Open-World Deepfake Attribution via Confidence-Aware Asymmetric Learning

Haiyang Zheng[1], **Nan Pu**[1,2*], **Wenjing Li**[2*], **Teng Long**[1], **Nicu Sebe**[1], **Zhun Zhong**[2]

[1]University of Trento,
[2]Hefei University of Technology
{haiyang.zheng, nan.pu}@unitn.it

Abstract

The proliferation of synthetic facial imagery has intensified the need for robust Open-World DeepFake Attribution (OW-DFA), which aims to attribute both known and unknown forgeries using labeled data for known types and unlabeled data containing a mixture of known and novel types. However, existing OW-DFA methods face two critical limitations: 1) A **confidence skew** that leads to unreliable pseudo-labels for novel forgeries, resulting in biased training. 2) An **unrealistic assumption** that the number of unknown forgery types is known *a priori*. To address these challenges, we propose a Confidence-Aware Asymmetric Learning (**CAL**) framework, which adaptively balances model confidence across known and novel forgery types. CAL mainly consists of two components: Confidence-Aware Consistency Regularization (**CCR**) and Asymmetric Confidence Reinforcement (**ACR**). *CCR* mitigates pseudo-label bias by dynamically scaling sample losses based on normalized confiden2ce, gradually shifting the training focus from high- to low-confidence samples. *ACR* complements this by separately calibrating confidence for known and novel classes through selective learning on high-confidence samples, guided by their confidence gap. Together, *CCR* and *ACR* form a mutually reinforcing loop that significantly improves the model's OW-DFA performance. Moreover, we introduce a Dynamic Prototype Pruning (**DPP**) strategy that automatically estimates the number of novel forgery types in a coarse-to-fine manner, removing the need for unrealistic prior assumptions and enhancing the scalability of our methods to real-world OW-DFA scenarios. Extensive experiments on the standard OW-DFA benchmark and a newly extended benchmark incorporating advanced manipulations demonstrate that *CAL* consistently outperforms previous methods, achieving new state-of-the-art performance on both known and novel forgery attribution.

Code — https://haiyangzheng.github.io/OWDFA-CAL
Extended version — https://arxiv.org/pdf/2512.12667

1 Introduction

The rapid advancement of image generation techniques, particularly diffusion models (Rombach et al. 2022a; CiteDrive 2024) and autoregressive modeling (Tian et al. 2024), has led to a surge in synthetic facial imagery on social media platforms. While such content often serves entertainment purposes, it raises serious concerns regarding identity misuse and the spread of misinformation. To mitigate these risks, DeepFake Attribution (DFA) (Yang et al. 2022; Yu et al. 2021; Guarnera et al. 2022) has been proposed not only to determine whether the content is fake, but also to identify the specific model architectures responsible for forgery generation. However, existing methods mainly focus on GAN-generated images, neglecting more advanced and realistic manipulations such as identity swaps (Xu et al. 2022b,a). Moreover, they often assume a closed-world setting, in which training and testing categories are shared—an assumption that rarely holds in open-world scenarios.

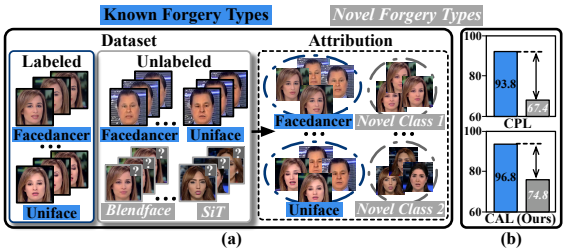

Figure 1: (a) Schema of the Open-World Deepfake Attribution. (b) Our CAL reduces the performance gap compared to the state-of-the-art method CPL (Sun et al. 2023), showing average results across all evaluation protocols.

To address these limitations, Sun *et al.* (Sun et al. 2023) introduce a new Open-World DFA (OW-DFA) task, as shown in Fig. 1(a), which requires models to identify both known and novel forgery types in an unlabeled set of manipulated face images by transferring knowledge from a labeled set containing only known forgeries. Existing OW-DFA methods (Sun et al. 2023, 2025; Zheng et al. 2025) typically adopt pseudo-labeling strategies to exploit unlabeled data and have achieved promising attribution performance. However, these methods exhibit a significant performance gap between known and novel forgery types. Through an in-depth analysis (detailed in Sec. 3) of this observation, we argue that this performance gap results from a ***confidence skew***: the model assigns lower confidence scores to predictions on novel forgery types, resulting in unreliable pseudo-

*Corresponding author.

labels. These inaccurate labels misguide the training objective, reinforcing the bias through a negative feedback loop and further amplifying the skew. Additionally, existing approaches assume that the number of forgery types is known *a priori*. However, such an assumption is unrealistic in real-world open-world scenarios, where generative forgery models evolve continuously. The number of forgery types in newly collected unlabeled data is inherently unknown and potentially unbounded, rendering these methods ill-suited for deployment in dynamic real-world environments.

To address these drawbacks, we propose a novel **Confidence-Aware Asymmetric Learning (CAL)** framework that adaptively mitigates the model's confidence imbalance between known and novel forgery types throughout training, consistently improving attribution accuracy on both seen and unseen forgery types. Our CAL framework consists of two key components: *Confidence-Aware Consistency Regularization (CCR)* and *Asymmetric Confidence Reinforcement (ACR)*. CCR effectively rectifies biased pseudo-label learning on unlabeled data by adaptively adjusting the regularization strength in a threshold-free manner. Specifically, during the early training stages, CCR mitigates the effect of noisy supervision by down-weighting low-confidence samples. This is achieved by scaling their loss contributions with normalized confidence scores. As training progresses, the model gradually shifts focus: it reduces the weight assigned to high-confidence samples and emphasizes learning from low-confidence ones, which predominantly correspond to novel forgery types. This adaptive strategy enables more stable and discriminative representation learning throughout the training process. Complementarily, ACR explicitly encourages the model to generate high-confidence predictions separately for known and novel categories via an asymmetric learning strategy. Specifically, we introduce a coefficient based on the model's confidence gap between known and novel categories to personalize the selection of high-confidence samples for known and novel forgery types. By enforcing learning on these high-confidence samples, ACR facilitates more effective consistency regularization in CCR, creating a mutually beneficial feedback loop. As shown in Fig. 1(b) and Fig. 2, our CAL largely improves the model's prediction confidence and accuracy on all classes. In addition, we introduce a *Dynamic Prototype Pruning (DPP)* strategy to estimate the number of novel forgery types with negligible computational overhead. We design a coarse-to-fine pruning mechanism to dynamically merge low-usage and redundant prototypes. This enables CAL to scale effectively to real-world OW-DFA deployments, where the number of attack types is unknown.

