Journal of Advanced College of Engineering and Management

Five shot and Ten Shot Analysis for Anomaly Detection of Video Scenes

[1]Deepak Kumar Singh, [2*]Dibakar Raj Pant
[1,2]*Department of Electronics and Computer Engineering, Pulchowk Campus, IOE*
Email: [1]deepak.singh@sagarmatha.edu.np
**Corresponding Author: [*]pdibakar@gmail.com*

DOI: 10.3126/jacem.v11i1.84526

Abstract
Anomaly detection in video surveillance is essential for maintaining public safety and identifying irregular events in real-world settings. This paper introduces a novel few-shot learning framework for video anomaly detection, titled *Five-Shot and Ten-Shot Analysis for Anomaly Detection from Video Scenes*. The proposed method leverages a meta-learning approach to achieve effective and adaptive performance under limited training data conditions. Specifically, this paper evaluated detection performance across 11 distinct anomaly categories—including water incidents, traffic accidents, shootings, fires, assaults, vandalism, explosions, fights, object falls, robberies, and human falls—captured across 14 diverse real-world scenarios such as malls, streets, offices, highways, and public parks. Our approach utilizes a five-shot one-query scheme during training and a ten-shot one-query scheme during testing, enabling strong generalization from minimal examples. The framework incorporates a hybrid backbone that combines spatial-temporal feature extraction using the Swin Transformer with Model-Agnostic Meta-Learning (MAML) for rapid task adaptation. Experimental results demonstrate that our method achieves robust anomaly detection performance with limited data and delivers competitive Area Under the Curve (AUC) scores.
Keywords *— Swin Transformer, Feedforward, MAML, Anomaly Detector Model, MSAD*

1. INTRODUCTION

Anomaly detection is critical across domains like video surveillance, healthcare, industry, and finance. In surveillance, real-time detection of incidents such as fires, assaults, or accidents is essential for public safety. However, existing models often lack generalization across varied real-world conditions and struggle with limited data availability. Most traditional methods rely on one-class classification trained only on normal data, making them ineffective in diverse environments due to limited scenario and anomaly variety. Additionally, they often overlook non-human anomalies like object falls or water leakage.

To address these challenges, our research is motivated by two core needs:

A. Generalization Across Diverse Scenarios: Traditional models fail to scale across various real-world contexts due to limited training diversity. Real-life anomaly detection must work not only within the same scene but also across different environments like malls, streets, highways, and offices.

B. Learning from Few Examples: In practical deployment, collecting large labeled datasets for each new environment is infeasible. Therefore, there is a

strong need for systems that can learn effectively from only a few examples of new scenarios or anomaly types.

This paper proposes a few-shot meta-learning approach using MAML to enable better adaptation to new scenarios and anomaly types with minimal data rapidly. To this end, this paper leverages the MSAD (Multi-Scenario Anomaly Detection) dataset, which comprises 14 different real-world scenarios and 11 distinct types of anomalies. The purpose of this research is to evaluate the capability of MAML where 5 shot and 10 shot approach has been used to detect anomalies with limited training data while ensuring high generalization across varied scenes. Through detailed analysis, this study demonstrates that the proposed method achieves competitive performance under both five-shot and ten-shot settings, offering a scalable and efficient solution for real-world video anomaly detection challenges.

2. LITERATURE REVIEW

Over the years, a wide range of techniques have been proposed for video anomaly detection across diverse sectors such as transportation, manufacturing, and surveillance. One significant direction is meta-learning, which focuses on rapid adaptation to new tasks using limited data.

Early efforts by Santoro et al. (2016) introduced Memory-Augmented Neural Networks (MANNs) to support fast learning by leveraging external memory, outperforming traditional LSTMs in few-shot scenarios [1]. Subsequently, Model-Agnostic Meta-Learning (MAML), introduced by Finn et al. (2017), became a foundational method due to its task-agnostic nature and its ability to fine-tune models with just a few gradient steps [2]. Ravi et al. (2017) further enhanced this idea by using LSTM-based meta-learners to guide model updates [3].

Additional meta-learning frameworks such as MetaNet (Munkhdalai et al., 2017) [4] and Meta-Transfer Learning (Soh & Cho, 2020) [5] emphasized generalization and rapid learning across tasks with limited supervision. MAML has since been applied in varied domains, including bearing fault detection [6], cyber security [7], and facility management systems [8], demonstrating its adaptability and efficiency.

