激光与光电子学进展 先进成像

基于多层级自注意力增强的遥感目标检测

魏谢根[1,2]，曹林[2,3]，田澍[3*]，杜康宁[3]，宋沛然[3]，郭亚男[3]

[1]北京信息科技大学仪器科学与光电工程学院，北京 100101；

[2]北京信息科技大学光电测试技术及仪器教育部重点实验室，北京 100101；

[3]北京信息科技大学信息与通信系统信息产业部重点实验室，北京 100101

摘要 随着遥感图像分辨率的不断提高，遥感图像目标检测技术获得了更广泛的关注。针对遥感图像中背景复杂噪声多、目标方向任意且目标尺寸变化大等问题，提出一种基于多层级局部自注意力增强的遥感目标检测算法。首先，在Oriented R-CNN骨干网络中引入Swin Transformer特征提取模块，使用具有移位窗口操作和层次设计的Transformer模块对特征提取的语义信息进行多层级局部信息建模。其次，使用Oriented RPN生成高质量的有向候选框。最后，将高斯分布之间的Kullback-Leibler divergence(KLD)作为回归损失函数，使得参数梯度能够根据对象的特征得到动态调整，更加准确地进行检测框的回归。所提算法在DOTA数据集和HRSC2016数据集上的平均精度均值(mAP)分别达77.2%和90.6%，和Oriented R-CNN算法相比，mAP分别提高了1.8个百分点和0.5个百分点。实验结果表明，所提算法能够有效地提高遥感图像目标检测精度。

关键词 旋转目标检测；遥感图像；Swin Transformer；高斯距离

中图分类号 TP753 **文献标志码** A **DOI**：10.3788/LOP223048

Remote Sensing Target Detection Based on Multilevel Self-Attention Enhancement

Wei Xiegen[1,2], Cao Lin[2,3], Tian Shu[3*], Du Kangning[3], Song Peiran[3], Guo Yanan[3]

[1]*School of Instrument Science and Opto-Electronics Engineering, Beijing Information Science & Technology University, Beijing* 100101, *China*;

[2]*Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument, Beijing Information Science & Technology University, Beijing* 100101, *China*;

[3]*Key Laboratory of Information and Communication Systems, Ministry of Information Industry, Beijing Information Science & Technology University, Beijing* 100101, *China*

Abstract Remote sensing image target detection technology has gained considerable attention with the improvement of remote sensing image resolution. This thesis proposes a remote sensing target detection algorithm based on multilevel local self-attention enhancement to solve such problems as complex background noise, arbitrary target direction, and large changes in target size in remote sensing images. First, the proposed algorithm adopts the Swin Transformer feature extraction module in an Oriented region-based convolutional neural network (R-CNN) backbone network, and the multilevel local information of feature-extracted semantic information is modeled using the Transformer module with shifted window operations and hierarchical design. Second, Oriented RPN is used to generate high-quality directed candidate boxes. Finally, the Kullback-Leibler divergence (KLD) between Gaussian distributions is regarded as the regression loss function, allowing the parameter gradient to be dynamically adjusted based on the object's characteristics for more accurate regression of the detection boxes. The mean average precision (mAP) of the proposed algorithm reaches 77.2% and 90.6% on the DOTA dataset and HRSC2016 dataset, respectively, and it is increased by 1.8 percentage points and 0.5 percentage points compared with the Oriented R-CNN algorithm. The results reveal that the proposed algorithm can effectively advance the target detection accuracy of remote sensing images.

Key words rotating object detection; remote sensing image; Swin Transformer; Gaussian distance

收稿日期：2022-11-14；**修回日期**：2022-12-23；**录用日期**：2023-01-04；**网络首发日期**：2023-02-07

基金项目：国家自然科学基金(62001032,62201066,62201066)、北京市教委科研计划(KZ202111232049,KM202111232014)

通信作者：*tianshu_0202@126.com

1 引 言

近年来，随着图像传感器技术的提升，高分辨率遥感图像在环境治理、军事侦察和智慧城市建设等领域得到了广泛的关注，因此对遥感图像目标检测算法进行研究与应用具有深远的意义[1]。

早期传统的目标检测算法主要是基于手工提取特征的，经典的算法有SIFTS[2]、HOG[3]等，之后使用支持向量机（SVM）等分类器对目标进行分类。传统算法具有计算量大、准确率低等问题，并且由于遥感图像成像方式的特殊性，成像区域中所包含的目标具有方向任意、尺度多样和背景复杂等特点。传统的检测方法在面对复杂的遥感图像时，对特征的描述能力十分有限[4]，不能够满足遥感图像目标检测精度不断提升的应用需求[5]。近年来，出现了学习能力强、可移植性好的深度学习算法。基于深度学习的遥感图像目标检测算法得到了广泛的关注。基于深度学习的目标检测算法包括单阶段检测算法[6]和双阶段检测算法。两者的差别在于是否有显式的区域特征提取过程。其中，单阶段检测器包括SSD[7]、RetinaNet[8]、YOLO[9]等经典算法，只需单次特征提取，模型就具有更快的检测速度，直接完成端到端的检测。在遥感图像检测领域，Li等[10]设计了一个用于检测像素级别标注的卫星船舶图像的单阶段检测器，该检测器能够对不同等级的特征图进行特征融合并提取多尺度特征，虽然可以准确定位目标，但是图像标注成本太高且算法检测速度一般。Pan等[11]提出了一种特征模块，使得模型可以根据大多数密集目标的组合方向调整神经元的感受野，提取到更多深层语义信息。单阶段算法虽然速度更快，但是一定程度上降低了检测精度。相比于单阶段算法，双阶段检测算法拥有显式的区域特征提取过程，具有更高的检测精度。双阶段检测器包括Fast R-CNN[12]、Faster R-CNN[13]等。基于遥感图像，郑哲等[14]提出了一种以Faster R-CNN为基础的框架，加入组合注意力机制，使用有向检测框来进行有向检测，以解决遥感图像目标紧密排列的问题，对高低层特征进行融合，有效地提高了检测精度。Lin等[15]以Mask R-CNN为基础框架，应用特征重用技术，来提升遥感图像目标检测算法的精度。

虽然主流方法在遥感图像目标检测任务中具有不错的效果，但是由于遥感图像具有密集排列、背景复杂、目标尺度变化大和目标方向不确定等特性，检测任务中仍然会出现目标漏检和错检等情况。针对上述问题，本文提出了一种基于多层级局部自注意力增强的遥感图像目标检测算法。以Oriented R-CNN模型框架[16]为基础，主要创新点包括两个方面。1）针对遥感图像密集排列、背景复杂的问题，引入Swin Transformer特征提取网络[17]。Swin Transformer基于移动窗口的分层特征提取，允许跨窗口连接，解决了普通卷积对局部几何特征提取能力不足的问题，使得遥感图像跨层级的局部几何特征能够得到更加准确的提取。2）针对遥感图像目标尺寸小、方向不一的问题，将损失模块融进网络模块中。引入Kullback-Leibler divergence（KLD）[18]高斯距离损失函数，将矩形框的表示转换成二维高斯分布，解决边界问题和类正方形检测问题，使得模型具有自适应调整能力，进一步提高遥感图像目标检测精度。

