基于多尺度胶囊Swin Transformer的SAR图像目标识别方法

侯宇超[1,2]，王洁[1]，李洪涛[1]，郝岩[3]，段晓旗[2]，黄凯文[2]，田有亮[2]
（1. 山西师范大学密码学与数据安全山西省重点实验室，山西 太原 030031；2. 贵州大学公共大数据国家重点实验室，贵州 贵阳 550025；
3. 太原师范学院数学与统计学院，山西 太原 030002）

摘　要：通过协同胶囊单元的语义特征编码和Swin Transformer的上下文特征图建模优势相结合，提出了一种多尺度胶囊Swin Transformer网络（MSCSTN），将胶囊编码和Swin Transformer联合应用于SAR图像目标识别。该网络集成3个并行的胶囊Swin Transformer编码结构，融合后对输入图像进行分类。每个结构通过基于膨胀卷积切片划分的胶囊令牌编码器和三维胶囊Swin Transformer模块构建，能捕获更深层次、更广泛的语义特征。在运动和静止目标的获取与识别（MSTAR）数据集及FUSAR-Ship数据集上的实验结果表明，MSCSTN在各种测试条件下均优于其他方法。结果表明，MSCSTN展现了良好的识别性能、泛化能力和应用潜力。

关键词：膨胀卷积切片分区；胶囊令牌编码器；三维胶囊Swin Transformer模块；多尺度胶囊Swin Transformer网络；SAR图像目标识别
中图分类号：TN957.52
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025045

Multi-scale capsule Swin Transformer-based method for SAR image target recognition

HOU Yuchao[1,2], WANG Jie[1], LI Hongtao[1], HAO Yan[3], DUAN Xiaoqi[2], HUANG Kaiwen[2], TIAN Youliang[2]
1. Shanxi Key Laboratory of Cryptography and Data Security, Shanxi Normal University, Taiyuan 030031, China
2. State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China
3. School of Mathematics and Statistics, Taiyuan Normal University, Taiyuan 030002, China

Abstract: A multi-scale capsule Swin Transformer network (MSCSTN) was proposed by synergizing the semantic feature encoding of capsule units with the context feature mapping of Swin Transformer. Capsule encoding and the Swin Transformer were jointly applied to SAR image target recognition. The network was integrated with three parallel capsule Swin Transformer encoding structures, which were fused to classify the input image. Each structure was constructed through a capsule token encoder based on expanded convolutional slice partition and a 3D capsule Swin Transformer module, which designed to capture of more profound and extensive semantic features. The experimental results on the moving and stationary target acquisition and recognition (MSTAR) dataset and FUSAR-Ship dataset were shown to demonstrate that MSCSTN outperformed other methods under various test conditions. The results demonstrate that MSCSTN exhibits excellent recognition performance, generalization ability, and potential for application.

Keywords: dilated convolution patch partition, capsule token encoder, three-dimensional capsule Swin Transformer module, multi scale capsule Swin Transformer network, SAR image target recognition

收稿日期：2024–11–07；修回日期：2025–02–18
通信作者：王洁，wjlkt@163.com
基金项目：国家自然科学基金资助项目（No.42461057, No.62272123, No.42371470）；山西省基础研究计划基金资助项目（No.202303021212164, No.202303021212255）；山西省高等学校科技创新基金资助项目（No.2022L405）；山西省研究生科研创新基金资助项目（No.2024KY474）；贵州省基础研究基金资助项目（No.2024129）
Foundation Items: The National Natural Science Foundation of China (No.42461057, No.62272123, No.42371470), The Fundamental Research Program of Shanxi Province (No.202303021212164, No.202303021212255), Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (No.2022L405), Postgraduate Education Innovation Program of Shanxi Province (No.2024KY474), Guizhou Provincial Basic Research Program (No.2024129)

0　引言

合成孔径雷达（SAR, synthetic aperture radar）具有全天候、远距离、高分辨率成像的能力[1-2]。由于其在遥感成像上的优势，已被广泛应用于地形分类[3]、变换检测[4]和灾害评估[5]等领域。然而，在SAR图像目标中，较小的纹理特征关注于目标的局部细节和背景的相干斑噪声，而较大的纹理特征通常存在于目标的轮廓中[6]。这些特点导致SAR图像人工解译效率较低。因此，发展SAR自动目标识别（SAR ATR, SAR automatic target recognition）技术具有重要的理论意义和巨大的实用价值。

SAR图像目标识别主要有模板匹配和模式识别2种方法。模板匹配是将测试样本与每类样本模板进行比较，通过相似度来识别图像目标[7]。模式识别是先借助物理或概念模型描述目标特征，然后利用分类器对图像目标进行识别[8]。值得注意的是，这些方法识别性能主要依赖于特征表示能力。当提取的特征不能凸显不同类型目标之间的差异时，识别性能会显著下降。

近年来，受深度学习技术成功的启发，研究人员将其引入SAR图像目标识别任务中，以提高特征表示能力[9]。Chen等[10]利用卷积层代替全连接层，首先提出一种全卷积神经网络（A-ConvNet, all-convolutional neural network）。Shang等[11]提出一种记忆卷积神经网络（M-Net, memory convolution neural network）。该方法在CNN的基础上增加一个信息记录器来记忆和存储样本空间特征，然后利用特征的空间相似性信息来预测测试样本标签。Liu等[12]充分利用SAR ATR的物理性质和深度判别特征，提出一种散射和深度特征融合网络（SDF-Net, scattering and deep feature fusion network）。这些基于CNN的方法极大地推进了SAR图像目标识别的研究。然而，这些方法仍然存在一些问题，很难解释图像中部分与整体之间的空间关系，标量卷积操作会丢失许多重要的特征信息使其对于输入微小变化不敏感。这些方法从卷积获得的局部特征中聚合全局信息，而不是直接对全局上下文进行编码，因此，从背景相干斑噪声复杂的SAR图像目标中难以获得清晰完整的全局信息。

相比于CNN，胶囊网络（CapsNet, capsule network）可以将神经层嵌入另一个图层中，集合图像的位姿信息和空间属性，更好地建模神经网络中内部知识表示的分层关系。动态路由机制可以自适应更新注意力权重，提高识别性能。Hou等[13]为了增强小样本数据下SAR图像特征提取能力，提出一种带类分离损失函数的多维并行胶囊网络。Yu等[14]引入联合分工协同训练框架作为特征过滤器，将CapsNet与广泛学习系统结合起来，提高模型识别性能。Swin Transformer可以借助更紧凑的特征表示和更丰富的语义信息对上下文特征图进行建模。He等[15]提出一种基于改进盲区的Swin Transformer网络，以减轻SAR图像中相干斑噪声的影响，提高识别性能。Li等[16]提出一种增强Swin Transformer检测网络的SAR图像目标识别方法。然而，目前提出的CapsNet的浅层多采用传统标量卷积层进行特征提取，对高层语义信息提取效果不佳。Swin Transformer多关注于图像的全局信息，对局部信息的理解有限。本文旨在发挥CapsNet和Swin Transformer的优势，首次尝试联合利用胶囊编码和Swin Transformer解决SAR图像目标识别问题。胶囊结构保留了图像中精细的局部特征，Swin Transformer增强了全局特征捕获能力。本文主要工作如下。

1) 提出一种多尺度胶囊Swin Transformer网络（MSCSTN, multi scale capsule swin transformer network），通过考虑来自不同尺度的代表性判别信息，更好地捕捉SAR图像目标特征。

2) 提出基于膨胀卷积切片分区的胶囊令牌编码器（DCPP-CTE, dilated-convolution-patch-partition-based capsule token encoder）。相比于传统令牌编码器，DCPP-CTE可以在保持像素相对空间位置不变、不丢失分辨率和不引入额外参数的前提下，通过设置膨胀率自由提取不同感受野区域信息，捕获更多微小的目标信息，最大限度地保留有用的内容，对特征进行更深层次的理解。

3) 提出一种三维胶囊Swin Transformer模块（3DCSTM, three-dimensional capsule swin transformer module）。相比于传统Swin Transformer模块，3DCSTM可以聚合来自多头注意力的信息，捕获更多语义特征。

4) 利用运动和静止目标的捕获与识别（MSTAR, moving and stationary target acquisition and recognition）数据集和FUSAR-Ship数据集进行

了大量的算法实验，实验结果表明，在多种测试条件下，所提MSCSTN优于绝大多数传统机器学习方法和基于深度学习的方法。

1　基本模型

1.1　胶囊网络

CNN从上一层至下一层传递的是标量。标量没有方向，对低层和高层特征的空间关系建模能力较弱，因此，CNN在识别具有空间关系特征时存在很大的局限性。与CNN不同的是，作为一种基于胶囊结构的新型神经网络，CapsNet中的胶囊代表了一组神经元，即向量[17]。胶囊的长度代表实体存在的概率，胶囊的输出方向代表除长度之外的角度、色调、位置等参数。低层胶囊$\boldsymbol{u}_i$使用动态路由算法将预测信息传递到高层胶囊$\boldsymbol{v}_j$中。低层胶囊和高层胶囊之间的连接如图1所示。

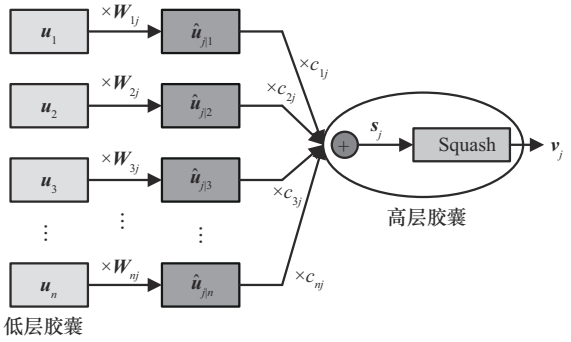
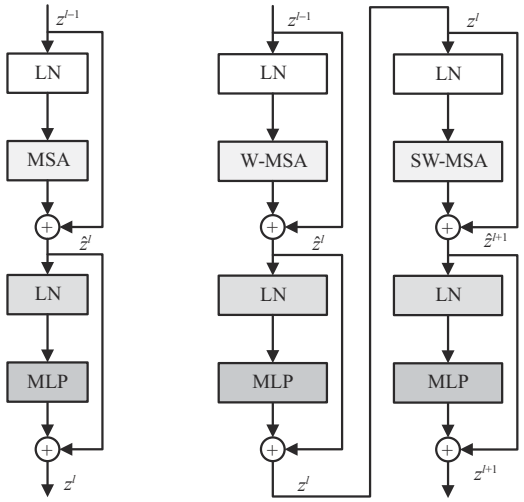


