Content Robust Image Generator Attribution

Anonymous authors
Paper under double-blind review

Abstract

Image generator attribution aims to identify what generator produced an image, if any. Prior work often focused on identifying new generators without requiring large amounts of labeled samples by searching for shifts in image distributions. However, these shifts may appear in other contests, such as a change in image content. Thus, an image may be attributed to the wrong generator because its image content changed from what was typically seen during training. To address this issue, we explore Content Robust imagE generator attrIbuTion (CREdIT), where a model is evaluated on its ability to attribute an image accurately even if the generators and/or image content is different than what was seen during training. After a thorough analysis, we created a carefully crafted yet simple baseline we refer to as FakesSense, which outperforms the state-of-the-art by 3-7%. This illustrates a significant shortcoming in prior work, demonstrating a need for more complex image generator attribution benchmarks like CREdIT.

1 Introduction

Detecting generated images (*e.g.*, Yu et al. (2018); Guo et al. (2023); Wang et al. (2023); Qian et al. (2020); Ni et al. (2022); Anas Raza & Mahmood Malik (2023); Hu et al. (2021); Wang et al. (2020a); Ojha et al. (2023); Tan et al. (2023); Liu et al. (2024c)) is an important component of defending against misinformation, but many applications require identifying an image's source (*e.g.*, copyright infringement). Watermarking and fingerprinting images can provide the means to recognize a generator (*e.g.*, Liu et al. (2022); Luo et al. (2010); Pereira & Pun (2000); Tancik et al. (2020); Ding et al. (2021b); Yu et al. (2019; 2021); Meng et al. (2025); Zhao et al. (2024)), but require altering the generator or the image samples that may not always be possible, such as for user uploaded content in social media. In contrast, image generator attribution methods can detect and attribute synthetic images without requiring image or model-based modifications. Most prior work has focused on addressing generator shifts, where attribution methods are tasked with identify samples from generators where they are provided few, if any, labeled examples (*e.g.*, Ni et al. (2022); Anas Raza & Mahmood Malik (2023); Hu et al. (2021); Sun et al. (2021); Xu et al. (2022); Zhong et al. (2021); Cao et al. (2022); Sun et al. (2023); Yang et al. (2023); Yu et al. (2018); Guo et al. (2023); Wang et al. (2023); Qian et al. (2020); Wang et al. (2022b); Li et al. (2020b); Dang et al. (2020b); Rossler et al. (2019); Liu et al. (2024a)). As illustrated in Fig. 1(a), these methods often evaluate in a setting where image content is consistent (Neves et al., 2020; Wang et al., 2019; Zhou et al., 2017; Wang et al., 2019; Dang et al., 2020a; Li et al., 2020a; Wang et al., 2022b), enabling them to assume that distribution shifts equate to new generators. However, this risks overfitting to an artificial restriction on image content since it often varies in practice.

This paper addresses this shortcoming by exploring a new task aligned to real-world needs, Content Robust imagE generator attrIbuTion (CREdIT). As shown in Fig. 1(b), the key difference in our task from those explored in most prior work is that image distribution shifts may be due to a new generator, its content, or both. However, as summarized in Tab. 1, most existing benchmarks either are limited in the types of generators (*e.g.*, diffusion, GAN, etc.), image content, or both. Thus, we propose DiverGen, a new benchmark that provides diverse generators that control both object category and attribute shifts. This enables us to measure the effect of multiple types of shifts on model performance, a glaring shortcoming of prior work.

Specifically, our experiments show that these multiple shift types causes methods from prior work to mistake content shifts with generator shifts. To illustrate, we use density-based clustering DBSCAN to group images

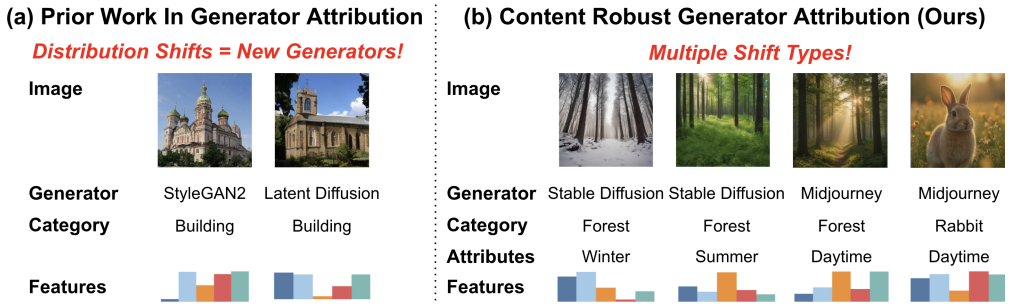


Figure 1: Image generator attribution methods aim to identify the source of an image. **(a)** Prior work often relies on the assumption that distribution shifts always indicate that a new generator has appeared (*e.g.*, Ni et al. (2022); Cao et al. (2022); Sun et al. (2023); Yang et al. (2023); Guo et al. (2023); Li et al. (2020b); Liu et al. (2024a)), overfitting to this simple setting. However, **(b)** illustrates our CREdIT task where distribution shifts can be due to image content or generator changes, better reflecting real-world attribution applications.

from the same generators in test set of our proposed DiverGen-Attr subset using features computed from CPL (Sun et al., 2023). Note that we cannot use a classifier as some generators are unseen during training. We use majority voting to label each cluster. However, this can result in multiple clusters being given the same label. To address this, only the cluster with the most samples gets the generator label, and each smaller cluster would have all samples marked incorrect as they indicate the model believes it is a different generator. We found that despite cheating by tuning DBSCAN's parameters to maximize accuracy on the test set, it only labeled 29.7% samples correct. Further, it only identified two generators in the test set when there are actually six. In comparison, in the same experiment our simple baseline reports an attribution accuracy of 55.2%, almost twice the performance, and with six detected generators. This shows that methods from prior work have difficulty understanding the differences between shifts in content and changes in generators.

Issues with multi-axis distribution shifts like CREdIT appears in other settings such as bias mitigation (Qraitem et al., 2024; 2023; Howard et al., 2024; Subramanian et al., 2021; Wang et al., 2020b) or when label noise appears alongside domain shifts (Humblot-Renaux et al., 2024; Qiao & Low, 2024; Seo et al., 2020). However, prior work in bias mitigation methods often requires subpopulation labels that are often unavailable, whereas domain shift problems just evaluate how robust methods are to these intersectional distribution shift problems. In contrast, we go a step farther and also investigate methods to address this problem without having labeled samples of all subpopulations that may appear.

Our contributions are summarized as follows:

- We explore Content Robust imagE generator attrIbuTion (CREdIT), a new task where the goal is to identify an image's source under both content *and* generator shifts.
- We introduce DiverGen new diverse dataset for image generator attribution. Tab. 1 shows our dataset has more shift types than those in prior work. We will make our dataset publicly available for academic use.
- We provide a thorough analysis on both exiting and our newly introduced datasets. We show that a simple baseline we refer to as FakesSense outperforms the state-of-the-art (Rizve et al., 2022; Zhong et al., 2021; Han et al., 2020; Cao et al., 2022; Yang et al., 2023; Sun et al., 2023; Liu et al., 2024b) by 3-7%, highlighting the need for further exploration of our task.

2 Related Work

Closed-set image generator attribution aims to identify which generator seen during training produced an image (*e.g.*, Yu et al. (2018); Guo et al. (2023); Wang et al. (2023); Qian et al. (2020); Wang et al. (2022b); Li et al. (2020b); Dang et al. (2020b); Rossler et al. (2019); Ni et al. (2022); Anas Raza & Mahmood Malik (2023); Hu et al. (2021)). Open-set methods expands on this by also detecting samples of unseen generators, i.e., those that do not appear during training (*e.g.*, Jain et al. (2021); Lee et al. (2021); Aneja & Nießner (2020); Rizve et al. (2022); Cao et al. (2022); Yang et al. (2023); Sun et al. (2023; 2021); Xu et al. (2022);

Dataset	Generator Diversity	Multi-Category	Attributes
iFakeFaceDB (Neves et al., 2020)	✗	✗	✗
TwoStream (Zhou et al., 2017)	✗	✗	✗
FakeSpotter (Wang et al., 2019)	✗	✗	✗
DFFD (Dang et al., 2020a)	✗	✗	✗
ForgeryNet (He et al., 2021)	✗	✗	✗
GenImage (Zhu et al., 2023)	✓	✓	✗
TWIGMA (Chen & Zou, 2023)	✗	✓	✗
SEMI-TRUTHS (Pal et al., 2024)	✗	✓	✗
RED140 (Guo et al., 2024)	✓	✓	✗
WildFake (Hong & Zhang, 2024)	✓	✓	✗
WILD (Bongini et al., 2025)	✓	✓	✗
DiverGen (ours)	✓	✓	✓

Table 1: Comparison of the properties of DiverGen to prior work on generated image detection and attribution. Our dataset is the only one to control for multiple content shift types.

Zhong et al. (2021)). To avoid correlations between generators and content, these methods typically restrict the generated images so they contain only faces (*e.g.*, Neves et al. (2020); Wang et al. (2019); Zhou et al. (2017); Wang et al. (2019); Dang et al. (2020a); Li et al. (2020a); Wang et al. (2022b)). As shown in Fig. 1(a), this enables these methods to assume that any shift in feature distributions is likely an unseen generator. However, many practical attribution applications cannot control what categories are seen at test time. While one could argue that a good attribution model would learn a content-agnostic features, most prior work purposefully does not evaluate on content shifts as mentioned earlier. CREdIT addresses this by evaluating the impact that shifts in image content as well as changes in generators has on performance, thereby better aligning to their real-world use. As we show, prior work underperforms a simple baseline that is explicitly designed to encourage learning content-robust representations.

While some recent attribution methods evaluate on images containing diverse object categories (*e.g.*, Liu et al. (2024b)) or generator types (*e.g.*, Zheng et al. (2025)), they assume they are provided a few labeled examples of each type. This may be sufficient when there are just a few generators with limited variations in image content. However, at even moderate scales it is impractical to assume we can be given examples of every generator in every variation of image content. Additionally, new generators are constantly being developed so labeled samples might simply be unavailable. As our results will show, even if we do provide these examples, these methods still underperform our simple baseline.

Some work on generated image detection considers the effect varying numbers categories has on performance (*e.g.*, Wang et al. (2020a); Ojha et al. (2023); Tan et al. (2024); Zhu et al. (2023)). However, detection's challenges are different than those of CREdIT. For example, in Fig. 1(b) the forest samples should have the same label in generated image detection (i.e., they are all generated images). In contrast, in CREdIT they should have different labels as they were produced from different generators, and separating them can be challenging due to similar image content. Thus, while generated image detection and attribution have some similarity, they posses different challenges that require specialized methods to address them.

3 Content Robust imagE generator attrIbuTion (CREdIT)

Given an image x_i our task is to identify the generator that produced it. We treat real and synthetic images the same, i.e., "real" is treated as one generator category. Following (Sun et al., 2023), we explore a setting where we are given a labeled and unlabeled set of images during training represented as $\mathcal{S}_l = \{(x_i, y_i)\}_{i=1}^{m}$ and $\mathcal{S}_{ul} = \{x_i\}$, respectively. We denote y_l as the generators in the labeled set and y_{ul} as the generators that only appear in the unlabeled set, i.e., $y_l \cap y_{ul} = \emptyset$. The labeled set of images are all labeled with their generator source, whereas the unlabeled set's images are produced from a mix of y_l and y_{ul} generators. This emulates a case where we curated a labeled training set, but also have in-the-wild unlabeled data available.

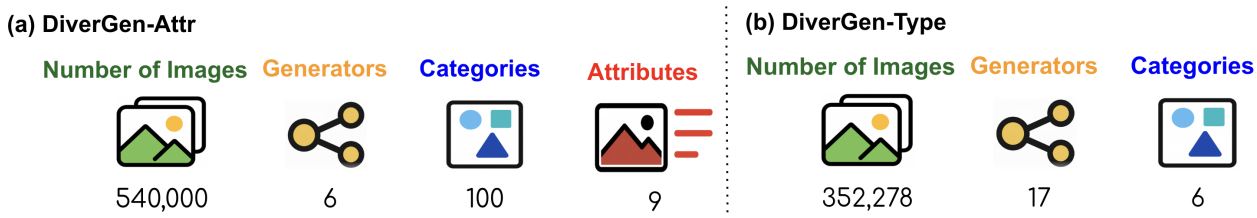


Figure 2: Summary of DiverGen in both the **(a)** DiverGen-Attr and **(b)** DiverGen-Type subsets. See Sec. 3.1 for construction details and the Appendix Sec. B for more detailed statistics.

Additionally, at test time, we further evaluate on generators $\mathcal{Y}_{us}$ that produce images $\mathcal{S}_{us}$ that are completely unseen during training (i.e., they appear in neither S_l nor S_{ul}).

Model attribution is critical for many applications, such as those investigating the source of a false news story or if any intellectual property rights were violated. To identify generators that are not in the labeled training set, most methods rely on identifying a distribution shift in samples from new generators (*e.g.*, Ni et al. (2022); Cao et al. (2022); Sun et al. (2023); Yang et al. (2023); Guo et al. (2023); Li et al. (2020b); Liu et al. (2024a)). As discussed in the introduction, this results in models that are unable to seperate content shifts and changes in generator. Thus, in this paper we explore our CREdIT task, where shifts can stem from both changes generators as well as content represented by an object category.

