研究主题：生成式人脸伪造检测

本学期工作总结汇报

汇报人：支雅鹏

汇报目录

汇报目录

一、研究背景与研究价值

研究背景

我们目前所处的2025年，是人工智能生成内容的爆炸性时代，AIGC在学术界和工业界引起了广泛的关注。在这场AIGC革命中，深度伪造可以被更加轻易地使用来操纵一个人的身份或控制人像中的面部表情和动作。不法分子可能利用该技术制作和传播虚假信息，从而带来伦理、法律和社会的多重挑战。因此，当前迫切需要研发有效的检测方法，能够精准识别这些经过伪造的内容，从而协助数字媒体平台筛选虚假信息，维护信息生态的健康与可信度。

伪造

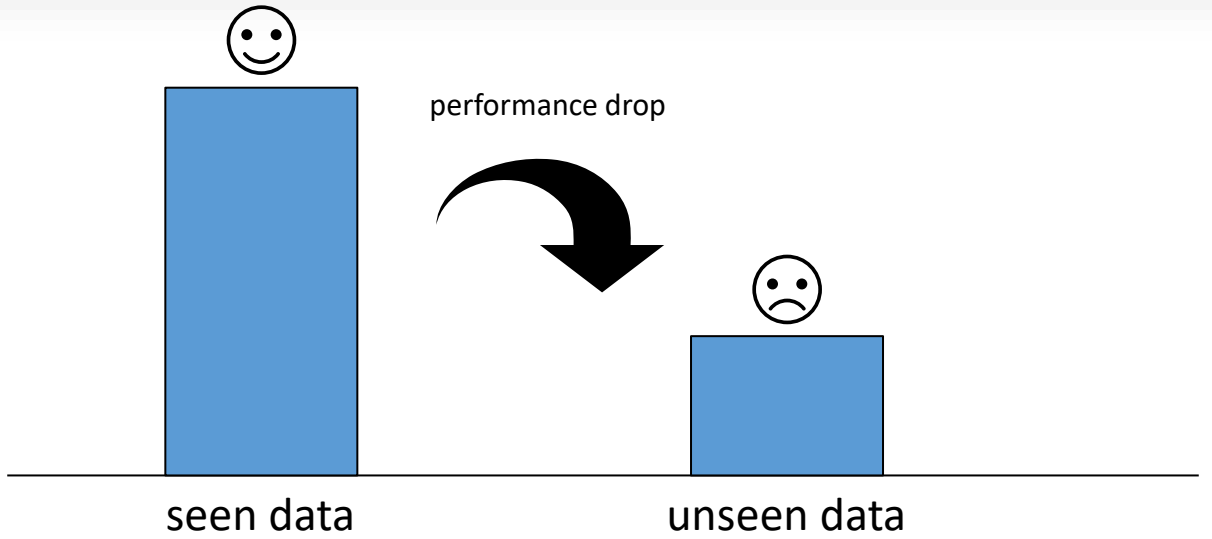
一、研究背景与研究价值

研究价值

随着深度学习的兴起，现有的检测方法重点已经转向基于深度学习的检测器，因为它们在特征提取能力方面表现更优。

不幸的是，现有这些方法都面存在严重的过拟合问题。虽然这些检测方法在遇到训练集中见过的伪造方法所伪造的图片时可以发挥较好检测效果，但是面临未见过的伪造方法所伪造的图片时现有检测器的检测效果都会急剧下降，这阻碍了这些检测方法的实际应用。

研究致力于缓解伪造检测现存的过拟合问题，优化伪造检测方法的泛化表现，促进伪造检测的落地应用。



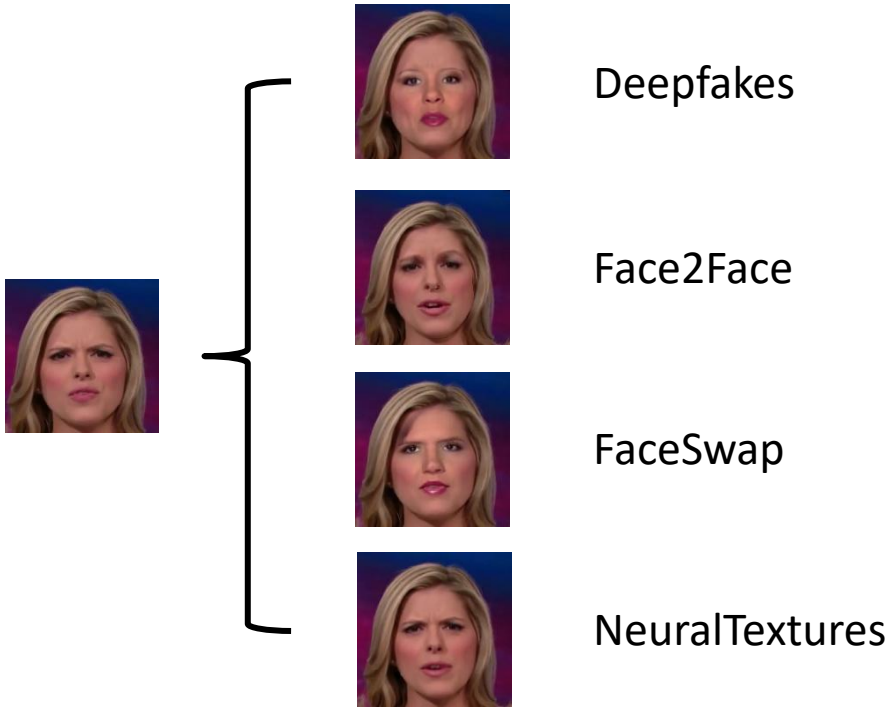
汇报目录

汇报目录

二、国内外研究成果调研

数据集

FaceForensics++: Learning to Detect Manipulated Facial Images (ICCV 2019)