Other directions in anomaly detection include reconstruction-based approaches using Autoencoders [9], and GAN-based one-class classification methods [10] for modeling normal behavior. Wu et al. (2021) proposed MetaFormer, a self-attention-based unsupervised framework for detecting complex anomalies by capturing both local and global patterns [11].

Recent advancements include explainable few-shot learning [12] and hybrid models such as MAVAE [8], which combine variational autoencoders with MAML for anomaly detection in dynamic environments. Hydra [13] tackled fast model selection for multivariate time series, while Li et al. (2023) explored zero-shot anomaly detection using batch normalization statistics [14]. Duan et al. (2024) applied meta-learning in AIOps for quick anomaly diagnosis [15].

Moreover, SAAD [16] and SA2D [17] introduced meta-learning based video anomaly detection frameworks that effectively adapt to new scenes and camera angles. These models were evaluated using the Multi-Scenario Anomaly Detection (MSAD) dataset, which offers greater diversity than older benchmarks like UCSD and ShanghaiTech.

In summary, existing literature shows the growing relevance of meta-learning especially MAML—in anomaly detection. Its ability to generalize across new tasks and scenarios with minimal training data makes it a promising solution for real-world, scenario-diverse applications like video surveillance.

3. METHODOLOGY

A. System Model

The system processes multiple video scenes (S1 to SN) by extracting frames, resizing, and normalizing them. Features are then extracted using a Swin Transformer, and the resulting embeddings are fed into a MAML-based anomaly detection model. The model performs binary classification (normal vs. anomalous) using five-shot and ten-shot one-way approaches.

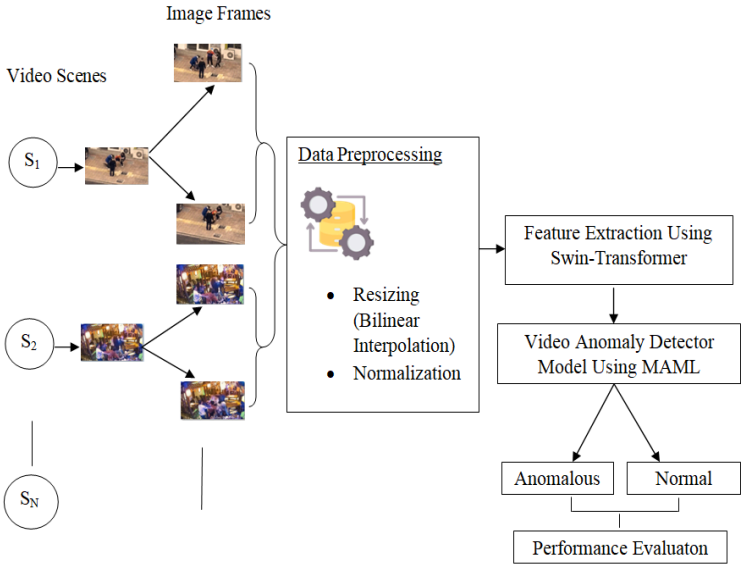

Figure 1: Block Diagram of Anomaly Detection using Meta Learning[18]

B. System Description

1. Dataset: MSAD (Multi Scenario Anomaly Detection) [16,17] datasets containing anomalies and normal instances has been used where the dataset has anomaly videos, normal testing and normal training datasets in video form. These datasets contain raw data collected from various sources such as sensors, logs, or databases. There are total of 720 videos as follows:
    a. Total of 240 videos of 11 different anomaly types from 13 different scenarios excluding highway scene.
    b. Total of 360 videos of normal training from 14 different scenario
    c. Total of 120 videos of normal testing

    The sample images from anomaly type dataset and normal dataset from different scenarios are in figures 2 and 3 below.

