Contrastive Pseudo Learning for Open-World DeepFake Attribution

Zhimin Sun\textsuperscript{1,2,3*}, Shen Chen\textsuperscript{2*}, Taiping Yao\textsuperscript{2}, Bangjie Yin\textsuperscript{2}, Ran Yi\textsuperscript{1†}, Shouhong Ding\textsuperscript{2†}, Lizhuang Ma\textsuperscript{1}
\textsuperscript{1}Shanghai Jiao Tong University, \textsuperscript{2}Tencent YouTu Lab, \textsuperscript{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 deep fake 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 \cite{2}, the malicious usage of fake content on social media has raised public concerns about face security and privacy. Dedicated research efforts \cite{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 attributing the source models of fake faces, has also significantly drawn widespread attentions \cite{39, 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 \cite{29, 17}.

Early approaches of sourcing attribution \cite{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 \cite{1, 2}, expression transfer \cite{5, 49}, attribute manipulation \cite{12, 13} and entire face synthesis \cite{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.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{owdfa.png}
\caption{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.}
\end{figure}
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 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.
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.

<table>
<thead>
<tr>
<th>Face Type</th>
<th>Labeled Sets</th>
<th>Unlabeled Sets</th>
<th>Source Dataset</th>
<th>Method</th>
<th>Tag</th>
<th>Labeled #</th>
<th>Unlabeled #</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ForgeryNet [29]</td>
<td>DeepFaceLab FaceShifter FSGAN</td>
<td></td>
<td></td>
<td>Known</td>
<td>1500</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ForgeryNet</td>
<td>FOMM ATVG-Net Talking-Head-Video</td>
<td></td>
<td></td>
<td>Novel</td>
<td>1500</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DFFD [15]</td>
<td>FaceAPP StarGAN</td>
<td></td>
<td></td>
<td>Novel</td>
<td>1500</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entire Face Synthesis</td>
<td>StyleGAN CycleGAN PGGAN StyleGAN2 [34]</td>
<td>StyleGAN CycleGAN PGGAN StyleGAN2</td>
<td>ForgeryNet</td>
<td>StyleGAN Novel</td>
<td>Known</td>
<td>1500</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DFFD</td>
<td>StyleGAN PGGAN</td>
<td></td>
<td></td>
<td>Novel</td>
<td>1500</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ForgeryNIR</td>
<td>CycleGAN StyleGAN2</td>
<td></td>
<td></td>
<td>Novel</td>
<td>1500</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real Face</td>
<td>Youtube-Real [47] Celeb-Real [41]</td>
<td>Youtube-Real Celeb-Real</td>
<td>FaceForensics++</td>
<td></td>
<td>Known</td>
<td>150000</td>
<td>5000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CelebDFv2 [41]</td>
<td></td>
<td></td>
<td></td>
<td>Novel</td>
<td>5000</td>
<td></td>
</tr>
</tbody>
</table>

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 \( D_l = \{ (x_i, y_i) \}_{i=1}^n \) and an unlabeled set \( D_u = \{ (x_i) \}_{i=1}^m \). We denote the classes in the labeled set as \( C_L \), and those in the unlabeled set as \( C_U \), where \( C_L \) contains only known categories, while \( C_U \) covers both known and novel categories, i.e., \( C_L \cap C_U \neq \emptyset \) and \( C_L \neq C_U \). We denote the known class as \( C_K = C_L \cap C_U \) and the novel class as \( C_N = C_U \setminus C_L \). The goal of OW-DFA task is utilizing both labeled sets \( D_l \) and unlabeled sets \( D_u \) to learn a feature extractor \( \phi(\cdot) \) and a classifier \( \sigma(\cdot) \), which can recognize source models for various fake 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.

