Supplementary Material for MULDE: Multiscale Log-Density Estimation via Denoising Score Matching for Video Anomaly Detection

Jakub Micorek Horst Possegger Dominik Narnhofer Horst Bischof Mateusz Koziński
Graz University of Technology, Austria
{jakub.micorek, possegger, dominik.narnhofer, bischof, mateusz.kozinski}@icg.tugraz.at

This supplementary material extends the results presented in the main manuscript with additional visualizations (Section 1) and detailed experiment results (Section 2).

1. Visualizations

Intuitive example of multiscale log-density estimation To provide more intuition about our neural log-density approximation, in Figure 1 we present a toy example that extends Figure 2 from the main manuscript. In Figure 1b, we plot our log-density approximation as a trajectory across a range of noise scales σ, for each normal and anomalous sample. Our log-density estimation separates normal and anomalous data well across a wide range of noise scales σ. We show the log-density estimation in Figure 1c.

Consistency of the anomaly score across different videos In Figures 2 and 3, we illustrate the trajectories of MULDE's anomaly score across different videos of the same scene, taken from the **ShanghaiTech** data set. The levels of the anomaly score for normal fragments of different videos are consistent, as are the levels of the score for anomalous fragments. The regularized MULDE exhibits better discrimination between normal and anomalous behavior than MULDE$_{\beta=0}$ without regularization.

2. Additional Results

In this section, we complement the results presented in the main manuscript for the frame-centric and object-centric setup. Furthermore, we provide further details on the choice of σ, the selection of L, and alternatives to the GMM fitting.

2.1. Frame-centric

In Table 1, we extend Table 2 of the main manuscript to include the less recent frame-centric VAD methods. The results of the competing methods were reproduced after the original publications, except for MSMA [13] which we re-implemented for processing videos, and AccI-VAD [17], which we adapted to frame-centric operation. Our method, MSMA, and AccI-VAD used the Hiera-L [18] features.

Method	ShanghaiTech		UCF-Crime		UBnormal	
	Micro	Macro	Micro	Macro	Micro	Macro
MNAD-Recon. [14]	70.5	-	-	-	-	-
Mem-AE. [8]	71.2	-	-	-	-	-
Frame-Pred. [11]	72.8	-	-	-	-	-
ClusterAE [4]	73.3	-	-	-	-	-
AMMCN [3]	73.7	-	-	-	-	-
MPN [12]	73.8	-	-	-	-	-
DLAN-AC [23]	74.7	-	-	-	-	-
BMAN [10]	76.2	-	-	-	-	-
CT-D2GAN [5]	77.7	-	-	-	-	-
CAC [21]	<u>79.3</u>	-	-	-	-	-
Scene-Aware [19]	74.7	-	72.7	-	-	-
BODS [20]	-	-	68.3	-	-	-
GODS [20]	-	-	70.5	-	-	-
GCL [24]	-	-	74.2	-	-	-
UBnormal [1]	-	-	-	-	68.5	80.3
FPDM [22]	78.6	-	74.7	-	62.7	-
AccI-VAD$_{\text{GMM}}$* [17]	76.2	82.9	60.3	84.5	66.8	83.2
AccI-VAD$_{\text{kNN}}$* [17]	71.9	83.1	53.0	82.7	65.2	82.5
MSMA* [13]	76.7	84.2	64.5	83.4	70.3	85.1
MULDE$_{\beta=0}$(ours)	78.4	**86.0**	<u>75.9</u>	<u>84.8</u>	<u>71.3</u>	**86.0**
MULDE(ours)	**81.3**	<u>85.9</u>	**78.5**	**84.9**	**72.8**	<u>85.5</u>

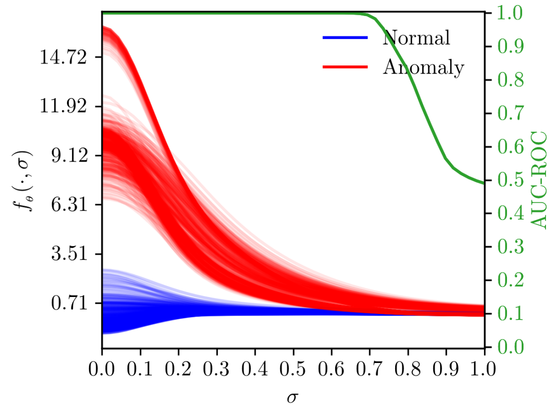
Table 1. Frame-centric results. Frame-level AUC-ROC (%) comparison (best marked **bold**, second best <u>underlined</u>). *implemented by us.

