Progressive Open Space Expansion for Open-Set Model Attribution

Tianyun Yang\textsuperscript{1,2}, Danding Wang\textsuperscript{1,2,*}, Fan Tang\textsuperscript{1,2}, Xinying Zhao\textsuperscript{1,2}, Juan Cao\textsuperscript{1,2}, Sheng Tang\textsuperscript{1,3}
\textsuperscript{1} Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
\textsuperscript{2}University of Chinese Academy of Sciences, Beijing, China
\textsuperscript{3}Research Institute of Intelligent Computing, Zhejiang Lab, Hangzhou, China
\{yangtianyun19z,wangdanding,tangfan,zhaoxinying21s,caojuan,ts\}@ict.ac.cn

Abstract

Despite the remarkable progress in generative technology, the Janus-faced issues of intellectual property protection and malicious content supervision have arisen. Efforts have been paid to manage synthetic images by attributing them to a set of potential source models. However, the closed-set classification setting limits the application in real-world scenarios for handling contents generated by arbitrary models. In this study, we focus on a challenging task, namely Open-Set Model Attribution (OSMA), to simultaneously attribute images to known models and identify those from unknown ones. Compared to existing open-set recognition (OSR) tasks focusing on semantic novelty, OSMA is more challenging as the distinction between images from known and unknown models may only lie in visually imperceptible traces. To this end, we propose a Progressive Open Space Expansion (POSE) solution, which simulates open-set samples that maintain the same semantics as closed-set samples but embedded with different imperceptible traces. Guided by a diversity constraint, the open space is simulated progressively by a set of lightweight augmentation models. We consider three real-world scenarios and construct an OSMA benchmark dataset, including unknown models trained with different random seeds, architectures, and datasets from known ones. Extensive experiments on the dataset demonstrate POSE is superior to both existing model attribution methods and off-the-shelf OSR methods. Github: https://github.com/ICTMCG/POSE

1. Introduction

Advanced generative modeling technology can create extremely realistic visual content, leading to dramatic changes in the field of AI-enhanced design, arts, and metaverse\cite{23, 41, 47}. Whereas, the broadcasting of malicious content generated by open source generation models has brought severe social impacts\cite{9, 18, 43}. Furthermore, new challenges have arisen for the ownership protection of copyrighted digital generative models. To solve these problems, model attribution, \textit{i.e.}, identifying the source model of generated contents, has drawn increasing attention recently\cite{5, 33, 48, 50, 53}.

Marra \textit{et al.}\cite{33} are among the first to point out that GAN models leave specific fingerprints in the generated images, just like camera devices. Further researches\cite{5, 14, 48, 50, 53} validate the existence of GAN fingerprints and show the feasibility of attributing fake images to a fixed and finite set of known models. However, most of these works focus on finding discriminative fingerprints among the contents generated by different GAN models following a simple closed-set setup. The ever-growing number of unseen source models in the real-world scenario appeal for a more generic approach. In this paper, we focus on
the problem of Open-Set Model Attribution (OSMA), i.e., simultaneously attributing images to known source models and identifying those from unknown ones.

An intuitive way to solve OSMA is to apply off-the-shelf open-set recognition (OSR) approaches to closed-set model attribution classifiers. Traditional OSR methods leverage the output logits to either reject or categorize the input images [2, 44]. However, following the discriminative line, the performance highly depends on the closed-set classifier [46]. The learned feature space is not rectified for open-set samples. Another mainstream of OSR methods is based on simulating open-set samples or features [7, 27, 35, 36, 58]. By the simulation of open space, the learned feature space is more compact for closed-set categories [58], leading the detection of unknown samples more valid. Nevertheless, existing works only leverage a single generator [7, 27, 35] or mechanism [58] for simulating open-set samples or features, which are not diverse enough to reduce the open space risk for OSMA. A generator could produce open-set samples of different semantics, but its fingerprint is fixed and thus not suitable for the expansion of open space.

