



Robust Multi-Modal Feature Fusion Network for Product Recognition in Live Streaming Scenes

Xiaoyun Huang 

School of Information Engineering, Guangxi Vocational College of Water Resources and Electric Power, Nanning 530023, China

Corresponding Author Email: sd2085043@163.com

Copyright: ©2026 The author. This article is published by IIETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.430426>

ABSTRACT

Received: 18 April 2026

Revised: 22 July 2026

Accepted: 3 August 2026

Available online: 31 August 2026

Keywords:

live-streaming product recognition, multi-modal feature fusion, weakly supervised saliency detection, cross-modal Transformer, uncertainty modeling, quality-aware temporal aggregation, degradation robustness

The large-scale development of live-streaming e-commerce has created an urgent demand for real-time product image recognition. However, products in live broadcasts often suffer from degradation such as motion blur, partial occlusion, and complex backgrounds due to camera movement, host manipulation, and scene switching. Moreover, fine-grained product categories exhibit subtle inter-class differences, which causes significant bottlenecks in both robustness and discriminability for recognition methods relying solely on single RGB appearance features. To address these challenges, this paper proposes the Saliency-Guided and Quality-Aware Multi-Modal Fusion Network (SG-QMFNet). The method jointly models four complementary cues: RGB appearance, edge structure, weakly supervised saliency, and on-screen Optical Character Recognition (OCR) text. It constructs a structure–appearance–semantic three-stream cross-modal alignment Transformer, where intra-modal self-attention and inter-modal cross-attention interact layer by layer to achieve deep alignment and complementary enhancement of heterogeneous features in a unified semantic space. To suppress interference from low-quality modalities, an uncertainty-driven dynamic gating fusion mechanism is designed. This mechanism uses classification confidence as a self-supervised signal to estimate per-frame reliability of each modality and adaptively adjusts fusion weights through a differentiable gating function. On the temporal dimension, a quality-aware temporal aggregation module is proposed, which jointly regresses frame-level quality scores using clarity metrics, saliency completeness, and detection confidence, and embeds them as attention biases into the temporal encoding process. This enables the selection of highly discriminative information from continuous frames while suppressing the negative impact of degraded frames. The four modules are progressively connected with degradation robustness as the core objective, forming a unified end-to-end recognition framework. Experiments on the self-built live-streaming product dataset LiveProduct-10K show that SG-QMFNet achieves a Top-1 accuracy of 85.7% and an mean Average Precision (mAP) of 87.5%, which are 3.3% and 2.8% higher than the best baseline method, respectively. On subsets with motion blur and partial occlusion, the improvement exceeds 7 percentage points. Generalization experiments on the public fine-grained product dataset Product-100 further verify the effectiveness of the method in static scenarios. Ablation studies and visualization analyses reveal the independent contributions and synergistic mechanisms of each module.

1. INTRODUCTION

The large-scale development of live-streaming e-commerce has made real-time product image recognition a key problem that urgently needs to be solved in the field of computer vision [1]. Unlike static product images, product targets in live-streaming frames are continuously in dynamic change. Camera movement, host hand gestures, and frequent scene switching cause degradation such as motion blur, defocus, and partial occlusion [2]. Overlapping bullet comments and decorative props in the background further increase interference [3]. At the same time, live-streaming product recognition usually targets fine-grained categories, such as lipsticks of different shades or snacks of different flavors,

where inter-class differences are subtle, posing higher requirements for feature discriminability [4]. Traditional recognition methods based on single RGB appearance features suffer a sharp performance drop under degradation conditions [5], making it difficult to meet the dual requirements of real-time performance and robustness in live-streaming scenarios. From the perspective of image processing, edge structure information can still maintain high stability under blur and low-light conditions [6]; saliency information helps suppress complex background interference [7]; and on-screen Optical Character Recognition (OCR) text can provide direct category semantic prior [8]. Therefore, how to effectively fuse these heterogeneous modalities [9] and dynamically select high-quality information in continuous frames [10] constitutes the

core scientific problem for improving live-streaming product recognition performance.

Existing research on product recognition still has obvious shortcomings in the above aspects [11]. First, the discriminability of single-modal features is severely constrained by degradation factors [12]. Most methods rely only on RGB appearance features. Under conditions such as motion blur, surface reflection, and occlusion, appearance texture information is largely lost, while complementary modalities such as edge structure and text semantics are not fully explored and utilized [13]. Second, foreground modeling under complex backgrounds is relatively weak [14]. In live-streaming frames, the product region usually occupies only a small part of the frame. Existing methods mostly use global features or features directly extracted from detection bounding boxes for classification, lacking fine-grained decoupling of the product foreground region [15]. This causes a large amount of background noise to mix into the discriminative features, significantly reducing recognition accuracy [16]. Third, multi-modal fusion strategies are relatively simple. Existing works have attempted to combine text and image modalities, but mostly adopt feature concatenation or fixed-weight addition [17], without considering the reliability differences of different modalities across different frames and under different degradation conditions, thus failing to achieve truly adaptive fusion [18]. Fourth, the utilization of temporal redundant information is insufficient. Live streaming itself is a continuous video stream, and there are significant quality differences and complementary cues between frames [19]. However, existing methods generally perform recognition based on a single frame, so single-frame misjudgments cannot be effectively corrected by temporal context, limiting overall recognition stability [20].

To address the above problems, this paper proposes a Saliency-Guided and Quality-Aware Multi-Modal Fusion Network (SG-QMFNet). The method jointly models four complementary cues: RGB appearance, edge structure, weakly supervised saliency, and on-screen OCR text. It constructs a structure–appearance–semantic three-stream cross-modal alignment Transformer, and achieves deep alignment and complementary enhancement of heterogeneous features through layer-by-layer attention interaction. To suppress interference from low-quality modalities, an uncertainty-driven dynamic gating fusion mechanism is designed, which uses classification confidence as a self-supervised signal to estimate the per-frame reliability of each modality and adaptively adjusts fusion weights. On the temporal dimension, a quality-aware temporal aggregation module is proposed, which jointly regresses frame-level quality scores using clarity metrics, saliency completeness, and detection confidence, and embeds them as attention biases into the temporal encoding process, thereby selecting highly discriminative information from continuous frames and suppressing the negative impact of degraded frames. The above modules are progressively connected with degradation robustness as the core objective, forming a unified end-to-end recognition framework.

The rest of this paper is organized as follows. Section 2 introduces the overall architecture of SG-QMFNet and the technical implementation of each core module in detail; Section 3 presents comparison experiments, ablation experiments, robustness analysis, and visualization discussions on the self-built live-streaming product dataset and the public fine-grained product dataset; Section 4 concludes the paper and looks forward to future research

directions.

2. METHOD

2.1 Overall framework

SG-QMFNet takes a continuous T -frame live-streaming video clip as input. Each frame I_t first passes through a lightweight object detector to extract candidate product regions, and simultaneously constructs four modal representations: RGB image patches, Canny edge maps, weakly supervised saliency maps, and OCR text sequences. The RGB branch adopts EfficientNet-B4 to extract appearance features. The edge branch uses a shallow convolutional network to capture structural information from binary edge maps. The saliency map performs foreground decoupling on RGB features to suppress background interference. The text sequence is encoded into semantic vectors by a lightweight Bidirectional Encoder Representations from Transformers (BERT). The decoupled appearance features, edge features, and text features are converted into token sequences of a unified dimension and fed into a cross-modal alignment Transformer for layer-by-layer interaction, enabling deep alignment of structure, appearance, and semantic information in a shared representation space. The aligned global representations of each modality are adaptively fused via an uncertainty-driven dynamic gating mechanism to obtain the frame-level representation $F_{fused}^{(t)}$. The quality-aware temporal aggregation module then performs weighted encoding on the fused features of T consecutive frames, outputs a video-level product representation, and feeds it into a classifier to complete the final recognition. The detailed design of each module is presented in turn below.

2.2 Multi-modal input construction and feature encoding

The original resolution of live-streaming frames is usually 1920×1080 or 1280×720 . Directly performing recognition on the full frame would introduce a large amount of background noise and incur high computational cost. This paper first uses a You Only Look Once version 5 small (YOLOv5s) detector pre-trained on the COCO dataset and fine-tuned on self-built live-streaming product detection data to extract all candidate product regions, with a confidence threshold of 0.3 and an Intersection over Union (IoU) threshold of 0.5 for non-maximum suppression. For each detected product bounding box, the margin is expanded outward by 10% to retain necessary contextual information, and then the region is cropped and resized to a resolution of 224×224 , yielding the RGB image patch I_t^{rgb} . When multiple product targets exist in a frame, each target independently enters the subsequent pipeline for separate recognition. For matching the same target across consecutive frames, a joint criterion of bounding box IoU and appearance feature similarity is used for simple tracking, ensuring that the features of each frame in the temporal aggregation stage correspond to the same product instance.

Motion blur, low lighting, and partial occlusion in live-streaming frames significantly damage RGB appearance features, causing a large loss of texture details. However, structural information such as the contour boundaries of products and the edges of packaging text is largely preserved. Based on this observation, this paper introduces an edge

structure modality to provide stable discriminative cues for degraded frames. For each RGB image patch I_t^{rgb} , it is first converted to a grayscale image and smoothed with a Gaussian filter with a kernel size of 5×5 and a standard deviation of 1.0 to suppress sensor noise. Then the Canny edge detector is applied with a low threshold of 50 and a high threshold of 150 to obtain a binary edge map $I_t^{edge} \in \{0,1\}^{224 \times 224}$. The edge branch uses a three-layer shallow convolutional network for feature extraction: the first layer uses a 3×3 convolution with stride 2 and 32 channels, followed by Batch Normalization (BatchNorm) and Rectified Linear Unit (ReLU); the second layer uses a 3×3 convolution with stride 2 and 64 channels, followed by BatchNorm and ReLU; the third layer uses a 3×3 convolution with stride 2 and 112 channels, followed by BatchNorm and ReLU. The output feature map $F_{edge} \in \mathbb{R}^{112 \times 28 \times 28}$ is strictly aligned in spatial size with the output of the third stage of the RGB backbone network. Since the edge branch network has limited depth, its computational

overhead is negligible while it can quickly respond to spatial changes in structural information.

