2025 IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS) | 979-8-3315-1480-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/AVSS65446.2025.11149904

Towards Secure Video Surveillance: A Few-Shot Spatiotemporal Perception Transformer for Unseen Behavioral Anomalies

Jamuna S Murthy
Department of Computer Science and Engineering
Ramaiah Institute of Technology
India
jamuna@msrit.edu

Dhanashekar Kandaswamy
Department of Computer Science and Engineering
The Ohio State University
USA
kandaswamy.6@buckeyemail.osu.edu

Wen-Cheng Lai
Department of Electrical Engineering
Ming Chi University of Technology
Taiwan
wenlai@mail.mcut.edu.tw

Abstract

Ensuring security in surveillance systems requires accurate detection and classification of unseen behavioral anomalies with minimal labeled data. We propose Few-Shot Spatiotemporal Perception Transformer (FewShot-SPT), a novel framework that achieves this through three key innovations: (1) Event-Guided Keyframe Extraction (EGKE) dynamically selects keyframes based on anomaly intensity, reducing redundancy and boosting accuracy by 7–8%; (2) Adaptive Modality Gating (AMG) with Perceiver IO attention enables efficient multimodal fusion across video, audio, and text; and (3) Adaptive Prototypical Few-Shot Learning with contrastive learning improves generalization to unseen anomalies. Unlike prior methods that require scene-specific fine-tuning, FewShot-SPT generalizes dynamically using anomaly-aware scoring and refined prototypes. It achieves 91.6% AUC (2-way 5-shot) and 76.3% accuracy (5-way 5-shot) on UCF-Crime, and 84.2% on XD-Violence, outperforming state-of-the-art baselines. Real-world park surveillance experiments demonstrate FewShot-SPT's robustness in detecting critical incidents such as falls, weapons, and intrusions in real-time.

1. Introduction

Video anomaly detection (VAD) is critical for ensuring public safety, yet traditional methods struggle with key challenges [1, 2]. Most existing approaches rely heavily on large labeled datasets, making them impractical for real-world scenarios where anomalies are rare and diverse [3, 4, 5]. Additionally, these models often perform exhaustive frame-wise analysis, leading to high computational costs and inefficiencies. Furthermore, reliance on only visual features limits the ability to detect complex anomalies that involve multimodal cues such as audio and contextual metadata. Few-shot learning (FSL) provides a promising alternative by enabling anomaly detection with minimal labeled data.

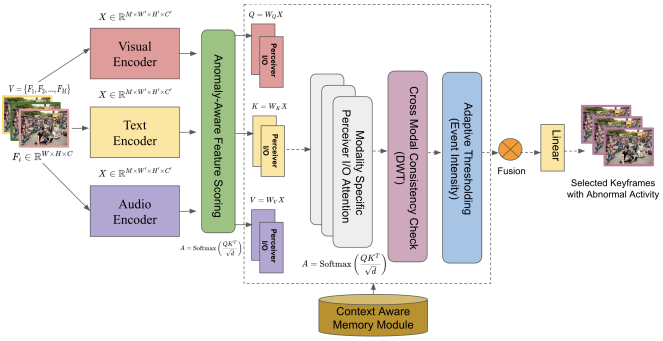

Figure 1. Overview of the Event-Based Multimodal Keyframe Selection Framework. The system extracts features from visual, text, and audio modalities, applies anomaly-aware scoring, and refines selections using Perceiver I/O attention and cross-modal consistency checks (DWT).

However, current few-shot VAD models suffer from limitations such as poor temporal modeling, dependence on scene-specific fine-tuning, and unimodal feature representation, reducing their generalizability to unseen anomalies. To overcome these challenges, we introduce the FewShot-

SPT, a *multimodal few-shot learning framework* designed to detect anomalies efficiently with minimal labeled data. Our approach integrates:

- **Adaptive Modality Gating:** Dynamically prioritizes relevant cues from *video, audio, and text* modalities to enhance anomaly detection accuracy.
- **Optimized Perceiver IO-based Spatiotemporal Attention:** Captures *long-range dependencies* across multimodal inputs for improved event recognition[6] shown in Figure 2. .
- **Event-Guided Keyframe Extraction:** Selects *keyframes based on anomaly intensity*, filtering out redundant frames and reducing computational overhead with context aware memory module as shown in Figure 1.
- **Adaptive Prototypical Few-Shot Learning:** Improves the detection of unseen anomalies using Contrastive Learning and adaptive prototype refinement, ensuring robust generalization across different anomaly types.