DeepFakes基于深度学习的自编码器方法，使用共享编码器训练源和目标面部的特征重建。

Face2Face使用视频中的表情参数，将源视频表情映射到目标视频的3D面部模型上。

FaceSwap基于图形学的方法，利用稀疏面部关键点提取面部区域，通过3D模板模型拟合源和目标的面部形状。

NeuralTextures通过学习目标人物的神经纹理并结合渲染网络生成表情修改结果。

文章还利用Xception训练了二分类模型进行伪造检测，该模型被后来的研究广泛地使用作为基准模型和骨干网络。

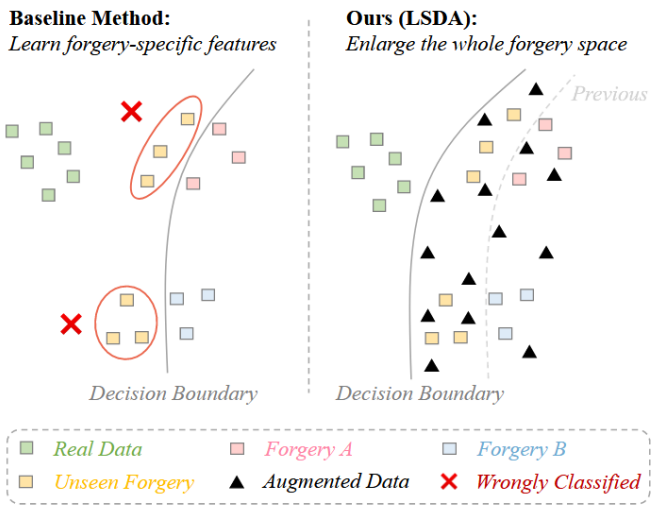
二、国内外研究成果调研

隐空间增强

Transcending Forgery Specificity with Latent Space Augmentation for Generalizable Deepfake Detection (CVPR 2024)

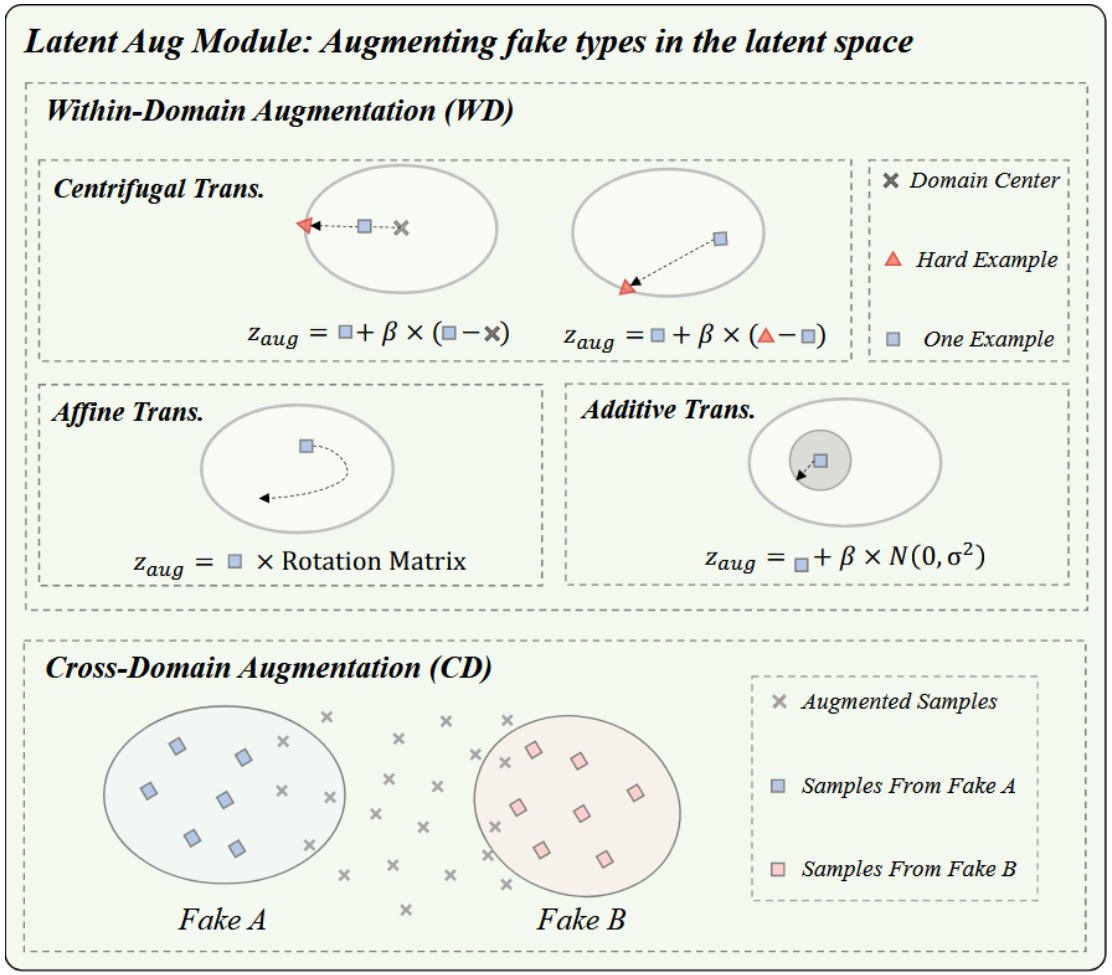
现有baseline可能过拟合特定伪造的特征因此不能很好地泛化至未见过的伪造

本文提出的方法通过隐空间增强扩大伪造空间来防止模型在特定伪造方法上的过拟合



域内增强：离心变换、仿射变换和加性变换。

跨域增强：Mixup，通过将两张图像及其对应的标签进行线性组合，生成新的训练样本，从而提高模型的泛化能力。



二、国内外研究成果调研

局部关系学习

Local Relation Learning for Face Forgery Detection (AAAI 2021)

