More results for paper: "*Perception Matters: Exploring Imperceptible and Transferable Anti-forensics for GAN-generated Fake Face Imagery Detection*"

Section I Experimental results on diverse datasets

Dataset 1. StyleGAN dataset [1] (*Karras et al. CVPR 2019*)

First, we report the TPR/TNR of 7 investigated forensic models in Table 1 to ensure forensic models can achieve good detection performance.

Table 1. Pretrained forensic models we evaluated and their performances measured by TPR and TNR on the StyleGAN dataset.

models	m1	m2	m3	m4	m5	m6	NDL
TPR (%)	98.6	94.1	91.4	95.8	90.8	99.6	98.6
TNR (%)	98.7	96.9	94.6	98.0	94.4	99.9	98.7

Next, we report the attack success rates from different DL-based source models and the quantitative visual performances in Table 2. The perturbation bounds were selected to make sure baseline attacks (FGSM and MIM) could achieve good attacking performance: FGSM (ε=5.5), MIM (ε=6.0), the proposed method ($\varepsilon^{[c]}$=2.5/6/6); With comparable ASR, we are going to compare the visual quality.

Table 2. Performance comparison of the attack success rate (%) and the visual quality when applying FGSM, MIM and the proposed method on the StyleGAN-generated fake face images. Best performances have been highlighted in bold.

models	Attack	m1	m2	m3	m4	m5	m6	NDL	Avg. ASR	NIQE	LPIPS	FSIMc
m1	FGSM	98.6	90.9	73.4	58.9	20.6	72.5	97.6	73.2	1.188	0.026	0.955
	MIM	98.6	91.6	77.4	63.4	21.1	81.2	98.6	76.0	1.032	0.028	0.952
	Prop.	98.6	91.2	83.3	78.3	40.9	63.3	98.6	**79.2**	**0.798**	**0.020**	**0.984**
m2	FGSM	89.0	94.1	68.5	40.3	12.4	39.1	98.6	63.1	1.321	0.027	0.962
	MIM	95.1	94.1	74.7	38.3	11.8	59.1	93.0	66.6	1.118	0.022	0.962
	Prop.	94.2	94.1	83.3	60.9	30.6	25.4	99.6	**69.6**	**0.841**	**0.018**	**0.987**
m3	FGSM	7.6	16.6	91.4	59.0	11.2	2.8	98.6	41.0	1.613	0.027	0.974
	MIM	8.7	17.8	91.4	61.9	10.0	2.7	98.6	41.6	1.580	0.024	0.977
	Prop.	10.9	24.9	91.4	76.1	22.0	0.7	98.6	**46.4**	**1.165**	**0.023**	**0.990**
m4	FGSM	4.5	8.9	60.2	40.5	25.2	1.6	98.6	34.2	1.588	0.026	0.974
	MIM	3.2	7.4	42.4	95.8	16.4	1.2	98.6	37.9	1.512	0.024	0.981
	Prop.	5.0	12.1	60.9	95.8	29.1	0.6	98.6	**43.2**	**1.130**	**0.021**	**0.990**
m5	FGSM	0.9	3.2	16.5	16.0	88.2	1.5	98.5	32.1	1.509	0.028	0.972
	MIM	0.7	2.1	12.9	10.4	90.8	1.0	97.3	30.7	1.313	0.026	0.962
	Prop.	1.3	5.3	28.7	30.2	90.8	0.6	98.6	**36.5**	**1.041**	**0.021**	**0.990**
m6	FGSM	39.3	28.1	41.8	29.8	18.7	99.6	97.0	50.6	1.531	0.052	0.952
	MIM	53.4	34.6	41.7	26.2	15.9	99.6	86.0	51.1	1.294	0.037	0.962
	Prop.	36.3	27.5	47.2	44.8	31.7	99.6	98.6	**55.1**	**0.908**	**0.031**	**0.988**

Based on Table 2, we have some observations: With comparable (or higher) averaged ASR with FGSM and MIM, the proposed method generates adversarial examples with much improved visual quality (i.e., lower NIQE/LPIPS and higher FSIMc). For example, when the source model is chosen as m1 (avg. ASR: 73.2% for FGSM, 76.0% for MIM, and **79.2**% for the Proposed method), the NIQEs are: 1.188 vs 1.032 vs. **0.798** (the lower the better); the LPIPS: 0.028 vs 0.026 vs **0.020**(the lower the better); FSIMc: 0.955 vs 0.952 vs **0.984** (the higher the better).

