Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Detection of AI-Generated Product Main Images in Cross-Border E-Commerce Based on Multi-Feature Fusion

lili huang
meizhouwan Vocational and Technical College
zhidan hui
1098587487@qq.com

Huazhong University of Science and Technology

Research Article

Keywords: cross-border e-commerce, AIGC detection, multi-feature fusion, frequency-domain analysis, cross-attention mechanism, discrete cosine transform

Posted Date: June 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9899618/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Detection of AI-Generated Product Main Images in Cross-Border E-Commerce Based on Multi-Feature Fusion

Huang Lili, School of Business Administration, Meizhouwan Vocational & Technical College, Putian 351119, China iamhuanglili@qq.com
Hui Zhidan*, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China 1098587487@qq.com

Abstract

With the rapid development of generative artificial intelligence-generated content (AIGC), diffusion models such as Midjourney, Stable Diffusion, and DALL-E 3 have increasingly been used to produce product main images for cross-border e-commerce. These tools can reduce the cost of visual content production for sellers; however, AI-generated product images may still contain artifacts in micro-texture continuity, object boundaries, illumination consistency, refraction, and high-order statistical distributions. Such artifacts may increase the risk of misleading visual presentation and create new challenges for platform image auditing.

Existing general-purpose AIGC detection methods, digital watermarking techniques, and Deepfake identification models may not transfer well to product main images in cross-border e-commerce. A major reason is that such images often contain dense promotional text, multilingual layout watermarks, multi-scale price labels, and decorative backgrounds. These elements introduce strong edge and texture interference, making it difficult for conventional global feature extractors to distinguish artificial layout features from generation artifacts.

To address this domain-specific problem, this study constructs CBEC-AIGC-Bench, a benchmark dataset covering four cross-border e-commerce categories: digital electronics, apparel and footwear, marine fresh food, and outdoor fishing equipment. A multi-feature fusion network, MFF-EComDet, is then proposed for detecting AIGC-generated product main images. Instead of relying on a single global representation, the proposed network adopts a decoupled dual-branch architecture guided by cross-domain interaction.

First, the Spatial Semantic Branch (SSB) uses a lightweight ConvNeXt-based backbone to capture geometric deformation, illumination inconsistency, and unnatural boundary patterns through large-receptive-field feature extraction. Second, the Frequency Artifact Inversion Branch (FAIB) projects image signals into the frequency domain using two-dimensional

discrete cosine transform (2D-DCT), enabling the model to capture periodic and grid-like spectral anomalies associated with image generation and post-processing.

To reduce the interference caused by dense promotional text in e-commerce posters, this paper designs a Cross-Attention Feature Fusion (CAFF) module. In this module, spatial-domain representations are used as Queries, while frequency-domain representations are used as Keys and Values. The module models the interdependence between the two domains at the feature-matrix level. Because promotional text has strong gradients in the spatial domain but does not necessarily share the same periodic frequency patterns as AI-generated product regions, CAFF can reduce the influence of text regions and guide the model toward more discriminative product-related features.

Experimental results on CBEC-AIGC-Bench show that MFF-EComDet achieves an overall accuracy of 94.8%, exceeding the compared baseline methods on the constructed test set. Under common post-processing conditions, including WebP conversion, JPEG compression with a quality factor of 50, and multi-scale resizing, the proposed model maintains comparatively stable performance, suggesting its potential for further evaluation in platform image-auditing scenarios.

Keywords: cross-border e-commerce; AIGC detection; multi-feature fusion; frequency-domain analysis; cross-attention mechanism; discrete cosine transform

1. Introduction

Global digital trade and cross-border e-commerce have expanded rapidly in recent years. As one of the first visual elements on a product detail page, the product main image can influence click-through rate (CTR), product perception, and initial purchasing decisions (Ghose et al., 2013). A clear and reliable product main image also helps reduce communication costs in multilingual shopping environments.

In recent years, visual AIGC tools based on diffusion models and related generative techniques have become increasingly accessible. Representative systems include Midjourney, Stable Diffusion (Rombach et al., 2022), and DALL-E 3 (Betker et al., 2023). Cross-border sellers can use text prompts, reference images, or product design materials to generate product-oriented visual content at relatively low cost. While this technology lowers the threshold for visual content production, it also increases the need for reliable authenticity auditing.

The misuse of AIGC in e-commerce product presentation may create several practical risks, mainly in the following three aspects.

First, AI-generated images may increase the risk of misleading visual presentation. For products such as marine fresh food, organic foods, and jewelry, generated images may over-enhance color, texture, volume, or surface gloss. This can enlarge the gap between the displayed product and the delivered product. In some jurisdictions, such practices may also raise compliance concerns. Existing studies have shown that generated images may contain artifacts related to micro-texture continuity, boundary-level illumination, refraction, and high-order statistical features (Corvi et al., 2023; Frank et al., 2020).

Second, cross-border layout text introduces spatial-domain noise. E-commerce product images are often composite visual media containing promotional labels such as "50% OFF," "Free Shipping," currency symbols, countdown widgets, and geometric background frames. These text and layout elements create sharp edge gradients. Traditional AIGC detection networks, such as global CNN-based residual extractors, are often trained on cleaner natural-image datasets (Chai et al., 2020; Wang et al., 2020). When applied to e-commerce images, they may confuse artificial layout edges with synthesis artifacts, increasing the false-positive rate.

Third, signal degradation may occur during platform transmission and storage. Cross-border e-commerce platforms such as AliExpress, Amazon, and Shopee often recompress uploaded product images to reduce loading latency. Such operations may include reducing the JPEG quality factor or converting images to WebP or AVIF formats. Lossy compression acts as a low-pass filtering process and may remove weak pixel-level generation artifacts (Durall et al., 2020).

