Data generation via diffusion models for crowd anomaly detection

Giulia Orru
giulia.orru@unica.it
Riccardo Lecca
r.lecca11@studenti.unica.it
Riccardo Puddu
ricpuric@gmail.com
Simone Maurizio La Cava
simonem.lac@unica.it
Marco Micheletto
marco.micheletto@unica.it
Gian Luca Marcialis
marcialis@unica.it

Department of Electrical and Electronic Engineering, University of Cagliari, Italy

Abstract

Crowd analysis is a critical aspect of public security and video surveillance. One of the primary challenges in developing effective crowd anomaly detectors is the lack of comprehensive training data. To address this issue, we investigate using synthetic data to enhance training for anomaly detection in crowded environments by generating a dataset of synthetic videos using two open-source diffusion models. Each synthetic video depicts typical crowded scenes that may be either normal or anomalous. To assess the effectiveness of our approach, we compare the model's performance across three training scenarios: using only real videos, only synthetic videos, and a combination of both. This preliminary analysis highlights the potential of data generated via diffusion models to improve crowd anomaly detectors' stability and classification capabilities.

1 Introduction

Crowd analysis is the discipline that studies the behavior of crowds. The constant growth of the population and the increase in urbanization at a global level have made the formation of groups and crowds increasingly frequent, with ever larger dimensions. To make the management of the numerous video surveillance cameras present in cities effective, numerous technical solutions have been proposed, in particular in the field of detection of anomalous events and automatic alarm systems [6]. The aim of anomalous event detection is to spot anomalous human behavior and situations that can endanger public safety [14].

However, the automatic detection of anomalies in crowded environments is a challenging task due to the variability in the types of anomalies and the characteristics of the data,

including environmental factors [23]. Consequently, crowd anomaly detection often relies on deep learning, specifically deep neural networks, which can automatically learn complex patterns from data [20]. These models can classify videos as anomalous or not based on the content they represent. Despite their effectiveness, training these networks requires a large dataset, a significant challenge in crowd analysis due to the abovementioned variability. Obtaining a sufficient number of labeled videos with the necessary characteristics for training is complex. To address the scarcity of real data in crowd analysis, synthetic data can be utilized. This approach has been successfully employed in other biometric-related application scenarios, such as face recognition, also aiding in addressing privacy concerns [24].

However, generating synthetic data for anomalous crowd behavior detection remains challenging, with available datasets often lacking realism [4]. This limitation reduces the effectiveness of systems developed using such data. Recent advances in generative deep learning offer potential solutions to this issue by enabling the creation of realistic synthetic crowded environments. Specifically, to our knowledge, the use of generative adversarial networks (GANs) and diffusion models for enhancing anomaly detection in this context has not yet been thoroughly investigated.

In this work, we explore the potential use of synthetic data generated via open-source diffusion models for crowd anomaly detection to mitigate the issue of data scarcity in specific application scenarios. To evaluate the effectiveness of our approach, we compare the performance of a state-of-the-art (SOTA) anomaly detector across three training scenarios: using only real videos, only synthetic videos, and a combination of both. Despite the complexity and variability of crowd behavior, we hypothesize that training neural networks on high-quality synthetic data can provide a solid foundation for effective anomaly detection. To validate our hypothesis, our main contributions to the SOTA are as follows: (i) the analysis of several publicly available generative diffusion models (GDMs) to determine their suitability for designing a crowd anomaly detector; (ii) the creation of a novel synthetic dataset generated through GDMs, containing videos depicting realistic crowds in both normal and anomalous situations; (iii) the evaluation of the proposed approach's effectiveness by comparing the performance of a SOTA anomaly detector across three training scenarios: using only real videos, only synthetic videos, and a combination of both.

The rest of the paper is structured as follows. Section 2 provides an overview of the SOTA in crowd anomaly detection, as well as the generation of synthetic data and its use for crowd analysis. Section 3 describes the approach we propose concerning the generation of realistic synthetic data and the analysis of their suitability for enhancing crowd anomaly detectors together with the resulting dataset. Experiments for its validation are presented together with the related results in Section 4. Finally, conclusions are drawn in Section 5.

