Beyond the Benchmark: Detecting Diverse Anomalies in Videos

Yoav Arad
The Hebrew University of Jerusalem, Israel
yoav.arad@mail.huji.ac.il

Michael Werman
The Hebrew University of Jerusalem, Israel
michael.werman@mail.huji.ac.il

ABSTRACT

Video Anomaly Detection (VAD) plays a crucial role in modern surveillance systems, aiming to identify various anomalies in real-world situations. However, current benchmark datasets predominantly emphasize simple, single-frame anomalies such as novel object detection. This narrow focus restricts the advancement of VAD models. In this research, we advocate for an expansion of VAD investigations to encompass intricate anomalies that extend beyond conventional benchmark boundaries. To facilitate this, we introduce two datasets, HMDB-AD and HMDB-Violence, to challenge models with diverse action-based anomalies. These datasets are derived from the HMDB51 action recognition dataset. We further present Multi-Frame Anomaly Detection (MFAD), a novel method built upon the AI-VAD framework. AI-VAD utilizes single-frame features such as pose estimation and deep image encoding, and two-frame features such as object velocity. They then apply a density estimation algorithm to compute anomaly scores. To address complex multi-frame anomalies, we add deep video encoded features capturing long-range temporal dependencies, and logistic regression to enhance final score calculation. Experimental results confirm our assumptions, highlighting existing models limitations with new anomaly types. MFAD excels in both simple and complex anomaly detection scenarios.

Keywords

Video Anomaly Detection, Computer Vision, Smart Surveillance Systems

1 INTRODUCTION

As the volume of recorded video content continues to grow, the need for robust and efficient video anomaly detection methods increases. The ability to automatically identify unusual events or behaviors within videos not only holds the promise of enhancing security but also offers the potential to reduce the manpower required for monitoring. However, achieving truly effective video anomaly detection remains a significant unsolved challenge, due to the diverse range of anomalies that can occur in real-world scenarios.

By nature, anomalous behaviors are rare. Thus, video anomaly detection (VAD) is often treated as a semi-supervised problem, where models are trained exclusively on normal videos and must subsequently distinguish between normal and abnormal videos during inference.

While current benchmark datasets vary in complexity, they share a common limitation in their narrow definition of anomalies. The three datasets, UCSD Ped2 [1], CUHK Avenue [2], and ShanghaiTech Campus [3], tend to limit anomalies primarily to novel object detection (Ped2, ShanghaiTech) or simple movement anomalies (Avenue).

Recent advancements in video anomaly detection predominantly relied on analyzing a few frames or even individual frames in isolation. Researchers predominantly choose between two approaches: reconstruction-based and prediction-based methods. Reconstruction-based methods [4–8] typically employ auto-encoders to learn representations of normal frames, reconstructing them accurately, while anomalous frames result in a higher reconstruction error. Prediction-based methods [3, 9–11] focus on predicting the next frame from a sequence of consecutive frames.

These few-frame based methods achieved impressive results, surpassing an AUC score of 99% [12, 13] on Ped2, over 93% [12] on Avenue, and exceeding 85% [12, 14] on ShanghaiTech, the most complex of the benchmark datasets.

Without a shift in research focus and assumptions, the existing datasets, results, and recurring research patterns may suggest that the field of video anomaly detection is nearing a plateau.

This paper emphasizes the necessity of broadening the scope of what constitutes an anomaly. We propose two novel datasets specifically designed to assess the detection of complex action-based anomalies.

https://www.doi.org/10.24132/CSRN.3401.5

These datasets, referred to as HMDB-AD and HMDB-Violence, build upon the HMDB51 action recognition dataset and define different actions as normal or abnormal activities. By analyzing the performance of various methods on our datasets, we underscore the limitations of existing approaches and advocate for further research on more comprehensive anomaly types.

Building upon the foundation laid by AI-VAD [12], we introduce Multi-Frame Anomaly Detection (MFAD), a novel method aimed at achieving balanced performance, excelling in both simple and complex anomaly detection. AI-VAD utilizes a two-step approach: first, it extracts multiple features and then employs density estimation algorithms to calculate anomaly scores. In their work, they rely on single-frame features like deep image encoding (using a pretrained encoder) and human pose estimations, along with two-frame features such as object velocity. We extend this method by introducing deep video encoding features to capture multi-frame, long-range temporal relationships. MFAD adheres to the AI-VAD framework, computing final scores for each feature using a density estimation algorithm. Additionally, we incorporate logistic regression to enhance the relationships between different feature scores and achieve more accurate final scores.

We extensively evaluate our method on classic benchmark datasets as well as on our newly proposed datasets. The experiments validate the added value of both video encoding features and the logistic regression module. Our method achieves competitive results on Ped2, Avenue, and ShanghaiTech, and greatly outperforms recent methods on HMDB-AD and HMDB-Violence. As a result, it offers a more versatile video anomaly detection solution capable of detecting a broader range of anomalies across various scenarios.

Our key contributions are:

- We highlight the limitations of current video anomaly detection benchmarks and advocate for further research in general video anomaly detection.
- We present MFAD, a novel method capable of effectively handling both simple, few-frame anomalies and complex, multi-frame anomalies.
- We provide two datasets designed for assessing a model's performance on multi-frame action-based anomalies.

2 RELATED WORK

2.1 Video Anomaly Detection Datasets

The datasets commonly used in video anomaly detection can be broadly categorized into two groups, reflecting the shift brought about by the advent of deep learning from approximately 2013 to 2018.

Early datasets are notably smaller and often considered practically solved, include Subway Entrance [15], Subway Exit [15], UMN [16], UCSD Ped1 [1], UCSD Ped2 [1], and CUHK Avenue [2]. Except UMN, these datasets feature only a single scene.

In contrast, more recent datasets have grown significantly in both scale and complexity. This newer group includes ShanghaiTech Campus [3], Street Scene [17], IITB Corridor [18], UBNormal [19], and the most recent and largest of them all, NWPU Campus [20].

It is worth noting that among these datasets, only three have gained popularity as benchmarks: UCSD Ped2, CUHK Avenue, and ShanghaiTech Campus. However, as discussed in this paper, each of these benchmarks has its own set of limitations that motivate the need for further research in the field of video anomaly detection.

