Self-Distilled Masked Auto-Encoders are Efficient Video Anomaly Detectors (Supplementary)

Nicolae-Cătălin Ristea[1,2,◇], Florinel-Alin Croitoru[1,◇], Radu Tudor Ionescu[1,3,*], Marius Popescu[1,3,]
Fahad Shahbaz Khan[4,5], Mubarak Shah[6]
[1]University of Bucharest, Romania, [2]NUST Politehnica Bucharest, Romania,
[3]SecurifAI, Romania, [4]MBZ University of Artificial Intelligence, UAE,
[5]Linköping University, Sweden, [6]University of Central Florida, US

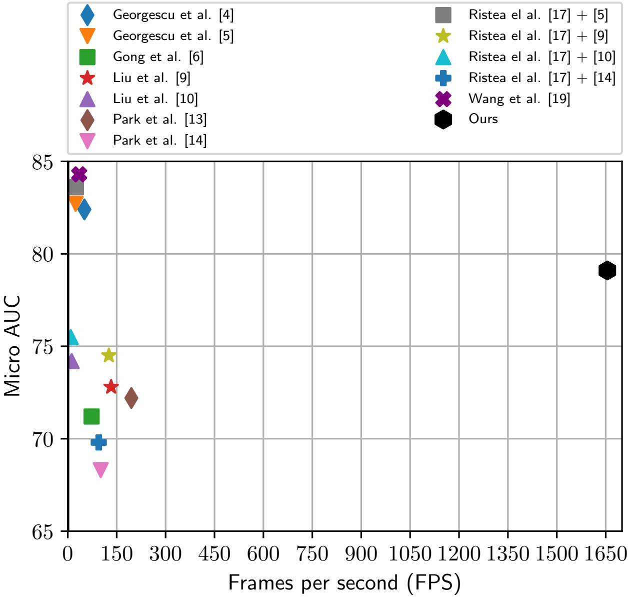


Figure 1. Performance versus speed trade-offs for our self-distilled masked AE and several state-of-the-art methods [4–6, 9, 10, 13, 14, 17, 19] (with open-sourced code), on the ShanghaiTech data set. The running times of all methods are measured on a computer with one Nvidia GeForce GTX 3090 GPU with 24 GB of VRAM. Best viewed in color.

Abstract

In the supplementary, we present localization results, as well as additional ablation and qualitative results. Finally, we discuss the connections between our approach and other frameworks based on masked auto-encoders.

1. Additional Results

Performance-speed trade-off. In the main article, we compared the performance-speed trade-off of our masked AE with other state-of-the-art methods on the Avenue data set. To demonstrate that our superior trade-off is maintained across data sets, we hereby analyze the trade-offs of several methods, including our own, on the ShanghaiTech data sets. The results illustrated in Figure 1 clearly indicate that our method is significantly faster than competing methods, while surpassing the other frame-level anomaly detection methods. This observation confirms the consistency of our trade-off across data sets.

Type	Method	Avenue		Shanghai		UBnormal		FPS
		RBDC	TBDC	RBDC	TBDC	RBDC	TBDC	
Object-centric	[2]	47.83	85.26	47.14	85.61	25.63	63.53	20
	[4]	57.00	58.30	42.80	83.90	19.71	55.80	51
	[5]	65.05	66.85	41.34	78.79	25.43	56.27	24
	[8]	15.77	27.01	20.65	44.54	-	-	-
	[10]	41.05	86.18	44.41	83.86	-	-	12
	[12] + [2]	49.01	85.94	47.73	85.68	-	-	20
	[12] + [5]	66.04	65.12	40.52	81.93	-	-	31
	[12] + [10]	46.49	86.43	45.86	84.69	-	-	10
	[17] + [5]	65.99	64.91	40.55	83.46	-	-	31
	[17] + [10]	62.27	89.28	45.45	84.50	-	-	10
Frame or cube level	[1]	-	-	-	-	0.04	0.05	37
	[9]	19.59	56.01	17.03	54.23	-	-	28
	[12] + [9]	23.79	66.03	19.13	61.65	-	-	26
	[15]	35.80	80.90	-	-	-	-	-
	[16]	41.20	78.60	-	-	-	-	-
	[17] + [9]	20.13	62.30	18.51	60.22	-	-	26
	[18]	-	-	-	-	0.01	0.01	56
	Ours	46.77	66.58	26.42	66.67	23.58	50.36	1655

Table 1. RBDC and TBDC scores (in %) of several state-of-the-art frame-level, cube-level and object-level methods versus our self-distilled masked AE on Avenue, ShanghaiTech and UBnormal. The top three scores for each category of methods are shown in red, green, and blue. All reported running times (including those of the baselines) are measured on a machine with an Nvidia GeForce GTX 3090 GPU with 24 GB of VRAM.