The RGB appearance branch adopts EfficientNet-B4 as the backbone network. The network is pre-trained on ImageNet and has good feature extraction capability and computational efficiency. This paper selects the output feature map of its third stage as the appearance representation, because this stage achieves a good balance between semantic abstraction level and spatial resolution. It not only retains sufficient local details to support fine-grained discrimination, but also has a certain semantic abstraction capability to cope with appearance changes. The input 224×224 image patch is processed by the third stage to obtain the feature map $F_{rgb} \in \mathbb{R}^{112 \times 28 \times 28}$. The parameters of the first two stages of the backbone network are frozen in the first training stage to stabilize the feature distribution, and the parameters of the third stage and subsequent stages are kept trainable.

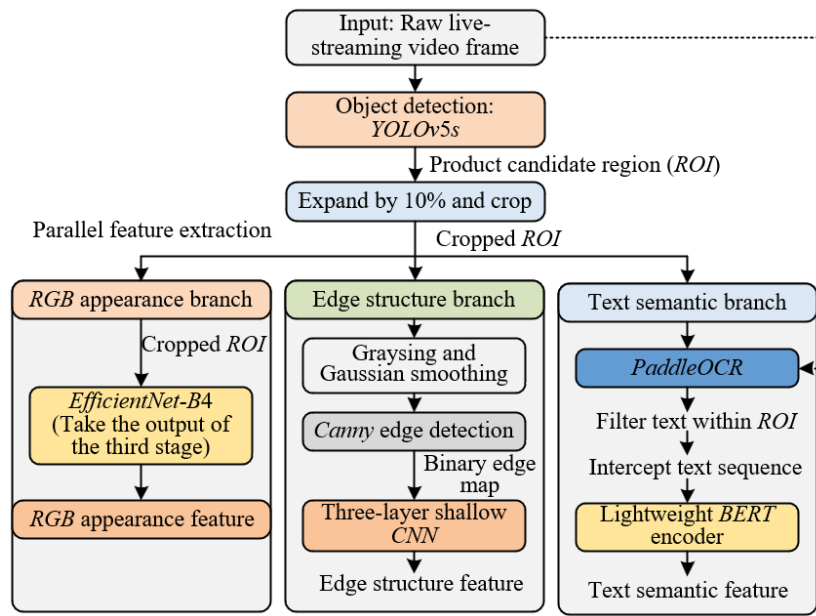


Figure 1. Schematic diagram of multi-modal input construction and feature extraction module

Note: ROI: Region of Interest; YOLOv5s: You Only Look Once version 5 small; BERT: Bidirectional Encoder Representations from Transformers; CNN: Convolutional Neural Network; PaddleOCR: Paddle Optical Character Recognition.

Live-streaming frames contain rich text information, including product titles, price tags, brand names, and high-frequency keywords in bullet comments. These texts have a strong semantic indication effect on product categories. This paper uses the PaddleOCR engine to perform text detection and recognition on the entire frame, supporting mixed Chinese and English scenarios. For each candidate product region, all text boxes whose IoU with its bounding box is greater than 0.2 are selected, sorted in descending order by detection confidence, and the top $L = 32$ text tokens are intercepted. If the number of valid tokens is less than 32, it is padded with a special padding token. Each text token is mapped to a vector through a lightweight BERT encoder. The encoder contains 2 Transformer layers, with a hidden dimension of 112 and 4 attention heads, and finally obtains the text sequence feature $F_{text} \in \mathbb{R}^{L \times 112}$. When the OCR engine fails to detect any text, a learnable empty text token is input, so that the text modality degenerates into a fixed category-independent prior, avoiding invalid information from interfering with the subsequent fusion process. Figure 1 shows the schematic diagram of

multi-modal input construction and feature extraction module.

2.3 Saliency-guided product region decoupling

In live-streaming scenarios, product regions are often partially occluded by host hand gestures, props, background decorations, and bullet comments. Directly performing global pooling on candidate regions will introduce a large amount of background noise unrelated to the product. To this end, this paper constructs a weakly supervised saliency generation network to generate a product foreground probability map using only category labels, and accordingly performs foreground enhancement and background suppression on RGB features. Figure 2 shows the principle of the RGB feature foreground decoupling mechanism based on weakly supervised saliency. The saliency decoder takes the feature map $F_{rgb} \in \mathbb{R}^{112 \times 28 \times 28}$ output from the third stage of the RGB backbone network as input. Its structure consists of two convolutional layers: the first layer is a 3×3 convolution with stride 1 and padding 1, reducing the number of channels from

112 to 64, followed by BatchNorm and ReLU; the second layer is a 3×3 convolution with stride 1 and padding 1, reducing the number of channels from 64 to 1, followed by Sigmoid activation. The output initial saliency map $S_{init} \in [0,1]^{28 \times 28}$, where the value at each spatial location indicates the probability that the location belongs to the product foreground. Since only category labels are provided during the training process, a weak supervision mechanism needs to be established to guide the saliency map to focus on discriminative regions. This paper uses the Gradient-weighted Class Activation Mapping (Grad-CAM) of the classification branch to generate a class activation map $M_{cam} \in [0,1]^{28 \times 28}$. The calculation method is as follows: F_{rgb} is processed by global average pooling and then sent to a fully connected layer to obtain classification logits. The gradient of the target category is back-propagated to F_{rgb} , and the global average of

the gradient for each channel is calculated to obtain channel weights. After weighted summation, ReLU and normalization are applied to obtain the class activation map. Then a center constraint is applied to the initial saliency map to keep its high-response region consistent with the high-response region of the class activation map. The constraint loss is defined as:

$$L_{center} = \frac{1}{|\Omega|} \sum_{p \in \Omega} (S_{init}(p) - \bar{M}_{cam}(p))^2 \quad (1)$$

where, Ω is the set of spatial locations, and $\bar{M}_{cam}$ is the normalized class activation map. This loss forces the main response of the saliency map to coincide with the classification discriminative region, thereby achieving foreground localization without pixel-level annotations.

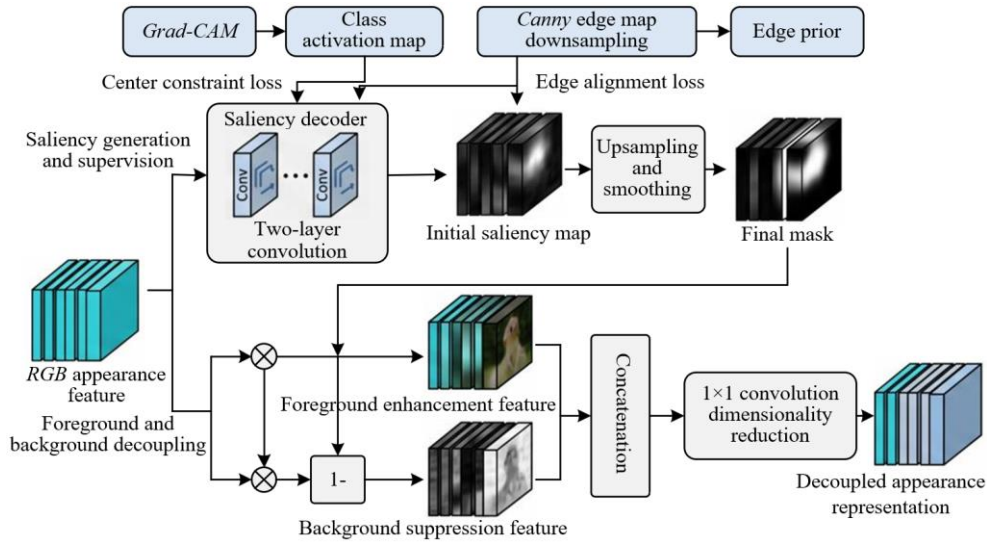


Figure 2. RGB feature foreground decoupling mechanism based on weakly supervised saliency

The initial saliency map is usually blurry at the boundary and difficult to accurately fit the product contour. To alleviate this problem, an edge prior constraint is introduced to align the spatial gradient distribution of the saliency map with the edge structure information. Specifically, the binary edge map I_t^{edge} is first downsampled to 28×28 , denoted as $I_t'^{edge}$, and then the edge alignment loss is defined as:

$$L_{edge} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \|\nabla S_{init}(p) - \nabla I_t'^{edge}(p)\|_2^2 \quad (2)$$

where, ∇ represents the spatial gradient calculated by the Sobel operator, and $I_t'^{edge}$ is the edge prior after downsampling the original binary edge map I_t^{edge} to 28×28 . This loss encourages the saliency map to produce a gradient change of corresponding amplitude at locations with strong edge response, so that the saliency boundary tends to the product contour. It should be noted that the Canny edge map simultaneously contains edges of background objects, but the center constraint loss has pulled the saliency main body to the product region. Under the joint action of the two, the saliency map will preferentially fit the boundary of the product main body. The final saliency map S_t is obtained by upsampling the initial map S_{init} to 224×224 and applying Gaussian smoothing,

with a smoothing kernel size of 3×3 and a standard deviation of 0.8. In the inference stage, the saliency map can be directly obtained through a single forward propagation without any additional annotations.

After obtaining the saliency map, it is downsampled to 28×28 to obtain S_t' . Foreground enhancement and background suppression operations are performed on the RGB features, specifically calculated as:

$$F_{rgb}^f = F_{rgb} \odot S_t', \quad F_{rgb}^{bg} = F_{rgb} \odot (1 - S_t') \quad (3)$$

where, $\odot$ denotes element-wise multiplication. The foreground enhancement feature F_{rgb}^f amplifies the feature response of the product main body region, and the background suppression feature F_{rgb}^{bg} retains the residual information of the non-main body region. The two are concatenated with the original feature F_{rgb} along the channel dimension to obtain a feature map with 336 channels, which is then reduced to 112 dimensions by a 1×1 convolution, forming the decoupled appearance representation $F_{rgb}^{dec} \in \mathbb{R}^{112 \times 28 \times 28}$. This representation significantly strengthens the foreground discriminative information while retaining the global context, providing a cleaner appearance feature for subsequent cross-modal alignment.

