transformer 在底层视觉应用上的发展综述

屠红艳[1] 管其杰[2]
TU Hongyan GUAN Qijie

摘 要 transformer，一种基于自注意机制的序列网络，首次被引入到自然语言处理范畴。由于其强大的长距离特征表示能力以及高性能和对视觉特定感应偏差的需求减少，基于 transformer 的模型被广泛应用于计算机视觉任务。在各种视觉基准测试中，基于 transformer 模型性能类似于或优于其他类型的网络，如卷积神经网络和递归神经网络。首先讨论基于 transformer 的不同底层视觉任务；然后分析了它们的优缺点；最后，总结了视觉 transformer 存在的问题，并提供了进一步的研究方向。

关键词 transformer；计算机视觉；深度学习；图像生成

doi：10.3969/j.issn.1672-9528.2024.02.011

0 引言

目前，相比于分割、分类和检测等高层视觉任务，基于 transformer 的底层视觉任务的内容相对较少，且更具挑战性。底层视觉任务通常将高分辨率或去噪等图像作为输出。transformer[1] 是由 Google 基于自注意机制提出的一种序列模型，目前已然成为自然语言处理领域（NLP）的主要架构。由于其强大的长距离特征提取能力，基于 transformer 框架的模型已被应用到计算机视觉任务。在各种视觉基准测试中，基于 transformer 的模型的性能类似于或优于其他类型的网络，如卷积神经网络和递归神经网络。由于 transformer 的高性能和对视觉特定感应偏差的需求减少，transformer 越来越受到计算机视觉界的关注。

在视觉应用中，一种模型结构是 CNN[2-3]，但 transformer 的强大性能使得未来它是 CNN 的潜在替代模型。Chen 等人 [4] 提出了基于 transformer 的序列方法来自动回归预测像素并应用在图像分类任务中，实验结果显示该模型取得了与 CNN 相当的性能。另一种应用于图像分类的主流视觉 transformer 模型是 Visual transformer（ViT）[5]，该模型将图像分割为若干个图像块并将其展平为一维序列作为模型的输入，它在多个图像识别基准上取得了一流的性能。除了视觉分类任务，其他各类视觉任务也开始大量使用基于 transformer 模型的方法，包括对象检测 [6-7]、语义分割 [8]、图像处理 [9] 和视频理解 [10]。

本文主要对 transformer 在底层视觉上的应用进行分类，并介绍各类底层视觉任务的相关算法。第 1 节介绍了视觉 transformer 的结构原理。第 2 节介绍了 transformer 模型在图像生成任务方面的应用发展。第 3 节介绍了 transformer 模型在超分辨率图像重建任务方面的应用。第 4 节介绍了 transformer 模型在图像增强任务方面的应用。第 5 节介绍了 transformer 模型在图像恢复任务方面的应用。第 6 节介绍了 transformer 模型在图像质量评估任务方面的应用。第 7 节总结并展望 transformer 存在的问题以及未来的研究方向。

1 视觉 transformer 的原理

Visual transformer（ViT）是一个直接应用于图像分类任务的 transformer 网络，它尽可能地遵循了 transformer 的原始设计。ViT 的框架模型流程为：首先，将输入的图像分割为若干个图像块（Patches）并将通过线性映射将每个图像块映射为一维向量，即 Token，在输入进编码器之前，对每个 Token 加入位置编码向量；然后，加入位置编码的 Token 输入到 L 个相同的 transformer 编码器；最后，将模型最后一个 transformer 编码器的 MLP 头部信息输出，即可获得分类结果，如图 1 所示。下面将从模型输入和 transformer 编码器简要概述 ViT 模型。

