DOI:10.11918/202302029

用于自监督表征学习的教师-学生互补掩码自动编码器

黄 靖[1,3],叶少雄[1],文元桥[2],朱立夫[1],黄亚敏[2]

(1.武汉理工大学 计算机与人工智能学院,武汉 430063; 2.武汉理工大学 智能交通系统研究中心,武汉 430063; 3.新一代人工智能技术应用交通运输行业研发中心,杭州 310013)

摘 要:针对自监督表征学习中掩码图像建模(MIM)方法存在上下游任务不匹配的问题,提出了一种称为教师-学生互补掩码自动编码器的新预训练模型,即TSCAE模型。该模型由具备互补掩码机制的教师模块和学生模块组成,其中教师模块基于Transformer结构,负责预测图像中掩码区域(如随机掩蔽输入图片的75%部分);学生模块则采用单一的编码器结构预测同一图像中剩余区域(如掩蔽输入图片余下的25%部分)。为从大量无标签数据中预训练出更丰富的视觉表征,TSCAE模型同时完成两类上游任务,分别是预测任务和对比任务,并在COCO和Tiny-ImageNet数据集上完成预训练。测试结果表明,在包括VOC在内的3个公有数据集和2个私有数据集上,TSCAE在图像分类、目标检测和语义分割等下游任务中,性能均优于经典的掩码自编码器(MAE)。特别地,TSCAE还在一定程度上缓解了预训练图像质量对视觉表征学习编码器的影响。

关键词:预训练模型;自监督学习;掩码图像建模;对比学习;编码器

中图分类号:TP399 **文献标志码**:A **文章编号**:0367-6234(2026)03-0074-14

Teacher-student complementary mask autoencoder for self-supervised representation learning

HUANG Jing[1,3], YE Shaoxiong[1], WEN Yuanqiao[2], ZHU Lifu[1], HUANG Yamin[2]

(1. School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430063, China; 2. Intelligent Transportation System Research Center, Wuhan University of Technology, Wuhan 430063, China; 3. Research and Development Center of Transport Industry of New Generation of Artificial Intelligence Technology, Hangzhou 310013, China)

Abstract: To address the problem of mismatch between upstream and downstream tasks exhibited by masked image modeling (MIM) methods in self-supervised representation learning, we proposed a novel pre-training model, called teacher-student complementary masked autoencoder, or in other words, the TSCAE model. The TSCAE model consists of two modules with complementary masked mechanisms, called teacher module and student module, respectively. The teacher module was designed as a Transformer-based structure to predict the masked region of an image (e.g., randomly masking 75% of the input image), while the student module employed a sole encoder to predict the remaining region of the same image (e.g., masking the remaining 25% of the input image). Meanwhile, to attain a richer visual representation from a large number of unlabeled data, the TSCAE model completed two kinds of upstream tasks, namely prediction and contrastive tasks. After that, the TSCAE model achieved the pre-training on COCO and Tiny-ImageNet datasets. The results demonstrate that across three public datasets including VOC and two private datasets, the proposed TSCAE model achieves better performance than the classical masked autoencoder (MAE) methods on downstream tasks such as image classification, object detection, and semantic segmentation. In particular, the TSCAE also alleviates the impact of the quality of the pre-training images on the visual representation learning encoder to a certain extent.

Keywords: pre-training model; self-supervised learning; masked image modeling; contrastive learning; encoder

在计算机视觉领域,基于自监督学习的预训练模型方法主要分为生成式和判别式两类。生成式方法常见于基于图像重建的通用自编码器或生成对抗网络(GAN)[1],这类方法直接操作于像素空间,计

收稿日期:2023-2-16;**录用日期**:2023-04-06;**网络首发日期**:2024-06-24
网络首发地址:https://link.cnki.net/urlid/23.1235.t.20240622.1054.019
基金项目:国家自然科学基金资助项目(52072287);浙江省科技计划项目(2021C01010);新一代人工智能技术应用交通运输行业研发中心开放基金(202302H);浙江省交通厅科技项目(2024006)
作者简介:黄 靖(1977—),男,副教授,硕士生导师;叶少雄(1999—),男,硕士研究生;文元桥(1975—),男,教授,博士生导师;朱立夫(1999—),男,硕士研究生;黄亚敏(1990—),男,研究员,博士生导师
通信作者:黄 靖,huangjing@whut.edu.cn

算耗时而且对于特征学习往往并非必要;判别式方法和常用的监督学习类似,通过特定的目标函数进行训练,但是输入和标签都是来自无标注数据,所以判别式方法也可属于自监督学习范畴。随着 Transformer[2] 在计算机视觉领域的广泛应用,研究者尝试将掩码图像建模(masked image modeling, MIM)应用于视觉领域并取得了一定进展,近期 MIM 的代表性工作包括 BEiT[3]、MAE[4] 和 CAE[5]。这些工作均采用基于 Transformer 的编码器–解码器架构,编码器能够通过 MIM 学习到良好的表征,从而在下游任务中实现优异的泛化性能。

然而,现有基于 MIM 的预训练模型有两个主要问题:一是上下游任务不匹配。BEiT 在预训练期间会随机掩蔽部分图像 Patch,并将损坏的图像 Patch 输入到编码器中;MAE 和 CAE 在预训练期间,编码器的输入只有可见图像块,而下游任务中需将完整图像输入编码器,导致上下游任务存在匹配间隙。二是编码器语义表征能力有限。由于未分离编码功能和 MIM 任务,预训练后编码器的语义表征能力一般,BEiT 的编码器需同时负责表征编码和掩码图像块预测,MAE 的解码器会进一步更新可见图像块的表征信息,承担了部分本应由编码器完成的表征学习任务。此外,实验发现经典 MIM 方法 MAE 在图片质量较差的数据集上,编码器的语义表征学习能力显著降低,其语义表征能力依赖于大规模、高质量的数据集。针对以上问题,从下面几个方面去解决。

1)为了增强编码器的语义表征能力,提出基于教师—学生网络[6] 的互补掩码预训练模型(TSCAE)。编码器只负责图像表征编码,教师网络中的解码器负责从可见图像块表征中预测掩码图像块的表征。预测得到的掩码图像块表征一方面经过掩码预测模块(3 层 MLP)预测真实像素,另一方面引入基于编码表征空间的对比学习作为自监督的前置任务,通过对比损失降低学生和教师分支输出表征的分布差异,实现教师表征信息的有效传递;多前置任务设计能够让编码器学习到更好的图像表征信息。

2)为了最大限度减少 MIM 方法中的上下游任务不匹配间隙,提出教师–学生互补掩码方法。将单张图像划分为可见图像块和掩码图像块作为输入,教师网络和学生网络的编码器输入由可见图像块和互补掩码图像块构成。从单个学生分支或者教师分支看,上下游任务仍存在一定不匹配,但是从模型整体看,这种互补掩码机制在图像输入结构和输入信息方面与下游任务更为接近,从而减小了上下游任务之间不匹配间隙。

1　相关工作

自监督学习一直是研究热点,研究者提出多种自监督学习方法并应用到自然语言处理和计算机视觉领域,取得了一系列理论和应用研究成果。自监督预训练模型最先在自然语言处理(NLP)领域取得突破,在多项下游任务中表现优异。Devlin 等[7] 提出的 BERT 模型表明,掩码建模方法能够大幅提升预训练模型性能。随着 ViT[8] 的提出和发展,Transformer 结构在计算机视觉领域得到广泛的应用,掩码图像模型(MIM)为视觉自监督学习开辟了新方向。在此之前,视觉自监督算法主要基于对比学习思路设计,其指导原则是通过自动构造相似实例和非相似实例,训练表示学习模型,使相似实例在投影空间中距离较近,非相似实例距离较远。近年来,基于对比学习的研究成果显著,部分方法性能超过了有监督学习,比如 MoCo 系列[9-11]、SimCLR 系列[12-13] 以及只采用正例进行对比学习的 BYOL[14]、SimSiam[15] 和 DINO[16]。对比学习的应用十分广泛,Tian 等[17] 提出的对比表示蒸馏(CRD),基于对比学习思路最大化教师和学生网络的互信息下界,提升输出变量的相关性;Chen 等[18] 提出的对比知识蒸馏方法,通过设计距离度量,使相似的样本更接近,同时分离不同样本。

近年来,基于 MIM 的自监督预训练方法发展迅速,Bao 等受到 BERT 的启发提出 BEiT,继承 ViT 的 Patch 划分策略,将预训练任务改为图像复原,将把可见图像块的颜色信息和掩码块(不包含掩码图像块的颜色信息)输入到 ViT 中,通过线性层完成预测;He 等提出基于掩码图像重建的预训练模型 MAE,采用编码器—解码器框架,输入图像经随机掩蔽后,编码器仅接收可见图像块,前置任务仅为图像像素复原;Chen 等提出一种新的 MIM 方法 CAE,通过完全分离表征学习和前置任务功能,使编码器学习更优表征,其预训练的前置任务包括图像复原和基于可见图像块表征的掩码图像块表征预测。相较于对比学习模型 MoCo v3(注意力图主要在图像的主体区域),CAE 能覆盖几乎所有 Patch;相比于对比学习类自监督学习方法,基于 MIM 的自监督表征学习方法更适用于下游任务。本文提出的 TSCAE 同样属于 MIM 类方法,其教师网络与学生网