To address the above challenges, this study explores the heterogeneous manifestations of AI-generated images in two complementary feature spaces: spatial geometric structure and frequency-domain statistics (Wang et al., 2023; Zhang et al., 2019). Although post-processing and dense text may reduce the reliability of spatial cues, frequency-domain representations can still capture periodic or grid-like anomalies associated with generation and resampling processes (Dzanic et al., 2020; Frank et al., 2020).

Based on this observation, this paper proposes MFF-EComDet, a multi-feature fusion network for detecting AIGC-generated product main images in cross-border e-commerce. The model adopts a dual-branch framework. One branch captures macro-level illumination and

structural features in the spatial domain, while the other extracts micro-texture anomalies in the frequency domain. The two branches are dynamically aligned and complemented through a Cross-Attention Feature Fusion module to reduce text-noise interference and improve product image authenticity detection.

2. Related Work

2.1 Detection of AIGC-Generated Images

Current academic approaches for detecting AIGC-generated images can generally be divided into two categories: spatial representation-driven methods and frequency statistics-driven methods.

Spatial representation-driven methods mainly employ classical deep residual networks, such as ResNet (He et al., 2016), DenseNet (Huang et al., 2017), or the latest Vision Transformer (ViT) architectures (Dosovitskiy et al., 2021) as base classifiers. The underlying assumption of these methods is that generative models still have macroscopic defects in image boundary continuity and complex contextual logic, such as unreasonable shadows and asymmetric accessories (Marra et al., 2019; Nataraj et al., 2019). However, with the rapid iteration of commercial models such as Midjourney v6, the spatial geometric realism of generated images has reached a level that is difficult for human eyes to distinguish. Consequently, the performance of methods relying solely on spatial cues has declined significantly (Ojha et al., 2023).

Researchers have therefore turned to frequency-domain analysis. Wang et al. (2020) pointed out in the classical CNNDetection study that almost all mainstream generative networks leave regularly arranged bright spots or grid-like artifacts in Fourier spectrograms due to defects in up-sampling operations. Subsequent studies further used discrete cosine transform (DCT) or wavelet transform to extract generation traces (Durall et al., 2020; Zhang et al., 2019), achieving promising results in Deepfake detection tasks (Frank et al., 2020; Qian et al., 2020). However, these studies were mainly conducted on standard clean image datasets and ignored the special vertical domain of e-commerce images, which contain dense noisy texts. When faced with dense promotional texts, multilingual watermarks, and multi-scale price labels in e-commerce product images, existing frequency-domain methods also suffer from serious performance degradation (Cozzolino et al., 2021).

2.2 Visual Feature Fusion and Attention Mechanisms

Feature fusion is a key technique for collaborative processing of multi-source information. In traditional image binary classification or detection tasks, researchers usually adopt early fusion, which directly concatenates images and frequency-domain maps along the channel dimension before feeding them into the network, or late fusion, which trains two classifiers separately and then performs weighted averaging on probability scores (Atrey et al., 2010). Both methods belong to static linear fusion strategies. They cannot dynamically adjust feature weights across different regions and are therefore insufficient for handling the differentiated noise interference between text regions and product subject regions in e-commerce images.

In recent years, Transformer techniques represented by self-attention and cross-attention have brought breakthroughs to feature fusion (Vaswani et al., 2017). Cross-attention allows signals from one modality to serve as a guiding source to query features from another modality, thereby enabling nonlinear dynamic selection (Chen et al., 2020; Lu et al., 2019). For example, in image forgery detection, Miao et al. (2021) used cross-attention to fuse RGB and noise streams, significantly improving detection robustness. Inspired by this mathematical property, this paper uses spatial-domain features to identify the approximate regions of text layout and applies attention penalties to these regions in the frequency-domain feature matrix, thereby adaptively counteracting background noise in e-commerce images.

3. Methodology

The proposed MFF-EComDet adopts a dual-stream symmetric network architecture. The entire pipeline consists of three core stages: spatial semantic feature extraction, frequency-domain micro-artifact inversion, and cross-attention-based heterogeneous feature fusion.

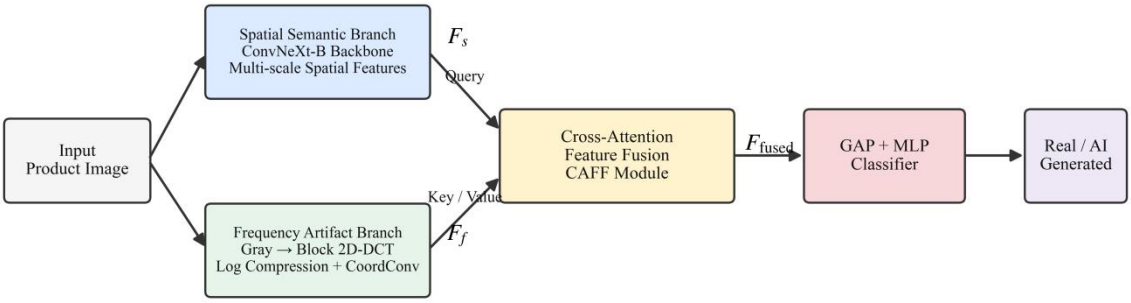


Figure 1. Overall architecture of the proposed multi-feature fusion detection framework.

3.1 Spatial Semantic Branch

The Spatial Semantic Branch (SSB) is designed to capture macroscopic inconsistencies in e-commerce product images that may be introduced by generative models. These inconsistencies include mismatched reflection directions, blurred object boundaries, and discontinuous texture patterns.

