INFP: INdustrial Video Anomaly Detection via Frequency Prioritization

Qianzi Yu[1] , **Kai Zhu**[1] , **Yang Cao**[1,2,✉] and **Yu Kang**[1,2]
[1]University of Science and Technology of China, Hefei, China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China
{yuqianzi, zkzy}@mail.ustc.edu.cn, {forrest, kangduyu}@ustc.edu.cn

Abstract

Industrial video anomaly detection aims to perform real-time analysis of video streams from industrial production lines and provide anomaly alerts. Conventional video anomaly detection methods focus more on the overall image, as they aim to identify anomalies among multiple normal samples appearing simultaneously. However, industrial scenarios, where the primary focus is on a single type of product, require attention to local areas to capture fine-grained details and specific patterns. Directly applying conventional methods to industrial scenarios can result in an inability to focus on products moving along fixed trajectories, ineffective utilization of their equidistant periodicity, and greater susceptibility to lighting variations. To address these issues, we propose INFP, an encoder-decoder framework that learns frequency-domain features from videos to capture periodic and dynamic characteristics, enhancing the model's robustness. Specifically, a trajectory filter is proposed that takes advantage of the significant difference between moving objects and static backgrounds in the frequency domain by assigning higher weights to fixed moving trajectories. Moreover, a multi-feature fusion module is proposed, in which the frequency domain features of the video are first extracted to leverage the unique equidistant periodicity information of videos from industrial production lines. The extracted frequency domain features are subsequently fused with spatio-temporal features and contextual information is further integrated from the fused representation, effectively mitigating the impact of lighting variations on production lines. Extensive experiments on the benchmark IPAD dataset demonstrate the superiority of our proposed method over the state-of-the-art.

1 Introduction

Video anomaly detection (VAD) [Zhong *et al.*, 2019; Zaheer *et al.*, 2020; Sultani *et al.*, 2018; Le and Kim, 2023; Gong *et al.*, 2019; Hasan *et al.*, 2016] aims to identify unusual or unexpected events in video sequences that deviate from normal patterns. In addition to traffic accidents, criminal activities, and illegal behaviors, the demand for applying video anomaly detection in industrial scenarios has continuously increased with the manufacturing industry's development. However, directly applying conventional video anomaly detection methods in industrial scenarios does not yield good performance because industrial scenarios differ from conventional ones in three key factors: (1) Products on industrial production lines are typically sparsely distributed within the frame, occupying a smaller visual proportion, whereas in conventional ones, vehicles and pedestrians are more densely distributed and occupy a larger portion of the frame. (2) Products on production lines exhibit an even distribution over time, demonstrating equidistant periodicity, which is not as apparent in conventional scenarios. (3) Compared to conventional scenarios, lighting variations in industrial ones are more pronounced and have a greater impact on detection results.

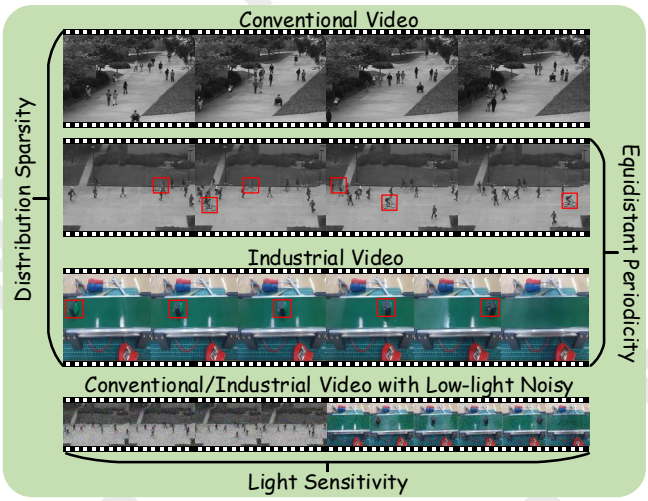

Figure 1: **Comparison of video anomaly detection in a conventional and industrial scenario.** Compared to the conventional scenario, industrial scenario video has the following characteristics: (1) The products in the video are sparsely distributed, occupying a smaller visual proportion. (2) The products appear on the production line at regular time intervals, exhibiting a distinct equidistant periodicity. (3) Lighting variation has a greater impact on the video.

To address the above issues, we propose INFP that contains a trajectory filter and a multi-feature module. Specifically, to solve the sparse distribution of products, an effective ap-

