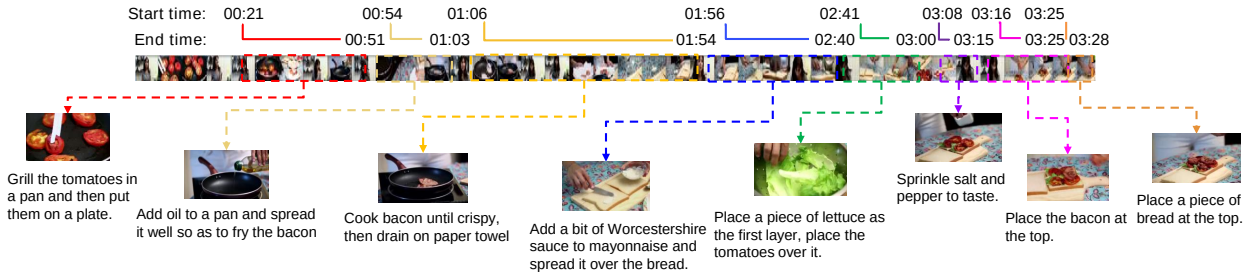

Figure 5: An example from the YouCook2 dataset on making BLT sandwiches. Each procedure step has time boundaries annotated and is described by an English sentence. Video from YouTube with ID: `4eWzsx1vAi8`.

We propose a deep-learning-based approach for procedure segmentation, called Procedure Segmentation Networks or ProcNets. ProceNets consist of three components, namely, Context-aware Embedding, Procedure Segment Proposal, and Sequential Prediction. We first uniformly downsample the input video into individual frames and extract the appearance feature through Convolution Neural Networks (i.e. ResNets). With the frame-wise feature in hand, we then model the context between features with Recurrent Neural Networks (i.e. Bidirectional LSTM), fulfilled by the Context-aware Embedding module. The features are fed into the Procedure Segment Proposal, where segment candidates are proposed. Finally, the third module outputs a most likely sequence of segments, given the pool of segment candidates.

For evaluation, given that no large-scale dataset is available for this problem, we collect our own. Cooking is a well-defined form of procedure, and hence we collect culinary videos and ensure the diversity of the video content with 89 recipes from all over the world. We name the dataset YouCook2 (download link in Sec. 7). The procedure steps for each video are annotated with temporal boundaries and described by imperative English sentences. All the videos are downloaded from YouTube and are unconstrained, can be performed by individual persons at their houses with unfixed cameras. YouCook2 is the largest task-oriented, instructional video dataset in the vision community, with fully-annotated 2000 long untrimmed videos.

Extensive experimental results on YouCook2 demonstrated that our proposed methods perform better than competitive existing methods on all the evaluation metrics. To better understand how each piece of our model functions, we conducted analytical experiments, such as ablation studies and sensitivity tests. Qualitative visualizations of the model output also suggested ProcNets have learned to capture the important & integrated cooking recipe steps (examples see Fig. 6). However, we did notice the model occasionally outputs semantically meaningful but unnecessary segments, such as the last recipe step in the Grilled Cheese example, in which all the previous recipe steps are repeated. A possible improvement would be adding a constraint to our model formulation such that only necessary segments for a procedure are generated.

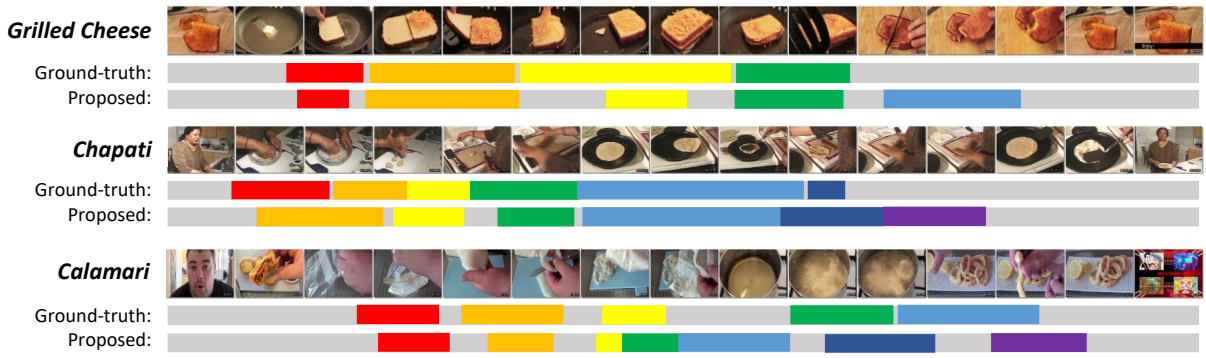

Figure 6: Qualitative results from test set. YouTube IDs: `BlTCkNkfmRY`, `jD4o_Lmy6bU` and `jrwHN188H2I`.

To conclude, our work is an effective attempt towards supervised segment detection/proposal. The class-agnostic procedure segment can be applied to downstream tasks, such as dense video captioning (recipe generation), action detection, and event parsing etc. Considering the current annotation is expensive to obtain, in our future work, we intend to explore low-cost approaches for this problem, possibly in a weakly-supervised or unsupervised fashion.

- L. Zhou, C. Xu, and **J. J. Corso**. Towards automatic learning of procedures from web instructional videos. In *Proceedings of AAAI Conference on Artificial Intelligence*, 2018.

3 Scientific Barriers

The scientific barrier of most importance relate to the actual feature space on which to compute the cross-reconstruction. In the initial work, optical flow is used as such a feature space. However, optical flow is not invariant to viewpoint, and hence, if the camera angle changes, then the optical flow signal will change. It is not clear that the cross-reconstruction will succeed in such a case. Plausible solutions would be to rely on a multitude of videos to establish a basis that can elicit common action modeling from many viewpoints. An alternative solution is to learn the feature space/representation in such a way that it is viewpoint invariant, but there is no clear way of doing this in the current literature. However, note that there remain numerous Army-relevant scenarios in which the viewpoint can be controlled and this barrier is not a problem, such as surveillance.

