Self-supervised Test-time Adaptation on Video Data

Fatemeh Azimi [1,2], Sebastian Palacio [1,2], Federico Raue[2], Joern Hees[2], Luca Bertinetto[3], Andreas Dengel[1,2]

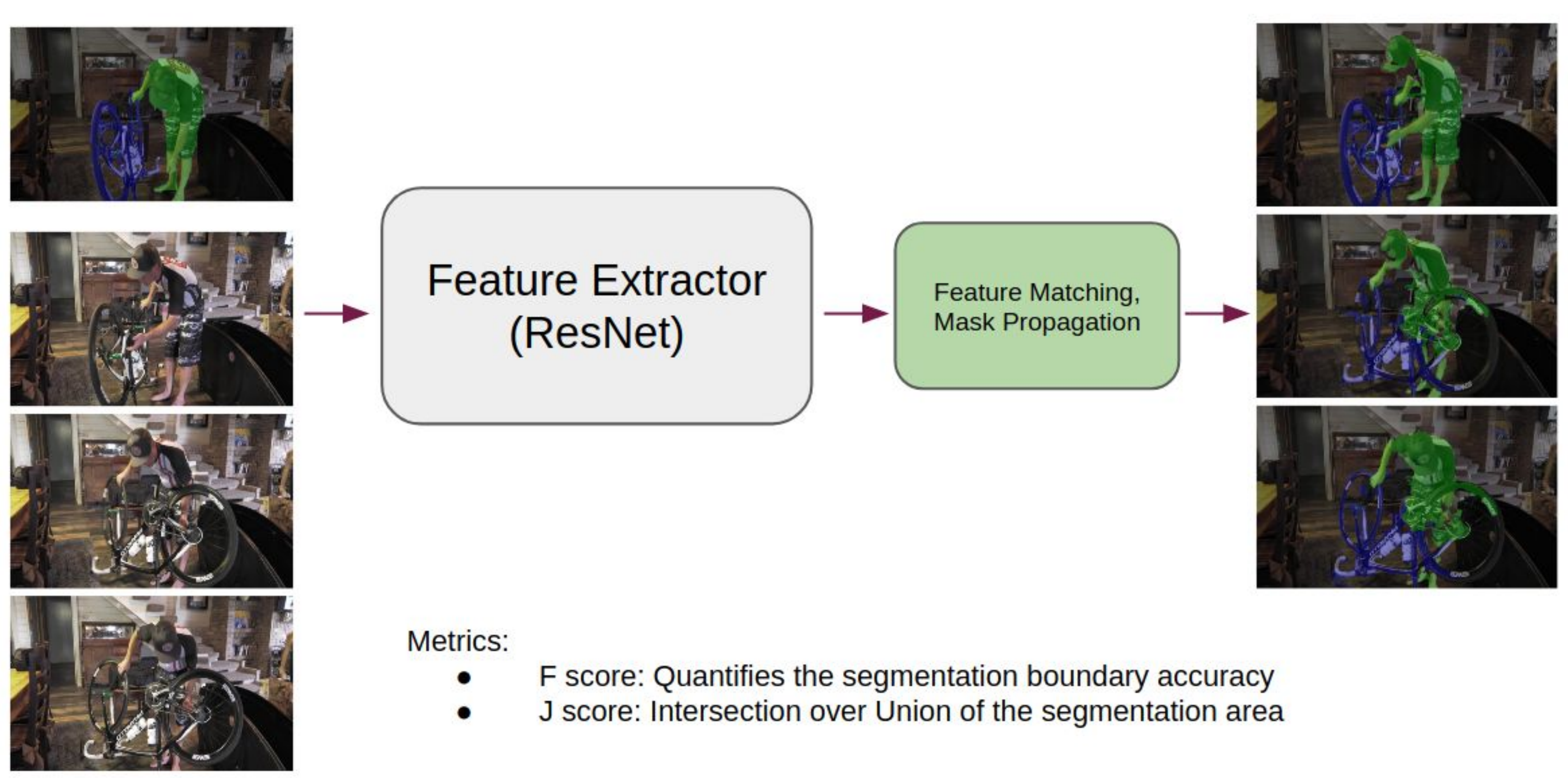
[1]TU Kaiserslautern, [2]DFKI, [3]Five AI Ltd.

