Episodic Meta-Learning with Cross-Modal Contrastive Regularization for Unknown-Type Forgery Detection

Javier Martínez Ruiz*

Universitat Politècnica de València,València,Spain

Email:javier.martinez.ruiz@ucm.es

Abstract—The rapid proliferation of generative adversarial networks and sophisticated image editing tools has made forged visual content increasingly easy to synthesize yet difficult to identify. Most existing detectors are trained under a closed-set assumption and degrade sharply when the forgery type at deployment was never seen during training, while the annotated samples of each novel manipulation are typically scarce. This paper addresses unknown-type forgery detection as an episodic few-shot learning problem and proposes a framework termed EM-CMCR. A dual-stream encoder jointly models high-level visual semantics from an RGB stream and low-level forensic fingerprints from a constrained-convolution noise stream, after which a cross-modal fusion module produces a unified representation. An episodic prototypical learner organizes training into N-way K-shot tasks so that the decision boundary adapts to novel forgery types from only a handful of exemplars. To prevent the two modalities from drifting into incompatible subspaces, a cross-modal contrastive regularization term pulls the visual and forensic embeddings of the same image together while pushing apart those of different images, and a symmetric prototype-distribution alignment term further enforces consistent class posteriors. Extensive experiments on four face forgery benchmarks and one generic manipulation benchmark show that EM-CMCR reaches 78.6% and 89.4% five-way accuracy under one-shot and five-shot settings, and obtains an average AUC of 86.2% across five unseen test domains, outperforming strong supervised and meta-learning baselines. Ablation studies confirm that both the noise stream and the contrastive regularization contribute substantially to the generalization gains.

Index Terms—Image forgery detection, deepfake, few-shot learning, meta-learning, contrastive learning, cross-modal fusion, open-set generalization.

I. Introduction

Visual content has become the dominant carrier of information on the modern Internet, but its trustworthiness is being steadily eroded by powerful generative and editing techniques that are rapidly becoming accessible to ordinary users. In the past, high-quality image manipulation often demanded professional photography skills[1-3], sophisticated editing software, and hours of careful manual adjustment, creating a high barrier for realistic forgery. Today, generative adversarial networks, diffusion models, and automatic face-swapping pipelines allow a non-expert to produce photorealistic forged images in seconds, with outputs

that are visually indistinguishable from authentic photographs to casual human observers[4-6]. These forged materials are not merely limited to trivial entertainment or artistic creation. Such content can be weaponized for large-scale misinformation campaigns, identity fraud, blackmail, and targeted reputational damage against individuals, public figures, and organizations[7-10]. As synthetic media continues to spread across social media platforms, instant messaging channels, and news ecosystems, reliable forgery detection has evolved from a niche computer vision research topic into a problem of growing societal importance, with implications for digital governance, media credibility, and personal privacy protection. Notably, the core idea of excavating subtle, physically grounded signal fingerprints from visual data has been proven effective in remote sensing imaging tasks. Ye et al.[4] (2024) designed a convolutional modulated scattering feature network to capture delicate scattering signatures for aircraft classification in SAR images, demonstrating the value of specialized feature extractors for revealing implicit signal characteristics hidden within image data. Extending this line of cross-modal signal analysis, Lan et al.[5] (2026) proposed a multi-level cross-modal edge filtering strategy for high-resolution optical–SAR image registration, showing that complementary information from heterogeneous image modalities can be fused to stabilize feature matching against distribution shifts. Further, Li et al.[6] (2026) developed EOCNet with elongated oriented convolutions to capture anisotropic structural patterns for high-aspect-ratio object detection in SAR imagery. Collectively, these studies illustrate that heterogeneous visual signals contain stable low-level statistical traces that are decoupled from intuitive semantic appearance[11-14]. This insight inspires our work: rather than relying only on RGB visual patterns, we seek to exploit complementary forensic fingerprints to detect synthetic forgeries under open-set conditions where the manipulation type is unseen.Following the release of large-scale synthetic media benchmarks in recent years, a rich line of forgery detectors has been developed by the research community, ranging from lightweight compact convolutional classifiers suitable for edge deployment to deep backbone networks augmented with handcrafted frequency and noise cues[15-17]. Most of these supervised detectors are trained on fixed collections of forgery categories, learning to capture distinctive residual artifacts introduced by specific generative models or editing workflows. Under closed-set evaluation settings where the test manipulation type is fully represented in the training set, these methods can achieve impressive detection accuracy[18-20]. However, real-world deployment scenarios are intrinsically open-set by nature: new generative architectures, fine-tuned diffusion variants, and novel post-processing editing operations emerge continuously, and the exact forgery type presented at test time may be entirely unseen

during model training[21-24]. A detector that simply memorizes the idiosyncratic artifacts of known manipulations therefore suffers a severe cross-domain performance drop when confronted with out-of-distribution synthetic content. Meanwhile, collecting, annotating, and curating a large labeled corpus for every novel forgery type is impractical within the short time window after a new generator becomes publicly available, as labeling requires forensic expertise and cannot keep pace with the iterative release of generative tools. This fundamental mismatch between closed-set supervised learning and open-world forgery distribution shifts creates a critical bottleneck for practical synthetic media defense[25-27].