Overall, our contributions are summarized as follows:

- We identify the importance of the confidence skew issue, which significantly degrades the overall performance of existing OW-DFA methods.
- We propose a new CAL framework that effectively reduces the confidence skew issue and promotes OW-DFA models toward unbiased learning.
- We design a novel DPP strategy that can automatically estimate the number of novel forgery types during training with negligible overhead.
- We extend the existing OW-DFA benchmark by incorporating advanced diffusion-based forgeries to build a challenging yet practical OW-DFA-40 benchmark. Extensive experiments show that our CAL achieves state-of-the-art performance on both benchmarks.

2 Related Work

Open-World DeepFake Attribution (OW-DFA). Motivated by growing concerns over privacy protection, DeepFake Attribution (DFA) (Yang et al. 2022; Yu et al. 2021; Guarnera et al. 2022) aims to detect manipulated content while simultaneously identifying the specific generative model architecture responsible for its creation. However, most existing GAN attribution methods (Yang et al. 2022; Yu et al. 2021; Guarnera et al. 2022) rely on model-specific fingerprints and operate under a *closed-world assumption*, where the training and testing distributions are aligned. This assumption often fails in real-world applications, where novel manipulation techniques continue to emerge. To overcome this limitation, the OW-DFA task was introduced in CPL (Sun et al. 2023) as a more realistic extension of DFA. The goal is to attribute each manipulated face in the unlabeled set, regardless of whether it originates from a known or novel generative model. Building upon CPL, CDAL (Zheng et al. 2025) leverages causal inference and counterfactual contrast to eliminate confounding biases in attention, thus improving the model's ability to identify discriminative generation patterns for robust attribution. *In this work, we follow the OW-DFA setting and further extend the existing benchmark by incorporating diffusion-based forgeries in response to the rapid evolution of generative techniques. Moreover, we propose a more challenging and practical evaluation protocol to comprehensively assess OW-DFA methods under realistic conditions.*

Generalized Category Discovery (GCD) (Han, Vedaldi, and Zisserman 2019) aims to cluster unlabeled data containing both known and unknown categories by leveraging prior knowledge from labeled categories. This setting has received increasing attention with methods such as GCD (Vaze et al. 2022), which leverages supervised and self-supervised contrastive learning on pre-trained DINO (Caron et al. 2021) features; SimGCD (Wen, Zhao, and Qi 2023), which adopts a parametric classifier with self-distillation; LegoGCD (Cao et al. 2024b), which introduces view-consistency regularization; ProtoGCD (Ma et al. 2025), which formulates prototype learning jointly for category discovery and outlier detection; and PALGCD (Wang, Zhong, and Gong 2025) introduces a prior-constrained association learning framework that integrates non-parametric semantic prototypes with a parametric classifier to effectively uncover the semantic structure of unlabeled data. **Open-World Semi-Supervised Learning (OWSSL)** shares an essentially identical problem setting with GCD but was independently introduced in the context of open-set recognition (Cao, Brbic, and Leskovec 2021). OWSSL methods follow similar principles, such as prototype-based alignment (Sun and Li 2022), pairwise similarity learning (Rizve et al. 2022), hierarchical semantic de-

composition (Wang et al. 2023), contrastive clustering (Ye et al. 2014), and conditional self-labeling (Niu et al. 2024). *Despite the shared assumptions, GCD and OWSSL tasks focus on understanding the global semantics of natural images, while OW-DFA requires attention to fine-grained facial details and forgery traces introduced by different deepfake methods. Moreover, existing methods and category estimation strategies are designed based on strong pre-trained models, making them less effective for the OW-DFA task.*

3 Confidence Skew in OW-DFA

Problem Setup. We consider open-world deepfake attribution with a labeled set generated by *known* methods, $\mathcal{D}_L = \{(\mathbf{x}_i^l, y_i^l) \in \mathcal{X} \times \mathcal{Y}_L\}_{i=1}^N$, and an unlabeled set generated by both known and *novel* methods, $\mathcal{D}_U = \{\mathbf{x}_i^u \in \mathcal{X}\}_{i=1}^M$. Here N and M are the numbers of samples in $\mathcal{D}_L$ and $\mathcal{D}_U$, respectively. Let $\mathcal{Y}_U$ be the label space covered by $\mathcal{D}_U$, with $\mathcal{Y}_L \subset \mathcal{Y}_U$. We define $K_L = |\mathcal{Y}_L|$ as the number of known classes and $K_U = |\mathcal{Y}_U|$ as the total number of classes. If K_U is known in advance, it can be directly utilized during training; otherwise, it must be estimated in training.

Definition and Observations. We define *confidence skew* as a consistent confidence gap between known and novel classes, leading to unreliable pseudo-labels for novel samples and reinforcing the skew. We make two key observations: **(O1)** Novel class samples under CPL show lower average confidence and a flatter distribution than known class samples (Fig. 2). **(O2)** CPL yields a higher ratio of pseudo-label noise in the low-confidence range $[0, 0.5)$ for novel-class samples compared to ours (Fig. 3).

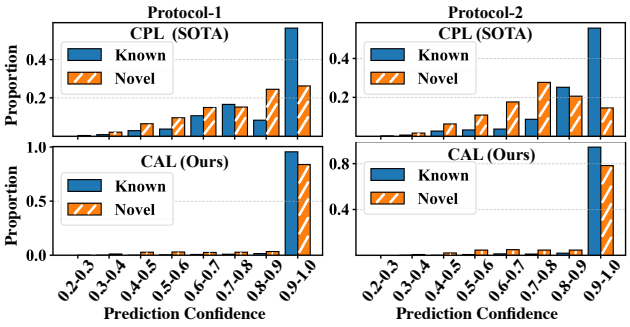

Figure 2: Distribution of sample confidence for known vs. novel classes.

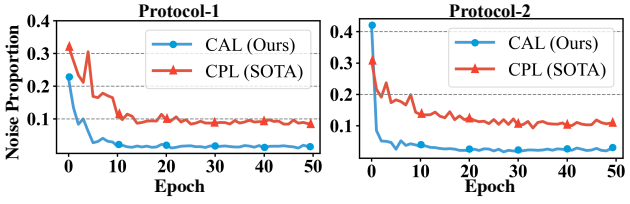

Figure 3: Proportion of pseudo-label noise in the low-confidence range $[0, 0.5)$ among all novel class samples.

