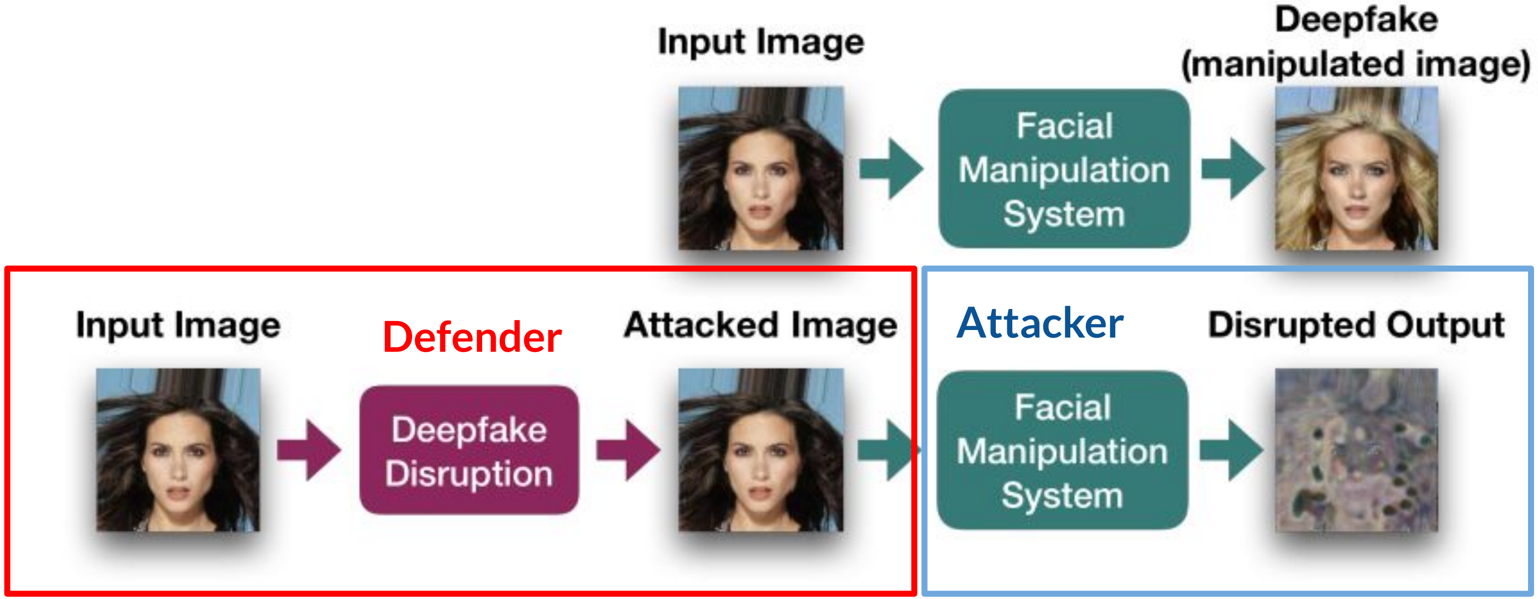
Attacking on proactive defense methods on Deepfake