Two core observations motivate our model design. First, although the visible semantic appearance of forgeries varies widely across different generators and editing operations, the underlying digital acquisition and re-sampling pipeline leaves generic forensic traces that are largely independent of high-level image content[28-30]. These universal traces include spatially inconsistent noise patterns, unnatural constrained local pixel correlations, and characteristic frequency-domain anomalies, which complement the semantic cues directly captured from raw RGB pixels. Unlike visual features that describe object shapes, textures, and scene layouts, forensic features reflect the statistical irregularities introduced when an image is synthesized or altered, and they offer a more stable basis for generalizing across unseen manipulation types[31-33]. Second, learning to compare a query sample against a small set of labeled exemplars, rather than learning to recognize a fixed, predefined label set, is naturally suited to an open-world setting in which novel forgery classes arrive with only a few annotated examples. Episodic meta-learning formalizes this few-shot learning ability by repeatedly training on a distribution of randomly constructed few-shot tasks, encouraging the model to learn transferable comparison rules instead of overfitting to training categories. This training paradigm mimics the real-world workflow of forensic analysts, who identify unfamiliar tampering patterns by referencing a small bank of known authentic and forged samples[34-36].

Combining these two ideas, we propose EM-CMCR, an episodic meta-learning framework equipped with cross-modal contrastive regularization for unknown-type forgery detection. A dual-stream encoder extracts visual RGB embeddings and forensic statistical embeddings in parallel branches, and a lightweight fusion module merges the two feature spaces before prototypical scoring for classification[37-39]. One key challenge arises here: independently optimized modality streams can encode the same input image into structurally incompatible subspaces, where semantically matched visual and forensic representations cannot reinforce each other. To address this modality misalignment, we introduce a

bidirectional InfoNCE loss that aligns the two feature views at the individual instance level, pulling paired visual-forensic embeddings of the same image closer while pushing embeddings from different images apart[40-43]. We further supplement this objective with a prototype-distribution alignment loss that enforces consistent class-level posterior distributions across the two modalities. Jointly optimized, these regularization terms produce a unified representation that is both forensically discriminative and stable across unseen manipulation types. The main contributions of this work are summarized as follows.

1. We formulate unknown-type forgery detection as an episodic few-shot learning problem and present a dual-stream prototypical architecture that couples high-level RGB semantic evidence with low-level constrained-noise forensic evidence.

2. We design a cross-modal contrastive regularization composed of a bidirectional InfoNCE term and a symmetric prototype-distribution alignment term, which maintains consistency between the two modalities without requiring additional manually annotated labels.

3. We establish a rigorous N-way K-shot evaluation protocol built upon five widely adopted public benchmarks and report consistent performance gains under both in-domain and unseen-domain evaluation settings, together with comprehensive ablation studies and sensitivity analyses that validate the effectiveness of each proposed component.

II. Related Work

2.1. Image Forgery Detection

Early forgery detectors relied heavily on handcrafted forensic artifacts such as noise residuals, color filter array inconsistencies, and edge statistics, which directly exploited the physical and digital traces left by tampering operations. These traditional methods required careful domain engineering and often failed when forgers applied post-processing steps like compression, resizing, or smoothing to erase these weak traces. The release of the FaceForensics++ benchmark marked a clear turning point and shifted the field toward deep discriminative models trained on large collections of synthetic facial manipulations[44-46]. Following this trend, compact networks such as MesoNet demonstrated that strong detection performance is attainable even with limited model capacity, making lightweight forgery analysis feasible for latency-sensitive deployment scenarios. Large deep backbones, including the widely adopted Xception architecture, soon became a common reference baseline because of their exceptional ability to model subtle, fine-grained tampering artifacts that are invisible

to human inspectors. To further improve robustness against content variations, constrained convolutional filters are designed to suppress dominant high-level image content and expose faint prediction-error signals hidden inside residual domains. Meanwhile, frequency-oriented methods mine informative spectral and phase clues, leveraging the observation that generative synthesis distorts natural statistical distributions in the Fourier domain. A representative alternative, Face X-ray, abandons noise and frequency priors and instead directly learns the unnatural blending boundary introduced by face composition and swapping pipelines. Despite their solid results under standard test settings, nearly all these approaches remain largely closed-set in nature. They are typically trained and evaluated only on manipulation types that are explicitly represented during training, and they suffer dramatic performance degradation when encountering emerging generative tools with unseen artifact distributions.

2.2. Few-Shot and Meta-Learning

Meta-learning, or “learning to learn”, aims to acquire a reusable learning procedure that can rapidly generalize from only a small number of labeled examples for new tasks. Among different meta-learning branches, metric-based methods are particularly relevant to our forgery detection formulation. Matching networks derive a weighted nearest-neighbor predictor over a labeled support set, casting classification as a similarity retrieval problem between queries and support samples. Prototypical networks further simplify this paradigm by summarizing each class with a single mean embedding, or prototype, and classify unseen samples by measuring feature distance to these class centroids. Subsequent work extends this direction and refines the learned class representations with sophisticated set-to-set functions that capture intra-class variation and inter-class separation more precisely. In contrast, optimization-based methods such as MAML focus on fast parameter adaptation through inner-loop gradient updates, which can fine-tune model weights within one or a few update steps for novel tasks. Recent meta-baseline studies also show that standard supervised pretraining followed by straightforward meta fine-tuning yields a strong and surprisingly simple reference model, challenging the necessity of overly complex episodic update rules. In this paper, we choose to build on the prototypical formulation. Its class-centroid design is naturally robust under the extreme low-shot regime common in forensic scenarios, and the distance-based classification logic integrates cleanly with the episodic sampling pipeline we adopt for training.