2 Recent works

2.1 Crowd anomaly detection

The research community has proposed various methods for automatically detecting irregularities, anomalies, or generally unrepresentative patterns in crowded environments. These methods can be categorized into those relying on hand-crafted features and those utilizing deep learning features [1]. Hand-crafted feature methods include textural and spatio-temporal features based on Gabor filters [8] and Optical Flow [31], as well as motion information methods that analyze the speed of group aggregation and disintegration [17]. How-

ever, designing effective hand-crafted descriptors is challenging due to the potential variability in anomaly dynamics and environmental factors.
Given these challenges, the research community has increasingly focused on deep classifiers for feature extraction, as they can automatically learn and adapt to complex patterns, improving the detection of anomalies in crowded environments. In this context, a common approach is to exploit dynamic information Long Short-Term Memory (LSTM) networks [22], which can detect anomalies by analyzing the evolution of crowd behavior thanks to its unique memory characteristics. However, LSTM networks require a time-consuming and laborious engineering process [13]. To face this drawback, it is possible to rely on simpler and more convenient 3D Convolutional Neural Networks (3D CNNs) [18]. In particular, 3D CNNs exploit spatio-temporal information by treating video data as a 3D entity, enabling the network to investigate complex dynamics considering the time as the third dimension [32].

2.2 Synthetic data for crowd analysis

In the era of deep learning, synthetic data is primarily used for training purposes, either alone or in combination with real data. This approach aims to increase the training set size, thereby improving performance and preventing overfitting. The growing interest in synthetic data within the biometric field is closely linked to the challenges of collecting large training sets of real data required by CNNs. These challenges include the significant manual annotation effort, the difficulty in gathering representative examples of target scenes or patterns of interest, and, more recently, concerns regarding privacy [4, 18].
In the context of crowd analysis, synthetic data has primarily been used to improve crowd counting [3, 30]. The use of synthetic data for anomaly detection in crowded environments, however, has been less explored. To our knowledge, the only study in this area is by Lin et al. [15], who demonstrated that augmenting the training set of a 3D CNN with synthetic video could enhance its performance. However, when the system was trained exclusively on synthetic data, it proved unreliable. This highlights the need to improve the realism of synthetic data to achieve better reliability.

2.3 Diffusion models for realistic synthetic data generation

Diffusion models were introduced in 2015 in [26]. Inspired by non-equilibrium statistical physics, they consist of systematically and slowly destroying structures in a data distribution through an iterative process of "forward diffusion". Subsequently, a reverse diffusion process restores the structure, producing a highly flexible generative model of the data. This diffusion process is also applicable to image generation [9]. A diffusion-based generative network begins with an initial image containing a specified amount of Gaussian noise. Rather than removing all the noise in a single pass, the network takes an iterative approach: in each step, it estimates and removes only a fraction of the total noise. The process involves calculating the expected amount of noise at each step, subtracting this from the current image to produce a denoised image, which becomes the starting point for the next iteration [27]. This gradual prediction and noise removal process is repeated iteratively.
Artificial intelligence (AI)-based image and video generators have made significant progress, opening new avenues for digital content creation. These systems utilize deep learning, particularly GANs and GDMs, to generate realistic images and videos from textual descriptions.

Cutting-edge text-to-image generation systems demonstrate the capabilities of these models, relying on sophisticated architectures, including hierarchical transformers and large lan-

guage models, to produce images with high resolution and diversity [25]. An example is CogView2 [6], a text-to-image generator using a hierarchical transformer-based approach to generate high-resolution images from textual descriptions. Research into text-to-video generation has also led to significant advancements in creating personalized videos from text inputs [11]. A leading model in this field is Phenaki [28], relying on an encoder-decoder model and a transformer model to generate temporally coherent and diverse videos conditioned on open-domain prompts. Another example is CogVideo [10], a large-scale text-to-video generative model of 9.4 billion parameters leveraging the knowledge acquired during the text-to-image pretraining phase of CogView2 to generate videos from natural language descriptions, allowing the control of the intensity of the changes through a multi-frame-rate hierarchical training strategy.

3 Synthetic data generation for crowd anomaly detection

The use of generative models, particularly GDMs, marks a significant advancement in technologies for simulating crowds. These models enable the detailed and realistic capture of collective behavior in complex scenarios, such as public events, emergency situations, and urban environments. The primary objective of this work is to generate videos that depict realistic crowds under both normal and anomalous situations. These synthetic videos are intended to support the training and enhancement of crowd anomaly detection systems. By providing high-quality datasets, the generated videos can facilitate the development of more robust and accurate models for identifying unusual behaviors in crowded environments.

