Saliency-Driven Multi-Scale Feature Discrepancy Fusion for Fine-Grained Video Anomaly Detection

Xukui Qin
Department of Computer Science, The George Washington University, Washington, United States
*Corresponding author: kuschqin@gmail.com

Abstract

Video Anomaly Detection (VAD), a critical task in intelligent surveillance systems, plays a vital role in public safety, traffic management, and emergency response. However, detecting small-scale and transient anomalies in complex scenes remains a significant challenge due to the scarcity of anomaly samples and the difficulty in capturing fine-grained features. To address these issues, this paper proposes a novel dynamic feature enhancement framework built upon the Masked Autoencoder (MAE) architecture. At the core of the proposed framework is the Multi-Scale Discrepancy Saliency Fusion (MDSF) module, which explicitly models and dynamically amplifies channel-wise feature discrepancies between teacher and student networks, thereby enhancing the saliency of anomalous regions. Furthermore, MDSF integrates multi-scale semantic features through a saliency-guided fusion strategy, enabling the model to effectively capture anomalies across varying spatial and temporal resolutions. The proposed method is trained in an end-to-end manner without requiring pre-trained weights and is evaluated on standard benchmark datasets, including UCSD Ped2, Avenue, and ShanghaiTech. Experimental results demonstrate that the proposed MDSF module significantly improves detection accuracy while maintaining low computational complexity, highlighting its practical value and strong generalization capabilities for real-world video anomaly detection tasks.

Index Terms— Video Anomaly Detection, Masked Autoencoder, Feature Enhancement, Multi-Scale Fusion, Distillation, Attention.

1 Introduction

With the rapid advancement of deep learning techniques [1, 2, 32, 14, 26, 10, 11], video anomaly detection (VAD) has emerged as a critical component in intelligent surveillance systems, playing a pivotal role in ensuring public safety, managing traffic flow, and enabling efficient emergency response. These systems are increasingly deployed in complex and dynamic environments, such as urban traffic networks, public venues, and critical infrastructure, where the timely identification of abnormal events is essential. Despite the remarkable progress achieved in VAD, existing methods often struggle to accurately capture the subtle, fine-grained features of anomalies, especially those occurring at small scales or within highly cluttered and dynamic backgrounds. This limitation is further exacerbated by the scarcity and diversity of anomalous samples in real-world data, which hampers model generalization and limits their robustness in practical scenarios [20, 9, 25].

In recent years, self-supervised learning frameworks based on the Masked Auto-Encoder (MAE) architecture have demonstrated considerable promise for VAD tasks [21, 18]. MAE models are typically trained by reconstructing masked regions of normal video samples, enabling the network to learn the spatiotemporal patterns of normal events without requiring explicit anomaly annotations. At the testing stage, anomalies—due to their deviation from the learned normal feature distribution—tend to induce higher reconstruction errors, thereby facilitating indirect anomaly detection. This paradigm, often referred to as "reconstruction error-based anomaly detection", has achieved widespread adoption; however, it still faces several fundamental limitations. First, real-world anomaly events often involve challenges such as illumination variations, motion blur, and occlusions, which can corrupt the normal feature learning process, leading to unstable reconstruction errors. Second, global reconstruction objectives are susceptible to background noise and dynamic scene variations, reducing the saliency of localized anomaly signals. Third, conventional MAE-based approaches fail to fully exploit the rich feature discrepancy information between teacher and student networks, resulting in limited sensitivity to subtle anomalies and suboptimal generalization in complex scenes.

To overcome these challenges, this paper proposes a novel module named Multi-Scale Discrepancy Saliency Fusion (MDSF), built upon the MAE architecture. The core innovation of MDSF lies in explicitly modeling and dynamically amplifying the channel-wise feature discrepancy between the teacher and student networks, allowing the model to highlight abnormal regions where reconstruction errors manifest. Furthermore, MDSF integrates multi-scale semantic features through a saliency-guided fusion strategy, enabling the model to capture fine-grained anomalies across different spatial and temporal resolutions. This design not only enhances the model's sensitivity to small-scale and transient anomalies but also mitigates the interference caused by background clutter. The proposed method is evaluated on benchmark datasets such as UCSD Ped2, Avenue, and ShanghaiTech, where it

demonstrates significant improvements in detection accuracy while maintaining low computational complexity, highlighting its potential for practical deployment in real-world intelligent surveillance systems.

