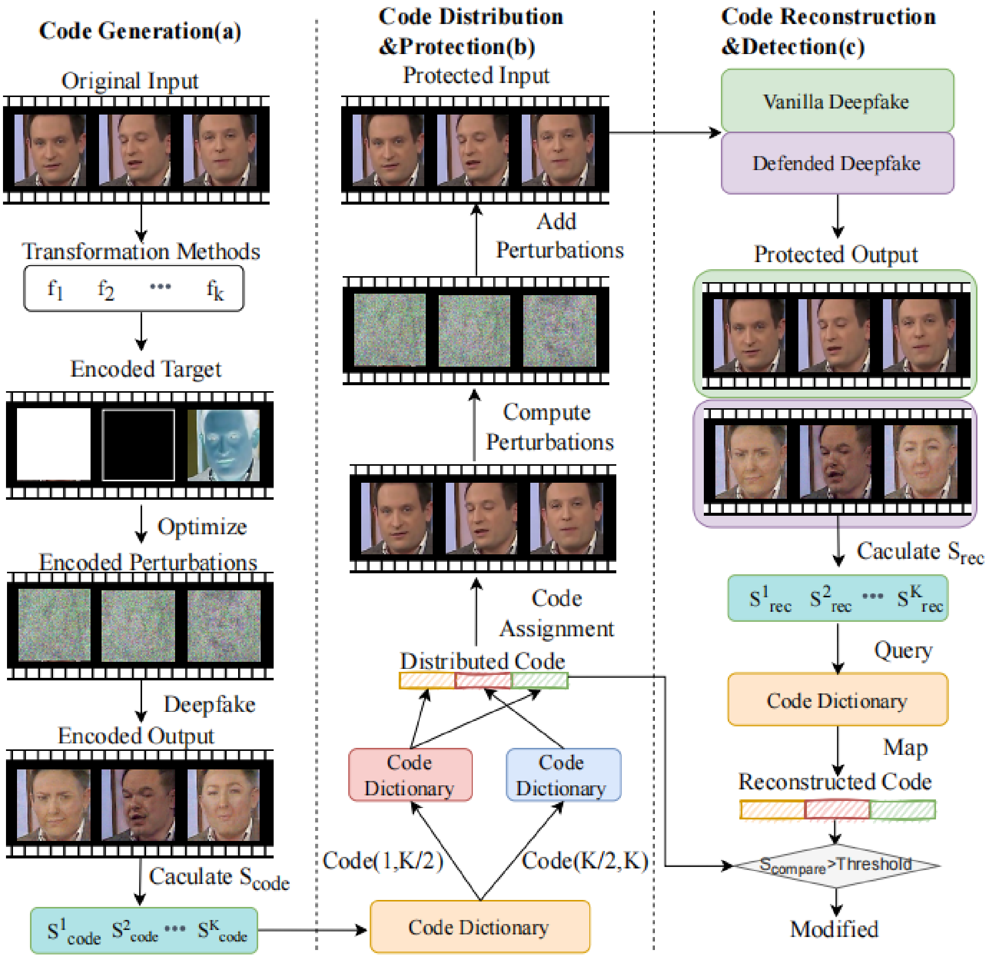
周报——2022.7.11

一、完成工作

这两周精读了论文Actively Encoded Perturbations as Protection against Deepfakes、Disrupting Image-Translation-Based DeepFake Algorithms with Adversarial Attacks。略读了论文Ensembling Off-the-shelf Models for GAN Training、StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation。

1. Actively Encoded Perturbations as Protection against Deepfakes

这是陈志凯师兄投到ACMMM的一篇论文。需要对这篇论文进行改进。具体的方法如下所示。其主要思想是在对抗样本当中添加新的特征，并且这个新的特征在相邻的视频帧之间存在着很大的差异。这篇论文考虑的是白盒的攻击场景，设置比较简单，后面可以考虑通过对抗样本的迁移性实现黑盒攻击，另外这篇论文的攻击方法是PGD，可以考虑通过GAN来实现，以便更好的控制需要的特征。另外可以考虑如何通过对抗样本将deepfake图片重构回原来的图片。具体的笔记[1]



2. Disrupting Image-Translation-Based DeepFake Algorithms with Adversarial Attacks

这篇论文是WACV2020年的论文，也可以算是使用对抗样本防御deepfake的开山之作，他与Disrupting Deepfakes的思想相似，实现比较简单，具体的笔记[2]

其他两篇论文的笔记[3] [4]

二、后面安排

后面主要需要的是对贺勇师兄的论文进行复现，然后尝试改进，实现陈志凯师兄里面的论文方法，然后尝试进行改进。

---

[1] Actively Encoded Perturbations as Protection against Deepfakes ↩

[2] Disrupting Image-Translation-Based DeepFake Algorithms with Adversarial Attacks ↩

[3] Ensembling Off-the-shelf Models for GAN Training ↩

[4] StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation ↩