APPENDIX

SUPPLEMENTARY RESULTS

In this appendix, we present further empirical results of the study. Further supplementary materials on hyperparameters can be found in this link.

A. Benchmark Deep Learning Models

Table VIII compares the configuration and study coverage of our LA3D approach and the related benchmark AN studies that employ DL models.

TABLE VIII
DEEP LEARNING-BASED BENCHMARK AN STUDIES

Model Name	Complexity	AN Type	PD	ReID	VAD / Action Recognition
SPACT [8]	DL	Full-Image	✓	×	✓
TED-SPAD [6]	DL	Full-Image	✓	×	✓
SG-GAN [51]	DL	Full-Body	×	✓	×
DEEPPRIVACY2 [20]	DL	Full-Body	×	✓	×
LA3D (ours)	Lightweight	Full-Body	✓	✓	✓

B. Further Evaluation on Person Re-Identification

We have further evaluated ReID for recognition attacks that employ non-anonymized data on either the query or the search gallery. Table IX and Table X present the ReID protection capability of the ANs when the anonymized image of a person is utilized to search for identification from a gallery of raw images and vice versa, respectively. In both cases, the adaptive ANs consistently outperform the corresponding baselines. Blurring is more susceptible to ReID attack when both the query and gallery are anonymized, and the protection drops by approximately 20%–26% for $\mathcal{G}^0$, and 10%–17% for $\mathcal{G}^{\text{a}}$ in the mAP. The $\mathcal{G}^{\text{a}}_{\text{max}}$ is the least affected with decline of 10%–11%.

C. Video Quality Assessment for Action Utility Preservation

We present a video quality analysis to evaluate the general capability of preserving the action utility of the AN. The employed conventional ANs are perturbation-based, not aimed at generating a realistic image (like GANs), and the purpose is to retain the visual utility for the end target VAD task through the action recognition I3D video encoder. Hence, we employed quality metrics, such as Fréchet Inception Distance (FID) and Kernel Inception Distance (KID), to measure the closeness of retained information before and after AN on I3D-encoded features of the videos.

Fig. 13 presents the average quality scores over the 290 test videos of the UCF-Crime dataset and 800 videos of the XD-Violence. The results overall correlate with the VAD performance, and the impact of AN exhibits consistency across the VAD and I3D models. However, the EDGED AN achieves the poorest quality scores. Despite these being in harmony with the VAD performance on the XD-Violence, the AN performed well for the MGFN on the UCF-Crime dataset (as shown in Fig. 10). The adaptive ANs have provided results that are relatively lower than their baseline, which is expected considering the stronger AN. Nonetheless, the FID and KID scores are generally low, indicating a promising preservation of the action utilities. We hypothesize that the AN methods can lead to varying efficacy levels for non-VAD action recognition tasks, and an in-depth investigation for a given target task is essential.

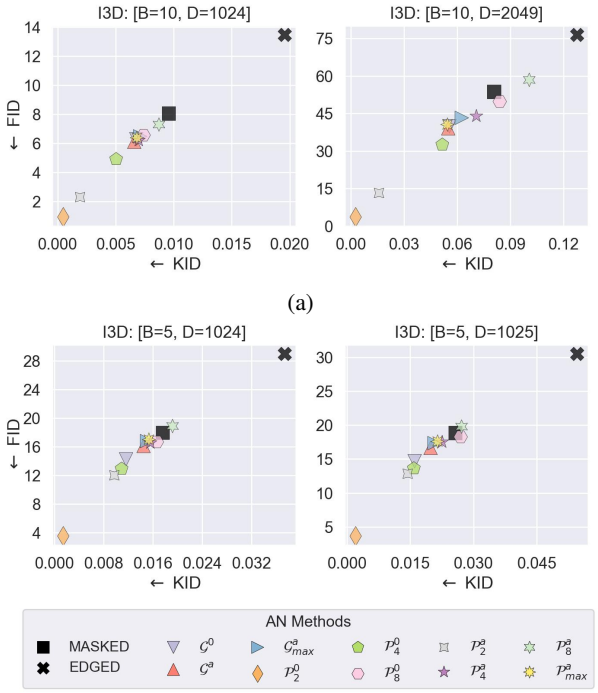
TABLE IX
REID (ANONYMIZED QUERY VS. RAW GALLERY) ON THE MARKET1501 DATASET [52] USING OSNET [53].