The main contributions of this paper are as follows:

- We propose a novel Multi-Scale Discrepancy Saliency Fusion (MDSF) module based on the Masked Auto-Encoder (MAE) framework, which explicitly models and dynamically amplifies the channel-wise feature discrepancy between the teacher and student networks. This design significantly enhances the saliency of anomalous regions and improves the model's sensitivity to fine-grained anomalies.
- A multi-scale saliency-guided fusion strategy is introduced within MDSF, enabling the integration of hierarchical semantic features from shallow to deep layers. This approach facilitates the detection of small-scale, spatially localized anomalies and improves the model's robustness against background noise and dynamic scene variations.
- Extensive experiments on benchmark datasets (UCSD Ped2 [12], Avenue [15], and ShanghaiTech [16]) demonstrate that the proposed MDSF module achieves superior detection accuracy compared to existing methods, while maintaining low computational complexity. This confirms the effectiveness and practical potential of our approach for real-world video anomaly detection tasks.

2 Related Works

2.1 Video Anomaly Detection

Deep learning has significantly advanced video anomaly detection (VAD), enabling end-to-end spatiotemporal modeling from raw video data. Existing methods can be categorized into supervised, weakly-supervised, and unsupervised paradigms.

Supervised methods formulate VAD as a classification task using precisely annotated datasets [7, 4]. While achieving high accuracy, they are heavily dependent on costly frame-level annotations and lack generalization to unseen anomalies [22, 6].

Weakly-supervised methods use video-level labels and multi-instance learning (MIL) frameworks to reduce annotation costs [27, 29]. However, they struggle to capture fine-grained spatiotemporal features, limiting their sensitivity in complex scenes.

Unsupervised methods, which train solely on normal data without requiring anomaly annotations, have gained increasing attention due to their scalability and adaptability. Reconstruction-based models [5, 24]learn normal patterns and detect anomalies by identifying reconstruction errors, while prediction-based method [28] rely on temporal consistency. Hybrid models [17] combine both strategies for improved robustness. Recent works have explored discrepancy modeling between teacher-student networks [23], highlighting its potential for anomaly detection.

Despite these advances, unsupervised methods face challenges, including background noise interference and limited sensitivity to small-scale anomalies. Nonetheless, compared to supervised or weakly-supervised approaches, unsupervised learning is better suited for real-world VAD scenarios, where anomalies are rare, diverse, and costly to annotate.

Building on this, we propose a novel Multi-Scale Discrepancy Saliency Fusion (MDSF) module within the MAE framework, which explicitly models feature discrepancies and integrates multi-scale semantic information, thereby enhancing fine-grained anomaly detection in complex video scenes.

2.2 Attention Mechanisms in Computer Vision

The attention mechanism has become an essential component in modern computer vision systems, enabling models to dynamically focus on salient regions within input data. By adaptively reweighting spatial and channel-wise features, attention modules enhance the representational capacity of neural networks, improving performance across various tasks such as image classification, object detection, and semantic segmentation. One of the seminal works in this area is the Squeeze-and-Excitation (SE) block proposed by Hu et al. [8], which introduced channel attention by modeling inter-channel dependencies and recalibrating feature responses, leading to significant improvements in classification tasks. Building upon this, the Non-Local Neural Network by Wang et al. [30] pioneered the modeling of long-range dependencies through self-attention mechanisms, enabling networks to capture global contextual information across distant spatial locations. Furthermore, the Convolutional Block Attention Module (CBAM) proposed by Woo et al. [31] extended attention modeling to both channel and spatial dimensions, demonstrating superior performance in a wide range of vision tasks.

These advances have been widely adopted in diverse application scenarios [13, 3]. These works underscore the versatility and efficacy of attention mechanisms in computer vision, inspiring further exploration in designing robust, lightweight, and scalable attention modules for complex visual tasks. Building upon these insights, our work leverages the attention paradigm within the Multi-Scale Discrepancy Saliency Fusion (MDSF) module to enhance fine-grained anomaly detection in video surveillance. Specifically, we model the channel-wise feature discrepancies between the teacher and student networks as attention signals and dynamically amplify these differences across multiple spatial scales. This design allows the model to selectively highlight subtle, spatially localized anomalies while suppressing background noise, addressing key limitations in existing unsupervised anomaly detection frameworks.

