# Contrastive Pseudo Learning for Open-World DeepFake Attribution

Zhimin Sun[1,2,3*] Shen Chen[2*] Taiping Yao[2] Bangjie Yin[2]
Ran Yi[1†] Shouhong Ding[2†] Lizhuang Ma[1]
[1]Shanghai Jiao Tong University [2]Tencent YouTu Lab
[3]Shanghai Key Laboratory of Computer Software Testing & Evaluating

## Abstract

*The challenge in sourcing attribution for forgery faces has gained widespread attention due to the rapid development of generative techniques. While many recent works have taken essential steps on GAN-generated faces, more threatening attacks related to identity swapping or expression transferring are still overlooked. And the forgery traces hidden in unknown attacks from the open-world unlabeled faces still remain under-explored. To push the related frontier research, we introduce a new benchmark called Open-World DeepFake Attribution (OW-DFA), which aims to evaluate attribution performance against various types of fake faces under open-world scenarios. Meanwhile, we propose a novel framework named Contrastive Pseudo Learning (CPL) for the OW-DFA task through 1) introducing a Global-Local Voting module to guide the feature alignment of forged faces with different manipulated regions, 2) designing a Confidence-based Soft Pseudo-label strategy to mitigate the pseudo-noise caused by similar methods in unlabeled set. In addition, we extend the CPL framework with a multi-stage paradigm that leverages pre-train technique and iterative learning to further enhance traceability performance. Extensive experiments verify the superiority of our proposed method on the OW-DFA and also demonstrate the interpretability of deepfake attribution task and its impact on improving the security of deepfake detection area.*

## 1. Introduction

With the rapid development of generative technologies such as Deepfakes [2], the malicious usage of fake content on social media has raised public concerns about face security and privacy. Dedicated research efforts [26, 48, 56] have been made in the real/fake detection task in recent years. Nonetheless, with its distinctive merits, DeepFake Attribution (DFA), known as identifying the source model of fake faces, has also significantly drawn widespread attentions [59, 61, 24]. On the one hand, DFA can be used for legal proceedings and provide interpretability to human beings, *i.e.*“why the face is fake.” On the other hand, with the nature of learning enhanced representation for different attacking types, DFA is also effective to boost the deepfake detection performance [29, 17].

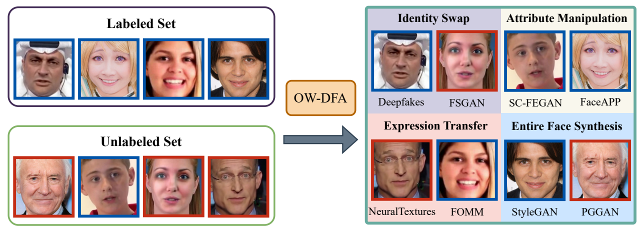



Figure 1. In the OW-DFA setting, the unlabeled dataset may contain attacks that have never been encountered in the labeled set. A feasible model should attribute the known attacks (images with blue border) and assign the unknown attacks (images with red border) to novel classes simultaneously.

Early approaches of sourcing attribution [59, 61, 24] mostly focus on the GAN-generated images rather than the more realistic and threatening attacks related to identity swapping or expression transferring. Meanwhile, most of them assume a closed scenario where the training set and test set share the same category distributions, which is not applicable to open-world scenarios since new types of forgery attacks emerge immensely. To this end, we introduce a new benchmark, Open-World DeepFake Attribution (OW-DFA), as shown in Figure 1. The OW-DFA benchmark consists of a labeled training dataset and an unlabeled dataset. The labeled dataset contains samples from known classes, while the unlabeled dataset includes samples from both known and unknown classes. More importantly, OW-DFA considers nearly 20 challenging and realistic forgery methods, including 4 widely-used forgery types, namely *identity swap* [1, 2], *expression transfer* [5, 49], *attribute manipulation* [12, 13] and *entire face synthesis* [32, 34]. The main challenge of OW-DFA is how to utilize unlabeled data in open-world scenes to improve the attribution performance for both known and unknown forged faces.

[*]Equal contribution. This work was done when Zhimin Sun was a research intern at Tencent YouTu Lab.
[†]Corresponding authors.

The OW-DFA is fundamentally different but closely related to Open-World Semi-Supervised Learning (OW-SSL), where some OW-SSL methods [27, 8, 25] demonstrate effectiveness in learning unknown categories through contrastive learning or pseudo-labeling strategies. However, since all classes of OW-DFA are face data whose disparity information relies on fine-grained forgery traces [10, 65], these OW-SSL methods that only focus on global information will be limited in attributing unknown forged faces. Moreover, Open-world GAN [18] discovers and refines unseen GANs with iterative algorithms. However, the fingerprint assumption relied on may not hold in the fake faces generated by non-GAN methods. Without a semi-supervised learning strategy, the features extracted by this model for different unknown attacks lack distinguishability.

In this paper, we propose a novel framework named Contrastive Pseudo Learning (CPL), which addresses the above issues from two perspectives: 1) We introduce a Global-Local Voting (GLV) module that guides inter-sample feature alignment by extracting both global and local information and adaptively highlights different manipulated regions through a spatially enhancing mechanism. By combining global and local similarity, we can filter and group together samples of the same attack type. 2) Besides the inter-sample relation, we also leverage the intra-sample information to enhance the class compactness using the pseudo-labeling technique. A Confidence-based Soft Pseudo-labeling (CSP) mechanism is proposed to mitigate the pseudo-noise induced by similar novel attack methods. Moreover, previous research [27, 55] has demonstrated the efficacy of pre-training techniques and iterative learning, so we extend the CPL framework with a multi-stage paradigm to further improve the attribution performance. Finally, extensive experimental results verify the superiority of our method on the OW-DFA benchmark. We also demonstrate the interpretability of the deepfake attribution task and its impact on improving the security of the deepfake detection area.

We summarize our contributions as follows:

(1) We present a new benchmark called Open-World DeepFake Attribution (OW-DFA), which aims to evaluate attribution performance against various types of fake faces under open-world scenarios.

(2) We propose a novel Contrastive Pseudo Learning (CPL) framework for OW-DFA task through 1) a Global-Local Voting module to guide the feature alignment of forged faces with different manipulated regions, 2) a Confidence-based Soft Pseudo-labeling strategy to mitigate pseudo-noise caused by similar methods in unlabeled set.

