Video Anomaly Detection via Spatio-Temporal Pseudo-Anomaly Generation : A Unified Approach

Supplementary Material

1. Datasets

Ped2 [6] dataset comprises of 16 training and 12 test videos and all videos have the same scene in the background. The videos with normal events consist of pedestrians only, whereas the videos with anomalous events include bikes, skateboards and carts apart from pedestrians.

Avenue [8] dataset comprises of 16 training and 21 test videos with every video having the same background scene. Normal events involve people routinely walking around while the abnormal instances include abnormal objects such as bikes and abnormal human actions such as unusual walking directions, running around or throwing things.

ShanghaiTech [9] dataset includes 330 training and 107 test videos recorded at 13 different background locations with complex lightning conditions and camera angles, making it the one of the largest one-class anomaly detection datasets. The test split captures a total of 130 anomalous events including running, riding a bicycle and fighting.

UBnormal [1] is a synthetic dataset with multi-scene backgrounds and a diverse set of anomalies. The dataset consists of training, validation and test split with both normal and abnormal events. The normal events include walking, talking on the phone, walking while texting, standing, sitting, yelling and talking with others. It is to be noted that abnormal events in each of the train, validation and test split are different to each other. The train split includes abnormal events like *falling, dancing, walking injured, running injured, crawling, and stumbling walk*. The validation split comprises *fighting, sleeping, dancing, stealing, and rotating 360 degrees*. All the evaluations are conducted on the validation set.

UBnormal data-split under OCC Setting. In order to use this dataset in the one class classification (OCC) setting, we train our model using only the normal 186 videos in the training split and the pseudo-anomalies (PAs) generated using them (i.e. totally ignoring the abnormal samples provided in the train set). We tested our model on all the videos in the validation split, comprising of 64 videos with both normal and abnormal events. Such a setting was chosen to keep consistency in evaluation as with other datasets under the OCC setting. The frame-level groundtruth annotation for validation set of UBnormal [1] was created using the script[1] provided by the authors.

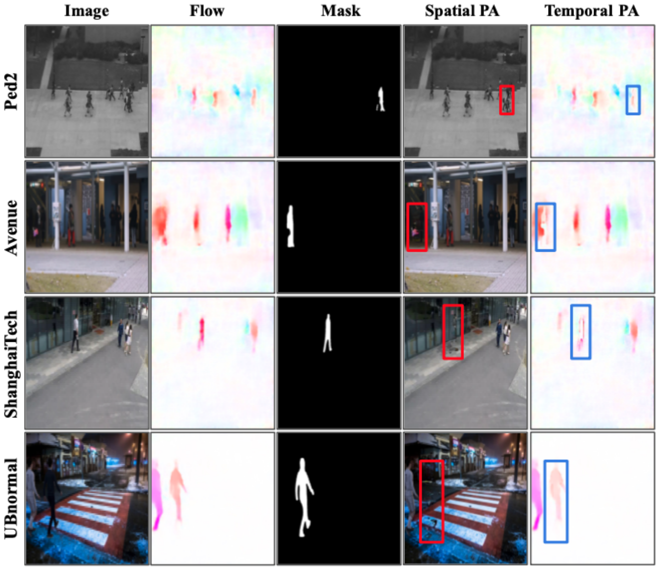


Figure 1. Qualitative Assessment : Visualisation of spatial and temporal PAs for all 4 datasets. Here we only show segmentation masks however the approach also works with random masks.

2. Additional Details and Insights

1: Pseudo-Anomaly Construction. We take an off-the-shelf Latent Diffusion Model [13] (LDM[2]) pre-trained on the Places dataset [16]. We do not perform any finetuning of the LDM on any video anomaly dataset and therefore it is "under-trained" on video data and hence capable of spatially distorting them. For inpainting the masked out regions of the images, 50 steps of inference were carried out. It is to be noted that due to lack of computational resources we did not experiment with other values of timesteps or any end-to-end finetuning. A very low number of timesteps may produce mostly noisy inpainting output while a very high value might result in inpainted images very close to the input image. The strategy for generation of random and segmentation masks was adopted from the code[3] provided by the authors of LAMA [14]. If segmentation mask was not detected for a frame, a random mask was selected instead. Figure 1 depicts more examples of generated PAs.