Anomaly localization results. To measure anomaly lo-

*corresp. author: raducu.ionescu@gmail.com; ◇equal contribution

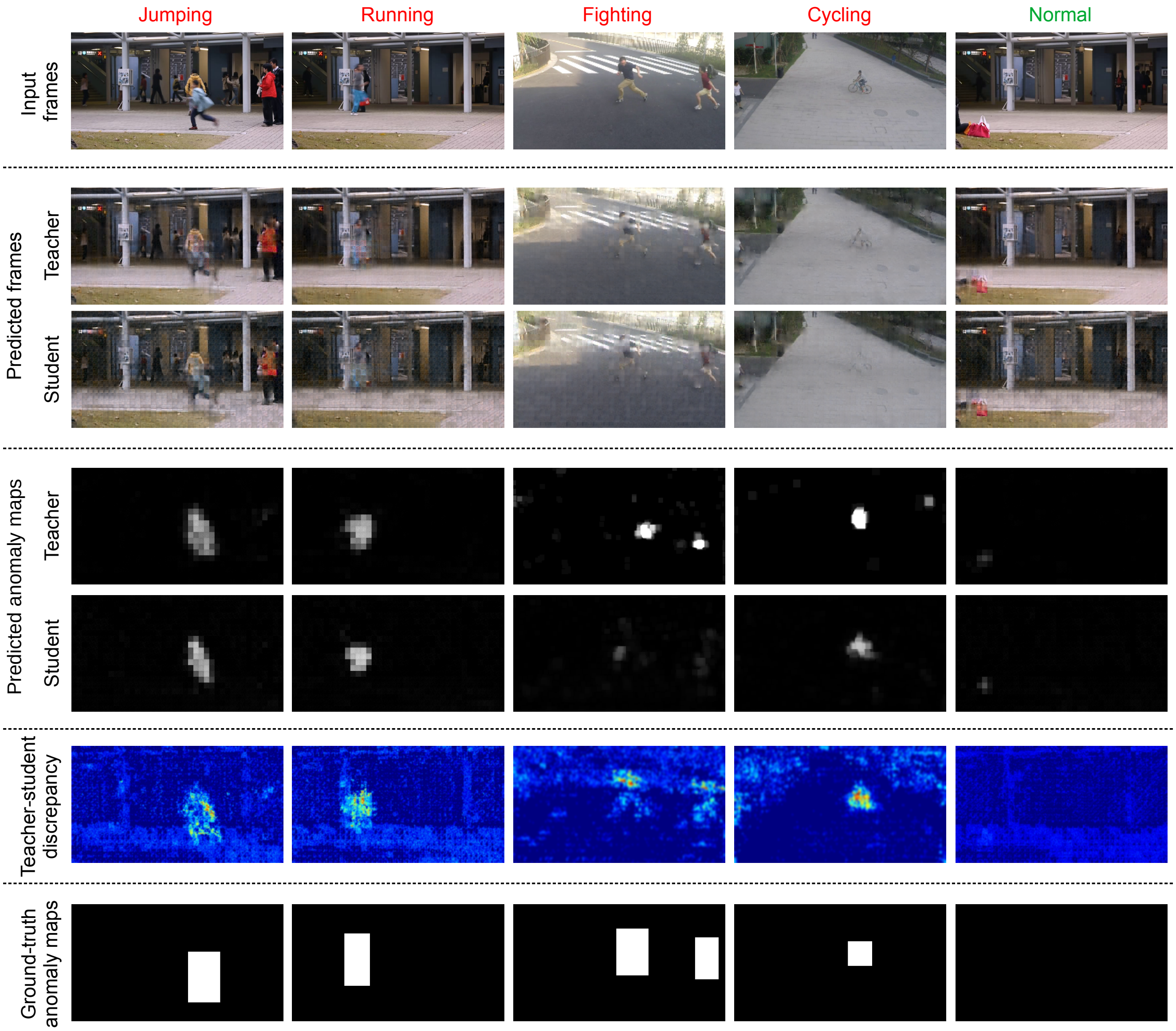


Figure 2. Examples of frames and anomaly maps reconstructed by our teacher and student models. Additionally, the differences (discrepancy maps) between the teacher and student outputs are shown in the sixth row. The first four columns correspond to abnormal examples from the Avenue and ShanghaiTech data sets, while the last column corresponds to a normal example. Best viewed in color.