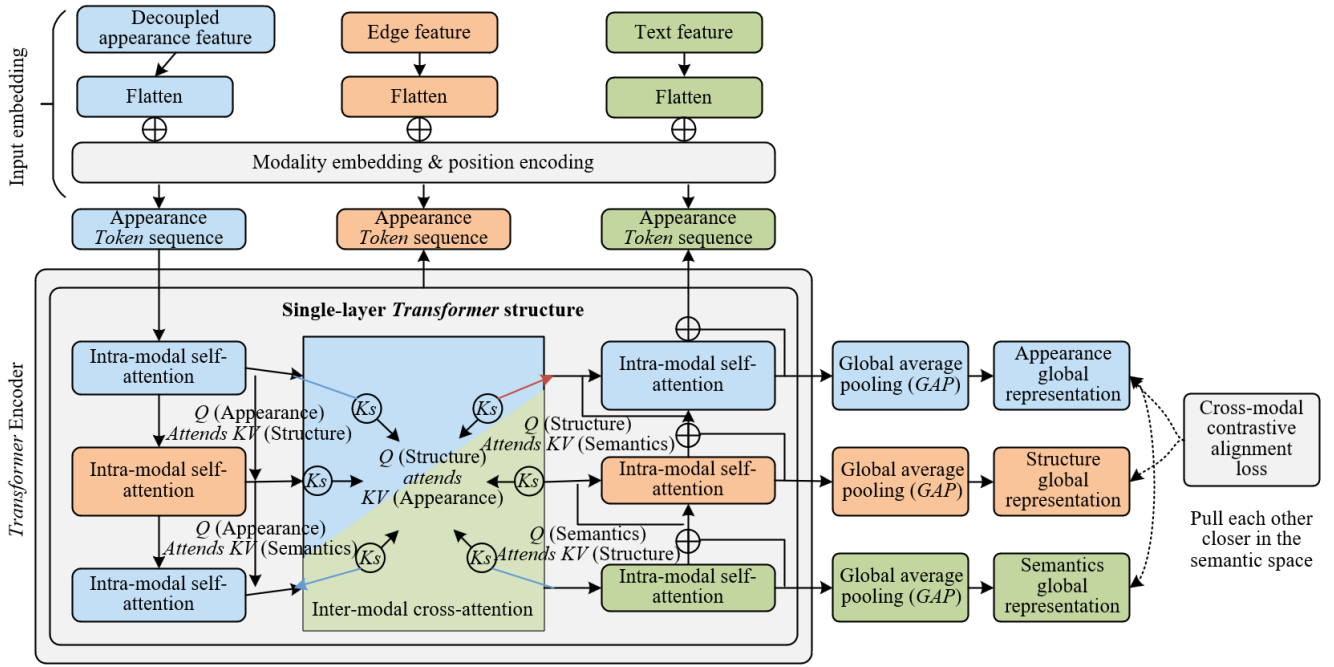


Figure 3. Architecture of the structure–appearance–semantics three-stream cross-modal alignment transformer

2.4 Structure–appearance–semantics three-stream cross-modal alignment transformer

To fuse the decoupled appearance features, edge structure features, and textual semantic features, this paper constructs a three-stream cross-modal alignment Transformer. It maps the three modalities into a unified token space and achieves information complementarity through layer-by-layer attention interaction. Figure 3 illustrates the architecture of the structure–appearance–semantics three-stream cross-modal alignment Transformer. First, the decoupled appearance feature $F_{rgb}^{dec} \in \mathbb{R}^{112 \times 28 \times 28}$ and the edge feature $F_{edge} \in \mathbb{R}^{112 \times 28 \times 28}$ are flattened into spatial token sequences $X^v = \{x_1^v, \dots, x_N^v\}$ and $X^e = \{x_1^e, \dots, x_N^e\}$, respectively, where $N = 28 \times 28 = 784$, and the dimension of each token is 112. The text feature $F_{text} \in \mathbb{R}^{L \times 112}$ is itself a sequence, with the number of tokens $L = 32$, and its dimension is consistent with the visual tokens. The three types of tokens share the same feature dimension, facilitating unified attention calculation in the subsequent steps. To distinguish the modality sources, learnable modality type embeddings $e^v, e^e, e^t \in \mathbb{R}^{112}$ are added to each token, corresponding to the appearance, structure, and semantics modalities, respectively. For the two types of visual tokens (appearance and structure), learnable spatial position encodings $P \in \mathbb{R}^{N \times 112}$ are also appended to preserve spatial structural information. The final token sequence representations input to the Transformer are:

$$\tilde{X}^v = X^v + e^v + P, \quad \tilde{X}^e = X^e + e^e + P, \quad \tilde{X}^t = X^t + e^t \quad (4)$$

The cross-modal alignment Transformer is stacked with $K = 3$ layers. Each layer contains an intra-modal self-attention sub-layer and an inter-modal cross-attention sub-layer, followed by a residual connection and LayerNorm for each sub-layer. The intra-modal self-attention separately applies standard multi-head self-attention to each modality to model the global dependencies within each modality. Taking the calculation of the appearance modality at layer as an example, its self-attention output is:

$$\tilde{X}^{v,(k)} = \text{LayerNorm}(\tilde{X}^{v,(k-1)} + \text{SelfAttn}(\tilde{X}^{v,(k-1)})) \quad (5)$$

where, $\text{SelfAttn}()$ is the multi-head self-attention, with the number of heads set to 4, and the dimension of each head set to 28. The self-attention calculation methods for the structure and semantics modalities are the same, but the projection parameters of each modality are independent of each other. In the inter-modal cross-attention sub-layer, pairwise interactions are performed between the modalities. Taking the case where appearance tokens act as queries and structure tokens act as keys and values as an example, the cross-attention is calculated as:

$$\text{CrossAttn}(\tilde{X}^v, \tilde{X}^e) = \text{softmax}\left(\frac{Q_v K_e^T}{\sqrt{d}}\right) V_e \quad (6)$$

where, $Q_v = \tilde{X}^{v,(k)} W_Q^v$, $K_e = \tilde{X}^{e,(k)} W_K^e$, $V_e = \tilde{X}^{e,(k)} W_V^e$, $W_Q^v, W_K^e, W_V^e \in \mathbb{R}^{112 \times 112}$ are learnable projection matrices, and $d = 112$ is the feature dimension. The output of the cross-attention is processed by a residual connection and LayerNorm to obtain the updated appearance tokens:

$$\tilde{X}^{v,(k)} = \text{LayerNorm}(\tilde{X}^{v,(k)} + \text{CrossAttn}(\tilde{X}^{v,(k)}, \tilde{X}^{e,(k)})) \quad (7)$$

Similarly, appearance tokens also perform cross-attention with semantic tokens, structure tokens interact with appearance and semantic tokens respectively, and semantic tokens interact with appearance and structure tokens respectively. In each layer, each modality performs cross-attention once with each of the other two modalities, resulting in a total of six pairwise interactions, thereby achieving sufficient inter-modal information flow at a controllable computational cost.

After $K = 3$ stacking layers, the aligned three types of token representations Z^v, Z^e, Z^t are obtained. The token sequences of each modality are average-pooled to obtain the global representation of that modality:

$$\bar{z}^v = \frac{1}{N} \sum_{i=1}^N z_i^v, \quad \bar{z}^e = \frac{1}{N} \sum_{i=1}^N z_i^e, \quad \bar{z}^t = \frac{1}{L} \sum_{i=1}^L z_i^t \quad (8)$$

To explicitly pull closer the distance between representations of the same product under different modalities while increasing the modal distance between different products, a cross-modal contrastive alignment loss is introduced. For the samples within a batch B , two contrastive terms—appearance vs. structure and appearance vs. semantics—are constructed. The loss function is defined as:

$$L_{align} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\text{sim}(\bar{z}_i^v, \bar{z}_i^e)/\tau) + \exp(\text{sim}(\bar{z}_i^v, \bar{z}_i^t)/\tau)}{\sum_{j=1}^B [\exp(\text{sim}(\bar{z}_i^v, \bar{z}_j^e)/\tau) + \exp(\text{sim}(\bar{z}_i^v, \bar{z}_j^t)/\tau)]} \quad (9)$$

where, $\text{sim}(a, b)$ denotes cosine similarity, and $\tau = 0.07$ is the temperature parameter. This loss encourages the appearance representation of the same product to be close to its structure and semantics representation in the embedding space, thereby enhancing the information consistency between different modalities and promoting the subsequent fusion stage to fully utilize multi-modal complementary cues. During training, the samples within a batch come from different categories, so negative sample pairs are formed naturally without requiring additional construction.

2.5 Uncertainty-driven dynamic modal gating fusion

After obtaining the three types of global representations from cross-modal alignment, the reliability of different modalities in the same frame often varies significantly. In motion-blurred frames, edge structure features are usually more stable than RGB appearance features; while when OCR text is missing or recognition is incorrect, the text semantic modality should not be given excessive trust. To adapt to this frame-by-frame variation in modal reliability, this paper introduces an uncertainty estimation branch for each modality. For the global representations of the appearance, structure, and semantics modalities $\bar{z}^v, \bar{z}^e, \bar{z}^t$, each is mapped to a scalar uncertainty score through a two-layer Multi-Layer Perceptron (MLP). The input dimension of the MLP is 112, the hidden

layer dimension is 64, ReLU activation is used, the output layer is 1-dimensional, and it is finally activated by softplus to ensure non-negativity. Taking the appearance modality as an example, its uncertainty score is calculated as:

$$u_v = \log(1 + \exp(\text{MLP}_v(\bar{z}^v))) \quad (10)$$

The uncertainty scores u_e, u_t for the structure and semantics modalities are obtained in the same way. The smooth property of the softplus activation is beneficial for gradient backpropagation and training stability. The larger the value of $u_m \geq 0$, the lower the reliability of that modality in the current frame. The introduction of this branch enables the network to quantitatively evaluate the credibility of each modality based on its own feature state, providing a basis for subsequent adaptive fusion.

Based on the above uncertainty scores, the gating weights corresponding to each modality are calculated. For the appearance, structure, and semantics modalities, the weights are defined as:

$$g_m = \frac{\exp(-u_m/\tau_g)}{\sum_{i \in \{v, e, t\}} \exp(-u_i/\tau_g)} \quad (11)$$

where, $\tau_g = 0.5$ is the temperature hyperparameter, used to control the smoothness of the weight distribution. When the uncertainty score of a certain modality is significantly higher than the others, its corresponding gating weight will be automatically reduced; conversely, modalities with lower uncertainty will obtain larger fusion weights. The sum of the three gating weights is always 1, ensuring the scale stability of the fused representation. The final frame-level fused feature is obtained by weighted summation of the global representations of each modality according to the gating weights, namely:

$$F_{fused} = g_v \cdot \bar{z}^v + g_e \cdot \bar{z}^e + g_t \cdot \bar{z}^t \quad (12)$$

This dynamic weighting method enables the fusion process to adaptively adjust according to the modal reliability of each frame, avoiding the suboptimal problems that fixed weights or simple concatenation may cause under different degradation conditions.

