DS-GA 3001 Course Project Report
Team 12: Learning View-Invariant Representation for Improved Multi-View Understanding in Hand-Centric Videos

Xu Zhang
New York University
Address
xz4863@nyu.edu

Sihang Li
New York University
Address
sl10496@nyu.edu

Swarali Borde
New York University
Address
sdb7897@nyu.edu

Team [12]

Abstract

Egocentric vision has suffered from the limited scale of egocentric data. Previous works focus on transferring knowledge from the exocentric domain to the egocentric domain (exo-ego transfer). In this work, we take a different perspective by introducing a self-supervised learning (SSL) pre-training stage before the training of the model that learns view-invariant representations. With such view-invariant representations, models trained with exocentric data can be applied to the egocentric domain without any labeled egocentric data. We showed that models trained with our refined representation can have improved performance in both egocentric and exocentric domains. We applied our method to the temporal action segmentation task in the Assembly101 dataset and show that it can be used in joint-view learning for improved performance.

1 Introduction

Training models with egocentric data suffers the limited scale of available data. A common solution is to transfer knowledge for exocentric (third-person) videos to the egocentric (first-person) domain. However, many exo-ego transfer methods sacrifice performance in the exocentric domain and are less favored in applications that require understanding data from multiple views. Instead of improving models after training, we take a different perspective and introduce a pre-training stage inspired by joint ego-exo learning methods [1] to improve performance in egocentric views without labeled egocentric data.

Different views of an action can contain different levels of information. For example, an egocentric video of a man walking can not show the full-body details presented in a corresponding exocentric video. We restrict our study to videos that are hand-centric and show full details of the object. For example, a video of a man assembling a toy can show roughly the same level of detail from different perspectives because the important information about this action is on the toy and the hands. Such hand-centric videos are still important in applications such as AR and robotics [2], where tracking human hands or robot arm movements is crucial to the tasks. In the hand-centric videos, since both egocentric and exocentric videos show the same details about the action, this means that the information about the action is independent of view perspectives. Hence, it is possible to learn a representation of the actions without any view-specific signal. In this project, we propose a novel framework for joint exo-ego self-supervised learning on hand-centric videos to learn a view-invariant representation. By jointly optimizing reconstruction and view-invariant losses, we force the SSL

process to learn a view-invariant representation that filters out the view-specific information. With our view-invariant representation, our method outperforms baseline approaches in action segmentation in both exocentric and egocentric view.

2 Related Work

Exo-Ego Transfer Exo-ego transfer aims to transfer knowledge from exocentric to egocentric views when a model is trained on exocentric data. Ho et al. [3] adapt visual signals in exocentric videos to egocentric ones by leveraging a semi-supervised domain adaptation technique in summarization. Li et al. [4] propose the pre-training approach Ego-exo to extract cues from exocentric videos helpful for the egocentric domain with pseudo-labels from off-the-shelf models through knowledge distillation. [5] is the most related to our work. They also leveraged temporally aligned video pairs of different views. However, they used knowledge distillation to make the egocentric model simulate the output of their pre-trained exocentric counterpart. Our work takes a different approach by pre-training a view-invariant representation to train models without view-specific noise.

Joint Ego-Exo Learning Learning from different views to obtain a joint ego-exo representation is not a new idea. Early work [1] [6] leverage exocentric and egocentric videos synchronized in time to learn a joint representation. Recent works relax the requirement of time synchronization and unlock the ability to learn from large-scale video data [7][8]. We note that in some applications, it's not difficult to collect synchronized videos from different views without labels, and cost of this is still lower than collecting labeled egocentric videos of the same scale. Furthermore, temporally aligned videos introduce the unique advantage of temporal alignment and make view-invariant representation learning easier. Our method takes temporally aligned videos to learn view-invariant representation. However, unlike previous methods that also require synchronized videos [1][6], our method uses the same feature extractor for data from different views, which relaxes assumptions about input data and makes the pipeline simpler at inference time. Furthermore, we make connection between the 3D representation and view-invariant representation to facilitate a more robust feature representation.

Temporal Action Segmentation The goal of action segmentation is to segment a video into non-overlaping time intervals and label each time segment with one of the given action labels. Action segmentation can be helpful for various downstream tasks such as video-to-text and action localization and used in robotics. Common datasets for action segmentation include Assembly101 [9] and Breakfast [10]. The most notable methods are C2F-TCN [11] and ASFormer [12]. Typically, the model passes the videos to an encoder such as ResNet. Then, it proposes an initial segmentation based on the embedding and sequentially refines the segmentation based on previous coarser segmentations. In this report, we call such a temporal action segmentation model a TAS model or a segmentation head. Our work uses action segmentation as a task to demonstrate the effectiveness of our approach though our method can be applied to other tasks.