Mechanism. *(1) Supervision imbalance ⇒ initial confidence skew.* Since the labeled set $\mathcal{D}_L$ only covers the known classes $\mathcal{Y}_L$, early optimization predominantly updates representations toward known classes, thereby biasing the feature space. *(2) Gumbel-Softmax on low-confidence samples ⇒ noisy pseudo-labels.* CPL generates pseudo-labels using Gumbel-Softmax sampling (Jang, Gu, and Poole 2016). When the predicted logits are flat, the sampling becomes highly sensitive to Gumbel noise, leading to unstable and noisy pseudo-labels. For novel forgery samples, which typically exhibit low confidence and flat logit distributions, this results in weak and unreliable supervision. *(3) Error-reinforcement loop ⇒ amplified skew.* Low confidence in novel samples leads to unreliable pseudo-labels, which in turn cause incorrect updates that hinder the learning of novel classes, thereby reinforcing and amplifying the skew.

4 Method

4.1 Framework Overview

As illustrated in Fig. 4, we propose a novel Confidence-Aware Asymmetric Learning (CAL) framework to accurately attribute deepfake generation methods. *First*, we introduce a Frequency-guided Feature Enhancement (FFE) module that emphasizes informative regions of manipulated content in the frequency domain, thereby enhancing the discriminative power of learned features. *Second*, we propose a Confidence-Aware Consistency Regularization (CCR) mechanism, which adaptively adjusts the regularization strength for unlabeled samples in a threshold-free manner based on their confidence scores (Fig. 4(b)). This design mitigates the impact of noisy pseudo-labels and stabilizes the learning process. *Third*, we incorporate an Asymmetric Confidence Reinforcement (ACR) strategy that explicitly guides the model to learn high-confidence prototypical classifiers for both known and novel categories through asymmetric optimization (Fig. 4(c)). *Finally*, we design a Dynamic Prototype Pruning (DPP) strategy to estimate the number of unseen categories by analyzing the confidence associated with each learned prototype (Fig. 4(d)).

Frequency-Guided Feature Enhancement. Given an input face image $\mathbf{x}_i$, we compute its frequency-domain representation using the Discrete Cosine Transform (DCT), defined as $\mathbf{f}_i = \mathrm{DCT}(\mathbf{x}_i)$. We then use a randomly initialized convolutional network $\mathcal{G}(\cdot)$ to generate a frequency-domain attention mask: $\mathbf{M}_i = \mathcal{G}(\mathbf{f}_i)$. We apply this mask to the frequency features via element-wise multiplication to obtain the weighted frequency representation, denoted as $\mathbf{F}_i = \mathbf{M}_i \odot \mathbf{f}_i$, which is expected to highlight discriminative frequency patterns indicative of different deepfake methods. We next transform the weighted frequency representation back to the spatial domain using the inverse Discrete Cosine Transform (IDCT), denoted as $\mathbf{A}_i = \mathrm{IDCT}(\mathbf{F}_i)$. The resulting map $\mathbf{A}_i$ captures the spatial localization of frequency-domain responses and serves as a *frequency-aware attention map*. To integrate these responses into the attribution process, we define the frequency-aware image representation as the multiplication of $\mathbf{A}_i$ and the spatial feature:

$$\mathbf{h}_i = Pooling(\mathbf{A}_i \odot \mathcal{E}(\mathbf{x}_i); 1 \times 1), \quad (1)$$

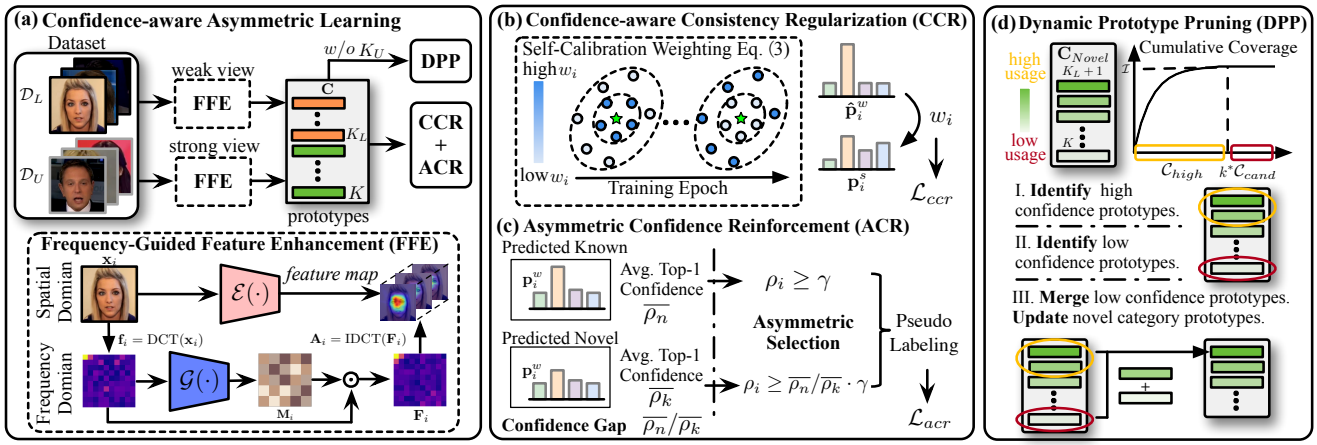

Figure 4: Framework of the proposed Confidence-Aware Asymmetric Learning (CAL).

where $\mathcal{E}(\cdot)$ denotes the image encoder and $Pooling(\cdot; 1\times1)$ denotes global 1×1 pooling. The resulting feature $\mathbf{h}_i$ fuses both spatial and frequency-domain cues, enabling more robust and accurate deepfake attribution.

Discussion. Prior studies (Zhang, Karaman, and Chang 2019; Ricker et al. 2022; Li et al. 2024) in deepfake detection have demonstrated that different types of deepfake generation methods often leave distinct artifacts in the frequency domain of synthesized images. To capture the unique frequency fingerprints of different deepfake methods, we introduce the FFE module to provide complementary frequency-domain insights to enhance spatial-domain features. Although frequency-domain features have been explored in related deepfake detection tasks (Tan et al. 2024; Li et al. 2024; Zhou et al. 2024; Cao et al. 2024a), how to inject frequency-domain knowledge into OW-DFA remains under-explored.

