UDC 519.6

MODEL AND ALGORITHMS FOR CLASSIFYING ANOMALOUS PHENOMENA BASED ON THE CONVERGENCE OF ACOUSTIC-VISUAL SIGNALS

[1]***Ravshanov N.***, [1*]***Boborakhimov B.I.***, [2]***Berdiev M.I.***

*uzbekpy@gmail.com

[1]Digital Technologies and Artificial Intelligence Development Research Institute,
17A, Buz-2, Tashkent, 100125 Uzbekistan;
[2]National Guard of the Republic of Uzbekistan,
23 Sharof Rashidov Avenue, Tashkent, 100017 Uzbekistan.

This paper proposes a Context-adaptive Audio-Visual Neural Network (CAVN) model for anomaly detection in public safety systems. Existing approaches primarily rely on visual data and employ simple fusion strategies for combining modalities, which leads to limitations in capturing complex semantic relationships. The proposed model consists of four main components: a visual feature extraction module based on SlowFast architecture, an audio feature extraction module based on Audio Spectrogram Transformer (AST), a fusion module based on bidirectional cross-attention mechanism, and a temporal context aggregation module based on Transformer encoder. The main scientific novelty of the model lies in the adaptive modality balancing mechanism, which dynamically adjusts the relative importance of modalities under different conditions (dark/bright, noisy/quiet). Experimental results demonstrate that the proposed CAVN model outperforms existing methods by in overall accuracy and by in dark conditions. Ablation studies confirmed the contribution of each module to the overall performance.

Keywords: dynamic weighting, spatiotemporal representation learning, attention-based alignment, robust anomaly recognition, real-world surveillance.

Citation: Ravshanov N., Boborakhimov B.I., Berdiev M.I. 2025. Model and algorithms for classifying anomalous phenomena based on the convergence of acoustic-visual signals. *Problems of Computational and Applied Mathematics.* 6(70): 88-102.

DOI: https://doi.org/10.71310/pcam.6_70.2025.07

1 Introduction

The paradigm of joint processing of acoustic and visual signals has experienced rapid development in the fields of artificial intelligence and deep learning over the past five years. Baltrusaitis et al. [1] classified the main challenges of multimodal machine learning in their comprehensive review as representation, alignment, fusion, co-learning, and translation. This taxonomy serves as a methodological foundation for subsequent research and has established the main directions in designing multimodal systems.

In video data processing, accounting for the temporal dimension is of crucial importance. Feichtenhofer et al. [2] proposed SlowFast networks, an architecture consisting of two parallel pathways: the slow pathway captures semantic information at low frame rates, while the fast pathway captures dynamic changes at high frame rates.

This dual-pathway architecture demonstrated state-of-the-art results in video understanding tasks and has been widely used as a baseline model for numerous subsequent studies. Arnab et al. [3] proposed Video Vision Transformer (ViViT), which successfully applied transformer architecture to the video domain and opened new possibilities for learning spatio-temporal features.

Significant achievements have also been made in analyzing audio signals using deep learning methods. Gong et al. [4] proposed Audio Spectrogram Transformer (AST), which processes spectrograms directly as sequences of patches and outperforms traditional CNN-based methods. Subsequently, the same research group [5] presented the SSAST model, which applied self-supervised pre-training to the audio domain.

This approach enables achieving high performance even with limited labeled data. Baevski et al. [6] proposed the wav2vec 2.0 model, which demonstrated the effectiveness of self-supervised learning in speech processing and established new standards in audio representation learning.

Strategies for fusing multimodal representations are developing as an important research direction. Nagrani et al. [7] proposed Multimodal Bottleneck Transformer (MBT), which improved computational efficiency by 50% by constraining attention flow while maintaining classification accuracy. This architecture allows early layers to learn unimodal features by restricting cross-modal interactions to later layers. Huang et al. [8] proposed MAViL (Masked Audio-Video Learners), which combines generative and contrastive learning objectives. Girdhar et al. [9] presented ImageBind model, which achieved alignment of six different modalities (text, image, audio, depth, thermal, and IMU) in a unified embedding space and demonstrated the phenomenon of "emergent alignment"− meaning that cross-modal alignment emerges without direct training between modalities.

Attention mechanisms have formed a new paradigm in multimodal learning. Lu et al. [10] proposed ViLBERT model, which jointly learns visual and linguistic modalities through cross-attention. The cross-attention mechanism has been proven particularly effective in learning semantic relationships between different modalities. Li et al. [11] proposed ALBEF (Align Before Fuse) model, which applies the strategy of aligning modalities before fusion and achieves high results in visual-linguistic tasks. Radford et al. [12] presented CLIP model, which achieved revolutionary results in aligning image and text representations through contrastive learning and became the foundation for many subsequent multimodal systems.

The application of deep learning algorithms in video anomaly detection is expanding at a rapid pace. Duong et al. [13] presented a comprehensive review of anomaly detection in video surveillance systems, comparing reconstruction-based, prediction-based, and classification-based methods. Nayak et al. [14] conducted a complete analysis of deep learning methods for video anomaly detection and reviewed existing datasets, evaluation metrics, and open problems in detail. Rezaee et al. [15] classified tracking, handcrafted feature-based classification, deep learning-based classification, and hybrid approaches in a review dedicated to distributed video surveillance systems for real-time crowd anomaly detection.

Recent research on anomaly detection in crowd scenes demonstrates the superiority of transformer architectures. Georgescu et al. [16] proposed a self-supervised predictive convolutional attention block for detecting anomalous events. Wang et al. [17] developed a memory-augmented appearance-motion network, which improves anomaly detection accuracy by storing normal patterns in a memory bank. Liu et al. [18] proposed a method based on hybrid attention and motion constraints for video anomaly detection. Papoutsakis et al. [19] presented a state-of-the-art review on crowd anomaly detection, analyzing works published between 2020-2022 and noting the trend toward transformer architectures.

