Learning Counterfactually Decoupled Attention for Open-world Model Attribution

Appendix

Table 1. Generalization to Modern Generative Models in OW-DFA. The extended content is marked as red.

Face Type	Labeled Sets	Unlabeled Sets	Source Dataset	Method	Tag	Labeled #	Unlabeled #
Identity Swap	Deepfakes [3] DeepFaceLab [1]	Deepfakes DeepFaceLab FaceSwap [5] FaceShifter [25] FSGAN [28]	FaceForensics++ [30]	Deepfakes	Known	7500	2500
				FaceSwap	Novel	-	7500
			ForgeryNet [16]	DeepFaceLab	Known	7500	2500
				FaceShifter	Novel	-	7500
				FSGAN	Novel	-	7500
Expression Transfer	Face2Face [33] FOMM [31]	Face2Face FOMM NeuralTextures [2] Talking-Head-Video [37] ATVG-Net [9]	FaceForensics++	Face2Face	Known	7500	2500
				NeuralTextures	Novel	-	7500
			ForgeryNet	FOMM	Known	7500	2500
				ATVG-Net	Novel	-	7500
				Talking-Head-Video	Novel	-	7500
Attribute Manipulation	MaskGAN [23] FaceAPP [4]	MaskGAN FaceAPP StarGAN2 [11] SC-FEGAN [18] StarGAN [10]	ForgeryNet	MaskGAN	Known	7500	2500
				StarGAN2	Novel	-	7500
				SC-FEGAN	Novel	-	7500
			DFFD [12]	FaceAPP	Known	7500	2500
				StarGAN	Novel	-	7500
Entire Face Synthesis	StyleGAN [20] CycleGAN [38] DiT-XL/2 [29]	StyleGAN CycleGAN PGGAN [19] StyleGAN2 [21] SiT-XL/2 [6] DDPM [17] RDDM [27] VQGAN [13]	ForgeryNet	StyleGAN2	Novel	-	7500
			DFFD	StyleGAN	Known	7500	2500
				PGGAN	Novel	-	7500
			ForgeryNIR [34]	CycleGAN	Known	7500	2500
				StyleGAN2	Novel	-	7500
			DF40 [35]	DiT-XL/2	Known	7500	2500
				SiT-XL/2	Novel	-	7500
				DDPM	Novel	-	7500
				RDDM	Novel	-	7500
				VQGAN	Novel	-	7500
Real Face	Youtube-Real [30]	Celeb-Real [26]	FaceForensics++	Youtube-Real	Known	75000	25000
			CelebDFv2 [26]	Celeb-Real	Novel	-	25000

1. More Information on Experimental Setting

To keep pace with modern visual generative models, we have extended the OW-DFA benchmark [32] with diffusion-based and flow-based generative models from the recent DF40 [35] dataset. The specifics on the extended benchmark is shown in Table 1.

Training and Testing Splits: We design two experimental settings to extend the OW-DFA dataset. The training and test sets were divided as 4:1 for both settings. In Setting 1, we added DiT-XL/2 [29], SiT-XL/2 [6], RDDM [27], and VQGAN [13] into both the training and test sets under the "Entire Face Synthesis" face type. Among these, DiT-XL/2 was used as a known attack comprising 10,000 images, of which 7,500 were labeled and 2,500 were unlabeled. The other three models (SiT-XL/2, RDDM, and VQGAN) were treated as novel attacks, each containing 7,500 unlabeled images. In Setting 2, we expanded the original dataset in OW-DFA Protocol 1 with 7,500 unlabeled images generated by DDPM [17], which were not used during training but were included in the testing phase as unseen novel attacks for evaluation. These images were randomly selected from the corresponding datasets to ensure representativeness and unbiased evaluation.

Training Details: For the benchmark experiments on OW-DFA [32], we followed [32] to consistently train the baseline methods [8, 14, 32] for 50 epochs with an initial learning rate of $3e^{-4}$, which is decayed by 80% every 10 epochs. The batch size is 128 for all experiments.