(3) Comprehensive experiments and visualization results demonstrate that our method achieves SOTA performance on OW-DFA. We also show that combining the deepfake attribution task with the deepfake detection task leads to better interpretability and face security.

## 2. Related Works

### 2.1. DeepFake Attribution

A plethora of works [14, 53, 44, 64, 11, 7, 52, 51, 19, 20, 23, 22, 21] for the real/fake detection task have been proposed in recent years. However, the generalization performance on novel attacks is still limited. As fake faces become visually realistic and need to be interpreted in legal proceedings, attribution of the source model of fake faces has gained widespread attention. Most existing works [59, 60, 61, 24] focus only on the problem of attributing GAN models, and a common strategy is to use the fingerprints of different GAN models to attribute those generated images. However, they only consider the close-world scenario where the training and test sets have the same category distribution. Such an assumption is not applicable to open-world scenarios since novel forgeries emerge greatly. The most relevant method, Open-world GAN [18], proposes an iterative algorithm to discover and refine unseen GANs in an open-world scenario. Although it has made some progress in open-world scenarios, the features extracted by this model for unknown attacks lack discriminability without proper use of unlabeled data. To this end, we propose a new benchmark OW-DFA, which contains samples from both known and unknown classes and then utilizes more challenging and realistic forgery methods. Furthermore, with the proposed CPL framework, we significantly boost the performance of deepfake attribution under the OW-DFA setup.

### 2.2. Open-World Semi-Supervised Learning

Open-World Semi-Supervised Learning (OW-SSL) [8] aims to leverage both labeled and unlabeled data in an open-world scenario, where novel classes may exist in the unlabeled datasets. Existing OW-SSL methods [8, 45, 25] bring instances of the same class in unlabeled datasets closer together based on inter-sample similarity. And assigning pseudo-labels to the high-confidence samples is another common technique [50, 62, 57, 45, 46]. Despite the promising performance of these methods on OW-DFA tasks, they still face some challenges. First, existing methods [8, 45] mainly focus on the global similarity of samples, neglecting the local consistency of forged face images that may indicate tampering. To address this, we propose a Global-Local Voting module that matches samples more accurately by considering both global and local features of facial attacks. Second, the threshold-based pseudo-labeling strategy [62, 57] can only handle samples with deterministic labels. To address this, we propose a probability-based pseudo-labeling strategy that imposes additional constraints on samples with low confidence.

Table 1. List of methods and corresponding datasets utilized in OW-DFA. Protocol 1 encompasses 20 challenging forgery techniques, with forgery types ranging from identity swap, expression transfer, attribute manipulation and entire face synthesis. The primary objective of Protocol 1 is to enhance the attribution of forgery attacks. Protocol 2 combines the forgery techniques from Protocol 1 with real faces to create a realistic open-set mixed attribution scenario that mimics real-life situations.

| Face Type | Labeled Sets | Unlabeled Sets | Source Dataset | Method | Tag | Labeled # | Unlabeled # |
|---|---|---|---|---|---|---|---|
| Identity Swap | Deepfakes [2]<br>DeepFaceLab [1] | Deepfakes<br>DeepFaceLab<br>FaceSwap [4]<br>FaceShifter [40]<br>FSGAN [42] | FaceForensics++ [47] | Deepfakes<br>FaceSwap | Known<br>Novel | 1500<br>- | 500<br>1500 |
| | | | ForgeryNet [29] | DeepFaceLab<br>FaceShifter<br>FSGAN | Known<br>Novel<br>Novel | 1500<br>-<br>- | 500<br>1500<br>1500 |
| Expression Transfer | Face2Face [54]<br>FOMM [49] | Face2Face<br>FOMM<br>NeuralTextures [5]<br>Talking-Head-Video [63]<br>ATVG-Net [9] | FaceForensics++ | Face2Face<br>NeuralTextures | Known<br>Novel | 1500<br>- | 500<br>1500 |
| | | | ForgeryNet | FOMM<br>ATVG-Net<br>Talking-Head-Video | Known<br>Novel<br>Novel | 1500<br>-<br>- | 500<br>1500<br>1500 |
| Attribute Manipulation | MaskGAN [38]<br>FaceAPP [3] | MaskGAN<br>FaceAPP<br>StarGAN2 [13]<br>SC-FEGAN [31]<br>StarGAN [12] | ForgeryNet | MaskGAN<br>StarGAN2<br>SC-FEGAN | Known<br>Novel<br>Novel | 1500<br>-<br>- | 500<br>1500<br>1500 |
| | | | DFFD [15] | FaceAPP<br>StarGAN | Known<br>Novel | 1500<br>- | 500<br>1500 |
| Entire Face Synthesis | StyleGAN [33]<br>CycleGAN [67] | StyleGAN<br>CycleGAN<br>PGGAN [32]<br>StyleGAN2 [34] | ForgeryNet | StyleGAN2 | Novel | - | 1500 |
| | | | DFFD | StyleGAN<br>PGGAN | Known<br>Novel | 1500<br>- | 500<br>1500 |
| | | | ForgeryNIR [58] | CycleGAN<br>StyleGAN2 | Known<br>Novel | 1500<br>- | 500<br>1500 |
| Real Face | Youtube-Real [47] | Celeb-Real [41] | FaceForensics++ | Youtube-Real | Known | 15000 | 5000 |
| | | | CelebDFv2 [41] | Celeb-Real | Novel | - | 5000 |

## 3. Open-World DeepFake Attribution

In this section, we first present the definition of the Open-World DeepFake Attribution (OW-DFA) task with labeled and unlabeled sets, known and novel categories, and the corresponding notation. We then list the dataset composition of OW-DFA and propose two challenging protocols. More preprocessing details for each dataset are provided in Supplementary Sec. B.