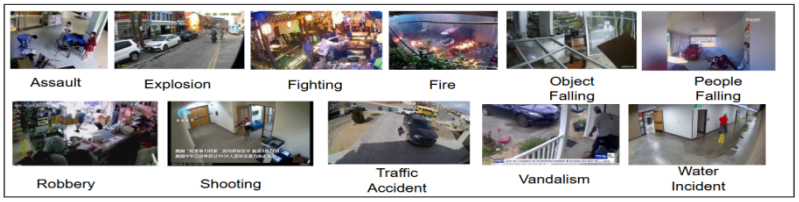

Figure 2: Samples from anomaly dataset

Figure 3: Samples from normal dataset

2. Data Preprocessing: It involves video frame resizing using Bilinear Interpolation method [20] where interpolation for y(height) and x(width) with coordinate points $A(y_1, x_1)$, $B(y_1, x_2)$, $C(y_2, x_1)$ and $D(y_2, x_2)$ includes formula for computation for X and Y for width dimension is given by:

$$X = A(1 - w_x) + Bw_x \quad (1)$$

$$Y = C(1 - w_x) + Dw_x \quad (2)$$

Then linear interpolation between the two interpolated values X and Y in the height dimension is given by:

$$Z = X(1 - w_y) + Y w_y$$

$$Z = A(1 - w_x)(1 - w_y) + Bw_x(1 - w_y) + C(1 - w_x)w_y + Dw_x w_y \quad (3)$$

Where,

$$w_x = (x - x_1)/(x_2 - x_1) \quad (4)$$

and

$$w_y = (y - y_1)/(y_2 - y1) \quad (5)$$

Using above equation, images were resized to required dimension i.e. 224X224 RGB image. The input images, originally in PIL format with dimensions [H×W×C] and pixel values ranging from 0 to 255 as unsigned integers, were transformed into floating-point tensors with dimensions [C×H×W] and values scaled to the [0,1] range. Since the images were in RGB mode, they were standardized after resizing by applying normalization based on the dataset's mean and standard deviation. This process helps ensure the model receives stable and consistent inputs during training. The model was initially trained on 9 scenarios (shop, frontdoor, parkinglot, office, pedestrian street, street highview, warehouse, road, park) with both normal and anomalous videos, validated on 2 scenarios (mall and train), and tested on 2 scenarios (restaurant and sidewalk).

3. Feature Extraction

For video frames (Image Input) of size 224 × 224 with 3 channels, the swin transformer [19] was used to extract feature vectors from the preprocessed data of the MSAD dataset which is illustrated in Figure 4.

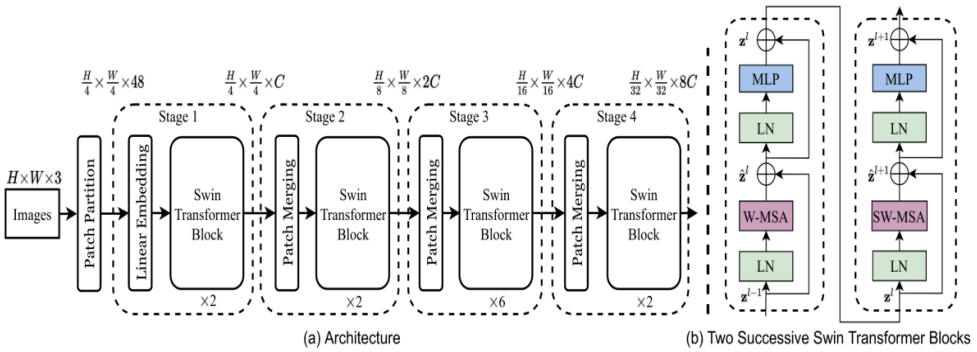

Figure 4: Feature Extraction using Swin Transformer [19]

The swin-s variant of the swin transformer [19] was employed due to its ability to construct hierarchical feature representations through shifted window self attention, which captures both fine-grained local details and broader global context.

Let the input image be:

$$I \in R^{224X224X3} \quad (6)$$

Patch Embedding: The image is first partitioned into non-overlapping 4×4 patches. Each patch is flattened and projected into a patch token T via a linear embedding [19]:

$$T = W_E \quad I \quad + \quad b_E \quad (7)$$

Where, W_E is a learnable embedding matrix and b_E is a bias term.

Hierarchical Feature Extraction: The patch tokens are processed through multiple stages, each consisting of swin transformer blocks. These blocks use Window-based Multi-Head Self-Attention (W-MSA) and Shifted Window MSA (SW-MSA) alternately to enable cross-window interactions [19]:

$$F^{(l)} = \text{SW-MSA}\ (\text{MLP}\ (F^{(l-1)})) \quad (8)$$

Where, $F^{(l)}$ is the feature map at layer l, and MLP denotes the feed-forward network within each transformer block.

Final Feature Representation: After the last stage, a global pooling operation aggregates the spatial features into a compact vector representation [19]:

$$F \quad = \quad \text{Pooling}(F^{(L)}) \quad (9)$$

Where, F denotes the final feature vector and L denotes the last layer, which for swin-s has 768 dimensions. These feature vectors were then served as inputs to the model for effective detection of anomalies.