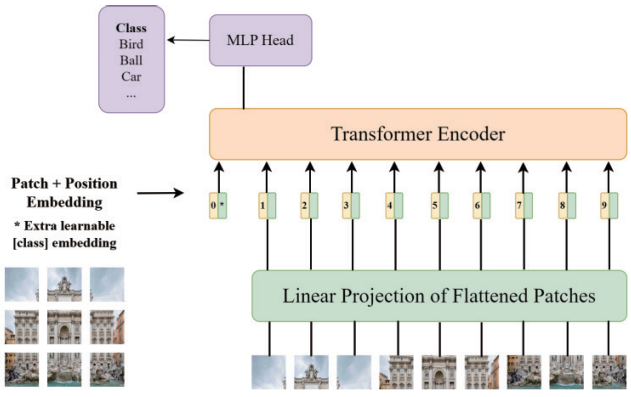

图 1 ViT 模型框架

1. 上海电力大学 上海 201306
2. 华东送变电工程有限公司 上海 201803

1.1 模型输入

与传统深度学习方法不同，ViT 网络将输入图片分为若干等同尺寸，以序列的形式输入到网络结构中。如图 1 所示，将原始输入图像$X \in R^{h\times w\times c}$展平为一个长度为 n 的一维分块序列，每一个小块图像（Patch）尺寸为$X_p \in R^{p\times p\times c}$，其中 h 和 w 分别表示原始输入图像的高和宽，p 表示分割后每个小块图像的尺寸，c 表示图像的通道数，n 表示原始输入图像被分割后的图像块数 $n=\frac{h\times w}{p^2}$。将得到的 n 个图像块经过一个线性映射层（linear project of flattened patches）后得到 n 个 Token embedding。与 BERT 的 [class] Token 类似 [11]，在嵌入 Patch 的序列增加一个额外的可学习的嵌入。这种嵌入的状态可以作为图像分类的表示。将得到的多个 Token embedding 和一个额外的 CLS Token 拼接（见图中的 * 表示），然后和位置编码（位置编码是一个可学习的参数，其维度和每个 Token 的维度相同）相加，构成完整的模块输入。

1.2 transformer encoder

transformer encoder 将含有位置编码的 Token 输入到 L 个相同的编码器中进行计算，每个编码器包含层归一化（layer normaliztion）、多头注意力（muti-head attention）、多层感知机（MLP）以及残差连接。其中多头注意力机制如图 2 虚线框部分，Q（Query）、K（Key）、V（Value）分别由同一个输入经过三个线性映射层得到。Scaled Dot-Product Attention 中的 Scale 代表缩放因子$\sqrt{d_k}$，h 表示有 h 组的 Q、K、V 计算的注意力机制，一组注意力头的计算公式为：