For the benchmark experiments on OSMA, we followed[36] to train the baseline methods [7, 36] for 40 epochs. The learning rate is set as 10^{-4}. The batch sizes are set to 8 and 32 in [36] and [7] respectively. All models are learned using the Adam [22] optimizer.

Table 2. Supplemented technical details of our CDAL.

Description	Output	Operations	Args
Feature Extraction	(B, C, H, W)	Backbone	–
Expert Routing Weight	(B, N)	MLP + Softmax	$C \to N$
Causal Expert Conv	$(B, M, \frac{H}{2}, \frac{W}{2})$	CondConv2d	kernel=3×3, stride=2 padding=1, experts=N
Standard Augment	(B, C, H, W)	Downscale → Upscale	×2 (bilinear)
		GaussianBlur2d	kernel=11×11 $\sigma = 7$
		AddNoise	$\varepsilon \sim \mathcal{N}(0, 0.1^2)$
Classification Head	(B, K)	MLP	$M \times C \to K$

2. More In-depth Technical Details

Here we provide more in-depth technical details of our CDAL to supplement the main pages.

2.1. Introductions of the Baseline Methods

Open-world Deepfake Attribution: In OW-DFA, we experimented on three baseline methods to validate the effectiveness of our proposed CDAL. Brief introductions of these baseline methods are listed as follows:

- **ORCA** [8] proposes an open-world self-supervised learning framework by constructing pairwise affinity through cosine similarity optimization, enforcing proximity between high-confidence matches. The $\mathcal{L}_{\text{original}}$ in this case contains $\mathcal{L}_S$ and $\mathcal{L}_P$ [8], which represent the supervised objective with an uncertainty adaptive margin, and a pairwise objective respectively.
- **NACH** [14] introduces a novel approach to filter out erroneous samples and synchronizes the learning pace between seen and unseen classes. The $\mathcal{L}_{\text{original}}$ in this case is an improved version of $\mathcal{L}_P$ from ORCA [8], which partitions the feature space and carefully processes unlabeled data to ensure that the model learns robustly from both labeled and unlabeled data.
- **CPL** [32] proposes global-local voting to align features of diversely manipulated forged faces and soft pseudo-labeling weighted by prediction confidence to suppress noise from similar manipulation patterns in unlabeled data. The $\mathcal{L}_{\text{original}}$ in this case are $\mathcal{L}_{GLV}$ and L_{CSP}. L_{GLV} combines global and local similarity to accurately match samples of the same attack type, while $\mathcal{L}_{CSP}$ assigns soft pseudo-labels based on the confidence of predictions to reduce the impact of pseudo-noise.

Open-world GAN Attribution: To evaluate the performance of CDAL in OSMA, we conducted experiments using the following baseline methods for comparison:

- **RepMix** [7] designs a representation mixing layer that synthesizes new data by interpolating data points in the feature space. It enhances the generalization to unseen semantics and transformations, which improves the robustness of synthesized image attribution. $\mathcal{L}_{\text{original}}$ in this case consists of L_{det} for detecting real from fake images, and L_{attr} for identifying the GAN architecture that generated the fake images.
- **POSE** [36] introduces a method to progressively simulate the open space of unknown models using lightweight augmentation models, which aims to expand the potential open space around the boundary of known models. The $\mathcal{L}_{\text{original}}$ in this case is mainly two parts. An augmented loss merges pixel reconstruction and embedding diversity for semantic fidelity and distinct samples. A diversity loss drives inter-model diversity to prevent overlaps and enhance the learned uniqueness.

2.2. More Technical Details on CDAL

As displayed in Table 2, we present more technical details to supplement the Approach Section in the main pages.

Feature Extraction: In the OW-DFA benchmark, we employ ResNet-50 [15] as the backbone network for the extraction of $\mathbf{X}$, which is also shared by baseline methods we experimented on [8, 14, 32]. In the OSMA benchmark, We follow [7, 36] to use their original feature extractors. Specifically, [36] uses a discrete cosine transform (DCT) transformation layer combined with a simple convolutional network, while [7] also uses ResNet-50 [15].