calization performance, we employ the recently proposed Region-Based Detection Criterion (RBDC) and Track-Based Detection Criterion (TBDC) [15]. Following Ramachandra *et al*. [15], we set the region overlap threshold to 0.1 and the track overlap threshold to 0.1, which allows us to directly compare with other methods reporting the RBDC and TBDC scores. In Table 1, we report the RBDC and TBDC scores of our method versus frame-level and object-centric methods, on the Avenue, ShanghaiTech and UBnormal data sets.

When compared with frame-level and cube-level methods, our approach obtains the best RBDC scores on all three data sets. Furthermore, our method outperforms all other frame-level and cube-level methods on ShanghaiTech and UBnormal, in terms of TBDC. The most dramatic differences in favor of our method are reported on the UBnormal data set. Notably, our method also outperforms some of the object-centric approaches, in terms of both RBDC and TBDC. Considering that our approach is a frame-level method, its anomaly localization results are remarkable. Not only that our method is generally better than frame-level and cube-level methods in terms of both RBDC and TBDC, but its processing speed is significantly higher.

Qualitative results. In Figure 2, we illustrate the frame re-

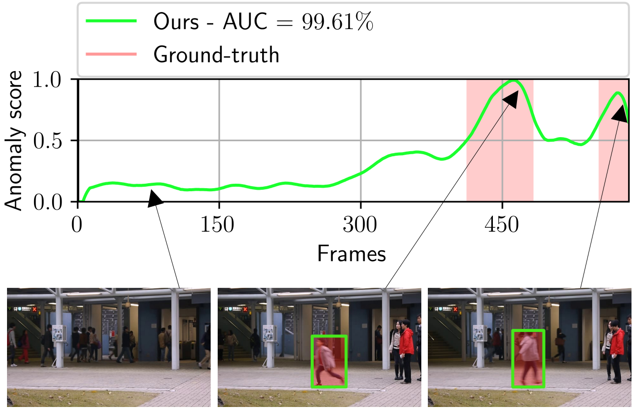


Figure 3. Predictions for test video 07 from Avenue. The abnormal bounding boxes are given by the convex hull of the patches labeled as abnormal. Best viewed in color.

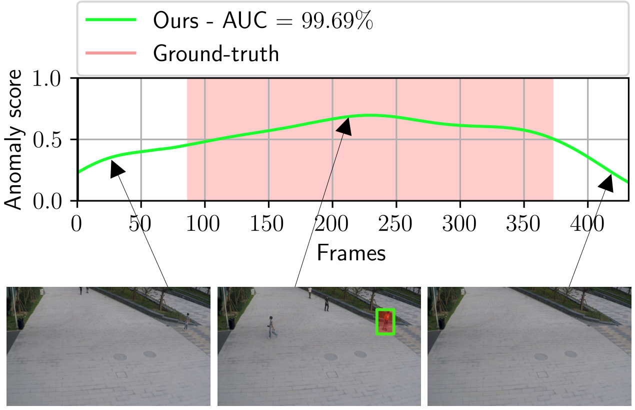


Figure 4. Predictions for test video 01_0015 from ShanghaiTech. The abnormal bounding boxes are given by the convex hull of the patches labeled as abnormal. Best viewed in color.

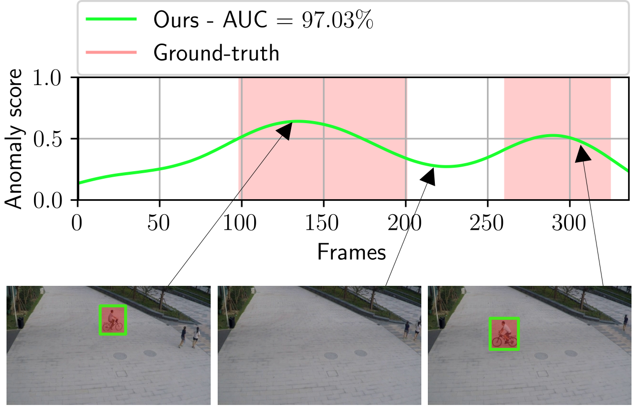


Figure 5. Predictions for test video 01_0051 from ShanghaiTech. The abnormal bounding boxes are given by the convex hull of the patches labeled as abnormal. Best viewed in color.