2 相关工作

2.1 Oriented R-CNN模型

遥感图像目标检测一般需要对目标的方向进行判别，所以使用有向的边界框，能够更好地贴合目标。同时，得益于通用检测的快速发展，当前大部分的旋转检测模型都是经典的通用检测器，例如，经典的针对遥感图像目标检测的二阶段检测算法Oriented R-CNN就是基于Faster R-CNN改进而来的。Oriented R-CNN引入了新的有向对象表示方案，称为中点偏移表示法，可以精确地表示出有向候选框信息，减少候选框生成过程的冗余计算。同时因为两阶段算法在第一阶段生成有向候选框，在第二阶段对有向候选框进行回归和分类，相比一阶段目标检测算法，多了有向候选框生成阶段，影响了两阶段算法的检测速度。由此可知，制约两阶段检测算法速度的问题往往来源于有向候选框生成阶段。针对以上问题，Oriented R-CNN模型以Faster R-CNN为模型结构的基础。第一阶段提出了Oriented RPN（Oriented region proposal network），产生有效、高质量的有向候选框，第二阶段为Oriented R-CNN检测头检测，以更快的速度对有向候选框进行精细回归和分类。

2.2 有向提取候选框网络

有向提取候选框（Oriented RPN）是基于RPN的，可以快速、准确地生成多个有向候选框。RPN主要用于生成候选区域，能够处理不同尺寸的输入图片，最后输出一组具有判别分数的候选区域。RPN的核心是锚点（anchor），可以利用锚点生成候选区域。锚点是特征图上当前滑窗的中心在原像素空间的映射点。以这个锚点为中心，生成一组大小和尺寸固定的候选窗口，同时在特征图上使用滑动窗口的操作方式来产生预测框，最后以此anchor为中心生成预设的k个预测框进行二分类和边框回归，生成最终的预测框。

Oriented RPN是在RPN的回归分支上增加角度信息和偏移量信息的，以此来生成有向矩形候选框，对于每个位置的anchor，Oriented RPN输出为$(x, y, w, h, \Delta\alpha, \Delta\beta)$，其中$x$和$y$表示有向候选框外接矩形中心点的横坐标和纵坐标，$w$和$h$表示外接矩形的宽和高，$\Delta\alpha$和$\Delta\beta$表示外接矩形中点和检测框顶点的偏移量。最后，利用Oriented R-CNN提出的全新的有向候

选框表示方法——中点偏移表示法，来得到有向候选框的四点坐标集。

中点偏移表示法是一种全新的旋转目标表示方法，如图 1 所示，该方法使用 6 个回归参数 $(x, y, w, h, \Delta\alpha, \Delta\beta)$ 来表示有向目标，在回归水平 (x, y, w, h) 的同时，预测水平候选框中点的偏移 $\Delta\alpha$ 和 $\Delta\beta$。随后，根据对称性原理就可以获得有向候选框 4 个顶点 v_1、v_2、v_3、v_4 的坐标值，公式为

$$\begin{cases} v_1 = (x, y - h/2) + (\Delta\alpha, 0) \\ v_2 = (x + w/2, y) + (0, \Delta\beta) \\ v_3 = (x, y + h/2) + (-\Delta\alpha, 0) \\ v_4 = (x - w/2, y) + (0, -\Delta\beta) \end{cases} 。 \tag{1}$$

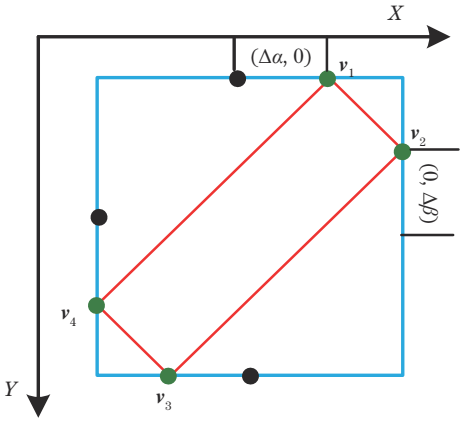

图 1　中点偏移表示法的原理

Fig. 1　Schematic of midpoint offset representation method

最后，就可以生成一个高质量的有向候选框。中点偏移表示法由于中点偏移的范围永远在边界框的边上，所以为有向候选框的回归提供了约束，同时避免了有向候选框生成网络的复杂设计，因此参数量远小于其他算法。

2.3　回归分类网络 Oriented R-CNN Head

Oriented R-CNN Head（Oriented R-CNN 检测头）是在传统 Faster R-CNN 检测框架的基础上进行改进的，通过在 Faster R-CNN 回归的分支上添加角度预测参数，进行精细回归和分类。由于 Oriented RPN 生成的有向候选框大部分为平行四边形，不利于进行量化操作，因此为了方便计算，检测头首先使用旋转感兴趣区域对齐（Rotated RoIAlign）模块提取有向候选框的边特征，在 Rotated RoIAlign 阶段，将生成的有向平行四边形候选框调整为有向矩形候选框，更加方便检测目标的量化；随后再将旋转不变的特征送入全连接层，进行精细回归和分类。如图 2 所示，调整的方法是延长平行四边形短对角线的长度，使其与长对角线保持一致，得到有向的矩形候选框。

传统的两阶段目标检测框架在算法的运行过程中通常使用 RoI Pooling 方法，根据预选框的位置坐标，

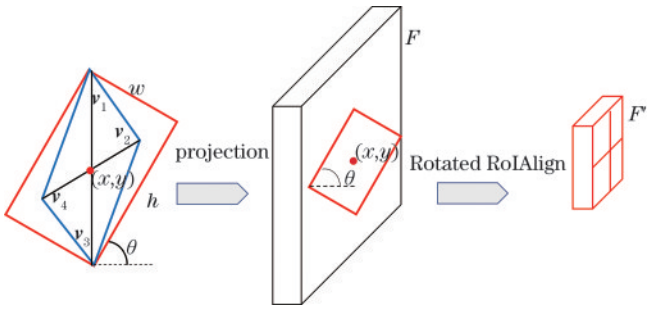

图 2　Rotated RoIAlign 原理

Fig. 2　Schematic of Rotated RoIAlign

在特征图中将相应区域池化为固定尺寸的特征图，方便后续的检测框回归和分类操作。RoI Pooling 这一操作存在两次量化过程，主要包括将选框边界量化为整数点坐标值和将量化后的边界区域平均分割成 $k \times k$ 个单元，对每一个单元的边界进行量化。上述两次量化过程会对候选框位置预测的精度有较大的影响，形成较大偏差。

为了解决上述缺点，在 RoIAlign 的基础上进行改进，得到 Rotated RoIAlign。RoIAlign 使用双线性插值方法获得坐标为浮点数的像素点上的图像数值，从而将整个特征聚集过程转化为一个连续的操作。Rotated RoIAlign 则加入了角度信息，能够有效地对目标方向进行特征聚集，减小目标候选框预测过程中的计算误差。