图1　低层胶囊和高层胶囊之间的连接

1.2　Swin Transformer

作为一种基于注意力的结构，Transformer超越了之前基于复杂递归和CNN的序列转换模型，被广泛应用于自然语言处理（NLP，natural language processing）的各个领域[18]。标准Transformer由多头自注意（MSA，multi-head self-attention）、多层感知机（MLP，multiple layer perceptron）和层归一化（LN，layer normalization）组成[19]，如图2(a)所示。MSA在建立输入和输出序列之间的全局依赖关系方面发挥了关键作用。Transformer凭借其在NLP领域取得的重大突破及卓越的性能，引起了计算机视觉（CV，computer vision）领域研究人员的广泛关注[20]。Dosovitskiy等[21]首次将纯Transformer网络结构应用于图像分类任务中，并取得了显著的分类效果。随后，基于Transformer的模型被广泛应用在目标检测[22]、图像分割[23]和图像生成[24]等各个领域，可以媲美甚至超越同时期基于CNN的方案。

然而，由于这些模型特征映射的分辨率较低，且计算复杂度随图像大小增加呈2次增长的特点，这些模型的架构并不适合在密集视觉任务或输入图像分辨率较高的情况下使用。因此，Liu等[25]提出了基于移位窗口策略的Swin Transformer。该模型抛弃了传统Transformer中的MSA，提出基于窗口的MSA（W-MSA，window based self-attention）和基于移位窗口的MSA（SW-MSA，shifted window based self-attention），2个连续的Swin Transformer如图2(b)所示。该模型将MSA的计算限制在非重叠窗口的同时允许跨窗口信息交互，对于不同分辨率的输入图像具有鲁棒性且计算复杂度随图像大小的增加呈线性增长。



(a) 标准Transformer　　(b) 2个连续的Swin Transformer

图2　Transformer结构

1.3　膨胀卷积

膨胀卷积的主要思想是在保持参数量不变的情况下在卷积核像素之间插入空洞以增加其感受野[26]。与传统标准卷积相比，膨胀卷积可以捕获更丰富的语义信息，提高识别准确率。当膨胀率为1时，膨胀卷积可视为标准卷积。而当膨胀率大于1时，卷积核将对特征图等膨胀率−1间隔采样。其计算式为

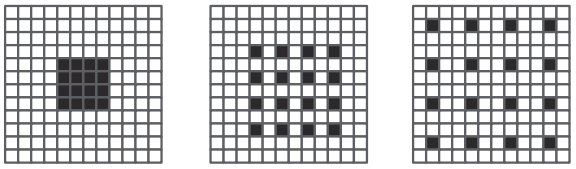
$$O_j^k(x,y)=\sigma\left(\sum_{e=0}^{w_d-1}\sum_{f=0}^{w_d-1}K_{ji}^k(e,f)\times O_i^{k-1}(x-r\cdot e,y-r\cdot f)+v_j^k\right) \tag{1}$$

其中，(x, y)为像素单元位置，$\sigma(\cdot)$为激活函数，K_{ji}^{k}为卷积核，其核大小为$w_d \times w_d$，r为膨胀率，v_j^k为偏置。

膨胀卷积的感受野P计算式为

$$P = w_d + \left(w_d - 1\right)\left(r - 1\right) \tag{2}$$

当卷积核大小为4×4，膨胀率分别为1、2、3时，膨胀卷积的感受野分别为4×4、7×7、10×10。在这3种情况下，膨胀卷积如图3所示。



(a) 膨胀率为1　(b) 膨胀率为2　(c) 膨胀率为3

图3　膨胀卷积

2　MSCSTN 模型

2.1　整体结构

MSCSTN 模型结构如图4所示。该网络并行了3条三维胶囊 Swin Transformer 编码结构，最后将它们整合在一起判定输入图像的类别。每个结构由一个 DCPP-CTE 和2个 3DCSTM 组成。具体地说，首先，将每幅大小为96×96的图像输入3个膨胀率，分别为1、2和3；将卷积核大小为4×4、步长为4的膨胀卷积切片分区层进行切片分区，并在通道方向展平，得到$\frac{96}{4} \times \frac{96}{4} = 24 \times 24$个特征维度为$4 \times 4 = 16$的切片。接着，将原始切片特征分别通过卷积核大小为3×3、步长为2的胶囊编码层后得到12×12个特征维度为16、向量长度为4的矢量胶囊编码特征图。然后，将矢量特征分别通过卷积核大小为3×3、步长为1的令牌编码层后得到12×12个特征维度为16、向量长度为8的矢量令牌编码特征图。再将上一阶段得到的矢量特征图分别输入2个连续的3DCSTM中，在3DCSTM中编码提取特征语义。最后，将3个结构输出的特征级联以确定其类别。

2.2　基于膨胀卷积切片分区的胶囊令牌编码器

传统令牌编码器采用的局部卷积切片分区忽略

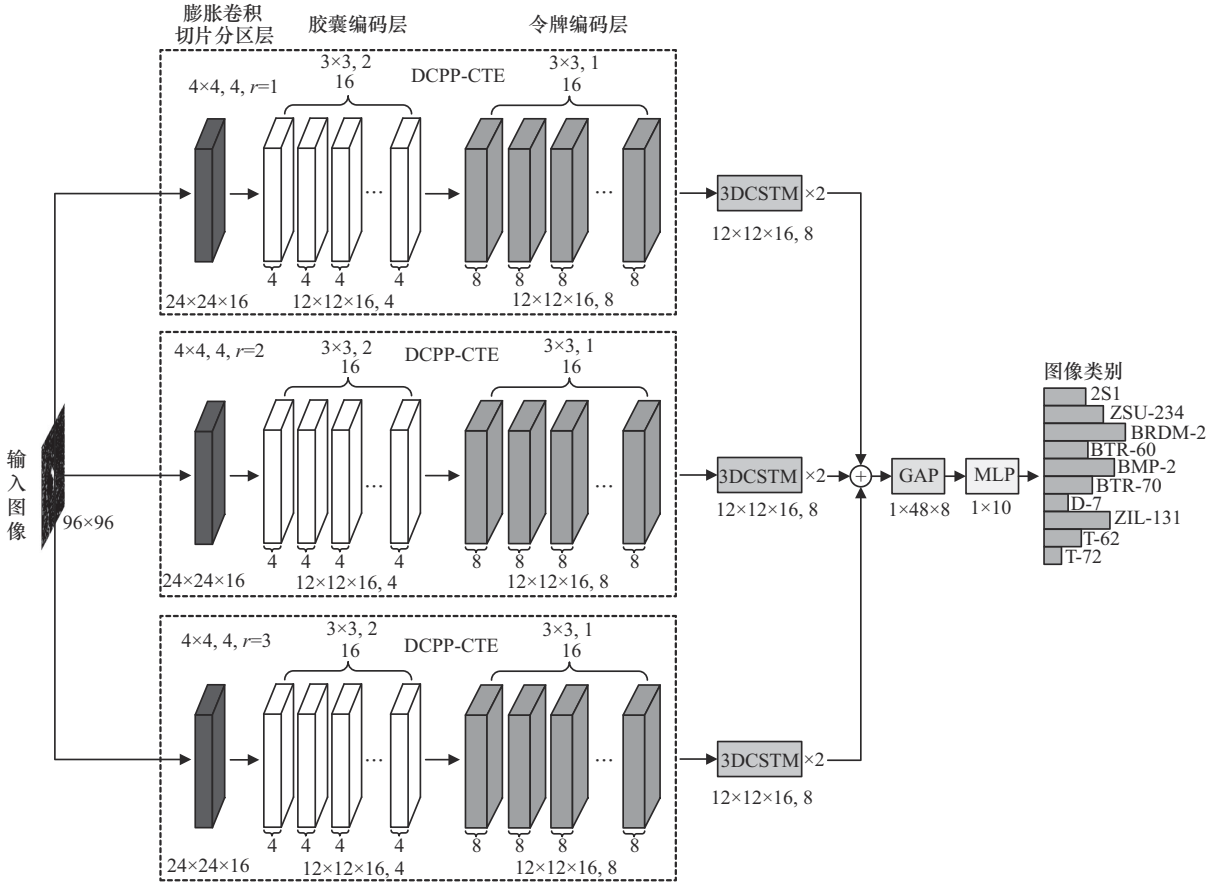


图4　MSCSTN 模型结构

了相邻图像区域的空间相似性，使其难以对全局空间特征建立长期的依赖关系，而标量卷积编码操作也会丢失许多重要的特征信息使其难以感受输入的细微变化。

为了解决这一问题，本文提出DCPP-CTE，结构如图5所示。DCPP-CTE由膨胀卷积切片分区层、胶囊编码层和令牌编码层3个部分组成。首先，将每幅大小为$h\times w$的图像输入膨胀率为r、卷积核大小为$w_d\times w_d$、步长为w_d的膨胀卷积切片分区层进行切片分区，并在通道方向展平，得到$\frac{h}{w_d}\times\frac{w}{w_d}$个特征维度为$w_d^2$的切片。接着，将原始切片特征通过卷积核大小为$w_c\times w_c$、步长为$w_c'$的胶囊编码层后得到$w_d^2$张大小为$\frac{h}{w_d\times w_c'}\times\frac{w}{w_d\times w_c'}$、向量长度为$l$的胶囊编码特征图。最后，将胶囊编码特征通过卷积大小为$w_t\times w_t$、步长为w_t'的令牌编码层后得到w_d^2张大小为$\frac{h}{w_d\times w_c'\times w_t'}\times\frac{w}{w_d\times w_c'\times w_t'}$、向量长度为$2l$的令牌编码特征图。在后续的实现中，DCPP-CTE并行了3个结构，每个结构中DCPP-CTE的膨胀率r是不同的。不同结构的切片分区具有不同大小的感受野区域，既能提取细微的局部信息，又能通过膨胀卷积收集丰富的空间特征，进而获得多尺度特征表示。