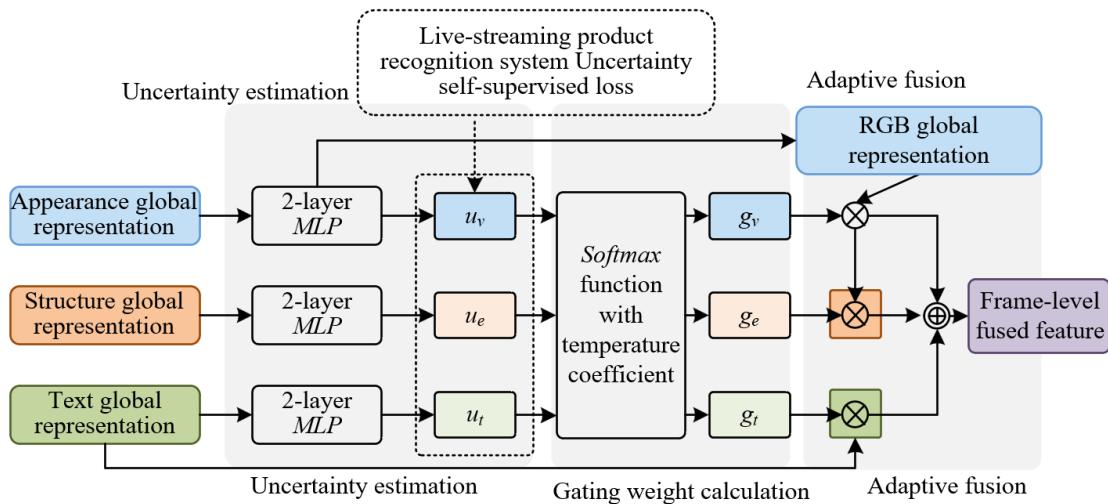


Figure 4. Uncertainty-driven dynamic modal gating fusion module

To give the uncertainty estimation a clear semantic meaning, classification confidence is used as a self-supervised signal for constraint. The global representations of the three modalities $\bar{z}^v, \bar{z}^e, \bar{z}^t$ are respectively fed into a shared fully connected classifier to obtain their respective predicted probability distributions p_v, p_e, p_t . Taking the predicted probability corresponding to the ground-truth category p_m^{gt} , the target uncertainty of this modality is defined as $1-p_m^{gt}$, which means that the lower the prediction confidence, the higher the uncertainty should be. Thus, a mean square error loss is constructed:

$$L_{unc} = \frac{1}{3} \sum_{m \in \{v, e, t\}} \left(u_m - (1-p_m^{gt}) \right)^2 \quad (13)$$

This loss guides the uncertainty estimation network to learn the correspondence between modal reliability and classification confidence. During training, the predicted probabilities of each modality are synchronously updated through the shared classifier, and the uncertainty branch adjusts its output accordingly. In the inference stage, the uncertainty score can be directly obtained by forward calculation of the MLP without relying on additional output from the classifier, thus not increasing inference complexity. Through the above self-supervised mechanism, dynamic gating fusion can perceive modal quality at the semantic level, thereby effectively suppressing the interference of low-quality modalities in the frame-level representation. Figure 4 shows the complete architecture of the uncertainty-driven dynamic modal gating fusion module.

2.6 Quality-aware temporal feature aggregation

The quality difference between consecutive frames of live-streaming video is significant. Some frames are clear and without occlusion, while some frames may suffer from blur, occlusion, or illumination degradation due to camera movement or target interaction. If all frames are aggregated with the same weight, the noise in low-quality frames will weaken the discriminative contribution of high-quality frames. To this end, this paper constructs a quality-aware temporal aggregation module, which estimates a comprehensive quality score for each frame and introduces this score as a bias in temporal attention to achieve selective aggregation. The calculation of the frame-level quality score is based on three complementary indicators. The sharpness score is measured by the variance of the Laplacian. The Laplacian operator response is calculated on the grayscale image patch I_t^{gray} and its variance is computed, namely:

$$q_t^{sharp} = \frac{1}{N_p} \sum_p \left(\nabla^2 I_t^{gray}(p) - \mu_{lap} \right)^2 \quad (14)$$

where, ∇^2 represents the 3×3 Laplacian convolution kernel, μ_{lap} is the mean of the Laplacian response, and N_p is the total number of pixels. A higher sharpness score indicates that the frame image is clearer. The saliency completeness is defined as the spatial mean of the saliency map $\bar{S}_t$, reflecting the proportion of the product foreground in the candidate region. When severe occlusion occurs, the foreground region shrinks and the mean decreases accordingly. The detection confidence $c_t \in [0, 1]$ is given by the object detector, indicating the reliability of the candidate box. The three indicators are concatenated into a three-dimensional vector and then input

into a lightweight regression network. The network is a three-layer MLP with an input dimension of 3, a hidden layer dimension of 16, and the output is activated by Sigmoid to obtain the frame-level comprehensive quality score:

$$q_t = \sigma \left(\text{MLP}([q_t^{sharp}, \bar{S}_t, c_t]) \right) \quad (15)$$

where, $q_t \in (0, 1)$, and a larger value indicates higher frame quality. The regression network is jointly trained with the overall framework without requiring additional annotations. The selection of the three indicators takes into account the quality attributes of image sharpness, foreground completeness, and detection reliability.

After obtaining the quality scores of each frame, they are embedded into the aggregation process as additive biases of temporal attention. The frame-level fused features of T consecutive frames $F_{fused}^{(1)}, \dots, F_{fused}^{(T)}$ are formed into a sequence and input to a single-layer Transformer encoder, with the dimension of each feature being 112. In the self-attention calculation, after the dot product of query and key is scaled, it is linearly added to the quality score vector. The attention weight calculation formula is:

$$\text{Attn}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d}} + \lambda_q \mathbf{1}q^T \right) V \quad (16)$$

where, Q, K, V are the query, key, and value matrices, respectively, $d = 112$ is the feature dimension, $q = [q_1, \dots, q_T]^T \in \mathbb{R}^{T \times 1}$ is the quality score vector, $\mathbf{1} \in \mathbb{R}^{T \times 1}$ is an all-ones vector, so $\mathbf{1}q^T \in \mathbb{R}^{T \times T}$ adds the quality score of each frame as an additive bias on the key dimension. λ_q is a learnable scaling factor, initialized to 0.1. The introduction of this bias term makes the attention weight tilt toward high-quality frames on the basis of the original content similarity, while preserving the global temporal dependency between frames, avoiding information discontinuity that may be caused by hard threshold screening. The Transformer encoder adopts 4 attention heads, the hidden dimension of the feed-forward network is 256, and the Dropout rate is 0.1, which is consistent with the training configuration of the overall framework.

The output sequence of the encoder is average-pooled to obtain the video-level product representation:

$$F_{video} = \frac{1}{T} \sum_{t=1}^T h_t \quad (17)$$

where, h_t is the output vector of the t -th frame after passing through the Transformer encoder. This video-level representation fuses multi-frame information and implicitly suppresses low-quality frames according to the quality scores, and then is fed into a fully connected classifier to obtain the final category prediction. Through this aggregation method, single-frame misjudgments can be effectively corrected by the temporal context, while the discriminative information of high-quality frames is enhanced, thereby improving recognition stability at the video level. The computational overhead of the entire quality-aware temporal aggregation module only adds a single-layer Transformer, which does not cause a significant burden on real-time inference. Figure 5 shows the complete principle of the quality-aware temporal feature aggregation module and video-level classification.

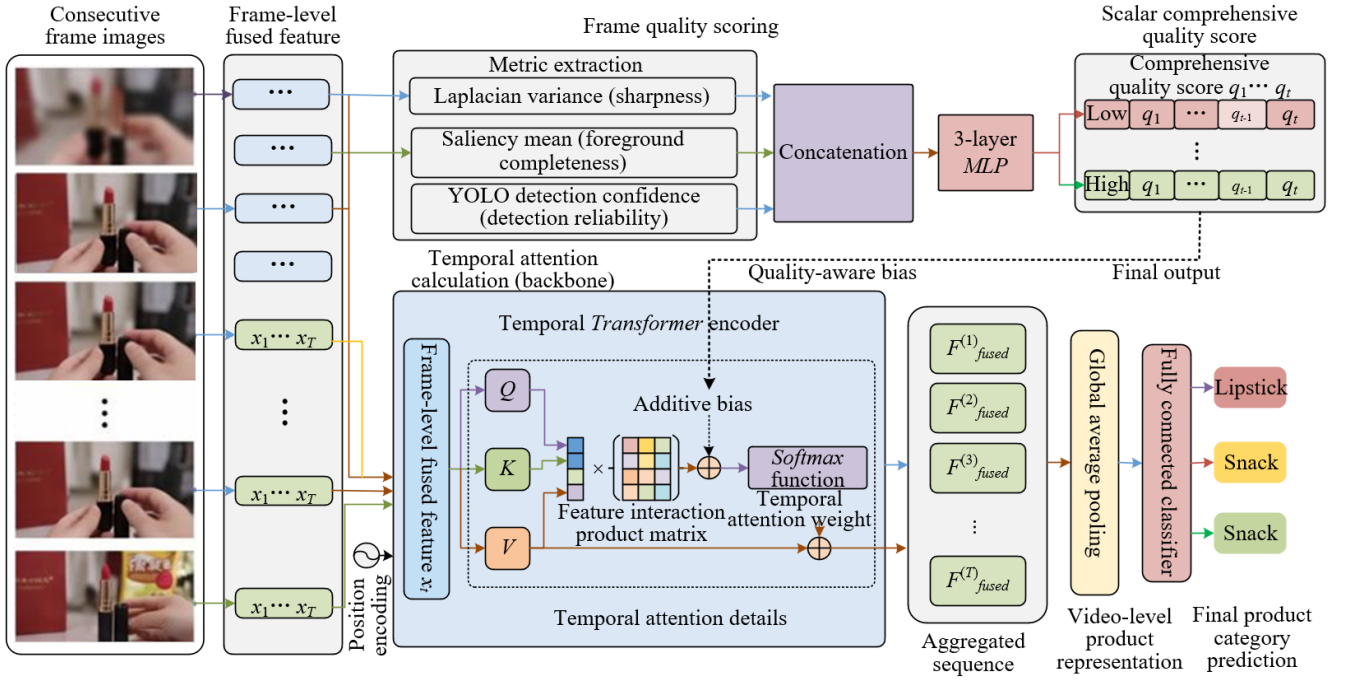


Figure 5. Quality-aware temporal feature aggregation module and video-level classification

Note: YOLO: You Only Look Once; MLP: Multi-Layer Perceptron.