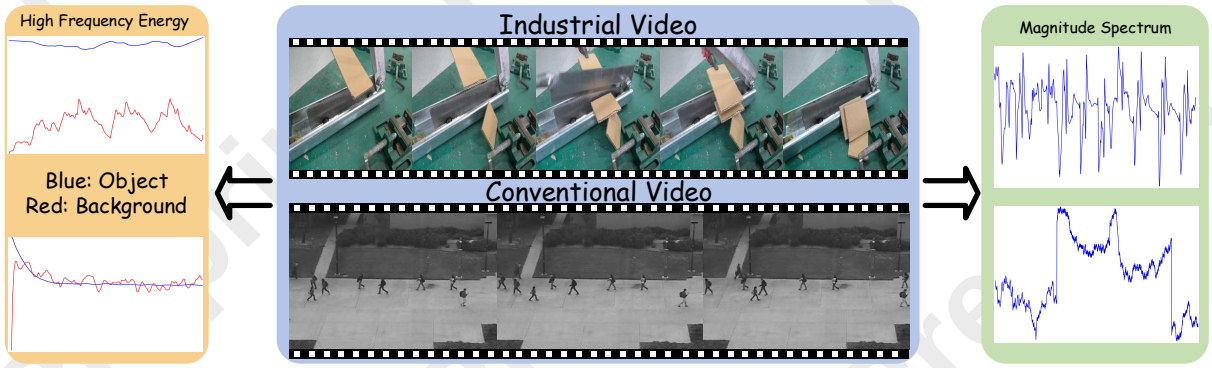

Figure 2: **Motivation.** The figure shows that in industrial scenario, the difference in high-frequency energy between moving objects and the background is more pronounced. Additionally, the magnitude spectrum of industrial scenes exhibits more obvious periodicity.

proach is to separate them from the background in industrial scenarios, allowing the detector to focus more on the product areas. As shown on the left side of Figure 2, we extract the high-frequency energy of moving objects and backgrounds in both conventional and industrial scenarios. The results reveal a clear distinction in high-frequency energy between the two in industrial scenarios, which enables us to separate the products from the background based on this difference. Therefore, a trajectory filter is designed to assign higher weights to regions with higher high-frequency energy. Assigning higher weights to moving objects during the training and inference steps highlights the moving object characteristics on fixed trajectories through frequency-domain weighting, effectively enhancing the model's robustness and adaptability to industrial video anomaly detection.

Moreover, to cope with periodic information and lighting variation, we propose a multi-feature fusion module. In this module, a frequency-domain feature extractor is designed to extract frequency-domain signals with distinct equidistant periodic spectra. After that, the fused features, formed by integrating frequency-domain features with spatiotemporal features, also exhibit distinct equidistant periodicity, providing robust and structured input for the subsequent decoder, thereby enhancing its ability to capture periodic information effectively. The use of periodic information can help identify speed anomalies and product absence in industrial scenarios. Subsequently, the contextual information of the fused features is fully integrated and activated, which makes full use of the complementarity of multiple features, reducing significant pixel variations in video frames caused by lighting changes.

Our main contributions are summarized as follows.

1) We introduce INFP, an encoder-decoder framework that learns frequency-domain features to address specific issues in the industrial video anomaly detection task.
2) A trajectory filter and a multi-feature fusion module are proposed to solve distribution sparsity, equidistant periodicity, and light sensitivity of products on the production line.
3) Extensive experiments on IPAD datasets demonstrate the superiority of our proposed method over SOTA. The metrics also indicate that our model has better generalization ability and robustness.

2 Related Work

2.1 Video Anomaly Detection

In video anomaly detection, researchers[Zhong *et al.*, 2019; Zaheer *et al.*, 2020; Sultani *et al.*, 2018] use normal and annotated anomaly video data to train models. Detailed annotation of enriched anomaly data typically helps improve the models' performance on the test set. However, due to the difficulty of obtaining anomaly video data in real-world scenarios, researchers[Le and Kim, 2023; Gong *et al.*, 2019; Liu *et al.*, 2020] train the model using only normal video data, which is called One-Class Classification (OCC) task, to get rid of dependence on anomaly data. The primary frameworks for the OCC problem can be generally divided into two categories: reconstruction-based models and prediction-based models. Reconstruction-based models are trained to reconstruct the input frame. Hasan[Hasan *et al.*, 2016] et al. first propose an auto-encoder structure to reconstruct video frames in the context of anomaly detection. MemAE[Gong *et al.*, 2019] propose a memory module to store feature information for better reconstruction. IPAD_VAD[Liu *et al.*, 2024] proposes a period memory module to explore periodic information in the reconstruction-based models. Prediction-based models use some previous frames to predict the future frames are normal or not. Astnet[Le and Kim, 2023] contains a channel-based decoder for focusing on important objects while predicting future frames. Doshi[Doshi and Yilmaz, 2021] et al. use a GAN[Goodfellow *et al.*, 2014] to predict whether future frames are normal. However, these methods struggle to perform well in industrial scenarios with complex lighting conditions.

2.2 Industrial Image Anomaly Detection

In recent years, image anomaly detection has gained increasing attention. Models are trained to determine whether an image contains anomalies and locate them. Anomaly detection in industrial images[Bergmann *et al.*, 2019; Wang *et al.*, 2024] plays a crucial role in adjusting industrial production lines. The training methods can be divided into four categories: deep feature embedding methods [Bergmann *et al.*, 2020; Deng *et al.*, 2024a; Salehi *et al.*, 2021], image reconstruction methods[Tsai *et al.*, 2022; Zhang *et al.*, 2023],