2.3　三维胶囊Swin Transformer模块

本文提出一种3DCSTM，该模块由一个利用远程全局特征交互的三维（移位）窗口多头自注意（3D(S)W-MSA，three-dimensional (shifted) window multi-head self-attention）、一个胶囊聚合（CA，capsule aggregation）和一个MLP组成。在3D(S)W-MSA和MLP前增加LN层进行线性归一化，并在CA和MLP后增加残差连接进行模型增强。3DW-MSA和3DSW-MSA分别被应用于2个连续的3DCSTM中，结构如图6所示。

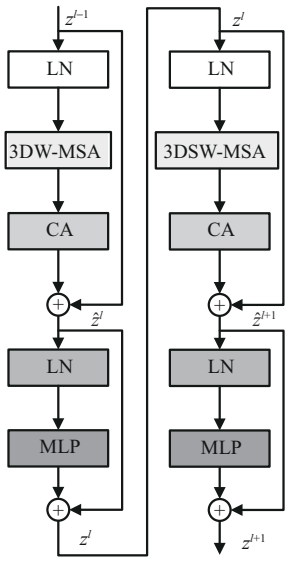


图6　2个连续的3DCSTM结构

连续的3DCSTM可表示为

$$\begin{aligned}\hat{z}^l&=\mathrm{CA}\left(\text{3DW-MSA}\left(\mathrm{LN}\left(z^{l-1}\right)\right)\right)+z^{l-1}\\ z^l&=\mathrm{MLP}\left(\mathrm{LN}\left(\hat{z}^l\right)\right)+\hat{z}^l\\ \hat{z}^{l+1}&=\mathrm{CA}\left(\text{3DSW-MSA}\left(\mathrm{LN}\left(z^l\right)\right)\right)+z^l\\ z^{l+1}&=\mathrm{MLP}\left(\mathrm{LN}\left(\hat{z}^{l+1}\right)\right)+\hat{z}^{l+1}\end{aligned}\tag{3}$$

其中，$\hat{z}^l$和z^l分别为第l层中3DW-MSA和MLP的输出，$\hat{z}^{l+1}$和z^{l+1}分别为第$l+1$层中3DSW-MSA和MLP的输出。

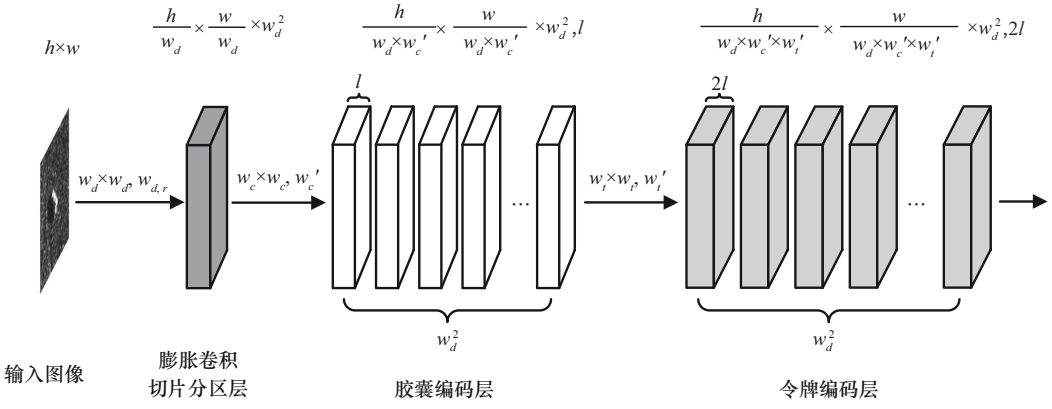


图5　DCPP-CTE结构

2.3.1　三维(移位)窗口多头自注意机制

作为一种高效的自注意力机制，本文在W-MSA的基础上提出了在局部三维窗口内计算自注意力的3DW-MSA。假设输入特征图大小为$h \times w \times d$，每个窗口切片大小为$M \times M \times M$。MSA在三维输入特征图下（3D-MSA）的计算复杂度为

$$\Omega(\text{3D-MSA})=4hwdC^2+2(hwd)^2C \tag{4}$$

3DW-MSA的计算复杂度为

$$\Omega(\text{3DW-MSA})=4hwdC^2+2hwdM^3C \tag{5}$$

其中，当窗口切片大小固定时，Ω(3D-MSA)的hwd是二次的，Ω(3DW-MSA)的hwd是线性的。相比于标准3D-MSA，3DW-MSA的窗口以不重叠的方式均匀地分割图像，一定程度上降低了计算量。

3DW-MSA缺乏跨窗口的连接，限制了它的建模能力。本文在不增加计算的情况下引入跨窗口交互，提出了3DSW-MSA。相比于3DW-MSA，3DSW-MSA将规则分区的窗口切片大小替换为$\frac{M}{2} \times \frac{M}{2} \times \frac{M}{2}$，并沿着3个方向坐标轴按$\left(\frac{M}{2},\frac{M}{2},\frac{M}{2}\right)$移动。同时，在计算注意力时将较小的窗口填充为$M \times M \times M$大小，屏蔽了填充值。3DSW-MSA中的自注意力计算跨越了3DW-MSA中先前窗口的边界，提供了它们之间的连接。3D移位窗口示例如图7所示。假设输入特征图大小为8×8×8，每个窗口切片大小为4×4×4。第l层采用了规则分区窗口，窗口数量为2×2×2=8。第$l+1$层将上一层规则分区的窗口切片大小替换为2×2×2，并沿着3个方向坐标轴按（2, 2, 2）移动，窗口数量变为3×3×3=27。虽然增加了窗口数量，但是通过循环移位，批处理窗口的数量仍然与规则分区的窗口数量保持一致，具有高效批计算能力。

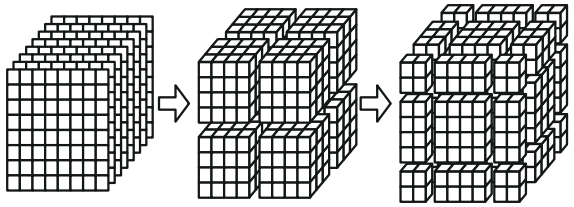


图7　3D移位窗口示例

值得注意的是，在计算自注意力时，本文为每个头部引入3D相对位置偏差$B_i \in \mathbb{R}^{M^2 \times M^2 \times M^2}$，第$i$个注意力头$u_i$的计算方法为

$$u_i = \text{Soft max}\left(\frac{\boldsymbol{Q}_i \boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{g}} + \boldsymbol{B}_i\right)\boldsymbol{V}_i, i \in [1,h] \tag{6}$$

其中，$\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i \in \mathbb{R}^{M^3 \times d}$为第$i$个头的查询矩阵、键矩阵和值矩阵；$d$为查询特征和键特征维度，$M^3$为3D窗口中的切片令牌的数量，$h$是注意力头的个数。由于每个轴的相对位置在$[-M+1, M-1]$的范围内，本文参数化了一个小型偏置矩阵$\hat{\boldsymbol{B}} \in \mathbb{R}^{(2M-1)\times(2M-1)\times(2M-1)}$，$\boldsymbol{B}_i$的值来自$\hat{\boldsymbol{B}}$。

2.3.2　胶囊聚合模块

本文使用的多头自注意力机制将注意力从一个唯一空间扩展到不同的表征子空间。这种映射过程旨在不同的子空间中独立探索可能的表示，因此可以避免由一个单一空间建模引起的风险。不同的注意力头可能在不同的子空间中携带不同的空间属性特征。然而，不同的子空间可能拥有相似的信息，进而导致语义冗余。将不同的注意力头直接连接在一起忽略了其冗余性，并可能将它们视为不同的语义而导致错误的后续操作，从而导致性能下降。本文利用胶囊结构设计一个胶囊聚合模块来排列和融合来自不同头部的语义和空间属性信息，提高模型特征提取能力。

胶囊编码旨在通过迭代动态路由过程将输入胶囊的信息聚类，并将每个聚类的代表性信息存储在输出胶囊中。本文在多头注意力机制之后插入一个胶囊聚合模块，对所有来自头部的信息进行梳理。在多头自注意力机制中，不同的注意力头可以看作是对同一物体不同视角的观察。输入胶囊层表示同一输入的不同空间属性特征。迭代路由过程可以更好地决定哪些信息及多少信息流向输出胶囊。理想情况下，每个输出胶囊代表输入的一个独特属性，并在组合时携带所有需要的信息。

胶囊聚合模块结构如图8所示。首先，利用预定义的空间转换矩阵$\boldsymbol{W}_{ij}$将第i个注意力头的输出$\boldsymbol{u}_i$转换为对特征胶囊j的预测$\hat{\boldsymbol{u}}_{j|i}$。转换公式为

$$\hat{\boldsymbol{u}}_{j|i} = \boldsymbol{W}_{ij}\boldsymbol{u}_i \tag{7}$$

然后，对胶囊$\hat{\boldsymbol{u}}_{j|i}$加权求和得到向量$\boldsymbol{s}_j$为

$$\boldsymbol{s}_j = \sum_i c_{ij}\hat{\boldsymbol{u}}_{j|i} \tag{8}$$

其中，c_{ij}为2个胶囊间的耦合系数。

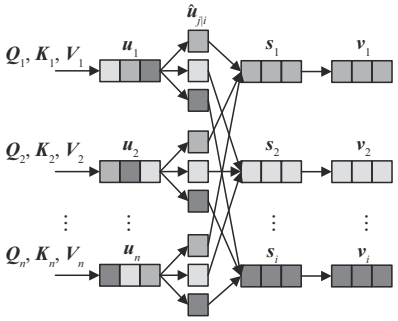


图8　胶囊聚合模块结构

最后，利用压缩（Squash）函数对胶囊输出进行非线性激活，使激活后的输出向量$\boldsymbol{v}_j$的长度保持在0和1范围内，且保证$\boldsymbol{v}_j$和输入向量$\boldsymbol{s}_j$方向一致。其公式为

$$\boldsymbol{v}_j = \frac{\left\|\boldsymbol{s}_j\right\|}{1+\left\|\boldsymbol{s}_j\right\|^2}\cdot\frac{\boldsymbol{s}_j}{\left\|\boldsymbol{s}_j\right\|} \tag{9}$$