Group 3: Ting-Chih Chen and Xiao Guo

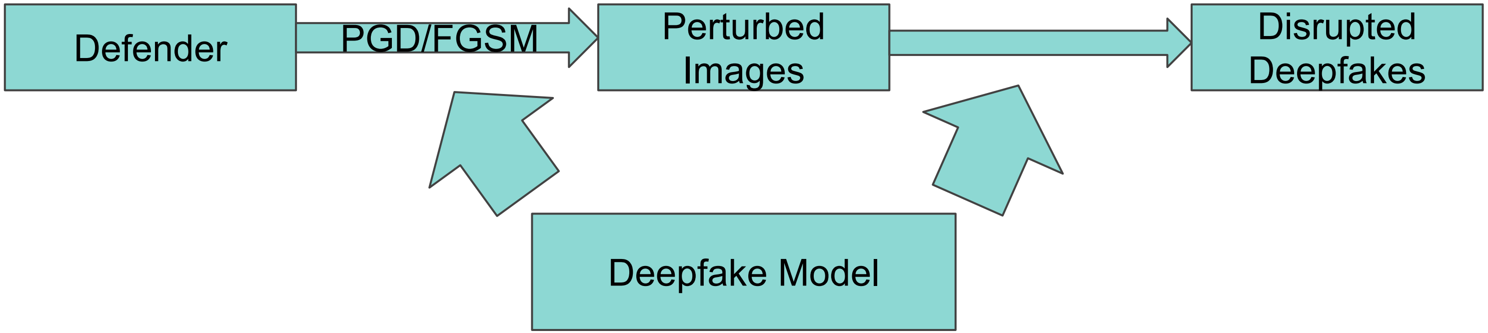
Recap

Target paper: Disrupting Deepfakes: Adversarial Attacks Against Conditional Image Translation Networks and Facial Manipulation Systems



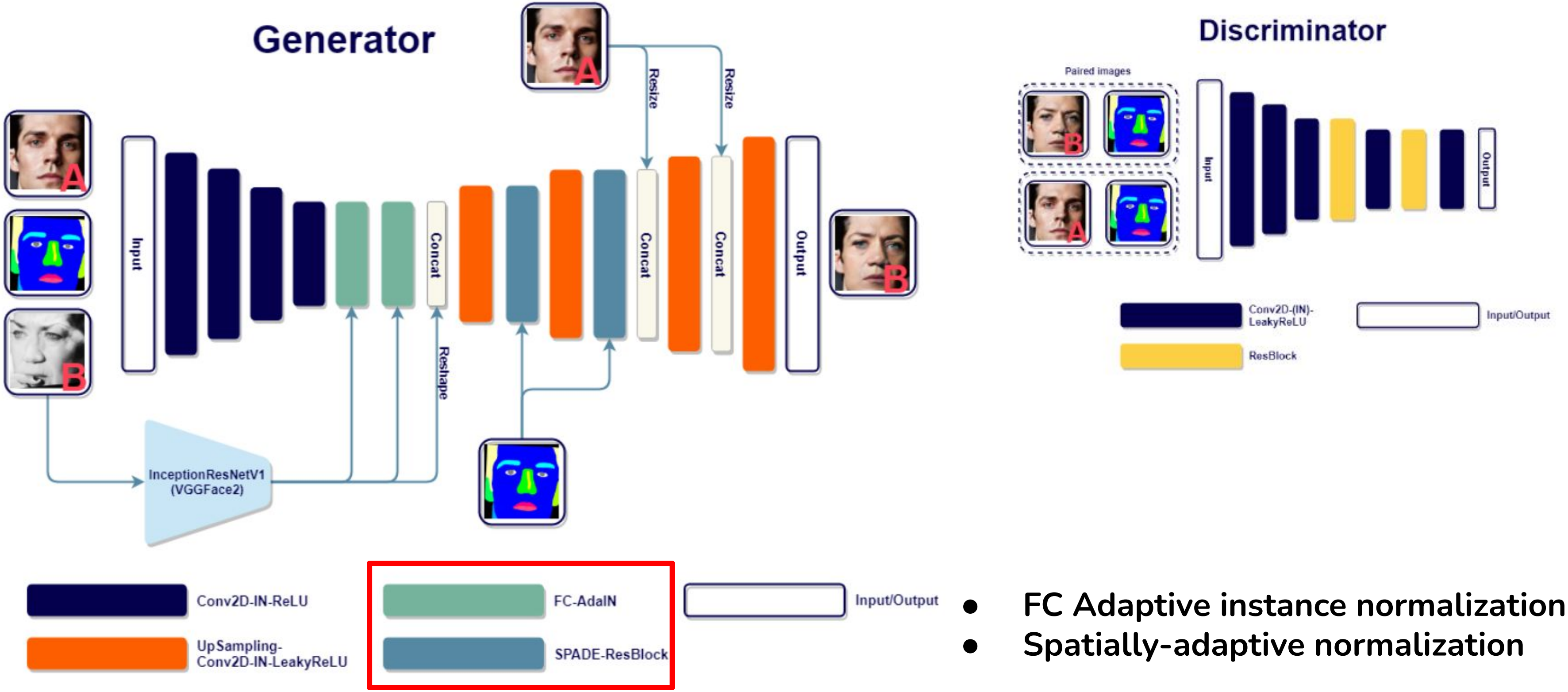
Task#1 Break down Defender - Transferability

