Self-supervised Test-time Adaptation on Video Data

Fatemeh Azimi [⋆†] **Sebastian Palacio** [⋆†] **Federico Raue** [⋆] **Jörn Hees** [⋆]

Luca Bertinetto [‡] **Andreas Dengel** [⋆†]

[⋆] German Research Center for Artificial Intelligence (DFKI) [‡] Five AI Ltd.
[†] TU Kaiserslautern
{firstname.lastnamel}@dfki.de, luca@robots.ox.ac.uk

Abstract

In typical computer vision problems on video data, pre-trained models are simply evaluated at test time without further adaptation. This general approach inevitably fails to capture potential distribution shifts that exist between training and test data. Adapting a pre-trained model to a new video encountered at test time could be essential to avoid the potentially catastrophic effects of such a shift, or to improve performance when the shift is mild. However, the lack of available annotations in test data prevents practitioners from using vanilla fine-tuning techniques. This paper explores whether the recent progress in self-supervised learning and test-time domain adaptation (TTA) in the image domain can be leveraged to efficiently adapt a model to a previously unseen and unlabelled video. We analyze the effectiveness of several recent self-supervised TTA techniques under the effect of both mild (but arbitrary) and severe domain shifts. From our extensive benchmark on multiple self-supervised dense tracking methods under various domain shifts, we find out that self-supervised TTA methods consistently improve the performance compared to baselines without adaptation, especially in presence of severe covariate shift.

1 Introduction

A fundamental assumption essential to the applicability of many machine learning solutions is the agreement between training and test data distribution. As this premise is often violated in real-world applications [6, 20], it is of high practical importance to seek solutions for adapting a pre-trained model to the test data distribution. Considering the high cost of labeling video data, unsupervised (and self-supervised) methods are of particular interest. Furthermore, the information inherent in the sequence of images could already be used to tailor a pre-trained model towards the data distribution encountered at test time.

In this paper, we analyze how unlabeled video data can be exploited for adapting pre-trained models to the test data distribution, explicitly studying the task of self-supervised dense tracking [9, 13, 27]. In this task, the goal is the frame-by-frame tracking of a pixel-wise mask, starting from a user-initialised mask provided by a hypothetical user in the first frame of the video. We consider the following real-world problem setup where 1) We make use of previously-trained self-supervised models, 2) During training, these models did not have access to data sampled from the test distribution, 3) No labels are used at test time. Adaptation is entirely based on self-suepervised objectives.

Recently, several approaches have studied a similar setup, but for the image domain, under the name of test-time adaptation [18, 23, 26, 28]. However, these methods usually assume the availability of a *diverse* batch of data from the test distribution (to be used for adaptation). In contrast, we study the

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

problem of domain shift for individual videos, where a batch may not contain diverse enough data, which can be problematic for neural networks using batch-normalization (BN) [8]).

Inspired by test-time adaptation methods in the image domain, we study multiple adaptation algorithms for the task of dense tracking by investigating their effectiveness and limitations when exploiting completely unlabeled video data together via a self-supervised objective. We consider two distinct scenarios: arbitrary and severe domain shift. In the former case, we perform test-time adaptations on unseen videos from an unknown distribution, which may be arbitrarily far from the training data. In the latter, we impose severe domain shift by adding artificial perturbations to the video frames.

2 Related Work

Domain Generalization and Test-time Adaptation. Domain generalization considers a scenario where the target data distribution is unavailable during the training phase [32]. The goal is to improve the performance on the target domain with a focus on enhancing the *training* process. In this respect, [3] proposes a multi-task setup and shows that adding an auxiliary self-supervised objective improves the generalization to unseen domains. [16] proposes a meta-learning approach in which the objective for improving the generalization is learned itself, in contrast with methods that utilize manually designed loss functions [1, 17]. Guo *et al.*[5] identify the BN layer as one of the factors that can lead to poor confidence calibration in the network output at test-time. Several works [2, 15, 24, 31] have studied this aspect in an attempt to adapt the normalization layer to the target distribution and improve the performance on the test data.

Unlike the approaches mentioned above, test-time adaptation only leverages the data available at *test* time. In this respect, Sun *et al.*[26] propose a multi-task setup using supervised and self-supervised objectives, where an auxiliary loss is used to further fine-tune the network during inference. In [28], the authors utilize entropy minimization [4, 21, 22, 25] to modify the modulation parameters of the BN layer to mitigate the impact of covariate shift between the training and testing data distributions. Furthermore, [18, 23] suggest updating the normalization statistics of the BN layer as an effective way for adapting the features to the target domain.

In this work, we build on test-time adaptation approaches, as we adapt the pre-trained models to the new unseen domains in video data. More details on algorithms employed in this paper can be found in subsection A.1.