In this study, we propose Progressive Open Space Expansion (POSE) tailored for open-set model attribution, which simulates the potential open space of challenging unknown models through involving a set of augmentation models progressively. For augmentation model construction, it can be daunting to consider all types of unknown models with a variety of architectures. Instead, lightweight networks composed of a few convolution layers are employed. They serve as “virtual” follow-up blocks of known models, augmenting closed-set samples to open-set samples surrounding them by modifying their fingerprints with reconstruction residuals. Despite the simple structure, these augmentation models show the potential to model traces of surrounding them by modifying their fingerprints with reconstruction residuals. Despite the simple structure, these augmentation models show the potential to model traces of unknown models. To enrich the simulated open space, multiple augmentation models are involved. Instead of training them independently, we design a progressive training mechanism to ensure the diversity of simulated open space across models in a computation-effective way.

To validate the effectiveness of POSE in the open world, we construct a benchmark dataset considering three challenging unknown scenarios as shown in Figure 1, which includes unknown models trained with either a different random seed, architecture or dataset from known models. Extensive experiments on the benchmark demonstrate POSE is superior to both existing GAN attribution methods and OSR methods. In summary, our contributions are:

- We tackle an important challenge for applying model attribution to open scenarios, the open-set model attribution problem, which attributes images to known models and identifies images from unknown ones.
- We propose a novel solution named POSE, which simulates the potential open space of unknown models progressively by a set of lightweight augmentation models, and consequently reduces open space risk.
- We construct an OSMA benchmark simulating the real-world scenarios, on which extensive experiments prove the superiority of POSE compared with existing GAN attribution methods and off-the-shelf OSR methods.

2. Related Work

Open-Set Recognition. Works on open-set recognition mainly follow a discriminative or generative line. The typical routine from the discriminative perspective first trains a K-way classifier on the closed set and then calibrates the confidence to a reasonable distribution for open-set discrimination [2, 44]. However, without rectification to the feature space, the performance depends heavily on the closed-set classifier [46]. Generative-based approaches employ generative models for two purposes: 1) to threshold the test-time reconstruction error as an open-set indicator [17, 38, 39, 45, 51, 57], or 2) to simulate the potential open space for augmenting classifier training [7, 8, 15, 27, 35, 36, 58]. For the open-set model attribution problem, it is sub-optimal to follow the reconstruction approach, as the intrinsic differences between samples generated by known and unknown models mainly lie in visually imperceptible traces, which are too subtle to be thresholded. For the second line, mixup mechanism [58], GAN [7, 27, 35], and auto-encoder [36] are leveraged for open space simulation. However, the open space risk for OSMA could be hardly reduced by a single generator or mechanism. The generators are proposed to generate images with different semantics, unable to produce diverse open-set fingerprints by its fixed weights and architecture. In this study, we simulate a rich open space for potential unknown models by involving a set of augmentation models progressively, and reduce the open space risk better.

Model Attribution. Model attribution aims to identify the source model of generated contents. Existing works tackle model attribution through an active [26, 54, 55] or passive manner [14, 16, 20, 33, 48, 49, 53]. Active methods embed artificial fingerprints [54, 55] or keys [26] into the generative model during the training process and then decouple them for source attribution in testing. Despite their capability of protecting particular models, requiring access to the training process limits their application in handling fake images generated by arbitrary models. With the verification of the existence of GAN fingerprints [33, 53], passive model attribution methods [14, 16, 20, 33, 48, 49, 53] are proposed by capturing the intrinsic differences in GAN fingerprints. Typically, a classifier is trained to distinguish images generated from a fixed and finite set of models, yet not applicable in the real world of an ever-growing number of unknown models. DNA-Det [50] and RepMix [5].
architectures by relaxing the attribution objective to the model architecture instead of the exact model. But they still do not consider unknown models with unseen architectures. Girish et al. [16] pioneers to consider open-world GAN attribution and discovery. However, their method follows a general semi-supervised learning pipeline, overlooking the intrinsic challenge of imperceptible trace discrimination. Our proposed POSE is tailored for this challenge and achieves better performance in not only open-set model attribution but also unknown model clustering.

3. Method

We tackle the OSMA problem by simulating the potential open space of unknown models around the boundary of known space through a set of augmentation models. At a high level, for open space simulation, two key components are 1) the simulation tool, i.e., how to construct the augmentation models, as it determines the scope of open space that could be expanded, and 2) the expansion mechanism. Given the augmentation models, the expansion mechanism decides how to increase the number of them to expand the open space in an efficient direction. In the following, we describe the simulation tool (Section 3.1) and expansion mechanism (Section 3.2) in detail.

