基于多尺度-多方向Transformer的图像识别

杨育婷[1),2),3)] 李玲玲[1),2),3)] 刘 旭[1),2),3)] 焦李成[1),2),3)]
刘 芳[1),2),3)] 马文萍[1),2),3)]

[1)](西安电子科技大学人工智能学院 西安 710071)
[2)](智能感知与图像理解教育部重点实验室 西安 710071)
[3)](智能感知与计算国际联合研究中心 西安 710071)

摘 要 有效的特征表示对提升深度学习模型的表征能力和图像识别性能至关重要。例如,多尺度特征表示方法能够捕捉不同尺度的丰富信息,有助于提高深度学习模型的图像识别性能。然而,当前的多尺度深度学习方法仍存在对图像方向特征建模不明确的局限,导致对具有方向性目标的误识别。为了更好地表示图像中蕴含的多方向特征,本文提出了一种基于多尺度-多方向Transformer的网络框架(MSMDFormer)。首先,该框架中设计了一种能够捕获并增强多个方向特征的多方向特征编码器。在此基础上,本文联合了不同尺度的Gabor表征与多头注意力机制,设计了一种多尺度多方向Transformer编码器,以有效地聚合图像的多尺度和多方向特征。最后,该框架对卷积特征和多尺度-多方向特征进行融合,然后将融合特征用于图像识别。实验结果表明,MSMDFormer在CIFAR10、CIFAR100和SVHN数据集上分别取得了95.65%、77.46%和96.87%的整体准确率,在与19种基准方法的对比中显示出具有竞争力的图像分类性能。与11种图像分割基准方法相比,MSMDFormer在ADE20K数据集上展现出0.33%至6.58%mIoU的性能增益。综上所述,本文提出的MSMDFormer在深度学习图像识别任务中展现了卓越的特征表示能力,具有广泛的应用前景。另外,探索更有效的方向特征表示方法将成为未来研究的重要方向。

关键词 Transformer;多尺度;多方向;特征表示;图像识别
中图法分类号 TP393 **DOI号** 10.11897/SP.J.1016.2025.00249

Multiscale and Multidirectional Transformer-Based Image Recognition

YANG Yu-Ting[1),2),3)] Li Ling-Ling[1),2),3)] LIU Xu[1),2),3)] JIAO Li-Cheng[1),2),3)]
LIU Fang[1),2),3)] MA Wen-Ping[1),2),3)]

[1)](School of Artificial Intelligence, Xidian University, Xi'an 710071)
[2)](Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xi'an 710071)
[3)](International Research Center for Intelligent Perception and Computation, Xi'an 710071)

Abstract Effective feature representation is crucial for improving the representational capacity of deep learning models and their image recognition performance. For example, multiscale feature representation methods can capture abundant information at different scales, which helps improve the image recognition performance of deep learning models. However, current multiscale deep learning methods still have limitations in the unclear modeling of directional features in images, leading to misrecognition of directional targets. To better represent the multidirectional features inherent in images, this paper proposes a multiscale and multidirectional Transformer

收稿日期:2024-02-26;在线发布日期:2024-09-10。本课题得到国家自然科学基金重点项目(61836009,62431020)和国家自然科学基金联合基金项目(U22B2054)资助。 **杨育婷**,博士,助理研究员(博士后),主要研究领域为计算机视觉、深度学习与多尺度几何分析。Email: yangyuting@xidian.edu.cn。**李玲玲**,博士,副教授,主要研究领域为量子进化计算、机器学习和深度学习。**刘 旭**,博士,副教授,主要研究方向为机器学习和图像处理。**焦李成**(通信作者),博士,教授,中国计算机学会(CCF)会士、IEEE Fellow,主要研究领域为图像处理、机器学习与智能信息处理。Email: lchjiao@mail.xidian.edu.cn。**刘 芳**,硕士,教授,博士生导师,主要研究领域为人工智能和模式识别、机器学习、图像感知和场景理解、进化计算和数据挖掘。**马文萍**,博士,教授,主要研究领域为自然计算与智能图像处理。

(MSMDFormer) framework. It first designs a multidirectional feature encoder to capture and enhance features from multiple directions. Based on this, we integrate Gabor representations of different scales with multi-head attention mechanism to design a multiscale and multidirectional Transformer encoder, effectively aggregating images' multiscale and multidirectional features. Finally, the framework fuses convolutional features with multiscale and multidirectional features, and then utilizes the fused features for image recognition. Experimental results show that MSMDFormer achieves overall accuracies of 95.65%, 77.46%, and 96.87% on the CIFAR10, CIFAR100, and SVHN datasets, respectively, demonstrating competitive image classification performance compared to 19 benchmark methods. Compared to 11 benchmark methods for image segmentation, MSMDFormer exhibits a performance gain of 0.33% to 6.58% mIoU on the ADE20K dataset. In summary, MSMDFormer demonstrates outstanding feature representation capabilities in deep learning image recognition tasks with broad application prospects. Additionally, exploring more effective methods for representing directional features will be an important direction for future research.

Keywords Transformer; multiscale; multidirectional; feature representation; image recognition

1 引 言

近年来，卷积神经网络（Convolutional Neural Networks，CNN）和Transformer框架在深度学习图像识别领域取得了重要研究进展，推动了图像分类、目标检测和图像分割等应用的进步[1]。通过挖掘视觉数据中的潜在特征，深度学习网络模型能够进行有效的表示、学习和推理，从而有助于实现准确预测[2]。关于生物视觉皮层的识别特性研究，如稀疏性、方向性和多尺度性等，为研究者提供了特征表示方法的灵感，有助于挖掘潜在特征并提升深度学习模型的性能[3-4]。例如，稀疏的特征表示方法可以降低数据维度、提高模型泛化能力并更好地捕捉关键信息；具有方向性的特征表示方法则关注数据和特征的方向信息，有助于捕捉数据结构和特征；而多尺度特征表示方法则能在不同尺度下提取和表示图像特征，可以更全面地捕捉图像信息。

多尺度和多方向特征表示方法因其能够捕捉不同细节级别和不同方向上的图像信息而备受关注[5]。它们为视觉数据提供了更为全面的特征表示，使得深度神经网络能够提取具有判别力的特征并且能够有效应对尺度、旋转和视角的变化。其中，多尺度特征的重要性在于其能够捕捉局部和全局信息。通过考虑多个尺度，深度学习网络能够有效地捕捉细节和更广泛的上下文信息，从而理解目标对象或场景的层次结构。例如，在目标识别任务中，小尺度特征可以捕捉边缘和角点等细节信息，而大尺度特征则能够捕捉整体形状和结构。将多尺度特征融入深度学习网络框架能够提高深度网络在不同尺度和分辨率下鲁棒地识别目标对象的能力。与多尺度特征表示不同，多方向特征表示是一种捕捉视觉数据中方向信息的常用方法。待识别对象和结构通常呈现出特定的方向模式，如纹理、边缘或梯度，这些模式可以被多方向特征有效地捕捉[6]。通过考虑不同方向特征，深度学习网络可以编码图像元素之间的空间排列和关系，使其更好地理解对象的几何形状和空间布局。这在场景理解等任务中尤为重要，对图像识别结果起着关键作用。

在计算机视觉领域，研究者提出了一系列多尺度卷积神经网络和多尺度Transformer模型。以下是一些常见的相关模型，具体包括特征金字塔网络（Feature Pyramid Networks，FPN）[7]、空间金字塔池化（Spatial Pyramid Pooling，SPP）[8]、多尺度深度神经网络（Multi-Scale Deep Neural Networks，MSDNN）[9]、多尺度注意网络（Multi-Scale Attention Network，MSAN）[10]。其中，FPN先提取CNN中的多尺度特征图，然后通过逐步提高其分辨率，从粗到细捕获各层级的语义信息；SPP通过金字塔卷积将输入图像分解为多个尺度的子图像，并对各尺度子图像进行卷积操作，以提取多尺度特征。MSDNN由粗尺度网络和细尺度网络构成，其中粗尺度网络预测整个图像映射图，而细尺度网络则在局部对预测结果进行细化；MSAN则是联合多尺度表征与大尺寸核注意力，获得不同尺度的注意力图。与多尺度CNN不同，多尺度Transformer主要

通过引入多个不同尺度的注意力头和注意力子层来处理多尺度输入数据，这使得Transformer在视觉任务中的性能得到了进一步的提升[11-12]。

虽然上述多尺度特征表示网络能够捕获图像的丰富多尺度信息，但它们在图像方向特征建模方面仍然有所欠缺。这意味着，在处理具有明显方向特征的图像时，以上多尺度深度网络无法有效利用方向信息。例如，对于具有明显方向属性（如纹理、边缘或形状等）的待识别目标，缺乏对其方向属性特征的利用可能导致深度网络无法准确表示和识别该目标[13]。为有效缓解这一问题，我们进一步探索了如何将方向特征建模纳入多尺度特征表示，旨在增强深度网络对方向信息的感知和利用能力。

通过感知不同方向和视角下的特征变化，多方向特征具有一定的旋转不变性和视角不变性，从而提高了图像识别任务的鲁棒性。例如，Gabor滤波器滤波是一种常用于捕捉方向特征的方法，能够有效捕获图像中的特定方向特征[14]。通过使用多个方向和频率的Gabor滤波器可以获得输入图像的一组多方向特征响应图。而这些特征响应图中就蕴含了图像的不同方向模式的信息。在现有工作中，Gabor卷积神经网络方法也得到了发展，陆续出现了Gabor卷积网络GCN[15]、Perez等人提出的Gabor卷积层增强网络鲁棒性的方法[16]、自适应的Gabor卷积网AGCNs[17]等模型。它们将Gabor特征表示引入卷积神经网络，在一定程度上增强了卷积神经网络的模型性能。然而，虽然Gabor滤波器可以捕捉不同方向和频率的特征响应，但是对全局特征的表示能力较弱。