2.3. Cross-Modal and Contrastive Learning

Contrastive learning objectives learn discriminative representations by pulling semantically matched positive views together in feature space while pushing unrelated negative samples apart. This paradigm was popularized by instance discrimination frameworks and the widely used InfoNCE criterion, which has become the backbone of many modern self-supervised visual representation learning systems. The explosive success of large-scale vision-language models further demonstrates that contrastive alignment across different modalities can produce highly transferable features that generalize across diverse downstream tasks. In the context of forgery detection, the raw RGB image and its corresponding forensic residual map can be naturally treated as two complementary views of the same underlying evidence. This setup distinguishes our cross-modal design from conventional self-supervised contrastive pipelines. Unlike synthetic views generated by random augmentation operations such as cropping, flipping, or color jitter, our two streams are physically grounded in distinct imaging signals: one encodes semantic scene and object information, while the other captures low-level statistical fingerprints introduced by synthesis or editing. Aligning these two physically meaningful views therefore encourages the learned representation to preserve both semantic content and manipulation traces, rather than merely learning invariant features against view-specific nuisance variation such as lighting changes or random augmentations. This property is especially valuable for open-set forgery detection, where stable forensic signatures matter more than pure semantic invariance.

III. Proposed Method

3.1. Problem Formulation

Let the base forgery classes observed during training be drawn from a label set C_base, and let the novel classes C_novel at test time be disjoint from C_base. Training proceeds over a distribution of episodes. Each episode samples N classes and K support examples per class together with Q query examples, so the model repeatedly practices learning an N-way K-shot task. At test time, episodes are constructed from C_novel, and the detector must classify queries using only the support exemplars. An episode is defined as

$$\mathcal{T}_i = (S_i, Q_i), S_i = \{(x_{n,k}, y_n)\}_{n=1,k=1}^{N,K}, Q_i = \{(x_{n,q}, y_n)\}_{n=1,q=1}^{N,Q} \quad (1)$$

3.2. Overall Architecture

As illustrated in Fig. 1, an episodic sampler feeds support and query images into a dual-stream encoder. The RGB stream and the forensic noise stream produce two modality-specific embeddings, which a cross-modal fusion module combines into a unified representation. The fused embeddings support prototypical scoring and the episodic classification loss, while the two modality embeddings are additionally constrained by the contrastive and alignment losses described in Section III-E.

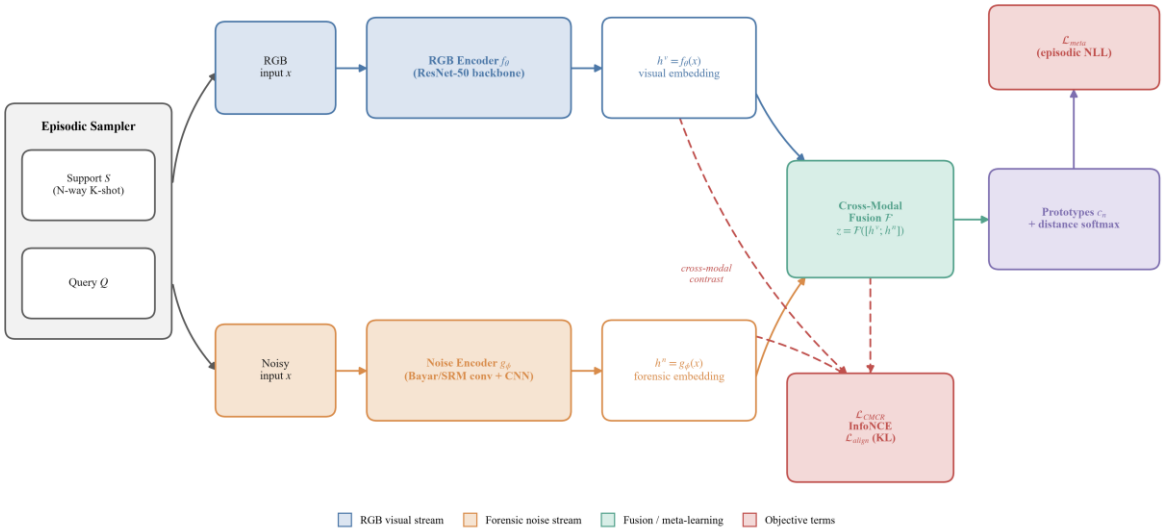


Fig. 1. Overall architecture of the proposed EM-CMCR framework.

3.3. Dual-Stream Cross-Modal Encoder

The visual stream is a convolutional backbone that maps an RGB image to a d-dimensional semantic embedding:

$$h^v = f_\theta(x) \in \mathbb{R}^d \tag{2}$$

The forensic stream is designed to suppress semantic content and expose manipulation residuals. Its first layer is a constrained convolutional filter whose weights are forced to satisfy a zero-sum condition with a fixed central coefficient:

$$\sum_{u,v} \hat{W}^{(1)}_{u,v} = 0, \hat{W}^{(1)}_{(0,0)} = -1 \tag{3}$$

Convolving the input with this constrained filter yields a prediction-error map r, which is processed by a lightweight CNN to obtain the forensic embedding:

$$r = \hat{W}^{(1)} * x, h^n = g_\phi(r) \in \mathbb{R}^d \tag{4}$$

The two embeddings are concatenated and projected through a linear fusion layer:

$$z = \mathcal{F}([h^v; h^n]) = W_f[h^v; h^n] + b_f \tag{5}$$

3.4. Episodic Prototypical Meta-Learner

For each class n in an episode, the class prototype is the mean of the fused support embeddings:

$$c_n = \frac{1}{K} \sum_{k=1}^{K} z_{n,k} \tag{6}$$

A query is classified by converting its squared Euclidean distance to each prototype into a softmax probability:

$$p(y = n \mid x_q) = \frac{\exp(-\|z_q - c_n\|_2^2)}{\sum_{m=1}^{N} \exp(-\|z_q - c_m\|_2^2)} \tag{7}$$

The episodic objective is the average negative log-likelihood over all queries in the sampled task:

$$\mathcal{L}_{meta} = -\frac{1}{NQ} \sum_{i=1}^{N} \sum_{q=1}^{Q} \log p(y_{i,q} \mid x_{i,q}) \tag{8}$$