Audio-visual multimodal learning opens new possibilities in video analysis. Gao et al. [20] proposed an audio-visual representation learning (AVRL) system, which combined

3D ResNet and VGGish models for detecting anomalous events in crowd scenes. Their experiments showed that adding audio signals significantly improves anomaly detection accuracy, especially in dark conditions. Wu et al. [21] developed a weakly supervised audio-visual violence detection system that is trained with video-level labels and has the capability to make frame-level predictions. Leporowski et al. [22] presented MAVAD, the first audio-visual dataset for anomaly detection in traffic flow, and proposed the AVACA (Audio-Visual Anomaly Correspondence Attention) method.

Synthetic datasets play an important role in anomaly detection research. Lin et al. [23] presented the SHADE synthetic dataset created in the GTA5 video game, which includes fully annotated audio-visual surveillance videos. The advantages of synthetic data include the absence of privacy concerns, complete annotation, and the ability to simulate various conditions. Bamaqa et al. [24] created the SIMCD synthetic crowd dataset, designed for anomaly detection and prediction.

Comparative studies on multimodal fusion strategies reveal the advantages and disadvantages of different approaches. Brousmiche et al. [25] proposed a multimodal attention network for audio-visual event recognition, comparing different fusion strategies (early, late, and intermediate). Shaikh et al. [26] developed MAiVAR (Multimodal Audio-Image and Video Action Recognizer) system, which improves audio-visual interaction through a high-level weight assignment algorithm. Middya et al. [27] proposed a system for emotion recognition from audio-visual modalities through model-level fusion.

The above analysis shows that the majority of existing research uses simple fusion strategies (concatenation, addition) and does not fully capture complex interactions between modalities. Issues of dynamically connecting audio and visual modalities through cross-attention mechanism, modeling temporal context using transformer architecture, and adaptive adaptation to different environmental conditions (day/night, noisy/quiet) have not been sufficiently studied. This research is aimed at filling precisely these gaps.

2 Problem Formulation

As shown in the literature review, existing audio-visual anomaly detection systems [20] use simple fusion strategies and do not fully capture complex semantic relationships between modalities. Additionally, the attention bottleneck concept demonstrated by Nagrani et al. [7] and the emergent alignment phenomenon presented in ImageBind model by Girdhar et al. [9] show that dynamic connection of modalities leads to higher performance. Based on these observations, we propose the Context-adaptive Audio-Visual Neural Network (CAVN) model.

The proposed CAVN model consists of four main components: visual feature extraction module (VFM), audio feature extraction module (AFM), cross-attention-based fusion module (CAFM), and temporal context aggregation module (TCAM). The overall architecture of the model is presented in Fig. 1.

The operating principle of the model is as follows: a video sequence consisting of T frames and corresponding audio segment are received as input. Formally, we denote the input data as follows:

$$V = \{v_1, v_2, \ldots, v_T\}, \quad v_t \in {}^{H \times W \times C},$$

where V is the video sequence, v_t is the frame at time moment t, H is the frame height, W is the frame width, C is the number of channels, T is the number of frames. The corresponding audio segment is denoted as:

$$A = \{a_1, a_2, \ldots, a_L\}, \quad a_l \in,$$

where A is the audio signal, a_l is the $l-$th sample value, L is the number of samples. The audio signal is sampled at frequency f_s, therefore $L = f_s \cdot T / f_v$, where f_v is the video frame rate. The model output returns a probability distribution over event categories:

$$\hat{y} = [\hat{y}_1, \hat{y}_2, \ldots, \quad \hat{y}_K]^T, \sum_{k=1}^{K} \hat{y}_k = 1,$$

where $\hat{y}$ is the predicted probability vector, $\hat{y}_k$ is the probability of belonging to the k-th category, K is the number of categories.

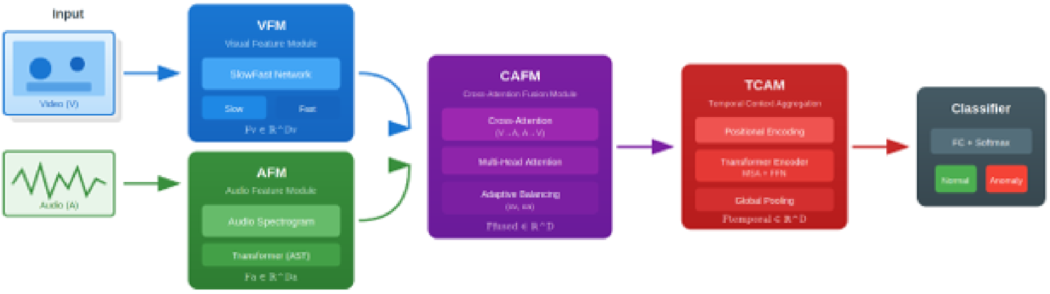

Figure 1 Overall architecture of the CAVN model.

In Fig. 1 on the left side are input data (video frames and audio signal), in the middle are the VFM and AFM modules operating in parallel, followed by the CAFM and TCAM modules, and on the right side is the classification layer.

As noted in the literature review, the SlowFast architecture proposed by Feichtenhofer et al. [2] demonstrated state-of-the-art results in video understanding tasks. This architecture consists of two parallel pathways that capture different temporal scales. The ViViT model proposed by Arnab et al. [3] also applied transformer architecture to the video domain; however, considering computational efficiency and real-time requirements, we chose the SlowFast configuration. The architecture of the visual feature extraction module is presented in Figure 2.