MULDE surpasses the baselines on all three data sets in terms of both the *micro* and the *macro* metric.

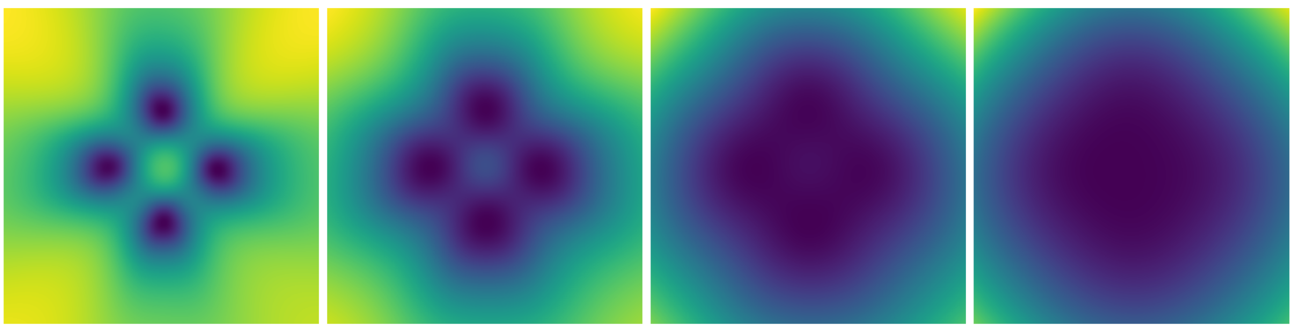
The choice of feature extractors In Table 2, we compare the performance of MULDE used with different video feature extractors in frame-centric VAD. For reference, the results reported in the main manuscript were obtained with Hiera-L [18]. We observe, that Hiera-H outperforms Hiera-L in certain experiments, but runs considerably slower. Hiera-B is the fastest feature extractor, but produces less discriminative features than Hiera-L and Hiera-H. Consequently, we opted for Hiera-L due to its favorable tradeoff between computation time and accuracy.

Comparing the *micro* performance attained by MULDE with I3D features (71.3%, bottom row of Table 2) to the re-

(a) Example dataset: Normal features and anomalous features.



(b) AUC-ROC across multiple noise-scales σ based on the negative log probability f_θ. The normal and anomalous samples are well separable.



$f_\theta(\cdot, \sigma = 0.001)$ $f_\theta(\cdot, \sigma = 0.1)$ $f_\theta(\cdot, \sigma = 0.5)$ $f_\theta(\cdot, \sigma = 1.0)$

(c) The log-density of normal training features is estimated with f_θ across multiple σ. MULDE leverages f_θ as a strong anomaly indicator.

Figure 1. The intuition behind the use of log-density estimation for anomaly detection. (a) Normal features, sampled from a mixture of four Gaussians, are shown in blue, while anomalous features are shown in red. (b) compares the values of f_θ for features from the normal and anomalous samples. Each graph shows the log-density at multiple noise scales for a single sample. Our anomaly indicator f_θ is well suited to separate anomalies from normal data. (c) shows the log-density approximations across noise scales.