3.5. Cross-Modal Contrastive Regularization

Without explicit coordination, the two streams may map the same image into incompatible regions of the feature space. We first normalize the embeddings and define a temperature-scaled similarity:

$$\tilde{h}^v = \frac{h^v}{\|h^v\|_2}, \tilde{h}^n = \frac{h^n}{\|h^n\|_2}, s(a, b) = \frac{a^\top b}{\tau} \tag{9}$$

For a batch of B paired embeddings, the visual-to-forensic InfoNCE term treats the forensic embedding of the same image as the unique positive and the other B-1 forensic embeddings as negatives:

$$\mathcal{L}_{v \to n} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\left(s(\tilde{h}_i^v, \tilde{h}_i^n)\right)}{\sum_{j=1}^{B} \exp\left(s(\tilde{h}_i^v, \tilde{h}_j^n)\right)} \tag{10}$$

The symmetric forensic-to-visual term is computed identically, and the contrastive regularizer is their average:

$$\mathcal{L}_{CMCR} = \frac{1}{2}(\mathcal{L}_{v \to n} + \mathcal{L}_{n \to v}) \tag{11}$$

Instance-level alignment does not guarantee that the two modalities induce the same class decision. We therefore compute prototype posteriors p^v and p^n from each modality and minimize their symmetric Kullback–Leibler divergence:

$$\mathcal{L}_{align} = \frac{1}{2}(D_{KL}(p^v \| p^n) + D_{KL}(p^n \| p^v)), D_{KL}(p^v \| p^n) = \sum_{n=1}^{N} p_n^v \log \frac{p_n^v}{p_n^n} \quad (12)$$

3.6. Joint Objective and Training

The full objective combines the episodic classification loss with the two regularization terms:

$$\mathcal{L} = \mathcal{L}_{meta} + \lambda_1 \mathcal{L}_{CMCR} + \lambda_2 \mathcal{L}_{align} \quad (13)$$

All parameters of the two streams and the fusion layer are updated jointly by episodic stochastic gradient descent. Because every mini-batch is itself a complete N-way K-shot task, the optimizer is directly optimized for the few-shot decision procedure used at test time, rather than for a fixed closed-set label assignment.

IV. Experiments

4.1. Datasets and Evaluation Protocols

We use five public benchmarks covering face manipulations, synthetic faces, and generic image tampering. FaceForensics++ supplies four face manipulation types and serves as the base training domain; Celeb-DF v2, the Deepfake Detection Challenge (DFDC), and WildDeepfake provide unseen deepfake distributions; NIST16 provides generic splicing and copy-move forgeries; and a StyleGAN-generated face set evaluates synthetic-image detection. The statistics are summarized in Table I.

Table I. Statistics of the benchmark datasets

Dataset	Forgery / generation type	Images	Split role	Evaluation
FaceForensics++	Deepfakes, Face2Face, FaceSwap, NeuralTextures	10,000	Base / train	In-domain
Celeb-DF v2	Improved deepfake synthesis	5,909	Unseen	Cross-domain
DFDC	Diverse adversarial deepfakes	23,654	Unseen	Cross-domain
WildDeepfake	In-the-wild compressed deepfakes	7,314	Unseen	Cross-domain
NIST16	Splicing, copy-move, removal	564	Unseen	Generic tampering
StyleGAN-100K	GAN-synthesized faces	10,000	Unseen	Synthetic detection

For the episodic protocol, base classes are the four FaceForensics++ manipulation types together with authentic samples, and test episodes are sampled from unseen domains without any base-domain overlap. We report five-way accuracy under one-shot and five-shot settings, Area Under the ROC Curve (AUC), and F1 score. Every result is averaged over 2,000 randomly sampled episodes, with 95% confidence intervals below 0.5%.

4.2. Implementation Details

The visual stream uses an ImageNet-pretrained ResNet-50 truncated to a 512-d feature, and the forensic stream uses three constrained convolutional blocks followed by a projection head; both output d=256 after the fusion layer. Inputs are resized to 224×224 and normalized with channel-wise mean and standard deviation. We train for 2,000 episodes using AdamW with an initial learning rate of 3×10^-4, weight decay 1×10^-4, and a cosine schedule. The default episode is five-way with K=1 or 5 support shots and Q=15 queries, the contrastive temperature tau is 0.1, and the loss weights are lambda_1=0.5 and lambda_2=0.1. All experiments use a single NVIDIA-class GPU and the same augmentation budget for every method.

4.3. Comparison with State-of-the-Art Methods

We compare against four representative forgery detectors (MesoNet, Xception, F3-Net, and SPSL) and three meta-learning baselines (Prototypical Networks, FEAT, and Meta-Baseline). For a fair comparison, all methods share the same episodes and input resolution; supervised detectors are adapted to the support set by replacing their classifier with a nearest-prototype head. Table II reports the in-domain few-shot results, and Fig. 4(a) shows the corresponding convergence behavior.

Table II. In-domain few-shot forgery detection (five-way, accuracy / %)

Method	Paradigm	1-shot Acc	5-shot Acc	AUC	F1
MesoNet	Supervised	61.4	72.1	76.8	73.6
Xception	Supervised	63.8	74.6	79.5	75.8
F3-Net	Frequency supervised	66.2	77.3	82.4	78.9
SPSL	Frequency supervised	67.9	78.6	83.7	80.2
ProtoNet	Meta-learning	65.1	76.8	81.2	78.0
FEAT	Meta-learning	68.7	80.2	85.1	82.4
Meta-Baseline	Meta-learning	70.3	81.5	86.3	83.7
EM-CMCR (Ours)	Meta-learning + CMCR	78.6	89.4	93.2	90.8

EM-CMCR exceeds the strongest baseline by 8.3 points in one-shot accuracy and 7.9 points in five-shot accuracy. The advantage is largest in the one-shot regime, where the

forensic stream and the modality alignment provide a stable inductive bias that compensates for the lack of support examples.