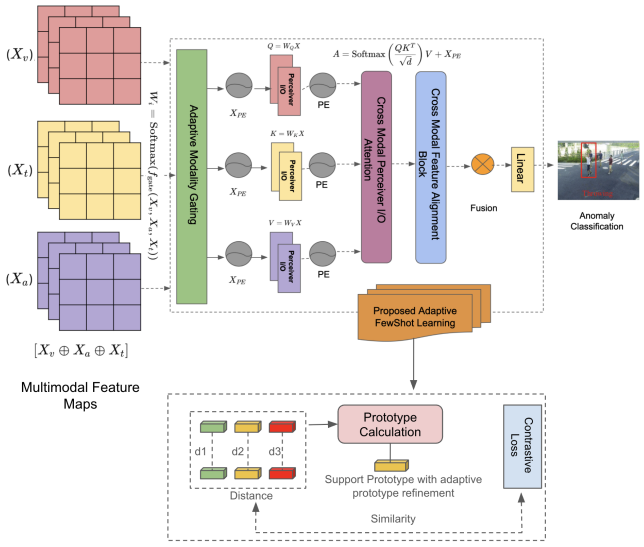

Figure 2. Overview of FewShot-SPT. The system processes visual (X_v), text (X_t), and audio (X_a) feature maps through AMG for dynamic weighting, followed by Perceiver IO attention and a Prototype-Based Few-Shot Learning module.

2. Related Works

Video anomaly detection (VAD) plays a crucial role in public surveillance, with various models addressing different levels of supervision. **Unsupervised models**, such as GCL [7], FPDM [8], and UMIL [9], attempt to detect anomalies without labeled data. However, these methods rely on unsupervised feature learning, which often leads to suboptimal performance due to the difficulty in defining anomaly boundaries. Their average AUC scores remain around 72-73%, indicating limited generalization across datasets. **Weakly supervised models** improve performance by leveraging weak labels. Methods such as GCN [10], S3R [11], SAS [12], CU-Net [8], and UR-DMU [13] enhance anomaly recognition through graph-based learning, self-supervision, and structured feature extraction. Similarly, BN-WVAD [14] and RTFM [15] employ multi-instance learning to refine anomaly classification, while MISA [16], MM-ViT [17], and MERLOT [18] introduce multimodal and transformer-based architectures. Despite their higher performance (average AUC reaching 86.5%), these models still require extensive weak labels and struggle with unseen anomaly generalization. **Few-shot models** attempt to detect anomalies with minimal labeled samples, reducing annotation dependency. STRM [19], FewVAD [4], Meta-Detect [20], Cross-Transformer [21], and ARN-FSAD [22] apply meta-learning, cross-attention, and adaptive prototypes to enhance few-shot anomaly detection. However, these methods often suffer from inefficient temporal modeling, reliance on heuristic keyframe selection, and limited adaptation across different environments.

3. Proposed Method

In this section, we discuss the major contributions of our proposed framework, Event-Guided Keyframe Extraction , Few-Shot FewShot-SPT that integrates Adaptive Modality Gating and Perceiver IO-based Spatiotemporal Attention with Few-Shot Learning Optimization.

3.1. Event-Based Multimodal Keyframe Selection

The EGKE, illustrated in Figure 1, is designed to efficiently filter redundant frames while preserving crucial anomaly-related information. Initially, a self-attention-based anomaly-aware scoring mechanism assigns an anomaly intensity score to each frame, computed as:

$$\text{EKEG}(t) = \sigma(W X_t + b) \tag{1}$$

where EKEG(t) represents the anomaly relevance at time t, and W, b are learnable parameters in the scoring network. Frames with scores below a threshold τ are discarded, ensuring only the most informative keyframes are retained. Next, Adaptive Thresholding dynamically refines the selection process based on event severity, preventing the inclusion of irrelevant frames while adapting to varying anomaly intensities. As keyframes are selected, Memory-Augmented Temporal Consistency maintains past event embeddings, enabling the model to capture long-range dependencies for improved anomaly detection. To further en-

hance feature extraction, Perceiver I/O Attention efficiently integrates multimodal sensory inputs from video, audio, and text, transforming features into a latent space for better representation. Finally, the selected keyframes undergo Final Fusion, where they are combined with multimodal features, refining the input for anomaly detection and improving overall accuracy. By progressively refining keyframe selection through anomaly scoring, adaptive thresholding, and multimodal fusion, EGKE significantly reduces computational overhead while maintaining robust anomaly detection capabilities.