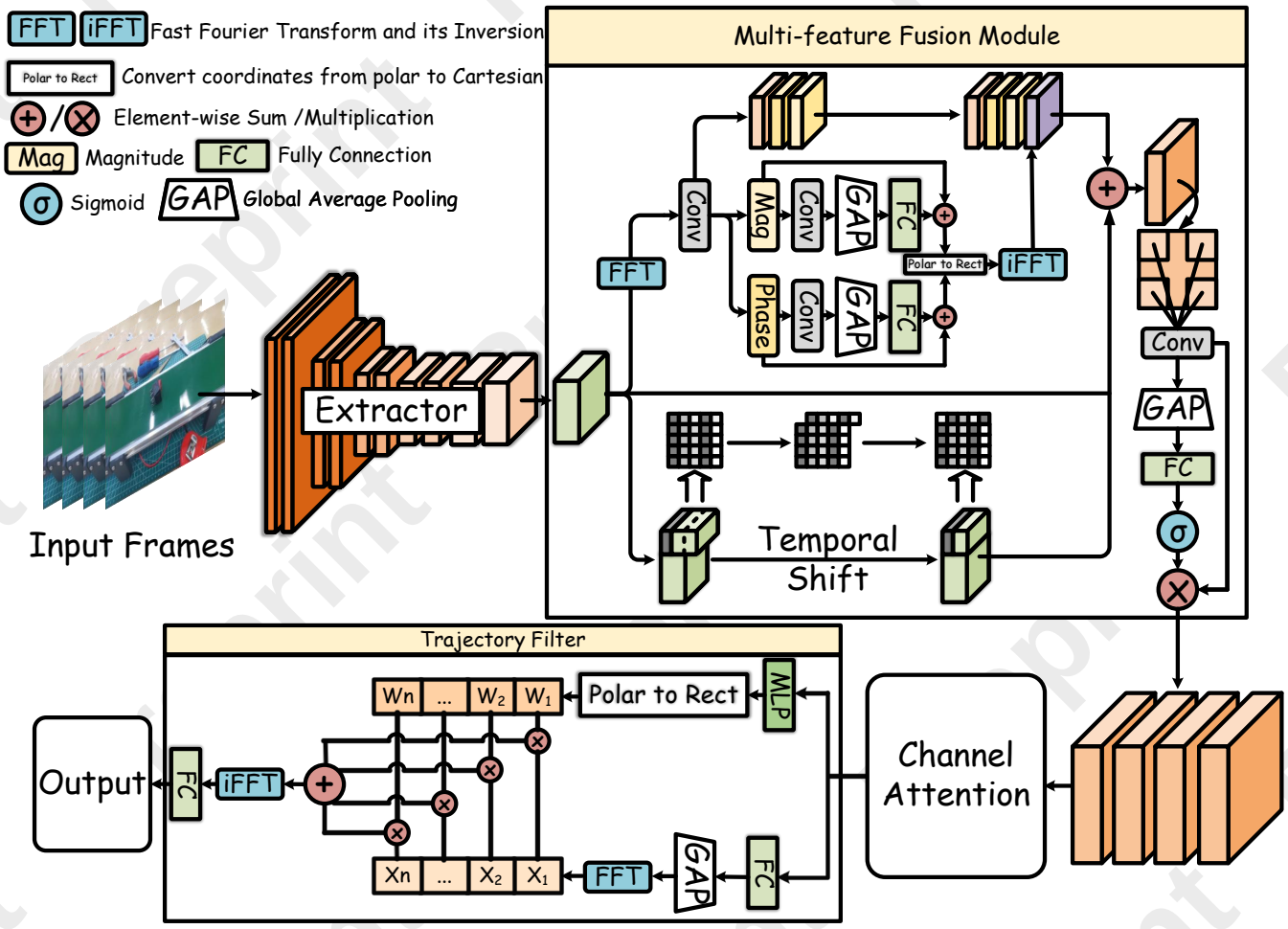

Figure 3: **Overall structure of our method.** It mainly consists of an encoder containing a feature extractor and a feature fusion module, and a decoder containing a channel attention module and a trajectory filter.

image generation methods[Gudovskiy *et al.*, 2022; Schlegl *et al.*, 2017], and self-supervision methods[Li *et al.*, 2021; Pirnay and Chai, 2022]. In addition, TF^2[Yu *et al.*, 2024] improves detection performance by generating more defect images. Deng[Deng *et al.*, 2024b] et al. propose VMAD (Visual-enhanced MLLM Anomaly Detection) that enhances MLLM (Multimodal Large Language Models) with visual-based IAD (Industrial Anomaly Detection) knowledge, simultaneously providing precise detection and comprehensive analysis of anomalies. However, these methods can only deal with still images, failing to fully take advantage of temporal and periodic information in the video. So it is difficult to directly apply these methods to industrial video anomaly detection scenarios.

3 Method

3.1 Model Overview

In this section, we will detail the overall framework of the proposed network. As shown in Figure 3, our model is based on a classic encoder-decoder network[Le and Kim, 2023] to complete the industrial video anomaly detection task. It consists of the multi-feature fusion encoder for captioning the multi-features and fusing them, and a decoder contains a trajectory filter for better determining future frames.

Our model uses a deep and wide convolutional neural network f_e to extract high-level features from the given frames I_v. Then a multi-feature extractor f_m transforms these features into three branches: spatial feature F_s, temporal feature F_t, and frequency feature F_f. A feature fusion module F_{fusion} is followed, which constructs a fusion feature combining multi-feature context. So the feature map M_e obtained after the encoder can be described as follows:

$$M_e = F_{fusion}(f_m(f_e(I_v); \theta)) \tag{1}$$

or

$$M_e = F_{fusion}(F_s, F_t, F_f; \theta)) \tag{2}$$

Here, θ is the learnable parameter of our encoder, consisting of the multi-feature extractor and feature fusion module.