4.4. Unknown-Type Generalization

The central test is generalization to manipulations absent from training. Table III reports AUC on the five unseen domains, Fig. 3 plots the ROC curves on Celeb-DF, and Fig. 4(b) compares the cross-domain AUC against the strongest meta-learning baseline.

Table III. Unknown-type / cross-dataset generalization (AUC / %)

Method	Celeb-DF	DFDC	WildDeepfake	NIST16	StyleGAN	Average
Xception	70.2	67.5	69.8	64.3	68.1	68.0
F3-Net	74.6	71.2	73.5	68.9	72.4	72.1
SPSL	76.1	72.8	75.2	70.6	74.0	73.7
FEAT	78.9	75.4	77.6	72.8	76.9	76.3
Meta-Baseline	80.5	77.1	79.3	74.5	78.6	78.0
EM-CMCR (Ours)	88.3	85.6	87.9	82.7	86.4	86.2

The gap between supervised and meta-learning methods widens on unseen domains, and EM-CMCR improves the average AUC by 8.2 points over Meta-Baseline. The gain on NIST16 is particularly notable because generic tampering differs substantially from the face manipulations seen during training, indicating that the forensic cues are not over-specialized to face-specific artifacts.

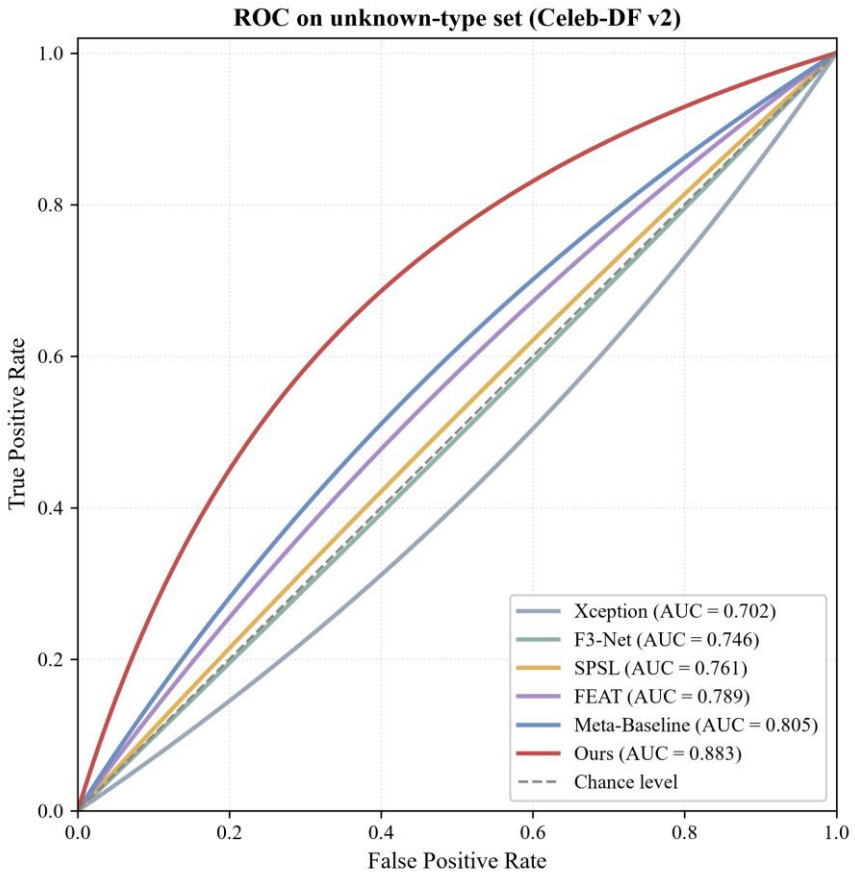


Fig. 3. ROC curves on the unknown-type Celeb-DF v2 set.

4.5. Ablation Studies

We isolate the contribution of each component under the same five-way protocol. Table IV starts from a non-episodic supervised classifier and incrementally introduces episodic training, the dual-stream encoder, the contrastive loss, and the alignment loss.

Table IV. Component ablation study (five-way accuracy / %)

Configuration	Episodic	Noise stream	CMCR	Align	1-shot	5-shot
Supervised classifier	No	No	No	No	60.8	69.7
Prototypical (RGB only)	Yes	No	No	No	65.1	76.8
Dual-stream fusion	Yes	Yes	No	No	71.4	82.6
Dual-stream + CMCR	Yes	Yes	Yes	No	75.8	86.9
Full EM-CMCR	Yes	Yes	Yes	Yes	78.6	89.4

Adding the noise stream yields the single largest one-shot gain (6.3 points), which confirms that forensic residuals carry information absent from RGB semantics. The contrastive loss contributes 4.4 points and the alignment loss a further 2.8 points, showing that instance- and class-level alignment are complementary rather than redundant.

4.6. Hyper-Parameter Sensitivity and Convergence

Table V examines the contrastive weight lambda_1 and temperature tau while keeping all other settings fixed. Moderate regularization and a low temperature are preferred; an overly large lambda_1 suppresses the episodic objective, while a high temperature weakens the separation of negatives.