3.1. Augmentation Model Construction

To simulate the space of unknown models, an intuitive way is to train a number of augmentation models approximating real unknown models. However, it is daunting to consider all types of unknown models of a variety of architectures. Empirically, we construct the augmentation models as lightweight networks by optimizing a reconstruction loss. Severing as “virtual” follow-up blocks of known models, augmentation models receive images from known models, maintain their semantics but modify the traces with reconstruction residuals. In this way, the augmented images would carry hybrid fingerprints of known classes and augmentation models as shown in Figure 3(b).

For the specific architecture choice, the candidate layer components could be upsampling and convolution. Upsampling layer would heavily influence high frequencies [6] and produce trivial open-set samples, which is thus not ideal to be included. As shown in [14], distinctions in GAN fingerprints exist in the frequency components. Equipped with rich learnable kernels, convolution layers are capable of modifying the frequencies, either suppressing or amplifying frequencies flexibly depending on the optimization objective [6]. Besides, [55] achieves GAN network fingerprinting by modulating convolution weights with a 128-bit fingerprint, reaching a capacity of around $10^{16}$ fingerprinted models in principle. Given these works suggesting the effect of convolution layers on GAN fingerprints, we use tiny networks with two convolution layers as augmentation models. Detailed ablation study is in Section 4.3.

Feasibility of tiny networks to simulate unknown model traces. It is non-trivial to measure the scope of open space that could be expanded. Motivated by existing work [14] showing that frequency spectrum contains rich information of model fingerprints, we adopt a simple but characteristic 1D representation of the Fourier power spectrum, azimuthal integration (AI) over the power spectrum [13], as a handcrafted criterion for the fingerprint space. Suppose $M_1$ is a known model and $M_2$ is a potential unknown model, to testify whether there exists an augmentation model $A$ able to expand the open space around model $M_1$ to cover model $M_2$ according to the criterion, $L_{\text{spectral}}$ is optimized:

$$L_{\text{spectral}}(\theta) = \|x - A(x)\| + \lambda\|\text{Freq}(A(x_1)) - \text{Freq}(x_2)\|,$$

(1)

where $\theta$ is the parameters of $A$. $x_1$, $x_2$ and $x$ are images generated by model $M_1$, model $M_2$, and all images. $\text{Freq}(\cdot)$ is spectrum azimuthal integration. $\lambda$ is to balance two parts of $L_{\text{spectral}}$. The first part is a pixel reconstruction loss constraining the reconstruction residual, and the second part is calculated in the frequency domain, converting the azimuthal integral spectrum of model $A$ to model $B$.

As shown in Figure 2, by optimizing $L_{\text{spectral}}$, augmentation models are able to transform the spectrum of one model to another model of different architecture (e.g.,
StyleGAN3 $\rightarrow$ StyleGAN2 or of different training dataset (e.g., StyleGAN.Bus $\rightarrow$ StyleGAN_Cat) with visually imperceptible traces. These results suggest the potential of lightweight networks to simulate rich unknown traces. We also emphasize that unknown models are actually out of access in the real world. Our solution is to utilize a set of augmentation models to cover as much open space of potential unknown models as possible around the boundary of known space through a progressive training manner.

3.2. Progressive Open Space Expansion

Figure 3 shows the framework of POSE. Our model consists of a task model and $N$ augmentation models progressively involved. The task model $M$ conducts open-set recognition task, including a feature extractor $F : \mathcal{X} \rightarrow \mathcal{Z}$ mapping images from the input space to the embedding space, and a classification head $H : \mathcal{Z} \rightarrow \mathcal{Y}$ for $K$-way known class classification, where $K$ is the number of known classes. The augmentation models $\{A_i\}_{i=0}^{N-1}$ are trained progressively one at an epoch, simulating the open space by converting input images $x$ to augmented images $\tilde{x}$.

**Training augmentation models.** Let $x_{\text{new}}$ be the augmented images generated by the present augmentation model (with a solid edge), and $x_{\text{old}}$ is generated by a randomly selected old augmentation model (with a dashed edge). The overall loss function for the present augmentation model is as follows:

$$L_{\text{aug}} = L_{\text{recons}}(x, \tilde{x}_{\text{new}}) + L_{\text{div}}(\tilde{x}_{\text{old}}, \tilde{x}_{\text{new}}),$$

(2)

where $L_{\text{recons}}$ is a reconstruction loss between augmented and input images in the pixel space, and $L_{\text{div}}$ is a diversity constraint to enlarge the difference between new and old augmented data in the embedding space.