Abstract

- In typical computer vision problems, 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.
- The lack of available annotations in test data prevents practitioners from using vanilla fine-tuning techniques.
- In this work, we explore 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.

Problem Formulation

- **Self-supervised Dense Tracking**
  - MAST[1]: Colorization
    - Search for correspondences by colorizing video frames.
    - Improved performance via using memory bank, LAB color space, and using regression instead of classification loss.
  - VideoWalk[2]: Contrastive Random Walk
    - Generate a palindrome from the video frames.
    - Divide each frame into multiple nodes (patches).
    - Track similar nodes via minimizing a cycle consistency objective.



Metrics:
- F score: Quantifies the segmentation boundary accuracy
- J score: Intersection over Union of the segmentation area

Problem Formulation

- **Test-time Adaptation**
  - Prediction-time BN[3]: Updates the BN statistics with a mi=omentum value between 0 and 1.

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

  - TENT*: Follows the proposed method in [4] where the affine parameters in the BN layer are updated, but self-supervised objective is employed instead of Entropy minimization.
  - Test-time Training (TTT)[5]: The whole network weights are tuned via minimizing the self-supervised objective.
- **Frame Selection**
  - Offline: All the frames are used for adaptation.
  - Online: The first half of the video is used for training and the second half for evaluation.

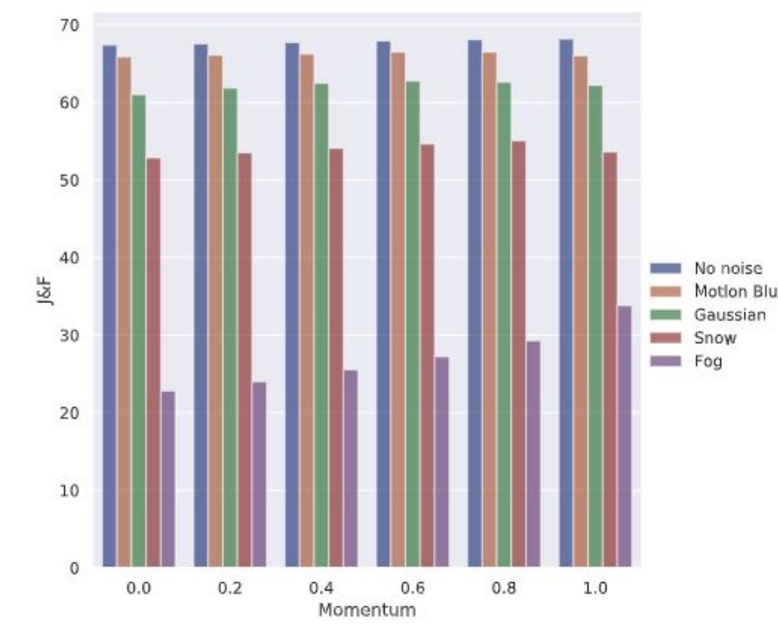
Experimental Results DAVIS2017

- **Arbitrary domain shift**
  - Each video considered as an individual domain and hypothesized to have arbitrary/mild domain shift wrt training data.
  - Marginal improvement, mainly due to updating the BN statistics.
- **Enforced domain shift**
  - Severe domain shift via manually adding noise to the input data.
  - Self-supervised TTA is highly effective in compensating for covariate shift.
  - The choice of TTA method depends on the perturbation variant.