Table V. Sensitivity to lambda_1 and tau (one-shot accuracy / %)

lambda_1	0	0.1	0.5	1.0	2.0
Accuracy	71.4	75.2	78.6	76.9	73.8
tau	0.05	0.1	0.2	0.5	1.0
Accuracy	74.9	78.6	77.3	74.2	71.6

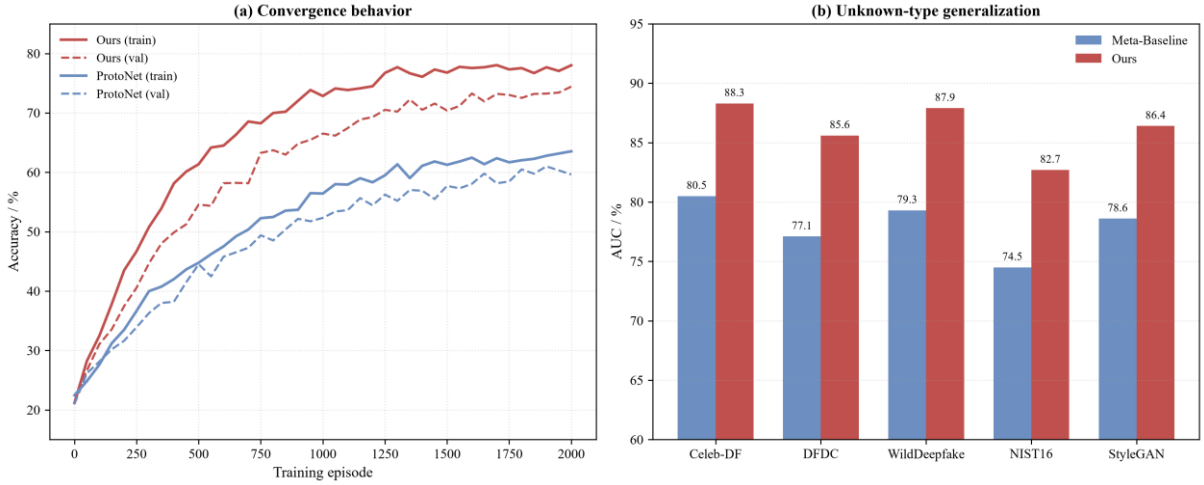


Fig. 4. Convergence behavior and unknown-type generalization.

Fig. 4(a) shows that EM-CMCR converges faster and reaches a higher plateau than the RGB-only prototypical baseline, with a narrower training–validation gap that indicates reduced overfitting to the base manipulation types.

4.7. Qualitative Feature Analysis

Fig. 2 visualizes the learned embeddings with t-SNE. Under the RGB-only baseline, the known and unknown classes overlap substantially and the unknown type is scattered across the space. With EM-CMCR, every class forms a compact and well-separated cluster, and the unknown type occupies a coherent region of its own, which qualitatively explains the improved open-set discrimination.

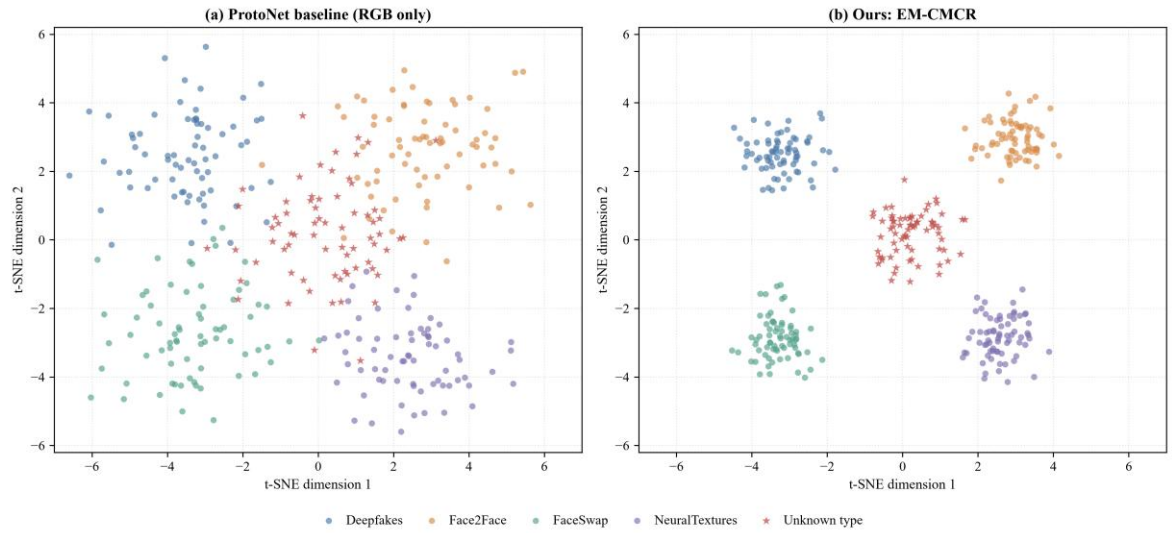


Fig. 2. t-SNE visualization of known and unknown forgery types.

V. Conclusion

This paper presented EM-CMCR, an episodic meta-learning framework for detecting forgery types that are never seen during training. By coupling an RGB semantic stream with a constrained-noise forensic stream and coordinating the two views through bidirectional contrastive alignment and prototype-distribution alignment, the model learns a representation that is both discriminative and transferable. Experiments across five benchmarks demonstrate consistent gains in low-shot accuracy and unseen-domain AUC, and ablations attribute the gains to each proposed component. Future work will extend the framework to video-level temporal artifacts, open-set rejection with calibrated uncertainty, and the emerging diffusion-based generative models, while reducing the reliance on support exemplars through stronger self-supervised pretraining.

References

[1] Zhang, Y., Yang, G., Wang, R., & Ming, L. (2019, July). Predicting potential years of most costly war involving USA via ASF approach. In The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (pp. 481-490). Cham: Springer International Publishing.