AN Method	mAP↓	$\Delta_{\text{N}}\downarrow$	$\Delta_0\downarrow$	CMC-R1↓	$\Delta_{\text{N}}\downarrow$	$\Delta_0\downarrow$
No-AN	0.826	–	–	0.942	–	–
MASKED	0.011	-98.7%	–	0.010	-98.9%	–
EDGED	**0.009**	**-98.9%**	–	**0.008**	**-99.2%**	–
$\mathcal{G}^0$	0.165	-80.0%	–	0.202	-78.6%	–
$\mathcal{G}^{\text{a}}(\alpha_l=0.5)$	0.042	-94.9%	-74.5%	0.042	-95.5%	-79.2%
$\mathcal{G}^{\text{a}}(\alpha_l=1.0)$	0.034	-95.9%	-79.4%	0.035	-96.3%	-82.7%
$\mathcal{G}^{\text{a}}_{\text{max}}$	0.022	-97.3%	-86.7%	0.022	-97.7%	-89.1%
$\mathcal{P}^0_2$	0.800	-3.1%	–	0.926	-1.7%	–
$\mathcal{P}^0_4$	0.687	-16.8%	–	0.832	-11.7%	–
$\mathcal{P}^0_8$	0.320	-61.3%	–	0.413	-56.2%	–
$\mathcal{P}^{\text{a}}_2(\alpha_l=0.5)$	0.079	-90.4%	-90.1%	0.092	-90.2%	-90.1%
$\mathcal{P}^{\text{a}}_4(\alpha_l=0.5)$	0.071	-91.4%	-89.7%	0.079	-91.6%	-90.5%
$\mathcal{P}^{\text{a}}_8(\alpha_l=0.5)$	0.073	-91.2%	-77.2%	0.079	-91.6%	-80.9%
$\mathcal{P}^{\text{a}}_2(\alpha_l=1.0)$	0.074	-91.0%	-90.8%	0.085	-91.0%	-90.8%
$\mathcal{P}^{\text{a}}_4(\alpha_l=1.0)$	0.060	-92.7%	-91.3%	0.059	-93.7%	-92.9%
$\mathcal{P}^{\text{a}}_8(\alpha_l=1.0)$	0.027	-96.7%	**-91.6%**	0.020	-97.9%	**-95.2%**
$\mathcal{P}^{\text{a}}_{\text{max}}$	0.027	-96.7%	–	0.020	-97.9%	–

TABLE X
REID (RAW QUERY VS. ANONYMIZED GALLERY) ON THE MARKET1501 DATASET [52] USING OSNET [53].

AN Method	mAP↓	$\Delta_{\text{N}}\downarrow$	$\Delta_0\downarrow$	CMC-R1↓	$\Delta_{\text{N}}\downarrow$	$\Delta_0\downarrow$
No-AN	0.826	–	–	0.942	–	–
MASKED	0.016	-98.1%	–	0.096	-89.8%	–
EDGED	**0.013**	**-98.4%**	–	**0.093**	**-90.1%**	–
$\mathcal{G}^0$	0.217	-73.7%	–	0.387	-58.9%	–
$\mathcal{G}^{\text{a}}(\alpha_l=0.5)$	0.065	-92.1%	-70.0%	0.172	-81.7%	-55.6%
$\mathcal{G}^{\text{a}}(\alpha_l=1.0)$	0.056	-93.2%	-74.2%	0.156	-83.4%	-59.7%
$\mathcal{G}^{\text{a}}_{\text{max}}$	0.036	-95.6%	-83.4%	0.132	-86.0%	-65.9%
$\mathcal{P}^0_2$	0.806	-2.4%	–	0.933	-1.0%	–
$\mathcal{P}^0_4$	0.704	-14.8%	–	0.895	-5.0%	–
$\mathcal{P}^0_8$	0.350	-57.6%	–	0.651	-30.9%	–
$\mathcal{P}^{\text{a}}_2(\alpha_l=0.5)$	0.080	-90.3%	-90.1%	0.262	-72.2%	-71.9%
$\mathcal{P}^{\text{a}}_4(\alpha_l=0.5)$	0.084	-89.8%	-88.1%	0.219	-76.8%	-75.5%
$\mathcal{P}^{\text{a}}_8(\alpha_l=0.5)$	0.084	-89.8%	-76.0%	0.221	-76.5%	-66.1%
$\mathcal{P}^{\text{a}}_2(\alpha_l=1.0)$	0.076	-90.8%	-90.6%	0.246	-73.9%	-73.6%
$\mathcal{P}^{\text{a}}_4(\alpha_l=1.0)$	0.059	-92.9%	**-91.6%**	0.193	-79.5%	-78.4%
$\mathcal{P}^{\text{a}}_8(\alpha_l=1.0)$	0.032	-96.1%	-90.9%	0.126	-86.6%	**-80.6%**
$\mathcal{P}^{\text{a}}_{\text{max}}$	0.032	-96.1%	–	0.126	-86.6%	–
*SG-GAN [51]	0.144	-82.6%	–	0.311	-67.0%	–
*DEEPPRIVACY2 [20]	0.085	-89.7%	–	0.447	-52.5%	–

The * are DL AN methods that employ body image generators.



