基于小波预提取特征的深度虚假人脸识别

郭嘉

摘要：

当前的深度学习模型已经能够生成以假乱真的人脸，在政治、商业、个人隐私等方面产生了很多安全隐患，因此设计能够自动判别出人脸图像是否由深度学习模型生成的模型成为了多媒体内容安全中一个重要的问题。本文利用小波变换使得高清人脸图像在降低分辨率后仍能有高频细节特征的辅助，并将变换结果输入 Swin transformer，希望通过大范围的关系建模弥补当前一些基于 CNN 的模型不能利用大范围特征关系捕获 GAN 的指纹特征的缺点。实验表明，利用 1 万张 FFHQ 数据集图片作为真实样本，1 万张 StyleGAN2 实验作为假脸样本的条件下，本文提出的模型能在预训练模型上微调 17 个 epoch 后在 12 万张测试样本上达到 99.998%的 AUC。(论文代码基于 Swin transformer 开源仓库。在本文附件 faceformer 文件夹的 Readme 第一部分说明了本文对其增加和修改的部分)

一、背景

由于对抗生成网络 GAN[1]的产生，以及深度学习的快速发展，目前基于深度学习的生成器[2][3][4]已经能产生非常逼真，且令人很难分辨的虚假人脸图像——虽然世界上并没有这个人（见图一，左边的图像是 StyleGAN2[4]生成的虚假人脸）。尽管这项技术在游戏制作、娱乐、艺术创作方面都有非常好的应用价值，但是这种技术引发的一系列安全问题也使人们担心并投入更加多的重视，虚假的人脸在进入社交网络后能助于散布不实的信息，对舆论造成影响，在政治、商业、个人隐私方面都能产生巨大的安全隐患。因此设计能够自动识别出图像中的人脸究竟是 GAN 生成的还是真的人脸的算法是当前多媒体内容安全中一个非常重要的课题。对于该问题，部分研究者采用手工特征提取的方法[10][11][12]，但由于 GAN 生成的图片质量越来越高，基于深度学习的虚假人脸判别方法逐渐得到人们的重视[8][13][14]。

然而目前基于深度学习的方法基本都使用卷积[8][13][14]，而一些工作[10][11]指出 GAN 生成的虚假图像的指纹信息是全局性的，因此使用卷积没办法快速融合全局信息作出判断。相反，利用 transformer[15][16][17]能快速融合全局信息，因此本文推测 transformer 能更好的辨别出虚假人脸。然而使用预训练 transformer 模型需要对图片进行降采样操作，会丢失高频信息，且 transformer 对于捕获局部信息的能力有可能不及卷积网络，因此本文利用小波变换提前提取高频分量，再对图片进行降采样，并将降采样后的图片与小波分量进行通道合并再使用 swin-transformer[16]预训练模型。本文实验证明，在仅使用 10000 张真人脸，10000 张假人脸的情况下，该方法能在 60000 张真人脸，60000 张假人脸

的测试集上达到 99.828%的准确率和 99.998%的 AUC。其中真人脸为 FFHQ 数据集中的 1024x1024 分辨率的 70000 张人脸，假人脸为利用 StyleGAN2[4]在 FFHQ 上训练后，生成的 70000 张 1024x1024 分辨率的虚假人脸。

二、相关技术介绍及其与本文关系

2.1 通过 GAN 生成虚假人脸

GAN[1]利用生成器与辨别器的对抗学习，通过大量数据训练能够得到一个从随机噪声向量产生图片的生成器。ProGAN[2]在此基础上，通过同时逐渐扩大生成器与辨别器的技巧加快了训练速度与训练的稳定性，使得 ProGAN 能够以假乱真的人脸图像。StyleGAN[3]将人脸生成技巧进一步往前推进，它们设计的学习过程能够在无监督条件下分离人脸的高层次特征和一些特征的随机性变化。StyleGAN2[4]重新设计了生成器的规范化操作，对生成器和辨别器逐渐生长的操作也进行了修改，使得生成人脸的质量进一步提升。