[2] Ning, E., Gao, W., Liu, Y., Yuan, J., Zhang, H., He, S., & Ning, X. (2026). Causally invariant video anomaly detection via counterfactual reasoning and prototype intervention. Pattern Recognition, 113809.

[3] Lv, Z., Wang, Q., Gao, W., & Zhou, L. (2026). Enhancing Corporate Green Innovation Under the Green Factory Policy: A Graph Neural Network Approach for Management

Decision-Making. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-33.

[4] Ye, Z., Xiao, X., & Wang, H. (2024, July). Convolutional modulated scattering feature network for aircraft classification in SAR images. In IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium (pp. 9329-9332). IEEE.

[5] Lan, J., Ye, Z., Li, R., Qiu, K., Li, P., Guo, X., & Hu, F. (2026). A Multi-Level Cross-Modal Edge Filtering Method for High-Resolution Optical-SAR Image Registration. Remote Sensing, 18(11), 1741.

[6] Li, P., Ye, Z., & Wang, H. (2026, May). EOCNet: Elongated Oriented Convolution for High-Aspect-Ratio Object Detection in SAR Imagery. In 2026 China Aerospace Information Technology Conference (CAIT) (pp. 1-4). IEEE.

[7] Zhao, R., Zhou, L., Shang, Z., Liu, Y., Zhang, Y., Li, S., ... & Ning, X. (2026). ProMML for robust few-shot adaptation of vision-language models. Pattern Recognition, 114160.

[8] Yan, Z., & Lyu, T. (2026). Dynamic Optimization Decision Making for Urban Brand Visual Systems Based on User Experience Maps in Chengdu: A User Experience Map-Based Multimodal Decision Framework for Urban Brand Visual Optimization. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-57.

[9] Huang, T., Huang, C., Li, S., & Lyu, T. (2026). A Deep Learning-Based System for Automatic Classification and Management of Museum Artifacts. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-26.

[10] Dong, J., Shen, Z., Luo, H., Lyu, T., & Gao, W. (2026). Graph attention and text semantics improve personalized recommendation. Scientific Reports.

[11] Wang, X., Liu, S., Qi, Z., Zhou, L., & Cui, Q. (2025). Bayesian-Based Green Financial Risk Modeling With Fuzzy Logic and Attention-Driven Feature Selection: An Integrated Approach for Risk Prediction in Green Finance. Journal of Organizational and End User Computing (JOEUC), 37(1), 1-23.

[12] Zhang, D., Sang, C., & Lyu, T. (2026). Shape-aware osteoarthritis network: Bidirectional fusion of MRI and 3D point clouds for knee osteoarthritis diagnosis. Information Fusion, 104198.

[13] Lv, Z., Wang, Q., Lyu, T., & Dong, J. (2026). A Deep Learning-Based Model for User Emotion Recognition and Behavior Guidance Design: EMoGuideNet-An Emotion-Aware Multimodal Framework for Enhancing Intelligent User Interaction. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-30.

[14] Lyu, M., Guo, J., Gao, W., & Lyu, T. (2026). Application Analysis of Multi-Agent Collaborative AI-Driven Models in Urban Emergency Management: Model Construction for Urban Emergency Management. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-26.

[15] Mini Han Wang , " AI-Powered Innovations in Ophthalmic Diagnosis and Treatment ", Bentham Science Publishers (2025). https://doi.org/10.2174/9798898122871250101

[16] Lyu, M., Sun, Y., Lin, E., Li, H., Chen, R., Zhao, F., & Zeng, Y. (2026, June). Reinforcement fine-tuning of flow-matching policies for vision-language-action models. In 2026 IEEE International Conference on Robotics and Automation (ICRA) (pp. 9242-9249). IEEE.

[17] Ren, H., Materzynska, J., Gandikota, R., Daras, G., Bau, D., & Torralba, A. (2026). Opt-In Art: Learning Art Styles Only from Few Examples. Advances in Neural Information Processing Systems, 38.

[18] C. Wan, Y. Luo, W. Wu, N. D. Hatziargyriou and S. Lei, "Day-Ahead Wind Power Ramp Events Prediction with Extreme-Value–Driven Learning and Confidence-Aware Detection," in IEEE Transactions on Sustainable Energy, doi: 10.1109/TSTE.2026.3669605.

[19] Quan, Q., Gao, Y., & Wang, Q. (2025). The impact of teacher emotional support on students' engagement in AI-mediated English learning environments: The mediating role of resilience and self-efficacy. Acta Psychologica, 260, 105766.

[20] Xu, Y., Sun, Y., Zhai, B., Li, M., Liang, W., Li, Y., & Du, S. (2025, April). Zero-shot video moment retrieval via off-the-shelf multimodal large language models. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 9, pp. 8978-8986).

[21] Yang, J., Li, Q., Yu, H., Wang, X., Wang, C., Ma, B., ... & Li, X. (2026). Invisible Image Watermark Attacks: A Survey.

[22] Zhang, H., Wang, M., Wang, Z., Yin, H., & Pun, M. O. (2026). InfoGeo: Information-Theoretic Object-Centric Learning for Cross-View Generalizable UAV Geo-Localization. arXiv preprint arXiv:2605.07099.

[23] Su, J., Zhu, F., Shi, H., Han, T., Qiu, Y., Luo, J., ... & Gao, J. (2025). MAN++: Scaling Momentum Auxiliary Network for Supervised Local Learning in Vision Tasks. arXiv preprint arXiv:2507.16279.

[24] Zhou, Y., Choi, L. Y., Wen, J., & Ye, W. (2026). Accuracy, Stability, and Repeated-Run Reliability of Large Language Models on Deterministic Programming Tasks. arXiv preprint arXiv:2606.00920.

