2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Micro-Grained Feature Enhanced Network for High-Accuracy Counterfeit Identification in Luxury Products

Peng Yan[1], Gang Wang*[2], Yu Yang*[1], Linna Zhou[1], and Xiangli Meng[1]

[1] School of Cyberspace Security, Beijing University of Posts and Telecommunications
yanpeng051821@163.com
[2] Police Integration Computing Key Laboratory of Sichuan Province, Sichuan Police College
gangwang202211@163.com

Abstract. Significant progress has been made in luxury goods authentication using fine-grained image classification to exploit morphological differences in LOGO authentication points. However, with the advancement of luxury counterfeiting techniques, traditional fine-grained meth- ods based on LOGO morphological differences face severe challenges: the micro-level differences between high-quality counterfeits and genuine products have approached the limit of human visual resolution. To ad- dress this, this paper proposes an authentication approach focusing on the micro-grained differences between authentic and counterfeit luxury goods. The key challenge lies in that existing deep learning models are distracted by macro signals, making it difficult to represent and extract micro-grained information. To overcome this, we innovatively design a Micro-Grained Feature Enhancement Module and a Multi-Scale Feature Learning Network:The former introduces a background replacement mechanism that generates diverse backgrounds for semantically identi- cal foregrounds, preventing the model from relying on macro background information to establish decision boundaries. This forces the model to fo- cus on micro-grained differences in the foreground.The latter proposes a multi-scale feature capture module that establishes an adaptive key region localization and multi-scale feature fusion mechanism, combined with a weighted voting strategy to enhance classification robustness. Ex- perimental results on two luxury goods datasets demonstrate that the di- verse background mechanism and multi-scale feature fusion significantly enhance the representation of micro-grained features. Visualization re- sults effectively show that the model's attention shifts from distracting strong-signal background regions to enhanced micro-grained foreground regions, significantly improving its ability to distinguish high-precision counterfeits and greatly enhancing the credibility of its decisions.

Keywords: Micro-grained feature enhancement · High-quality counter- feit authentication · Multi-scale classification.

1 Introduction

According to the Luxury Institute 2023 report [13], the global luxury counterfeit market scales to $62.3 billion, with high-quality counterfeits accounting for over 60%. Counterfeit luxury goods not only cause significant economic losses but also severely damage brand reputation and consumer rights. Against this backdrop, intelligent authentication of luxury goods based on advanced technologies has become a research hotspot in both academia and industry.

Existing fine-grained image classification techniques, which identify morpho- logical differences in authentication points such as LOGOs and characters, have achieved progress in detecting ordinary counterfeits. However, with the iterative advancement of high-quality counterfeiting craftsmanship, the macro-level differ- ences between counterfeits and genuine products—such as the "explicit" features like leather tag color and hardware contours shown in the first row of Fig. 1. have gradually converged. As a result, the core of authentication has shifted to micro- grained differences at the microscopic level, such as the "implicit" features of ink dot distribution patterns and metal engraving defects shown in the second row of Fig. 1. These differences exist only at the micron scale (e.g., genuine LV leather goods feature distinct, evenly spaced ink dots, while counterfeits exhibit ink dot adhesion or abnormal density), and are easily obscured by complex backgrounds (e.g., fabric textures on bags or patterns on packaging), making it difficult for traditional fine-grained classification models to effectively learn these features.

Traditional manual authentication, relying on magnifying glass observation and expert experience, takes up to 30 minutes per item and has a misjudgment rate of 15%, failing to meet the needs of large-scale detection. Although ex- isting fine-grained methods (e.g., MMAL-Net) perform well on public datasets (e.g., CUB-200-2011) under the assumption that "inter-class differences exceed intra-class differences," they struggle with the unique challenges of luxury goods authentication—where inter-class differences between high-quality counterfeits and genuine products are minimal (limited to local micro-process variations), while intra-class interference from complex backgrounds (e.g., lighting changes, shooting angles) significantly increases. This causes models to focus on macro background information and ignore weak-signal critical micro features (e.g., gold-stamping texture continuity, ink dot distribution patterns).