图一 你能分清上面两张图像哪张是由 StyleGAN2 生成的么？

2.2 虚假人脸数据集

为了促进对虚假人脸取证等其他任务的研究，研究人员利用上文所述的 ProGAN[2]和 StyleGAN[3]生成了一些虚假人脸数据集。100K-Generated-Images[3]数据集是 StyleGAN[3]在 FFHQ[3]数据集上训练后产生的 100000 张虚假人脸。100K-Faces [7]数据集同样利用 StyleGAN[3]生成人脸，但他们的训练数据的背景有所不同，这使得生成人脸的背景中没有出现能令人辨识出的异常。Dang 等人[17]利用 100000 张 ProGAN 生成的人脸与 200000 张 StyleGAN 生成的人脸组成 DFFD 数据集。iFakeFaceDB[16] 由 80000 张 ProGAN 生成的人脸与 250000 张 StyleGAN 生成的人脸组成，但他们对这些生成的人脸移出了原本 GAN 的一些指纹信息，这使得已有的一些虚假人脸取证模型很难再利用这些 GAN 的指纹信息进行取证，对取证模型进一步改进提出了更高的要求。

考虑到这些工作没有利用到最新的人脸生成技术，本文利用 StyleGAN2[4]生

成 70000 张图片作为虚假图片作为数据集的一部分，实验证明本文应用的方法能非常容易鉴别利用到最新的人脸生成技术产生的虚假人脸。

2.3 虚假人脸检测

虚假人脸检测的方法主要包括两大类，一类基于手工提取特征，另一类基于深度学习。在第一大类方法中，McCloskey 等人 [10]认为 GAN 生成的人脸由于最后一层为 1x1 的卷积，其卷积核有可能会使生成的 RGB 颜色强度相关性较强，而相机生成的照片则不会这样，并且卷积网络有一些规范化操作，从而不会生成过曝或者欠曝的区域，基于这两个发现，他们提取了相关的特征送入 SVM 进行真假分类。Wang 等人[11]认为将人脸送入人脸识别网络后每层神经元的激活信息都能一定程度上反应人脸的真假，因此他们把网络每层神经元激活数作为特征送入 SVM 分类。Guarnera[12]等人利用 EM 算法提取出人脸图像特征，并将提取出的特征送入 KNN,LDA,SVM 进行最终检测。在基于深度学习的方法中，Yu[13]等人利用卷积网络提取出图像的指纹特征进行分类，Marra[14]等人利用增量学习的方法让深度网络能辨别不同结构的 GAN 生成虚假图像，使其有更强的鲁棒性。Dang[8]等人利用注意力机制使网络注意到对判别真假更具有效果的脸部部位。

目前暂时没有利用 Transformer 分析 GAN 的指纹信息从而判别出虚假人脸的工作，本文考虑到从模型的浅层就融合全局信息对指纹信息的提取很有用处，因此本文作出尝试。

2.4 Transformer

Transformer[19]在图像分类，分割和检测方面取得了令人瞩目的进展[15][16][17]，transformer 使得神经网络能尽早地融合全局信息，而卷积网络只能逐步融合局部信息。本文猜测 GAN 生成的虚假图像的指纹特征是全局性的，因此本文利用[16]来进行虚假图像检测。

2.5 离散小波变换

离散小波变换在图像处理中有非常重要的应用[18]，本文利用小波变换主要是为了降采样适应 Transformer 预训练模型的同时不丢失人脸的细节信息。

三、技术方案

3.1 模型概述

我们的推测是目前基于 CNN 的方法无法尽快融合全局的信息，我们认为使用 transformer 模型[19]的全局注意力机制能使模型在浅层网络处就能建模获取 GAN 的指纹信息。为了提高模型的计算速度我们采用 Swin transformer[16]。为了使用 Swin transformer 预训练模型进行虚假人脸判别，最简单的方法是直接对输入图像降采样，并将最后一层全连接层输出维数改为 2 维，下面在 3.2 节说

明这种方式，然而这种方式会导致细节信息的丢失，在 3.3 节我们说明采用更大的分块能保持原图的分辨率输入，然而这导致模型参数和计算量的大幅增加，在 3.4 节我们将指出对原图像进行离散小波变换进行预处理既能保留原有图像高频信息且对计算量和参数量增加不大。（为了提高文章的易读性，后面尽量不使用符号而使用具体的数字进行模型的叙述，这样有两个好处，一是避免使用符号时需要阐述符号含义导致文章比较冗长，二是避免了后文对这些参数取值的说明从而提高文章简洁性。）3.5 节将指出舍弃数据增强的必要性 3.6 节说明我们如何获取训练和测试所需的数据集。