$$Attention(Q,K,V)=softmax(\frac{Q\cdot K^T}{\sqrt{d_k}})\cdot V \quad (1)$$

将 h 组的注意力值通过 concat 拼接，计算公式为：

$$MutiHeadAttention(Q,K,V)=Concat(head_1,head_2,...,head_h) \quad (2)$$

式中：$head_i$ 表示第 i（$1\le i\le h$）组注意力值。

最终，第 l 层（$1\le l\le L$）Encoder 计算流程为：

$$Z_l^{'}=MHSA\left(LN(Z_{l-1})\right)+Z_{l-1} \quad (3)$$

$$Z_l=MLP\left(LN(Z_l^{'})\right)+Z_l^{'} \quad (4)$$

式中：Z_l' 和 Z_l 分别表示第 l 层编码器的注意力模块和多层感知机的输出特征。

2 基于 transformer 图像生成方法

将 transformer 应用到图像生成任务中的一个简单而有效的方法是直接将架构从 CNN 转换为 transformer，如图 3（a）所示。Parmar 等人 [12] 提出了一种图像生成方法 Image transformer。仅将最邻近生成的像素作为解码器的输入，从实验结果中得出，Image transformer 的性能可以和基于 CNN 的图像生成模型的性能相媲美，也反映了基于 transformer 为基层框架的模型在图像生成任务上的可用性和有效性。Esser 等人 [13] 提出了 Taming transformers，网络由两个部分组成：VQGAN 和 transformer。VQGAN 是 VQVAE [14] 的一种变体，它使用判别器和感知损失来提高视觉质量。通过 VQGAN，图像可以被表示为一系列含有上下文丰富的离散向量，因此这些向量可以很容易地通过自回归的 transformer 模型进行预测，该模型可以学习生成高分辨率图像的远程交互作用。因此，所提出的 Taming transformers 在各种图像合成任务上取得了最先进的结果。Jiang 等人 [15] 提出了基于 transformer 的 GAN 网络结构 TransGAN。由于计算内存的限制，难以按像素级直接生成高分辨率的图像，因此需要在不同阶段逐步提高特征图的分辨率。相应地，设计了一个多尺度的判别器来处理不同阶段的输入大小的输入。现有的 GAN 正则化方法与自注意相互作用不佳，导致训练过程中的严重不稳定。Lee 等人 [16] 提出了 ViTGAN，该方法将新的正则化方法引入生成器和鉴别器，以稳定训练过程和收敛。在自注意力模块中引入欧氏距离来增强 transformer 判别器的敏感性，并提出了自调制层范数和隐式神经表示来加强生成器的训练。ViTGAN 是第一个可以实现与最先进的基于 CNN 的 GAN 相当性能的方法。图 3(b) 是文本到图像的生成框架 [17]，不在本文的调查范围内。