2: Extracting ViFi-CLIP Features. For the training split of the benchmark datasets and their corresponding spatial pseudo-anomalies, we extract frame level features using the

[1] https://github.com/lilygeorgescu/UBnormal/tree/main/scripts

[2] https://github.com/CompVis/latent-diffusion/tree/main

[3] https://github.com/advimman/lama/tree/main/saicinpainting

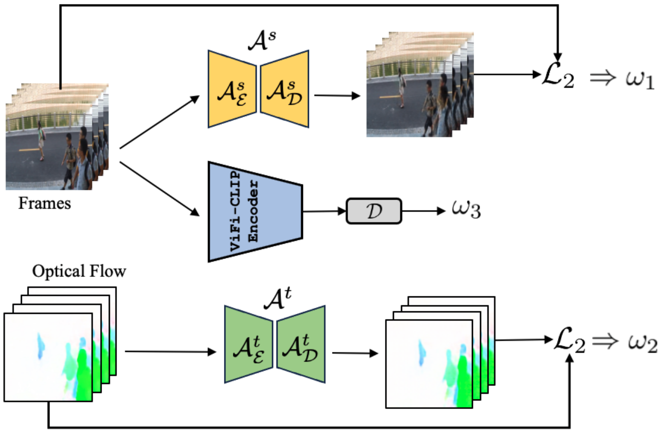

$\mathcal{A}^s$
$\mathcal{A}^s_{\mathcal{E}}$
$\mathcal{A}^s_{\mathcal{D}}$
$\mathcal{L}_2 \Rightarrow \omega_1$
Frames
ViFi-CLIP Encoder
$\mathcal{D}$
ω_3
Optical Flow
$\mathcal{A}^t$
$\mathcal{A}^t_{\mathcal{E}}$
$\mathcal{A}^t_{\mathcal{D}}$
$\mathcal{L}_2 \Rightarrow \omega_2$

Figure 2. During inference, aggregate anomaly score is computed by calculating the weighted sum (eq 4) of all the three types of anomaly information; reconstruction quality ω_1 (eq 2), temporal irregularity ω_2 (eq 3) and semantic inconsistency ω_3.

ViFi-CLIP [12] model. The input to the ViFi-CLIP model has size : $\mathcal{B}' \times \mathcal{T}' \times 3 \times 224 \times 224$, where $\mathcal{B}'$ (batch size) was set to 1 and $\mathcal{T}'$ (# of frames) was set to 16. All frames were passed into ViFi-CLIP in a sliding window fashion with a stride of 16 therefore we obtain a 512-dimensional feature for every frame. ViFi-CLIP uses the backbone of ViT-B/16 [4] and is pre-trained on Kinetics-400 [5]. It is to be noted that the ViFi-CLIP model performs temporal pooling of the CLIP [11] features, however we do not perform temporal pooling and use the frame level representations as during inference we evaluate our pipeline using frame level micro AUC scores. For the frames of the videos in test split (Ped2, Avenue, ShanghaiTech) and validation split (UBnormal), we follow the same procedure for feature extraction.

3: Effect of changing the probability of sampling PAs. We conduct an experimental study by varying the probability of sampling spatial and temporal PAs (p_s, p_t) on Ped2 during training between 0.1 to 0.5 and measuring micro AUC scores during inference. Figure 4 shows that the model achieves best performance when $p_s = 0.4$ and $p_t = 0.5$.