Causal Expert Convolution: In CE Convolution, to dynamically generate combination weights for each expert, we map the feature channels to the number of experts N using an MLP, followed by Softmax normalization. This enables adaptive feature allocation for different regions.

Standard Augmentation: In Causal Attention Augmentation, we conduct a series of standard data augmentation

Table 3. Detailed quantitative results (%) of GAN discovery (Protocol 2) on OSMA [36], which are averaged among five splits.

Method	Close-set	Unseen Seed			Unseen Architecture			Unseen Dataset			Unseen All		
	ACC	Purity	NMI	ARI	Purity	NMI	ARI	Purity	NMI	ARI	Purity	NMI	ARI
RepMix [7]	93.69	23.71	33.06	15.21	50.94	64.73	40.86	28.52	34.75	13.93	31.53	51.60	18.71
POSE [36]	94.81	29.54	32.37	13.83	62.77	70.16	49.67	41.52	48.89	25.50	41.04	60.59	26.39
RepMix + Ours	94.01	24.55	**33.75**	**16.09**	56.45	67.17	44.40	35.79	39.97	20.79	37.96	52.08	20.66
Improvement	+0.32	+0.84	+0.69	+0.88	+5.51	+2.44	+3.54	+7.27	+5.22	+6.86	+6.43	+0.48	+1.95
POSE + Ours	**95.25**	**30.32**	33.14	14.57	**67.95**	**73.39**	**55.36**	**50.31**	**55.13**	**33.53**	**48.93**	**61.89**	**29.65**
Improvement	+0.44	+0.78	+0.77	+0.74	+5.18	+3.23	+5.69	+8.79	+6.24	+8.03	+7.89	+1.30	+3.26

operations. Specifically, we use two bilinear interpolations to simulate resolution changes, an 11×11 Gaussian blur kernel with a standard deviation of 7 to blur high-frequency details, and Gaussian noises with a standard deviation of 0.1.

Classification Head: When computing the causal effect, we use a shared classifier δ to map the fused feature dimensions to the target class numbers K, and output the final classification probability logits.

Attention Weight: The normalized energy distribution of factual attention channels is a Categorical Distribution. Each element in the Categorical Distribution is computed by summing the feature values across the spatial dimensions of the feature map $\mathbf{F}$, followed by a square root operation and normalization:

$$\mathbf{w} = \text{Norm}\left(\sqrt{\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{F}h,w}\right). \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^M$ represents the normalized energy distribution across M channels. Mathematically, this categorical distribution is formulated as:

$$\begin{aligned} p(\mathbf{w}) &= \text{Cat}(w_1, w_2, \ldots, w_M) \\ &= \begin{cases} 1 & \text{if } \sum_{i=1}^{M} w_i = 1 \text{ and } w_i \geq 0 \text{ for all } i \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{2}$$

with $\sum_{i=1}^{M} w_i = 1$ and $w_i \geq 0$ for all $i \in 1, 2, \ldots, M$.

Hyper-parameters: The hyper-parameters η_1, η_2, and η_3 for the total loss functions are set as 0.5, 0.5 and 0.2 respectively.

Alternative Counterfactual Attentions in Table 5(d): The alternative counterfactual attentions we compare with in Table 5(d) are derived as follows:

- **Random Attentions:** We create random attention by sampling from a uniform distribution over [0, 2] to make an even spread of attention values.
- **Uniform Attentions:** We create uniform attention by setting all attention weights to a fixed value of 0.5.
- **Reversed Attentions:** We create reversed attention by computing the element-wise subtraction between an all-ones tensor and the factual attentions extracted by our CE-Conv.
- **Shuffled Attentions:** We create shuffled attention by randomly reordering the factual attentions values extracted by CE-Conv. Specifically, we flatten the attentions, apply a random permutation, and then reshape them into the original shape.