Self-supervised Dense Tracking. In recent years there has been a surge of interest in self-supervised methods for different applications [10], including dense tracking. In [27], the authors use video colorization as the self-supervised objective. Multiple works [11, 12, 13, 14, 29, 30] improved this algorithm with various modifications such as and incorporating memory and cycle consistency into the architecture and the training process. In a different line of work, [9] suggests a framework where the video is processed into a graph by dividing each frame into multiple patches (nodes). They train the embeddings by performing a random walk on the constructed graph using a cycle consistency objective, creating a palindrome from the video frames. In this paper, we employ [9] and [13] as our baselines and provide further details about these algorithms in subsection A.2.

3 Problem Setup

This section discusses our proposed problem formulation and experimental setup. Our primary focus lies on studying the impact of covariate shift in the task of self-supervised dense tracking and the possible remedies utilizing the unlabeled video data inspired by test-time adaptation literature from the image domain [18, 23, 26, 28].

We initially contemplate a hypothesis where each video is considered as an individual domain with an arbitrary distribution shift w.r.t. the training data. Next, we experiment with enforced domain shift by manually adding various perturbations to the test videos, similarly to what is done in [6] (though for images). To this end, we ask the following questions:

- Assuming each video represents a specific domain, how effective are the current test-time adaptation methods when applied to the task of dense tracking?

- Considering the self-supervised setups for dense tracking, can finetuning the model on the target video (essentially overfitting to a specific video domain using the self-supervised objective) further improve performance?
- In the case of clear domain shift such as the one described by the perturbations in [6], how effective are these adaptation methods for recovering the performance in self-supervised dense tracking tasks?

To answer to these questions, we experiment with modified variants of three recent approaches for test-time adaptation in image domain, namely **Prediction-time BN** [18, 23], **TENT** [28], and **TTT** [26]. Our setup is different from the one assumed by these methods in the following ways: **First**, The mentioned methods are developed for image classification and assume a diverse batch of data from the target distribution is available at test time. In our setup, each video is considered an individual domain, and the frames sampled from a single video comprise the batch, meaning the batch might not contain enough diversity. **Second**, All these methods build on top of models trained in a supervised manner, while we examine baselines that are trained in a self-supervised fashion. **Third**, We use a modified version of TENT [28] where the self-supervised objective substitutes the entropy loss to alleviate the dependency to the first frame label during the adaptation phase. We refer to this adapted version as TENT*. **Fourth**, Unlike the prediction-time BN scenario in [18, 23], the captured statistics from a video sequence may not be diverse enough, so replacing the normalization statistics in the normalization layer with those collected from the video frames might hurt the performance. Therefore, we experiment with different momentum values as:

$$\hat{x} = (1 - \alpha) \times x_{old} + \alpha \times x_{new} \tag{1}$$

where $\alpha \in [0, 1]$ is the momentum value and x_{old} and x_{new} are the collected statistics from the training data and the video under test, respectively. In our experimental setup, we consider two different scenarios, apt for offline and online applications. In the former, all the video frames are available prior to inference. In the latter, we have access to a limited amount of data from the test domain, but not to the test frames themselves. In this paper, we use the recent self-supervised methods VideoWalk [9] and MAST [13] as our baselines. Further details about these algorithms are provided in subsection A.2.

4 Experiments

In this section, we illustrate the experimental results obtained on the DAVIS-2017 dataset [19], a standard benchmark for evaluating dense tracking. Following the usual procedure in dense tracking [19], we report the J and F scores of the segmented object masks. These metrics indicate the intersection-over-union and object boundary accuracy respectively. According to section 3, Table 1 presents the results for offline and online setup. In each block, the first row shows the average of J and F scores of the baselines without any adaptation in cursive. The remaining rows show the difference w.r.t. to the first row when using each of the adaptation approaches.

In the first block of rows, we investigate the efficacy of test-time adaptation with a self-supervised objective on the test data with arbitrary domain shifts (without any added perturbation). As we are working with self-supervised baselines, it is interesting to understand whether further tuning on a specific video is helpful and to which extent it can improve the performance on the downstream task. Next, we study a scenario with a substantial domain shift between the training and testing data distributions. In this respect, we follow the procedure in [7] and impose a synthetic covariate shift to the video frames. In particular, we experiment with Gaussian noise, Motion Blur, Fog, and Snow perturbations (with the maximum level of severity) as described in [6].

As can be seen from the results in Table 1, self-supervised test-time adaptation on data without perturbation slightly improves the results, while it considerably decreases the adverse effect of covariate shift for data with severe domain shift. The behavior in an arbitrary domain shift scenario (without perturbation) implies that in situations with mild distribution shift, overfitting to the current self-supervised objectives does not fully transfer to the downstream task and only marginally improves the performance. However, these methods can successfully adapt the features to the target domain when there is a severe distribution shift between the training and testing data. Interestingly, in most cases, updating the normalization statics (BN column) has an equal or superior positive impact on the dense tracking accuracy despite its simplicity. However, we note that the "fog" perturbation is an exception where both MAST and VideoWalk methods achieve considerably better accuracy with