### 3.1. Definition

The Open-World DeepFake Attribution (OW-DFA) task consists of a labeled set $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^n$ and an unlabeled set $\mathcal{D}_u = \{(x_i)\}_{i=1}^m$. We denote the classes in the labeled set as $\mathcal{C}_L$, and those in the unlabeled set as $\mathcal{C}_U$, where $\mathcal{C}_L$ contains only known categories, while $\mathcal{C}_U$ covers both known and novel categories, *i.e.*, $\mathcal{C}_L \cap \mathcal{C}_U \neq \emptyset$ and $\mathcal{C}_L \neq \mathcal{C}_U$. We denote the known class as $\mathcal{C}_K = \mathcal{C}_L \cap \mathcal{C}_U$ and the novel class as $\mathcal{C}_N = \mathcal{C}_U \backslash \mathcal{C}_L$. The goal of OW-DFA task is utilizing both labeled sets $\mathcal{D}_l$ and unlabeled sets $\mathcal{D}_u$ to learn a feature extractor $\phi(\cdot)$ and a classifier $\sigma(\cdot)$, which can recognize source models for various faces types. Unlike previous work [18, 24, 59, 60, 61] that only considered GAN-generated images, OW-DFA also includes more threatening attacks related to identity swap or expression transfer.

### 3.2. Protocols

We create the OW-DFA benchmark based on several deepfake datasets, including FF++ [47], CelebDF [41], ForgeryNet [29], DFFD [15] and ForgeryNIR [58]. These datasets are widely used in the deepfake detection task with large-scale data and various types of forged faces, which can be roughly divided into 5 face types: *identity swap*, *expression transfer*, *attribute manipulation*, *entire face synthesis* and *real face*.

As can be seen in Table 1, we define two protocols for OW-DFA to evaluate the performance in real-world scenarios: 1) **Protocol-1** aims to evaluate the attribution performance of the forgery method, which includes 20 manipulation methods across 4 mainstream forgery face types: *identity swap*, *expression transfer*, *attribute manipulation* and *entire face synthesis*. Under this setting, all labeled and unlabeled data are fake faces. 2) **Protocol-2** includes additional real faces from different domains on top of Protocol-1, taking into account the fact that real faces may appear on social platforms. Specifically, we introduce real faces from the FaceForensics++ and Celeb-DF datasets in labeled sets and unlabeled sets respectively. Compared to each forgery type, the amount of real data is larger to simulate the distribution of faces in real scenes.

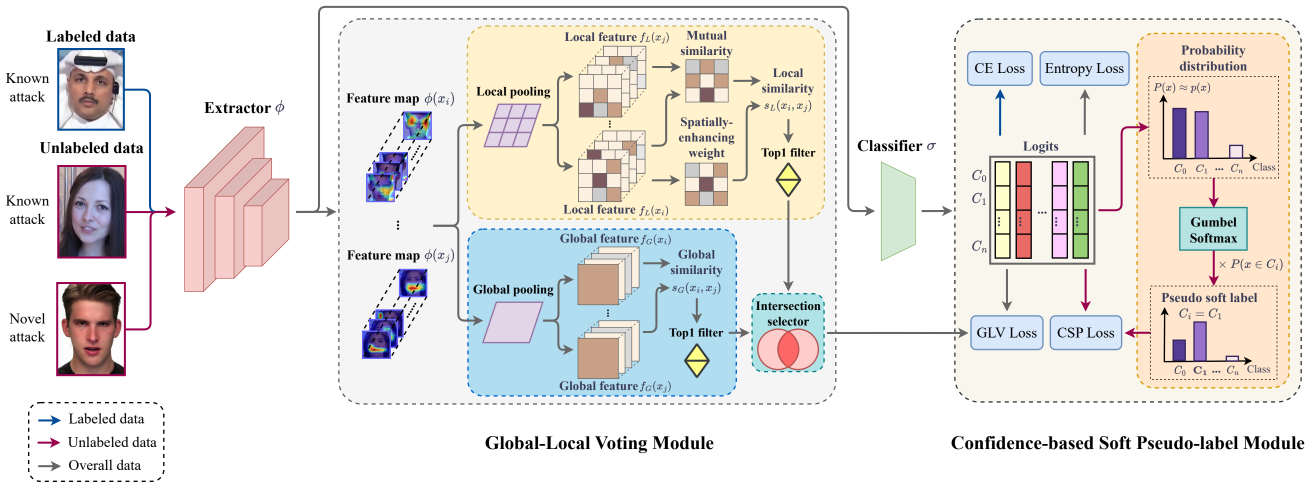



Figure 2. Contrastive Pseudo Learning (CPL) framework for Open-World DeepFakes Attribution task.

# 4. Contrastive Pseudo Learning Framework

The key challenge of OW-DFA is to use labeled and unlabelled sets to jointly learn discriminative representations of known and novel attacks. To this end, we proposed a novel Contrastive Pseudo Learning (CPL) framework, as shown in Figure 2. The CPL framework includes two key components: 1) a Global-Local Voting (GLV) module to guide the feature alignment of different forgery types. 2) a Confidence-based Soft Pseudo-labeling (CSP) module to mitigate the pseudo-noise caused by similar forgery methods in unlabeled sets. Then we summarise all relevant loss functions. Finally, we combine the proposed CPL framework with a pretraining technique and iterative learning to further improve the performance under the OW-DFA setup.

## 4.1. Global-Local Voting Module

To facilitate the representation compactness of novel attacks in unlabeled sets, one feasible strategy is the contrast learning [8, 27], which aims to transform the unsupervised clustering problem into a similarity measurement problem. In particular, given an input face image $x_i$ and label $y_i^l$ for labeled sample, we use $\phi(x_i)$ to extract the corresponding feature map, and a Pooling layer is applied to obtain the global representation, which is formulated as:

$$f_G(x_i) = Pooling(\phi(x_i); 1 \times 1), \tag{1}$$

where $f_G(x_i) \in \mathbb{R}^d$, and $d$ denotes the feature dimensions. For each pair $\{(x_i, x_j) : i, j \in (0, \cdots, n+m)\}$, the inter-sample relation is measured by the cosine similarity of their global features:

$$s_G(x_i, x_j) = \frac{f_G(x_i) \cdot f_G(x_j)}{\|f_G(x_i)\| \|f_G(x_j)\|}. \tag{2}$$

Given a mini-batch containing both $n$ labeled and $m$ unlabeled samples, we use the above strategy to compute the similarity between each sample $x_i$ and all other samples. Then we bring $x_i$ closer to its most similar sample $\widetilde{x}_i$ by a variant of BCE loss, *i.e.* global relation constraints:

$$\mathcal{L}_{\mathrm{GR}} = -\frac{1}{n+m} \sum_{x_i \in \mathcal{D}_l \cup \mathcal{D}_u} \log\langle \sigma(f_G(x_i)), \sigma(f_G(\widetilde{x}_i)) \rangle, \tag{3}$$

where $\sigma$ outputs the probability for each sample.