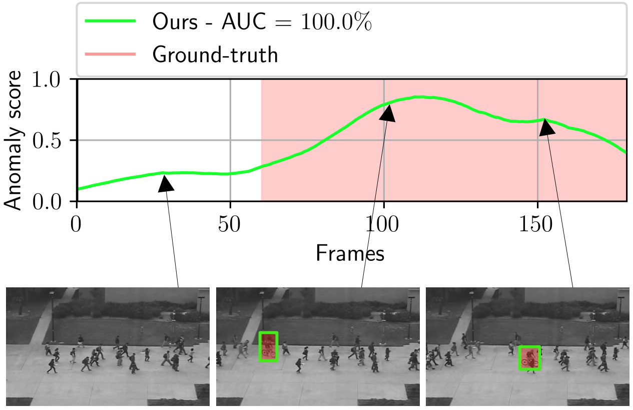


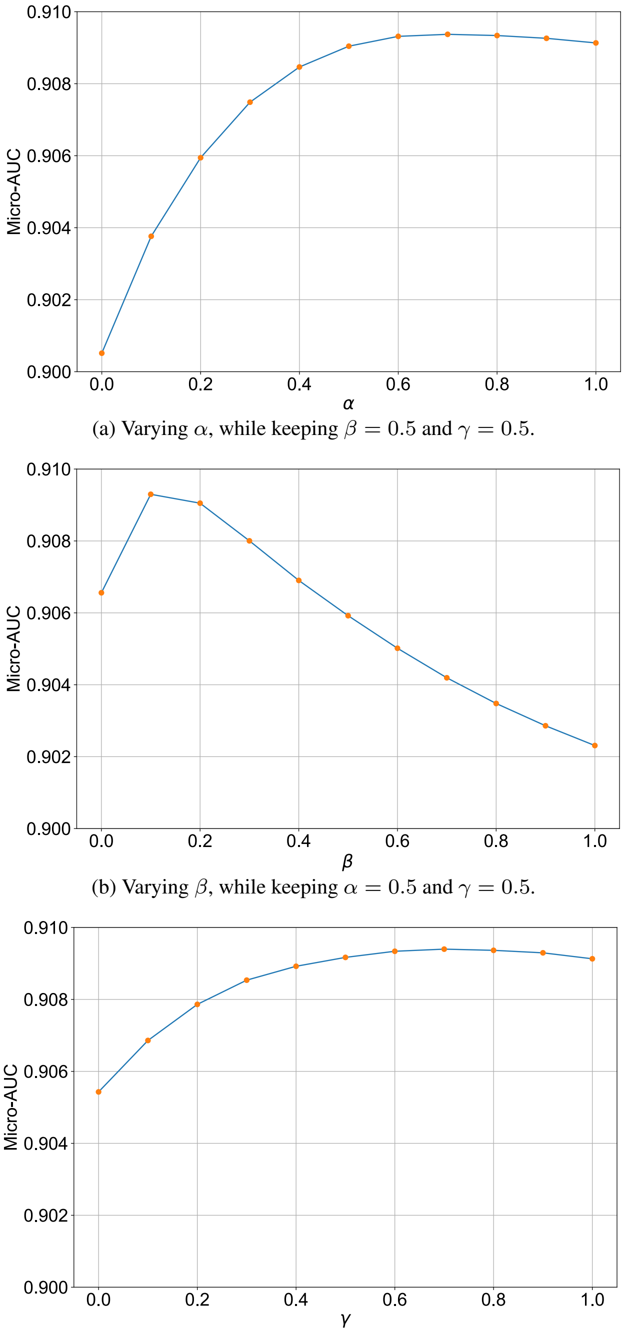
Figure 6. Predictions for video Test001 from UCSD Ped2. The abnormal bounding boxes are given by the convex hull of the patches labeled as abnormal. Best viewed in color.

constructions and the anomaly maps returned by the teacher and student models for five input frames. We keep the same five examples as in the main paper, essentially adding the outputs from the student model, as well as the discrepancy maps between the teacher and the student. For the first four examples, which are abnormal, we can see that the frame reconstructions of both teacher and student models are deficient in the anomalous regions, as desired. Moreover, in the fourth example, the student entirely removes the bicycle from its reconstructed output, which triggers a true positive detection. The anomaly maps generated by the teacher are generally better than the ones generated by the student. The latter maps are well aligned with the ground-truth anomalies, but the predicted anomalies cover a smaller than expected area. However, the discrepancy maps exhibit intense disagreements in the anomalous regions, indicating that the discrepancy maps are good indicators for abnormal events. For the normal example depicted in the fifth column, the anomaly and discrepancy maps do not show any pixels with high anomaly scores, confirming that our method yields the desired effect.