- Author’s defense scheme needs to utilize the image translation model that the attacker choose to conduct either PGD/FGSM attack.
- However, which model to use is not decided by the defenders(authors)



Target Image

Method-1 fewshot-face-translation-GAN



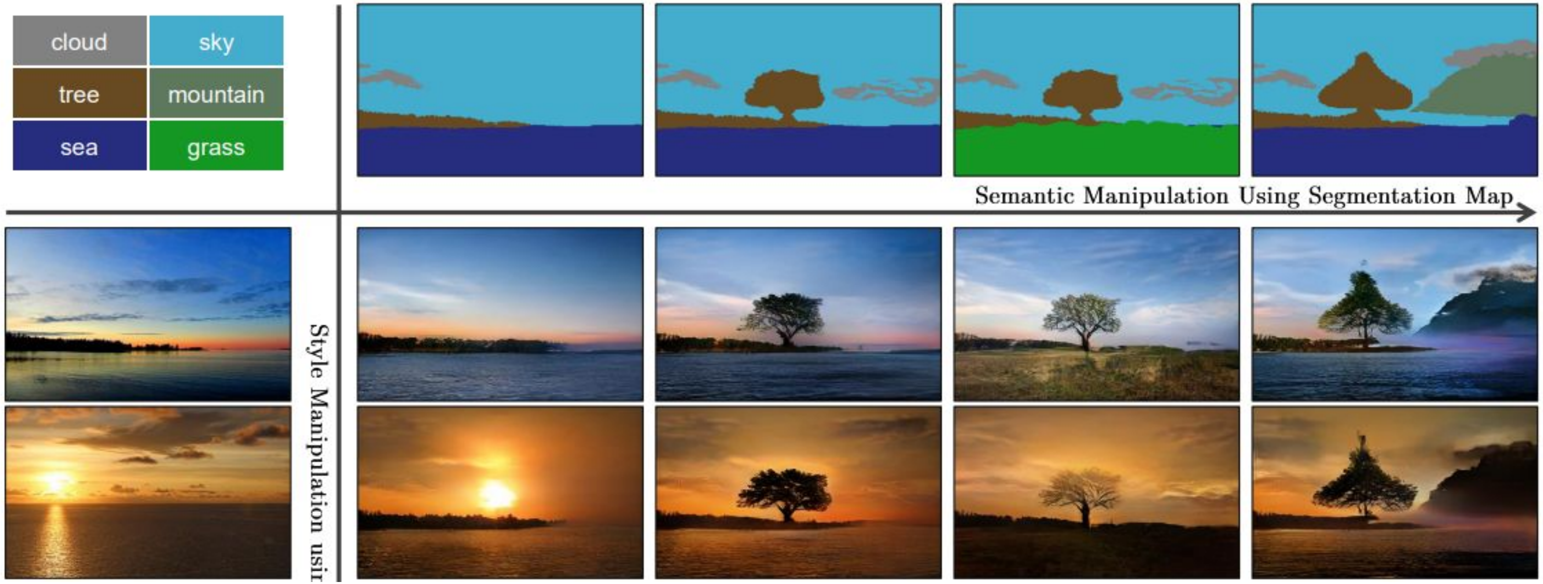
- FC Adaptive instance normalization
- Spatially-adaptive normalization

Adaptive instance normalization(FC-AdaIN)

- The AdaIN residual block is a residual block using the AdaIN as the normalization layer
- AdaIN normalizes the activations of a sample in each channel to have a zero mean and unit variance
- Then, it scales the activations using a learned affine transformation consisting of a set of scalars and biases
- The affine transformation can be used to obtain global appearance info
- Ex:
    - Latent representation(object appearance) -> Decoder with AdaIN -> obtain the content image(locations of eyes)