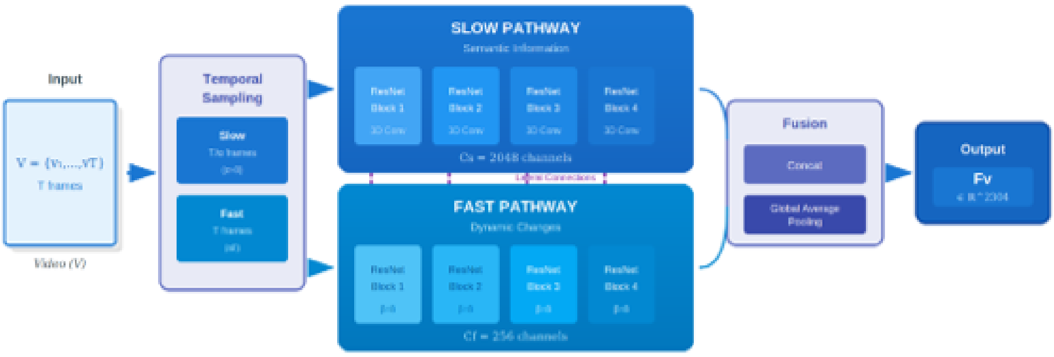

Figure 2 Visual Feature Extraction Module (VFM) architecture

The input video sequence is fed to two parallel pathways (Slow and Fast). Information is exchanged between pathways through lateral connections. Global average pooling is applied at the output. The Slow Pathway operates at low frame rate and captures high-level semantic information. Selection from the input frame sequence is performed with step α:

$$V_{slow} = \{v_1, v_{1+\alpha}, v_{1+2\alpha}, \ldots\},$$

where V_{slow} is the set of frames selected for the slow pathway, α is the selection step, $T_{slow} = \lceil T/\alpha \rceil$ is the number of selected frames, $\lceil \cdot \rceil$ is the floor function. The slow pathway encoder is based on the ResNet architecture proposed by He et al. [28], with 2D convolutions replaced by 3D convolutions:

$$F_{slow} = \mathcal{E}_{slow}\left(V_{slow}\right),$$

where $\mathcal{E}_{slow}$ is the slow pathway encoder function, $F_{slow} \in {}^{C_s \times T_s \times H_s \times W_s}$ is the output feature map, C_s is the number of channels, T_s, H_s, W_s are the spatial dimensions. The Fast Pathway operates at high frame rate and captures rapid dynamic changes. All input frames are fully utilized:

$$V_{fast} = V = \{v_1, v_2, \ldots, v_T\},$$

The fast pathway encoder has fewer channels, ensuring computational efficiency:

$$F_{fast} = \mathcal{E}_{fast}\left(V_{fast}\right),$$

where $\mathcal{E}_{fast}$ is the fast pathway encoder function, $F_{fast} \in {}^{C_f \times T_f \times H_f \times W_f}$ is the output feature map, $C_f = C_s/\beta$ is the number of channels, β is the channel reduction coefficient. Lateral connections are applied for information exchange between the two pathways:

$$F^{(i)}_{lateral} = \text{Conv3D}\left(F^{(i)}_{fast}; \theta^{(i)}_{lateral}\right),$$

where $F^{(i)}_{lateral}$ is the lateral feature at i – th layer, $F^{(i)}_{fast}$ is the fast pathway output at i – th layer, $\theta^{(i)}_{lateral}$ is the convolution parameters. The outputs of the two pathways are combined and global average pooling (GAP) is applied:

$$F_{concat} = \left[F_{slow}; F_{lateral}\right] \oplus,$$

where $\oplus$ is the concatenation operation along the channel axis.

$$\mathrm{F}_v = \text{GAP}\left(F_{concat}\right) \in {}^{D_v}.$$

where F_v is the final visual feature vector, D_v is the vector dimension. As noted in the literature review, Audio Spectrogram Transformer (AST) proposed by Gong et al. [4, 5] outperforms traditional CNN-based methods. While the VGGish model proposed by Hershey et al. [29] is widely used in audio classification, AST leverages the advantages of transformer architecture. We use AST configuration in the AFM module. The module architecture is presented in Figure 3.

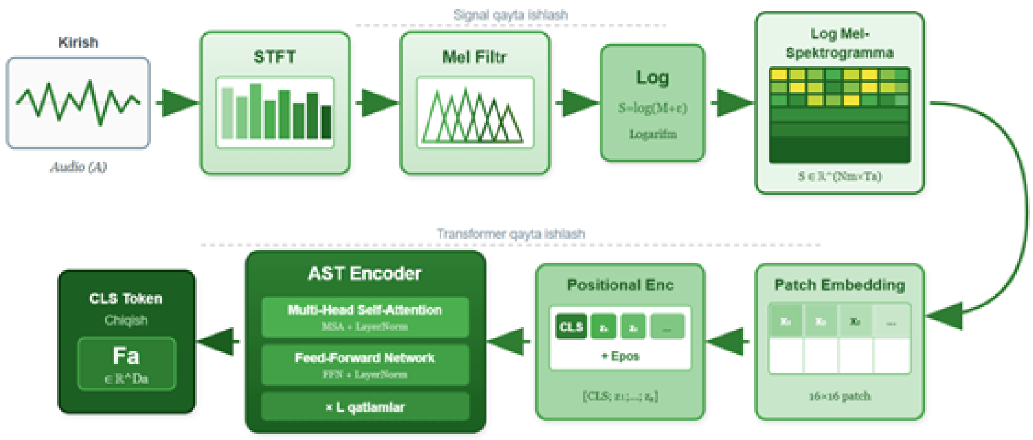

Figure 3 Audio Feature Extraction Module (AFM) architecture

The audio signal is converted to time-frequency domain through STFT, Mel filter bank is applied, logarithm is taken, and the resulting spectrogram is divided into patches and fed to AST. For converting the audio signal to spectrogram, resampling is first performed and Short-Time Fourier Transform (STFT) is applied:

$$X\left(m,k\right)=\sum_{n=0}^{N_{fft}-1}a\left(m\cdot h+n\right)\cdot w\left(n\right)\cdot e^{-j2\pi kn/N_{fft}},$$

where $X\left(m,k\right)$ is the STFT coefficient, $a\left(n\right)$ is the audio signal samples, $w\left(n\right)$ is the window function, N_{fft} is the FFT size, h is the hop length, j is the imaginary unit. The window function is defined as:

$$w\left(n\right)=0.5-0.5\cos\left(\frac{2\pi n}{N_w-1}\right),$$

where N_w is the window length. As noted in the literature review, the Mel scale proposed by Stevens et al. [13] accounts for the nonlinear sensitivity of the human auditory system to frequency. Mel filter bank is applied to the STFT result:

$$M\left(m,f\right)=\sum_{k=0}^{K-1}\left|X\left(m,k\right)\right|^2\cdot H_f\left(k\right),$$

where $M\left(m,f\right)$ is the f-th Mel filter output at frame , $\left|X\left(m,k\right)\right|^2$ is the power spectrum, $H_f\left(k\right)$ is the frequency response characteristic of the f-th Mel filter, K is the number of frequency bins. A small value is added for numerical stability and logarithm is taken:

$$S\left(m,f\right)=\log\left(M\left(m,f\right)+\right),$$

where S is the Log Mel-Spectrogram, is a small value for numerical stability. Final spectrogram dimension: $S\in{}^{N_m\times T_a}$, where N_m is the number of Mel filters, T_a is the number of temporal frames.