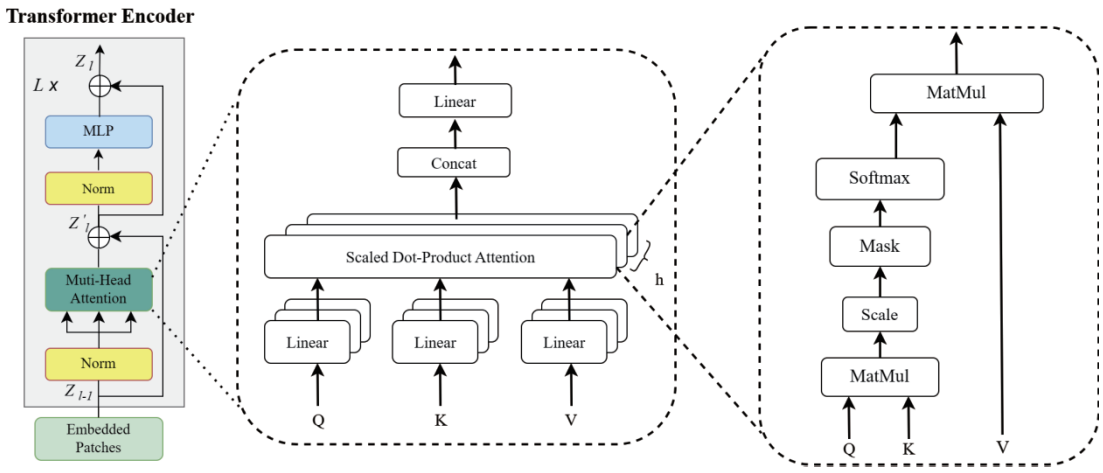

图 2 transformer Encoder 内部结构

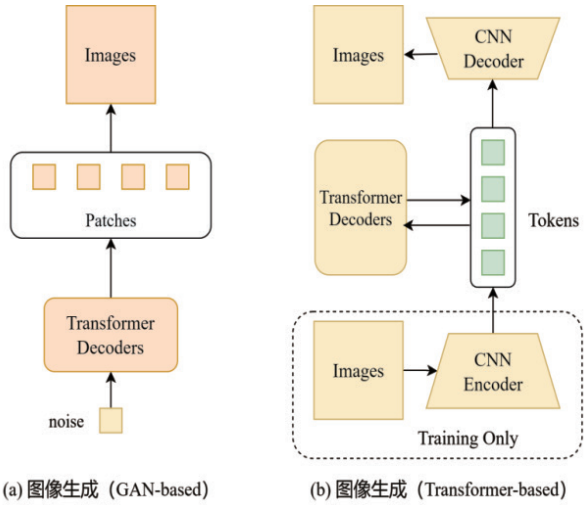

图 3 基于 transformer 图像生成方法的通用框架

3 基于 transformer 超分辨率图像重建方法

图像超分辨率重建是一种从低分辨率图像中恢复高分辨率图像的底层视觉任务。不同于单图像的超分辨率 SISR，现有的超分辨率重建方法 [18-20] 忽略了从注意力机制的角度来学习参考图像中的纹理特征并转移到高分辨率（HR）图像中。Yang 等人 [21] 提出了基于 transformer 的图像超分辨率的纹理网络（TTSR）。TTSR 模型由四个交互密切模块构成用于特征融合，多模块的特征融合设计使得从 LR 图像和 Ref 图像更好地进行特征交互，从而可以学习生成高质量的纹理特征。此外，TTSR 模型凭借跨尺度方式可以进一步堆叠，因此，能够恢复出不同级别的纹理特征。实验表明，TTSR 在定量和定性评估方面都比最先进的方法取得了显著改进。由于视觉 transformer 需要将图像进行分割，因此具有以下两个缺点：（1）由于原始图像需要分割为若干个图像块（patch），因此每个 patch 的边界像素重建无法使用原始邻近像素；（2）重建图像的 patch 周围存在边界伪影。Liang 等人基于 Swin-transformer[22] 提出了 SwinIR[23] 方法，该模型由浅层特征提取、深度特征提取和高质量图像重建模块构成。图像内容和注意力权重之间基于内容的交互，可以解释为空间变化卷积。网络表现出更好的性能和更少的参数。由于 SwinIR 的中间特征图中观察到明显的块效应，而这些伪影是由窗口分区机制引起的，这种现象表明移位窗口机制无法有效地建立跨窗口连接。因此，Chen 等人 [24] 提出了一种新颖的混合注意力 transformer（HAT）。该模型结合了通道注意力和自注意力方法，此外，为了更好地聚合跨窗口信息，HAT 引入了重叠交叉注意模块（OCAB）来增强相邻窗口特征之间的交互。在训练阶段，额外提出了一种相同任务的预训练策略，以带来进一步的改进。大量实验证明了所提出模块的有效性，整体方法明显优于最先进的方法 1 dB 以上。虽然 transformer 模型弥补了 CNNs 的不足，但其计算复杂度随着空间分辨率的增加而二次增长，因此不适用于大多数涉及高分辨率图像的图像重建任务。Zamir 等人 [25] 提出了一个有效的 transformer 模型，Restoration transformer 通过对基础模块（多头前馈网络）的几个关键设计，使它能够捕捉远距离像素间的相互作用，同时仍然适用于大图像。

4 基于 transformer 图像增强方法

图像增强旨在突出图像中重要的特征信息，抑制不感兴趣的特征信息。提升图像质量存在两个挑战：第一是必须要用有限的计算资源处理高分辨率图像；第二是需要结合输入低质图像的结构信息和全局信息来输出稳定的高质量图像。IPT 模型和 STAR 模型 [26] 是基于 transformer 架构的图像增强方法。IPT 是一种预训练模型，它需要同时处理多个任务，因此模型设计多头和多尾结构来进行不同任务的图像处理。在 Urban100 数据集上，对比当前最好的超分辨率算法，IPT 模型展现出巨大的优势。但是，IPT 模型有 114M 参数以及 33G 的 FLOPS。相比之下，STAR 被设计为轻量级的，并且可以实现实时性能，如图 4 所示。