2.7 Loss function

The training of SG-QMFNet is completed by jointly optimizing five losses. The total loss is defined as:

$$L_{total} = L_{cls} + \lambda_1 L_{sal} + \lambda_2 L_{align} + \lambda_3 L_{unc} + \lambda_4 L_{temp} \quad (18)$$

where, L_{cls} is the video-level classification cross-entropy loss, calculated as $-\log p_{video}^{gt}$, and p_{video}^{gt} represents the probability corresponding to the ground-truth category in the video-level prediction. L_{sal} is the weakly supervised saliency loss, composed of the center constraint loss L_{center} and the edge alignment loss L_{edge} , which respectively constrain the main response location and boundary gradient distribution of the saliency map. L_{align} is the cross-modal contrastive alignment loss, used to pull closer the representation distances of the same product under appearance, structure, and semantic modalities. L_{unc} is the uncertainty self-supervised loss, which uses the classification confidence of each modality as the target signal to guide the uncertainty estimation branch to output scores consistent with modal reliability. L_{temp} is the temporal consistency loss, defined as:

$$L_{temp} = \frac{1}{T-1} \sum_{t=1}^{T-1} \|F_{fused}^{(t)} - F_{fused}^{(t+1)}\|_2^2 \quad (19)$$

This term constrains the fused features of adjacent frames to change smoothly, suppressing drastic fluctuations in inter-frame representations. The five loss terms act on different stages such as saliency generation, cross-modal alignment, gated fusion, and temporal aggregation, forming a complete supervision for the overall framework.

The weight coefficients of each loss term are determined by grid search on the validation set, and are finally set to $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$, $\lambda_4 = 0.1$. Training is conducted in two stages: in the first stage, the parameters of the first two stages of the

backbone network are frozen, and only the saliency decoder, edge branch, text encoder, gated fusion module, and classifier are trained for 10 epochs; in the second stage, all trainable parameters are unfrozen for end-to-end fine-tuning for 30 epochs. The optimizer uses Adam with Weight Decay (AdamW), the batch size is set to 32, the initial learning rate is 1×10^{-4} , the weight decay is 0.05, and the learning rate decays according to the cosine annealing strategy. The number of input frames is fixed to $T=5$, and the resolution of image patches is uniformly 224×224 . The data augmentation strategies include random cropping, horizontal flipping, color jittering, and random occlusion simulation. The latter approximates local occlusion caused by host hand gestures and props by randomly covering rectangular areas in the image patch, enhancing the model's robustness to occlusion scenarios.

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Dataset and experimental setup

There is a lack of fine-grained product recognition benchmarks for live-streaming scenarios in existing public datasets. This paper constructs the LiveProduct-10K dataset. The data is sourced from public replay videos of multiple real e-commerce live-streaming platforms, covering 50 fine-grained product categories, including 12 categories of lipstick, 10 categories of snacks, 8 categories of beverages, 7 categories of phone cases, and 13 categories of skincare products. For each category, video clips containing the target product are intercepted from different live rooms and different time periods. Each clip lasts about 2 seconds, from which 5 frames are sampled at equal intervals as representative frames. Each frame is annotated with product bounding boxes and category labels, and the degradation type is recorded. The dataset contains a total of 10,000 clips, which are divided into training, validation, and test sets in an 8:1:1 ratio. Table 1 shows the detailed statistics of the dataset.

Table 1. Statistics of the LiveProduct-10K dataset

Statistic Item	Value
Total number of clips	10,000
Total number of frames	50,000
Number of categories	50
Average number of clips per category	200
Average number of product bounding boxes per frame	1.3
Training/Validation/Test clips	8,000/1,000/1,000
Proportion of clips with motion blur	22.4%
Proportion of clips with partial occlusion	31.7%
Proportion of clips with low lighting	18.9%
Proportion of clips with complex background	45.2%
Proportion of clips without degradation	27.6%

As can be seen from Table 1, the dataset exhibits significant diversity in the distribution of degradation types. More than 70% of the clips contain at least one degradation scenario, with complex backgrounds and partial occlusion having the highest proportions, reaching 45.2% and 31.7%, respectively. This distribution characteristic is highly consistent with the visual conditions in real live-streaming scenarios and can effectively test the recognition capability of the method under non-ideal conditions.

To verify the generalization performance, experiments were also conducted on the public fine-grained product dataset Product-10K. A subset named Product-100 was formed by selecting 100 categories containing at least 50 samples, totaling approximately 50,000 static product images. Since this dataset consists of static images, the experiment runs in single-frame inference mode and does not enable the quality-aware temporal aggregation module to ensure a fair comparison.

The evaluation metrics include Top-1 accuracy, Top-5 accuracy, and mean Average Precision (mAP). Frame-level inference speed (FPS) is also reported to evaluate computational efficiency. All experiments were completed on a single NVIDIA RTX 3090 GPU, implemented based on PyTorch 1.12. The backbone network EfficientNet-B4 is pre-trained on ImageNet, the edge branch and saliency decoder are randomly initialized, and the lightweight BERT encoder uses pre-trained weights from Chinese Wikipedia. The number of input frames is $T=5$, and the resolution of image patches is 224×224 . Training uses the AdamW optimizer with a batch size of 32, an initial learning rate of 1×10^{-4} , cosine annealing scheduling, and a total of 40 epochs.

3.2 Comparison experiments with existing methods

To comprehensively evaluate the performance of SG-QMFNet, three types of representative methods are selected for comparison: single-modal image classification methods include Residual Network (ResNet)-50, Efficient Network-B4, and Swin Transformer (Swin-T); fine-grained recognition methods include Neural Teacher Student Network (NTS-Net), Contrastive Attention Learning (CAL), and Transformer for Fine-Grained Recognition (TransFG); multi-modal fusion methods include Vision-and-Language Transformer (ViLT) and ALign Before Fuse (ALBEF).

All methods use the same data split and augmentation strategy, and are retrained to convergence on LiveProduct-10K. Table 2 shows the Top-1, Top-5, mAP, and FPS of each method on the test set.

As can be seen from Table 2, SG-QMFNet achieves the best

results on all three metrics of Top-1, Top-5, and mAP, reaching 85.7%, 96.2%, and 87.5%, respectively, which are 3.3 percentage points, 1.4 percentage points, and 2.8 percentage points higher than the best baseline TransFG. This improvement margin is of practical significance in fine-grained recognition tasks, indicating that multi-modal complementary information and degradation robustness design can bring stable performance gains. It is worth noting that ViLT and ALBEF, which introduce the text modality, do not significantly outperform pure vision methods. ViLT's Top-1 is even 1.2 percentage points lower than EfficientNet-B4. The reason is that the OCR text in live-streaming frames contains a large amount of noise and redundant information, and simple image-text joint encoding makes it difficult to effectively screen out semantic cues that are valuable for category discrimination. SG-QMFNet dynamically adjusts the contribution weight of the text modality through uncertainty gating, fully utilizing the semantic prior when the OCR result is reliable, and automatically reducing its influence when text is missing or erroneous, thereby avoiding the interference of noisy text on the fused representation. In terms of inference speed, the frame-level FPS of SG-QMFNet is 67. Although it is lower than lightweight methods such as ResNet-50 and EfficientNet-B4, it still maintains high single-frame processing efficiency. The efficiency loss mainly comes from the cross-modal Transformer and quality-aware temporal aggregation. Thanks to the lightweight design of each module, the overall computational overhead remains within an acceptable range.

Table 2. Comparison results with existing methods on the LiveProduct-10K test set

Method	Top-1 (%)	Top-5 (%)	mAP (%)	FPS (frames/s)
<i>ResNet-50</i>	76.3	91.2	78.5	156
<i>EfficientNet-B4</i>	79.8	93.5	82.1	142
<i>Swin-T</i>	81.2	94.1	83.6	98
<i>NTS-Net</i>	80.5	93.8	82.9	85
<i>CAL</i>	81.8	94.3	84.0	76
<i>TransFG</i>	82.4	94.8	84.7	63
<i>ViLT</i>	78.6	92.7	80.4	88
<i>ALBEF</i>	80.9	93.6	82.8	79
<i>SG-QMFNet</i>	85.7	96.2	87.5	67

Note: ResNet-50: Residual Network 50; EfficientNet-B4: Efficient Network B4; Swin-T: Swin Transformer Tiny; NTS-Net: Neural Teacher Student Network; CAL: Contrastive Attention Learning; TransFG: Transformer for Fine-Grained Recognition; ViLT: Vision-and-Language Transformer; ALBEF: ALign BEfore Fuse; SG-QMFNet: Saliency-Guided Quality-aware Multi-Modal Fusion Network.

3.3 Ablation experiments

To verify the independent contribution of each core module, five groups of ablation variants are designed. *w/o Saliency* removes the saliency decoupling module, and directly feeds the original RGB features into the cross-modal Transformer. *w/o Edge* removes the edge structure modality, retaining only the appearance and text modalities. *w/o Text* removes the text semantic modality, retaining only the appearance and edge modalities. *w/o Gating* replaces the dynamic gating with fixed average fusion, namely $F_{fused} = \frac{1}{3}(\bar{z}^v + \bar{z}^e + \bar{z}^t)$. *w/o Temporal* removes the quality-aware temporal aggregation module, and only uses single-frame fused features for classification. Figure 6 shows the Top-1 accuracy and mAP of each variant on the test set.

As can be seen from Figure 6, the removal of each module causes a performance drop, and the magnitude of the drop is consistent with the functional positioning of the module in the overall framework. The removal of the temporal aggregation module causes the largest performance loss, with a Top-1 drop of 2.7 percentage points and an mAP drop of 2.7 percentage points, indicating that quality screening and selective aggregation between consecutive frames play an irreplaceable role in recognition stability. The contribution of the saliency decoupling module is second, with a Top-1 drop of 2.5 percentage points after removal, verifying the necessity of foreground enhancement under complex background and occlusion conditions. The removal of dynamic gating causes a drop of 1.6 percentage points, indicating that the adaptive fusion strategy is superior to the fixed-weight scheme. The removal of edge modality and text modality causes drops of 1.8 and 1.3 percentage points respectively, showing that the two types of complementary information have independent supplementary effects on RGB appearance. Among them, the contribution of structural information is slightly higher than

that of semantic information, which is consistent with the fact that edge structure is more stable under degradation conditions in live-streaming frames. The five groups of ablation experiments together show that the relationship between the modules is not simply additive, but forms a synergistic effect through successive connections. Only the complete SG-QMFNet can fully release the potential of each module.