4.2 Confidence-Aware Asymmetric Learning

Motivation. To comprehensively address the issue of *confidence skew* in OW-DFA methods, we design two components targeting complementary aspects: *1) Progressive emphasis on low-confidence samples (→CCR).* In the early training stages, low-confidence predictions—primarily from novel forgeries—are highly prone to label noise. Thus, we initially *down-weight* their gradient contributions. As the model evolves, these weights are progressively increased, enabling the network to effectively learn from previously disregarded samples. *2) Confidence-skew-aware pseudo-label selection (→ACR).* Although reliable pseudo-labels typically originate from high-confidence predictions, the confidence distributions of known and novel classes are inherently misaligned. To account for this, we select pseudo-labels *independently* within each partition using asymmetric, dynamically updated thresholds that reflect the evolving confidence gap. This strategy ensures clean supervision across both partitions while mitigating further bias.

Confidence-Aware Consistency Regularization (CCR). Given the lack of ground-truth labels for the unlabeled dataset, we employ a consistency regularization-based training strategy (Sohn et al. 2020) to encourage similar predictions for perturbed versions of the same image. Specifically, we build a prototypical classifier with K learnable prototypes $\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$, where each prototype represents a "fingerprint" characteristic of a deepfake generation method. Here, K can be set to the ground-truth number of forgery types or estimated by our method in Sec. 4.3. For the feature representation $\mathbf{h}_i$ of the i-th face image, its similarity to the k-th prototype is defined as $\mathbf{s}_{i\to k} = \mathbf{h}_i \cdot \mathbf{c}_k$. The similarities between image $\mathbf{x}_i$ and all K prototypes form the similarity vector $\mathbf{s}_i = [\mathbf{s}_{i\to1}, \ldots, \mathbf{s}_{i\to K}]$. We define the predicted probability as $\mathbf{p}_i = \sigma(\mathbf{s}_i)$, with σ denoting the softmax function. For each input image $\mathbf{x}_i$, we generate two augmented views: a weakly augmented version $\mathbf{x}_i^w$ and a strongly augmented version $\mathbf{x}_i^s$, and extract the corresponding feature embeddings $\mathbf{h}_i^w$ and $\mathbf{h}_i^s$. Following FixMatch (Sohn et al. 2020), we define the consistency regularization loss as:

$$\mathcal{L}_{ws} = \frac{1}{|\mathcal{B}|}\sum_i \mathbb{1}(\max(\hat{\mathbf{p}}_i^w) > \delta)\ell(\hat{\mathbf{p}}_i^w, \mathbf{p}_i^s), \tag{2}$$

where $\mathcal{B}$ denotes a mini-batch, δ is a confidence threshold used to filter high-confidence samples, $\ell(\cdot)$ denotes the cross-entropy loss, and $\hat{\mathbf{p}}_i^w = \sigma(\mathbf{s}_i^w/\tau)$ represents the sharpened predicted probability for the weakly augmented view, with the temperature parameter τ set to 0.1.

Limitation. The regularization encourages the model to learn meaningful representation structures by enforcing prediction consistency under different augmentations for unlabeled data. However, directly applying regularization to the OW-DFA task introduces two key limitations. *First*, in the absence of ground-truth labels for novel categories, the model's predictions tend to be biased toward known categories, and such regularization may further aggravate this bias. *Second*, the threshold-based filtering used in $\mathcal{L}_{ws}$ is sensitive to the choice of the hyperparameter δ, making it inflexible to adapt to varying prediction confidence throughout the training process.

To address these limitations, we propose a **self-calibration weighting** strategy that dynamically rectifies learning

strength based on prediction confidence, in a **threshold-free** fashion. Initially, the model assigns higher weights to high-confidence samples primarily from known forgery types, facilitating the learning of reliable attribution cues from known types. As training progresses, the model gradually shifts emphasis from high-confidence to low-confidence samples, thereby enhancing the learning of novel categories that generally exhibit lower confidence. Specifically, given the sharpened predicted probability $\hat{\mathbf{p}}_i^w$ of the weakly augmented view for the i-th face image, we compute its confidence with respect to the most similar category as $\hat{\rho}_i = \max(\hat{\mathbf{p}}_i^w)$. We then assign a weight to each sample using a scheduling parameter that increases linearly with training epochs, defined as $\frac{e}{E}$, where e denotes the current training epoch and E represents the maximum number of training epochs. The final sample weight is computed as:

$$w_i = (1 - \frac{e}{E}) \cdot \hat{\rho}_i + \frac{e}{E} \cdot (1 - \hat{\rho}_i). \tag{3}$$

This dynamic weighting scheme enables the model to focus on high-confidence samples from known categories in early training, and gradually shift toward low-confidence samples from novel categories as training progresses. Formally, the CCR loss is:

$$\mathcal{L}_{ccr} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_i \cdot \ell\big(\hat{\mathbf{p}}_i^w, \mathbf{p}_i^s\big). \tag{4}$$

Asymmetric Confidence Reinforcement (ACR). To further address the confidence skew issue, we propose ACR, which selects and debiases hard pseudo-labels to reinforce the model's predictions based on the current confidence gap between known and novel classes. Specifically, after a warm-up period of e_0 epochs, we collect the model's top-1 predictions and corresponding confidence scores for all unlabeled samples, and compute the average top-1 confidence for samples predicted as known and novel classes, respectively. Formally, the model's average top-1 prediction confidence for samples predicted as known classes, denoted as $\overline{\rho_k}$, is defined as the mean over the set $\{\rho_i = \max(\mathbf{p}_i^w) \mid \hat{y}_i = \arg\max(\mathbf{p}_i^w) \leq K_L\}$. Similarly, the average top-1 confidence for samples predicted as novel classes is denoted as $\overline{\rho_n}$. We define the model's confidence gap between known and novel classes as $\overline{\rho_n}/\overline{\rho_k}$. Given this confidence gap, we apply asymmetric thresholds for selecting hard pseudo-labels for known and novel categories. For samples predicted as known categories, we apply a high confidence threshold γ. For samples predicted as novel categories, we adopt a relaxed threshold $\overline{\rho_n}/\overline{\rho_k} \cdot \gamma$. This rule is encoded by the indicator:

$$\eta_i = \mathbb{I}\Big[(\hat{y}_i \leq K_L \wedge \rho_i \geq \gamma) \vee (\hat{y}_i > K_L \wedge \rho_i \geq \tfrac{\overline{\rho_n}}{\overline{\rho_k}} \cdot \gamma)\Big]. \tag{5}$$

