E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TN911; TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)08-2758-17

论文引用格式:Chen X L, Du Z L, Zhang X G and Wang X. 2025. Distortion-adaptive and position-aware network for salient object detection in 360° omnidirectional image. Journal of Image and Graphics, 30(8):2758-2774(陈晓雷, 杜泽龙, 张学功, 王兴. 2025. 畸变自适应与位置感知的360°全景图像显著目标检测网络. 中国图象图形学报, 30(8):2758-2774)[DOI:10.11834/jig.240592]

畸变自适应与位置感知的360°全景图像显著目标检测网络

陈晓雷*,杜泽龙,张学功,王兴

兰州理工大学电气工程与信息工程学院,兰州　730050

摘　要:**目的**　现有360°全景图像显著目标检测方法一定程度上解决了360°全景图像投影后的几何畸变问题,但是这些方法面对复杂场景或是前景与背景对比度较低的场景时,容易受到背景干扰,导致检测效果不佳。为了同时解决几何畸变和背景干扰,提出一种畸变自适应与位置感知网络(distortion-adaptive and position-aware network,DPNet)。**方法**　提出两个对畸变和位置敏感的自适应检测模块:畸变自适应模块(distortion-adaptive module,DAM)和位置感知模块(position-aware module,PAM)。它们可以帮助模型根据等矩形投影的特点和具体图像决定应该关注图像的哪些区域。在此基础上,进一步提出一个显著信息增强模块(salient information enhancement module,SIEM),该模块用高级特征指导低级特征,过滤其中的非显著信息,防止背景干扰对360°显著目标检测效果的影响。**结果**　实验在2个公开数据集(360-SOD,360-SSOD)上与13种新颖方法进行了客观指标和主观结果的比较,在8个评价指标上的综合性能优于13种对比方法。本文还设置了泛化性实验,采用交叉验证的方式表明了本文模型优秀的泛化性能。**结论**　本文所提出的360°全景图像显著目标检测模型DPNet,同时考虑了360°全景图像投影后的几何畸变问题和复杂场景下的背景干扰问题,能够有效地、完全自适应地检测显著目标。

关键词:360°全景图像;显著目标检测(SOD);畸变自适应;位置感知;抗背景干扰

Distortion-adaptive and position-aware network for salient object detection in 360° omnidirectional image

Chen Xiaolei*, Du Zelong, Zhang Xuegong, Wang Xing

College of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou　730050, China

Abstract: **Objective** Salient object detection (SOD) in the field of computer vision originates from the study of human visual attention mechanisms. Its goal is to emulate the human ability to focus on specific objects or areas in complex scenes that naturally capture the interest of human vision. SOD, as a foundational research area in computer vision, is important for various downstream tasks, such as object tracking, semantic segmentation, person re-identification, camouflaged object detection, and image retrieval. In recent years, the advancements in Virtual reality (VR) and augmented reality (AR) technologies have expanded SOD beyond traditional 2D images to encompass 360° omnidirectional images (or pan-

收稿日期:2024-10-15;**修回日期**:2024-12-11;**预印本日期**:2024-12-18

*通信作者:陈晓雷　chenxl703@lut.edu.cn

基金项目:国家自然科学基金项目(61967012)

Supported by: National Natural Science Foundation of China (61967012)

oramic images). The application of 360° SOD serves as a crucial preprocessing step for enhancing the efficiency of subsequent advanced visual tasks. These tasks include coding, editing, stitching, quality assessment, and transmission of 360° omnidirectional images. In contrast to traditional 2D images, 360° omnidirectional images exhibit the following core differences: 360° omnidirectional images are spherical. Given that no encoder has been specifically developed for spherical images, 360° omnidirectional images need to be projected into 2D images for further processing. Common projection methods mainly encompass equirectangular projection (ERP), cube-map projection, and octahedron projection. Regardless of the projection method used, geometric distortion is inevitable. This geometric distortion severely impacts the performance of SOD, which results in a significant decline in performance when traditional 2D SOD methods are directly applied to 360° omnidirectional images. Therefore, addressing the challenge of geometric distortion generated by 360° omnidirectional image projection is the core problem in the field of SOD in 360° omnidirectional images (360° SOD). In recent years, some 360° SOD methods have been established to solve the problem of geometric distortion caused by projection and have achieved good detection results to a certain extent. However, their approaches are either limited in effectiveness or rely on artificially designed features, which restricts the ability of the model to detect salient objects in 360° omnidirectional images. Meanwhile, most of the models have poor detection results when facing complex scenes or scenes with low contrast between foreground and background, which are easily interfered with by the background. This study introduces a distortion-adaptive and position-aware network (DPNet) for 360° SOD to solve the abovementioned problems. The aim is to further solve the problem of background interference in complex scenes by considering geometric distortion of ERP image, which helps better detect salient objects in 360° omnidirectional images. **Method** DPNet combines vision transformer (ViT) and convolutional neural networks (CNNs) to build the basic framework of the network. It uses ViT and CNNs to design the encoder and constructs a combination decoder based on U-Net architecture to decode the features from the two encoders step by step. This way combines the global coding advantages of ViT and the local coding advantages of CNN. Compared with previous dual parallel structures, the two encoder backbones of our network are parallel, and the ViT backbone plays a guiding role for the CNN backbone. In other words, the CNN backbone can complement the detail information based on the semantic features extracted by the ViT backbone. On the one hand, this study proposes two distortion-adaptive detection modules, namely, distortion-adaptive module (DAM) and position-aware module (PAM), to solve the geometric distortion problem caused by ERP. DAM models geometric distortion in feature maps through channel-by-channel deformable convolution. PAM calculates spatial weights along the latitude and longitude, which directs the network to adaptively focus on salient regions in the image. Specifically, the global features extracted by the ViT backbone are processed by the DAM to model the geometric distortion. Then, two branches are extracted: one branch is sent to the decoder, and the other branch is sent to PAM to provide position prior information. PAM is placed in the shallow layer of CNN backbone and is responsible for fusing the position prior information with the information extracted in the shallow layer of CNN backbone to guide the subsequent feature extraction. In this way, DPNet can decide which regions of the 360° omnidirectional images should be focused on according to the characteristics of ERP and specific input images. On the other hand, a salient information enhancement module (SIEM) is proposed to further solve the problem of background interference in complex scenes. Currently, most SOD methods use structures such as U-Net to simply aggregate feature maps at different scales. This process inevitably treats a large amount of non-salient information contained in the low-level features as useful information, which leads to poor detection results. For addressing this issue, SIEM uses high-level features to guide low-level features, filters non-salient information, and prevents the influence of background interference on the effectiveness of 360° SOD. **Result** We compare our model with 13 state-of-the-art methods on 2 public datasets, namely, 360-SOD and 360-SSOD. Notably, its overall performance on 8 evaluation metrics is better than those of the latest 13 methods. In addition, the generalization experiment is set up, and the excellent generalization performance of the model is confirmed by cross-validation. Then, an ablation experiment is conducted to verify the performance of the proposed module. Finally, a set of complexity comparison experiments proves that the proposed model DPNet achieves a good balance in terms of detection accuracy and model complexity. **Conclusion** The existing 360° SOD methods cannot effectively address the geometric distortion problem after projection and the background interference problem in complex scenes. Thus, we propose a distortion-adaptive and position-aware 360° SOD network (DPNet) based on ViT and CNNs. The proposed DAM and PAM play a pivotal

role in guiding the network to focus on areas requiring attention based on the distinctive characteristics of ERP and specific input images. In addition, the proposed SIEM works to guide low-level features with high-level features, which effectively filters out non-salient information present in low-level features and enhances the salient information. These capabilities can help the model effectively deal with the background interference problem in complex scenes. Through an extensive set of experiments, we demonstrate that our method outperforms 13 state-of-the-art SOD methods, which establishes its superiority in 360° SOD applications.

Key words: 360° omnidirectional image; salient object detection (SOD); distortion-adaptive; position-aware; anti-background interference

0 引 言

计算机视觉领域的显著目标检测(salient object detection, SOD)起源于对人类视觉注意力机制的研究,旨在通过计算机模拟人类的视觉关注能力,即在复杂的场景中找到人眼最感兴趣的目标或区域。作为计算机视觉中基础性的研究工作,SOD在视觉相关的下游任务中发挥着重要作用,如目标跟踪(Hong等,2015)、语义分割(Hoyer等,2019)、行人重识别(Zhao等,2017)、伪装目标检测(Fan等,2020)、图像检索(Liu和Fan,2013)和图像识别(Ye等,2017)等。随着虚拟现实(virtual reality, VR)和增强现实(augmented reality, AR)技术的发展,SOD不再局限于传统的2D图像,而是逐渐拓展到了360°全景图像(360° omnidirectional image)(陈晓雷 等,2023)。360°全景图像的显著目标检测(以下简称360°SOD)可以作为其他高级视觉任务的预处理步骤来提高处理效率,如360°全景图像的编码(Luz等,2017)、编辑(Serrano等,2017)、拼接(Li等,2020b)、质量评估(Zhou等,2021)和导航(Maugey等,2017)等。

360°全景图像相比于传统2D图像有以下最核心的不同点:360°全景图像是一种球形图像,由于目前还没有专门针对球形图像的编码器,因此需要将360°全景图像投影为2D图像再进行处理,常用的投影方式(Chen等,2018)有:等矩形投影(equirectangular projection, ERP)、立方体投影(cube-map projection, CMP)以及八面体投影(octahedron projection, OHP)等。无论何种投影方式都不可避免地会引起图像畸变,增加处理难度。

一些360°SOD方法考虑了投影畸变问题,并获得了较好的检测效果。Li等人(2020a)构造了一个失真自适应模块来处理由ERP引起的失真,该方法将360°全景图像裁剪成几部分,分别处理后再拼接合并,其消除畸变问题的能力比较有限。Huang等人(2020)和Cong等人(2024)提出将ERP图像和CMP图像同时作为网络输入的方法,旨在结合ERP图像和CMP图像各自的优点降低畸变对于360°SOD的影响。然而CMP图像虽然在投影的6个面上几何畸变较小,但是可能会导致严重的对象分割和边界不连续问题,从而限制了模型的整体检测性能。Zhao等人(2023)探索了ERP图像的特点,为360°SOD开发了基于Transformer的模型,并提出两个畸变自适应模块使模型更适应ERP图像中的畸变。此外,他们还设计了一个关系矩阵将先验信息嵌入模型以应对ERP图像赤道部分大物体的鱼眼效应,但是这种先验信息对于两极地区物体的检测效果有一定的削减作用。