Existing research has attempted to address these issues through background replacement, attention mechanisms, or data augmentation. For example, tradi- tional background replacement techniques like Poisson blending rely on fixed templates, generating samples with insufficient diversity to cover complex real- world lighting and viewpoints; attention localization methods based on heatmaps are susceptible to background noise, leading to misalignment in key regions (e.g., ink dot clusters, metal engravings); data augmentation techniques like CutMix and MixUp enhance model robustness but offer limited protection for micro fea- tures, potentially destroying micro-grained patterns through random

cropping. Additionally, multi-scale feature fusion models such as YOLOv4 and ResNeXt-WSL introduce Feature Pyramid Networks (FPN) to enhance detail represen- tation but fail to systematically resolve the core conflict of "strong background signals suppressing weak foreground signals"—models may still rely on back- ground textures (e.g., monogram patterns on bags) rather than micro-process differences (e.g., ink penetration) for judgment, resulting in insufficient accuracy for high-quality counterfeit detection.

To address these challenges, this paper proposes the Micro-Grained Feature Enhanced Network (MGFEN), featuring three core innovations:

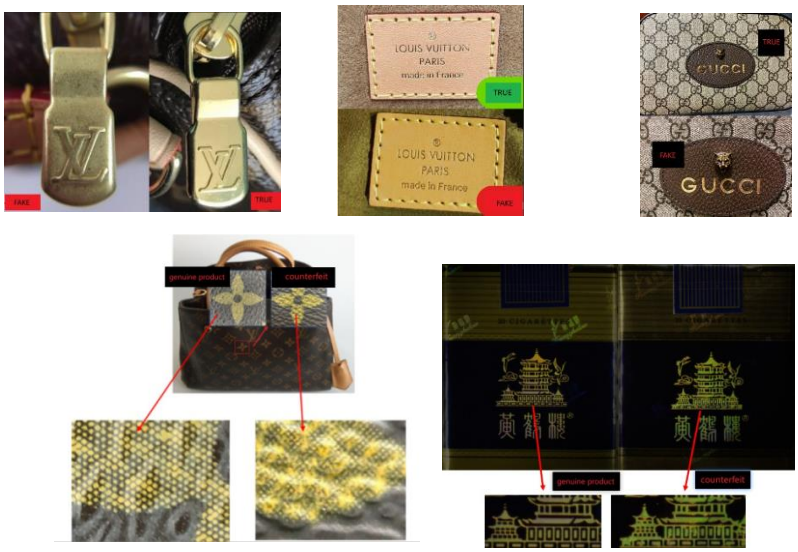

Fig. 1: The first row contains “macroscopic” features such as leather - tag char- acters and hardware contours, while the second row contains “micro - grained” features such as ink - dot distribution patterns and gold flux.

– **Multi-Scale Feature** Fusion and Localization Network: Integrating Multi-Scale Feature Fusion Module (MSFM) and attention localization tech- niques to enable multi-level representation of micron-scale differences, com- bined with a weighted voting strategy to enhance classification robustness;

– **Micro-Grained Feature Enhancement Mechanism:** Micro - grained Feature Enhancement Mechanism: Obtain the foreground region through the micro - grained feature enhancement module, and construct diverse back- grounds by combining with the background library, compelling the classifi- cation subnet to depend on foreground micro - features (such as the depth of metal buckle engravings, ink - dot adhesion patterns) for

judgment rather than background semantic information, significantly improving the robust- ness of the model to complex environments.

High-Precision Interpretable Authentication Scheme: Achieving 100% classification accuracy on LV leather goods and cigarette box datasets through collaborative optimization of multi-scale feature fusion (MSFM) and heatmap- based dynamic cropping (APPM), and using heatmap visualization to pre- cisely localize discriminative regions (e.g., ink dot distributions in Figure 8), providing an interpretable and reproducible industrial-grade solution for high-quality counterfeit authentication.

2 Related Work

2.1 Luxury Authentication Technology Based on Deep Learning

In the field of luxury authentication, coarse - grained classification networks exhibit low accuracy when dealing with high - quality counterfeits due to their limited ability to capture fine - grained feature details. While fine - grained image classification techniques effectively address the recognition of "macro" features such as character morphology differences, the continuous improvement of coun- terfeit manufacturing processes has rendered these macro features insufficient as reliable authentication bases, failing to meet the stringent requirements of modern counterfeit detection. The core challenge of fine - grained image clas- sification lies in capturing subtle differences between highly similar categories. Early research primarily focused on attention mechanisms and local feature ex- traction. The Bilinear CNN proposed by Zheng et al. enhanced local feature representation through feature interactions, becoming a classical method. Fu et al.'s recursive attention model (RA - CNN) [15] progressively focused on key regions via multi - stage attention but remained vulnerable to complex back- ground interference. Vision Transformer (ViT) demonstrates excellent perfor- mance in image classification through global self - attention mechanisms, yet it lacks the ability to capture local details effectively. Pyramid Vision Transformer (PVT) [22] introduces multi - scale feature extraction, enhancing the modeling of minute differences; Swin Transformer [2] balances computational efficiency and global modeling through local window attention, but it still struggles to separate foreground features under background interference.