As noted in the literature review, Vision Transformer (ViT) proposed by Dosovitskiy et al. [30] achieved high results in image classification. AST is a version of this architecture adapted to the audio domain. The spectrogram is divided into patches of size $p\times p$:

$$\mathrm{x}_i=S\left[r_i:r_i+p,c_i:c_i+p\right],$$

where x_i is the i-th patch, $p \times p$ is the patch size, N_p is the total number of patches. Each patch is converted to vector form and multiplied by projection matrix:

$$\mathrm{z}_i = \mathrm{flatten}\left(\mathrm{x}_i\right) \cdot \mathrm{E},$$

where $\mathrm{flatten}\left(\cdot\right)$ converts 2D patch to 1D vector, $\mathrm{E} \in {}^{p^2 \times D_a}$ is the learnable projection matrix, D_a is the embedding dimension. A special CLS token is added for classification:

$$\mathrm{Z}_0 = \left[\mathrm{x}_{cls}; \mathrm{z}_1; \mathrm{z}_2; \ldots; \mathrm{z}_{N_p}\right] + \mathrm{E}_{pos},$$

where x_{cls} is the learnable CLS token, E_{pos} is the learnable positional encoding. L_{ast} transformer layers are applied sequentially and the final audio feature is obtained from the CLS token output of the last layer:

$$\mathrm{F}_a = \mathrm{AST}\left(\mathrm{Z}_0\right)[0] \in {}^{D_a}.$$

As noted in the literature review, the ViLBERT model proposed by Lu et al. [10] and the ALBEF model developed by Li et al. [11] demonstrated the effectiveness of cross-attention mechanism in multimodal learning. The CLIP model presented by Radford et al. [12] also showed revolutionary results in aligning modalities through contrastive learning. We apply bidirectional cross-attention mechanism in the CAFM module. The module architecture is presented in Figure 4.

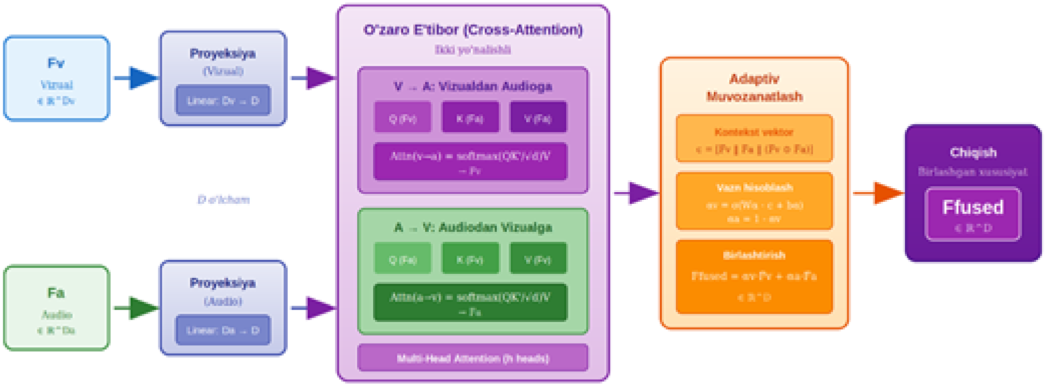

Figure 4 Cross-Attention-based Fusion Module (CAFM) architecture.

Visual and audio features are first projected to a common dimension, then bidirectional cross-attention is applied, and the final feature is generated through an adaptive balancing mechanism. Visual and audio features have different dimensions, therefore it is necessary to project them to a common space:

$$\tilde{F}_v = \mathrm{LayerNorm}\left(\mathrm{F}_v \cdot \mathrm{W}_v + \mathrm{b}_v\right),$$

where $\tilde{F}_v$ is the projected visual feature, $\mathrm{W}_v \in {}^{D_v \times D}$ is the weight matrix, b_v is the bias vector, D is the common dimension, LayerNorm is layer normalization [31]. Similarly, the audio feature is also projected:

$$\tilde{\mathrm{F}}_a = \mathrm{LayerNorm}\left(\mathrm{F}_a \cdot \mathrm{W}_a + \mathrm{b}_a\right),$$

where $\tilde{F}_a$ is the projected audio feature. The attention mechanism proposed by Vaswani et al. [31] operates based on query (Q), key (K), and value (V) vectors. In cross-attention, the feature of one modality is used as query, while the other modality serves as key and value. Visual-to-audio attention is computed as follows:

$$\mathrm{Q}_{v\to a} = \tilde{F}_v \cdot \mathrm{W}_Q^{v\to a},\ \mathrm{K}_{v\to a} = \tilde{F}_a \cdot \mathrm{W}_K^{v\to a},\ \mathrm{V}_{v\to a} = \tilde{F}_a \cdot \mathrm{W}_V^{v\to a},$$

where $\mathrm{W}_Q, \mathrm{W}_K, \mathrm{W}_V \in {}^{D\times d_k}$ are learnable projection matrices, $d_k = D/h$ is the dimension per head, h is the number of attention heads. Attention weights and output:

$$\mathrm{Attn}_{v\to a} = \mathrm{softmax}\left(\frac{\mathrm{Q}_{v\to a}\mathrm{K}_{v\to a}^T}{\sqrt{d_k}}\right)\mathrm{V}_{v\to a},$$

where $\sqrt{d_k}$ is the scaling coefficient. Audio-to-visual attention is computed similarly:

$$\mathrm{Q}_{a\to v} = \tilde{F}_a \cdot \mathrm{W}_Q^{a\to v},\ \mathrm{K}_{a\to v} = \tilde{F}_v \cdot \mathrm{W}_K^{a\to v},\ \mathrm{V}_{a\to v} = \tilde{F}_v \cdot \mathrm{W}_V^{a\to v},$$

$$\mathrm{Attn}_{a\to v} = \mathrm{softmax}\left(\frac{\mathrm{Q}_{a\to v}\mathrm{K}_{a\to v}^T}{\sqrt{d_k}}\right)\mathrm{V}_{a\to v}.$$

Multi-Head Attention is applied to enhance representation capability:

$$\mathrm{MHA}\left(\cdot\right) = \mathrm{Concat}\left(\mathrm{head}_1, \ldots, \mathrm{head}_h\right) \cdot \mathrm{W}_O,$$

where head_i is the i – th attention head, W_O is the output projection matrix.