Table 4. Results of experiment on handcrafted input features [24] on **Protocol-1** of OW-DFA [32].

Method	Known	Novel			All		
	ACC	ACC	NMI	ARI	ACC	NMI	ARI
CPL [32]	98.68	75.21	73.19	65.71	86.25	85.58	82.35
CPL [32] + MHFs [24]	98.90	78.83	76.94	69.81	88.25	87.88	84.09
CPL [32] + Ours	98.90	86.02	82.19	76.98	92.06	90.60	87.66
CPL [32] + MHFs [24] + Ours	98.90	86.58	85.12	79.41	92.37	91.85	88.74

3. More Experimental Results and Analysis

Here we provided additional quantitative and visualization results to supplement those in the main pages.

3.1. More Quantitative Results

Results of GAN discovery: Table 3 demonstrates more in-depth performances of different GAN discovery methods. Our proposed CDAL significantly enhances both RepMix [7] and POSE [36]. When combined with RepMix, our method achieves the highest improvement in Purity by 7.27% on the unseen architectures. Similarly, with POSE, CDAL enhances the Purity by 8.79% which further validates its generalization capability. These results highlight the effectiveness of CDAL in improving adaptability to unseen data on various aspects.

Results of Applying CDAL to Handcrafted Input Features: While our proposed CDAL focus on the improvement on previous learning strategies including handcrafted design of region partition and feature space, a recent work [24] highlights an alternative approach that leverages handcrafted features for open-set model attribution. We accordingly conducted experiment on implementing the Multi-Directional High-Pass Filters (MHFs) in [24] to concatenate the input features of baseline model with handcrafted features, denoted as CPL+MHFs. From the results

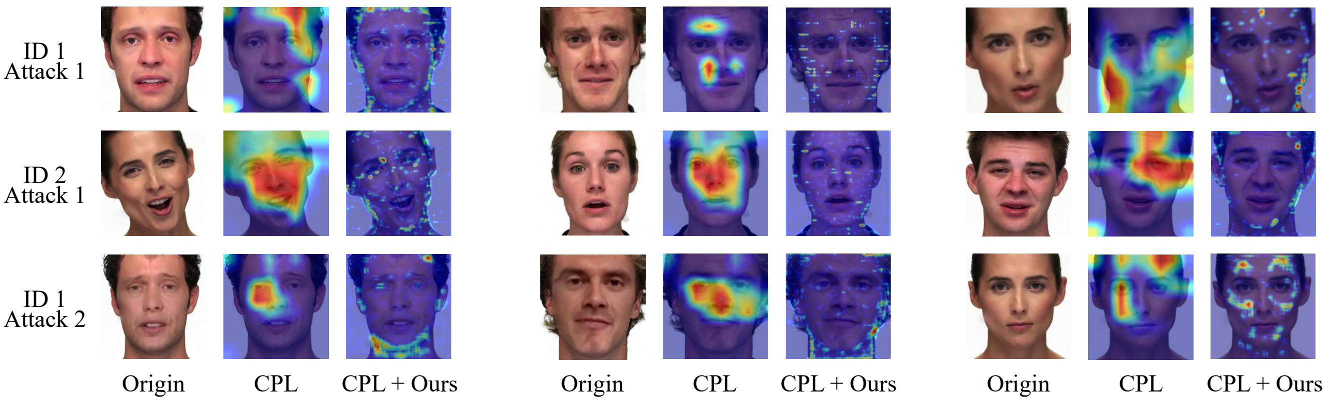


Figure 1. Additional motivative examples.