20884
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 unlabelled 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 $y_i$ 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:

$$f_G(x_i) = \text{Pooling}(\phi(x_i); 1 \times 1),$$  \hspace{1cm} (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 intersample relation is measured by the cosine similarity of their global features:

$$s_G(x_i, x_j) = \frac{f_G(x_i) \cdot f_G(x_j)}{\|f_G(x_i)\| \|f_G(x_j)\|}.$$  \hspace{1cm} (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 $\bar{x_i}$ by a variant of BCE loss, i.e. global relation constraints:

$$\mathcal{L}_{GR} = -\frac{1}{n + m} \sum_{x_i \in D_l \cup D_u} \log(\sigma(f_G(x_i)), \sigma(f_G(\bar{x_i}))),$$  \hspace{1cm} (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:

$$f_L(x_i) = \text{Pooling}(\phi(x_i); q \times q),$$  \hspace{1cm} (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:

$$s^k_L(x_i, x_j) = \frac{f^k_L(x_i) \cdot f^k_L(x_j)}{\|f^k_L(x_i)\| \|f^k_L(x_j)\|},$$  \hspace{1cm} (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 $f^k_L(x_i)$. We first calculate the priority weight of $k$-th patch for each sample $x_i$ as follows:

$$w^k_i = \frac{\|f^k_L(x_i)\|_2}{\sum_{k=1}^{q^2} \|f^k_L(x_i)\|_2}.$$  

(6)

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

$$s_L(x_i, x_j) = \sum_{k=1}^{q^2} w^k_i \cdot s^k_L(x_i, x_j).$$  

(7)

Next, we propose a voting strategy to take the global and local similarities into consideration. Given an unlabeled sample $x^u_i$, we can find the two most similar samples $\tilde{x}^u_i$ and $\tilde{x}^u_i$ 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., $\tilde{x}^u_i = \tilde{x}^u_i$, then the pair $(x^u_i, \tilde{x}^u_i)$ is constrained to be close. For labeled sample $x^l_i$, we randomly select another sample $\tilde{x}^l_i$ that belongs to the same class $y^l_i$ in the same batch. The ultimate loss function for Global-Local Voting module $L_{GLV}$ is formulated as below:

$$L_{GLV} = -\frac{1}{n} \sum_{x_i \in D_l} \log \langle \sigma(f_G(x_i^l)), \sigma(f_G(\tilde{x}^l_i)) \rangle - \frac{1}{m} \sum_{x_i \in D_u} \log \langle \sigma(f_G(x^u_i)), \sigma(f_G(\tilde{x}^u_i)) \rangle.$$  

(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 $x^u_i$, we first obtain the class probability through $p^u_i = \sigma(f_G(x^u_i))$, where $p^u_i \in \mathbb{R}^{[C_N \cup C_N]}$. Then we introduce the Gumbel Softmax [30] to generate pseudo-label $\tilde{y}^u_i$ based on the probability $p^u_i$ as follows:

$$\tilde{y}^u_i = \text{GumbelSoftmax}(p^u_i).$$  

(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 $\lambda^u_i = \hat{p}^u_i c$, where $c = \arg \max \tilde{y}^u_i$. Finally, we apply soft pseudo-labels of unlabeled data by cross-entropy loss as follows:

$$L_{CSP} = -\frac{1}{n} \sum_{x_i \in D_u} \sum_{c \in C_N} \lambda^u_i \cdot \hat{y}^u_{ic} \log \hat{p}^u_{ic}.$$  

(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 $L_{CE}$, and a regularization term $R$ to avoid a trivial solution of assigning all instances to the same class, which formulated as follows:

$$L_{CE} = -\frac{1}{n} \sum_{x_i \in D_l} \sum_{c \in C_K} y^l_{ic} \log p^l_{ic},$$  

(11)

$$R = KL \left( \frac{1}{n+m} \sum_{x_i \in D_l \cup D_u} \sigma(f_G(x_i))||P(y) \right),$$  

(12)

where $p^l_{ic} = \sigma(f_G(x^l_i))$ is class probability and $P$ denotes a prior probability distribution of labels $y$. The final loss function is given by:

$$L = L_{CE} + \eta_1 L_{GLV} + \eta_2 L_{CSP} + \eta_3 R,$$  

(13)