Spatially-adaptive normalization(SPADE)

- SPADE is a layer for synthesizing photorealistic images given an input semantic layout



Results

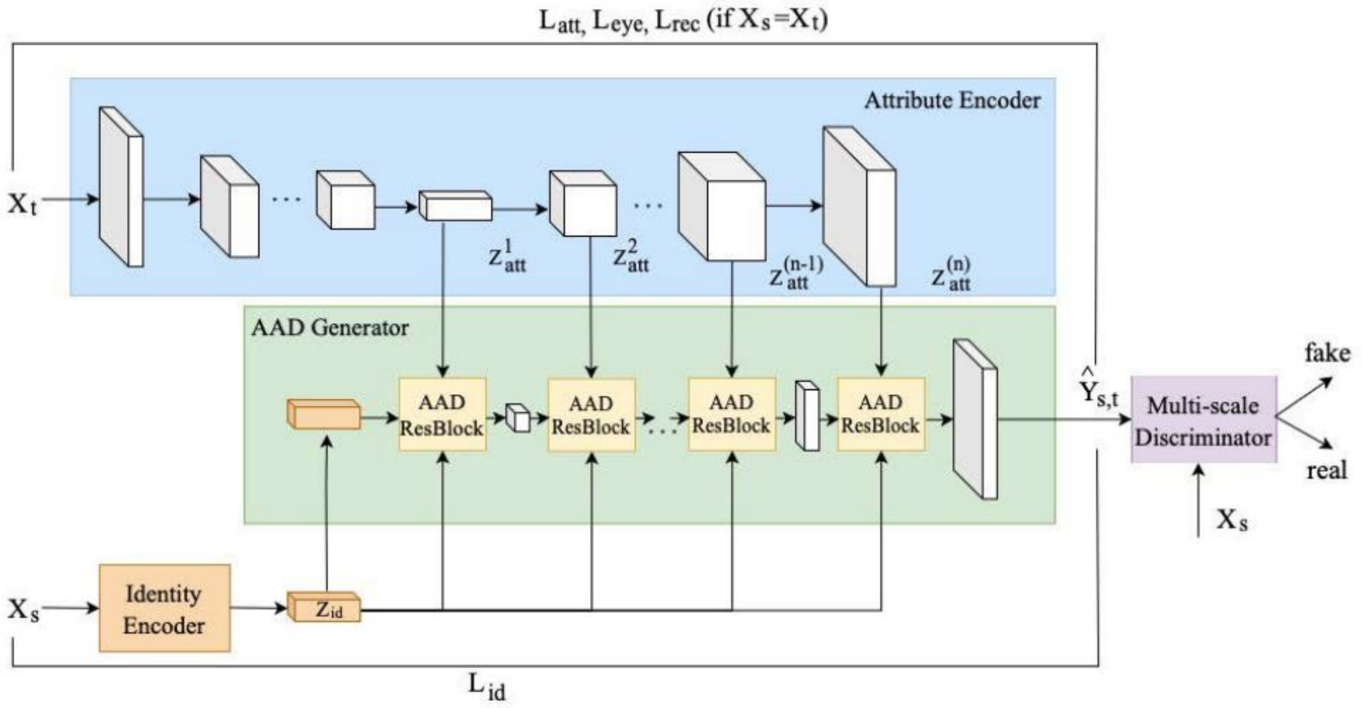
Method-2 GHOST

- Identity encoder is a pre-trained ArcFace model that extracts the features from the source images
- Attribute encoder is a model with a U-Net architecture that extracts attributes from the target images
    - U-Net is fully convolutional network. Authors think this can yield more precise segmentations
- AAD generator is a model to mix attributes and the identities. Then, it will generate new face with source identity and target attribute features

Architecture

Target images

Source images



Results

Method-3 StyleGAN-NADA

- CLIP-Guided Domain Adaptation of Image Generators

- StyleGAN-NADA converts a pre-trained generator to new domains using only a textual prompt and no training data