The output of the encoder is used as the input of the decoder, which consists of a channel attention module F_{ca} and a trajectory filter F_{filter}. The channel attention module consists of an average pooling and two convolutional layers. Each of the two convolutional layers is followed by a ReLU and a sigmoid activation function, respectively. The output of

the channel attention M_{ca} is computed as follows,

$$M_{ca} = F_{ca} \otimes M_e \tag{3}$$

where $\otimes$ denotes element-wise product. The final output of our network is computed as follows,

$$M = F_{filter}(M_{ca}; \theta') \tag{4}$$

Here, θ' is the learnable parameter of our decoder, consisting of the channel attention module and the trajectory filter.

The goal of our model is to predict the future frame P_{t+1} based on the previous frames $\{P_1, P_2, ..., P_t\}$. We choose L_2 loss to constrain the similarity of every pixel from the predicted frame and the ground-truth frame.

$$L_2(P, \hat{P}) = \frac{1}{n}\sum_{i=1}^{n}(P - \hat{P})^2 \tag{5}$$

What is more, a gradient constraint is added to get rid of the potential blur and lighting variant in the frame, which calculates the difference in absolute values of the gradients along the two dimensions in 2-dimensional space.

$$L_{2d} = \sum_{i,j}(\sqrt{(|P_{i+1,j+1} - P_{i,j+1}| - |\hat{P}_{i+1,j+1} - \hat{P}_{i,j+1}|)^2} + \sqrt{(|P_{i+1,j+1} - P_{i+1,j}| - |\hat{P}_{i+1,j+1} - \hat{P}_{i+1,j}|)^2}) \tag{6}$$

Hence, the final loss of our network is defined as follows,

$$L((P, \hat{P})) = \alpha L_2(P, \hat{P}) + (1 - \alpha)L_{2d}(P, \hat{P}) \tag{7}$$

where α is the coefficient that controls the weights of these two losses.

3.2 Multi-feature Fusion

To cope with periodic information and lighting variation, we propose a multi-feature fusion module. We design a frequency feature extractor to transform periodic information into the frequency domain to highlight its characteristics. First, we perform a Fourier transform on the previously extracted features $f_e(I_v)$.

$$z = \int_{-\infty}^{+\infty} f_e(I_v)e^{-j\omega t}\, dt \tag{8}$$

Meanwhile, z can be expressed as:

$$z = \hat{a} + \hat{b}j \tag{9}$$

$$z = |z|e^{j \cdot angle(z)} \tag{10}$$

where $|z| = \sqrt{\hat{a}^2 + \hat{b}^2}$, $angle(z) = arg(z) = arctan(\frac{\hat{b}}{\hat{a}})$. We further extract features from the real and imaginary parts separately.

$$a = \hat{a} + \delta(\overline{\omega}(f(\hat{a}))) \tag{11}$$

$$b = \hat{b} + \delta(\overline{\omega}(f(\hat{b}))) \tag{12}$$

where $f(\cdot), \overline{\omega}(\cdot), \delta(\cdot)$ denote convolutional neural network (CNN), LeakyReLU, global average pooling followed by the fully connected network, respectively. We apply the inverse Fourier transform to the newly obtained $z = a + bj$.

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} ze^{j\omega t}d\omega \tag{13}$$

So the final frequency feature can be written as:

$$F_{frq} = [\gamma(z); f(t)] \tag{14}$$

where γ denotes convolutional neural network (CNN), z originates from equation (8).

In addition to frequency domain information, we also need to utilize the spatiotemporal information of industrial videos to help us make better decisions. We use [Lin *et al.*, 2019] to exploit temporal information in the industrial video.

$$F_{st} = f_e(I_v) + tsm(f_e(I_v)) \tag{15}$$

So the multi-feature can be written as:

$$F_{mul} = F_{frq} + F_{st} \tag{16}$$

To address the impact of lighting variations on videos, we further process the multi-feature and extract deeper information by combining feature contexts. First, we predefine a region $A \in R^{a\times a\times c}$ on the multi-feature $F_{mul} \in R^{h\times w\times c}$, with four sub-regions $\widetilde{a}$ located in one of the four corners of A. We perform convolution operations on the predefined four sub-regions, allowing us to focus only on the predefined areas while ignoring information from other regions. The values obtained through convolution are used as the results of region A after combining with contextual information. Since we need to maintain the same feature dimensions before and after combining with contextual information, zero padding is required when defining A in the four corners and along the four edges of F_{mul}. Then the output is passed through a ReLU activation to obtain activation map M_{con}. Although our method incorporates contextual information, it may generate biased activation maps. Therefore, we need to correct any potential biases.

$$M'_{con} = M_{con} \otimes \sigma(Z_2 \cdot (\delta(Z_1 \cdot (M_{con})))) \tag{17}$$

where Z_1 and Z_2 represent fully connected network, δ and σ represent ReLU and Sigmoid activation, respectively. The modified M'_{con} then is passed into the decoder.