络的编码器输入由可见图像块和互补掩码图像块构成,预训练的前置任务除图像像素复原外,还引入教师网络与学生网络输出表征的对比学习。通过与现有主要模型的对比,验证了本文模型的有效性。

2 教师–学生互补掩码预训练模型

2.1 网络概述

本文提出的教师–学生互补掩码预训练模型(TSCAE)由教师模块和学生模块组成。如图 1 所示,整个预训练模型是一个非对称的教师–学生网络结构。在掩码图像建模方法中,图像掩码策略的设计至关重要,以往的图像掩码方法均存在上下游任务不匹配的问题,而 TSCAE 中的互补掩码方法对输入图片使用互补掩蔽(黑色色块为掩码块,对于输入网络不可见;像素块对输入网络可见),教师模块和学生模块对图像的掩蔽是互补的,这种设计从整体上减少了上下游任务的不匹配间隙。教师网络与学生网络中的编码器是 ViT 结构,两者网络结构与参数完全相同,得益于其强大的自注意力机制,编码器可以学习到更好的图片表征。教师模块比学生模块多了解码器部分,用于从编码器输出的可见图像块表征中预测掩码图像块的表征;解码器预测得到的表征将与学生网络学习得到的表征进行对比学习。由于教师与学生模块的输入对图像的掩蔽是互补的,两者输出的表征存在较大的差异,而基于对比的知识蒸馏方法能够提高教师–学生模块输出表征的相关程度,使教师分支与学生分支在训练过程中实现信息互补,有利于提高编码器的表征学习能力。

在自监督学习中,可以通过设置各种上游任务训练编码器,最大限度挖掘编其表征学习能力。上游任务是自监督学习的核心策略,能够通过数据本身定义的伪标签从数据中学习表征。如图 1 所示,TSCAE 的教师分支和学生分支设计了 3 个上游任务:1)预测图像的真实像素,使用 MSE 损失函数;2)基于编码表征的对比学习,使教师与学生模块输出的对应表征块在空间上尽量靠近;3)基于编码表征的全局特征信息对比,使用交叉熵损失函数。通过这 3 个上游任务可以让编码器学习到更优的图像表征,从而在下游任务中实现更好的泛化性能。

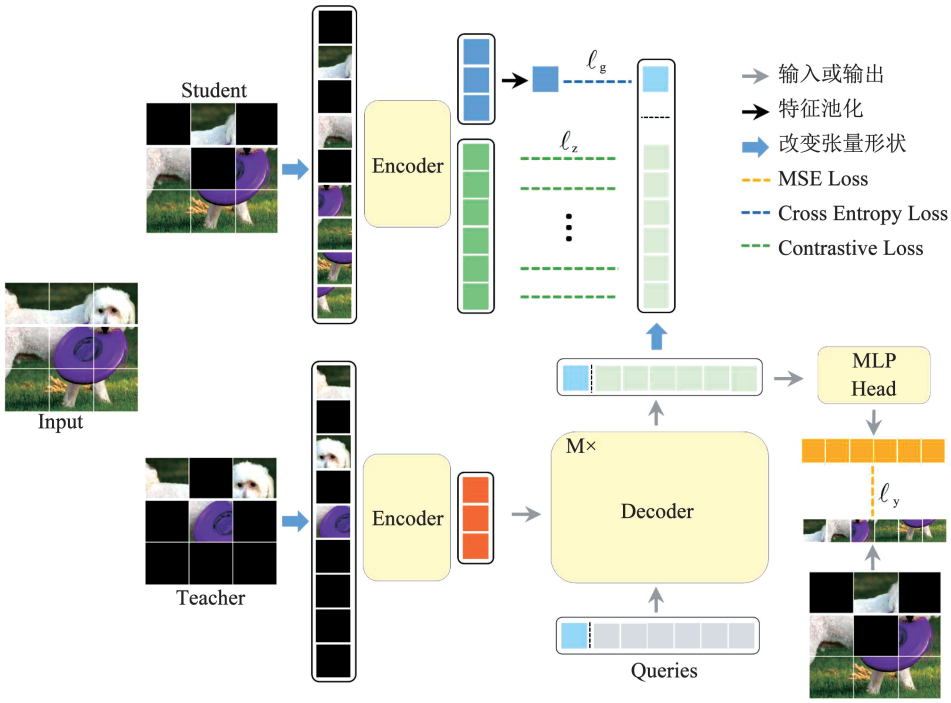


图 1 教师–学生互补掩码自监督预训练模型的结构图

Fig. 1 Structural diagram of teacher-student complementary masked self-supervised pre-training model

相关自监督预训练模型的计算如图 2 所示,(a)代表本文提出的 TSCAE,(b)代表 CAE,(c)代表 MAE,蓝色部分用于损失函数的计算。t 表示教师分支,s 表示学生分支,m 表示掩码图像块,v 表示可见图像块。(a)中黑色编码器 $\mathcal{F}$ 输入教师网络的可见块 $\boldsymbol{X}_{\mathrm{v}}^{\mathrm{t}}$ 和掩码块 $\boldsymbol{Q}_{\mathrm{m}}^{\mathrm{t}}$,输出可见块的表征信息 $\boldsymbol{Z}_{\mathrm{v}}^{\mathrm{t}}$;解码器 g 用于从表征 $\boldsymbol{Z}_{\mathrm{v}}^{\mathrm{t}}$ 中预测出掩码块的表征信息 $\boldsymbol{Z}_{\mathrm{m}}^{\mathrm{t}}$ 和掩码块的全局特征信息 $\boldsymbol{G}_{\mathrm{m}}^{\mathrm{t}}$;$\mathcal{R}$ 用于从掩码块表征信息 $\boldsymbol{Z}_{\mathrm{m}}^{\mathrm{t}}$ 预测真实的掩码块像素 $\boldsymbol{Y}_{\mathrm{m}}$;蓝色编码器 $\mathcal{F}$ 输入学生网络的可见块 $\boldsymbol{X}_{\mathrm{v}}^{\mathrm{s}}$ 和掩码块 $\boldsymbol{Q}_{\mathrm{m}}^{\mathrm{s}}$,输出的 $\boldsymbol{Z}_{\mathrm{v}}^{\mathrm{s}}$ 和 $\boldsymbol{G}_{\mathrm{m}}^{\mathrm{s}}$ 分别表示学生网络可见块的表征信息和掩码块的全局特征信息;ℓ_{y}、ℓ_{z} 和 ℓ_{g} 为损失

函数。(b)中编码器 $\mathcal{F}$ 输入可见块 $\boldsymbol{X}_{\mathrm{v}}$,输出潜在表征信息 $\boldsymbol{Z}_{\mathrm{v}}$;潜在上下文回归器 $\mathcal{H}$ 从 $\boldsymbol{Z}_{\mathrm{v}}$ 中预测掩码块的潜在表征 $\boldsymbol{Z}_{\mathrm{m}}$,$\boldsymbol{Z}_{\mathrm{m}}$ 用于预测掩码块的表征 $\bar{\boldsymbol{Z}}_{\mathrm{m}}$;解码器 $\mathcal{L}$ 负责从 $\boldsymbol{Z}_{\mathrm{m}}$ 来预测目标 $\boldsymbol{Y}_{\mathrm{m}}$;$\ell_{\mathrm{y}}$ 和 ℓ_{z} 为损失函数。(c)中编码器 $\mathcal{F}$ 输入可见块 $\boldsymbol{X}_{\mathrm{v}}$,输出潜在表征信息 $\boldsymbol{Z}_{\mathrm{v}}$;解码器 $\mathcal{H}$ 输入表征信息 $\boldsymbol{Z}_{\mathrm{v}}$ 和掩码块 $\boldsymbol{Q}_{\mathrm{m}}$,输出预测的掩码块真实像素 $\boldsymbol{Y}_{\mathrm{m}}$;$\ell_{\mathrm{y}}$ 为损失函数。从计算图可以直观地看出,本文提出的 TSCAE 与其他模型在网络结构和自监督任务上有较大差异。在预训练完成后,网络中的编码器拥有较强的语义表征能力。在下游任务中,图像块全部输入编码器,输出的图像语义表征可分别用于图像分类、目标检测和语义分割等视觉任务。