3 Proposed Approach

Consider a video representation extracted by a off-the-shelf feature extractor $f(V) \in \mathbb{R}^{T \times d}$ where T is the number of frames in the video, d is the dimension of the latent representation, and $V \in \mathbb{R}^{T \times H \times W}$ is the raw video. Our method trains a feature refiner that extracts view-invariant information from V using self-supervised learning. With the refined feature, we can train a segmentation head that performs better in both egocentric and exocentric domains even if the head is exposed with only labeled exocentric videos.

3.1 Task Definition

Given a video $V \in \mathbb{R}^{T \times H \times W}$ where T is the number of frames, H and W are the height and width of a video frame, a off-the-shelf feature extractor extracts its latent representation $f(V) \in \mathbb{R}^{T \times d}$. Let $D^{exo}_{train} = \{(f(V_i)^{exo}, S^{exo}_i)\}^n_{i=1}$ be a collection of n exocentric video presentations and labels, where S^{exo}_i is the manually labeled segmentation for the i^{th} exocentric video V^{exo}_i. We define exo-ego view-joint pre-training as the problem of training a feature refiner r that is applied to $f(V)$ using an additional collection of unlabeled pairs of exocentric and egocentric videos $D^{pair}_{transfer} = \{(V^{exo}_j, V^{ego}_j)\}^m_{j=1}$. Given D^{exo}_{train} as training data for segmentation, the goal is to use the feature

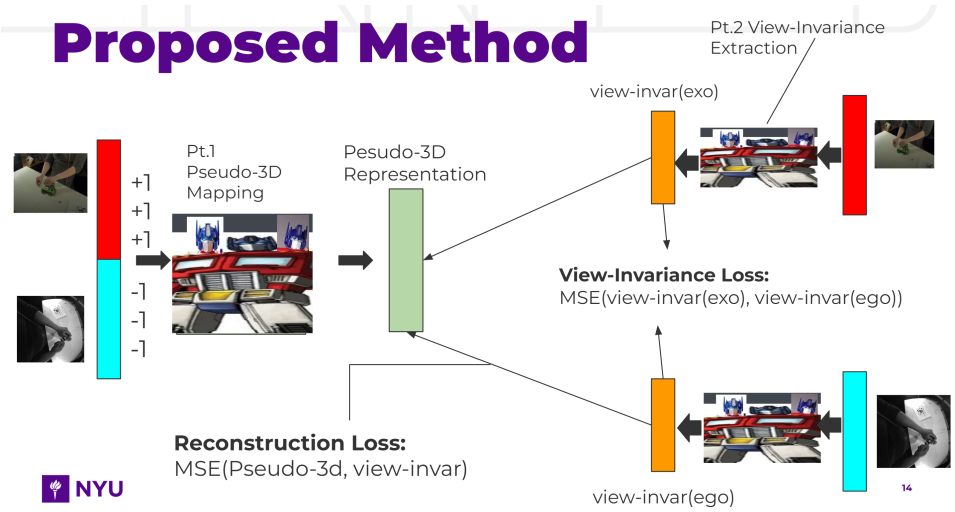

Figure 1: Our proposed SSL framework consists of two parts: a pesudo-3D stream jointly model the representation from different views and a view-invariant stream that extracts view-invariant information from specific views. Two streams are connected by a reconstruction loss. Furthermore, we pose a view-invariant loss that encourages view-invariant representation from different views to be similar.

refiner r to improve the performance of an action segmentation algorithm ψ to generalize over both $D^{ego}_{test} = \{(V^{ego}_h, S^{ego}_h)\}^q_{h=1}$ (of egocentric videos V^{ego}_h and ground truth segmentations S^{ego}_h) and D^{exo}_{test}, defined similarly.

3.2 Self-Supervised Learning for Feature Refiner

As shown in Figure 1. Given a pair of temporally aligned exocentric and egocentric video features in frame t, $f(V^t)^{exo}$ and $f(V^t)^{ego}$, we put them in two different streams. In the pseudo-3D stream, $f(V^t)^{exo}$ and $f(V^t)^{ego}$ were concatenated and sent to an encoder that models the joint representation of the different views. In the view-invariant stream, representations of each view were sent individually to the view-invariant feature refiner. The refined feature output from the refiner will be used to reconstruct the joint representation from the pseudo-3D stream. After the training, we discard the pseudo-3D stream and take the view-invariant feature refiner as the end product of our pre-training process. Now we will introduce each component of our framework in details.