Dataset 2. StyleGAN2 dataset [2] (*Karras et al. CVPR 2020*)

In Table 3, we report the TPR/TNR of seven investigated forensic models on clean styleGAN2 dataset (following the same dataset split procedure as the StyleGAN dataset during training/val/test).

Table 3. Pretrained forensic models we evaluated and their performances measured by TPR and TNR on StyleGAN2 dataset.

models	m1	m2	m3	m4	m5	m6	NDL
TPR (%)	98.8	99.0	98.1	98.5	96.2	99.9	99.5
TNR (%)	99.2	99.4	98.5	98.5	97.2	99.9	99.4

To ensure different attack methods could achieve comparable ASR, we used the following parameters: FGSM (ε=6.0), MIM (ε=7.5), the proposed method ($\varepsilon^{[c]}$=2/6/6). The comparison results have been reported in Table 4.

Table 4. Performance comparison of the attack success rate (%) and the visual quality when applying FGSM, MIM and the proposed method on the StyleGAN2-generated fake face images. Best performances have been highlighted in bold.

models	Attack	m1	m2	m3	m4	m5	m6	NDL	Avg. ASR	NIQE	LPIPS	FSIMc
m1	FGSM	98.8	98.9	84.2	94.6	5.0	2.6	62.5	63.8	1.728	0.034	0.969
	MIM	98.8	99.0	97.5	97.5	6.4	23.3	41.0	66.2	1.660	0.036	0.965
	Prop.	98.8	99.0	98.1	98.5	12.9	14.4	92.6	**73.5**	**1.029**	**0.018**	**0.992**
m2	FGSM	98.7	99.0	89.3	92.7	4.5	0.9	67.5	64.7	1.694	0.033	0.970
	MIM	98.8	99.0	97.0	95.7	5.4	3.3	48.3	63.9	1.514	0.031	0.970
	Prop.	98.8	99.0	98.1	98.5	9.8	1.5	88.8	**70.6**	**0.953**	**0.016**	**0.993**
m3	FGSM	93.7	87.8	98.1	90.4	5.5	0.9	91.3	66.8	1.772	0.029	0.975
	MIM	97.7	95.8	98.1	96.8	6.7	0	39.1	62.0	1.849	0.031	0.975
	Prop.	98.8	99.0	98.1	98.5	13.8	0	76.1	**69.2**	**1.140**	**0.019**	**0.994**
m4	FGSM	90.8	82.2	53.7	98.5	4.9	0	74.7	57.8	1.764	0.029	0.975
	MIM	96.9	90.6	64.1	98.5	5.3	0	25.3	54.4	1.819	0.027	0.975
	Prop.	98.8	99.0	96.9	98.5	10.0	0	57.3	**65.8**	**1.084**	**0.018**	**0.994**
m5	FGSM	38.0	30.0	17.9	46.4	96.2	0	90.2	45.5	1.632	0.032	0.971
	MIM	27.2	23.6	13.2	38.4	96.2	0	72.6	38.7	1.553	0.027	0.975
	Prop.	87.1	71.3	52.2	92.0	96.2	0	83.2	**68.9**	**0.948**	**0.015**	**0.994**
m6	FGSM	46.9	57.4	22.9	36.8	5.1	99.9	99.5	52.6	1.503	0.049	0.953
	MIM	58.7	80.2	30.3	39.6	6.3	99.8	99.5	59.2	1.429	0.047	0.953
	Prop.	83.1	89.2	54.2	93.1	13.4	99.9	99.5	**76.1**	**0.848**	**0.027**	**0.991**

In Table 4, we can find that, on StyleGAN2, even with higher averaged ASR (e.g. more than 7.3% higher when source model is m1), the proposed method still achieves much improved visual quality. For instance, with source model as m1, the NIQE metric comparison: 1.728 for FGSM vs 1.660 for MIM and **1.029** for the proposed method; the LPIPS comparison: 0.034 for FGSM vs. 0.036 for MIM and **0.018** for the proposed method; FSIMc comparison: 0.969 for FGSM vs. 0.965 for MIM and **0.992** for the proposed one. Finally, to avoid possible confusions, we would also clarify that to make a fair comparison between the proposed method and baselines on individual models, we need to keep their visual quality comparable: e.g., to have comparably high visual quality, on m6 (source model as m1), for FGSM and MIM, their ASRs are about 0.3% and 0.8% which almost fail, while the proposed method achieves **14.4**% (though still not high enough, yet a large improvement over baselines). This better explains the compromise between ASRs and visual quality, and the proposed method has largely alleviated/reconciled this trade-off.