However, the tempered region varies for different forged types, *e.g.*, the GAN-generated images are forged at every pixel, whereas expression transfer tends to manipulate in the mouth region. When comparing the similarity of face samples, not considering local fine-grained traces may lead to incorrect contrastive constraints. Previous works [10, 66] have shown that integrating global and local information can enhance feature learning. Building upon these findings, we further incorporate local information as well as a voting mechanism to select high-quality pairs. Specifically, we slice the feature map $\phi(x_i)$ for each sample $x_i$ into $q \times q$ regions and the corresponding local representation is obtained as follows:

$$f_L(x_i) = Pooling(\phi(x_i); q \times q), \tag{4}$$

where $f_L(x_i) \in \mathbb{R}^{d \times q \times q}$. Then we calculate the patch-wise similarity of each sample pair at the same location by cosine similarity:

$$s_L^k(x_i, x_j) = \frac{f_L^k(x_i) \cdot f_L^k(x_j)}{\|f_L^k(x_i)\| \|f_L^k(x_j)\|}, \tag{5}$$

where $k$ represents the $k$-th patch in $f_L(x_i)$.

Given that manipulated areas may vary across different forged faces, we further introduce a spatially enhancing mechanism to adjust the priority of patch-wise similarities. MAT [64] has shown that manipulated areas of forged faces tend to have a higher response, while norm-based analysis [36] demonstrates the effectiveness of $L_2$-norm based attention modules. Inspired by these findings, we use $L_2$-norm to reflect the response of local blocks $f_L^k(x_i)$. We first calculate the priority weight of $k$-th patch for each sample $x_i$ as follows:

$$w_i^k = \frac{\|f_L^k(x_i)\|_2}{\sum_{k=1}^{q^2} \|f_L^k(x_i)\|_2}. \tag{6}$$

Combining with spatially enhancing weights, the local similarity $s_L(x_i, x_j)$ is obtained:

$$s_L(x_i, x_j) = \sum_{k=1}^{q^2} w_i^k \cdot s_L^k(x_i, x_j). \tag{7}$$

Next, we propose a voting strategy to take the global and local similarities into consideration. Given an unlabeled sample $x_i^u$, we can find the two most similar samples $\widetilde{x}_i^u$ and $\widehat{x}_i^u$ based on Top-1 global similarity $s_G$ and local similarity $s_L$, respectively. If the results of two Top-1 sample are consistent, *i.e.*, $\widetilde{x}_i^u = \widehat{x}_i^u$, then the pair $(x_i^u, \widetilde{x}_i^u)$ is constrained to be close. For labeled sample $x_i^l$, we randomly select another sample $\widetilde{x}_i^l$ that belongs to the same class $y_i^l$ in the same batch. The ultimate loss function for Global-Local Voting module $\mathcal{L}_{\text{GLV}}$ is formulated as below:

$$\begin{aligned}\mathcal{L}_{\text{GLV}} = &-\frac{1}{n}\sum_{x_i \in \mathcal{D}_l} \log\langle \sigma(f_G(x_i^l)), \sigma(f_G(\widetilde{x}_i^l))\rangle \\ &-\frac{1}{m}\sum_{x_i \in \mathcal{D}_u} \mathbb{I}\left(\widetilde{x}_i^u = \widehat{x}_i^u\right) \log\langle \sigma(f_G(x_i^u)), \sigma(f_G(\widetilde{x}_i^u))\rangle.\end{aligned} \tag{8}$$

## 4.2. Confidence-based Soft Pseudo-labeling Module

With the contrastive learning described above, faces of the same forgery type can be grouped, but some samples with similar manipulated regions may be mixed with other classes without proper supervision. Pseudo-labeling is a feasible solution that uses the predicted category with the highest probability as classification supervision. However, from the study in Figure 3, we found that the second and the third predictions still have a high probability of being the correct class. Therefore, only considering the Top-1 prediction would introduce noisy samples.

Inspired by the study, we propose a Confidence-based Soft Pseudo-labeling module that assigns a pseudo-label for each unlabeled sample based on the output probability of all classes. For each unlabeled sample $x_i^u$, we first obtain the class probability through $p_i^u = \sigma(f_G(x_i^u))$, where $p_i^u \in \mathbb{R}^{|\mathcal{C}_K \cup \mathcal{C}_N|}$. Then we introduce the Gumbel Softmax [30]

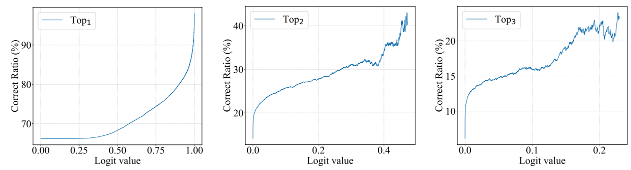



Figure 3. Study of the correlation between the Top-3 value of logits and correct ratio. The results indicate that the second and third predictions still have a high probability of being the correct class.

to generate pseudo-label $\widetilde{y}_i^u$ based on the probability $p_i^u$ as follows:

$$\widetilde{y}_i^u = GumbelSoftmax\left(p_i^u\right). \tag{9}$$

We further use the probability of the pseudo-label as a weight to reduce the impact of pseudo-noise when the probability of the assigned pseudo-label is low, and vice versa. The dynamic weight can be calculated through $\lambda_i^u = p_{ic}^u$, where $c = \arg\max \widetilde{y}_i^u$. Finally, we apply soft pseudo-labels of unlabeled data by cross-entropy loss as follows:

$$\mathcal{L}_{\text{CSP}} = -\frac{1}{m}\sum_{x_i \in \mathcal{D}_u}\sum_{c \in \mathcal{C}_K \cup \mathcal{C}_N} \lambda_i^u \cdot \widetilde{y}_{ic}^u \log p_{ic}^u. \tag{10}$$

## 4.3. Loss Functions and Multi-stage Paradigm

Besides the above constraints, we include two loss functions widely used in semi-supervised learning: a cross-entropy loss for labeled data $\mathcal{L}_{\text{CE}}$, and a regularization term $\mathcal{R}$ to avoid a trivial solution of assigning all instances to the same class, which formulated as follows:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{n}\sum_{x_i \in \mathcal{D}_l}\sum_{c \in \mathcal{C}_K} y_{ic}^l \log p_{ic}^l, \tag{11}$$

$$\mathcal{R} = KL\left(\frac{1}{n+m}\sum_{x_i \in \mathcal{D}_l \cup \mathcal{D}_u} \sigma(f_G(x_i)) \| \mathcal{P}(y)\right), \tag{12}$$

where $p_i^l = \sigma(f_G(x_i^l))$ is class probability and $\mathcal{P}$ denotes a prior probability distribution of labels $y$. The final loss function is given by:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \eta_1 \mathcal{L}_{\text{GLV}} + \eta_2 \mathcal{L}_{\text{CSP}} + \eta_3 \mathcal{R}, \tag{13}$$

with hyper-parameters $\eta_1$, $\eta_2$ and $\eta_3$.