Dataset	Features	Micro							Macro						
		MULDE				MSMA*	AccI-VAD*		MULDE				MSMA*	AccI-VAD*	
		β_0	$\beta_{0.01}$	$\beta_{0.1}$	β_1		GMM	kNN	β_0	$\beta_{0.01}$	$\beta_{0.1}$	β_1		GMM	kNN
ShanghaiTech	Hiera-B	72.1	<u>73.4</u>	73.1	**73.7**	71.5	67.9	68.2	<u>83.1</u>	**83.6**	82.7	81.9	79.8	80.6	79.0
ShanghaiTech	Hiera-L	78.4	80.7	<u>81.3</u>	**81.4**	76.7	76.2	71.9	**86.0**	84.9	<u>85.9</u>	84.5	84.2	82.9	83.1
ShanghaiTech	Hiera-H	79.4	79.8	<u>79.9</u>	**81.7**	77.4	74.7	72.7	**88.3**	87.0	<u>87.6</u>	86.8	86.4	86.0	83.5
UBnormal	Hiera-B	70.2	<u>71.8</u>	71.6	**72.4**	70.7	66.0	63.1	84.0	<u>84.1</u>	<u>84.1</u>	83.7	**85.4**	83.6	81.9
UBnormal	Hiera-L	71.3	**72.9**	<u>72.8</u>	72.5	70.3	66.8	65.2	**86.0**	85.2	<u>85.5</u>	84.7	85.1	83.2	82.5
UBnormal	Hiera-H	70.5	<u>72.7</u>	<u>72.7</u>	**72.8**	71.2	67.7	63.0	<u>86.9</u>	**87.1**	**87.1**	86.3	<u>86.9</u>	85.7	84.5
UCF-Crime	Hiera-B	**74.2**	72.2	71.9	<u>72.4</u>	69.2	69.4	68.1	85.1	**85.6**	<u>85.2</u>	84.6	83.5	85.1	84.0
UCF-Crime	Hiera-L	75.9	76.6	**78.5**	<u>77.2</u>	64.5	60.3	53.0	84.8	<u>84.9</u>	<u>84.9</u>	**85.5**	83.4	84.5	82.7
UCF-Crime	Hiera-H	74.8	**76.7**	<u>75.0</u>	74.9	71.4	60.4	57.3	**87.4**	85.4	85.0	85.0	<u>86.7</u>	84.6	82.7
UCF-Crime	I3D	67.6	<u>70.8</u>	69.9	**71.3**	68.2	63.5	64.2	**87.6**	<u>87.5</u>	87.3	86.5	87.1	**87.6**	<u>87.5</u>

Table 2. Frame-level AUC-ROC (%) comparison. For each input feature representation Hiera-B(ase), Hiera-L(arge), Hiera-H(uge), and I3D, we mark the best scores **bold** and <u>underline</u> the second-best. *adapted from image-based anomaly detection (MSMA) and object-centric VAD (AccI-VAD) to frame-centric VAD.