This adaptive scheme ensures that, even when the confidence for novel classes is relatively low, suitably reliable novel samples are still incorporated during training. As the model gradually improves its understanding of novel categories, the ratio $\overline{\rho_n}/\overline{\rho_k}$ increases adaptively, causing the selection threshold to tighten automatically. Consequently, progressively higher-quality novel samples are fed back into training, steadily closing the confidence gap between known and novel categories. The pseudo-label loss based on the selected pseudo-labels is defined as:

$$\mathcal{L}_{acr} = \frac{1}{|\mathcal{B}_U|} \sum_{i \in \mathcal{B}_U} \eta_i \cdot \ell\big(\hat{y}_i, \mathbf{p}_i\big), \tag{6}$$

where $\mathcal{B}_U$ denotes the unlabeled subset in the sampled mini-batch.

Total Loss. Following CPL (Sun et al. 2023), we adopt a supervised classification loss $\mathcal{L}_{ce}$ on the labeled dataset $\mathcal{D}_L$ for both weak and strong views, along with a regularization term $\mathcal{R}$ on all training data to avoid the trivial solution of assigning all instances to the same class. During the model training process, the total loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{R} + \mathcal{L}_{acr} + \alpha \cdot \mathcal{L}_{ccr}, \tag{7}$$

where α is the weight of $\mathcal{L}_{ccr}$.

4.3 Dynamic Prototype Pruning for Unknown Category Number Estimation

To address scenarios where the number of forgery methods is unknown, we propose a Dynamic Prototype Pruning (DPP) strategy in a multi-stage fashion, as illustrated in Fig. 4 (d). DPP dynamically prunes prototypes of novel classes at each training epoch, ultimately enabling automatic estimation of the number of novel forgery types. To accommodate potentially unknown classes, we allocate a large prototype budget K, which is much larger than the number of known classes K_L. We provide a theoretical analysis of the impact of this setting on the prototypes for novel classes. Let $u_j^{(t)} = \sum_{i=1}^{M} \mathbb{1}\big[\arg\max \mathbf{p}_i^{w(t)} = j\big]$ denote the usage count of the j-th prototype after the t-th update. Under conditions such as bounded intra-class noise of novel forgery types, it can be shown that there exist $\varepsilon \in (0, 1)$ and an iteration upper bound T_0 such that, for all $t \leq T_0$, at least $\big(1 - \frac{K_U}{K}\big)K$ prototypes satisfy $u_j^{(t)} \leq \varepsilon M/K$. The formal statement is given in Lemma 4.1.

Lemma 4.1. *Assume bounded step sizes, angular cluster separability, and bounded feature noise. If the prototype budget satisfies $K \gg K_U$, then there exist $\varepsilon \ll 1$ and $T_0 > 0$ such that, for every $t \leq T_0$,*

$$\mathbb{E}\Big[\big|\{\, j \mid u_j^{(t)} \leq \varepsilon M/K\}\big|\Big] \geq \Big(1 - \tfrac{K_U}{K}\Big)K. \tag{8}$$

Based on this analysis, we further propose a coarse-to-fine strategy that dynamically merges redundant low-usage prototypes during training.

Stage-I: High-confidence Prototype Identification. We perform a coarse partition by selecting high-confidence prototypes and treating the remaining ones as candidate low-confidence prototypes. The index set of novel-class prototypes is initialized as $\mathcal{C}_{novel} = \{K_L + 1, \ldots, K\}$. The usage counts of novel-class prototypes are sorted in descending order as $\mathcal{U} = \{u_k \mid k \in \mathcal{C}_{novel}\}_{\downarrow}$. We leverage the cumulative coverage of prototype usage to identify high-confidence prototypes, considering the overall distribution of samples in the unlabeled dataset. The cumulative coverage for the k-th prototype in $\mathcal{C}_{novel}$ is defined as $r_k = \big(\sum_{j=1}^{k} u_j\big)/\big(\sum_{u_j \in \mathcal{U}} u_j\big)$. Given a target coverage ratio $\mathcal{I}$, the condition $r_k \geq \mathcal{I}$ implies that the top-k prototypes collectively account for at least $\mathcal{I}$ proportion of the samples predicted as novel forgery types. Inspired by the *"2σ rule under a normality assumption"*, we set the coverage threshold to $\mathcal{I} = 95.44\%$ to identify a core set of frequently used

Method	Protocol-1							Protocol-2							Protocol-3						
	All			New			Known	All			New			Known	All			New			Known
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	ACC	NMI	ARI	ACC	NMI	ARI	ACC	ACC	NMI	ARI	ACC	NMI	ARI	ACC
SimGCD	73.3	82.5	54.5	65.0	75.0	57.6	81.7	74.3	81.3	68.4	62.0	76.1	55.8	86.4	85.6	86.3	82.9	78.5	79.7	72.3	90.4
OwMatch	74.1	82.1	81.7	61.2	74.3	53.6	96.2	72.6	80.1	64.2	61.2	74.7	54.4	84.1	84.3	88.4	87.9	66	77.6	63.3	94.8
LPS	76.1	79.0	75.4	61.2	69.5	53.1	90.9	75.1	79.2	78.5	59.9	73.7	53.2	92.8	80.0	86.9	85.9	57.8	70.5	52.5	92.5
LegoGCD	70.9	76.3	57.6	61.0	69.2	52.9	82.6	72.2	81.0	58.3	64.5	77.1	57.5	80.8	82.0	81.9	75.3	68.6	74.8	64.8	86.8
ProtoGCD	62.9	79.7	37.7	59.8	71.1	51.3	66.8	61.8	70.0	41.5	58.1	68.1	48.7	69.3	78.2	81.0	63.4	73.1	75.9	67.3	82.1
PALGCD	76.5	80.0	79.0	54.2	66.5	45.1	95.9	77.6	80.6	79.2	65.0	75.4	56.9	92.3	83.3	82.5	79.3	63.0	71.4	61.7	89.5
CPL	82.6	85.5	85.1	64.3	73.2	56.3	96.2	79.6	83.7	86.3	63.6	75.3	55.3	95.0	86.3	88.1	79.2	74.4	79.4	70.4	90.3
CDAL	84.3	87.4	86.0	70.0	77.7	62.5	95.3	80.2	83.9	85.7	65.7	76.2	57.3	94.1	87.3	88.7	83.4	74.4	78.3	68.7	91.6
Ours (w/o K_U)	87.5	89.7	90.2	74.5	**82.3**	70.9	**98.2**	80.0	86.1	88.7	63.0	78.2	59.1	97.5	**91.5**	**92.1**	**91.7**	**83.2**	**85.4**	**79.6**	**95.0**
Ours	**88.3**	**90.3**	**91.5**	**76.5**	82.2	**72.4**	98.0	**82.8**	**87.2**	**90.0**	**67.5**	**80.2**	**63.5**	**97.6**	91.0	91.8	91.6	80.3	83.9	77.3	94.8