3　所提方法内容

3.1　网络结构

在 Oriented R-CNN 研究基础上，提出了一种多层级局部自注意力增强和候选框自适应调整的遥感图像目标检测算法。网络结构如图 3 所示，主要由 Swin Transformer 骨干（BackBone）网络、特征金字塔网络（FPN）[19]、Oriented RPN、Rotated RoIAlign 和损失函数模块（KLD）组成。Oriented R-CNN 使用多尺度结构来检测目标，通过改进有向生成候选区域网络，在遥感图像目标检测方向取得了一定的成效。然而，由于遥感图像背景十分复杂并且方向任意，存在浅层特征提取能力较差且目标检测精度不高的问题。因此本文运用以 Transformer 为代表的基于自注意力机制的 Swin Transformer 网络。

所提目标检测框架主要分为 4 个阶段。第 1 阶段，通过引进骨干网络 Swin Transformer，改善了在浅层特征提取能力不足的问题，提高模型的层级信息感知能力。骨干网络的主要作用是生成输入图片的特征图，获得更加准确的特征信息；FPN 使用特征金字塔结构，在增加极少计算量的情况下，对多尺度问题有很好的改善效果。第 2 阶段，利用 Oriented RPN 模块，根据设定好的交并比（IoU）生成有向候选框。第 3 阶段，结合 Rotated RoIAlign 模块，基于候选框进行分类与

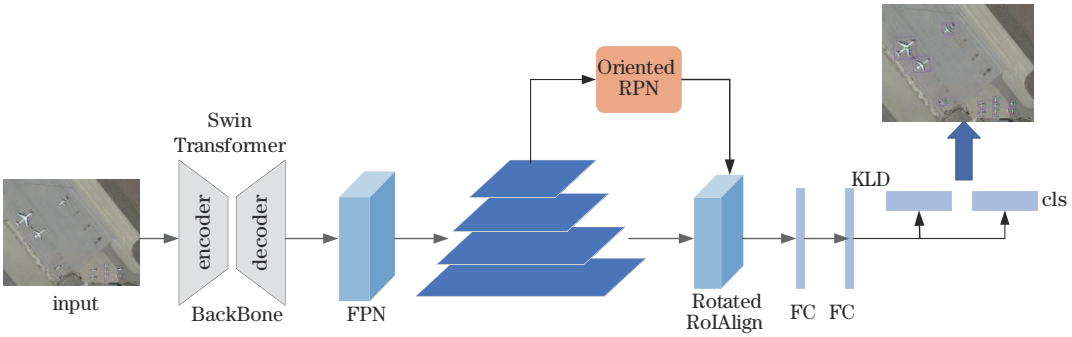

图 3　所提算法的整体框架

Fig. 3　Overall framework of the proposed algorithm

位置回归。最后阶段，利用损失模块进行定位损失函数计算，通过引入高斯损失距离函数（KLD），自适应地提高对候选框变化的辨识能力，最终提高目标检测精度。

3.2　基于多层级局部自注意力的特征提取网络

遥感图像目标往往与局部背景具有密切关系，同类目标可能会出现在不同的遥感图像场景中，增加了对多类目标检测的难度。卷积神经网络（CNN）虽能在一定程度上增加感受野范围，在提取中间层特征和视觉结构上具有一定的优势，但是在提取底层特征的依赖关系和层次关系时仍具有局限性。因此，引入 Swin Transformer 特征提取模块，旨在快速进行层次构建，通过提升层次信息交互能力，提升对目标不同层级局部特征的表示能力，进而提升对多类遥感图像目标检测的精度。Swin Transformer 结构由 Swin Transformer Block 串联组成，同时 Swin Transformer Block 由 S-MSA[20] 和 SW-MSA 组成。

1）Swin Transformer Block

传统的 Transformer 都是基于全局信息来计算注意力的，复杂度较高，而 Swin Transformer 则在每个窗口内计算注意力，减少了计算量。Swin Transformer 由多个 Swin Transformer Block 组成，图 4 展示了基于 Transformer 编码器构建的 Swin Transformer Block，主要包括移位窗口多头自注意力层（SW-MSA）、窗口多头自注意力层（W-MSA）和包含 GELU 非线性函数的多层感知机（MLP）。各个模块之前使用 Layer-Norm（LN）进行归一化处理，同时在每个 Block 之后使用残差连接进行计算。Swin Transformer 计算方法为

$$\begin{cases} \hat{z}^{l}=\text{W-MSA}\left[\text{LN}\left(z^{l-1}\right)\right]+z^{l-1} \\ z^{l}=\text{MLP}\left[\text{LN}\left(\hat{z}^{l}\right)\right]+\hat{z}^{l} \\ \hat{z}^{l+1}=\text{SW-MSA}\left[\text{LN}\left(z^{l}\right)\right]+z^{l} \\ z^{l+1}=\text{MLP}\left[\text{LN}\left(\hat{z}^{l+1}\right)\right]+\hat{z}^{l+1} \end{cases}, \tag{2}$$

主要流程为：输入到 Swin Transformer Block 的特征首先经过 z^{l-1} 进行归一化，再经过 SW-MSA 进行特征学习；其次，进行残差操作得到 $\hat{z}^{l}$；最后经过 1 个归一化层、1 个 MLP 和 1 个残差连接，得到这一层的输出特征。

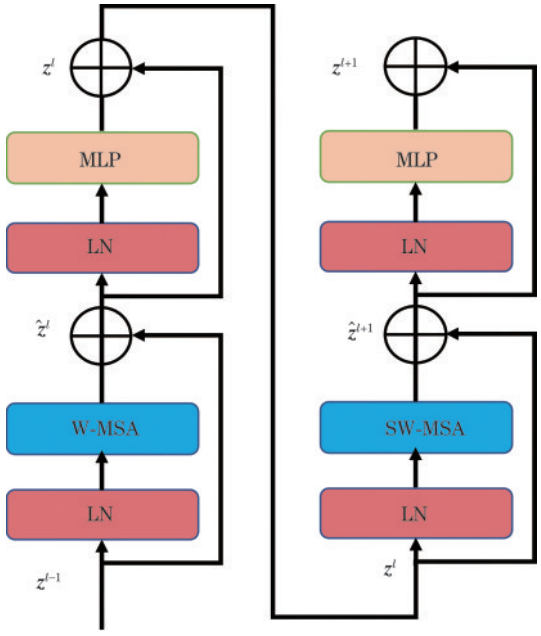

图 4　Swin Transformer Block 的结构

Fig. 4　Swin Transformer Block structure

2）W-MSA

Swin Transformer 中 Block 模块 Layer-Norm 和残差连接与 Transformer 中的完全相同，区别就是两个 MSA（multi-head self-attention）变换成 W-MSA（window multi-head self-attention）和 SW-MSA。

由于 MSA 模块计算量过大，所以引入 W-MSA 减少计算量，如图 5 所示，左侧使用的是经典的 MSA 模块，在自注意力机制中，特征图的每个像素都需要和图中其他像素进行计算比较。但使用 W-MSA 后，首先会将特征图按 $M\times M$（$M=2$）尺寸划分成单独窗口，然后再对每个窗口进行自注意力计算。