3.3 Trajectory Filter

To focus attention on the fixed trajectories of object motion rather than the surrounding background, we introduce a trajectory filter that assigns higher weights to pixels on the fixed trajectories and lower weights to the background. The features $F_d \in \mathbb{R}^{B\times H\times W\times C}$ processed by equation (3) are fed into the trajectory filter. First, we perform a Fourier transform on the F_d to extract the frequency feature.

$$X = \{x_1, x_2, ...x_n\} = \int_{-\infty}^{+\infty} \delta(Z_3(F_d))e^{-j\omega t}\, dt \tag{18}$$

where Z_3 represents fully connected network, δ represents ReLU activation, respectively. Then, we use global average pooling and an MLP to compute the trajectory weights.

$$R = softmax(Z_4(GAP(F_d))) \tag{19}$$

Method	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	R01	R02	R03	R04	Avg.
conAE[Hasan *et al.*, 2016]	63.1	47.7	53.0	34.7	82.9	46.6	58.3	69.1	53.0	55.3	39.8	50.4	77.6	64.3	40.7	70.1	56.7
memAE[Gong *et al.*, 2019]	63.2	50.6	65.6	49.5	78.8	45.9	57.9	84.7	65.7	59.9	49.4	50.7	77.9	65.0	41.6	70.7	61.1
AstNet[Le and Kim, 2023]	67.7	52.0	61.0	51.6	80.4	54.1	54.5	82.6	59.8	55.7	47.8	60.8	79.8	66.8	42.1	67.6	61.5
DMAD[Liu *et al.*, 2023]	55.9	55.3	47.9	47.9	69.3	61.0	66.9	87.5	69.7	67.0	56.0	55.8	79.5	68.5	43.1	63.1	62.2
V-Swin-T[Liu *et al.*, 2022]	68.2	60.0	66.6	54.7	85.6	53.3	59.5	88.5	69.7	60.5	54.8	**69.1**	81.1	74.1	42.3	75.5	66.5
IPAD_VAD[Liu *et al.*, 2024]	69.5	**63.9**	**70.6**	**58.3**	86.2	61.2	60.6	85.6	**71.2**	62.2	**60.9**	67.1	84.4	75.4	43.5	**76.7**	68.6
Ours	**77.2**	62.8	**70.6**	56.0	**89.7**	**73.4**	**78.1**	**96.3**	66.6	**71.9**	56.9	60.9	**86.5**	**77.9**	**48.4**	75.3	**71.8**

Table 1: **Main Result on the IPAD** (industrial period video dataset). Each result is a frame-level AUC score. The best results are highlighted in **bold** and the second-best result is underlined. The last column indicates the mean value of AUC under all cases.

Category	Baseline	+Fusion	+Filter	+Fusion&Filter
S01	69.8	75.3	72.9	77.2
S02	52.5	57.1	55.8	62.8
S03	62.4	68.1	66.2	70.6
S04	50.8	53.5	52.7	56.0
S05	79.9	85.4	84.2	89.7
S06	53.5	68.1	68.8	73.4
S07	55.4	73.4	72.2	78.1
S08	81.8	94.2	92.5	96.3
S09	59.8	62.2	64.1	66.6
S10	56.2	68.8	64.7	71.9
S11	47.8	54.8	53.3	56.9
S12	55.5	57.9	56.8	60.9
R01	81.0	84.8	83.5	86.5
R02	67.5	75.4	73.6	77.9
R03	41.2	45.4	43.2	48.4
R04	66.6	70.9	71.7	75.3
Average	61.3	68.5	67.2	71.8

Table 2: **The results of ablation studies.** '+Fusion' and '+Filter' represent the baseline with only multi-feature fusion and the baseline with only trajectory filter, respectively. The last column represents our method.

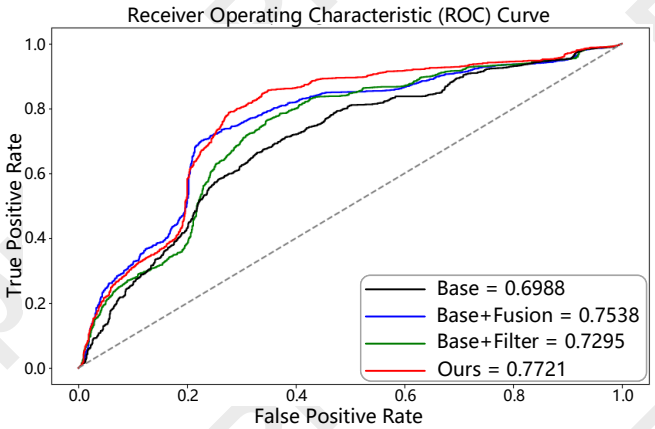

Figure 4: ROC Curve of ablation study, taking category S01 as an example.

where Z_4 represents fully connected network, $GAP(F_d) = \frac{1}{H \cdot W} \sum_{h=1}^{H} \sum_{w=1}^{W} (F_d)_{h,w}$. We generate a random complex weight $W = W_{real} + j \cdot W_{imag}$ and combine it with the trajectory weight.

$$W = \{w_1, w_2, ..., w_n\} = \sum_{n=1}^{N} R_{b,n,c} \cdot W_{h,w,n} \tag{20}$$

We multiply X and W element-wise in the frequency domain to obtain the final weighted result, and the final output is defined as follows:

$$X_{out} = \delta(Z_5(\frac{1}{2\pi} \int_{-\infty}^{\infty} (X \otimes W) e^{j\omega t} d\omega)) \tag{21}$$

where Z_5 represents fully connected network, δ represents ReLU activation, respectively.

4 Experiments

4.1 Dataset and Evaluation Protocol

IPAD[Liu *et al.*, 2024] is the first video anomaly detection dataset that focuses on industrial scenarios, which contains a total of 597,979 frames, with 430,867 frames allocated for training data and 167,112 frames for the test data. The industrial processes in this dataset are selected through factory site visits and discussions with engineers. This dataset includes 16 different types of industrial devices and contains over six hours of video footage, comprising both synthetic and real-world scenes.