Moreover, previous studies [27, 55] have established the effectiveness of pre-training techniques and iterative learning, hence we extend the CPL framework with a multi-stage paradigm, as outlined in Algorithm 1. In Stage 1, we first pre-train on labeled data to achieve robust performance for known classes. In Stage 2, we apply the CPL approach on labeled and unlabeled data to discover and enhance the representation of novel attacks. In Stage 3, we leverage the Semi-Supervised $k$-means algorithm [55] to cluster unlabeled samples and assign pseudo-labels based on cluster assignments. We then fine-tune the model with labels in both

**Algorithm 1:** Multi-stage Paradigm for OW-DFA.

**Data:** Labeled set $\mathcal{D}_L = (x_i^l, y_i^l)_{i=1}^n$, Unlabeled set $\mathcal{D}_U = (x_i^u)_{i=1}^m$.
**Input:** Feature extractor $\phi(\cdot)$, Classifier $\sigma(\cdot)$, Iteration times $T_1, T_2, T_3$.

```
Initialize φ(·) with ImageNet pre-trained weights;
Initialize σ(·) randomly;
▷ Stage 1: Pre-training on labeled-set
for t in range(T_1) do
    for (x_i^l, y_i^l) ∈ D_L do
        Update φ(·) and σ(·) with Eq. 11;
    end
end
▷ Stage 2: Contrastive Pseudo Learning
for t in range(T_2) do
    for (x_i^l, y_i^l) ∈ D_L, x_i^u ∈ D_U do
        Update φ(·) and σ(·) with Eq. 13;
    end
end
▷ Stage 3: Iterative Learning
S_L = (φ(x_i^l), y_i^l)_{i=1}^n; S_U = (φ(x_i^u))_{i=1}^m;
D̃_U = Semi-Sup k-means(S_L, S_U);
for t in range(T_3) do
    for (x_i^l, y_i^l) ∈ D_L , (x_i^u, ỹ_i^u) ∈ D̃_U do
        Update φ(·) and σ(·) with Eq. 11;
    end
end
return φ, σ
```

labeled set $\mathcal{D}_L$ and generated pseudo-labels set $\widetilde{\mathcal{D}}_U$. With the multi-stage paradigm, we can further improve the attribution performance on the OW-DFA task. More details are provided in Supplementary Sec. C.

# 5. Experiments

**Implementation Details.** We implement the proposed approach via PyTorch. All the models are trained on 1 NVIDIA 3090Ti GPU. We use ResNet-50 [28] pre-trained on ImageNet [16] as our feature extractor, and a fully-connected layer as the classifier. We resize the input image to $256 \times 256$, and train the network with Adam [35] optimizer, a learning rate of $2e^{-4}$, a batch size of 128 and 50 epochs. The learning rate decreases to 0.2 of the original every 10 epochs. We use dlib[1] as the face detector and expand the region by 1.2 times to include more facial information. The temperature $\tau$ in Gumbel Softmax [30] is set to 1. For the Semi-supervised $k$-means [55] used in Stage 3, 10 clusters are initialized using K-Means++ [6], with the tolerance of $1e^{-4}$ and max iteration times of 100.

[1] https://github.com/davisking/dlib

**Evaluation Metrics.** Following [8, 27, 18], we use three metrics to evaluate the performance of all methods on the OW-DFA task, *i.e.* Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand index (ARI). We align the predicted labels with ground-truth labels using the Hungarian algorithm [37]. Unless specified, the results we report are obtained through the CPL framework only.

## 5.1. Benchmark Evaluation

**Compared Methods.** We provide baselines for the OW-DFA task by modifying previous works on GAN attribution [18, 59] and Open-World Semi-Supervised Learning (OW-SSL) [27, 8, 45]. We also include the newly released method NACH [25] in our evaluation. To ensure a fair comparison, we use ResNet-50 [28] as the feature extractor and apply consistent hyperparameters across all approaches. We exclude strong and weak augmentation strategies due to their inapplicability to the OW-DFA task. Additionally, we provide a lower bound based on supervised learning on the labeled set, and an upper bound based on supervised learning on the overall data from both labeled and unlabeled sets. More details are provided in Supplementary Sec. D.

**Results on Protocol-1.** We present the results of Protocol-1 in Table 2, demonstrating that CPL outperforms all GAN attribution methods and OW-SSL methods on both novel and overall classes. These results highlight the effectiveness of CPL, surpassing the previous state-of-the-art method NACH [25] by approximately 1.10-2.74% absolute improvement on different evaluations for novel classes and 1.09% improvement on ACC for overall classes. The lower bound experiment, trained only on labeled data, achieves extremely high accuracy for known categories but exhibits poor generalization. Despite the impact of learning novel attacks on prediction results for known attacks, the prediction accuracy of CPL for the known classes remains higher than most OW-SSL methods and only slightly lower than RankStats [27]. However, there is a significant performance gap between RankStats [27] and CPL on the novel and the overall classes. It is worth noting that DNA-Det [59], a closed-set approach, does not perform well across all classes, as the GAN fingerprints it assumes may not be present in forgery images generated by non-GAN methods. Open-world GAN [18] exceeds the lower bound but does not benefit from semi-supervised learning, limiting the further improvement of its results.