MSA 和 W-MSA 总体计算量为

$$\begin{cases} \Omega_{\text{MSA}}=4hwC^{2}+2(hw)^{2}C \\ \Omega_{\text{W-MSA}}=4hwC^{2}+2M^{2}hwC \end{cases}, \tag{3}$$

式中：h 代表特征图的高度；w 代表特征图的宽度；C 代表特征图的深度；M 代表每个窗口的大小。由式（3）可

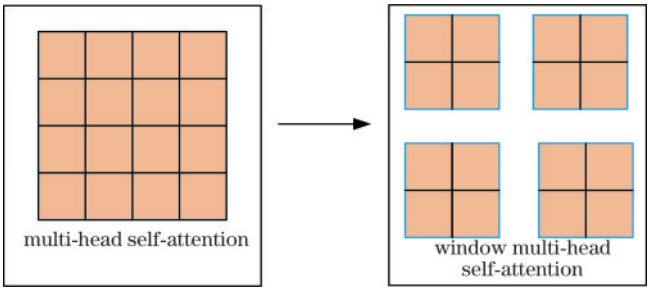

图 5　MSA 和 W-MSA 原理对比

Fig. 5　Comparison between MSA principle and W-MSA principle

得出：第一个等式中，特征图的长宽具有二次复杂度；第二个等式中，在每个蓝色框内计算自注意力，将计算区域控制在以窗口为单位的区域，能够极大地减轻网络的计算量。且当 M 固定时，计算具有线性复杂度，表明 W-MSA 可以减少计算量，具有良好的扩展性。

3) SW-MSA

W-MSA 在算法运行过程中会计算每个窗口内的自注意力，所以无法进行窗口间的信息连接。为了解决这个问题，引入了 SW-MSA，即滑动的 W-MSA。如图 6 所示，左侧是 W-MSA，右侧是 SW-MSA，对两张图进行对比可以发现窗口的偏移，窗口从左上角分别向下方还有右侧偏移了 $M/2$ 个像素。从图 6 展示的 SW-MSA 运行机理可以清晰看出，自注意力窗口偏移后可以让原本没有信息传递的两个窗口进行信息交流，解决了不同窗口之间无法进行信息传递的问题。

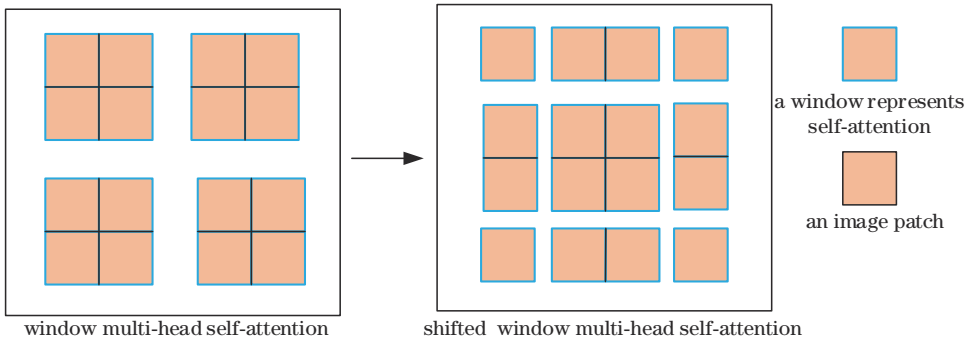

图 6　W-MSA 和 SW-MSA 原理对比

Fig. 6　Comparison betwwen W-MSA principle and SW-MSA principle

3.3　基于自适应点偏移的高斯损失函数

现有的旋转检测器大多都是由水平检测器发展而来的，能够满足多数图像检测的要求。然而，遥感图像具有目标尺度变化大、目标方向任意及背景复杂的特点，导致现有的水平回归损失对遥感图像目标检测有一定的局限性，使得在高精度检测中这些检测器不够突出。为了解决旋转回归损失中耦合参数设定问题，在 Oriented R-CNN 的基础上，引入了一种高斯损失函数（KLD）。KLD 方法原理如下，首先将旋转矩形 (x, y, w, h, θ) 转换为二维高斯分布 $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，转换公式为

$$\begin{cases} \boldsymbol{\mu}=(x,y)^{\mathrm{T}} \\ \boldsymbol{\Sigma}^{\frac{1}{2}}=\boldsymbol{R}\boldsymbol{\Lambda}\boldsymbol{R}^{\mathrm{T}}=\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}\frac{w}{2} & 0\\ 0 & \frac{h}{2}\end{pmatrix}\begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix}=\begin{pmatrix}\frac{w}{2}\cos^2\theta+\frac{h}{2}\sin^2\theta & \frac{w-h}{2}\cos\theta\sin\theta\\ \frac{w-h}{2}\cos\theta\sin\theta & \frac{w}{2}\sin^2\theta+\frac{h}{2}\cos^2\theta\end{pmatrix} \end{cases}, \tag{4}$$

式中：$\boldsymbol{R}$ 代表旋转矩阵；$\boldsymbol{\Lambda}$ 代表对角矩阵。然后，通过高斯分布之间的 KLD 计算回归损失。两个二维高斯函数之间的关系为

$$D_{\mathrm{KL}}\left(\mathcal{N}_p\|\mathcal{N}_t\right)=\frac{1}{2}\left(\boldsymbol{\mu}_p-\boldsymbol{\mu}_t\right)^{\mathrm{T}}\boldsymbol{\Sigma}_t^{-1}\left(\boldsymbol{\mu}_p-\boldsymbol{\mu}_t\right)+\frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Sigma}_p\right)+\frac{1}{2}\ln\frac{\left|\boldsymbol{\Sigma}_t\right|}{\left|\boldsymbol{\Sigma}_p\right|}-1, \tag{5}$$

或者

$$D_{\mathrm{KL}}\left(\mathcal{N}_t\|\mathcal{N}_p\right)=\frac{1}{2}\left(\boldsymbol{\mu}_p-\boldsymbol{\mu}_t\right)^{\mathrm{T}}\boldsymbol{\Sigma}_p^{-1}\left(\boldsymbol{\mu}_p-\boldsymbol{\mu}_t\right)+\frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_t\right)+\frac{1}{2}\ln\frac{\left|\boldsymbol{\Sigma}_p\right|}{\left|\boldsymbol{\Sigma}_t\right|}-1。 \tag{6}$$

式(5)主要由预测框的长度、宽度和角度线性组成，通过链式的方式实现参数耦合，第一部分公式为

$$\left(\boldsymbol{\mu}_p-\boldsymbol{\mu}_t\right)^{\mathrm{T}}\boldsymbol{\Sigma}_t^{-1}\left(\boldsymbol{\mu}_p-\boldsymbol{\mu}_t\right)=\frac{4\left(\Delta x\cos\theta_t+\Delta y\sin\theta_t\right)^2}{w_t^2}+\frac{4\left(\Delta y\cos\theta_t-\Delta x\sin\theta_t\right)^2}{h_t^2}, \tag{7}$$

式中：$\Delta x=x_p-x_t$，$\Delta y=y_p-y_t$，$\Delta\theta=\theta_p-\theta_t$。可以通过梯度来分析，令 $\theta_t=0°$，发现当目标很小或者对应的边长很短时，模型会增大相应方向的梯度，进行点位偏移的自适应。

