Supplementary Material for "Cross-Domain Video Anomaly Detection without Target Domain Adaptation"

Abhishek Aich[⋆], Kuan-Chuan Peng[†], Amit K. Roy-Chowdhury[⋆]
[⋆]University of California, Riverside, USA, [†]Mitsubishi Electric Research Laboratories, USA
{aaich001@, amitrc@ece.}ucr.edu, kpeng@merl.com

1. Additional Details for `zxVAD`

Additional Implementation Details. We implement `zxVAD` in PyTorch [1]. We resize the input frames to 256×256 and normalize them to the range of $[-1,1]$. The generator, the discriminator, and the normalcy classifier are trained with the learning rates of 0.0002, 0.00002, and 0.00002, respectively with the Adam [2] optimizer ($\beta_1 = 0.5, \beta_2 = 0.999$), following [3]. The generator takes 4 frames as input and outputs one frame. We drop the last sigmoid layer of $\mathcal{N}(\cdot)$ as suggested in [4]. We extracted the frames of all TI datasets at 30 frames/sec. The batch size is set as 8. The training iterations for both SHT and UCFC are set as 5000 in all settings of combinations with TI datasets. Unless otherwise specified, we use the default PyTorch parameters. The average training time is ∼2 hours for VAD datasets and ∼24 hours for the experiments involving TI datasets on the Nvidia Titan Xp GPUs.

Table 1: **Augmentation parameters.** `K` denotes `kornia.augmentation`.

Operation	Kornia Parameters
`K.ColorJitter`	0.1,0.1,0.1,0.1
`K.RandomAffine`	$degrees = 360$
`K.RandomPerspective`	$distortion\_scale = 0.2$

Augmentation Parameters. Our relative attention affirmation loss $\mathcal{L}_{\text{RAA}}$ requires augmentation of normal frames $\boldsymbol{v}$ using Kornia [5] to create augmented normal frames $g(\boldsymbol{v})$. We use `kornia.augmentation.AugmentationSequential` to apply these augmentation operations sequentially whose parameters are listed in Tab. 1. `kornia.augmentation.ColorJitter` has four parameter values that represent factors of "*brightness*," "*contrast*," "*saturation*," and "*hue*." All the operations have probability parameter $p = 1.0$.

Location and Size Parameters in Pseudo-Anomaly Synthesis Module. Our untrained CNN based Pseudo-anomaly synthesis module $\boldsymbol{\mathcal{O}}$ creates pseudo-anomalies $\tilde{\boldsymbol{v}}$ by pasting cropped object $\boldsymbol{M_x}$ at random location r_z with random size $r_x \times r_y$. We start by initializing a temporary tensor $\overline{\boldsymbol{v}}$ with $\boldsymbol{v}$. The random location r_z is a rectangular box with coordinates (b_1, b_2, b_3, b_4) [6]. These are computed as $b_1 = b_x - {}^{b_w}/2, b_2 = b_x + {}^{b_w}/2, b_3 = b_y - {}^{b_h}/2$, and $b_4 = b_y + {}^{b_h}/2$, where (b_x, b_y, b_w, b_h) are uniformly sampled as follows. If H and W are height and width of $\boldsymbol{v}$ respectively, then $b_x \sim \text{Unif}(0, W), b_y \sim \text{Unif}(0, H), b_w = W\sqrt{1-\beta}, b_h = H\sqrt{1-\beta}$. Here, $b_2 > b_1$ and $b_4 > b_3$. We then resize $\boldsymbol{M_x}$ and $\boldsymbol{M}$ to size $(b_2 - b_1) \times (b_4 - b_3)$. Finally, only the pixels corresponding to regions where $\boldsymbol{M}_{(i,j)} = 1$ are replaced in $\overline{\boldsymbol{v}}$ to create anomaly frame $\tilde{\boldsymbol{v}}$. To handle boundary conditions where $0 \leq b_x, b_w \leq W$ and $0 \leq b_x, b_w \leq H$, we clip the values to be in the range of $[0, W]$ and $[0, H]$, respectively. Here, $\beta \sim \text{Unif}(0, 1)$.

Evaluation criteria. For anomaly scores, we follow [3, 7] and compute Peak Signal to Noise Ratio (PSNR) [8] scores per frame and normalize PSNR of all frames in each testing video to the range [0, 1] in order to compare with ground-truth binary labels. Note that we observed such normalization practice (adopted from [3]) impacts anomaly scores.

2. Additional Results on `zxVAD`

Impact of the amount of TI Data. We analyzed the impact of the amount of TI data on our `zxVAD` framework in extreme settings. Particularly, we evaluated `zxVAD` when the amount of videos of TI datasets (HMDB and UCF101) is close to the number of training videos available in the VAD datasets. With 0.5%, 1%, 2%, 4%, and 8% of HMDB data, we observed an average cross-domain AUC performance of 74.99% on Ped1, 93.82% on Ped2, and 79.49% on Ave. A similar observation was made on UCF101 (0.0625%, 0.125%, 0.315%, 0.63%, and 1.25% of data resulted in average cross-domain AUC performance of 74.61% on Ped1, 94.17% on Ped2, and 79.46% on Ave). This demonstrates that almost SOTA cross-domain performance on the current VAD datasets is achievable even with an extremely low amount of TI data.