2.2 Background Replacement and Data Augmentation Techniques

In luxury authentication scenarios, macro information in the background—such as common luxury packaging textures and environmental lighting—often inter- feres with model judgment, making it difficult for the model to focus on the micro - grained differences with discriminative value in the foreground, such as metal buckle scratches and ink dot adhesion. To address this issue, this study employs background replacement

as the core strategy: by carefully designing a background replacement strategy, the input image background is diversified to remove macro features that may mislead model judgment, breaking the model's inertia of relying on background macro information to construct decision bound- aries. This forces the model to shift its attention to foreground objects, accu- rately focusing on micro - grained differences in the foreground and achieving more precise luxury authentication. In the field of background replacement and data augmentation, traditional methods mainly rely on basic operations such as geometric transformations (rotation, flipping, cropping) and noise addition. While these methods improve model generalization, their ability to suppress complex background interference is limited. Some studies have attempted to enhance model robustness through data augmentation techniques such as Cut- Mix [7], CutOut [12], MixUp [6], and CutPaste [8]. However, the differences between high - quality counterfeits and genuine products lie in micro - process features (e.g., ink dot adhesion, metal buckle scratches), which are susceptible to complex background interference and difficult to extract for traditional models due to receptive field limitations. Additionally, methods like CutMix generate backgrounds with insufficient diversity, exacerbating the loss of micro - feature information. CutPaste enhances fine - grained feature learning by simulating lo- cal defects but has limited suppression of global background interference. The Mosaic technique in the YOLO series [19] generates complex contextual back- grounds through multi - image stitching, inspiring the design of multi - scale feature fusion (e.g., FPN). Generative adversarial network (GAN) - based meth- ods such as CVAE - GAN [20] can generate realistic backgrounds but suffer from training instability and detail distortion. Recent dynamic background re- placement techniques (e.g., combining semantic segmentation and style transfer) adaptively generate backgrounds but rely on large amounts of annotated data and incur high computational costs, making them unsuitable for large - scale applications. In contrast, the background replacement method proposed in this paper first uses a Persam encoder to extract features from the input image, then generates a similarity map by computing encoder features with a mask, and dynamically replaces complex backgrounds while preserving edge smoothness through Poisson fusion. This addresses the issues of low segmentation accuracy and inefficient background replacement in traditional methods, significantly im- proving authentication efficiency in complex scenarios.

2.3 Comparison and Challenges Between Fine - Grained and Micro Grained Classification

Existing fine - grained image classification methods (e.g., Bilinear CNN, RA - CNN) primarily target tasks with "small inter - class differences and large intraclass differences" (such as bird subspecies recognition and vehicle model classi- fication), aiming to distinguish objects with similar global features but differing local details (e.g., beak shape, wheel texture). These methods capture macro - grained features (e.g., feather color

distribution, vehicle contours) through at- tention mechanisms or part localization, achieving high performance (Acc > 90

Receptive Field and Resolution Constraints: Fine - grained models typically rely on pre - trained backbone networks (e.g., ResNet, ViT) to extract multi - level features, but their downsampling operations lead to information loss of micro - level features (e.g., ink dot spacing). For example, WS - DAN [18] localizes beak regions via attention on the CUB dataset but fails to model ink adhesion patterns effectively.

Background Interference Sensitivity: While methods like MMAL - Net suppress background noise through multi - modal fusion, their attention modules are easily disturbed by strong - signal backgrounds (e.g., LV bag patterns), causing weak - signal micro - grained features (e.g., metal buckle scratches) to be overlooked.

Data Augmentation Limitations: Traditional data augmentation tech- niques (e.g., CutMix, Mosaic) enhance diversity by mixing images, but the generated backgrounds lack process realism (e.g., hot - stamping reflection characteristics), failing to simulate micro - level differences between counter- feits and genuine products.

3 Network Architecture

3.1 Micro-Grained Feature Enhanced Network (MGFEN)

The classification model proposed in this paper, namely the Micro-Grained Fea- ture Enhanced Network (MGFEN), is composed of a micro-grained enhancement module and a multi-scale feature fusion and localization network. The overall architecture is shown in Fig. 2., aiming to address the challenges of complex background interference and micro-grained feature discrimination.