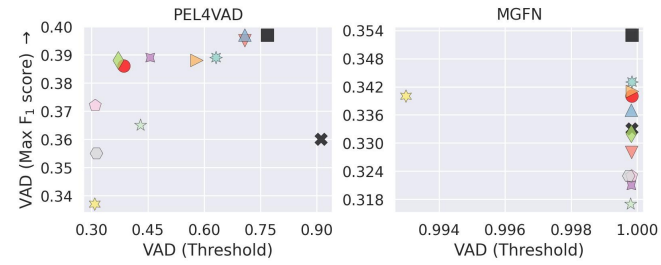
(b)

Fig. 13: Video quality evaluation of the anonymized videos: a) UCF-Crime dataset, and b) XD-violence. The FID and KID scores are computed on the extracted I3D features (with dimension of $B \times D$) of the videos, and the distance is measured from the I3D features of the No-AN videos.

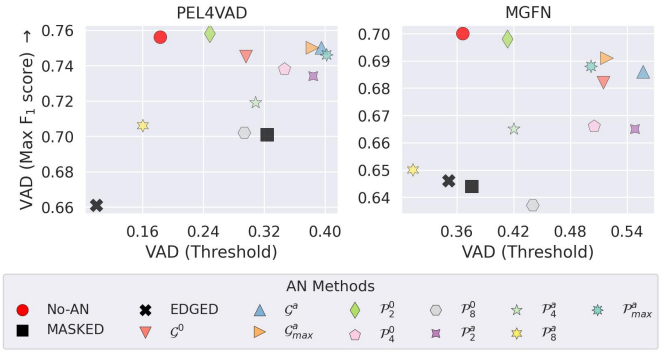
D. Video Feature Processing for Video Anomaly Detection

The VAD models were trained on I3D-RGB extracted features of the raw videos of the UCF-Crime dataset with a frame size of $[320, 240]$. The I3D video encoder operates on an input video frame size of $[340, 256]$ and generates features using a clip sequence length $V = 16$, and crop augmentations $B = 10$ with a crop size of $[224, 224]$. The PEL4VAD utilizes feature dimensions of $D = 1024$ using I3D encoder from Ref. [43], whereas MGFN adopts $D = 2049$ using non-local I3D from Ref. [44].

Benchmark DL AN studies, such as TED-SPAD [6], employ frame normalization for the MGFN that scales into the range of $[0, 1]$, where the value is divided by 255. However, we found that their reported VAD performance (AUC $= 0.78$) is considerably lower than the reported scores by the original MGFN study in Ref. [33] (AUC $= 0.86$ with $V = 32$). Following the implementation of Ref. [44], we have achieved better scores



(a)



(b)

Fig. 14: VAD maximum F_1 scores and their thresholds: a) UCF-Crime dataset, and b) XD-violence. The plots highlight the need to adjust the detection threshold to achieve optimal performance when applying AN to a VAD.

(AUC $= 0.83$) by standardizing the frames as:

$$\bar{\mathbf{I}} = \frac{\mathbf{I} - \mu}{\sigma}, \tag{16}$$

where the $\bar{\mathbf{I}} \in \mathbb{R}$ is the standardized data of the input frame $\mathbf{I} \in \mathbb{Z}$, and the $\mu = 114.75$ and $\sigma = 57.375$ are image scaling mean and standard deviation factors, respectively, that are derived from the Kinetics400 dataset [44], [48].

We normalize the frames into $[-1, 1]$ for PEL4VAD model using:

$$\bar{\mathbf{I}} = \frac{2 \times \mathbf{I}}{255} - 1. \tag{17}$$

For the XD-Violence dataset, the I3D encoder [43] operates on input video frames with a size of $[340, 256]$ normalized into $[-1, 1]$. The encoder utilizes a clip sequence length $V = 16$, and crop augmentations $B = 5$ with a crop size of $[224, 224]$. Both the VAD models employ the same encoded features with a dimension of $D = 1024$, but the MGFN incorporates one more feature through its feature amplification mechanism, which increases its final dimension to $D = 1025$.

E. VAD Confidence Score Evaluation

This subsection presents an investigation of AN and VAD through temporal plots of anomaly confidence scores across video frames. Fig. 14 depicts the best F_1 score along with its decision threshold for each AN method. Some of the AN methods, such as heavier $\mathcal{P}$, exhibit a lower confidence score and require a lower decision threshold when generating VAD flags, whereas others, such as MASKED and $\mathcal{G}$, need much higher thresholds. The plots demonstrate the necessity of adjusting thresholds when applying AN to a VAD for optimal rates of false positives and false negatives.