上述模型一定程度上解决了360°全景图像投影后的几何畸变问题,对于360°SOD起到了很大的推动作用,但是它们解决几何畸变问题的方法要么效果有限,要么采用了人为设计的特征,限制了模型对360°全景图像显著目标的检测能力。同时,大部分模型在面对复杂场景或是前景与背景对比度较低的场景时检测效果都不佳,容易受到背景的干扰。

为了进一步应对360°全景图像投影后的畸变问题和复杂场景下的背景干扰问题,本文在已有工作基础上,提出一种畸变自适应和位置感知网络(distortion-adaptive and position-aware network, DPNet)。DPNet采用视觉Transformer(vision Transformer, ViT)结合卷积神经网络(convolutional neural network, CNN)的方式构建编码器,并基于U-Net构造了一个组合解码器,逐步对来自两个编码器的特征进行解码,从而能够结合ViT的全局编码优势和CNN的局部编码优势。与以往双流并行结构不同,本文网络的两个编码器主干不仅是并行的结构,ViT主干还对CNN主干起到了指导作用,从而使CNN主干可以在ViT主干提取的语义特征的基础上补全细节信息。

一方面，为了解决ERP引起的几何畸变问题，本文提出畸变自适应模块（distortion-adaptive module，DAM）和位置感知模块（position-aware module，PAM）以减轻其对360°SOD的影响。具体而言，将ViT主干提取的全局特征经过DAM对几何畸变进行建模，然后引出两路分支，一路分支送入解码器，另一路分支送入PAM以提供位置先验信息，PAM放置在CNN的浅层，负责将位置先验信息与浅层提取的信息进行融合后指导后续特征的提取，如此便能够让DPNet根据ERP的特点和具体输入图像决定应该关注360°全景图像的哪些区域。

另一方面，为了进一步解决复杂场景下的背景干扰问题，本文提出显著信息增强模块（salient information enhancement module，SIEM）。目前，大多数SOD方法使用U-Net等结构简单地聚合不同尺度的特征图，不可避免地将低级特征中包含的大量非显著信息也作为有用信息，导致检测效果不佳（Yuan等，2023）。本文提出的SIEM，用高级特征指导低级特征，过滤低级特征中的非显著信息，增强其中的显著信息，从而能够有效应对复杂场景下的背景干扰问题。

图1展示了本文方法和4种对比方法对4幅复杂场景图像的检测效果，其中，DATFormer（distortion-aware transformer）（Zhao等，2023）、LDNet（light-weight distortion-aware network）（Huang等，2023）和FANet（features adaptation network）（Huang等，2020）是专为360°SOD设计的方法，TSCNet（texture-semantic collaboration network）（Li等，2024）是目前较新的2D SOD方法。从图1可以看到，在场景较为复杂的情况下，特别是显著目标与背景对比度极低的情况下（尤其是第1、2行的图像），4种对比方法均表现不佳，而本文方法可以有效地抑制背景干扰，性能明显优于其他方法。

本文主要贡献如下：1）提出一个畸变自适应和位置感知网络DPNet，该网络同时考虑了360°全景图像投影后的几何畸变问题和复杂场景下的背景干扰问题，能够有效地、完全自适应地检测显著目标。实验证明在两个公开的360°SOD数据集上DPNet综合性能优于13种先进的SOD方法。2）为了使模型能够根据ERP的特点和具体输入图像决定应关注360°全景图像的哪些区域，本文提出两个对畸变敏感的自适应检测模块：畸变自适应模块DAM和位置感知模块PAM。DAM通过可变形卷积逐通道对特征图中的几何畸变进行建模；PAM分别沿着经纬度方向计算空间权重，并根据空间权重引导网络自适应地关注图像中的显著区域，继而定位显著目标。3）为了增强低级特征中的显著信息，本文提出显著信息增强模块SIEM，该模块从高级特征的前景和背景中深度挖掘显著信息形成空间注意力图，引导网络过滤低级特征中的非显著信息，从而增强显著信息。实验表明，该模块与PAM相结合可以有效应对复杂场景下的背景干扰问题。

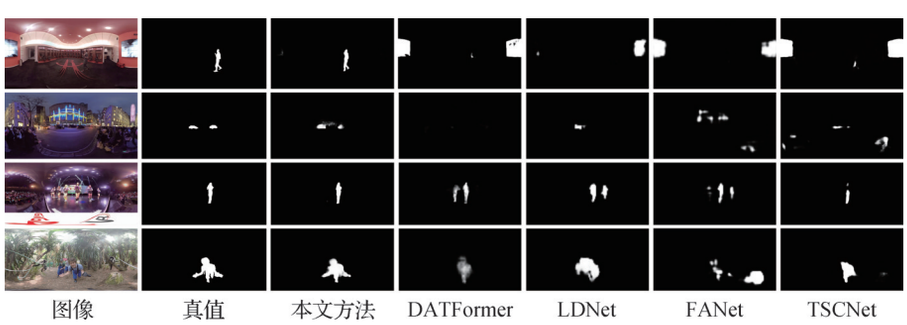

图1　复杂场景下不同模型的主观结果对比

Fig. 1　Comparison of subjective results of different models in complex scenarios

1　相关工作

1.1　2D图像显著目标检测

基于深度学习的2D SOD获得了比传统方法更为优秀的性能，主要分为基于CNN的方法和基于Transformer的方法。基于CNN的方法中，经典的工作有EGNet（edge guidance network）（Zhao等，2019）、GateNet（gated network）（Zhao等，2020）、F^3Net（fusion，feedback and focus network）（Wei等，2020a）和LDF

(label decoupling framework)(Wei等,2020b)。EGNet提出经典的联合边缘信息指导2D SOD的方法。GateNet设计了一种新的门控网络结构,采用多级门单元平衡每个编码器的贡献。F^3Net提出一个交叉特征模块和一个级联反馈解码器来处理不同层次特征之间的差异,并细化多层次特征。LDF提出用于2D SOD的标签解耦框架,通过将显著标签解耦为细节图和主体图以引导网络同时学习更好的边缘特征和避免边缘附近像素的干扰。在最近的工作中,何伟和潘晨(2022)探讨了注意力机制在显著目标检测中的应用潜力,提出新的显著检测模型。SeaNet(lightweight network based on semantic matching and edge alignment)(Li等,2023)提出一种基于语义匹配和边缘对齐的新型轻量级网络,编码器通过动态语义匹配模块和边缘自对齐模块分别提取高级特征和低级特征,然后从最高级别特征开始,根据两个模块输出中包含的准确位置和精细细节推断显著对象。叶欣悦等人(2024)分析了现有网络基本卷积组结构中线性修正单元的选通行为,并进一步提出一种互补信息交互融合模块,将单一模态的"冗余"特征用于辅助另一模态特征。

除了以上基于CNN的方法外,还有一些基于Transformer的工作。VST(visual saliency transformer)(Liu等,2021)从序列到序列的角度重新思考2D SOD,为RGB SOD和RGB-D SOD开发了一种基于纯Transformer的新颖统一模型,并且还提出一种新的Token上采样方法和多任务解码器。Selfreformer(self-refined network with transformer)(Yun和Lin,2022)提出一种新的基于Transformer的网络,可以根据全局和局部上下文进行自我引导。HRTransNet(high-resolution transformer)(Tang等,2023)研究了基于高分辨率Transformer的双模态2D SOD,通过注入一种补充模态有效地结合两种模态线索,该方法还将模型应用到了RGB-D、RGB-T和光场显著目标检测,并取得了优异的性能。另外,Xie等人(2022)、Yao等人(2023)和Yuan等人(2024)探索了结合CNN和Transformer进行SOD的方法,例如Yuan等人(2024)通过使用并行双编码器结构和特征迭代融合模块,有效地结合了CNN和Transformer,获得了优越的性能。

上述2D SOD方法由于未考虑到360°全景图像自身特性的影响,在直接应用到360°SOD领域时效果较差。但是由于其较为深厚的研究积累,2D SOD相关方法对360°SOD仍有一定的借鉴价值。除此之外,上述结合CNN和Transformer的方法虽然使用了双编码器结构,然而两个编码器主干都是单独使用,缺乏编码器之间的交互,一定程度上限制了模型的性能。

1.2 360°全景图像显著目标检测

360°SOD是一个较新的研究方向,相关研究工作还较少,截止目前有以下工作:DDS(distortion-adaptive network with deep supervision)(Li等,2020a)构造了一个失真自适应模块处理由ERP引起的失真,并引入了多尺度上下文集成块感知和区分360°全景图像中丰富的场景和对象。FANet(Huang等,2020)同时将ERP图像和6幅相应的CMP图像作为输入,旨在利用这两种投影方式各自的优势。MPFR-Net(multi-projection fusion and refinement network)(Cong等,2024)提出一种比FANet(Huang等,2020)更好的方法,该方法将立方体展开为一组平面图像以最大化保持物体的连续性,然后将ERP图像和4幅立方体展开图像作为360°SOD网络的输入。此外,CSMA-Net(channel-spatial mutual attention network)(Zhang等,2022)仔细探索了ERP图像和CMP图像之间的互补信息,设计了通道空间相互关注模块,该模块能够有效融合基于360°多投影的瓶颈特征。HPNet(bi-branch hybrid projection network)(Zhang等,2023)设计了一种混合投影特征融合模块,以有效地结合从不同层提取的CMP和ERP特征,最后使用渐进式预测模块细化特征并逐步定位显著目标。SCFANet(semantics and context feature aggregation network)(He等,2024)通过充分探索ERP图像和对应的CMP图像之间的交互性,提出语义和上下文特征聚合网络。Ma等人(2020)将360°SOD分为多阶段任务,并提出基于对象级语义显著性排序的360°全景图像视觉失真处理方法。Wu等人(2023)受人类观察过程的启发,提出一种基于样本自适应视图变换器模块的视图感知360°SOD方法。MFFPANet(multi-level feature fusion and progressive aggregation network)(Chen等,2024)探索了一种360°SOD的多阶段解决方案,该解决方案考虑了RGB图像和互补的对象级语义信息对定位显著对象的影响。LDNet(Huang等,2023)提出一种新颖且高效的轻量级框架,以更小的参数量实现了与其

他基于CNN的大型方法相比具有竞争力的性能。DATFormer(Zhao等,2023)则根据ERP的特点设计了一个纯Transformer网络。MIDP-Net(multi-scale interaction and densely-connected prediction network)(Dai等,2023)提出一种多尺度交互和密集连接预测网络来抵消失真并提取360°全景图像中的多尺度信息,该网络使用多级特征交互模块聚合相邻级别不同感受野的特征和上层解码器的语义线索,并使用双通道解码器模块通过二次解码进一步强化语义信息。

