2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Video Anomaly Latent Training GAN (VALT GAN): Enhancing Anomaly Detection through Latent Space Mining

Anikeit Sethi[*], Krishanu Saini[*], Rituraj Singh[*], Sumeet Saurav[†], Aruna Tiwari[*], Sanjay Singh[†], Vikas Chauhan[‡]

[*]Computer Science and Engineering, Indian Institute of Technology Indore, Indore, India
[†]Intelligent System Groups, CSIR-CEERI, Pilani, India
[‡]Electrical and Computer Science, National Taipei University of Technology, Taipei City 106, Taiwan

***Abstract*—Anomaly detection in video data plays a crucial role in numerous applications, such as industrial monitoring and automated surveillance. This paper presents a novel method for video anomaly detection (VAD) using Generative Adversarial Networks (GANs). The proposed method called VALT-GAN combines two separate branches, one for spatial information and the other for temporal information, to capture relevant features from video data. The framework is utilized to learn the normal features from the training video dataset, enabling the generator to produce realistic samples. However, existing GAN-based methods face challenges in detecting subtle or unseen anomalies. To address this, we introduce latent mining for adversarial training which allows us to train a robust GAN model with high anomaly detection (AD) capability. We exploit the latent space following the continuous nature of the generator using the Iterative Fast Gradient Signed Method (IFGSM) which improves the quality of the generated images. Experimental evaluations show the effectiveness of VALT-GAN as compared to traditional methods on UCSD (University of California, San Diego) Peds2, CUHK (Chinese University of Hong Kong) Avenue, and ShanghaiTech datasets.**

***Index Terms*—Generative Adversarial Networks, Latent Space Mining, Anomaly Detection**

I. Introduction

Security cameras are extensively deployed in numerous public locations and the need for automated anomaly detection in surveillance videos has grown. Identifying unusual incidents like accidents or crimes manually is difficult and time-consuming, especially in long video sequences. Thus, there is a demand for methods that can detect potential anomalies at the frame level. Several approaches are proposed to address the challenge of video anomaly detection (VAD). These include analyzing entire video frames [1]–[4] or smaller image patches within frames [5]–[9]. Recent advancements in deep learning and computer vision have revolutionized this field, enabling more sophisticated and data-driven methods. VAD can be categorized into unsupervised and weakly-supervised techniques. Unsupervised VAD assumes only videos with normal activities are available for training. One promising approach is using Generative Adversarial Networks (GANs) for VAD, leveraging their ability to generate realistic data. Another common approach is learning a latent representation or code that captures the essence of normal behavior and reconstructs input data, aiming to create a structured manifold where normal events lie closely [10], [11].

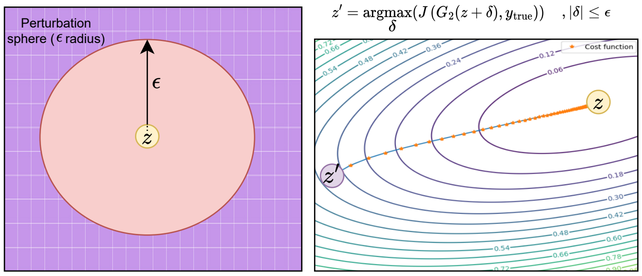

Fig. 1. Selecting adversarial samples for robust training of the model.

Within the structured manifold representation, anomalies are identified as deviations from the established structure. This manifold is often visualized as an epsilon space or circle, where normal events cluster tightly within a small epsilon neighborhood or circle. Perturbed [12], [13] examples generated through adversarial attacks are also considered, as they challenge the model's robustness as shown in Fig. 1. Anomalies, including both naturally occurring anomalies and perturbed instances, manifest as points or instances lying outside this epsilon space, indicating a deviation from the expected normal behavior. By training the model using latent mining [10], we ensure that it accurately predicts both legitimate and perturbed examples in the same direction, promoting robustness against adversarial attacks.