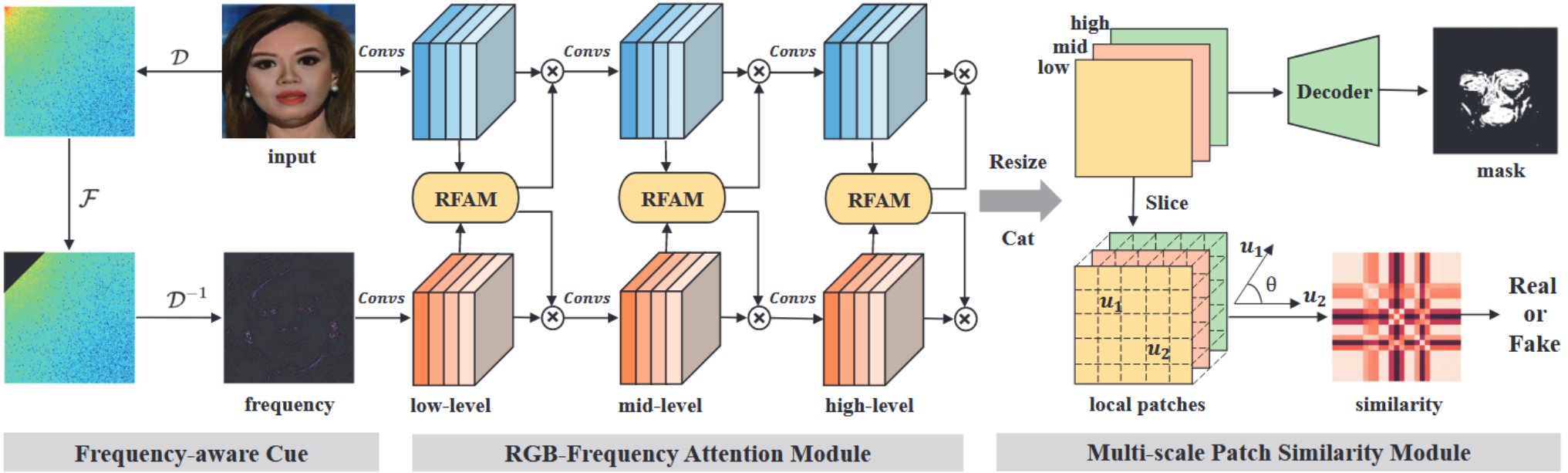
现有方法将问题建模为全局监督的二分类问题，没有考虑局部区域之间的相关性，全局监督不足以学习到一个泛化的特征，并且容易产生过拟合。

提出局部关系学习，用于有效捕获伪造痕迹，如异常纹理和高频噪声。

利用离散余弦变换将图像转换到频域，通过滤波器分离低中高频。

RGB-Frequency Attention Module（RFAM）模块在不同的语义层协同融合RGB和频域信息，以促进局部区域特征的学习。

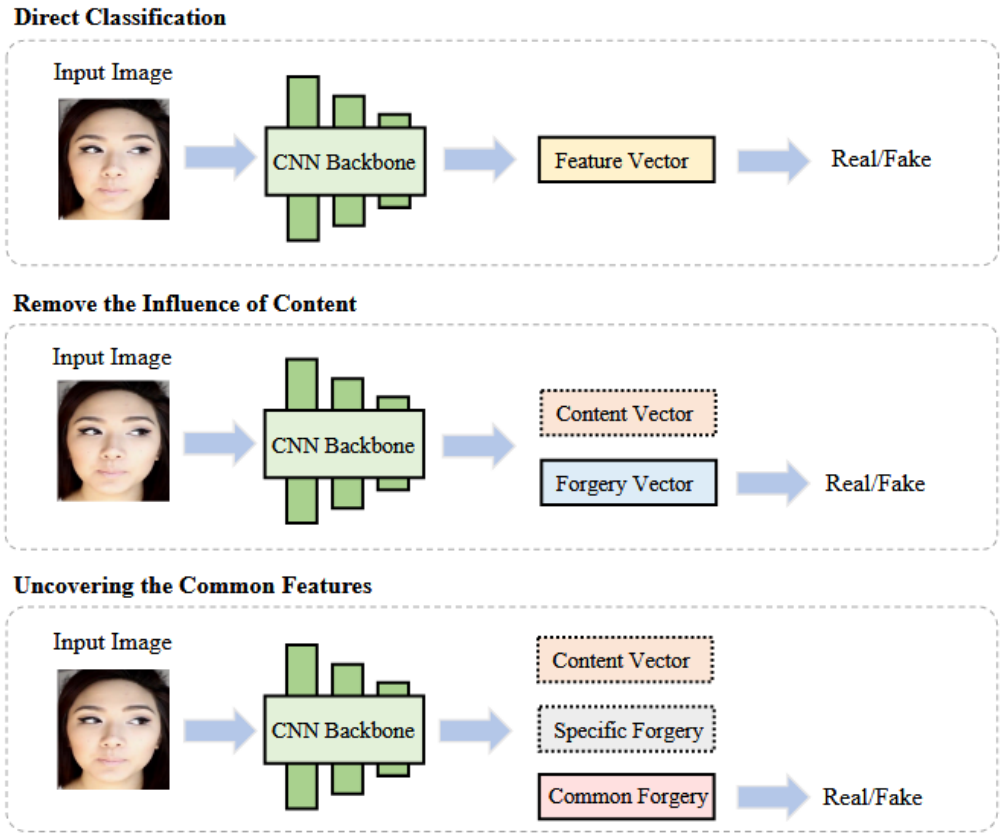
Multi-scale Patch Similarity Module（MPSM）融合多尺度特征，将feature转换为patch，再flatten为一维向量计算余弦相似度s，较低的s表明patch之间的差异较大，判定为伪造，反之真实。



二、国内外研究成果调研

解耦学习

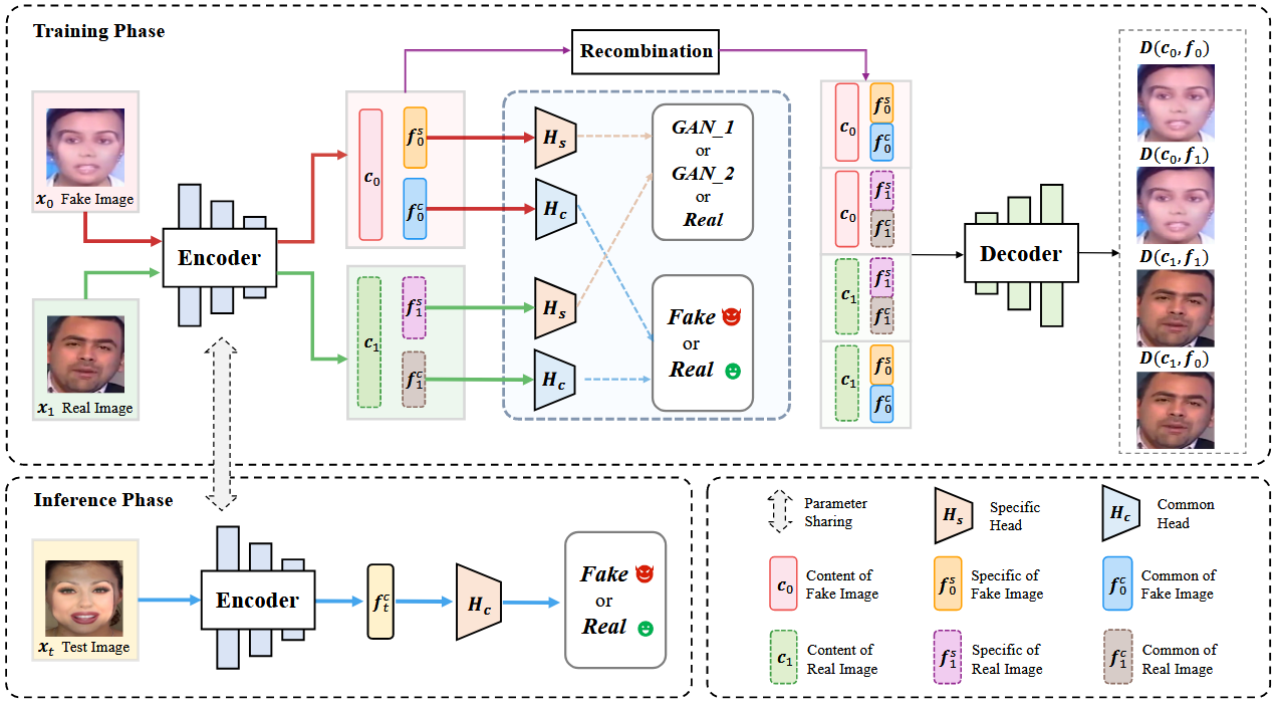
UCF: Uncovering Common Features for Generalizable Deepfake Detection (ICCV 2023)