上述360°SOD方法为该领域奠定了重要的基础,它们从各个方面探索了减轻360°全景图像投影畸变对360°SOD影响的方案。就采用ERP图像作为输入的模型而言,其专注于解决ERP畸变的影响而没有关注到复杂场景对显著目标检测的干扰,导致模型在处理一些场景复杂的360°全景图像时表现不佳。所以本文致力于在解决ERP几何畸变影响的同时强化模型处理复杂场景图像的能力。

2 本文模型

2.1 网络总体结构

本文提出的模型结构如图2所示。该模型包含1个ViT编码器、1个CNN编码器以及1个组合解码器。为了平衡预测精度和模型复杂度,ViT编码器采用DilateFormer(multi-scale dilated transformer)(Jiao等,2023)作为主干,DilateFormer是最近提出的一个ViT网络,其改进了原始ViT网络浅层自注意力机制建模的冗余,使用多尺度扩张注意力提取浅层特征,不仅减小了计算代价还提高了准确性。ViT编码器每一层(图2中表示为DFblock)的输出在图2中标识为$\{\boldsymbol{D}_i|i=1,2,3,4\}$,其中$\boldsymbol{D}_4$经过本文提出的DAM后分为两路,一路送入解码器,另一路送入本文提出的位置感知模块PAM用于指导特征(instructional feature, IF)。CNN编码器使用ResNet(residual network)(He等,2016)作为主干,其每一层(图2中表示为Rblock)的输出在图2中标识为

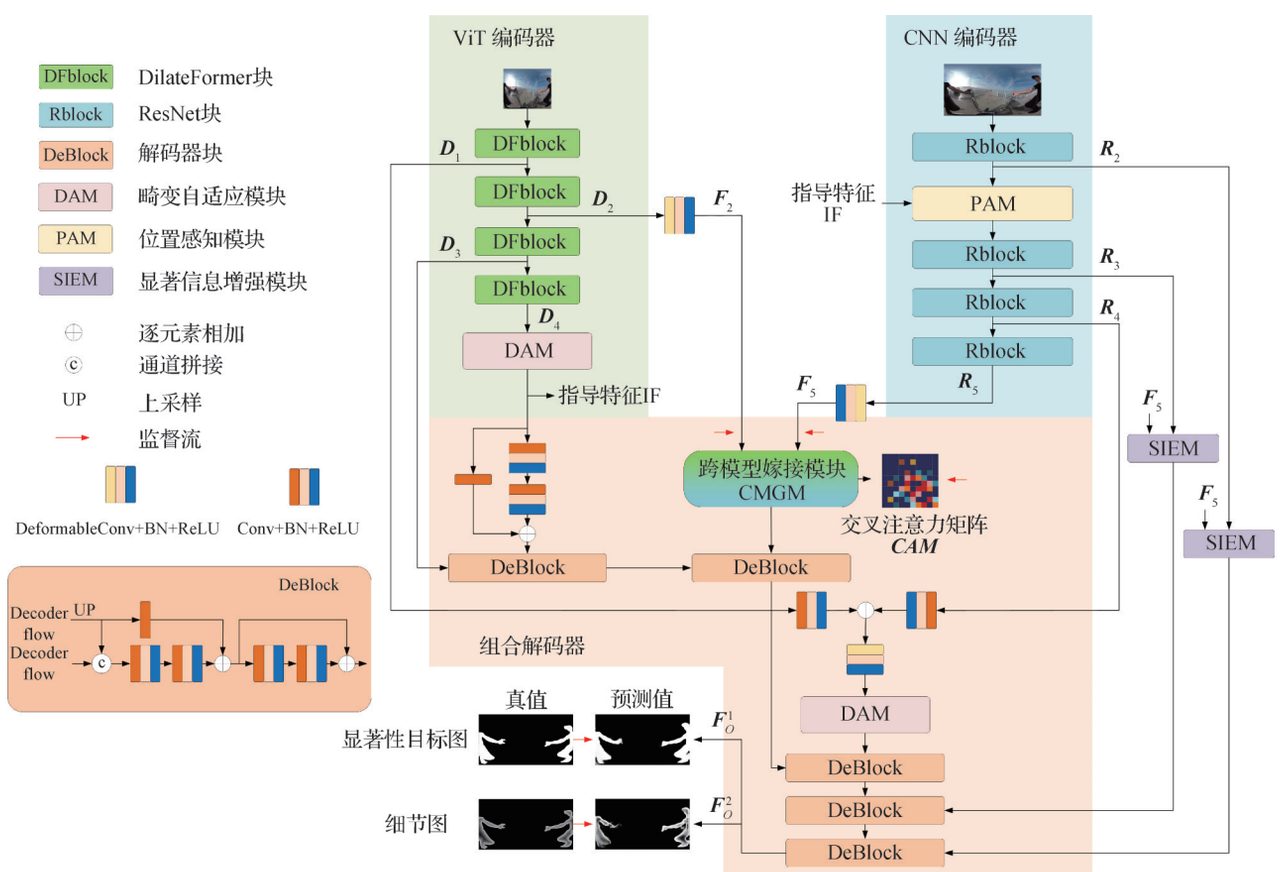

图2 DPNet结构图

Fig. 2 Structural diagram of the proposed DPNet

$\{\boldsymbol{R}_i | i = 2, 3, 4, 5\}$，其中 $\boldsymbol{R}_1$ 由于包含的显著信息较少，所以在图2中未给出，$\boldsymbol{R}_5$ 经过可变形卷积后得到 $\boldsymbol{F}_5$。CNN编码器的浅层加入了PAM，PAM能够在指导特征IF的指导下帮助网络定位显著目标。

在DAM和PAM的帮助下，DPNet能够根据ERP的特点和具体输入图像决定应该关注360°全景图像的哪些区域。对于网络输入，ViT编码器将图像分辨率下采样到固定低分辨率大小，CNN编码器输入原始高分辨率ERP图像，这样ViT编码器可以在低分辨率条件下捕获准确的语义信息，CNN编码器可以在高分辨率条件下捕获更多的细节信息，从而使整个模型能够在不同尺度的特征融合下预测显著目标。

组合解码器利用Xie等人(2022)提出的跨模型嫁接模块(cross-model grafting module，CMGM)来消除ViT编码器和CNN编码器共同的显著预测错误，并逐步解码及融合ViT编码器和CNN编码器各自编码的特征。CMGM模块有两个输出，一个输出是交叉注意力矩阵(cross attention matrix，CAM)，另一个输出是特征图，本文使用 ***CAM*** 作为损失函数监督，使用特征图输出指导低级特征的融合。除此之外，本文还提出一个显著信息增强模块SIEM，使用受监督的 $\boldsymbol{F}_5$ 特征图增强低层特征中的显著信息，减小非显著信息以及噪声对模型预测性能的影响。为了使模型更容易在畸变条件下检测到显著目标，本文在模型的关键位置使用可变形卷积代替普通卷积，并在网络结尾使用DAM进一步加强模型对于畸变的适应能力。最终，网络分别输出了显著性目标图 $\boldsymbol{F}_0^1$ 和细节图 $\boldsymbol{F}_0^2$，并使用相应的真值标签监督其训练。

2.2 畸变自适应模块

本文提出的畸变自适应模块DAM可以帮助网络对ERP造成的几何畸变进行全局性的适应，为特征的后续处理提供方便。该模块结构如图3所示，输入特征图首先经过通道注意力计算每个通道的重要程度，然后将通道权重乘到输入特征图上。考虑到ERP的畸变影响，采用可变形卷积对畸变进行建模。通道加权后的特征图经过由Deformable Conv、Batch Norm、ReLU(rectified linear unit)、sigmod组成的子网络逐通道捕获空间区域的重要程度后获得空间权重矩阵，空间权重矩阵再与通道加权后的特征图进行逐通道的元素间相乘就得到了空间加权特征图，接着将空间加权特征图与原始特征图相加得到与原始特征图维度大小相等的特征图，其维度为 (C, H, W)。受Lin等人(2022b)启发，本文将相加后结果的通道数压缩为1获得注意力图 ***att***，其维度为

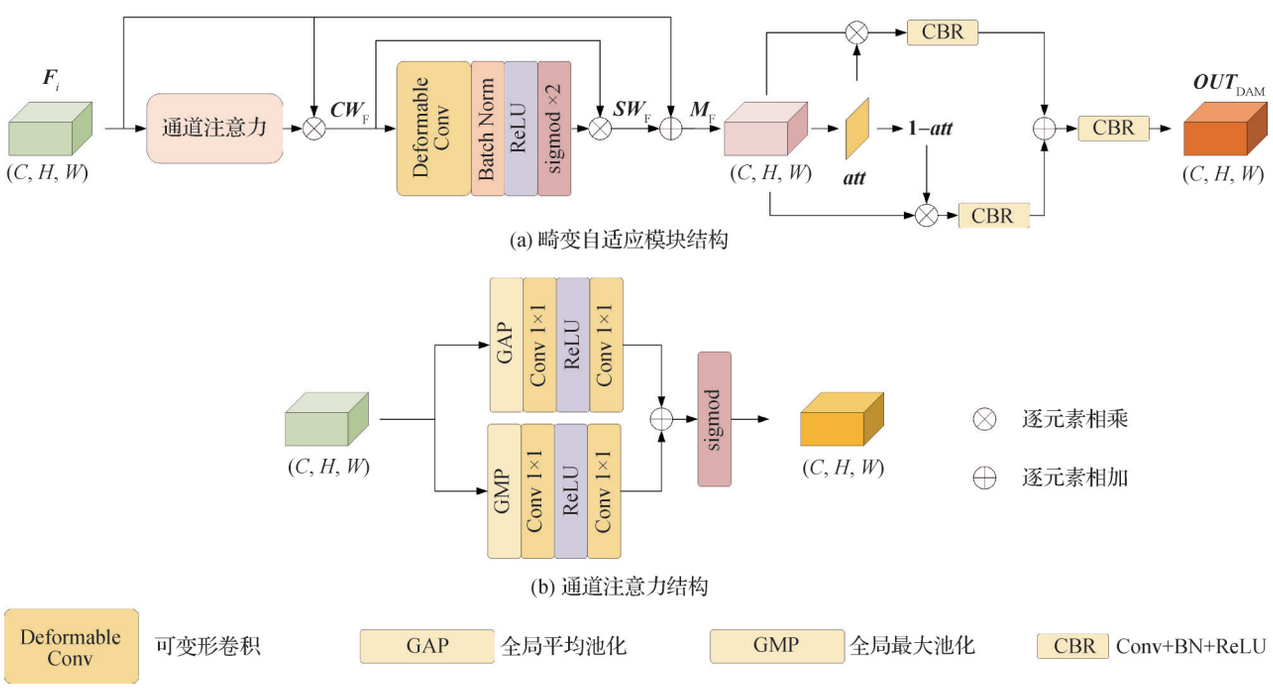

图3 畸变自适应模块结构总框图

Fig. 3 General structure diagram of distortion-adaptive module

((a) structure diagram of the distortion-adaptive module; (b) structure diagram of channel attention)