Sec. 3.1 describes our DiverGen dataset, which contains diverse categories, controls for attributes, and various generator types to help better understand our task. Sec. 3.2 describes our simple FakesSense baseline method that identifies issues with prior work in generator attribution and then addresses them by adapting components from other tasks such as domain adaptation or learning with noisy labels.

3.1 DiverGen Dataset

To enable our exploration of CREdIT we create a new benchmark to provide content-controlled generated images across generators. As we summarized in Tab. 1, prior work does not provide controlled samples across multiple types of content shifts. This is a significant shortcoming in an age where text-to-image models are plentiful. In addition, even for benchmarks like GenImage (Zhu et al., 2023) that provides a diverse set of generators and content categories, they typically have few images per-task. Specifically, for each object category GenImage provides only 162 training and 6 testing samples. Thus, we propose a content-Diverse Generator (DiverGen) benchmark for attribution containing two distinct subsets that address the shortcomings of datasets in prior work. Sec. 3.1.1 describes our DiverGen-Attr subset which focuses on providing a wide variety of content shift types using recent diffusion and transformer-based generators. Sec. 3.1.2 describes our DiverGen-Type subset focuses on scaling the number and variety of generators and provides a large number of images per content category. This enables DiverGen to provide a strong attribution benchmark across a diverse set of experimental settings. Fig. 2 summarizes our dataset's statistics.

3.1.1 DiverGen-Attr subset

DiverGen-Attr subset provides a large variety of content shift types over a set of strong, diverse generators. Specifically, we select six models from recent work that includes both transformer and diffusion generators: CogView (Ding et al., 2021a), HyperSD (Ren et al., 2024), Flux (Labs, 2024), Midjourney (Midjourney, 2024), Stable Diffusion XL (Podell et al., 2023), and Stable Diffusion 3M (Esser et al., 2024). Each generator was tasked to produce 100 images containing for every object-attribute pair. We choose to adopt CIFAR-100 (Krizhevsky, 2012) for our list of object categories and select a set of generalized attributes focused on environmental and seasonal variations that can control the general image content. Specifically, we allow unconstrained generations as well as Indoor, Outdoor, Spring, Summer, Autumn, Winter, Night, and Sunny. The result is 100 images per class $\times$ 100 classes $\times$ 6 generators $\times$ 9 attributes = 540,000 images in DiverGen-Attr, providing a large scale dataset with significant diversity. In the next section we will increase the number of samples per content type to better understand the relationship between data scale and performance.

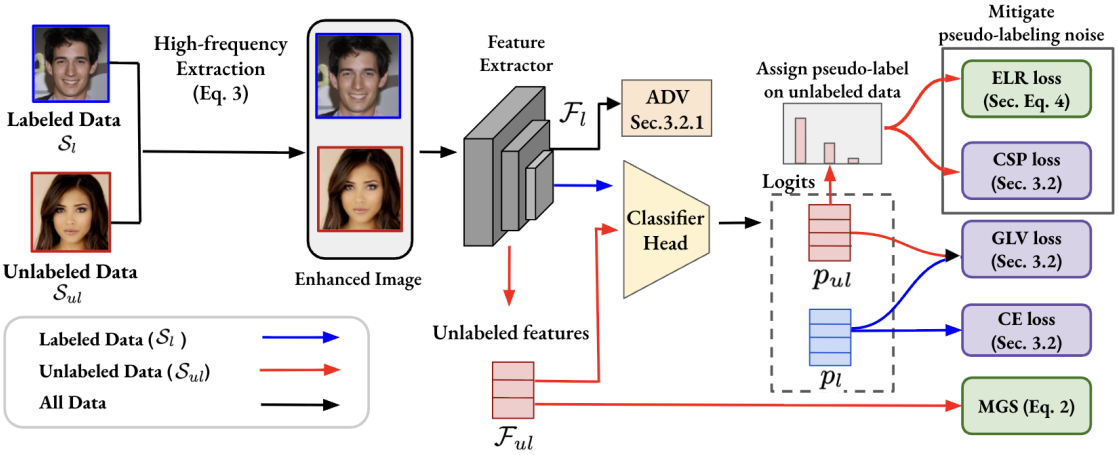


Figure 3: FakesSense **framework**. Both the labeled ($\mathcal{S}_l$) and unlabeled images ($\mathcal{S}_{ul}$) are first enhanced by HFE with eq. (7) (only used during training). The enhanced images are fed into the feature extractor to get labeled features ($\mathcal{F}_l$) and unlabeled features ($\mathcal{F}_{ul}$). $\mathcal{F}_l$ is then sent to the ADV classifier to encourage elimination of content features (Sec. 3.2.1) and $\mathcal{F}_{ul}$ is used to calculate the MGS loss (eq. (6)). We then feed the feature representations into the classification head to obtain labeled logits (p_l) and unlabeled logits p_{ul}. p_l is used to calculate both the GLV loss and CE loss (Sec. 3.2) and p_{ul} is used to calculate the GLV and CSP losses (Sec. 3.2) in addition to ELR (see eq. (8)).

3.1.2 DiverGen-Type **subset**

The DiverGen-Type subset focuses on scaling the number of generators and images per content category. We generated at least 10K images each from a set of 17 image generators including GAN (StyleGAN2 (Karras et al., 2020), StyleGAN3 (Karras et al., 2021)), diffusion/transformer based models (TamingTransformer (Esser et al., 2021), LatentDiffusion (Rombach et al., 2021), GuidedDiffusion (Kim et al., 2022), LSGM (Vahdat et al., 2021)). The focus of our new images is to increase object diversity, representing six different category domains (faces, furniture, buildings, art, vehicles, animals), resulting in 171,451 generated images. The remaining 180,827 images (352,278 total) and generators are sourced from existing datasets that generated human faces including celeb-real, youtube-real images from CelebDFv2 (Li et al., 2020a), StyleGAN (Karras et al., 2019), FaceAPP (FaceApp, 2024), starGAN (Choi et al., 2018), PGGAN (Karras et al., 2017) from DFFD (Dang et al., 2020a), and CycleGAN (Zhu et al., 2020), StyleGAN (Karras et al., 2019), PGGAN (Karras et al., 2017), StyleGAN2 (Karras et al., 2020) from ForgeryNir (Wang et al., 2022b). See the Appendix Sec. B for additional details and dataset statistics.

3.2 **Baseline** FakesSense **Framework**

In this section we describe our FakesSense baseline, which we summarize in Fig. 3. We generally follow the approach of CPL (Sun et al., 2023), which we describe briefly below, but make several modifications to adapt it to our task. Specifically, as noted in the introduction, the new challenge in CREdIT is that attribution models need to be able to tell the difference between shifts in image content and changes in generator. Thus, our improvements focus on reducing training noise or explicitly disentangles content shifts from generators. While most of these model components have appeared in some form in prior work, they are either methods that have never been used in generator attribution or we modified to make them more generalizable.

More formally, given our training images $\mathcal{S}_l, \mathcal{S}_{ul}$ representing the labeled and unlabeled images respectively, we use a feature extractor to obtain corresponding feature sets $\mathcal{F}_l$ and $\mathcal{F}_{ul}$. Both $\mathcal{F}_l$ and $\mathcal{F}_{ul}$ are passed into a classifier to obtain a set of logits p_l and p_{ul}, respectively. Then we compute cross-entropy loss over p_l as our primary task loss:

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

where y^l_{ic} is the ground-truth label and $p^l_{ic} = \sigma(\phi(x^l_i))_c$ is the predicted probability for class $c \in \mathcal{C}_K$.

Additionally, labeled and unlabeled samples p_l, p_{ul} use Global-Local Voting (GLV) module that encourages global-local prediction consistency. In other words, GLV is defined as:

$$\mathcal{L}_{\text{GLV}} = -\frac{1}{n} \sum_{x_i \in \mathcal{D}_l} \log\langle \sigma(f_G(x^l_i)),\, \sigma(f_G(\widetilde{x}^l_i)) \rangle - \frac{1}{m} \sum_{x_i \in \mathcal{D}_u} \mathbb{I}(\widetilde{x}^u_i = \widehat{x}^u_i) \log\langle \sigma(f_G(x^u_i)),\, \sigma(f_G(\widetilde{x}^u_i)) \rangle. \tag{2}$$

where $\tilde{x}_i$ and $\hat{x}_i$ denote the nearest neighbor of x_i under global and local similarity respectively, and the indicator $\mathbb{I}(\tilde{x}^u_i = \hat{x}^u_i)$ retains only pairs where both agree.

CPL also uses a confidence-based soft pseudo-labeling (CSP) loss, which requires periodically computing a set of pseudo labels for the unlabeled images $\mathcal{S}_{ul}$ during training. Specifically, CSP applies a confidence-weighted cross-entropy over soft pseudo-labels on the unlabeled set:

$$\mathcal{L}_{\text{CSP}} = -\frac{1}{m} \sum_{x_i \in \mathcal{D}_u} \sum_{c \in \mathcal{C}_K \cup \mathcal{C}_N} \lambda^u_i \cdot \tilde{y}^u_{ic} \log p^u_{ic}, \tag{3}$$

where $\tilde{y}^u_i = \text{GumbelSoftmax}(p^u_i)$ is the soft pseudo-label from (Jang et al., 2017), $c = \arg\max_c \tilde{y}^u_{ic}$ is the most likely pseudo-class, and $\lambda^u_i = p^u_{ic^*}$ is a confidence weight that down-weights uncertain assignments.

Finally, CPL also uses a regularization term $\mathcal{R}$ to prevent assigning all samples to one single class:

$$\mathcal{R} = \text{KL}\left(\frac{1}{n+m} \sum_{x_i \in \mathcal{D}_l \cup \mathcal{D}_u} \sigma(f_G(x_i)) \;\middle\|\; \mathcal{P}(y) \right), \tag{4}$$

where $\mathcal{P}(y)$ is a prior over class labels y. Thus, the full objective of CPL is:

$$\mathcal{L}_{\text{CPL}} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{GLV}} + \lambda_2 \mathcal{L}_{\text{CSP}} + \lambda_3 \mathcal{R}, \tag{5}$$

where λ_{1-3} are tunable hyper-parameters. Please see Sun et al. (2023) for additional details on these components. In the next section, we discuss our five improvements.

3.2.1 Reducing Noise and Improving Generalization

Reducing the effect of distribution shifts. While modules we will discuss in Sec. 3.2.1 aims to improve performance over unlabeled images $\mathcal{S}_{ul}$, we find their impact on unseen generators $\mathcal{S}_{us}$ can be limited. Thus, we also explore mechanisms aimed at generalizing to new image distributions, i.e., by integrating multi-source domain generalization techniques. In particular, we leverage SWA-f, a variation of Stochastic Weight Averaging (SWA) (Izmailov et al., 2019) that only averages weights on the feature extractor. Unlike the traditional SWA approach, SWA-f does not average the weights on the classification head to enhance the generalization abilities of the feature extractor while maintaining the learning dynamics of the classification head. As SWA-f requires no specific loss function or domain labels, it is highly generalizable. As we will show, this is more effective than more recent methods like URM (Krishnamachari et al., 2024), which focuses on learning features that is more evenly represented across the training distributions. However, this relies on having labeled distribution shifts, information not available in the unlabeled set.

Disentangling content shifts from generator changes. We directly improve our model's ability to separate content shifts from generator changes by using an Adversarial (ADV) method to teach the model to ignore content shifts. This is a stark contrast with prior work in generator attribution that created an artificial evaluation setting without content shifts (Neves et al., 2020; Wang et al., 2019; Zhou et al., 2017; Wang et al., 2019; Dang et al., 2020a; Li et al., 2020a; Wang et al., 2022b), which we show does not generalize. Specifically, for each content type (i.e., objects or attributes), we train a linear classifier using the same features used by the generator attribution classifier to predict its content category. For example, on DiverGen-Attr we would train two classifiers, one that predicts object category and one that predicts attribute category, whereas other datasets we rely only on object category as attributes are not controlled for. Then we employ gradient reversal (Ganin & Lempitsky, 2015) to encourage the model to eliminate content features, helping to reduce confusion from shifts in content.

Maximizing generator separation (MGS). We found prior work like CPL (Sun et al., 2023) on CREdIT can learn a fragmented feature space where samples from the same generator can appear across the feature space. This suggests that generalization can be challenging as the features between generators is not well separated. Thus, we encourage our model to learn a more cohesive representation for each generator, enabling us to be more robust to shifts in content. This is accomplished by maximizing the distance between the center of mass for centriods computed for each generator in $\mathcal{S}_{ul}$ according to its pseudo-label. Formally, given d generators and their respective centroids, we maximize generator separation (MGS) via:

$$\mathcal{L}_{\text{MGS}} = -\sum_{i=1}^{d}\sum_{j=i+1}^{d}\frac{C_i \cdot C_j}{\|C_i\|\|C_j\|}, \tag{6}$$

where d is the number of clusters and C_i, C_j are the centroids of generators i & j.

Reducing image noise via High-frequency Extraction. Prior work extracts the high-frequency component of an image where generative artifacts are more pronounced to make them easier to learn and then use during inference (Corvi et al., 2022; Zhu et al., 2023). However, these methods tend to compute statistics over a set of training images, which we show overfits to observed shifts, making them ill-suited to CREdIT where we must generalize to two types of distribution shifts. Thus, we use a version of High-Frequency Extraction (HFE), a straightforward image augmentation approach to amplify the most salient high-frequency image component and learn features that are more discriminative of the manipulation type.

