Test Time Adaptation for Egocentric Vision

Simone Alberto Peirone[1,*], Mirco Planamente[1,2,3], Barbara Caputo[1,3] and Giuseppe Averta[1]

[1]Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino, 10129, Italy
[2]Italian Institute of Technology, Genova, Italy
[3]Consortium Cini, Italy

Abstract

In the last few years, the technological advancement of wearable cameras has led to an increasing interest in egocentric (first-person) vision, thanks to its ability to capture activities from the user's perspective with applications in a variety of different tasks, from human-object interaction to action prediction and anticipation. In contrast, continuous head movement, variations in lighting conditions and differences in the way humans complete the same task represent a source of bias that strengthens the coupling between the model's predictions and the training domain, affecting its ability to generalize to new environments. Several Domain Adaptation (DA) techniques have been proposed to make models more robust. Among these, Unsupervised Domain Adaptation (UDA) combines labeled source data and unlabeled target data to close the gap between different domains. However, real-world applications require more flexibility, as target samples are often scarce, unrepresentative or even private. Test Time Adaptation (TTA) appears to be a viable solution to these issues, with adaptation performed directly at test time under the simple assumption that input samples provide clues on the actual distribution of the target domain which could be used to improve predictions. With TTA, models undergo multiple adaptation steps at test time by minimizing an adaptation loss on target data and updating normalization statistics. This work provides a comparative analysis of multiple adaptation techniques on the EPIC-Kitchens dataset. Experiments indicate strong accuracy improvements over the unadapted baselines, suggesting that TTA effectively improves model performance in dynamic environments.

Keywords

Egocentric Vision, Domain Adaptation, Test-Time Adaptation

1. Introduction

In recent years, applications of deep learning to computer vision have spread to countless tasks, from image classification and object detection, to action recognition and video generation. Recently, the availability of light and cheap wearable cameras has made first-person video recording much more accessible, marking the birth of *egocentric vision*. The first-person perspective opens up new opportunities for a more in-depth study of how humans interact with the world, paving the way for more human-aware robots. In this work, we focus on Egocentric Action Recognition (EAR), addressing one of the main issues that make this task challenging on egocentric videos.

Model predictions tend to be deeply correlated with the surrounding environment, a problem known as *domain shift*, that may lead to sharp performance drops as the environment changes. On one hand, this motivated the search for additional sources of information, such as audio and optical flow, that are unevenly affected by these variations and can integrate complementary insights. On the other hand, even multiple modalities may not be sufficient when the gap between the data used for training (*source*) and the new samples (*target*) becomes too large. With Unsupervised Domain Adaptation (UDA), the model learns how to adapt to the target domain as part of the training process. UDA assumes that the target samples are available at training time but, depending on the application, this data may not yet exist or may be private. In addition, training must be repeated for each new domain, a serious limitation in terms of required computing power and time. Domain Generalization (DG) faces the same problem from a different perspective that does not require access to the target data during training. The goal of DG is to produce models that generalize better by learning the same concepts across multiple domains. Although the resulting models are more robust, variations in the data distribution are so ubiquitous and subtle in egocentric vision that effective generalisation is difficult to achieve. Even though UDA is not realistic in the egocentric setting, the idea of adapting models to the target samples is powerful.

In this work, we address the application of Test Time Adaptation (TTA) to egocentric vision. TTA moves this

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy.

*Corresponding author. This work was supported in part by the IIT HPC Infrastructure for the Availability of High Performance Computing (Franklin) and CINECA Award through the ISCRA Initiative, in part by FAIR - Future Artificial Intelligence Research, and in part by European Union Next-Generation EU through PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, under Grant PE00000013.

simone.peirone@polito.it (S. A. Peirone); mirco.planamente@polito.it (M. Planamente); barbara.caputo@polito.it (B. Caputo); giuseppe.averta@polito.it (G. Averta)

adaptation process to the test phase, under the simple assumption that input samples provide clues on the actual distribution of the target domain which could be used to improve predictions. Compared to UDA, TTA allows for continuous adaptation which is suitable for the dynamic nature of egocentric vision.