遥感图像由于具有图像尺度比例大、背景复杂等

特点，可能因一个细微的角度错误，精度严重降低，因此可以通过引入KLD损失函数来解决这个问题。KLD能够根据对象的特征动态调整参数权重，这种机制非常适合高精度检测，同时KLD又是尺度不变性的，更加符合遥感图像的特征。

4 实验与结果分析

4.1 实验环境与参数配置

本文实验在Linux操作系统下进行，使用Oriented R-CNN模型作为基本配置，所有实验均在1块NVIDIA GTX2080Ti（11 GB显存）上进行，PyTorch版本为1.70。

实验时输入图像尺寸设为1024×1024，设置剪切步长为824，初始学习率设为0.0001，动量设为0.9，衰减系数设为0.05，batch size设为4，使用随机梯度下降（SGD）优化器。

4.2 数据集描述

1）DOTA数据集

DOTA数据集[20]由2806张航空图像组成，图像来源于多种传感器和平台的不同分辨率遥感图像。总共包含188282个包围框标注的实例目标，在实验中，主要采用矩形框的标注形式。DOTA数据集的类别主要包括飞机（PL）、棒球场（BD）、桥梁（BR）、田径场（GTF）、小型车辆（SV）、大型车辆（LV）、轮船（SH）、网球球场（TC）、篮球场（BC）、储油罐（ST）、足球场（SBF）、环形交叉路口（RA）、港口（HA）、游泳池（SP）和直升机（HC）这15种类别。部分样例如图7所示。

DOTA数据集中训练集、验证集和测试集的图像数量占总图像数的1/2、1/6和1/3，分别有1411、458和937幅图像。将训练集与验证集合并为统一训练集，每张图片的大小都控制在800×800到4000×4000内。同时由于该数据集中的图像类别多样、方向分布不均匀、目标尺度变化大，因此是最具有挑战性且最具有代表性的遥感数据集之一。

2）HRSC2016数据集

HRSC2016数据集[21]是一个高分辨率遥感船舰图像数据集，全称为High Resolution Ship Collection 2016，由西北工业大学于2016年发布，采用矩形框标注格式。该数据集中所有的图片都是从Google Earth平台中获得的。这些图像的分辨率为2～0.4 m，大小范围为300×300～1500×900像素，其中大部分都超过了1000×600像素。HRSC2016数据集中的部分图像如图8所示。

图7 DOTA数据集部分样例

Fig. 7 Some samples in DOTA dataset

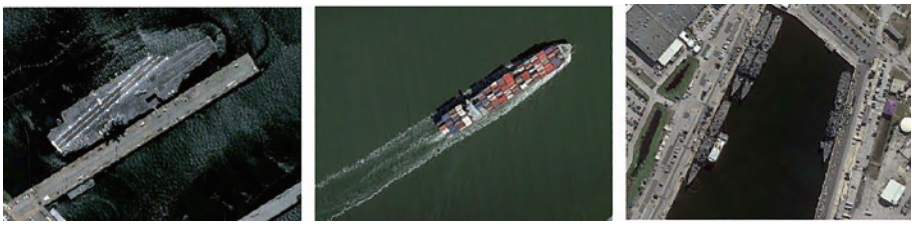

图8 HRSC2016数据集部分样例

Fig. 8 Some samples in HRSC2016 dataset

HRSC2016数据集共有1061幅图像，其中包括70张海面图像和991张港口图像，这两类图像分别标注了90和2886个待测样本，总共包括2976个目标。HRSC2016中训练集、验证集和测试集分别有436幅、181幅和444幅图像，三个集合分别标注了1207、541和1228个目标船舰。

4.3 评价指标

采用平均精度均值（mAP）、帧率（s）、模型参数量（Parameters）和浮点运算量（FLOPs）作为遥感图像目标检测中常用的网络模型效果的评价指标。mAP表示不同种类识别精度的平均值，是对目标检测的精确率和召回率的综合评价，具有很好的衡量性。FPS展示了网络模型检测的速度。同时交并比（IoU）也是模型检测精度的重要基础指标，是计算检测结果边框与目标边框之间的交集区域与并集区域的比例，实际计算中设置的阈值为0.5，只要比例超过0.5，认为检测结果是正确的。假设输出的检测结果为A，真实结果为B，IoU计算公式为

$$R_{\mathrm{IoU}}=\frac{A\cap B}{A\cup B}。\tag{8}$$

精确率(P)表示预测为正类样本中真实正类样本的比例,计算公式为

$$P=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}},\tag{9}$$

式中:N_{TP}为检测为正确分类的正类样本数;N_{FP}为错误分类为正类样本的负类样本数。召回率(R)表示正确检测到的正类样本目标在所有正样本中的比例,计算公式为

$$R=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}},\tag{10}$$

式中:N_{FN}为错误分为负类的正类样本数。利用平均精度(AP)作为评价标准,AP是遥感图像目标检测领域比较通用的评价方法,AP包含两个主要的指标,精度和召回率。精度和召回率是一对相对的度量评价指标。一般来说,精度低,召回率往往偏高;而召回率低,精度往往会偏高。AP的计算方式为

$$P_{\mathrm{AP}}=\int_0^1 P(R)\mathrm{d}t。\tag{11}$$

对于多分类问题,通常引入mAP,公式为

$$P_{\mathrm{mAP}}=\frac{\sum_{n=0}^{N}P_{\mathrm{AP}n}}{N}。\tag{12}$$

帧率(s)数值展示了算法的检测速度的快慢,公式为

$$s=\frac{1}{t},\tag{13}$$

式中:t表示对每张图的处理时间。

4.4　实验结果分析

4.4.1　检测结果分析

1) 基于DOTA数据集分析

在相同的实验条件下,对所提算法与目前先进的几种遥感图像目标检测算法进行实验对比,其中包括经典的ICN算法[22]、Faster R-CNN算法、根据四个点的偏移量来表示旋转检测的Gliding Vertex算法[23]、将空间转换应用在RoI上的RoI Transformer算法[24]、当前先进的二阶段检测算法Oriented R-CNN、一阶段检测算法Rotated RetinaNet和S^2A-Net[25]。在DOTA数据集上,比较不同算法对15个类别的检测结果和mAP。检测速度均在相同的环境中测试,实验结果如表1所示,其中加粗字体表示最优值。

表1　不同算法在DOTA数据集上的AP

Table 1　AP of different algorithms on DOTA dataset　　unit: %

Category	ICN	Faster R-CNN-O	Gliding Vertex	RoI Transformer	Oriented R-CNN	Proposed algorithm
PL	81.36	89.12	**89.64**	88.65	89.19	89.24
BD	74.30	83.06	**85.00**	82.60	82.53	82.88
BR	47.70	50.26	52.26	52.53	51.86	**52.95**
GTF	70.32	67.49	**77.34**	70.87	72.21	75.50
SV	64.89	78.64	73.01	77.93	**78.86**	78.85
LV	67.82	73.44	73.14	76.67	81.87	**84.26**
SH	69.98	85.97	86.82	86.87	87.91	**88.24**
TC	90.76	90.89	90.74	90.71	90.90	**90.91**
BC	79.06	84.58	79.02	83.83	86.70	**86.91**
ST	78.20	82.92	**86.81**	82.51	85.13	86.08
SBF	53.64	54.34	59.55	53.95	63.85	**64.57**
RA	62.90	66.09	70.91	67.61	65.85	67.89
HA	67.02	66.22	72.94	74.67	73.24	**75.33**
SP	64.17	68.99	70.86	68.75	68.76	**70.97**
HC	50.23	58.52	59.32	61.03	56.07	**63.43**
mAP	68.16	73.37	75.02	74.61	75.66	**77.20**

Note: -O indicates that the algorithm is used for rotating target detection.