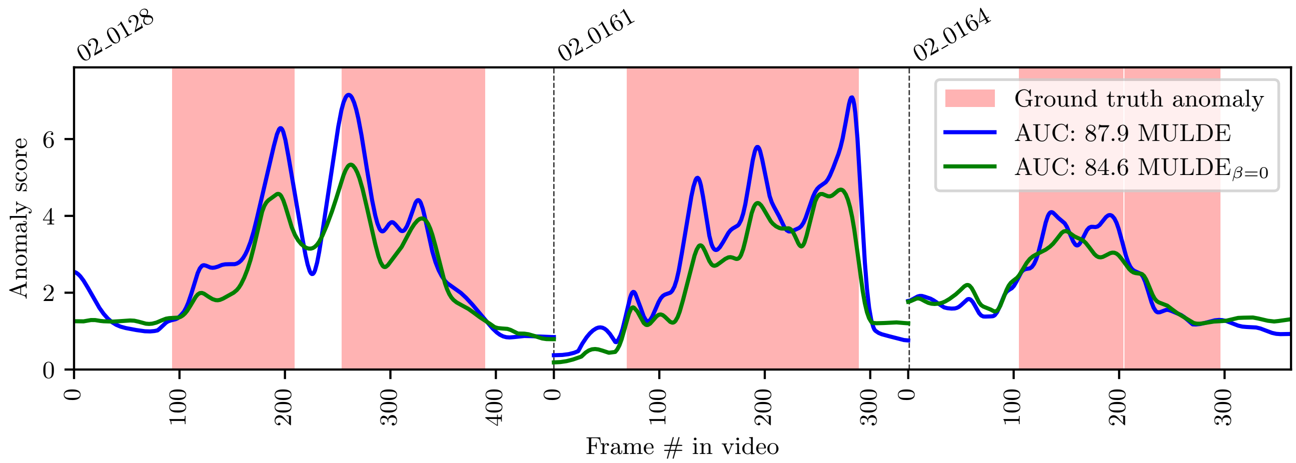


Figure 2. The value of MULDE's anomaly score computed for each frame of the test scene 02 of the **ShanghaiTech** data set (videos 02_0128, 02_0161, and 02_0164) and the resulting *micro* AUC score. MULDE with regularization has an advantage (+3.3%) over the non-regularized MULDE$_{\beta=0}$.

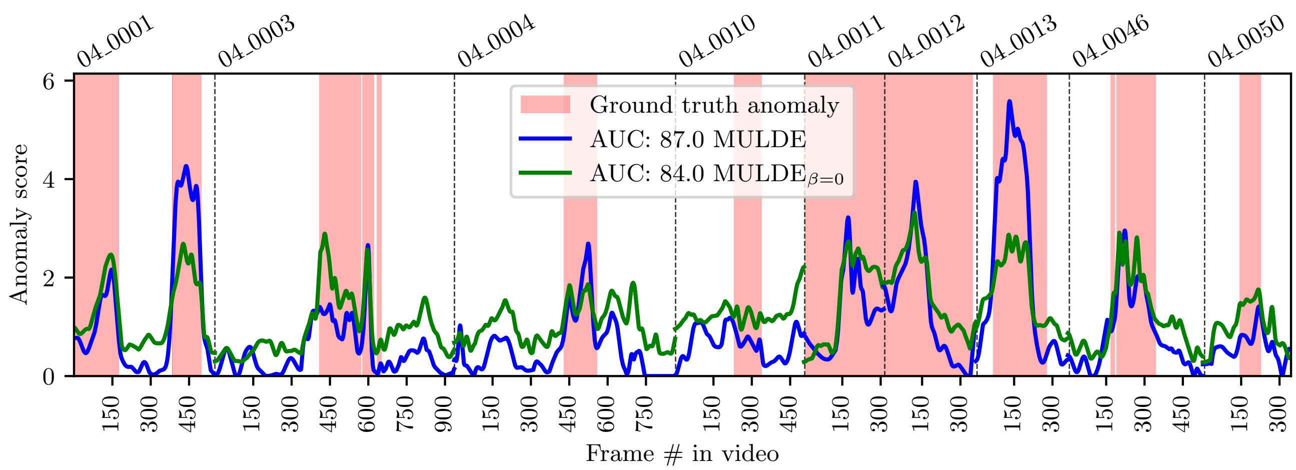


Figure 3. MULDE's anomaly score computed for each frame of the test scene 04 of the **ShanghaiTech** data set (videos 04_0001, 04_0003, etc.) and the resulting *micro* AUC score. MULDE with regularization outperforms the non-regularized MULDE by 3 percent points.

sults of BODS (68.3%) and GODS (70.5%, Table 1), which both also use I3D features, we see that MULDE is a favorable anomaly detector. In particular, this shows that our approach is truly feature-agnostic and works well with any input feature representation.