Another interesting remark is that the reconstructed frames returned by the student are worse than those of the teacher. This happens because the student learns to reconstruct the teacher's output frames instead of the original input frames. Nevertheless, the reconstruction power of the student is less important to us, *i.e.* we care more about obtaining discrepancy maps that are highly correlated with the abnormal events. As discussed above, our student works as expected, helping the teacher to better predict the anomalies.

In Figure 3, we illustrate the anomaly scores for test video 07 from the Avenue data set. On this test video, our model reaches an AUC higher than 99%, being able to accurately identify the person running and jumping around.

In Figure 4, we showcase the anomaly scores for video 01_0015 from the ShanghaiTech test set. As in the previous example, our model obtains an AUC higher than 99%, returning higher anomaly scores when the skateboarder



(a) Varying α, while keeping $\beta = 0.5$ and $\gamma = 0.5$.

(b) Varying β, while keeping $\alpha = 0.5$ and $\gamma = 0.5$.

(c) Varying γ, while keeping $\alpha = 0.5$ and $\beta = 0.5$.

Figure 7. Micro AUC scores on the Avenue data set, while varying the hyperparameters α, β and γ controlling the anomaly score contributions of the teacher decoder, the teacher-student discrepancy, and the classification head, respectively. Each hyperparameter is varied between 0 and 1, while keeping the others fixed to 0.5.

passes through the pedestrian area.

In Figure 5, we present the anomaly scores for video 01_0051 from the ShanghaiTech test set. Our model reaches an AUC of 97.03% on this video, being able to flag and locate the abnormal event, namely riding a bike into a pedestrian area.

CvT block type	AUC		FPS
	Micro	Macro	
MLP [20]	89.2	88.1	1454
Pointwise convolutions (ours)	91.3	90.9	1655

Table 2. Micro and macro AUC scores (in %) on Avenue [11] with pointwise convolutional layers versus fully connected layers in the CvT transformer blocks.

In Figure 6, we illustrate the anomaly scores for video Test001 from UCSD Ped2. Here, our model reaches an AUC of 100%, being able to perfectly differentiate between normal and abnormal events.

Ablating pointwise convolutions. We next assess the impact of replacing the fully connected layers inside the vanilla CvT blocks [20] with pointwise convolutions. The results presented in Table 2 show that our minor architectural change leads to a speed boost of 211 FPS and an increase of 2.1% in terms of the micro AUC. The results confirm that the pointwise convolutions provide a superior trade-off between accuracy and speed.

Ablating anomaly score components. In Figure 7, we illustrate the impact of α, β, and γ on the micro AUC score computed on the Avenue data set. These hyperparameters are the weights associated to the three anomaly score components, namely the teacher decoder, the teacher-student discrepancy, and the classification head. We note that all weight configurations lead to micro AUC scores higher than 90%, indicating that our method is fairly robust to suboptimal tuning of α, β, and γ. Indeed, the vast majority of combinations lead to micro AUC scores that are higher than the micro AUC scores of all other frame-level and cube-level methods evaluated on Avenue (see Table 1 from the main paper). Nonetheless, we generally observe that the teacher decoder and the classification head should have higher weights than the teacher-student discrepancy.

2. Extended Related Work

Driven by the goal of learning better high-level representations, some studies, such as [3, 21], tried to modify the pretraining phase of the masked AE [7]. Since these methods [3, 21] may appear to be related to our approach, we discuss the differences in detail below.

Chen *et al.* [3] argued that the pretraining procedure of the vanilla masked AE [7] is suboptimal because learning to reconstruct low-level information is not necessarily beneficial for tasks such as classification. Hence, they propose a procedure to reconstruct the high-level representations of the masked tokens instead. The training is performed by maximizing the cosine similarity between teacher and student representations. The teacher is an encoder given by the exponential moving average of past versions of the student encoder. In their case, this training process is called self-distillation because the student learns from aggregated past versions of itself. In our case, self-distillation refers to the fact that the teacher and the student have a shared (identical) encoder. Hence, there is a large difference in terms of the architecture and the training procedure between our model and that of Chen *et al*. [3]. This is also confirmed by the fact that Chen *et al*. [3] does not even cite the work of Zhang *et al*. [22], which introduces the form of self-distillation that inspired our work.