在现有研究中，Yeung等人提出的ABFormer中的边界感知空间注意模块能够获取边界特征和上下文特征，并将其作为自注意力计算的输入，从而实现对图像边界特征的关注[18]。在SF_MSFormer中，Yang等人提出的纹理增强器将获取的图像纹理特征和低频主要分量作为自注意力模块的输入，实现了对纹理特征的增强[11]。另外，P-MHSA[19]通过池化操作获取不同尺度的特征，结合多头注意力机制降低了序列长度，并获取了强健的上下文信息。可以发现，联合边界、纹理、尺度特征与Transformer模型可以实现更加强健特征表示。

与以往工作不同的是，本文尝试将多尺度多方向特征与Transformer自注意力联合并整合到神经网络中，以提高深度网络对多尺度多方向特征的表示能力。联合多尺度多方向的Gabor表征与多头注意力，不仅可以捕获多尺度多方向特征，还可以在一定程度上对Gabor特征进行全局建模。

本文提出了一种基于多尺度-多方向Transformer的深度网络框架，旨在联合方向与尺度特征进一步增强深度网络的特征表示能力。通过将方向特征引入自注意力机制中，实现方向特征的增强。另外，所提出的多尺度多方向Transformer编码器能够有效地整合图像的多尺度和多方向特征。最后，我们在CIFAR10、CIFAR100和SVHN这三个小型数据集上，以及ImageNet和ADE20K这两个大规模数据集上进行了实验验证。实验结果表明，本文提出的方法有效且在同类方法中展现出一定的性能优势，同时也显示出在大规模数据集和图像分割任务上的良好可扩展性。

总的来说，本文的主要创新和贡献如下：

（1）在自注意力机制中引入方向特征，设计了一种多方向特征编码器，更有效地表示与增强了图像的多方向特征；

（2）进一步联合Gabor表征与多头自注意力构建了多尺度多方向Transformer编码器，有效地聚合了图像的多尺度与多方向特征；

（3）提出了一种新颖的多尺度-多方向Transformer框架，实现了高性能的图像识别。在适当的参数量下，它实现了具有竞争力的图像分类性能并在ADE20K数据集上取得了0.33%至6.58%的mIoU图像分割性能增益。

2　相关工作

本文提出了一种多尺度-多方向的Transformer网络框架。该框架涉及卷积神经网络和Transformer、方向表征与Gabor卷积神经网络方面的相关工作。

2.1　卷积神经网络和Transformer

卷积神经网络（Convolutional Neural Networks，CNN）和Transformer是广泛应用于计算机视觉和自然语言处理任务中的深度学习模型[20-21]。受生物视觉系统启发，CNN通过多层的卷积与池化操作可以有效表示图像的局部特征。虽然它因具有良好的局部特征表示能力在图像识别领域占据了一定的主导地位，但是在处理远距离依赖（即全局特征表示）方面仍存在一定不足[22]。

与CNN不同，Transformer是一种无卷积的深度学习模型框架。它是一种基于自注意力机制的模

型，最初在机器翻译任务中取得了优异的性能表现。通过在输入序列内建立全局依赖关系，它能够有效地捕捉顺序数据中的远距离依赖[23]。它主要使用自注意力机制计算输入序列中不同位置之间的注意力权重，这使得其在处理顺序数据时更加有效[24]。它的结构核心是多头自注意力机制（Multi-Head Self-Attention，MHSA）和前馈神经网络。其中，MHSA机制允许模型在不同表示子空间中学习不同的特征[25]。

起初，Transformer在自然语言处理领域取得了显著成果。视觉Transformer模型ViT[26]的出现是Transformer在图像处理领域的初步尝试，并且在性能上可以与卷积神经网络相媲美。随后，大量的Transformer模型涌现，如Swin Transformer[27]、PvT[28]和TNT-S[29]等方法，在图像分类、目标检测、图像分割等领域展现了优越性能。研究者发现，将CNN和Transformer结合起来，可有效整合局部和全局的特征并实现更强健的特征表示，如LGLFormer[30]和HIRI-ViT[31]等工作。另外，一些研究者提出联合边界、纹理、尺度特征与Transformer模型可以实现更加强健特征表示，如SF-MSFormer[11]、ABFormer[18]和P-MHSA[19]等方法。

然而，以往的CNN和Transformer模型在处理缺乏方向特征的序列数据和具有方向性目标的图像时存在一定的局限性[32]。为了解决这一问题，研究者曾尝试利用图卷积网络（Graph Convolutional Neural Network，GCNN）和循环神经网络（Recurrent Neural Network，RNN）去处理缺乏方向特征的数据。这两种网络架构分别通过引入图结构和循环连接，有效地捕获了顺序数据中的方向特征和上下文依赖关系。具体地，GCNN使用图数据中的节点之间的连接传播信息，而RNN通过循环连接捕捉顺序关系。虽然GCNN和RNN能够一定程度上捕捉数据的方向特征，但是前者适合处理图结构数据，后者适合处理顺序数据，捕捉图像方向特征有限。然而，在图像识别中，对图像方向特征的建模不够明确，这可能导致深度网络对具有方向性目标的错误识别。因此，图像方向特征的表示仍需进一步研究。

总的来说，CNN和Transformer是深度视觉表征学习中的两种重要框架且各具优势。但是，它们在处理缺乏方向特征的图像数据时存在局限性。同时，方向特征对图像识别具有重要的研究意义。它们能够捕捉到图像中的纹理、结构和形状等信息，对于图像处理、特征提取、目标识别等任务起着关键作用。因此，本文探索了联合方向特征学习的Transformer，构建了多尺度多方向Transformer图像识别框架。

2.2 方向特征表示与Gabor卷积神经网络

方向表示神经网络是一种专门用于捕捉和处理方向性信息的神经网络架构[33]，它利用特定的机制捕捉并利用输入数据中的方向性特征。现有方向表示神经网络包含方向感知神经网络[34]、方向选择性神经网络[35-36]、方向卷积神经网络[37]等类型。具体来说，方向感知神经网络的结构设计旨在建模方向性信息。它往往通过方向感知滤波器和激活函数来提取并响应输入数据中的方向特征。方向选择性神经网络侧重点在于选择和增强特定的方向特征。它利用方向选择性滤波器，使神经网络对输入的特定方向特征的感知更加敏感[38]。方向卷积神经网络主要利用方向特征对CNN进行了进一步扩展[39]。它利用具有方向感知的特定卷积核和池化层，增强网络对方向性信息的感知能力。上述方向性特征表示学习的神经网络均适用于需要方向感知的深度学习任务，并被广泛应用于图像识别、纹理分析和目标检测等领域。

Gabor滤波器则是一种常用的方向感知滤波器，主要用于特定方向的特征提取[40]。Gabor卷积神经网络架构[15,41]通过在卷积层中采用Gabor滤波器作为卷积核来提取图像的特征。这些滤波器在不同方向上具有不同的响应，能够从图像中提取方向特征。Gabor神经网络通常由多个卷积层和池化层组成，逐步提取和组合图像的局部和全局特征。使用Gabor神经网络的主要优势在于对方向性和纹理特征的敏感性。由于Gabor滤波器具有方向选择性，网络能够更好地捕捉图像中的边缘、纹理和其他方向特征，从而提高图像识别和分类的准确性[42]。

在现有研究中，Luan等人率先将Gabor滤波器集成到深度CNN中，从而增强了深度卷积神经网络对特征方向和尺度变化的鲁棒性[15]。接着，Yuan等人提出了自适应Gabor卷积网络AGCNs[43]，将卷积核与Gabor滤波器自适应相乘，使得Gabor函数的尺度和方向等参数随神经网络训练一起学习。随后，Reyes等人提出嵌入Gabor滤波器的U-Net网络，有效地增强了深度学习特征的鲁棒性[44]。2023年，Fan等人提出了一种用于无监督视频目标分割任务的Gabor Transformer模型。他们构建的Gabor滤波Transformer模块能够有效地挖掘目标

的结构特征与纹理细节特征，进而显著提高视频物体分割的准确性[45]。总而言之，Gabor特征的引入能够增强神经网络对尺度和方向的敏感性，提升深度卷积神经网络在特征学习中的鲁棒性。此外，Gabor卷积神经网络在深度学习中的图像处理领域也得到了广泛应用。

3　多尺度-多方向Transformer网络

多尺度-多方向Transformer网络（Multi-Scale and Multi-Directional Transformer，MSMDFormer）通过联合多尺度和多方向的特征表示，增强深度网络的表征学习能力，以进行高性能的影像解译。为了实现多方向特征表示，本文构建了多方向特征编码器，可以捕捉并增强多个方向的特征。随后，考虑到不同尺度的方向特征的差异，本文进一步将多尺度特征与多方向特征联合起来构建了多尺度多方向Transformer编码器。本节重点介绍了MSMDFormer网络的整体框架，并详细描述了多方向特征编码器和多尺度多方向Transformer编码器的工作原理。

3.1　网络整体框架

以输入图像尺寸为32×32时的图像识别分类任务为例，本文提出的MSMDFormer整体网络框架如图1所示。MSMDFormer框架由卷积特征学习分支和多尺度多方向特征学习分支两部分组成。其中，卷积特征学习分支主要利用传统的ResNet 18骨干网络进行图像卷积特征学习。它包含ResNet 18卷积层以及常规的层1、层2、层3和层4。多尺度多方向特征学习分支设计主要通过级联多尺度多方向Transformer编码器，以捕获学习图像的多尺度多方向特征。可以看出，一个多尺度多方向Transformer编码器具有2倍的下采样效果。随后，该框架利用自注意力机制将两分支学习到的卷积特征与多尺度多方向特征进行计算，从而得到融合特征。最后，我们将融合特征送入全连接层，用于最终的图像识别分类。