2.2. Object-centric

In the object-centric experiments, reported in Table 1 of the main manuscript, we used MULDE in combination with the feature extraction pipeline proposed by Reiss and Hoshen [17]. It detects objects in each video frame and extracts deep, velocity, and human pose features for each detected object, as detailed in the manuscript. Here, we complement these results with the performance attained by MULDE, MSMA [13], and AccI-VAD [17] using the pose (P), deep (D), and velocity (V) features separately. AccI-VAD pools the highest anomaly scores from each frame for P, D, and V and normalizes each feature type by its min/max training counterpart. Finally, the scores of P, D, and V are added up using pairs of the feature types and the triplet. MULDE differs in that regard, instead of min-/max-normalization, we standardize by the training statistics, then clip negative values which are normal, and add up. We follow AccI-VAD's ablation protocol for MULDE and present the results for the **Avenue** data set in Table 3. Table 4 contains the results obtained for **ShanghaiTech**. For **Avenue**, the combination

P	D	V	Micro				Macro			
			AccI-VAD $kNN_{P,D}$ GMM_V	MSMA*	MULDE	$MULDE_{\beta=0}$	AccI-VAD $kNN_{P,D}$ GMM_V	MSMA*	MULDE	$MULDE_{\beta=0}$
✓			73.8	84.2	**84.6**	84.5	76.2	**87.0**	86.5	86.0
	✓		85.4	87.7	**89.0**	87.5	87.7	87.9	88.0	**88.3**
		✓	86.0	83.6	86.6	**87.4**	89.6	87.2	91.8	**92.7**
✓	✓		89.3	89.5	**91.5**	90.6	88.8	89.5	**91.0**	**91.0**
	✓	✓	93.0	89.4	**93.1**	92.5	**95.5**	91.2	94.7	94.7
✓		✓	87.8	86.7	**91.1**	90.2	93.0	90.5	95.3	**95.8**
✓	✓	✓	93.3	90.2	**94.3**	93.1	**96.2**	92.5	96.1	96.1
Best			93.3	90.2	**94.3**	93.1	**96.2**	92.5	96.1	96.1

Table 3. Detailed results for object-centric setup for the **Avenue** dataset on the **micro** and **macro** frame-level AUC-ROC evaluation. Combinations of the object-centric pose (P), deep features (D), and velocities (V) features following the AccI-VAD [17] ablations. For every object-centric feature and its combinations, we mark the best scores **bold** and underline the second-best. *adapted from image-based anomaly detection (MSMA) to object-centric VAD.

P	D	V	Micro				Macro			
			AccI-VAD $kNN_{P,D}$ GMM_V	MSMA*	MULDE	$MULDE_{\beta=0}$	AccI-VAD $kNN_{P,D}$ GMM_V	MSMA*	MULDE	$MULDE_{\beta=0}$
✓			74.5	76.4	**78.5**	76.0	81.0	82.2	**83.6**	82.1
	✓		72.5	74.6	**76.6**	74.9	82.5	78.8	82.3	**83.0**
		✓	**84.4**	81.5	82.0	82.4	84.8	86.1	88.1	**88.2**
✓	✓		76.7	81.5	**82.6**	80.5	84.9	**89.5**	88.8	87.5
	✓	✓	**84.5**	79.3	82.2	81.7	**88.7**	83.5	87.9	87.4
✓		✓	85.9	84.1	**86.6**	86.4	88.8	90.0	**91.5**	91.0
✓	✓	✓	85.1	83.7	**86.7**	84.8	89.6	90.2	**90.6**	89.8
Best			85.9	84.1	**86.7**	86.4	89.6	90.2	**91.5**	91.0

Table 4. Detailed results for object-centric setup for the **ShanghaiTech** dataset on the **micro** and **macro** frame-level AUC-ROC evaluation. Combinations of the object-centric pose (P), deep features (D), and velocities (V) features following the AccI-VAD [17] ablations. For every object-centric feature and its combinations, we mark the best scores **bold** and underline the second-best. *adapted from image-based anomaly detection (MSMA) to object-centric VAD.

of all three feature types gives the best micro and macro scores. Similarly, for **ShanghaiTech**, P, D, V leads to the best micro score, and the combination of P and V leads to the best macro score. These results show that it might be beneficial to combine features, encoding complementary information. MULDE makes this combination easy as it is feature-agnostic.