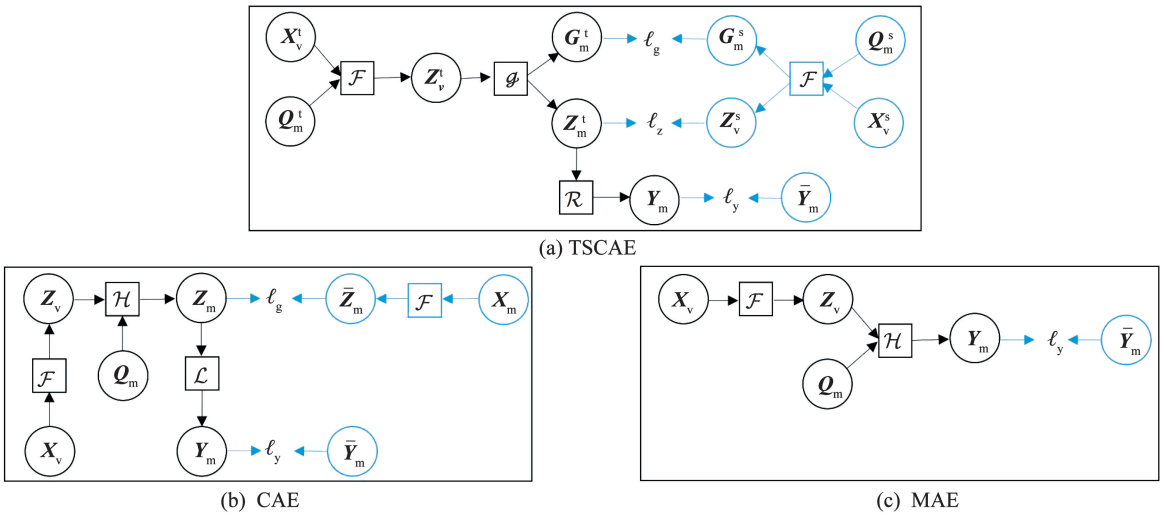


图 2　相关预训练模型的计算图

Fig. 2　Computational display of related pre-trained models

2.2　互补掩码机制

在基于掩码图像重建(MIM)的自监督预训练模型中,设计合理的掩蔽策略十分重要,本文提出了一种基于教师–学生网络的互补掩码方法。首先采用 ViT 的做法,将输入图像分为无重叠的 Patch 块(如 ViT-Base 将图片划分成 16 × 16 的图像块),然后通过均匀采样策略对这些 Patch 进行采样,未被采样的 Patch 会被掩蔽并进行正态分布的初始化(均匀采样可避免图像中心附近出现更多的掩码 Patch 块)。教师分支与学生分支的输入均进行上述掩码操作,且两者对图像的掩蔽呈互补关系,例如从展平的 Patch 块中随机采样 25% 作为教师网络的可见 Patch 块,剩余的 75% 对教师网络不可见,学生网络则相反。

如图 1 所示,输入图像被分为多个 Patch 块,其中黑色的色块对对应的网络输入不可见,而带有图像的色块对对应的网络输入可见。从单个学生网络或教师网络来看,上下游仍有一定不匹配间隙。从整体上看,TSCAE 在输入形式和输入信息上是高度互补的,上下游输入的匹配度显著提升。互补掩码方法能够在教师网络和学生网络的输入形式上保持一致性,也能够更多地保留每一个 Patch 块的相关信息,有利于教师网络与学生网络之间的信息互补,同时也为两者的表征对比奠定基础,最终助力编码器学习到更优的图像表征。图 13 是 3 种掩码方法的对比图,表 5 是对应掩码方法的对比实验结果。实验表明,相比其他掩码方法,互补掩码策略在下游任务中表现更优,在一定程度上改进了上下游任务的匹配性。

2.3　教师–学生模块

教师与学生模块是整个网络的核心,主要分为三大组件:教师模块(Teacher)、学生模块(Student)和掩码预测模块(MLP Head)。教师模块和学生模块中的编码器部分由 ViT-Base 组成,其中 block 迭代 12 层,dim 为 768 维,且两者网络结构和模型参数完全一致。教师–学生模块的输入采用互补掩码的策略。此外,鉴于目前基于对比表示的知识蒸馏方法已具备较好性能,本文在 TSCAE 中引入教师模块与学生模块之间的对比学习,以进一步提高编码器的表征学习能力。

1)教师模块

如图 3 所示,教师网络中的解码器基于 Transformer 结构设计,与 DETR[19] 中的解码器类似。Queries 的初始化为参数随机,依次进行多头自注意

力计算(Multi-Head Self-Attention)、多头交叉注意力计算(Multi-Head Cross-Attention)和前馈网络(FFN)。为构建更深层模型,每个模块周围都采用残差连接,随后接入层归一化模块,解码器通过迭代多层更新 Queries 参数。

解码器的输入有 3 个部分:第 1 部分是编码器学习得到的可见表征块,用于预测教师网络输入不可见的掩码块表征;第 2 部分是随机初始化的 Queries 块,其个数与教师网络不可见的掩码块个数相同,也与学生网络可见的图像块个数相同,最终经过解码器迭代更新得到 Queries 块,将与学生网络学习到的可见表征块进行一对一的对比学习,用于预测图像的真实像素,Queries 块前设有一个信息块,用于表示解码器迭代更新中学习到的 Queries 块的全局特征信息,该信息将与学生网络学习到的掩码块表征进行全局信息对比;第 3 部分是位置嵌入编码(Positional Encoding),用于表示各序列块的位置信息。为使教师模块预测的表征与学生模块学习到的表征信息保持高度一致,解码器的位置嵌入编码与编码器的位置编码设置为相同,不再重新初始化,这有助于上游任务中基于编码表征空间的对比学习。

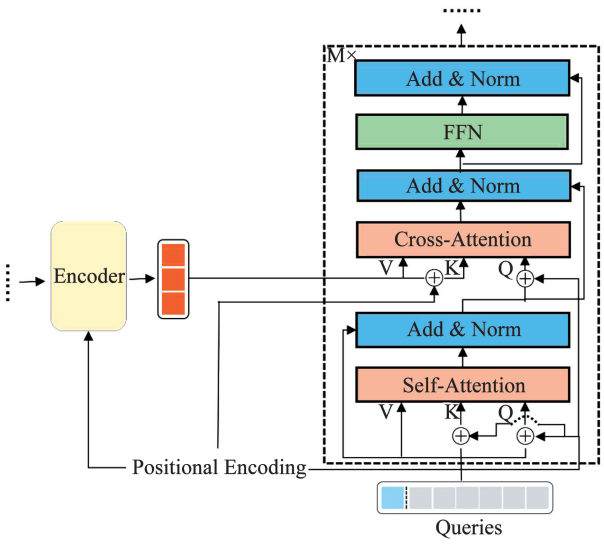


图 3 基于 Transformer 结构的解码器模块

Fig. 3 Decoder module based on Transformer structure

2)学生模块

TSCAE 中的学生模块包括两个基于表征对比的上游任务。如图 4 所示,L 代表表征块的数量,教师模块预测得到的所有表征块和学生模块学习得到的所有表征块,都将参与基于对比学习的 SimSiam 损失函数计算(该对比损失只采用正例进行对比学习,且有较好的性能)。本文对该对比损失进行改进:不同于对整张图像提取的特征进行对比学习,TSCAE 中教师分支和学生分支输出的每一个特征块都进行对比学习,最后的对比损失为 L 个损失之和。此外,学生模块输出的掩码块表征与解码器预测得到的掩码块表征是互补的,来自同一张样本图片的两个不同部分。学生模块输出的掩码块表征经过特征的最大池化后,与教师模块预测得到的 Queries 的全局特征(如图 1 中的浅蓝色块所示)进行交叉熵损失对比,使得学生和教师两个分支输出的 k 维概率分布尽量相近。通过设计两个基于表征对比的上游任务,编码器能够持续学习表征信息,这种设计几乎将整张图片都作为自监督信号,使教师分支与学生分支在训练过程中实现信息互补,不断完善彼此缺失的信息。

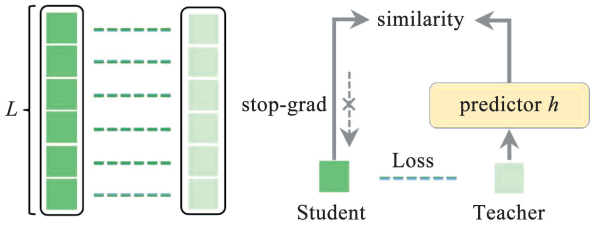


图 4 基于编码表征的对比学习结构图

Fig. 4 Contrastive learning structure based on encoded representations

3)掩码预测模块

教师分支中的掩码预测模块对应图 1 中的 MLP Head。教师网络解码器预测得到的掩码图像块表征,一方面和学生网络的输出进行对比学习,另一方面经过 MLP Head 模块进行图像像素的回归预测。MLP Head 模块采用 3 层线性层预测真实的像素值,输入层、隐藏层和输出层的维度分别是 768、1 024、768,激活函数使用的是 ReLU 函数。经过 3 层线性层后,预测值与真实像素块进行 MSE 损失函数的计算。图 5 是 TSCAE 在 COCO[20] 训练集(118 287 张图片)和私有的医学肝脏数据集(10 000 张图片)上经过 400 轮的自监督预训练后,在 COCO 验证集和非自然光的医学肝脏验证集上使用 MLP Head 模块得到的可视化结果。每组图中,左边是原图,中间是随机遮住 75% 像素块的实际输入图,右边是掩码预测模块的预测结果图。可以看出,无论是自然光图片还是非自然光图片,该模块在随机遮住图片 75% 像素的情况下仍有较好的图像复原效果,目标的轮廓和颜色均能被清楚还原。