To achieve the objective of creating realistic crowd videos, several generative models have been analyzed. Each model brings unique strengths or limitations to the synthesis process, which we evaluated to determine their suitability for our purposes. Therefore, we examined the following open-source generators:

Deforum Stable Diffusion (DSD)[1]: Originally for text-to-image generation [21], this model has been adapted to video synthesis. It uses a diffusion process to iteratively refine noisy sequences into high-quality videos.

Zeroscope_v2 576w (ZS): based on ModelScope [29], this model enhances video generation by employing a denoising process and attention maps to improve quality and consistency. It is trained on 9,923 clips and 29,769 tagged frames at 24 frames per second (fps), with a resolution of 576×320.

VideoCrafter2 (VC2) [7]: This advanced model employs a video variational autoencoder and a video latent diffusion model. Additionally, it integrates temporal convolution from ModelScope to further enhance temporal consistency. It generates high-quality videos at a resolution of 1024×576.

AnimateDiff (AD) [7]: An open-source tool that enables the creation of animations by starting with an existing image and iteratively applying text prompts to modify it.

For each generative model, we evaluated a set of prompts to determine their ability to produce realistic-looking videos suitable for analyzing crowd behaviors and detecting anomalies. The prompt design was inspired by the appearance of crowds in SOTA public datasets such as Motion-Emotion (MED) [19] and UCSD Ped2 [16]. The prompts used for generating normal videos are (i) *people walking, cctv view, security camera view*, (ii) *group of people walking on the street, security camera view*, (iii) *top-down perspective of people walking on the*

[1] https://github.com/deforum-art/deforum-stable-diffusion

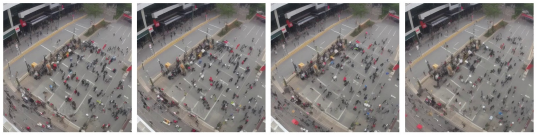

Figure 1: Frames extracted from a video generated using Animatediff with the prompt *"people fighting, cctv view, security camera view"*. From left to right, one frame every 5 is shown, for a total of 15 frames.

Figure 2: Frames extracted from a video generated using Deforum Stable Diffusion with the prompt *"people fighting, security camera view"*. From left to right, one frame every 25 is shown, for a total of 100 frames.

sidewalk, and their variations. The prompts used for generating anomaly videos are (i) *people fighting, cctv view, security camera view* , (ii) *top-down perspective of a group of people fighting on the sidewalk*, (iii) *top-down view of a group of people in a square suddenly dispersing in panic*, and their variations.

The AD synthetic videos consist of sequences featuring perspective variations or changes in the depicted subject. However, the movement of individuals was limited or absent (Fig. 1). Consequently, it was concluded that such content does not meet the requirements for our predefined objectives, making it unsuitable for our purposes.

Despite the capabilities of DSD to generate videos of up to 1000 frames to adjust numerous parameters, it exhibited significant limitations in the realism and temporal stability of the generated sequences, as illustrated in Figure 2. These images highlight the lack of temporal stability, with the perspective shifting dramatically from a security camera-like view to a close-up within just four seconds.

Figures 3 and 4 clearly show that the videos generated with ZS and VC2 exhibit notable temporal stability. The camera shot is static, and there are no abrupt changes in scene, colors, or subjects, as well as inconsistency with the prompt or abrupt motion, which have instead been detected for the other models. For these reasons, ZS and VC2 were selected to generate a synthetic dataset to address the following question: *is it possible to exploit videos generated with GDMs to improve the performance of modern crowd anomaly detector systems?* In fact, these models are particularly advantageous because they are open-source, capable of generating sufficiently realistic videos, and feature a simple generation procedure that can be parallelized through code.

3.1 The proposed synthetic dataset

To maintain the realism of the generated images, we noted that videos should be generated with similar characteristics to those on which the model was trained. For this reason, ZS videos were generated at a resolution of 576×320 pixels, while VC2 videos were produced at 512×320 pixels. Additionally, achieving a balance between the video duration and re-

Figure 3: Frames extracted from a video generated using Zeroscope with the prompt *"people walking, cctv view, security camera view"*. From left to right, one frame every 40 is shown, for a total of 120 frames.

Figure 4: Frames extracted from a video generated using VideoCrafter2 with the prompt *"people walking on the sidewalk, static view from above"*. From left to right, one frame every 5 is shown, for a total of 15 frames.

alism led us to select 32 frames for ZS and 14 frames for VC, both with a frame rate of 8 fps. In fact, exceeding the number of frames per video on which the models were trained significantly compromises the quality of video generation.