Our proposed VALT-GAN framework improves VAD by combining a unified two-stream model, which leverages both spatial and temporal information, with a novel training strategy. The networks are trained on regular samples to learn the features of normal video frames and accurately predict the next video frame. The unified model incorporates convolutional layers for spatial features and temporal segment networks for motion information. By iteratively updating latent variables using z optimization, we enhance the quality of generated sam-

The authors would like to thank the Ministry of Electronics and Information Technology (MeiTY) to grant this project titled "Resource Constrained Artificial Intelligence" with grant number: 4(16)/2019-ITEA.

978-0-7381-4408-5/23/$31.00 ©2023 IEEE

ples. Our integration greatly enhances the ability of model to differentiate between normal and anomalous events, improving accuracy and robustness in VAD. The main contributions of our work are summarized below:

1) We propose a VALT-GAN comprised of a dual-stream generator model, considering both spatial and temporal features simultaneously.
2) By utilizing latent representations and epsilon space, our approach incorporates the intrinsic structure and distribution of normal events enabling the improvement of the mapping of latent variables to generate more realistic and meaningful video frames.
3) The proposed VALT-GAN outperforms recent SOTA models in terms of Area Under Curve (AUC) in UCSD Peds2, CUHK Avenue, and ShanghaiTech datasets.

Through extensive experiments and evaluations, we demonstrate and validate the efficacy and resilience of our proposed method to detect anomalies in surveillance videos. The remaining sections of this paper are structured as follows: Section II delves into the relevant literature review, while Section III elaborates on our proposed methodology. We present the experimental results of our approach in Section IV, and finally, Section V encapsulates the conclusion of our study.

II. Related Work

In this section, we provide related work of anomaly detection in video. The two main approaches to tackle video anomaly detection is using reconstruction and prediction-based methods.

Reconstruction-based methods. The reconstruction-based method works on the principle of learning the features from the video frame to use it for reconstruction. Nguyen et al. [14] proposed a deep Convolutional Neural Network (CNN) for anomaly detection by employing auto-encoder principles and sparse combination learning. Abati et al. [15] utilize a deep autoencoder combined with a parametric density estimator to acquire the probability distribution of latent representations. The utilization of clustering to determine the distance of features extracted by the encoder enhance the quality of features for reconstruction [16]. Park et al. [17] proposed a memory module to learn the most common normal features for reconstruction. Further, the memory module is utilized to extract the most similar features of normal video frames for the reconstruction [18]. Cho et al. [19] proposed the use of two encoders one for learning the spatial features and the other for learning the temporal features which are then used by the decoder to reconstruct the video frame. Further, the dual network is utilized to learn the spatial and temporal features for reconstruction and image-to-image translation [20]. In reconstruction-based methods, anomalies are detected by comparing the reconstruction quality or error, while prediction-based methods compare the predicted future frame with the ground-truth frame to identify anomalies across multiple frames. These deep learning-based approaches offer the potential for video representation using unsupervised learning and have gained attention in the field of AD [21].

Prediction-based methods. The prediction-based method works on the principle of learning the features from the video frame to use it for producing the future video frame. Yu et al. [22] proposed a method that takes a sequence of video frames to predict the future video frame. Park et al. [17] utilizes the memory module to construct the future video frame from the sequence of video frames as input. Tang et al. [23] proposed to utilize both reconstruction and future frame prediction methods to enhance the quality of video frames constructed and improve anomaly detection. The multi-path anomaly detection framework is proposed by Wang et al. [24] to capture both spatial and temporal dependencies. Cai et al. [25] proposed to use the temporal constraints to enhance anomaly detection in the video, while Cross U-net [21] is another prediction-based method used for anomaly detection in videos.

III. VALT-GAN

This section introduces the VALT-GAN method, designed for future frame prediction using latent mining. The method comprises a Generator (G) and Discriminator (D), as illustrated in Figure 2.