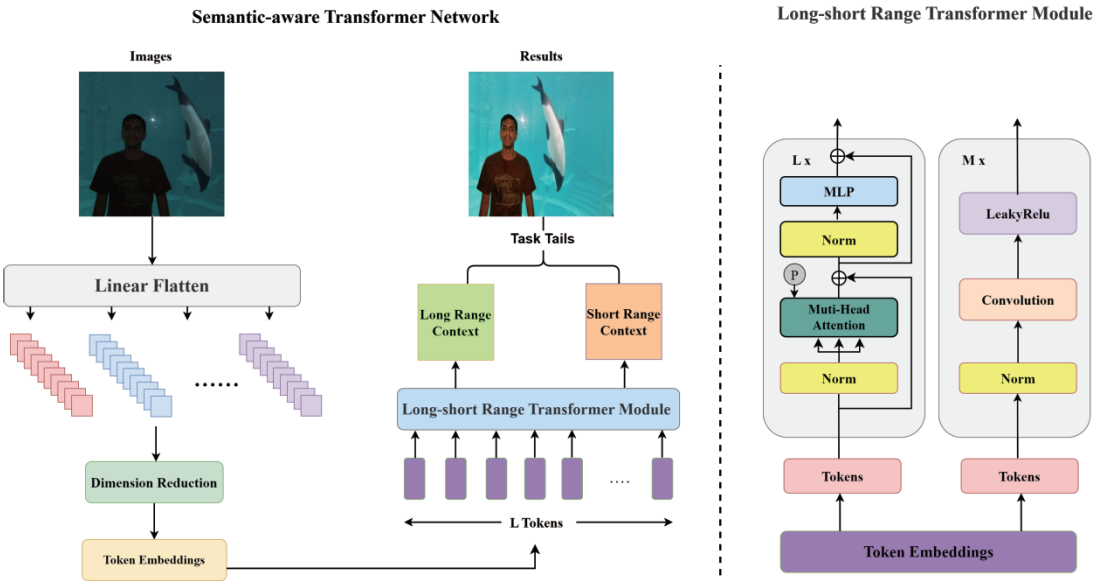

图 4 STAR 模型结构

STAR 主要包含 MSA 和 FFN，避免了卷积模块的堆叠，能更有效地提取结构信息。在 STAR 中图像块被映射成 Token，与直接计算像素之间的依赖不同，STAR 直接学习 Token 之间的依赖关系，也可以隐式地学习语义结构，从而比 CNN 提供更多有意义的语义结果。作者提出了一个特殊设计的双路 long short Range transformer 确保 STAR 能够集中捕获图像全局上下文信息（global contexts）并且减少计算量。STAR 可以很轻易地适用于不同底层视觉任务，例如：光照提升、自动白平衡、图像修饰，并且其模型的复杂度很低，效果也达到 SOTA。

5 基于 transformer 的图像恢复方法

图像复原旨在从低质量的图片恢复出对应的高质量图片。Wan 等人 [27] 首次将 transformer 用于图像修复，第一阶段先将缺损图降采样（降低计算量）并用 transformer 修复来初步重建结构和粗纹理作为视觉先验；第二阶段在第一步的先验上加入原分辨率的残缺图做引导，使用常规的 CNN 完成细节纹理的填充。与确定性补全方法相比，图像保真度的性能大大提高。Wang 等人 [28] 提出了基于 transformer 的架构 Uformer 方法。Uformer 有两个核心设计：第一个关键要素是局部增强的 window transformer 块，使用基于非重叠窗口的自注意来减少计算要求，并在前馈网络中采用深度卷积来进一步提高其捕获局部环境的能力；第二个关键因素是，设计了三种 skip-connection 方案，可以有效地将信息从编码器传递到解码器。在这两个设计的支持下，Uformer 在捕捉图像恢复的有用依赖关系方面拥有很高的能力。Feng 等人 [29] 提出了基于 transformer 的图像文档恢复的模型，以解决文档图像的几何和光照失真问题。在基于 transformer 强大的长距离以及全局特征提取能力的基础上，模型采用几何矫正和光照恢复方法。通过设置一组已学习的查询嵌入，几何扭曲 transformer 通过自注意机制捕获文档图像的全局特征，并对像素逐位解码来矫正几何扭曲。在几何解扭曲之后，照明矫正 transformer 进一步消除了阴影伪像，以提高视觉质量和 OCR 准确性。Chen 等人 [30] 注意到 transformer 全局注意力计算需要耗费大量资源，然而基于局部分块的窗口注意力方法缺乏局部注意力之间的特征交互。为了建立一个适合于图像恢复的高效 transformer 模型，提出了一个矩形窗口自注意（Rwin-SA）来替代 ViT 中普通的多头自注意机制，形成交叉聚合 transformer 块（CATB）的方法。大量的实验表明，该模型在几个图像恢复应用上优于最近的先进方法。