认为现有检测方法泛化能力差是由于过拟合了伪造无关内容和特定伪造相关特征。

基于解耦学习，解耦学习是一种将复杂特征分解为更简单、更精确定义的变量的方法，并将它们编码为具有高判别力的独立维度。

通过多任务解耦框架将图像信息分解为三部分:与伪造无关的特征、特定伪造方法特征以及通用伪造特征。利用通用伪造特征进行分类能一定程度上缓解过拟合导致的泛化表现差的问题。



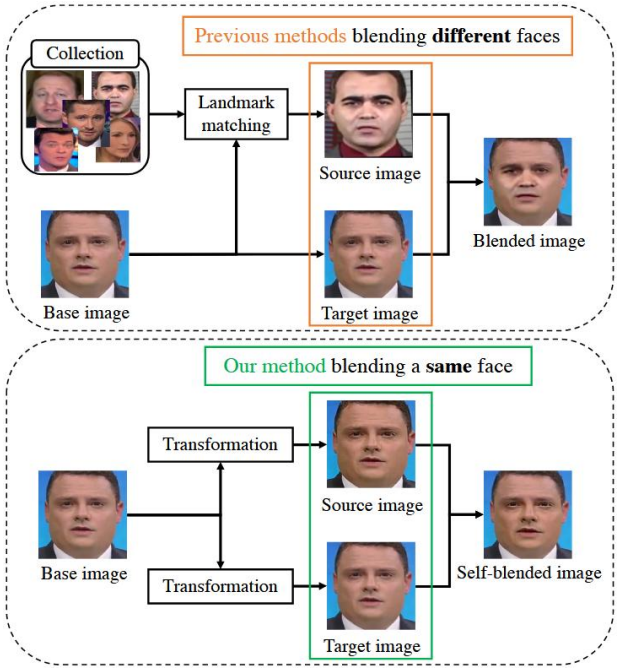
二、国内外研究成果调研

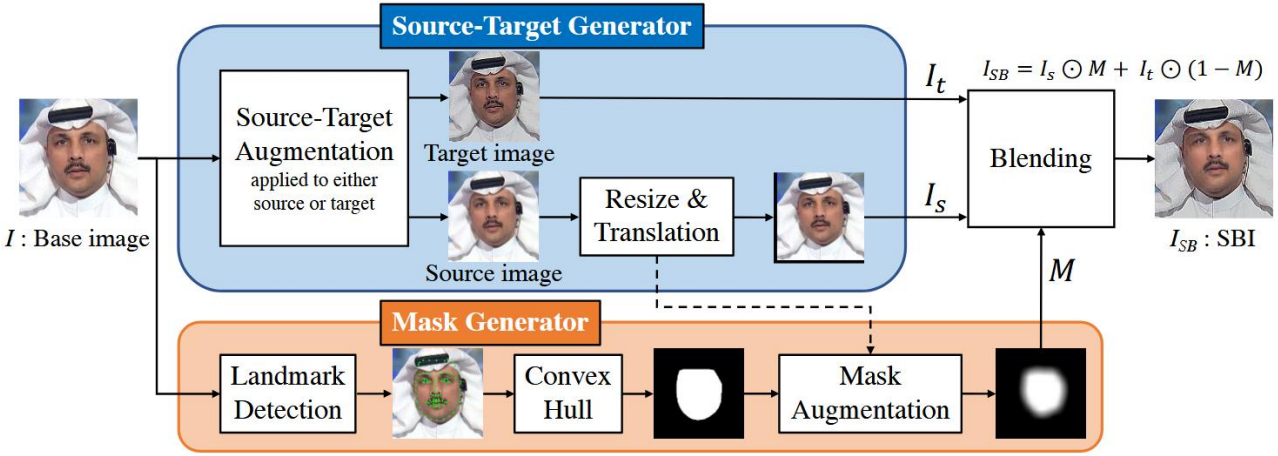
自混合监督

Detecting Deepfakes with Self-Blended Images (CVPR 2022)

利用图片自混合的方法替代传统训练方法，摆脱对伪造样本的依赖。

✸方法特点：只使用真实人脸训练，通过图片自混合产生更加难以识别的伪造样本，鼓励模型学习更加鲁棒的表示，同时避免过拟合特定的伪造方法





STG模块从单张真实图像生成一对伪源图像和伪目标图像，以产生统计不一致和空间错配。模块随机应用调整RGB通道值、色相、饱和度、亮度和对比度，图像降采样或锐化处理，以模拟伪造图像中常见的色彩偏差和频率异常。