3.2 Swin transformer 及使用预训练模型进行虚假人脸识别需要进行的调整

Swin transformer[16]是基于[15]改进的视觉模型，通过局部的自注意力机制以及高效的滑窗融合相邻特征块信息。使用 Swin transformer 进行虚假人脸识别最直接的做法是将模型最后的全连接层的输出维数改为 2 维，其他层的参数使用其预训练模型的参数。为了使用其预训练模型，我们必须对高清的人脸输入（1024x1024）降采样至（224x224），如图 2。这种方式会使人脸的高频细节特征丢失，而 GAN 的指纹信息很有可能包含这些高频的部分，因此这种方式可能会影响最后分类的准确率，后面的实验也将验证这点，因此我们需要在能够使用预训练模型的情况下考虑如何保留这些高频特征。本节所述的方式在 20000 样本的训练集上训练并在 120000 样本的测试集上测试已经能够达到 99.520%的分类准确率，下面几节将通过考虑保留这些细节信息从而进一步提升模型表现。

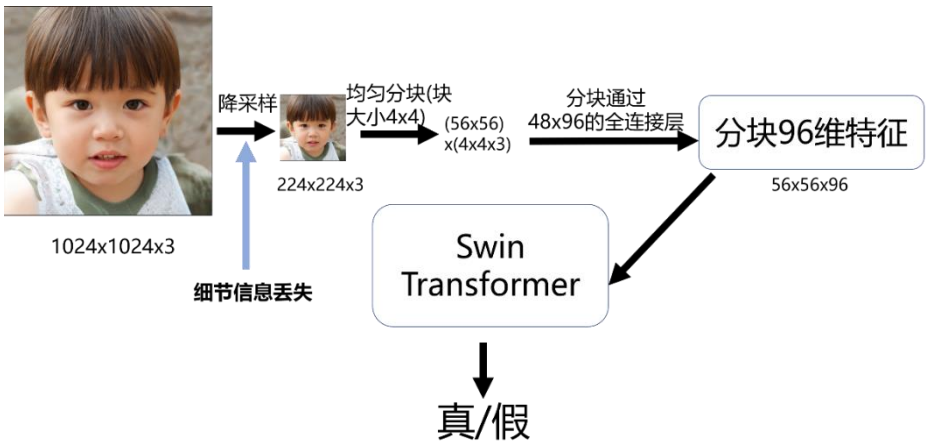


图 2 直接降采样利用预训练模型，导致细节信息丢失

3.3 大分块策略

为了使 1024x1024 的图像对应到 56x56 个图像块，并且不损失高清人脸的任何高频特征，可以选择一个能够整除 56 且比 1024 大的整数 N，把 1024x1024 的图像升采样到 NxN，这里我们选择 N=1120，并且把分块大小从原预训练模型对应的 4x4，变成(4x(1120/224))x(4x(1120/224)),这里将分块边长扩大 5 倍的原

因是输入图像的边长也变成了原预训练模型的 5 倍，如图 3 所示。原模型的 4x4 分块将对应到一个 16 维的特征，由于有 R,G,B 三个通道，每个分块将会有 48=16x3 维的特征，通过全连接层将该 48 维的块语义信息映射为 96 维语义信息。如果采取大分块策略，每个分块将对应 48x5x5=1200 维的语义特征，原模型的的块特征提取模块将会需要将近 25 倍的参数，这大大增加了计算负担。由于这种大分块策略会使训练时间增加到一个难以承受的地步——一个 epoch 的训练在 4 张 2080ti 的机器上都要花上 1 天，因此后面实验部分将会放弃这种做法。