其中，当$\boldsymbol{s}_j$较长时，第一项近似为1；类似地，当$\boldsymbol{s}_j$很短时，第一项近似为0；第二项是对$\boldsymbol{s}_j$进行单位化操作。

注意力头i与胶囊j之间的权重由动态路由机制进行更新，计算步骤如下。

首先，初始化连接概率b_{ij}，并计算耦合系数c_{ij}为

$$c_{ij} = \frac{\exp\left(b_{ij}\right)}{\sum_k \exp\left(b_{ik}\right)} \tag{10}$$

其次，利用Squash函数激活胶囊输出，得到输出胶囊$\boldsymbol{v}_j$。

然后，将所有的注意力头分别转换得到的胶囊$\hat{\boldsymbol{u}}_{j|i}$与胶囊$\boldsymbol{v}_j$点乘，并更新胶囊间连接概率$b_{ij}$为

$$b_{ij} = b_{ij} + \hat{\boldsymbol{u}}_{j|i}\boldsymbol{v}_j \tag{11}$$

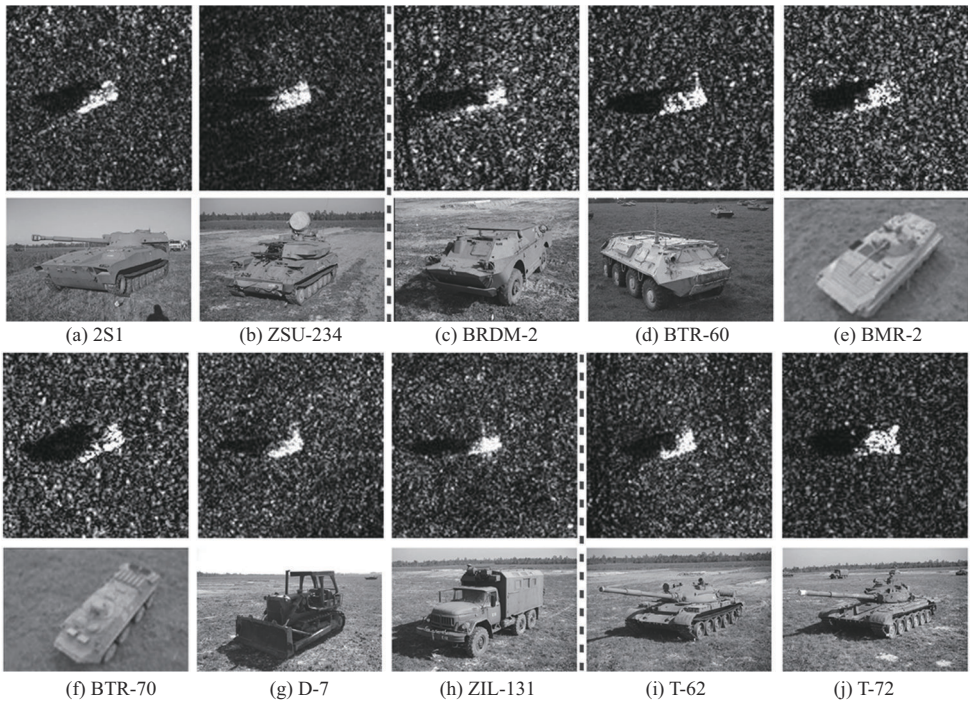
最后，达到路由迭代次数时，循环停止，否则返回第一步重新计算系数c_{ij}。

3　实验结果与分析

3.1　实验数据集

本文使用的SAR图像数据集为MSTAR数据集和FUSAR-Ship数据集，这2个数据集的详细信息如下。

MSTAR数据集是由美国国防部先进研究项目局和空军实验室共同资助，使用桑迪亚国家实验室X波段SAR传感器平台采集的，分辨率为0.3 m×0.3 m，全方位角覆盖0°～360°。该数据集包括10类军用车辆目标（火炮：2S1和ZSU-234；卡车：BRDM-2、BTR-60、BMP-2、BTR-70、D-7和ZIL-131；坦克：T-62和T-72），MSTAR数据集中的10种图像如图9所示。这些SAR图像来源于不同的场景。基于SAR图像采集条件的差异，MSTAR数据集可分为标准操作



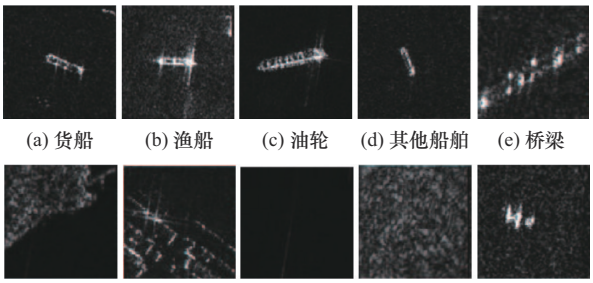
(a) 2S1　(b) ZSU-234　(c) BRDM-2　(d) BTR-60　(e) BMR-2

(f) BTR-70　(g) D-7　(h) ZIL-131　(i) T-62　(j) T-72

图9　MSTAR数据集中的10种图像

条件（SOC，standard operating condition）和扩展操作条件（EOC，extended operating condition）。

FUSAR-Ship数据集来自高分3号（GF-3）卫星，专为SAR海洋监测设计[27]。该数据集由126张原始图像构成，涵盖了海洋、陆地、海岸、河流和岛屿等多种场景。本文选择的船舶包括货船、渔船、油轮和其他船舶；陆地包括桥梁、沿海土地和陆地块；海洋包括海洋块和海杂波；强虚警目标等10种具有代表性的场景进行评估，FUSAR-Ship数据集中的10种图像如图10所示。



(a) 货船　(b) 渔船　(c) 油轮　(d) 其他船舶　(e) 桥梁

(f) 沿海土地　(b) 陆地块　(h) 海洋块　(i) 海杂波　(j) 强虚警目标

图10　FUSAR-Ship数据集中的10种图像

3.2　实验设置

MSTAR数据集和FUSAR-Ship数据集中不同类别的图像大小不同。为了避免尺寸与噪声的影响，本文直接将所有SAR图像的中心裁剪为96×96大小的输入模型，在此过程中未进行任何姿态修正。

本文在Intel i7 9800X CPU、NVIDIA RTX 2080Ti GPU、16 GB内存、CUDA 11.1、CuDNN 7.6.5、Python 3.7.3语言、Pytorch框架的计算机上进行实验，采用随机梯度下降算法[28]。训练时设置算法迭代50次，学习率为0.01，动量系数为0.95，批次大小为16，权重衰减系数为0.000 1，批次归一化常数ε为0.000 01，随机失活率为0.2，耦合系数c_{ij}为3[26]。

3.3　标准操作条件下的识别性能

本节在SOC下对该模型进行评估[10]。实验使用的训练集为17°俯仰角的车辆目标，测试集为15°俯仰角的车辆目标，SOC下的训练集和测试集如表1所示。

表1　SOC下的训练集和测试集

类型	训练集数量/个	测试集数量/个
2S1	299	274
ZSU-234	299	274
BRDM-2	298	274
BTR-60	256	195
BMP-2	233	195
BTR-70	233	196
D-7	299	274
ZIL-131	299	274
T-62	299	273
T-72	232	196

本节分别使用STN、MSSTN和MSCSTN这3种模型进行实验并对比实验结果。STN与MSSTN模型结构如图11和图12所示。在使用相同实验设置并经过充分训练的情况下，STN、MSSTN和MSCSTN的详细测试结果如表2~表4所示。MSCSTN的整体识别准确率达到了99.92%。相比于STN和MSSTN，MSCSTN的整体识别准确率分别提高了0.79%和0.46%。MSCSTN在10类车辆目标中只有BMP-2、BTR-70这2种车辆分类存在错误，且这两种车辆识别准确率也都高于99%。

表5将MSCSTN与一些方法进行了对比。相比于SVM [17]、SRC [18]、A-ConvNet [10]、CNN＋SVM[29]、CNN-TL-bypass [30]、CNN＋属性散射中心（ASC，attributed scattering center）[31]、轻量卷积神经网络（LCNN，lightweight CNN）＋视觉注意力（VA，visual attention）[32]、贝叶斯特征匹配（BFM，bayesian feature matching）[33]、分层融合全局和局部特征（FGL，hierarchically fusing global and local feature）[33]、M-Net [11]、IFTS-Net [22]、SDF-Net [33]、FEC [34]、特征生成和校准的局部分类（LcFGC，local classification with feature generation and calibration）[35]、多模态特征融合学习（MoFFL，multimodal feature fusion learning）[36]和