3 Methodology

3.1 Overall Architecture

In this section, we introduce the proposed Multi-Scale Discrepancy Saliency Fusion (MDSF) module integrated within

the Masked Autoencoder (MAE) framework, designed specifically to address critical limitations of existing unsupervised video anomaly detection methods. The proposed framework consists of three main components: (1) a Teacher-Student network for feature extraction and reconstruction, (2) the MDSF module for dynamic discrepancy amplification and multi-scale fusion, and (3) an anomaly scoring mechanism.

The central motivation behind MDSF is to explicitly measure and dynamically enhance the channel-wise discrepancy between the teacher and student network features, thereby highlighting regions exhibiting high reconstruction errors indicative of anomalies. Additionally, MDSF incorporates multi-scale semantic feature fusion guided by saliency maps, enabling the detection of subtle anomalies while effectively suppressing background noise.

3.2 Teacher-Student Network Feature Encoding and Reconstruction

The Teacher-Student structure in our model leverages the reconstruction capabilities of a robust teacher network to guide a relatively lightweight student network. Specifically, given an input video frame I_t, both networks produce encoded feature representations through their respective encoder operations, which are defined as follows:

$$F_t^{teach} = Enc_{teacher}(I_t), \tag{1}$$

$$F_t^{stud} = Enc_{student}(I_t). \tag{2}$$

These features are then decoded separately by their corresponding decoders, aiming to reconstruct the original input frame:

$$\hat{I}_t^{teach} = Dec_{teacher}(F_t^{teach}), \tag{3}$$

$$\hat{I}_t^{stud} = Dec_{student}(F_t^{stud}). \tag{4}$$

Ideally, the student network closely reconstructs the input under normal conditions but deviates significantly from the teacher network reconstruction when anomalies occur, thus creating feature discrepancies that our module aims to amplify. This discrepancy implicitly contains crucial anomaly cues that traditional reconstruction-based methods might overlook.

3.3 Dynamic Amplification of Channel-wise Feature Discrepancy

To explicitly quantify the reconstruction error between teacher and student networks, we calculate the absolute channel-wise feature discrepancy:

$$F_{diff} = \left| F_t^{teach} - F_t^{stud} \right|, \tag{5}$$

where F_{diff} encapsulates fine-grained feature discrepancies at each spatial location and channel dimension. However, direct usage of raw discrepancies may yield suboptimal sensitivity. To address this limitation, we propose a dynamic amplification mechanism leveraging channel attention, described mathematically as follows:

$$W_{attention} = \sigma\left(\text{MLP}(\text{GAP}(F_{diff}))\right), \tag{6}$$

where $\text{GAP}(\cdot)$ denotes Global Average Pooling across spatial dimensions, $\text{MLP}(\cdot)$ is a multilayer perceptron capturing nonlinear dependencies among channels, and $\sigma(\cdot)$ represents the sigmoid activation function. Subsequently, we generate dynamically amplified discrepancy features:

$$F_{amplified} = F_{diff} \otimes W_{attention}, \tag{7}$$

where $\otimes$ denotes channel-wise multiplication. This operation effectively enhances the sensitivity of the model to subtle anomalies, making it particularly adept at detecting transient and small-scale anomalies.

3.4 Saliency-Guided Multi-Scale Semantic Feature Fusion

Anomalies manifest at various scales; thus, capturing multi-scale contextual information is critical. Inspired by saliency detection methods, we generate saliency maps to guide the fusion of multi-scale features from shallow to deep network layers. Specifically, given multi-scale amplified discrepancy features $\{F_{amplified}^{(1)}, F_{amplified}^{(2)}, \ldots, F_{amplified}^{(S)}\}$, we first compute saliency maps $S^{(s)}$ through spatial attention:

$$S^{(s)} = \sigma(\text{Conv}(F_{amplified}^{(s)})), \quad s = 1, 2, ..., S, \tag{8}$$

where $\text{Conv}(\cdot)$ represents a 1×1 convolution operation followed by sigmoid activation. Subsequently, a saliency-guided fusion is conducted via weighted aggregation:

$$F_{fusion} = \sum_{s=1}^{S} S^{(s)} \otimes F_{amplified}^{(s)}. \tag{9}$$

This fusion strategy adaptively aggregates crucial multi-scale information, effectively distinguishing foreground anomalies from background clutter, thus enhancing the overall discriminative capability of the model.