The TTA methods addressed in this work can be divided in two groups: *backpropagation free* and *loss based* methods. The former are based uniquely on the update of the parameters of the network that do not require gradient updates, like the normalization statistics of the Batch Normalization layers of the network. Updates of the normalization statistics ensure that the activation maps within the internal layers of the network remain in the usual range seen during training. The second group, *loss-based* methods, are based on the minimization of a loss function computed over the model predictions or on the intermediate outputs of the network. This work provides an overview of these TTA methods with applications on action recognition and on optical flow estimation.

2. Related Works

Video Domain Adaptation The goal of Unsupervised Domain Adaptation (UDA) is to learn how to bridge the domain gap between a source labeled domain and a non-annotated target domain directly from data. UDA in video tasks focuses on the alignment of the temporal dimension across different domains [1, 2], on exploiting complementary information between different modalities [3] and on learning invariant representations across domains and modalities [4, 5]. The objective of DG is to learn predictive functions that are more robust to out-of-distribution data, using multiple source domains during training but without accessing target data. In this context, RNA-Net [6] introduces a loss to align the feature norms of different modalities. VideoDG [7] aligns the temporal relations between different frames across domains.

Test Time Adaptation Test Time Adaptation (TTA) defers the adaptation phase to test time. Previous works focused on updating the statistics of the normalization layers with those of the target samples [8, 9]. Authors of [10] introduce an auxiliary task that is jointly trained with the main task during training, and finetuned on the target data during the test phase. ViTTA [11] aligns the statistics of different temporal augmentations of the same video to those seen during training. T3A [12] computes pseudo-prototypes for each class label and classifies samples based on the distance from these prototypes.

Table 1
Comparison of several Domain Adaptation approaches.

	Training phase		Testing phase	
	Source	Target	Target	Online
UDA	✓	✓	-	✓
DG	✓	-	-	✓
TTA	-	-	✓	✓

3. Method

The goal of TTA is to update the predictive hypothesis learned by the model at test time, directly on the test data. Unlike UDA and DG, TTA does not require a training process with access to source and target data (Table 1). As labels are not available for the target domain, it is assumed that the two label spaces coincide. In addition, the absence of target labels limits the number of techniques that can be applied to improve the quality of the model assumptions. TTA approaches can be classified into two main categories, depending on whether or not they require gradient updates through backpropagation.

The first approach is based on updating the Batch Normalization statistics collected by the model at training time. Replacing normalization statistics collected on the training data with online estimates of the target data has been shown to improve the robustness of the model in presence of covariance shift [8]. The second approach, loss-based methods, use gradient updates on the model. Among these methods, class losses improve the quality of the predictions, e.g. by optimizing some side properties of the class distribution produced by the model, like its entropy. Feature level losses attempts to adapt the model by operating on its intermediate outputs.

3.1. Batch normalization

The first class of methods for TTA derives from the assumption that domain shift primarily results in a deviation from the batch normalization statistics collected at training time. A natural solution is to replace the running statistics of the normalization layers with the mean and variance of the target samples [8, 9]. These update ensure that the output of the normalization layers remains in the same range encountered at training time, as this is the only portion of the input space explored during training and model behavior may become unpredictable outside this region. The authors of [13] propose α-BN to mix the source and target statistics, to reduce the discrepancy between source and target activations without moving too far from the activations seen during training. For each BN layer i, a new set of normalization statistics $(\tilde{\mu}^{(i)},\ \tilde{\sigma}^{(i)})$ is computed as a weighted average of the old

estimates and the new running estimates:

$$\begin{aligned}\hat{\mu}^{(i)} &= (1-\alpha)\hat{\mu}(i) + \alpha\mathbb{E}(x_t)\\ \hat{\sigma}^{(i)} &= (1-\alpha)\hat{\sigma}^{(i)} + \alpha\sqrt{\text{var}(x_t)},\end{aligned} \tag{1}$$

where $\alpha \in [0,1]$ controls the balance between the old and new statistics. This method does not require back-propagation and can be easily combined with the other adaptation techniques.

