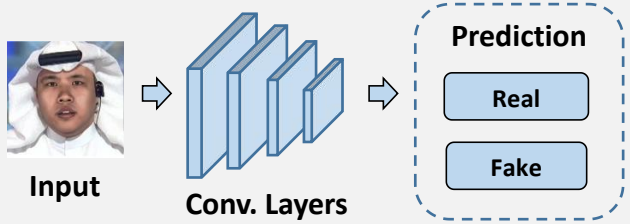
Self-Supervised Adversarial Training for Robust Face Forgery Detection

Yueying Gao, Weiguo Lin, Junfeng Xu, Wanshan Xu, Peibin Chen

BMVC 2023

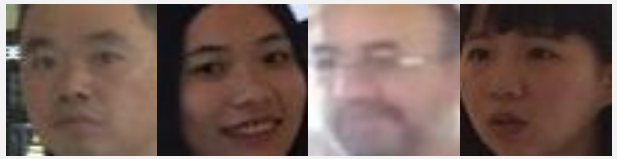
What is Deepfakes Detection

- Auto-Encoders and GAN boost Deepfakes
- Detectors performed well on standerd datasets



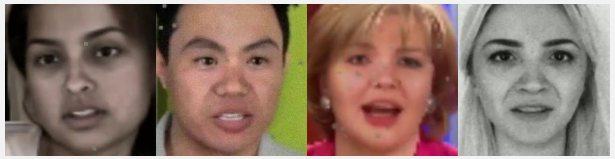
Detect Forgery Faces in Real Life

Low Resolution



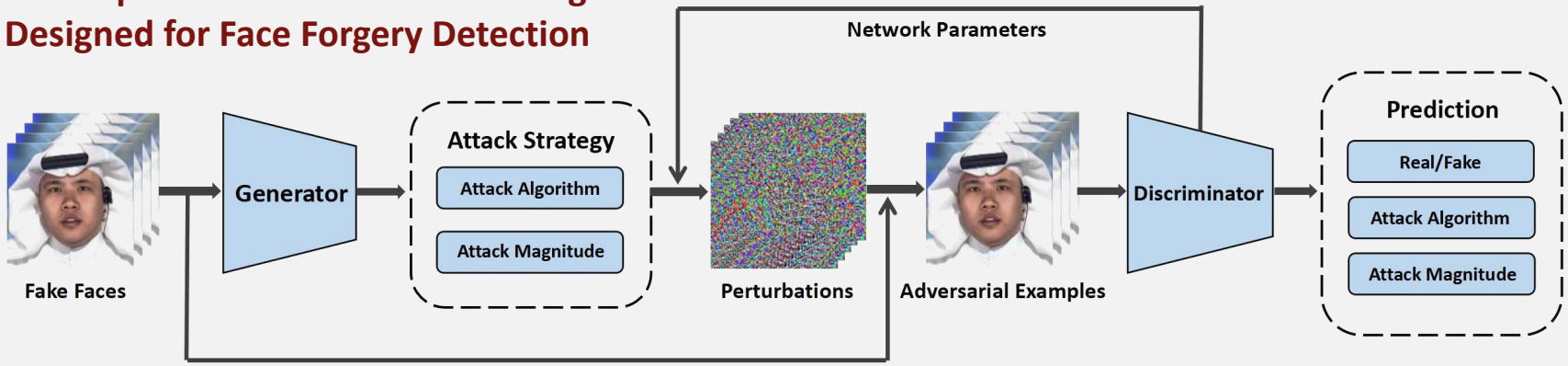
Multiple Faces

Image Corruptions



- Consider introducing adversarial training to boost the robustness of detectors

Self-Supervised Adversarial Training Designed for Face Forgery Detection



Adversarial Attacks

- BIM
- C&W
- Deepfool
- NES

Adversarial Training

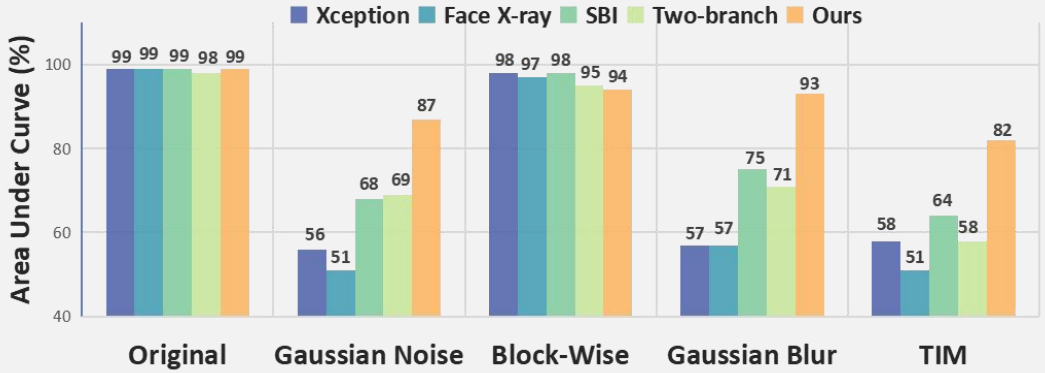
$$\min_{w}\max_{\theta} Loss(\theta, w) = L_{main} + \mu L_T + \lambda L_M$$

L_main & L_T : AM-Softmax L_M : L_1

Fast Training Strategy

- Real images shuffled by frame
  Fake images shuffled by video
- Save the net parameters for adversarial examples generating

Robustness to Perturbed Images



Robustness to Compressed Images

Methods	RAW	HQ	LQ
Xception	0.989	0.961	0.895
Face X-ray	0.988	0.866	0.631
SBI	**0.994**	0.950	0.903
Two-branch	0.981	0.958	0.890
Cvit	-	0.937	-
SAT(ours)	0.992	**0.988**	**0.934**