Supplementary Material for Progressive Open Space Expansion for Open-Set Model Attribution

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

The supplementary material is organized as follows:

- Section 1 provides the open-set discrimination result on images generated by a stable-diffusion model.
- Section 2 gives an analysis on two situations for open-set model attribution: unseen seed model and fine-tuned model.
- Section 3 gives robustness analysis against common image perturbations.
- Section 4 shows the full five splits of the OSMA benchmark.
- Section 5 visualizes randomly selected samples from the OSMA benchmark.

1. Evaluation on Diffusion Model

We evaluate samples generated by the newly arisen stable-diffusion model [18]. We use CoCo [13] captions to generate 1k stable-diffusion samples and test POSE's open-set discrimination performance on these samples. Randomly selected samples are shown in Figure 3. As shown in Figure 1, the AUC point between closed-set and unseen stable-diffusion samples is 92.40, indicating that POSE is able to capture the difference in traces of known models and stable-diffusion model, and recognize stable diffusion samples as from a new model.

2. Unseen Seed and Finetuned Model

For open-set model attribution, there exist two situations near the known space boundary, *i.e.*, models trained with only seed different, and models fine-tuned from the known models. To analyze how POSE reacts in the two situations, we train a 2-way POSE classifier on {celeba, ProGAN_celeba_seed0}, and test the classifier on seven unseen models including ProGAN_celeba_seed1, and six models finetuned from ProGAN_celeba_seed0 on the celeba

*Corresponding author

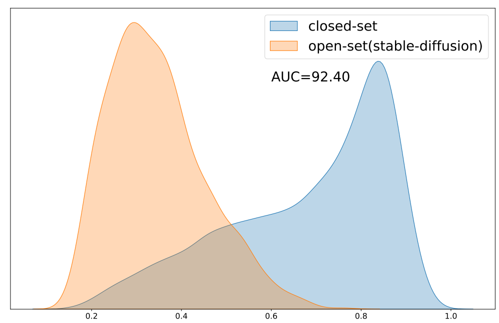


Figure 1. Confidence histograms on unseen stable diffusion data and closed-set data.

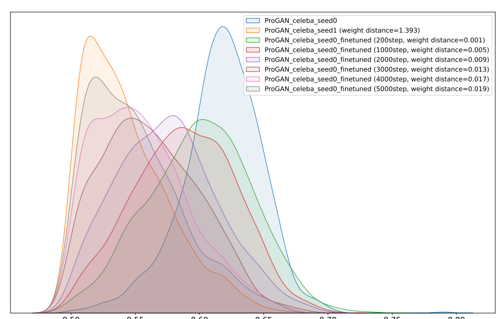


Figure 2. Confidence histograms on data from ProGAN_celeba_seed0, unseen seed model (ProGAN_celeba_seed1), and finetuned models (ProGAN_celeba_seed0_finetuned).

dataset. We plot in Figure 2 the confidence histograms for these models and calculate the weight distance between seen ProGAN_celeba_seed0 model and seven unseen models. Specifically, the weight distance between two models is calculated as follows:

$$D(W_1, W_2) = \frac{1}{N}\sum_{i=1}^{N}\frac{\|W_{2,i} - W_{1,i}\|}{\|W_{1,i}\|}, \quad (1)$$

where W_1 and W_2 are weights of two models with the same architecture. N is the number of layers that are equipped with learnable weights.

As shown in Figure 2, the POSE classifier is able to separate samples generated by an unseen seed model (ProGAN_celeba_seed1) from seen ProGAN_celeba_seed0 model. With the finetune step increases (from 200 to 5000), the weight distance between the finetuned model and the original ProGAN_celeba_seed0 model increases followingly (from 0.001 to 0.019). When the weight distance reaches 0.019, POSE achieves a clear separation between the finetuned model and the original model. These results indicate that POSE is sensitive to trace changes brought by model weight changes, and is suitable for scenarios requiring strict attribution.