Yang *et al*. [21] modified the vanilla masked AE to learn a spatio-temporal representation. The architecture attaches an additional decoder, which is trained to reconstruct the motion gradients. Unlike Yang *et al*. [21], we do not attempt to reconstruct the motion gradients. Instead, we leverage the motion gradient information to make our model focus on reconstructing tokens which correspond to higher motion. This is necessary to avoid reconstructing the static background scene, which is predominant in anomaly detection data sets.

Aside from the technical differences, another aspect that creates an even higher gap between our method and those of Chen *et al*. [3] and Yang *et al*. [21] is the target task. Indeed, our masked AE is specifically designed for abnormal event detection in video, while the masked AEs proposed in [3, 21] are focused on improving the pretraining procedure.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding? In *Proceedings of ICML*, 2021. 1

[2] Antonio Bărbălău, Radu Tudor Ionescu, Mariana-Iuliana Georgescu, Jacob Dueholm, Bharathkumar Ramachandra, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. SSMTL++: Revisiting Self-Supervised Multi-Task Learning for Video Anomaly Detection. *Computer Vision and Image Understanding*, 229: 103656, 2023. 1

[3] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. SdAE: Self-distillated Masked Autoencoder. In *Proceedings of ECCV*, pages 108–124, 2022. 4, 5

[4] Mariana-Iuliana Georgescu, Antonio Bărbălău, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly Detection in Video via Self-Supervised and Multi-Task Learning. In *Proceedings of CVPR*, pages 12742–12752, 2021. 1

[5] Mariana Iuliana Georgescu, Radu Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A Background-Agnostic Framework with Adversarial Training for Abnormal Event Detection in Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4505–4523, 2022. 1

[6] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton Van Den Hengel. Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection. In *Proceedings of ICCV*, pages 1705–1714, 2019. 1

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of CVPR*, pages 16000–16009, 2022. 4

[8] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-Centric Auto-Encoders and Dummy Anomalies for Abnormal Event Detection in Video. In *Proceedings of CVPR*, pages 7842–7851, 2019. 1

[9] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future Frame Prediction for Anomaly Detection – A New Baseline. In *Proceedings of CVPR*, pages 6536–6545, 2018. 1

[10] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction. In *Proceedings of ICCV*, pages 13588–13597, 2021. 1

[11] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal Event Detection at 150 FPS in MATLAB. In *Proceedings of ICCV*, pages 2720–2727, 2013. 4

[12] Neelu Madan, Nicolae-Catalin Ristea, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised masked convolutional transformer block for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1): 525–542, 2024. 1

[13] Chaewon Park, MyeongAh Cho, Minhyeok Lee, and Sangyoun Lee. FastAno: Fast anomaly detection via spatio-temporal patch transformation. In *Proceedings of WACV*, pages 2249–2259, 2022. 1

[14] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning Memory-guided Normality for Anomaly Detection. In *Proceedings of CVPR*, pages 14372–14381, 2020. 1

[15] Bharathkumar Ramachandra and Michael Jones. Street Scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of WACV*, pages 2569–2578, 2020. 1, 2

[16] Bharathkumar Ramachandra, Michael Jones, and Ranga Vatsavai. Learning a distance function with a Siamese network to localize anomalies in videos. In *Proceedings of WACV*, pages 2598–2607, 2020. 1

[17] Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B. Moeslund, and Mubarak Shah. Self-Supervised Predictive Convolutional Attentive Block for Anomaly Detection. In *Proceedings of CVPR*, pages 13576–13586, 2022. 1

[18] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-World Anomaly Detection in Surveillance Videos. In *Proceedings of CVPR*, pages 6479–6488, 2018. 1

[19] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *Proceedings of ECCV*, pages 494–511, 2022. 1

[20] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing Convolutions to Vision Transformers. In *Proceedings of ICCV*, pages 22–31, 2021. 4

[21] Haosen Yang, Deng Huang, Bin Wen, Jiannan Wu, Hongxun Yao, Yi Jiang, Xiatian Zhu, and Zehuan Yuan. Self-supervised video representation learning with motion-aware masked autoencoders. *arXiv preprint arXiv:2210.04154*, 2022. 4, 5

[22] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-Distillation: Towards Efficient and Compact Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, 2022. 5