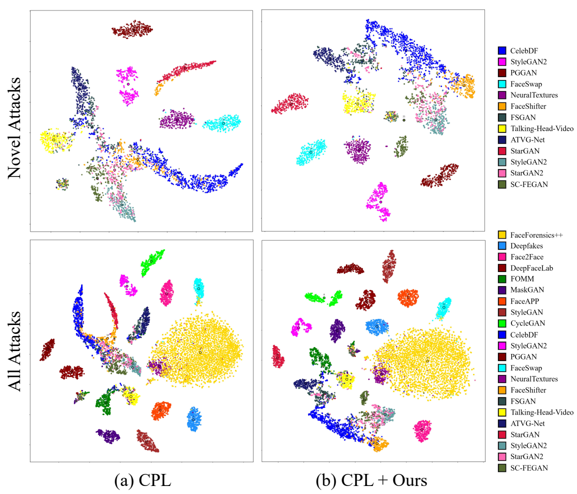


Figure 2. t-sne visualization of Protocol 2 in OW-DFA.

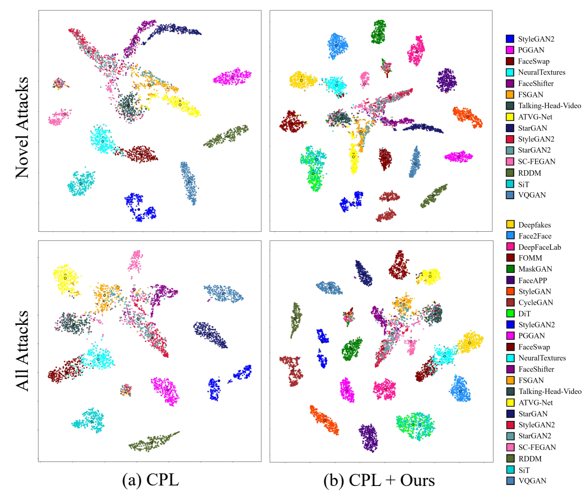


Figure 3. t-sne visualization of Setting 1 on our extended OW-DFA benchmark.

in Table 4, we can see that while CPL+MHFs is able to enhance the performances upon CPL, the further incorporation of our method (CPL+MHFs+Ours) achieves significantly better performance across all metrics compared with CPL+MHFs. Also, for the baseline method [32], solely introducing CDAL also brings notably larger performance gains compared with solely introducing MHFs. These results demonstrate the versatility of our proposed method, whose effectiveness can be imposed upon various input features, as well as various learning strategies.

Results of Robustness Against Adaptive Attack: We employed FGSM (budget $5e^{-4}$) as an adaptive attack to input samples and observe obviously smaller drops with CDAL in Table 5, which verifies the robustness of our method.

Table 5. FGSM attacks on **Protocol 1** of OW-DFA

FGSM	CPL			CDAL		
	ACC	NMI	ARI	ACC	NMI	ARI
w./o. attack	75.21	73.19	65.71	86.02	82.19	76.98
w./ attack	68.73	65.58	57.08	82.29	79.74	73.92
Δ	-6.48	-7.61	-8.63	**-3.73**	**-2.45**	**-3.06**