4 Scientific Significance

This is the first work we are aware of to propose joint spatiotemporal discovery of common actions across multiple videos with no supervision. Furthermore, it is the first such work in vision to relax the assumption that only one action is occurring at any given time in a video. The proposed work builds on our earlier work in temporal discovery of common actions, which, at the time, was the first work on the problem (2012). Although the project makes the basic inquiry by using the problem of discovering common actions across multiple videos, the basic formulation is broadly applicable to situations in which space-time patterns are present across many signal samples. Such additional applications include but are not limited to human behavior patterns in social networks, common structure discovery in dynamic robot mapping problems, and discovering equity factor-groups in financial markets. Hence, the project has a high significance both in its core technology and the potential application areas.

5 Plans for Next Year

N/A as this is a final report.

6 Collaborations and Leveraged Funding

Our team has had a collaboration with the Army Research Labs, Adelphi MD and with TARDEC in Warren, MI that were leveraged in discussion during this project.

The PI has leveraged funding from Google and Samsung to help develop the large instructional video dataset that is being used in this research project.

7 Technology Transfer

YouCook2. As part of our efforts in creating large video data sets, we have released the YouCook2 dataset, which contains more than 2000 videos of procedures. It is available online at `http://youcook2.eecs.umich.edu/`.

ProcNets. We have released the source code for the method in our AAAI 2018 paper. It reproduces our experimental results and is available at `https://github.com/LuoweiZhou/ProcNets-YouCook2`.

Grouping Process Model. The software underlying the grouping process model for enhanced semantic segmentation in video, which is a possible mid-layer representation to underly the coaction discovery problem, has been released on the web. `http://web.eecs.umich.edu/~jjcorso/r/a2d/`

8 Full Publication List

We summarize all of the publications that were in part or full supported by this award; this is an all-inclusive list from the project's start.

[1] X. Sun, R. Szeto, and **J. J. Corso**. A Temporally-Aware Interpolation Network for Video Frame Inpainting. In *Proceedings of Asian Conference on Computer Vision (ACCV)*, 2018.

[2] L. Zhou, C. Xu, and **J. J. Corso**. Towards automatic learning of procedures from web instructional videos. In *Proceedings of AAAI Conference on Artificial Intelligence*, 2018.

[3] L. Zhou, Y. Zhou, **J. J. Corso**, R. Socher, and C. Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[4] Y. Yan, C. Xu, D. Cai, and **J. J. Corso**. Weakly supervised actor-action segmentation via robust multi-task ranking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[5] P. Koch and **J. J. Corso**. Sparse factorization layers for neural networks with limited supervision. *CoRR*, abs/1612.04468, 2016.

[6] C. Xu and **J. J. Corso**. Actor-action semantic segmentation with grouping-process models. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[7] W. Chen and **J. J. Corso**. Action detection by implicit intentional motion clustering. In *Proceedings of IEEE International Conference on Computer Vision*, 2015.

This project also partially funded the doctoral dissertation of Chenliang Xu, entitled "Scale-Adaptive Video Understanding." Dr. Xu is now an Assistant Professor in Computer Science at the University of Rochester.

References

[BC+08] Y-lan Boureau, Yann L Cun, et al. Sparse feature learning for deep belief networks. In *Advances in neural information processing systems*, pages 1185–1192, 2008.

[C**Corso**15] W. Chen and **J. J. Corso**. Action detection by implicit intentional motion clustering. In *Proceedings of IEEE International Conference on Computer Vision*, 2015.

[GBB11] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Aistats*, volume 15, page 275, 2011.

[GKHE10] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[Joa06] Thorsten Joachims. Training linear svms in linear time. In *ACM SIGKDD Conferences on Knowledge Discovery and Data Mining*, 2006.

[JRSR10] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *NIPS*, 2010.

[K**Corso**16] P. Koch and **J. J. Corso**. Sparse factorization layers for neural networks with limited supervision. *CoRR*, abs/1612.04468, 2016.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LRKT14] L. Ladicky, C. Russell, P. Kohli, and P.H. Torr. Associative hierarchical random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1056–1077, 2014.

[MBP12] Julien Mairal, Francis Bach, and Jean Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012.

[SS**Corso**18] X. Sun, R. Szeto, and **J. J. Corso**. A Temporally-Aware Interpolation Network for Video Frame Inpainting. In *Proceedings of Asian Conference on Computer Vision (ACCV)*, 2018.

[TMSV16] Snigdha Tariyal, Angshul Majumdar, Richa Singh, and Mayank Vatsa. Greedy deep dictionary learning. *arXiv preprint arXiv:1602.00203*, 2016.

[TZY16] Yi-Hsuan Tsai, Guangyu Zhong, and Ming-Hsuan Yang. Semantic co-segmentation in videos. In *European Conference on Computer Vision*, 2016.

[X**Corso**16] C. Xu and **J. J. Corso**. Actor-action semantic segmentation with grouping-process models. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[YXC**Corso**17] Y. Yan, C. Xu, D. Cai, and **J. J. Corso**. Weakly supervised actor-action segmentation via robust multi-task ranking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[ZX**Corso**18] L. Zhou, C. Xu, and **J. J. Corso**. Towards automatic learning of procedures from web instructional videos. In *Proceedings of AAAI Conference on Artificial Intelligence*, 2018.

[ZZ**Corso**+18] L. Zhou, Y. Zhou, **J. J. Corso**, R. Socher, and C. Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018.