3. Robustness Analysis

Generated images may undergo post-processings in real-world scenarios. We evaluate the robustness of POSE against five common image perturbations, which are Blurring with Gaussian, JPEG compression, Lighting, additive Gaussian noise, crop, and resize. We evaluate the original version and immunized version of POSE. The original version indicates the perturbation is not included in model training, and the immunized version indicates that the perturbation is included as a kind of data augmentation in model training. We plot the OSCR results w.r.t the strength of each perturbation in Figure 4. As seen, without immunization, image perturbations would largely influence the model attribution performance. Nevertheless, with image perturbations included as data augmentation operations in model training, the performance drop is largely relieved. Specifically, the immunized version is rather robust to Lighting, Noise, and Crop perturbations. For JPEG compression quality $\sim$ [80, 100], and blur kernel size $\sim$ [0, 3], the performance drop could maintain within a 10% range.

4. Full Dataset Splits

We provide the full five splits of the OSMA benchmark in Table 1, Table 2, Table 3, Table 4, and Table 5, in which Table 1 is the same as Table 1 in the main text.

5. Visualization of Dataset Samples

We provide randomly selected samples in the benchmark for models trained on CelebA, Face-HQ, ImageNet, Youtube, LSUN-Bedroom, LSUN-Cat, and LSUN-Bus, which are shown in Figure 5.

References

[1] Faceswap. https://faceswap.dev. 3, 4

[2] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018. 3, 4

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 3, 4

[4] 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. 3, 4

[5] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *TIP*, 28(11):5464–5478, 2019. 3, 4

[6] Minguk Kang and Jaesik Park. Contragan: Contrastive learning for conditional image generation. In *NeurIPS*, 2020. 3, 4

[7] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 3, 4

[8] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 3, 4

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3, 4

[10] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 3, 4

[11] Kwot Sin Lee, Ngoc-Trung Tran, and Ngai-Man Cheung. Infomax-gan: Improved adversarial image generation via information maximization and contrastive learning. In *WACV*, 2021. 3, 4

[12] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. In *CVPR*, 2020. 3, 4

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[14] Mario Lučić, Michael Tschannen, Marvin Ritter, Xiaohua Zhai, Olivier Bachem, and Sylvain Gelly. High-fidelity image generation with fewer labels. In *ICML*, 2019. 3, 4

[15] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 3, 4

[16] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*, 2019. 3, 4

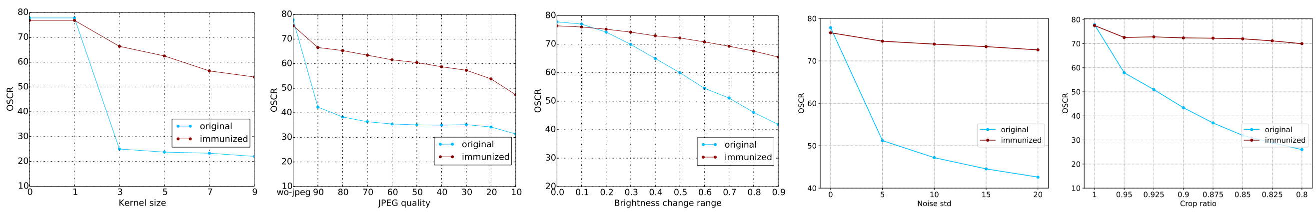
[17] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM MM*, 2020. 3, 4

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[19] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 3, 4

(a) A blue train in the middle of a forest. (b) A chain link fence contains a building and rubbish. (c) A clock sits high on a wall while below it on the wall features mosaic tiles. (d) A group of people in a field smile holding frisbees. (e) A soccer player trying to score a goal.

Figure 3. Randomly selected samples generated by the stable diffusion model.



(a) Blurring (b) JPEG compression (c) Lighting (d) Noise (e) Crop and resize

Figure 4. Robustness analysis. The results are evaluated on Split 1 of the benchmark.

Table 1. Split 1 of the OSMA benchmark.