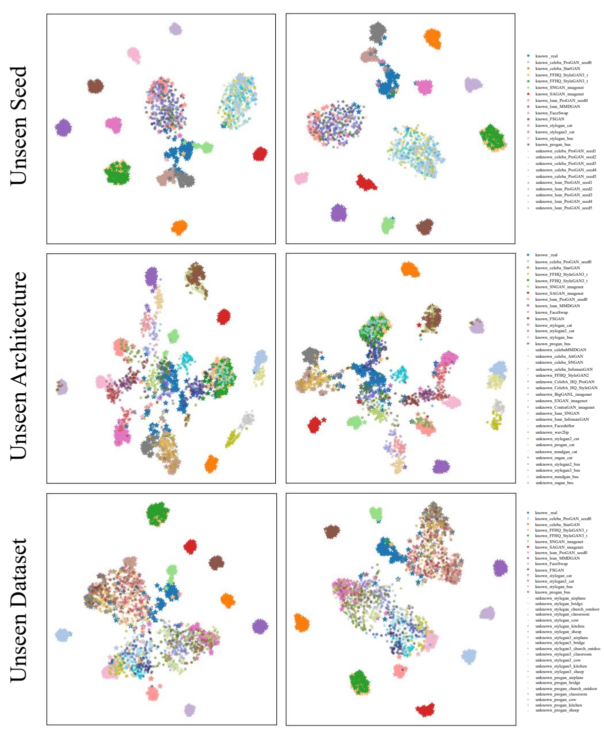
3.2. More Visulization Results

Additional Motivative Examples: Figure 1 provides additional visualizations of motivative examples of our CDAL as in Figure 1 of the main pages. The "Original" columns show that forgery images originating from the same identity exhibit high semantic similarity from source identities in facial features, which causes them to naturally group together in the feature space. The "CPL" columns show that current baseline method [32] still struggles with these source biases faced with unseen novel attacks, rather than effectively capturing model-specific traces. In contrast, the "CPL+Ours" columns demonstrate that our method aims to spot those subtle forgery traces which are crucial for model attribution.

More t-SNE Visualization of OW-DFA: As illustrated in Figure 2, we further present the t-SNE results of Protocol 2 in OW-DFA. Our method maintains better discriminative capability even when confronted with real faces from Celeb-DF [26]. our method also successfully clusters new attacks like FSGAN [28] and FaceShifter [25] from ForgeryNet [16] which are hard to distinguish due to similar source bias features.

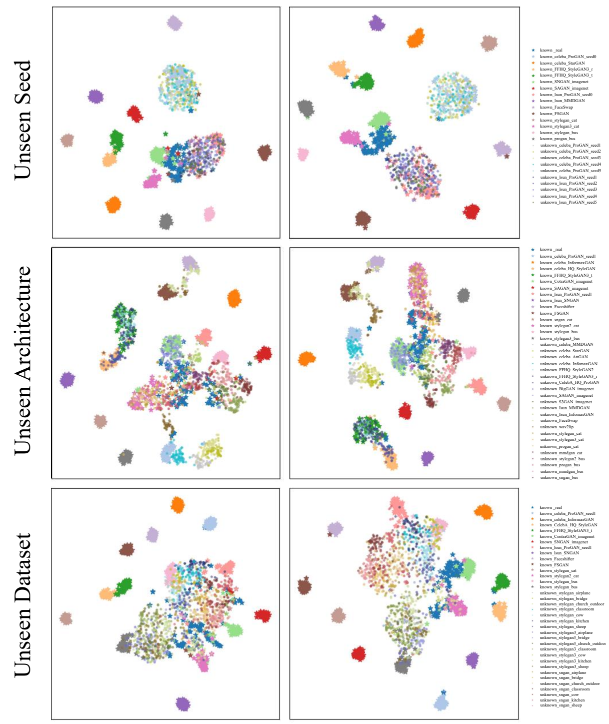
We also provide t-SNE visualization for Setting 1 on our extended OW-DFA benchmark in Figure 3, where our method demonstrates notable superiority over the baseline method [32].

t-SNE Visualization of OSMA: To visually compare the performance differences between CDAL and the baseline method POSE, Figure 4 presents the t-SNE visualization results of split 1 in OSMA. In the Unseen architecture, POSE exhibits limited ability to distinguish feature from different generative model architectures, resulting in significant overlap between clusters. In contrast, CDAL forms clusters with high separation, by effectively extracting discriminative features from different architectures, which demonstrates stronger adaptability to unseen architectural variations. For the Unseen seed and Unseen dataset, the feature representations generated by POSE are either scattered or significantly overlapping, highlighting its limitations in handling random variations and data diversity. In comparison, CDAL consistently forms clusters with clear boundaries, which showcases its robustness and generalization capability. The visualization results for splits 2, 3, 4, and 5 are shown in Figures 5, 6, 7, and 8.