Relevancy among VAD data. We followed [9, 10] for the relevancy analysis between the TI to target domain (Ave, Ped1/2) VAD data. We observed higher relevancy scores among SHT (to Ave: 0.241, to Ped1/2: 0.250) and UCFC (to Ave: 0.201, to Ped1/2: 0.167) compared to average TI (to Ave: 0.186, to Ped1/2: 0.138). This confirms: TI data is indeed less relevant to VAD data.

More results on the impact of randomly initialized networks for Pseudo-Anomaly Synthesis. We analyzed the impact of the randomly initialized network $\mathcal{R}(\cdot)$ on our untrained CNN based pseudo-anomaly synthesis module. In Fig. 1, it can be observed that our `zxVAD` method outperforms the state-of-the-art (SOTA) xVAD works on the Ped1 and Ped2 datasets in the zero-shot settings when the source is SHT irrespective of kind of randomly initialized network $\mathcal{R}(\cdot)$ employed to extract objects from all our TI datasets.

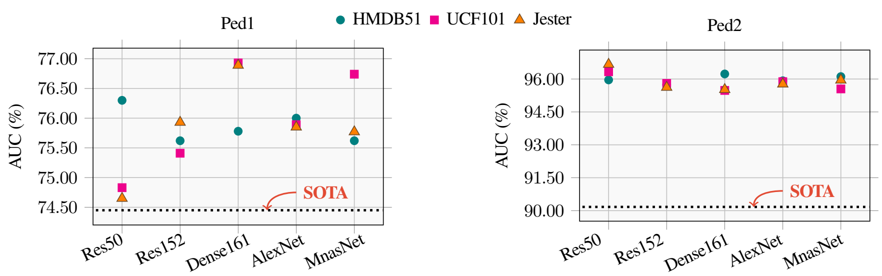


Figure 1: **Impact of $\mathcal{R}(\cdot)$ in `zxVAD`.** The source is SHT.

More results on same-dataset testing. We beat our baselines in the same-dataset testing in all VAD and TI combination scenarios as shown in Tab. 2. We also compare with more state-of-the-art unsupervised VAD methods under the same-dataset setting in Tab. 3.

Table 2: **Same-dataset testing on the $\mathbf{SHT_{dc}}$ dataset.** We beat our baselines in all the source domain data settings.

VAD Training data	Input to $\mathcal{O}$	Method	AUC (%) on $\mathrm{SHT_{dc}}$
$\mathrm{SHT_{dc}}$	N/A	rGAN [11] (paper)	70.11
$\mathrm{SHT_{dc}}$	N/A	MPN [7] (code)	67.47
$\mathrm{SHT_{dc}}$	$\mathrm{SHT_{dc}}$	`zxVAD` **(ours)**	70.73
$\mathrm{SHT_{dc}}$	HMDB	`zxVAD` **(ours)**	**70.85**
$\mathrm{SHT_{dc}}$	UCF101	`zxVAD` **(ours)**	70.80
$\mathrm{SHT_{dc}}$	Jester	`zxVAD` **(ours)**	70.50

Table 3: **Additional same dataset testing comparison.** The best and second best AUC are marked in **bold** and <u>underline</u>, respectively.

Methods	Ped2	Ave	SHT
MPPCA [12]	69.3	-	-
MPPC+SFA [12]	61.3	-	-
MDT [13]	82.9	-	-
ConvAE [14]	85.0	80.0	60.9
TSC [15]	91.0	80.6	67.9
StackRNN [15]	92.2	81.7	68.0
MT-FRCN [16]	92.2	-	-
Unmasking [17]	82.2	80.6	-
Frame-Pred [3]	95.4	85.1	72.8
AMC [18]	96.2	86.9	-
MemAE [19]	94.1	83.3	71.2
SDOR [20]	83.2	-	-
rGAN [11]	96.2	85.8	**77.9**
LMN [21]	**97.0**	<u>88.5</u>	70.5
MPN [7]	96.9	**89.5**	<u>73.8</u>
`zxVAD`	<u>96.95</u>	83.8	71.6

Ablation analysis. `zxVAD` is not too sensitive to the loss ratios and Table (on *right*) validates this point. For our backbone GAN, we use exact same ratios as suggested in [3]. For the proposed normalcy classifier, we *do not* use ratios for our losses $\mathcal{L}_{\mathrm{AA}}$ and $\mathcal{L}_{\mathrm{RAA}}$ (*i.e.* set as 1). Finally, the effect of ratios α_{n} on $\mathcal{L}_{\mathrm{N}}$ and α_{rn} on $\mathcal{L}_{\mathrm{RN}}$ is shown. All cases show better AUC than SOTA MPN [7].