A. Latent Adversarial Mining for Quality

Latent adversarial mining [10] plays a crucial role in improving the quality of the generator in a GAN by:

1) Discovering informative latent codes: Latent adversarial mining involves searching for latent codes that fool the discriminator into classifying generated samples as real. This process helps identify latent representations that capture important features and characteristics, leading to higher-quality generated samples.
2) Promoting diverse and realistic output: By encouraging the generator to explore a broader range of latent codes through adversarial mining, it helps prevent mode collapse and promotes the generation of diverse and realistic samples. This diversity enhances the quality of the generator's output by capturing a more comprehensive representation of the target distribution.
3) Enhancing training stability: Latent adversarial mining can provide additional regularization during GAN training. By iteratively refining the latent codes through the interplay between the generator and discriminator, it helps stabilize the training process and improve the overall quality and convergence of the generator.

The process of latent space mining requires the selection of a latent sample in a small perturbation region that maximizes the loss J as shown below:

$$\min_{\theta}\left[E_{(x,y)\in D}\left[\max_{\delta\in\psi} J\left(G_2(G_1(x)+\delta);\theta\right), y_{\text{true}}\ )\right]\right] \tag{1}$$

where θ represents the model parameter and ψ is the ϵ perturbation circle. In order to find the optimized δ value, we utilize IFGSM [10] technique shown below.