Dataset 3. ProGAN dataset [3] (*Karras et al. ICLR 2018*)

In Table 5, we report the TPR/TNR of seven investigated forensic models on clean ProGAN dataset (following the same dataset split procedure as the StyleGAN dataset during training/val/test).

Table 5. Pretrained forensic models we evaluated and their performances measured by TPR and TNR on ProGAN dataset.

models	m1	m2	m3	m4	m5	m6	NDL
TPR (%)	98.2	97.8	96.5	97.6	95.3	99.9	99.9
TNR (%)	98.2	98.0	96.7	97.6	95.3	100	99.6

In the ProGAN experiment, the perturbation bounds were: FGSM (ε=7.5), MIM (ε=8.0), the proposed method ($\varepsilon^{[c]}$=3/6/6). The comparison results have been reported in Table 6.

Table 6. Performance comparison of the attack success rate (%) and the visual quality when applying FGSM, MIM and the proposed method on the ProGAN-generated fake face images. Best performances have been highlighted in bold.

models	Attack	m1	m2	m3	m4	m5	m6	NDL	Avg. ASR	NIQE	LPIPS	FSIMc
m1	FGSM	98.2	30.7	32.5	3.3	18.5	0	99.9	40.4	2.101	0.053	0.953
	MIM	98.2	97.6	80.1	3.6	15.7	0.6	96.1	56.0	1.428	0.026	0.965
	Prop.	98.2	97.8	85.1	5.1	26.1	1.7	99.8	**59.1**	**1.101**	**0.019**	**0.987**
m2	FGSM	9.5	97.2	27.0	2.5	17.9	0	99.7	36.3	2.109	0.050	0.958
	MIM	97.0	97.8	70.4	2.4	15.4	0.2	93.9	53.9	1.410	0.025	0.971
	Prop.	97.7	97.8	78.7	4.0	28.0	0.9	99.2	**58.0**	**1.050**	**0.020**	**0.988**
m3	FGSM	0	0.3	96.5	2.9	16.7	0	99.7	30.9	2.003	0.050	0.959
	MIM	2.5	6.6	96.5	2.7	14.1	0	80.6	29.0	1.738	0.029	0.972
	Prop.	0	3.8	96.5	3.8	26.1	0.2	99.6	**32.9**	**1.328**	**0.025**	**0.987**
m4	FGSM	0	0	8.1	97.6	37.3	0.7	99.9	34.8	1.712	0.048	0.956
	MIM	0	0	10.7	97.6	37.5	0.5	99.8	35.2	1.588	0.037	0.964
	Prop.	0	0	10.9	97.6	51.6	0.9	99.7	**37.2**	**1.262**	**0.030**	**0.987**
m5	FGSM	0	0	3.2	1.9	93.6	0.1	98.9	28.2	1.716	0.045	0.959
	MIM	0	0	2.9	1.2	95.3	0	96.5	28.0	1.453	0.025	0.975
	Prop.	0	0	3.1	2.2	95.3	0	96.2	**28.1**	**1.217**	**0.024**	**0.987**
m6	FGSM	0	0.9	8.4	17.6	27.8	99.8	99.9	36.3	1.462	0.061	0.945
	MIM	10.6	13.0	17.7	17.6	23.7	99.9	99.1	40.2	1.118	0.035	0.959
	Prop.	7.5	9.7	15.6	27.1	37.5	99.9	98.9	**42.3**	**0.911**	**0.029**	**0.982**

Similarly with the observations as reported on the StyleGAN and StyleGAN2 datasets, from Table 6, we can also see that the proposed method can achieve improved visual quality with comparable (or higher) averaged ASRs on the fake face imagery antiforensic task.

Dataset 4. StyleGAN dataset [1] (*Karras et al. CVPR 2019*) with larger image resolution

In Table 7, we report the TPR/TNR of seven investigated forensic models on clean StyleGAN dataset (image resolution as 512x512), by following the same dataset split procedure as Datasets 1-3 during training/val/test.

On the clean test dataset, the performance of each model is reported in Table 7 as follows.