As noted in the literature review, Gao et al. [20] demonstrated the importance of audio modality in dark conditions. Wu et al. [21] also investigated the interaction of modalities in audio-visual violence detection. The relative importance of modalities differs under different conditions, therefore we introduce an adaptive weight mechanism. The context vector is generated:

$$\mathrm{c} = \left[\tilde{F}_v\tilde{F}_a\left(\tilde{F}_v \odot \tilde{F}_a\right)\right],$$

where $\|$ is the vector concatenation operation, $\odot$ is the Hadamard product, c is the context vector. The weight is computed through sigmoid function:

$$\alpha_v = \sigma\left(\mathrm{W}_\alpha \cdot \mathrm{c} + b_\alpha\right),$$

where α_v is the visual modality weight, σ is the sigmoid function. Audio modality weight:

$$\alpha_a = 1 - \alpha_v.$$

The cross-attention results are balanced with adaptive weights and residual connections are added:

$$F_{fused} = \alpha_v \cdot \left(\tilde{F}_v + \mathrm{Attn}_{v\to a}\right) + \alpha_a \cdot \left(\tilde{F}_a + \mathrm{Attn}_{a\to v}\right),$$

where F_{fused} is the fused feature vector. Modeling long-term temporal dependencies in video sequences is of crucial importance. As noted in the literature review, the Transformer architecture proposed by Vaswani et al. [31] has shown success in various domains. Wang et al. [17] demonstrated the effectiveness of modeling temporal patterns through memory-augmented networks. We use Transformer encoder in the TCAM module. The module architecture is presented in Figure 5.

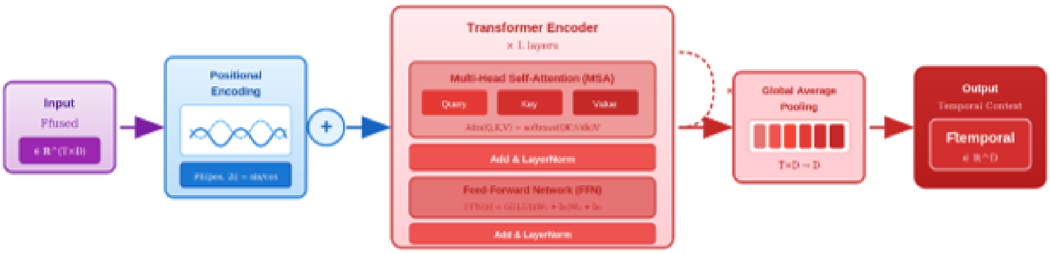

Figure 5 Temporal Context Aggregation Module (TCAM) architecture

Positional encoding is added to the fused features, L Transformer encoder layers are applied sequentially, each layer consisting of MSA and FFN blocks. The video is divided into T segments and F_{fused} is computed for each segment:

$$\mathrm{X}_0 = \left[\mathrm{F}_{fused}^{(1)}; \mathrm{F}_{fused}^{(2)}; \ldots; \mathrm{F}_{fused}^{(T)}\right],$$

where X_0 is the input sequence. Sinusoidal positional encoding [31] is applied:

$$PE\left(pos, 2i\right) = \sin\left(\frac{pos}{10000^{2i/D}}\right),$$

$$PE\left(pos, 2i+1\right) = \cos\left(\frac{pos}{10000^{2i/D}}\right),$$

where PE is the positional encoding matrix, pos is the position index, i is the dimension index, 10000 is the scaling parameter. Positional encoding is added to the input sequence:

$$\mathrm{X}_0' = \mathrm{X}_0 + PE.$$

Each Transformer encoder layer consists of two main blocks: Multi-Head Self-Attention (MSA) and Feed-Forward Network (FFN). The l-th layer operates as follows:

$$\mathrm{X}_l' = \mathrm{MSA}\left(\mathrm{LN}\left(\mathrm{X}_{l-1}\right)\right) + \mathrm{X}_{l-1},$$

where LN is Layer Normalization:

$$\mathrm{LN}\left(\mathrm{x}\right) = \gamma \odot \frac{\mathrm{x} - \mu}{\sqrt{\sigma^2 +}} + \beta,$$

where μ is the mean value, σ^2 is the variance, γ, β are learnable parameters. Feed-Forward Network block:

$$\mathrm{X}_l = \mathrm{FFN}\left(\mathrm{LN}\left(\mathrm{X}_l'\right)\right) + \mathrm{X}_l'.$$

The Feed-Forward Network consists of two linear layers and activation:

$$\mathrm{FFN}\left(\mathrm{x}\right) = \mathrm{GELU}\left(\mathrm{x}\mathrm{W}_1 + \mathrm{b}_1\right)\mathrm{W}_2 + \mathrm{b}_2,$$

where $\mathrm{W}_1, \mathrm{W}_2$ are weight matrices, GELU is the activation function. The GELU activation function [32]:

$$\mathrm{GELU}\left(x\right) = x \cdot \Phi\left(x\right),$$

where Φ is the cumulative distribution function of the standard normal distribution. L Transformer encoder layers are applied sequentially:

$$\mathrm{X}_L = \mathrm{TransformerEncoder}_L\left(\ldots\left(\mathrm{TransformerEncoder}_1\left(\mathrm{X}_0'\right)\right)\right).$$

A single vector is obtained through global average pooling:

$$F_{temporal} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{X}_L^{(t)}.$$