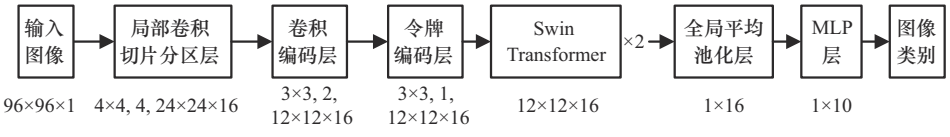


图11　STN模型结构

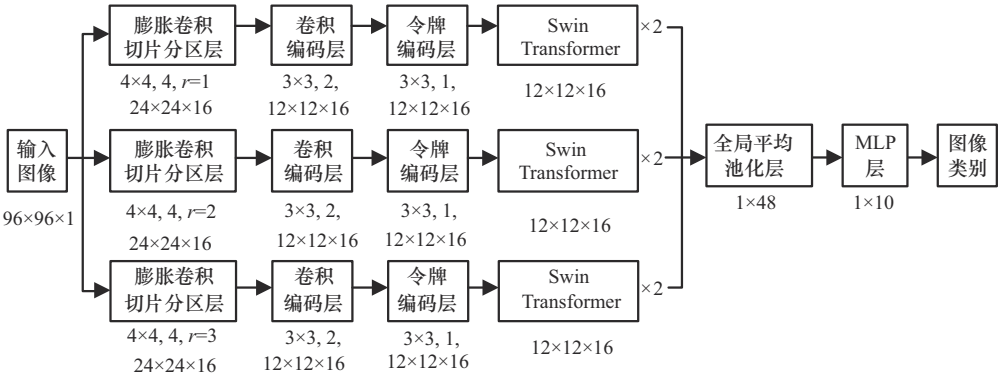


图 12　MSSTN 模型结构

表 2　　**SOC 下 STN 的实验结果**

类型	2S1	ZSU-234	BRDM-2	BTR-60	BMP-2	BTR-70	D-7	ZIL-131	T-62	T-72	总计
2S1	**273**	0	0	0	0	0	0	0	0	0	
ZSU-234	0	**272**	**2**	0	0	0	0	**2**	0	0	
BRDM-2	0	**1**	**270**	0	0	**1**	0	0	0	0	
BTR-60	0	0	0	**192**	0	0	0	0	0	0	
BMP-2	**1**	0	**1**	**2**	**192**	0	**1**	**1**	0	0	
BTR-70	0	0	**1**	0	**1**	**194**	0	0	0	0	
D-7	0	0	0	**1**	0	**1**	**272**	0	0	0	
ZIL-131	0	0	0	0	**1**	0	0	**271**	0	0	
T-62	0	**1**	0	0	**1**	0	**1**	0	**273**	**1**	
T-72	0	0	0	0	0	0	0	0	0	**195**	
识别准确率	99.64%	99.27%	98.54%	98.46%	98.46%	98.98%	99.27%	98.91%	100%	99.49%	**99.13%**

表 3　　**SOC 下 MSSTN 的实验结果**

类型	2S1	ZSU-234	BRDM-2	BTR-60	BMP-2	BTR-70	D-7	ZIL-131	T-62	T-72	总计
2S1	**273**	0	0	0	0	**1**	0	0	0	0	
ZSU-234	0	**274**	0	**1**	0	0	0	0	**1**	**1**	
BRDM-2	0	0	**273**	0	0	0	0	**1**	0	0	
BTR-60	0	0	0	**192**	0	0	0	0	0	**1**	
BMP-2	0	0	0	0	**195**	**1**	0	0	0	0	
BTR-70	0	0	0	0	0	**193**	0	0	0	0	
D-7	**1**	0	0	**1**	0	0	**274**	0	0	0	
ZIL-131	0	0	0	**1**	0	0	0	**271**	0	0	
T-62	0	0	0	0	0	0	0	0	**272**	0	
T-72	0	0	**1**	0	0	**1**	0	**2**	0	**194**	
识别准确率	99.64%	100%	99.64%	98.46%	100%	98.47%	100%	98.91%	99.63%	98.98%	**99.46%**

目标阴影掩膜辅助学习（TSMAL，target-shadow mask assistance learning）[37]，MSCSTN 的识别准确率分别提高了 3.14%、3.63%、0.79%、0.77%、0.83%、0.51%、0.38%、4.30%、0.84%、0.21%、1.02%、0.34%、0.33%、0.74%、0.17% 和 0.55%。实验结果表明，MSCSTN 具有良好的识别性能。

表4　　SOC下MSCSTN的实验结果

类型	2S1	ZSU-234	BRDM-2	BTR-60	BMP-2	BTR-70	D-7	ZIL-131	T-62	T-72	总计
2S1	**274**	0	0	0	0	0	0	0	0	0	
ZSU-234	0	**274**	0	0	0	0	0	0	0	0	
BRDM-2	0	0	**274**	0	0	**1**	0	0	0	0	
BTR-60	0	0	0	**195**	0	0	0	0	0	0	
BMP-2	0	0	0	0	**194**	0	0	0	0	0	
BTR-70	0	0	0	0	0	**195**	0	0	0	0	
D-7	0	0	0	0	**1**	0	**274**	0	0	0	
ZIL-131	0	0	0	0	0	0	0	**274**	0	0	
T-62	0	0	0	0	0	0	0	0	**273**	0	
T-72	0	0	0	0	0	0	0	0	0	**196**	
识别准确率	100%	100%	100%	100%	99.49%	99.49%	100%	100%	100%	100%	**99.92%**

表5　SOC下MSCSTN与现有方法的识别准确率对比

方法	识别准确率
SVM [17]	96.78%
A-ConvNet [10]	99.13%
CNN-TL-bypass[30]	99.09%
LCNN＋VA [32]	99.54%
FGL [33]	99.08%
IFTS-Net [22]	98.90%
FEC [34]	99.59%
MoFFL [36]	99.75%
MSCSTN	**99.92%**
SRC [18]	96.29%
CNN＋SVM[29]	99.15%
CNN＋ASC[31]	99.41%
BFM [33]	95.62%
M-Net [11]	99.71%
SDF-Net [33]	99.58%
LcFGC [35]	99.18%
TSMAL [37]	99.37%

3.4　扩展操作条件下的识别性能

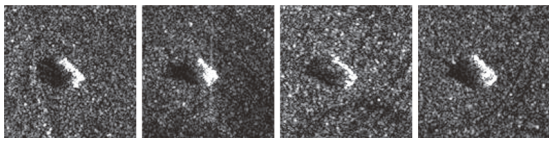
为了进一步验证所提方法的可靠性、稳定性和泛化能力，本节在几种EOC下对该模型进行评估。

3.4.1　EOC 1：大俯仰角变化

本节选用2S1、ZSU-234、BRDM-2和T-72这4种车辆目标进行实验。训练集和测试集的俯仰角分别为17°和30°，EOC 1下的训练集和测试集如表6所示。图13展示了不同俯仰角下的SAR图像。表7给出了EOC 1下MSCSTN的实验结果。表8将MSCSTN与几种具有代表性的方法进行了对比。相比于SVM[17]、SRC[18]、A-ConvNet[10]、CNN＋SVM[29]、LCNN＋VA[32]、M-Net[11]、SCN[38]、IFTS-Net[22]、SDF-Net[33]、LcFGC[35]、MoFFL[36]和TSMAL[37]，MSCSTN的识别准确率分别提高了7.21%、0.73%、3.01%、1.92%、1.65%、1.65%、0.35%、2.27%、1.97%、2.27%、0.52%和1.13%。

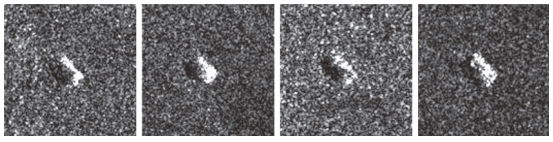
表6　　EOC 1下的训练集和测试集

类型	训练集数量/个	测试集数量/个
2S1	299	288
ZSU-234	299	288
BRDM-2	298	287
T-72	299	288



(a1) 2S1　(a2) ZSU-234　(a3) BRDM-2　(a4) T-72

(a) 俯仰角为17°



(b1) 2S1　(b2) ZSU-234　(b3) BRDM-2　(b4) T-72

(b) 俯仰角为30°

图13　不同俯仰角下的SAR图像

表 7　EOC 1 下 MSCSTN 的实验结果

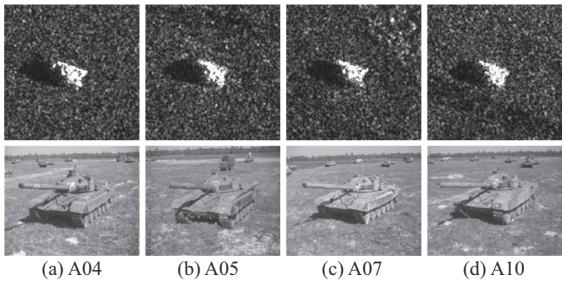
类型	2S1	ZSU-234	BRDM-2	T-72	总计
2S1	**286**	**4**	0	0	
ZSU-234	**2**	**284**	0	0	
BRDM-2	0	0	**286**	**3**	
T-72	0	0	**1**	**285**	
识别准确率	99.31%	98.61%	99.65%	98.96%	**99.13%**

表 8　EOC 1 下 MSCSTN 与现有方法的识别准确率对比

方法	识别准确率
SVM [17]	91.92%
A-ConvNet [10]	96.12%
LCNN＋VA [32]	97.48%
SCN [38]	98.78%
SDF-Net [33]	97.16%
MoFFL [36]	98.61%
MSCSTN	**99.13%**
SRC [18]	98.4%
CNN＋SVM [29]	97.21%
M-Net [11]	97.48%
IFTS-Net [22]	96.86%
LcFGC [35]	96.86%
TSMAL [37]	98.00%

3.4.2　EOC 2：型号变化

图 14 展示了 4 种不同配置的 T-72 图像。EOC 2 下的训练集（俯仰角为 17°）和测试集（俯仰角为 15°和 17°）如表 9 所示，用于测试的 BMP-2 和 T-72 的配置不包括在训练集中。表 10 给出了详细的实验结果。表 11 将 MSCSTN 与几种具有代表性的方法进行了对比。MSCSTN 在每一类图像目标上的识别准确率均达到 98% 以上，且整体识别准确率要高于其他对比方法。实验结果表明，在型号变化下，MSCSTN 可以捕获同一目标不同配置间的差异，取得较为理想的效果。