3.2. Loss-based TTA

Entropy Minimization (ENT) The *entropy* of a probability distribution measures the uncertainty of its outcomes. In a classification task, it is preferable for the model to produce predictions with very low entropy as higher entropy indicates *confusion* in the predictive function learned by the model. A natural solution might be to update the network weights along the direction that minimizes the entropy of the predictions, as an effort to strengthen the network's confidence and obtain more clear decision boundaries.

$$\ell_{ent}(\mathbf{y}) = -\frac{1}{n}\sum_i\sum_c y_{i,c}\log y_{i,c} \tag{2}$$

Minimization of prediction entropy does not guarantee the best solution, as the model may become even more radicalized on incorrect predictions or produce always the same prediction.

Information Maximization (IM) Information Maximization (IM) mitigates the detrimental effect of the entropy loss in the presence of an unbalanced test dataset by enforcing diversification in the model predictions. The IM loss ℓ_{im} is defined as the sum of the average entropy of the predictions $y_{i,\cdot}$ and the negative entropy of the average prediction $\tilde{p}$.

$$\ell_{im}(\mathbf{y}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C} y_{i,c}\log y_{i,c} + \sum_{c=1}^{C}\tilde{p}_c\log\tilde{p}_c, \tag{3}$$

where n is the batch size and C the number of classes. The first reduces the uncertainty of the predictions, while the second ensures that they remain globally different to avoid the pitfalls of entropy minimization alone.

Minimum Class Confusion (MCC) *Minimum Class Confusion* (MCC) loss [14] targets a reduction of the *pairwise class confusion* of the model predictions, i.e. the situation in which a sample is ambiguously classified into two different classes class c_i and c_j with an equally high probability. The degree of pairwise confusion between two classes i and j is measured using a similarity function, e.g. the dot product, of the predictions $y_{\cdot,i}$ and $y_{\cdot,j}$, which represent the probabilities that the samples in the current batch belong to classes i and j, respectively. More compactly, this is equivalent to computing a confusion matrix defined as:

$$\mathbf{C}_{i,j} = \mathbf{y}_{\cdot,i}^T W_{ii}\mathbf{y}_{\cdot,j}. \tag{4}$$

where W_{ii} controls the contribution of each sample to give more importance to those with a low entropy, i.e. higher confidence. This coefficient is computed as:

$$W_{ii} = \frac{n(1+\exp(-\mathbb{H}(\mathbf{y}_{i,\cdot})))}{\sum_{i=1}^{n}(1+\exp(-\mathbb{H}(\mathbf{y}_{i,\cdot})))}, \tag{5}$$

where n is the batch size and $\mathbb{H}(\mathbf{y}_{i,\cdot})$ is the entropy of the predictions. To rebalance the contributions of the different classes, the $\mathbf{C}$ is row normalized and the loss is defined as the sum of all the off-diagonal entries of the matrix:

$$\ell_{MCC} = \frac{1}{C}\sum_{i=1}^{C}\sum_{j\neq i}^{C}\mathbf{C}_{i,j}. \tag{6}$$

Indeed, the confusion matrix should be ideally close to the identity matrix, indicating low *between-class* confusion and strong prediction confidence.

Relative Norm Alignment (RNA) The Relative Norm Alignment (RNA) loss [6] tackles the *norm-unbalance problem* in the target domain by re-balancing the contributions of the different modalities during the testing phase [15]. Since RNA is an unsupervised loss, its formulation can be trivially extended to the TTA setting:

$$\ell_{RNA} = \left(\frac{\mathbb{E}[h(X^{m_1})]}{\mathbb{E}[h(X^{m_2})]} - 1\right)^2 \tag{7}$$

where X^{m_1} and X^{m_2} are the features extracted from two different modalities m_1 and m_2 and $h(\cdot)$ is the usual L2 norm.