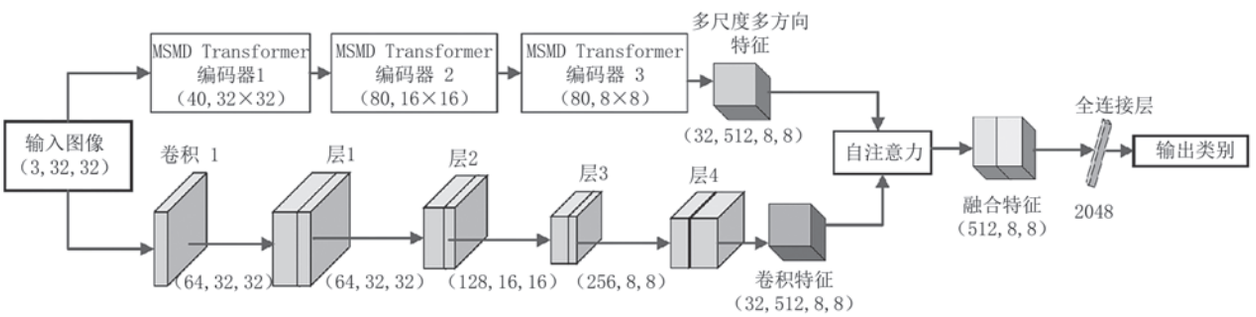

图1　基于多尺度-多方向Transformer的图像识别框架

通过上述步骤，MSMDFormer框架能够有效地学习并聚合图像的多尺度和多方向特征，增强深度卷积网络的特征表示能力，进一步提升图像识别性能。

3.2　多尺度多方向特征表示

本文提出的多尺度多方向特征表示学习由多个多尺度多方向Transformer编码器级联构成。其中，多尺度多方向Transformer编码器由不同尺度的Gabor滤波器组和多头自注意力模块构成。它首先考虑多方向特征学习，然后进一步考虑联合多尺度特征学习。具体地，我们构建了多方向特征编码器和多尺度多方向Transformer编码器。

3.2.1　多方向特征编码器

多方向特征编码器主要考虑将不同的方向特征引入图像理解中。在空间域上，Gabor滤波器的响应仅取决于局部区域的像素，它在图像中的各个位置都能独立地计算特征响应。在频率域上，Gabor滤波器通过调整频率和方向参数，可以选择性地响应特定频率和方向上的纹理特征。因此，Gabor表征存在缺乏全局性特征表示的局限。本文通过利用自注意力机制对方向特征进行进一步学习以弥补该局限。同时，受纹理增强编码器[11]的启发，如果将自注意力的输入适当设置为方向特征可实现特定方向的特征增强。如图2(a)所示，单方向特征编码器的输出DF可以表示为：

$$DF = Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Softmax((\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}})/\sqrt{d})\boldsymbol{V} \tag{1}$$

其中，$1/\sqrt{d}$表示一个缩放因子，$Attention()$操作表示自注意力计算操作，查询$\boldsymbol{Q}$、键$\boldsymbol{K}$和值$\boldsymbol{V}$设置为图像自身与图像的方向特征。具体而言，对于输入$\boldsymbol{X}$而言，$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$的设置如下：

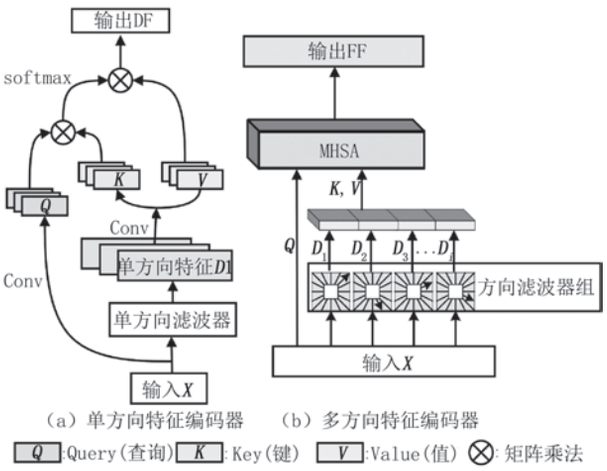

图2 单/多方向特征编码器

$$\boldsymbol{Q}=Conv(\boldsymbol{X}),\ \boldsymbol{K}=\boldsymbol{V}=Conv(\boldsymbol{D}1) \tag{2}$$

其中，$Conv()$表示常规卷积操作，$\boldsymbol{D}1$为通过特定方向滤波器获得的图像$\boldsymbol{X}$的某一个方向的特征。

在单方向特征编码的基础上，我们进一步考虑了多方向特征的表示学习。本文提出的多方向特征编码器的设计如图2(b)所示。对于输入图像$\boldsymbol{X}$，多方向特征编码器重点关注了多个方向特征，利用自注意力机制实现不同方向特征的增强。首先，通过方向滤波器组可以获得多方向特征MDF，可以表示为

$$MDF=\{\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{D}_3,\cdots,\boldsymbol{D}_i\}=DFB(\boldsymbol{X}) \tag{3}$$

其中，$\boldsymbol{D}_i$表示第i个方向特征，$DFB()$为方向滤波器组滤波操作。

进一步地，多方向特征编码器模块的输出特征FF可以表示为

$$FF=MHSA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}), \tag{4}$$

其中，$\boldsymbol{Q}=\boldsymbol{X},\boldsymbol{K}=\boldsymbol{V}=\boldsymbol{MDF}$。其中，多头自注意力$MHSA$的表达式为

$$MHSA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Cat}(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_i)\boldsymbol{W}^{\circ} \tag{5}$$

其中，Cat表示级联操作。$\boldsymbol{W}^{\circ}$表示用于将多个注意力头部输出进行线性变换和融合的权重矩阵。它能够将不同注意力头部的特征进行聚合，从而帮助模型学习有效的特征表示。其中，第i个头部自注意力特征$\boldsymbol{h}_i$表示为

$$\boldsymbol{h}_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^{Q},\boldsymbol{K}\boldsymbol{W}_i^{K},\boldsymbol{V}\boldsymbol{W}_i^{V}) \tag{6}$$

其中，$\boldsymbol{W}^{Q}$、$\boldsymbol{W}^{K}$、$\boldsymbol{W}^{V}$表示$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$的线性投影矩阵。

3.2.2 多尺度多方向Transformer编码器

Gabor滤波器在方向选择性、尺度适应性、局部化特性、生物学合理性和特征表示能力等方面具有优势，使其成为一种常用的多尺度多方向滤波器[43]。为聚合不同尺度与不同方向特征，本文联合Gabor滤波器与多头注意力机制进一步构建了多尺度多方向Transformer编码器。首先，使用一组Gabor滤波器在不同尺度下对输入图像进行卷积操作，从而提取多尺度的Gabor特征。不同尺度的Gabor滤波器组均包括多个方向(例如0°、45°、90°、135°等)的滤波器，可以捕捉图像不同方向的特征。其次，对于不同尺度的Gabor特征，使用多方向特征编码器对其进行特征增强。具体地，我们按照尺度$scale=1,2,3\cdots$依次对于输入进行n个方向的Gabor特征提取，然后将获得的$scale\times n$个多尺度多方向特征送入多头自注意力学习。多尺度多方向Transformer编码器可以对多个尺度多个方向上的特征进行建模和整合，从而提取更丰富的图像特征。具体过程如下：

首先，2D Gabor函数用于获取二维图像的方向特征，过程可以表示为

$$G_{s,\theta}(x,y;\sigma)=\frac{1}{2\pi\sigma^2}\bullet\mathrm{e}^{-\frac{1}{2}\left(\frac{x'^2}{\sigma^2}+\frac{y'^2}{s^2\sigma^2}\right)} \tag{7}$$

其中，

$$\begin{aligned}x'&=x\cos\theta+y\sin\theta\\ y'&=-x\sin\theta+y\cos\theta\end{aligned} \tag{8}$$

其中，x和y分别表示图像像素的水平和垂直坐标。σ表示高斯滤波器的标准差，s表示尺度因子，θ表示方向系数。

对于2D图像而言，v个尺度u个方向的Gabor卷积定义如下：

$$\boldsymbol{C}_{i,u}^{v}=\boldsymbol{C}_{i,o}\circ\boldsymbol{G}_{u,v} \tag{9}$$

其中，$\boldsymbol{C}_{i,o}$表示传统卷积操作，$\boldsymbol{G}_{u,v}$表示v个尺度u个方向的Gabor滤波器系数，$\circ$表示卷积计算。

如图3所示，多尺度多方向Gabor编码器输出的多尺度多方向特征可以表示为$MSMDF=MSMDGabor(\boldsymbol{X})$，其具体可以表示为

$$MSMDF=MHSA(\boldsymbol{Q}=\boldsymbol{X},\boldsymbol{K}=\boldsymbol{V}=\boldsymbol{C}_{i,u}^{v}\circ\boldsymbol{X}) \tag{10}$$

进一步地，多尺度多方向Transformer编码器可以对输入图像X进行编码，输出多尺度多方向特征OF。它可以表示为$OF=LN(LN(\boldsymbol{X}+MSMDGabor(\boldsymbol{X}))+FFN(LN(\boldsymbol{X}+MSMDGabor(\boldsymbol{X}))))$。其中，$LN$表示层归一化操作；$FFN$表示前馈网络；+表示逐元素相加操作。