(1, H, W)，然后反转该注意力图获得反向注意力图 $\mathbf{1}-\boldsymbol{att}$，再分别将它们与对应特征图相乘后再相加得到最终结果，其维度仍然为(C, H, W)。这一处理步骤能够引导DPNet进一步从高置信度显著区域和低置信度背景区域挖掘显著信息。

以上处理过程可以表示为

$$\boldsymbol{CW}_{\mathrm{F}} = f_{\mathrm{ChannelAttention}}(\boldsymbol{F}_i) \otimes \boldsymbol{F}_i \tag{1}$$

$$\boldsymbol{SW}_{\mathrm{F}} = 2 \times sigmoid(ReLU(BN(DConv(\boldsymbol{CW}_{\mathrm{F}})))) \otimes \boldsymbol{CW}_{\mathrm{F}} \tag{2}$$

$$\boldsymbol{M}_{\mathrm{F}} = \boldsymbol{SW}_{\mathrm{F}} + \boldsymbol{F}_i \tag{3}$$

$$\boldsymbol{att} = sigmoid(CBR(\boldsymbol{M}_{\mathrm{F}})) \tag{4}$$

$$\boldsymbol{OUT}_{\mathrm{DAM}} = CBR(CBR(\boldsymbol{att} \otimes \boldsymbol{M}_{\mathrm{F}}) + CBR((\mathbf{1}-\boldsymbol{att}) \otimes \boldsymbol{M}_{\mathrm{F}})) \tag{5}$$

式中，$\boldsymbol{F}_i$ 表示DAM模块的输入特征图，$f_{\mathrm{ChannelAttention}}(\cdot)$ 表示通道注意力，$\boldsymbol{CW}_{\mathrm{F}}$ 表示通道加权的特征图，$DConv(\cdot)$ 表示可变形卷积，BN 表示Batch Normalization，$ReLU(\cdot)$ 表示ReLU激活函数，$sigmoid(\cdot)$ 表示sigmoid激活函数，$\boldsymbol{M}_{\mathrm{F}}$ 表示作为中间结果的特征图，$\boldsymbol{SW}_{\mathrm{F}}$ 表示空间加权的特征图，$CBR(\cdot)$ 表示Conv+Batch Normalization+RELU，$\boldsymbol{att}$ 表示将 $\boldsymbol{M}_{\mathrm{F}}$ 通道压缩到1生成的注意力图，$\boldsymbol{OUT}_{\mathrm{DAM}}$ 表示DAM最终的输出特征图。在式(2)中给 $sigmoid(\cdot)$ 乘以了系数2，这样 $\boldsymbol{SW}_{\mathrm{F}}$ 的像素值就在[0, 2]之间，值大于1表示更重要，而小于1表示不太重要(Zhao等，2023)。

2.3 位置感知模块

本文提出的位置感知模块PAM主要是借助ViT编码主干生成的指导特征IF(带有畸变信息的高级语义特征)帮助CNN编码主干定位显著目标的位置，使得CNN编码主干能够在此基础上侧重于局部细节的提取。PAM的结构如图4所示，该模块有两个输入，分别是指导特征IF和CNN编码器第2层的输出 $\boldsymbol{R}_2$。

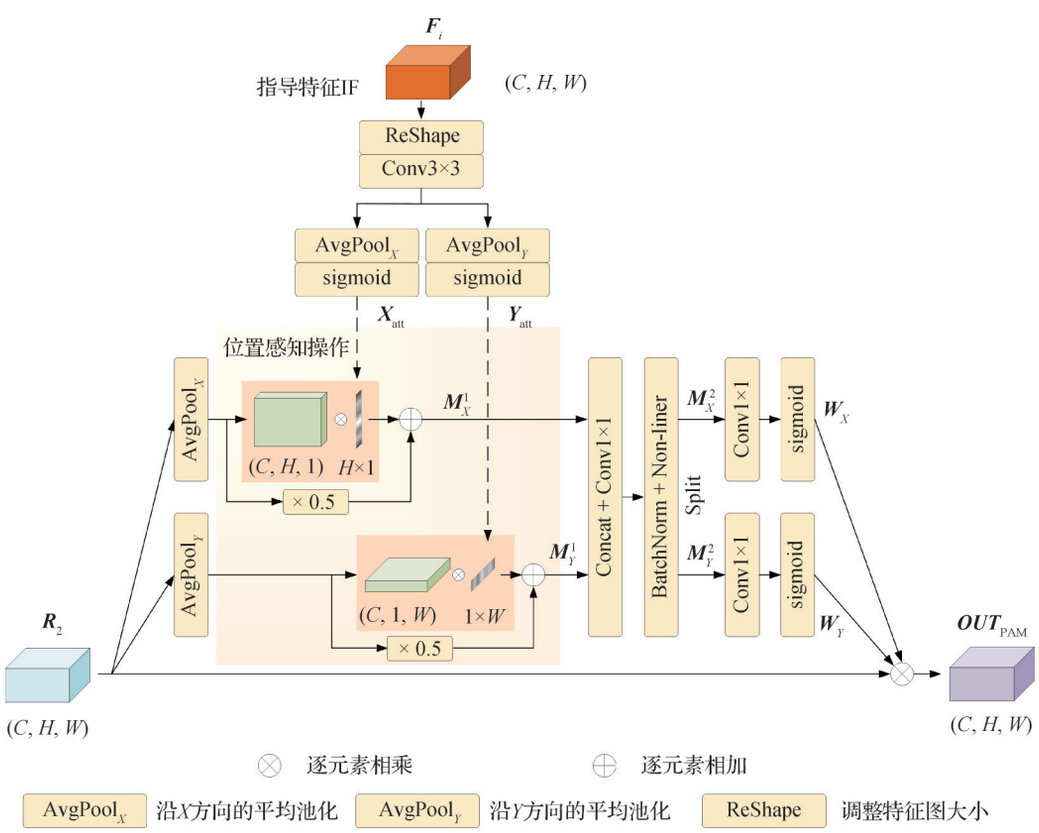

图4 位置感知模块结构图

Fig. 4 Structure diagram of the position-aware module

指导特征IF来自于ViT编码器，它首先调整空间大小以与 $\boldsymbol{R}_2$ 保持一致，并经过一个3×3卷积将通道变换到1，然后分别沿 X 方向和 Y 方向做平均池化再经过sigmoid函数，分别得到大小为 $H \times 1$ 的注意力图和 $1 \times W$ 的注意力图。$\boldsymbol{R}_2$ 同样经过沿 X 方向和 Y 方向的平均池化后分别得到维度为(C, H, 1)和(C, 1, W)的特征图，然后将它们与之前生成的空间大小相同的注意力图相乘，再进行系数为0.5的残

差连接，这一步骤称为位置感知操作（position-aware operation）。接着，将X方向和Y方向分支的特征图转换拼接为维度为$(C, 1, W+H)$的特征图，再经过卷积、批归一化和非线性处理后恢复到各自原来的形状，即$(C, H, 1)$和$(C, 1, W)$。最后将其经过sigmoid函数处理后与$\boldsymbol{R}_2$相乘，相乘结果送入CNN编码器下一层。经过这一系列操作后，由ViT编码器提取的全局语义信息就融合进了CNN编码器中，使其在几何畸变条件下能够感知到显著目标的位置。并且，由于在位置感知操作中加入了残差连接，即使ViT编码器出现了错误的预测，也不会造成CNN编码器的深层网络完全继承这些错误。以上处理过程可以表示为

$$\boldsymbol{X}_{\text{att}} = sigmoid(AvgPool_X(Conv_{3\times3}(ReShape(\boldsymbol{F}_1)))) \tag{6}$$

$$\boldsymbol{Y}_{\text{att}} = sigmoid(AvgPool_Y(Conv_{3\times3}(ReShape(\boldsymbol{F}_1)))) \tag{7}$$

$$\boldsymbol{M}_X^1 = AvgPool_X(\boldsymbol{R}_2) \otimes \boldsymbol{X}_{\text{att}} + AvgPool_X(\boldsymbol{R}_2) \times 0.5 \tag{8}$$

$$\boldsymbol{M}_Y^1 = AvgPool_Y(\boldsymbol{R}_2) \otimes \boldsymbol{Y}_{\text{att}} + AvgPool_Y(\boldsymbol{R}_2) \times 0.5 \tag{9}$$

$$\boldsymbol{M}_X^2, \boldsymbol{M}_Y^2 = Split(NL(BN(Conv_{1\times1}(Concat(\boldsymbol{M}_X^1, \boldsymbol{M}_Y^1))))) \tag{10}$$

$$\begin{aligned} \boldsymbol{W}_X &= sigmoid(Conv_{1\times1}(\boldsymbol{M}_X^2)) \\ \boldsymbol{W}_Y &= sigmoid(Conv_{1\times1}(\boldsymbol{M}_Y^2)) \end{aligned} \tag{11}$$

$$\boldsymbol{OUT}_{\text{PAM}} = \boldsymbol{R}_2 \otimes \boldsymbol{W}_X \otimes \boldsymbol{W}_Y \tag{12}$$

式中，$\boldsymbol{F}_1$表示指导特征，$AvgPool_X(\cdot)$表示沿X方向的平均池化，$AvgPool_Y(\cdot)$表示沿Y方向的平均池化，$\boldsymbol{X}_{\text{att}}$、$\boldsymbol{Y}_{\text{att}}$表示由指导特征生成的注意力图，$\boldsymbol{M}_X^1$、$\boldsymbol{M}_X^2$、$\boldsymbol{M}_Y^1$、$\boldsymbol{M}_Y^2$表示沿$X$方向和$Y$方向处理的中间结果，$Concat(\cdot)$和$Split(\cdot)$分别表示空间维度的拼接操作和分离操作，$\boldsymbol{OUT}_{\text{PAM}}$表示PAM最终的输出特征图。需要注意的是，指导特征生成的注意力图$(\boldsymbol{X}_{\text{att}}, \boldsymbol{Y}_{\text{att}})$的像素值介于[0, 1]之间，加上0.5倍的输入特征图，新特征图的像素值就会分布在1的左右，由此可以强调大于1的部分，削弱小于1的部分，但不完全消除这些信息。