For the ZS model, which uses a Denoising Diffusion Implicit Model sampler [27], we set a relatively high number of denoising steps at 35. The "Classifier Free Guidance" parameter was configured with a significantly higher value of 35, as preliminary observations indicated better model alignment with the prompt compared to the standard range of 10 to 17. Due to the low frame rate of 8 fps, we decided to interpolate the videos to 32 fps using linear interpolation. To ensure uniformity in the generated videos, specific prompts were used for each model. For ZS, normal videos were generated using the prompts *"people walking, security camera view"* and *"people moving, cctv view, security camera view"*. Anomalous videos were created with the prompts *"people fighting, security camera view"* and *"people battling, cctv view, security camera view"*.

For VC2, the prompt *"top-down perspective of a group of people walking on the sidewalk"* is used for normal videos. Anomalous videos were generated with the prompts *"top-down perspective of a brawl on the sidewalk"* and *"top-down perspective of a group of people fighting a man on the sidewalk"*. These specific prompts were chosen based on previous analyses, aiming to capture the complexity of human behavior in public contexts. It was also observed that using more specific prompts, such as "people running away from a central point", did not produce videos that adequately reflected the employed prompt. The limited duration of the videos (1.4 to 3.3 seconds) prevents the representation of significant transitions between normal and anomalous situations. Consequently, each video captures a single scenario and is labeled as either "normal" or "abnormal" based on the prompt used.

We generated 240 videos with a frame rate of 32 fps: 60 abnormal and 60 normal videos with ZS (3.3 seconds, 576×320), for a total of 12672 frames, and 60 abnormal and 60 normal with VC2 (1.4 seconds, 512×320), for a total of 5376 frames. The videos are available from the corresponding author, upon request.

4 Experimental analysis

4.1 Experimental protocol

In our experimental analysis, we assessed the impact of synthetic data on the performance of crowd anomaly detectors by employing a 3D CNN, specifically the 3D-ShuffleNet (SNet) network [12], which was pre-trained on the Kinetics-600 dataset for human action recognition. We further fine-tuned the model for crowd anomaly detection using three distinct approaches: real data, synthetic data, and a combination of both.

For real data, we utilized the Motion Emotion Dataset (MED), which comprehensively captures various crowd behaviors, both normal and anomalous. The MED comprises 31 video sequences, totaling approximately 44000 frames (554×235). These videos were recorded at 30 fps using a fixed overhead camera to monitor individual paths. The dataset is annotated with labels for behaviors such as panic, fight, congestion, obstacle, and neutral, following the labeling protocol established by [17]. Hence, we used non-overlapping groups of 20 frames, treating each batch as anomalous if one or more frames contained any anomaly, and as normal otherwise. Considering that the 3D ShuffleNet network (learning rate 0.01, num. epochs 60) accepts input groups of 16 frames, from each 20-frame batch, we selected 16 frames starting from a random one within the batch.

To assess the model's performance, we employed a k-fold cross-validation with $k = 31$ for each fine-tuning process. For real data, this corresponded to a leave-one-out setup. For synthetic data, our primary focus was to evaluate the overall impact of synthetic videos without introducing bias toward any specific synthetic video generation method. Therefore, we ensured each fold contained an equal portion of anomalous and normal videos, randomly selecting the source models to maintain diversity. When training with both real and synthetic data, we faced time constraints that made it impractical to explore all possible combinations. Instead, we ensured a balanced and representative training set by randomly pairing each fold of real data with a random fold of synthetic data.

4.2 Results

The goal of this investigation is to assess the effectiveness of incorporating additional synthetic data during the training stage to enhance the performance of crowd anomaly detectors. From Figure 5, it is possible to observe that the model trained on real data performs acceptably on real videos but slightly worse on synthetic videos. Conversely, the model trained on synthetic data is very effective on synthetic videos but unreliable on real ones.

More interesting is the outcome of the model trained on both real and synthetic data, which demonstrates reliable performance on both types of data. In particular, this model shows a slight enhancement in performance in both scenarios compared to the one employing only real data for fine-tuning ($AUC = 0.79$ on real data and $AUC = 0.98$ on synthetic data).