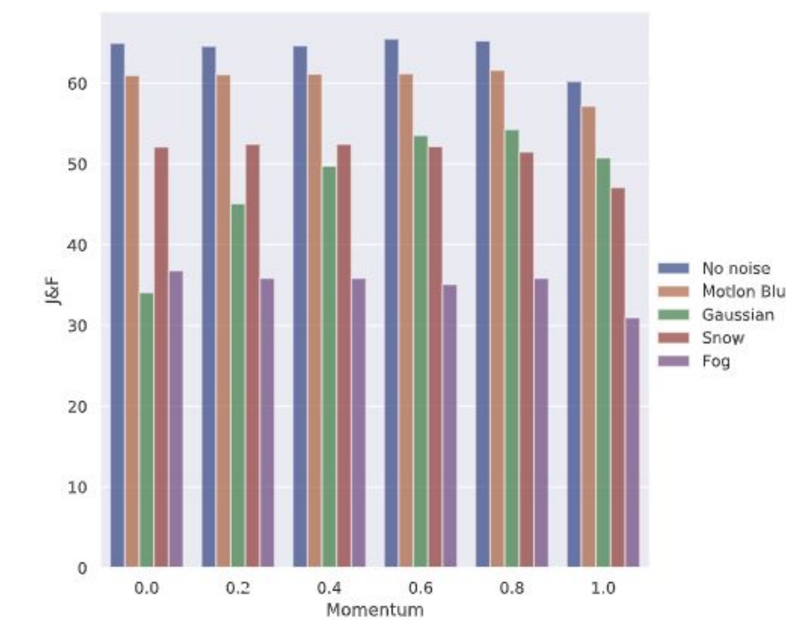
Dense Tracking (Offline)				Dense Tracking (Online)				Test-time Adaptation			
VideoWalk		MAST		VideoWalk		MAST		BN	TENT*	TTT	Noise
J	*F*	*J*	*F*	*J*	*F*	*J*	*F*				
64.38	*70.40*	*62.95*	*66.94*	*69.46*	*74.43*	*67.11*	*70.85*				—
+1.00	**+0.56**	**+0.47**	+0.62	+0.67	**+0.99**	**+1.04**	**+1.04**	✓			
+1.04	+0.50	+0.32	**+0.65**	**+0.70**	+0.97	+0.20	+0.30		✓		
+1.17	+0.47	+0.09	+0.34	+0.64	+0.84	+0.27	+0.39			✓	
58.40	*63.08*	*32.70*	*35.48*	*64.43*	*67.89*	*41.51*	*43.36*				Gaussian
+1.85	+2.16	**+19.82**	**+20.54**	+2.07	+2.58	**+18.21**	**+19.26**	✓			
+1.91	+2.44	+17.98	+18.77	**+3.73**	**+3.91**	+15.90	+17.17		✓		
+2.67	**+2.97**	+18.06	+18.15	+2.11	+2.20	+15.37	+16.58			✓	
62.97	*68.75*	*58.49*	*63.45*	*67.69*	*72.50*	*64.54*	*69.99*				Motion Blur
+0.69	**+0.51**	**+0.49**	**+0.80**	+1.01	+1.62	**+0.35**	**+0.10**	✓			
+0.41	+0.34	-0.10	+0.13	**+1.04**	**+1.69**	-0.21	-0.22		✓		
+0.18	+0.11	+0.12	-0.18	+0.97	+1.28	-0.58	-0.43			✓	
50.89	*54.77*	*51.12*	*53.08*	*56.44*	*59.20*	*58.51*	*59.68*				Snow
+1.63	+2.78	**+0.83**	**+0.77**	**+2.60**	**+2.80**	+0.51	+0.46	✓			
+1.99	+2.80	+0.14	+0.34	+2.43	+2.52	**+0.77**	**+0.99**		✓		
+2.79	**+3.92**	+0.32	+0.39	+1.98	+1.91	+0.15	+0.38			✓	
19.27	*26.32*	*35.55*	*38.05*	*24.76*	*30.76*	*43.42*	*45.03*				Fog
+11.23	+10.76	0.00	0.00	+11.54	+9.860	0.00	0.00	✓			
+12.01	+12.23	+3.09	+2.66	+9.67	+9.22	+3.83	+3.51		✓		
+18.70	**+18.42**	**+9.85**	**+8.50**	**+14.07**	**+14.21**	**+9.24**	**+9.54**			✓	

Ablation on Momentum in Prediction-time BN

- We experimentally observed that replacing the BN statistics with the once collected from the test video leads to suboptimal performance.
- This could be due to the fact that a single video may not capture diverse-enough scenes.



(a) VideoWalk



(b) MAST

Visualization of studied perturbations

Conclusion

- We investigated the role of TTA in alleviating the impact of covariate shift in self-supervised VOS.
- Based on practical considerations, we studied two scenarios namely offline and online TTA.
- Our results demonstrate while self-supervised TTA marginally improves the performance for arbitrary domain shift, it is highly effective when dealing with severe data distribution shift in both online and offline stups.

References

[1] Lai, Z., Lu, E. and Xie, W., 2020. MAST: A memory-augmented self-supervised tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6479-6488).

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

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

[4] Wang, D., Shelhamer, E., Liu, S., Olshausen, B. and Darrell, T., 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726.*

[5] Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A. and Hardt, M., 2020, November. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning* (pp. 9229-9248). PMLR.