3.3. Beyond Action Recognition

Optical flow is a robust modality for action recognition but its computation is expensive [16]. Several approaches have been proposed to solve the optical flow estimation task directly from data [17, 18] with real-time performance [17] and may be suitable for online EAR [19]. A EAR pipeline could integrate rough estimates of the optical flow at test time, while still relying on more expensive algorithms for training. However, this introduces a gap in the distribution of the optical flow seen during training, compared to the test phase. In this work, we evaluate TTA techniques to adapt existing models to the optical flow estimated with different approaches and to improve the estimation process itself.

TTA for optical flow estimation Optical flow models are typically trained on synthetic datasets [20], which may result in noisy estimates in real-world scenarios. In this section, we propose a loss function inspired by [21] to align the boundaries of the estimated optical flow to the visual edges of the objects in the frame. For horizontal edges, the loss is defined as:

$$\ell_{smooth,x} = \frac{1}{n}\sum_{i}\exp\left(-\frac{\lambda}{3}\sum_{c}\left|\frac{\partial I_c^{(1)}}{\partial x}\right|\right)\left|\frac{\partial V}{\partial x}\right|, \qquad (8)$$

while an equivalent loss is defined for vertical edges.

4. Experiments

All experiments proposed in this work were conducted on the EPIC-Kitchens-55 dataset [22], following the experimental protocol defined by [4], which restricts its analysis to the three largest kitchens in terms of number of labeled samples, hereafter referred to as D1, D2 and D3. Samples are annotated using one of eight verb classes. Unless otherwise specified, models are pre-trained using RNA-Net [6] on two *source* domains D_i and D_j and tested on a possibly different *target* domain D_k. Accuracy is averaged across of all unique combinations $D_i, D_j \longrightarrow D_k$.

Input Five clips are uniformly sampled from the video, each consisting of 16 frames adjacent RGB frames and 16 optical flow frames, with stride 2. The visual samples are augmented using random crops, scale jitters and horizontal flips. RGB and Flow feature extractors use an I3D model [23]. Audio is processed from chunks of 1.28-seconds, converted into a log-spectrogram and fed to a BN-Inception model [24]. Features are then fed to a classifier and the outputs of different clips are averaged to compute the final predictions.

Adaptation protocol The target data set is processed in batches of 32 action samples, each consisting of 5 clips. Clips are adapted according to their index within each sample and an adaptation step is performed for each group to update the model parameters, i.e. the first clips are processed, then the second ones and so on. Once all the clips have been processed, the model is evaluated over the same 5 clips and the predictions are averaged.

4.1. Do we need Time Time Adaptation?

Table 2 reports the Top-1 accuracy for *seen* and *unseen* domains, highlighting the dramatic drop in performances when evaluating on the latter. It must be noted that not all the modalities behave the same. RGB is clearly the worst because of its dependence on visual appearances. Audio is the most robust modality, even though the drop is still quite significant while optical flow has the best accuracy in both configurations.

Table 2
Top-1 accuracy (%) of RNA-Net [6] on training domains (*seen*) and new domains (*unseen*).

	Top-1 Acc. (%)		
	Seen domains	*Unseen* domains	Difference
RGB	53.86	36.64	-17.22
Flow	61.00	50.53	-10.47
Audio	52.34	43.32	-9.02

4.2. Test Time Adaptation

Experiments using BN updates and class losses indicate that different modalities exhibit quite different behaviours after one adaptation step (Table 3). RGB is the worst modality in terms of domain shift, but it improves significantly up to +2.91 percentage points with IM loss. The updates are fairly consistent between the different losses. The only exception is Audio, which suffers a drop in accuracy regardless of the adaptation technique. The adaptation phase may be repeated more than once before producing the final predictions. Although multi-step adaptation may not be realistically applicable due to the increased latency, it is interesting from a theoretical point of view to analyze the robustness of the adaptation techniques and the adaptability of the different modalities. Table 4 shows that RGB+Flow is quite robust to multi-stage adaptation, while the accuracy of RGB+Audio decreases rapidly, as also shown in Fig. 1.

Table 3
Top-1 accuracy (%) after one adaptation step on *unseen* domains. Best in **bold**, second best underlined.