4. Anomaly Detector Model

The architecture illustrated in Figure 5 employs a multi-layer perceptron (MLP) tailored for anomaly detection, where the input consists of 1024 dimensional image embeddings extracted from the Swin-S model. These embeddings are processed within the Anomaly Detector module through three sequential blocks, each composed of a fully connected layer, batch normalization, and a ReLU activation function to introduce non-linearity. The first block reduces the feature dimension from 1024 to 512, the second maintains the 512-dimensional representation, and the third further compresses it to 256. A final linear layer then maps the 256-dimensional features into a single scalar logit. During inference, the forward pass sequentially processes the input embeddings through these layers, producing an output tensor of size [batch size, 1], where each value corresponds to a logit score. This score is evaluated using the Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss) function to classify the input as either anomalous or normal.

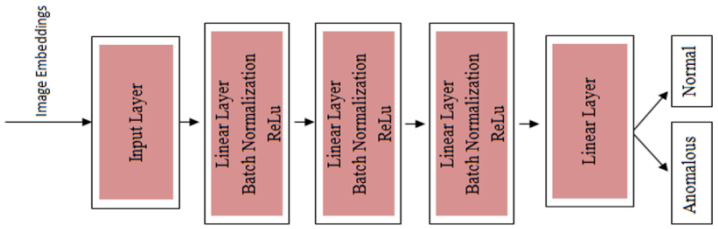

Figure 5: Anomaly Detector Model Layered Architecture [18]

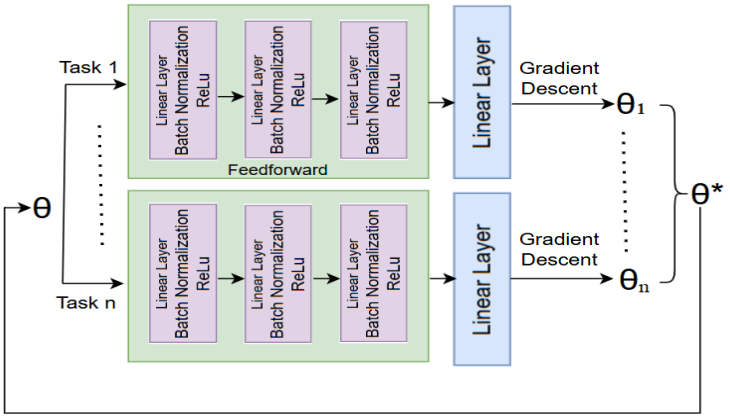

Figure 6: MAML Training Architecture [18]

The meta-training process in figure 6 uses MAML to optimize parameters for rapid adaptation, enabling the detector to generalize across tasks and perform well on unseen scenarios with minimal labeled data [18].

Task sampling created balanced datasets by randomly selecting scenarios, drawing k-shot and k-query samples for normal and anomalous classes, and

dividing them into shuffled support and query sets for training and evaluation.

The inner loop fine-tunes a temporary model copy on the support set using manual gradient updates with a fixed learning rate, adapting it to the current task while preserving the computational graph for outer updates.

The outer loop updates meta-parameters by averaging query losses across tasks after inner-loop adaptation, then optimizes the model via backpropagation to learn a generalizable initialization for rapid task adaptation.

Meta-validation evaluates the model by adapting on support sets, testing on query sets, averaging task losses, and applying early stopping to guide training and model selection [18].

5. Model Agnostic Meta Learning[18]
   For adaptive anomaly detection on MASD and new datasets, Swin Transformer embeddings with MAML were used to train an anomaly detector that learns optimal parameters and generalizes well to unseen data figure 7.