Table 1: Results on the OW-DFA-40 benchmark. Best results are in **bold**, second-best are underlined.

prototypes, analogous to selecting statistically significant regions in a normal distribution. We define the index set of high-confidence prototypes as $\mathcal{C}_{high} = \{j \mid j \leq k^\star\}$, where $k^\star = \min\{k \mid r_k \geq \mathcal{I}\}$. The remaining prototypes are regarded as candidate low-confidence prototypes, denoted as $\mathcal{C}_{cand} = \mathcal{C}_{novel} \setminus \mathcal{C}_{high}$.

Stage-II: Low-confidence Prototype Filtering. We adopt a strict strategy to further filter out potential noisy entries. For all candidate prototypes, we compute the average usage count $\bar{u}$ and the first-order difference of the cumulative coverage, defined as $\Delta r_k = r_k - r_{k-1}$. We then calculate the average first-order difference $\overline{\Delta r}$ across all candidate prototypes. Low-confidence prototypes are defined as those with small usage counts and minimal changes in cumulative coverage, with the corresponding index set defined as $\mathcal{C}_{\text{low}} = \left\{ k \in \mathcal{C}_{\text{cand}} \,\middle|\, u_k < \bar{u} \ \wedge \ \Delta r_k < \overline{\Delta r} \right\}$.

Stage-III: Similarity-driven Prototype Merge. To avoid information loss caused by hard pruning, we merge low-confidence prototypes into their most similar high-confidence prototypes. The updated index set of novel-class prototypes $\mathcal{C}_{novel}$ is defined as $\mathcal{C}_{novel} \leftarrow \mathcal{C}_{novel} \setminus \mathcal{C}_{low}$.

5 Experiment

5.1 Experiment Setup

OW-DFA-40 Benchmark. We construct a new OW-DFA-40 benchmark comprising 40 deepfake generation methods covering five mainstream facial forgery categories. In addition to the 20 methods from the original OW-DFA benchmark (Sun et al. 2023), we include 20 newly added state-of-the-art techniques spanning Face Swapping (2022b; 2022a; 2023; 2023; 2023), Face Reenactment (2021; 2022; 2022; 2022; 2023; 2023; 2023), Face Editing (2021), Entire Face Synthesis (2021; 2022; 2022b), and Diffusion-based Generation (2023; 2024; 2024; 2024). These fake face images are generated using real facial data from two widely-used datasets: FaceForensics++ (Rossler et al. 2019) and Celeb-DF (Li et al. 2020).

Evaluation Protocol. We define three evaluation protocols. **Protocol-1** includes 40 forgery methods—19 known and 22 unknown—along with real face data, yielding a total of 41 attribution classes. **Protocol-2** builds on Protocol-1 by treating all methods under *Entire Face Synthesis* and *Diffusion-based Generation* as unknown, resulting in 13 known and 28 unknown classes. This setup simulates the emergence of new attack paradigms. **Protocol-3** also extends Protocol-1, incorporating more known methods across forgery categories, leading to 29 known and 12 unknown classes. For all protocols, we follow the data split strategy of the original OW-DFA benchmark: 80% of the data is used for training and 20% for testing, forming the labeled dataset $\mathcal{D}_L$ and the unlabeled dataset $\mathcal{D}_U$. Unlike the original OW-DFA benchmark, we unify real face data from FaceForensics++ and Celeb-DF into a single category, denoted as *real*, which is included as an attribution class in each protocol.

Implementation Details. For a fair comparison, we use a ResNet-50 pre-trained on ImageNet (Deng et al. 2009) as the image encoder, following CPL (Sun et al. 2023). For face preprocessing, we resize all input images to 256×256 and detect faces using dlib. The optimizer is Adam with a learning rate of 2×10^{-4}. All models are trained for 50 epochs with a batch size of 128. The module $\mathcal{G}(\cdot)$ is implemented as a lightweight convolutional network with two layers of 3×3 kernels. The warm-up epoch for $\mathcal{L}_{acr}$ is set to $e_0 = 5$. The weak augmentation includes only `RandomHorizontalFlip` with a probability of 0.5, while the strong augmentation combines `RandomHorizontalFlip`, `RandomResizedCrop`, and brightness adjustment, each applied with a probability of 0.2. The initial number of attribution prototypes is set to $K = 10 \times K_L$. We set the hyperparameters to $\alpha = 0.2$ and $\gamma = 0.9$, determined based on the labeled dataset under Protocol-1. To avoid complex fine-tuning, we use this hyperparameter setting for all experiments. All models are trained on a single NVIDIA A100 GPU, and results are averaged over three runs using random seeds $\{0, 1, 2\}$.

Compared Methods. We compare our method with CPL (Sun et al. 2023) and CDAL (Zheng et al. 2025), along with strong baselines from the Generalized Category Discovery (GCD) and Open-World Semi-Supervised Learning

(OWSSL) settings. Specifically, we include SimGCD (Wen, Zhao, and Qi 2023), LegoGCD (Cao et al. 2024b), and ProtoGCD (Ma et al. 2025) from GCD, as well as Owmatch (Niu et al. 2024), LPS (Ye et al. 2014), and PALGCD (Wang, Zhong, and Gong 2025) from OWSSL. All compared methods are configured according to the settings of CPL (Sun et al. 2023).