More formally, given input image $\boldsymbol{X}$, we extract the high-frequency component $HF(\boldsymbol{X}) = \text{abs}(\boldsymbol{X} - \text{GF}(\boldsymbol{X}))$, where GF is the Gaussian blur filter. Then, the enhanced version of the input image $\boldsymbol{X}$ is computed as:

$$\hat{\boldsymbol{X}} = \boldsymbol{X} + \nu HF(\boldsymbol{X}), \tag{7}$$

where $\nu > 0$ is a hyperparameter controlling the high-frequency component amplification level. ν is set via grid search using the validation set. See Sec. A for an example of HFE enhancement.

Minimizing the impact of pseudo-label noise. We also take from the learning with noisy labels (LNL) literature to mitigate the effect of noise due to using pseudo-labels on our unlabeled set. Specifically, ELR (Liu et al., 2020b) mitigates the effect of incorrect labels by regularizing the model to early stages of training, before the model memorized false labels. Formally, ELR is defined as maximizing the inner product between current classifier outputs p_i and those made early during training p_i^* for sample x_i, i.e.,

$$\mathcal{L}_{\text{ELR}} = \log(1 - \langle p_i, p_i^* \rangle). \tag{8}$$

Note that we will also experiment with more recent LNL methods, but find that they do not perform well in CREdIT. This is likely due, in part, to their demonstrated sensitivity to shifts in image domain (Humblot-Renaux et al., 2024; Wang et al., 2026), like those in CREdIT. Thus, FakesSense's objective is:

$$\mathcal{L} = \mathcal{L}_{\text{CPL}} + \lambda_4 \mathcal{L}_{\text{MGS}} + \lambda_5 \mathcal{L}_{\text{ELR}}, \tag{9}$$

where λ_{4-5} are tunable hyper-parameters.

4 Experiments

Implementation details. We implement FakesSense using PyTorch and all the models are trained on a single NVIDIA A100 GPU. When extracting high-frequency components using HFE in eq. (7), we used $\nu = 0.5$. Each compared method uses a ResNet-50 backbone trained for 15 epochs with a batch size of 20. For FakesSense, our model is trained with Adam optimizer with a learning rate of 2×10^{-4} and weight decay of 10^{-5}. For all other models, we use their default parameters.

Evaluation procedure and metrics. During evaluation, all images from $\mathcal{S}_l, \mathcal{S}_{ul}, \mathcal{S}_{us}$ generators representing the images labeled, unlabeled, and unseen generators, respectively, are fed into the feature extractor of each model. Then, we use KMeans to separate samples into generator clustering and compute the final

Dataset	Labeled (ID)	Unlabeled (ID)	Unseen (OOD)
GenImage	ADM, Stable Diffusion V1.4	GLIDE, Wukong	VQDM, Stable Diffusion V1.5, Midjourney, BigGAN
DiverGen-Attr	Stable Diffusion XL, HyperSD	CogView, Midjourney	Flux, Stable Diffusion 3M
DiverGen-Type	Celeb-real, DAGAN, FOMM, LatentDiff, Taming, PGGAN	LIA, Maxine, LSGM, StyleGAN3, StyleGAN2, FaceAPP	Youtube-real, StyleHeat, TPS, GuidedDiff, StyleGAN, StarGAN, CycleGAN

Table 2: Generators used in each split for different datasets. We report the labeled in-distribution (ID), unlabeled ID, and unseen out-of-distribution (OOD) generators for each benchmark.

	Content	In-Distribution			Out-of-Distribution		
	Generator	**Labeled (ID)**	**Unlabeled (ID)**	**Unseen (OOD)**	**Seen (ID)**	**Unseen (OOD)**	**ALL Acc**
(a)	NCL (Zhong et al., 2021)	87.6±3.8	55.7±4.5	16.6±1.7	67.1±3.7	17.7±4.8	52.5±1.6
	openLDN (Rizve et al., 2022)	78.7±2.3	82.9±0.5	13.0±6.6	75.7±1.9	13.6±6.5	57.1±1.2
	RankStat (Han et al., 2020)	86.6±4.5	76.8±4.1	18.4±1.3	70.2±0.7	17.8±2.2	57.8±0.6
	ORCA (Cao et al., 2022)	99.0±0.6	52.8±7.0	26.1±8.8	71.2±1.4	27.0±7.6	58.4±3.5
	VL2V-ADiP (Addepalli et al., 2024)	99.0±0.1	91.2±2.2	37.6±2.8	30.6±2.2	53.6±1.0	64.5±0.4
	POSE (Yang et al., 2023)	92.5±2.3	**97.8±0.1**	34.8±4.0	84.3±0.8	35.3±3.7	72.6±0.6
	CDAL[1] (Zheng et al., 2025)	87.2±2.3	96.1±1.0	46.4±5.0	77.6±1.8	50.1±4.7	74.1±0.6
(b)	CPL (Sun et al., 2023)	95.2±2.2	89.0±0.4	29.1±2.0	**86.4±1.6**	27.4±3.9	69.4±0.5
	+URM (Krishnamachari et al., 2024)	**99.7±0.1**	89.0±1.5	35.3±2.9	31.4±1.3	54.0±0.1	64.1±0.3
	+MIRO (Cha et al., 2022)	99.1±0.4	93.4±1.1	40.4±4.3	29.8±2.1	**55.1±0.7**	65.4±0.3
	+LSL (Kim et al., 2024)	86.8±0.9	83.9±1.4	28.3±1.6	76.9±0.2	26.7±1.9	64.0±0.5
(c)	FakesSense (ours)	99.3±0.1	90.0±0.3	51.6±3.3	80.6±2.4	52.3±3.8	**77.2±0.8**
	w/o MGS	87.9±1.5	77.4±0.7	32.6±0.4	71.4±1.2	35.8±3.1	64.1±0.7
	w/o ADV	98.8±0.6	91.7±1.5	**52.5±4.0**	80.0±8.2	49.8±3.7	76.9±0.9
	w/o ELR	84.1±5.7	82.5±1.8	35.8±4.7	75.4±2.6	38.1±5.3	66.2±0.7
	w/o HFE	95.9±0.9	89.8±0.3	39.8±2.4	86.7±1.0	40.9±1.5	74.0±0.2

[1]Uses a few labeled samples from all generators and content types (including OOD)

Table 3: Attribution accuracy across content shifts and generator changes on DiverGen-Attr. We report performance for generators **Seen** during training consisting of the in-distribution **Labeled** ($\mathcal{S}_l$) and **Unlabeled** ($\mathcal{S}_{ul}$) generators in addition to out-of-distribution generators **Unseen** during training ($\mathcal{S}_{us}$). Similarly, we separately report performance for in-distribution content (object categories and attributes both seen during training) and out-of-distribution (object categories and attributions both unseen during training). All results are on held-out evaluation-only images. We find FakesSense outperforms prior work by 3% on average. See Tab. 4 for results isolating content shift type.

accuracy for each generator. We average results over three independent trials and report the mean and standard deviation. The generators used in each set are shown in Tab. 2.

For GenImage (Zhu et al., 2023) experiments, we selected 50 image categories as in-distribution c_l that are seen during training and 50 different image categories as c_{us} to test the model's robustness during evaluation. Some experiments also introduce new unlabeled categories c_{ul} in unlabeled images $\mathcal{S}_{ul}$. We selected ADM (Dhariwal & Nichol, 2021), Stable Diffusion V1.4 (Rombach et al., 2022) as the labeled generators ($\mathcal{S}_l$) and Glide (Nichol et al., 2022), Wukong (Wukong, 2024) as the unlabeled generators ($\mathcal{S}_{ul}$). During testing, we introduced VQDM (Gu et al., 2022), Stable Diffusion V1.5 (Rombach et al., 2022), Midjourney (Midjourney, 2024) and BigGAN (Brock et al., 2019) as the unseen generators ($\mathcal{S}_{us}$).

4.1 Results

Tab. 3 reports attribution accuracy on DiverGen-Attr where object and attributes are changed together (i.e., the object category and attribute are novel for OOD content). When examining Tab. 3(a) which summarizes prior work, we make two primary observations. First, the CDAL method performs the best, but this is due, in part, to the fact that CDAL has an advantage over all other methods since it is provided with some samples of the unseen generators and OOD content. Yet, when compared to our approach on the first line

Content	In-Distribution			Out-of-Distribution		
Generator	**Labeled (ID)**	**Unlabeled (ID)**	**Unseen (OOD)**	**Seen (ID)**	**Unseen (OOD)**	**ALL Acc**
(a) Shift in Attributes						
ORCA (Cao et al., 2022)	82.6±23.1	79.2±21.3	35.2±12.5	79.3±11.5	33.3±12.9	64.9±3.6
CPL (Sun et al., 2023)	95.3 ±0.5	84.1±2.1	38.2±6.0	86.9±1.4	36.5±6.4	71.4±1.2
POSE (Yang et al., 2023)	87.4±1.1	**94.6±0.7**	3.9±0.9	5.2±0.7	44.9±0.6	40.4±0.3
FakesSense (ours)	**95.3±1.8**	89.2±0.4	**46.1±4.6**	**87.3±1.1**	**45.1±4.8**	**75.2±0.9**
(b) Shift in Category						
ORCA(Cao et al., 2022)	66.0±22.7	78.9±21.1	38.9±7.4	71.0±0.4	37.2±7.8	60.5±2.7
CPL (Sun et al., 2023)	**97.3±1.3**	81.9±1.1	35.0±9.0	86.3±2.3	33.5±8.5	70.1±1.9
POSE (Yang et al., 2023)	89.2±1.8	**96.3±0.6**	2.1±1.5	0.9±0.2	49.0±1.5	43.5±0.3
FakesSense (ours)	90.7±0.8	88.8±0.3	**56.6±2.7**	**88.8±0.7**	**53.3±1.7**	**77.8±0.5**

Table 4: Attribution accuracy over isolated content shifts and generator changes on DiverGen-Attr. We report performance for generators **Seen** during training consisting of the in-distribution **Labeled** ($\mathcal{S}_l$) and **Unlabeled** ($\mathcal{S}_{ul}$) generators in addition to out-of-distribution generators **Unseen** during training ($\mathcal{S}_{us}$). We report isolated content shifts in **(a)** attributes and **(b)** categories. All results are on held-out evaluation-only images. FakesSense outperforms prior work by 4-7% over both shift types.

Content	In-Distribution			Out-of-Distribution		
Generator	**Labeled (ID)**	**Unlabeled (ID)**	**Unseen (OOD)**	**Seen (ID)**	**Unseen (OOD)**	**ALL Acc**
openLDN (Rizve et al., 2022)	39.4±4.3	11.0±2.5	5.9±2.5	9.4±1.4	4.3±1.6	11.2±0.5
NCL (Zhong et al., 2021)	34.3±1.0	39.9±2.3	9.8±1.0	21.2±6.5	11.4±5.7	19.9±0.3
RankStat (Han et al., 2020)	42.7±2.5	34.1±2.3	**25.3±1.3**	19.0±1.8	**26.3±1.2**	27.2±0.2
POSE (Yang et al., 2023)	39.6±4.2	44.9±4.1	15.8±0.8	**31.8±1.2**	17.3±2.1	26.8±0.1
OCC-CLIP[1] (Liu et al., 2024b)	13.1±0.2	14.5±2.0	12.2±4.0	12.0±1.9	15.5±2.5	13.4±0.4
CPL (Sun et al., 2023)	50.0±1.9	1.3±0.9	1.1±0.6	6.1±2.6	1.4±0.4	8.6±0.4
FakesSense (ours)	**54.4±0.1**	**47.4±1.4**	21.7±3.7	25.0±1.3	21.6±3.3	**29.8±0.2**

[1]Uses a few labeled samples from all generators and content types (including OOD)

Table 5: Attribution accuracy across content shifts and generator changes on GenImage (Zhu et al., 2023). We report performance for generators **Seen** during training consisting of the in-distribution **Labeled** ($\mathcal{S}_l$) and **Unlabeled** ($\mathcal{S}_{ul}$) generators in addition to out-of-distribution generators **Unseen** during training ($\mathcal{S}_{us}$). Similarly, we separately report performance for in-distribution and out-of-distribution content. All results are on held-out evaluation-only images. We find FakesSense outperforms prior work by 3% on average.

of Tab. 3(c), our approach reports a 3% gain. Our second Tab. 3(a) observation is that many methods from prior work conflate shifts in content with changes in generators, resulting in many methods reporting better performance on Unseen generators when the content is also OOD. This also explains why some methods like VL2V-ADiP (Addepalli et al., 2024) reports better performance on Unseen generators than Seen generators on OOD data- the model is predicting many samples from OOD content as an Unseen generator.

In Tab. 3(b) we report performance of CPL (Sun et al., 2023), which our FakesSense method is based upon, and alternative candidate methods for improvement. For example, as the key challenge in CREdIT stems from challenges in generalizing to unseen image content, one could try to correct it by using methods from Domain Generalization (DG) like MIRO (Cha et al., 2022) and URM (Krishnamachari et al., 2024). Similarly, ELR could be argued as a very dated Learning from Noisy Labels (LNL) method, so we could use more recent approaches like LSL (Kim et al., 2024). However, all of these methods actually hurt performance on how task, highlighting how the challenges we face are different than those being benchmarked in the DG and LNL literature, and care must be used to ensure that any components help address our core challenges.