		CelebA	Face-HQ	ImageNet	Youtube	LSUN-Bedroom	LSUN-Cat	LSUN-Bus
Seen Real		CelebA	Face-HQ	ImageNet	Youtube	LSUN-Bedroom	LSUN-Cat	LSUN-Bus
Seen Fake		StarGAN [4], ProGAN_seed0 [7]	StyleGAN3-r [8], StyleGAN3-t	SAGAN [19], SNGAN	FSGAN [16], FaceSwap [1]	ProGAN_seed0, MMDGAN	StyleGAN, StyleGAN3	ProGAN, StyleGAN
Unseen Fake	**Unseen Seed**	ProGAN (seed1,2,3,4,5)	-	-	-	ProGAN (seed1,2,3,4,5)	-	-
	Unseen Architecture	SNGAN [15], AttGAN [5], MMDGAN [2], InfoMaxGAN [11]	StyleGAN2 [10], ProGAN, StyleGAN [9]	S3GAN [14], BigGAN [3], ContraGAN [6]	Wav2Lip [17], FaceShifter [12]	SNGAN, InfoMaxGAN	SNGAN, ProGAN, MMDGAN, StyleGAN2	SNGAN, MMDGAN, StyleGAN2, StyleGAN3
	Unseen Dataset	ProGAN, StyleGAN, StyleGAN3 (Cow, Sheep, Classroom, Bridge, Kitchen, Airplane, Church)						
Unseen Real		Coco, Summer						

Table 2. Split 2 of the OSMA benchmark

		CelebA	Face-HQ	ImageNet	Youtube	LSUN-Bedroom	LSUN-Cat	LSUN-Bus
Seen Real		CelebA	Face-HQ	ImageNet	Youtube	LSUN-Bedroom	LSUN-Cat	LSUN-Bus
Seen Fake		InfoMaxGAN [11], ProGAN_seed1 [7]	StyleGAN [9], StyleGAN3-t [8]	ContraGAN [6], SNGAN	FSGAN [16], FaceShifter [12]	SNGAN, ProGAN_seed1	SNGAN, StyleGAN2	StyleGAN, StyleGAN3
Unseen Fake	**Unseen Seed**	ProGAN (seed0,2,3,4,5)	-	-	-	ProGAN (seed0,2,3,4,5)	-	-
	Unseen Architecture	SNGAN [15], AttGAN [5], MMDGAN [2], StarGAN [4]	ProGAN, StyleGAN2 [10], StyleGAN3-r	S3GAN [14], BigGAN [3], SAGAN [19]	Wav2Lip [17], FaceSwap [1]	MMDGAN, InfoMaxGAN	ProGAN, MMDGAN, StyleGAN, StyleGAN3	ProGAN, SNGAN, MMDGAN, StyleGAN2
	Unseen Dataset	SNGAN, StyleGAN, StyleGAN3 (Cow, Sheep, Classroom, Bridge, Kitchen, Airplane, Church)						
Unseen Real		Coco, Summer						