通过蒙版按比例混合伪源图像和伪目标图像来生成最终图像。

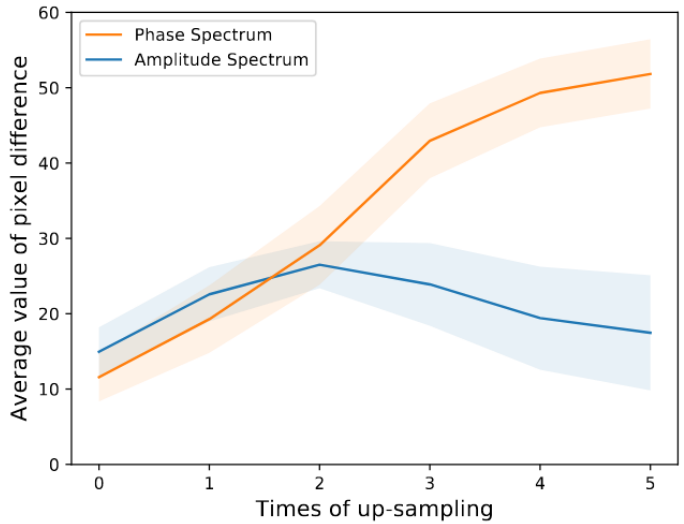
二、国内外研究成果调研

频域线索

Spatial-phase shallow learning: rethinking face forgery detection in frequency domain (CVPR 2021)

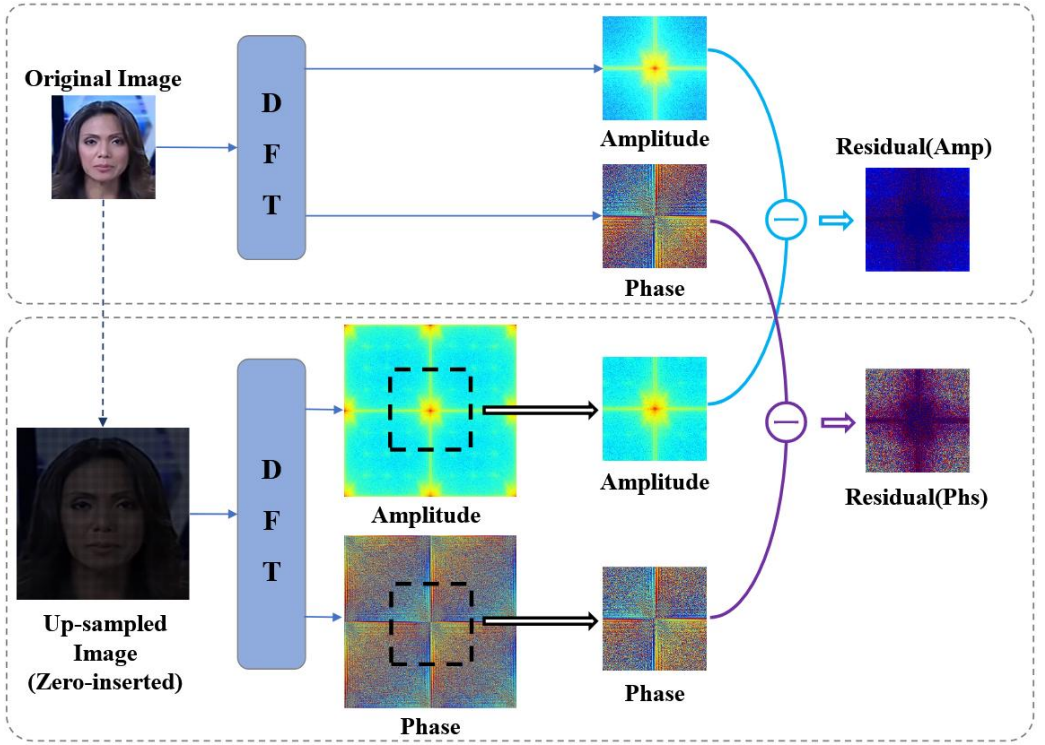
图像不仅在像素空间（即空间域）表现出特征，在频域中也包含独特的伪造线索。基于频域的检测器关注图像的频率分布，如低频（大致轮廓）和高频（细节信息）的异常，以应对传统空间域检测的局限性。

SPSL基于这样的一个观察：上采样是大多数人脸伪造技术的必要步骤，而累计上采样会在频域中，特别是相位谱中，导致显著变化。



SPSL还证明了相位谱对上采样操作的敏感度比幅度谱更高，使用相位谱能够更好地捕捉到因上采样操作所产生的伪影。为此SPSL结合空间图像和相位谱来捕捉人脸伪造的上采样伪影，从而提高伪造检测的泛化能力。

此外，对于人脸伪造检测任务而言，作者认为局部纹理信息比高层语义信息更重要。因为高层语义信息包含了原始和伪造人脸图像的许多共同特征。因此，SPSL还通过浅化网络来减小感受野，抑制高层特征，聚焦于局部区域。

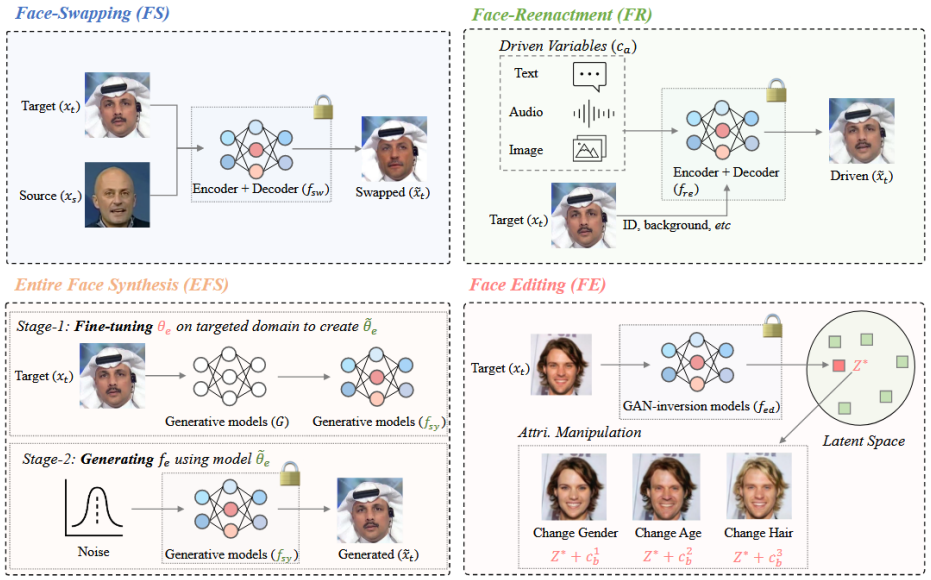


二、国内外研究成果调研

新一代深度伪造检测

DF40: Toward Next-Generation Deepfake Detection (NeurIPS 2024)