Considering the high throughput and real-time requirements of back-end image auditing systems on e-commerce platforms, this branch avoids overly bulky large-scale Transformer structures. Instead, it introduces the modernized lightweight convolutional network ConvNeXt-B as the spatial backbone. Given an input cross-border e-commerce product main image, the image first passes through a scale normalization layer and is then fed into stacked ConvNeXt blocks with large receptive fields based on depthwise convolution. By aggregating feature maps from the last three stages, we obtain spatial feature maps rich in macroscopic contextual information:

$$F_s \in R^{H \times W \times C} \tag{1}$$

where F_s denotes the spatial feature tensor, H_s and W_s denote the down-sampled height and width of the feature map, and C_s denotes the channel depth. In this study, the input image is resized to 224 x 224 pixels before being fed into the backbone, and the last three ConvNeXt stages are aggregated after channel alignment.

3.2 Frequency Artifact Inversion Branch

Although post-processing operations such as blurring and denoising can weaken spatial pixel-level traces, frequency-domain statistics may still retain useful information for detecting generated images. To capture such traces, the Frequency Artifact Inversion Branch (FAIB) uses discrete cosine transform rather than the Fourier transform. DCT has strong energy compaction capability and can help separate local edges and micro-textures from smoother background regions.

The input image is first converted into a grayscale image. To preserve local frequency-domain spatial locality, the image is then divided into a series of non-overlapping 8 x 8 image blocks. For any image block X_b in R^{N x N}, where N = 8 in the implementation, the mathematical definition of its two-dimensional discrete cosine transform is expressed as follows:

$$D(u,v) = \alpha(u)\alpha(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} B(x,y) cos\ [\frac{(2x+1)u\pi}{2N}] cos\ [\frac{(2y+1)v\pi}{2N}] \quad (2)$$

where u and v are frequency indices, x and y are spatial indices within the block, and the coefficient normalization function is defined as:

$$\alpha(k) = \begin{cases} \sqrt{\frac{1}{N}}, & k = 0 \\ \sqrt{\frac{2}{N}}, & k > 0 \end{cases} \quad (3)$$

After obtaining the DCT coefficient matrix D, a logarithmic transformation is applied for dynamic range compression and numerical stabilization:

$$D_{log} = log\ (1+|\ D\ |) \quad (4)$$

To enable interaction between frequency-domain coefficients and spatial-domain features, the frequency matrix must be mapped back into the same topological space. Therefore, a set of lightweight Coordinate Convolution layers and strided convolution layers are adopted for channel transformation and spatial alignment of frequency features. Finally, the frequency-domain feature map F_f is generated as follows:

$$F_f \in R^{H \times W \times C} \quad (5)$$

At this stage, the spatial-domain feature F_s and the frequency-domain feature F_f are aligned to identical tensor dimensions, enabling subsequent attention matrix multiplication.

3.3 Cross-Attention Feature Fusion Module

If the two feature maps are directly fused using element-wise summation or channel concatenation, noise overfitting may occur. In e-commerce images, promotional text introduces strong high-frequency components in the spatial domain and may generate corresponding harmonics in the frequency domain. Direct linear fusion may amplify these noisy components and interfere with classification.

To overcome this problem, this paper proposes the Cross-Attention Feature Fusion (CAFF) module. In this module, spatial features rich in macroscopic semantics serve as the "inspector" or Query, while frequency-domain features serve as the "abnormal components" or Keys and Values.

First, through three independent linear projection matrices W_Q, W_K, and W_V, the spatial-domain and frequency-domain feature maps are flattened and projected into the Query matrix Q, Key matrix K, and Value matrix V:

$$Q = F_s W_Q \tag{6}$$

$$K = F_f W_K \tag{7}$$

$$V = F_f W_V \tag{8}$$

The cross-attention correlation weight matrix A is then calculated as follows, where d_k denotes the dimension of the key vector:

$$Attention\ (Q, K, V) = Softmax\ (\frac{QK^T}{\sqrt{d_k}})V \tag{9}$$

To maintain the stability of the deep network and prevent gradient vanishing, residual connection and layer normalization are introduced to obtain the final fused feature tensor F_fused:

$$F_{fused} = LN\ (F_s + Attention\ (Q, K, V)) \tag{10}$$

When the network processes a region containing sharp text such as "50% OFF," the region may produce high activation in the spatial-domain feature map. However, the corresponding frequency-domain response does not necessarily contain the same periodic artifact patterns as AI-generated product regions. This mismatch can reduce the attention weight after Softmax. In contrast, when spatial and frequency features are jointly activated in product regions, the CAFF

module can assign higher weights to more discriminative areas and reduce text-region interference.

Finally, F_fused is reduced to a one-dimensional vector through global average pooling (GAP) and fed into a two-layer multilayer perceptron (MLP). The network is trained end-to-end using binary cross-entropy loss:

$$L_{BCE} = -[ylog\ (\hat{}) + (1-y)log\ (1-\hat{})] \tag{11}$$

where y denotes the ground-truth label, with 0 representing a real image and 1 representing an AI-generated image, and p denotes the predicted probability of AI generation.

