Attacking on proactive defense methods on Deepfake

Oct 04, 2022

Team 3: Ting-Chih Chen, Xiao Guo

Background

- Deepfake has long been weaponized to cause negative effects on society and individuals
- Researchers proposed multiple defense scheme to detect and disrupt this misuse(passive, proactive)
- Proactive methods aims to inject perturbation into images to intercept generating
- Human eyes can’t catch the perturbation

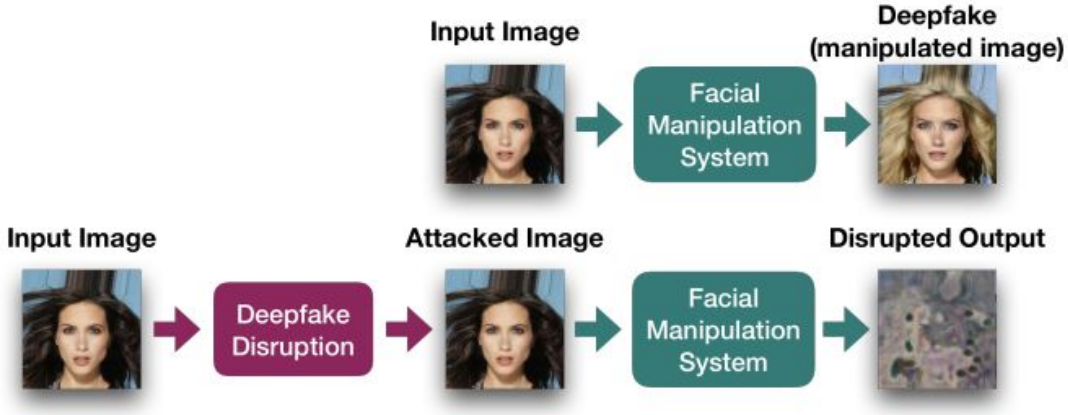


Fig 1.Illustration of deepfake disruption with a real example.[1]

Motivation

- Focusing on Disrupting Deepfakes:
  - Adversarial attacks against conditional image translation networks and facial manipulation systems
- Strong but fake assumption about control over which GAN to use - > **Transferability**
- Finding ways to eliminate the perturbation -> **Denoise and reconstruction**

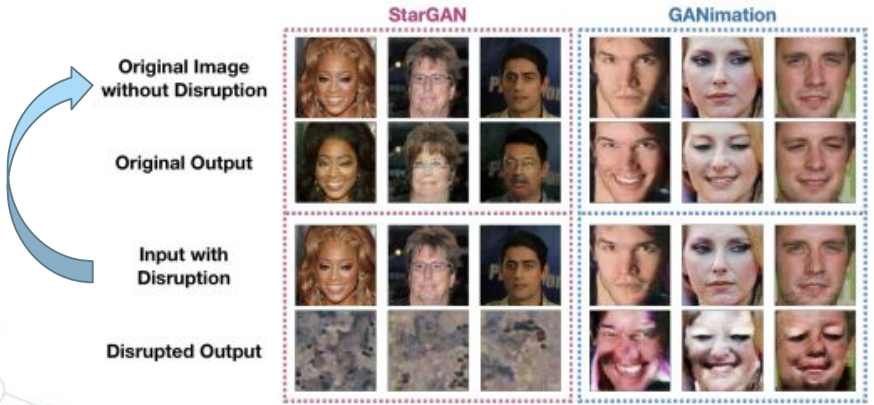


Fig 1. An example of our deepfake disruptions on StarGAN and GANimation[1]

Related Work

- Target Paper(Proactive defense, Perturbation)

[1] Ruiz, Nataniel, Sarah Adel Bargal, and Stan Sclaroff. "Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems." *European Conference on Computer Vision*. Springer, Cham, 2020.

- Proactive Detection (Watermarking)

[2] Yang, Yuankun, et al. "FaceGuard: Proactive Deepfake Detection." *arXiv preprint arXiv:2109.05673* (2021).

- Passive Detection (Neural Network)

[3] He, Yang, et al. "Beyond the spectrum: Detecting deepfakes via re-synthesis." *arXiv preprint arXiv:2105.14376* (2021).

- Study on evading detection

[4]Cao, Xiaoyu, and Neil Zhenqiang Gong. "Understanding the Security of Deepfake Detection." *International Conference on Digital Forensics and Cyber Crime*. Springer, Cham, 2022.