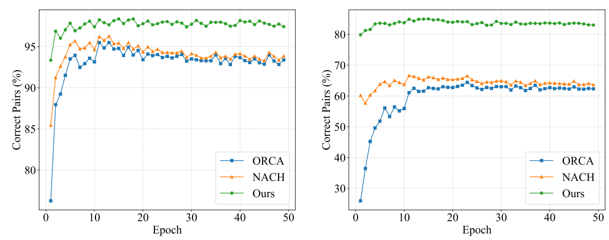
**Results on Protocol-2.** We conduct further experiments on Protocol-2, which incorporates real faces, making the attribution task more challenging and closer to real-world scenarios. Our observations on Protocol-2 are similar to those on Protocol-1, with CPL showing a more significant improvement in the performance of attributing novel and all classes. As shown in Table 2, CPL significantly outperforms NACH [25] and ORCA [8] by approximately 5.02-

Table 2. Benchmark Evaluation on **Protocol-1** and **Protocol-2**.

| Method | Protocol-1: Fake | | | | | | | Protocol-2: Real & Fake | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Known | Novel | | | All | | | Known | Novel | | | All | | |
| | ACC | ACC | NMI | ARI | ACC | NMI | ARI | ACC | ACC | NMI | ARI | ACC | NMI | ARI |
| Lower Bound | **99.96** | 40.96 | 46.43 | 24.05 | 46.90 | 63.18 | 36.35 | **99.80** | 46.48 | 48.44 | 31.49 | 65.73 | 68.91 | 65.75 |
| Upper Bound | 98.21 | 95.36 | 91.57 | 92.14 | 96.68 | 93.94 | 93.59 | 98.57 | 94.15 | 91.93 | 93.11 | 96.83 | 93.80 | 95.05 |
| DNA-Det [59] | 74.47 | 34.82 | 44.22 | 19.35 | 34.99 | 55.55 | 24.89 | 89.13 | 28.44 | 25.97 | 8.18 | 54.37 | 50.10 | 31.45 |
| Openworld-GAN [18] | 99.57 | 38.93 | 45.89 | 41.52 | 57.62 | 57.63 | 47.47 | 99.60 | 46.68 | 53.66 | 45.82 | 69.26 | 58.60 | 61.09 |
| RankStats [27] | 98.58 | 49.94 | 56.05 | 39.76 | 72.49 | 73.63 | 66.49 | 96.84 | 45.26 | 52.44 | 30.17 | 74.39 | 72.21 | 81.66 |
| ORCA [8] | 97.17 | 66.32 | 63.00 | 53.30 | 80.81 | 79.23 | 74.05 | 95.04 | 53.81 | 60.01 | 38.91 | 78.99 | 78.04 | 83.80 |
| OpenLDN [45] | 97.42 | 45.83 | 51.05 | 38.12 | 63.94 | 71.38 | 62.53 | 96.40 | 42.23 | 50.66 | 28.86 | 71.19 | 73.26 | 82.51 |
| NACH [25] | 96.88 | 70.13 | 67.10 | 56.63 | 82.61 | 81.98 | 76.41 | 96.19 | 53.92 | 58.49 | 38.73 | 79.53 | 77.91 | 84.53 |
| **CPL** | 97.50 | **71.89** | **68.20** | **59.37** | **83.70** | **82.31** | **77.64** | 95.64 | **59.92** | **63.90** | **43.75** | **81.10** | **80.23** | **84.99** |

Table 3. Ablation study on each component of CPL on **Protocol-1**. Each component of CPL contributes towards final performance.

| CE | GR | GLV | CSP | Known | Novel | | | All | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | ACC | ACC | NMI | ARI | ACC | NMI | ARI |
| ✓ | | | | **99.96** | 40.96 | 46.43 | 24.05 | 46.90 | 63.18 | 36.35 |
| ✓ | ✓ | | | 97.17 | 66.32 | 63.00 | 53.30 | 80.81 | 79.23 | 74.05 |
| ✓ | | ✓ | | 96.64 | 68.16 | 66.16 | 57.33 | 81.54 | 81.39 | 77.60 |
| ✓ | ✓ | | ✓ | 96.29 | 69.08 | 67.72 | 55.66 | 81.81 | 82.09 | 75.06 |
| ✓ | | ✓ | ✓ | 97.50 | **71.89** | **68.20** | **59.37** | **83.70** | **82.31** | **77.64** |




(a) Known-known pairs (b) Novel-novel pairs

Figure 4. The ratio of correctly selected pairs for known-known pairs and novel-novel pairs with different approaches.

6.00% and 3.89-6.11% respectively on different evaluations of novel classes, while achieving an absolute improvement of 1.57% and 2.11% on ACC for overall classes. Our experiments on Protocol-2 demonstrate that CPL is successful in attributing forged attacks in realistic scenarios containing real data, and is more adept at exploring unlabeled data than existing approaches.

## 5.2. Ablation Study

**Components of CPL.** Our analysis of the different components in CPL is presented in Table 3. We first evaluate the performance with cross-entropy loss only, which is consistent with the lower bound experiment. Next, we assess the impact of GLV loss by comparing it to GR loss. Results show that GLV achieves a significant improvement on novel classes, making it the most critical component of our proposed framework. Although GLV loss sacrifices known class performance, we still achieve a substantial enhancement on overall classes. Building on pairwise similarity learning, we further explore the effect of CSP. Comparing the second and fourth rows, we observe that CSP brings an improvement on NMI of 4.72% and 2.86% for novel and overall classes respectively, indicating that CSP can enhance overall performance. Finally, we combine GLV and CSP to obtain the CPL framework, achieving optimal results on both novel and overall classes. In conclusion, this extensive ablation study empirically validates the effectiveness of different components in CPL.

**Ablation on GLVM.** We analyze the correct rate of pairs obtained by different similarity measurement methods for known classes and unknown classes respectively, as shown in Figure 4. From the results, we can see that the GLV loss proposed by us can significantly improve the correct rate of sample matching, both for the known class and for the unknown class. Even in the early stage of training, we can still achieve a correct rate of $\sim$80% in the accuracy of unknown class matching. Although our recovery rate becomes lower due to the filtering of GLV, it has better results for the overall training by introducing fewer noise samples.