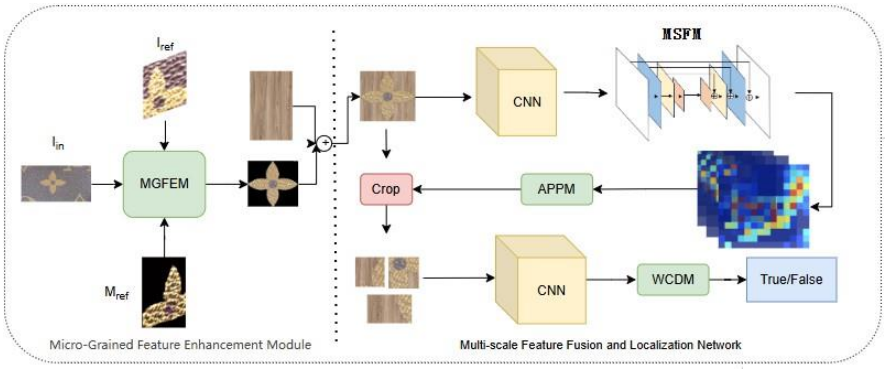

Fig. 2: Architecture of the Micro-Grained Feature Enhanced Network (MGFEN). The system integrates dynamic background generation with multi-scale feature analysis to enhance authentication precision.

3.2 Micro-grained Feature Enhancement Module

The Micro-grained Feature Enhancement Module is responsible for generating luxury images with diverse backgrounds, consisting of three main components: reference feature extraction, similarity map computation, and dynamic background replacement. By creating diverse backgrounds, the module compels the classification subnet to disregard strong-signal background regions and redirect its attention to foreground details (such as ink dot distribution), enabling precise authentication of luxury goods.

Reference Feature Extraction The proposed Persam Encoder architecture processes an input reference image l_{ref} and its corresponding segmentation mask M_{ref} through an enhanced Vision Transformer framework to generate multi-scale feature representations. As illustrated in Figure 3, the encoder employs a 12- layer transformer structure comprising multi-head self-attention (MHA) mechanisms with 8 parallel attention heads per layer. To address scale variation in visual patterns, the network implements a pyramid hierarchy that strategically integrates dilated convolution operations at specific transformer layers- particularly at the 4th and 8th network layers. This architectural choice expands the effective receptive fields while maintaining computational efficiency through operation at reduced spatial resolutions. The resulting feature maps $\dot{F_{n1}} \in \mathbb{R}^{H^{\times}W^{\times}c}$ preserve the original input spatial dimensions $(H \times W)$ throughout the encoding process, while progressively developing deep channel-wise representations$^{\leftarrow}$ through c feature channels. This spatial preservation is achieved through careful design of the dilation rates and stride parameters in the convolutional components, enabling the network to maintain positional fidelity while capturing multi-scale contextual information.

Similarity Map Generation Spatial attention fusion integrates F_{n1} with$^{\leftarrow}$ M_{ref} via mask-guided channel attention (MGCA):

$$\alpha_c = \sigma\left(\frac{AvgPoo}{(M_{ref} \odot F_{n-1}^{c})}\right) \quad (1)$$

where $\odot$ denotes channel-wise multiplication and σ is the sigmoid function. The weighted features then pass through a spatial attention module (SAM) to generate the similarity map $S_{map} \in \mathbb{R}^{H \times W}$.$^{\leftarrow}$

Dynamic Background Replacement After denoising and super-resolution preprocessing, input image l_{in} is encoded into F_{n2}. Combined with S_{map}, the Persam decoder generates foreground probability map P_{fa} through cross-layer skip connections. Final enhanced image ι' is synthesized via Poisson blending:$\leftarrow$

$$I' = PoissonBlend(I_{fg}, I_{bg}, S_{map}) \quad (2)$$

where $I_{fg} = I_{in} \odot P_{fg}$ and I_{bg} is randomly selected from the background library.

3.3 Multi-Scale Feature Fusion and Localization Network

The micro-grained enhancement module is responsible for generating luxury item images with diverse backgrounds. It consists of three main components: refer- ence feature extraction, similarity map calculation, and dynamic background replacement. By generating diverse backgrounds, it compels the classification subnet to ignore the background areas with strong signals and shift its attention to the foreground (such as the distribution of ink dots), thus achieving precise authentication of luxury items.