Table 7. Pretrained forensic models we evaluated and their performances measured by TPR and TNR on StyleGAN (512x512) dataset.

models	m1	m2	m3	m4	m5	m6	NDL
TPR (%)	99.9	99.9	100	99.9	99.7	100	99.9
TNR (%)	100	100	100	100	99.9	100	99.8

As shown in Table 7, forensic models achieve very good performance on the StyleGAN (512x512) dataset. Compared with StyleGAN (128x128) dataset, all forensic models have higher detection accuracy. This is because there are indeed more forensic traces to be made use of on larger images.

Table 8. Performance comparison of the attack success rate (%) and the visual quality when applying FGSM(ε=5.5), MIM(ε=8.0) and the proposed method ($\varepsilon^{[c]}$=2/6/6) on the fake face images with source model as m1 (trained on the StyleGAN2 dataset). Best performances have been highlighted in bold.

Attack	m1	m2	m3	m4	m5	m6	NDL	Avg. ASR	NIQE	LPIPS	FSIMc
FGSM	99.9	0	5.6	99.9	10.1	0	98.7	44.9	2.124	0.043	0.995
MIM	99.9	99.9	16.5	99.9	23.3	0	97.3	62.4	1.973	0.050	0.995
Prop.	99.9	99.7	23.7	99.9	24.1	0.1	97.9	**63.6**	**1.279**	**0.033**	**0.999**

As some examples, we evaluate the baselines and the proposed method with source model as m1. First, we use a similar set of parameters (i.e. perturbation bound) as in StyleGAN (128x128), and the comparison results are shown in Table 8. We observe the adversarial images (512x512) from FGSM, MIM and the proposed method appear to have improved visual metrics than those from 128x128 images. Yet we observe their averaged ASRs also decreased, correspondingly. This is probably because, on larger images, the computed perturbations on some pixels equal zero (or almost zero but thresholded as zero). Despite that, the proposed method still achieves the best visual quality given comparable (or higher) averaged attack success rates.

In addition, we experimented with larger perturbation bounds on StyleGAN (512x512), and we report the comparison results as in Table 9. As shown in Table 9, the averaged ASRs are largely improved for each of the three attacks at the expense of reduced visual quality. Nevertheless, we can still conclude that, compared with baselines, the proposed method can *still achieve much better visual quality given comparable (or higher) averaged attack success rates*.

Table 9. Performance comparison of the attack success rate (%) and the visual quality when applying FGSM(ε=16.0), MIM(ε=16.0) and the proposed method ($\varepsilon^{[c]}$=5/12/12) on the fake face images with source model as m1 (trained on the StyleGAN2 dataset). Best performances have been highlighted in bold.

Attack	m1	m2	m3	m4	m5	m6	NDL	Avg. ASR	NIQE	LPIPS	FSIMc
FGSM	99.9	99.9	11.7	99.6	96.7	5.6	99.6	73.3	3.179	0.200	0.966
MIM	99.9	99.9	19.8	99.9	86.3	1.2	98.7	72.2	3.179	0.152	0.980
Prop.	99.9	99.9	34.6	99.9	94.3	2.0	98.5	**75.6**	**2.557**	**0.134**	**0.993**

Section II More examples for visual quality comparison on diverse datasets

With comparable ASRs, in this section we show more visual examples generated from different datasets. The perturbation bounds of each attack is the same as for comparison tables in Tables 2, 4, 6, 9.

1) On StyleGAN (128x128) dataset [1] (Karras et al. CVPR 2019)

Some comparison examples are shown in Fig. 1.

Figure 1. Examples of fake face images for visual quality comparisons on FGSM, MIM and the proposed method on StyleGAN. We recommend to zoom in the digital images for better visual comparison (pdf conversion may lose some noisy details e.g. noisy textures in facial/background regions of baseline attacks).

2) On StyleGAN2 dataset [2] (Karras et al. CVPR 2020)

In Fig. 2, we show some comparison examples on StyleGAN2.

3) On ProGAN dataset [3] (Karras et al. ICLR 2018)

In Fig. 3, we show some comparison examples on ProGAN.

4) On StyleGAN (512x512) dataset [1] (Karras et al. CVPR 2019)

In Fig. 4, we show some comparison examples on StyleGAN (512x512).

Figure 2. Examples of fake face images for visual quality comparisons on FGSM, MIM and the proposed method on StyleGAN2. We recommend to zoom in the digital images for better visual comparison (pdf conversion may lose some noisy details e.g. noisy textures in facial/background regions of baseline attacks).