In Tab. 3(c) we report the results of our FakesSense baseline as well as a leave-one-out ablation of our model components (see Sec. A for detailed ablation) showcasing their usefulness. Further, we report a 3% overall gain when comparing our full model in the first line of Tab. 3(c) to prior work in Tab. 3(a,b). While one can

Generator	Labeled (ID)	Unlabeled (ID)	Unseen (OOD)	ALL Acc
VL2V-ADiP (Addepalli et al., 2024)	**89.3±2.6**	45.0±1.2	25.2±4.7	51.7±0.6
openLDN (Rizve et al., 2022)	1.0±0.3	16.7±3.5	34.0±3.2	18.1±0.1
NCL (Zhong et al., 2021)	26.1±9.6	29.3±13.5	38.8±13.3	31.8±3.9
RankStats (Han et al., 2020)	62.6±7.1	44.5±6.0	49.7±3.3	52.2±0.5
ORCA (Cao et al., 2022)	85.9±2.2	58.9±2.4	38.5±3.4	59.9±0.6
POSE (Yang et al., 2023)	69.6±6.2	**72.2±6.6**	47.7±5.3	60.1±2.4
OCC-CLIP[1] (Liu et al., 2024b)	27.2± 1.7	46.0± 2.0	53.8± 1.5	42.9± 1.0
CPL (Sun et al., 2023)	85.5±2.3	59.0±3.5	37.6±2.9	59.5±1.2
FakesSense (ours)	81.0±0.8	57.8±2.0	**54.8±1.8**	**64.0±0.3**

[1]Uses a few labeled samples from all generators and content types (including OOD)

Table 6: Attribution accuracy across content shifts and generator changes on DiverGen-Type. We report performance for generators **Seen** during training consisting of the in-distribution **Labeled** ($\mathcal{S}_l$) and **Unlabeled** ($\mathcal{S}_{ul}$) generators in addition to out-of-distribution generators **Unseen** during training ($\mathcal{S}_{us}$). Note that all reported results have content shifts from training and reported results are on held-out evaluation-only images. FakesSense outperforms prior work by 4% on average.

Generator	Labeled (ID)	Unlabeled (ID)	Unseen (OOD)	ALL Acc
AE (Corvi et al., 2022; Zhu et al., 2023)	**70.7±4.1**	**73.5±1.2**	60.7±3.3	67.8±0.7
HFE	63.0±1.3	72.6±1.3	**70.8±0.3**	**68.9±0.1**

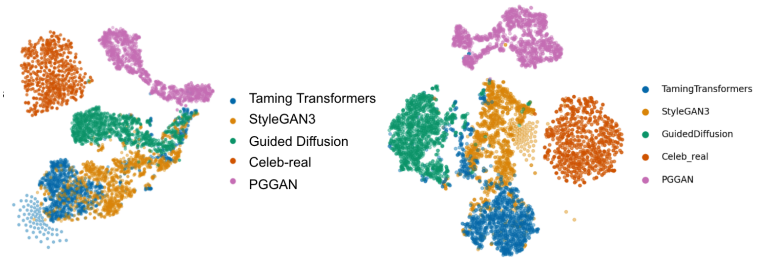
Table 7: Comparison of HFE with existing artifact extraction (AE) methods. HFE outperforms prior work overall despite an unfair comparison where HFE is extracted per-image, but prior work computes image statistics using 1K labeled images per generator.

argue that our approach simply combines components that manifest themselves in prior work, most of these were used in other tasks (MGS, ADV, ELR) or required adaptation (HFE) and not generator attribution. This suggests that the new evaluation setting provided by CREdIT is necessary to ensure effective attribution models are developed, which we further validate below.

Specifically, let us consider three other settings. First, Tab. 4 reports results on DiverGen-Attr where the shifts in content are less pronounced as we isolate the effect of attribute and category shifts. Second, Tab. 5 assess performance on GenImage (Zhu et al., 2023), where there are many content shift categories, but few samples per category. Third, Tab. 6 outlines results on DiverGen-Type, which contains diverse generator types where there are many samples per content category. Comparing across these tables shows that the performance from most methods from prior work collapses in one or more settings. For example, POSE (Yang et al., 2023) performed relatively well in Tab. 3(a) where content shifts were more pronounced, but performs about half as well as other methods in Tab. 4 when content shifts are more nuanced. Similarly, CPL (Sun et al., 2023), the method we build off of, shows vulnerabilities in Tab. 5 when few samples of a large variety of content categories are provided. openLDN (Rizve et al., 2022) also performs poorly in both Tab. 5 and Tab. 6. However, our simple baseline performs 3-7% better than prior work across all these settings.

Artifact analysis. Prior work in artifact extraction (AE) created distinct fingerprints generated by different generators (Corvi et al., 2022; Zhu et al., 2023). However, these methods are rarely utilized in attribution tasks due to using 1K labeled samples to construct noise residuals for each generator. However, in CREdIT we only have labeled samples for a subset of generators. Thus, as we show in Tab. 7, AE overfits to seen generators, resulting in our HFE approach (which is applied per-image) obtaining a 10% improvement on unseen generators, and a 1% gain overall that is also more stable.

tSNE visualization. Fig. 4 provides a visualization of the feature distribution of the unseen category during evaluation that compares FakesSense against CPL (Sun et al., 2023). Based on Fig. 4, it shows that



(a) CPL (Sun et al., 2023) (b) FakesSense (ours)

Figure 4: tSNE (van der Maaten & Hinton, 2008) comparison over the unseen generators in DiverGen-Type. We find our approach improves generalization by creating more compact and better-separated generator clusters.

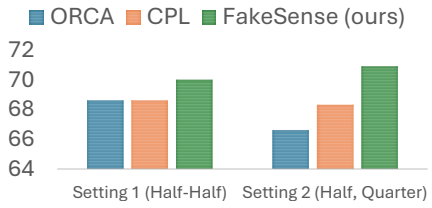


Figure 5: Accuracy across all generator types when the unlabeled split is imbalanced in Setting 1 and Setting 2. FakesSense outperforms prior work by 1-2%. See Sec. 4.1 for discussion.

our method effectively promotes features from the same generator class into their respective compact feature space to promote separability. Thus, FakesSense significantly improves the quality of feature grouping and generalization on novel, unseen attacks.

Imbalanced data distribution. Given that real-world data distributions are rarely balanced, we introduced an imbalanced training set on the unlabeled generators to analyze the robustness of our proposed method. Initially, each of the four unlabeled generators contained 1300 training samples. In Setting 1, the number of samples per generator was adjusted: LSGM (Vahdat et al., 2021) was reduced to 1000 samples, FaceAPP (FaceApp, 2024) to 800, CycleGAN (Zhu et al., 2020) to 600 samples, while StyleGAN2 (Karras et al., 2020) remained at 1300 samples. In Setting 2, LSGM had 1000 samples, FaceAPP had 800 samples, StyleGAN2 remained at 1300, but CycleGAN was further reduced to only 300 training samples. Fig. 5 shows FakesSense outperforms prior work (Sun et al., 2023; Cao et al., 2022) even in the presence of imbalanced data distribution. This highlights FakesSense's robustness in maintaining high attribution ability even under such challenging training conditions.

4.2 Additional Analysis on Content Shifts

How does the number of training categories affect performance? Tab. 8 reports performance when varying the number of categories during training, but keep the number of categories at test time fixed. We find adding more categories reports gains until the training set has the same number of categories as the test set, but decreases when we add more categories. This suggests that at some point the model begins to overfit to the category shifts, making it hard to generalize across generator shifts.

To better understand the effect the number of training images per content category has on performance, Tab. 9 keeps the number of training images fixed at 1.5K per generator, effectively reducing the number of

# Categories	Content In-Distribution			Content Out-of-Distribution		
	Labeled (ID)	Unlabeled (ID)	Unseen (OOD)	Seen (ID)	Unseen (OOD)	ALL Acc
5	**79.2±1.3**	13.2±1.0	9.9±1.2	9.5±1.0	22.0±2.1	21.9±0.2
10	27.1±1.0	24.3±2.3	19.8±2.0	22.6±1.3	21.9±2.2	22.5±0.2
25	24.4±3.2	31.6±4.5	25.0±2.5	**26.6±3.3**	14.3±1.4	23.5±0.2
50	42.7±2.5	34.1±2.3	**25.3±1.3**	19.0±1.8	**26.3±1.2**	**27.2±0.2**
100	27.3±1.8	**49.8±5.0**	10.7±1.1	13.9±1.4	19.5±2.0	20.7±0.2

Table 8: Measuring the effect the number of content categories has on attribution using GenImage (Zhu et al., 2023). See Sec. 4.2 for discussion.

# Categories	#Per Category	Content In-Distribution			Content Out-of-Distribution		
		Labeled (ID)	Unlabeled (ID)	Unseen (OOD)	Seen (ID)	Unseen (OOD)	ALL Acc
10	150	26.2±0.3	37.4±0.4	25.7±0.3	20.2±0.2	11.7±0.1	22.3±0.2
15	100	27.6±0.3	18.6±0.2	21.6±0.2	11.8±0.1	24.7±0.3	20.3±0.2
50	30	**40.6±0.4**	**38.9±0.4**	23.3±0.2	10.5±0.1	18.7±0.2	23.1±0.2
100	15	21.1±0.2	37.6±0.4	**26.4±0.3**	**29.3±0.3**	12.6±0.1	**24.4±0.2**
300	5	36.8±0.4	19.2±0.2	19.0±0.2	10.8±0.1	**29.5±0.3**	21.8±0.2

Table 9: Compares the effect the number of image categories and data samples using GenImage (Zhu et al., 2023). The total number of samples is maintained at 1,500 for overall consistency. See Sec. 4.2 for discussion.

# Categories	Content In-Distribution			Content Out-of-Distribution		
	Labeled (ID)	Unlabeled (ID)	Unseen (OOD)	Seen (ID)	Unseen (OOD)	ALL Acc
10 (animals, ID)	41.3±6.2	**29.7±7.1**	**24.1±3.5**	24.4±1.8	**16.0±1.9**	25.0±0.6
10 (food, OOD)	**46.6±3.4**	26.3±5.7	22.7±4.4	**29.6±1.8**	15.0±3.7	**25.9±0.4**

Table 10: Contrasting the effect of adding 10 new images during training that are in-distribution (animals) or 10 out-of-distribution (food) categories affects performance on GenImage (Zhu et al., 2023). We find adding OOD categories is more effective. See Sec. 4.2 for discussion.

images per category as we increase the number of categories. However, as in our previous experiment, we use the same 50 categories for inference. We find that there is less variance between the number of categories, suggesting a minimum number of images per category is needed to learn a a good representation.

Does it matter what types of categories are added? Tab. 10 compares adding 10 categories similar to those in our training set or selected to be diverse. Specifically, we used an initial set of 50 categories containing entirely animal classes, and we either added more animals or included food categories. We find a small benefit to adding diverse categories, although this mainly stems from gains on labeled samples.

Does it matter to what image generator subset new categories appear? Tab. 11 reports the performance of adding 25 new categories to the various image subsets. We find that adding new categories only to the labeled boosts performance on seen and unlabeled samples, whereas adding them only to the unlabeled images only benefits generalization to unseen generators. Adding the categories to both seen and unlabeled generators obtains the best performance, just as adding them only to the unseen generators images hurts performance overall as the model confuses category shifts for generator changes.

5 Conclusion

In this paper, we explore Content Robust imagE generator attrIbuTion (CREdIT), which contains both category and generator shifts in image generator attribution. To investigate our task we introduce DiverGen-a new benchmark constructed with diverse generators over multiple types of content shifts. We find state-of-the-art in attribution method's performance often collapses in settings explored CREdIT. Instead, we introduce a simple baseline we refer to as FakesSense that boosts performance by 3-7%. Future work could focus on efficient adaptation methods to quickly adapt to emerging threats from new image generators.

Img Subset	Labeled (ID)	Unlabeled (ID)	Unseen (OOD)	ALL Acc
Original	67.0±2.6	59.9±3.9	48.6±1.8	56.0±0.3
Labeled	**83.4±1.1**	**63.8±2.4**	44.5±1.4	59.0±0.7
Unlabeled	59.3±4.2	52.3±4.9	**62.4±4.0**	59.1±0.6
Seen	81.4±2.0	55.1±1.6	61.5±1.5	**64.9±0.1**
Unseen	51.4±1.0	53.6±2.8	56.2±1.1	54.8±0.6

Table 11: Effect of introducing 25 new image categories to different subsets on GenImage (Zhu et al., 2023). Adding new image categories to the Seen (Labeled+Unlabeled) generators during training results in the highest performance. Original means no new image categories are introduced. See Sec. 4.2 for discussion.

References

Sravanti Addepalli, Ashish Ramayee Asokan, Lakshay Sharma, and R. Venkatesh Babu. Leveraging vision-language models for improving domain generalization in image classification, 2024. URL https://arxiv.org/abs/2310.08255.

Muhammad Anas Raza and Khalid Mahmood Malik. Multimodaltrace: Deepfake detection using audiovisual representation learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 993–1000, 2023. doi: 10.1109/CVPRW59228.2023.00106.