2.4 显著信息增强模块

本文提出的显著信息增强模块SIEM结构如图5所示，该模块可以深度挖掘高级特征中的前景和背景信息，并根据它们过滤掉跳跃连接部分低层次特征中的噪声干扰，同时增强低层次特征中的显著信息。由于$\boldsymbol{F}_5$是CNN编码器最后一层的输出，并且经过可变形卷积后受到了真值图的监督，其中包含着比较丰富的显著信息，所以使用$\boldsymbol{F}_5$增强低层次特征中的显著信息。首先，$\boldsymbol{F}_5$经过1×1卷积进行通道变换并经过sigmoid函数获得空间权重。接着将其通道数压缩到1以生成$\boldsymbol{att}$和$\boldsymbol{1}-\boldsymbol{att}$，它们分别包含着高级特征中的前景和背景信息。然后将低层次特征$\boldsymbol{LF}$分别与$\boldsymbol{att}$和$\boldsymbol{1}-\boldsymbol{att}$相乘，再经过CBR操作进一步提取蕴含在前景和背景中的显著特征。最后将结果相加，经过$\boldsymbol{CBR}$操作后作为最终输出。以上处理过程可以表示为

$$\boldsymbol{att} = sigmoid(Conv_{1\times1}(\boldsymbol{F}_5)) \tag{13}$$

$$\begin{aligned} \boldsymbol{OUT}_{\text{SIEM}} = CBR(&CBR(\boldsymbol{att} \otimes \boldsymbol{LF}) + \\ &CBR((\boldsymbol{1}-\boldsymbol{att}) \otimes \boldsymbol{LF})) \end{aligned} \tag{14}$$

式中，$\boldsymbol{LF}$表示低层次特征，$\boldsymbol{att}$表示由$\boldsymbol{F}_5$生成的注意力图，$\boldsymbol{OUT}_{\text{SIEM}}$表示SIEM模块输出的特征图。

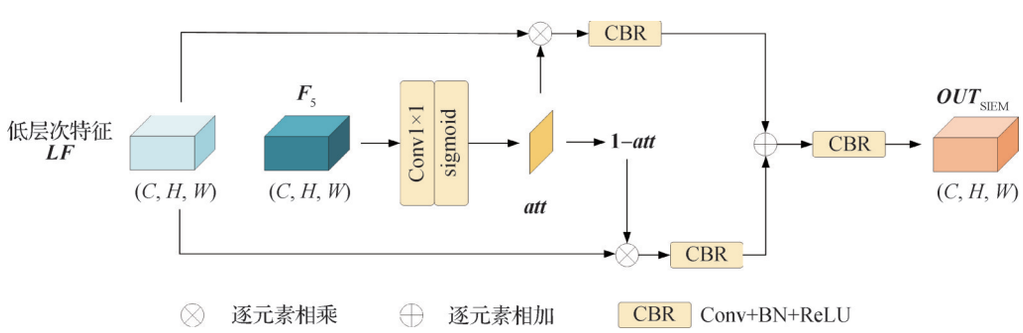

图5 显著信息增强模块结构图

Fig. 5 Structure diagram of salient information enhancement module

2.5 损失函数

为了让模型快速地收敛，本文采用多级监督的形式，分别在DPNet的输出$\boldsymbol{F}_0^1$、$\boldsymbol{F}_0^2$、CNN编码器的最后一层输出$\boldsymbol{F}_5$、ViT编码器的第2层输出$\boldsymbol{F}_2$以及跨模型嫁接模块输出的交叉注意力矩阵$\boldsymbol{CAM}$处设置损失监督。其中，$\boldsymbol{F}_0^1$、$\boldsymbol{F}_0^2$分别使用真值图和细节图监督，$\boldsymbol{F}_5$、$\boldsymbol{F}_2$使用真值图监督，$\boldsymbol{CAM}$使用Xie等人（2022）提出的注意力引导损失进行监督。360°全景

图像经过ERP投影后的RGB图像、真值图、细节图和边缘图的示例如图6所示。其中，细节图相比传统的边缘图具有更广泛的监督范围，更有利于细化边界。

图6　原图及监督图像示例

Fig. 6　Examples of original and supervision images

具体的损失函数采用BCE（binary crossentropy）损失函数（De Boer等，2005）和IOU（intersection over union）损失函数（Máttyus 等，2017）来对模型输出 $\boldsymbol{F}_O^1$、$\boldsymbol{F}_O^2$、$\boldsymbol{F}_5$、$\boldsymbol{F}_2$ 进行监督。DPNet具体损失函数组成可以表示为

$$loss_1 = \alpha l_{\text{bce+iou}}(\boldsymbol{F}_O^1, \boldsymbol{G}) + \beta l_{\text{bce+iou}}(\boldsymbol{F}_O^2, \boldsymbol{DG}) \quad (15)$$

$$loss_2 = \gamma l_{\text{bce+iou}}(\boldsymbol{F}_5, \boldsymbol{G}) \quad (16)$$

$$loss_3 = \varepsilon l_{\text{bce+iou}}(\boldsymbol{F}_2, \boldsymbol{G}) \quad (17)$$

$$loss_4 = l_{\text{AG}}(\boldsymbol{CAM}, \boldsymbol{G}) \quad (18)$$

$$Loss = loss_1 + loss_2 + loss_3 + loss_4 \quad (19)$$

式中，$l_{\text{bce+iou}}$ 表示BCE损失函数和IOU损失函数，l_{AG} 表示注意力引导损失，$\boldsymbol{F}_O^1$ 表示DPNet输出的显著性目标图，$\boldsymbol{F}_O^2$ 表示DPNet输出的细节图，$\boldsymbol{G}$ 表示真值图，$\boldsymbol{DG}$ 表示对应细节图，$\boldsymbol{CAM}$ 表示交叉注意力矩阵，$\alpha, \beta, \gamma, \varepsilon$ 分别表示各部分损失函数的系数，本文开展实验时，$\alpha, \beta, \gamma, \varepsilon$ 的值分别为1.0、0.2、0.125、0.125。$Loss$ 表示最终损失函数，可以看到它由4个子部分组成，$loss_1$ 表示对模型最终输出的特征图 $\boldsymbol{F}_O^1$、$\boldsymbol{F}_O^2$ 进行监督的损失函数，$loss_2$ 表示对CNN编码器的最后一层输出 $\boldsymbol{F}_5$ 进行监督的损失函数，$loss_3$ 表示对ViT编码器的第2层输出 $\boldsymbol{F}_2$ 进行监督的损失函数，$loss_4$ 表示对跨模型嫁接模块输出的交叉注意力矩阵 $\boldsymbol{CAM}$ 进行监督的损失函数。

3　实验与分析

3.1　数据集及评价指标

本文使用两个公开数据集360-SOD（Li等，2020a）和360-SSOD（Ma等，2020）测试DPNet的性能。其中，360-SOD包含500幅高分辨率ERP全景图像，360-SSOD包含1 105幅高分辨率ERP全景图像。使用以下8种评价指标对模型性能进行评估：平均绝对误差（mean absolute error，MAE）（Perazzi等，2012）、F-measure（max-F、mean-F、adp-F）（Achanta等，2009）、E-measure（max-Em、mean-Em、adp-Em）（Fan等，2018）和结构测度（structure-measure，S_m）（Fan等，2017）。

3.2　实验设置

本文使用PyTorch训练DPNet，所有实验均在一台配有RTX 4070 GPU的台式主机上完成。在实验中，CNN编码器主干网络采用ResNet34，ViT编码器主干网络采用DilateFormer（base），它们均在ImageNet上进行了预训练。优化器采用随机梯度下降（stochastic gradient descent，SGD），动量参数和权重衰减分别设置为0.9和0.000 5，网络的最大学习率设置为0.01，周期设置为20，batchsize设置为10。ViT主干输入分辨率为224×224像素的ERP图像，CNN主干输入分辨率为1 024×512像素的ERP图像。实验中数据集的划分如下：360-SOD数据集中400幅图像用于训练，100幅图像用于测试；360-SSOD数据集中850幅图像用于训练，255幅图像用于测试。训练过程中，使用随机翻转、裁剪和多尺度输入图像对数据集进行增强。

3.3　实验对比

本文将所提出的模型与13种目前先进的SOD方法进行对比，包括LDF（Wei等，2020b）、F^3Net（Wei等，2020a）、VST（Liu等，2021）、RRNet（relational reasoning network）（Cong 等，2022）、PGNet（pyramid grafting network）（Xie等，2022）、MSCNet（multi-scale context network）（Lin 等，2022a）、SeaNet（Li 等，2023）、BSCGNet（boundary-semantic collaborative guidance network）（Feng 等，2023）、TSCNet（Li 等，2024）、FANet（Huang 等，2020）、MPFRNet（multi-projection fusion and refinement network）（Cong 等，2024）、LDNet（Huang 等，2023）和DATFormer（Zhao等，2023）。为了公平比较，所有模型均使用官方代码或作者提供的显著图在同一软硬件环境中测试，并且都进行了微调以获得最好结果。

各模型在360-SOD数据集的客观指标对比如表1所示，表中↑表示越大越好，↓表示越小越好（下同）。可以看到，本文方法相较于对比方法取得了最好的性能。

360-SOD数据集上的主观结果比较如图7所示，本文方法能够检测到分布在ERP图像各个位置的

表1　360-SOD数据集上各模型客观指标对比

Table 1　Comparison of objective metrics of all models on the 360-SOD dataset

类别	方法(年份)	MAE ↓	max-F ↑	mean-F ↑	adp-F ↑	max-Em ↑	mean-Em ↑	adp-Em ↑	Sm ↑
2D SOD	LDF(2020)	0.023 8	0.709 4	0.691 4	0.655 4	0.876 8	0.864 2	0.863 1	0.792 0
	F^3Net(2020)	0.023 0	0.705 4	0.677 0	0.641 1	0.873 9	0.863 8	0.855 5	0.793 4
	VST(2021)	0.026 4	0.693 3	0.632 9	0.531 0	0.889 2	0.835 2	0.768 9	0.787 8
	RRNet(2021)	0.026 1	0.634 2	0.615 8	0.610 4	0.792 9	0.754 6	0.805 6	0.765 0
	PGNet(2022)	0.0 257	0.687 1	0.673 4	0.643 9	0.838 4	0.823 1	0.855 9	0.780 0
	MSCNet (2022)	0.046 4	0.724 3	0.656 1	0.626 9	0.882 3	0.834 3	0.836 3	0.735 4
	SeaNet(2023)	0.056 6	0.465 7	0.451 9	0.461 2	0.802 8	0.744 3	0.820 6	0.649 7
	BSCGNet(2023)	0.024 9	0.681 1	0.675 4	0.677 7	0.855 9	0.831 8	0.865 1	0.784 0
	TSCNet(2023)	0.024 3	0.713 7	0.703 9	0.688 7	0.888 0	0.877 1	0.887 6	0.798 3
360°SOD	FANet(2020)	0.026 1	0.685 5	0.661 2	0.596 9	0.879 4	0.858 8	0.837 3	0.777 5
	MPFRNet(2023)	0.019 1	0.765 2	0.755 3	0.744 7	0.885 1	0.874 7	0.890 4	0.841 7
	LDNet(2023)	0.028 9	0.656 2	0.639 1	0.617 1	0.865 5	0.841 4	0.857 6	0.767 9
	DATFormer(2023)	**0.017 6**	0.780 1	0.762 2	0.727 2	0.900 7	0.888 3	0.903 0	0.846 6
	本文	0.018 8	**0.803 8**	**0.788 4**	**0.755 8**	**0.921 7**	**0.910 3**	**0.910 0**	**0.850 2**