Table 3. Split 3 of the OSMA benchmark

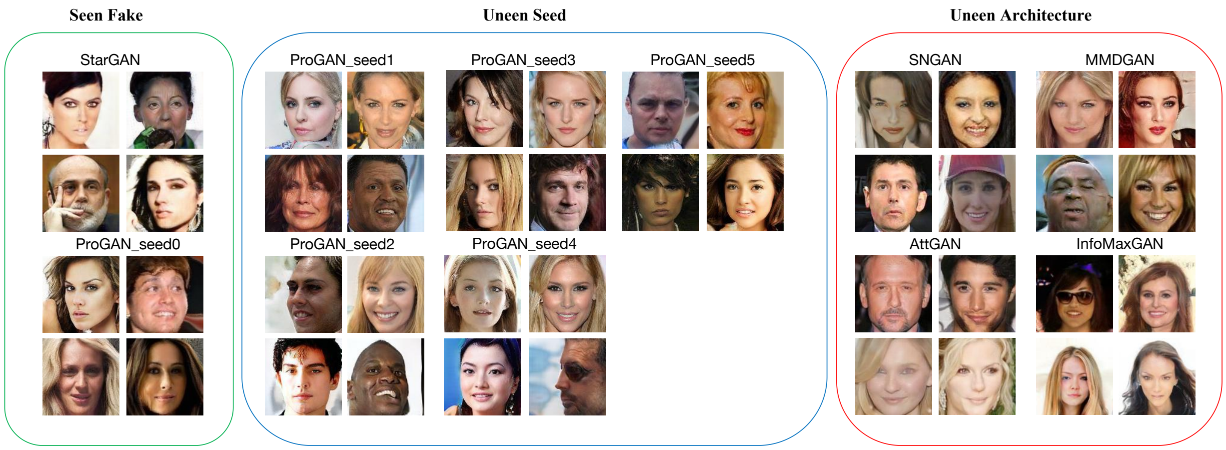
Seen Real		CelebA	Face-HQ	ImageNet	Youtube	LSUN-Bedroom	LSUN-Cat	LSUN-Bus
Seen Fake		AttGAN [5], ProGAN_seed2 [7]	ProGAN, StyleGAN3-t [8]	S3GAN [14], SNGAN	FaceSwap [1], FaceShifter [12]	InfoMaxGAN, ProGAN_seed2	SNGAN, ProGAN	ProGAN, MMDGAN
Unseen Fake	**Unseen Seed**	ProGAN (seed0,1,3,4,5)	-	-	-	ProGAN (seed0,1,3,4,5)	-	-
	Unseen Architecture	SNGAN [15], InfoMaxGAN [11], MMDGAN [2], StarGAN [4]	StyleGAN [9], StyleGAN2 [10], StyleGAN3-r	ContraGAN [6], BigGAN [3], SAGAN [19]	Wav2Lip [17], FSGAN [16]	MMDGAN, SNGAN	StyleGAN2, MMDGAN, StyleGAN, StyleGAN3	SNGAN, StyleGAN, StyleGAN3, StyleGAN2
	Unseen Dataset	SNGAN, MMDGAN, ProGAN (Cow, Sheep, Classroom, Bridge, Kitchen, Airplane, Church)						
Unseen Real		Coco, Summer						

Table 4. Split 4 of the OSMA benchmark

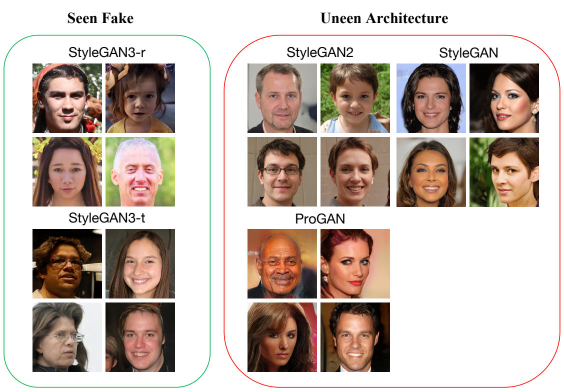
Seen Real		CelebA	Face-HQ	ImageNet	Youtube	LSUN-Bedroom	LSUN-Cat	LSUN-Bus
Seen Fake		SNGAN [15], ProGAN_seed3 [7]	ProGAN, StyleGAN3-r [8]	ContraGAN [6], BigGAN [3]	Wav2Lip [17], FSGAN [16]	SNGAN, ProGAN_seed3	ProGAN, MMDGAN	SNGAN, MMDGAN
Unseen Fake	**Unseen Seed**	ProGAN (seed0,1,2,4,5)	-	-	-	ProGAN (seed0,1,2,4,5)	-	-
	Unseen Architecture	AttGAN [5], InfoMaxGAN [11], MMDGAN [2], StarGAN [4]	StyleGAN [9], StyleGAN2 [10], StyleGAN3-t	S3GAN [14], SNGAN, SAGAN [19]	FaceSwap [1], FaceShifter [12]	MMDGAN, InfoMaxGAN	SNGAN, StyleGAN, StyleGAN2, StyleGAN3	ProGAN, StyleGAN, StyleGAN2, StyleGAN3
	Unseen Dataset	SNGAN, ProGAN, MMDGAN (Cow, Sheep, Classroom, Bridge, Kitchen, Airplane, Church)						
Unseen Real		Coco, Summer						