Shivangi Aneja and Matthias Nießner. Generalized zero and few-shot transfer for facial forgery detection, 2020.

Pietro Bongini, Sara Mandelli, Andrea Montibeller, Mirko Casu, Orazio Pontorno, Claudio Vittorio Ragaglia, Luca Zanchetta, Mattia Aquilina, Taiba Majid Wani, Luca Guarnera, Benedetta Tondi, Giulia Boato, Paolo Bestagini, Irene Amerini, Francesco G. B. De Natale, Sebastiano Battiato, and Mauro Barni. Wild: a new in-the-wild image linkage dataset for synthetic image attribution. *CoRR*, abs/2504.19595, April 2025. URL https://doi.org/10.48550/arXiv.2504.19595.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019. URL https://arxiv.org/abs/1809.11096.

Kaidi Cao, Maria Brbić, and Jure Leskovec. Open-world semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. *European Conference on Computer Vision (ECCV)*, 2022.

Yiqun Chen and James Zou. Twigma: A dataset of ai-generated images with metadata from twitter, 2023. URL https://arxiv.org/abs/2306.08310.

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, 2018.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models, 2022.

H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain. On the detection of digital face manipulation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5780–5789, Los Alamitos, CA, USA, jun 2020a. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.00582. URL https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00582.

Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil Jain. On the detection of digital face manipulation, 2020b.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers, 2021a. URL https://arxiv.org/abs/2105.13290.

Yuzhen Ding, Nupur Thakur, and Baoxin Li. Does a gan leave distinct model-specific fingerprints? In *BMVC*, 2021b.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12868–12878, 2021. doi: 10.1109/CVPR46437.2021.01268.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

FaceApp. https://faceapp.com/app, 2024. Accessed: 2024-06-05.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, 2015.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis, 2022. URL https://arxiv.org/abs/2111.14822.

Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization, 2023.

Xiao Guo, Vishal Asnani, Sijia Liu, and Xiaoming Liu. Tracing hyperparameter dependencies for model parsing via learnable graph pooling network, 2024. URL https://arxiv.org/abs/2312.02224.

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.

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, 2021. URL https://arxiv.org/abs/2103.05630.

F. Hong, L. Zhang, L. Shen, and D. Xu. Depth-aware generative adversarial network for talking head video generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3387–3396, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.00339. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00339.

Yan Hong and Jianfu Zhang. Wildfake: A large-scale challenging dataset for ai-generated images detection, 2024. URL https://arxiv.org/abs/2402.11843.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11975–11985, June 2024.

Jiashang Hu, Shilin Wang, and Xiaoyong Li. Improving the generalization ability of deepfake detection via disentangled representation learning. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 3577–3581, 2021. doi: 10.1109/ICIP42928.2021.9506730.

Galadrielle Humblot-Renaux, Sergio Escalera, and Thomas B Moeslund. A noisy elephant in the room: Is your out-of-distribution detector robust to label noise? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22626–22636, 2024.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization, 2019.

Anubhav Jain, Pavel Korshunov, and Sébastien Marcel. Improving generalization of deepfake detection by training for attribution. In *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, 2021. doi: 10.1109/MMSP53017.2021.9733468.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.00813. URL https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00813.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ArXiv*, abs/1710.10196, 2017. URL https://api.semanticscholar.org/CorpusID:3568073.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:235606261.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2416–2425, 2022. doi: 10.1109/CVPR52688.2022.00246.

Noo-Ri Kim, Jin-Seop Lee, and Jee-Hyong Lee. Learning with structural labels for learning with noisy labels. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27600–27610, 2024. URL https://api.semanticscholar.org/CorpusID:272724665.

Kiran Krishnamachari, See-Kiong Ng, and Chuan-Sheng Foo. Uniformly distributed feature representations for fair and robust learning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=PgLbS5yp8n.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Sangyup Lee, Shahroz Tariq, Junyaup Kim, and Simon S. Woo. Tar: Generalized forensic framework to detect deepfakes using weakly supervised learning, 2021.

Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen. Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24070–24079, 2023.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3204–3213, 2020a. doi: 10.1109/CVPR42600.2020.00327.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics, 2020b.

Fengyuan Liu, Haochen Luo, Yiming Li, Philip Torr, and Jindong Gu. Which model generated this image? a model-agnostic approach for origin attribution. In *European Conference on Computer Vision (ECCV)*, 2024a.

Fengyuan Liu, Haochen Luo, Yiming Li, Philip Torr, and Jindong Gu. Which model generated this image? a model-agnostic approach for origin attribution, 2024b. URL https://arxiv.org/abs/2404.02697.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024c.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33: 20331–20342, 2020a.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels, 2020b.

Xinwei Liu, Jian Liu, Yang Bai, Jindong Gu, Tao Chen, Xiaojun Jia, and Xiaochun Cao. Watermark vaccine: Adversarial attacks to prevent watermark removal. In *European Conference on Computer Vision (ECCV)*, 2022.

Lixin Luo, Zhenyong Chen, Ming Chen, Xiao Zeng, and Zhang Xiong. Reversible image watermarking using interpolation technique. *IEEE Transactions on Information Forensics and Security*, 5(1):187–193, 2010.

Zheling Meng, Bo Peng, and Jing Dong. Latent watermark: Inject and detect watermarks in latent diffusion space. *IEEE Transactions on Multimedia*, pp. 1–12, 2025.

Midjourney, 2024. URL `https://www.midjourney.com/home/`. Accessed: 2024-06-05.

João C. Neves, Ruben Tolosana, Ruben Vera-Rodriguez, Vasco Lopes, Hugo Proença, and Julian Fierrez. Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1038–1048, 2020. doi: 10.1109/JSTSP.2020.3007250.

Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. Core: Consistent representation learning for face forgery detection, 2022.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022. URL `https://arxiv.org/abs/2112.10741`.

NVIDIA. Nvidia maxine. `https://developer.nvidia.com/maxine`, 2024. Accessed: 2024-06-19.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023.

Anisha Pal, Julia Kruk, Mansi Phute, Manognya Bhattaram, Diyi Yang, Duen Horng Chau, and Judy Hoffman. Semi-truths: A large-scale dataset of ai-augmented images for evaluating robustness of ai-generated image detectors, 2024. URL `https://arxiv.org/abs/2411.07472`.

S. Pereira and T. Pun. Robust template matching for affine resistant image watermarks. *IEEE Transactions on Image Processing*, 9(6):1123–1129, 2000.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL `https://arxiv.org/abs/2307.01952`.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues, 2020.

Rui Qiao and Bryan Kian Hsiang Low. Understanding domain generalization: A noise robustness perspective. In *The Twelfth International Conference on Learning Representations*, 2024.

Maan Qraitem, Kate Saenko, and Bryan A. Plummer. Bias mimicking: A simple sampling approach for bias mitigation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Maan Qraitem, Kate Saenko, and Bryan A. Plummer. From fake to real: Pretraining on balanced synthetic images to prevent spurious correlations in image recognition. In *European Conference on Computer Vision (ECCV)*, 2024.

Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis, 2024. URL `https://arxiv.org/abs/2404.13686`.

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*, 2022.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2021. URL https://api.semanticscholar.org/CorpusID:245335280.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner. Faceforensics++: Learning to detect manipulated facial images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society. doi: 10.1109/ICCV.2019.00009. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00009.

Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *European Conference on Computer Vision (ECCV)*, 2020.

Aliaksandr Siarohin, Stéphane Lathuilière, S. Tulyakov, Elisa Ricci, and N. Sebe. First order motion model for image animation. In *Neural Information Processing Systems*, 2020. URL https://api.semanticscholar.org/CorpusID:202767986.

Shivashankar Subramanian, Xudong Han, Timothy Baldwin, Trevor Cohn, and Lea Frermann. Evaluating debiasing techniques for intersectional biases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2492–2498, November 2021.

Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Liang Jilin, and Rongrong Ji. Dual contrastive learning for general face forgery detection. *ArXiv*, abs/2112.13522, 2021. URL https://api.semanticscholar.org/CorpusID:245502532.

Zhimin Sun, Shen Chen, Taiping Yao, Bangjie Yin, Ran Yi, Shouhong Ding, and Lizhuang Ma. Contrastive pseudo learning for open-world deepfake attribution, 2023.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12105–12114, 2023.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the upsampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28130–28139, June 2024.

Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space, 2021.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Run Wang, Felix Juefei-Xu, L. Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. In *International Joint Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:212976079.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020a.

Siqi Wang, Aoming Liu, and Bryan A. Plummer. Noise-aware generalization: Robustness to in-domain noise and out-of-domain generalization. In *International Conference on Learning Representations (ICLR)*, 2026.

Yaohui Wang, Di Yang, François Brémond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *ArXiv*, abs/2203.09043, 2022a. URL `https://api.semanticscholar.org/CorpusID:247518586`.

Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7278–7287, 2023. doi: 10.1109/CVPR52729.2023.00703.

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, 2022b. doi: 10.1109/TIFS.2022.3146766.

Zeyu Wang, Klint Qinami, Ioannis Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020b.

Wukong. `https://xihe.mindspore.cn/modelzoo/wukong`, 2024. Accessed: 2024-06-05.

Wangmeng Xiang, Chao Li, Biao Wang, Xihan Wei, Xian-Sheng Hua, and Lei Zhang. Spatiotemporal self-attention modeling with temporal patch shift for action recognition, 2022.

Ying Xu, Kiran Raja, and Marius Pedersen. Supervised contrastive learning for generalizable and explainable deepfakes detection. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 379–389, 2022. doi: 10.1109/WACVW54805.2022.00044.

T. Yang, D. Wang, F. Tang, X. Zhao, J. Cao, and S. Tang. Progressive open space expansion for open-set model attribution. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15856–15865, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.01522. URL `https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01522`.

Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan, 2022.

Ning Yu, Larry S. Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7555–7565, 2018. URL `https://api.semanticscholar.org/CorpusID:201058738`.

Ning Yu, Larry Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. doi: 10.1109/TIP.2017.2662206.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative ai. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8643–8672, 2024.

Yu Zheng, Boyang Gong, Fanye Kong, Yueqi Duan, Bingyao Yu, Wenzhao Zheng, Lei Chen, Jiwen Lu, and Jie Zhou. Learning counterfactually decoupled attention for open-world model attribution. In *International Conference on Computer Vision (ICCV)*, 2025.

				In-Distribution			Out-of-Distribution		
MGS	ADV	ELR	HFE	**Labeled (ID)**	**Unlabeled (ID)**	**Unseen (OOD)**	**Seen (ID)**	**Unseen (OOD)**	**ALL Acc**
				95.2±2.2	89.0±0.4	29.1±2.0	**86.4±1.6**	27.4±3.9	69.4±0.5
✓				92.3±1.8	89.7±0.7	44.6±2.4	84.9±1.2	43.2±0.3	73.8±0.1
	✓			86.4±2.4	69.0±1.8	40.9±0.9	70.8±0.5	42.6±1.3	64.2±0.1
		✓		99.8±0.1	84.8±1.8	52.4±8.6	76.4±4.5	49.9±9.5	74.8±1.7
			✓	83.3±3.9	82.6±1.7	39.1±1.9	74.4±1.1	43.4±2.6	67.3±0.2
✓		✓		89.0±0.6	83.8±2.3	45.1±0.9	77.8±0.5	41.1±0.7	69.8±0.5
✓			✓	92.6±2.9	86.6±1.7	48.7±0.5	84.4±0.7	46.1±1.6	74.2±0.2
✓	✓			92.8±2.8	88.8±0.7	37.9±6.3	86.4±1.7	32.6±3.2	71.0±0.7
		✓	✓	99.6±0.4	85.1±2.7	49.6±2.0	74.3±2.6	49.8±2.1	73.9±0.2
	✓		✓	88.8±2.7	80.7±0.2	37.0±7.2	75.0±2.1	37.9±7.3	66.8±1.4
	✓	✓		94.0±1.3	89.4±0.8	41.0±1.9	84.4±1.3	43.5±2.6	73.6±0.1
	✓	✓	✓	87.9±1.5	77.4±0.7	32.6±0.4	71.4±1.2	35.8±3.1	64.1±0.7
✓		✓	✓	98.8±0.6	91.7±1.5	52.5±4.0	80.0±8.2	49.8±3.7	76.9±0.9
✓	✓		✓	84.1±5.7	82.5±1.8	35.8±4.7	75.4±2.6	38.1±5.3	66.2±0.7
✓	✓	✓		95.9±0.9	89.8±0.3	39.8±2.4	86.7±1.0	40.9±1.5	74.0±0.2
✓	✓	✓	✓	99.3±0.1	90.0±0.3	**51.6±3.3**	80.6±2.4	52.3±3.8	**77.2±0.8**

Table 12: Ablation study on different components of FakesSense on DiverGen-Attr. We find each component contributes to the overall performance. See Tab. 13 for an ablation over DiverGen-Type.

Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10862–10870, 2021. doi: 10.1109/CVPR46437.2021.01072.

P. Zhou, X. Han, V. I. Morariu, and L. S. Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1831–1839, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. doi: 10.1109/CVPRW.2017.229. URL `https://doi.ieeecomputersociety.org/10.1109/CVPRW.2017.229`.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image, 2023. URL `https://arxiv.org/abs/2306.08571`.