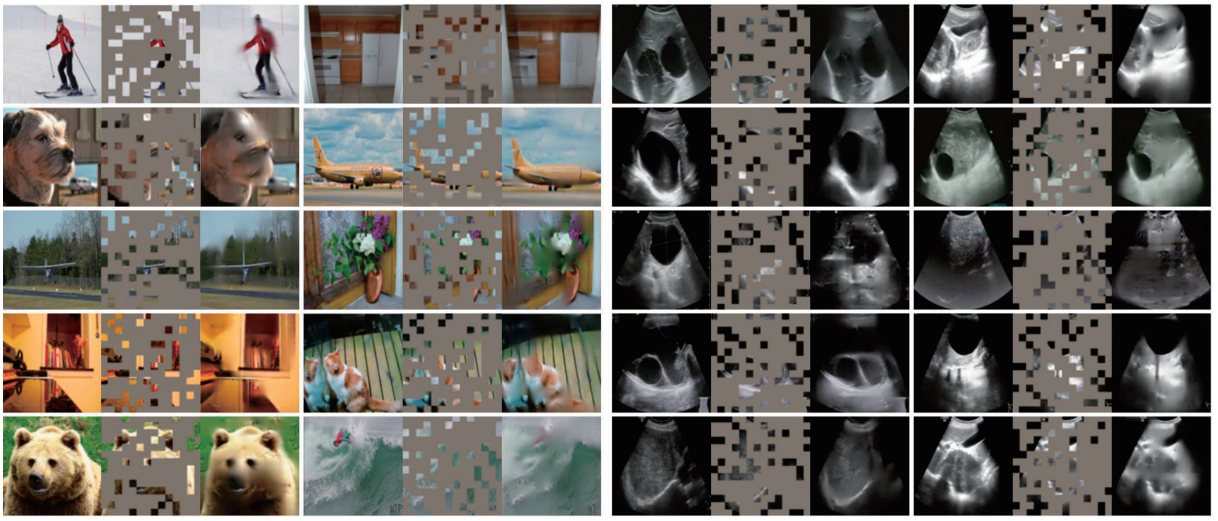


图 5　预训练模型 TSCAE 在 COCO 验证集上(左)和在医学肝脏验证集上(右)的像素复原效果图

Fig. 5　Pixel restoration renderings of pre-training model TSCAE on the COCO validation set (left) and medical liver validation set (right)

2.4　损失函数

TSCAE 中的 3 个上游任务包括基于 MIM 方法的图像重建和基于教师-学生网络之间的表征对比,能从图片中发掘更多的自监督信号。教师分支与学生分支在训练过程中能够实现信息互补,因此编码器能够拥有更强的表征学习能力。3 个上游任务分别对应 3 个损失函数,如图 2(a)蓝色部分中的 ℓ_y、ℓ_z、ℓ_g。损失函数 ℓ_y 的任务是对图像像素进行预测,损失函数为 MSE 均方误差,公式为

$$\ell_y(\boldsymbol{Y}_m,\bar{\boldsymbol{Y}}_m)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y}_i)^2 \tag{1}$$

式中:n 为教师分支中掩码块序列的长度;$\boldsymbol{Y}_m$ 为掩码预测模块预测得到的像素块序列(如图 1 中的黄色色块,y_i 对应块中每个像素值);$\bar{\boldsymbol{Y}}_m$ 为掩码块序列对应的真实像素块序列(图 1 中黄色色块下方的像素块,$\bar{y}$ 对应该块中每个像素值)。

损失函数 ℓ_z 对应的任务是对比学习解码器预测得到的掩码图像块表征和学生网络学习得到的图像块表征。为了让对应 Patch 块的特征信息能够相互靠近,并且消除负样本寻找对模型训练的限制,TSCAE 使用基于 SimSiam 的对比学习损失函数。与之不同的是,TSCAE 中每个训练批次会对应一个 Patch 块序列,序列中的每个块都要与对应的正样本进行对比学习(图 4 中一对深绿色块和浅绿色块为一对正样本)。在计算损失函数时,需要将教师分支与学生分支输出张量的形状转化为二维张量,再进行后续对比损失计算,具体推导公式为:

$$\boldsymbol{z}_1=\boldsymbol{Z}_m^t\times\mathrm{reshape}(N\times L,D) \tag{2}$$

$$\boldsymbol{z}_2=\boldsymbol{Z}_v^s\times\mathrm{reshape}(N\times L,D) \tag{3}$$

$$\boldsymbol{p}_1,\boldsymbol{p}_2=h(\boldsymbol{z}_1),h(\boldsymbol{z}_2) \tag{4}$$

$$\ell_z(\boldsymbol{Z}_m^t,\boldsymbol{Z}_v^s)=\frac{1}{2}S(\boldsymbol{p}_1,\boldsymbol{z}_2)+\frac{1}{2}S(\boldsymbol{p}_2,\boldsymbol{z}_1) \tag{5}$$

式中:N 为批处理大小;L 为序列块的长度(即图 4 中的 L);D 为张量的维度;$\boldsymbol{Z}_m^t\in\mathbb{R}^{N\times L\times D}$ 为教师网络解码器预测得到的表征序列(图 4 中的浅绿色块序列);$\boldsymbol{Z}_v^s\in\mathbb{R}^{N\times L\times D}$ 为学生网络编码器学习得到的表征序列(图 4 中的深绿色块序列)。$\boldsymbol{Z}_m^t$ 和 $\boldsymbol{Z}_v^s$ 经过变化后记为 $\boldsymbol{z}_1\in\mathbb{R}^{N\times L\times D}$,$\boldsymbol{z}_2\in\mathbb{R}^{N\times L\times D}$;$h(*)$ 是一个 prediction MLP 层,输入形状与输出形状一致,用于防止训练模型坍塌;$\boldsymbol{p}_1$、$\boldsymbol{p}_2$ 是 $\boldsymbol{z}_1$ 和 $\boldsymbol{z}_2$ 分别经过这个 MLP 层的结果。最后的损失通过计算两个余弦相似性得到,余弦相似性的计算公式为

$$S(\boldsymbol{p},\boldsymbol{z})=-\frac{\boldsymbol{p}}{\|\boldsymbol{p}\|}\cdot\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|} \tag{6}$$

损失函数 ℓ_g 是一个交叉熵损失函数。教师网络解码器预测得到的掩码图像块表征与学生网络编码器学习到的掩码图像块表征作为一张样本图片的两个不同部分,应属于同一类。$\boldsymbol{G}_m^t\in\mathbb{R}^{N\times D}$ 表示教师网络解码器迭代更新学习到的 Queries 块的全局特征信息(图 1 中的浅蓝色块);$\boldsymbol{G}_m^s\in\mathbb{R}^{N\times D}$ 表示学生网络学习得到的掩码块表征序列经特征池化后的特征信息(图 1 中的深蓝色块)。为了进行全局特征信息对比,将这两个包含图像块序列全局特征信息的表征块作为交叉熵损失函数的输入,该损失函数

公式为

$$\ell_{\mathrm{g}} = (\boldsymbol{G}_{\mathrm{m}}^{\mathrm{t}}, \boldsymbol{G}_{\mathrm{m}}^{\mathrm{s}}) = -\frac{1}{M}\sum(\boldsymbol{G}_{\mathrm{m}}^{\mathrm{t}} \log \boldsymbol{G}_{\mathrm{m}}^{\mathrm{s}}) \quad (7)$$

式中 M 表示掩码块的数量。TSCAE 最后的损失函数为上面的 3 个损失函数相加,如公式(8)。使用 3 个损失函数可以让编码器从不同角度学习到更丰富的表征信息。

$$\ell = \ell_{\mathrm{y}}(\boldsymbol{Y}_{\mathrm{m}}, \bar{\boldsymbol{Y}}_{\mathrm{m}}) + \ell_{\mathrm{z}}(\boldsymbol{Z}_{\mathrm{m}}^{\mathrm{t}}, \boldsymbol{Z}_{\mathrm{v}}^{\mathrm{s}}) + \ell_{\mathrm{g}}(\boldsymbol{G}_{\mathrm{m}}^{\mathrm{t}}, \boldsymbol{G}_{\mathrm{m}}^{\mathrm{s}}) \quad (8)$$

3 实验与分析

3.1 数据准备

为了验证 TSCAE 的有效性,本文在图像分类、目标检测和语义分割 3 个下游任务中进行微调实验。为比较预训练的图片质量对模型表征学习能力的影响,实验使用 Tiny-ImageNet[22](11 万张,图片大小为 64 × 64)和 COCO(约 11 万张)2 个数据集(样本数量相当,但是 COCO 数据集的图片质量更好)。下游任务在 3 个经典的数据集[23-24]和两个私有数据集上进行微调:私有数据集包括非自然光场景下用于图片分类的医学肝脏数据集,以及自然场景下用于目标检测的水上目标数据集,部分图片见图 7 和图 8。医学肝脏数据集共 10 000 张有标签图片,包括 16 种病灶,分别为 AE1、AE2、AE3、BLA、CE1、CE2、CE3、CE4、CE5、CL、FNH、HCC、HH、ICC、LHC、LM。将这些图片按照 8:2 的比例分为训练集和验证集。水上目标数据集包含 12 000 张图片,分为 6 类,包括集装箱运货船、邮轮、帆船、散货船和其他类船只及岛礁。其中训练集有 10 800 张,验证集有 1 200 张。