- Reconstruction

[5] Chen, Zhikai, et al. "Magdr: Mask-guided detection and reconstruction for defending deepfakes." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

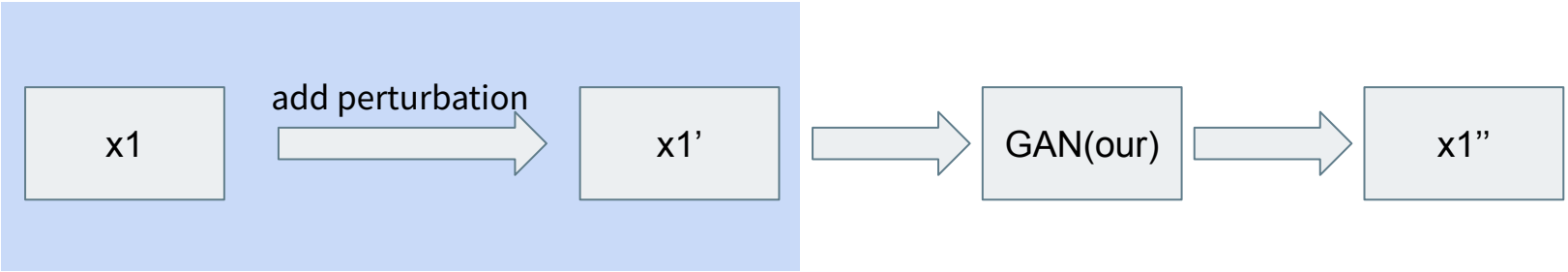
- GANs to use(Alternatives adversary may consider)

[6] Chu, Wenqing, et al. "Learning to caricature via semantic shape transform." *International Journal of Computer Vision* 129.9 (2021): 2663-2679. (Semantic-CariGANs)

[7] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019. (StyleGAN)

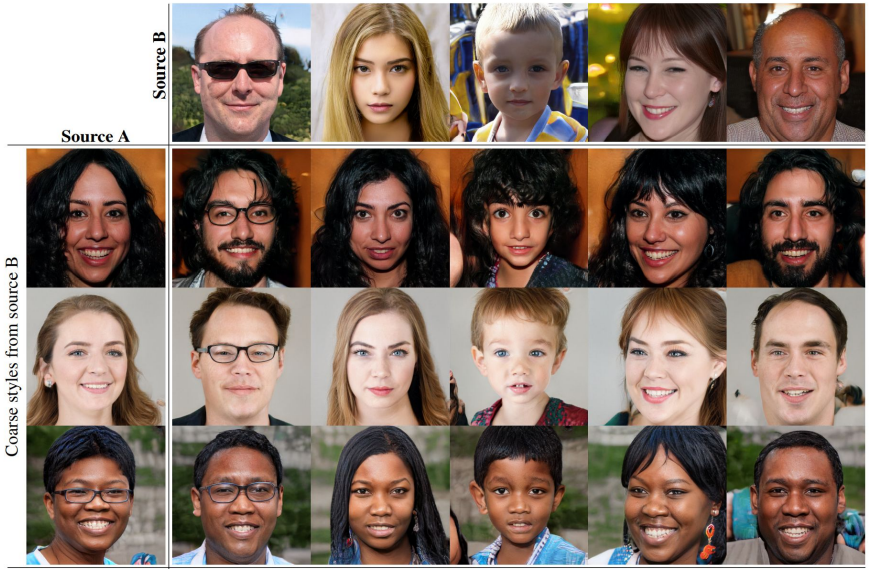
Approach - 1

- We find other GAN to breakdown this disruption
  - Prove x1 is close to x1’’ that mean we remove the perturbation truly