注:加粗、下划线和点式下划线字体分别表示各列最优、次优和第3的结果。

显著目标,对于接近极点的目标(第4、5、7行)也有很好的检测效果。另外,本文方法能够有效地将显著目标从背景中分离出来而不会引入背景干扰,如第8行图像中的人物目标,其他方法均误将背景信息当做显著信息检测了出来,而本文方法则没有产生类似错误。最后,本文方法能够很好地捕捉图像细节,如第6行图像中的飞机螺旋桨叶。

各模型在360-SSOD数据集的客观指标对比如表2所示。可以看到,本文方法在所有指标上均取得了最佳值,总体综合性能优于所有对比方法。

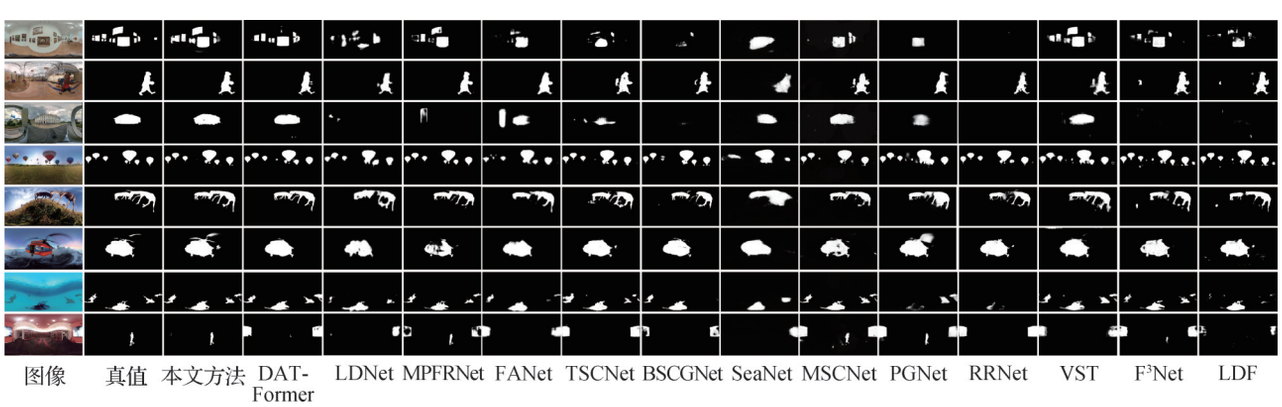

图7　360-SOD数据集上的主观结果比较

Fig. 7　Comparison of subjective results on the 360-SOD dataset

360-SSOD数据集上的主观结果比较如图8所示,相较于其他方法,本文方法能够较好地屏蔽背景信息的干扰(如第1、2、9行),而其他方法或多或少地受到了背景信息的干扰,存在多检或漏检的现象。本文方法对显著目标的细节检测更完整、更接近真值图,如第7行图像中建筑物的电线杆,第8行中人物的形态。另外,本文方法对小目标的检测效果也非常好,如第4、5行中的物体,大部分方法产生了错误的预测结果,或没有检测出任何物体,而本文方法可以在强背景干扰的条件下准确地检测出小目标。

表 2　360-SSOD 数据集上各模型客观指标对比

Table 2　Comparison of objective metrics of all models on the 360-SSOD dataset

类别	方法(年份)	MAE ↓	max-F ↑	mean-F ↑	adp-F ↑	max-Em ↑	mean-Em ↑	adp-Em ↑	Sm ↑
2D SOD	LDF(2020)	0.032 8	0.608 1	0.587 2	0.532 6	0.858 2	0.832 4	0.778 8	0.743 7
	F^3Net(2020)	0.032 3	0.605 7	0.584 5	0.556 8	0.857 5	0.840 3	0.817 1	0.737 6
	VST(2021)	0.035 6	0.596 2	0.526 8	0.447 9	0.849 2	0.766 7	0.700 3	0.741 1
	RRNet(2021)	0.034 1	0.420 4	0.395 3	0.416 3	0.695 3	0.596 6	0.700 7	0.645 4
	PGNet(2022)	0.032 6	0.567 1	0.554 3	0.554 7	0.783 1	0.758 9	0.821 0	0.719 7
	MSCNet (2022)	0.050 5	0.597 7	0.568 1	0.531 6	0.830 6	0.795 4	0.778 7	0.711 8
	SeaNet(2023)	0.041 1	0.394 3	0.343 0	0.401 1	0.766 0	0.575 8	0.772 7	0.596 1
	BSCGNet(2023)	0.029 2	0.596 9	0.585 1	0.595 6	0.830 0	0.791 3	0.826 1	0.735 0
	TSCNet(2023)	0.029 0	0.634 6	0.625 1	0.607 9	0.864 1	0.851 0	0.852 0	0.749 3
360°SOD	FANet(2020)	0.029 1	0.607 1	0.584 7	0.537 4	0.846 0	0.827 1	0.795 5	0.740 6
	MPFRNet(2023)	–	–	–	–	–	–	–	–
	LDNet(2023)	0.034 2	0.586 2	0.567 2	0.537 6	0.839 0	0.818 7	0.822 6	0.724 5
	DATFormer(2023)	0.031 8	0.599 5	0.579 2	0.578 5	0.820 8	0.769 2	0.834 4	0.742 4
	本文	**0.028 4**	**0.671 2**	**0.657 9**	**0.645 2**	**0.872 1**	**0.851 2**	**0.856 8**	**0.769 4**

注：加粗、下划线和点式下划线字体分别表示各列最优、次优和第 3 的结果。“–”表示因没有源代码而无法测试的值。

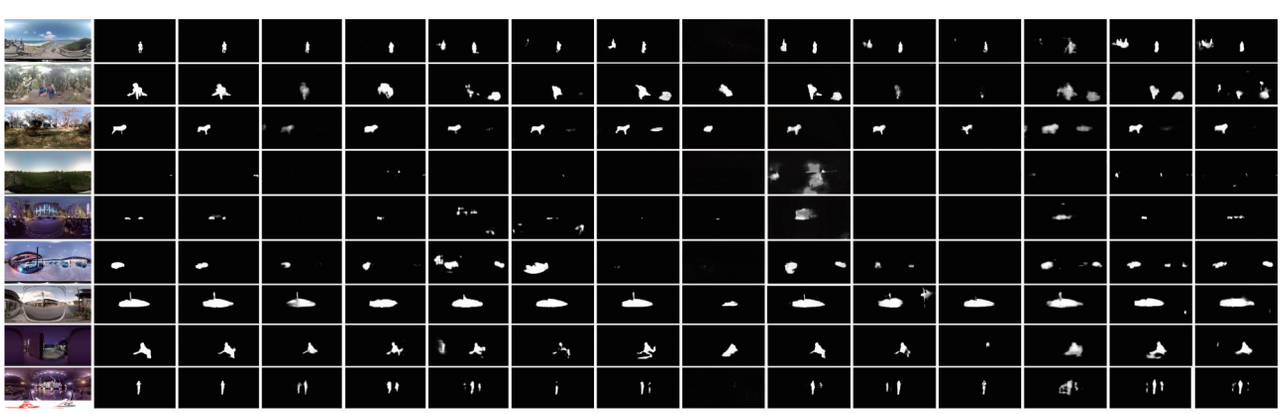

图 8　360-SSOD 数据集上的主观结果比较

Fig. 8　Comparison of subjective results on the 360-SSOD dataset

3.4　泛化性实验对比

为了验证本文模型的泛化性能，与现有先进模型进行实验对比。各模型在 360-SSOD 数据集上进行训练、在 360-SOD 数据集上进行测试的客观指标对比如表 3 所示；各模型在 360-SOD 数据集上进行训练、在 360-SSOD 数据集上进行测试的客观指标对比如表 4 所示。从表 3 和表 4 中可以看到，本文方法相较于对比方法取得了最好的综合性能。

3.5　消融实验

为了更好地测试本文提出各个模块的有效性，在 360-SOD 数据集上进行消融实验，客观指标对比如表 5 所示，主观结果对比如图 9 所示。其中 baseline 表示基线模型，DAM、PAM、SIEM 分别代表本文提出的畸变自适应模块、位置感知模块、显著信息增强模块，w/o 表示从完整模型中删除某一模块。

各个模块的消融实验结果分析如下：

1）畸变自适应模块消融实验分析。畸变自适应

表3　各模型在360-SSOD数据集上进行训练、在360-SOD数据集上进行测试的客观指标对比

Table 3　Comparison of objective metrics for each model trained on the 360-SSOD datase and tested on the 360-SOD dataset

类别	方法(年份)	MAE ↓	max-F ↑	mean-F ↑	adp-F ↑	max-Em ↑	mean-Em ↑	adp-Em ↑	Sm ↑
2D SOD	LDF(2020)	0.032 7	0.614 3	0.594 6	0.537 3	0.834 8	0.808 5	0.798 7	0.740 1
	F^3Net(2020)	0.031 5	0.622 2	0.598 5	0.563 0	0.839 2	**0.821 6**	0.828 6	0.737 3
	VST(2021)	0.036 6	0.625 0	0.534 6	0.461 9	**0.842 1**	0.749 9	0.731 0	0.733 2
	RRNet(2021)	0.037 8	0.316 8	0.292 8	0.319 2	0.624 7	0.484 4	0.631 5	0.593 4
	PGNet(2022)	0.029 7	0.585 7	0.565 1	0.587 9	0.755 9	0.704 1	0.791 0	0.726 9
	MSCNet (2022)	0.050 9	0.600 4	0.570 1	0.523 6	0.813 0	0.781 4	0.783 6	0.700 2
	SeaNet(2023)	0.042 6	0.363 7	0.305 3	0.380 3	0.736 4	0.540 7	0.766 4	0.575 7
	BSCGNet(2023)	0.034 1	0.513 6	0.497 4	0.530 6	0.749 6	0.661 4	0.778 1	0.682 1
	TSCNet(2023)	0.034 4	0.608 2	0.601 9	0.601 9	0.821 9	0.790 5	0.830 1	0.722 3
360°SOD	FANet(2020)	0.034 4	0.596 6	0.575 2	0.526 7	0.799 2	0.782 3	0.804 1	0.724 7
	MPFRNet(2023)	–	–	–	–	–	–	–	–
	LDNet(2023)	–	–	–	–	–	–	–	–
	DATFormer(2023)	0.031 9	0.645 1	0.621 1	0.618 2	0.809 0	0.750 4	0.826 2	0.749 0
	本文	**0.028 9**	**0.670 7**	**0.657 9**	**0.677 3**	0.831 5	0.791 6	**0.841 2**	**0.763 1**