Other datasets that can be considered are UCF-Crime [21] and XD-Violence [22]. These datasets are built for fully supervised VAD learning and therefore are orders of magnitude larger than current benchmarks for unsupervised VAD such as this work. We follow previous studies and don't use them for our comparisons.

2.2 HMDB51 Action Recognition Dataset

The HMDB51 [23] dataset, originally designed for action recognition (AR), is relatively small in scale. It is a collection of 6,766 video clips distributed across 51 distinct categories. Most other datasets are significantly larger and more diverse: SSv2 [24], Kinetics-400 [25], Kinetics-600 [26], Kinetics-700-2020 [27] each consist of hundreds of thousands of frames and hundreds of different classes.

The HMDB51 dataset draws content from various sources, ensuring diversity. In this dataset, each class consists of no less than 101 video clips.

2.3 Video Anomaly Detection Methods

Hand-crafted feature based methods

Numerous methods, spanning both classical and contemporary approaches, adhere to a two-stage anomaly detection framework. This framework involves an initial step of extracting hand-crafted features, specifically selected by the researcher and not learned through a deep neural network model. Subsequently, another algorithm is applied to compute anomaly scores.

Early techniques used classic image and video features, including the histogram of oriented optical flow (HOF) [28–31], histogram of oriented gradients (HOG) [31], and SIFT descriptors [32]. In more recent developments, the proliferation of deep learning has facilitated the adoption of off-the-shelf models, such as object detectors, for feature extraction. For instance, in the case of AI-VAD [12], a combination of pose estimations, optical flow predictions, object detection, and deep image encodings is used to construct robust feature representations.

Following feature extraction, classical methodologies often employed scoring techniques such as density estimation algorithms [33–35]. Recent approaches have demonstrated the effectiveness of integrating these features with another learning model [13].

Reconstruction and prediction based methods

In recent years, the increasing prominence of deep learning has driven the widespread adoption of both reconstruction and prediction based methods in video anomaly detection.

Reconstruction-based [4–8] approaches often utilize auto-encoders to learn representations of normal video frames and subsequently detect abnormal frames by identifying higher reconstruction errors. However, the powerful generalization ability of modern auto-encoders can often also reconstruct anomalies. Thus, making it harder to differentiate normal and abnormal frames.

Prediction-based [3, 9–11] models forecast the subsequent frame by leveraging a sequence of preceding frames, employing time sensitive architectures such as LSTMs, memory networks, 3D auto-encoders and transformers. This predictive approach often yields superior results compared to similar reconstruction-based techniques [11], as it captures more complex forms of anomalies. Nevertheless, with the minimal differences between consecutive video frames, these methods face similar challenges to reconstruction-based approaches with respect to modern generators.

Auxiliary tasks methods

Expanding beyond reconstruction and prediction, some models incorporate diverse self-supervised auxiliary tasks, with task success determining frame anomaly scores. These tasks include jigsaw puzzles [36], time direction detection [37], rotation prediction [38], and more. SSMTL++ [14, 39] train a single deep backbone on multiple self-supervised tasks and achieve state-of-the-art results on the benchmark datasets.

3 PROPOSED DATASETS

We introduce two novel datasets designed to assess the capability of various models in detecting forms of anomalies not covered by existing benchmarks. These datasets emphasize action-based anomalies, a category absent in current benchmarks. The first dataset, referred to as HMDB-AD, aligns with the conventional definition of normal activities (walking and running) but challenges models with abnormal behaviors that demand a broader context for detection (climbing and performing a cartwheel). In contrast, the larger and more intricate HMDB-Violence dataset divides 16 action categories into 7 violent (abnormal) and 9 non-violent (normal) activities. This categorization necessitates models to consider a wide range of behaviors when classifying events as either normal or abnormal, making it a closer representation of real-world scenarios.

HMDB-AD dataset

HMDB-AD is the simpler dataset among the two introduced in this paper. It consists of 995 video clips, divided into 680 training videos and 315 testing videos. Normal activities within this dataset are running and walking, aligning with their respective HMDB51 classes. Abnormal activities are climbing and performing a cartwheel. The training dataset contains only of normal videos: 207 running videos and 473 walking videos. Meanwhile, the test dataset has both abnormal videos and randomly selected normal videos; 107 cartwheel videos, 108 climbing videos, 25 running videos, and 75 walking videos. Frames from the videos can be viewed in Appendix A.1.

HMDB-Violence dataset

HMDB-Violence is the larger and more complex of the two datasets presented in this paper. It has 2,566 videos, with a distribution of 1,601 training videos and 965 testing videos. The train set has nine normal categories: running (221 videos), walking (517), waving (98), climbing (104), hugging (110), throwing (96), sitting (134), turning (222), and performing a cartwheel (99). In the test set, there are seven abnormal categories: falling (136), fencing (116), hitting (127), punching (126), using a sword (127), shooting (103), and kicking (130). Additionally, the test set includes 100 videos randomly sampled from the various normal categories: turning (18), walking (31), running (11), sitting (8), hugging (8), performing a cartwheel (8), climbing (4), throwing (6), and waving (6). The abnormal activities in HMDB-Violence are characterized by their violent nature. Examples can be viewed in Appendix A.2.

Annotations

We maintain a consistent labeling for every frame within a video. If a video represents a normal action category, all its frames are labeled as normal. Conversely, if it belongs to an abnormal action category, all frames are marked as abnormal. This simple labeling approach works, as the actions within these videos effectively occupy the entire duration, leaving minimal room for unrelated "spare" frames.

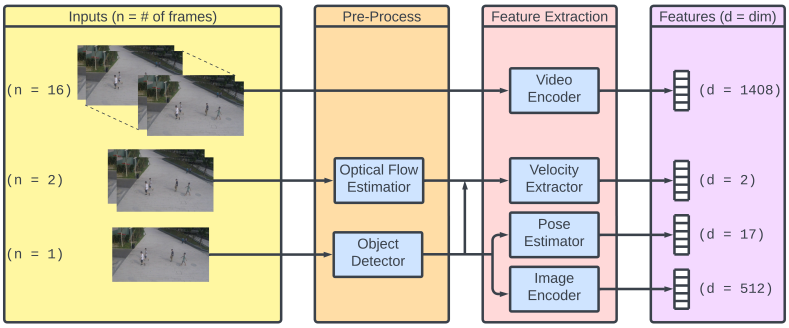

Figure 1: An overview of our feature extraction process.