6 基于 transformer 的图像质量评估方法

感知图像质量评估（IQA）本质上是一项识别任务，即识别质量图像的优劣。IQA 可分为三类：全参考图像质量指标（full reference）、半参考图像质量指标（reduced reference）和无参考图像质量指标（no reference）。You 等人 [31] 提出了基于 ViT 架构的无参考评价方法（TRIQ），模型定义了具有足够长度的位置编码来覆盖数据集中的最大图像分辨率，因此可以处理不同分辨率和不同宽高比的图像。Cheon 等人 [32] 提出了一种基于 transformer 架构的全参考评价方法（IQT）。模型使用 CNN 来提取参考和失真图像的特征，提取的特征图被投影到固定大小特征并进行扁平化。特征中添加了可训练的额外 Token 嵌入与位置编码。失真图像特征输入至 transformer 解码器，参考图像特征输入至 transformer 解码器。获取 transformer 输出的 Token 输入至 MLP 得到评估结果。Ke 等人 [33] 提出了一个多尺度的图像质量方法（MUSIQ）来处理不同尺寸、不同长宽比的原分辨率图像。通过多尺度的图像表示，该模型可以捕捉到不同粒度（granularity）的图像质量。另外，该网络引入一种新型的基于哈希的（hash-based）二维空间嵌入方法和一种尺度嵌入，来作为多尺度表示中的位置嵌入。MUSIQ 只改变输入编码，因此它可以适应任何 transformer 变体。

7 总结与展望

相比应用于视觉任务的卷积神经网络，以 transformer 为框架的方法展示出良好的性能和巨大的潜力，在视觉任务主流方法中已经占领了一席之地。近年来已经提出了许多方法，这些方法在广泛的视觉任务中都表现出良好的性能。但是，目前基于 transformer 架构的大多数计算机视觉方法仍存在下列情况。

（1）基于 transformer 框架的视觉模型处理的任务有限，比较单一化，无法同时处理多种任务。目前，自然语言处理领域如 GPT-3 [34] 和 GPT-4[35] 已经实现了基于 transformer 框架的方法在一个模型中处理多个任务。Perceiver [36] 和 Perceiver IO [37] 是开创性的模型，可以在多个领域工作，包括图像、音频、多模态、点云。因此，下一步如何设计将多视觉任务融合在一个模型中是未来重要研究方向。

（2）由于 transformer 模型缺乏归纳偏置的能力，因此在训练网络时需要大量的数据，网络性能才能超越 CNN。此外，在大量数据的条件下，训练网络需要消耗的算力资源也是非常庞大的。因此，如何减少 transformer 模型对数据的依赖以及减低计算开销是未来重要的研究方向。

参考文献：

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL].(2017-06-12)[2023-08-06].https://arxiv.org/abs/1706.03762.

[2] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE,

2016: 770-778.

[3] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. In IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6):1137-1149.

[4] CHEN M, RADFORD A, CHILD R, et al. Generative pretraining from pixels[C]//ICML' 20: Proceedings of the 37th International Conference on Machine Learning.New York:ACM,2020:13-18.

[5] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL].(2020-10-22)[2023-07-29].https://arxiv.org/abs/2010.11929.

[6] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[EB/OL].(2020-05-26)[2023-07-21].https://arxiv.org/abs/2005.12872.

[7] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: deformable transformers for end-to-end object detection[EB/OL].(2020-10-28)[2023-08-16].https://arxiv.org/abs/2010.04159.

[8] ZHENG S X, LU J C, ZHAO H S, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[EB/OL].(2020-11-30)[2023-09-01].https://arxiv.org/abs/2012.15840.

[9] CHEN H T, WANG Y H, GUO T Y, et al. Pre-trained image processing transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE,2021, 12299-12310.

[10] ZHOU L W, ZHOU Y B, CORSO J, et al. End-to-end dense video captioning with masked transformer[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Piscataway:IEEE,2018: 8739-8748.

[11] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL].(2018-10-11)[2023-08-27].https://arxiv.org/abs/1810.04805.

[12] PARMAR N, VASWANI A, USZKOREIT J, et al. Image transformer[EB/OL].(2018-03-15)[2023-07-29].https://arxiv.org/abs/1802.05751.