(a) A04　(b) A05　(c) A07　(d) A10

图 14　4 种不同配置的 T72 图像

表 9　EOC 2 下的训练集和测试集

类型	训练集		测试集	
	序列号	数量/个	序列号	数量/个
BMP-2	SN9563	233	SN9566	428
			SNC21	429
BRDM-2	E71	298	—	0
BTR-70	C71	233	—	0
T-72	132	232	SN812	426
			A04	573
			A05	573
			A07	573
			A10	567

表 10　EOC 2 下 MSCSTN 的实验结果

类型	序列号	BMP-2	BRDM-2	BTR-70	T-72	识别准确率
BMP-2	SN9566	**420**	**4**	**3**	**1**	98.13%
	SNC21	**425**	**2**	**2**	0	99.06%
T-72	SN812	**4**	0	0	**422**	99.06%
	A04	**2**	0	0	**571**	99.65%
	A05	**2**	0	**2**	**569**	99.30%
	A07	**4**	0	**2**	**567**	98.95%
	A10	**3**	**1**	0	**563**	99.29%
总计						**99.10%**

表 11　EOC 2 下 MSCSTN 与现有方法的识别准确率对比

方法	识别准确率
SVM [17]	96.64%
A-ConvNet [10]	98.12%
LCNN＋VA [32]	98.66%
FGL [33]	98.46%
IFTS-Net [22]	95.46%
FEC [34]	98.48%
MoFFL [36]	98.57%
MSCSTN	**99.10%**
SRC [18]	95.12%
CNN＋SVM [29]	98.09%
BFM [33]	97.90%
M-Net [11]	98.74%
SDF-Net [33]	98.76%
LcFGC [35]	98.40%
TSMAL [37]	98.20%

3.4.3　EOC 3：噪声污染

本节根据预定义的信噪比（SNR，signal-to-noise ratio）向所有测试样本添加加性白高斯噪声来模拟受随机噪声干扰的SAR图像，而对训练样本不做任何修改[34]。本节实验使用的训练集与测试集和3.3节相同。图15显示了不同SNR下的加噪图像。随着噪声污染的恶化，越来越多的目标特征被淹没在噪声中，这必然会增加识别难度。图16显示了不同方法在不同SNR下的识别准确率对比。从图16中可以看出，当SNR从10 dB降低到−10 dB时，所有方法的识别性能均发生显著下降。MSCSTN在不同SNR下都有最高的识别准确率，并且在SNR为−10 dB时识别准确率仍可达到92.87%，远高于其他方法。实验结果表明，MSCSTN在受到不同程度噪声污染场景中鲁棒性较强。

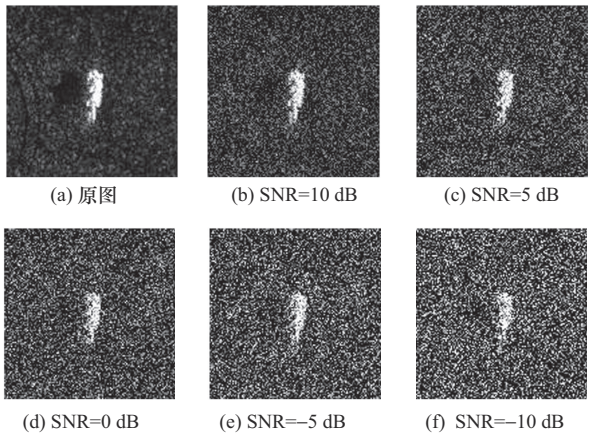


图15　不同SNR下的加噪图像

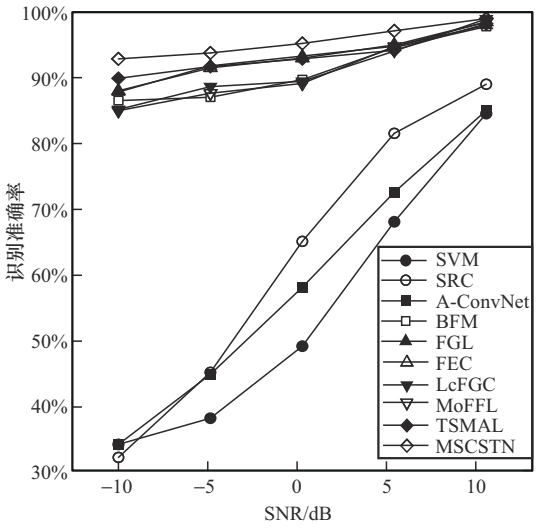


图16　不同SNR下不同方法的识别准确率对比

3.4.4　EOC 4: 分辨率变化

本节根据预定义分辨率从频域中心提取指定比例的数据，得到不同分辨率的测试样本图像[34]，如图17所示。在此过程中对训练样本不做任何修改。本节实验使用的训练集与测试集和3.3节相同。图18展示了不同分辨率下不同方法的识别准确率对比。从图18中可以看出，随着分辨率的降低，目标信息逐渐不清晰，所有方法的识别性能逐渐下降，但MSCSTN的识别性能始终优于其他对比方法。结果表明，MSCSTN在处理测试集中样本分辨率变化问题时具有一定的潜力。

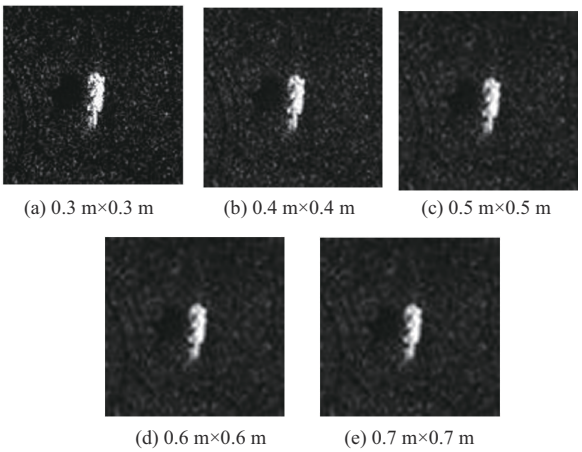


图17　不同分辨率下的SAR图像

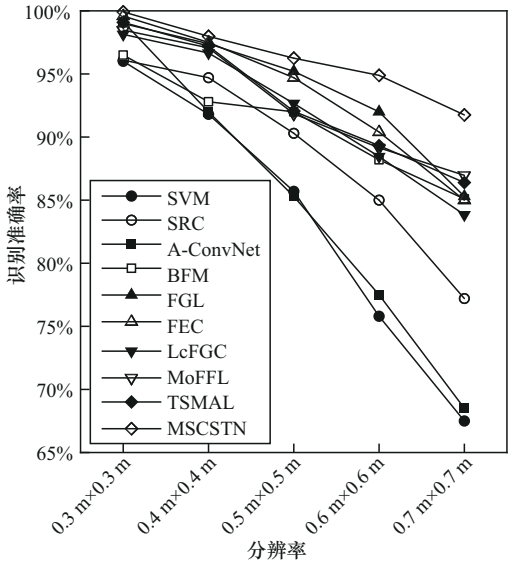


图18　不同分辨率下不同方法的识别准确率对比

3.4.5　EOC 5: 训练样本数量有限

本节选取俯仰角为17°下的部分图像作为训练集和俯仰角为15°下的全部图像作为测试集。图19显示

了不同训练样本比例下不同方法的识别准确率对比。A-ConvNet[9]、M-Net[11]、SCN[38]、IFTS-Net[12]、SDF-Net[33]、FEC[34]、LcFGC[35]、MoFFL[36]和TSMAL[37]在10%训练样本下的识别准确率分别为73.44%、82.0%、79.2%、80.2%、76.4%、85.1%、85.79%、85.63%和84.41%。此时MSCSTN的识别准确率仍能达到96.58%，远高于其他方法的识别准确率。实验结果表明，当训练样本数量有限时，MSCSTN可以取得相对满意的识别性能。

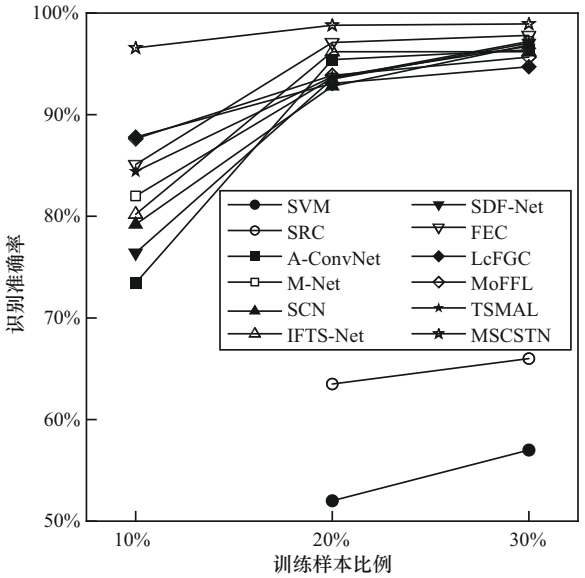


图19　不同训练样本比例下不同方法的识别准确率对比

3.5　FUSAR-Ship数据集下的识别性能

本节利用FUSAR-Ship数据集对该模型进行评估，其训练样本和测试样本的详细信息如表12所示。表13给出了MSCSTN的实验结果。表14将MSCSTN与几种具有代表性的方法进行了对比。在FUSAR-Ship数据集下，MSCSTN获得了93.57%的识别准确率。相比于SVM[17]、SRC[18]、A-ConvNet[10]、CNN+SVM[29]、LCNN+VA[32]、BFM[33]、M-Net[11]、IFTS-Net[22]、SDF-Net[33]、FEC[34]、LcFGC[35]、MoFFL[36]和TSMAL[37]，MSCSTN的识别准确率分别提高了14.66%、12.42%、12.66%、7.55%、9.96%、3.77%、6.50%、3.47%、5.61%、4.53%、3.47%、2.38%和3.06%。实验结果表明，在FUSAR-Ship数据集下，MSCSTN仍然可以取得相对满意的识别性能。