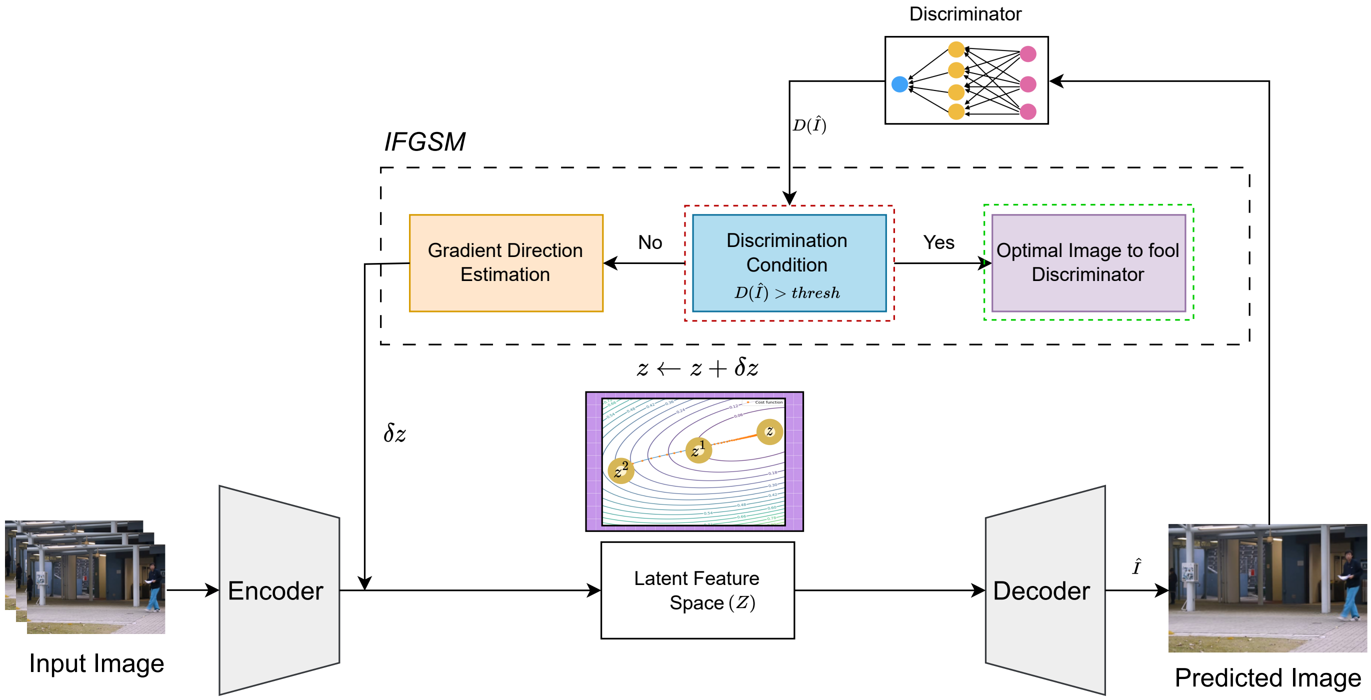

Fig. 2. Overall architecture of proposed VALT-GAN consisting of encoder-decoder network with latent mining to generate future video frame from Generator. The latent codes are iteratively updated (z, z^1, z^2) through IFGSM. This process improves the adversarial training by allowing the Generator to generate a realistic frame resembling the Discriminator's true distribution.

$$x^{(t+1)} = \text{clip}_{x,\epsilon}\left(x^{(t)} + \alpha \cdot \text{sign}\left(\nabla_x J\left(x, y_{\text{true}}\right)\right)\right) \quad (2)$$

where, $x^{(t)}$ represents the updated input at iteration t, α is the step size of the learning rate set to ϵ/t, $J(x, y_{true})$ is the loss function with respect to input X and target y_{target}. $\nabla_x J(x, y_{true})$ denotes the gradient of the loss function with respect to the input x, $sign(.)$ represents the sign function, and $clip_{x,\epsilon}(.)$ restricts the values of the input within the range defined by x and ϵ. Here, J is selected as the adversarial loss of generated image.

B. Model Architecture

The AE architecture used by VALT-GAN consists of a decoder (G_2) and an encoder (G_1). G_1 is a two-stream network that captures both spatial and temporal aspects of the input video frames, while G_2 predicts the next frame using the data retrieved from G_1. The generator's model architecture is shown in Table I.

The proposed VALT GAN for VAD incorporates a Wide Residual Network (WRN) [26], [27] architecture with a latent sample mining method to enhance the model's robustness. The WRN, with a widening factor (k) of 4, serves as the foundation for capturing spatial details and understanding the visual structure of the input sequence. The network can learn more complex and representative features because of its wider structure, which enhances detection capabilities.

A temporal shift module [28] is also introduced as part of the dual-branch approach. The last feature map obtained from the WRN is passed through two streams to understand spatial and motion information. Firstly, the temporal branch applies a shift to a subset of features over multiple frames. This technique enables the model to capture temporal features and dependencies across multiple input frames, enhancing its understanding of dynamic patterns and improving performance in handling sequential data. To maintain spatial information, features derived from each frame are concatenated in the spatial branch.

After the encoder stage, a gradient-based perturbation technique is applied in the latent space using AdvLatGAN [13] to enhance the quality of the generated samples. By leveraging the gradients of the generator network, the latent representation is iteratively updated to optimize the desired quality criteria which depends on discriminator score threshold or fixed number of iterations. This iterative process utilizes objective functions related to quality improvements, such as adversarial loss or pixel-wise difference. The generator produces improved quality and fidelity samples by fine-tuning the latent representation using these gradients as shown in Equation 3.

$$z^* = \arg\min_{z} \; E_{|z-z^0|\leq\epsilon}\left[\log\left(1 - J(G_2(z))\right]\right., \quad z = G_1(x) \quad (3)$$

Where z^0 is the latent space generated by the encoder. The optimal z^* is then passed to the decoder which restores the predicted frame's spatial details and resolution. The decoder consists of multiple layers, each comprising a sequence of blocks. Each layer uses deconvolution to upsample the features