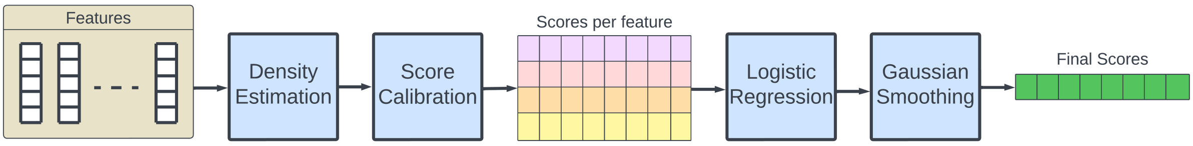

Figure 2: An overview of our anomaly score calculation during inference.

4 MFAD: MULTI-FRAME ANOMALY DETECTION

Our method, MFAD, consists of three key stages: feature extraction, per-feature score computation, and logistic regression. We extract four types of features: object velocities, human pose estimations, deep image encodings, and deep video encodings. For each of these features, we independently calculate density scores. We then employ a logistic regression model to optimally fuse the scores across these four feature kinds. Lastly, we smooth, Gaussian, to produce the final anomaly scores. An overview of our method can be found in Fig. 1, Fig. 2.

4.1 Feature Extraction

Few-Frame Features

In line with AI-VAD [12], HF2-VAD [13], we extract object bounding boxes and optical flows from each frame. We then extract human pose estimations, object velocities, and deep image encodings. These features are derived from individual frames (pose and image encoding) or pairs of frames (velocity) enabling the detection of straightforward anomalies such as novel objects.

Multi-Frame Features

Recognizing the necessity for detecting complex anomalies that span multiple frames, we introduce a deep video encoder. This encoder captures features in a manner similar to deep image encoding but takes into account longer frame sequences (in our case, 16 frames). For this purpose, we leverage VideoMAEv2 [46], a state of the art video foundation model. Subsequently, we process these features in a fashion similar to AI-VAD [12].

4.2 Density Score Calculation

We employ a Gaussian Mixture Model (GMM) for the two-dimensional velocity features and the k-nearest neighbors (kNN) algorithm for the high-dimensional pose, image encoding, and video encoding features. Subsequently, we compute the minimum and maximum density scores for the training set and use them to calibrate the test scores during inference.

Max Feature

We add a fifth feature, denoted as max. After calculating the density scores per feature, we aggregate them into a new feature that holds the maximum feature score per frame.

$$\text{max} = \max\{\text{P, V, IE, VE}\} \in [0,1]^{\#frames}$$

Our experiments show the added value of this feature.

4.3 Logistic Regression

To improve the accuracy of our final anomaly score computation, we incorporate logistic regression as the final step of our method. In this setup, we denote $X \in [0,1]^{\#frames \times \#features}$ as the feature matrix and $y \in \{0,1\}^{\#frames}$ as our ground truth labels. Our final scores are:

$$h_\theta(X) = \sigma(WX + B)$$

Method	Ped2	Avenue	ShanghaiTech	HMDB-AD	HMDB-Violence
HF2-VAD [13]	**99.3%**	91.1%	76.2%	–	–
AED [40]	98.7%	92.3%	82.7%	–	–
HSC-VAD [41]	98.1%	**93.7%**	83.4%	–	–
DLAN-AC [42]	97.6%	89.9%	74.7%	–	–
SSMTL [39]	97.5%	91.5%	82.4%	–	–
LBR-SPR [43]	97.2%	90.7%	72.6%	–	–
AMMCNet [44]	96.6%	86.6%	73.7%	–	–
AI-VAD [12]	99.1%	93.3%	**85.9%**	*70.1%*	*70.5%*
Jigsaw Puzzles [36]	99.0%	92.2%	84.3%	53.8%	52.7%
MNAD [11]	97.0%	88.5%	70.5%	56.3%	51.3%
MPN [45]	96.9%	89.5%	73.8%	58.8%	53.7%
MFAD (Ours)	*99.0% ± 0.5%*	*92.9% ± 0.5%*	*84.8% ± 0.4%*	**90.0% ± 0.4%**	**80.2% ± 0.2%**
MFAD w/o IE (Ours)	98.4% ± 0.7%	90.7% ± 0.5%	85.0% ± 0.4%	86.9% ± 0.5%	76.0% ± 0.2%

Table 1: Comparison to frame-level AUC. Best (bold), second (underlined), and third (italic). IE, denotes image encoding features.

where $\sigma(t) = \frac{1}{1+e^{-t}}$ is the sigmoid function and $\theta = (W, B)$ are the parameters we want to optimize. Our loss function is:

$$L(h_\theta(X), y) = -y\log(h_\theta(X)) - (1-y)\log(1-h_\theta(X))$$

During its training phase, we randomly sample a small fraction of the test frames for model training, while the remainder is used for evaluation. It is crucial to emphasize that the frames utilized for training are excluded from the evaluation process for our reported results, ensuring the validity of our findings.

The final step in our method is applying Gaussian smoothing to the anomaly scores.

5 EXPERIMENTS

5.1 Datasets

In addition to HMDB-AD and HMDB-Violence, we evaluate MFAD on the three benchmark video anomaly detection datasets: UCSD Ped2, CUHK Avenue, and ShanghaiTech Campus. These datasets are primarily outdoor surveillance camera footage, with the sole normal activity being pedestrian movement.

UCSD Ped2

The UCSD Ped2 dataset has 16 training videos and 12 testing videos, all situated within a single scene. Abnormal events in this dataset include the appearance of skateboards, bicycles, or cars within the video frame. Videos are standardized to a resolution of 240 × 360 pixels.

CUHK Avenue

The CUHK Avenue dataset has 16 training videos and 21 testing videos, all within a single scene. Anomalies within this dataset are activities such as running, throwing objects, and bike riding. All videos have a resolution of 360 × 640 pixels.