Dense Tracking (offline)		Dense Tracking (online)		Test-time Adaptation			
VideoWalk	MAST	VideoWalk	MAST	BN	TENT*	TTT	Noise
67.39	*64.95*	*71.95*	*68.98*				—
+0.78	**+0.55**	+0.83	**+1.04**	✓			
+0.77	+0.49	**+0.84**	+0.25		✓		
+0.82	+0.22	+0.74	+0.33			✓	
60.74	*34.09*	*66.16*	*42.44*				Gaussian
+2.01	**+20.18**	+2.33	**+18.74**	✓			
+2.18	+18.38	**+3.82**	+16.54		✓		
+2.82	+18.11	+2.20	+15.98			✓	
65.86	*60.97*	*70.10*	*67.26*				Motion Blur
+0.60	**+0.65**	+1.32	**+0.23**	✓			
+0.38	+0.02	**+1.37**	-0.21		✓		
+0.15	-0.03	+1.13	-0.51			✓	
52.83	*52.1*	*57.82*	*59.10*				Snow
+2.21	**+0.80**	**+2.70**	+0.49	✓			
+2.40	+0.24	+2.48	**+0.88**		✓		
+3.36	+0.36	+1.95	+0.27			✓	
22.80	*36.8*	*27.76*	*44.23*				Fog
+10.99	0.00	+10.70	0.00	✓			
+12.12	+2.86	+9.45	+3.76		✓		
+18.56	**+9.18**	**+14.14**	**+9.39**			✓	

Table 1: The average of J and Fscores for VideoWalk [9] and MAST [13] self-supervised dense tracking methods on the DAVIS-2017 validation set in offline and online setups. Results in italic correspond to absolute metrics, followed by rows indicating the absolute gain in performance when using one of the test-time adaptation methods. For each column (within a block), best results are outlined in bold.

TENT* and TTT algorithms. Furthermore, the results show a similar pattern in offline and online scenarios, suggesting that performing test-time adaptation is beneficial for both circumstances.

For the results shown in column BN, we experimented with different momentum values and updated the normalization statistics according to Equation 1. Here the results are provided with the best-found momentum, and additional results are provided in subsection A.4. We experimentally observed that partially updating the normalization statistics with those from the target domain alleviates the impact of covariate shift. However, completely replacing them (momentum value of 1 in Equation 1) can deteriorate the performance. This behavior is likely due to the lack of diversity across video frames within the same batch. Finally, we observe different trends in the behaviour of VideoWalk and MAST. For instance, for the "fog"-type perturbation, VideoWalk benefits from updating the normalization statistics, whereas for MAST it is better to keep the statistics unchanged. This can result from different training objectives in these approaches: the self-supervised loss in MAST is purely based on color information, while VideoWalk additionally utilizes semantic information.

5 Conclusion

In this work, we investigated the role that self-supervision can have in alleviating the harmful effect of distribution mismatch between train and test video data. We considered two scenarios of practical relevance. One, for offline applications, in which the entire video sequence is available in advance. Another, for online applications, in which instead we are interested in real-time inference and only have access to some unlabeled data from the target domain prior to inference. We studied the behavior of two best-performing self-supervised dense tracking algorithms in the presence of several domain shifts. Our experimental results confirm that self-supervised test-time adaptation is an effective method for decreasing the impact of covariate shift in dense tracking. For future work, we plan to employ transductive approaches as an effective way for utilizing the information of the unlabeled video data.

References

[1] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in Neural Information Processing Systems*, 31:998–1008, 2018.

[2] Collin Burns and Jacob Steinhardt. Limitations of post-hoc feature alignment for robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2525–2533, 2021.

[3] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019.

[4] Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. *CAP*, 367: 281–296, 2005.

[5] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[6] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *iclr*, 2019.

[7] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[9] Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. *arXiv preprint arXiv:2006.14613*, 2020.

[10] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[11] Shu Kong and Charless Fowlkes. Multigrid predictive filter flow for unsupervised learning on videos. *arXiv preprint arXiv:1904.01693*, 2019.

[12] Zihang Lai and Weidi Xie. Self-supervised learning for video correspondence flow. *arXiv preprint arXiv:1905.00875*, 2019.

[13] Zihang Lai, Erika Lu, and Weidi Xie. Mast: A memory-augmented self-supervised tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2020.

[14] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. *arXiv preprint arXiv:1909.11895*, 2019.

[15] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

[16] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *International Conference on Machine Learning*, pages 3915–3924. PMLR, 2019.

[17] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013.

[18] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.

[19] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[20] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

[21] Subhankar Roy, Aliaksandr Siarohin, Enver Sangineto, Samuel Rota Bulo, Nicu Sebe, and Elisa Ricci. Unsupervised domain adaptation using feature-whitening and consensus loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9471–9480, 2019.

[22] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019.

[23] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33, 2020.

[24] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 68–83. Springer, 2020.

[25] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018.

[26] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *icml*, 2020.

[27] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 391–408, 2018.

[28] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[29] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019.

[30] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, 2021.

[31] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *European Conference on Computer Vision*, pages 420–439. Springer, 2020.

[32] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *arXiv preprint arXiv:2103.02503*, 2021.