To expand the simulated open space, an intuitive way is to increase the number of augmentation models. However, naively increasing the number of augmentation models by training them independently could result in similar simulated open space across models as shown in Figure 5(top). Thus, we propose to increase the diversity progressively by constraining the discrepancy of old and new augmented data at each iteration. Simply increasing the embedding distance of $x_{\text{new}}$ and $x_{\text{old}}$ without direction would produce trivial samples also far from the known data. Hence, we apply a regularization term narrowing the embedding between new augmented images $x_{\text{new}}$ and known images $x$, to increase the diversity towards a direction approaching the known space:

$$L_{\text{div}} = \alpha F_{\text{cos}}(\tilde{z}_{\text{new}}, \tilde{z}_{\text{old}}) - \beta \min(F_{\text{cos}}(\tilde{z}_{\text{new}}, z), d),$$

(3)

where $F_{\text{cos}}$ refers to the cosine similarity loss. $d$ is a distance margin to avoid the augmented data totally overlapping with known data. $\alpha$ and $\beta$ are two hyperparameters.

**Training the task model.** Given the known images $x$, new and old augmented images $\tilde{x}_{\text{old}}$ and $\tilde{x}_{\text{new}}$. The overall loss function $L_{\text{task}}$ for the task model is formulated as:

$$L_{\text{task}} = L_{\text{cls}}(x, \tilde{x}_{\text{old}}) + L_{\text{metric}}(x, \tilde{x}_{\text{old}}) + L_{\text{metric}}(x, \tilde{x}_{\text{new}}),$$

(4)

where $L_{\text{cls}}$ is a cross-entropy loss for known class classification, and $L_{\text{metric}}$ is a triplet loss to distinguish the embedding of augmented data from known data, and simultaneously separating different known and augmented classes. $L_{\text{metric}}$
is calculated on both new and old augmented data to avoid forgetting old open-set samples. Specifically, different from existing works taking all simulated open-set samples as a single \( K + 1 \) open class [36, 58], we assign different labels to augmented images from different known classes, as they carry diverse hybrid fingerprints (e.g., for input images of class \( i \) and class \( j \), we assign the augmented images by sending them into an augmentation model to different extended class labels \( K + i \) and \( K + j \)). Then we sample triplets \( T \) from the merged images of input and augmented data. \( \mathcal{L}_{\text{metric}} \) is calculated on triplet sets as follows:

\[
\mathcal{L}_{\text{metric}} = \sum_{T \in T} \left[ \left\| x_{a}^{n} - x_{p}^{n} \right\|^2 - \left\| x_{a}^{n} - x_{n}^{j} \right\|^2 + m \right] + ,
\]

where \( T = (x_{a}^{n}, x_{p}^{n}, x_{n}^{j}) \) is one triplet consisted of an anchor sample \( x_{a}^{n} \), a positive sample \( x_{p}^{n} \) and a negative sample \( x_{n}^{j} \). \( [\cdot]_+ = \max(0, \cdot) \) denotes the hinge loss function. \( m \) is the violate margin that requires the distance of negative pairs to be larger than the distance of positive pairs, by at least a margin \( m \). By optimizing \( \mathcal{L}_{\text{metric}} \), known and augmented data are distinguished, and different known and augmented classes are separated simultaneously.

4. Experiments

4.1. Experimental Details

Datasets. We construct a benchmark for open-set model attribution in Table 1ootnote{Here we only present one way to split this benchmark. The other four splits of the benchmark are in the supplementary material.}. There are four groups of data: Seen Real, Seen Fake, Unseen Real, and Unseen Fake, where Seen and Unseen mean whether images in this group are used for model attribution training or not. Detailed descriptions are as follows:

- **Seen Real**: Seven sets of real images: CelebA [31], Faces-HQ (1024 resolution face images containing CelebA-HQ [22] and FFFHQ [24]), ImageNet (11), Youtube (real Youtube images from Faceforensics++ [42]), Bedroom, Cat and Bus images of LSUN [52].
- **Seen Fake**: Seen models trained on the seen real datasets in the same columns. Each real dataset trains two models of different architectures.
- **Unseen Real**: Images from two real datasets non-overlapping with the seen reals, including CoCo [30] and Summer Yosemite images [59].
- **Unseen Fake**: Unseen models trained by partly the same settings with seen fake ones, but with different seeds, architectures, or training datasets. We control partly the same settings to consider the most challenging three scenarios. Please refer to Table 2 for the relationship between Seen and Unseen Fake. Three categories of Unseen Fake are:
  - Unseen Seed: Unseen models of the same architecture and training data as certain seen fakes, but trained with different random seeds.
  - Unseen Architecture: Unseen models trained on the same datasets as seen fakes, but with different architectures.
  - Unseen Dataset: Unseen models of the same architecture as certain seen fakes, but trained on different datasets.

In total, the benchmark contains 15 known classes including real and 14 seen models, and 53 unknown classes including 10 unseen seed models, 22 unseen architecture models and 21 unseen dataset models. To be realistic, as shown in Table 3, in the testing set, Unseen Real are also considered as closed-set samples as it should be taken as the same known ‘real’ class as Seen Real, and Unseen Fake are taken as open-set classes.

Training, testing, and evaluation. The overall training progress is shown in Algorithm 1. In testing, the test image is fed into the feature extractor \( F \) and classification head \( H \), getting predicted confidence scores for each class after a softmax function. In practice, if the max confidence score is larger than a threshold \( \theta \), the image is recognized as the known category corresponding to the index of the score. Otherwise, it is detected as unknown. For evaluation, we follow works in OSR [7, 46] to evaluate in two aspects: accurately classifying the known models, and distinguishing known/unknown models correctly, which are measured by accuracy and the area under ROC curve (AUC) respectively. We also report Open Set Classification Rate (OSCR) [12] measuring the trade-off between the two aspects.

Implementation details. For the task model, the feature extractor includes a discrete cosine transform (DCT) transformation layer and a simple convolution network same as DNA-Det [50]. The classification head is an MLP network with two linear layers. The augmentation models have two convolution layers, whose weight shapes are \( 3 \times 32 \times 3 \times 3 \) and \( 32 \times 3 \times 3 \times 3 \), respectively. For optimization, we use the Adam optimizer with a step scheduler. The initial learning rate is set to \( 10^{-4} \) for task models and \( 10^{-2} \) for

---

Algorithm 1 Progressive Open-Space Expansion (POSE)

**Input**: Known class images \( x \); Number of epochs \( N 

**Output**: Learned task model \( M \); Learned augmentation models \( \{A_i\}_{i=0}^{N-1} \)

1. for \( i = 0, \ldots, N \) do
2. Generate augmented images \( \tilde{x}_{\text{new}} \) with \( A_i(x) \)
3. if \( i = 0 \) then
4. Train \( A_0 \) by \( \mathcal{L}_{\text{recom}}(x, \tilde{x}_{\text{new}}) \)
5. else
6. Randomly select \( j \) from \( 0 \) to \( i - 1 \)
7. Generate augmented image \( \tilde{x}_{\text{old}} \) with \( A_j(x) \)
8. Train \( A_i \) by \( \mathcal{L}_{\text{recom}}(x, \tilde{x}_{\text{new}}) \) and \( \mathcal{L}_{\text{div}}(\tilde{x}_{\text{old}}, \tilde{x}_{\text{new}}) \)
9. end if
10. Train the task model \( M \) using \( \mathcal{L}_{\text{task}} \)
11. end for
Different Same Same
Same Different

seen Real, seen Fake, unseen Seed, and unseen Architecture are trained on the corresponding seen Real dataset in the same column. unseen Dataset include a total of 21 ProGAN, StyleGAN, and StyleGAN models trained on 7 datasets non-overlapping with seen Real.

Unseen Seed
- ProGAN
  (seed1,2,3,4,5)
- StyleGAN
- InfoMaxGAN

Unseen Architecture
- SNGAN
  [34]
- AttGAN
  [19]
- MMDGAN
  [3]
- Infogan
  [28]
- StyleGAN
  [25]
- BigGAN
  [4]
- ContraGAN
  [21]
- Wav2Lip
  [40]
- FaceShifter
  [29]
- InfoMaxGAN
- SNGAN
- MMDGAN
- StyleGAN
- StyleGAN2

Unseen Dataset
- ProGAN, StyleGAN, StyleGAN3 (Cow, Sheep, Classroom, Bridge, Kitchen, Airplane, Church)

Table 2. The relationship between unseen fake and seen fake.