ShanghaiTech Campus

ShanghaiTech Campus stands as the largest and most complex dataset among the three, featuring 330 training videos and 107 testing videos distributed across 13 distinct scenes. Notably, two of these scenes involve non-stationary cameras, resulting in varying angles between videos of the same scene. Abnormal events primarily include running and the presence of cars and bikes. All videos have a resolution of 480 × 856 pixels.

5.2 Implementation details

We adopt the code from AI-VAD for extracting velocity, pose, and deep image encoding features. For our new deep video features, we leverage the state-of-the-art video foundation model, VideoMAEv2 [46], with the publicly available pretrained weights, fine-tuned on the SSv2 dataset (*vit_g_hybrid_pt_1200e_ssv2_ft*). Our encoding process is carried out on non-overlapping consecutive blocks of 16 frames, extracting Temporal Action Detection (TAD) features for each block. In our experiments, we found no difference in results between non-overlapping blocks and sliding-window blocks. When employing the nearest neighbors algorithm to the video encoding features, we set $k = 1$.

We employ AlphaPose for pose estimation, derive object velocity through optical flows computed via FlowNet2, and utilize YOLOv3 for object detection. For deep image encoding, we leverage CLIP, using a ViT-32 backbone.

Our code and a setup guide are available on `https://github.com/yoavarad/MFAD`.

5.3 Anomaly Detection Results

Our results are based on our optimal model configuration, see Section 5.4. This configuration involves leveraging all four feature types and the max feature while training a logistic regression model on a random 2% of the test set frames for computing final anomaly scores. It is crucial to note that the data used for training the

Configuration	Ped2	Avenue	ShanghaiTech	HMDB-AD	HMDB-Violence
VE	80.3%	87.9%	71.3%	84.9%	75.8%
P + V [12]	98.7%	86.8%	**85.9%**	54.2%	56.1%
P + V + IE [12]	**99.1%**	**93.5%**	85.1%	71.2%	67.9%
P + V + VE	95.8%	91.0%	83.5%	77.8%	70.3%
P + V + IE + VE	96.8%	92.6%	83.0%	82.9%	75.2%
P + V + IE + VE + max	97.0%	92.8%	82.1%	**85.1%**	**76.7%**

Table 2: Comparison of different model configurations, evaluating the impact of various feature types, including pose features (P), velocity features (V), image encoding features (IE), and video encoding features (VE), on the model's performance. max is the max value between {P, V, IE, VE}. Best and second best results are in bold and underlined, respectively.

logistic regression model is not included in the evaluation. To ensure reliability, we repeat this final step 100 times and report the mean AUC result along with the standard deviation. The consistently low standard deviation across all datasets underscores the stability of our method.

MFAD demonstrates competitive results on the well-established benchmark datasets, with modest differences of approximately -0.3%, -0.8%, and -1.1% from the state-of-the-art results on Ped2, Avenue, and ShanghaiTech, respectively. The true strength of our approach becomes evident when applied to the newly introduced datasets, HMDB-AD and HMDB-Violence. On these datasets, we achieve substantial improvements of 19.9% and 9.7%, respectively.

MFAD was tested against four different methods on these new datasets, including AI-VAD [12], upon which our work is built and is the state-of-the-art on the ShanghaiTech dataset. This substantial enhancement highlights the generalizability of our approach to various complex anomalies, without majorly impacting our detection ability of simple anomalies, underscoring the significance of our contributions. For detailed comparison see Table 1. We further report the configuration of MFAD without image encoding (IE) features, improving results on ShanghaiTech by 0.2%.

MFAD faces similar challenges to previous methods when evaluated against the benchmark datasets. Particularly, the object-oriented aspect of MFAD struggles when confronted with scenarios involving closely clustered pedestrians.

In addition to quantitative evaluations, we conducted qualitative analyses on videos from the ShanghaiTech dataset, which feature more complex anomalies beyond novel object detection. These anomalies are shown in Appendix B. The positive impact of our method is clearly evident in Fig. 3, where abnormal frames receive higher anomaly scores, while normal frames receive lower anomaly scores, further validating our method.

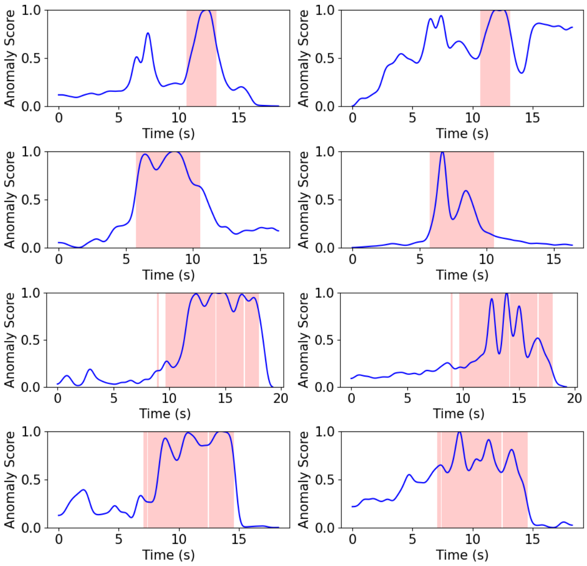

Figure 3: Qualitative results from four ShanghaiTech videos: 01_0028, 03_0032, 03_0039, 07_0008 (respectively). In each pair, MFAD (left) is compared to AI-VAD [12] (right). Anomalous sections are highlighted in red, while the anomaly scores, ranging from 0 to 1, are the blue line. These videos feature complex, behavior-based anomalies rather than novel object detection scenarios, that are more common in this dataset. Clearly, MFAD improves both detecting anomalies and accurately assessing normal parts of the video. Best viewed in color.

5.4 Ablation Study

We perform an ablation study to determine two factors: the added benefit of the video encoding feature, and the most favorable configuration for the logistic regression module.