Part	Layer	Output Shape	Details
Encoder (G_1)	Input	Window of $[H, W, 3]$	Input video frame
	WRN	$Z = [H/8, W/8, 4096]$	Expansion factor(k) = 4
	Spatial (F) Conv1 $[H/8 \times W/8 \times 2048]$		Temporal (F') Feature Shift $[H/8 \times W/8 \times 2048]$
Bottleneck	Sum	$[H/8 \times W/8 \times 2048]$	$F \oplus F'$
Decoder (G_2)	DeConv1	$[H/4, W/4, 1024]$	2×2, strides 2, BN, LeakyReLU
	DeConv2	$[H/2, W/2, 512]$	2×2, strides 2, BN, LeakyReLU
	DeConv3	$[H, W, 256]$	2×2, strides 2, BN, LeakyReLU
	Conv2	$[H, W, 3]$	1×1, strides 1, BN, LeakyReLU

TABLE I
GENERATOR NETWORK ARCHITECTURE OF PROPOSED VALT-GAN. THE H DENOTES THE HEIGHT AND W DENOTES THE WIDTH OF THE VIDEO FRAME WHILE WRN DENOTES THE WIDERRESNET BACKBONE

and recover fine-grained information lost during encoding. Batch normalization is used to normalize intermediate features, promoting training stability. The Rectified Linear Unit (ReLU) activation function introduces non-linearity and enhances representation capabilities. The methodology progressively reconstructs the spatial details by combining deconvolution, batch normalization, and ReLU activation in the decoder layers, resulting in visually realistic and high-resolution outputs. By conducting the latent transform during training, the methodology compensates for the challenge of aligning the pace between generators and discriminators. This is crucial because training the generator is inherently more difficult than training the discriminator, as the generator struggles to map a continuous distribution to a disconnected one with multiple modes.

The discriminator uses a down-sampling convolution architecture and penalizes structure at the patch level. Isola et al. [29] demonstrates that each $N \times N$ patch is evaluated to determine authenticity, assuming independence between pixel blocks. The discriminator loss drives adversarial learning.

C. Loss function for VALT

The adversarial loss in a GAN can be expressed using the latent space, which is the space where the encoder produces a latent space that is used by the decoder to generate the corresponding samples. Equation 4 shows the adversarial loss.

$$L_{adv}(D,G) = E_{p_{\text{real}}}\left[\log(D(x))\right] + E_{p_{\text{fake}}}[\log(1 - D(G(x)))] \tag{4}$$

Here, x represents real samples, $G(x)$ represents generated samples, $D(x)$ represents the output of the discriminator for real samples, and $D(G(x))$ represents the output of the discriminator for generated samples. We also utilize the L_1 loss, which measures the average absolute difference between the generated samples and the actual samples. It encourages G to produce frames that are close to the actual data distribution in terms of pixel-wise similarity as shown in Equation 5.

$$L_1 = \|G(x) - x\|_1 \tag{5}$$

Here, $G(x)$, x represents the generated frame and the real inputs respectively. $||_1$ denotes the L_1 norm. We also utilize the L_2 loss between predicted and real samples. This loss emphasizes the larger dissimilarities. Equation 6 represents the L_2 loss.

$$L_2 = \|G(x) - x\|_2 \tag{6}$$

Thus, the overall loss function for VALT-GAN is the combination of L_1, L_2, and L_{adv} loss as shown in Equation 7.

$$\mathbf{L_{overall}} = L_2 + L_1 + L_{adv} \tag{7}$$

D. Anomaly Detection

To evaluate model prediction for normal occurrences, we use metrics like Mean Squared Error (MSE) and Peak Signal Noise Ratio (PSNR). PSNR calculates the generated frame's deviation from the actual frame. Normalized PSNR values calculate the anomaly score $P(t)$, as shown in Equation 8:

$$P(t) = \frac{PSNR_t - \min(PSNR)}{\max(PSNR) - \min(PSNR)} \tag{8}$$

Frames are classified as normal or abnormal using a threshold in the AD process, which compares computed anomaly scores with τ.

IV. EXPERIMENTATION

Our proposed VALT-GAN efficiently detects anomalies in the video in real time. We first describe our datasets with the experimental setup and quantitative analysis.