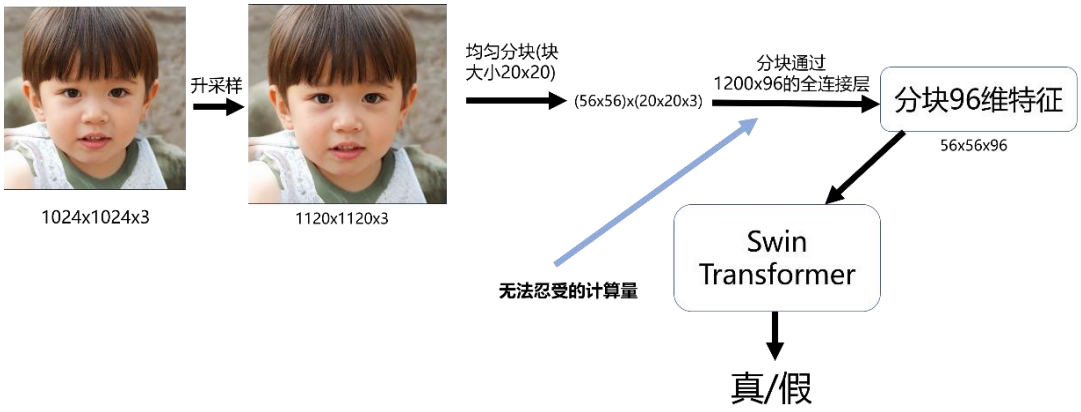


图 3 升采样，并将初始块特征全连接层输入维数增大至原来的 25 倍，计算量过大

3.4 通过离散小波变换保留细节特征

如图 5 所示，对于 1024x1024x3 的高清人脸图像，我们分别对每个通道进行离散小波变换，这里使用 Haar 小波。这里我们保留了 2 级的细节，这是基于以下考虑，首先保留更多级的细节会增加预处理的计算量，而且会增加块特征提取的通道数，如果保留 3 级细节则需要 10x3x16=480 个通道，每多一级细节特征就需要增加 3x3x16=256 个通道，如果我们保留 6 级细节，那么计算量将会和上节所提出的方式几乎持平。第二是 2 级的细节对应 256x256 的大小，与降采样后 224x224 的大小相类似。这样 RBG 每个通道将会对应 7 个通道的特征，于是输入通道数为 7x3=21，如图 4 所示。

也许读者会认为小波变换也会增加计算量，其实并不是如此，在训练阶段我们把小波变换预处理的结果存储起来，训练时直接使用。在推理时会有一些计算量的需求，但这比上节提出模型所需的计算时间要少太多。

3.5 不使用数据增强

图像分类常使用颜色调整，随机块去除，加入噪声等数据增强方式，这种考虑是为了增强模型的鲁棒性，且基于在这些数据增强后图像对应类别不会改变，

然而对于虚假人脸识别，这些操作将可能会把原来应该被模型认定为真的人脸，

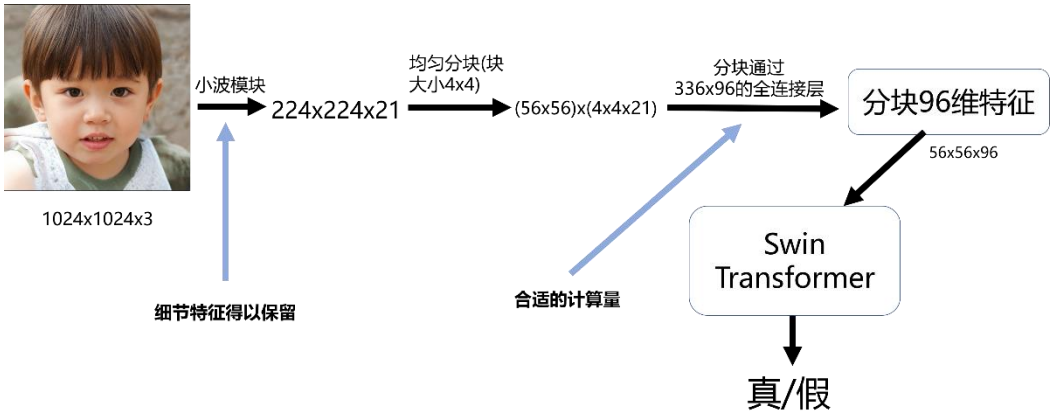


图 4 使用离散小波变换提取至 2 级细节，全连接层输入维数增大至原来的 7 倍

变为假的人脸，因为经过数据增强后这确实不应该是真实人脸对应的图像。

3.6 数据集生成

为了体现以上方案的优势，我们使用目前生成人脸效果逼真的 Stylegan2[4]在 FFHQ 数据集[3]上训练生成 70000 张 1024x1024 分辨率的虚假人脸，真实人脸使用 FFHQ 数据集[3]70000 张 1024x1024 分辨率的人脸。