Feature Selection

In their ablation study, AI-VAD [12] demonstrated the incremental value of their three distinct feature types: pose estimation, deep image encoding, and velocity features, as well as the added effect of Gaussian

Configuration	Ped2	Avenue	ShanghaiTech	HMDB-AD	HMDB-Violence
$\alpha = 0\%$	96.8%	92.6%	83.0%	82.9%	75.2%
$\alpha = 1\%$	98.5% ± 1.1%	92.5% ± 0.6%	84.5% ± 0.6%	89.6% ± 0.6%	79.6% ± 0.3%
$\alpha = 2\%$	99.0% ± 0.6%	92.7% ± 0.7%	84.7% ± 0.4%	89.9% ± 0.4%	79.7% ± 0.2%
$\alpha = 3\%$	99.2% ± 0.5%	92.7% ± 0.7%	84.8% ± 0.4%	89.9% ± 0.3%	79.7% ± 0.1%
$\alpha = 4\%$	99.4% ± 0.3%	92.7% ± 0.7%	84.7% ± 0.3%	89.9% ± 0.3%	79.8% ± 0.1%
$\alpha = 5\%$	99.4% ± 0.4%	93.0% ± 0.6%	84.8% ± 0.3%	90.0% ± 0.2%	79.8% ± 0.1%
$\alpha = 10\%$	99.5% ± 0.2%	92.9% ± 0.6%	84.8% ± 0.2%	90.1% ± 0.2%	79.8% ± 0.1%
$\alpha = 20\%$	99.6% ± 0.1%	93.1% ± 0.5%	84.8% ± 0.2%	90.2% ± 0.2%	79.8% ± 0.1%
$\alpha = 50\%$	**99.7% ± 0.1%**	93.1% ± 0.3%	84.8% ± 0.2%	90.2% ± 0.2%	79.7% ± 0.3%
$\alpha = 90\%$	99.7% ± 0.3%	**93.2% ± 0.6%**	84.8% ± 0.7%	90.2% ± 0.5%	79.8% ± 0.8%
$\alpha = 0\%$ + max	97.0%	92.8%	82.1%	85.1%	76.7%
$\alpha = 1\%$ + max	98.5% ± 0.8%	92.5% ± 0.6%	84.5% ± 0.6%	89.8% ± 0.5%	80.2% ± 0.3%
$\alpha = 2\%$ + max	99.0% ± 0.5%	92.9% ± 0.5%	84.8% ± 0.4%	90.0% ± 0.4%	**80.2% ± 0.2%**
$\alpha = 3\%$ + max	99.1% ± 0.7%	92.9% ± 0.6%	85.0% ± 0.3%	90.0% ± 0.3%	**80.2% ± 0.2%**
$\alpha = 4\%$ + max	99.3% ± 0.5%	93.0% ± 0.5%	85.1% ± 0.3%	90.1% ± 0.3%	**80.2% ± 0.2%**
$\alpha = 5\%$ + max	99.3% ± 0.4%	93.0% ± 0.4%	85.1% ± 0.3%	90.2% ± 0.3%	**80.2% ± 0.2%**
$\alpha = 10\%$ + max	99.5% ± 0.2%	93.0% ± 0.4%	85.2% ± 0.2%	90.2% ± 0.2%	**80.2% ± 0.2%**
$\alpha = 20\%$ + max	99.6% ± 0.1%	93.0% ± 0.3%	85.2% ± 0.2%	90.3% ± 0.2%	80.1% ± 0.2%
$\alpha = 50\%$ + max	**99.7% ± 0.1%**	93.0% ± 0.3%	**85.3% ± 0.2%**	**90.4% ± 0.2%**	80.1% ± 0.3%
$\alpha = 90\%$ + max	99.7% ± 0.2%	93.1% ± 0.6%	85.3% ± 0.6%	90.4% ± 0.5%	80.1% ± 0.8%

Table 3: Performance comparison between various model configurations, with different amounts of training data for the logistic regression model. α represents the proportion of test set frames employed for the training, with these frames excluded from model evaluation. We repeat the process 100 times, and both mean and standard deviation values are reported. The first half uses the basic four features, and the second half also uses the extra max feature. Best and second best results are highlighted in bold and underlined, respectively. The minimal difference in results between different values of $\alpha > 0\%$ is evident.

smoothing. Expanding upon their work, we test the impact of incorporating deep video encoding features in different forms. Our study consists of tests involving both video encoding features in isolation, the combination of all four feature types, and the addition of an extra max feature, that has the value of the maximum feature score per each frame. Furthermore, we explore the substitution of image encoding features with video encoding features due to their semantic similarity.

As presented in Table 2, utilizing solely video encoding features yields impressive performance for multi-frame anomalies. However, this specialization comes at the cost of lower performance on the traditional video anomaly detection datasets. On the other hand, employing all four feature types results in a comprehensive and well-balanced model that performs admirably across all datasets, even though it may not achieve the top rank in any specific dataset.

Logistic Regression

When incorporating the logistic regression model, we conducted experiments to assess the impact of varying amounts of additional training data extracted from the testing set. Specifically, we explored using 1-5%, 10%, 20%, 50%, 90% of the frames for the training. We used both the configuration using only the four basic features and the configuration also using the additional max feature. The results, as presented in Table 3, indicate that the amount of extra training data has minimal effects, as long as there is some extra data. We repeated each configuration 100 times and reported both the mean and standard deviation values. The consistently low standard deviation values observed across all configurations and datasets underscore the robustness of our approach. We chose 2% extra data with the max feature as the optimal trade-off between extra data and efficacy.

6 CONCLUSION

Our paper introduces a broader interpretation of anomalies, encompassing both simple anomalies, commonly found in existing benchmarks, and multi-frame complex anomalies. Building upon the foundation laid by AI-VAD [12], we present a novel method that achieves state-of-the-art performance on our proposed datasets while remaining competitive with recent methods on benchmark datasets. We introduce two new datasets of varying complexity, designed to assess the ability of future models to detect complex action-based anomalies.

In future work, we aim to explore even more intricate types of anomalies, such as location and time-based anomalies (e.g. detecting normal actions occurring at abnormal locations or times) thus further advancing the field of general anomaly detection in videos.

REFERENCES

[1] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. San Francisco, CA, USA: IEEE, Jun. 2010, pp. 1975–1981. [Online]. Available: http://ieeexplore.ieee.org/document/5539872/

[2] C. Lu, J. Shi, and J. Jia, "Abnormal Event Detection at 150 FPS in MATLAB," in *2013 IEEE International Conference on Computer Vision*, Dec. 2013, pp. 2720–2727, iSSN: 2380-7504.