在上述条件下,TSCAE 与现有主要模型(文献[3]、文献[4]、文献[5]和文献[16])进行对比,所有预训练模型所使用的预训练数据和实验环境都保持一致。作为一种新的 MIM 类方法,TSCAE 还与 MAE 进行更加详细的对比实验。此外,为了验证 TSCAE 中部分模块的有效性,进行了相应的消融实验。

3.2 实验设置

使用 Pytorch 搭建教师-学生互补掩码自监督预训练模型,在 GPU 为 2 台 RTX2080Ti 的设备上进行实验。实验设置参照 MAE,实验中使用 ViT-B/16 作为编码器。教师网络中的解码器模块的迭代层数设置为 6,解码器模块的 Queries 输入长度设置为教师模块中掩码块的序列长度 L 加 1,使用正态分布初始化。损失函数为图 1 中 3 个 Loss 之和,即 $\ell = \ell_{\mathrm{y}} + \ell_{\mathrm{z}} + \ell_{\mathrm{g}}$。使用 AdamW 优化器进行优化,基本学习率设为 0.000 3,权重衰减设为 0.005。

3.3 分类实验

TSCAE 中编码器的基准模型(Baseline)为 ViT-Base(ViT-B/16)。在不同数据集上预训练后,分别在 Tiny-ImageNet 数据集和医学肝脏数据集上微调,以评估提出的预训练模型的图片表征学习能力。

1)数据集 Tiny-ImageNet 数据集实验

Tiny-ImageNet 包含 200 个类别,训练集 100 000 张图片,验证集 10 000 张图片,测试集 10 000 张图片,图片像素大小均为 64 × 64。不使用额外的图片数据集进行预训练,使用训练集和测试集共 11 万张图片进行无标签自监督预训练,然后在下游任务中微调,评估 TSCAE 学习到图片表征信息质量。

直接使用 ViT-B/16 进行有监督训练 200 轮,Top-1 识别准确率为 53.2%;TSCAE 预训练 400 轮后使用 ViT-B/16 微调 50 轮,Top-1 识别准确率达到 68.1%。可以看出,经过 TSCAE 预训练后,只需要微调少量轮次就能获得较高的准确率,编码器在预训练后能够得到较好的图像表征,有利于下游分类任务。

TSCAE 主要和基于掩码图像建模(MIM)方法的代表模型 MAE 和基于对比学习的代表模型 DINO 进行对比。图 6(a)是各模型在 Tiny-ImageNet 上 150 轮微调训练准确率图和训练 Loss 下降图,从图中可以很明显地看出 TSCAE 在下游分类任务中表现更好:相比 DINO 和 MAE,TSCAE 的识别准确率上升更快,Loss 曲线下降更快,模型收敛更快,识别准确率更高。具体实验结果如表 1 所示,Epochs 为模型预训练的轮次,TSCAE 使用 Tiny-ImageNet 作为预训练数据时,预训练 400 轮的分类准确率为 72.2%,高于其他预训练模型;使用 COCO 作为预训练数据时,预训练 400 轮的分类准确率为 75.1%,高于 MAE 的 69.8%。此外,TSCAE 还优于在 Tiny-ImageNet 数据集上分类准确率较高的有监督网络模型[25-27]。

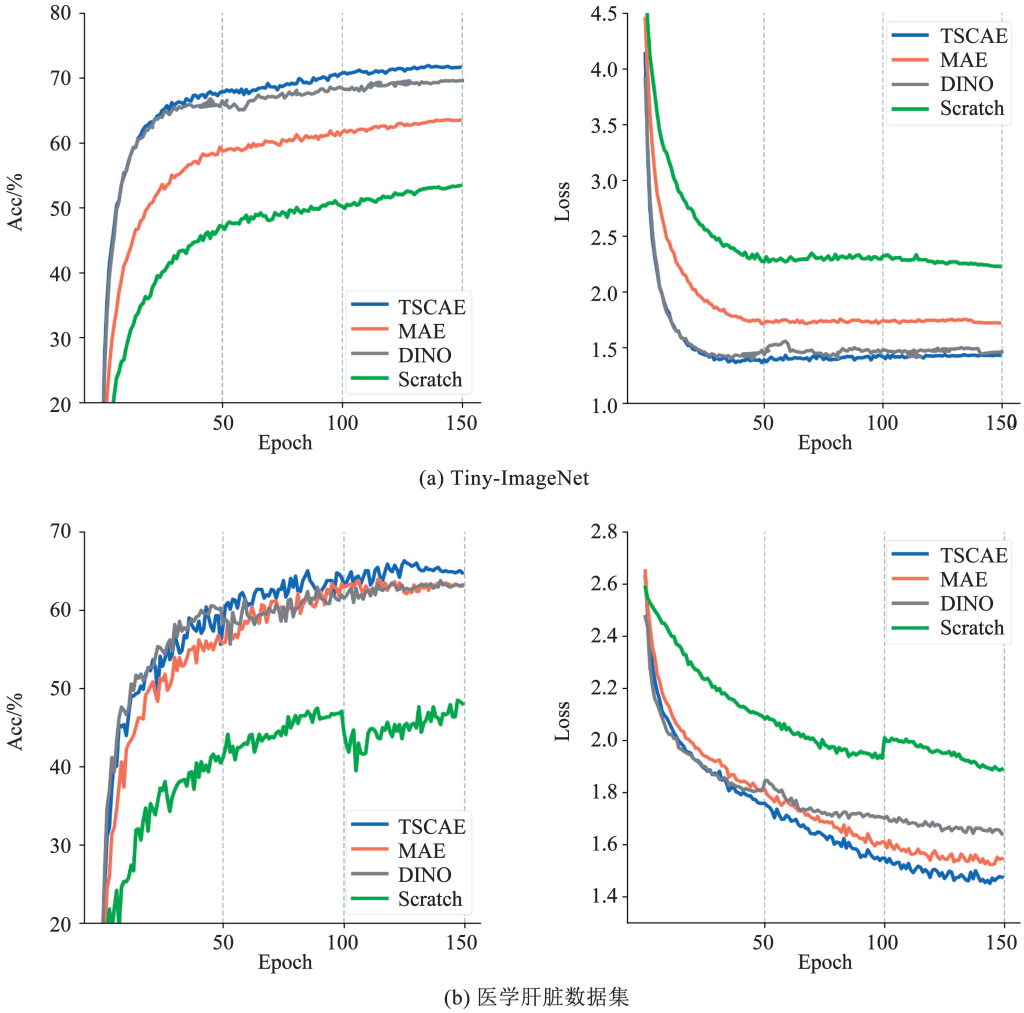


图 6　相关模型在 Tiny-ImageNet 和医学肝脏数据集上的微调准确率和 Loss 曲线

Fig. 6　Fine-tuned accuracy and Loss curves of related models on Tiny-ImageNet and medical liver dataset

表 1　相关模型在 Tiny-ImageNet 数据集上微调的分类准确率

Tab. 1　Classification accuracy of related models fine-tuned on Tiny-ImageNet dataset

算法模型	主干网络	Epochs	预训练数据	Top-1 Acc/%
Scratch	ViT-B/16	—	NO	55.2
UPANets	—	—	NO	67.7
AutoMix	ResNetXt – 50	—	NO	70.7
SAMix	ResNetXt – 50	—	NO	72.2
DINO	ViT-B/8	400	Tiny-ImageNet	69.1
BEiT	ViT-B/16	400	Tiny-ImageNet	64.1
MAE	ViT-B/16	200	Tiny-ImageNet	62.6
MAE	ViT-B/16	400	Tiny-ImageNet	64.3
MAE	ViT-B/16	400	COCO	69.8
CAE	ViT-B/16	400	Tiny-ImageNet	71.3
TSCAE	ViT-B/16	200	Tiny-ImageNet	70.1
TSCAE	ViT-B/16	400	Tiny-ImageNet	72.2
TSCAE	ViT-B/16	400	COCO	75.1

2)医学肝脏数据集上的实验

在非自然光的医学肝脏数据集上进行实验。医学数据集的特点是有标签的数据样本少,图片中目标存在分辨率低噪声大关键信息占比小的问题。实验数据集中有标签的医学肝脏图片共有 7 183 张,分为 16 类,将这些图片作为自监督学习的训练数据,下游微调时把训练集与验证集的图片比例设置为8:2。与 Tiny-ImageNet 上的实验环境一致,同样使用 ViT-Base(ViT-B/16)作为预训练模型中的骨干网络。

图 7　来自医学肝脏数据集的部分肝脏图片

Fig. 7　Some liver images from the medical liver dataset

图 6(b)为各模型在医学肝脏数据集上微调的每轮准确率图和训练 Loss 下降对比图。由于非自然光图片分辨率低、噪音大,所以不使用预训练权重

直接训练 ViT 很难收敛,并且识别准确率不高;而使用 TSCAE 进行预训练后,对 ViT 进行微调时并不需要太多的训练轮次就可让网络尽快收敛,并且识别的准确率大幅提升。表 2 为相关自监督模型在私有肝脏数据集上的实验结果,不使用预训练方法,直接训练 ViT 在验证集上的最高 Top-1 准确率为 50.3%,而 TSCAE 在验证集上微调后的 Top-1 准确率能够达到 66.3%,比 MAE 高出 2.6 个百分点,比基于对比学习方法的 DINO 高出 2.3 个百分点。