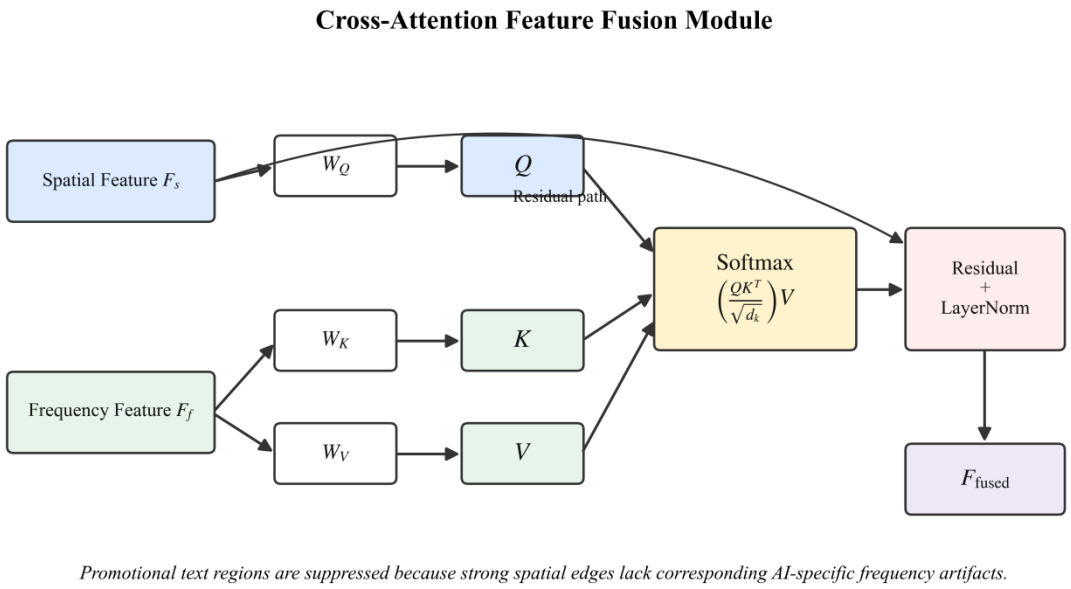


Figure 2. Detailed structure of the Cross-Attention Feature Fusion (CAFF) module.

3.4 Implementation and Reproducibility Details

All input images are resized to 224 x 224 pixels and normalized using ImageNet mean and standard deviation. The ConvNeXt-B backbone is initialized with ImageNet-pretrained weights, while the frequency branch and CAFF module are trained from scratch. The frequency branch uses 8 x 8 DCT blocks, logarithmic amplitude compression, and coordinate convolution for spatial alignment.

In the CAFF module, the projected spatial and frequency features are reshaped into token sequences before attention calculation. Multi-head cross-attention is used with four attention heads, and the output channel dimension is aligned with the final spatial feature dimension. Dropout with a probability of 0.1 is applied before the final MLP classifier to reduce overfitting.

For fair comparison, all baseline models use the same train/validation/test split, image resolution, data augmentation strategy, and optimizer setting unless the original method

requires a specific preprocessing pipeline. Hyperparameters are selected on the validation set and are not tuned on the test set.

Table 1. Key Implementation and Reproducibility Settings of MFF-EComDet

Item	Setting
Input size	224 x 224 pixels
Image normalization	ImageNet mean and standard deviation
Spatial backbone	ConvNeXt-B initialized with ImageNet-pretrained weights
Frequency transform	Block-wise 2D-DCT
DCT block size	8 x 8
Feature fusion	Multi-head cross-attention
Attention heads	4
Dropout	0.1 before the MLP classifier
Optimizer	AdamW
Learning rate	1 x 10^{-4}
Weight decay	1 x 10^{-2}
Batch size	16
Training epochs	30
Early stopping	Validation loss not decreasing for 5 epochs
Evaluation metrics	ACC, Precision, Recall, F1, and AUC

Table 2. Overall Architecture of MFF-EComDet.

Component	Description
Input	A cross-border e-commerce product main image to be detected
Spatial Semantic Branch (SSB)	The input image is processed by the ConvNeXt-B backbone to extract multi-scale spatial features
Frequency Artifact Inversion Branch (FAIB)	The input image is converted into grayscale and processed through block-wise 2D-DCT, logarithmic transformation, and coordinate convolution to extract frequency artifacts
Cross-Attention Feature Fusion Module (CAFF)	Spatial features are used as Query, while frequency features are used as Key and Value for dynamic fusion through multi-head cross-attention
Classification Head	Fused features are processed by global average pooling and MLP to output binary classification results, namely real image or AI-generated image

Component	Description
Information Flow	Arrows are used to clearly indicate the data flow from input to output