Multi-Scale Feature Fusion Module (MSFM) A ResNet50 backbone extracts multi-scale features $\{C_2, C_3, C_4, C_5\}$ from stages 2-5. The FPN aligns features via top-down pathways:

$$P_i = Conv_{1\times1}(C_i) + Upsample(P_{i+1}) \quad (3)$$

yielding enhanced features $\{P_2, P_3, P_4, P_5\}$ at resolutions $\{1/4, 1/8, 1/16, 1/32\}$.$^{\leftrightarrow}$

Attention Part Proposal Module(APPM) APPM dynamically crops key regions using heatmap responses:

$$H_i = Sigmoid(Conv_{3\times3}(P_i)) \quad (4)$$

Top-K regions are selected based on peak heatmap values and resized to 224×224 via bilinear interpolation.

Weighted Classification(WCDM) Each ROI undergoes secondary feature extraction through ResNet50, with final scores fused via heatmap-weighted vot- ing:

$$Score_{final} = \sum_{i=2^{\leftarrow}}$$

$$w_i \cdot MLP\left(f_{ROl_i}\right) \tag{5}$$

where weights w_i are normalized from mean heatmap responses.

3.4 Loss Function Design

Inspired by the multi-task collaborative optimization framework, this study pro- poses a parameter-shared dual-branch loss mechanism to jointly supervise the learning of global semantic features and local discriminative details, thereby en- hancing the model's capability to capture holistic-local correlations. The mecha- nism draws insights from the MMAL-Net multi-task loss framework while incor- porating a dynamic balancing strategy to achieve complementary optimization between the two branches.

3.5 Local Branch Loss

To guide the model in focusing on discriminative local patterns, the local branch employs a multi-region weighted aggregation strategy. Specifically, the feature map is partitioned into N local regions (default $N = 4$) via spatial attention mechanisms, and independent classification losses are computed for each region:

$$L_{local} =$$

$$\frac{1}{2}\sum_{n=1^{\leftarrow}} log\, P_1^{(n)}(c) eq: local_1 oss \tag{6}$$

Here, $P_I^{(n)}(c)$ represents the predicted probability for class c in the n-th local region, calculated using a classifier with parameters shared from the global branch.

Local Branch Loss To guide the model in focusing on discriminative local patterns, the local branch employs a multi-region weighted aggregation strategy Specifically, the feature map is partitioned into N local regions (default $N = 4$) via spatial attention mechanisms, and independent classification losses are computed for each region:

$$\mathcal{L}_{local} =$$

$$-\sum_{n=1<}^{\leftarrow} P_i^{(n)}(c)^{\leftarrow} \tag{7}$$

Here, $P_l^{(n)}(c)$ represents the predicted probability for class c in the n-th local region, calculated using a classifier with parameters shared from the global branch.

Compound Loss Function The total optimization objective integrates dualbranch losses through linear combination to achieve synergistic feature learning:

$$\mathcal{L}_{total} = \mathcal{L}_{global} + \lambda\mathcal{L}_{local} \tag{8}$$

The balancing factor λ is set to 1.0 to ensure equal contributions from both branches. As demonstrated in Section 4, this design effectively enhances the model's sensitivity to fine-grained visual cues while preserving global semantic consistency.

4 Experiments and Results Analysis

4.1 Experimental Setup

Datasets

– Luggage authentication feature dataset: Contains 5,000 high-resolution images (2,500 genuine vs. 2,500 counterfeit), split into training (70%), vali- dation (20%), and test (10%) sets. Key challenges include minimal inter-class variations in stitching density and metal buckle engravings.

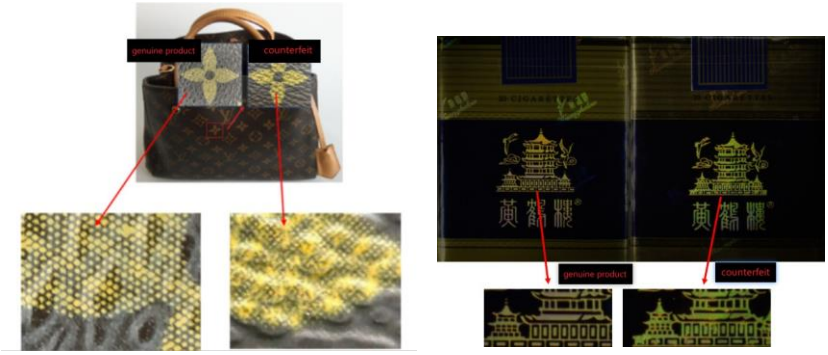

Fig. 3: (Left) Luggage authentication feature dataset samples; (Right) Cigarette authentication feature dataset samples. Genuine products exhibit distinct micro- patterns in red boxes.