3.4 Robustness analysis under different degradation conditions

The degradation types in live-streaming scenarios are diverse. To evaluate the robustness of the method under various degradation conditions, the test set is divided into four subsets according to degradation types: motion blur, partial occlusion, low lighting, and complex background, with each subset containing at least 200 clips. Figure 7 shows the Top-1 accuracy of SG-QMFNet and two representative baseline methods, EfficientNet-B4 and ALBEF, on each subset.

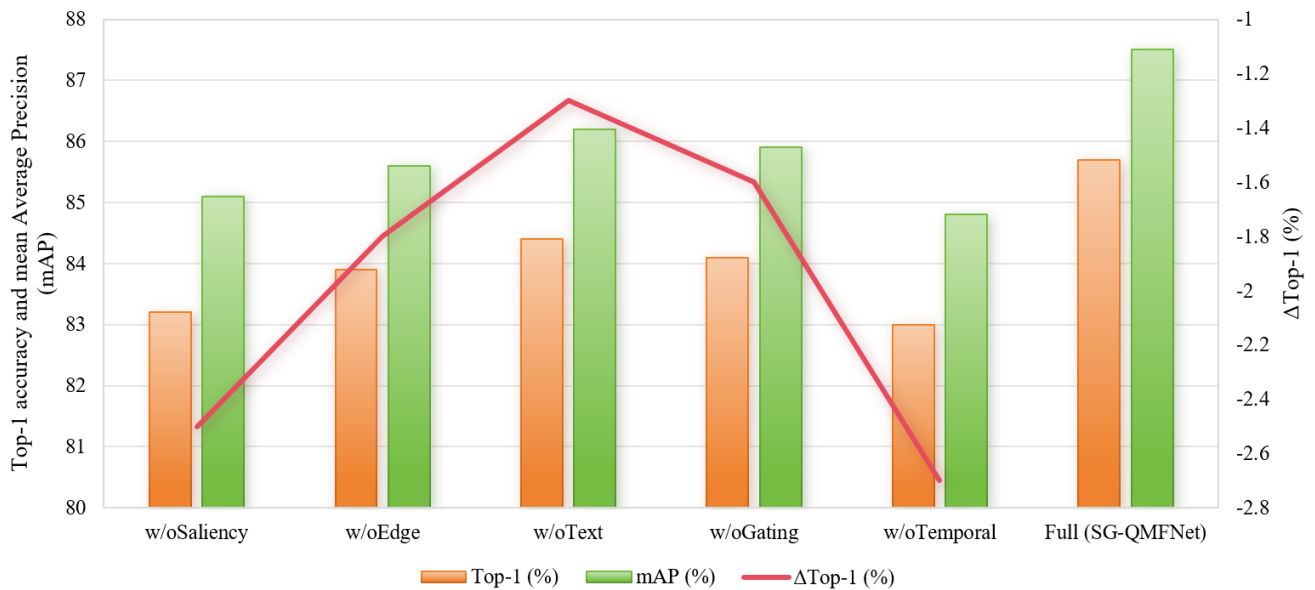


Figure 6. Ablation experiment results
Note: SG-QMFNet: Saliency-Guided Quality-aware Multi-Modal Fusion Network.

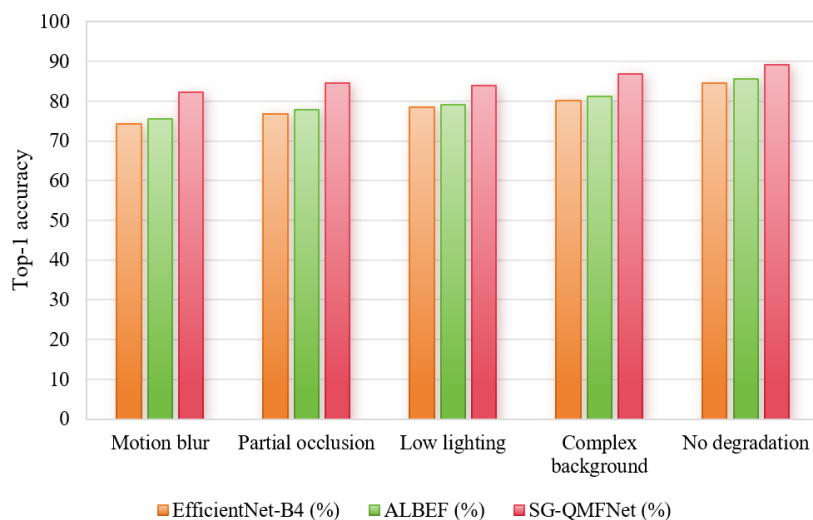


Figure 7. Comparison of Top-1 accuracy under different degradation conditions
Note: ALBEF: ALign BEfore Fuse; SG-QMFNet: Saliency-Guided Quality-aware Multi-Modal Fusion Network.

As can be seen from Figure 7, SG-QMFNet maintains the lead on all degradation subsets, and the lead margin significantly expands under degradation conditions. On the motion blur subset, SG-QMFNet improves by 8.1 percentage points over EfficientNet-B4 and 6.7 percentage points over ALBEF, which is the largest improvement margin among all subsets. This result is directly attributed to the fact that the edge structure modality can still maintain the structural stability of contours and packaging text under blur conditions, as well as the effective suppression of blurred frames by quality-aware temporal aggregation. On the partial occlusion subset, the Top-1 of SG-QMFNet reaches 84.5%, which is 7.7 and 6.6 percentage points higher than the two baselines, respectively. The saliency decoupling module effectively isolates the noise interference in occluded regions through foreground enhancement. On the low lighting and complex background subsets, SG-QMFNet achieves 84.0% and 86.8%, respectively, with improvement margins of 5.5 and 6.7 percentage points, indicating that the multi-modal complementary mechanism is also robust to illumination changes and background interference. On the subset without degradation, SG-QMFNet still maintains a lead of 4.6 percentage points, showing that each module does not introduce performance loss under clear conditions, and the method has good versatility. It is worth noting that ALBEF is only slightly better than EfficientNet-B4 under various degradation conditions, which again shows that simple image-text fusion cannot fundamentally solve the degradation robustness problem. In contrast, SG-QMFNet achieves substantial robustness improvement through the combined effect of structural modality, saliency decoupling, and temporal quality screening.

3.5 Temporal aggregation effectiveness analysis

The performance of the temporal aggregation module is directly affected by the number of input frames $T = 1, 3, 5, 7$. To determine the optimal frame number configuration, models are trained with respectively and evaluated on the test set. When $T = 1$, the model degenerates to single-frame recognition and does not use the temporal Transformer. Table 3 shows the Top-1 accuracy, inference time per clip, and clip throughput under different frame number settings.

Table 3. Performance comparison of different temporal frame numbers

Frame Number T	Top-1 (%)	Inference Time (ms/Clip)	Clip Throughput (Clips/s)
1	83.0	12.1	82.6
3	84.6	28.7	34.8
5	85.7	41.2	24.3
7	85.9	55.3	18.1

As can be seen from Table 3, the Top-1 accuracy shows a clear trend of diminishing marginal returns as the frame number increases. When T increases from 1 to 3, the accuracy increases by 1.6 percentage points; from 3 to 5, it further increases by 1.1 percentage points; while when T increases from 5 to 7, the improvement is only 0.2 percentage points. This trend indicates that 5 frames are sufficient to capture the temporal redundant information of the same product target within a short time window, and continuing to increase the frame number cannot bring effective discriminative gains. In terms of inference time, when $T = 5$, the time per clip is 41.2 ms, corresponding to about 24.3 clips/s, achieving a good balance between accuracy and computational efficiency. When $T = 7$, the time increases to 55.3 ms, and the clip throughput drops to 18.1 clips/s, while the accuracy only increases by 0.2 percentage points, so the benefit of continuing to increase the frame number is limited. Considering the trade-off between accuracy and efficiency, $T = 5$ is the optimal configuration, and this setting is adopted for all subsequent experiments.

3.6 Dynamic gating mechanism analysis

To deeply analyze the contribution of the uncertainty-driven dynamic gating mechanism, it is compared with four common fusion strategies. Average fusion assigns equal weights to the three modalities. Concatenation fusion concatenates the global representations of the three modalities and reduces the dimension through a fully connected layer. Fixed-weight fusion obtains the optimal weight combination $g_v = 0.5$, $g_e = 0.3$, $g_t = 0.2$ through grid search on the validation set. Attention fusion uses a learnable attention module to compute softmax weights based on the three modal representations. Figure 8 shows the Top-1 accuracy of each fusion strategy on the overall test set and the motion blur subset.

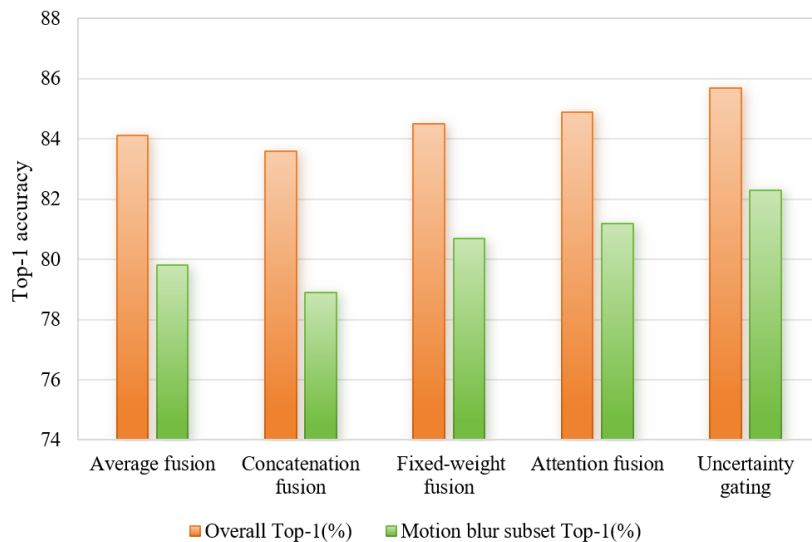


Figure 8. Performance comparison of different fusion strategies

As can be seen from Figure 8, uncertainty gating fusion achieves the best results on both the overall test set and the motion blur subset. On the overall test set, uncertainty gating improves by 1.6 percentage points over average fusion and 0.8 percentage points over attention fusion. On the motion blur subset, the gap between the strategies is further amplified: uncertainty gating improves by 2.5 percentage points over average fusion and 1.1 percentage points over attention fusion. This result shows that under degradation conditions, the reliability difference between modalities is more significant, and the importance of dynamically adjusting fusion weights increases accordingly. Concatenation fusion performs the worst, even lower than average fusion, indicating that directly concatenating high-dimensional features easily introduces redundancy and noise, which is not conducive to the discriminability of the fused representation. Fixed-weight fusion is better than average fusion, but cannot adapt to the frame-by-frame variation in modal reliability. Attention fusion can adaptively compute weights according to the input, but lacks explicit uncertainty modeling, making it prone to overfitting on noisy modalities. Uncertainty gating learns modal reliability through self-supervision of classification confidence, constraining the basis of weight allocation at the semantic level, thus showing stronger stability under degradation conditions.