4 实验结果与分析

本节展示了实验结果及相关分析，验证了所提

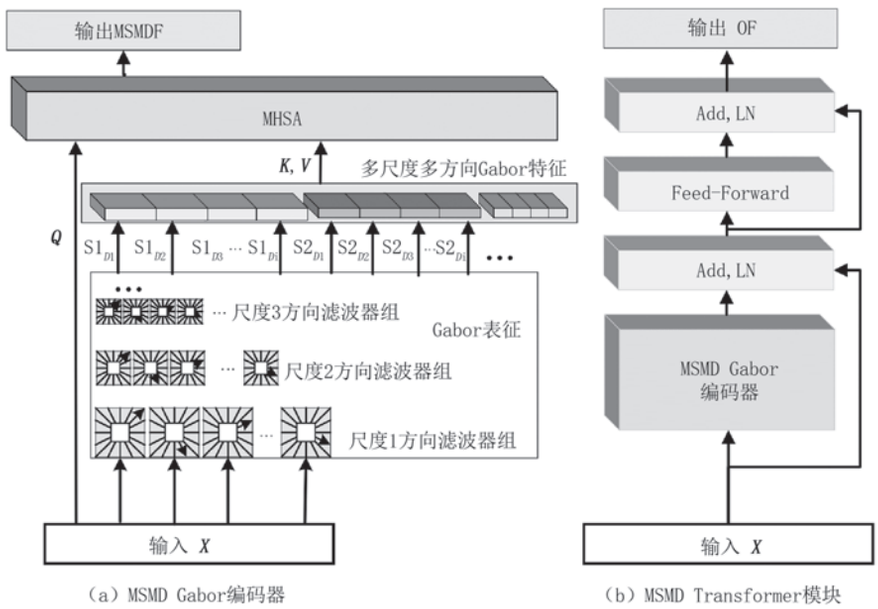

图3　多尺度-多方向Transformer特征编码器

出的多尺度多方向Transformer网络的有效性。此外，我们进行了多项消融实验，以展示多尺度多方向编码器模块的优势，并探讨了不同组成部分对网络整体性能的影响。

4.1　实验设置

本小节主要介绍了实验数据集、评估指标、设置细节和主要对比方法。

4.1.1　数据集

本文首先在CIFAR10、CIFAR100和SVHN数据集上对提出方法进行了验证。其中，CIFAR10和CIFAR100数据集是由Krizhevsky等人[①]提出，均包含60 000张训练图像。其中，CIFAR10中的图像有10类，而CIFAR100中的图像有100类。街景数据集SVHN包含了73 257个训练样本和26 032个测试样本，类别数为10，具体对应数字0到9。以上提及的三个数据集的图像分辨率均为32×32。

另外，扩展实验中使用的ImageNet和ADE20K数据集的图像尺寸均为224×224。其中，ImageNet数据集规模庞大、多样，并且提供了丰富的标签信息。该数据集包含约1000个类别，各类别都有大约1000张标记的图像。其图像来自各种场景和领域，包括动物、物体、自然景观、日常生活等。ADE20K数据集是图像分割任务的常用实验数据集之一，涵盖了150个目标对象的类别。该数据集包含超过20 000张训练图像、2000张验证图像和3000张测试图像，涵盖了多种室内和室外环境、物体和场景。每张图像都对应给出了对象的类别信息和像素级语义分割标签，具有大量的注释信息。

4.1.2　评价指标

关于实验性能评价，本文使用了常用的整体准确率(Overall Accuracy，OA)作为模型图像分类性能评价指标。为了评估模型的复杂性，实验中统计了模型参数数量。通过使用这些指标，我们可以评估提出方法的性能和复杂性。OA具体可以表示为

$$OA = T / TT \tag{11}$$

其中，T为正确分类的样本数，TT为总样本数。

除此以外，分割任务评价中采用了平均交并比(mIOU)和像素准确率(Pixel Accuracy)作为评价指标。其中，$mIOU$可以表示为

$$mIOU = \frac{1}{N}\sum_{i=1}^{N} IoU_i \tag{12}$$

其中，N为类别数，IoU_i表示第i个类别的IoU值。其中，$IoU = \frac{预测分割结果 \cap 真实分割掩码}{预测分割结果 \cup 真实分割掩码}$，$\cap$表示交集，$\cup$表示并集。另外，像素准确率的计算方式为

$$PixelAccuracy = \frac{CP}{TP} \tag{13}$$

其中，CP代表正确预测的像素数量，TP代表总像素数量。

① Learning Multiple Layers of Features from Tiny Images, http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf, 2009,4,8.

4.1.3 实验细节与损失函数

实验中使用了随机梯度下降(Stochastic Gradient Descent,SGD)优化器对网络进行优化学习。CIFAR和SVHN数据集训练迭代次数为300个epochs,初始学习率为0.03,在第150和200个epoch时进行学习率衰减。ImageNet数据集的训练迭代次数为90个epochs。另外,ADE20K数据集的训练迭代次数为120个epochs。实验中的批量大小(Batch Size)均设置为32。关于损失函数,本文提出方法在训练过程中使用了标准的交叉熵损失。另外,所有实验均在Nvidia Tesla V100 4xGPU和PyTorch 1.7.0的平台环境中进行。

4.1.4 对比方法

在图像分类实验中,我们主要选取ResNet18[46]基准网络、ResNet50以及基于Garbor的神经网络模型作为对比算法。其中,基于Garbor的神经网络模型包括了使用Gabor滤波器和伪逆学习自编码器实现快速图像识别的GF+PILAE[47]方法、Gabor卷积网络GCN[15]、混合Gabor卷积网络HGCN[48]、深度Gabor卷积网Gimg+Conv[49]、Perez等人提出的Gabor卷积层增强网络鲁棒性的方法[16]、基于可学习Gabor参数的Gabor层(即卷积层)替换各种深度架构的前几层的可变Gabor特征网络DGFNs(R18)[50]、使用Gabor滤波器作为稀疏表征输入后续网络的Gabor-AlexNet[51](6个方向)、自适应的Gabor卷积网AGCNs(5x5)[43]、Gabor散射网络Scat+WRN(Gabor)[17]。和Rivas等人提出的Gabor滤波器初始化的卷积网络[52]。除此以外,本文还给出了一些Transformer相关模型作为对比方法,包括了经典的ViT-Small[26]、Swin-T[27]、PvT-Small[28]等模型。同时,本文还选取了MLP模型[53]、扩散模型[54]和VNAS模型[55]三种框架方法作为对比方法。关于图像分割,本文选取了2017-2024年一些经典的模型,如多路径细化网络RefineNet[56]、特征金字塔PSPNet[57]、自适应尺度卷积网络SAC[58]、上下文编码网络EncNet[59]、动态结构语义传播网络DSSPN[60]、多任务的统一感知解析网络UperNet[61]、逐点空间注意网络PSANet[62]、多流密集连接网络MDCN[63]、多阶段上下文细化网络MCRNet[64]、基于公式的监督学习方法SegRCDB[65]、多池化上下文网络MPCNet[66],作为对比方法。

4.2 实验结果

在实验中,我们将MSMDFormer模型与Gabor卷积网络及一些与Transformer相关的方法进行了性能对比。表1详细列出了这些方法在SVHN、CIFAR10和CIFAR100数据集上的图像分类精度,以及统计了模型在CIFAR100数据集上的训练参数量。实验结果表明,本文提出的MSMDFormer模型在SVHN、CIFAR10和CIFAR100数据集上性能均优于ResNet18模型,分别实现了1.20%、1.13%和1.92%的整体精度提升。与ResNet18和ResNet50通过增加层数以提升性能的方式不同,MSMDFormer引入了多尺度多方向的表示学习分支。在参数量少于ResNet50的情况下,所提出的方法仍然展示出更加优异的模型性能。与HGCN、Gabor-AlexNet、AGCNs等Gabor相关的表征方法相比,MSMDFormer具有明显的性能优势。在相同实验平台、训练迭代次数和学习率衰减设置下,MSMDFormer与微调后的GCN模型(GCN*)实现了具有竞争力的性能,在SVHN和CIFAR10数据集上展现出2.48%和1.68%的整体精度增益。除此以外,与经典的MLP模型[53]、扩散模型[54]和VNAS模型[55]相比,提出的方法也依然展现了具有一定竞争力的性能。

4.3 消融实验

为了探究不同模块和参数对MSMDFormer网络模型的影响,本文进行了多项消融实验,涉及多尺度多方向编码器的数量、嵌入位置与方式、不同特征融合方式、多头注意力头数量,以及Gabor表征的尺度与方向数量等方面。

4.3.1 多尺度多方向编码器数量对模型性能的影响

本文提出的多尺度多方向特征学习分支由多尺度多方向Transformer编码器级联构成。一个多尺度多方向Transformer编码器可以实现特征的2倍下采样。实验选择的数据集输入尺寸为32×32,可进行多次下采样。这也就意味着我们可以选取多个多尺度多方向特征编码器级联来获取多尺度多方向特征。因此,我们对于不同编码器数量进行了消融实验(无MLPs),实验结果如表2所示。

实验结果表明,当MSMD编码器数量为2时候,MSMDFormer在SVHN、CIFAR10和CIFAR100上分别展示出96.68%、94.96%和76.39%的整体精度。同时,编码器数量为2时比编码器数量为1或者3时的模型具有更高的整体精度。这可能是因为编码器在经历两次编码之后获得了更多尺度和更多方向的特征。在进行第三次尺度分解后的图像