[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*, 2018, pp. 6536–6545. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Liu_Future_Frame_Prediction_CVPR_2018_paper

[4] 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*, 2019, pp. 1273–1283. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2019/html/Nguyen_Anomaly_Detection_in_Video_Sequence_With_Appearance-Motion_Correspondence_ICCV_2019_paper

[5] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2017, pp. 439–444, iSSN: 1945-788X.

[6] 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," Aug. 2019, arXiv:1904.02639 [cs]. [Online]. Available: http://arxiv.org/abs/1904.02639

[7] Y. Fan, G. Wen, D. Li, S. Qiu, and M. D. Levine, "Video Anomaly Detection and Localization via Gaussian Mixture Fully Convolutional Variational Autoencoder," May 2018, arXiv:1805.11223 [cs]. [Online]. Available: http://arxiv.org/abs/1805.11223

[8] 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*, 2016, pp. 733–742. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Hasan_Learning_Temporal_Regularity_CVPR_2016_paper

[9] Y. Lu, K. M. Kumar, S. s. Nabavi, and Y. Wang, "Future Frame Prediction Using Convolutional VRNN for Anomaly Detection," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Sep. 2019, pp. 1–8, iSSN: 2643-6213.

[10] G. Yu, S. Wang, Z. Cai, E. Zhu, C. Xu, J. Yin, and M. Kloft, "Cloze Test Helps: Effective Video Anomaly Detection via Learning to Complete Video Events," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 583–591. [Online]. Available: https://dl.acm.org/doi/10.1145/3394171.3413973

[11] 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*, 2020, pp. 14 372–14 381. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Park_Learning_Memory-Guided_Normality_for_Anomaly_Detection_CVPR_2020_paper

[12] T. Reiss and Y. Hoshen, "Attribute-based Representations for Accurate and Interpretable Video Anomaly Detection," Dec. 2022, arXiv:2212.00789 [cs]. [Online]. Available: http://arxiv.org/abs/2212.00789

[13] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 588–13 597. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Liu_A_Hybrid_Video_Anomaly_Detection_Framework_via_Memory-Augmented_Flow_Reconstruction_ICCV_2021_paper

[14] A. Barbalau, R. T. Ionescu, M.-I. Georgescu, J. Dueholm, B. Ramachandra, K. Nasrollahi, F. S. Khan, T. B. Moeslund, and M. Shah, "SSMTL++: Revisiting Self-Supervised Multi-Task Learning for Video Anomaly Detection," Feb. 2023, arXiv:2207.08003 [cs]. [Online]. Available: http://arxiv.org/abs/2207.08003

[15] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 3, pp. 555–560, Mar. 2008.

[16] U. of Minnesota, "Unusual crowd activity dataset of university of minnesota," 2006. [Online]. Available: http://mha.cs.umn.edu/proj_events.shtml

[17] B. Ramachandra and M. J. Jones, "Street Scene: A new dataset and evaluation protocol for video anomaly detection," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2020, pp. 2558–2567, iSSN: 2642-9381.

[18] R. Rodrigues, N. Bhargava, R. Velmurugan, and S. Chaudhuri, "Multi-timescale Trajectory Prediction for Abnormal Human Activity Detection," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2020, pp. 2615–2623, iSSN: 2642-9381.

[19] A. Acsintoae, A. Florescu, M.-I. Georgescu, T. Mare, P. Sumedrea, R. T. Ionescu, F. S. Khan, and M. Shah, "UBnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 143–20 153. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Acsintoae_UBnormal_New_Benchmark_for_Supervised_Open-Set_Video_Anomaly_Detection_CVPR_2022_paper.html

[20] C. Cao, Y. Lu, P. Wang, and Y. Zhang, "A New Comprehensive Benchmark for Semi-Supervised Video Anomaly Detection and Anticipation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 392–20 401. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Cao_A_New_Comprehensive_Benchmark_for_Semi-Supervised_Video_Anomaly_Detection_and_CVPR_2023_paper

[21] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Sultani_Real-World_Anomaly_Detection_CVPR_2018_paper

[22] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision," in *European Conference on Computer Vision (ECCV)*, vol. 12375. Cham: Springer International Publishing, 2020, pp. 322–339, series Title: Lecture Notes in Computer Science. [Online]. Available: https://link.springer.com/10.1007/978-3-030-58577-8_20

[23] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A Large Video Database for Human Motion Recognition," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[24] Qualcomm, "Moving Objects Dataset: Something-Something v. 2," 2018. [Online]. Available: https://developer.qualcomm.com/software/ai-datasets/something-something

[25] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset," May 2017, arXiv:1705.06950 [cs]. [Online]. Available: http://arxiv.org/abs/1705.06950

[26] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A Short Note about Kinetics-600," Aug. 2018, arXiv:1808.01340 [cs]. [Online]. Available: http://arxiv.org/abs/1808.01340

[27] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu, and A. Zisserman, "A Short Note on the Kinetics-700-2020 Human Action Dataset," Oct. 2020, arXiv:2010.10864 [cs]. [Online]. Available: http://arxiv.org/abs/2010.10864

[28] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 1932–1939, iSSN: 1063-6919.

[29] J. Pers, V. Sulic, M. Kristan, M. Perse, K. Polanec, and S. Kovacic, "Histograms of optical flow for efficient representation of body motion," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1369–1376, Aug. 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865510001121

[30] R. V. H. M. Colque, C. Caetano, M. T. L. de Andrade, and W. R. Schwartz, "Histograms of Optical Flow Orientation and Magnitude and Entropy to Detect Anomalous Events in Videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 673–682, Mar.

2017, conference Name: IEEE Transactions on Circuits and Systems for Video Technology.

[31] B. Sabzalian, H. Marvi, and A. Ahmadyfard, “Deep and Sparse features For Anomaly Detection and Localization in video,” in *2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, Mar. 2019, pp. 173–178, iSSN: 2049-3630.

[32] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[33] L. J. Latecki, A. Lazarevic, and D. Pokrajac, “Outlier Detection with Kernel Density Functions,” in *Machine Learning and Data Mining in Pattern Recognition*, ser. Lecture Notes in Computer Science, P. Perner, Ed. Berlin, Heidelberg: Springer, 2007, pp. 61–75.