3. Evaluation Criteria

To measure the reconstruction quality, we follow the recent works of [3, 7, 10], which utilised normalized Peak Signal to Noise Ratio (PSNR) P_t between an input frame and its reconstruction to calculate the anomaly score. This is illustrated in the following equation.

$$P_t = 10 \log_{10} \frac{M^2_{\hat{\mathbf{x}}_\mathbf{t}}}{\frac{1}{R}||\hat{\mathbf{x}}_\mathbf{t} - \mathbf{x}_\mathbf{t}||^2_2}, \tag{1}$$

$$\omega_1^{(t)} = 1 - \frac{P_t - \min_t(P_t)}{\max_t(P_t) - \min_t(P_t)}, \tag{2}$$

where $\mathbf{x}_\mathbf{t}$ is the input frame at time t, $\hat{\mathbf{x}}_\mathbf{t}$ represents reconstruction of $\mathbf{x}_\mathbf{t}$, R denotes the total number of pixels in $\hat{\mathbf{x}}_\mathbf{t}$ and $M_{\hat{\mathbf{x}}_\mathbf{t}}$ is the maximum possible pixel value of $\hat{\mathbf{x}}_\mathbf{t}$. The anomaly score $\omega_1^{(t)}$ is an indicator of reconstruction quality of the input frame. For measuring the temporal irregularity, we compute the normalised $\mathcal{L}_2$ loss between input optical flow at time t and its reconstruction given by the equation:

$$\omega_2^{(t)} = \frac{1}{R'}||\hat{\phi}(\mathbf{x}_\mathbf{t}, \mathbf{x}_{(\mathbf{t+1})}) - \phi(\mathbf{x}_\mathbf{t}, \mathbf{x}_{(\mathbf{t+1})})||^2_2, \tag{3}$$

where $\phi(\mathbf{x}_\mathbf{t}, \mathbf{x}_{(\mathbf{t+1})})$ is the input optical flow frame calculated using consecutive frames $\mathbf{x}_\mathbf{t}$ and $\mathbf{x}_{(\mathbf{t+1})}$, $\hat{\phi}(\mathbf{x}_\mathbf{t}, \mathbf{x}_{(\mathbf{t+1})})$ represents the reconstruction of $\phi(\mathbf{x}_\mathbf{t}, \mathbf{x}_{(\mathbf{t+1})})$, R' denotes the total number of pixels in $\hat{\phi}(\mathbf{x}_\mathbf{t}, \mathbf{x}_{(\mathbf{t+1})})$. To measure the semantic inconsistency, the input frames sequence is fed into $\mathcal{D}$ in a sliding window fashion with a window size of 16. The output probability ($\omega_3^{(t)}$) of a frame at time t to be anomalous is computed using its ViFi-CLIP feature representation.

A higher value of $\omega_1^{(t)}$, $\omega_2^{(t)}$ and $\omega_3^{(t)}$ represents higher reconstruction error for frame and optical flow and high anomaly probability at time t in the test videos during inference. Alternatively, they are indicators of poor reconstruction quality, temporal irregularity and semantic inconsistency and their aggregation can aid in determining real-world anomalies. The aggregate anomaly score is given by the following equation :

$$\omega_{agg}^{(t)} = \begin{cases} \eta_1\omega_1^{(t)} + \eta_2\omega_2^{(t)} + \eta_3\omega_3^{(t)}, & \text{w/ } \mathcal{D} \\ \eta_1\omega_1^{(t)} + \eta_2\omega_2^{(t)}, & \text{w/o } \mathcal{D}; (\eta_3 = 0) \end{cases} \tag{4}$$

where η_1, η_2, η_3 are weights assigned to $\omega_1^{(t)}$, $\omega_2^{(t)}$ and $\omega_3^{(t)}$ respectively. The values of η_1, η_2 and η_3 lies in the interval $[0, 1]$ and their sum is equal to 1. We manually tune the values of η_1, η_2, η_3 for all the datasets. The values of (η_1, η_2, η_3) for all the datasets are given by - Ped2 (0.65,0.25,0.1), Avenue (0.45,0.5,0.05), Shanghai (0.85, 0.13, 0.02) and UBnormal (0.4, 0.5, 0.1). In all of the cases, any of the three component can be excluded during evaluation by setting the corresponding weight (η_1, η_2, η_3) to zero.