3.2.1 Pseudo-3D Stream

Pseudo-3D Stream first takes the concatenated feature representation $[f(V^t)^{ego}, f(V^t)^{exo}] \in \mathbb{R}^{2d}$ into a transformer encoder followed by a linear layer that maps the representation to $\mathbb{R}^{3d}$, where an encoder models the joint representation of different views in $\mathbb{R}^{3d}$. We chose to model the joint representation in $\mathbb{R}^{3d}$ since this is isomorphic to the triplet (R^d, R^d, R^d) and many classic computer vision methods for reconstruction require at least 3 images. The modeled joint representation will be mapped back to $\mathbb{R}^d$ as $f(V^t)^{joint}$, for the view-invariant features to reconstruct.

3.2.2 View-Invariant Stream

In the end, we want to train a feature refiner that filters out view-specific information of the video representation. We use a transformer encoder as our architecture. Since its output is a view-invariant feature and the 3D representations are view-invariant, this view-invariant feature can be used to reconstruct the pseudo-3D representation. Since we restrict our focus to hand-centric videos, the information about the action in different views remains the same level. Thus, we expect the output of view-invariant feature refiner for the video frame of different views to be highly similar.

3.2.3 Training Objective and Loss Formulation

Since the refined features from different views of the same frame are expected to be highly similar, we explicitly pose this constraint as a view-invariant loss:

$$L_{invr}(V^t) = MSE(r(f(V^t))^{ego}, r(f(V^t))^{exo}) \tag{1}$$

Since the pseudo-3D representation $f(V^t)^{joint}$ is already mapped to $\mathbb{R}^d$ for reconstruction, the reconstruction loss can simply be MSE loss:

$$L_{recn}(f(V^t)^{joint}, r(f(V^t))^{exo/ego}) = MSE(f(V^t)^{joint}, r(f(V^t))^{exo/ego}) \tag{2}$$

Finally, we combine all losses together to obtain

$$Loss = Loss_{recn}(f(V^t)^{joint}, r(f(V^t))^{exo}) + Loss_{recn}(f(V^t)^{joint}, r(f(V^t))^{ego}) + \lambda L_{invr}(V^t) \tag{3}$$

where $\lambda = ((100 + cur\_epoch)/(100 + (cur\_epoch^{1.18})))$

Here, λ regulates the importance of L_{invr} in training. Since the view-invariant information for videos from different views is not exactly the same, L_{invr} will never be 0 and will dominate in later epochs of pre-training. Thus, we propose this factor to encourage more focus on reconstruction loss in later epochs.

4 Datasets and Evaluation

Assembly101 Assembly101 dataset consists of 4321 videos of participants assembling and disassembling toys such as trucks. It features cameras from different view points: 8 exocentric and 4 egocentric cameras totaling 513 hours of recording. Assembly101's task include action recognition and temporal action segmentation with 100K coarse labels for its temporal action segmentation task. Since assembly activities are hand-centric in nature and the dataset has synchronized egocentric and exocentric videos, Assembly101 is the ideal dataset to test our method. We considered coarse-grained action annotations for the temporal action segmentation task. For our experiment, we chose the most presentative camera views e4 for the egocentric videos and v3 for exocentric videos. We removed videos with missing segmentation labels and report the performance on the validation set as in [9][13] for our experiment. We chose 50% (90 videos) of videos in the official training data as the synchronized video pairs for self-supervised learning without the use of their labels, whereas the rest of the training data will be used to train the segmentation head and only has exocentric videos and labels available to the model.

Implementation Details Following the authors [9], we use TSM [14] as the off-the-shelf feature extractor f and C2F-TCN [15] as the temporal action segmentation model. For the training of feature refiner, we ran 250 epochs with pytorch SGD optimizer learning rate 1e-3. The pre-training took 8 hours on a single Nvidia A100 GPU.

Mesaures We use standard measures in Assembly101 temporal action segmentation task: Edit distance, F1 score, and Mean of Frames(MoF). All results are reported in percentage (0 - 100 range). See [16] for details on how these measures are defined.