**Ablation on PPLM.** We replace the PPLM in the CPL framework with several pseudo-labeling techniques and show the results in Table 4. We notice that directly assigning labels [39] will introduce noisy samples, resulting in a significant decrease in the overall effect. Dynamic-threshold approaches [62, 57] have certain improvements on known classes, but they tend to ignore samples with high uncertainty, causing low performances on novel classes, and fix-threshold approach [50] also achieves limited improvement. Meanwhile, ST Gumbel Softmax [30] ignores the uncertainty of low-confidence samples. In contrast, our CSP takes all prediction results into account, while reducing the effect of noise by introducing confidence-based weights, achieving the best results on both novel and overall class.

Table 4. Ablation study on pseudo label strategy on **Protocol-1**.

| Pseudo-label Strategy | Known | Novel | | | All | | |
|---|---|---|---|---|---|---|---|
| | ACC | ACC | NMI | ARI | ACC | NMI | ARI |
| GLV | 96.64 | 68.16 | 66.16 | 57.33 | 81.54 | 81.39 | 77.60 |
| GLV + Pseudo-label [39] | 96.61 | 65.55 | 65.18 | 55.66 | 80.14 | 80.78 | 76.78 |
| GLV + FixMatch [50] | 96.14 | 67.21 | 66.18 | 56.33 | 80.81 | 80.82 | 76.33 |
| GLV + FlexMatch [62] | 96.64 | 66.16 | 65.57 | 56.10 | 80.48 | 81.06 | 76.99 |
| GLV + FreeMatch [57] | 97.02 | 67.74 | 67.00 | 56.45 | 81.50 | 81.43 | 76.85 |
| GLV + Gumbel-Softmax [30] | 96.33 | 68.27 | 67.92 | 57.69 | 81.46 | 82.11 | 77.49 |
| **GLV + CSP (CPL)** | **97.50** | **71.89** | **68.20** | **59.37** | **83.70** | **82.31** | **77.64** |

Table 5. Results of multi-stage paradigm on **Protocol-1**.

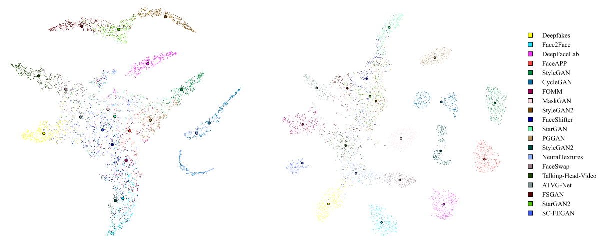
| Stage | Known | Novel | | | All | | |
|---|---|---|---|---|---|---|---|
| | ACC | ACC | NMI | ARI | ACC | NMI | ARI |
| S1-Pretrain | **99.96** | 40.96 | 46.43 | 24.05 | 46.90 | 63.18 | 36.35 |
| S2-CPL | 97.33 | 71.75 | 67.68 | 58.03 | 83.59 | 82.36 | 77.47 |
| S3-IL | 97.08 | **72.78** | **70.87** | **59.10** | **84.20** | **84.05** | **77.58** |

**Multi-stage Paradigm.** We further conduct an ablation study to evaluate the performance of the Multi-stage Paradigm in Algorithm 1 and report the results in Table 5. Stage 1 is exactly the lower bound of our method in Table 2. The results of Stage 2 suggest that initializing the model with pre-trained weights on the labeled set accelerates the semi-supervised learning process, while providing a slight improvement in effectiveness compared to direct training based on weights pre-trained on ImageNet [16]. For the iterative learning in Stage 3, we use Semi-supervised K-Means [55] to generate pseudo labels and apply fine-tuning on the previous model, further improving the performance of the model on both novel and overall classes. Table 5 demonstrates that each stage in the paradigm contributes to the high performance of our method.

**t-SNE Visualization.** In order to compare the performance of CPL more intuitively, we performed t-SNE [43] visualization for Open-World GAN [18] and CPL in Figure 5. We observe that CPL has greatly improved the clustering performance compared to Open-World GAN [18]. Given the satisfactory results of known classes, CPL is capable of isolating novel classes of lower difficulty into distinct classes, *e.g.*, StyleGAN2 [34] and SC-FEGAN [31]. For novel attacks of higher difficulty, CPL is also effective in clustering these samples. On the other hand, the gap between different attack types is significantly larger, even for data within the same dataset, *e.g.*, Deepfakes [2], FaceSwap [4] and NeuralTextures [5] in FF++ [47], and this can be attributed to the fact that CPL concentrates on combining patch-wise local similarity with global similarity.

### 5.3. Real/Fake Detection

To further verify the significance of the deepfake attribution task for deepfake detection, we conduct additional experiments for comparison based on Protocol-2. We compare the results of three approaches: a) deepfake binary classification, b) deepfake multi-classification, and c) our CPL framework. Approaches a) and b) are trained on labeled data, while the CPL framework utilizes both labeled and unlabeled data. We construct unlabeled sets using various combinations of known, new fake, and new real faces. The AUC results of these approaches are evaluated and presented in Table 6. We observe that the performance of b) is consistently higher than that of a), especially when new images are introduced. Compared to b) and c), the CPL framework achieves a significant improvement with $\sim$9.5% AUC on new fake and new real set. These results clearly illustrate that the introduction of the deepfake attribution task can further enhance the security of the deepfake detection task.




(a) Openworld-GAN[18] (b) CPL

Figure 5. t-SNE visualization on **Protocol-1**.

Table 6. AUC results for real/fake detection on **Protocol-2**.

| Data | | | Approach | | |
|---|---|---|---|---|---|
| Known | New Fake | New Real | a) Binary | b) Multi | c) CPL |
| ✓ | | | 99.95 | 100.00 | **100.00** |
| ✓ | ✓ | | 93.06 | 94.80 | **99.91** |
| ✓ | ✓ | ✓ | 94.99 | 95.84 | **96.28** |
| | ✓ | ✓ | 73.91 | 76.47 | **85.97** |

## 6. Conclusion

We introduce a novel benchmark, Open-World DeepFake Attribution (OW-DFA), which aims to enhance attribution performance against various types of fake faces in open-world scenarios. Our proposed framework, Contrastive Pseudo Learning (CPL), introduces a Global-Local Voting module to guide the inter-sample relations of forged faces with different manipulated regions. A probability-based pseudo-label strategy is also employed to mitigate the pseudo-noise caused by similar attack methods. Furthermore, we extend the CPL framework with a multi-stage paradigm that incorporates pre-training techniques and iterative learning to further improve traceability performance. Extensive experiments demonstrate the superiority of CPL on the OW-DFA benchmark. We also highlight the interpretability and security of the DFA task and its impact on the deepfake detection field.