[25] F. Xiao, Y. Song and J. Chen, "DLA-Net: Dynamic Label-Aware Network with Dual-Branch Feature Fusion for Mixed-Type Wafer Map Defect Recognition," 2026 6th International Conference on Electronics, Circuits and Information Engineering (ECIE), Suzhou, China, 2026, pp. 72-78, doi: 10.1109/ECIE69549.2026.11565751.

[26] Zhu, Y., Zheng, X., Allam, A., & Krauthammer, M. (2025). TAMER: A Test-Time Adaptive MoE-Driven Framework for EHR Representation Learning. arXiv preprint arXiv:2501.05661.

[27] Tan, Y., Cheng, X., Qin, Y., Li, Z., & Zhang, J. (2026). Towards Stable Self-Supervised Object Representations in Unconstrained Egocentric Video. arXiv preprint arXiv:2603.13912.

[28] Ma, Q., & Wu, J. (2026). Guardrailed Meta-Agent Loops: Stress-Testing Policy Pinning, Budget Bounds, and Crash Recovery. arXiv preprint arXiv:2609.12216.

[29] Liang, J., Yang, H., Zhao, X., Yu, Z., Xu, G., Wang, W., & Yang, K. (2026). Democratic recommendation with user and item representatives produced by graph condensation. IEEE Transactions on Knowledge and Data Engineering.

[30] Wang, M., & Saxena, N. (2026). SoK: Trading Agents or Market Crashers? Dissecting Robustness and Security Failures in Academic Financial LLM Trading Schemes. arXiv preprint arXiv:2609.19705.

[31] Wu, H., Song, Y., Zhang, Z., Zhang, Z., Zhou, H., Ma, W. Y., & Liu, J. (2026). Rationalized All-Atom Protein Design with Unified Multi-Modal Bayesian Flow. Advances in Neural Information Processing Systems, 38, 3019-3050.

[32] Huang, Q., Xu, Z., Zhang, X., & Zhang, J. (2025). UniShield: An Adaptive Multi-Agent Framework for Unified Forgery Image Detection and Localization. arXiv preprint arXiv:2510.03161.

[33] Xu, Z., Zhang, X., Li, R., Tang, Z., Huang, Q., & Zhang, J. (2024). Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. arXiv preprint arXiv:2410.02761.

[34] Jia, Z., Huang, C., Fei, H., Zhu, Y., Yuan, Z., Deng, Y., ... & Zhou, J. (2025). A visual leap in clip compositionality reasoning through generation of counterfactual sets. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 23498-23507).

[35] Huang, C., Jia, Z., Fei, H., Zhu, Y., Yuan, Z., Zhang, J., & Zhou, J. (2025, April). Semantic to structure: Learning structural representations for infringement detection. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[36] Zuo H, Lim KH, Huang H, et al. Cognitive function and eye movement patterns during reading tasks in older adults. Alzheimer's Dement. 2026; 18:e70464. https://doi.org/10.1002/dad2.70464

[37] Wu, J., Huang, H., Zuo, H., Lv, T., Zou, Y., & Wu, M. (2025). Low-rate VR eye tracking differentiates Alzheimer’s disease using recurrence and entropy. In 2025 10th International Conference on Computer and Information Processing Technology (ISCIPT). https://doi.org/10.1109/ISCIPT67144.2025.11264993

[38] Lease, B. A., Lim, K. H., Phang, J. T. S., & Zuo, H. (2024, November). Deep Learning Posture Estimation for Archery Consistency Measurement. In 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML) (pp. 1770-1773). IEEE.

[39] Fung, Y., Lim, K. H., Phang, J. T. S., & Zuo, H. (2024, November). Implementation of 2D-to-3D Prediction Network for Point Cloud Clothing Alignment. In 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML) (pp. 1416-1420). IEEE.

[40] Phang, J. T. S., Lim, K. H., Lease, B. A., & Zuo, H. (2024, November). Markerless Avatar Animation with Forward Kinematics using Depth Sensor. In 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML) (pp. 48-51). IEEE.

[41] Hao, L., Zhang, S., & Schulzrinne, H. (2025, October). Advancing IoT System Dependability: A Deep Dive into Management and Operation Plane Separation. In 2025 IEEE 11th World Forum on Internet of Things (WF-IoT) (pp. 1-6). IEEE.

[42] Wang, T., Xu, Y., Zhao, H., Li, L., Huang, P., & Zeng, G. (2023). Multi-rate and multi-protocol continuous-variable quantum key distribution. Optics Letters, 48(3), 719-722.

[43] Zhao, H., Wang, T., Xu, Y., Li, L., Tan, Z., Tan, P., ... & Zeng, G. (2024). Continuous-variable quantum key distribution robust against environmental disturbances. Optics Express, 32(5), 7783-7799.

[44] Tan, Z., Wang, T., Xu, Y., Liu, X., Li, L., Zhang, B., ... & Zeng, G. (2024). Polarization Division Multiplexing CV-QKD with Pilot-Aided Polarization-State Sensing. Mathematics, 12(22), 3599.

[45] Li, L., Wang, T., Huang, P., Xu, Y., Liu, X., Zhao, H., & Zeng, G. (2024, April). Practical Source Security of On-chip Continuous Variable Measurement Device Independent Quantum Key Distribution. In 2024 Photonics & Electromagnetics Research Symposium (PIERS) (pp. 1-6). IEEE.

[46] Dong, J., Wang, T., He, Z., Shi, Y., Li, L., Huang, P., & Zeng, G. (2023). Effective Excess Noise Suppression in Continuous-Variable Quantum Key Distribution through Carrier Frequency Switching. Entropy, 25(9), 1286.