表1　本文方法与其他方法性能对比

方法 \ 数据集	出版者(年份)	SVHN	CIFAR10	CIFAR100	参数量(M)
ResNet18 [46]	ICCV (2016)	95.67	94.52	75.54	11.20
ResNet50 [46]	ICCV (2016)	96.30	95.20	77.30	22.56
GF + PILAE [47]	ICONIP (2018)	-	47.02	-	-
GCN* [15]	IEEE TIP (2018)	94.39	93.97	77.20	17.60
HGCN [48]	PR Letters (2018)	96.40	93.89	73.02	4.50
Gimg+Conv [49]	Neurocomputing (2020)	-	61.26	-	-
Perez et al. [16]	ECCV (2020)	96.70	91.35	76.86	-
DGFNs(R18) [50]	WACV (2021)	-	91.03	-	3.40
Gabor-AlexNet(6d) [51]	ICICSP (2021)	-	84.23	-	-
ViT-Small [26]	ICLR (2021)	-	81.90	47.40	48.80
TNT-S [29]	NeurIPS (2021)	-	85.80	71.50	23.80
Swin-T [27]	ICCV (2021)	-	89.80	74.80	27.50
PvT-Small [28]	ICCV (2021)	-	91.10	76.30	24.50
Rivas et al. [52]	ICCVW (2023)	-	80.41	72.00	-
MLP Mixer [53]	NeurIPS (2021)	-	95.63	76.31	19.00
AGCNs(5x5) [43]	PR (2022)	94.82	89.78	-	1.20
Scat + WRN (Gabor) [17]	NMITCON (2023)	-	90.50	-	17.02
Wang et al. [54]	PMLR (2023)	95.56	**95.74**	75.22	17.02
VNAS [55]	IJCV (2024)	-	94.35	73.07	3.50
MSMDFormer	2024	**96.87**	95.65	**77.46**	18.87

注：*代表微调后的模型，最佳性能已加黑

特征中虽然尺度和方向特征增加，但是存在较多的低分辨率特征，在一定程度上并不有利于最终图像分类。总体而言，当编码器数量为2或3时，模型的性能优于编码器数量为1时的表现。可以发现，并非多尺度多方向特征编码器的数量越多越好。在实际实验时需要选择合适的编码器数量以获得较好的模型性能。

表2　不同MSMD编码器数量时的模型性能对比

编码器数量	SVHN/%	CIFAR10/%	CIFAR100/%	参数量/M
1	96.23	94.45	73.24	18.71
2	**96.68**	**94.96**	**76.39**	18.92
3	96.56	94.88	75.89	19.56

4.3.2　不同嵌入位置与方式对于模型性能的影响

为了研究多尺度多方向特征在不同的嵌入位置或嵌入方式上的模型性能，我们主要设计了三种不同的模型结构。三种模型结构具体涉及：将MSMD特征模块嵌入CNN之前的位置、在卷积网络层2和层3之间嵌入MSMD特征模块和如图1所示的两分支学习方式。实验模型性能如表3所示。

实验结果表明，与基准CNN网络相比，将多尺度多方向特征作为CNN的先验输入的学习方式使

表3　MSMD特征在不同嵌入位置时的模型性能对比

嵌入方式	SVHN/%	CIFAR10/%	CIFAR100/%	参数量/M
基准CNN	95.67	94.52	75.54	11.20
CNN前	96.68	95.37	75.96	13.33
层2与层3之间	96.37	94.49	74.13	43.98
MSMDFormer	**96.87**	**95.65**	**77.46**	18.87

得模型在SVHN、CIFAR10和CIFAR100数据集上分别获得了1.01%、0.85%和0.42%的整体精度提升。在层2和层3之间的位置嵌入提出的多尺度多方向Transformer编码器时，在SVHN、CIFAR10和CIFAR100数据集上分别展现了96.37%、94.49%和74.13%的整体准确率。结果表明，MSMDFormer框架将多尺度方向特征和卷积特征进行两分支融合时的模型在SVHN、CIFAR10和CIFAR100数据集上分别展现了96.87%、95.65%和77.46%的整体准确率。这在一定程度上表明，将多尺度多方向特征作为独立的学习分支能够有效帮助卷积神经网络提升模型性能。

4.3.3　不同特征融合方式对于模型性能影响

特征融合是提出MSMDFormer的重要组成部分，旨在将多尺度、多方向的特征表示与卷积神经网络的特征进行有效融合，以实现最终的图像分类。

在实验过程中，我们主要考虑了级联和自注意力两种特征融合方式对于模型性能的影响。采用不同融合方式时的模型性能如图4所示。

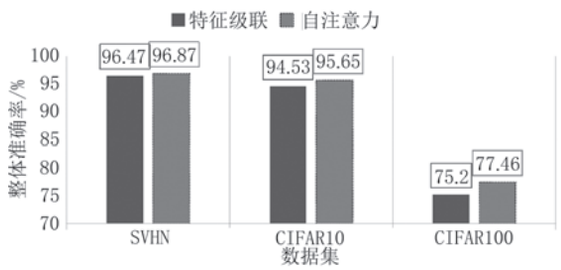

图4 不同融合方式的模型分类精度对比

实验结果表明，采用特征级联方式进行特征融合时的模型在SVHN、CIFAR10和CIFAR100数据集上分别达到了96.47%、94.53%和75.20%的整体精度；而使用自注意力机制进行特征融合的模型在这三个数据集上的整体精度分别为96.87%、95.65%和77.46%。相比之下，使用自注意力机制进行特征融合的模型性能优于采用级联方式的模型性能。

4.3.4 多头注意力的头部数量对于模型性能影响

多尺度多方向编码器结合了自注意力机制中的关键模块—多头注意力(MHSA)。在实验中，MHSA的头部数量对模型性能也会产生影响。为了探究这一参数对模型性能的具体影响，我们进行了相应的消融实验。图5呈现了四种MHSA头部数量时的模型整体准确率。

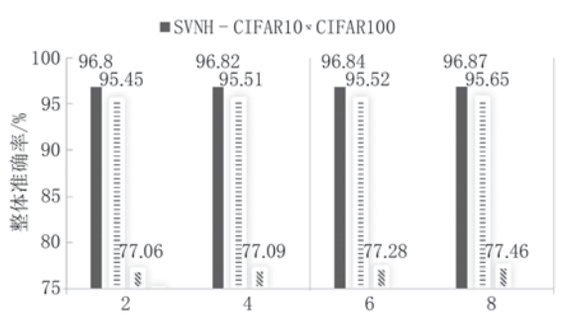

图5 不同MHSA头部数量时的模型整体准确率

实验结果表明，MHSA头部数量为2时，MSMDFormer模型在SVHN、CIFAR10和CIFAR100数据集上分别展现了96.80%、95.45%和77.06%的整体精度。MHSA头部数量为4的时候，模型在SVHN、CIFAR10和CIFAR100数据集上分别展现了96.82%、95.51%和77.09%的整体精度。与头部数量为4的时候相比，头部数量为6和8的时候的模型分别在SVHN、CIFAR10和CIFAR100数据集上展现了相对稳定的小幅度的整体精度提升。这主要是因为MHSA头部数量增加可以使得模型能够细致地关注输入的不同部分，捕捉到更多的局部特征和上下文信息，从而实现更好的表示能力，提高识别性能。

4.3.5 Gabor表征尺度与方向数对模型性能的影响

通过Gabor滤波器组可以得到多个不同方向的特征，因此通过提出的多尺度多方向编码器可以实现对于多个不同方向的特征学习。利用Gabor滤波器进行特征表示时候，我们可以发现不同方向的特征不同，但是互补的两个方向的特征相似。因此，在对于Gabor表征方向数量选择时，我们主要选取了四个方向的特征进行学习。为探究不同尺度、方向数量对于模型性能的影响，我们列出了几种不同尺度-方向数量时候的模型性能，具体如表4所示。

表4 不同尺度-方向数时的模型性能对比

尺度数与方向数	SVHN/%	CIFAR10/%	CIFAR100/%
尺度数=1 & 方向数=2	96.7	95.65	77.13
尺度数=1 & 方向数=4	96.8	95.45	77.06
尺度数=1 & 方向数=6	96.84	95.53	77.35
尺度数=4 & 方向数=1	96.83	**95.71**	77.27
尺度数=6 & 方向数=1	96.79	95.65	77.28
尺度数=6 & 方向数=4	**96.87**	95.65	**77.46**

实验结果表明，当Gabor表征尺度数为1方向数量为4时，模型在SVHN、CIFAR10和CIFAR100数据集上分别展现了96.80%、95.45%和77.06%的整体精度。当Gabor表征尺度数为1方向数量为6时，模型在SVHN、CIFAR10和CIFAR100数据集上分别展现了96.84%、95.53%和77.35%的整体精度。在尺度数为1时，对比方向数为2、4和6时候的模型性能可以发现，随着方向数的增加，模型性能提升。当尺度数为6时，方向数为4时候的模型比方向数为1时候的模型性能有所提升。当Gabor表征尺度数为6方向数量为4时候，模型在SVHN、CIFAR10和CIFAR100数据集上分别展现了96.87%、95.65%和77.46%的整体精度。相比之下，尺度数与方向数增加，模型性能有少许增益。

4.3.6 多尺度特征与多方向特征对模型性能的影响

MSMDFormer聚合了多尺度和多方向特征，为

探究多尺度特征与多方向特征谁对模型提升的作用更大，我们进行了进一步的消融实验。具体实验结果如图6所示。实验结果表明，与基准模型相比，增加多尺度特征后的模型在SVHN、CIFAR10和CIFAR100数据集上分别展现了96.80%、95.45%和77.06%的整体精度，增加多方向特征后的模型在以上三个数据集上分别展现了0.99%、1.13%和1.62%的整体精度提升。与同时增加多尺度特征和多方向特征后的模型精度相比，在CIFAR100和SVHN数据集上，多方向特征对于最终提出的MSMDFormer模型性能的影响较大。总的来说，多尺度特征和多方向特征对于特征表示都具有重要的影响，而具体的影响程度会因应用场景和任务的不同而有所差异。