[13] ESSER P, ROMBACH R, OMMER B. Taming transformers for high-resolution image synthesis[EB/OL].(2020-11-17)[2023-07-17].https://arxiv.org/abs/2012.09841.

[14] OORD A, VINYALS O, KAVUKCUOGLU K. Neural discrete representation learning[EB/OL].(2017-11-02)[2023-07-29].https://arxiv.org/abs/1711.00937.

[15] JIANG Y F, CHANG S Y, WANG Z Y. TransGAN: two pure transformers can make one strong GAN, and that can scale up[EB/OL].(2014-02-14)[2023-08-23].https://arxiv.org/abs/2102.07074.

[16] LEE K, CHANG H W, JIANG L, et al. ViTGAN: training GANs with vision transformers[EB/OL].(2021-07-29)[2023-08-02].https://arxiv.org/abs/2107.04589.

[17] RASHME A, PAVLOV M, GOH G, et al. Zero-Shot Text-to-Image Generation[J]. arXiv preprint arXiv: 2102.12092, 2021.

[18] DONG C, LOY C C, HE K M, et al. Image super-resolution using deep convolutional networks[EB/OL].(2014-12-31)[2023-08-15]. https://arxiv.org/abs/1501.00092.

[19] KIM J, LEE J K, LEE K M. Accurate image super-resolution using very deep convolutional networks[EB/OL].(2015-11-31)[2023-07-25].https://arxiv.org/abs/1511.04587.

[20] KIM S, KANG I, KWAK N. Semantic sentence matching with densely-connected recurrent and co-attentive information[EB/OL].(2023-01-03)[2023-09-01].https://ar5iv.labs.arxiv.org/html/1805.11360.

[21] YANG F Z, YANG H, FU J L, et al. Learning texture transformer network for image super-resolution[EB/OL].(2020-06-07)[2023-08-15].https://arxiv.org/abs/2006.04139.

[22] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[EB/OL].(2021-05-25)[2023-09-02].https://arxiv.org/abs/2103.14030.

[23] LIANG J Y, CAO J Z, SUN G L, et al. SwinIR: image restoration using swin transformer[EB/OL].(2021-08-23)[2023-08-02].https://arxiv.org/abs/2108.10257.

[24] CHEN X Y, WANG X T, ZHOU J T, et al. Activating more pixels in image super-resolution transformer[EB/OL].(2022-05-09)[2023-09-20].https://arxiv.org/abs/2205.04437.

[25] ZAMIR S W, ARORA A, KHAN S, et al. Restormer: efficient transformer for high-resolution image restoration[EB/OL].(2021-11-18)[2023-08-26].https://arxiv.org/abs/2111.09881.

[26] ZHANG Z Y, JIANG Y T, JIANG J, et al. Star: a structure-aware lightweight transformer for real-time image enhancement[C]//International Conference on Computer Vision. Piscataway:IEEE, 2021:4086-4095.

[27] WAN Y Z, ZHANG J B, CHEN D D, et al. High-fidelity pluralistic image completion with transformers[EB/OL].(2021-05-25)[2023-07-29].https://arxiv.org/abs/2103.14031.

[28] WANG Z D, CUN X D, BAO J M, et al. Uformer: a general

（下转第 57 页）

[6] BLEI D M, NG Y A, JORDAN M. Latent dirichlet allocation[J].The journal of machine learning research, 2003(3): 993-1022.

[7] GRIFFITHS T, MARK S. Finding scientific topics[J]. Proceedings of the national academy of sciences of the United States of America, 2004,101: 5228 - 5235.

[8] 张国生 . 大数据驱动的多视点软件需求规约 [J]. 中国电子科学研究院学报 ,2020,15(2):147-151+158.

[9] 李飞 , 高晓光 , 万开方 . 基于动态 Gibbs 采样的 RBM 训练算法研究 [J]. 自动化学报 ,2016,42(6):931-942.