Performance in terms of the region- and track-based detection criteria The region- and track-based detection criteria (RBDC and TBDC) were introduced by Ramachandra and Jones [15] to assess the anomaly localization capabilities of VAD methods, which are not captured by the more common frame-level AUC-ROC scores.

RBDC and TBDC require pixel-level anomaly scores. AccI-VAD [17], SSMTL [6], BA-AED [7], MULDE apply the anomaly score to the bounding-box, *i.e.* the region obtained by the object detector. We computed the RBDC and TBDC metrics for MULDE using the code and annotations released by Georgescu et al. [7]. We provide RBDC and TBDC scores for regions based on deep features (D)

Method	Avenue		ShanghaiTech	
	RBDC	TBDC	RBDC	TBDC
Ramachandra *et al.* [15]	35.8	**80.9**	-	-
Ramachandra *et al.* [16]	41.2	78.6	-	-
Frame-Pred. [11]	-	-	17.0	54.2
CAE-SVM [9]	-	-	20.6	44.5
BA-AED [7]	65.0	67.0	41.3	78.8
SSMTL [6]	57.0	58.3	42.8	83.9
SSMTL [6]+UBnormal [1]†	61.1	61.4	47.2	**86.2**
$MULDE_D$ (ours)	**73.1**	74.4	48.9	81.2
$MULDE_V$ (ours)	13.8	46.8	**55.0**	85.6
$MULDE_{D,V}$ (ours)	71.8	79.2	52.7	83.6

Table 5. Localization-based evaluation using RBDC and TBDC scores [15]. We provide scores for regions based on deep features (D) only, velocity (V) only and the combination of D, V. †extended training data used.

L	4	8	16	32	64
AUC-ROC	72.16	72.80	72.89	72.95	72.99

Table 6. Frame-centric results on **UBnormal** for a different number of noise scales L. Frame-level micro AUC-ROC (%).

only, velocity (V) only, and the combination of D and V in Table 5. Pose (P) is not used for this evaluation, as the features provided by Reiss and Hoshen [17] are already normalized to the top left image corner and thus, can not be attributed to a specific location within the frame. For comparison, we report the results of the baseline methods after [1]. MULDE clearly outperforms previous approaches in terms of the region-based RBDC. In terms of the TBDC, MULDE is among the top-performing approaches, outperformed only by [15] on **Avenue** and by [1] (which requires additional training data) on **ShanghaiTech**.

2.3. Parameter selection and alternatives to GMM

In this section, we discuss the selection of the noise range σ, the selection of the number of noise scales L, details on the GMM fitting and alternative approaches to the GMM fitting.

Noise range selection of σ Even though our method eliminates the need to select the noise range $[\sigma_{\text{low}}, \sigma_{\text{high}}]$ used for training, the range of employed noise scales should be sufficiently large to cover anomalies that would be seen at test time. Currently, there is no automatic way to select the upper limit of the noise range. We circumvent this limitation by standardizing the features component-wise and using a fixed, wide range of noise scales. We set $\sigma_{\text{high}} = 1.0$ to make it equal to the standard deviation of the distribution of training video features. The $\sigma_{\text{low}} = 0.001$ was selected to make the interval wide. We kept this range for all data sets, even though Figure 4 of the main manuscript (frame-centric micro score on **ShanghaiTech**) suggests that such a wide interval might not be necessary: When used with a single σ, MULDE performs best for $\sigma = 0.33$, and $\sigma < 0.2$ or $\sigma > 0.5$ lead to much lower scores. Initial experiments showed promising results across all the datasets; thus, we did not fine-tune these hyperparameters.

Selection of number of noise scales L In all the reported experiments, we decimated the range of noise scales into $L = 16$ points. Testing other values of L in the frame-centric VAD on **UBnormal** (results reported in Table 6) reveals that MULDE is not sensitive to the number of noise scales used.