Following the prior work[Hasan *et al.*, 2016; Liu *et al.*, 2018; Liu *et al.*, 2024], we choose the frame-level area under the curve (AUC) to evaluate the performance of our proposed method and other state-of-the-art video anomaly detection methods. AUC, a widely used evaluation metric for binary classification models, represents the area beneath the model's receiver operating characteristic (ROC) curve. We concatenate the video frames in the test set and calculate the AUC values. Higher AUC values indicate better anomaly detection methods' performance.

4.2 Implement Details

Each video frame is resized as 224 $\times$ 288, the intensity of which is normalized to the range of $[-1, 1]$ before being fed into the model. The learning rate is set as 2e-4 initially and decrease to 1e-4 at epoch 120. The Adam optimizer is used to train our network. A sequence of five video frames is randomly selected from the training set, with the first four frames serving as input and the fifth frame serving as the ground truth. The frame predicted by the model is then compared to the ground truth frame to calculate the anomaly score.

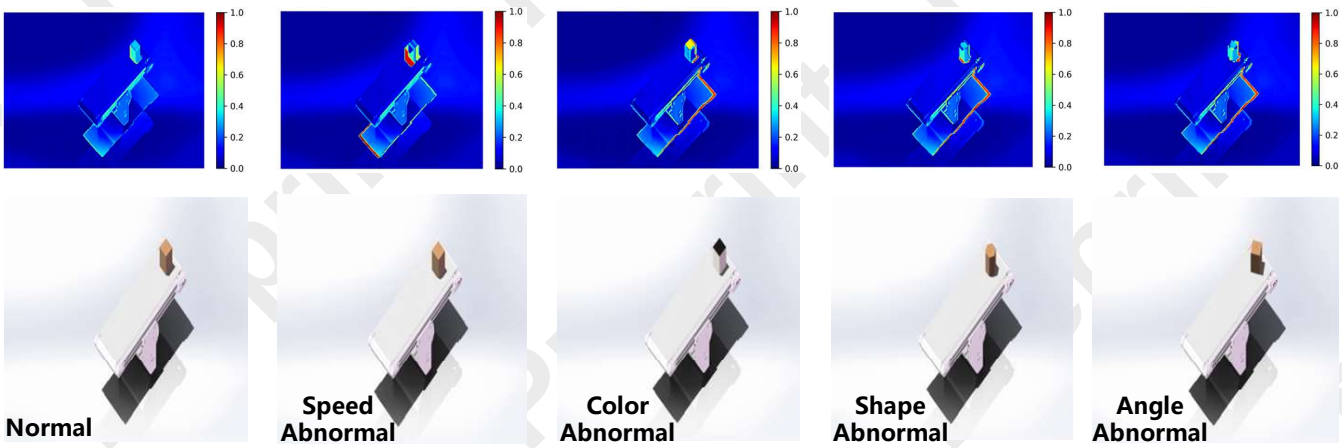

Figure 5: **Difference in heat map between the predicted frame and the ground truth frame.** Taking category S01 as an example, the images show the heat maps under different conditions: normal, speed anomaly, color anomaly, shape anomaly, and angle anomaly.

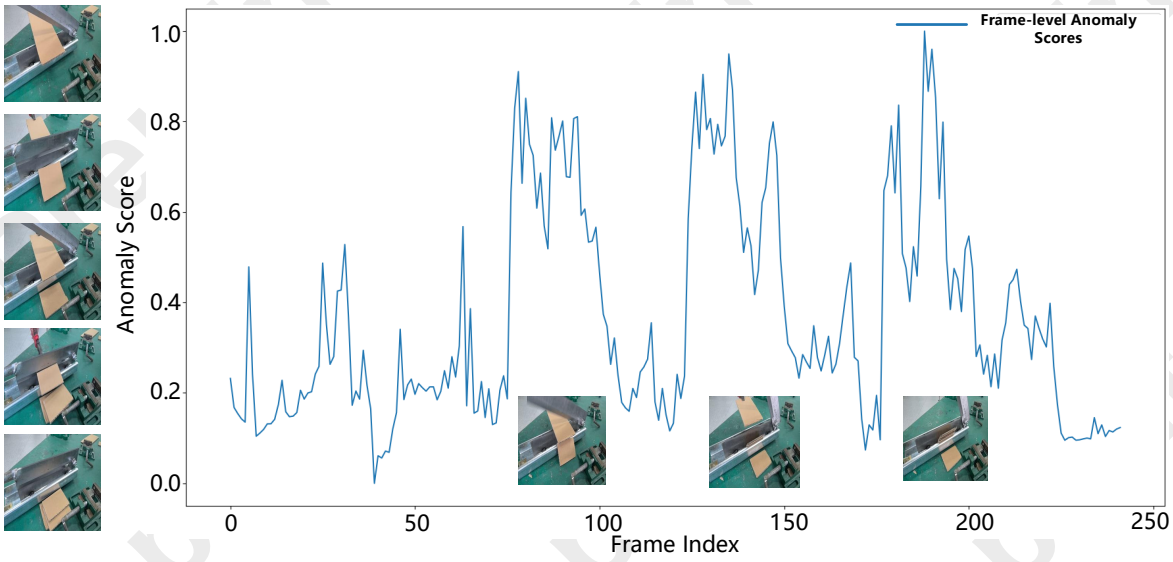

Figure 6: **Visualization of Anomaly Score.** The ground truth is displayed on the left side from top to bottom. It can be observed that the three peaks in the anomaly score correspond to the occurrence of three abnormal events.