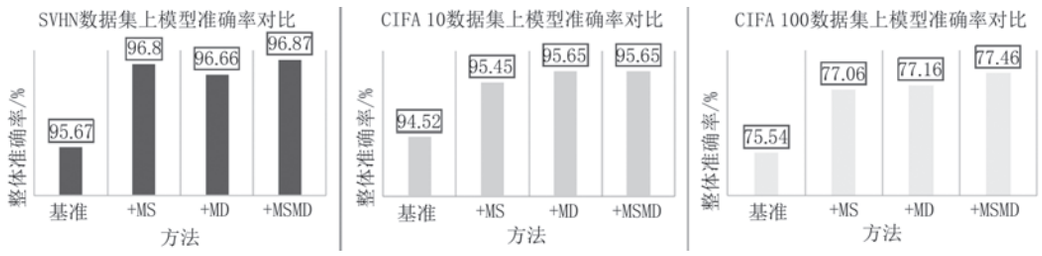

图6　多尺度与多方向特征对模型性能的影响

4.3.7　各模块对模型性能的影响

本文提出的网络框架主要在基准CNN网络上增加了多尺度多方向特征学习分支、自注意力融合模块和用于分类的MLPs。在这里，我们对以上几个模块进行了消融实验。几种不同模块添加到基准神经网络模型中的最终的模型性能如表5所示。

表5　各模块对模型性能的影响

模块	SVHN/%	CIFAR10/%	CIFAR100/%
基准CNN	95.67	94.52	75.54
＋MLPs	96.68	95.37	75.96
＋MLPs＋MSMD特征学习分支	**96.87**	**95.65**	**77.46**

实验结果表明，基准CNN在SVHN、CIFAR10和CIFAR100数据集上分别展现了95.67%、94.52%和75.54%的整体精度。将简单的全连接FC层换成由两层FC构成的MLPs用于最终分类时，模型在SVHN、CIFAR10和CIFAR100数据集上分别展现了96.68%、95.37%和75.96%的整体精度。然后，同时增加MLPs和多尺度多方向编码器分支时构成的模型在SVHN、CIFAR10和CIFAR100数据集上分别展现了96.87%、95.65%和77.46%的整体准确率。可以看出，多尺度多方向特征学习分支对基准网络模型的性能有着明显的提升效果。

4.4　扩展实验

为了进一步验证提出模型在大规模数据集上的有效性，我们在ImageNet上进行了实验。其模型性能如图7所示。可以看出，ResNet18基准模型在ImageNet数据集上展现了66.49%的Top1准确率。将多尺度多方向特征学习分支集成到模型中后，其在ImageNet数据集上的Top1准确率提高了2.03个百分点，达到了68.52%的整体准确率。这表明有效引入多尺度多方向特征能够提升卷积神经网络在大规模数据集上的图像识别性能。

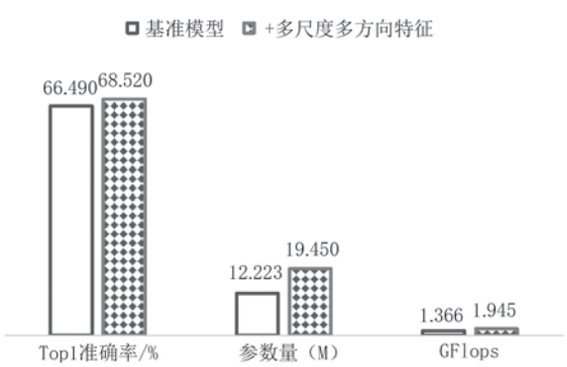

图7　提出方法与基准模型在ImageNet数据集上的性能对比

除此以外，为了进一步验证提出模型在其它视觉任务上的有效性，我们在图像分割数据集ADE20K上进行了实验。模型性能如表6所示，其中R152与R101分别表示骨干网络为ResNet152和ResNet101，Swin表示骨干网络为Swin Transformer。实验结果显示，本文提出的方法在ADE20K数据集展现了44.98%mIOU与82.86%的像素准确率。与现有的图像分割模型（如RefineNet[56]、PSPNet[57]、SAC[58]、EncNet[59]、DSSPN[60]、UperNet[61]、PSANet[62]、MDCN[63]、MCRNet[64]、SegRCDB[65]和

MPCNet[66])相比，MSMDFormer展现出了一定的性能优势。与其中的SAC[58]、EncNet[59]、MDCN[63]方法相比，本文提出的方法在mIOU性能上仍展现出竞争力，同时其像素准确率性能优于其他方法。这也表明了提出方法在图像分割任务上的有效性和泛化性。

表6 ADE20K数据集上的模型分割性能对比

方法	mIOU/%	Pixel. Acc/%
RefineNet(R152)[56]	40.70	-
PSPNet(R101)[57]	43.29	81.39
SAC(R101)[58]	44.30	81.86
EncNet (R101)[59]	44.65	81.69
DSSPN(R101)[60]	43.68	81.13
UperNet(R101)[61]	42.66	81.01
PSANet(R101)[62]	43.77	81.51
MDCN(R101)[63]	44.06	-
MCRNet(R101)[64]	41.25	**88.64**
SegRCDB(R101)[65]	39.56	51.48
SegRCDB(Swin)[65]	41.51	52.58
MPCNet(R101)[66]	38.04	82.55
MSMDFormer(R101)	**44.98**	82.86

4.5 特征可视化与收敛性分析

为了进一步理解并明晰本文提出的MSMDFormer网络特征表示效果，我们对于模型的特征进行了可视化。以CIFAR10中的样本“马”为例，我们对于Gabor滤波器与其提取的图像特征进行可视化。图8(a)所示为Gabor滤波器在[0，Π]上的8个方向6个尺度的基特征表示。图8(b)给出了由图8(a)Gabor滤波器组滤波后的图像特征。可视化特征表明，方向与尺度不同的Gabor滤波器捕获的图像特征不同，这将对图像识别结果产生的影响不同。具体地，从视觉观测角度可以看出，方向为Π/2时候的Gabor滤波器表示的特征更加完备，更加利于模型对示例样本数据的识别。与其它尺度时的Gabor滤波器相比，尺度为1时的Gabor滤波器可以获得更加丰富的特征。

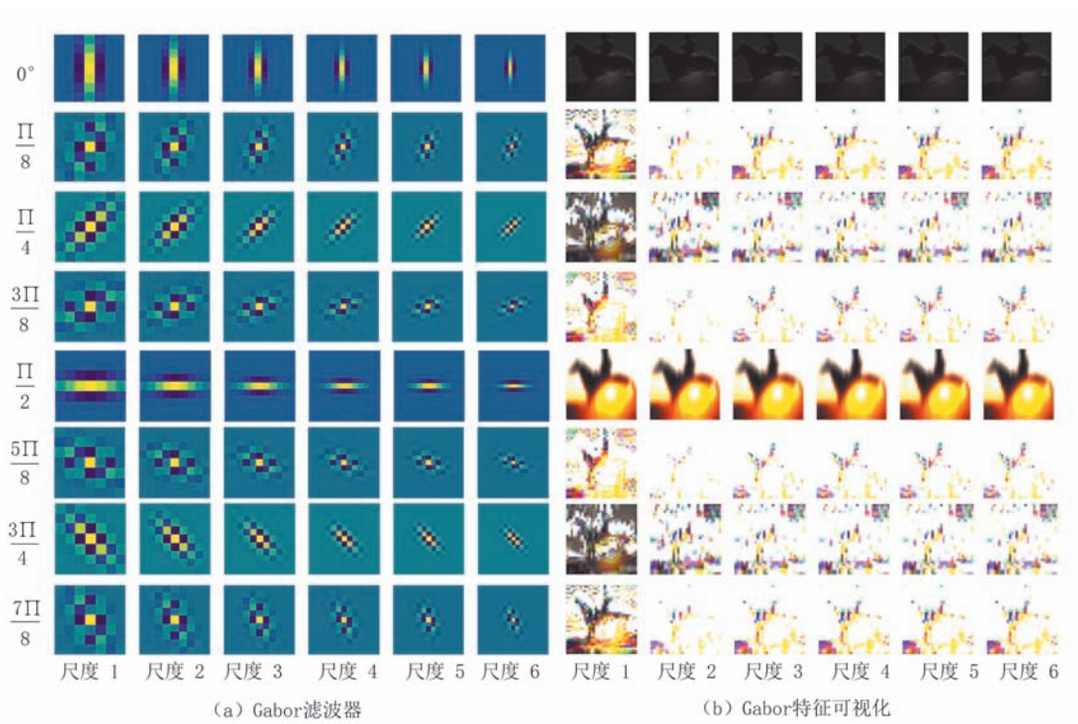

图8 Gabor滤波器基特征与Gabor特征可视化

另外，本节给出了在CIFAR10数据集上训练的模型的卷积分支特征、多尺度多方向编码器特征以及将两者融合的模型特征图的可视化结果。如图9所示，卷积特征能够有效地学习图像中的目标，并将注意力集中在目标识别上，但仍存在背景干扰的问题。相比之下，多尺度多方向编码器特征虽然没有明显的注意力效果，但能够突出显示待识别的物体。通过结合这两种特征，模型能够有效聚焦于待识别的目标物体，从而促进最终的图像识别。

为呈现模型收敛性，我们在此给出了MSMDFormer

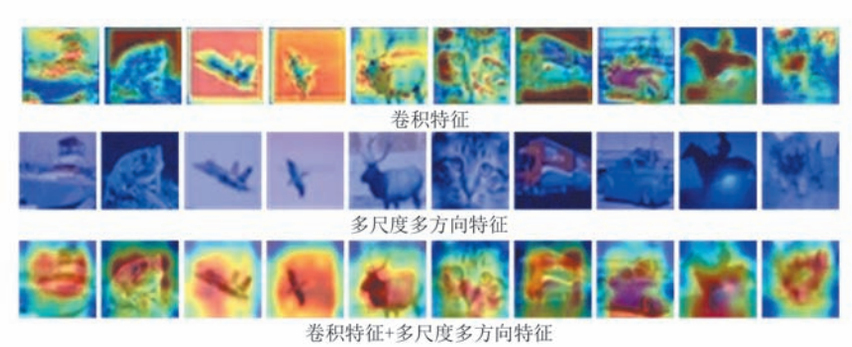

图9　MSMDFormer模型的两分支特征类别激活图

与基准 Baseline 在 CIFAR10 和 CIFAR100 数据集上的训练损失曲线和验证集精度曲线。如图10所示，与基准网络相比，MSMDFormer展现出更快的收敛速度和明显的精度优势。这在一定程度上也显示出本文提出的多尺度多方向特征表示分支对模型收敛速度和精度的积极影响。