5 Experiment

Baselines We compare our method with 2 baselines: (1) Naive baseline without pre-training. Naive baseline does not utilize any synchronized videos, it's trained with labeled exocentric videos and tested on both egocentric and exocentric data. (2) Synchronization KD baseline is adapted from [5]. Since the experiment setting is different in [5], we switched the teacher and student in knowledge distillation. In particular, we train an encoder that takes exocentric latent representation to output the egocentric representation during pre-training. The idea is that, the learned representation will take an exocentric representation and output a representation that simulates its egocentric counterpart. Then when we train the temporal action segmentation model, the TAS model is trained on the pseudo-egocentric representation simulated from exocentric data. Hence, when we shift the domain to actual egocentric domain, the TAS model will face a smaller domain gap.

Method	Edit	F1@10	F1@25	F1@50	MoF	total
No Pre-training	22.61	22.62	19.13	12.53	28.62	105.52
Synchronization KD	**25.19**	25.19	21.32	14.24	31.45	117.39
Ours Method	24.74	**26.62**	**22.50**	**15.40**	**32.05**	121.31

Table 1: Perfomance on Egocentric data

Method	Edit	F1@10	F1@25	F1@50	MoF	total
No Pre-training	24.52	25.24	21.70	15.16	31.29	117.92
Synchronization KD	27.32	27.74	23.35	16.10	33.42	127.91
Ours Method	**27.55**	**28.57**	**25.15**	**17.97**	**35.14**	134.39

Table 2: Perfomance on Exocentric data

Main Results In Table 1, we show the experiment results for egocentric video data. Both pre-training methods have significant improvement compared with the baseline without pre-training. Our method outperforms both baselines except for the edit score. A Similar pattern can be seen in exocentric results, as shown in Table 2, where our method outperforms baselines in all measures. The performance gap between the models' performance on egocentric and that of exocentric is significant. This is not surprising since the models only have access to labeled exocentric videos during training of TAS models and egocentric camera has lower resolution and is grayscale. Improvements in the Synchronization KD baseline have different reasons in the two view points. In egocentric data, we presented the TAS model with the simulated egocentric representation based on exocentric input during training. So the segmentation head is trained on simulated egocentric data distribution. When we move the domain to egocentric data, the decreased domain gap between simulated and actual egocentric data contributes to the improved performance. In exocentric data, we evaluate the performance of the model in the same domain of training data. We suspect that Synchronization KD's feature refiner still learned some feature-invariant information during the knowledge distillation even if our proxy task is to reconstruct data representation of one view point from the data of the other view point. For our method, since the feature refiner uses the shared weights for both views, our feature refiner transfers the input representations from different views to the same underlying view-invariant representation. Thus, the TAS model is trained on the view-invariant representation and has no view-point domain gap at the evaluation time. This shows that it is possible to learn a single feature extractor that extracts features from different view points to obtain view-invariant representations, and this view-invariant representation can improve the performance of the segmentation head due to its better representation ability.

Ablation Study We did two types of ablation study to show the components that contribute the most in our design. First, we have the ablation of pseudo-3D stream architecture. We compare our pseudo-3D stream architecture with a simple implementation using pytorch transformer encoder of the same scale followed a fully-connected layer, as shown in Table 3. The result shows our proposed architecture contributed to the improved performance in exocentric data. In the second ablation experiment, we compare our pipeline with the conventional exo-ego transfer task where the TAS model is first trained with exocentric data, and we use unlabeld exocentric and egocentric video pairs to improve the performance of the TAS model. As shown in Table 4, when we apply the feature refiner without tuning the segmentation head, we do not observe any improvement in performance. This is not surprising since the TAS model is trained on exocentric data distribution, but the feature refiners transfer the data to another domain which introduces domain gap for the TAS model. This also shows the TAS model training is an indispensable component of our pipeline.