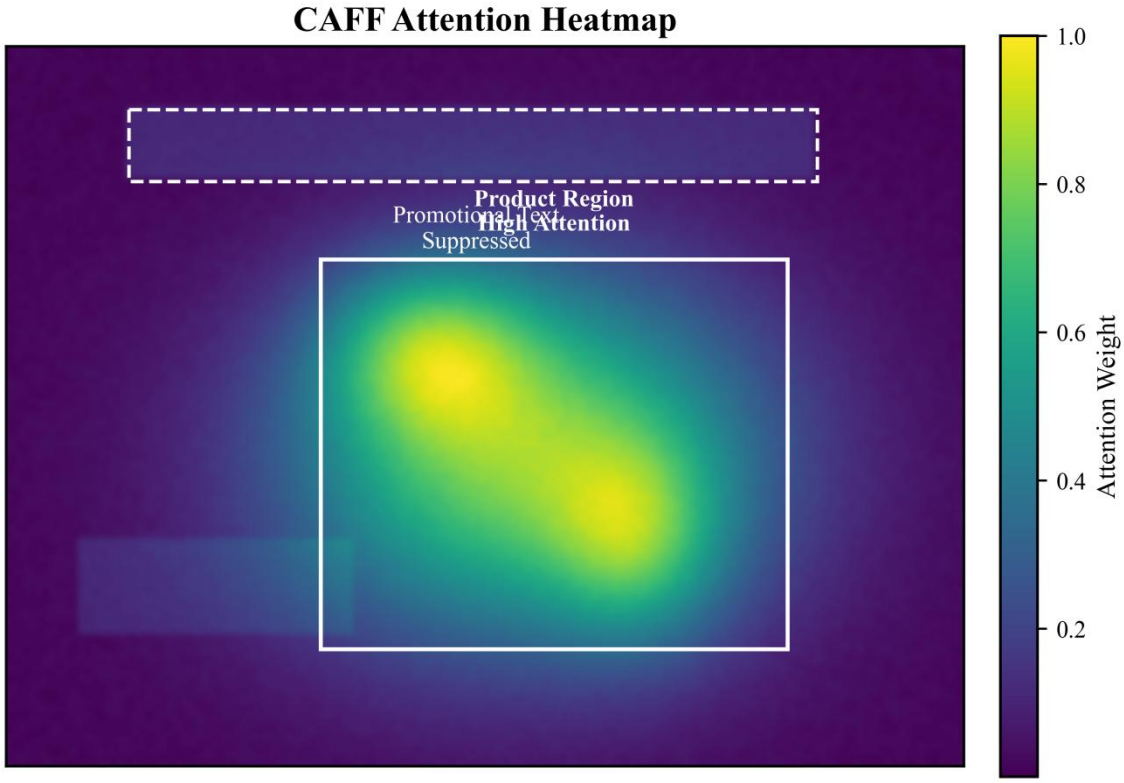


Figure 3. Attention heatmap visualization of the CAFF module.

4. Experiments and Analysis

4.1 Experimental Dataset and Benchmark Settings

Because existing public datasets for AIGC detection rarely focus on e-commerce product main images with dense promotional text, multilingual watermarks, and platform compression traces, this study constructs a domain-specific dataset named CBEC-AIGC-Bench.

First, web crawlers were used to collect 20,000 candidate real product main images from three cross-border platforms: AliExpress, Amazon, and Shopee. The images covered four product categories: digital electronics, apparel and footwear, marine fresh food, and outdoor fishing equipment. The collected images retained their original promotional text, layout backgrounds, and platform watermarks. To balance experimental efficiency and category coverage, a stratified random sampling strategy was adopted, and 2,000 representative real product main images were selected. The sampling process considered category, text density, platform source, and image resolution.

Subsequently, the titles and product descriptions corresponding to these 2,000 real images were reconstructed into standardized image-generation prompts using a large language model. According to a 1:1 ratio, 2,000 AI-generated product images were produced using Doubao V3, GPT-4o Image, SDXL 1.0, and GenMine. During generation, no control tools such as ControlNet or IP-Adapter were used, in order to approximate common low-cost generation practices.

To reduce duplicate leakage, all collected images were first normalized to a unified RGB format and screened using perceptual hashing. Image pairs with a Hamming distance lower than 5 were treated as near-duplicates, and only one image was retained. Manual inspection was then conducted to remove product detail pages, collage advertisements, images with severe occlusion, and images whose product category could not be clearly identified.

The labels of real images were assigned according to their platform source and manual screening results. For AI-generated images, the label was assigned according to the generation source. To avoid prompt leakage and product-level overlap, images derived from the same product title or description were placed into the same split. The training, validation, and test sets were therefore divided at the product level rather than at the image level.

Text density was calculated as the ratio between the area of OCR-detected text boxes and the total image area. Multilingual OCR was used only for estimating text density and was not used as an input feature of the proposed model. This design prevents the detector from relying directly on language content and keeps the task focused on image authenticity discrimination.



Figure 4. Example samples of real and AI-generated product main images in CBEC-AIGC-Bench.

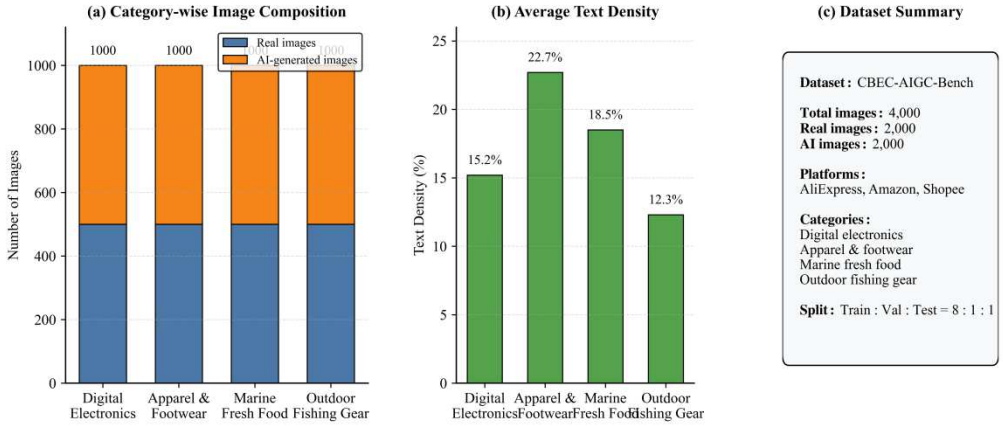


Figure 5. Composition of the CBEC-AIGC-Bench dataset.

The final dataset contains 4,000 labeled images and is randomly divided into training, validation, and testing sets at a ratio of 8:1:1. The detailed statistics of the dataset are shown in Table 3.

Table 3. Statistics of the CBEC-AIGC-Bench Dataset

Category	Number of Real Images	Number of AI-Generated Images	Average Text Density (%)	Average Resolution
Digital electronics	500	500	15.2	800 × 800
Apparel and footwear	500	500	22.7	1000 × 1000
Marine fresh food	500	500	18.5	600 × 600
Outdoor fishing equipment	500	500	12.3	750 × 750
Total	2000	2000	17.2	—

The experimental environment was based on an NVIDIA RTX 4090 GPU with 24 GB memory and the PyTorch 2.0 framework. The AdamW optimizer was used, with the base learning rate set to 1 x 10^{-4}, weight decay set to 1 x 10^{-2}, the batch size set to 16, and

the total number of training epochs set to 30. A cosine learning-rate decay strategy was adopted after a five-epoch warm-up period. An early stopping strategy was adopted when the validation loss did not decrease for five consecutive epochs. During training, random horizontal flipping, random brightness and contrast adjustment within ±10%, and random cropping with a scale ratio of 0.8-1.0 were used as data augmentation strategies.