表1记录了所提算法与其他先进遥感图像目标检测算法的精度比较。可以得出:由于引入了KLD损失函数,所提算法可以根据目标的长宽比来动态调整权重,所提算法在LV(大型汽车)、SH(船只)、HA(港口)等大长宽比目标的类别中检测精度最高;得益于Swin Transformer的分层次特征提取,所提改进Oriented R-CNN算法在TC(网球场)和BC(篮球场)等具有等比例大小并且朝向大致相同的类别中也有最佳精度;由于飞机的特征明显,目标较大,检测出来较为容易,因此可以发现所提算法与Oriented R-CNN算法在PL(飞机)检测中精度相当。所提算法对ST(储油罐)的检测效果不佳,这可能是因为油罐形状单一,都为圆形油罐,存在背景中有圆形物体干扰,同时Swin Transformer特征提取网络是基于移动窗口进行分层特征提取的,允许跨

窗口连接，容易受相邻目标和背景轮廓影响，对相似背景给出较低相似度的错判，该问题可以通过修改IoU阈值来进行优化。在DOTA数据集上，改进Oriented R-CNN算法的mAP值为77.20%，较经典的ICN算法、RoI Transformer算法和Oriented R-CNN算法分别提高了9.04个百分点、2.59个百分点和1.54个百分点。所提改进算法对15类目标的mAP具有明显优势，结果表明，所提方法具有稳定的特征提取能力以及较强的鲁棒性，能稳定地从浅层目标中提取特征。

为了更加直接地体现不同算法检测结果的区别，在测试集上对ICN算法、Oriented R-CNN算法和所提算法进行定性对比，检测结果如图9所示。

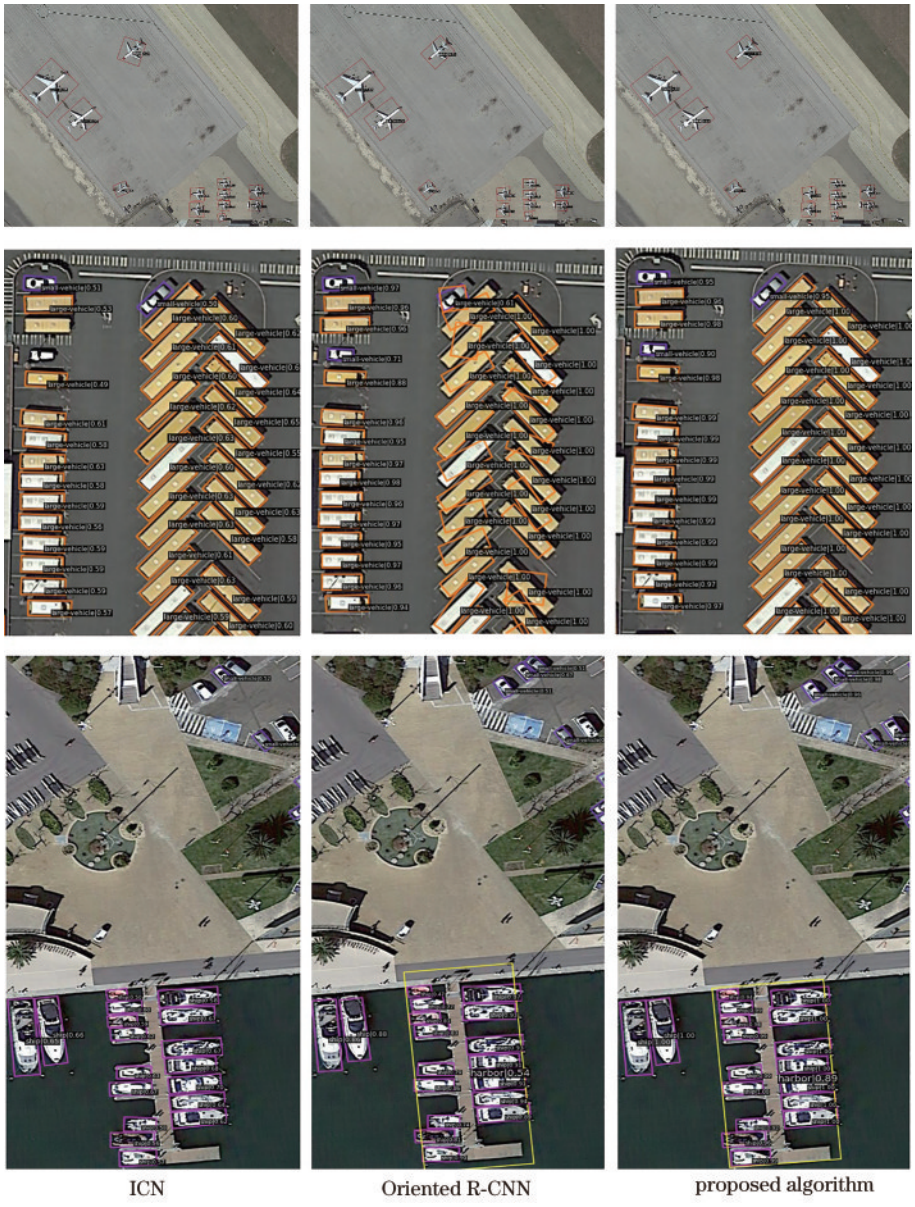

图9　不同算法在DOTA数据集上的检测结果

Fig. 9　Detection results of different algorithms on DOTA dataset

在大部分情况下，所提算法针对具有不同属性的多类遥感目标都取得了比较满意的检测结果。结果表明，通过引入Swin Transformer模块对全局特征进行有效提取，优化损失函数KLD提高对目标尺度比的敏感度，所提算法提高了遥感图像目标检测精度。从图9第1行可以看出，所提算法和其他先进算法都能够准确地检出飞机，这是因为飞机特征明显，并且朝向也大径相同。第2行图像中，对于排列密集的车辆，ICN算法出现了漏检的情况，Oriented R-CNN算法一定程度上存在定位不精确的问题，相比之下，所提算法对目标的定位效果优于另外两个算法，同时未出现误检等情况。从图9的最后一行图片可以看出，所提算法能够在复杂多类别背景下检测出密集排列的目标，检测出了右上角并不明显的小型汽车和港口目标，而ICN算法出现了漏检的情况，同时基础算法Oriented R-CNN对最右侧白色汽车也出现了漏检情况。由于Swin