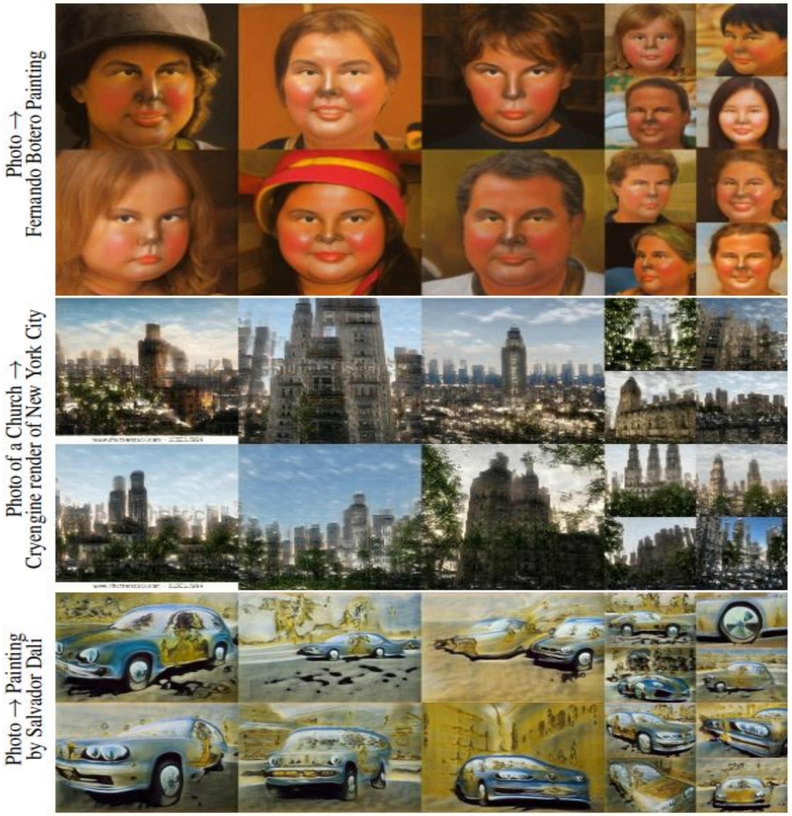


Fig 1. Examples of text-driven, out-of-domain generator adaptations induced by our method[1]

[1]Gal, Rinon, et al. "Stylegan-nada: Clip-guided domain adaptation of image generators." ACM Transactions on Graphics (TOG) 41.4 (2022): 1-13.

Method-3 StyleGAN-NADA

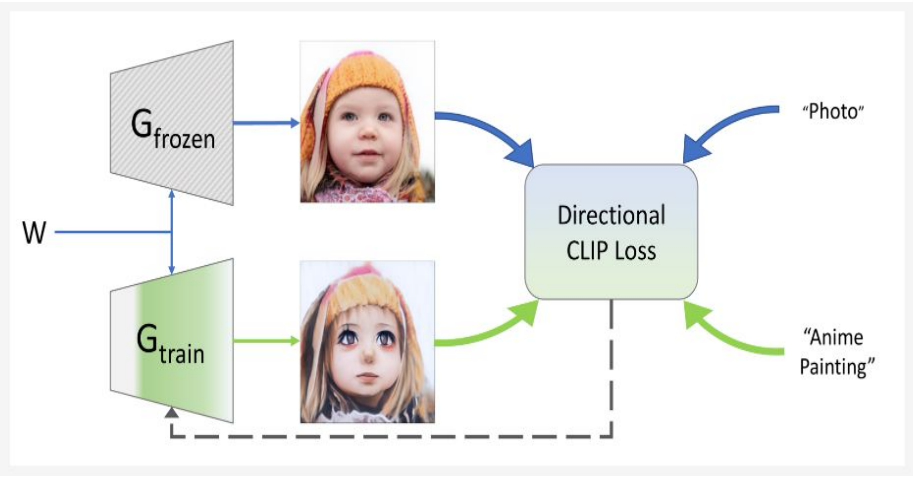


Fig 1. Model Concept[1]