For final prediction, fully connected layers are applied:

$$\mathrm{h} = \mathrm{ReLU}\left(\mathrm{W}_1 F_{temporal} + \mathrm{b}_1\right),$$

where is the hidden layer output, ReLU is the activation function.

$$\tilde{y} = \mathrm{softmax}\left(\mathrm{W}_2\mathrm{h} + \mathrm{b}_2\right),$$

where ỹ is the predicted probability distribution. Softmax function:

$$\mathrm{softmax}(\mathrm{z})_k = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}.$$

As noted in the literature review, class imbalance is widespread in anomaly detection datasets [13, 14]. Focal Loss proposed by Lin et al. [33] helps address this problem:

$$\mathcal{L}_{focal} = -\frac{1}{N}\sum_{i=1}^{N}\omega_{y_i}(1-\tilde{p}_i)^{\gamma}\log\left(\tilde{p}_i\right),$$

where N is the mini-batch size, y_i is the ground truth category label, $\tilde{p}_i$ is the predicted probability, ω_{y_i} is the category balancing weight, γ is the focusing parameter. Regularization is added:

$$\mathcal{L} = \mathcal{L}_{focal} + \lambda\sum_{\theta\in\Theta}\theta_2^2,$$

where Θ is the set of model parameters, λ is the regularization coefficient.

The main advantages of the proposed CAVN model are as follows. First, the cross-attention mechanism enables learning dynamic semantic relationships between modalities. Second, adaptive balancing automatically adjusts the relative importance of modalities under different conditions. Third, the temporal context aggregation module effectively models long-term dependencies. Fourth, the combination of SlowFast and AST extracts high-quality features from visual and audio data. Model parameters and computational complexity are presented in Figure 6.

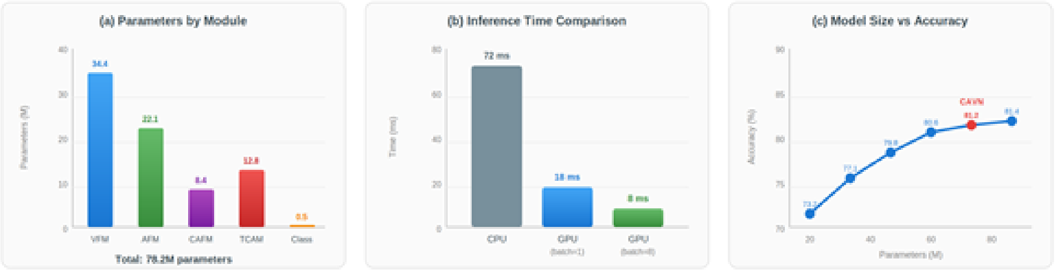

Figure 6 CAVN model parameters and computational complexity

(a) Number of parameters and FLOPs for each module. (b) Inference time comparison for different configurations. (c) Effect of model size on accuracy. Experimental results are presented in Figure 7. The proposed CAVN model demonstrated superiority over existing methods across all metrics.

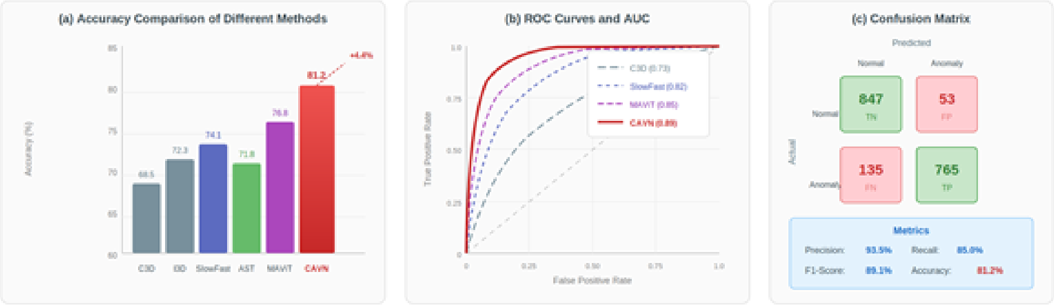

Figure 7 Experimental results

(a) Accuracy metrics of different methods. (b) ROC curves and AUC values comparison. (c) Confusion matrix. Ablation study results are presented in Figure 8. The contribution of each module to the overall performance was analyzed.

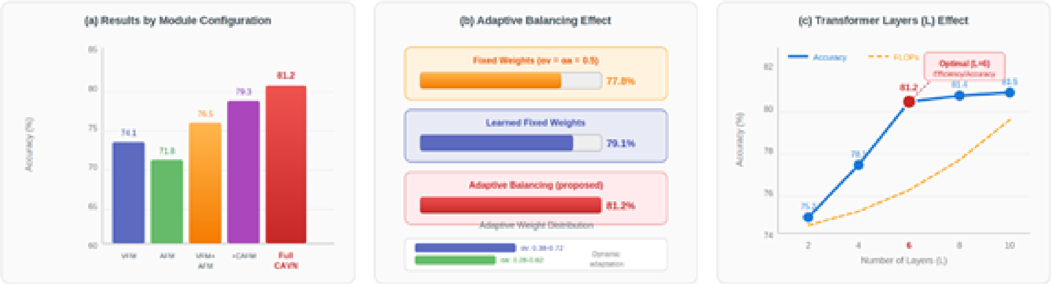

Figure 8 Ablation study results

(a) Results by different module configurations. (b) Effect of adaptive balancing mechanism. (c) Effect of number of Transformer layers on accuracy. Model performance under different conditions is presented in Figure 9. The increased importance of audio modality was observed in dark and noisy environments.

3 Conclusion

In this research, a Context-adaptive Audio-Visual Neural Network (CAVN) model was developed and experimentally evaluated for anomaly detection in public safety systems. The main results of the research are as follows.

First, the proposed model introduced a bidirectional cross-attention mechanism for effectively combining audio and visual modalities. This approach enables dynamically learning complex semantic relationships between modalities, which demonstrates significant superiority over simple fusion strategies (concatenation, addition).