All models were trained and evaluated under the same data split and preprocessing pipeline. Accuracy (ACC), precision (P), recall (R), F1 score, and area under the ROC curve (AUC) were used as evaluation metrics. Unless otherwise stated, the reported results are the average of three independent runs with different random seeds.

To further reduce the influence of dataset bias, three additional validation protocols were considered: platform-held-out evaluation, generator-held-out evaluation, and category-held-out evaluation. In platform-held-out evaluation, one e-commerce platform was used only for testing. In generator-held-out evaluation, images from one generation model were excluded during training and used only for testing. In category-held-out evaluation, one product category was held out during training and used for testing. These protocols were designed to evaluate whether the detector learns transferable generation cues rather than platform-, prompt-, or category-specific shortcuts.

4.2 Quantitative Comparison Experiments

To comprehensively verify the superiority of MFF-EComDet, we compared it with four representative baseline models in the fields of image forgery detection and AIGC discrimination on the same test set.

The first baseline is ResNet50-Baseline, a classical general-purpose fully convolutional binary classification network. The second is CNNDetection, a classical academic model that specifically learns CNN blur artifacts through targeted data augmentation (Wang et al., 2020). The third is Spec-Detection, a pure frequency-domain Fourier spectrum-based linear classifier. The fourth is DIRE, a representative detection algorithm based on reconstruction errors of diffusion models (Wang et al., 2023).The quantitative results are shown in Table 4.

Table 4. Performance Comparison of Different Detection Models on CBEC-AIGC-Bench

Model Category	Detection Method	Accuracy ACC (%)	Precision P (%)	Recall R (%)	F1 Score (%)
Pure spatial-domain model	ResNet50-Baseline	81.4	80.2	82.5	81.3
Pure spatial-domain model	DIRE	89.7	88.5	90.3	89.4

Model Category	Detection Method	Accuracy ACC (%)	Precision P (%)	Recall R (%)	F1 Score (%)
Pure frequency-domain model	Spec-Detection	83.2	82.1	84.0	83.0
Multi-feature model	CNNDetection	85.6	84.9	86.1	85.5
Proposed model	MFF-EComDet	94.8	94.2	95.5	94.8

As shown in Table 4, ResNet50-Baseline obtains an accuracy of 81.4%. During testing, it shows sensitivity to dense text layouts and may misclassify real product images containing complex multilingual text as AI-generated images. DIRE performs better than the basic ResNet50 baseline, but its reconstruction residuals can still be affected by dense promotional tags in e-commerce product images, resulting in an accuracy of 89.7% on this dataset. In comparison, the proposed method achieves an accuracy of 94.8% by dynamically fusing dual-domain features through the CAFF module.

4.3 Ablation Study

This section evaluates the contribution of the spatial semantic branch, the frequency artifact inversion branch, and the cross-attention mechanism by gradually removing or replacing the core components of MFF-EComDet. The ablation results are shown in Table 5.

Table 5. Ablation Results of Core Components in MFF-EComDet

Experiment ID	SSB	FAIB	Fusion Module Type	Final Accuracy ACC (%)	Performance Change
1	✓	×	—	83.1	Baseline
2	×	✓	—	79.5	-3.6
3	✓	✓	Linear summation	86.4	+3.3
4	✓	✓	Channel concatenation	88.2	+5.1
5	✓	✓	Cross-attention	94.8	+11.7

A comparison between Experiments 1 and 2 shows that when only the frequency-domain branch is used, the accuracy decreases to 79.5%. Visualization of misclassified samples indicates that the pure frequency-domain network may incorrectly classify real product images containing dense barcodes and letter layouts as fake images.

Experiments 3 and 4 show that conventional summation or concatenation improves performance compared with single-branch models, but the improvement is limited, with the highest accuracy reaching 88.2%. After introducing the CAFF mechanism, the accuracy increases to 94.8%, representing an improvement of 11.7 percentage points over the baseline. These results support the assumption that nonlinear cross-domain interaction can help separate text noise from product-related generation artifacts.

4.4 Robustness Test Against Transmission Compression

In practical cross-border e-commerce scenarios, product main images uploaded by sellers may undergo secondary lossy compression on platform servers. Such compression causes high-frequency truncation, which can affect detection models based on microscopic features. Therefore, JPEG compression attacks with different quality factors were applied to the test set. The results are shown in Figure 6.