The improved robustness on real data is also confirmed by other performance metrics in Table 4.2, showing an appreciable increase in precision (from 64.29% to 69.13%) and in accuracy (from 72.38% to 75.03%). The reasons behind such improvements are mainly related to the reduced number of incorrectly classified normal samples. Still, it is interesting to observe that, despite the overall better performance, introducing synthetic data in the training phase leads to a higher frequency of unrecognized anomalous samples, resulting in higher FRR (from 30.75% to 38.89%).

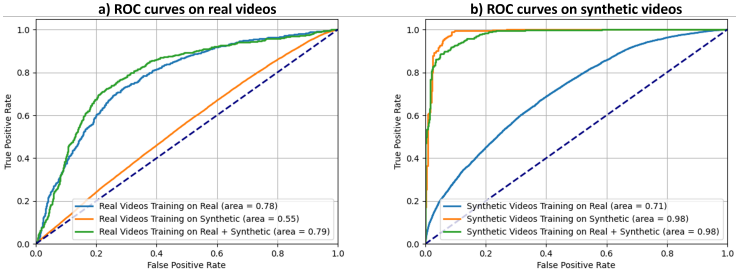

Figure 5: 3D-CNN ROC curves on the real (a) and synthetic (b) test set with the three different fine-tuning: real samples (blue), synthetic samples (orange), and both (green).

Training samples	Precision [%]	Accuracy [%]	Recall [%]	F1-score [%]	FAR [%]	FRR [%]
Real	64.29	72.38	60.24	62.20	20.27	**30.75**
Synthetic	39.89	53.72	44.87	42.23	40.91	55.13
Real+Synthetic	**69.13**	**75.03**	**61.11**	**64.87**	**16.53**	38.89

Table 1: 3D-CNN results on the real test set with the three different fine-tuning sets.

Despite being less interesting for real-world application scenarios, we also included performance related to tests on synthetic data in Table 4.2 for completeness.

While visualizing the system's overall performance is useful, it is equally important to examine its behavior at a finer granularity, particularly at the frame level. This section details how classification changes frame by frame, highlighting differences in anomaly prediction, robustness, and stability among the differently fine-tuned models using anomalous sequences from the MED dataset.

Classification of anomalous sequences First, we analyze a sequence from video number 10 (Fig. 6). The sequence from the 1,800th frame to the end, spanning 86 seconds, is labeled as anomalous. Specifically, this sequence depicts a scene where a firework is lit in the crowd. Initially, there is a slight cluster of curiosity around the firework, but the crowd begins dispersing in two waves, with the second wave occurring after frame 2,250 being more pronounced. Consequently, the classification value should exceed 0.5. It is evident that the model trained solely on synthetic data performs the worst, approaching the 0.5 threshold only once, thus misclassifying the scene as normal. The model trained with real data shows a more appreciable performance, recognizing some scenes as anomalies. Interestingly, there is a hint of response around frame 2,050, where the crowd starts dispersing, but it does not lead to an anomaly classification. A clear response is seen only in frame 2,250, when the scene is almost empty, with the firework in the middle. In contrast, the model trained with both synthetic and real data exhibits the best performance among the three. It highlights responses at all critical points: the ignition of the firework (around frame 1,800), the first dispersion wave (between frames 2,000 and 2,100), and the final dispersion wave after frame 2,250.

Classification of normal sequences To analyze a normal case, we selected frames 300 to 800 from video 20 of the MED dataset, covering 13 seconds (Fig. 7). The sequence depicts various groups of people interacting with one another. This scenario clearly demonstrates the unsuitability of the model trained only on synthetic samples for real videos, resulting in predominantly incorrect classifications. The analysis becomes more interesting when comparing the model fine-tuned on real samples with the one fine-tuned on both real and synthetic samples. Several peaks nearing the 0.5 threshold appear before frame 300 (8-9 seconds) and after frame 400 (14 seconds), leading the models to classify these frames as

Training samples	Precision [%]	Accuracy [%]	Recall [%]	F1-score [%]	FAR [%]	FRR [%]
Real	65.85	63.50	53.11	58.80	26.49	46.88
Synthetic	**96.14**	**93.25**	**89.83**	**92.88**	**3.46**	**10.16**
Real+Synthetic	94.62	91.84	88.44	91.42	4.86	11.58

Table 2: 3D-CNN results on the synthetic test set with the three different fine-tuning sets.