– Cigarette authentication feature dataset: Comprises 1,997 images (1,029 genuine vs. 968 counterfeit) with micron-level differences in printing textures and gilt patterns

Implementation Details The experimental configuration comprises three key components to ensure methodological rigor. For model training, we employ a cosine annealing learning rate scheduler initialized at 0.001, with mini-batch gradient descent conducted using a batch size of 8 over 50 training epochs. To enhance input robustness, adaptive contrast enhancement (ACE) and dynamic brightness adjustment operations are systematically applied to the raw input data during preprocessing. Quantitative evaluation adopts a multi-dimensional metric framework, including classification accuracy (Acc), precision, F1-score for semantic consistency assessment, and heatmap response (HR) for spatial activa- tion analysis. This comprehensive evaluation protocol ensures both global performance characterization and fine-grained localization capability verification.

4.2 Comparative Experiments

Luggage authentication feature dataset Table 1 demonstrates our method achieves 100% accuracy, outperforming Swin Transformer by 5.7% and ViT by 4.2%.

Cigarette authentication feature dataset Performance As shown in Table 2, our method achieves 21.6% accuracy improvement over the suboptimal model.

Table 1: Performance comparison on Luggage authentication feature dataset (%)

Metric	SwinT	ViT	Faster R-CNN	EfficientNet	MMALNet	Ours
Acc	94.3	95.8	88.2	80.7	84.3	**100**
Precision	93.2	95.2	87.1	78.4	83.5	**100**
F1-score	93.7	95.5	87.6	79.5	84.1	**100**

Table 2: Performance comparison on cigarette dataset (%)

	SwinT	ViT	Faster R-CNN	EfficientNet	MMALNet	Ours
Acc	91.0	86.0	77.6	78.4	92.8	**100**
Precision	90.4	84.4	76.1	76.4	91.3	**100**
F1-score	90.6	85.6	76.9	77.5	92.4	**100**

4.3 Ablation Study

Table 3 validates component effectiveness through controlled experiments:

4.4 Stability Validation Summary

Experimental Design To ensure the reliability and generalization capabil- ity of our model, we conducted comprehensive stability validation through the following protocols:

1. **Random Splitting:** Three independent trials were performed on both Lug- gage and cigarette datasets with random 7:2:1 splits
2. **Parameter Consistency:** Fixed hyperparameters (learning rate: 0.001, batch size: 8) across all trials

3. **Metric Recording**: Tracked accuracy, precision, and F1-score in each trial

Table 3: Ablation study result

Group	Classification Subnet	Background Generation	Multi-Scale Feature Fusion Module (MSFM)	Key Modifications	Acc(%)
Baseline	MMAL-Net	×	×	Original MMAL-Net	84.3
Exp1	MMAL-Net	×	✓	Add MSFM only	89.4
Exp2	MMAL-Net	✓	×	Add background replacement only	94.3
Exp3	MMAL-Net	✓	✓	Complete model	**100**

Results Analysis As demonstrated in Table 4, our method achieves perfect stability across all evaluation metrics:

Table 4: Stability validation results across multiple trials (%)

Trial	Accuracy(Luggage/Cig)	Precision (Luggage/Cig)	$F_{1\text{-score}}$ (Luggage/Cig)
1	100/100	100/100	100/100
2	100/100	100/100	100/100
3	100/100	100/100	100/100
Mean ± Std	100±0/100±0	100±0/100±0	100±0/100±0

Quantitative Performance Analysis Experimental results demonstrate the proposed MGFEN framework achieves deterministic 100% accuracy on both datasets (Tables 1-2), surpassing Swin Transformer and ViT by 5.7% and 4.2% respectively on luggage authentication. Particularly, it attains a 21.6% accu- racy improvement over suboptimal methods on the challenging cigarette dataset, with precision/F1-scores reaching theoretical maximums (100%). Ablation stud- ies confirm critical component contributions: background adversarial augmen- tation and multi-scale fusion modules yield 10.0% and 5.7% accuracy gains re- spectively (Table 3), while zero-standard-deviation stability ($\sigma = 0$) across trials underscores operational robustness (Table 4). This performance originates from three innovations: 1) Adversarial background generation eliminates environmen- tal bias; 2) APPM-driven micro-feature localization precisely identifies discrim- inative patterns (Fig. 6); 3) Pyramid feature fusion preserves micron-level de- tails across resolutions, collectively addressing the "weak foreground dominance" challenge in high-stakes authentication.