Figs. 15b and. 15c portray a visual illustration of the VAD confidence scores of the AN methods on sample videos from the UCF-Crime and XD-Violence datasets, respectively. The PEL4VAD has demonstrated promising anomaly localization across the AN methods, while the MGFN struggles with some of the videos. The figures also show that different AN methods have varying impacts on the strength of the anomaly scores. Methods, including the MASK, EDGED, and $\mathcal{G}$, have an increasing effect, whereas the $\mathcal{P}$ can diminish the scores.

F. Limitations and Future Research Directions

Despite the encouraging performance of the proposed approaches, we outline potential limitations and future study considerations below:

- *Optimization of hyperparameters*: The adaptive Θ^{a} initializes with the base Θ^0. This may lead to varying performance depending on the choice of the base parameters of Θ^0, e.g., the varying capacity of $\mathcal{P}^{\text{a}}_{d\in\{2,4,8\}}$ on the AN and VAD. We have found $\mathcal{G}^0_{k\in\{10,\ldots,15\}}$ and $\mathcal{P}^0_4$ provide good base parameters at image size of $\mathbf{z}_{\text{ref}} = [320, 240]$ with the scaling factor of $\alpha_r = \mathbf{z}/\mathbf{z}_{\text{ref}}$ for larger resolutions. Although the $\text{AN}_{\text{max}} = 1$, that enforces the maximum adaptive AN, avoids the dependency on the hyperparameters, it may deteriorate the AN for $\mathcal{P}^{\text{a}}_{\text{max}}$ and increases the computational cost for $\mathcal{G}^{\text{a}}_{\text{max}}$. Thus, we recommend further study in auto-optimization of base parameters to enhance performance.
- *Target depth vs. surface area*: We have considered the segmentation mask area as a rough approximation of depth, i.e., entities at shallow depth will have higher areas and vice versa. Although this performs in most cases, occluded entities violate this assumption. Depth estimation models offer more accurate assessments of object depth in images, thereby enhancing strategies for addressing this challenge. Another potential solution is to employ dense pose estimation, a method that helps segment human body parts and has been utilized for generating synthetic images [55]. This technique can also assist in detecting occlusions by identifying missing body parts. However, the implementation of these DL approaches in computationally limited and real-time settings is often constrained due to their slow processing speeds and limited accuracy [20] (see Fig. 12). We have tested lightweight depth estimation models, such as LITE-MONO models [56], and found their accuracy inadequate for depth estimation of human subjects; they are excessively sensitive to human skin and clothing. Our study has attempted to address the challenge using damped scaling for decent occlusions (see Eq. (10)). We have also introduced $\text{AN}_{\text{max}} = 1$, which ensures the maximum AN with relatively significant computational leverage. But, it would result in non-uniform AN for targets at the same depth with varying areas due to occlusion. For future research work on improved adaptive AN methods, we recommend integrating enhanced lightweight depth estimation with the surface area of target subjects. For instance, the adaptive scaling factor r in Eq. (10) can further be weighted with a depth score.
- *Miss-detections*: We have employed an image-level segmentation, via the YOLO medium model, as a high-accuracy human detector. But, some targets can still be missed due to lower image quality, such as occlusion, motion, lighting, weather, and image resolution. Object segmentation study is a separate domain, and further analysis on it is beyond the scope of the AN research. Nevertheless, we recommend using larger YOLO models with higher accuracy, employing data augmentation, or utilizing video-level detectors with tracking capability to fill the gap and improve detection with additional overhead [57]–[59].

(a) UCF-Crime dataset: Burglary033_x264.mp4

(b) UCF-Crime dataset: Shooting022_x264.mp4

(c) XD-Violence dataset: Fast.Five.2011__#00-32-56_00-33-26_label_B2-0-0.mp4

(d) XD-Violence dataset: Deadpool.2016__#0-18-58_0-19-20_label_B1-0-0.mp4

Fig. 15: VAD anomaly score illustration on sample videos: (top) confidence score of the PEL4VAD [32], (middle) confidence score of the MGFN [33], and (bottom) video frame sampled at the red dot markers of the score plots). The PEL4VAD has provided better anomaly localization across the AN methods, whereas the MGFN has struggled to localize in (b). Both VAD models have achieved good anomaly localization in (c) and (d), with PEL4VAD offering superior separation of the anomaly regions. The EDGED AN limits the performance in (d) for both models.