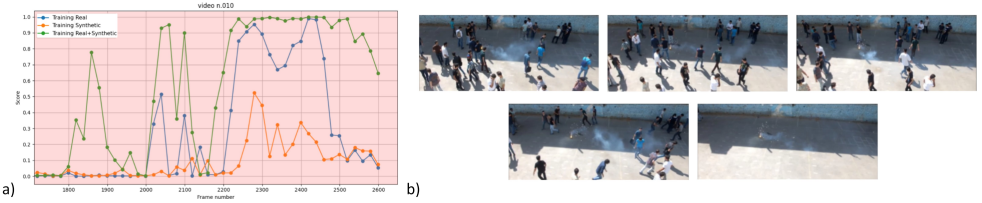

Figure 6: 3D-CNN classification on real video n.10 after fine-tuning using real (blue), synthetic (orange), and both (green) videos, and example frames 1800, 1950, 2010, 2250, and 26000.

anomalous. These peaks are caused by physical contact or people being very close to each other. For instance, just before frame 300 and near frame 400, a group of people in the foreground is seen hugging and patting each other as a greeting. These behaviors trigger the anomaly classification. After these peaks, the frames from 450 to 750 show that the model trained on real samples alone provides entirely incorrect interpretations, resulting in false alarms. In conclusion, this scenario demonstrates that the model trained on both real and synthetic samples offers better classification performance, although it still gets confused by some particular situations.

Although these results have been confirmed by the analysis of other videos, not reported for reasons of space, it is important to highlight that sometimes the model trained on both real and synthetic data detects the anomaly only when it is evident. In this cases, the model trained on real data has more sensibility and is more balanced. This demonstrates the need for synthetic videos that are as realistic as possible for the purpose of training or fine-tuning crowd anomaly detectors.

5 Conclusions

This work aims to analyze the effectiveness of using synthetic data in the design and training phase of crowd anomaly detectors, thus facing the challenges deriving from using real data, such as the lack of large-scale datasets, computational complexity in labeling, and difficulty finding diverse and particular scenarios. We analyzed various text-to-video generation methods to evaluate which ones suit the purpose. The analysis highlighted the limits of these models, often linked to the lack of relationship between textual prompt and output, temporal inconsistency, and the presence of artifacts.

Based on these results, we proposed a novel synthetic dataset generated through two diffusion models, Zeroscope_v2 576w and VideoCrafter2, containing scenes of crowds both with anomalies, such as fights and brawls, and with no anomalies. We used this synthetic dataset alongside the SOTA MED dataset to evaluate if joint training with synthetic videos yields a more robust model than training with real data alone.

Our findings revealed that the model trained on both real and synthetic data outperformed those trained only on real data. Specifically, the model trained only on real data exhibited

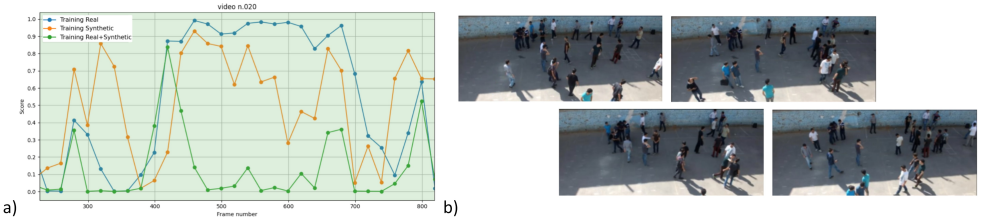

Figure 7: 3D-CNN classification on real video n.20 after fine-tuning using real (blue), synthetic (orange), and both (green) videos, and example frames 50, 240, 450, and 720.

high sensitivity to scene changes, leading to abrupt score changes and potential misclassifications. Including synthetic data in the training set demonstrated greater stability during transitions. This highlights the potential of synthetic data when coupled with real data.

In the future, it is expected that novel synthetic data generation models, specialized in anomalous events in crowded environments, will overcome current limitations, thus starting a significant step ahead in the problem of training an anomaly detection system properly.

Acknowledgment

This work is supported by the European Union – Next Generation EU through the Italian Ministry of University and Research (MUR) within the National Recovery and Resilience Plan (PNRR) - Mission 4 "Education and Research" - C2/1.1 - Fund for Relevant Projects of National Interests (PRIN) - "BullyBuster 2 - the ongoing fight against bullying and cyberbullying with the help of artificial intelligence for the human wellbeing" (Proj. Code: P2022K39K8 - CUP: F53D23009240001).