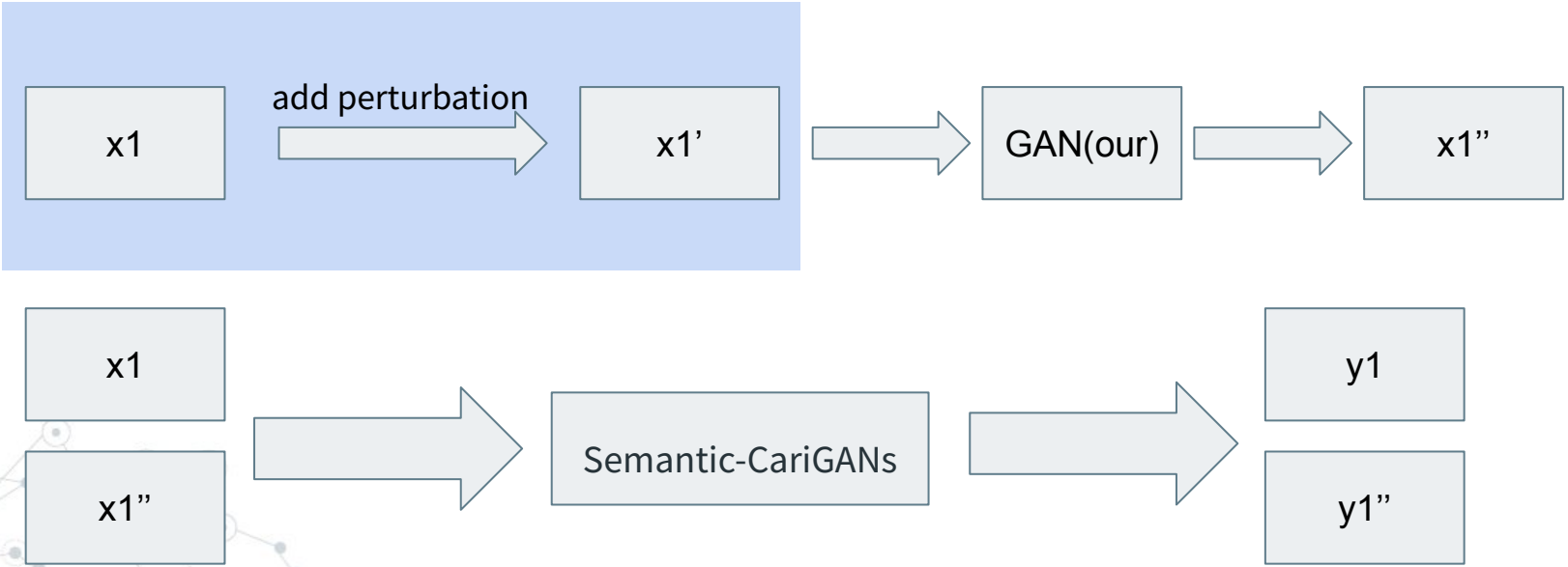
Approach - 1

- StyleGAN
  - StyleGAN can mainly generate the images with style-transform



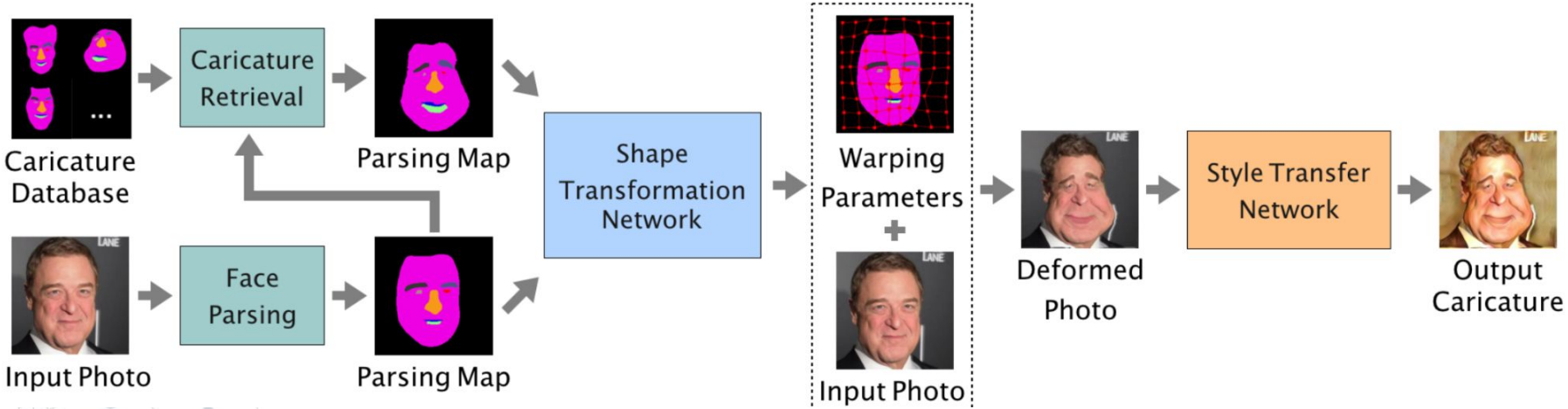
Approach - 2

- We build up a GAN to revert the images also test on caricature style
  - To test caricature whether GAN can catch out the same face features in different images



Approach - 2

- Semantic-CariGANs



Approach - 2

Photo Hand-drawn Our Results with Diverse Shapes and Styles

Fig. 1 Examples of normal photos, hand-drawn caricatures, and a set of caricature outputs generated by the proposed method. Our approach is able to render a diverse set of visually pleasing caricatures.

Evaluation

- Metrics: L1, L2 and MSE(image similarity)
- Approach - 1,2:
  - L1 and L2 metric < 0.05
  - MSE will close to 0

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error

n = number of data points

Y_i = observed values

$\hat{Y}_i$ = predicted values

Questions?