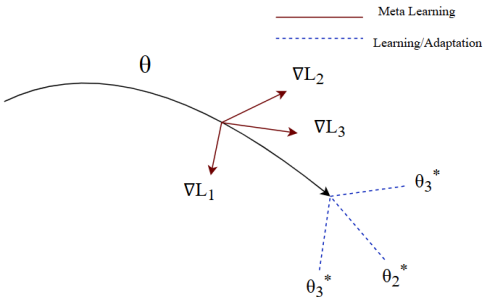

Figure 7: The Architecture of MAML [18]

Algorithm: Generic MAML algorithm [2] Require:

D(T): Distribution over tasks

Require: α and β: learning rates

a. Randomly initialize parameter ω
b. While not done do
c. Sample batch of task *Ti ~ D(T)*
d. For all T_i

e. Evaluate with $\nabla_{\omega} L(\omega)$ with respect to k samples
f. Calculate adapted parameters using gradient descent:
   $\omega_i = \omega - \alpha \nabla_{\omega} L(\omega)$
g. end for
h. Update $\omega \leftarrow \omega - \beta \nabla_{\omega} L(\omega - \alpha \nabla_{\omega} L(\omega))$
i. end while

4. RESULT AND DISCUSSION

The study used the MSAD video dataset, extracting and preprocessing frames by resizing to 224×224 and normalizing them. Features were extracted via a Swin Transformer, producing 768-dimensional embeddings, which were fed into an anomaly detection model for binary classification of normal and anomalous behavior.

The model output is illustrated in given figure 8.

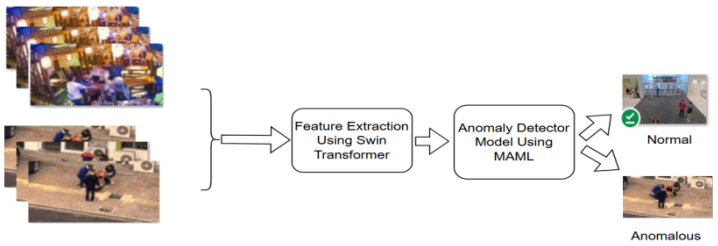

Figure 8: Output of the Model [18]

The experimental setup includes two different approaches including 5 shot and 10 shot approach, where parameters given in table 1 below were used for model training and validation using 5 shot approach.

Table 1: The MAML parameters and hyper-parameters used in the implementation

Parameter	**Value**
Epochs	100
K-Shot	5
K-Query	1
Optimizer	adam
Inner Learning Rate	0.1
Outer Learning Rate	0.01
Batchsz	2000
Input Image Size	224 X 224
No. of Iterations	400000
Feature Embedding Size	768
Device	cuda

The model's performance, evaluated after 400,000 iterations using a confusion matrix given in figure 9, that showed 371,431 true positives, 363,737 true negatives, 36,263 false positives, and 28,569 false negatives. These values were used to calculate the model's key performance metrics.

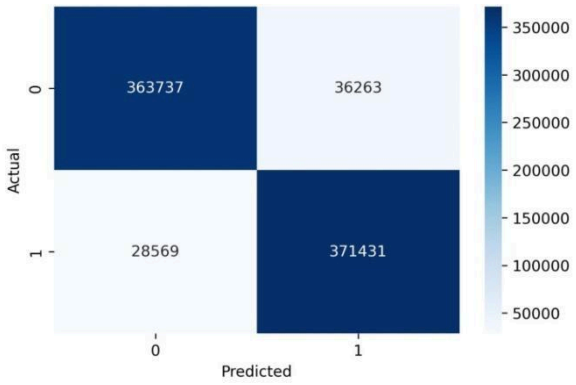

Figure 9: Confusion Matrix for Model Evaluation

Performance matrix:

Accuracy = (TP + TN) / (TP + TN + FP + FN)
= (371431 + 363737) / (371431 + 363737 + 36263 + 28569)
= 0.918

Precision = (TP) / (TP + FP)
= 371431 / (371431 + 36263)
= 0.911

Recall = TP / (TP + FN)
= 371431 / (371431 + 28569)
= 0.928

F1-Score = 2 ∗ (Precision ∗ Recall) / (Precision + Recall)
= 0.919

The ROC curve shows the model's strong detection ability, with a steep initial rise and an AUC near 1, indicating high classification performance well above the random-guess baseline. The model evaluation metrics is given in table 2.

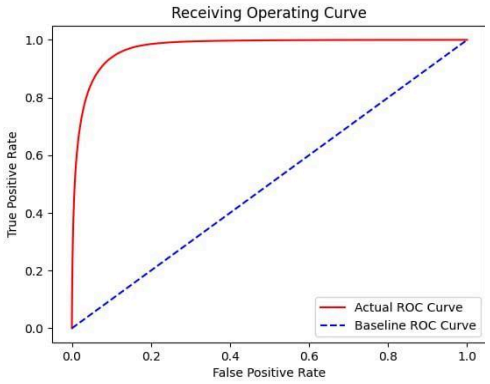

Figure 10: ROC Curve for Model Evaluation

Table 2: Evaluation Metrics

Metric	Value
Loss	0.209
Accuracy	0.918
Precision	0.911
Recall	0.928
F1-Score	0.919
Avg. Precision	0.970
AUC Score	0.974

The experimental set up for model evaluation using 10 shot approach for 400000 number of iteration with MAML parameters are as illustrated in table 3.