4.3 Comparison with State-of-the-arts

We compare our method with six other methods which can be divided into three types: (1) Methods for conventional video anomaly detection: conAE[Hasan *et al.*, 2016], memAE[Gong *et al.*, 2019], AstNet[Le and Kim, 2023], V-Swin-T[Liu *et al.*, 2022]; (2) Method for industrial image anomaly detection: DMAD[Liu *et al.*, 2023]; (3) Method for industrial video anomaly detection: IPAD_VAD[Liu *et al.*, 2024]. It is worth mentioning that V-Swin-T is a model that contains an auto-encoder reconstruction structure utilizing the Video Swin Transformer as the feature extractor. To the best of our knowledge, IPAD_VAD is currently the only model specifically designed for industrial video anomaly detection, aside from our method. Table 1 shows the main experimental results, which indicate the superiority of our proposed network. Our method achieved an average AUC of 71.8% on the IPAD dataset, surpassing the existing state-of-the-art (SOTA) methods.

4.4 Ablation Study

We conduct several ablation experiments on the IPAD dataset as shown in Table 2 to demonstrate the effectiveness of our architecture. The performance of our network is mainly attributed to the two modules: the multi-feature fusion and the trajectory filter. Note that we progressively add additional components to the baseline, enabling us to gauge the performance improvement obtained by each. It can be obviously seen that the multi-feature fusion module and the trajectory filter module bring about 7.2% and 5.9% AUC scores, respectively. And these two modules boost the overall performance, resulting in a 10.5% improvement. Figure 4 shows the ROC curves when evaluating the category S01, using the trained

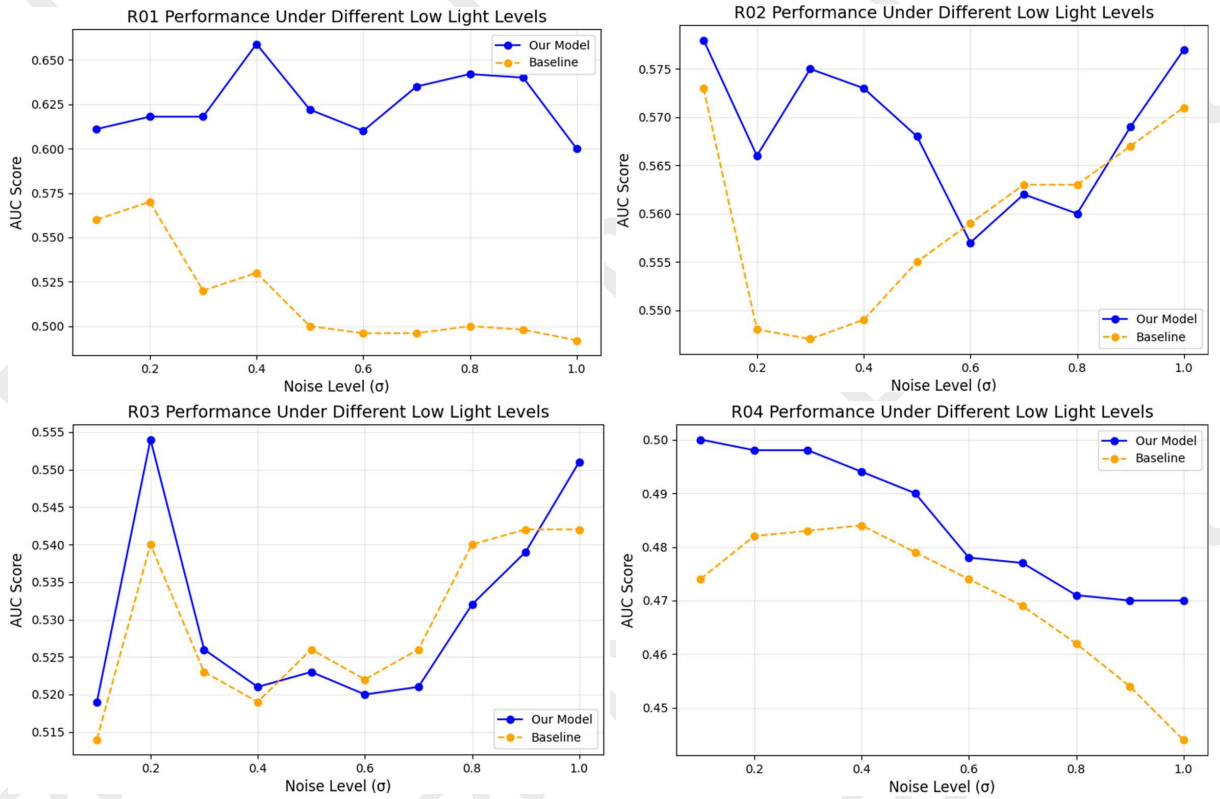

Figure 7: **Robustness Comparison in Low Light Conditions.** Four sub-figures respectively illustrate the AUC score between our model and the baseline model under different level low-light disturbances on {R01-R04}.

baseline, baseline with the multi-feature fusion module, baseline with the trajectory filter module, and baseline with both two modules.