## Acknowledgements

This project is supported by National Natural Science Foundation of China (No.72192821,61972157,62272447), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Science and Technology Commission (21511101200), Shanghai Sailing Program (22YF1420300), CCF-Tencent Open Research Fund (RAGR20220121), Young Elite Scientists Sponsorship Program by CAST (2022QNRC001), Beijing Natural Science Foundation (L222117), the Fundamental Research Funds for the Central Universities (YG2023QNB17)

## References

[1] Deepfacelab. https://github.com/iperov/DeepFaceLab. Accessed: 2023-2-28. 1, 3

[2] Deepfakes. https://github.com/deepfakes/faceswap. Accessed: 2023-2-28. 1, 3, 8

[3] Faceapp. https://faceapp.com/app. Accessed: 2023-2-28. 3

[4] Faceswap. https://github.com/MarekKowalski/FaceSwap/. Accessed: 2023-2-28. 3, 8

[5] Neuraltextures. https://github.com/SSRSGJYD/NeuralTexture. Accessed: 2023-2-28. 1, 3, 8

[6] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006. 6

[7] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2022. 2

[8] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *International Conference on Learning Representations*, 2022. 2, 4, 6, 7

[9] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7832–7841, 2019. 3

[10] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1081–1088, 2021. 2, 4

[11] Zhaoyu Chen, Bo Li, Shuang Wu, Jianghe Xu, Shouhong Ding, and Wenqiang Zhang. Shape matters: deformable patch attack. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 529–548. Springer, 2022. 2

[12] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 3

[13] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 3

[14] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 2

[15] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 5781–5790, 2020. 3

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6, 8

[17] Shichao Dong, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. Explaining deepfake detection by analysing image matching. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, pages 18–35. Springer, 2022. 1

[18] Sharath Girish, Saksham Suri, Sai Saketh Rambhatla, and Abhinav Shrivastava. Towards discovery and attribution of open-world gan generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14094–14103, 2021. 2, 3, 6, 7, 8

[19] Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 735–743, 2022. 2

[20] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3473–3481, 2021. 2

[21] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, and Lizhuang Ma. Delving into the local: Dynamic inconsistency learning for deepfake video detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 744–752, 2022. 2

[22] Zhihao Gu, Taiping Yao, Yang Chen, Shouhong Ding, and Lizhuang Ma. Hierarchical contrastive inconsistency learning for deepfake video detection. In *European Conference on Computer Vision*, pages 596–613. Springer, 2022. 2

[23] Zhihao Gu, Taiping Yao, C Yang, Ran Yi, Shouhong Ding, and Lizhuang Ma. Region-aware temporal inconsistency learning for deepfake video detection. In *Proceedings of the 31th International Joint Conference on Artificial Intelligence*, volume 1, 2022. 2

[24] Luca Guarnera, Oliver Giudice, Matthias Nießner, and Sebastiano Battiato. On the exploitation of deepfake model recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 61–70, 2022. 1, 2, 3

[25] Lan-Zhe Guo, Yi-Ge Zhang, Zhi-Fan Wu, Jie-Jing Shao, and Yu-Feng Li. Robust semi-supervised learning when not all

classes have labels. In *Advances in Neural Information Processing Systems*, 2022. 2, 6, 7

[26] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049, 2021. 1

[27] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 4, 5, 6, 7

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[29] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4360–4369, 2021. 1, 3

[30] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. 5, 6, 7, 8

[31] Youngjoo Jo and Jongyoul Park. Sc-fegan: Face editing generative adversarial network with user's sketch and color. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 3, 8

[32] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 1, 3

[33] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[34] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 1, 3, 8

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[36] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 7057–7075. Association for Computational Linguistics (ACL), 2020. 5

[37] HW Kuhn et al. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. 6

[38] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[39] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 7, 8

[40] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 3

[41] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020. 3

[42] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7184–7193, 2019. 3

[43] Pavlin G. Poličar, Martin Stražar, and Blaž Zupan. opentsne: a modular python library for t-sne dimensionality reduction and embedding. *bioRxiv*, 2019. 8

[44] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*, pages 86–103. Springer, 2020. 2

[45] Mamshad Nayeem Rizve, Navid Kardan, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Openldn: Learning to discover novel classes for open-world semi-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 382–401. Springer, 2022. 2, 6, 7

[46] Mamshad Nayeem Rizve, Navid Kardan, and Mubarak Shah. Towards realistic semi-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 437–455. Springer, 2022. 2

[47] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 3, 8

[48] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022. 1

[49] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 1, 3

[50] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 2, 7, 8

[51] Ke Sun, Hong Liu, Taiping Yao, Xiaoshuai Sun, Shen Chen, Shouhong Ding, and Rongrong Ji. An information theoretic approach for attention-driven face forgery detection. In

*European Conference on Computer Vision*, pages 111–127. Springer, 2022. 2

[52] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2316–2324, 2022. 2

[53] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 2

[54] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 3

[55] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7492–7501, 2022. 2, 5, 6, 8

[56] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 1

[57] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Zhen Wu, and Jindong Wang. Freematch: Self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations*, 2023. 2, 7, 8

[58] Yukai Wang, Chunlei Peng, Decheng Liu, Nannan Wang, and Xinbo Gao. Forgerynir: deep face forgery and detection in near-infrared scenario. *IEEE Transactions on Information Forensics and Security*, 17:500–515, 2022. 3

[59] Tianyun Yang, Ziyao Huang, Juan Cao, Lei Li, and Xirong Li. Deepfake network architecture attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4662–4670, 2022. 1, 2, 3, 6, 7

[60] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7556–7566, 2019. 2, 3

[61] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 14448–14457, 2021. 1, 2, 3

[62] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 2, 7, 8

[63] Sibo Zhang, Jiahong Yuan, Miao Liao, and Liangjun Zhang. Text2video: Text-driven talking-head video synthesis with phonetic dictionary. *arXiv preprint arXiv:2104.14631*, 2021. 3

[64] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. 2, 5

[65] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning to recognize patch-wise consistency for deepfake detection. *arXiv preprint arXiv:2012.09311*, 6, 2020. 2

[66] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15023–15033, 2021. 4

[67] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 3