Table 3: Meta Learning Parameters used in Ten Shot Experiments

Parameter	Value
K-Shot	10
K-Query	1
No. of Iterations	400000
Input Image Size	224 X 224
No. of Layers	3
Feature Embedding Size	768
Device	cuda

After 400,000 iterations, the model achieved 375,455 true positives, 368,276 true negatives, 31,724 false positives, and 24,545 false negatives, forming the basis for its key evaluation metrics.

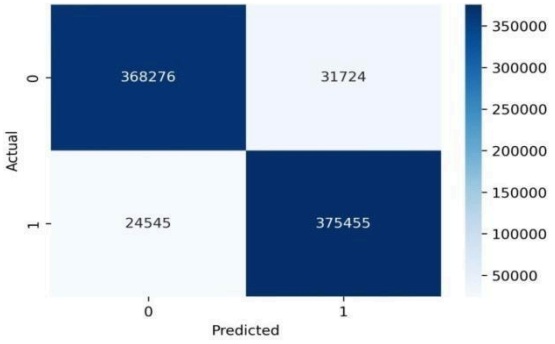

Figure 11: Confusion Matrix for Model Evaluation

Performance matrix

Accuracy = (TP + TN) / (TP + TN + FP + FN)
= 743731 / 800000
= 0.929

Precision = TP / (TP + FP)
= 375455 / (375455 + 31724)
= 0.922

Recall = TP / (TP + FN)
= 375455 / (375455 + 24545)
= 0.938

F1-Score = 2 ∗ (Precision ∗ Recall) / (Precision + Recall)
= 0.929

The ROC curve in figure 12 shows the model's strong anomaly detection performance, with a sharp initial rise indicating high true positive rates at low false positives. Its AUC near 1 confirms excellent overall classification capability. The model evaluation metric is given in table 4.

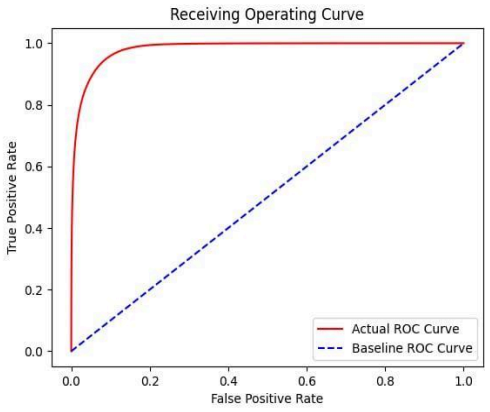

Figure 12: ROC Curve for Model Evaluation

Table 4: Model Evaluation Metric

Metric	Value
Loss	0.168
Accuracy	0.929
Precision	0.922
Recall	0.938
F1-Score	0.930
Avg. Precision	0.980
AUC Score	0.982

In the context of video anomaly detection using the MSAD dataset, the classification was evaluated under three few-shot learning scenarios: 5-shot, and 10-shot. In 5-shot setting, where the model was given five examples per class, there was a significant improvement in classification performance across all metrics. The F1-score rises to 91.9%, indicating that the model benefits substantially from the additional data, which helps it capture a wider range of anomaly patterns. In the 10-shot setting, the model's performance improved even further, achieving an F1-score of 93%. This indicated that while adding more examples continues to enhance performance, the rate of improvement diminished slightly compared to the jump from 5-shot to 10-shot. Overall, these results demonstrated the effectiveness of few-shot learning in video anomaly detection, with notable gains in performance as more labeled data became available, even in small quantities.