3.7 Hyperparameter sensitivity analysis

This section examines the impact of two key hyperparameters on model performance: the weight coefficient of the cross-modal contrastive alignment loss λ_2 and the gating temperature parameter τ_g . Figure 9 shows the Top-1 accuracy of the model on the validation set under different values.

As can be seen from Figure 9, λ_2 achieves the optimal validation set accuracy of 85.6% at 0.3, and deviations from this value lead to performance degradation. When λ_2 is reduced to 0.1, the binding force of cross-modal alignment is insufficient, the representations of the three modalities fail to

be sufficiently pulled closer, the utilization efficiency of complementary information decreases, and the accuracy drops to 84.3%. When λ_2 is increased to 0.7, the dominant effect of the contrastive loss on training becomes too strong, which may interfere with the optimization of the main objective of the classification task, and the accuracy falls back to 84.7%. The temperature parameter τ_g also reaches the optimum at 0.5. Too small a τ_g makes the gating weight distribution sharp, approaching hard selection behavior, and the training process becomes unstable; too large a τ_g makes the weights tend to be uniform, degenerating into approximate average fusion and losing the adaptive adjustment capability. $\tau_g = 0.5$ achieves a good balance between discriminability and smoothness, verifying the rationality of the default setting.

3.8 Generalization verification on the Product-100 dataset

To verify the generalization ability of the method on static product images, experiments are conducted on the Product-100 dataset. Since this dataset does not contain temporal information, SG-QMFNet adopts single-frame inference mode and does not enable the quality-aware temporal aggregation module. Figure 10 shows the Top-1 and Top-5 accuracy of each method on this dataset.

As can be seen from Figure 10, SG-QMFNet achieves a Top-1 accuracy of 80.3% in single-frame inference mode, which is 1.8 percentage points higher than the best baseline TransFG and 4.2 percentage points higher than EfficientNet-B4. This result shows that the effectiveness of the multi-modal fusion and saliency decoupling modules does not depend on video temporal information, and can also bring stable performance gains on static product images. ALBEF still performs worse than the pure vision fine-grained method TransFG on this dataset, further confirming the limitations of simple image-text fusion strategies. SG-QMFNet achieves refined utilization of text semantic information through cross-modal alignment and uncertainty gating, enabling the text modality to also play a positive complementary role in static scenarios.

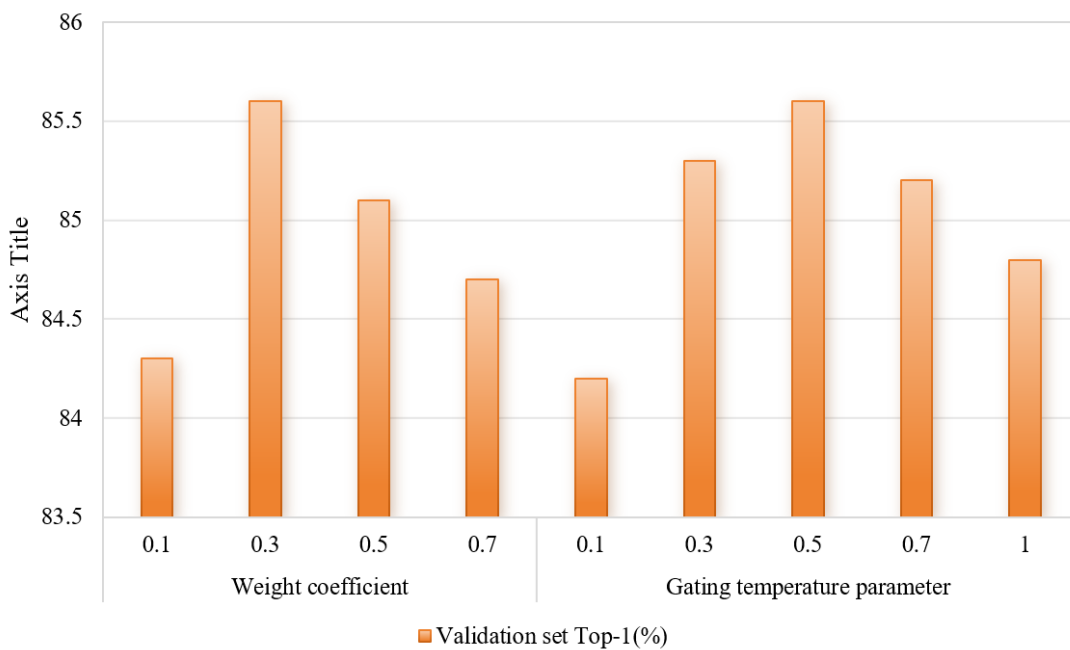


Figure 9. Hyperparameter sensitivity analysis

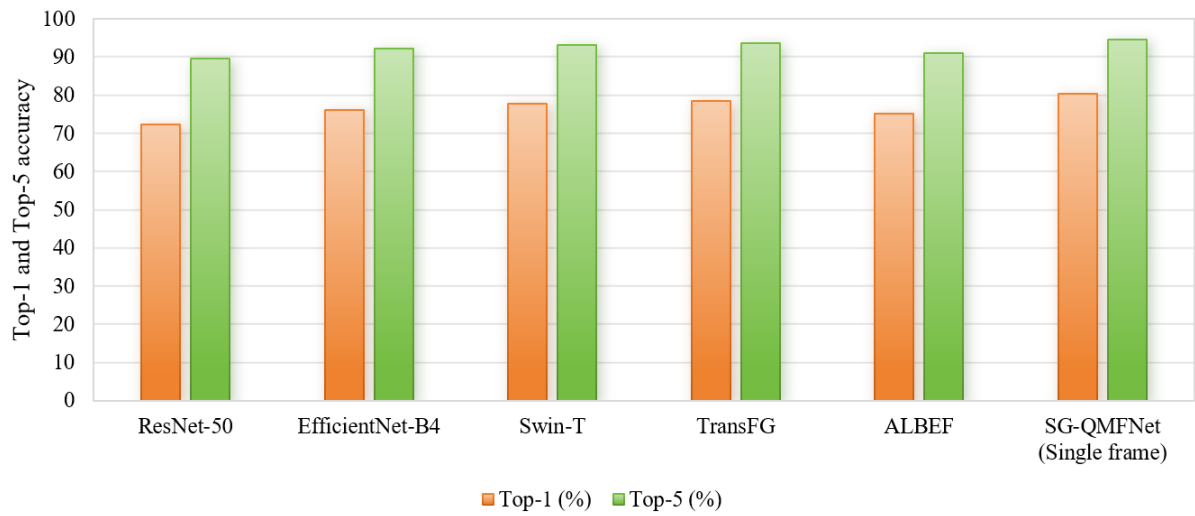


Figure 10. Comparison results on the Product-100 dataset

Note: ResNet-50: Residual Network 50; Swin-T: Swin Transformer Tiny; TransFG: Transformer for Fine-Grained Recognition; ALBEF: ALign Before Fuse; SG-QMFNet: Saliency-Guided Quality-aware Multi-Modal Fusion Network.

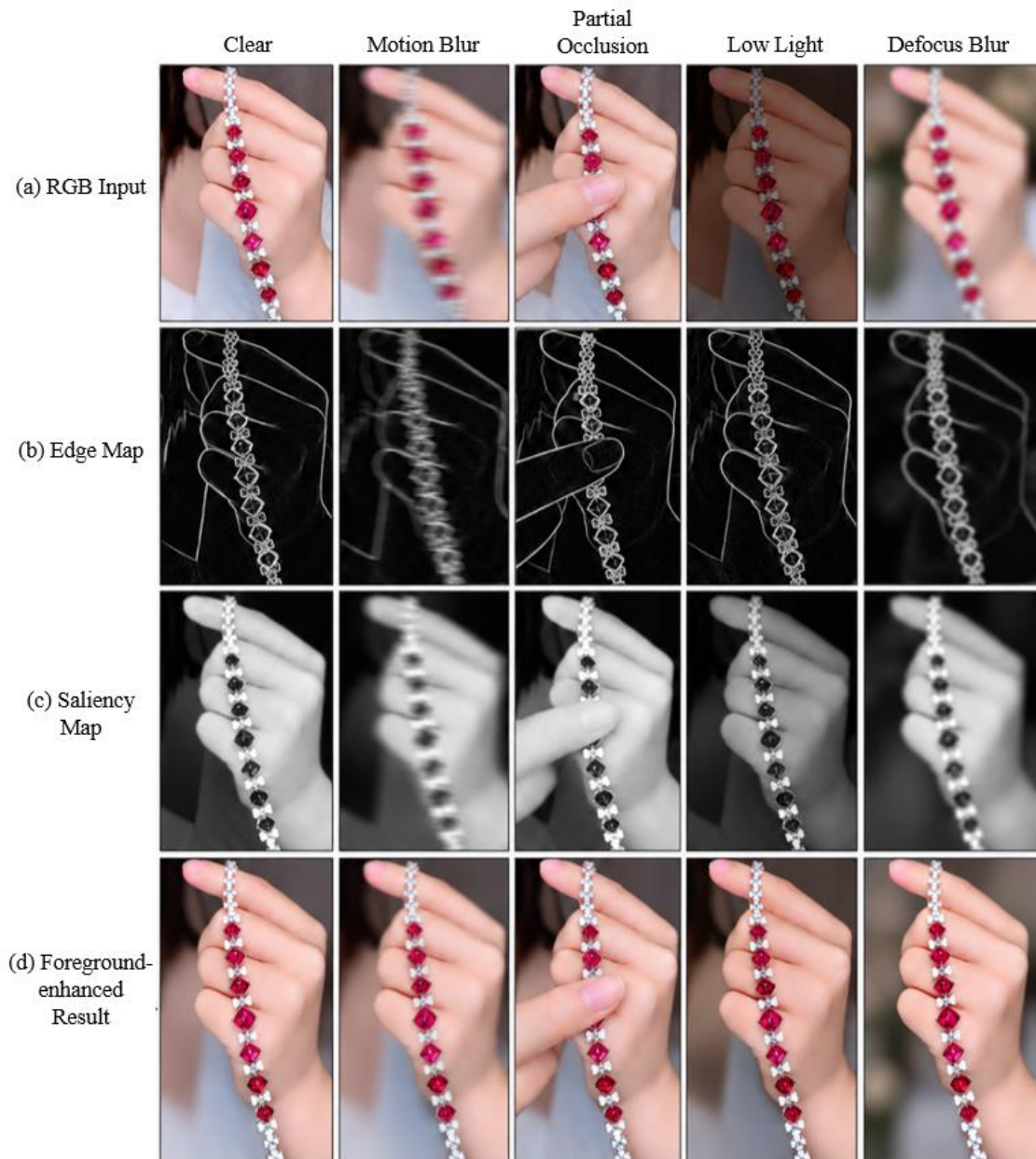


Figure 11. Comparison of multi-modal image processing results under different degradation conditions