表 2 相关模型在医学肝脏数据集上微调的分类准确率

Tab. 2 Classification accuracy of related models fine-tuned on medical liver dataset

算法模型	主干网络	Epochs	Top-1 Acc/%	Top-5 Acc/%
Scratch	ViT-B/16	—	50.3	95.4
DINO	ViT-B/8	400	64.0	98.2
BeiT	ViT-B/16	400	63.5	98.1
MAE	ViT-B/16	400	63.7	98.2
CAE	ViT-B/16	400	65.5	98.2
TSCAE	ViT-B/16	400	66.3	98.1

如表 2 所示,TSCAE 在下游两个分类数据集上表现优异。同样是在图片质量较差的 Tiny-ImageNet 上进行预训练,TSCAE 比经典的图像掩码建模方法 MAE 在下游分类任务上表现更优,一方面可能是因为预训练的图片质量对 MAE 中编码器的图片表征学习能力影响更大,另一方面可能得益于 TSCAE 中的互补掩码机制能够在一定程度上改进上下游任务的匹配性,使上游学习的表征更有利于下游任务的微调。此外,自监督学习过程中的多个前置任务可以充分发掘图像中的有用信息,进而让编码器获得更强的表征学习能力。

3.4 目标检测与语义分割

实验选择图片质量更优、在下游任务上表现更好的 COCO 训练集进行相关模型的预训练,然后分别在 2 个公有数据集和 1 个私有数据集上进行微调,以验证提出的自监督预训练模型 TSCAE 在目标检测和语义分割两个下游任务上的泛化性能。

在目标检测的下游任务实验中,使用目标检测模型 Mask R-CNN[28] 在 PASCAL VOC 和私有的水上目标数据集上进行微调。具体方案:使用 COCO 训练集中的约 11 万张图片进行自监督预训练,在 PASCAL VOC 2007 + 2012(16 551 张)上进行微调,在 VOC 2007 验证集(4 952 张)上进行评估验证;在水上目标训练集(10 800 张)上进行微调,在水上目标验证集(1 200 张)上进行验证。目标检测模型 Mask R-CNN 使用 ViT 作为骨架网络,与 FPN[29] 联合使用。实验结果如表 3 所示,在相同实验环境下,TSCAE 模型在预训练 400 轮后,在 VOC2007 和水上目标验证集上微调 100 轮后的 mAP 分别是 61.2、47.9,比 MAE 表现更优,也优于基于对比学习的模型 DINO。图 10(a)是各模型预训练后在 VOC 2007 验证集上微调 100 轮的 mAP 曲线图。TSCAE 在特定场景水上目标检测的可视化效果如图 9 所示,使用 TSCAE 预训练 400 轮后对 Mask R-CNN 微调 100 轮,检测器能够正确分类且定位到水上的各目标。

图 8 来自水上目标数据集的部分图片

Fig. 8 Selected images from water target dataset

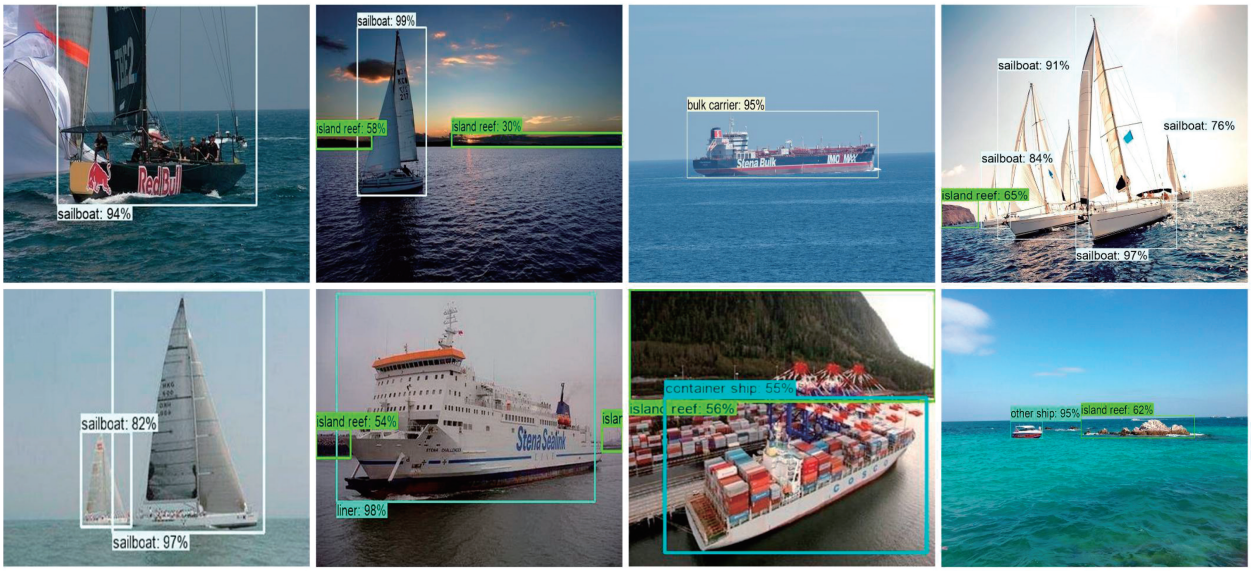


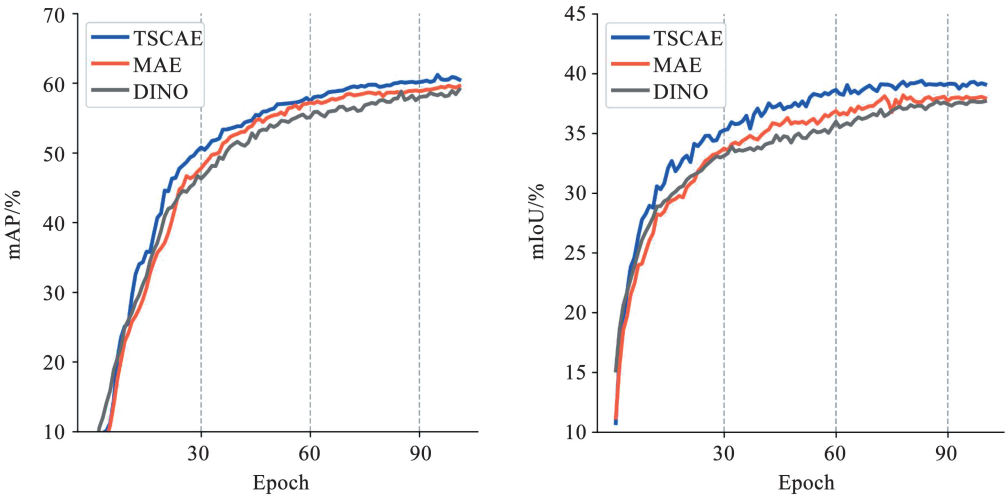
图 9 预训练模型 TSCAE 在水上目标数据集上微调的检测可视化

Fig. 9 Detection visualization of pre-training model of TSCAE fine-tuned on water target dataset

表 3　相关模型在 PASCAL VOC 上目标检测的微调实验结果对比

Tab. 3　Comparison of experimental results of related models fine-tuned in terms of target detection on PASCAL VOC

算法模型	主干网络	Epochs	预训练数据	VOC 2007/%			水上目标/%		
				AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
DINO	ViT-B/8	400	COCO	60.4	82.0	67.4	46.7	75.3	48.1
MAE	ViT-B/16	400	COCO	60.6	82.1	67.6	46.8	75.7	48.6
TSCAE	ViT-B/16	400	COCO	**61.2**	**83.5**	**68.2**	**47.9**	**77.2**	**49.2**