5.2 Comparison with State of the Art

Results on OW-DFA-40. We compare our method with the aforementioned baselines on the OW-DFA-40 benchmark, as shown in Tab. 1. The proposed CAL consistently outperforms all state-of-the-art competitors across nearly all metrics. Compared to CPL (Sun et al. 2023) and CDAL (Zheng et al. 2025), CAL achieves average All ACC improvements of 4.5% and 3.4%, respectively. More importantly, CAL surpasses CPL by an average of 7.3% in Novel ACC across all protocols, indicating its superior ability to discover novel deepfake methods and to mitigate model bias toward known classes. *Even when the number of novel forgery types K_U* ***is unknown****, our CAL still outperforms CPL by an average of 3.5% in All ACC, demonstrating its applicability in real open-world scenarios.* OWSSL and GCD methods are primarily designed for capturing global visual similarity in natural images, making them less effective for deepfake attribution. Specifically, compared to two state-of-the-art OWSSL methods, CAL improves All ACC by 10.4% over OwMatch and 10.3% over LPS, on average across all protocols. Furthermore, CAL surpasses four GCD methods in average All ACC: by 9.6% over SimGCD, 19.7% over ProtoGCD, 12.3% over LegoGCD, and 8.2% over PALGCD.

Results on OW-DFA (Sun et al. 2023). We further evaluate our method on the original OW-DFA benchmark, where CAL achieves an average All ACC improvement of 3.4% and a notable 7.1% gain in Novel ACC over CPL across the two evaluation protocols.

5.3 Ablation Study

Ablation on Components. We conduct an ablation study of the proposed components in our CAL on Protocol-1 of the OW-DFA-40 benchmark, as shown in Tab. 2. The components under evaluation include the base loss $\mathcal{L}_{ce} + \mathcal{R}$, our proposed Confidence-Aware Consistency Regularization loss $\mathcal{L}_{ccr}$, our Asymmetric Confidence Reinforcement loss $\mathcal{L}_{acr}$, and Frequency-Guided Feature Enhancement (FFE) module. Models I, III, and V illustrate the step-by-step addition of training components. Using only the base loss enables the model to effectively learn known classes, but yields only 33.2% in Novel ACC. Adding $\mathcal{L}_{ccr}$ results in an 18.4% improvement in All ACC and a substantial 32.8% improvement in Novel ACC, highlighting the critical role of self-calibration weighting in facilitating novel deepfake discovery. Building on Model III, incorporating $\mathcal{L}_{acr}$ further enhances performance by 3.7% in All ACC and 10.5% in Novel ACC. The comparisons between Models II and III, as well as between Models IV and V, demonstrate the effect of FFE. The inclusion of FFE yields an average improvement of 2.7% in All ACC in both comparisons, underscoring the benefit of incorporating frequency-domain information and suggesting that different types of deepfake methods generate images with distinctive frequency characteristics.

	$\mathcal{L}_{ce}$ $+\mathcal{R}$	$\mathcal{L}_{ccr}$	$\mathcal{L}_{acr}$	FFE	All			New			Known
					ACC	NMI	ARI	ACC	NMI	ARI	ACC
I	✓			✓	66.2	71.3	74.5	33.2	43.0	18.8	95.6
II	✓	✓			80.1	83.3	85.6	54.8	68.4	47.3	98.1
III	✓	✓		✓	84.6	87.9	89.6	66.0	76.5	61.5	98.3
IV	✓	✓	✓		87.4	88.9	90.5	73.4	79.3	69.9	**98.6**
V	✓	✓	✓	✓	**88.3**	**90.3**	**91.5**	**76.5**	**82.2**	**72.4**	98.0

Table 2: Ablation study on training components. The best results are marked in **bold**.

Method	backbone	**Protocol-1**		**Protocol-2**		**Protocol-3**	
		K	Err(%)	K	Err(%)	K	Err(%)
GCD	DINO-ViT-B	19	53.7	13	68.3	33	19.5
GCD	ResNet-50	20	51.2	13	68.3	35	14.6
Ours	ResNet-50	39	**4.8**	37	**9.7**	40	**2.4**

Table 3: Comparison with class number estimation method from GCD.

5.4 Evaluation

Evaluation of Class Number Estimation. We compare our DPP strategy with the class number estimation method proposed in GCD (Vaze et al. 2022), as shown in Tab. 3. GCD treats clustering accuracy on the labeled dataset as a black-box scoring function and selects the class number that yields the highest accuracy as the final estimate—a strategy widely adopted in both GCD and OWSSL tasks. For a fair comparison, we evaluate both methods using DINO-ViT-B and an ImageNet-pretrained ResNet-50 as backbones. Our DPP method consistently achieves lower estimation errors across all protocols. In contrast, GCD's approach heavily depends on the pretrained backbone's ability to extract meaningful features from unknown classes, which proves inadequate in the OW-DFA setting.

6 Conclusion

We propose Confidence-Aware Asymmetric Learning (CAL), a novel framework for Open-World DeepFake Attribution (OW-DFA), which addresses key challenges of confidence skew and unknown category estimation. By combining Confidence-Aware Consistency Regularization (CCR) for adaptive pseudo-label regularization and Asymmetric Confidence Reinforcement (ACR) for asymmetric confidence calibration, CAL effectively mitigates bias toward known forgery types. Furthermore, the proposed Dynamic Prototype Pruning (DPP) strategy enables dynamic estimation of novel categories without knowing the number of novel forgery types, improving the scalability of CAL. Experiments on standard and extended OW-DFA benchmarks show that CAL achieves state-of-the-art performance on both known and novel forgeries.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (No. 62572166 & No. 62402157) and the Fundamental Research Funds for the Central Universities (No. JZ2025HGTB0219). This work was also supported by the EU Horizon project "ELIAS - European Lighthouse of AI for Sustainability" (No. 101120237) and the FIS project GUIDANCE (Debugging Computer Vision Models via Controlled Cross-modal Generation) (No. FIS2023-03251). We further acknowledge CINECA and the ISCRA initiative for providing high-performance computing resources.

References

Bounareli, S.; Tzelepis, C.; Argyriou, V.; Patras, I.; and Tzimiropoulos, G. 2023. Hyperreenact: one-shot reenactment via jointly learning to refine and retarget faces. In *ICCV*.

Cao, J.; Zhang, K.-Y.; Yao, T.; Ding, S.; Yang, X.; and Ma, C. 2024a. Towards unified defense for face forgery and spoofing attacks via dual space reconstruction learning. *IJCV*.

Cao, K.; Brbic, M.; and Leskovec, J. 2021. Open-world semi-supervised learning. *arXiv preprint arXiv:2102.03526*.