A Additional Experiments

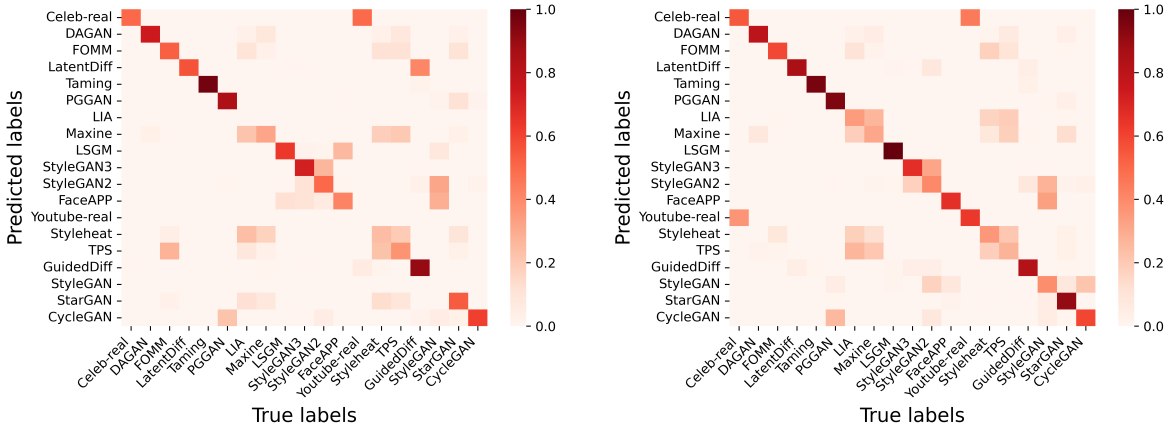
Ablation study. We examine how different components of FakesSense contribute to the overall performance in Tab. 12 and Tab. 13. We find that all individual components of FakesSense contribute to FakesSense overall performance. Removing each component from FakesSense reveals the roles of each component. Removing ADV or ELR results in a drop in performance on labeled and unlabeled in-distribution generators. Removing HFE leads to a drop in performance across all categories, showing that HFE is essential for learning stable feature representations. When combining all 4 components of FakesSense (MGS, ADV, ELR, HFE), we achieve the best overall accuracy with a 3.2% boost on DiverGen-Type.

Generator confusion. Fig. 6 compares confusion matrix for CPL (Sun et al., 2023) and FakesSense. We find CPL confuses some models from similar architectures (StyleGAN2, StyleGAN3, Guided Diffusion, and Latent Diffusion), as well as Youtube-real images. When compared with our FakesSense, we observe a significant gain in attribution accuracy, especially for classes that CPL struggles most with.

Training on a mix of category domains. As illustrated in Tab. 14, training on multiple category domains helps to improve overall performance, This is largely due to challenges in identifying the generators for new

MGS	ADV	ELR	HFE	Labeled (ID)	Unlabeled (ID)	Unseen (OOD)	ALL Acc
				85.5±**2.3**	**59.0**±**3.5**	37.6±2.9	59.5±1.2
	✓	✓	✓	76.8±3.2	51.5±1.7	53.6±0.9	60.3±0.3
✓		✓	✓	75.3±3.4	53.4±3.1	52.2±1.0	59.9±0.4
✓	✓		✓	79.5±0.5	49.5±3.4	54.5±3.1	60.8±0.3
✓	✓	✓		66.8±4.4	38.0±0.6	53.8±2.6	52.9±0.5
✓	✓	✓	✓	79.2±1.8	53.1±0.2	**56.7**±**1.7**	**62.7**±**0.1**

Table 13: Ablation study on different components of FakesSense on DiverGen-Type validation set. In the first row, we report the performance of the baseline experiment without introducing any key components of FakesSense. FakesSense shows the best overall performance after introducing MGS, ADV, ELR, and HFE.



(a) CPL (Sun et al., 2023)

(b) FakesSense

Figure 6: Comparing confusion matrices on DiverGen-Type. We find that much of the gain made by FakesSense is due to significant improvements in a few categories where CPL performs poorly.

object categories. This echoes similar observations made in the generated image detection literature (Wang et al., 2020a).

How does the number of images per category during training affect performance?

Tab. 15 reports the performance when varying the number of images per object categories during training. Based on the performance, We find that adding more images per category results in significant improvements across all generator categories. This suggests that the model demonstrates significant performance gains through more diverse data samples. Thus, the number of images per object category contributes significantly to attribution accuracy. For experiment that uses 2,000 to 10,000 samples, we used a variable number of images per category to make the model more robust. Notably, the highest overall accuracy of 64% was achieved with the greatest number of images per category at variable number. The results highlight the importance of having a comprehensive dataset with sufficient number of images for each object category, highlighting the critical characteristic of DiverGen-Type in developing effective attribution methods.

Artifact analysis. In previous works (Corvi et al., 2022; Zhu et al., 2023), artifact analysis has been used to examine the distinct fingerprints generated by different generators. We employed the same procedure by first extracting 1,000 noise residuals using the denoising filter $f(\cdot)$ in DnCNN (Zhang et al., 2017) on input image $\boldsymbol{X}$ to get $R_i = \boldsymbol{X}_i - f(\boldsymbol{X}_i)$. We then averaged these 1000 image residuals to obtain$\hat{F} = \frac{1}{1000}\sum_{i=1}^{1000} R_i$. Finally, we applied Fourier transform on the averaged residuals to perform the artifact analysis as illustrated in Fig. 10.

Method	Face	Furni.	Build.	Art	Anim.	Vehi.
FakesSense-Face	51.2	78.6	51.5	27.1	36.3	4.3
FakesSense-Multi	62.6	53.6	72.9	74.9	62.8	66.5

Table 14: Per-category accuracy across object category types on DiverGen-Type comparing training only on faces vs. multiple categories. We find multi-category training is key for good performacne.

# Per Categories	**Labeled (ID)**	**Unlabeled (ID)**	**Unseen (OOD)**	**ALL Acc**
150	51.7±1.2	43.5±2.0	47.9±1.1	47.7±1.4
500	71.9±0.9	56.1±1.4	44.9±2.1	57.0±0.6
1000	67.5±1.2	**60.0±1.4**	53.5±2.1	60.0±0.7
Full Dataset	**81.0±0.8**	57.8 ±2.0	**54.8±1.8**	**64.0±0.3**

Table 15: Evaluate the effects of the number of images per category on the **Labeled** (ID), **unlabeled** (ID), and **Unseen** (OOD) generators on DiverGen-Type.

Based on Fig. 10, we observe interesting artifact patterns from LSGM, CycleGAN, and PGGAN generators. However, images from celeb-real and stylegan3, as well as between GuidedDiffusion and Youtube-real, demonstrate some degree of resemblance, posing a greater challenge for attribution tasks. Since the artifacts show similar patterns across different generators, detecting patterns in individual images from different image categories can be a greater challenge. Enhancing specific patterns on each individual input through high-frequency extraction (HFE) can be more effective in capturing key characteristic features in the raw images.

However, such artifact analysis (Corvi et al., 2022; Zhu et al., 2023) are rarely utilized in attribution tasks due to the need for large image samples to construct noise residuals per generator. This approach becomes more challenging with insufficient training samples. Our proposed HFE method effectively mitigates the need for generator-specific artifact construction. Instead, HFE captures characteristics of individual images, allowing it to be applied on a per-image basis without relying on large datasets.

Sensitivity analysis. We provide analysis in Fig. 7 on how different hyperparameter values for each individual component of FakesSense can affect overall performance. Specifically, the ranges for each hyperparameter are as follows: MGS loss weight in $[10^{-3}, 10^{2}]$, ADV loss weight in [0.1, 0.4], scaling factor (ν) for high-frequency extraction (HFE) in [0.5, 2.5], and lambda value in $[10^{-3}, 10^{1}]$ for ELR (Liu et al., 2020a). Based on the results in Fig. 7, the weight for MGS loss has only a small effect on the overall accuracy, thus allowing diverse values to be used for the weight during training. ELR also shows similar behavior as MGS loss where only a large weight value of 10 significantly affects overall performance. For ADV, the highest performance is reached when using a coefficient of 0.3. HFE also achieves the best performance when the scaling factor is set to 1.0.

Complex category domain shift. To evaluate FakesSense ability to handle complex, real-world shifts we trained and tested FakesSense on different object categories to supplement our experiments on GenImage (Zhu et al., 2023) in the main paper. In Tab. 16, we provide the detailed information about the image categories and generators used in both the training and evaluation sets.

As shown in Tab. 17, FakesSense demonstrates improved performance compared to prior works, particularly on the $\mathcal{S}_{ul}$ generators across different object categories from the training set with a 20% improvement in accuracy compared to CPL (Sun et al., 2023). FakesSense outperforms prior works further highlights its ability to generalize to more complex real-world shifts.

Reducing the effect of pseudo-label noise. Tab. 18 has analogous observations of incorporating methods from the Learning with Noisy Labels literature as seen with the DG methods. Specifically, that more sophisticated recent methods (*e.g.*, Kim et al. (2024)) underperform the regularization-based techniques from older work like ELR (Liu et al., 2020a). This is further validated in Fig. 11, where we report the pseudo label F1-score for samples in our unlabeled set during training. We find that combining ELR with

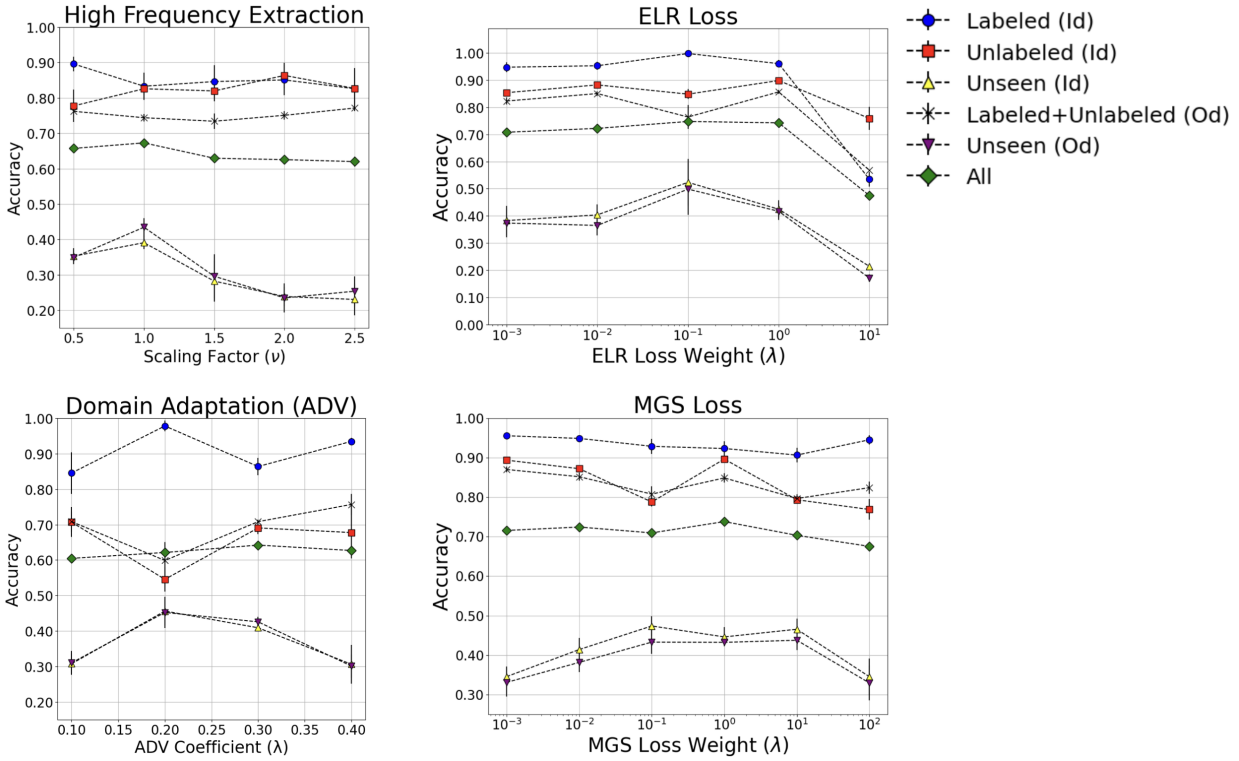


Figure 7: Sensitivity analysis for hyperparameter values. We find that FakesSense is relatively stable across multiple settings.

Train	Evaluation (**Labeled(ID)+Unlabeled(ID)**)	Evaluation (**Unseen(OD)**)
Guided diffusion: bed (labeled)	Guided diffusion: horse	LSGM: face
StyleGAN3: face (labeled)	StyleGAN3: animal	StyleGAN3: art
StyleGAN2: car (unlabeled)	StyleGAN2: horse	StyleGAN2: church
Taming transformer: cat (unlabeled)	Taming transformer: face	Latent diffusion: bed

Table 16: Categorical and generator information on training and evaluation set. During evaluation, we either used generators seen during training, Labeled(ID) and Unlabeled(ID) but with new image categories, or we introduced unseen generators or object categories for Unseen(OD).

our MGS loss significantly improves the quality of pseudo-labels by reducing its high amount of label noise, boosting performance.