Unseen Seed
- Different
- Same

Unseen Architecture
- Different
- Same

Unseen Dataset
- Same
- Different

augmentation models. The hyper-parameter $\lambda$ in $L_{\text{spectral}}$ is set to $10^{-4}$. For the hyper-parameters in $L_{\text{dev}}$, we empirically set $\alpha = 10^{-4}$, $\beta = 10^{-2}$, and $\rho = 0.95$. The violate margin $m$ in $L_{\text{metric}}$ is set to 0.3. Input images are all resized to 128px before sending into the network.

Compared methods. We compare against three groups of baselines: 1) GAN attribution, 2) GAN discovery and 3) open-set recognition methods. Most of them are not designed for the open-set model attribution problem. Compared methods and modifications to them are as follows:

- **GAN attribution:** PRNU [33], Yu et al. [53], DCT-CNN [14], DNA-Det [50], and RepMix [5]. We train them on the closed-set, and then obtain classification confidence scores on both the closed-set and open-set following the regular routine of OSR. Comparison results are in Table 4.
- **GAN discovery:** Girish et al. [16] focus on discovering newly arisen unseen GANs. To compare with this work, we first train on the closed-set, and extract features for both the closed-set and open-set following the setup in [16], which are clustered by K-Means with the true cluster number and predicted cluster number by [16]. Results are in Table 5.
- **Open-set recognition:** The classical discriminative OSR method OpenMax [2] and three state-of-the-art generative OSR methods: PROSER [58], ARPL+CS [7], and DIAS [35]. Comparison results are in Table 6.

4.2. Evaluation of Open-Set Model Attribution

Comparison with GAN attribution methods. As shown in Table 4, POSE outperforms existing fake image attribution methods by a large margin in terms of AUC and OSCR. Specifically, the AUC points on unseen architecture, unseen dataset, and unseen all data are improved by 3.24%, 9.45%, and 11.02%, respectively, indicating the superiority of POSE in open-set discrimination. PRNU [33] achieves slightly higher AUC on seed seed, which may be because filters to extract PRNU noise are rather sensitive to differences in traces brought by model weights. We can also observe that unseen seed is the hardest open-set, unseen architecture the second, and unseen dataset is the easiest, which is consistent with our intuition as unseen seed and unseen architecture data only differ from seen data in visually imperceptible traces, and unseen seed data even shares the same model architecture. An exception case is DNA-Det [50], which performs well on unseen architecture, but poorly on unseen dataset. This is due to the fact that DNA-Det is designed to extract architecture traces regardless of the training dataset, yet data in unseen dataset share the same model architectures as seen data.

Comparison with the GAN discovery method. To compare with Girish et al. [16] in discovering unseen models, we train POSE on seen data and obtain features for both the seen and unseen data, which are then clustered for two different values of cluster number $k$. $k = 68$ corresponds to the true number of classes for seen and unseen data. $k = 49$ represents the number of clusters that [16] returns after four iterations. Average purity, Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) are calculated to evaluate the clustering performance. Results are shown in Table 5. As seen, without iterative training on both seen and unseen data like [16], POSE is able to outperform in terms
Table 4. Comparison with existing GAN attribution methods. Results are averaged among five splits of the benchmark.

<table>
<thead>
<tr>
<th>Method</th>
<th>Closed-Set</th>
<th>Unseen Seed</th>
<th>Unseen Architecture</th>
<th>Unseen Dataset</th>
<th>Unseen All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>AUC</td>
<td>OSCR</td>
<td>AUC</td>
<td>OSCR</td>
</tr>
<tr>
<td>PRNU [33]</td>
<td>55.27</td>
<td><strong>69.20</strong></td>
<td>49.16</td>
<td>70.02</td>
<td>49.49</td>
</tr>
<tr>
<td>Yu et al. [53]</td>
<td>85.71</td>
<td>53.14</td>
<td>50.99</td>
<td>69.04</td>
<td>64.17</td>
</tr>
<tr>
<td>DCT-CNN [14]</td>
<td>86.16</td>
<td>55.46</td>
<td>52.68</td>
<td>72.56</td>
<td>67.43</td>
</tr>
<tr>
<td>DNA-Det [50]</td>
<td>93.56</td>
<td>61.46</td>
<td>59.34</td>
<td>80.93</td>
<td>76.45</td>
</tr>
<tr>
<td>RepMix [5]</td>
<td>93.69</td>
<td>54.70</td>
<td>53.26</td>
<td>72.86</td>
<td>70.49</td>
</tr>
</tbody>
</table>

Table 5. Comparison with Girish et al. [16] for GAN discovery.