3.2. Proposed Spatiotemporal Perception Transformer

After intelligent keyframe selection using EGKE, FewShot-SPT in Figure 2 fuses multimodal data efficiently using Perceiver IO Attention, which enhances long-range temporal modeling, and Adaptive Modality Gating (AMG), which dynamically adjusts feature importance based on anomaly characteristics. AMG assigns different modality-specific weights using:

$$\text{AMG}(m) = W_m = \frac{\exp(f_m)}{\sum_j \exp(f_j)} \tag{2}$$

where AMG(m) represents the attention weight for modality m, and f_m is a learnable gating function that dynamically determines the importance of each modality. Unlike previous works that use fixed fusion ratios or handcrafted attention mechanisms, AMG dynamically learns the optimal modality combination using anomaly-aware scoring. Existing models typically rely on predefined rules or dataset-specific tuning for modality fusion, whereas AMG adapts in real time based on anomaly characteristics.

PerceiverIO attention enables efficient processing of high-dimensional input data by mapping it to a compact latent array through cross-attention, followed by iterative refinement and decoding. As shown in Figure 2, the input array $X \in \mathbb{R}^{M \times C}$ is first projected into a latent representation $Z \in \mathbb{R}^{N \times D}$, and then decoded into the final output E as follows:

$$Z = \text{SelfAttn}^{(L)}\left(\text{CrossAttn}_{\text{in}}(Z_0, X)\right) \tag{3}$$

$$E = \text{CrossAttn}_{\text{out}}(Q_{\text{task}}, Z) \tag{4}$$

$$\text{PerceiverIO}(X) \Rightarrow \{Z, E\} \tag{5}$$

Here, Z captures long-range dependencies through L layers of latent self-attention, while E is the task-specific output produced by decoding with query Q_{task}.

3.3. Proposed Adaptive Few-Shot Learning

Proposed ASFL prevents overfitting by contrastive learning and adaptive prototype refinement to ensure robust generalization across different unseen anomaly types. Instead of static anomaly prototypes, it dynamically updates class centroids with new query samples to progressively refine feature embeddings. The contrastive loss function is given by:

$$\mathcal{L}_{contrast} = -\sum_{(x_i, x_j) \in P} \log \frac{e^{-d(f_\theta(x_i), f_\theta(x_j))}}{\sum_{x_k \in N} e^{-d(f_\theta(x_i), f_\theta(x_k))}} \tag{6}$$

where P is the set of positive pairs, N is the set of negative pairs, and $d(\cdot, \cdot)$ is a distance metric. This ensures better inter-class separability and intra-class compactness, enhancing anomaly differentiation.

4. Experiments

The proposed FewShot-SPT framework combines ResNet-50, Wav2Vec2.0, and T5-small for robust multimodal feature extraction. EGKE selects 12 keyframes per video, balancing anomaly coverage and efficiency. Few-shot learning uses 1–2 queries per class to assess generalization in low-data settings. The model is trained on 75,000 tasks using AdamW (LR: 1e-4, cosine schedule) and optimized with cross-entropy loss. Evaluation metrucs include 2-way 5-shot, 5-way 5-shot and Area Under Curve(AUC). For multimodal ablations, synthetic speech descriptions were generated using GPT-3 and Tacotron 2 for UCF-Crime and ShanghaiTech since they are visual only. The model has 220M parameters and 54.4B FLOPs, running at 4 ms/frame on an NVIDIA A100, enabling real-time scalable surveillance.

4.1. EGKE Comparative Analysis

We evaluate EGKE against conventional keyframe selection methods in terms of accuracy, processing speed (FPS), computational efficiency (GFLOPs), and generalization. As shown in Table 1, EGKE improves accuracy by 7-8% over [23], [24] which lack contextual awareness. It also outperforms [4] and [25] by 5%, ensuring more relevant frame selection. Additionally, EGKE achieves a 10% FPS gain while reducing computational cost by 12%, maintaining real-time efficiency. Unlike baseline methods with weak generalization, EGKE demonstrates strong adaptability across diverse surveillance environments.