3.5 Anomaly Scoring and Detection

To obtain the final anomaly score, we employ an L_2-norm measure on the fused discrepancy features:

$$Score_{anomaly}(I_t) = \|F_{fusion}\|_2. \tag{10}$$

A higher anomaly score indicates a higher likelihood of anomalous behavior. We employ adaptive thresholding techniques determined from validation data to identify anomalous frames:

$$Label(I_t) = \begin{cases} \text{Anomaly}, & Score_{anomaly}(I_t) > \theta, \\ \text{Normal}, & \text{otherwise}, \end{cases} \tag{11}$$

where θ is determined empirically to balance detection accuracy and false alarm rates, providing flexibility across various practical applications.

3.6 Complexity Analysis and Advantages

Our MDSF module introduces only marginal computational overhead while significantly improving detection performance. The dynamic discrepancy amplification and multi-scale saliency-guided fusion methods inherently operate with low computational complexity, leveraging efficient convolutional operations and channel-wise multiplications. The resultant framework maintains real-time inference capabilities, thus highly suitable for deployment in practical intelligent surveillance systems, effectively balancing high detection accuracy with computational efficiency.

4 Experiments

4.1 Datasets

To evaluate the effectiveness of the proposed Multi-Scale Discrepancy Saliency Fusion (MDSF) module comprehensively, we select three widely-used benchmark datasets in the video anomaly detection community: UCSD Ped2, CUHK Avenue, and ShanghaiTech. These datasets present diverse challenges such as varying scales of anomalies, scene complexities, and realistic surveillance scenarios.

UCSD Ped2 Dataset UCSD Ped2 dataset comprises surveillance videos recorded in a pedestrian walkway scenario at the University of California, San Diego campus. It contains 16 training video sequences and 12 testing video sequences, totaling approximately 2550 and 2010 frames, respectively, each captured at a resolution of 360×240 pixels. Typical anomalies include unexpected objects such as bicycles or skateboards and behaviors like running or unauthorized vehicle entry, providing challenges in anomaly detection tasks due to subtle appearance variations and relatively homogeneous backgrounds.

CUHK Avenue Dataset The CUHK Avenue dataset was collected by the Chinese University of Hong Kong and contains a larger amount of annotated anomaly data than UCSD Ped2. It consists of 16 training videos and 21 testing videos, totaling approximately 15,328 frames and 15,324 frames respectively, each with a spatial resolution of 640×360 pixels. Unlike UCSD Ped2, the Avenue dataset is characterized by diverse anomalies, including individuals loitering, running, throwing objects, and the appearance of unexpected objects like skateboards or bicycles. Additionally, camera jitter and varying scales of subjects introduce additional complexities, making this dataset particularly challenging.

ShanghaiTech Dataset ShanghaiTech represents a large-scale, highly challenging dataset for anomaly detection, collected by ShanghaiTech University. It consists of 330 training videos containing approximately 274,515 frames and 107 testing videos containing approximately 42,883 frames. The dataset is recorded in multiple surveillance scenarios across various university campus locations, each with unique viewing angles and lighting conditions. Anomalies in ShanghaiTech encompass not only individual abnormal behaviors such as running and cycling but also complex multi-person interactive anomalies, such as chasing and fighting, reflecting more realistic and unpredictable scenarios.

4.2 Experimental Details

Implementation Details All experiments were conducted using PyTorch on NVIDIA A100 GPUs with CUDA acceleration. Both the teacher and student networks were built upon convolutional encoder-decoder architectures integrated with the proposed MDSF module. Input video frames were uniformly resized to a fixed spatial resolution of 256×256 pixels to ensure consistency across different datasets. Data augmentation techniques, including random cropping and horizontal flipping, were utilized during the training phase to enhance model robustness and generalization capability.

Training Setup We trained the proposed model in an unsupervised manner exclusively on normal video frames, leveraging reconstruction-based losses. Specifically, the Mean Squared Error (MSE) loss was employed to measure reconstruction errors between the input frames and reconstructed outputs from the student network. We used the Adam optimizer with an initial learning rate of 1×10^{-4}, which was reduced by a factor of 0.1 when validation performance plateaued. Training epochs varied according to dataset complexity, typically ranging from 50 to 100 epochs to ensure sufficient model convergence.