[10] HU Q, SHEN J J, WANG K,et al. A Web service clustering method based on topic enhanced gibbs sampling algorithm for the dirichlet multinomial mixture model and service collaboration graph[J]. Information sciences,2022,586:1-10.

[11] 陈广明 , 姒依萍 . 高校在线课程的多元评价体系建设研究 [J]. 宁波职业技术学院学报 ,2021,25(1):85-88+104.

[12] PU X M,YAN G X,YU C Q,et al. Sentiment analysis of online course evaluation based on a new ensemble deep learning mode: evidence from Chinese[J]. Applied sciences, 2021,11(23): 11313-11331.

[13] 金贤 , 戚华 . 基于质量立体模型的实践教学质量评价指标体系的构建 : 以顶岗实习为例 [J]. 南京广播电视大学学报 , 2019(2): 55-58.

[14] 尼格拉木·买斯木江 . 面向慕课在线课程质量评价指标提取及情感分析研究 [D]. 乌鲁木齐 : 新疆师范大学 , 2021.

[15] PILACUANBONETE L, GALINDOVILLARDÓN P, DELGADOÁLVAREZ F. HJ-biplot as a tool to give an extra analytical boost for the latent dirichlet assignment (LDA) model: with an application to digital news analysis about COVID-19[J]. Mathematics, 2022, 10(14): 2529-2529.

[16] XIE X P, LI D D, ZHU W W,et al. Drug efficacy prediction in tumors based on lda model[C]//Chinese Control Conference. Piscataway: IEEE, 2022: 410-415.

[17] EGWOM O J,HASSAN M,TANIMU J J,et al. An LDA–SVM machine learning model for breast cancer classification[J]. Biomedinformatics,2022,2(3): 345-358.

【作者简介】

张成（1996—），男，江苏东台人，研究生，助理工程师，研究方向：教育信息化。

（收稿日期：2023-10-18）

（上接第 51 页）

u-shaped transformer for image restoration[EB/OL].(2021-06-06)[2023-09-06].https://arxiv.org/abs/2106.03106.

[29] FENG H, WANG Y C, ZHOU W G, et al. Doctr: document image transformer for geometric unwarping and illumination correction[EB/OL].(2021-10-25)[2023-09-17].https://arxiv.org/abs/2110.12942.

[30] CHEN Z, ZHANG Y L, GU J J, et al. Cross aggregation transformer for image restoration[EB/OL].(2022-11-24)[2023-09-07].https://arxiv.org/abs/2211.13654.

[31] YOU J Y, KORHONEN J. Transformer for image quality assessment[EB/OL].(2020-10-30)[2023-08-29].https://arxiv.org/abs/2101.01097.

[32] CHEON M, YOON S J, KANG B, et al. Perceptual image quality assessment with transformers[EB/OL].(2021-04-05)[2023-09-06].https://arxiv.org/abs/2104.01778.

[33] KE JJ, WANG Q, WANG Y L, et al. MUSIQ: multi-scale image quality transformer[EB/OL].(2021-08-12)[2023-08-26].https://arxiv.org/abs/2108.05997.

[34] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[EB/OL].(2020-05-28)[2023-09-08].https://arxiv.org/abs/2005.14165.

[35] OPENAI. GPT-4 Technical Report[EB/OL].(2023-05-18)[2023-06-26].https://arxiv.org/abs/2303.08774.

[36] JAEGLE A, GIMENO F, BROCK A, et al. Perceiver: general perception with iterative attention[EB/OL].(2021-05-04)[2023-09-20]. https://arxiv.org/abs/2103.03206.

[37] JAEGLE A, BORGEAUD S, ALAYRAC J B, et al. Perceiver IO: a general architecture for structured inputs & outputs[EB/OL].(2021-07-30)[2023-08-20]. https://arxiv.org/abs/2107.14795.

【作者简介】

屠红艳（1996—），女，江苏扬州人，助理工程师，研究方向：深度学习、图像重建。

管其杰（1995—），通信作者（email: 18208021@mail.shiep.edu.cn），女，江苏南京人，助理工程师，研究方向：深度学习、图像处理。

（收稿日期：2023-11-01）