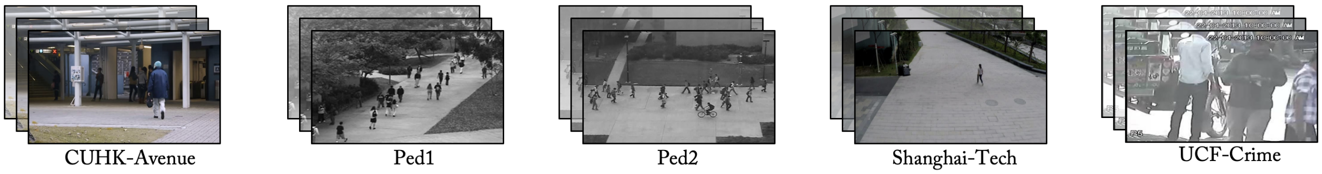
Ratios	$\mathbf{SHT}_{dc}$
MPN [7]	67.47
$(\alpha_{\mathrm{n}},\alpha_{\mathrm{rn}})=(1,0.01)$	**70.85**
$(\alpha_{\mathrm{n}},\alpha_{\mathrm{rn}})=(1,0.1)$	69.49
$(\alpha_{\mathrm{n}},\alpha_{\mathrm{rn}})=(0.1,0.1)$	69.95
$(\alpha_{\mathrm{n}},\alpha_{\mathrm{rn}})=(0.01,0.01)$	70.37

3. Examples from Datasets

We provide some video examples of the VAD datasets (SHT, UCFC, Ped1, Ped2, and Ave in Fig.2(a)) and TI datasets (HMDB, UCF101, and Jester in Fig.2(b)) listed in `Tab.2` of the main manuscript.

4. More Qualitative Results

We show additional examples of pseudo-abnormal frames created using our pseudo-anomaly module in Fig. 3 and difference maps from three different datasets indicating anomalies in Fig. 4.



(a) VAD datasets

(b) TI datasets

Figure 2: **Examples from VAD and TI datasets.** We visualize some examples of videos used for experiments in our paper.

Figure 3: **Pseudo-abnormal frames.** We present examples of pseudo-abnormal frames generated using our proposed untrained CNN based pseudo-anomaly synthesis module.

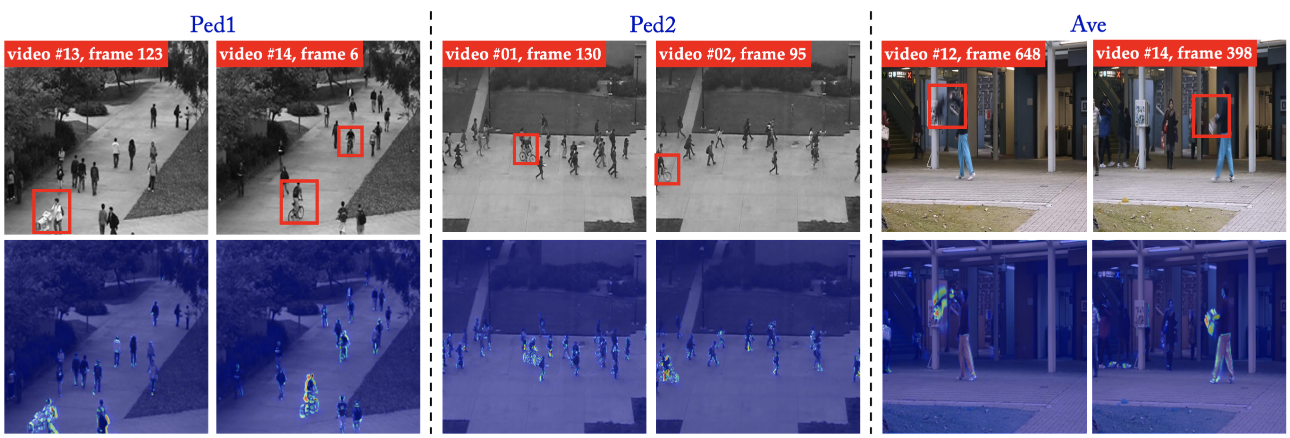


Figure 4: **Difference maps.** We show more examples of difference maps obtained from `zxVAD` (source: SHT). The lighter colors in difference map mean larger prediction error indicating anomalies. Red boxes indicate ground truth anomalies. Best viewed in color.

References

[1] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[3] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018.

[4] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

[5] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020.

[6] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[7] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15425–15434, 2021.

[8] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[9] Kuan-Chuan Peng, Ziyan Wu, and Jan Ernst. Zero-shot deep domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 764–781, 2018.

[10] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot object recognition by semantic manifold distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2635–2644, 2015.

[11] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive anomaly detection. In *European Conference on Computer Vision*, pages 125–141. Springer, 2020.

[12] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *2009 IEEE conference on computer vision and pattern recognition*, pages 2921–2928. IEEE, 2009.

[13] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981. IEEE, 2010.

[14] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.

[15] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017.

[16] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3619–3627, 2017.

[17] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu. Unmasking the abnormal events in video. In *Proceedings of the IEEE international conference on computer vision*, pages 2895–2903, 2017.

[18] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1273–1283, 2019.

[19] 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 the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019.

[20] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12173–12182, 2020.

[21] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14372–14381, 2020.