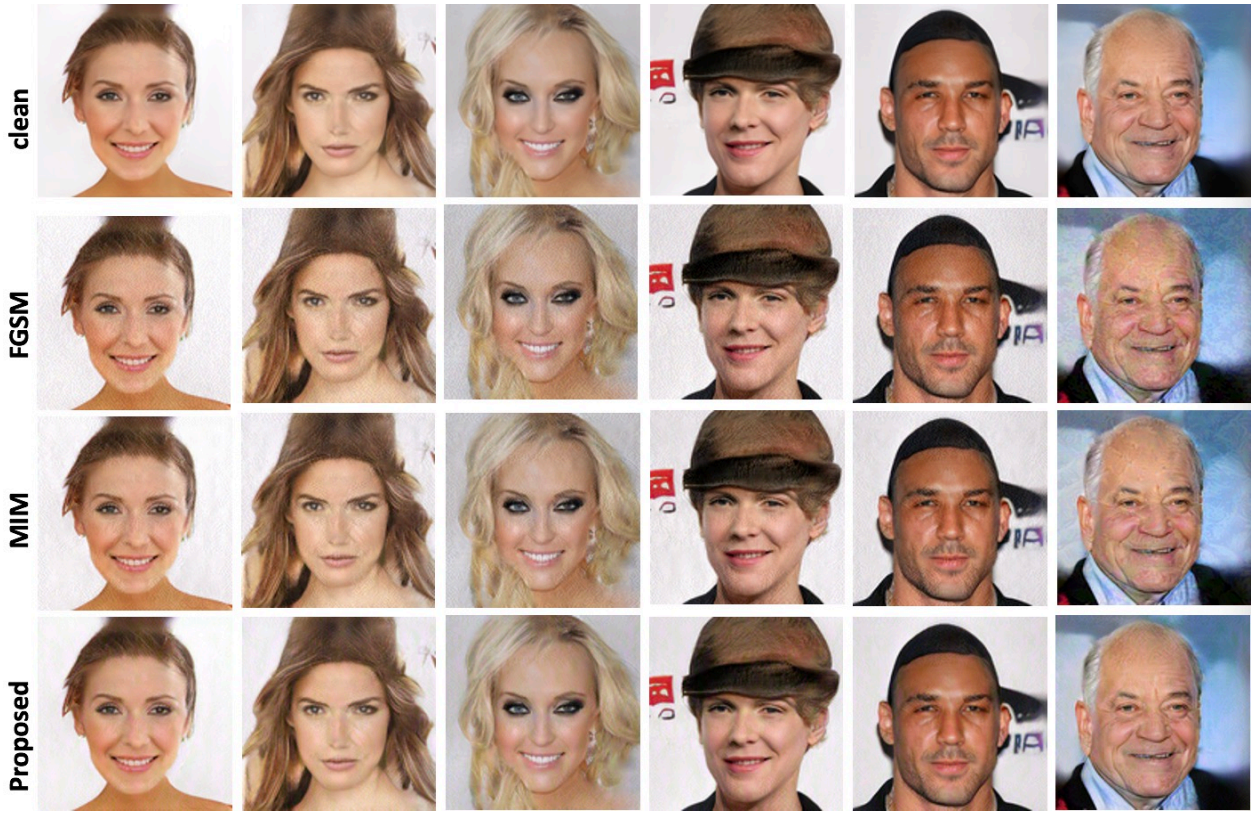

Figure 3. Examples of fake face images for visual quality comparisons on FGSM, MIM and the proposed method on ProGAN. We recommend to zoom in the digital images for better visual comparison (pdf conversion may lose some noisy details e.g. noisy textures in facial/background regions of baseline attacks).