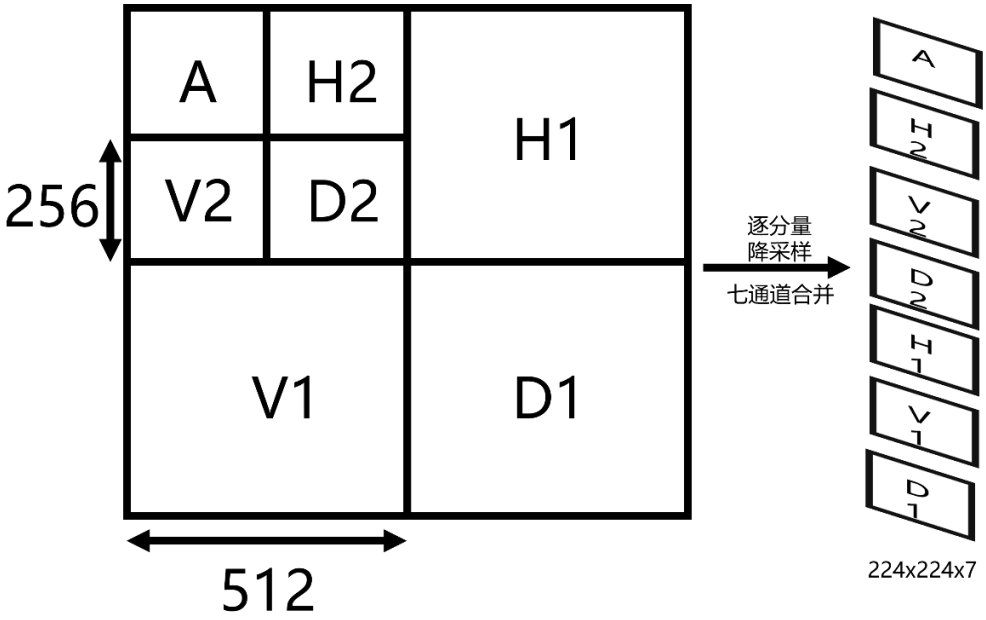


图 5 对 1024x1024 的图像做离散小波变换,提取至 2 级细节，并将所有分量降采样至 224x224

四、实验

4.1 模型能利用少量训练样本获取 StyleGAN2 的指纹特征

（后文中训练样本和测试样本中真假人脸各占 50%。）表一展示了训练样本数与模型表现的关系。我们使用 Swin transformer tiny 的预训练模型，在训练集上进行微调。从表一可以看出当训练样本只有 20000 的时候，仅仅训练 6 个 epochs 也能在 120000 样本的测试集上达到 99.905%的 AUC。这说明 StyleGAN2 的指纹特征很容易被 Swin transformer 捕获。而基于小波变换的模型表现在短短的 6 个 epochs 训练后不及直接降采样的表现，我们推测这是因为基于小波变换的输入有更多的参数量，且与预训练模型所预期的输入模式不太匹配导致的，但 2.2 节将会说明经过足够 epochs 的训练，基于小波预处理的模型将有更高的准确率。

训练样本数	测试样本数	准确率	AUC	epoch	输入策略
64000	76000	99.337%	99.985%	6	直接降采样
20000	120000	98.286%	99.905%	6	直接降采样
20000	120000	97.054%	99.763%	6	小波变换

表一

4.2 小波预处理的优势

从表 2 中可以看出经过小波预处理的模型经过足够的训练，从准确率和 AUC 方面都比直接降采样的模型高。这说明 StyleGAN2 的指纹特征有一些高频的分量，但成分可能不太多，小波变换保留的这些高频特征能使 Swin transformer 更好地辨别出这些指纹特征从而判断出虚假人脸。