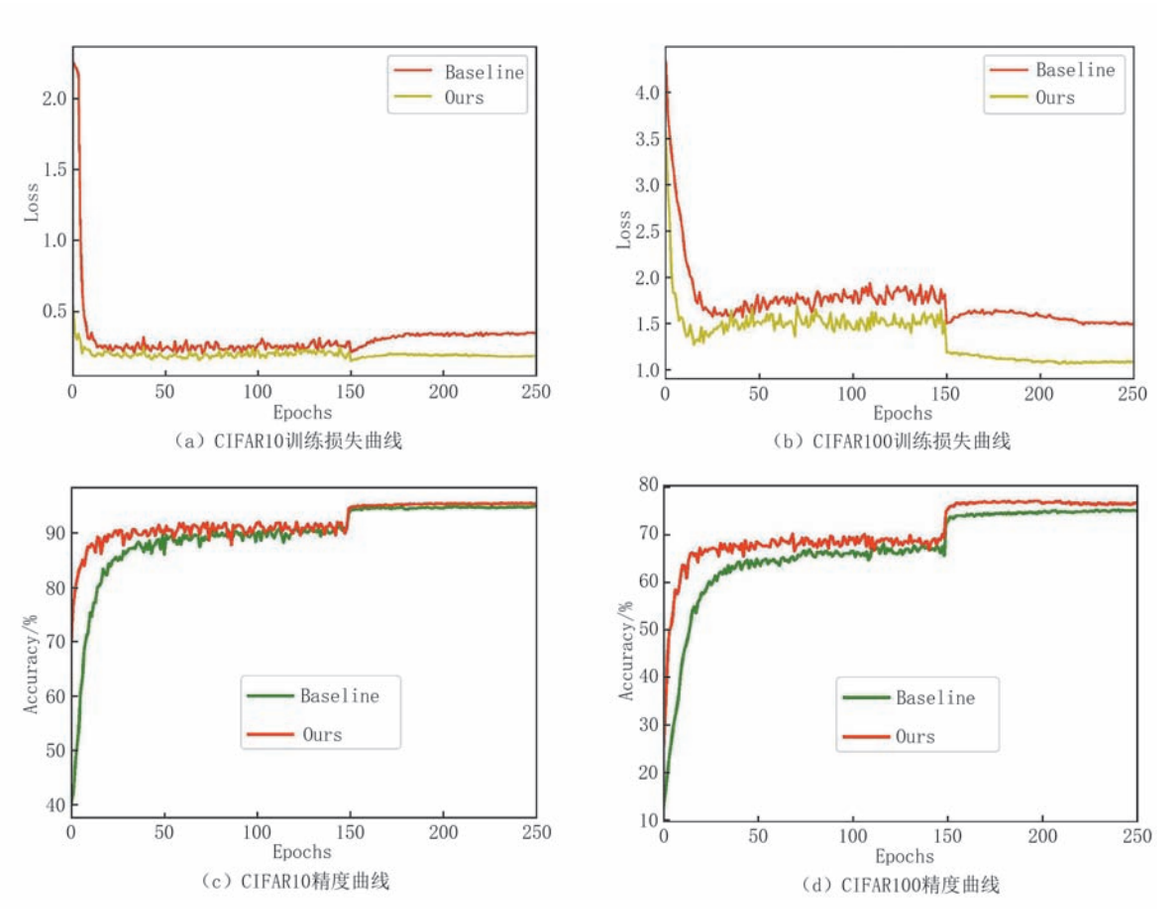

图10　MSMDFommer与基准网络的损失/精度曲线

5　总结与讨论

本工作以构建强健的特征表示网络为目标，探索了多尺度多方向特征表示在图像识别任务中的重要性与应用。具体地，本文提出了一种基于多尺度-多方向 Transformer 的深度学习网络框架。其中，提出的多方向特征编码器能够捕捉不同方向和多方向特征。实验在自然场景数据集 CIFAR10、CIFAR100 和街景数据集 SVHN 上进行了验证。实

验结果表明，MSMDFormer方法能够有效地提高卷积神经网络的图像识别准确性，并展现出一定的性能优势。同时，该方法可以进一步扩展到大规模数据集，如ImageNet和ADE20K图像分割数据集，以验证其在大规模数据集上的有效性以及在图像分割任务上的泛化性。总之，本文提出的基于多尺度-多方向Transformer的深度网络框架在深度图像识别任务中展现出了强大的特征表示能力。我们的研究为深度学习领域的图像识别任务提供了一种有益的方法，具有广泛的应用前景。

当然，在本研究过程中，我们发现模型仍然存在一些局限性。例如，该模型无法直接应用于大尺寸数据集（如常见的尺寸为224×224图像数据集）。这主要是因为提出方法中使用了自注意力计算。因此，本文在实验过程中主要探索了模型在小尺寸数据集上进行了性能验证与消融实验。如何使得模型更好地适应大尺寸数据集的训练和测试仍然是有待研究的问题。另外，Gabor表征只是获取方向特征的一种较为简单有效的实现方式，因此探究其他方向特征表示方法也将是我们未来的研究方向之一。

参 考 文 献

[1] Zhang S, Gong Y H, Wang J J. The development of deep convolution neural network and its applications on computer vision. Chinese Journal of Computers, 2019, 42(3): 453-482 (in Chinese)
(张顺，龚怡宏，王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用. 计算机学报, 2019, 42(3): 453-482)

[2] Duan Y Q, Zheng Y, Lu J W, et al. Structural relational reasoning of point clouds//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019, 949-958

[3] Liu X, Li L L, Liu F, et al. Scattering graph convolutional network-based PolSAR image classification. SCIENTIA SINICA Informationis, 2022, 52(10): 1900-1914 (in Chinese)
(刘旭，李玲玲，刘芳 等. 基于散射图卷积网络的PolSAR影像地物分类. 中国科学：信息科学, 2022, 52(10): 1900-1914)

[4] Fan D P, Ji G P, Qin X B, et al. Cognitive vision inspired object segmentation metric and loss function. SCIENTIA SINICA Informationis, 2021, 51(9): 1475-1489 (in Chinese)
(范登平，季葛鹏，秦雪彬 等. 认知规律启发的物体分割评价标准及损失函数. 中国科学：信息科学, 2021, 51(9): 1475-1489)

[5] Jiao L C, Gao J, Liu X, et al. Multi-scale representation learning for image classification: A survey. IEEE Transactions on Artificial Intelligence, 2021, 4(1): 23-43

[6] Zhang W H, Jiao L C, Liu F, et al. Adaptive contourlet fusion clustering for sar image change detection. IEEE Transactions on Image Processing, 2022, 31: 2295-2308

[7] Lin T Y, Doll'ar P, Girshick R, et al. Feature pyramid networks for object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017, 2117-2125

[8] He K M, Zhang X Y, Ren S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916

[9] Ren W Q, Liu S, Zhang H, Pan J S, et al. Single image dehazing via multiscale convolutional neural networks// Proceedings of the 14th European Conference on Computer Vision, 2016. Amsterdam, The Netherlands, 2016: 154-169

[10] Wang L, Shen J, Tang E, Zheng S N, et al. Multi-scale attention network for image super-resolution. Journal of Visual Communication and Image Representation, 2021, 80: 103300

[11] Yang Y T, Jiao L C, Liu F, et al. An explainable spatial-frequency multi-scale transformer for remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-15.

[12] Lin C H, Shen C, Deng J Y, et al. Digitally forged face content creation and detection. Chinese Journal of Computers, 2023, 46(3): 469-498 (in Chinese)
(蔺琛皓，沈超，邓静怡 等. 虚假数字人脸内容生成与检测技术. 计算机学报, 2023, 46(3): 469-498)

[13] Lv Z Q, Cheng Z S, Li J B, et al. Treecn: Time series prediction with the tree convolutional network for traffic prediction. IEEE Transactions on Intelligent Transportation Systems, 2023, 3751-3766

[14] Hou Z L, Liu Y X, Zhang L. Pos-gift: A geometric and intensity-invariant feature transformation for multimodal images. Information Fusion, 2024, 102: 102027

[15] Luan S Z, Chen C, Zhang B C, et al. Gabor convolutional networks. IEEE Transactions on Image Processing, 2018, 27(9): 4357-4366

[16] P'erez J C, Alfarra M, Jeanneret G, et al. Gabor layers enhance network robustness// Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 450-466

[17] Rao S, Varma V. Alternate approaches to scattering networks in image classification// Proceedings of the 2023 International Conference on Network, Multimedia and Information Technology (NMITCON). Bengaluru, India, 2023, 1-5

[18] Yeung C C, Lam K M. Attentive boundary-aware fusion for defect semantic segmentation using transformer. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 3271723

[19] Wu Y H, Liu Y, Zhan X, et al. P2T: Pyramid pooling transformer for scene understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(11): 12760-12771

[20] Cai M L, Wang J X, Liu J P, et al. Transformer-GAN architecture for anomaly detection in multivariate time series. SCIENTIA SINICA Informationis, 2023, 53(5): 972-992 (in Chinese)

(蔡美玲，汪家喜，刘金平 等. 基于 Transformer GAN 架构的多变量时间序列异常检测. 中国科学：信息科学，2023，53(5)：972-992)