	Adaptation	Top-1 Acc. (%)
RGB	-	36.64
	ENT	39.49 (+2.85)
	MCC	39.51 (+2.87)
	IM	**39.55** (+2.91)
Flow	-	50.53
	ENT	52.84 (+2.31)
	MCC	**52.88** (+2.35)
	IM	52.83 (+2.30)
Audio	-	43.32
	ENT	**42.87** (-0.45)
	MCC	42.83 (-0.49)
	IM	42.80 (-0.52)
RGB+Flow	-	51.34
	RNA	**53.64** (+2.30)
RGB+Audio	-	51.07
	RNA	**52.27** (+1.20)

Table 4
Comparison of class losses on top-1 accuracy after 5 steps. Best in **bold**, second best underlined.

	Adaptation	Top-1 Acc. (%) 1 step	5 steps
RGB+Flow	-	51.36	
	ENT	53.63 (+2.29)	53.35 (+1.99)
	MCC	**53.66** (+2.32)	**53.40** (+2.04)
	IM	53.57 (+2.23)	53.27 (+1.93)
RGB+Audio	-	51.07	
	ENT	52.25 (+1.18)	51.80 (+0.73)
	MCC	**52.54** (+1.47)	**52.11** (+1.04)
	IM	52.21 (+1.14)	51.41 (+0.34)

Comparison with UDA TTA is much more limited compared to UDA as it exploits only a small batch of target data to adapt the model. On the other side, this allows TTA to adapt the model more specifically for the current batch. Therefore, the adaptation steps are based on a very narrow view of the target domain, as all updates are discarded once a new batch of data is available. The combination of RNA-Net and the proposed TTA techniques outperforms state-of-the-art methods on both RGB-Flow and RGB-Audio (Table 5).

Table 5
Comparison with state-of-the-art methods. Best in **bold**, second best underlined.

	Adaptation	Top-1 Acc. (%)
RGB+Flow	Source Only	45.87
	STCDA [5] (UDA)	51.20
	RNA-Net [6] (Multi-DG)	51.34
	Our (MCC, 4 steps)	**53.85**
RGB+Audio	Source Only	45.87
	MM-SADA [4] (UDA)	47.75
	RNA-Net [6] (Multi-DG)	51.07
	Our (MCC, 1 step)	**52.54**

Figure 1: Accuracy gains over multiple adaptation steps.

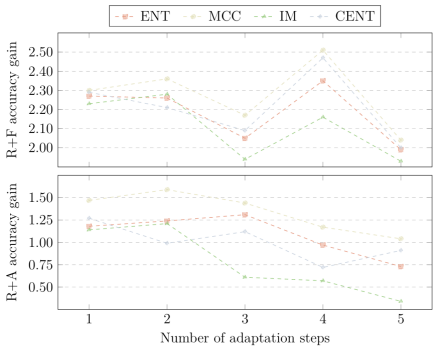

Table 6
Effect of TTA on action recognition using estimated optical flow. The estimated optical flow is denoted as $F^{\dagger}$.

	Adaptation	Top-1 Acc. (%)
Flow	-	50.53
$Flow^{\dagger}$	-	25.04
	IM, 1 step	32.15

Figure 2: Effect of TTA on optical flow estimation.

FlowFormer [18] FlowFormer [18] + TTA

TTA for real-time Optical Flow estimates A model trained on the optical flow computed using the TV-L1 algorithm [16] experience a pronounced performance drop when testes on the estimated optical flow produced by PWC-Net [17], as the accuracy drops from 50.53% to 25.04%. Since TTA has proven to be a viable approach to improve model accuracy in the presence of domain shift, we explore the applicability of TTA to reduce this drop (Table 6). The best TTA approach, 1 step of IM and BN updates, is able to effectively bridge part of the gap between the two sources, recovering more than 7%.

TTA for Optical Flow Estimation Figure 4.2 shows the effect of the L1 smoothness loss on the optical flow frames estimated by FlowFormer [18]. Adaptation is performed online, e.g. without resetting the model after each batch, using batch size 1 and learning rate 0.0001. The FlowFormer model was pretrained on the Sintel dataset [20]. The output of FlowFormer [18] is clean but still exhibits noisy edges and imperfections. With TTA, the estimated flow becomes more sharp and clean.