Transformer模块能够快速地获取浅层特征，所提算法对不明显的小目标也能进行正确识别，以上说明所提算法具有优秀的鲁棒性。

2）基于HRSC2016数据集分析

为了验证所提算法在遥感图像上的有效性，继续在HRSC2016数据集上进行对比实验，得到5种算法在数据集上的各性能指标，结果如表2所示，其中AP_{50}和AP_{75}分别表示IoU阈值为0.5和0.75的AP值。

从表2可以看出：当IoU阈值为0.5时，所提算法的AP值为90.6%，比经典的RoI Transformer和Oriented R-CNN分别提高了4.5个百分点和0.5个百分点，所提算法使用的骨干网络Swin Transformer能够提取更深层的语义信息，同时利用KLD损失函数进行高斯优化，进一步提升了检测精度。IoU阈值越大，算法的准确度越高。从表2可以看出，随着IoU阈值的增大，AP值有所降低，但是所提算法的AP值还是优于其他算法的，表明所提算法的整体定位精度更高。

表2　不同算法在HRSC2016数据集上的检测结果

Table 2　Detection results of different algorithms on HRSC2016 dataset

Method	BackBone	AP_{50} /%	AP_{75} /%
RetinaNet-O	ResNet-50	84.8	59.9
RoI Transformer	ResNet-101	86.1	65.3
S^2A-Net	ResNet-50	89.7	74.6
Oriented R-CNN	ResNet-50	90.1	76.9
Proposed method	Swin Transformer	90.6	79.8

Note: -O indicates that the algorithm is used for rotating target detection.

为了直观地展示实验算法的检测结果，选取另外两种算法与所提算法进行可视化对比，检测结果如图10所示，其中红色旋转矩形框为正确检测结果。可以发现其他算法都存在漏检、定位不精确等情况，所提算法无漏检等情况，说明所提算法具有良好的适应性，更加有利于遥感图像的目标检测。

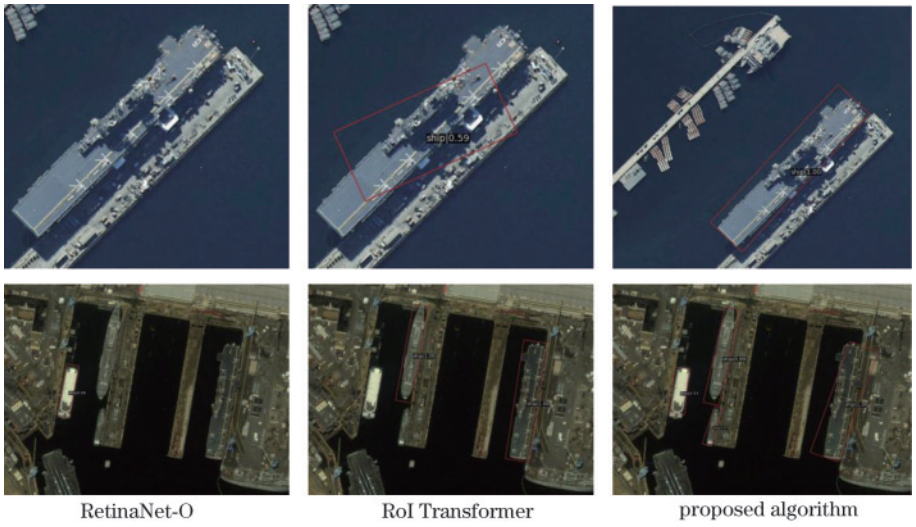

图10　不同算法在HRSC2016数据集上的检测结果

Fig. 10　Detection results of different algorithms on HRSC2016 dataset

4.4.2　模型速度与精度分析

在相同的实验条件下，对先进的单阶段检测算法和双阶段检测算法的检测精度和检测速度进行实验，结果如表3和图11所示。从表3可以看出：所提算法具有最优的mAP，为77.20%，Oriented R-CNN次之，RetinaNet-O表现最差；从速度来看，一阶段经典算法RetinaNet-O速度最快，所提算法几乎与S^2A-Net算法持平，并且远远超过同为二阶段算法的RoI Transformer算法。从图11可以发现，所提算法在取得较高检测精度的情况下，检测速度也是十分优秀的。这是因为所提算法引入了Swin Transformer特征提取网络。Swin Transformer使用了一种拥有移动窗口的自注意力模型，通过串联窗口多头自注意力层（W-MSA）和移位窗口多头自注意力层（SW-MSA），具有全局注意力能力的同时，将关于图像大小的平方关系降为线性关系，减少了计算量。所提算法大幅减少了运算量，提高了检测速度。

表3　不同算法在DOTA数据集上的速度和精度

Table 3　Speed and accuracy of different algorithms on DOTA dataset

Method	Framework	s /(frame·s^{-1})	mAP /%
RetinaNet-O	One-stage	16.8	69.79
S^2A-Net	One-stage	15.5	73.85
RoI Transformer	Two-stage	14.3	74.61
Gliding Vertex	Two-stage	15.3	75.02
Oriented R-CNN	Two-stage	15.0	75.86
Proposed method	Two-stage	15.2	77.20

Note: -O indicates that the algorithm is used for rotating target detection.

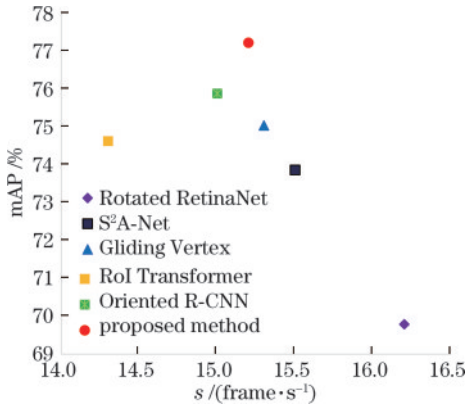

图 11　不同算法在 DOTA 数据集上的检测速度与准确率

Fig. 11　Detection speed and accuracy of different algorithms on DOTA dataset

4.4.3　模型性能分析

基于 DOTA 数据集，在相同的实验条件下比较了先进的单阶段检测算法和双阶段检测算法的参数量和浮点运算量两种性能指标。由表 4 可知，相比双阶段算法，单阶段算法参数规模和浮点运算量较低，原因是单阶段算法只有特征提取过程。与基础算法 Oriented R-CNN 相比，所提算法在增加了少量浮点运算量的情况下，参数量降低了 0.04×10^6，精度提升了 1.8%。因此，相比其他算法，所提方法在参数量和运算量几乎不变的情况下，提升了目标检测精度。

表 4　主要算法在 DOTA 数据集上参数量和浮点运算量对比

Table 4　Comparison of Parameters and FLOPs of main algorithms on DOTA dataset

Model	Framework	mAP / %	Parameters / 10^6	FLOPs / 10^9
RetinaNet-O	One-stage	69.79	43.56	213.87
S^2A-Net	One-stage	73.85	44.66	215.36
RoI Transformer	Two-stage	74.61	59.76	231.69
Gliding Vertex	Two-stage	75.02	64.98	240.24
Oriented R-CNN	Two-stage	75.86	42.14	213.43
Proposed method	Two-stage	77.20	42.10	213.58

Note: -O indicates that the algorithm is used for rotating target detection.