人脸伪造的方法非常多样大致可以分为face-swapping (FS), face-reenactment (FR), entire face synthesis (EFS), and face editing (FE)四大类，现存的许多检测器只能应对训练集中见过的伪造方法，而难以泛化至未见过的伪造方法。这些检测器在训练集上普遍存在过拟合伪造不相关特征的现象，严重影响模型的泛化能力。



文章提出了超大型数据集，包含了40种伪造方法，弥补了旧数据集的伪造方法过时的缺陷。

文章发现语言-视觉跨模态模型CLIP在DF40的评估拥有最好的表现，强于现有的为深度伪造检测专门设计的网络。

文章提出了5个开放性的问题。

1. BlendFake的优化使用：如何发掘BlendFake在深伪检测训练中的真正潜力？能否将BlendFake和Deepfake数据结合以提升训练效果，并且有效应对哪些类型的伪造？

2. 增量学习框架的设计：随着深伪数据多样性的增加，如何设计一个增量学习框架来有效地学习多种不同的伪造？特别是如何在该框架中处理真实数据？

3. 基于类别的深伪分类：由于现有技术难以涵盖所有伪造方法，是否可以基于伪造特征（如面部交换FS、面部重现FR等）对深伪进行分类？是否能定义一个“度量标准”将不同伪造方法归类，以便共同学习伪造特征？

4. 扩展CLIP-large模型的应用：CLIP-large在图片检测方面表现优异，能否进一步扩展该模型用于深伪视频检测？

5. 利用多模态技术提升检测效果：是否可以运用当前多模态技术来开发更好的真实面部特征表示，从而提升深伪检测的效果

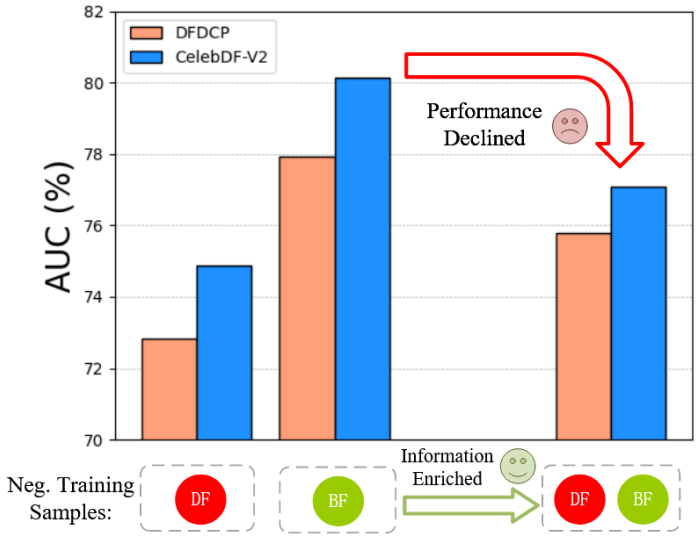
二、国内外研究成果调研

混合训练

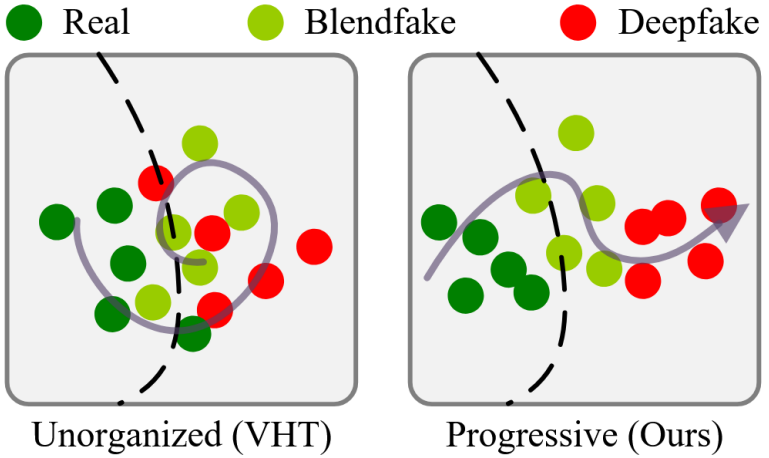
Can We Leave Deepfake Data Behind in Training Deepfake Detector? (NeurIPS 2024)

针对DF40中提出问题：如何发掘BlendFake在深伪检测训练中的真正潜力？能否将BlendFake和DeepFake数据结合以提升训练效果？

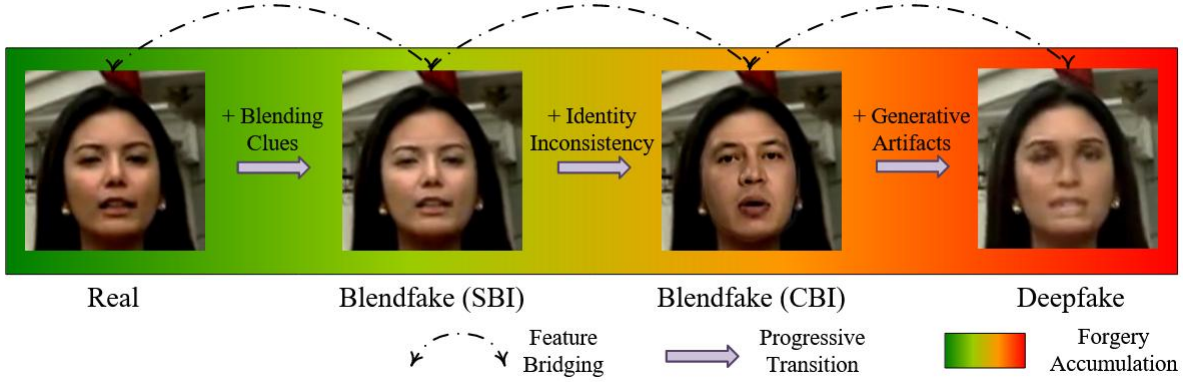
文章发现简单地将BlendFake和DeepFake进行混合训练会会产生反直觉的检测表现下降问题。



作者将这种反直觉的检测表现下降问题归因于隐空间缺乏组织，精心组织的隐空间分布被证明有利于网络性能



通过三元组二分类训练策略构建特征桥，模拟真实到混合伪造到深度伪造的渐进转变过程



汇报目录

汇报目录

三、代表性工作复现进展

深度伪造检测平台

DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection (NeurIPS 2023)