A. Dataset

We test VALT-GAN with three publicly available SOTA datasets: UCSD Peds2, CUHK Avenue, and ShanghaiTech. The description of the datasets are:

UCSD Peds2. This dataset [30] is focused on capturing walkways that exclusively feature pedestrians. For the purpose of evaluation, we specifically considered Peds2, which contains 4,560 frames. In this dataset, the anomaly of interest involves the presence of non-pedestrian objects. The training set consists of 16 clips, while the test set contains 12 clips.

CUHK Avenue. This Dataset [31] was gathered within a campus avenue, containing 16 clips of 15,328 frames exclusively focused on training the model to recognize normal events. The evaluation set consists of 21 clips comprising 15,324 frames, encompassing various anomalous events. These events include uncommon behavior, diverse movements in terms of speed and direction, as well as the occurrence of vehicles.

ShanghaiTech. This dataset [7] is specifically designed for crowd counting and analysis. It comprises two subsets, Part A and Part B, featuring highly congested areas. Part A focuses on stationary crowds, while Part B includes stationary and moving crowds. Part A consists of 482 images with 241,677 annotated individuals, while Part B comprises 716 images with 88,488 annotated individuals.

B. Evaluation Metric

To evaluate the performance of VALT GAN we utilized the area under the receiver operating curve (AUC). For the purpose of detecting anomalies, the ground truth labels and frame-level scores acquired by the network were compared. The AUC score is commonly utilized for anomaly detection studies and quantifies the model's ability to differentiate between normal and anomalous events. It is based on the receiver operating characteristics (ROC) curve of *true positive rate* (TPR) v.s. *false positive rate* (FPR).

$$TPR = TP/(FN + TP) \tag{9}$$

$$FPR = FP/(FP + TN) \tag{10}$$

Anomalies are determined by comparing frame scores with a threshold (refer equation 8), creating a confusion matrix, and the thresholds are modified repeatedly. TPR and FPR; see equations 9 and 10 are calculated from the matrix to plot the ROC curve.

C. Implementation details

In our experiments, we utilized Adam optimizer with a learning rate of 0.0002 to train our model on Nvidia-tesla V100 GPU. The IFGSM parameter thresh is taken as 0.9 (close to 1), and the maximum iteration is set to 10. We reshape the frames to 160 × 160 and set a window size of 4. Our implementation is done on Pytorch (version 2.0).

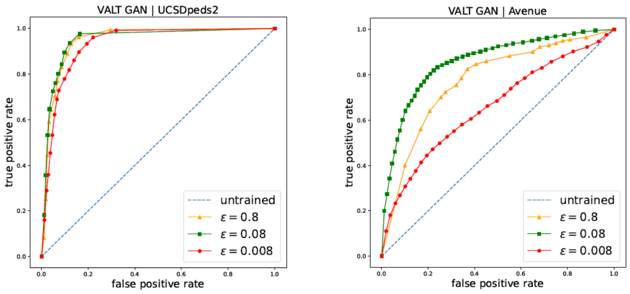

Fig. 3. ROC curve of VALT-GAN on UCSDped2 and Avenue dataset.

D. Comparison Study

This section compares VALT GAN with related work in AD, including traditional reconstruction and prediction-based methods that have given benchmark results in recent years. The AUC scores of different models are given in Tables II for the benchmark datasets.

Method	Peds2 (%)	Avenue (%)	Shanghai (%)
HybridAE [14]	92.2	81.7	68.0
Auto-reg [15]	95.4	-	72.5
CDD-AE [16]	96.5	86.0	73.3
Mem-guided [17]	88.3	83.1	70.5
MemAE [18]	94.1	83.3	71.2
AEP [22]	97.3	90.2	64.1
ITAE+NFs [19]	97.3	85.8	74.7
AMC [20]	96.2	86.9	-
Mem-guided [17]	97.0	88.5	70.5
UNet-inte [23]	96.3	85.1	73.0
ROADMAP [24]	96.3	88.3	76.6
Msm-net [25]	96.8	87.3	74.2
Cross U-net [21]	97.0	90.8	72.0
Ours	**97.2**	**91.1**	**77.4**