Improve Joint-View TAS Performance In all settings, the input dimension to the TAS model is the same (2048 in our case). That means, it's a better representation of the videos that contributes to the

Method	Edit	F1@10	F1@25	F1@50	MoF	total
Simple Arch (ego)	**25.36**	26.05	22.50	15.33	31.97	121.22
Our Pseudo-3D Arch (ego)	24.74	**26.62**	22.50	**15.40**	**32.05**	121.31
Simple Arch (exo)	25.04	27.78	23.56	16.41	32.67	125.47
Our Pseudo-3D Arch (exo)	**27.55**	**28.57**	**25.15**	**17.97**	**35.14**	134.39

Table 3: Ablation 1: Compared with simple architecture, our more advanced architecture contributed to improved performance in exocentric view

Method	Edit	F1@10	F1@25	F1@50	MoF	total
Baseline without SSL(ego)	22.61	22.62	19.13	12.53	28.62	105.52
Synchronization KD (ego)	23.47	23.86	19.63	12.94	28.98	108.87
Our Method (ego)	18.14	18.31	16.05	10.81	25.14	88.44

Table 4: Ablation 2: In conventional exo-ego transfer, the segmentation head can not be tuned with labeled data, apply our method naively does not introduce any improvement

improved performance of the TAS model. This opens up the potential to use our framework to improve the performance of joint-view temporal action segmentation task. In the original Assembly101 temporal action segmentation task, model has access to videos from all the views. We test whether our framework can be used to improve the performance of an arbitrary TAS model in the Assembly101 temporal action segmentation task. We use the same TAS model as in our main experiments and include more view points in our pre-trained framework: e4, v3, and v4. We give our method access to labeled data from all view points during both pre-training and training stage. Then we compare the performance of the baseline without pre-training and our method. As shown in Table 5, we see some small improvement for the final performance with pre-training. Our method is TAS model-agnostic and can be applied to any TAS model. Thus, this offers potential to further improve state-of-the-art performance by incorporating our framework with the best-performing TAS models.

6 Conclusion

We investigated the problem of using view-invariant representation to improve model's performance on different views. We proposed a pre-training framework that improves the TAS model's performance in all views by having an extra pre-training stage that does not requires labels before its training. Our experiments show that our framework can improve models' performance significantly. Furthermore, we showed that, by incorporating our framework, the same TAS model can reach better performance when trained with the same data. This fundamental technology has great promise in applications that require effective understanding from multiple views such as video captioning with multiple cameras in TV shows. Further works include incorporating temporal information in pre-training process and exploiting explicit 3D priors for better representation learning.

Method	Edit	F1@10	F1@25	F1@50	MoF	total
Baseline	31.13	33.06	28.43	19.82	37.97	150.41
Our Method	31.35	34.09	30.09	21.24	38.35	155.12

Table 5: Improved performance on Assembly101 TAS task by incorporating our framework.

References

[1] Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Proceedings of the 1st Workshop and Challenge on Comprehensive Video Understanding in the Wild*, CoVieW'18, page 3, New York, NY, USA, 2018. Association for Computing Machinery.

[2] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. AssemblyHands: towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12999–13008, 2023.

[3] Hsuan-I Ho, Wei-Chen Chiu, and Y. Wang. Summarizing first-person videos from third persons' points of views. In *European Conference on Computer Vision*, 2017.

[4] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6939–6949, 2021.

[5] Camillo Quattrocchi, Antonino Furnari, Daniele Di Mauro, Mario Valerio Giuffrida, and Giovanni Maria Farinella. Synchronization is all you need: Exocentric-to-egocentric transfer for temporal action segmentation with unlabeled synchronized video pairs. In *European Conference on Computer Vision*, pages 253–270. Springer, 2024.

[6] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. What i see is what you see: Joint attention learning for first and third person video co-analysis. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1358–1366, New York, NY, USA, 2019. Association for Computing Machinery.

[7] Zihui (Sherry) Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 53688–53710. Curran Associates, Inc., 2023.

[8] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bootstrap your own views: Masked ego-exo modeling for fine-grained view-invariant video representations. *arXiv preprint arXiv:2503.19706*, 2025.

[9] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022.

[10] H. Kuehne, A. B. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of Computer Vision and Pattern Recognition Conference (CVPR)*, 2014.

[11] Dipika Singhania, Rahul Rahaman, and Angela Yao. C2f-tcn: A framework for semi-and fully-supervised temporal action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11484–11501, 2023.

[12] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. In *British Machine Vision Conference*, 2021.

[13] Emad Bahrami, Gianpiero Francesca, and Juergen Gall. How much temporal long-term context is needed for action segmentation? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10351–10361, 2023.

[14] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7082–7092, 2019.

[15] Dipika Singhania, Rahul Rahaman, and Angela Yao. Coarse to fine multi-resolution temporal convolutional network. *arXiv preprint arXiv:2105.10859*, 2021.

[16] Guodong Ding, Fadime Sener, and Angela Yao. Temporal action segmentation: An analysis of modern techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1011–1030, 2023.