Note : We also experimented with the learnt weights for the three anomaly indicators but there was a marginal decrease in the performance compared to manually tuning their weights.

Evaluation Metric. For evaluation, we follow the standard metric of frame-level area under the ROC curve (micro-AUC) as in [15]. We obtain the ROC curve by varying the anomaly score thresholds to plot False Positive Rate and True Positive Rate for the whole test set for a given dataset. Higher AUC values indicate better performance and more accurate detection of anomalies.

Table 1. Discriminator ($\mathcal{D}$) architecture details

Layers	(Input size, Output size)
Linear Layer 1	(512,128)
ReLU	-
Linear Layer 2	(128,1)

Table 2. Autoencoder ($\mathcal{A}^s$ and $\mathcal{A}^t$) architecture details

	Layer	Input Channels	Output Channels	Filter Size	Stride	Padding	Negative Slope
Encoder	Conv3D	3	96	(3,3,3)	(1,2,2)	(1,1,1)	-
	BatchNorm3D	-	-	-	-	-	-
	LeakyReLU	-	-	-	-	-	0.2
	Conv3D	96	128	(3,3,3)	(2,2,2)	(1,1,1)	-
	BatchNorm3D	-	-	-	-	-	-
	LeakyReLU	-	-	-	-	-	0.2
	Conv3D	128	256	(3,3,3)	(2,2,2)	(1,1,1)	-
	BatchNorm3D	-	-	-	-	-	-
	LeakyReLU	-	-	-	-	-	0.2
	Conv3D	256	256	(3,3,3)	(2,2,2)	(1,1,1)	-
	BatchNorm3D	-	-	-	-	-	-
	LeakyReLU	-	-	-	-	-	0.2
Decoder	ConvTranspose3D	256	256	(3,3,3)	(2,2,2)	(1,1,1)	-
	BatchNorm3D	-	-	-	-	-	-
	LeakyReLU	-	-	-	-	-	0.2
	ConvTranspose3D	256	128	(3,3,3)	(2,2,2)	(1,1,1)	-
	BatchNorm3D	-	-	-	-	-	-
	LeakyReLU	-	-	-	-	-	0.2
	ConvTranspose3D	128	96	(3,3,3)	(2,2,2)	(1,1,1)	-
	BatchNorm3D	-	-	-	-	-	-
	LeakyReLU	-	-	-	-	-	0.2
	ConvTranspose3D	96	3	(3,3,3)	(1,2,2)	(1,1,1)	-
	Tanh	-	-	-	-	-	-