B Details of DiverGen Dataset

Existing datasets on attribution tasks (e.g., (Neves et al., 2020; Zhou et al., 2017; Wang et al., 2019; Dang et al., 2020a; He et al., 2021)) comprises mainly of face images generated by generative adversarial networks (GAN). Based on Tab. 19, many of these datasets have the limitations of containing only a single image category (faces) and lacking new generative models. While GenImage (Zhu et al., 2023) consists of 1K image classes, each of the object class contains only 162 training images and 6 testing images. The limited number of images per category makes it hard to train effective attribution models. Furthermore, current datasets mainly consider shifts in generators, neglecting more diverse shifts such as shift in category or attribute.

	Labeled (ID)	Unlabeled (ID)	Unseen (OOD)	ALL Acc
ORCA (Cao et al., 2022)	43.7±11.6	46.5±1.3	63.8±8.6	54.4±2.6
CPL (Sun et al., 2023)	**65.0±3.1**	44.0±5.5	56.7±3.3	55.6±1.9
Ours	64.1±2.5	**66.0±1.9**	**57.6±0.7**	**61.3±0.4**

Table 17: Performance comparison under category shifts. We find FakesSense outperforms prior work by 6-7% under these complex domain shifts.

Figure 8: Example of proposed high-frequency extraction approach. Left to right: input image, the low-frequency image component extracted with a Gaussian blur filter, our high-frequency detail enhancement with $\nu = 1.5$.

B.1 DiverGen-Attr subset

To consider more robust distribution shifts, we selected six generators: CogView (Ding et al., 2021a), HyperSD (Ren et al., 2024), Flux (Labs, 2024), Midjourney (Midjourney, 2024), Stable Diffusion XL (Podell et al., 2023), and Stable Diffusion 3M (Esser et al., 2024). Each generator was tasked to produce 100 images for every object-attribute pair using the same 100 object classes as CIFAR-100 (Krizhevsky, 2012). We then introduced different environmental and seasonal attributes to control image content that are Default, Indoor, Outdoor, Spring, Summer, Autumn, Winter, Night, and Sunny. Thus, we have 100 images per class × 100 classes × 6 generators × 9 attributes = 540,000 images in DiverGen-Attr.

B.2 DiverGen-Type subset

To further overcome limitations in existing datasets, our proposed DiverGen-Type dataset addresses these shortcomings by improving the diversity of generative models in the dataset and introducing multiple image categories to address the domain shift issue in open-set problems. Furthermore, DiverGen offers substantially more images per object category (roughly 5K on average vs. 168 for GenImage (Zhu et al., 2023)).

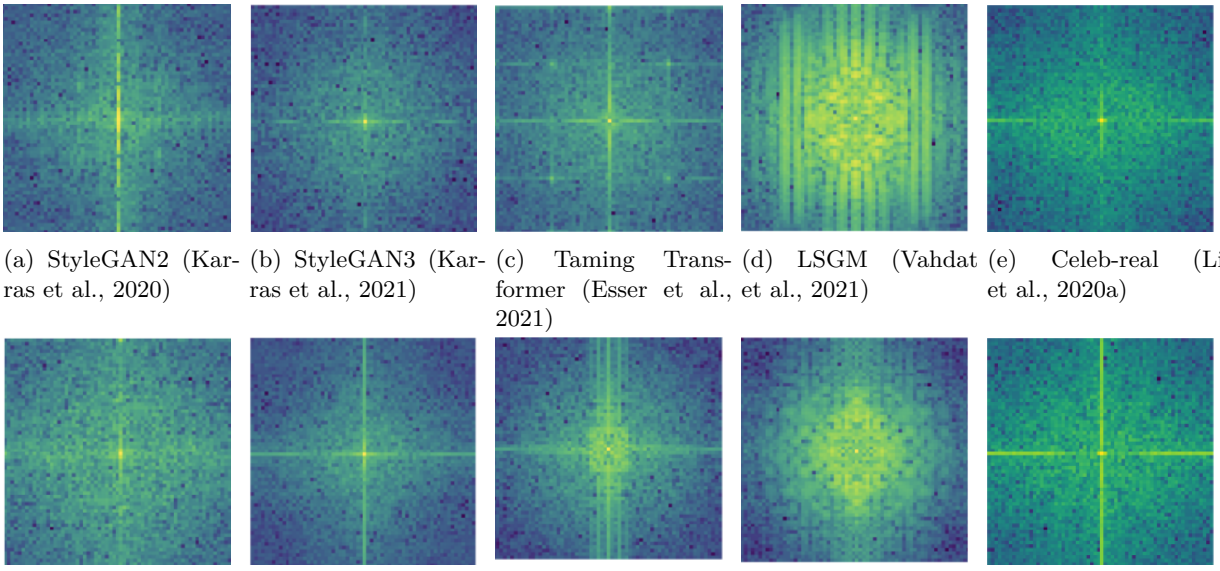
In addition to incorporating new generators, we have also included three existing datasets for additional variability. Details follow:

CelebDFv2 (Li et al., 2020a) is a dataset comprising a total of 5,639 fake videos generated using an improved synthesis method comprising a diverse range of genders, age groups, and ethnicity.

DFFD (Dang et al., 2020a) consists a total of 0.8M of real faces and 1.8M of fake faces created by a diverse range of facial manipulation methods.

(a) StyleGAN3 (Karras et al., 2021) (b) PGGAN (Karras et al., 2017) (c) Celeb Real (Li et al., 2020a) (d) Guided Diffusion (Kim et al., 2022) (e) Taming Transformer (Esser et al., 2021)

Figure 9: Sample images from DiverGen dataset. DiverGen contains a wide range of GAN and transformer/diffusion-based generators that produce high-quality images that are even challenging for humans to distinguish from real ones.



(a) StyleGAN2 (Karras et al., 2020) (b) StyleGAN3 (Karras et al., 2021) (c) Taming Transformer (Esser et al., 2021) (d) LSGM (Vahdat et al., 2021) (e) Celeb-real (Li et al., 2020a)

(f) Latent Diffusion (Rombach et al., 2021) (g) Guided Diffusion (Kim et al., 2022) (h) CycleGAN (Zhu et al., 2020) (i) PGGAN (Karras et al., 2017) (j) Youtube-real (Li et al., 2020a)

Figure 10: From top to bottom we show the artifact generated from 1000 noise residual of stylegan2 (Karras et al., 2020), stylegan3 (Karras et al., 2021), taming transformer (Esser et al., 2021), lsgm (Vahdat et al., 2021), celeb-real (Li et al., 2020a), latent diffusion (Rombach et al., 2021), guided diffusion (Kim et al., 2022), cyclegan (Zhu et al., 2020), pggan (Karras et al., 2017), and youtube-real (Li et al., 2020a).

ForgeryNIR (Wang et al., 2022b) contains a total of 50,000 fake and real identities in the dataset where all images are generated by GAN models. Various perturbations are then applied to simulate real-world scenarios.

New Generators. DiverGen includes both commonly used GAN models in image generation detection tasks, including StyleGAN (Karras et al., 2019), StyleGAN2 (Karras et al., 2020), SyleGAN3 (Karras et al., 2021), StarGAN (Choi et al., 2018), PGGAN (Karras et al., 2017), DAGAN (Hong et al., 2022), CycleGAN (Zhu et al., 2020), as well as more robust and recent generators FaceAPP (FaceApp, 2024), LSGM (Vahdat et al., 2021), FOMM (Siarohin et al., 2020), LIA (Wang et al., 2022a), LatentDiffusion (Rombach et al., 2021), Maxine (NVIDIA, 2024), TPS (Xiang et al., 2022), StyleHEAT (Yin et al., 2022), GuidedDiffusion (Kim et al., 2022), and TamingTransformer (Esser et al., 2021). DiverGen comprises of 17 generator types resulting in 268,269 generated images as shown in Tab. 20 with at least 10K new images generated by each generator. DiverGen also includes 84,009 real human faces images selected from celeb-real and youtube-real subsets in CelebDFv2 dataset (Li et al., 2020a).

Generator	Labeled (ID)	Unlabeled (ID)	Unseen (OOD)	ALL Acc
DISC (Li et al., 2023)	34.8±2.1	40.1±4.7	**51.5±5.6**	42.7±0.5
LSL (Kim et al., 2024)	76.8±4.8	**64.7±0.2**	25.2±4.7	49.7±1.8
ELR (Liu et al., 2020a)	**79.0±1.2**	53.0±0.5	46.1±3.9	**58.7±0.9**

Table 18: Comparison with LNL methods combined with CPL (Sun et al., 2023) on DiverGen evaluating performance on the labeled generators ($\mathcal{S}_l$), unlabeled generators seen during training ($\mathcal{S}_{ul}$) and novel generators unseen during training ($\mathcal{S}_{us}$).

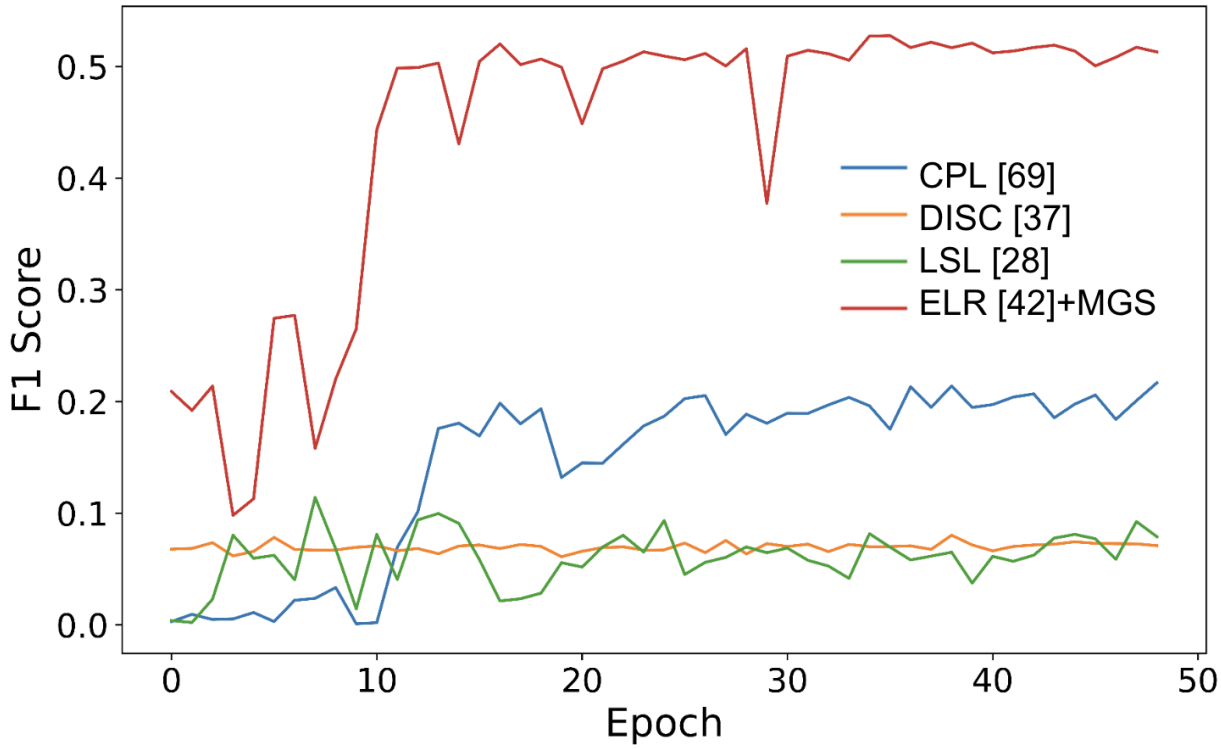


Figure 11: FakesSense The F1-measure of pseudo-labels of the unlabeled data shows that the combination of ELR and MGS achieves the highest F1 score, outperforming alternative methods such as CPL (Sun et al., 2023), DISC (Li et al., 2023), and LSL (Kim et al., 2024).

Tab. 20 provides a detailed break down of DiverGen-Type dataset.

B.3 Dataset Samples

DiverGen-Type improves object diversity by introducing six different category domains (faces, furniture, buildings, art, vehicles, animals). In DiverGen-Attr the number of object categories is 100 which follows CIFAR-100 (Krizhevsky, 2012).

See Fig. 12 for sample images in DiverGen-Attr and Figs. 13 to 21 for DiverGen-Type dataset.

B.4 Benchmark Evaluation Training and Evaluation Data Splits.

In this section, we discuss the detailed training and evaluation data used by FakesSense in the experiments section of the main paper.

Dataset	# Generators	Categories	# Real	# Fake	Generator Category		
					GAN	Diffusion	Transformer
iFakeFaceDB (Neves et al., 2020)	1	Face	0	87,000	✓	✗	✗
TwoStream (Zhou et al., 2017)	2	Face	1,400	2,010	✓	✗	✗
FakeSpotter (Wang et al., 2019)	2	Face	6,000	5,000	✓	✗	✗
DFFD (Dang et al., 2020a)	7	Face	58,703	240,336	✓	✗	✗
ForgeryNet (He et al., 2021)	15	Face	1,438,201	1,457,861	✓	✗	✗
GenImage (Zhu et al., 2023)	8	General	1,331,167	1,350,000	✓	✓	✗
TWIGMA (Chen & Zou, 2023)	unknown	General	0	800,000	✗	✓	✗
SEMI-TRUTHS (Pal et al., 2024)	8	General	27,635	1,470,000	✗	✓	✗
WildFake (Hong & Zhang, 2024)	21	General	1,013,446	2,557,278	✓	✓	✗
WILD (Bongini et al., 2025)	10	Face	0	20,000	✓	✓	✓
DiverGen (ours)	**23**	General	**84,009**	**808,269**	✓	✓	✓

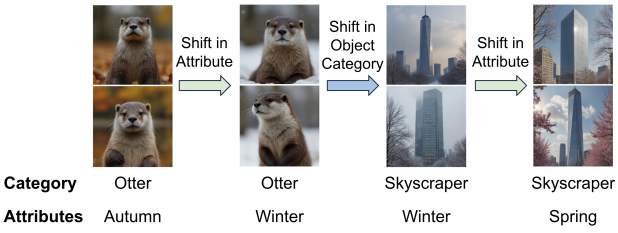
Table 19: Dataset statistics of DiverGen comprising of both DiverGen-Attr and DiverGen-Type subsets compared with those from prior work on generated image detection and attribution. Our dataset provides samples from a more diverse set of generators and categories than prior work.