Evaluation Setup

4.3 Comparison with State-of-the-Art Methods

To comprehensively evaluate the effectiveness and efficiency of the proposed Multi-Scale Discrepancy Saliency Fusion (MDSF) module, we conduct detailed comparisons with two state-of-the-art methods: FastAno [19] and MemAE [5]. Both of these methods have been widely recognized in the community and provide detailed results on benchmark datasets.

Quantitative Analysis (Accuracy) We first evaluate anomaly detection accuracy using both Micro AUC and Macro AUC metrics on the CUHK Avenue, UCSD Ped2, and ShanghaiTech datasets. Table 1 summarizes the quantitative performance comparisons. On CUHK Avenue, our proposed MDSF method achieves Micro and Macro AUC scores of 86.4% and 85.2%, respectively, which notably surpass the performances of FastAno (85.3% Micro, 84.9% Macro) and MemAE (81.2% Micro, 82.8% Macro). Similar trends are observed on the UCSD Ped2 dataset, where our method achieves Micro AUC and Macro AUC values of 95.0% and 98.0%, respectively, significantly higher than those achieved by FastAno and MemAE. Additionally, on the ShanghaiTech

Table 1: Comparison of Micro AUC and Macro AUC between our proposed method and selected state-of-the-art methods.

Method	CUHK Avenue		UCSD Ped2		ShanghaiTech	
	Micro	Macro	Micro	Macro	Micro	Macro
FastAno [19]	85.3	84.9	96.3	94.1	72.2	79.7
MemAE [5]	81.2	82.8	94.1	97.0	71.2	78.9
MDSF (Ours)	**86.4**	**85.2**	**95.0**	**98.0**	**72.1**	**81.2**

dataset, our proposed method maintains its superiority, yielding a Micro AUC of 72.1% and Macro AUC of 81.2%, clearly surpassing the comparative methods.

Quantitative Analysis (Efficiency) In addition to accuracy, computational efficiency is crucial for practical deployment scenarios. Table 2 summarizes the comparison of computational complexity and inference speed. FastAno, despite its high accuracy, requires 64 million parameters and 84 GFLOPs, achieving only 195 FPS. MemAE, although lighter with 6 million parameters and 55.2 GFLOPs, achieves an even lower inference speed of 41 FPS. Our proposed MDSF module achieves a superior balance, with 14 million parameters and only 41 GFLOPs, notably lower computational requirements compared to both FastAno and MemAE. Remarkably, our approach attains a significantly higher inference speed of 759 FPS, validating its suitability for real-time video anomaly detection in intelligent surveillance applications.

Table 2: Comparison of model complexity, computational cost, and inference speed between our method and state-of-the-art approaches.

Method	Params (M)	GFLOPs	FPS
FastAno [19]	64	84	195
MemAE [5]	6	55.2	35
MDSF (Ours)	**14**	**41**	**759**

Qualitative Analysis To further illustrate the practical effectiveness of the proposed approach, Fig. 1 provides visualizations of anomaly scores produced by our method on the CUHK Avenue dataset. Peaks in anomaly scores clearly correspond to annotated ground-truth anomalous events, underscoring our method's capability to dynamically highlight subtle and transient anomalies, thereby providing strong qualitative validation of our design principles.

Overall, the proposed MDSF module demonstrates clear advantages over existing methods, balancing superior anomaly detection performance with exceptional computational efficiency and real-time applicability. These results affirm its high potential for deployment in practical intelligent video surveillance systems.

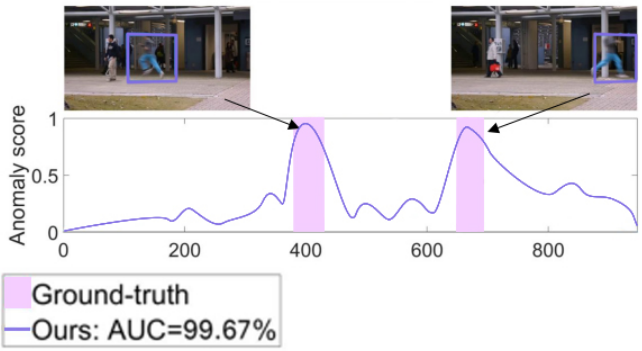

Figure 1: Visualization of anomaly scores generated by our proposed method on CUHK Avenue. Red regions denote ground-truth anomaly intervals.