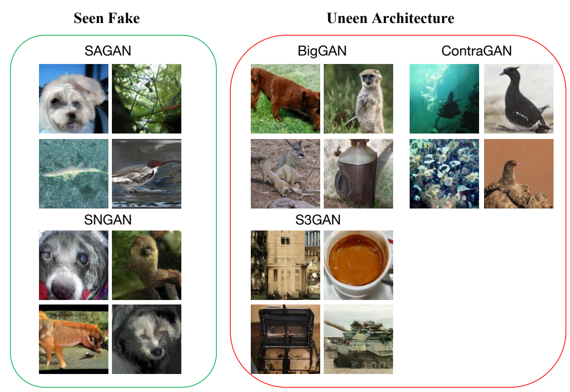
Table 5. Split 5 of the OSMA benchmark

Seen Real		CelebA	Face-HQ	ImageNet	Youtube	LSUN-Bedroom	LSUN-Cat	LSUN-Bus
Seen Fake		AttGAN [5], ProGAN_seed4 [7]	StyleGAN [9], StyleGAN3-t [8]	ContraGAN [6], BigGAN [3]	FaceSwap [1], Wav2Lip [17]	InfoMaxGAN, ProGAN_seed4	StyleGAN, ProGAN	ProGAN, MMDGAN
Unseen Fake	**Unseen Seed**	ProGAN (seed0,1,2,3,5)	-	-	-	ProGAN (seed0,1,2,3,5)	-	-
	Unseen Architecture	SNGAN [15], InfoMaxGAN [11], MMDGAN [2], StarGAN [4]	ProGAN, StyleGAN2 [10], StyleGAN3-r	S3GAN [14], SNGAN, SAGAN [19]	FaceShifter [12], FSGAN [16]	MMDGAN, SNGAN	StyleGAN2, MMDGAN, SNGAN, StyleGAN3	SNGAN, StyleGAN, StyleGAN3, StyleGAN2
	Unseen Dataset	ProGAN, MMDGAN, StyleGAN (Cow, Sheep, Classroom, Bridge, Kitchen, Airplane, Church)						
Unseen Real		Coco, Summer						