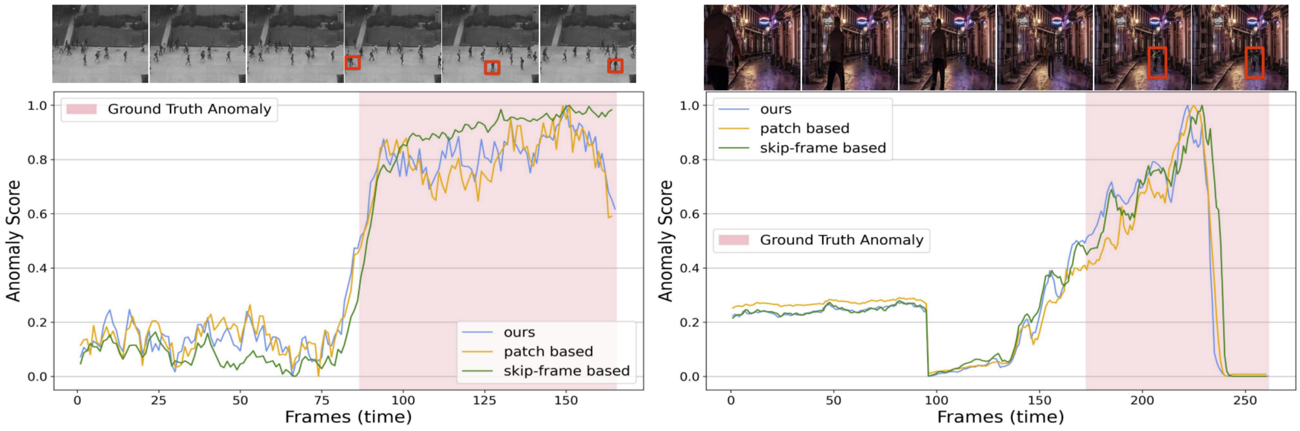
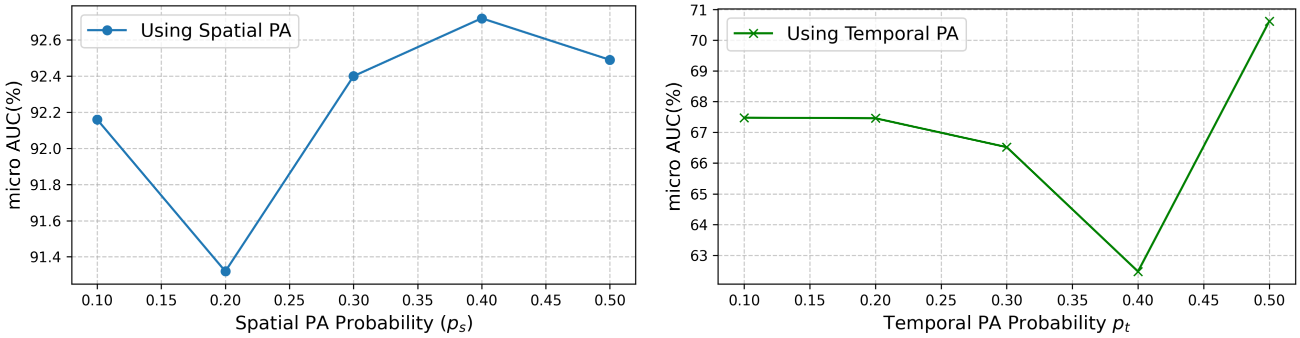


Figure 3. Qualitative Assessment : Visualization of anomaly score over time for sample videos in Ped2 (left) and UBnormal (right). Compared with other PAs generator and reconstruction based methods in LNTRA [2] - patch and skip-frame based.



(a) (b)

Figure 4. Comparison of micro-AUC scores on Ped2 dataset calculated from output of $\mathcal{A}^s$ ($\mathcal{A}^t$) trained on a range of values of p_s (p_t) between $\{0.1,0.5\}$. We observe that setting $p_s = 0.4$ and $p_t = 0.5$ yields the best performance as shown in (a) and (b) respectively. These probability values are fixed for all other experiments.

References

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20143–20153, 2022. 1

[2] Marcella Astrid, Muhammad Zaigham Zaheer, Jae-Yeong Lee, and Seung-Ik Lee. Learning not to reconstruct anomalies. In *BMVC*, 2021. 3

[3] Fei Dong, Yu Zhang, and Xiushan Nie. Dual discriminator generative adversarial network for video anomaly detection. *IEEE Access*, 8:88170–88176, 2020. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[5] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[6] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32, 2014. 1

[7] 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. 2

[8] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *2013 IEEE International Conference on Computer Vision*, pages 2720–2727, 2013. 1

[9] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. *ICCV, Oct*, 1(2):3, 2017. 1

[10] 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. 2

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[12] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554, 2023. 2

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[14] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 1

[15] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14193, 2020. 2

[16] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1