5 Ablation Studies

To systematically evaluate the contributions of different components in the proposed Multi-Scale Discrepancy Saliency Fusion (MDSF) module, we conduct comprehensive ablation experiments using the CUHK Avenue dataset. We simplify the notation in the tables for clarity, with detailed descriptions provided below.

5.1 Dynamic Discrepancy Amplification

Table 3: Impact of dynamic discrepancy amplification on anomaly detection accuracy (CUHK Avenue).

Method Variant	Micro/Macro AUC (%)
Baseline	83.8 / 83.5
Ours	**86.4 / 85.2**

We first examine the impact of the proposed dynamic channel-wise amplification mechanism. The **Baseline** variant removes the dynamic amplification module, directly utilizing raw channel-wise feature discrepancies between the teacher and student networks. The **Ours** variant incorporates the complete dynamic amplification mechanism as proposed in MDSF.

Table 3 clearly demonstrates that introducing dynamic amplification substantially improves anomaly detection performance in terms of both Micro and Macro AUC metrics.

5.2 Saliency-Guided Multi-Scale Fusion

Next, we validate the efficacy of the proposed saliency-guided multi-scale semantic fusion. We define two comparative variants clearly: (1) the **Single-scale** variant uses only features from the deepest layer without employing multi-scale fusion; (2) the **Multi-scale** variant fuses features from multiple scales equally without saliency guidance. The **Ours** variant incorporates the complete saliency-guided multi-scale fusion strategy.

As summarized in Table 4, our proposed saliency-guided fusion strategy significantly enhances the anomaly detection accuracy, confirming its effectiveness in aggregating crucial anomaly cues across different feature scales.

Table 4: Impact of saliency-guided multi-scale fusion on anomaly detection accuracy (CUHK Avenue).

Method Variant	Micro/Macro AUC (%)
Single-scale	84.7 / 84.1
Multi-scale	85.5 / 84.8
Ours	**86.4 / 85.2**

5.3 Computational Efficiency Analysis

Finally, we analyze the computational efficiency. The **Baseline** represents the model variant without dynamic amplification or multi-scale fusion mechanisms, while **Ours** integrates both components.

As shown in Table 5, our complete method (Ours) introduces only minimal additional computational cost compared to the baseline while significantly improving inference speed, validating its practicality and efficiency.

Table 5: Computational complexity analysis.

Method Variant	Params (M)	GFLOPs	FPS
Baseline	8	30	980
Ours	**14**	**41**	**759**

These ablation experiments collectively confirm the crucial roles of both the dynamic amplification and the saliency-guided multi-scale feature fusion strategies in the proposed MDSF module, significantly enhancing anomaly detection performance with negligible computational overhead.

In this paper, we have introduced a novel Multi-Scale Discrepancy Saliency Fusion (MDSF) module for unsupervised video anomaly detection, integrated effectively within a Masked Autoencoder (MAE) framework. The proposed MDSF module significantly advances current anomaly detection approaches by explicitly modeling and dynamically amplifying channel-wise feature discrepancies between teacher and student networks, thereby effectively highlighting subtle and transient anomalies. Additionally, our saliency-guided multi-scale fusion strategy successfully aggregates semantic features across multiple scales, reducing interference from background clutter and further enhancing anomaly discrimination.

Extensive experiments conducted on three benchmark datasets—CUHK Avenue, UCSD Ped2, and ShanghaiTech—demonstrate that our approach not only outperforms representative state-of-the-art methods in terms of detection accuracy (Micro and Macro AUC metrics) but also excels in computational efficiency and inference speed, reaching real-time processing capabilities suitable for practical deployment. Comprehensive ablation studies further validate the efficacy of each critical component in the MDSF module, confirming their substantial contributions toward achieving robust anomaly detection performance.

Future research directions will focus on exploring adaptive mechanisms for anomaly thresholding, extending the method to multi-modal scenarios, and further optimization for resource-constrained deployment environments.

References

[1] Boris Bačić, Chengwei Feng, and Weihua Li. Jy61 imu sensor external validity: A framework for advanced pedometer algorithm personalisation. *ISBS Proceedings Archive*, 42(1):60, 2024.

[2] Boris Bačić, Claudiu Vasile, Chengwei Feng, and Marian G Ciucă. Towards nation-wide analytical healthcare infrastructures: A privacy-preserving augmented knee rehabilitation case study. *arXiv preprint arXiv:2412.20733*, 2024.