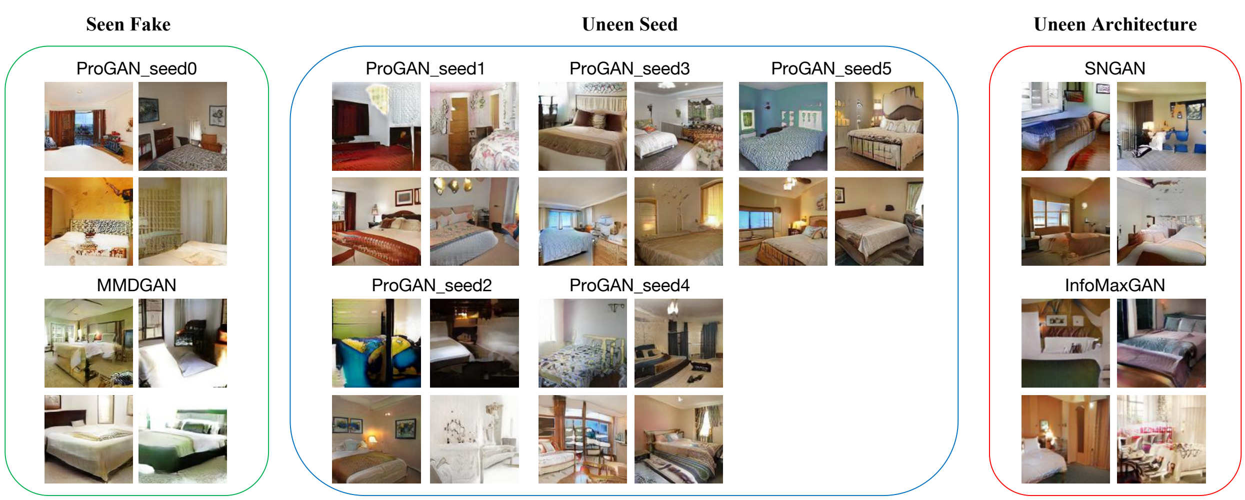
(a) CelebA





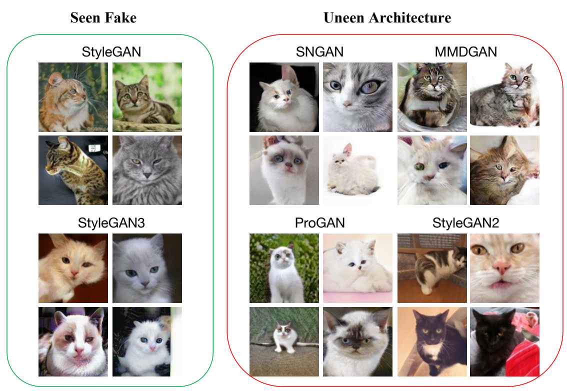
(b) Face-HQ

(c) ImageNet

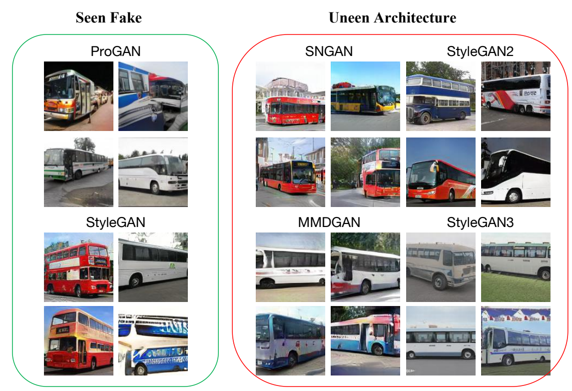


(d) LSUN-Bedroom

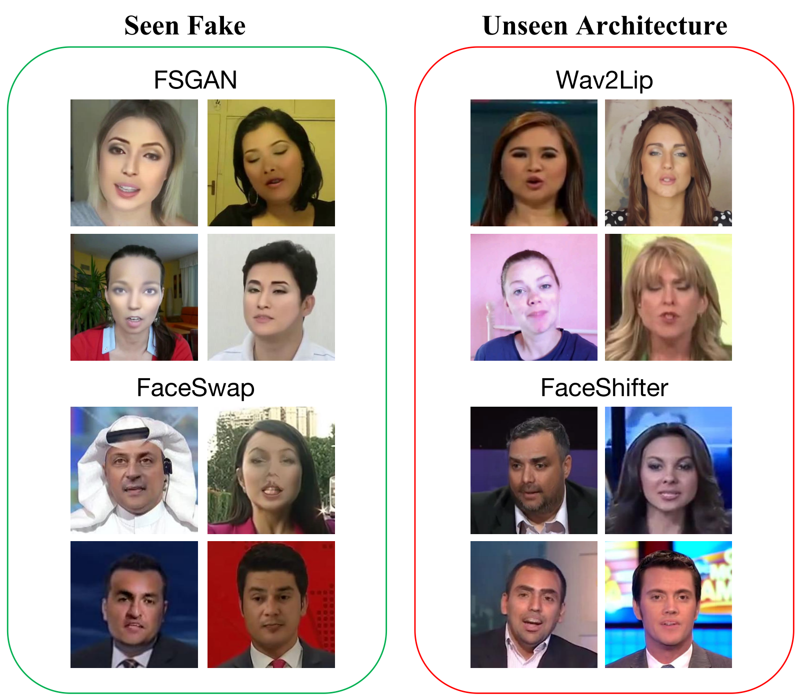
Figure 5. Randomly selected samples for models trained on CelebA (a), Face-HQ (b), ImageNet (c), LSUN-Bedroom (d), LSUN-Cat (e), LSUN-Bus (f), and Youtube (g) dataset. Seen Fake, Unseen Seed and Unseen Architecture are based on Split 1 of the benchmark.



(e) LSUN-Cat



(f) LSUN-Bus



(g) Youtube

Figure 5. Randomly selected samples for models trained on CelebA (a), Face-HQ (b), ImageNet (c), LSUN-Bedroom (d), LSUN-Cat (e), LSUN-Bus (f), and Youtube (g) dataset. Seen Fake, Unseen Seed and Unseen Architecture are based on Split 1 of the benchmark. *(cont.)*