[34] M. Glodek, M. Schels, and F. Schwenker, “Ensemble Gaussian mixture models for probability density estimation,” *Computational Statistics*, vol. 27, pp. 127–138, Dec. 2013.

[35] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, “A Geometric Framework for Unsupervised Anomaly Detection,” in *Applications of Data Mining in Computer Security*, ser. Advances in Information Security, D. BarbarÃ¡ and S. Jajodia, Eds. Boston, MA: Springer US, 2002, pp. 77–101. [Online]. Available: https://doi.org/10.1007/978-1-4615-0953-0_4

[36] G. Wang, Y. Wang, J. Qin, D. Zhang, X. Bao, and D. Huang, “Video Anomaly Detection byÂ Solving Decoupled Spatio-Temporal Jigsaw Puzzles,” in *Computer Vision - ECCV 2022*, ser. Lecture Notes in Computer Science, S. Avidan, G. Brostow, M. CissÃ©, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 494–511.

[37] D. Wei, J. J. Lim, A. Zisserman, and W. T. Freeman, “Learning and Using the Arrow of Time,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8052–8060. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Wei_Learning_and_Using_CVPR_2018_paper.html

[38] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised Representation Learning by Predicting Image Rotations,” Mar. 2018, arXiv:1803.07728 [cs]. [Online]. Available: http://arxiv.org/abs/1803.07728

[39] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, “Anomaly Detection in Video via Self-Supervised and Multi-Task Learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 742–12 752. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Georgescu_Anomaly_Detection_in_Video_via_Self-Supervised_and_Multi-Task_Learning_CVPR_2021_paper.html

[40] M. I. Georgescu, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, “A Background-Agnostic Framework With Adversarial Training for Abnormal Event Detection in Video,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4505–4523, Sep. 2022, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[41] S. Sun and X. Gong, “Hierarchical Semantic Contrast for Scene-Aware Video Anomaly Detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 846–22 856. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Sun_Hierarchical_Semantic_Contrast_for_Scene-Aware_Video_Anomaly_Detection_CVPR_2023_paper

[42] Z. Yang, P. Wu, J. Liu, and X. Liu, “Dynamic Local Aggregation Network with Adaptive Clusterer for Anomaly Detection,” Jul. 2022, arXiv:2207.10948 [cs]. [Online]. Available: http://arxiv.org/abs/2207.10948

[43] G. Yu, S. Wang, Z. Cai, X. Liu, C. Xu, and C. Wu, “Deep Anomaly Discovery From Unlabeled Videos via Normality Advantage and Self-Paced Refinement,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 987–13 998. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Yu_Deep_Anomaly_Discovery_From_Unlabeled_Videos_via_Normality_Advantage_and_CVPR_2022_paper.html

[44] R. Cai, H. Zhang, W. Liu, S. Gao, and Z. Hao, “Appearance-Motion Memory Consistency Network for Video Anomaly Detection,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, pp. 938–946, May 2021, number: 2. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16177

[45] H. Lv, C. Chen, Z. Cui, C. Xu, Y. Li, and J. Yang, “Learning Normal Dynamics in

Videos with Meta Prototype Network," May 2021, arXiv:2104.06689 [cs]. [Online]. Available: http://arxiv.org/abs/2104.06689

[46] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao, "VideoMAE V2: Scaling Video Masked Autoencoders With Dual Masking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 549–14 560. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Wang_VideoMAE_V2_Scaling_Video_Masked_Autoencoders_With_Dual_Masking_CVPR_2023_paper.html

A SAMPLES FROM PROPOSED DATASETS

A.1 HMDB-AD Examples

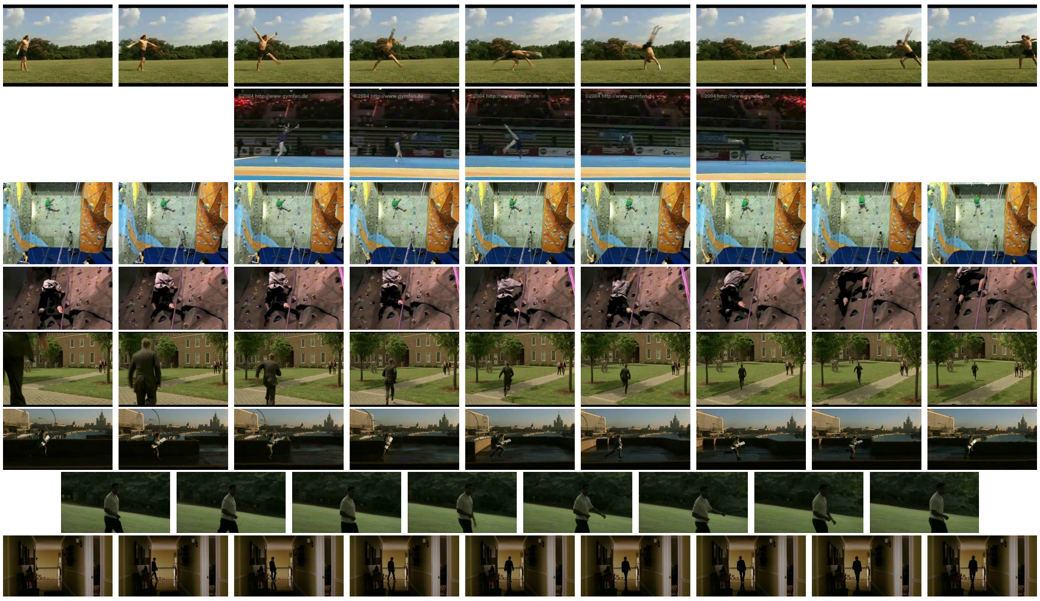

Figure 4: HMDB-AD examples: cartwheel (2), climb (2), run (2), walk (2).