References

[1] Amnah Aldayri and Waleed Albattah. Taxonomy of anomaly detection techniques in crowd scenes. *Sensors*, 22(16):6080, 2022.

[2] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. URL https://arxiv.org/abs/2401.09047.

[3] Rita Delussu, Lorenzo Putzu, and Giorgio Fumera. Scene-specific crowd counting using synthetic training images. *Pattern Recognition*, 124:108484, 2022.

[4] Rita Delussu, Lorenzo Putzu, and Giorgio Fumera. Synthetic data for video surveillance applications of computer vision: A review. *International Journal of Computer Vision*, pages 1–37, 2024.

[5] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022.

[6] Omar Elharrouss, Noor Almaadeed, and Somaya Al-Maadeed. A review of video surveillance systems. *Journal of Visual Communication and Image Representation*, 77:103116, 2021.

[7] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024.

[8] Yu Hao, Zhi-Jie Xu, Ying Liu, Jing Wang, and Jiu-Lun Fan. Effective crowd anomaly detection through spatio-temporal texture analysis. *International Journal of Automation and Computing*, 16(1):27–39, 2019.

[9] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(1), jan 2022. ISSN 1532-4435.

[10] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[11] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18219–18228, 2022.

[12] Okan Kopuklu, Neslihan Kose, Ahmet Gunduz, and Gerhard Rigoll. Resource efficient 3d convolutional neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[13] Chengyang Li, Liping Zhu, Dandan Zhu, Jiale Chen, Zhanghui Pan, Xue Li, and Bing Wang. End-to-end multiplayer violence detection based on deep 3d cnn. In *Proceedings of the 2018 VII international conference on network, communication and computing*, pages 227–230, 2018.

[14] 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, 2013.

[15] Wei Lin, Junyu Gao, Qi Wang, and Xuelong Li. Learning to detect anomaly events in crowd scenes from synthetic data. *Neurocomputing*, 436:248–259, 2021.

[16] 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, 2010. doi: 10.1109/CVPR.2010.5539872.

[17] Giulia Orrù, Davide Ghiani, Maura Pintor, Gian Luca Marcialis, and Fabio Roli. Detecting anomalies from video-sequences: a novel descriptor. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4642–4649. IEEE, 2021.

[18] Giulia Orrù, Elia Porcedda, Simone Maurizio La Cava, Roberto Casula, and Gian Luca Marcialis. Human-centered evaluation of anomalous events detection in crowded environments. In *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6. IEEE, 2023.

[19] Hamidreza Rabiee, Javad Haddadnia, Hossein Mousavi, Maziyar Kalantarzadeh, Moin Nabi, and Vittorio Murino. Novel dataset for fine-grained abnormal behavior understanding in crowd. In *13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 95–101. IEEE, 2016.

[20] Khosro Rezaee, Sara Mohammad Rezakhani, Mohammad R Khosravi, and Mohammad Kazem Moghimi. A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. *Personal and Ubiquitous Computing*, 28(1):135–151, 2024.

[21] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.

[22] Mohammad Sabih and Dinesh Kumar Vishwakarma. A novel framework for detection of motion and appearance-based anomaly using ensemble learning and lstms. *Expert Systems with Applications*, 192:116394, 2022.

[23] Francisco Luque Sánchez, Isabelle Hupont, Siham Tabik, and Francisco Herrera. Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. *Information Fusion*, 64: 318–335, 2020.

[24] Hatef Otroshi Shahreza et al. Sdfr: Synthetic data for face recognition competition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2024.

[25] Aditi Singh. A survey of ai text-to-image and ai text-to-video generators. In *2023 4th International Conference on Artificial Intelligence, Robotics and Control (AIRC)*, pages 32–36. IEEE, 2023.

[26] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[28] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru

Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vOEXS39nOF.

[29] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[30] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8198–8207, 2019.

[31] Yu Zhao, Yu Qiao, Jie Yang, and Nikola Kasabov. Abnormal activity detection using spatio-temporal feature and laplacian sparse representation. In *Neural Information Processing: 22nd International Conference, ICONIP 2015, November 9-12, 2015, Proceedings, Part IV 22*, pages 410–418. Springer, 2015.

[32] Shifu Zhou, Wei Shen, Dan Zeng, Mei Fang, Yuanwang Wei, and Zhijiang Zhang. Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Processing: Image Communication*, 47:358–368, 2016.