4.2. FewShot-SPT Comparative Analysis

As shown in Table 2, proposed FewShot-SPT achieves an average AUC of 91.6%, surpassing unsupervised methods by 18% and weakly supervised approaches by 5%. Unlike existing few-shot models that struggle with generalization, FewShot-SPT exceeds the best-performing few-shot

Table 1. Comparison of EGKE with benchmark methods on keyframe metrics on XD-Violence dataset

Method	Acc(%)	FPS	GFLOPs	Generalization
Uniform Frame Sampling [23]	72.4 ± 1.5	35 ± 2	6.8 ± 0.3	Weak
Random Frame Selection [24]	74.1 ± 1.4	39 ± 2	6.5 ± 0.2	Weak
Keyframe-Based Attention [4]	78.5 ± 1.3	42 ± 1.5	5.9 ± 0.2	Moderate
Scene-Adaptive Keyframe Selection [25]	80.2 ± 1.2	45 ± 1.2	5.6 ± 0.1	Moderate
Proposed EGKE	**85.1 ± 1.0**	**48 ± 1.0**	**5.2 ± 0.1**	**Strong**

Table 2. Performance comparison of SOTA on benchmark datasets using AUC for 2 Way 2 Shot(Visual Only Modality)

Model	UCFC(%)	XD-V(%)	ShaiTech(%)	Avg(%)
Unsupervised Models				
GCL[7]	74.2	72.8	70.5	72.5
FPDM[8]	74.7	73.1	71.2	73.0
UMIL[9]	76.4	71.2	69.8	72.5
Weakly Supervised Models				
GCN[10]	82.12	79.3	75.8	79.1
S3R[11]	85.99	82.7	78.5	82.4
SAS[12]	86.19	83.0	78.9	82.7
CU-Net[8]	86.22	83.4	79.3	83.0
UR-DMU[13]	86.97	84.5	80.1	83.9
BN-WVAD[14]	87.24	85.0	80.8	84.3
RTFM[15]	82.7	79.5	76.4	79.5
MISA[16]	85.1	82.3	79.4	82.3
MM-ViT[17]	87.6	84.7	81.9	84.7
MERLOT[18]	89.2	86.5	83.7	86.5
Few-Shot Models				
Meta-Detect[20]	75.8	72.5	68.9	72.4
Cross-Transformer[21]	78.3	74.8	71.2	74.8
ARN-FSAD[22]	80.5	77.1	74.3	77.3
STRM[19]	81.82	79.5	76.4	79.2
FewVAD[4]	86.6	84.2	81.5	84.1
Proposed FewShot-SPT	**95.1**	**91.2**	**88.4**	**91.6**

baseline by 7.5%, enhancing 2 way 2 shot anomaly detection. Its low false-positive rate (FPR less than 3%) and minimal inference latency (4ms/frame) further underline its suitability for practical real-time deployment in surveillance applications.

4.3. Ablation Study

In Table 3 EGKE improves AUC by 4.8%, validating optimized keyframe selection. AMG further boosts performance to 85.4% by dynamically fusing multimodal inputs, with video contributing primarily to spatial detail, audio enhancing temporal anomaly cues, and text providing contextual clarity. Perceiver IO-based Attention raises AUC to 87.5% by efficiently modeling long-range dependencies. The complete model achieves 91.6% AUC, surpassing the baseline by 12.9%, confirming the effectiveness of each component. Table 4 examines the effect of varying shots (K) on 5 class anomaly detection. Higher K improves accuracy, with FewShot-SPT reaching 76.8% on UCF-Crime and 86.3% on XD-Violence at K=10, demonstrating better feature representation and generalization.