表12　FUSAR-Ship数据集中的训练样本和测试样本

类型	训练集数量/个	测试集数量/个
货船	366	156
渔船	248	106
油轮	150	64
其他船舶	312	133
桥梁	1023	438
沿海土地	707	303
陆地块	1137	487
海洋块	1250	535
海杂波	1378	590
强虚警目标	299	128

表13　FUSAR-Ship数据集下MSCSTN的实验结果

类型	货船	渔船	油轮	其他船舶	桥梁	沿海土地	陆地块	海洋块	海杂波	强虚警目标	总计
货船	141	**4**	**4**	**2**	0	0	0	0	0	0	
渔船	0	96	**2**	**4**	0	0	0	0	**3**	0	
油轮	**7**	**2**	53	**1**	0	0	0	0	0	**1**	
其他船舶	**4**	**4**	**3**	124	**3**	0	**2**	0	**2**	0	
桥梁	**4**	0	0	**1**	415	**6**	**8**	**6**	0	0	
沿海土地	0	0	**1**	0	**5**	285	**6**	**5**	**6**	0	
陆地块	0	0	**1**	0	**8**	**5**	466	**7**	**10**	**2**	
海洋块	0	0	0	0	**5**	**2**	**1**	505	**9**	**2**	
海杂波	0	0	0	0	**2**	**3**	**1**	**10**	551	**8**	
强虚警目标	0	0	0	**1**	0	**2**	**3**	**2**	**9**	115	
识别准确率	90.38%	90.57%	82.81%	93.23%	94.75%	94.06%	95.69%	94.39%	93.39%	89.84%	**93.57%**

表14　FUSAR-Ship下MSCSTN与现有方法识别准确率对比

方法	识别准确率
SVM [17]	78.91%
A-ConvNet [10]	80.91%
LCNN＋VA [32]	83.61%
M-Net [11]	87.07%
SDF-Net [33]	87.96%
LcFGC [35]	90.10%
TSMAL [37]	90.51%
SRC [18]	81.15%
CNN＋SVM [29]	86.02%
BFM [33]	89.80%
IFTS-Net [22]	90.10%
FEC [34]	89.04%
MoFFL [36]	91.19%
MSCSTN	**93.57%**

3.6　消融分析

3.6.1　DCPP-CTE有效性验证

本节将DCPP-CTE中膨胀卷积切片分区替换为简单的局部卷积切片分区，对比和验证膨胀卷积切片分区的有效性。不同条件下2种分区方法的识别准确率对比如图20所示。值得注意的是，在EOC 3、EOC 4、EOC 5下选择了最苛刻的条件用作对比，EOC 3下SNR为−10 dB，EOC 4下分辨率为0.7 m × 0.7 m，EOC 5下训练样本比例为10%。当使用局部卷积切片分区时，模型识别准确率普遍要低于使用膨胀卷积切片分区时模型识别准确率。膨胀卷积切片分区有助于模型自由提取不同感受野区域信息，最大化提高其识别性能。

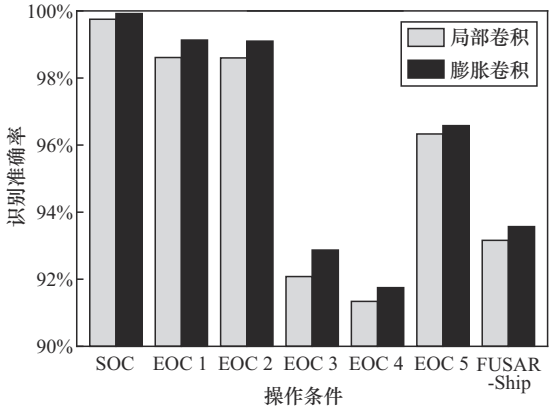


图20　不同条件下2种分区方法的识别准确率对比

本节将DCPP-CTE中矢量胶囊编码替换为简单的标量卷积编码，对比和验证矢量胶囊编码的有效性。不同条件下2种编码方法的识别准确率对比如图21所示。EOC 2下使用标量卷积编码时模型识别准确率要略高于使用矢量胶囊编码时模型识别准确率。而在其他条件下，使用标量卷积编码时模型识别准确率要低于使用矢量胶囊编码时模型识别准确率。矢量胶囊编码有助于模型捕获微小目标信息，深入理解特征。

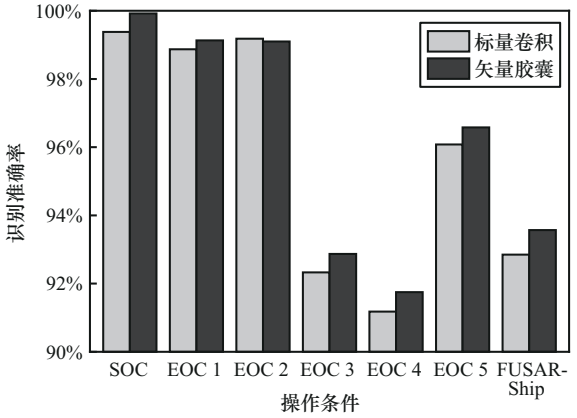


图21　不同条件下2种编码方法的识别准确率对比

不同分支可以通过不同膨胀率的膨胀卷积提取不同感受野区域信息。每个分支的膨胀卷积膨胀率对识别性能有着重要的影响。选择(1, 1, 1)、(1, 1, 2)、(1, 2, 2)、(2, 2, 2)、(2, 2, 3)、(2, 3, 3)、(3, 3, 3)和(1, 2, 3)共8组膨胀率大小组合评估其对性能的影响，识别准确率对比如图22所示。当选择(1, 2, 3)这种组合时，模型识别准确率普遍较高。不同膨胀率的膨胀卷积可以提取图像中不同维度的空间特征，提高模型识别性能。

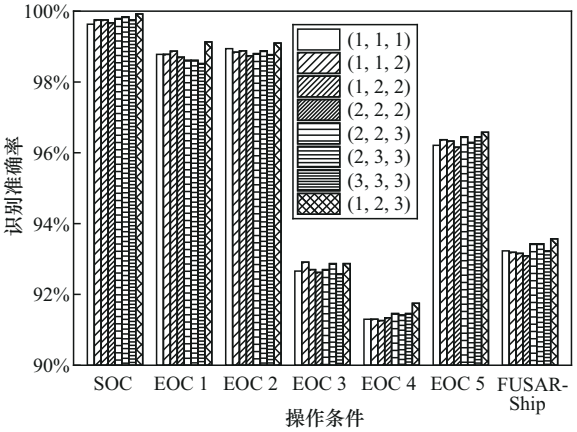


图22　不同条件下多种膨胀率大小组合的识别准确率对比

3.6.2　3DCSTM有效性验证

本节将3DCSTM替换为普通的Swin Transformer，对比和验证3DCSTM的有效性。不同条件下2种模块的识别准确率对比如图23所示。EOC 2和EOC 3下使用Swin Transformer和3DCSTM取得了相似的模型识别准确率。而在其他条件下使用Swin Transformer时模型识别准确率要低于使用3DCSTM时模型识别准确率。3DCSTM有助于模型梳理来自多头注意力的信息，提高语义特征提取能力。

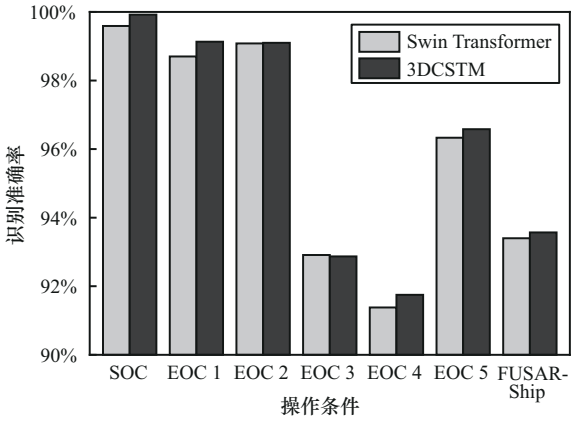


图23　不同条件下2种模块的识别准确率对比

3.7　不同版本MSCSTN的性能分析

本节将MSCSTN模型的3条并行编码结构缺省为1条、2条，得到3个版本的MSCSTN，并将其标记为MSCSTN(Base)、MSCSTN(Large)和MSCSTN(Huge)。通过对比不同方法的识别准确率和推理时间，探究在不同编码分支数量下MSCSTN模型的识别与实时性能。为了公平起见，所有的测试实验均在3.2节介绍的环境中进行且在推理分析时批次大小设为1。不同版本MSCSTN与现有方法的性能对比如表15所示。MSCSTN(Base)与A-ConvNet[10]在SOC、EOC1、EOC2和FUSAR-Ship等条件下推理时间相近，略低于CNN＋SVM[29]、LCNN＋VA[32]的推理时间，仅为MSCSTN(Huge)推理时间的40%左右，但却可以获得99.59%、98.08%、98.71%、89.42%的识别准确率，超过了A-ConvNet[10]、CNN＋SVM[29]、LCNN＋VA[32]在同条件下的识别准确率，在一定程度上验证了提出的DCPP-CTE和3DCSTM的有效性。MoFFL[36]需要将从相位历史和散射域学习到的特征与从现成的深度特征提取器获得的图像特征融合，计算量较大，推理时间远大于MSCSTN(Huge)，然而，在多种条件下其识别准确率却低于MSCSTN(Huge)。TSMAL[37]利用目标信息与阴影信息学习卷积互补表示，推理时间略大于MSCSTN(Huge)，然而，在多种条件下其识别准确率也低于MSCSTN(Huge)。可以发现，使用适中的推理时间，MSCSTN(Huge)在多种测试条件下获得了最高的识别准确率。这也证明了提出的MSCSTN(Huge)的优越性。当计算资源充足时，可以选择本文提出的模型MSCSTN(Huge)。当计算资源受限时，MSCSTN(Base)、MSCSTN(Large)也是不错的选择。