A.2 HMDB-Violence Examples

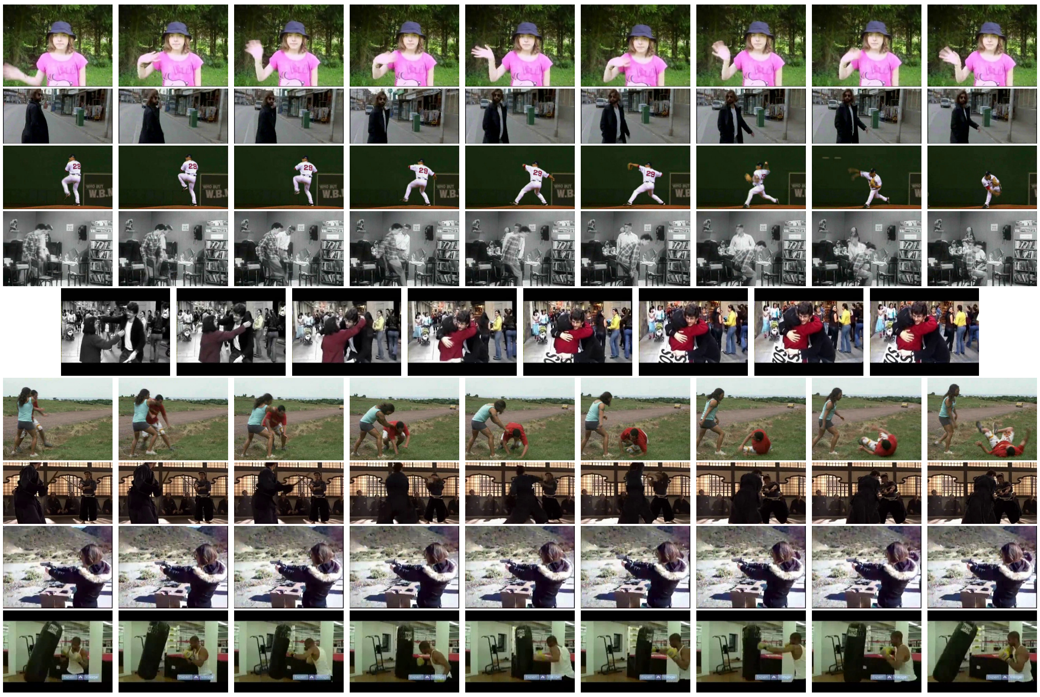

Figure 5: HMDB-Violence examples: wave, turn, throw, sit, hug, fall, sword, shoot, punch.

B VIDEOS FOR QUALITATIVE ANALYSES

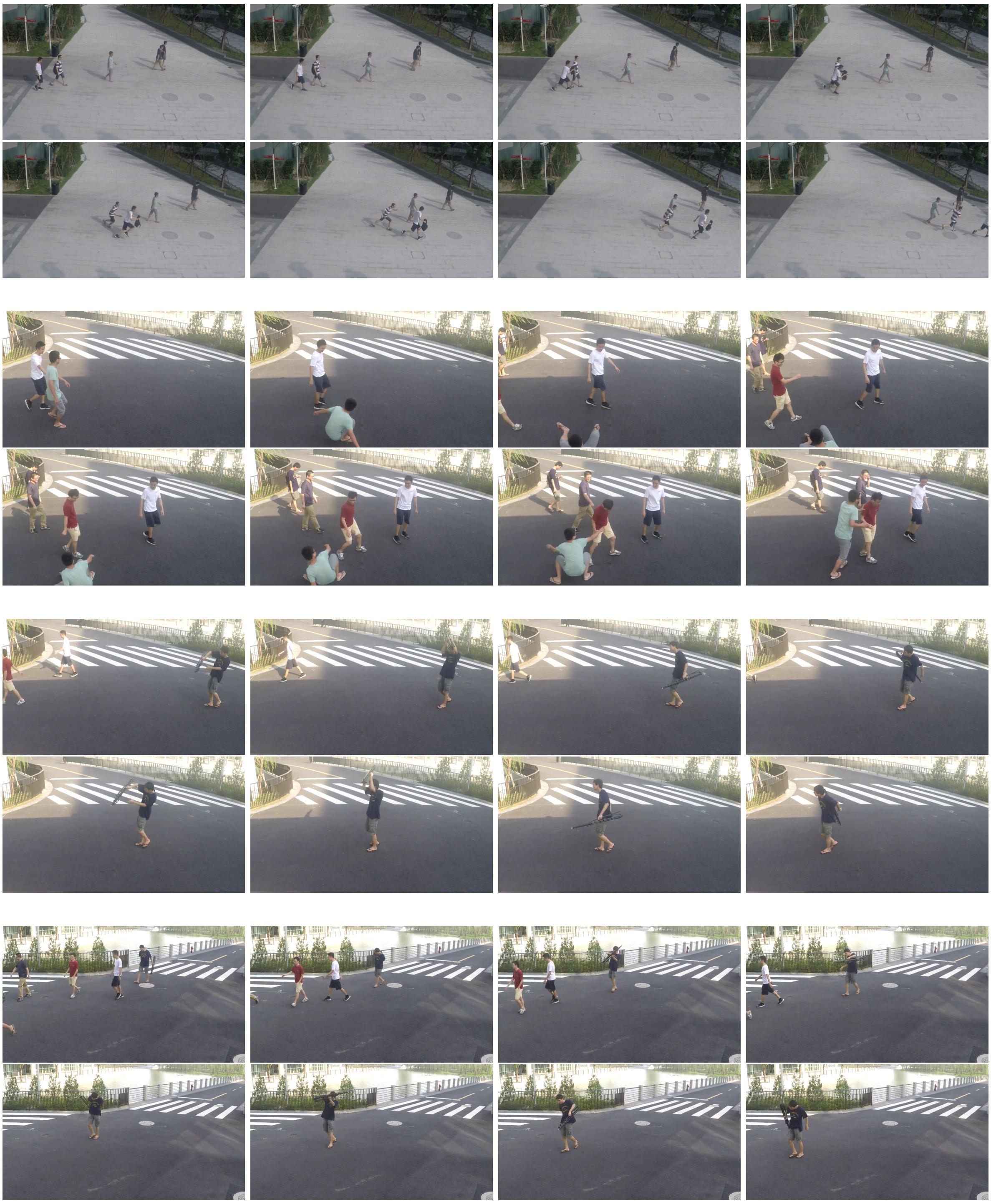

Figure 6: The anomalies from 01_0028, 03_0032, 03_0039, 07_0008 (top to bottom, respectively) videos from ShanghaiTech Campus dataset.