TABLE II
COMPARISON RESULTS OF AUC VALUES FOR UCSD PEDS2, CUHK AVENUE AND SHANGHAITECH DATASETS WITH SOTA MODELS

To illustrate the effect of IFGSM on the AD model, we perform the quantitative study on various ϵ perturbation radii. Figure 3 displays the ROC curves generated using different epsilon values of 0.8, 0.08, and 0.008 on UCSD Peds2 and Avenue. The curves visualize the relationship between the TPR and FPR at varying thresholds, allowing for a comprehensive assessment of the performance of our VALT-GAN. In our experiments, we observe that a larger radius leads to mode collapse whereas a smaller radius does not provide sufficient robustness, which is reflected in the ROC curves.

E. Ablation Study

1) Run time analysis: We compare the VALT-GAN execution time of the Peds2 Dataset on various SOTA methods. As shown in Table III, The average computation time for the test frame reported by our proposed method on the peds2 dataset is **48ms**.

2) Generator network: The AUC performance (%) of the VALT-GAN using WiderResnet as the backbone across various combinations of components as shown in Table IV on CUHK Avenue, UCSD Ped2, and ShanghaiTech datasets.

Methods	Processor	Peds2
FFP+MC [3]	GPU	46
CDD-AE [16]	GPU	60
MemAE [18]	GPU	38
Msm-net [25]	GPU	50
VALT-GAN (Ours)	GPU	**36**

TABLE III
COMPARISONS BETWEEN VALT-GAN AND OTHER APPROACHES IN TERMS OF TIME TAKEN TO COMPUTE EACH TESTING FRAME IN MILLISECONDS (MS)

Methods	Peds2	Avenue	ShanghaiTech
ResNet-50	95.1	83.3	71.9
Se-ResNext-50	96.1	84.8	73.5
WiderResNet	97.2	91.1	77.4

TABLE IV
COMPARISON OF THE PROPOSED VALT-GAN WITH DIFFERENT DCNN AS BACKBONE IN TERMS OF AUC (%). THE UTILIZATION OF WIDERRESNET AS THE BACKBONE ARCHITECTURE IN THE PROPOSED FRAMEWORK DEMONSTRATES BEST PERFORMANCE.

V. CONCLUSION

We proposed the VALT-GAN model to detect anomalies in video datasets. It consists of a two-stream generator G and a discriminator D. The prime contribution of our work is to select adversarial samples in adversarial learning trained in end-to-end procedure to achieve the modulation between the generalization ability of G and the discriminating capability of D. The experimental results show that our model performs better in accuracy and stability than the majority of competing models, proving the potential and dependability of the Latent space mining approach in adversarial learning.

REFERENCES

[1] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, “Learning temporal regularity in video sequences,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 733–742, 2016.

[2] W. Li, V. Mahadevan, and N. Vasconcelos, “Anomaly detection and localization in crowded scenes,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 18–32, 2013.

[3] W. Liu, W. Luo, D. Lian, and S. Gao, “Future frame prediction for anomaly detection–a new baseline,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6536–6545, 2018.

[4] Y. Zhang, H. Lu, L. Zhang, X. Ruan, and S. Sakai, “Video anomaly detection based on locality sensitive hashing filters,” *Pattern Recognition*, vol. 59, pp. 302–311, 2016.

[5] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang, “Video anomaly detection and localization using hierarchical feature representation and gaussian process regression,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2909–2917, 2015.

[6] J. Kim and K. 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*, pp. 2921–2928, IEEE, 2009.

[7] W. Luo, W. Liu, and S. Gao, “A revisit of sparse coding based anomaly detection in stacked rnn framework,” in *Proceedings of the IEEE international conference on computer vision*, pp. 341–349, 2017.