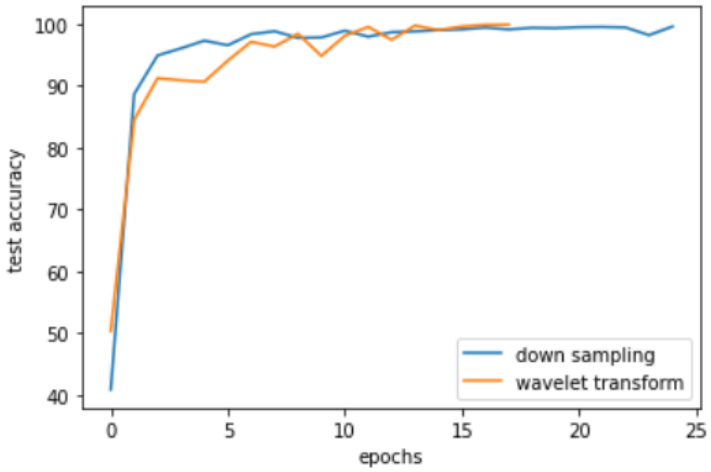
图六展示了训练的过程，预训练模型在降采样输入时准确率并不是 50%左右，而我们知道基于随机策略判别真假人脸大概是 50%的准确率，而基于小波变换的输入可以认为模型一开始满足这个条件，这符合我们的认知，因为降采样输入基于纯空间域，预训练模型对这些输入会有偏见，然而预训练模型开始很难辨认基于小波变换的输入，因此一开始准确率是 50%左右。

训练样本数	测试样本数	准确率	AUC	输入策略
20000	120000	99.520%	99.987%	直接降采样
20000	120000	99.828%	99.998%	小波变换

表二

4.3 训练和测试时间

图六中的所有数据利用 4 张 2080ti 训练 10 天可以获得。70000 张虚假人脸训练样本利用单张 2080ti 需要 2 天时间生成。



图六

五、结论

小波预处理能保留提取 StyleGAN2 指纹信息的高频分量，并且能高效利用 Swin transformer 预训练的模型对 StyleGAN2 生成的虚假人脸进行甄别，这表示甄别 StyleGAN2 的指纹信息可以通过大范围的注意力机制有效获取，假以部分的高频信息能够一定程度上提高这种指纹信息的获取准确度。

六、参考文献

[1] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[2]Karras T, Aila T, Laine S, et al. Progressive growing of gans for improved quality, stability, and variation[J]. arXiv preprint arXiv:1710.10196, 2017.

[3]Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4401-4410.

[4]Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8110-8119.

[5]Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild[C]//Proceedings of the IEEE international conference on computer vision. 2015: 3730-3738.

Yi D, Lei Z, Liao S, et al. Learning face representation from scratch[J]. arXiv preprint arXiv:1411.7923, 2014.

[6]Cao Q, Shen L, Xie W, et al. Vggface2: A dataset for recognising faces across

pose and age[C]//2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, 2018: 67-74.
[7]100,000 Faces Generated by AI, 2018. [Online]. Available: https: //generated.photos/
[8]Dang H, Liu F, Stehouwer J, et al. On the detection of digital face manipulation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition. 2020: 5781-5790.
[9]Neves J C, Tolosana R, Vera-Rodriguez R, et al. Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(5): 1038-1048.
[10]McCloskey S, Albright M. Detecting gan-generated imagery using color cues[J]. arXiv preprint arXiv:1812.08247, 2018.
[11]Wang R, Juefei-Xu F, Ma L, et al. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces[J]. arXiv preprint arXiv:1909.06122, 2019.
[12]Guarnera L, Giudice O, Battiato S. Deepfake detection by analyzing convolutional traces[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 666-667.
[13]Yu N, Davis L, Fritz M. Attributing fake images to gans: Analyzing fingerprints in generated images[J]. arXiv preprint arXiv:1811.08180, 2018, 2.
[14]Marra F, Saltori C, Boato G, et al. Incremental learning for the detection and classification of gan-generated images[C]//2019 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2019: 1-6.
[15]Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
[16]Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[J]. arXiv preprint arXiv:2103.14030, 2021.
[17] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European Conference on Computer Vision. Springer, Cham, 2020: 213-229.
[18]Akansu A N, Haddad R A, Haddad P A, et al. Multiresolution signal decomposition: transforms, subbands, and wavelets[M]. Academic press, 2001.
[19]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.