4　结束语

本文在通用MSTAR数据库上验证了MSCSTN的有效性。在SOC下，MSCSTN的识别准确率为

表15　不同版本MSCSTN与现有方法的性能对比

方法	编码分支数量	SOC		EOC 1		EOC 2		FUSAR-Ship	
		识别准确率	推理时间/s	识别准确率	推理时间/s	识别准确率	推理时间/ s	识别准确率	推理时间/s
A-ConvNet[10]	—	99.13%	0.002 8	96.12%	0.002 6	98.12%	0.002 7	80.91%	0.002 8
CNN＋SVM[29]	—	99.15%	0.004 7	97.21%	0.004 8	98.09%	0.004 8	86.02%	0.004 8
LCNN＋VA[32]	—	99.54%	0.004 5	97.48%	0.004 2	98.66%	0.004 4	83.61%	0.004 3
M-Net[11]	—	99.71%	0.006 4	97.48%	0.006 2	98.74%	0.006 4	87.07%	0.006 4
MoFFL[36]	—	99.75%	0.013 8	98.61%	0.013 3	98.57%	0.013 4	91.19%	0.013 3
TSMAL[37]	—	99.37%	0.009 2	98.00%	0.009 0	98.20%	0.009 2	90.51%	0.009 1
MSCSTN(Base)	1	99.59%	0.003 0	98.08%	0.002 9	98.71%	0.003 0	89.42%	0.003 0
MSCSTN(Large)	2	99.84%	0.005 8	98.95%	0.005 7	98.99%	0.005 8	93.40%	0.005 6
MSCSTN(Huge)	3	99.92%	0.008 5	99.13%	0.008 6	99.10%	0.008 5	93.57%	0.008 5

99.92%，优于绝大多数已经提出的方法。在大俯仰角变化、型号变化、噪声污染、分辨率变化和训练样本数量有限等EOC与FUSAR-Ship数据集下，MSCSTN的识别准确率都优于其他方法。因此，MSCSTN具有有效性和鲁棒性，在SAR ATR系统中具有很大的应用潜力。

在今后的研究中，鉴于SAR ATR对实时性的高要求，需专门探索模型轻量化设计，平衡参数规模与任务复杂度，减少冗余参数，以提升效率。此外，尽管GPU环境加速了深度网络运算，但在SAR ATR中，向DSP与FPGA等并行环境移植仍面临挑战，需深入研究解决方案。

参考文献：

[1] MOREIRA A, PRATS-IRAOLA P, YOUNIS M, et al. A tutorial on synthetic aperture radar[J]. IEEE Geoscience and Remote Sensing Magazine, 2013, 1(1): 6-43.

[2] 赵泉华, 王肖, 李玉, 等. 基于多特征加权的SAR影像舰船检测优化方法[J]. 通信学报, 2020, 41(3): 91-101.
ZHAO Q H, WANG X, LI Y, et al. Ship detection optimization method in SAR imagery based on multi-feature weighting[J]. Journal on Communications, 2020, 41(3): 91-101.

[3] WANG H X, YANG H R, HUANG Y B, et al. Classification of land cover in complex terrain using Gaofen-3 SAR ascending and descending orbit data[J]. Remote Sensing, 2023, 15(8): 2177.

[4] ZHANG W H, JIAO L C, LIU F, et al. Adaptive contourlet fusion clustering for SAR image change detection[J]. IEEE Transactions on Image Processing, 2022, 31: 2295-2308.

[5] DENG J W, LI M D, CHEN S W. Urban damage-level estimation with reconstructed quad-pol SAR data from dual-pol SAR mode[J]. IEEE Geoscience and Remote Sensing Letters, 2024, 21: 4007305.

[6] ZHU M Z, HU X R, FENG Z P, et al. Unveiling SAR target recognition networks: Adaptive perturbation interpretation for enhanced understanding[J]. Neurocomputing, 2024, 600: 128137.

[7] OWIRKA G J, VERBOUT S M, NOVAK L M. Template-based SAR ATR performance using different image enhancement techniques[C]//Algorithms for Synthetic Aperture Radar Imagery VI. Bellingham: SPIE Press, 1999: 302-319.

[8] KECHAGIAS-STAMATIS O, AOUF N. Automatic target recognition on synthetic aperture radar imagery: a survey[J]. IEEE Aerospace and Electronic Systems Magazine, 2021, 36(3): 56-81.

[9] LI J W, YU Z T, YU L, et al. A comprehensive survey on SAR ATR in deep-learning era[J]. Remote Sensing, 2023, 15(5): 1454.

[10] CHEN S Z, WANG H P, XU F, et al. Target classification using the deep convolutional networks for SAR images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(8): 4806-4817.

[11] SHANG R H, WANG J M, JIAO L C, et al. SAR targets classification based on deep memory convolution neural networks and transfer parameters[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2018, 11(8): 2834-2846.

[12] LIU Z G, WANG L F, WEN Z D, et al. Multilevel scattering center and deep feature fusion learning framework for SAR target recognition[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5227914.

[13] HOU Y C, XU T, HU H P, et al. MdpCaps-csl for SAR image target recognition with limited labeled training data[J]. IEEE Access, 2020, 8: 176217-176231.

[14] YU C L, ZHAI Y K, HUANG H F, et al. Capsule broad learning system network for robust synthetic aperture radar automatic target recognition with small samples[J]. Remote Sensing, 2024, 16(9): 1526.

[15] HE X, CHEN Y S, HUANG L B, et al. Swin transformer with improved blind-spot network for SAR target classification[C]//Proceedings of the IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium. Piscataway: IEEE Press, 2024: 7268-7271.

[16] LI K Y, ZHANG M, XU M P, et al. Ship detection in SAR images based on feature enhancement swin transformer and adjacent feature fusion[J]. Remote Sensing, 2022, 14: 3186.

[17] SABOUR S, FROSST N, HINTON G E, et al. Dynamic routing between capsules[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 3859-3869.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv Preprint, arXiv: 1706.03762, 2017.

[19] PATWARDHAN N, MARRONE S, SANSONE C. Transformers in the real world: a survey on NLP applications[J]. Information, 2023, 14 (4): 242.

[20] LIU Y, ZHANG Y, WANG Y X, et al. A survey of visual transformers[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(6): 7478-7498.

[21] 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.

[22] RAO W Q, GAO L R, QU Y, et al. Siamese transformer network for hyperspectral image target detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5526419.

[23] CHENG B W, MISRA I, SCHWING A G, et al. Masked-attention mask transformer for universal image segmentation[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 1280-1289.

[24] DING M, YANG Z, HONG W, et al. Cogview: mastering text-to-image generation via transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 19822-19835.

[25] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 9992-10002.

[26] WU H K, ZHANG J G, HUANG K Q, et al. FastFCN: rethinking dilated convolution in the backbone for semantic segmentation[J]. arXiv Preprint, arXiv: 1903.11816, 2019.

[27] HOU X Y, AO W, SONG Q, et al. FUSAR-Ship: building a high-resolution SAR-AIS matchup dataset of Gaofen-3 for ship detection and recognition[J]. Science China Information Sciences, 2020, 63(4): 140303.

[28] KENDALL A, BADRINARAYANAN V, CIPOLLA R. Bayesian SegNet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding[J]. arXiv Preprint, arXiv: 1511.02680, 2015.

[29] GAO F, HUANG T, SUN J P, et al. A new algorithm for SAR image target recognition based on an improved deep convolutional neural network[J]. Cognitive Computation, 2019, 11(6): 809-824.
[30] HUANG Z L, PAN Z X, LEI B. Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data[J]. Remote Sensing, 2017, 9(9): 907.
[31] JIANG C J, ZHOU Y. Hierarchical fusion of convolutional neural networks and attributed scattering centers with application to robust SAR ATR[J]. Remote Sensing, 2018, 10(6): 819.
[32] SHAO J Q, QU C W, LI J W, et al. A lightweight convolutional neural network based on visual attention for SAR image target classification[J]. Sensors, 2018, 18(9): 3039.
[33] DING B Y, WEN G J, MA C H, et al. An efficient and robust framework for SAR target recognition by hierarchically fusing global and local features[J]. IEEE Transactions on Image Processing, 2018, 27(12): 5983-5995.
[34] ZHANG J S, XING M D, XIE Y Y. FEC: a feature fusion framework for SAR target recognition based on electromagnetic scattering features and deep CNN features[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(3): 2174-2187.
[35] WANG S Y, WANG Y H, LIU H W, et al. A few-shot SAR target recognition method by unifying local classification with feature generation and calibration[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 62: 5200319.
[36] WEN Z D, YU Y L, WU Q. Multimodal discriminative feature learning for SAR ATR: a fusion framework of phase history, scattering topology, and image[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 62: 5200414.
[37] GUO S, CHEN T, WANG P H, et al. TSMAL: target-shadow mask assistance learning network for SAR target recognition[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 17: 18247-18263.
[38] GUO Y R, PAN Z X, WANG M M, et al. Learning capsules for SAR target recognition[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 4663-4673.

[作者简介]

侯宇超（1995– ），男，山西吕梁人，博士，山西师范大学副教授，贵州大学访问学者、硕士生导师，主要研究方向为SAR影像智能解译、模式识别。

王洁（1977– ），女，山西临汾人，博士，山西师范大学教授、硕士生导师，主要研究方向为网络信息安全、机器学习等。

李洪涛（1984– ），男，山东临沂人，博士，山西师范大学教授、硕士生导师，主要研究方向为网络信息安全、图像处理等。

郝岩（1993– ），女，吉林白山人，博士，太原师范学院讲师、硕士生导师，主要研究方向为SAR影像智能解译、模式识别。

段晓旗（1990– ），男，山东泰安人，博士，贵州大学讲师、硕士生导师，主要研究方向为SAR影像智能解译、时空数据分析与挖掘。

黄凯文（1995– ），女，河南南阳人，博士，贵州大学讲师、硕士生导师，主要研究方向为网络信息安全、图像处理等。

田有亮（1982– ），男，贵州六盘水人，博士，贵州大学教授、博士生导师，主要研究方向为密码学与安全协议、大数据安全与隐私保护等。