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

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
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-

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

Data: Labeled set $D_L = (x^l_i, y^l_i)_{i=1}^m$, Unlabeled set $D_U = (x^u_i)_{i=1}^m$.
Input: Feature extractor $\phi(\cdot)$, Classifier $\sigma(\cdot)$, iteration times $T_1, T_2, T_3$.
1 Initialize $\phi(\cdot)$ with ImageNet pre-trained weights;
2 Initialize $\sigma(\cdot)$ randomly;
3 Stage 1: Pre-training on labeled-set
4 for $t$ in range $(T_1)$ do
5 for $(x^l_i, y^l_i) \in D_L$ do
6 Update $\phi(\cdot)$ and $\sigma(\cdot)$ with Eq. 11;
7 end
8 end
9 Stage 2: Contrastive Pseudo Learning
10 for $t$ in range $(T_2)$ do
11 for $(x^l_i, y^l_i) \in D_L, x^u_i \in D_U$ do
12 Update $\phi(\cdot)$ and $\sigma(\cdot)$ with Eq. 13;
13 end
14 end
15 Stage 3: Iterative Learning
16 $S_L = (\phi(x^l_i), y^l_i)_{i=1}^m; S_U = (\phi(x^u_i))_{i=1}^m$;
17 $D_U = \text{Semi-Sup} k$-means$(S_L, S_U)$;
18 for $t$ in range $(T_3)$ do
19 for $(x^l_i, y^l_i) \in D_L, (x^u_i, y^u_i) \in D_U$ do
20 Update $\phi(\cdot)$ and $\sigma(\cdot)$ with Eq. 11;
21 end
22 end
23 return $\phi, \sigma$

labeled set $D_L$ and generated pseudo-labels set $D_U$. 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 dlib1 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.

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.

---

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

---
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, while achieving an absolute improvement in the accuracy of unknown classes. 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 certainty of low-confidence samples. In contrast, our CSP takes all prediction results into account, while reducing the certainty of low-confidence samples.