3.9 Visualization results

To intuitively verify the adaptability of the proposed method to typical degradation scenarios in live-streaming product recognition and the multi-modal collaborative processing effect, this paper conducts a visual analysis of the multi-stage processing results of product images under five representative conditions: clear, motion blur, partial occlusion, low lighting, and defocus blur (Figure 11). Subfigure (a) shows that the original inputs exhibit significant apparent quality differences under different degradation conditions. Among them, motion blur and defocus blur cause obvious attenuation of texture details and local text information; partial occlusion directly destroys the complete visibility of the product structure; and low lighting weakens the contrast between the target region and the background. These phenomena jointly indicate that a single RGB appearance feature is difficult to stably support fine-grained product discrimination. Subfigure (b) further shows that the edge structure branch can stably preserve product contours, gem arrangement relationships, and key geometric boundaries under various degradation interferences. Especially in blurred scenes, it can still maintain a relatively clear structural response, indicating that the structure modality can provide an effective supplement to the recognition process when appearance information is damaged. The saliency maps given in Subfigure (c) show that weakly supervised saliency modeling can continuously focus on the product main body region. Even in the presence of occlusion or limited brightness, the model can still accurately distinguish the foreground from the background, thereby reducing the contamination of discriminative features by irrelevant regions. The foreground enhancement and background suppression results shown in Subfigure (d) further prove that after saliency guidance, the color, edge, and local texture of the product main body are significantly enhanced, while the background response is effectively weakened, ultimately obtaining a more discriminative target representation. Based on a comprehensive view of all subfigures, it can be seen that the proposed method does not rely on local improvements of a single modality. Instead, through the synergy of RGB appearance, edge structure, and saliency prior, it achieves continuous optimization from target localization, structure preservation to discriminative enhancement under degradation scenarios. This result provides strong qualitative support for the core conclusion of the paper that multi-modal feature fusion can improve the robustness and fine-grained recognition capability of live-streaming product image recognition.

4. CONCLUSION

This paper addresses the core challenges of product recognition in live-streaming scenarios, including motion blur, partial occlusion, complex backgrounds, and subtle fine-grained differences, and proposes a SG-QMFNet. With degradation robustness as the main design thread, the method constructs four closely connected modules: weakly supervised saliency decoupling, structure–appearance–semantics three-stream cross-modal alignment, uncertainty-driven dynamic gating fusion, and quality-aware temporal aggregation. Saliency decoupling achieves fine-grained extraction of the product foreground without pixel-level annotations,

effectively isolating background noise. The cross-modal alignment Transformer maps edge structure, RGB appearance, and OCR semantics into a unified representation space, and fully exploits the complementary cues between modalities through layer-by-layer attention interaction. Uncertainty gating uses classification confidence as a self-supervised signal to achieve frame-by-frame adaptive adjustment of fusion weights. Quality-aware temporal aggregation uses frame-level quality scores to guide attention allocation, selecting highly discriminative information from continuous frames and suppressing the interference of degraded frames. Experiments on the self-built dataset LiveProduct-10K show that the method achieves a Top-1 accuracy of 85.7% and an mAP of 87.5%, which are 3.3 percentage points and 2.8 percentage points higher than the best baseline, respectively, with improvements exceeding 7 percentage points on the motion blur and partial occlusion subsets. Ablation experiments confirm the independent contribution of each module, generalization experiments verify the applicability of the method on static product images, and visualization analysis further reveals the internal mechanism of the performance gain.

The work in this paper shows that robust recognition for degradation scenarios should not rely solely on the enhancement of a single appearance feature, but should be systematically designed from multiple levels such as multi-modal complementarity, foreground decoupling, and temporal quality screening. Future research can be carried out in two directions. First, further compress the model size while maintaining recognition accuracy, and explore lightweight architectures for mobile deployment to meet the actual needs of edge computing in live-streaming platforms. Second, incorporate the audio modality into the fusion framework. The host's oral broadcast information in live streaming often contains direct category hints, and audio-visual joint modeling is expected to further improve recognition robustness and coverage.

ACKNOWLEDGEMENT

This paper was funded by the 2022 Annual Guangxi University Young and Middle-aged Teachers' Basic Research Ability Improvement Project "Applied Research on the Dissemination and Promotion of Intangible Cultural Heritage from the Perspective of New Media" (Grant No.: 2022KY1094).

REFERENCES

- [1] Wang, Y., Lu, Z., Cao, P., Chu, J., Wang, H., Wattenhofer, R. (2022). How live streaming changes shopping decisions in E-commerce: A study of live streaming commerce. *Computer Supported Cooperative Work (CSCW)*, 31(4): 701-729. <https://doi.org/10.1007/s10606-022-09439-2>
- [2] Han, L., Yin, Z. (2022). Global memory and local continuity for video object detection. *IEEE Transactions on Multimedia*, 25: 3681-3693. <https://doi.org/10.1109/tmm.2022.3164253>
- [3] Li, Q., Peng, X., Cao, L., et al. (2020). Product image recognition with guidance learning and noisy supervision. *Computer Vision and Image Understanding*, 196: 102963.

- <https://doi.org/10.1016/j.cviu.2020.102963>
- [4] Santra, B., Shaw, A.K., Mukherjee, D.P. (2022). Part-based annotation-free fine-grained classification of images of retail products. *Pattern Recognition*, 121: 108257. <https://doi.org/10.1016/j.patcog.2021.108257>
 - [5] Flusser, J., Lébl, M., Šroubek, F., Pedone, M., Kostková, J. (2023). Blur invariants for image recognition. *International Journal of Computer Vision*, 131(9): 2298-2315. <https://doi.org/10.1007/s11263-023-01798-7>
 - [6] Yang, W., Wang, W., Huang, H., Wang, S., Liu, J. (2021). Sparse gradient regularized deep Retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30: 2072-2086. <https://doi.org/10.1109/tip.2021.3050850>
 - [7] Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., Yang, R. (2021). Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3239-3259. <https://doi.org/10.1109/TPAMI.2021.3051099>
 - [8] Pettersson, T., Riveiro, M., Löfström, T. (2024). Multimodal fine-grained grocery product recognition using image and OCR text. *Machine Vision and Applications*, 35(4): 1-20. <https://doi.org/10.1007/s00138-024-01549-9>
 - [9] Huang, W., Wang, D., Ouyang, X., Wan, J., Liu, J., Li, T. (2024). Multimodal federated learning: Concept, methods, applications and future directions. *Information Fusion*, 112: 102576. <https://doi.org/10.1016/j.inffus.2024.102576>
 - [10] Wu, Z., Li, H., Xiong, C., Jiang, Y., Davis, L.S. (2020). A dynamic frame selection framework for fast video recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4): 1699-1711. <https://doi.org/10.1109/tpami.2020.3029425>
 - [11] Melek, C.G., Sönmez, E.B., Varlı, S. (2024). Datasets and methods of product recognition on grocery shelf images using computer vision and machine learning approaches: An exhaustive literature review. *Engineering Applications of Artificial Intelligence*, 133: 108452. <https://doi.org/10.1016/j.engappai.2024.108452>
 - [12] Li, C., Guo, C., Han, L., et al. (2021). Low-light image and video enhancement using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9396-9416. <https://doi.org/10.1109/tpami.2021.3126387>
 - [13] Oucheikh, R., Pettersson, T., Löfström, T. (2022). Product verification using OCR classification and Mondrian conformal prediction. *Expert Systems with Applications*, 188: 115942. <https://doi.org/10.1016/j.eswa.2021.115942>
 - [14] Futagami, T., Hayasaka, N. (2020). Automatic product region extraction based on analysis of images uploaded to C2C online market. *Journal of Organizational Computing and Electronic Commerce*, 30(4): 323-334. <https://doi.org/10.1080/10919392.2020.1788359>
 - [15] Han, J., Yao, X., Cheng, G., Feng, X., Xu, D. (2019). P-CNN: Part-based convolutional neural networks for fine-grained visual categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2): 579-590. <https://doi.org/10.1109/tpami.2019.2933510>
 - [16] Jiao, S., Goel, V., Navasardyan, S., et al. (2023). Collaborative content-dependent modeling: A return to the roots of salient object detection. *IEEE Transactions on Image Processing*, 32: 4237-4246. <https://doi.org/10.1109/tip.2023.3293759>
 - [17] Yang, X., Feng, S., Wang, D., Zhang, Y. (2020). Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 23: 4014-4026. <https://doi.org/10.1109/tmm.2020.3035277>
 - [18] Wang, Y., He, J., Wang, D., Wang, Q., Wan, B., Luo, X. (2023). Multimodal transformer with adaptive modality weighting for multimodal sentiment analysis. *Neurocomputing*, 572: 127181. <https://doi.org/10.1016/j.neucom.2023.127181>
 - [19] He, F., Li, Q., Zhao, X., Huang, K. (2022). Temporal-adaptive sparse feature aggregation for video object detection. *Pattern Recognition*, 127: 108587. <https://doi.org/10.1016/j.patcog.2022.108587>
 - [20] Wang, X., Huang, Z., Liao, B., Huang, L., Gong, Y., Huang, C. (2021). Real-time and accurate object detection in compressed video by long short-term feature aggregation. *Computer Vision and Image Understanding*, 206: 103188. <https://doi.org/10.1016/j.cviu.2021.103188>