注：加粗、下划线和点式下划线字体分别表示各列最优、次优和第3的结果。“–”表示因没有源代码而无法测试的值。

表4　各模型在360-SOD数据集上进行训练、在360-SSOD数据集上进行测试的客观指标对比

Table 4　Comparison of objective metrics for each model trained on the 360-SOD dataset and tested on the 360-SSOD dataset

类别	方法(年份)	MAE ↓	max-F ↑	mean-F ↑	adp-F ↑	max-Em ↑	mean-Em ↑	adp-Em ↑	Sm ↑
2D SOD	LDF(2020)	0.050 7	0.438 6	0.430 9	0.417 7	0.720 9	0.701 7	0.728 4	0.647 7
	F^3Net(2020)	0.054 5	0.431 5	0.425 1	0.422 0	0.696 7	0.691 6	0.712 0	0.643 5
	VST(2021)	0.057 6	0.436 8	0.410 3	0.372 0	0.724 6	0.700 3	0.673 3	0.650 6
	RRNet(2021)	**0.050 0**	0.386 8	0.375 3	0.392 8	0.669 5	0.620 6	0.674 2	0.619 1
	PGNet(2022)	0.054 5	0.373 0	0.365 0	0.370 1	0.641 6	0.634 7	0.678 4	0.625 3
	MSCNet (2022)	0.083 1	0.444 0	0.412 7	0.405 6	0.720 5	0.686 2	0.701 7	0.615 0
	SeaNet(2023)	0.079 3	0.347 2	0.334 5	0.341 6	0.710 4	0.710 4	0.724 4	0.588 2
	BSCGNet(2023)	0.051 6	0.426 0	0.418 7	0.422 9	0.712 9	0.686 7	0.720 7	0.642 0
	TSCNet(2023)	0.052 8	0.445 9	0.439 6	0.429 2	**0.747 5**	**0.740 7**	**0.743 7**	0.652 0
360°SOD	FANet(2020)	0.053 1	0.411 3	0.403 0	0.390 3	0.692 8	0.683 6	0.702 8	0.640 5
	MPFRNet(2023)	–	–	–	–	–	–	–	–
	LDNet(2023)	–	–	–	–	–	–	–	–
	DATFormer(2023)	0.050 4	0.453 1	0.448 2	0.446 1	0.694 5	0.689 7	0.700 5	0.658 5
	本文	0.052 4	**0.464 2**	**0.459 3**	**0.463 2**	0.721 0	0.713 3	0.737 4	**0.663 0**

注：加粗、下划线和点式下划线字体分别表示各列最优、次优和第3的结果。“–”表示因没有源代码而无法测试的值。

表 5　360-SOD 数据集上不同模型组合的客观指标对比

Table 5　Objective metrics comparison of different model combinations on the 360-SOD dataset

方法	MAE ↓	max-F ↑	mean-F ↑	adp-F ↑	max-Em ↑	mean-Em ↑	adp-Em ↑	Sm ↑
baseline	0.021 7	0.763 6	0.740 1	0.708 0	0.888 1	0.879 5	0.880 4	0.825 0
w/o DAM	0.019 8	0.766 7	0.752 4	0.728 9	0.885 4	0.875 7	0.886 8	0.827 5
w/o PAM	0.019 1	0.791 8	0.767 9	0.736 4	0.902 5	0.896 1	0.899 5	0.841 3
w/o SIEM	0.020 8	0.781 6	0.764 3	0.733 2	0.909 2	0.897 5	0.895 0	0.836 5
本文	**0.018 8**	**0.803 8**	**0.788 4**	**0.755 8**	**0.921 7**	**0.910 3**	**0.910 0**	**0.850 2**

注：加粗、下划线和点式下划线字体分别表示各列最优、次优和第 3 的结果。

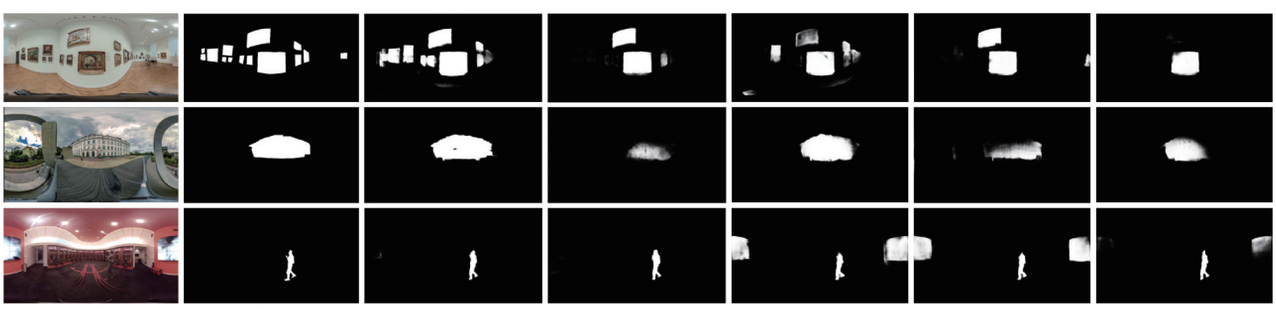

图 9　360-SOD 数据集上不同模型组合的主观结果比较

Fig. 9　Comparison of subjective results of different model combinations on the 360-SOD dataset

模块在本文模型中主要用于从全局角度对 ERP 投影引起的畸变进行建模，其对后续特征的提取至关重要。由表 5 可以看出，从完整模型中删除畸变自适应模块后各项客观评价指标均有较大程度的下降，同时如图 9 第 1、2 行所示，模型对于显著目标（第 1 行图像中的多个相框、第 2 行图像中的中心建筑物）的检测能力大大下降，并且容易受到几何畸变的影响，这验证了本文提供的畸变自适应模块的有效性。

2）位置感知模块消融实验分析。位置感知模块主要用于将 ViT 主干提取的包含畸变的全局特征信息传输到 CNN 主干中，起到信息复用的效果，从而指导后续特征的提取。由表 5 可以看出，从完整模型中删除位置感知模块后各项评价指标均有较大程度的下降，同时如图 9 所示，模型对于显著目标的检测能力大大下降，并且容易受到背景信息的干扰而出现误检，这验证了本文提供的位置感知模块的有效性。

3）显著信息增强模块消融实验分析。显著信息增强模块主要用于补充模型深层特征丢失的细节信息，并且过滤掉非显著信息，增强显著信息。由表 5 可以看出，从完整模型中删除显著信息增强模块后各项评价指标均有较大程度的下降。如图 9 所示，模型对于显著目标的检测能力大大下降，并且容易受到背景信息的干扰而出现误检，这验证了本文提供的显著信息增强模块的有效性。

本文模型在 360-SSOD 数据集上也得到类似的消融实验结果，限于篇幅，不再赘述。

3.6　复杂度实验对比

本文选择模型计算量 FLOPs（floating point of operations per second）以及模型的参数量 Params 来对比不同模型的复杂度，复杂度对比实验结果如表 6 所示。其中，MSCNet（2022）、SeaNet（2023）、LDNet（2023）是专门设计的轻量化方法，所以 FLOPs 和 Params 都非常小。本文方法与其他非轻量化模型相比，计算量较小，参数量较大，这可能是因为本文方法采用双编码器主干，但是从之前的实验来看，本文方法获得了最好的客观指标和主观结果，并且泛化性能也最好。上述实验结果表明，本文模型在保证检测精度的同时较好地控制了模型的复杂度。

4　结　论

现有 360°SOD 方法尚不能很好地应对 360°全景

表6　各模型复杂度对比

Table 6　Comparison of complexity of each model

方法(年份)	FLOPs/G	Params/M
LDF(2020)	15.572 4	25.150 1
F^3Net(2020)	16.494 5	25.536 7
VST(2021)	31.089 2	83.054 9
RRNet(2021)	451.409 3	75.693 6
PGNet(2022)	42.970 5	72.666 4
MSCNet(2022)	15.466 3	3.264 4
SeaNet(2023)	1.809 4	2.745 1
BSCGNet(2023)	86.463 5	26.993 3
TSCNet(2023)	117.408 6	101.203 6
FANet(2020)	123.721 2	25.399 3
MPFRNet(2023)	–	–
LDNet(2023)*	2.900 0	3.400 0
DATFormer(2023)	38.178 9	29.568 1
本文	22.410 6	78.485 4

注:“–”表示因没有源代码而无法测试的值,“*”表示从原论文中获取的数值。

图像投影后的几何畸变问题和复杂场景下的背景干扰问题。为此,本文提出一种基于CNN和Vision Transformer的畸变自适应和位置感知360°SOD网络DPNet。DPNet中的畸变自适应模块DAM和位置感知模块PAM可以根据ERP的特点以及具体输入图像引导网络关注需要关注的地方。此外,本文提出的显著信息增强模块SIEM能够用高级特征指导低级特征,过滤低级特征中的非显著信息和增强其中的显著信息,从而能够帮助模型有效应对复杂场景下的背景干扰问题。大量实验表明,本文方法的性能优于现有13种代表性先进SOD方法。

参考文献(References)

Achanta R, Hemami S, Estrada F and Susstrunk S. 2009. Frequency-tuned salient region detection//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 1597-1604 [DOI: 10.1109/CVPR.2009.5206596]

Chen G, Shao F, Chai X L, Jiang Q P and Ho Y S. 2024. Multi-stage salient object detection in 360° omnidirectional image using complementary object-level semantic information. IEEE Transactions on Emerging Topics in Computational Intelligence, 8(1): 776-789 [DOI: 10.1109/TETCI.2023.3259433]

Chen X L, Zhang P C, Lu Y B and Cao B N. 2023. Saliency detection of panoramic images based on robust vision transformer and multiple attention. Journal of Electronics and Information Technology, 45(6): 2246-2255 (陈晓雷, 张鹏程, 卢禹冰, 曹宝宁. 2023. 基于鲁棒视觉变换和多注意力的全景图像显著性检测. 电子与信息学报, 45(6): 2246-2255) [DOI: 10.11999/JEIT220684]