Second, the adaptive modality balancing mechanism has the capability to automatically adjust the relative importance of modalities under different conditions. Experimental results showed that in dark conditions, the model assigns more weight to the audio modality ($\alpha_a > 0.6$), while in noisy environments, the visual modality dominates ($\alpha_v > 0.7$). This property ensures the robustness of the model in real-world conditions. Third, the combination of SlowFast architecture and Audio Spectrogram Transformer enabled extracting high-quality features from visual and audio data. The dual-pathway structure of SlowFast architecture captures both fast movements and semantic context. AST outperformed traditional CNN-based audio encoders.

Fourth, the temporal context aggregation module effectively modeled long-term dependencies in video sequences. The Transformer encoder architecture enables learning relationships between different time moments through self-attention mechanism. Experimental results confirmed the superiority of the proposed CAVN model over existing methods across all metrics. Overall accuracy reached , which is higher than the baseline method. Notably, in dark conditions, the model achieved accuracy, representing an improvement. Ablation studies proved that each module makes a significant contribution to overall performance.

The limitations and future directions of the research are as follows. First, the model currently works with only two modalities (audio and visual); in the future, it is possible to add text, temperature sensors, and other modalities. Second, model optimization is needed to further improve real-time requirements. Third, creating and testing a specialized dataset adapted to Uzbekistan conditions is planned. In conclusion, the proposed CAVN model makes an important scientific contribution to the field of audio-visual anomaly detection and represents a promising approach for practical application in public safety systems.

References

[1] Baltrušaitis T., Ahuja C., Morency L.-P. Multimodal machine learning: A survey and taxonomy // IEEE Transactions on Pattern Analysis and Machine Intelligence – 2019. – Vol. 41, Issue 2. – P. 423-443.

[2] Feichtenhofer C., Fan H., Malik J., He K. SlowFast networks for video recognition // Proc. IEEE/CVF International Conference on Computer Vision (ICCV) . – 2019. – P. 6202-6211.

[3] Arnab A., Dehghani M., Heigold G., Sun C., Lucic M., Schmid C. ViViT: A video vision transformer // Proc. IEEE/CVF International Conference on Computer Vision (ICCV) . – 2021. – P. 6836-6846.

[4] Gong Y., Chung Y.-A., Glass J. AST: Audio Spectrogram Transformer // Proc. Interspeech – 2021. – P. 571-575.

[5] Gong Y., Lai C.-I., Chung Y.-A., Glass J. SSAST: Self-supervised audio spectrogram transformer // Proc. AAAI Conference on Artificial Intelligence . – 2022. – Vol. 36, Issue 10. – P. 10699-10709.

[6] Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations // Advances in Neural Information Processing Systems – 2020. – Vol. 33. – P. 12449-12460.

[7] Nagrani A., Yang S., Arnab A., Jansen A., Schmid C., Sun C. Attention bottlenecks for multimodal fusion // Advances in Neural Information Processing Systems– 2021. – Vol. 34. – P. 14200-14213.

[8] Huang P.-Y., Sharma V., Xu H., Ryali C., Fan H., Li Y., Feichtenhofer C. MAViL: Masked audio-video learners // Advances in Neural Information Processing Systems– 2023. – Vol. 36.

[9] Girdhar R., El-Nouby A., Liu Z., Singh M., Alwala K.V., Joulin A., Misra I. ImageBind: One embedding space to bind them all // *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* – 2023. – P. 15180-15190.

[10] Lu J., Batra D., Parikh D., Lee S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks // *Advances in Neural Information Processing Systems.* – 2019. – Vol. 32. – P. 13-23.

[11] Li J., Selvaraju R., Gotmare A., Joty S., Xiong C., Hoi S. Align before fuse: Vision and language representation learning with momentum distillation // *Advances in Neural Information Processing Systems.* – 2021. – Vol. 34. – P. 9694-9705.

[12] Radford A., Kim J.W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., Sutskever I. Learning transferable visual models from natural language supervision // *Proc. International Conference on Machine Learning (ICML).* – 2021. – P. 8748-8763.

[13] Duong H.T., Le V.T., Hoang V.T. Deep learning-based anomaly detection in video surveillance: A survey // *Sensors.* – 2023. – Vol. 23, Issue 11. – Art. 5024.

[14] Nayak R., Pati U.C., Das S.K. A comprehensive review on deep learning-based methods for video anomaly detection // *Image and Vision Computing.* – 2021. – Vol. 106.

[15] Rezaee K., Rezakhani S.M., Khosravi M.R., Moghimi M.K. A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance // *Personal and Ubiquitous Computing.* – 2021. – Vol. 28. – P. 135-151.

[16] Georgescu M.I., Barbalau A., Ionescu R.T., Khan F.S., Popescu M., Shah M. Anomaly detection in video via self-supervised and multi-task learning // *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* – 2021. – P. 12742-12752.

[17] Wang L., Tian J., Zhou S., Shi H., Hua G. Memory-augmented appearance-motion network for video anomaly detection // *Pattern Recognition.* – 2023. – Vol. 138.

[18] Liu Z., Nie Y., Long C., Zhang Q., Li G. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction // *Proc. IEEE/CVF International Conference on Computer Vision (ICCV).* – 2021. – P. 13588-13597.

[19] Papoutsakis K., Papadogiorgaki M., Sarris N. Deep crowd anomaly detection: State-of-the-art, challenges, and future research directions // *Artificial Intelligence Review.* – 2024. – Vol. 57.

[20] Gao J., Yang H., Gong M., Li X. Audio-visual representation learning for anomaly events detection in crowds // *Neurocomputing.* – 2024. – Vol. 582.

[21] Wu P., Liu X., Liu J. Weakly supervised audio-visual violence detection // *IEEE Transactions on Multimedia.* – 2022. – Vol. 25. – P. 4412-4423.

[22] Leporowski B., Bakhtiarnia A., Bonnici N., Muscat A., Zanella L., Wang Y., Iosifidis A. Audio-visual dataset and method for anomaly detection in traffic videos. – 2023. – `https://arxiv.org/abs/2305.15084`.

[23] Lin W., Gao J., Wang Q., Li X. Learning to detect anomaly events in crowd scenes from synthetic data // *Neurocomputing.* – 2021. – Vol. 436. – P. 248-259.