4.5 Discussion

Heat Map Visualization Figure 5 presents the difference in heat maps between the predicted frames and the ground truth frames under various conditions, including normal, speed anomaly, color anomaly, shape anomaly, and angle anomaly scenarios. Under normal conditions, the heat map shows minimal and evenly distributed differences, indicating that the model accurately predicts the video frames with only minor deviations. Regarding the speed anomaly, the heat map highlights significant differences in motion-related regions, demonstrating the model's ability to detect speed variations. For the color anomaly, the heat map reveals intensified differences in specific local areas where color shifts occur, showcasing the model's sensitivity to such changes. The shape anomaly results in noticeable differences around the contours of objects, reflecting the model's capacity to identify anomalies caused by alterations in object shapes. The angle anomaly produces distinct differences in areas affected by angular deviations, indicating the effectiveness of the model in capturing variations caused by object rotation. Overall, the heat maps visually demonstrate the model's capability to detect and localize various types of anomalies in industrial video scenarios.

Anomaly Score Visualization Figure 6 shows the visualization results of the anomaly score for the 18th video in the test set of category R04. We extract some frames under normal conditions as references, arranging them sequentially from top to bottom on the left side of the figure. It can be observed that when anomalies occur three times, our anomaly score also shows three corresponding significant increases. This demonstrates the effectiveness of our method and highlights its ability to adapt to subtle temporal and spatial anomalies.

Robustness To demonstrate the robustness of our method against lighting variations, we test our model and the baseline model under different levels of low-light disturbances on the {R01-R04} dataset, evaluating their performance using AUC scores. Figure 7 shows the results. Compared to baseline methods, our method maintains a relatively high and stable AUC score under varying levels of low-light disturbances, demonstrating its strong robustness.

5 Conclusion

In this work, we propose INFP, an innovative approach that fully leverages frequency-domain information to address the challenges of industrial video anomaly detection. The encoder-decoder architecture consisting of a multi-feature fusion module and a trajectory filter, effectively addresses the unique challenges of industrial scenarios and enhances the robustness of the model. Experimental results on the dataset demonstrate that our method outperforms other approaches.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant (62033012).

References

[Bergmann *et al.*, 2019] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

[Bergmann *et al.*, 2020] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020.

[Deng *et al.*, 2024a] Huilin Deng, Hongchen Luo, Wei Zhai, Yang Cao, and Yu Kang. Prioritized local matching network for cross-category few-shot anomaly detection. *IEEE Transactions on Artificial Intelligence*, 2024.

[Deng *et al.*, 2024b] Huilin Deng, Hongchen Luo, Wei Zhai, Yang Cao, and Yu Kang. Vmad: Visual-enhanced multimodal large language model for zero-shot anomaly detection. *arXiv preprint arXiv:2409.20146*, 2024.

[Doshi and Yilmaz, 2021] Keval Doshi and Yasin Yilmaz. Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition*, 114:107865, 2021.

[Gong *et al.*, 2019] 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.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[Gudovskiy *et al.*, 2022] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 98–107, 2022.

[Hasan *et al.*, 2016] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.

[Le and Kim, 2023] Viet-Tuan Le and Yong-Guk Kim. Attention-based residual autoencoder for video anomaly detection. *Applied Intelligence*, 53(3):3240–3254, 2023.

[Li *et al.*, 2021] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674, 2021.

[Lin *et al.*, 2019] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7083–7093, 2019.

[Liu *et al.*, 2018] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018.

[Liu *et al.*, 2020] Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J Radke, and Octavia Camps. Towards visually explaining variational autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8642–8651, 2020.

[Liu *et al.*, 2022] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[Liu *et al.*, 2023] Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-measurable anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12147–12156, June 2023.

[Liu *et al.*, 2024] Jinfan Liu, Yichao Yan, Junjie Li, Weiming Zhao, Pengzhi Chu, Xingdong Sheng, Yunhui Liu, and Xiaokang Yang. Ipad: Industrial process anomaly detection dataset. *arXiv preprint arXiv:2404.15033*, 2024.

[Pirnay and Chai, 2022] Jonathan Pirnay and Keng Chai. Inpainting transformer for anomaly detection. In *International Conference on Image Analysis and Processing*, pages 394–406. Springer, 2022.

[Salehi *et al.*, 2021] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021.

[Schlegl *et al.*, 2017] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.

[Sultani *et al.*, 2018] 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.

[Tsai *et al.*, 2022] Chin-Chia Tsai, Tsung-Hsuan Wu, and Shang-Hong Lai. Multi-scale patch-based representation learning for image anomaly detection and segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3992–4000, 2022.

[Wang *et al.*, 2024] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22883–22892, 2024.

[Yu *et al.*, 2024] Qianzi Yu, Kai Zhu, Yang Cao, Feijie Xia, and Yu Kang. Tf 2: Few-shot text-free training-free defect image generation for industrial anomaly inspection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[Zaheer *et al.*, 2020] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 358–376. Springer, 2020.

[Zhang *et al.*, 2023] Ji Zhang, Xiao Wu, Zhi-Qi Cheng, Qi He, and Wei Li. Improving anomaly segmentation with multi-granularity cross-domain alignment. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8515–8524, 2023.

[Zhong *et al.*, 2019] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge 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*, pages 1237–1246, 2019.