Method	Loss	Accuracy	Precision	Recall	F1-Score	AUC
5-Shot	0.209	0.918	0.911	0.928	0.919	0.974
10-Shot	0.168	0.929	0.922	0.938	0.930	0.982

5. CONCLUSION

In this study, two different methods were presented to video anomaly detection by leveraging a multi-scenario anomaly detection dataset where 14 different scenarios included 11 different anomaly types and normal scenes in the dataset. The dataset was first used for extracting the frames, which were prepared for processing by resizing an normalizing. After which optimal format for the input to model was ensured. Then anomaly detection approach demonstrated the effectiveness of swin

transformer in capturing spatial dependencies in video frames while detector model using MAML (Five Shot and Ten Shot Approach) was employed to feature map for classifying video frames into normal and anomalous. The results demonstrated significant improvement in anomaly detection performance, showing the power of combining swin transformer for feature extraction with MAML for meta-learning. This approach not only improved classification accuracy but also ensured better generalization to novel and complex datasets.

REFERENCES

1. A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), New York, NY, USA, Jun. 2016, pp. 1842–1850.
2. C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in Proc. Int. Conf. Mach. Learn. (ICML), Sydney, Australia, Aug. 2017, pp. 1126–1135.
3. S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in Proc. Int. Conf. Learn. Represent. (ICLR), Toulon, France, Apr. 2017.
4. T. Munkhdalai and H. Yu, "Meta networks," in Proc. Int. Conf. Mach. Learn. (ICML), Sydney, Australia, Aug. 2017, pp. 2554–2563.
5. Y. Soh and Y. Cho, "Meta-transfer learning for few-shot learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2020, pp. 403–412.
6. Y. Zhang, Y. Zhang, L. Zhang, and J. Xu, "Few-shot bearing fault diagnosis using MAML-based meta-learning," *IEEE Access*, vol. 8, pp. 202612–202620, 2020, doi: 10.1109/ACCESS.2020.3036223.
7. J. Smith, A. Kumar, and M. Abdelbaki, "Few-shot cyber attack detection with model-agnostic meta-learning," in Proc. Int. Conf. Cyber Secur. Cryptogr. (ICC), Paris, France, Dec. 2020, pp. 88–93.
8. J. Moon, K. Shin, and H. Cho, "A hybrid MAML-VAE approach for anomaly detection in facility management systems," Expert Syst. Appl., vol. 223, p. 119933, Oct. 2023, doi: 10.1016/j.eswa.2023.119933.
9. D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in Proc. Asian Conf. Comput. Vis. (ACCV), Perth, Australia, Dec. 2018, pp. 622–638.
10. X. Chen, Y. Li, and Y. Jin, "Adversarial learning for one-class classification," in Proc. AAAI Conf. Artif. Intell., Vancouver, Canada, Feb. 2021, pp. 3555–3562.
11. Z. Wu, H. Fu, and X. Wei, "MetaFormer: Self-attention-based framework for unsupervised video anomaly detection," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, Canada, Oct. 2021, pp. 11932–11942.
12. H. Meng, K. Liu, and X. Wang, "Explainable few-shot learning via prototype learning and attention visualization," Neural Netw., vol. 162, pp. 453–465, Aug. 2023, doi: 10.1016/j.neunet.2023.05.014.
13. J. Navarro, D. Cortés, and A. Lozano, "Hydra: Few-shot fast model selection for multivariate time series," in Proc. Int. Conf. Time Series (TSConf), Munich, Germany, Jul. 2023.

14. Y. Li, Z. Xu, and K. Wu, “Zero-shot anomaly detection using batch normalization statistics,” in Proc. Int. Conf. Learn. Represent. (ICLR), Kigali, Rwanda, May 2023.
15. Y. Duan, J. Shen, and B. Tang, “Meta-learning for rapid anomaly diagnosis in AIOps systems,” in Proc. AAAI Conf. Artif. Intell., Vancouver, Canada, Feb. 2024.
16. X. Zhu, Z. Wang, and J. Liu, “SAAD: Scene-adaptive anomaly detection via meta-learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2024.
17. X. Zhu and Z. Wang, “SA2D: Self-adaptive anomaly detection with meta-learned scene priors,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Jun. 2024, doi: 10.1109/TPAMI.2024.3401234.
18. D. K. Singh, D. R. Pant, G. Gautam, and B. Shrestha, “Meta-learning approach for adaptive anomaly detection from multi-scenario video surveillance,” *Appl. Sci.*, vol. 15, no. 12, p. 6687, Jun. 2025, doi: 10.3390/app15126687.
19. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, and B. Guo, “Swin Transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 10012–10022.
20. C. Ji, “Bilinear resize,” [Online]. Available: https://chao-ji.github.io/jekyll/update/2018/07/19/BilinearResize.html. [Accessed: Jan. 28, 2025].