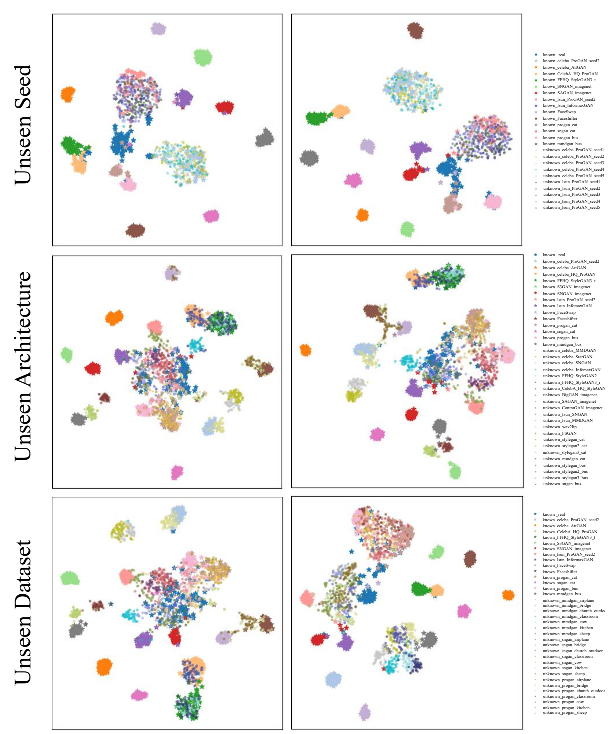
(a) POSE (b) POSE + Ours

Figure 4. t-sne visualization of split 1 in OSMA.



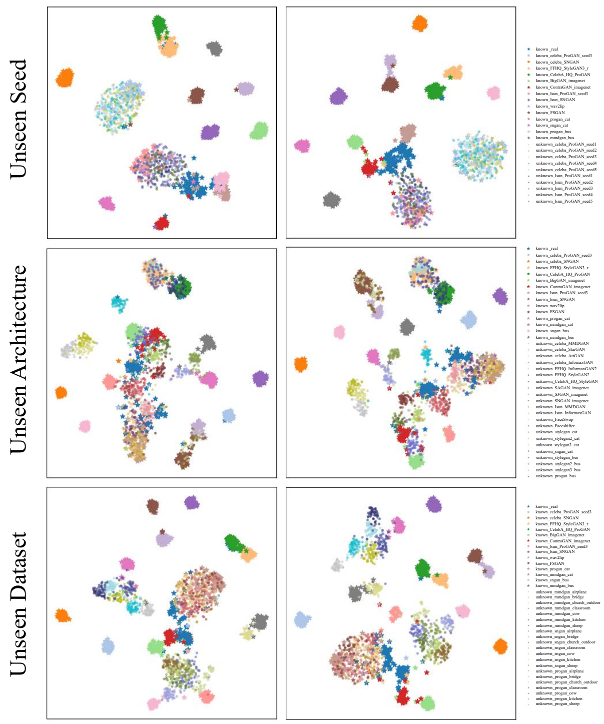
(a) POSE (b) POSE + Ours

Figure 5. t-sne visualization of split 2 in OSMA.



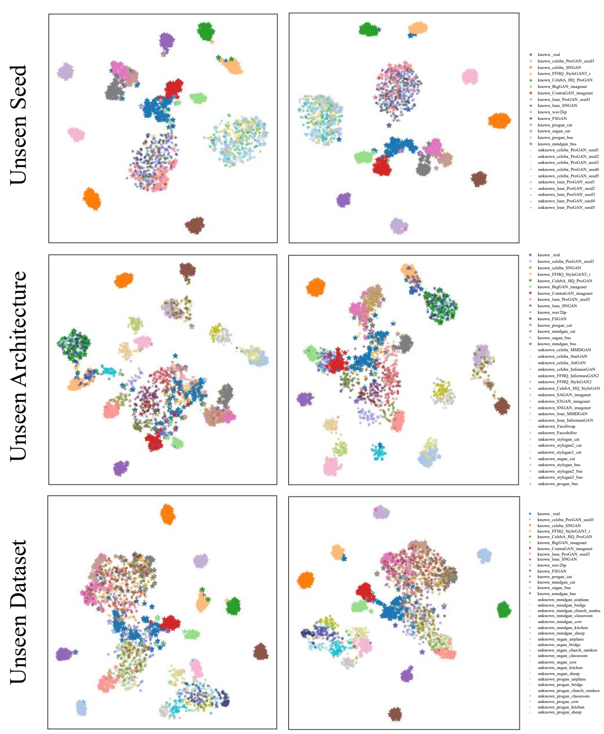
(a) POSE (b) POSE + Ours

Figure 6. t-sne visualization of split 3 in OSMA.