### Table 3. Ablation study on each component of CPL on Protocol-1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Known ACC</th>
<th>Novel ACC</th>
<th>All ACC</th>
<th>Known NMI</th>
<th>Novel NMI</th>
<th>All NMI</th>
<th>Known ARI</th>
<th>Novel ARI</th>
<th>All ARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower Bound</td>
<td>99.96</td>
<td>40.96</td>
<td>46.43</td>
<td>24.05</td>
<td>46.90</td>
<td>63.18</td>
<td>36.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper Bound</td>
<td>99.21</td>
<td>95.36</td>
<td>91.57</td>
<td>92.14</td>
<td>96.88</td>
<td>93.94</td>
<td>93.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA-Net [59]</td>
<td>94.47</td>
<td>74.32</td>
<td>43.22</td>
<td>79.35</td>
<td>64.99</td>
<td>53.35</td>
<td>24.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Openword-GAN [18]</td>
<td>90.13</td>
<td>28.44</td>
<td>32.97</td>
<td>18.15</td>
<td>34.31</td>
<td>56.16</td>
<td>31.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RankStats [23]</td>
<td>98.58</td>
<td>49.94</td>
<td>56.08</td>
<td>59.76</td>
<td>74.29</td>
<td>73.63</td>
<td>66.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORCA [8]</td>
<td>97.17</td>
<td>66.32</td>
<td>63.00</td>
<td>53.30</td>
<td>80.81</td>
<td>79.23</td>
<td>74.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OpenLDN [45]</td>
<td>97.42</td>
<td>45.83</td>
<td>51.05</td>
<td>38.12</td>
<td>63.94</td>
<td>71.38</td>
<td>62.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NACH [25]</td>
<td>96.88</td>
<td>70.13</td>
<td>67.10</td>
<td>56.63</td>
<td>82.61</td>
<td>81.98</td>
<td>76.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CPL</strong></td>
<td>97.36</td>
<td>71.89</td>
<td>68.20</td>
<td>59.37</td>
<td>83.70</td>
<td>82.31</td>
<td>77.64</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Known ACC</td>
<td>Novel ACC</td>
<td>All ACC</td>
<td>Known NMI</td>
</tr>
<tr>
<td>Lower Bound</td>
<td>99.96</td>
<td>40.96</td>
<td>46.43</td>
<td>24.05</td>
</tr>
<tr>
<td>Upper Bound</td>
<td>99.21</td>
<td>95.36</td>
<td>91.57</td>
<td>92.14</td>
</tr>
<tr>
<td>DNA-Net [59]</td>
<td>94.47</td>
<td>74.32</td>
<td>43.22</td>
<td>79.35</td>
</tr>
<tr>
<td>Openword-GAN [18]</td>
<td>90.13</td>
<td>28.44</td>
<td>32.97</td>
<td>18.15</td>
</tr>
<tr>
<td>RankStats [23]</td>
<td>98.58</td>
<td>49.94</td>
<td>56.08</td>
<td>59.76</td>
</tr>
<tr>
<td>ORCA [8]</td>
<td>97.17</td>
<td>66.32</td>
<td>63.00</td>
<td>53.30</td>
</tr>
<tr>
<td>OpenLDN [45]</td>
<td>97.42</td>
<td>45.83</td>
<td>51.05</td>
<td>38.12</td>
</tr>
<tr>
<td>NACH [25]</td>
<td>96.88</td>
<td>70.13</td>
<td>67.10</td>
<td>56.63</td>
</tr>
<tr>
<td><strong>CPL</strong></td>
<td>97.36</td>
<td>71.89</td>
<td>68.20</td>
<td>59.37</td>
</tr>
</tbody>
</table>

Figure 4. The ratio of correctly selected pairs for known-known pairs and novel-novel pairs with different approaches.
Table 4. Ablation study on pseudo label strategy on Protocol-1.

<table>
<thead>
<tr>
<th>Pseudo-label Strategy</th>
<th>Known</th>
<th>Novel</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACC</td>
<td>NMI</td>
<td>ARI</td>
</tr>
<tr>
<td>GLV</td>
<td>96.64</td>
<td>68.16</td>
<td>66.16</td>
</tr>
<tr>
<td>GLV + FixMatch [58]</td>
<td>96.14</td>
<td>67.21</td>
<td>66.18</td>
</tr>
<tr>
<td>GLV + FaceSwap [62]</td>
<td>97.02</td>
<td>67.74</td>
<td>68.00</td>
</tr>
<tr>
<td>GLV + FreeMatch [57]</td>
<td>96.33</td>
<td>68.27</td>
<td>67.92</td>
</tr>
<tr>
<td>GLV + Gumbel-Softmax [30]</td>
<td>96.33</td>
<td>68.27</td>
<td>67.92</td>
</tr>
<tr>
<td>GLV + OSP [4]</td>
<td>97.50</td>
<td>71.38</td>
<td>68.20</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Stage</th>
<th>Known</th>
<th>Novel</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACC</td>
<td>NMI</td>
<td>ARI</td>
</tr>
<tr>
<td>S1-Pretrain</td>
<td>99.96</td>
<td>40.96</td>
<td>46.43</td>
</tr>
<tr>
<td>S2-CPL</td>
<td>97.33</td>
<td>71.75</td>
<td>67.68</td>
</tr>
<tr>
<td>S3-IL</td>
<td>97.08</td>
<td>72.78</td>
<td>70.87</td>
</tr>
</tbody>
</table>

**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 [3] and NeuralTextures [5] in FF++, 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 ~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.

**6. Conclusion**

We introduce a novel benchmark, Open-World Deep-Fake 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

[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
[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 \textit{Advances in Neural Information Processing Systems}, 2022. 2, 6, 7


[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