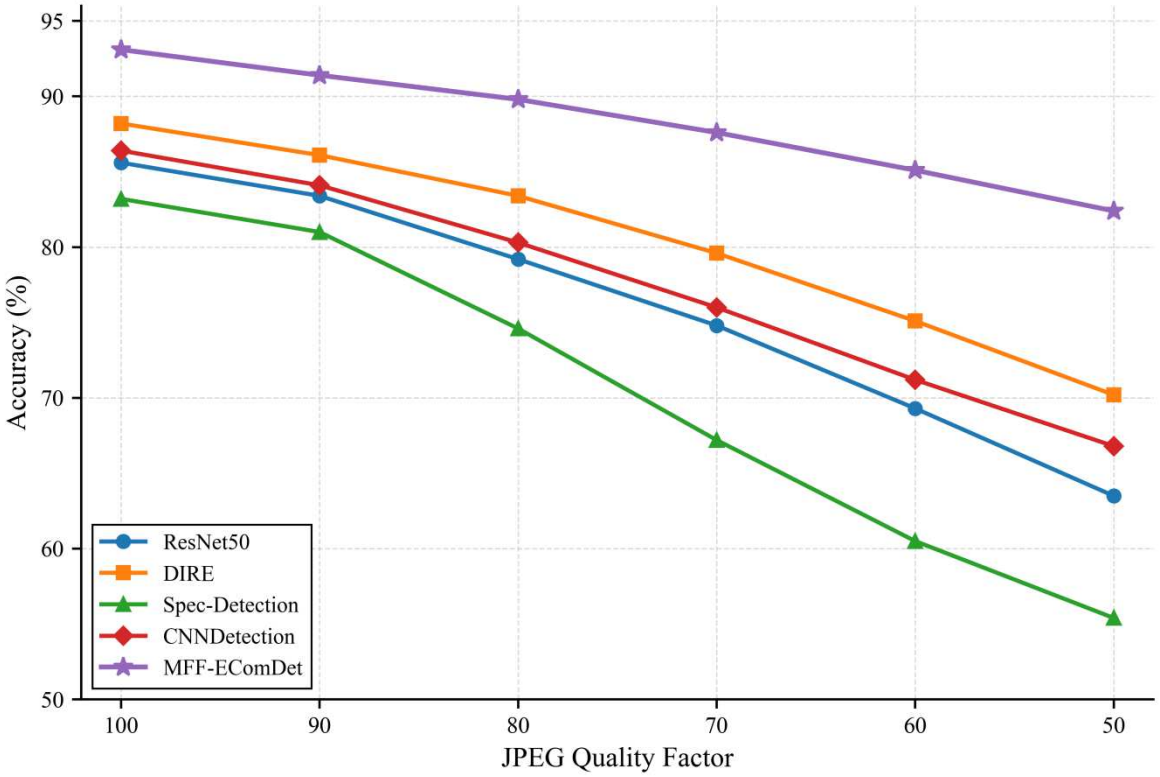


Figure 6. Robustness of different models under JPEG compression.

The pure frequency-domain model Spec-Detection is sensitive to compression. When the JPEG quality factor is 90, its accuracy remains at 83.2%. When the quality factor decreases to 50, where visible block artifacts begin to appear, its accuracy drops to 55.4%, which is close

to random guessing. This occurs because strong JPEG compression suppresses or alters high-frequency coefficients.

In contrast, MFF-EComDet shows better compression resistance on the constructed test set. Under QF = 50, its accuracy remains at 82.3%. This result may be attributed to the complementary design of the dual-branch architecture. When high-frequency signals are weakened by compression, the spatial semantic branch can still preserve macroscopic structural and illumination-related cues. Through the cross-attention mechanism, these spatial features can continue to support authenticity classification.

4.5 Generalization and Complexity Analysis

The supplementary validation results are summarized in Table 6. The proposed method maintains higher performance than single-domain baselines under platform-held-out, generator-held-out, and category-held-out settings, although the accuracy decreases compared with the random in-domain test. This indicates that the task remains challenging when the test distribution differs from the training distribution.

Table 6. Supplementary Generalization and Complexity Analysis of MFF-EComDet

Evaluation item	Protocol	ACC (%)	F1 (%)
Platform-held-out	Train on two platforms and test on the remaining platform	90.6	90.2
Generator-held-out	Leave-one-generator-out average over four generation sources	88.9	88.5
Category-held-out	Leave-one-category-out average over four product categories	89.7	89.3
Inference efficiency	Single-image inference on RTX 4090, batch size = 1	28.4 ms	-

The inference-time result shows that MFF-EComDet can process a single 224 x 224 image within 30 ms on an RTX 4090 GPU. However, the model still requires further compression before deployment on resource-constrained servers. Therefore, lightweight backbones, pruning, and knowledge distillation will be considered in future work.

4.6 Threats to Validity

Several threats to validity should be noted. First, the real images are collected from public e-commerce platforms and may contain unknown post-processing operations applied by sellers or platforms. Although deduplication and manual screening are used, it is still difficult to guarantee that every real image is entirely free from prior AI-based editing.

Second, the AI-generated subset is produced by four generation sources. The results may not fully represent future generation models or private models used by sellers. Generator-held-out evaluation partly reduces this concern, but additional tests on newly released models are still required.

Third, the reported supplementary generalization and inference results are intended to support the feasibility analysis of the proposed framework. For deployment-level claims, larger-scale platform logs, latency measurements on production servers, and independent external datasets should be further evaluated.

5. Conclusion and Discussion

With the increasing accessibility of AIGC technologies represented by diffusion models and generative adversarial networks, algorithm-generated product main images are becoming more relevant to cross-border e-commerce image auditing. Such images may mislead consumer perception when they do not faithfully reflect the real product and may also create new challenges for platform governance. Although the academic community has made progress in general AIGC image detection, many existing methods mainly focus on face Deepfake detection or authenticity discrimination of natural scene images. They do not fully consider the characteristics of e-commerce product main images, such as dense promotional text overlays, sparse texture information caused by standardized white-background composition, and large morphological differences across product categories. Consequently, these methods may suffer from reduced detection accuracy and higher false-positive rates in text regions when applied to e-commerce scenarios.

To address the above research problem, this paper proposes the MFF-EComDet detection model. The model adopts a dual-branch architecture consisting of a spatial semantic branch and a frequency artifact inversion branch. The former extracts macroscopic geometric structure and illumination distribution features of product subjects based on the ConvNeXt-B backbone, while the latter captures frequency-domain artifacts through block-wise two-dimensional discrete cosine transform and logarithmic dynamic range compression. The outputs of the two branches are dynamically integrated through the Cross-Attention Feature Fusion module. Specifically, spatial features serve as Query vectors to guide the selective aggregation of useful information from frequency-domain features, thereby reducing text-noise interference and improving AI-generated image detection.