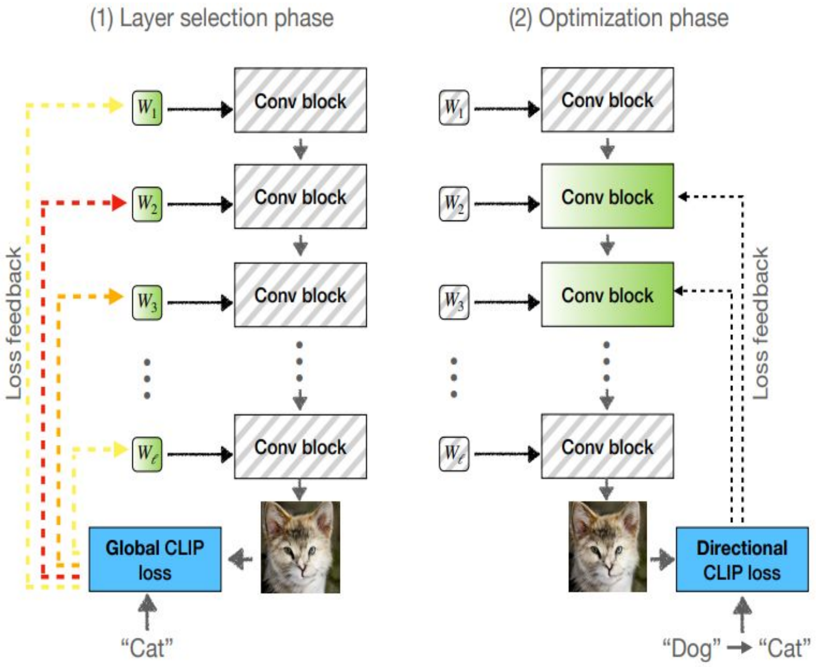


Fig 2. The adaptive layer-freezing mechanism has two phases[1]

[1]Gal, Rinon, et al. "Stylegan-nada: Clip-guided domain adaptation of image generators." ACM Transactions on Graphics (TOG) 41.4 (2022): 1-13.

Method-3 Results

Evaluation

- **MSE**
  - the most widely used and also the simplest full reference metric which is calculated by the squared intensity differences of distorted and reference image pixels
- **SSIM**(structural similarity index measure)
  - gives normalized mean value of structural similarity between the two images, considers luminance, contrast, structure
- **PSNR**(peak signal-to-noise ratio)
  - the ratio between the maximum possible signal power and the power of the distorting noise which affects the quality of its representation

Evaluation

	MSE	SSIM	PSNR
Ghost	**0.737**	**0.998**	**50.396**
Fewshot	57.684	0.951	32.295
StyleGAN-NADA	247.152	0.777	25.259
StyleGAN-Baseline	226.91	0.799	26.611

Task 2 - Removing Perturbations

- Initial Idea: using a denoising autoencoder to denoise the perturbed images
- Model:
- Input: Perturbed images(from CelebA, perturbed using StarGAN)
- Target: Original image(from CelebA)
- 500 Epochs, 64 batch_size, time constraints
- Denoising example:

```
Model: "model_5"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_6 (InputLayer)        [(None, 256, 256, 3)]     0

 conv2d_15 (Conv2D)          (None, 256, 256, 32)      896

 max_pooling2d_10 (MaxPoolin  (None, 128, 128, 32)     0
 g2D)

 conv2d_16 (Conv2D)          (None, 128, 128, 32)      9248

 max_pooling2d_11 (MaxPoolin  (None, 64, 64, 32)       0
 g2D)

 conv2d_transpose_10 (Conv2D  (None, 128, 128, 32)     9248
 Transpose)

 conv2d_transpose_11 (Conv2D  (None, 256, 256, 32)     9248
 Transpose)

 conv2d_17 (Conv2D)          (None, 256, 256, 3)       867

=================================================================
Total params: 29,507
Trainable params: 29,507
Non-trainable params: 0
_________________________________________________________________
```

Future work

1. Formalize Task 1 result, polish the conclusion
2. Try image translation network(pix2pix) with paired labeled image examples