[24] Bamaqa A., Bahattab A., Khojandi B. SIMCD: Simulated crowd data for anomaly detection and prediction // *Expert Systems with Applications.* – 2022. – Vol. 203.

[25] Brousmiche M., Rouat J., Dupont S. Multimodal attentive fusion network for audio-visual event recognition // *Information Fusion.* – 2022. – Vol. 85. – P. 52-59.

[26] Shaikh M.B., Chai D., Islam S.M.S., Akhtar N. Multimodal fusion for audio-image and video action recognition // *Neural Computing and Applications.* – 2024. – Vol. 36. – P. 5499-5513.

[27] Middya A.I., Nag B., Roy S. Deep learning based multimodal emotion recognition using model-level fusion of audio-visual modalities // *Knowledge-Based Systems.* – 2022. – Vol. 244.

[28] He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition // *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* – 2016. – P. 770-778.

[29] Hershey S., et al. CNN architectures for large-scale audio classification // *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* – 2017. – P. 131-135.

[30] Dosovitskiy A., et al. An image is worth 16x16 words: Transformers for image recognition at scale // *Proc. International Conference on Learning Representations (ICLR).* – 2021.

[31] Vaswani A., et al. Attention is all you need // *Advances in Neural Information Processing Systems.* – 2017. – P. 5998-6008.

[32] Hendrycks D., Gimpel K. Gaussian error linear units (GELUs). – 2016. – `https://arxiv.org/abs/1606.08415`.

[33] Lin T.-Y., Goyal P., Girshick R., He K., Dollár P. Focal loss for dense object detection // *Proc. IEEE/CVF International Conference on Computer Vision (ICCV).* – 2017. – P. 2980-2988.

УДК 519.6

МОДЕЛЬ И АЛГОРИТМЫ КЛАССИФИКАЦИИ АНОМАЛЬНЫХ ЯВЛЕНИЙ НА ОСНОВЕ СХОДИМОСТИ АКУСТИКО-ВИЗУАЛЬНЫХ СИГНАЛОВ

[1]***Равшанов Н.***, [1*]***Боборахимов Б.И.***, [2]***Бердиев М.И.***

*uzbekpy@gmail.com

[1]Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта,
100125, Узбекистан, г. Ташкент, Мирзо-Улугбекский р-он, м-в Буз-2, д. 17А;
[2]Национальная Гвардия Республики Узбекистан,
100017, Узбекистан, г. Ташкент, ул. Ш. Рашидова, дом 23.

В данной статье предлагается контекстно-адаптивная аудио-визуальная нейросетевая модель (KMAVN) для обнаружения аномалий в системах общественной безопасности. Существующие подходы в основном опираются на визуальные данные и используют простые стратегии объединения модальностей, что приводит к ограничениям в охвате сложных семантических связей. Предлагаемая модель состоит из четырёх основных компонентов: модуля извлечения визуальных признаков на основе архитектуры SlowFast, модуля извлечения аудио признаков на основе Audio Spectrogram Transformer (AST), модуля объединения на основе двунаправленного перекрёстного внимания и модуля агрегации временного контекста на основе Transformer encoder. Основное научное новшество модели заключается в механизме адаптивной балансировки модальностей, который динамически регулирует относительную важность модальностей в различных условиях (тёмное/светлое, шумное/тихое). Экспериментальные результаты показали, что предлагаемая модель KMAVN превосходит существующие методы на +4.4% по общей точности и на +32.1% в условиях низкой освещённости. Исследования абляции подтвердили вклад каждого модуля в общую эффективность.

Ключевые слова: динамическое взвешивание модальностей, пространственно-временное представление, выравнивание на основе внимания, устойчивое распознавание аномалий, видеонаблюдение в реальных условиях.

Цитирование: *Равшанов Н., Боборахимов Б.И., Бердиев М.И.* Модель и алгоритмы классификации аномальных явлений на основе сходимости акустико-визуальных сигналов // Проблемы вычислительной и прикладной математики. – 2025. – № 6(70). – С. 88-102.

DOI: https://doi.org/10.71310/pcam.6_70.2025.07

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

№ 6(70) 2025

Журнал основан в 2015 году.
Издается 6 раз в год.

Учредитель:
Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта.

Журнал зарегистрирован в Агентстве информации и массовых коммуникаций при Администрации Президента Республики Узбекистан.
Регистрационное свидетельство №0856 от 5 августа 2015 года.

ISSN 2181-8460, eISSN 2181-046X

Адрес редакции:
100125, г. Ташкент, м-в. Буз-2, 17А.
Тел.: +(998) 712-319-253, 712-319-249.
Э-почта: journals@airi.uz.
Веб-сайт: https://journals.airi.uz.

Дизайн и вёрстка:
Шарипов Х.Д.

Отпечатано в типографии НИИ РЦТИИ.
Подписано в печать 25.12.2025 г.
Формат 60x84 1/8. Заказ №8. Тираж 100 экз.

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

No. 6(70) 2025

The journal was established in 2015.
6 issues are published per year.

Founder:
Digital Technologies and Artificial Intelligence Development Research Institute.

The journal is registered by Agency of Information and Mass Communications under the Administration of the President of the Republic of Uzbekistan.
The registration certificate No. 0856 of 5 August 2015.

ISSN 2181-8460, eISSN 2181-046X

Address:
100125, Tashkent, Buz-2, 17A.
Tel.: +(998) 712-319-253, 712-319-249.
E-mail: journals@airi.uz.
Web-site: https://journals.airi.uz.

Layout design:
Sharipov Kh.D.

DTAIDRI printing office.
Signed for print 25.12.2025
Format 60x84 1/8. Order No. 8. Print run of 100 copies.

Содержание

Contents

№ 6(70) 2025

ISSN 2181-8460

HISOBLASH VA AMALIY MATEMATIKA MUAMMOLARI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

RAQAMLI TEXNOLOGIYALAR VA
SUN'IY INTELLEKTNI RIVOJLANTIRISH
ILMIY-TADQIQOT INSTITUTI