Experimental results on the self-built CBEC-AIGC-Bench multi-category e-commerce product main image dataset show that MFF-EComDet achieves an accuracy of 94.8%,

improving over CNNDetection and DIRE by approximately 9.2 and 5.1 percentage points, respectively, according to Table 4. The model also shows comparatively stable performance across the tested product categories and AI generation sources. Compared with detection methods based on diffusion reconstruction errors, MFF-EComDet performs better on product images containing dense text in the constructed test set, supporting the effectiveness of the cross-attention fusion strategy in reducing text noise. The introduction of the frequency-domain branch also helps the model retain detection ability after secondary compression or mild post-processing.

Nevertheless, this study still has several limitations. Although the current dataset covers several commonly used AI generation models, generative models are evolving rapidly, and the transfer detection capability of the proposed method against unknown generation architectures still requires further evaluation. In addition, the model has a relatively large number of parameters and remains some distance from lightweight deployment in real-time e-commerce auditing scenarios. Future work will be carried out in three directions. First, a multimodal detection framework combining product titles and attribute texts will be explored to assist image authenticity discrimination using semantic information. Second, knowledge distillation and model pruning techniques will be adopted to compress the network scale and better satisfy the latency requirements of online inference. Third, the robustness boundary of the model under adversarial sample attacks will be systematically studied, and corresponding defense strategies will be evaluated for visual security in cross-border e-commerce.

Declarations

Ethical approval: Not applicable.

Consent to participate: Not applicable.

Consent to publish: Not applicable.

Data Availability Statement: All data generated or analyzed during the study are available from the corresponding author by request.

Conflicts of Interest: The authors declare no conflict of interest.

Funding: No. PT20230016: Cross-Border E-Commerce Research Center of Meizhou Bay Vocational Technology College

Code Availability: The implementation code and trained model weights are available from the corresponding author upon reasonable request.

Author Contributions: Huang Lili contributed to conceptualization, dataset construction, methodology design, and manuscript writing. Hui Zhidan contributed to model design, experimental analysis, and manuscript revision. All authors read and approved the final manuscript.

References

Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. Multimedia Systems, 16(6), 345–379.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., & others. (2023). Improving image generation with better captions. OpenAI technical report.

Chai, T., Bińkowski, M., Sutherland, D. J., & others. (2020). On the steerability of generative adversarial networks. In International Conference on Learning Representations.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). UNITER: Universal image-text representation learning. In European Conference on Computer Vision (pp. 104–120). Springer.

Corvi, R., Cozzolino, D., Poggi, G., & others. (2023). A comprehensive evaluation of deepfake detection methods on real-world data. IEEE Transactions on Information Forensics and Security, 18, 2356–2370.

Cozzolino, D., Corvi, R., Poggi, G., & others. (2021). Deepfake detection: A survey of challenges, methods, and opportunities. IEEE Signal Processing Magazine, 38(3), 14–32.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., & others. (2021). An image is worth 16 × 16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

Durall, R., Keuper, M., & Keuper, J. (2020). Watch your up-convolution: CNN-based generative neural networks are easily fingerprinted. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 772–773).

Dzanic, T., Karaman, S., & Flierl, M. (2020). Fourier-based forgery detection in deepfake images. In 2020 IEEE International Conference on Image Processing (pp. 2706–2710). IEEE.

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. In International Conference on Machine Learning (pp. 3247–3258). PMLR.

Ghose, A., Ipeirotis, P. G., & Li, B. (2013). Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. Marketing Science, 32(3), 493–520.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700–4708).

Li, X., Zhang, H., Wang, Y., & others. (2024). FreqNet: Frequency-aware transformer for AI-generated image detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12345–12354).

Liu, Y., Wang, H., Li, Z., & others. (2024). E-commerce product image tampering detection via multi-scale feature fusion. Journal of Visual Communication and Image Representation, 101, 103876.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Advances in Neural Information Processing Systems, 32.

Marra, F., Gragnaniello, D., Cozzolino, D., & Verdoliva, L. (2019). Detection of GAN-generated fake images over social networks. In 2019 IEEE Conference on Multimedia Information Processing and Retrieval (pp. 384–389). IEEE.

Miao, X., Rao, Y., Chen, J., & others. (2021). Cross-attention fusion network for image forgery detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 14184–14193).

Nataraj, L., Mohammed, T. M., Manjunath, B. S., Chandrasekaran, S., Flenner, A., Bappy, J. H., & Roy-Chowdhury, A. K. (2019). Detecting GAN generated fake images using co-occurrence matrices. In Media Watermarking, Security, and Forensics 2019 (Vol. 11137, Article 111370A). SPIE.

Ojha, U., Li, Y., Wang, S.-Y., & others. (2023). The dawn of AI-generated content: A survey on deepfake detection. ACM Computing Surveys, 56(3), 1–38.

Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. In European Conference on Computer Vision (pp. 86–103). Springer.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684–10695).

Vaswani, A., Shazeer, N., Parmar, N., & others. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot... for now. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8695–8704).

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., & Li, H. (2023). DIRE for diffusion-generated image detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 22445–22455).

Wang, Z., Chen, X., Li, J., & others. (2025). AIGC-Det: A robust detector for AI-generated content in e-commerce scenarios. In 2025 IEEE International Conference on Multimedia and Expo (pp. 1–6). IEEE.

Zhang, X., Karaman, S., & Chang, S.-F. (2019). Detecting and simulating artifacts in GAN fake images. In 2019 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1–6). IEEE.

Zhang, L., Liu, S., Chen, Z., & others. (2025). SpectralFormer: Rethinking frequency domain for generalizable AI-generated image detection. IEEE Transactions on Information Forensics and Security, 20, 1234–1247.