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg. Purity</th>
<th>NMI</th>
<th>ARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Girish et al. [16] (k=49)</td>
<td>32.89</td>
<td>61.89</td>
<td>21.05</td>
</tr>
<tr>
<td>POSE (k=49)</td>
<td>39.16</td>
<td><strong>61.91</strong></td>
<td><strong>27.48</strong></td>
</tr>
<tr>
<td>POSE (k=68)</td>
<td>41.04</td>
<td>60.59</td>
<td>26.39</td>
</tr>
</tbody>
</table>

Figure 4. Feature space visualization (top) and confidence histograms (bottom) for Base method and our proposed POSE.

of purity and ARI, and get close NMI, which validates the efficiency of POSE in reasoning about unseen GAN classes. **Comparison with OSR methods.** We take optimizing the plain task model by a cross-entropy loss as a Base method, and compare with OpenMax [2], PROSER [58], ARPL+CS [7], and DIAS [35] by employing their proposed techniques and simulated open-set samples to augment the Base classifier. As shown in Table 6, applying augmentation models (Base+AM) is superior to other OSR methods, and involving \( \mathcal{L}_{div} \) (Base+AM+\( \mathcal{L}_{div} \)) gains further improvement. These results prove that the simulated open space by POSE is more suitable for OSMA than off-the-shelf OSR methods.

4.3. Ablation Study

**Validation of training with augmentation models.** As shown in Table 6, compared with the Base method, applying augmentation models (Base+AM) not only improves closed-set classification accuracy (from 90.68% to 93.41%) but also helps open-set discrimination (from 73.78% to 80.31% in AUC on unseen all data). This indicates sim-

Figure 5. (a-d) Feature space visualization of progressive open space expansion process without (top) and with \( \mathcal{L}_{div} \) (bottom). **Blue** triangles are data of one known class, and circles in other colors are augmented data for this class by different augmentation models. (e) Pairwise similarities of open space expanded by 20 augmentation models. Blue color indicates higher similarity. **Calculating open space by augmentation models helps more compact feature extraction not only among close-set classes but also in between close/open-set. Meanwhile, observing the feature space visualization cases in Figure 4(top), the augmented data simulates a rich open space enclosing the known data points, resulting in a clear better close/open discrimination. Better open-set discrimination is also reflected in the prediction confidence scores in Figure 4(below), again justifying the effect of augmentation models. **Validation of \( \mathcal{L}_{div} \).** Figure 5(a-d) give an intuitive understanding of how \( \mathcal{L}_{div} \) affects the open space expansion progress. As seen, with \( \mathcal{L}_{div} \) added, the diversity of open space simulated by different augmentation models is largely increased, which is also indicated in the decreased pairwise similarity in Figure 5(e). Further, with the help of \( \mathcal{L}_{div} \), the AUC points on unseen seed, architecture, dataset and unseen all increase by 1.98%, 1.96%, 3.20%, and 2.45%, respectively. These results prove that \( \mathcal{L}_{div} \) increases the diversity of simulated open space and reduces the open space risk better. **Validation of the architecture of augmentation models.** We investigate the impact that the augmentation model’s architecture has on open-set discrimination performance from three aspects: 1) Upsampling layer. As discussed in Section 3.1, including upsampling in the augmenta-
Table 6. Comparison with existing OSR methods and ablation experiments. Results are averaged among five splits of the benchmark.