Details on GMM fitting Once the network f_θ is trained, we compute the multi-scale log-density approximation for each video feature $\mathbf{x}$ in the training set $\mathcal{T}$. This results in a data set of vectors $\{[f_\theta(\mathbf{x}, \sigma_1), \dots f_\theta(\mathbf{x}, \sigma_L)]\}_{\mathbf{x} \in \mathcal{T}}$. In other words, each single $d-$dimensional video feature $\mathbf{x}$ is evaluated at L noise scales which results in a new $L-$dimensional feature vector. We then fit a $L-$dimensional GMM with one, three, and five components to this set of vectors using the Expectation–maximization algorithm.

At test time, our neural network takes a vector of a video feature and produces a multi-scale vector of log-density approximations, which is then input to the GMM yielding a negative log-likelihood which we use as the anomaly score. Finally, like in previous work [1, 2, 6, 7, 9, 17], these scores are temporally smoothed with a 1d-Gaussian filter to obtain the final anomaly score.

Alternatives to GMM As discussed in section 3.3 of the manuscript, in theory, a log-likelihood estimation at a well-chosen noise level is sufficient to detect anomalies. This is confirmed by the result presented in Fig. 4 of the paper, where we see that using the best, single noise level yields a micro score on par with the GMM. However, the choice of the optimal noise scale is not trivial: the noise should be high enough to blend modes of the probability density function originating from individual training samples but not so high as to distort the shape of the original, noise-free distribution. It is difficult to determine the optimal noise level without anomalous validation data, which prompted us to use the GMM. Theoretically, we could substitute the GMM with an alternative aggregation method, for example, max-, average-, or median-pooling. We evaluated these methods as follows: Before pooling, we equalized the log-density estimates by standardizing them across each noise scale with their respective means and standard deviations computed over the training set. We evaluate the alternative pooling method to the frame-centric experiment on **ShanghaiTech**. As reported in Table 2 in the manuscript, MULDE$_{\beta=0}$ with the GMM attains a Micro score of 78.4. This result decreases to 76.30 with max-pooling, 76.02 with average-pooling, and 75.83 with median-pooling instead of the GMM.

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 *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[2] Antonio Barbalau, 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, 2023.

[3] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. Appearance-motion memory consistency network for video anomaly detection. In *Proc. of the Conference on Artificial Intelligence (AAAI)*, 2021.

[4] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. Clustering driven deep autoencoder for video anomaly detection. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2020.

[5] Xinyang Feng, Dongjin Song, Yuncong Chen, Zhengzhang Chen, Jingchao Ni, and Haifeng Chen. Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection. In *Proc. of the International Conference on Multimedia*, 2021.

[6] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[7] Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 44(9), 2021.

[8] 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 *Proc. of the International Conference on Computer Vision (ICCV)*, 2019.

[9] 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 *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Bman: Bidirectional multi-scale aggregation networks for abnormal event detection. *Transactions on Image Processing*, 29, 2019.

[11] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning Normal Dynamics in Videos with Meta Prototype Network. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[13] Ahsan Mahmood, Junier Oliva, and Martin Andreas Styner. Multiscale score matching for out-of-distribution detection. *Proc. of the International Conference on Learning Representations (ICLR)*, 2021.

[14] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proc. of the Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[16] Bharathkumar Ramachandra, Michael Jones, and Ranga Vatsavai. Learning a distance function with a siamese network to localize anomalies in videos. In *Proc. of the Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[17] Tal Reiss and Yedid Hoshen. Attribute-based representations for accurate and interpretable video anomaly detection. *arXiv preprint arXiv:2212.00789*, 2022.

[18] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. *Proc. of the International Conference on Machine Learning (ICML)*, 2023.

[19] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. Scene-aware context reasoning for unsupervised abnormal event detection in videos. In *Proc. of the International Conference on Multimedia*, 2020.

[20] Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2019.

[21] Ziming Wang, Yuexian Zou, and Zeming Zhang. Cluster attention contrast for video anomaly detection. In *Proc. of the International Conference on Multimedia*, 2020.

[22] Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2023.

[23] Zhiwei Yang, Peng Wu, Jing Liu, and Xiaotao Liu. Dynamic local aggregation network with adaptive clusterer for anomaly detection. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2022.

[24] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.