- 统一平台：DeepfakeBench 提供了首个全面的深度伪造检测基准，解决了该领域缺乏标准化和一致性的问题。
- 数据管理：DeepfakeBench 提供统一的数据管理系统，确保所有检测模型使用一致的输入。
- 集成框架：DeepfakeBench 提供了一个集成框架，用于实现最先进的检测方法。
- 标准化评估：DeepfakeBench 引入了标准化的评估指标和协议，增强性能评估的透明度和可重复性。

数据集名	备注
Celeb-DF-v1	-
Celeb-DF-v2	-
FaceForensics++, DeepfakeDetection, FaceShifter	c23版本
UADFV	-
Deepfake Detection Challenge (Preview)	-
Deepfake Detection Challenge	测试集

三、代表性工作复现进展

局部关系学习

Local Relation Learning for Face Forgery Detection (AAAI 2021)

DeepfakeBench中实现了lrl_detector代码，但运行测试出错。我已解决并提交了Github issue：
https://github.com/SCLBD/DeepfakeBench/issues/108

DeepfakeBench的测试框架存在内存管理方面存在问题，在大型数据集上测试时会导致内存爆炸。我已解决并提交了Github issue：
https://github.com/SCLBD/DeepfakeBench/issues/109

根据DeepfakeBench标准的跨域泛化评估协议，我在FF++上对模型进行训练，并在其余数据集上对模型进行了测试。

	ACC	AUC	ERR	AP
Celeb-DF-v1	0.726	0.794	0.283	0.872
Celeb-DF-v2	0.714	0.758	0.319	0.846
DFDCP	0.655	0.691	0.362	0.798
DFD	0.815	0.830	0.253	0.976
DFDC	0.622	0.677	0.377	0.696
UADFV	0.770	0.904	0.177	0.904

DeepfakeBench还支持视频级AUC指标，对于图像深度伪造检测模型，它利用同一个视频的不同帧的预测平均值进行计算，测试所得视频级AUC如表格所示。

数据集	Celeb-DF-v1	Celeb-DF-v2	DFDCP	DFD	DFDC	UADFV
视频级AUC	0.841	0.828	0.721	0.871	0.698	0.956

三、代表性工作复现进展

解耦学习

UCF: Uncovering Common Features for Generalizable Deepfake Detection (ICCV 2023)

DeepfakeBench复现ucf_detector，在FF++上进行训练并测试

	ACC	AUC	ERR	AP
Celeb-DF-v1	0.727	0.811	0.271	0.887
Celeb-DF-v2	0.649	0.772	0.303	0.861
DFDCP	0.595	0.693	0.358	0.792
DFD	0.756	0.821	0.258	0.974
DFDC	0.658	0.731	0.336	0.751
UADFV	0.818	0.920	0.164	0.925

视频级AUC指标

数据集	Celeb-DF-v1	Celeb-DF-v2	DFDCP	DFD	DFDC	UADFV
视频级AUC	0.861	0.837	0.706	0.867	0.751	0.955

三、代表性工作复现进展

自混合监督

Detecting Deepfakes with Self-Blended Images
(CVPR 2022)

DeepfakeBench复现sbi_detector，在FF++上进行训练并测试

	ACC	AUC	ERR	AP
Celeb-DF-v1	0.597	0.705	0.354	0.800
Celeb-DF-v2	0.587	0.756	0.310	0.845
DFDCP	0.542	0.721	0.332	0.826
DFD	0.583	0.778	0.302	0.969
Fsh	0.534	0.695	0.357	0.661
UADFV	0.818	0.920	0.164	0.925

视频级AUC指标

数据集	Celeb-DF-v1	Celeb-DF-v2	DFDCP	DFD	Fsh	UADFV
视频级AUC	0.756	0.818	0.781	0.830	0.720	0.984