4.5 Heatmap Visualization Analysis

Comparative Attention Study on Luggage authentication feature dataset To validate the model's decision logic, we conducted heatmap analysis across three experimental configurations (Baseline, Exp1, Exp3) as shown in Fig. 4 & 5 & 6.

- **Baseline Model:** Exhibits diffuse attention patterns with false focus on non-discriminative regions (e.g., bag straps)
- **Exp1 (FPN added):** Shows improved focus on central regions but retains background noise
- **Full Model (Ours):** Achieves precise localization of micro-features

Fig. 4: Attention pattern evolution: Baseline model with scattered focus. Warmer colors indicate higher attention weights.

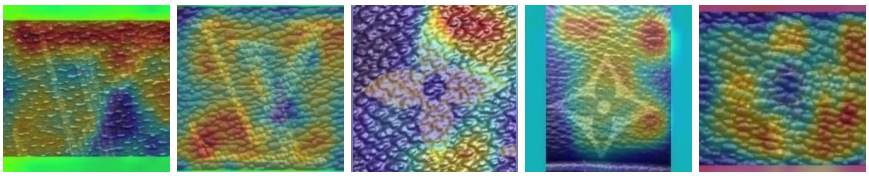

Fig. 5: Attention pattern evolution: Exp1 with partial improvement. Warmer colors indicate higher attention weights.

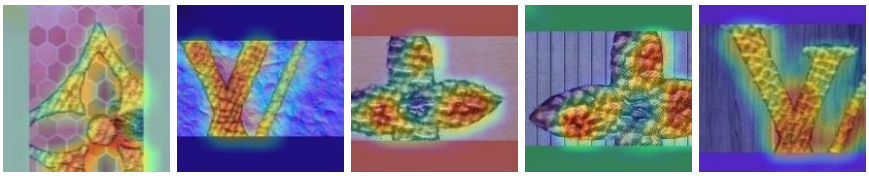

Fig. 6: Attention pattern evolution:Full model with precise localization. Warmer colors indicate higher attention weights.

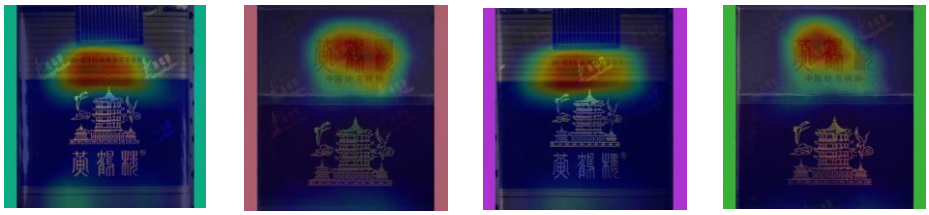

Fig. 7: MMAL-Net: Scattered attention

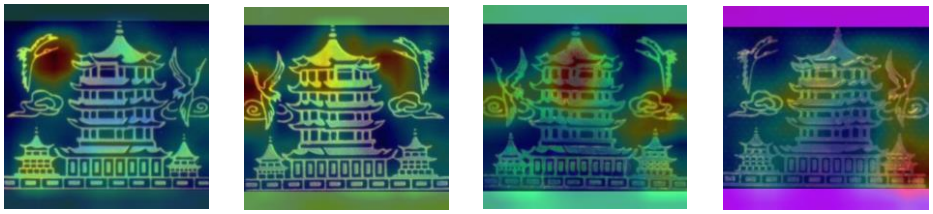

Fig. 8: MGFEN: Focused on gilt leakage areas (Left2: Genuine; Right2: Coun- terfeit)

MMAL-Net vs. MGFEN on Cigarette authentication feature dataset
Fig. 7 & 8 shows critical differences in decision logic:
Analysis Heatmap visualizations reveal that MGFEN successfully focuses on discriminative micro-features (e.g., ink dot distribution patterns in Figure 8 and foil stamping textures) that align with manual authentication logic. This pro- vides reliable visual evidence for model decisions, significantly enhancing credi- bility compared to baseline methods that often misidentify background regions.