[8] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, “Learning deep representations of appearance and motion for anomalous event detection,” *arXiv preprint arXiv:1510.01553*, 2015.

[9] B. Zhao, L. Fei-Fei, and E. P. Xing, “Online detection of unusual events in videos via dynamic sparse coding,” in *CVPR 2011*, pp. 3313–3320, IEEE, 2011.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[11] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, “Deep autoencoding gaussian mixture model for unsupervised anomaly detection,” in *International conference on learning representations*, 2018.

[12] Y. Kishi, T. Ikegami, S.-i. O’uchi, R. Takano, W. Nogami, and T. Kudoh, “Perturbative gan: Gan with perturbation layers,” *arXiv preprint arXiv:1902.01514*, 2019.

[13] Y. Li, Y. Mo, L. Shi, and J. Yan, “Improving generative adversarial networks via adversarial learning in latent space,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 8868–8881, 2022.

[14] T. N. Nguyen and J. Meunier, “Hybrid deep network for anomaly detection,” *arXiv preprint arXiv:1908.06347*, 2019.

[15] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, “Latent space autoregression for novelty detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 481–490, 2019.

[16] Y. Chang, Z. Tu, W. Xie, and J. Yuan, “Clustering driven deep autoencoder for video anomaly detection,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pp. 329–345, Springer, 2020.

[17] H. Park, J. Noh, and B. Ham, “Learning memory-guided normality for anomaly detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14372–14381, 2020.

[18] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. 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*, pp. 1705–1714, 2019.

[19] M. Cho, T. Kim, W. J. Kim, S. Cho, and S. Lee, “Unsupervised video anomaly detection via normalizing flows with implicit latent features,” *Pattern Recognition*, vol. 129, p. 108703, 2022.

[20] T.-N. Nguyen and J. Meunier, “Anomaly detection in video sequence with appearance-motion correspondence,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1273–1283, 2019.

[21] Y. Kim, J.-Y. Yu, E. Lee, and Y.-G. Kim, “Video anomaly detection using cross u-net and cascade sliding window,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 3273–3284, 2022.

[22] J. Yu, Y. Lee, K. C. Yow, M. Jeon, and W. Pedrycz, “Abnormal event detection and localization via adversarial event prediction,” *IEEE transactions on neural networks and learning systems*, vol. 33, no. 8, pp. 3572–3586, 2021.

[23] Y. Tang, L. Zhao, S. Zhang, C. Gong, G. Li, and J. Yang, “Integrating prediction and reconstruction for anomaly detection,” *Pattern Recognition Letters*, vol. 129, pp. 123–130, 2020.

[24] X. Wang, Z. Che, B. Jiang, N. Xiao, K. Yang, J. Tang, J. Ye, J. Wang, and Q. Qi, “Robust unsupervised video anomaly detection by multipath frame prediction,” *IEEE transactions on neural networks and learning systems*, vol. 33, no. 6, pp. 2301–2312, 2021.

[25] Y. Cai, J. Liu, Y. Guo, S. Hu, and S. Lang, “Video anomaly detection with multi-scale feature and temporal information fusion,” *Neurocomputing*, vol. 423, pp. 264–273, 2021.

[26] Z. Wu, C. Shen, and A. Van Den Hengel, “Wider or deeper: Revisiting the resnet model for visual recognition,” *Pattern Recognition*, vol. 90, pp. 119–133, 2019.

[27] S. Zagoruyko and N. Komodakis, “Wide residual networks,” *arXiv preprint arXiv:1605.07146*, 2016.

[28] J. Lin, C. Gan, and S. Han, “Tsm: Temporal shift module for efficient video understanding,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7083–7093, 2019.

[29] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[30] A. Chan and N. Vasconcelos, “Ucsd pedestrian database,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, 2008.

[31] C. Lu, J. Shi, and J. Jia, “Abnormal event detection at 150 fps in matlab,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2720–2727, 2013.