Source	Types	# Total Samples
CelebDFv2 (Li et al., 2020a)	Celeb-real	53,708
	Youtube-real	30,301
DFFD (Dang et al., 2020a)	StyleGAN (Karras et al., 2019)	15,001
	FaceAPP (FaceApp, 2024)	11,801
	StarGAN (Choi et al., 2018)	15,001
	PGGAN (Karras et al., 2017)	15,011
ForgeryNIR (Wang et al., 2022b)	CycleGAN (Zhu et al., 2020)	10,001
	StyleGAN (Karras et al., 2019)	10,001
	PGGAN (Karras et al., 2017)	10,001
	StyleGAN2 (Karras et al., 2020)	10,001
New to DiverGen-Type	DAGAN (Hong et al., 2022)	13,001
	FOMM (Siarohin et al., 2020)	13,001
	LIA (Wang et al., 2022a)	13,001
	Maxine (NVIDIA, 2024)	13,001
	StyleHEAT (Yin et al., 2022)	13,001
	TPS (Xiang et al., 2022)	13,001
	Guided Diffusion (Kim et al., 2022)	12,975
	Latent Diffusion (Rombach et al., 2021)	19,552
	LSGM (Vahdat et al., 2021)	10,249
	StyleGAN2 (Karras et al., 2020)	13,345
	Taming Transformer (Esser et al., 2021)	19,267
	StyleGAN3 (Karras et al., 2021)	18,057

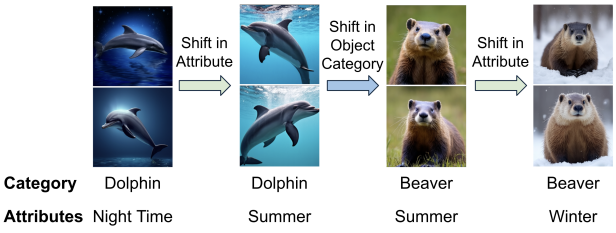
Table 20: Summary of DiverGen-Type distribution. DiverGen-Type provides a more comprehensive dataset by introducing multiple image domains and recent generative models.

Training data splits. To study both attribute and object level shifts, we construct various setting that separates in-distribution (ID) and out-of-distribution (OOD) factors across two aspects: object category and attribute.

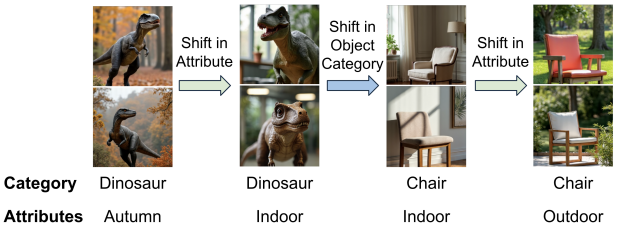
Shift in attributes. When considering only attribute shift, the in-distribution (ID) attributes include default, indoor, daytime, spring, and autumn. The out-of-distribution (OOD) attributes include outdoor, night-time, summer, and winter. These OOD attributes introduce significant changes in appearance, such as lighting, background, seasonal variation, and scene context.



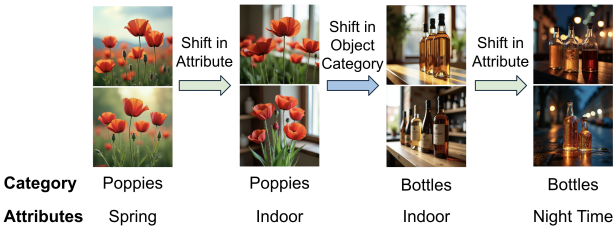
(a) Sample images in DiverGen-Attr generated by Midjourney (Midjourney, 2024).



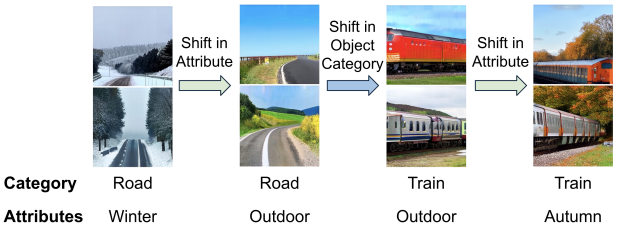
(b) Sample images in DiverGen-Attr generated by CogView (Ding et al., 2021a).



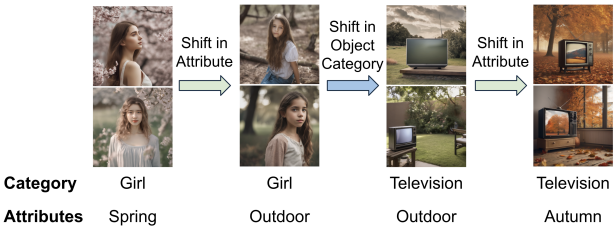
(c) Sample images in DiverGen-Attr generated by Flux (Labs, 2024).



(d) Sample images in DiverGen-Attr generated by HyperSD (Ren et al., 2024).



(e) Sample images in DiverGen-Attr generated by Stable Diffusion 3M (Esser et al., 2024).



(f) Sample images in DiverGen-Attr generated by Stable Diffusion XL (Podell et al., 2023).

Figure 12: Image samples from DiverGen-Attr.

(a) Bed (b) Horse (c) Vehicle

Figure 13: Bed → Horse → Vehicle samples generated by Guided Diffusion (Kim et al., 2022).

Shift in categories. We randomly select 60 object categories as the in-distribution (ID) class set and reserve the remaining 40 categories as out-of-distribution (OOD) classes. This allows us to assess how well the model generalizes to novel object categories that never appear during training.

Shift in both attributes and categories. For the most challenging evaluation setting, we consider samples that simultaneously shift in both object category (OOD class) and generative attribute (OOD attribute).

(a) Car (b) Cat (c) Face

(d) Horse (e) Church (f) Vehicle

Figure 14: Car → Cat → Face → Horse → Church → Vehicle samples generated by StyleGAN2 (Karras et al., 2020).

We use the same ID/OOD attribute and category used in shift in attributes and shift in categories. This combined setting tests the model's robustness under more robust distribution shift.

Train–test split. We use an 80 to 20 train–evaluation split applied across each object–attribute class. ID attributes and ID categories are used only for training, while OOD attributes, OOD categories, are only used for evaluation to ensure a fair analysis of OOD generalization.

(a) Bed (b) Church

(c) Vehicle (d) Face

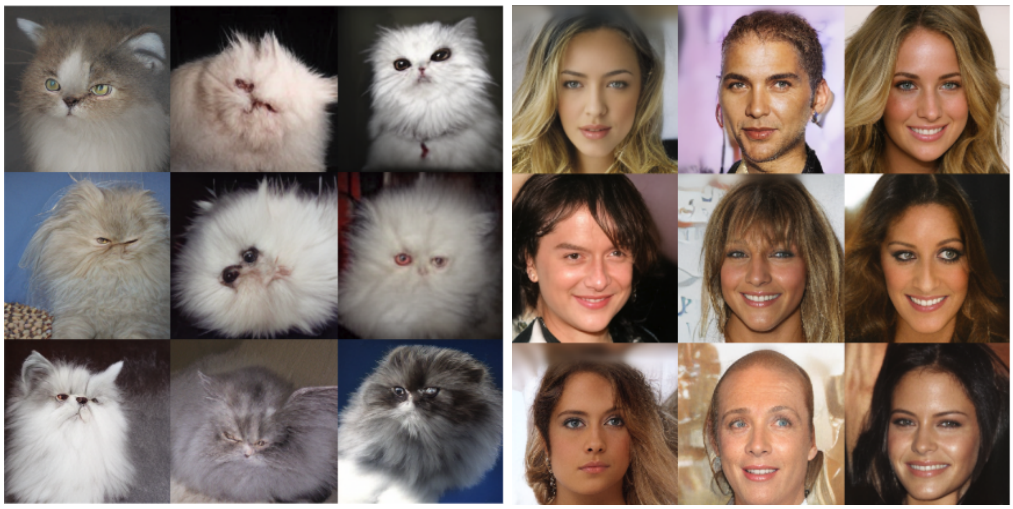
Figure 15: Bed → Church → Vehicle → Face samples generated by Latent Diffusion (Rombach et al., 2021).

Figure 16: Sample images generated by LSGM Vahdat et al. (2021).

(a) Art (b) Cat (c) Dog (d) Fox

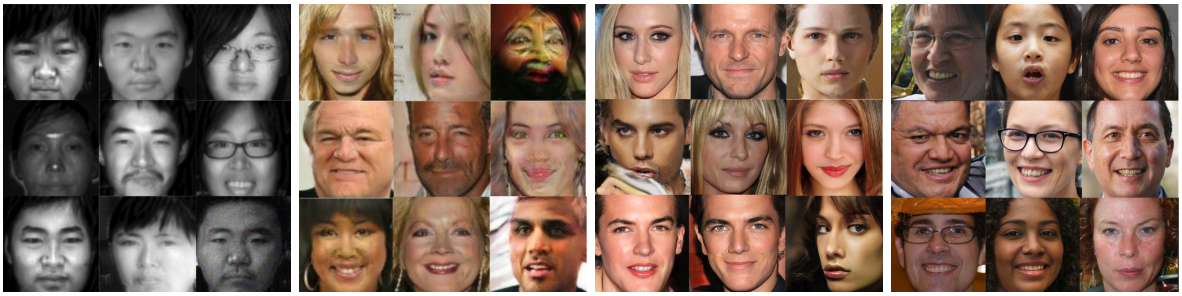
(e) Face (f) Tiger (g) Leopard

Figure 17: Art → Cat → Dog → Fox → Face → Tiger → Leopard samples generated by StyleGAN3 (Karras et al., 2021).



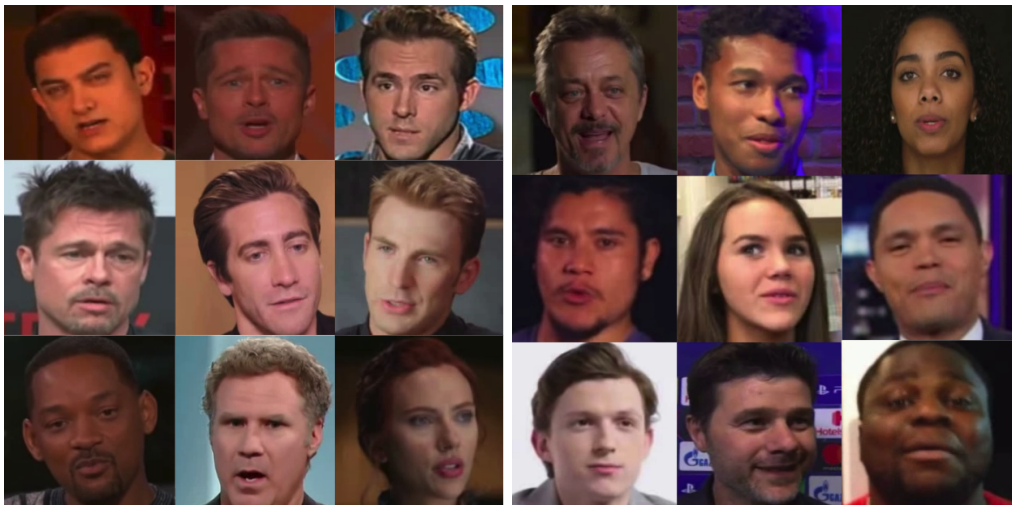
(a) Cat (b) Face

Figure 18: Cat → Face samples generated by Taming Transformer (Esser et al., 2021).



(a) CycleGAN (Zhu et al., 2020) (b) StarGAN (Choi et al., 2018) (c) PGGAN (Karras et al., 2017) (d) StyleGAN (Karras et al., 2019)

Figure 19: Sample images generated by GAN-based models.



(a) Celeb-real (b) Youtube-real

Figure 20: Samples of real faces from Celeb-real and Youtube-real (Li et al., 2020a).

(a) DAGAN (Hong et al., 2022) (b) FOMM (Siarohin et al., 2020) (c) TPS (Xiang et al., 2022)

(d) StyleHeat (Yin et al., 2022) (e) LIA (Wang et al., 2022a) (f) Maxine (NVIDIA, 2024)

Figure 21: Sample images generated by DAGAN, FOMM, TPS, StyleHeat, LIA and Maxine (Hong et al., 2022; Siarohin et al., 2020; Wang et al., 2022a; NVIDIA, 2024; Yin et al., 2022; Xiang et al., 2022).