5. Conclusion

Experiments indicate strong performance drops when models are tested on unseen domains. TTA is able partially cover this gap and to further increase the accuracy of DG models, improving by a significant margin over the best state-of-the-art UDA models when the models are evaluated on unseen domains. Furthermore, TTA demonstrates promising results when applied to the optical flow estimation task and is able to reduce the performance drop experienced by EAR models when tested with the optical flow estimated by real-time approaches.

References

[1] M.-H. Chen, Z. Kira, G. AlRegib, J. Yoo, R. Chen, J. Zheng, Temporal attentive alignment for large-scale video domain adaptation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6321–6330.

[2] A. Jamal, V. P. Namboodiri, D. Deodhare, K. Venkatesh, Deep domain adaptation in action space., in: BMVC, volume 2, 2018, p. 5.

[3] L. Yang, Y. Huang, Y. Sugano, Y. Sato, Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14722–14732.

[4] J. Munro, D. Damen, Multi-modal domain adaptation for fine-grained action recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 122–132.

[5] X. Song, S. Zhao, J. Yang, H. Yue, P. Xu, R. Hu, H. Chai, Spatio-temporal contrastive domain adaptation for action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9787–9795.

[6] M. Planamente, C. Plizzari, E. Alberti, B. Caputo, Domain generalization through audio-visual relative norm alignment in first person action recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1807–1818.

[7] Z. Yao, Y. Wang, J. Wang, P. Yu, M. Long, Videodg: generalizing temporal relations in videos to novel domains, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).

[8] Z. Nado, S. Padhy, D. Sculley, A. D'Amour, B. Lakshminarayanan, J. Snoek, Evaluating prediction-time batch normalization for robustness under covariate shift, arXiv preprint arXiv:2006.10963 (2020).

[9] S. Schneider, E. Rusak, L. Eck, O. Bringmann, W. Brendel, M. Bethge, Improving robustness against common corruptions by covariate shift adaptation, Advances in Neural Information Processing Systems 33 (2020) 11539–11551.

[10] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, M. Hardt, Test-time training with self-supervision for generalization under distribution shifts, in: International conference on machine learning, PMLR, 2020, pp. 9229–9248.

[11] W. Lin, M. J. Mirza, M. Kozinski, H. Possegger, H. Kuehne, H. Bischof, Video test-time adaptation for action recognition, arXiv preprint arXiv:2211.15393 (2022).

[12] Y. Iwasawa, Y. Matsuo, Test-time classifier adjustment module for model-agnostic domain generalization, Advances in Neural Information Processing Systems 34 (2021) 2427–2440.

[13] F. You, J. Li, Z. Zhao, Test-time batch statistics calibration for covariate shift, arXiv preprint arXiv:2110.04065 (2021).

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

[15] M. Plananamente, C. Plizzari, B. Caputo, Test-time adaptation for egocentric action recognition, in: International Conference on Image Analysis and Processing, Springer, 2022, pp. 206–218.

[16] B. K. Horn, B. G. Schunck, Determining optical flow, Artificial intelligence 17 (1981) 185–203.

[17] D. Sun, X. Yang, M.-Y. Liu, J. Kautz, Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8934–8943.

[18] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, H. Li, Flowformer: A transformer architecture for optical flow, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII, Springer, 2022, pp. 668–685.

[19] G. Goletto, M. Planamente, B. Caputo, G. Averta, Bringing online egocentric action recognition into the wild, IEEE Robotics and Automation Letters 8 (2023) 2333–2340.

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

[21] R. Jonschkowski, A. Stone, J. T. Barron, A. Gordon, K. Konolige, A. Angelova, What matters in unsupervised optical flow, in: European Conference on Computer Vision, Springer, 2020, pp. 557–572.

[22] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al., Scaling egocentric vision: The epic-kitchens dataset, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 720–736.

[23] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[24] E. Kazakos, A. Nagrani, A. Zisserman, D. Damen, Epic-fusion: Audio-visual temporal binding for egocentric action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5492–5501.