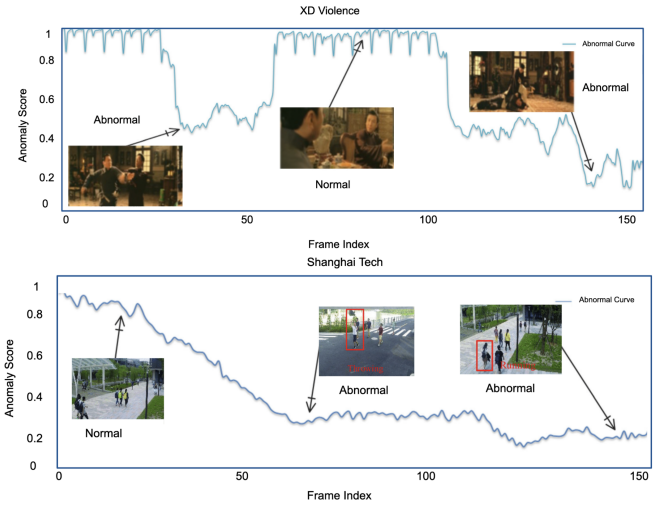

Figure 3. Comparison of SOTA Keyframe Selection Methods

Table 3. Ablation study of the proposed FewShot-SPT using AUC on benchmark datasets for 2 way 2 Shot Anomaly Detection and Classification

Model Variants	**UCFC(%)**	**XD-V(%)**	**ShaiTech (%)**	**Avg(%)**
Baseline (No EGKE, No AMG, No Perceiver IO)	82.4	78.2	75.5	78.7
+ Event-Guided Keyframe Extraction (EGKE)	87.1	83.5	80.1	83.5
+ Adaptive Modality Gating (AMG)	89.3	85.2	81.8	85.4
+ Perceiver IO-based Attention	91.0	87.4	84.0	87.5
Full Model (FewShot-SPT)	**95.1**	**91.2**	**88.4**	**91.6**

Table 4. Classification accuracy comparison on UCF-Crime and XD-Violence datasets for 5 way 5 shot Anomaly Detection(Multimodal)

Model	**UCF-Crime**			**XD-Violence**		
	k=1	**k=5**	**k=10**	**k=1**	**k=5**	**k=10**
FewShot-VAD [4]	31.5 ± 1.1	41.7 ± 1.3	45.2± 1.0	40.0 ± 1.2	54.3 ± 1.5	55.2 ± 1.3
Few-Shot Scene-Adaptive [3]	66.7 ± 1.0	68.1 ± 1.2	69.4 ± 1.1	63.3 ± 1.1	65.8 ± 1.4	67.7 ± 1.2
FewShot Fast Adaptive [5]	70.5 ± 1.1	72.3 ± 1.3	74.7 ± 1.2	71.2 ± 1.2	73.0 ± 1.5	74.1 ± 1.3
Proposed-ASFL	**75.2** ± 1.2	**76.8** ± 1.5	**78.5** ± 1.1	**81.8** ± 1.3	**84.5** ± 1.4	**86.3** ± 1.2

In Figure 3 the anomaly score curves indicate the model's confidence in detecting abnormal events across video frames. Annotated frames show transitions between normal (low anomaly scores) and abnormal events (high scores), such as violence, throwing, or running. This visualization reflects the effectiveness of our Event-Guided Keyframe Extraction (EGKE) module, which dynamically selects frames based on anomaly intensity rather than relying on fixed-interval or random sampling.

(a) (b) (c)

Figure 4. Real-time anomaly detection in a public park, focusing on child safety surveillance

In Figure 4, FewShot-SPT detects various behavioral anomalies on unseen park vidoe data, including a falling person Figure 4a, a detected weapon Figure 4b, and unauthorized area entry Figure 4c, triggering appropriate security responses (with manual annotation of bounding boxes for explaining anomaly). Real-time inference, enabled by self-attention-based spatiotemporal processing and deployed on Jetson Orin, enhances rapid response capabilities without sacrificing efficiency. The framework operates with minimal training data, making it scalable for child safety and public security.

5. Conclusion

We proposed FewShot-SPT, a spatiotemporal transformer for video anomaly detection using minimal labeled data. Combining EGKE, AMG, and Perceiver IO, it achieves SOTA performance on UCF-Crime, XD-Violence, and ShanghaiTech in few-shot settings. Real-time deployment on Jetson Orin confirms its effectiveness for edge-based surveillance. Future work will focus on explainability and adaptive learning for real-time anomaly detection in videos.