[21] Huang J J，Li P W，Peng M，et al. Review of deep learning-based Topic Model. Chinese Journal of Computers，2020，43(5)：827-855 (in Chinese)
(黄佳佳，李鹏伟，彭敏 等. 基于深度学习的主题模型研究. 计算机学报. 2020，43(5)：827-855)

[22] Li Z W，Liu F，Yang W J，et al. A survey of convolutional neural networks：Analysis，applications，and prospects. IEEE Transactions on Neural Networks and Learning Systems，2021，33(12)：6999-7019

[23] Yang Y T，Jiao L C，Liu X，et al. Transformers meet visual learning understanding：A comprehensive review. arXiv preprint arXiv：2203.12944，2022

[24] Vaswani A，Shazeer N，Parmar P，et al. Attention is all you need//Proceedings of the Advances in Neural Information Processing Systems. Long Beach，USA ，2017，5998-6008

[25] Tan H C，Liu X P，Yin B C，et al. Mhsa-net：Multihead self-attention network for occluded person re-identification. IEEE Transactions on Neural Networks and Learning Systems，2022，34(11)：8210-8224

[26] Dosovitskiy A，Beyer L，Kolesnikov A，et al. An image is worth 16x16 words：Transformers for image recognition at scale// Proceedings of the International Conference on Learning Representations. Vienna，Austria，2021，1-21

[27] Liu Z，Lin Y，Cao Y，et al. Swin transformer：Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal，Canada，2021，10012-10022

[28] Wang W，Xie E，Li X，et al. Pyramid vision transformer：A versatile backbone for dense prediction without convolutions// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal，Canada，2021，568-578

[29] Han K，Xiao A，Wu E，et al. Transformer in transformer // Advances in Neural Information Processing Systems，Virtual-only Conference.2021，34：15908-15919

[30] Yang Y，Jiao L，Li L，et al. LGLFormer：Local-global lifting transformer for remote sensing scene parsing. IEEE Transactions on Geoscience and Remote Sensing，2024，62：3344116

[31] Yao T，Li Y，Pan Y，et al. HIRI-ViT：Scaling vision transformer with high resolution inputs. IEEE Transactions on Pattern Analysis and Machine Intelligence，2024(01)：1-12

[32] Zaheer M，Guruganesh G，Dubey K A，et al. Big bird：Transformers for longer sequences//Proceedings of the Advances in Neural Information Processing Systems. Vancouver，Canada，2020，33：17283-17297

[33] Poulenard A and Ovsjanikov M. Multi-directional geodesic neural networks via equivariant convolution. ACM Transactions on Graphics (TOG)，2018，37(6)：1-14

[34] Hu X W，Fu C W，Zhu L，et al. Direction-aware spatial context features for shadow detection and removal. IEEE Transactions on Pattern Analysis and Machine Intelligence，2019，42(11)：2795-2808

[35] Tao S C，Zhang X L，Hua Y X，et al. A novel artificial visual system for motion direction detection with completely modeled retinal direction-selective pathway. Mathematics，2023，11(17)：3732

[36] Yue S G and Fu Q B. Modeling direction selective visual neural network with on and off pathways for extracting motion cues from cluttered background//Proceedings of the International Joint Conference on Neural Networks. Anchorage，USA，2017，831-838

[37] Shen Y，Zhu S J，Yang T J，et al. Bdanet：Multiscale convolutional neural network with cross-directional attention for building damage assessment from satellite images. IEEE Transactions on Geoscience and Remote Sensing，2021，60：1-14

[38] Milocco L and Isaac S C I. A method to predict the response to directional selection using a Kalman filter. Proceedings of the National Academy of Sciences，2022，119(28)：e2117916119

[39] Garcia A，Musallam M A，Gaudilliere V，et al. D. Lspnet：A 2d localization-oriented spacecraft pose estimation neural network// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，Nashville，TN，USA，2021，2048-2056

[40] Zhou Z T，Zhou L，Hu D W. Scene recognition combining structural and textural features. Science China Information Sciences，2012，42(6)：687-702

[41] Alekseev A，Bobe A. Gabornet：Gabor filters with learnable parameters in deep convolutional neural network//Proceedings of the International Conference on Engineering and Telecommunication (EnT). Xi'an，China，2019，1-4

[42] Chen C，Zhou K N，Qi S Y，et al. A learnable gabor convolution kernel for vessel segmentation. Computers in Biology and Medicine，2023，158：106892

[43] Yuan Y，Wang L N，Zhong G Q，et al. Adaptive gabor convolutional networks. Pattern Recognition，2022，124：108495

[44] Reyes A A，Paheding S，Deo M，et al. Gabor filter-embedded u-net with transformer-based encoding for biomedical image segmentation//Proceedings of the International Workshop on Multiscale Multimodal Medical Imaging. Singapore，2022，76-88

[45] Fan J Q，Su T K，Zhang K H，et al. Temporally efficient gabor transformer for unsupervised video object segmentation// Proceedings of the 31st ACM International Conference on Multimedia. Ottawa，Canada，2023，3394-3402

[46] He K M，Zhang X Y，Ren S Q，et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，USA，2016，770-778

[47] Deng X D，Feng S B，Guo P，et al. Fast image recognition with gabor filter and pseudoinverse learning autoencoders//Neural Information Processing：25th International Conference. Siem Reap，Cambodia，2018，501-511

[48] Liu C L，Ding W R，Wang X D，et al. Hybrid gabor convolutional networks. Pattern Recognition Letters，2018，116：164-169

[49] Yuan Y，Zhang J A，Wang Q. Deep gabor convolution network for person re-identification. Neurocomputing，2020，378：

387-398

[50] Gong X, Xia X, Zhu W T, et al. Deformable gabor feature networks for biomedical image classification//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2021, 4004-4012

[51] Lu Z Z, Xiao Liang X, Yang G A, et al. Small-scale convolutional neural networks with learnable gabor filter for image classifications//Proceedings of the 2021 4th International Conference on Information Communication and Signal Processing (ICICSP). Shanghai, China, 2021, 425-431

[52] Rivas P, Rai M. Gabor filters as initializers for convolutional neural networks: A study on inductive bias and performance on image classification//LatinX in AI Workshop at ICML, Hawaii, USA,2023,1-6

[53] Tolstikhin I O, Houlsby N, Kolesnikov A, et al. Mlp-mixer: An all-mlp architecture for vision// Proceedings of the Advances in Neural Information Processing Systems. Virtual-only Conference, 2021, 34: 24261-24272

[54] Wang Z, Pang T, Du C, et al. Better diffusion models further improve adversarial training// Proceedings of the International Conference on Machine Learning. Hawaii, USA, 2023, 36246-36263

[55] Ma B, Zhang J, Xia Y, et al. VNAS: Variational neural architecture search. International Journal of Computer Vision, 2024, 132: 3689-3713

[56] Lin G, Milan A, Shen C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA,2017,1925-1934

[57] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017, 2881-2890

[58] Zhang R, Tang S, Zhang Y, et al. Scale-adaptive convolutions for scene parsing//Proceedings of the IEEE International Conference on Computer Vision. Hawaii, USA, 2017, 2031-2039

[59] Zhang H, Dana K, Shi J, et al. Context encoding for semantic segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake, USA, 2018, 7151-7160

[60] Liang X, Zhou H, Xing E. Dynamic-structured semantic propagation network//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake, USA, 2018, 752-761

[61] Xiao T, Liu Y, Zhou B, et al. Unified perceptual parsing for scene understanding// Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018, 418-434

[62] Zhao H, Zhang Y, Liu S, et al. Psanet: Point-wise spatial attention network for scene parsing//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018 267-283

[63] Jia D, Cao J, Pan J, et al. Multi-stream densely connected network for semantic segmentation. IET Computer Vision, 2022, 16(2): 180-191

[64] Liu Q, Dong Y, Li X. Multi-stage context refinement network for semantic segmentation. Neurocomputing, 2023, 535: 53-63.

[65] Shinoda R, Hayamizu R, Nakashima K, et al. SegRCDB: Semantic segmentation via formula-driven supervised learning//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023, 20054-20063

[66] Liu Q, Dong Y, Jiang Z, et al. Multi-pooling context network for image semantic segmentation. Remote Sensing, 2023, 15 (11): 1-15

YANG Yu-Ting, Ph. D., research assistant postdoctoral researcher. Her research interests include computer vision, deep learning and multiscale geometric analysis.

LI Ling-Ling, Ph. D., associate professor. Her current research interests include quantum evolutionary optimization, machine learning and deep learning.

LIU Xu, Ph. D., associate professor. His research interests include machine learning and image processing.

JIAO Li-Cheng, Ph. D., professor. His research interests include image processing, natural computation, machine learning, and intelligent information processing.

LIU Fang, M. S., professor, Ph. D. supervisor. Her research interests include signal and image processing, synthetic aperture radar image processing, multi-scale geometry analysis, learning theory and algorithms, optimization problems, and data mining.

MA Wen-Ping, Ph. D., professor. Her research interests include natural computing and intelligent image processing.

Background

The research problem of this paper belongs to the visual representation in computer vision. Convolutional neural networks and Transformer frameworks have made significant advancements in computer vision. However, powerful feature representation still holds important research significance in deep image recognition. In addition to multiscale features, the anisotropy of image features plays a crucial role in image recognition. Multidirectional feature learning has excellent potential in capturing image edges, textures, shapes, and structures and has become an important research focus.

This paper proposes a deep network framework called multiscale and multidirectional Transformer (MSMDFormer), which achieves more powerful feature representation. Specifically, we introduce a multidirectional encoder to enhance the multidirectional features. Furthermore, we design a multiscale and multidirectional Transformer encoder that effectively aggregates the multiscale and multidirectional features of the images. This study is part of the national natural science foundation key project-brain-like cognitive machine learning and remote sensing interpretation application project, and it primarily investigates the influence of directionality in vision on recognition tasks.