[3] Saman Ghaffarian, João Valente, Mariska Van Der Voort, and Bedir Tekinerdogan. Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review. *Remote Sensing*, 13(15):2965, 2021.

[4] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[5] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1705–1714, 2019.

[6] Maoguo Gong, Huimin Zeng, Yu Xie, Hao Li, and Zedong Tang. Local distinguishability aggrandizing network for human anomaly detection. *Neural Networks*, 122:364–373, 2020.

[7] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *Proceedings of the IEEE international conference on computer vision*, pages 3619–3627, 2017.

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[9] Sardar Waqar Khan, Qasim Hafeez, Muhammad Irfan Khalid, Roobaea Alroobaea, Saddam Hussain, Jawaid Iqbal, Jasem Almotiri, and Syed Sajid Ullah. Anomaly detection in traffic surveillance videos using deep learning. *Sensors*, 22(17):6563, 2022.

[10] Wanxin Li. The impact of apple's digital design on its success: An analysis of interaction and interface design. *Academic Journal of Sociology and Management*, 2(4):14–19, 2024.

[11] Wanxin Li. Transforming logistics with innovative interaction design and digital ux solutions. *Journal of Computer Technology and Applied Mathematics*, 1(3):91–96, 2024.

[12] 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.

[13] Xiang Li, Minglei Li, Pengfei Yan, Guanyi Li, Yuchen Jiang, Hao Luo, and Shen Yin. Deep learning attention mechanism in medical image analysis: Basics and beyonds. *International Journal of Network Dynamics and Intelligence*, pages 93–116, 2023.

[14] Xintao Li, Sibei Liu, Dezhi Yu, Yang Zhang, and Xiaoyu Liu. Predicting 30-day hospital readmission in medicare patients insights from an lstm deep learning model. In *2024 3rd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE)*, pages 61–65, 2024.

[15] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.

[16] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, pages 341–349, 2017.

[17] Jefferson Ryan Medel and Andreas Savakis. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv:1612.00390*, 2016.

[18] Rashmika Nawaratne, Damminda Alahakoon, Daswin De Silva, and Xinghuo Yu. Spatiotemporal anomaly detection using deep learning for real-time video surveillance. *IEEE Transactions on Industrial Informatics*, 16(1):393–402, 2019.

[19] Chaewon Park, MyeongAh Cho, Minhyeok Lee, and Sangyoun Lee. Fastano: Fast anomaly detection via spatio-temporal patch transformation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2249–2259, 2022.

[20] Karishma Pawar and Vahida Attar. Deep learning based detection and localization of road accidents from traffic surveillance videos. *ICT Express*, 8(3):379–387, 2022.

[21] Jing Ren, Feng Xia, Yemeng Liu, and Ivan Lee. Deep video anomaly detection: Opportunities and challenges. In *2021 international conference on data mining workshops (ICDMW)*, pages 959–966. IEEE, 2021.

[22] AR Revathi and Dhananjay Kumar. An efficient system for anomaly detection using deep learning classifier. *Signal, Image and Video Processing*, 11:291–299, 2017.

[23] Nicolae-C Ristea, Florinel-Alin Croitoru, Radu Tudor Ionescu, Marius Popescu, Fahad Shahbaz Khan, Mubarak Shah, et al. Self-distilled masked auto-encoders are efficient video anomaly detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15984–15995, 2024.

[24] Mohammad Sabokrou, Mahmood Fathy, and Mojtaba Hoseini. Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. *Electronics Letters*, 52(13):1122–1124, 2016.

[25] Erkan Şengönül, Refik Samet, Qasem Abu Al-Haija, Ali Alqahtani, Badraddin Alturki, and Abdulaziz A Alsulami. An analysis of artificial intelligence techniques in surveillance video anomaly detection: A comprehensive survey. *Applied Sciences*, 13(8):4956, 2023.

[26] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023.

[27] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.

[28] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *international conference on machine learning*, pages 3560–3569. PMLR, 2017.

[29] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2020.

[30] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[31] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[32] Yang Zhang, Fa Wang, Xin Huang, Xintao Li, Sibei Liu, and Hansong Zhang. Optimization and application of cloud-based deep learning architecture for multi-source data prediction, 2024.