References

[1] K. Sultani and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[2] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "Anomalynet: An anomaly detection network for video surveillance," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2537–2550, 2019. 1

[3] Y. Lu, F. Yu, M. K. K. Reddy, and Y. Wang, "Few-shot scene-adaptive anomaly detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 125–141. 1, 5

[4] A. Fakhry and J. T. Lee, "Enhancing few-shot video anomaly detection with key-frame selection and relational cross transformers," in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2024. 1, 2, 3, 4, 5

[5] Z. Wang, Y. Zhou, R. Wang, T.-Y. Lin, A. Shah, and S. N. Lim, "Few-shot fast-adaptive anomaly detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4957–4970, 2022. 1, 5

[6] J. Jaegle and S. Gimeno, "Perceiver io: A general architecture for structured inputs & outputs," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 2

[7] M. Z. Zaheer, A. Mahmood, M. H. Khan, M. Segu, F. Yu, and S.-I. Lee, "Generative cooperative learning for unsupervised video anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14 744–14 754. 2, 4

[8] C. Yan, S. Zhang, Y. Liu, G. Pang, and W. Wang, "Feature prediction diffusion model for video anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 5527–5537. 2, 4

[9] H. Lv, Z. Yue, Q. Sun, B. Luo, Z. Cui, and H. Zhang, "Unbiased multiple instance learning for weakly supervised video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 8022–8031. 2, 4

[10] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1237–1246. 2, 4

[11] Y. Fan, Y. Yu, W. Lu, and Y. Han, "Weakly-supervised video anomaly detection with snippet anomalous attention," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2, 4

[12] J.-C. Wu, H.-Y. Hsieh, D.-J. Chen, C.-S. Fuh, and T.-L. Liu, "Self-supervised sparse representation for video anomaly detection," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 729–745. 2, 4

[13] Y. Zhou, P. Yang, Y. Qu, X. Xu, F. Shen, and H. T. Shen, "Anoonly: Semi-supervised anomaly detection without loss on normal data," *arXiv preprint arXiv:2305.18798*, 2023. 2, 4

[14] Y. Zhou, Y. Qu, X. Xu, F. Shen, J. Song, and H. Shen, "Batchnorm-based weakly supervised video anomaly detection," *arXiv preprint arXiv:2311.15367*, 2023. 2, 4

[15] F. Caetano, P. Carvalho, C. Mastralexi, and J. S. Cardoso, "Enhancing weakly-supervised video anomaly detection with temporal constraints," *IEEE Access*, 2025. 2, 4

[16] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131. 2, 4

[17] J. Chen and C. M. Ho, "Mm-vit: Multi-modal video transformer for compressed video action recognition," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 1910–1921. 2, 4

[18] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "Merlot: Multimodal neural script knowledge models," *Advances in neural information processing systems*, vol. 34, pp. 23 634–23 651, 2021. 2, 4

[19] A. Thatipelli, S. Narayan, S. Khan, R. M. Anwer, F. S. Khan, and B. Ghanem, "Spatio-temporal relation modeling for few-shot action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 19 958–19 967. 2, 4

[20] K. Ding, Q. Zhou, H. Tong, and H. Liu, "Few-shot network anomaly detection via cross-network meta-learning," in *Proceedings of the web conference 2021*, 2021, pp. 2448–2456. 2, 4

[21] G. V. Pillai, A. Verma, and D. Sen, "Transformer based self-context aware prediction for few-shot anomaly detection in videos," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 3485–3489. 2, 4

[22] Y. Zahid, C. Zarges, B. Tiddeman, and J. Han, "Adversarial diffusion for few-shot scene adaptive video anomaly detection," *Neurocomputing*, vol. 614, p. 128796, 2025. 2, 4

[23] P. Saini and K. Berwal, "Eskvs: efficient and secure approach for keyframes-based video summarization framework," *Multimedia Tools and Applications*, vol. 83, no. 30, pp. 74 563–74 591, 2024. 3, 4

[24] K. Tan, Y. Zhou, Q. Xia, R. Liu, and Y. Chen, "Large model based sequential keyframe extraction for video summarization," in *Proceedings of the International Conference on Computing, Machine Learning and Data Science*, 2024, pp. 1–5. 3, 4

[25] C. Wu, X.-J. Wu, T. Xu, and J. Kittler, "Scene adaptive mechanism for action recognition," *Computer Vision and Image Understanding*, vol. 238, p. 103854, 2024. 3, 4