4.4.4　消融实验

为了进一步验证所提改进 Oriented R-CNN 算法的有效性，在数据量更多的 DOTA 数据集上进行消融实验，将 Swin Transformer 网络和 KLD 损失分别嵌入到原始模型中，验证不同模型结构下的算法的检测性能，结果如表 5 所示。可以看出在加入各个模块后，均能获得比原来模型更高的检测性能。相比原始模型，单独嵌入 Swin Transformer 模块和 KLD 损失函数后，mAP 分别提升了 1.15% 和 0.46%，在同时加入两个模块的情况下，mAP 有 1.34 个百分点的提升，效果最好。由于损失函数方面引入了高斯分布的 KLD 损失函数，在取得了与 Baseline 参数量和浮点数运算相近性能的情况下，mAP 增加了 0.35 个百分点，运算复杂度没有升高的情况下，提高了检测精度。通过引入全局语义增强的自注意力机制 Swin Transformer，模型能够有效进行特征提取，精确定位目标轮廓，有效提高精度。从表 5 可以看出，加入了 Swin Transformer 模块和 KLD 模块后均有相应的性能提升，其中，同时加入两个模块后，性能提升最大，表明了所提方法的有效性。

表 5　消融实验结果

Table 5　Results of ablation experiment

Baseline	Swin Transformer	KLD	mAP / %	Parameters / 10^6	FLOPs / 10^9
√			75.86	42.14	213.43
√	√		76.73	44.75	215.66
√		√	76.21	41.15	213.44
√	√	√	77.20	42.10	213.58

5　结　　论

遥感图像目标检测的研究具有十分重要的意义，在 Oriented R-CNN 的基础上，针对遥感图像目标检测算法浅层特征提取能力不足和目标尺寸动态调整能力缺乏的问题，提出了一种基于全局语义增强与局部自适应的遥感图像目标检测算法。首先，针对遥感图像背景复杂的问题，嵌入多层级局部自注意力增强模块 Swin Transformer，加强对浅层的特征提取能力；其次，针对遥感图像目标方向任意、目标尺寸变化大的问题，对损失函数进行了优化，引入高斯损失函数 KLD，使得参数的梯度可以根据目标的特征得到动态调整；最后，在 DOTA 数据集和 HRSC2016 数据集上进行实验，结果表明，所提改进算法有效提高了遥感图像中对目标的检测精度，能够较好地完成遥感图像目标检测任务。虽然所提改进算法的检测性能有一定的提高，能够识别出一些复杂背景下的目标，但从实验结果发现，所提改进算法针对目标稀疏、方向多变的遥感图像时检测精度仍有待提高。在后续工作中，将对网络进行针对性优化，进一步提升检测性能，构造拥有更强鲁棒性的遥感图像目标检测算法。

参　考　文　献

[1] Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 10778-10787.

[2] Joo H B, Jeon J W. Feature-point extraction based on an improved SIFT algorithm[C]//2017 17th International Conference on Control, Automation and Systems (ICCAS), October 18-21, 2017, Jeju, Republic of Korea. New York: IEEE Press, 2017: 345-350.

[3] Mukhtar A, Tang T B. Vision based motorcycle detection using HOG features[C]//2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), October 19-21, 2015, Kuala Lumpur, Malaysia. New York: IEEE Press, 2016: 452-456.

[4] 张磊, 张永生, 于英, 等. 遥感图像倾斜边界框目标检测研究进展与展望[J]. 遥感学报, 2022, 26(9): 1723-1743.
Zhang L, Zhang Y S, Yu Y, et al. Survey on object detection in tilting box for remote sensing images[J]. Journal of Remote Sensing, 2022, 26(9): 1723-1743.

[5] 聂光涛, 黄华. 光学遥感图像目标检测算法综述[J]. 自动化学报, 2021, 47(8): 1749-1768.
Nie G T, Huang H. A survey of object detection in optical remote sensing images[J]. Acta Automatica Sinica, 2021, 47(8): 1749-1768.

[6] 刘金香, 班伟, 陈宇, 等. 融合多维度CNN的高光谱遥感图像分类算法[J]. 中国激光, 2021, 48(16): 1610003.
Liu J X, Ban W, Chen Y, et al. Multi-dimensional CNN fused algorithm for hyperspectral remote sensing image classification[J]. Chinese Journal of Lasers, 2021, 48(16): 1610003.

[7] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[8] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2999-3007.

[9] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[10] Li M J, Guo W W, Zhang Z H, et al. Rotated region based fully convolutional network for ship detection[C]//2018 IEEE International Geoscience and Remote Sensing Symposium, July 22-27, 2018, Valencia, Spain. New York: IEEE Press, 2018: 673-676.

[11] Pan X J, Ren Y Q, Sheng K K, et al. Dynamic refinement network for oriented and densely packed object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11204-11213.

[12] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2016: 1440-1448.

[13] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[14] 郑哲, 雷琳, 孙浩, 等. FAGNet: 基于MAFPN和GVR的遥感图像多尺度目标检测算法[J]. 计算机辅助设计与图形学学报, 2021, 33(6): 883-894.
Zheng Z, Lei L, Sun H, et al. FAGNet: multi-scale object detection method in remote sensing images by combining MAFPN and GVR[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(6): 883-894.

[15] Lin Y D, He H J, Yin Z K, et al. Rotation-invariant object detection in remote sensing images based on radial-gradient angle[J]. IEEE Geoscience and Remote Sensing Letters, 2015, 12(4): 746-750.

[16] Xie X X, Cheng G, Wang J B, et al. Oriented R-CNN for object detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 3500-3509.

[17] Liu Z, Lin Y T, Cao Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 9992-10002.

[18] Yang X, Yang X J, Yang J R, et al. Learning high-precision bounding box for rotated object detection via Kullback-Leibler divergence[EB/OL]. (2021-06-03)[2022-10-08]. https://arxiv.org/abs/2106.01883.

[19] 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), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[20] Xia G S, Bai X, Ding J, et al. DOTA: a large-scale dataset for object detection in aerial images[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 3974-3983.

[21] Liu Z K, Wang H Z, Weng L B, et al. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds[J]. IEEE Geoscience and Remote Sensing Letters, 2016, 13(8): 1074-1078.

[22] Azimi S M, Vig E, Bahmanyar R, et al. Towards multi-class object detection in unconstrained remote sensing imagery[M]//Jawahar C V, Li H D, Mori G, et al. Computer vision-ACCV 2018. Lecture notes in computer science. Cham: Springer, 2019, 11363: 150-165.

[23] Xu Y C, Fu M T, Wang Q M, et al. Gliding Vertex on the horizontal bounding box for multi-oriented object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(4): 1452-1459.

[24] Ding J, Xue N, Long Y, et al. Learning RoI transformer for detecting oriented objects in aerial images[EB/OL]. (2018-12-01)[2022-10-08]. https://arxiv. org/abs/1812.00155.

[25] Han J M, Ding J, Li J, et al. Align deep features for oriented object detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5602511.