汇报目录

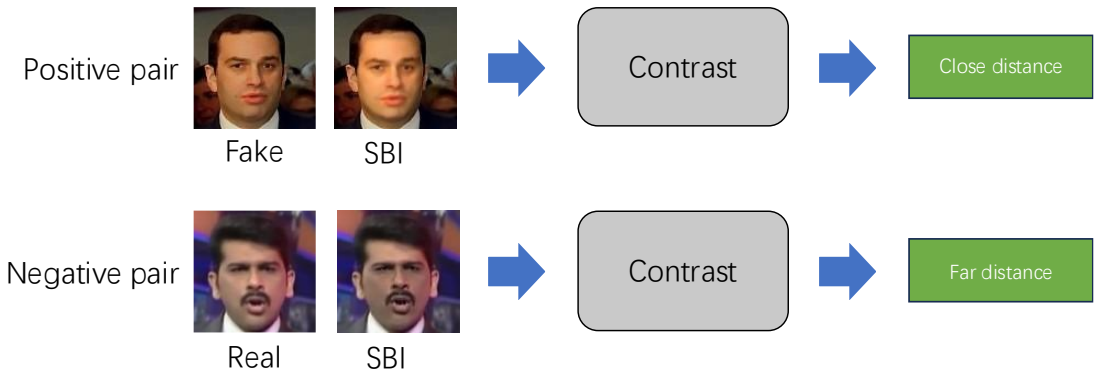
汇报目录

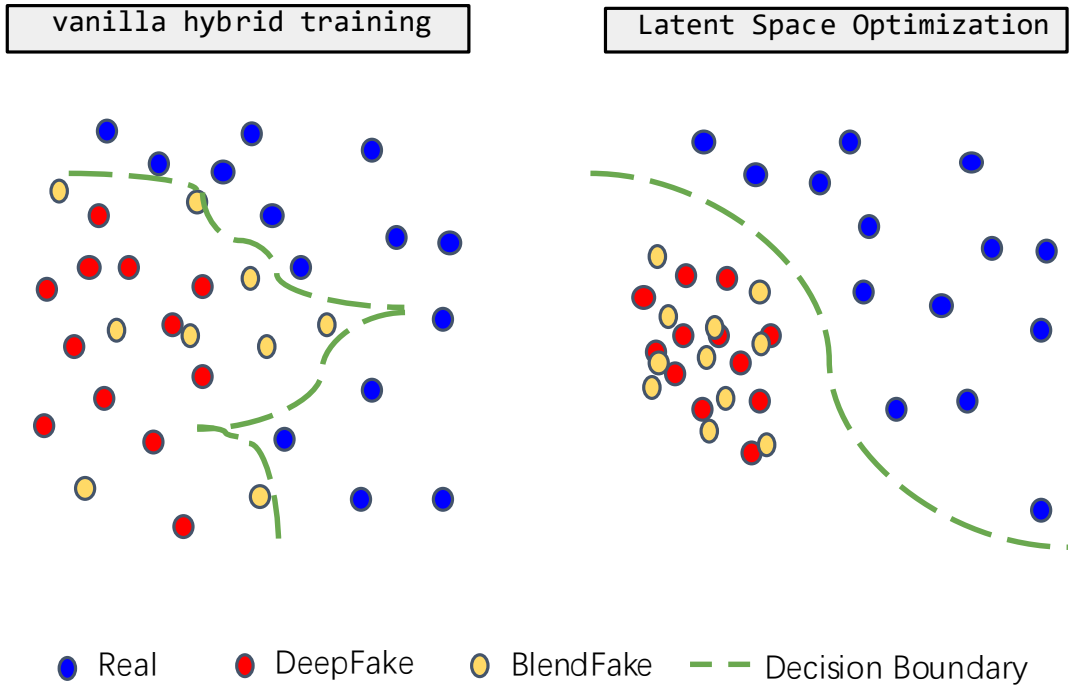
四、创新目标和初步思路

混合训练

新的基础视觉模型作为骨干网络能更好地编码特征，选取对比学习领域最强的模型DINOv2作为骨干网络。

自混合到深度伪造的渐进过程的组织是没有必要的，通过拉远隐空间中真实样本与自混合样本的距离、拉近自混合样本和深度伪造样本的距离同样能够很好地组织隐空间分布。





四、创新目标和初步思路

混合训练

对比损失L_{con}，输入N个图像$X = \{x_0, x_1, \dots, x_{n-1}\}$，图像对应的标签$T = \{t_0, t_1, \dots, t_{n-1}\}$，$t_i = 0$表示真实，$t_i = 1$表示伪造，对输入图像作自混合变换得到的混合伪造图像记为$S = \{s_0, s_1, \dots, s_{n-1}\}$，当$t_i = 0$时$x_i$与$s_i$不相似，当$t_i = 1$时$x_i$与$s_i$相似。$F$表示编码器，通过编码器能够得到输入数据的隐空间特征表示，α为距离阈值超参数。

$$L_{con} = \frac{1}{N}\sum_{i=0}^{N-1}\left[t_i \|F(x_i) - F(s_i)\|_2 + (1 - t_i)\max(0, \alpha - \|F(x_i) - F(s_i)\|_2)\right]$$

通过对辅助对比损失的优化可以在隐空间内同时拉远负样本对距离（真实和伪造）、拉近正样本对距离（BlendFake和DeepFake）。

四、创新目标和初步思路

混合训练

基于DeepfakeBench标准评估协议的评估结果。

	FF++	FF-F2F	FF-DF	FF-FS	FF-NT	FSh	DFD	CDFv1	CDFv2	DFDC	DFDCP	UADFV	Avg.
Ours	0.961	0.971	0.980	0.980	0.912	0.699	0.813	0.847	0.846	0.759	0.757	0.880	0.867

其他待补充实验：

跨伪造方法评估（ff++域内3类训练1类评估）
o/w FF-DF
o/w FF-FS
o/w FF-F2F
o/w FF-NT

T-SNE可视化分析

Method	Venues	CDFv1	CDFv2	DFDC	DFDCP	C-Avg.
LRL	AAAI'21	0.794	0.757	0.676	0.691	0.730
SBI	CVPR'22	0.705	0.756	0.714	0.720	0.654
UCF	ICCV'23	0.779	0.753	0.719	**0.759**	0.753
Ours		**0.847**	**0.846**	**0.759**	0.757	**0.802**

参考文献

[1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner. FaceForensics++: Learning to Detect Manipulated Facial Images[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 1-11.

[2] Yan, Zhiyuan and Luo, Yuhao and Lyu, Siwei and Liu, Qingshan and Wu, Baoyuan. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 8984-8994.

[3] Chen, Shen and Yao, Taiping and Chen, Yang and Ding, Shouhong and Li, Jilin and Ji, Rongrong. Local relation learning for face forgery detection[C]// Proceedings of the AAAI conference on artificial intelligence. Menlo Park, CA: AAAI, 2021: 1081-1088.

[4] Yan, Zhiyuan and Zhang, Yong and Fan, Yanbo and Wu, Baoyuan. UCF: Uncovering Common Features for Generalizable Deepfake Detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2023: 22412-22423.

[5] Liu, Honggu and Li, Xiaodan and Zhou, Wenbo and Chen, Yuefeng and He, Yuan and Xue, Hui and Zhang, Weiming and Yu, Nenghai. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 772-781.

[6] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, Baoyuan Wu. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection[C]// Advances in Neural Information Processing Systems. La Jolla, California: MIT Press, 2023, 36: 4534–4565.

[7] Shiohara, Kaede and Yamasaki, Toshihiko. Detecting deepfakes with self-blended images[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 18720-18729

[8] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, Li Yuan. DF40: Toward Next-Generation Deepfake Detection[C]// Advances in Neural Information Processing Systems. La Jolla, California: MIT Press, 2024.

[9] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuanhao Luo, Zhongyuan Wang, Chen Li. Can We Leave Deepfake Data Behind in Training Deepfake Detector?[C]// Advances in Neural Information Processing Systems. La Jolla, California: MIT Press, 2024.

[10] Mathilde Caron, Julien Mairal, Hugo Touvron, Ishan Misra, Piotr Bojanowski, Hervé Jegou and Armand Joulin. Emerging properties in self-supervised vision transformers[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 9650-9660.

[11] Huy V. Vo, Marc Szafraniec, Vasil Khalidov et al. DINOv2: Learning Robust Visual Features without Supervision[J]. Transactions on Machine Learning Research, 2023.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy et al. Learning transferable visual models from natural language supervision[C]// Proceedings of the International Conference on Machine Learning. New York: ACM, 2021: 8748-8763.

[13] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 3204-3213.

[14] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures[J]. Transactions on Graphics, 2019, 38(4): 1–12.

[15] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2019: 46-52.

汇报目录

汇报目录

五、下学期工作开展计划

进度安排

进 度	具 体 安 排
阶段1	2月-3月：完成第一个创新点论文中文初稿的编写
阶段2	3月-4月：完成第一篇论文的所有工作
阶段3	4月后：构思第二个创新点，暂定为解决深度伪造领域训练过程中的样本极度不均衡问题

2025.1.16