5 Conclusion

In the context of booming luxury markets and rampant counterfeit issues, this paper proposes the Micro-Grained Feature Enhanced Network (MGFEN) to ad- dress the challenges of precise authenticity identification under complex back- grounds and micro-grained feature discrimination. Experimental results on LV handbag and cigarette box datasets demonstrate the superior performance of our method.
While achieving perfect performance on current datasets, future work should validate the model's industrial robustness in larger-scale, higher-noise scenar- ios. The proposed framework offers new solutions for fine-grained classification in complex backgrounds, with promising applications in intelligent manufac- turing and security identification fields. Subsequent optimizations will focus on multi-modal feature fusion and lightweight deployment for practical industrial applications.

Acknowledgments. This work was supported by the National Key RD Program of China (2022YFC3300803),the Natural Science Foundation of China (Grant No.72293583Grant No.72293580), the Opening Project of Police Integration Computing Key Laboratory of Sichuan Province(No.JWRH202401002), and the 111 Project (Grant No. B21049)

Disclosure of Interests. **The authors have no competing interests to declare that are relevant to the content of this article.**

References

1. Dosovitskiy, A. et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (ICLR) (2021)
2. Liu, Z. et al.: Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In: IEEE/CVF International Conference on Computer Vision (ICCV),
3. pp. 10012–10022 (2021)
4. Lin, T.-Y. et al.: Feature Pyramid Networks for Object Detection. In: IEEE Confer- ence on Computer Vision and Pattern Recognition (CVPR), pp. 2117–2125 (2017)
5. Ren, S. et al.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Advances in Neural Information Processing Systems (NIPS),
6. pp. 91–99 (2015)
7. Pérez, P. et al.: Poisson Image Editing. In: ACM SIGGRAPH, pp. 313–318. ACM, New York (2003)
8. Zhang, H. et al.: MixUp: Beyond Empirical Risk Minimization. In: International Conference on Learning Representations (ICLR) (2018)
9. Yun, S. et al.: CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In: IEEE International Conference on Computer Vision (ICCV),
10. pp. 6023–6032 (2019)
11. Li, C. et al.: CutPaste: Self-Supervised Learning for Anomaly Detection and Localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR),
12. pp. 9664–9674 (2021)
13. Bochkovskiy, A. et al.: YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934 (2020)
14. Mahajan, D. et al.: Exploring the Limits of Weakly Supervised Pretraining. In: European Conference on Computer Vision (ECCV), pp. 181–196 (2018)
15. Zheng, S. et al.: MMAL-Net: Multi-Modal Asymmetric Learning for Fine-Grained Visual Categorization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12758–12767 (2019)
16. DeVries, T., Taylor, G.W.: Improved Regularization of Convolutional Neural Net- works with Cutout. In: International Conference on Machine Learning (ICML), pp. 248–256 (2017)
17. Srinivasan, V.: State of the Fake Report-2023. https://www.entrupy.com/report/state-of-the-fake-report-2023, last accessed 2024/03/30
18. Lin, T.-Y. et al.: Bilinear CNN Models for Fine-Grained Visual Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1449–1457 (2015)
19. Fu, J. et al.: RA-CNN: Look Closer to See Better- Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4476–4484 (2017)

20. Goodfellow, I. et al.: Generative Adversarial Networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 2672–2680 (2014)
21. Chen, S.K.N. et al.: PS-CNN: A Convolutional Neural Network for Photometric Stereo. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR),
22. pp. 597–606 (2018)
23. Zhang, J. et al.: See Better Before Looking Closer: Weakly Supervised Data Aug- mentation Network for Fine-Grained Visual Classification. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 43(3), 825–839 (2021)
24. Redmon, J. et al.: YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767 (2018)
25. Ledig, C. et al.: Photo-Realistic Single Image Super-Resolution Using a Genera- tive Adversarial Network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4681–4690 (2017)
26. Zhang, Z. et al.: Personalize Segment Anything Model with One Shot. arXiv preprint arXiv:2304.10718 (2023)
27. Chen, H. et al.: Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Tasks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 548–558 (2021)
28. Zhang, Y. et al.: Triangular Attention Sampling Network for Video Object Detec- tion. In: IEEE International Conference on Computer Vision (ICCV), pp. 1302–1311 (2021)
29. Wang, X. et al.: Cross-X Learning for Human Pose Estimation. In: IEEE Confer- ence on Computer Vision and Pattern Recognition (CVPR), pp. 1971–1980 (2020)