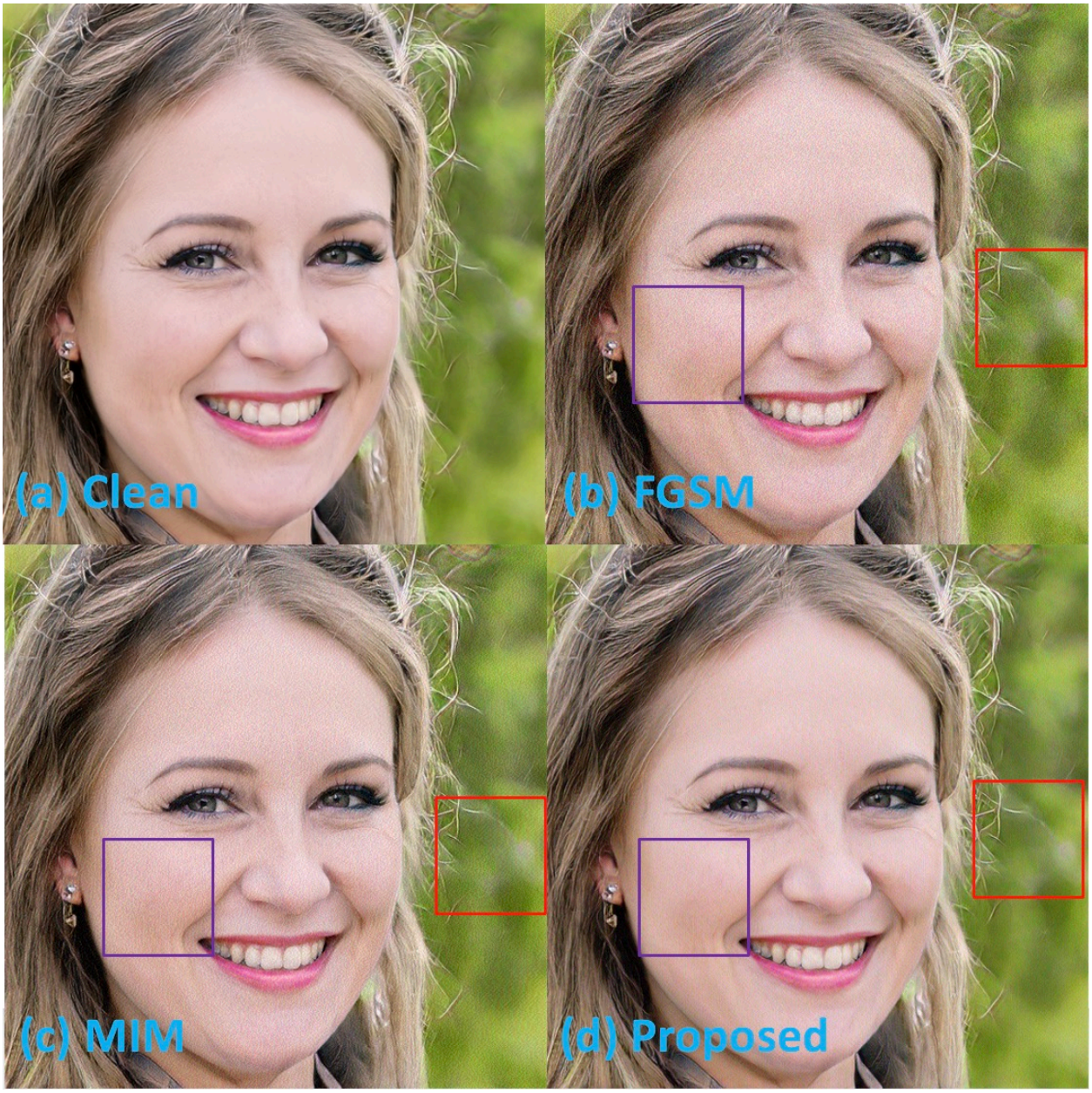

Figure 4. Examples of fake face images for visual quality comparisons on FGSM, MIM and the proposed method on StyleGAN (512x512). We recommend to zoom in the digital images for better visual comparison (pdf conversion may lose some noisy details e.g. noisy textures in facial/background regions of baseline attacks).

Section III Some observations and analysis

The transferability is asymmetric between forensic models. For instance, in Table 1, with source model as m1, we can see that the ASRs on m5 or m6 are lower than the rest forensic models. This phenomenon is mainly due to the differences in the network architecture (i.e. module components or network depth etc), which can influence the attack transferability. Generally, a source model transfers more easily to a target model when they adopt a similar architecture; or vice versa. To be specific, m5 is a *lightweight* network architecture particularly designed for the mobile setting. It uses some specially designed modules (e.g. *inverted residual blocks*) while m1 does not adopt such sophisticated modules. As a result, the ASR on m5 is lower than that of the rest DL-based models. As for m6, it adopts the resnet50 as its backbone (50 layers), while the rest architectures are within 10 layers. The much differences in layers can also make the adversarial examples harder to transfer between different models. The asymmetry in network models can also account for the differences in the averaged ASRs when we choose different forensic model as the source model.

References

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

[2] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan. CVPR 2020.

[3] Karras T, Aila T, Laine S, et al. Progressive growing of gans for improved quality, stability, and variation. ICLR 2018.