(a) POSE (b) POSE + Ours

Figure 7. t-sne visualization of split 4 in OSMA.



(a) POSE (b) POSE + Ours

Figure 8. t-sne visualization of split 5 in OSMA.

References

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

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

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

[4] Faceapp. https://faceapp.com/app, . Accessed: 2023-2-28. 1

[5] Faceswap. https://github.com/MarekKowalski/FaceSwap/, . Accessed: 2023-2-28. 1

[6] Sara Atito, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer. *arXiv preprint arXiv:2104.03602*, 2021. 1

[7] Tu Bui, Ning Yu, and John Collomosse. Repmix: Representation mixing for robust attribution of synthesized images. In *ECCV*, pages 146–163, 2022. 1, 2, 3

[8] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *ICLR*, 2022. 1, 2

[9] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, pages 7832–7841, 2019. 1

[10] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 1

[11] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 1

[12] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *CVPR*, pages 5781–5790, 2020. 1

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 1

[14] Lan-Zhe Guo, Yi-Ge Zhang, Zhi-Fan Wu, Jie-Jing Shao, and Yu-Feng Li. Robust semi-supervised learning when not all classes have labels. In *NeurIPS*, 2022. 1, 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[16] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *CVPR*, pages 4360–4369, 2021. 1, 5

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 1

[18] Youngjoo Jo and Jongyoul Park. Sc-fegan: Face editing generative adversarial network with user's sketch and color. In *ICCV*, pages 1745–1753, 2019. 1

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 1

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1

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

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[23] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 1

[24] Jialiang Li, Haoyue Wang, Sheng Li, Zhenxing Qian, Xinpeng Zhang, and Athanasios V Vasilakos. Are handcrafted filters helpful for attributing ai-generated images? In *ACM MM*, pages 10698–10706, 2024. 3

[25] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 1, 5

[26] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *CVPR*, pages 3207–3216, 2020. 1, 4

[27] Jiawei Liu, Qiang Wang, Huijie Fan, Yinong Wang, Yandong Tang, and Liangqiong Qu. Residual denoising diffusion models. In *CVPR*, pages 2773–2783, 2024. 1

[28] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *ICCV*, pages 7184–7193, 2019. 1, 5

[29] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 1

[30] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11, 2019. 1

[31] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, pages 7137–7147, 2019. 1

[32] Zhimin Sun, Shen Chen, Taiping Yao, Bangjie Yin, Ran Yi, Shouhong Ding, and Lizhuang Ma. Contrastive pseudo learning for open-world deepfake attribution. In *ICCV*, pages 20882–20892, 2023. 1, 2, 3, 4, 5

[33] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, pages 2387–2395, 2016. 1

[34] Yukai Wang, Chunlei Peng, Decheng Liu, Nannan Wang, and Xinbo Gao. Forgerynir: deep face forgery and detection in near-infrared scenario. *TIFS*, 17:500–515, 2022. 1

[35] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. In *NeurIPS*, 2024. 1

[36] Tianyun Yang, Danding Wang, Fan Tang, Xinying Zhao, Juan Cao, and Sheng Tang. Progressive open space expansion for open-set model attribution. In *CVPR*, pages 15856–15865, 2023. 1, 2, 3

[37] Sibo Zhang, Jiahong Yuan, Miao Liao, and Liangjun Zhang. Text2video: Text-driven talking-head video synthesis with phonetic dictionary. *arXiv preprint arXiv:2104.14631*, 2021. 1

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *ICCV*, 2017. 1