(a) VOC 2007　　(b) ADE20K

图 10　相关模型在 VOC 2007 和 ADE20K 上微调的训练曲线

Fig. 10　Training curves of related models fine-tuned on VOC 2007 and ADE20K

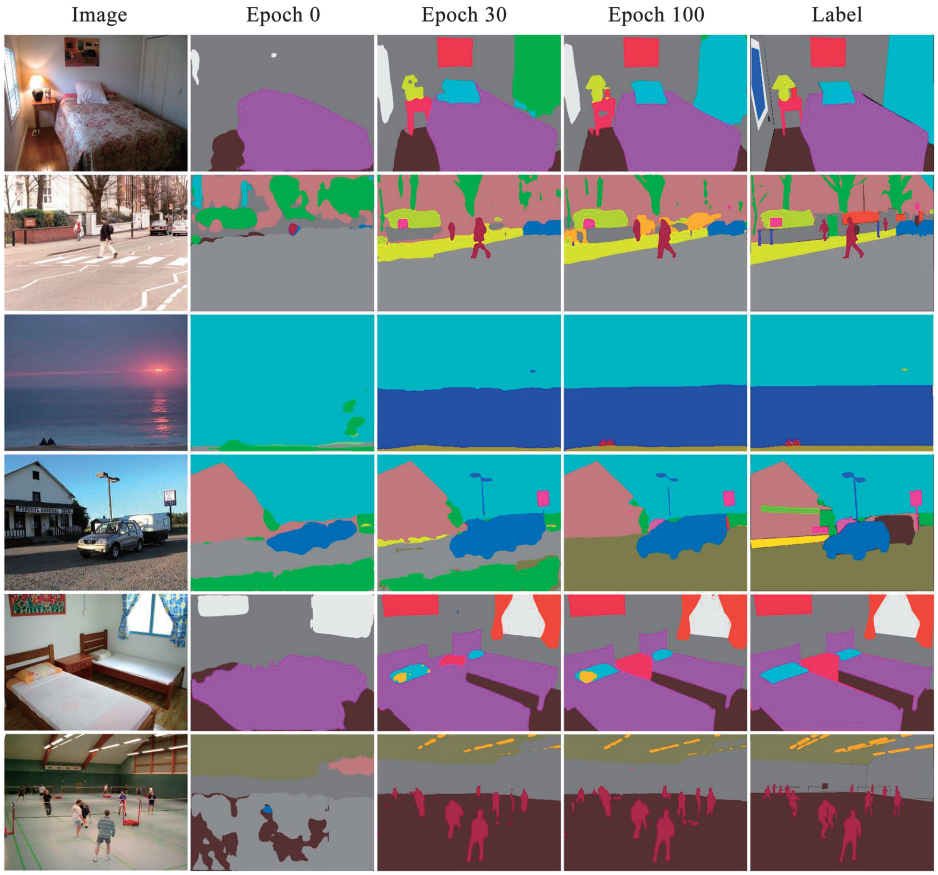


图 11　预训练模型 TSCAE 在 ADE20K 上微调的语义分割可视化

Fig. 11　Semantic segmentation visualization of pre-training model TSCAE fine-tuned on ADE20K

在语义分割的下游任务实验中,选择分割模型 UperNet[30] 和公开数据集 ADE20K 作为基准。相关自监督模型在 COCO 训练集上进行预训练,下游使用分割网络 UperNet 进行微调,实验结果如表 4 所示。TSCAE 在 ADE20K 上微调后得到的 mIoU 为 39.3,比 MAE 表现更优,也优于基于对比学习的预训练方法 DINO。图 10(b)是各模型在 ADE20K 上微调 100 轮得到的训练 mIoU 曲线图,可以看出 TSCAE 在下游语义分割任务上有较好的表现。TSCAE 在 ADE20K 上微调的第 0 轮、30 轮、100 轮分割结果的可视化如图 11 所示,随着下游微调轮次增加,语义分割的效果越来越好:经过 TSCAE 预训练后,编码器已具备一定的表征能力,微调第 0 轮时,部分像素可以分类正确;微调 30 轮后,物体形状大致能够分割出来(如人、天空、海水和床等);微调 100 轮后,物体的分割更加精细准确(如第 1 组图中的枕头、第 5 组图中的桌柜)。TSCAE 因为关注图片中的每一块图像的特征学习,所以在微调开始时不能马上分割出图中的主体区域,在经过少量微调后,能够逐渐分割出图中的主体目标,在充分微调后可实现更准确的物体分割。

表 4 相关模型在 ADE20K 上语义分割的微调实验结果对比

Tab. 4 Comparison of experimental results of related models fine-tuning in terms of semantic segmentation on ADE20K

算法模型	主干网络	Epochs	预训练数据	mIoU/%
DINO	ViT-B/8	400	COCO	38.1
MAE	ViT-B/16	400	COCO	38.2
TSCAE	ViT-B/16	400	COCO	39.3

3.5 注意力图可视化

图 12 为 TSCAE 在 COCO 上预训练后,分类标记作为 ViT 编码器最后一层中不同头的查询时的注意力图。可以看出,不同注意力头对图像的关注区域不同,且能够较为明显地区分出图像的主体和背景。TSCAE 除了会在图像的主体区域有较高响应,在图像的背景和细节区域也有响应。正因为 TSCAE 几乎关注图像中的每一块区域,所以在各类下游任务中表现优异。TSCAE 学习到的表征不仅包括图像的全局语义,还包括图像的细节和非主体区域的信息,这种更加丰富的表征信息更有利于下游任务进行微调。

图 12 使用分类标记作为最后一层中不同头的查询时的注意力图

Fig. 12 Attention maps when using the classfication tokens as queries for different heads in final layer

3.6 消融实验

消融实验分别从提出的互补掩码方法、预训练中的前置任务设计和预训练图片质量对表征学习的影响 3 个方面进行一系列的对比实验验证 TSCAE 及相关模块的性能。

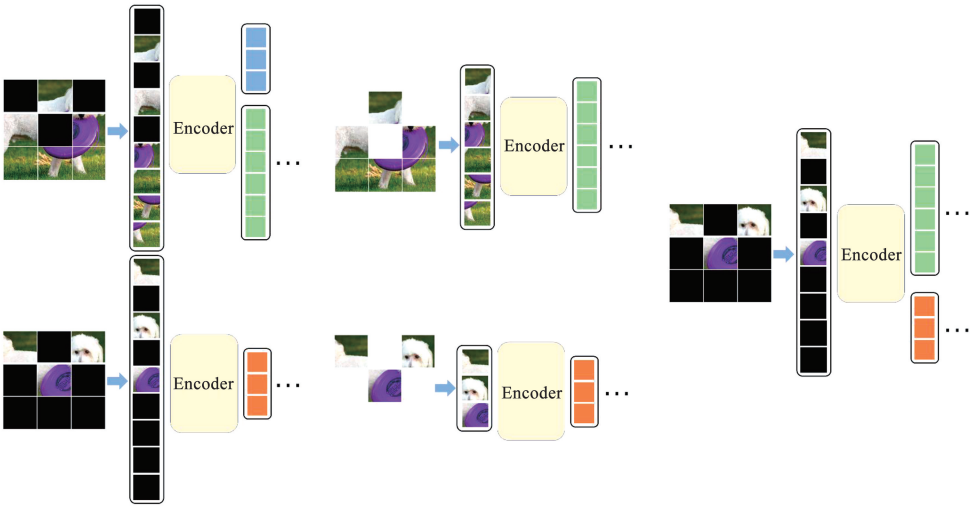
1)互补掩码机制

为证明 TSCAE 中互补掩码机制的优势,设计一组对比实验。图 13 中,(a)表示 TSCAE 中基于教师—学生网络的互补掩码方法;(b)表示不使用掩码方法,即简单的把图像块分成两部分,分别输入编码器中,CAE 模型中的输入就是这种方式;(c)表示单掩码方法,整个图片遮住一部分输入到编码器中。本文使用这 3 种方法分别在下游分类、目标检测和语义分割任务上微调,实验环境均与 3.3 节和 3.4 节保持一致,损失函数均设置为 $\ell = \ell_y + \ell_z$。实验结果如表 5 所示,使用互补掩码方法(a)在下游各任务上性能表现最优;使用单掩码方法(c)的实验结果最差,说明仅靠一个分支网络不足以进行有效的

对比学习,编码器的表征学习能力较弱;使用无掩码方法(b)的实验结果优于单掩码方法,但差于互补掩码方法,这可能是因为输入中缺少部分图像信息,与下游任务的输入间隙较大。上下游匹配度越高,上游学习到的表征往往在下游任务中能取得更好的效果。互补掩码方法中教师分支和学生分支的输入和下游输入保持一致,虽然每一个分支的输入都会掩蔽部分图像,但是高度互补的两个分支可以互相弥补彼此缺失的信息,在一定程度上减少了图像掩码建模方法中的上下游任务不匹配的间隙。互补掩码方法在下游任务上更好的表现,说明互补掩码方法拥有较好的匹配性。此外,在相同的上游任务下,不同掩码方法上下游任务匹配的程度不一样,导致下游任务性能存在差异,而本文提出的互补掩码策略相比于其他两种掩码方法表现更好,进一步证明了该方法能够改进上下游任务的匹配性。

2)预训练中的上游任务设计

为了验证 TSCAE 中上游任务的有效性,设计一组实验分析各个上游任务对编码器表征学习能力的影响。分别使用 ℓ_y、$\ell_y+\ell_z$、$\ell_y+\ell_g$ 和 $\ell_y+\ell_z+\ell_g$ 4 种损失搭配在下游各种任务中进行实验,实验环境与 3.3 节和 3.4 节保持一致。实验结果如表 6 所示,可以看出使用 3 种损失函数相加在下游图像分类、目标检测和语义分割任务中获取的分数最高;单独使用 ℓ_y 损失函数时,下游任务分数最低;加入损失函数 ℓ_z 后性能提升较为明显;加入损失函数 ℓ_g 后有小幅提升。说明将教师网络与学生网络表征之间的对比学习作为上游任务加入 TSCAE,能够大大提升编码器的表征学习能力。