Cao, X.; Zheng, X.; Wang, G.; Yu, W.; Shen, Y.; Li, K.; Lu, Y.; and Tian, Y. 2024b. Solving the catastrophic forgetting problem in generalized category discovery. In *CVPR*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *CVPR*.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2024. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *ICLR*.

CiteDrive, I. 2024. Midjourney. https://www.midjourney.com/home/. Accessed: 2024-04-20.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Guarnera, L.; Giudice, O.; Nießner, M.; and Battiato, S. 2022. On the exploitation of deepfake model recognition. In *CVPR*.

Han, K.; Vedaldi, A.; and Zisserman, A. 2019. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*.

Hong, F.-T.; and Xu, D. 2023. Implicit identity representation conditioned memory compensation network for talking head video generation. In *ICCV*.

Hong, F.-T.; Zhang, L.; Shen, L.; and Xu, D. 2022. Depth-aware generative adversarial network for talking head video generation. In *CVPR*.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-free generative adversarial networks. *NeurIPS*.

Li, Y.; Bammey, Q.; Gardella, M.; Nikoukhah, T.; Morel, J.-M.; Colom, M.; and Von Gioi, R. G. 2024. MaskSim: Detection of synthetic images by masked spectrum similarity analysis. In *CVPR*.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *CVPR*.

Liu, J.; Wang, Q.; Fan, H.; Wang, Y.; Tang, Y.; and Qu, L. 2024. Residual denoising diffusion models. In *CVPR*.

Liu, Z.; Li, M.; Zhang, Y.; Wang, C.; Zhang, Q.; Wang, J.; and Nie, Y. 2023. Fine-grained face swapping via regional gan inversion. In *CVPR*.

Ma, N.; Goldstein, M.; Albergo, M. S.; Boffi, N. M.; Vanden-Eijnden, E.; and Xie, S. 2024. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *ECCV*.

Ma, S.; Zhu, F.; Zhang, X.-Y.; and Liu, C.-L. 2025. ProtoGCD: Unified and Unbiased Prototype Learning for Generalized Category Discovery. *IEEE TPAMI*.

Niu, S.; Lin, L.; Huang, J.; and Wang, C. 2024. OwMatch: Conditional Self-Labeling with Consistency for Open-World Semi-Supervised Learning. *NeurIPS*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *ICCV*.

Ricker, J.; Damm, S.; Holz, T.; and Fischer, A. 2022. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*.

Rizve, M. N.; Kardan, N.; Khan, S.; Shahbaz Khan, F.; and Shah, M. 2022. Openldn: Learning to discover novel classes for open-world semi-supervised learning. In *ECCV*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Rosberg, F.; Aksoy, E. E.; Alonso-Fernandez, F.; and Englund, C. 2023. Facedancer: Pose-and occlusion-aware high fidelity face swapping. In *WACV*.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. In *CVPR*.

Sauer, A.; Schwarz, K.; and Geiger, A. 2022. Stylegan-xl: Scaling stylegan to large diverse datasets. In *SIGGRAPH*.

Shiohara, K.; Yang, X.; and Taketomi, T. 2023. Blendface: Re-designing identity encoders for face-swapping. In *ICCV*.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*.

Sun, Y.; and Li, Y. 2022. Opencon: Open-world contrastive learning. *arXiv preprint arXiv:2208.02764*.

Sun, Z.; Chen, S.; Yao, T.; Yi, R.; Ding, S.; and Ma, L. 2025. Rethinking open-world deepfake attribution with multi-perspective sensory learning. *IJCV*.

Sun, Z.; Chen, S.; Yao, T.; Yin, B.; Yi, R.; Ding, S.; and Ma, L. 2023. Contrastive pseudo learning for open-world deepfake attribution. In *ICCV*.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *AAAI*.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *NeurIPS*.

Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; and Cohen-Or, D. 2021. Designing an encoder for stylegan image manipulation. *SIGGRAPH*.

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2022. Generalized category discovery. In *CVPR*.

Wang, M.; Zhong, Z.; and Gong, X. 2025. Prior-Constrained Association Learning for Fine-Grained Generalized Category Discovery. In *AAAI*.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*.

Wang, Y.; Yang, D.; Bremond, F.; and Dantcheva, A. 2022. Latent image animator: Learning to animate images via latent space navigation. *ICLR*.

Wang, Y.; Zhong, Z.; Qiao, P.; Cheng, X.; Zheng, X.; Liu, C.; Sebe, N.; Ji, R.; and Chen, J. 2023. Discover and align taxonomic context priors for open-world semi-supervised learning. *NeurIPS*.

Wen, X.; Zhao, B.; and Qi, X. 2023. Parametric classification for generalized category discovery: A baseline study. In *ICCV*.

Xu, C.; Zhang, J.; Han, Y.; Tian, G.; Zeng, X.; Tai, Y.; Wang, Y.; Wang, C.; and Liu, Y. 2022a. Designing one unified framework for high-fidelity face reenactment and swapping. In *ECCV*.

Xu, Z.; Hong, Z.; Ding, C.; Zhu, Z.; Han, J.; Liu, J.; and Ding, E. 2022b. Mobilefaceswap: A lightweight framework for video face swapping. In *AAAI*.

Yang, T.; Huang, Z.; Cao, J.; Li, L.; and Li, X. 2022. Deepfake network architecture attribution. In *AAAI*.

Ye, B.; Gan, K.; Wei, T.; and Zhang, M.-L. 2014. Bridging the gap: Learning pace synchronization for open-world semi-supervised learning. In *IJCAI*.

Yu, N.; Skripniuk, V.; Abdelnabi, S.; and Fritz, M. 2021. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *ICCV*.

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2023. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *CVPR*.

Zhang, X.; Karaman, S.; and Chang, S.-F. 2019. Detecting and simulating artifacts in gan fake images. In *WIFS*.

Zhao, J.; and Zhang, H. 2022. Thin-plate spline motion model for image animation. In *CVPR*.

Zheng, Y.; Gong, B.; Kong, F.; Duan, Y.; Yu, B.; Zheng, W.; Chen, L.; Lu, J.; and Zhou, J. 2025. Learning Counterfactually Decoupled Attention for Open-World Model Attribution. *arXiv preprint arXiv:2506.23074*.

Zhou, J.; Li, Y.; Wu, B.; Li, B.; Dong, J.; et al. 2024. FreqBlender: Enhancing DeepFake Detection by Blending Frequency Knowledge. *NeurIPS*.