| Method                | Closed-Set Accuracy | Unseen Seed AUC | Unseen Seed OSCR | Unseen Architecture AUC | Unseen Architecture OSCR | Unseen Dataset AUC | Unseen Dataset OSCR | Unseen All AUC | Unseen All OSCR |
|-----------------------|---------------------|-----------------|-------------------|--------------------------|---------------------------|--------------------|---------------------|----------------|----------------|}
| Base                  | 90.68               | 62.02           | 60.58             | 76.03                    | 72.92                     | 77.01              | 73.88               | 73.78          | 70.97          |
| Base+OpenMax [2]      | 91.11               | 63.27           | 61.60             | 76.40                    | 73.29                     | 75.33              | 72.32               | 73.50          | 70.70          |
| Base+PROSER [58]      | 92.12               | 63.32           | 62.19             | 79.55                    | 76.57                     | 81.43              | 78.64               | 77.22          | 74.66          |
| Base+ARPL+CS [7]      | 91.77               | 54.94           | 54.17             | 79.09                    | 75.97                     | 80.48              | 77.52               | 75.08          | 72.47          |
| Base+DIAS [35]        | 92.77               | 62.15           | 61.02             | 79.34                    | 76.49                     | 84.14              | 81.13               | 78.00          | 75.41          |
| Base+AM               | 93.41               | 66.17           | 65.04             | 82.21                    | 79.42                     | 85.04              | 82.20               | 80.31          | 77.80          |
| Base+AM+$L_{div}$ (POSE) | 94.81              | 68.15           | 67.25             | 84.17                    | 81.62                     | 88.24              | 85.64               | 82.76          | 80.50          |

Figure 6. Ablation study on the architecture of augmentation models. We consider whether to include a upsample layer, layer number and kernel size of convolution layers.

- Upsampling would heavily influence high frequencies and tend to produce trivial open-set samples. We empirically include upsampling layer in the augmentation model (Down+Conv+Up+Conv), resulting in much lower performance than two pure convolution layers (Conv+Conv) shown in Figure 6(a), which verifies our former assumption.
- Layer number. In Figure 6(b), the AUC reaches the summit when the convolution layer number equals 2 and decreases with more layers. This may be due to that a single layer is hard to simulate large open space, while more layers lead to larger reconstruction residuals and produce easier open-set samples far from known space.
- Kernel size. Figure 6(c) shows the AUC w.r.t. kernel size of convolution layers, indicating a $3 \times 3$ kernel size is the best. This could be because most generative models use a $3 \times 3$ kernel size.

Progressive vs. joint training. It is computation-expensive to jointly train a large number of augmentation models one-off and constrain the diversity among them. We compare the OSR performance of progressive and joint training with 9 augmentation models (the maximum number of augmentation models could be trained jointly on a 32G V100 GPU) on one split. At the 9th epoch with the same number of augmentation models involved, AUCs on unseen all data under progressive and joint training are 80.17% and 79.42%, while the latter takes five times longer to train than the former, which indicates progressive training provides an effective way to expand the open space.

Influence of the number of augmentation models. We study the effect of the number of augmentation models involved ($N$) in improving open-set recognition performance in Figure 7. We can observe that the performance converges at early epochs without $L_{div}$, while the AUC of POSE increases continually until about 19 epochs, indicating the necessity to increase the diversity of simulated open space.

5. Conclusions and Future Work

This paper tackles an important problem in applying model attribution to the real world, the open-set model attribution problem, to attribute images to known source models and identify those from unknown ones. We propose a novel progressive open space expansion solution, whose core idea is to simulate the potential open space of unknown models by involving augmentation models progressively. We consider three challenging types of unknown models to construct a benchmark. The superior experimental results indicate that it is promising to model the boundary of known models by artificially constructed models. Studying how fingerprints are influenced by network components and simulating them through artificial tiny networks could be productive in the future. Such research could lead to a more generic fingerprint space representation.

6. Acknowledgements

The research work is supported by the China Postdoctoral Science Foundation (2022TQ0344), the Project of Chinese Academy of Sciences (E141020), the National Natural Science Foundation of China (62203425), and the Zhejiang Provincial Key Research and Development Program of China (No.2021C01164).
References


[32] Mario Lu{c}i{c}, Michael Tschannen, Marvin Ritter, Xiaohua Zhai, Olivier Bachem, and Sylvain Gelly. High-fidelity image generation with fewer labels. In ICML, 2019. 6

[33] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In MIPR, 2019. 1, 2, 6, 7


[50] Tianyun Yang, Ziyao Huang, Juan Cao, Lei Li, and Xirong Li. Deepfake network architecture attribution. In AAAI, 2022. 1, 2, 5, 6, 7
[53] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In ICCV, 2019. 1, 2, 6, 7
[58] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In CVPR, 2021. 2, 5, 6, 7, 8