(a) 互补掩码的方法　　(b) 不使用掩码方法　　(c) 单掩码的方法

图 13　不同掩码方法的示意图

Fig. 13　Different masking methods

表 5　几种掩码方法的实验结果对比

Tab. 5　Comparison of experimental results of several masking methods

方法	主干网络	Epochs	Tiny-ImageNet Top-1 Acc/%	肝脏数据 Top-1 Acc/%	VOC 2007 mAP/%	ADE20K mIoU/%
(c)单掩码方法	ViT-B/16	400	66.0	60.5	54.1	33.9
(b)无掩码方法	ViT-B/16	400	70.1	64.6	58.4	38.3
(a)互补掩码方法	ViT-B/16	400	**71.8**	**65.8**	**59.6**	**39.1**

表 6　对比不同损失函数的实验结果对比

Tab. 6　Comparison of experimental results of different loss functions

损失函数	主干网络	Epochs	Tiny-ImageNet Top-1 Acc/%	肝脏数据 Top-1 Acc/%	VOC 2007 mAP/%	ADE20K mIoU/%
$\ell_y+\ell_z+\ell_g$	ViT-B/16	400	72.2	66.3	61.2	39.3
ℓ_y	ViT-B/16	400	64.0	63.1	57.1	36.2
$\ell_y+\ell_z$	ViT-B/16	400	71.8	65.8	59.6	39.1
$\ell_y+\ell_g$	ViT-B/16	400	66.1	64.5	57.6	36.9

表7 相关模型在不同数据集上预训练的实验结果对比

Tab.7 Comparison of experimental results of related models pre-trained on different datasets

算法模型	主干网络	预训练数据	Tiny-ImageNet Top-1 Acc/%	VOC 2007 mAP/%	ADE20K mIoU/%
MAE	ViT-B/16	COCO	69.8(↑5.5)	60.6(↑7.1)	38.2(↑5.3)
MAE	ViT-B/16	Tiny-ImageNet	64.3	53.5	32.9
TSCAE	ViT-B/16	COCO	75.1(↑2.9)	61.2(↑4.4)	39.3(↑2.8)
TSCAE	ViT-B/16	Tiny-ImageNet	72.2	56.8	36.5

3)预训练图片质量对表征学习的影响

为证明预训练数据集中图片质量对表征学习的影响,设计一组实验:TSCAE 和 MAE 在不同的预训练数据集上进行预训练,然后使用预训练好的权重在 Tiny-ImageNet、VOC 2007 和 ADE20K 上进行微调。预训练数据为 COCO 训练集(11 万张,图片质量较好)和 Tiny-ImageNet 训练集加上验证集(11 万张,图片大小均为 64×64,像素较低、质量较差)。

实验结果如表 7 所示,相比于在 Tiny-ImageNet 上进行预训练,TSCAE 在 COCO 上进行预训练后,下游各任务性能分别提升 2.9、4.4、2.8 个百分点;而 MAE 在 COCO 上预训练后,下游各任务性能分别提升 5.5、7.1、5.3 个百分点。相比 MAE,使用 TSCAE 进行预训练时,图片质量对编码器的表征能力影响更小。而在实际的工程实践中,很多场景下的数据量并不多,且图片质量比较差(如医学图像的特点是数据样本少、图片分辨率低、噪声大、关键信息占比小),如何充分利用这些数据是关键问题,本文提出的自监督预训练模型 TSCAE 能够在一定程度上解决这个问题。

4 结 语

本文提出一种用于自监督表征学习的教师-学生互补掩码自动编码器(TSCAE)。该模型中的互补掩码方法能够在一定程度上减少掩码图像建模方法中上下游任务不匹配的间隙,基于教师-学生网络编码表征空间的对比学习作为上游任务,能从图像中发掘更多的自监督信号,增强编码器的语义表征能力。实验表明,TSCAE 在下游各任务上的表现相对于目前前沿的自监督学习方法具有一定的竞争力,图像质量对编码器的表征能力的影响相对于 MAE 更小。本研究旨在为自监督表征学习的探索提供更多的思路,也为某些特定数据集质量较差的场景提供参考。

参考文献

[1] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139. DOI: 10.1145/3422622

[2] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Long Beach, CA, USA: Curran Associates, Inc., 2017: 5998. DOI:10.48550/arXiv.1706.03762

[3] BAO H, DONG L, PIAO S H, et al. BEiT: BERT pre-training of image transformers[J]. arXiv preprint, 2021. arXiv: 2106.08254. DOI: 10.48550/arXiv.2106.08254

[4] HE K, CHEN X, XIE S, et al. Masked autoencoders are scalable vision learners[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 16000. DOI: 10.1109/CVPR52688.2022.01553

[5] CHEN X K, DING M Y, WANG X D, et al. Context autoencoder for self-supervised representation learning[EB/OL]. arXiv preprint, 2022. arXiv:2202.03026. DOI: 10.48550/arXiv.2202.03026

[6] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. arXiv preprint, 2015. arXiv: 1503.02531. DOI: 10.48550/arXiv.1503.02531

[7] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171. DOI: 10.18653/v1/N19-1423

[8] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C]//International Conference on Learning Representations (ICLR 2021). Virtual Event: OpenReview.net, 2021. DOI: 10.48550/arXiv.2010.11929

[9] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 9729. DOI:10.1109/CVPR42600.2020.00975

[10] CHEN XL, FAN H Q, GIRSHICK R B, et al. Improved baselines with momentum contrastive learning[EB/OL]. arXiv preprint, 2020. arXiv:2003.04297. DOI: 10.48550/arXiv.2003.04297

[11] CHEN X L, XIE S N, HE K M. An empirical study of training self-supervised vision transformers[EB/OL]. arXiv preprint, 2021. arXiv:2104.02057. DOI: 10.48550/arXiv.2104.02057

[12] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//

Proceedings of the 37th International Conference on Machine Learning (ICML 2020). Vienna, Austria: PMLR, 2020: 1597. DOI: 10.48550/arXiv.2002.05709

[13] CHEN T, KORNBLITH S, SWERSKY K, et al. Big self-supervised models are strong semi-supervised learners [J]. Advances in Neural Information Processing Systems, 2020, 33: 22243. DOI:10.48550/arXiv.2006.10029

[14] GRILL J B, STRUB F, ALTCHÉ F, et al. Bootstrap your own latent-a new approach to self-supervised learning [J]. Advances in Neural Information Processing Systems, 2020, 33: 21271. DOI: 10.48550/arXiv.2006.07733

[15] CHEN X, HE K. Exploring simple siamese representation learning [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021: 15750. DOI:10.1109/CVPR46437.2021.01549

[16] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 9650. DOI:10.1109/ICCV48922.2021.00951

[17] TIAN Y, KRISHNAN D, ISOLA P. Contrastive representation distillation[C]//International Conference on Learning Representations (ICLR 2020). Virtual Event: OpenReview.Net, 2020. DOI:10.48550/arXiv.1910.10699

[18] CHEN L, WANG D, GAN Z, et al. Wasserstein contrastive representation distillation [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021: 16296. DOI: 10.1109/CVPR46437.2021.01603

[19] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//Computer Vision-ECCV 2020: 16th European Conference. Glasgow, UK: Springer, Cham, 2020: 213. DOI:10.1007/978-3-030-58452-8_13

[20] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//Computer Vision-ECCV 2014: 13th European Conference . Zuric, Switzerland: Springer, Cham, 2014: 740. DOI: 10.1007/978-3-319-10602-1_48

[21] ZHOU B, ZHAO H, PUIG X, et al. Semantic understanding of scenes through the ADE20K dataset[J]. International Journal of Computer Vision, 2019, 127(3): 302. DOI:10.1007/s11263-018-1140-0

[22] LE Y, YANG X. Tiny-ImageNet visual recognition challenge[R]. Stanford, USA: CS231n, Stanford University, 2015.

[23] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The PASCAL visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303. DOI: 10.1007/s11263-009-0275-4

[24] EVERINGHAM M, ESLAMI S M A, VAN GOOL L, et al. The PASCAL visual object classes challenge: A retrospective [J]. International Journal of Computer Vision, 2015, 111(1): 98. DOI: 10.1007/s11263-014-0733-5

[25] TSENG C H, LEE S J, FENG J N, et al. UPANets: Learning from the universal pixel attention networks[EB/OL]. arXiv preprint, 2021. DOI:10.48550/arXiv.2103.08640

[26] LIU Z, LI S, WU D, et al. Unveiling the power of mixup for stronger classifiers [EB/OL]. arXiv preprint, 2021. DOI: 10.48550/arXiv.2103.13027

[27] LI S, LIU Z, WU D, et al. Boosting discriminative visual representation learning with scenario-agnostic mixup [EB/OL]. arXiv preprint, 2021. DOI:10.48550/arXiv.2111.15454

[28] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 2980. DOI:10.1109/ICCV.2017.322

[29] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA, 2017: 936. DOI: 10.1109/CVPR.2017.106

[30] XIAO T, LIU Y, ZHOU B, et al. Unified perceptual parsing for scene understanding [C]//Computer Vision-ECCV 2018: 15th European Conference. Munich, Germany: Springer, Cham, 2018: 418. DOI:10.1007/978-3-030-01228-1_26

(编辑　丁晓清)