Chen Z Z, Li Y M and Zhang Y X. 2018. Recent advances in omnidirectional video coding for virtual reality: projection and evaluation. Signal Processing, 146: 66-78 [DOI: 10.1016/j. sigpro. 2018. 01.004]

Cong R M, Huang K, Lei J J, Zhou X, Huang Q M and Meur O L. 2024. Multi-projection fusion and refinement network for salient object detection in 360 omnidirectional image. IEEE Transactions on Neural Networks and Learning Systems, 35(7): 9495-9507 [DOI: 10.1109/TNNLS.2022.3233883]

Cong R M, Zhang Y M, Fang L Y, Li J, Zhao Y and Kwong S. 2022. RRNet: relational reasoning network with parallel multiscale attention for salient object detection in optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-11 [DOI: 10.1109/TGRS.2021.3123984]

Dai H W, Bao L X, Shen K Y, Zhou X F and Zhang J Y. 2023. 360° omnidirectional salient object detection with multi-scale interaction and densely-connected prediction//Proceedings of the 12th International Conference on Image and Graphics. Nanjing, China: Springer: 427-438 [DOI: 10.1007/978-3-031-46305-1_35]

De Boer P T, Kroese D P, Mannor S and Rubinstein R Y. 2005. A tutorial on the cross-entropy method. Annals of Operations Research, 134(1): 19-67 [DOI: 10.1007/s10479-005-5724-z]

Fan D P, Cheng M M, Liu Y, Li T and Borji A. 2017. Structure-measure: a new way to evaluate foreground maps//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4558-4567 [DOI: 10.1109/ICCV.2017.487]

Fan D P, Gong C, Cao Y, Ren B, Cheng M M and Borji A. 2018. Enhanced-alignment measure for binary foreground map evaluation//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: AAAI Press: 698-704

Fan D P, Ji G P, Sun G L, Cheng M M, Shen J B and Shao L. 2020. Camouflaged object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2774-2784 [DOI: 10.1109/CVPR42600.2020.00285]

Feng D J, Chen H Y, Liu S N, Liao Z Y, Shen X Y, Xie Y K and Zhu J. 2023. Boundary-semantic collaborative guidance network with dual-stream feedback mechanism for salient object detection in optical remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 61: #4706317 [DOI: 10.1109/TGRS. 2023. 3332282]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning

for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He W and Pan C. 2022. The salient object detection based on attention-guided network. Journal of Image and Graphics, 27(4): 1176-1190 (何伟, 潘晨. 2022. 注意力引导网络的显著性目标检测. 中国图象图形学报, 27(4): 1176-1190) [DOI: 10.11834/jig.200658]

He Z T, Shao F, Chen G, Chai X L and Ho Y S. 2024. SCFANet: semantics and context feature aggregation network for 360° salient object detection. IEEE Transactions on Multimedia, 26: 2276-2288 [DOI: 10.1109/TMM.2023.3293994]

Hong S, You T, Kwak S and Han B. 2015. Online tracking by learning discriminative saliency map with convolutional neural network//Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France: JMLR.org: 597-606

Hoyer L, Munoz M, Katiyar P, Khoreva A and Fischer V. 2019. Grid saliency for context explanations of semantic segmentation//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #580

Huang M K, Li G Y, Liu Z and Zhu L C. 2023. Lightweight distortion-aware network for salient object detection in omnidirectional images. IEEE Transactions on Circuits and Systems for Video Technology, 33(10): 6191-6197 [DOI: 10.1109/TCSVT.2023.3253685]

Huang M K, Liu Z, Li G Y, Zhou X F and Le Meur O. 2020. FANet: features adaptation network for 360 omnidirectional salient object detection. IEEE Signal Processing Letters, 27: 1819-1823 [DOI: 10.1109/LSP.2020.3028192]

Jiao J Y, Tang Y M, Lin K Y, Gao Y P, Ma A J, Wang Y W and Zheng W S. 2023. DilateFormer: multi-scale dilated transformer for visual recognition. IEEE Transactions on Multimedia, 25: 8906-8919 [DOI: 10.1109/TMM.2023.3243616]

Li G Y, Bai Z and Liu Z. 2024. Texture-semantic collaboration network for ORSI salient object detection. IEEE Transactions on Circuits and Systems II: Express Briefs, 71(4): 2464-2468 [DOI: 10.1109/TCSII.2023.3333436]

Li G Y, Liu Z, Zhang X P and Lin W S. 2023. Lightweight salient object detection in optical remote-sensing images via semantic matching and edge alignment. IEEE Transactions on Geoscience and Remote Sensing, 61: #5601111 [DOI: 10.1109/TGRS.2023.3235717]

Li J, Su J M, Xia C Q and Tian Y H. 2020a. Distortion-adaptive salient object detection in 360 omnidirectional images. IEEE Journal of Selected Topics in Signal Processing, 14(1): 38-48 [DOI: 10.1109/JSTSP.2019.2957982]

Li J, Zhao Y F, Ye W H, Yu K W and Ge S M. 2020b. Attentive deep stitching and quality assessment for 360 omnidirectional images. IEEE Journal of Selected Topics in Signal Processing, 14(1): 209-221 [DOI: 10.1109/JSTSP.2019.2953950]

Lin Y H, Sun H, Liu N Z, Bian Y T, Cen J and Zhou H Y. 2022a. A lightweight multi-scale context network for salient object detection in optical remote sensing images//Proceedings of the 26th International Conference on Pattern Recognition (ICPR). Montreal, Canada: IEEE: 238-244 [DOI: 10.1109/ICPR56361.2022.9956350]

Lin Y H, Sun H, Liu N Z, Bian Y T, Cen J and Zhou H Y. 2022b. Attention guided network for salient object detection in optical remote sensing images//Proceedings of the 31st International Conference on Artificial Neural Networks. Bristol, UK: Springer: 25-36 [DOI: 10.1007/978-3-031-15919-0_3]

Liu G H and Fan D P. 2013. A model of visual attention for natural image retrieval//Proceedings of 2013 International Conference on Information Science and Cloud Computing Companion. Guangzhou, China: IEEE: 728-733 [DOI: 10.1109/ISCC-C.2013.21]

Liu N, Zhang N, Wan K Y, Shao L and Han J W. 2021. Visual saliency transformer//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 4722-4732 [DOI: 10.1109/ICCV48922.2021.00468]

Luz G, Ascenso J, Brites C and Pereira F. 2017. Saliency-driven omnidirectional imaging adaptive coding: modeling and assessment//2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP). Luton, UK: IEEE: 1-6 [DOI: 10.1109/MMSP.2017.8122228]

Ma G X, Li S, Chen C L Z, Hao A M and Qin H. 2020. Stage-wise salient object detection in 360 omnidirectional image via object-level semantical saliency ranking. IEEE Transactions on Visualization and Computer Graphics, 26(12): 3535-3545 [DOI: 10.1109/TVCG.2020.3023636]

Máttyus G, Luo W J and Urtasun R. 2017. DeepRoadMapper: extracting road topology from aerial images//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 3438-3446 [DOI: 10.1109/ICCV.2017.372]

Maugey T, Le Meur O and Liu Z. 2017. Saliency-based navigation in omnidirectional image//2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP). Luton, UK: IEEE: 1-6 [DOI: 10.1109/MMSP.2017.8122229]

Perazzi F, Krähenbühl P, Pritch Y and Hornung A. 2012. Saliency filters: contrast based filtering for salient region detection//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 733-740 [DOI: 10.1109/CVPR.2012.6247743]

Serrano A, Sitzmann V, Ruiz-Borau J, Wetzstein G, Gutierrez D and Masia B. 2017. Movie editing and cognitive event segmentation in virtual reality video. ACM Transactions on Graphics, 36(4): #47 [DOI: 10.1145/3072959.3073668]

Tang B, Liu Z Y, Tan Y C and He Q. 2023. HRTransNet: HRFormer-driven two-modality salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 33(2): 728-742

[DOI: 10.1109/TCSVT.2022.3202563]

Wei J, Wang S H and Huang Q M. 2020a. F^{3}Net: fusion, feedback and focus for salient object detection//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press: 12321-12328 [DOI: 10.1609/aaai.v34i07.6916]

Wei J, Wang S H, Wu Z, Su C, Huang Q M and Tian Q. 2020b. Label decoupling framework for salient object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13022-13031 [DOI: 10.1109/CVPR42600.2020.01304]

Wu J J, Xia C Q, Yu T S and Li J. 2023. View-aware salient object detection for 360 omnidirectional image. IEEE Transactions on Multimedia, 25: 6471-6484 [DOI: 10.1109/TMM.2022.3209015]

Xie C X, Xia C Q, Ma M C, Zhao Z R, Chen X W and Li J. 2022. Pyramid grafting network for one-stage high resolution saliency detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11707-11716 [DOI: 10.1109/CVPR52688.2022.01142]

Yao C L, Feng L, Kong Y Q, Xiao L and Chen T. 2023. Transformers and CNNs fusion network for salient object detection. Neurocomputing, 520: 342-355 [DOI: 10.1016/j.neucom.2022.10.081]

Ye L W, Liu Z, Li L N, Shen L Q, Bai C and Wang Y. 2017. Salient object segmentation via effective integration of saliency and objectness. IEEE Transactions on Multimedia, 19(8): 1742-1756 [DOI: 10.1109/TMM.2017.2693022]

Ye X Y, Zhu L, Wang W W and Fu Y. 2024. RGB_D salient object detection algorithm based on complementary information interaction. Journal of Image and Graphics, 29(5): 1252-1264 (叶欣悦, 朱磊, 王文武, 付云. 2024. 互补特征交互融合的RGB_D实时显著目标检测. 中国图象图形学报, 29(5): 1252-1264) [DOI: 10.11834/jig.230583]

Yuan J B, Zhu A Q, Xu Q Z, Wattanachote K and Gong Y Y. 2024. CTIF-Net: a CNN-transformer iterative fusion network for salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 34(5): 3795-3805 [DOI: 10.1109/TCSVT.2023.3321190]

Yuan Y, Gao P and Tan X Y. 2023. M^{3}Net: multilevel, mixed and multistage attention network for salient object detection [EB/OL]. [2024-10-15]. https://arxiv.org/pdf/2309.08365.pdf

Yun Y K and Lin W S. 2022. SelfReformer: self-refined network with transformer for salient object detection [EB/OL]. [2024-10-15]. https://arxiv.org/pdf/2205.11283.pdf

Zhang J, Zhang Q D, Shen X L and Wang X. 2023. Salient object detection on 360° omnidirectional image with bi-branch hybrid projection network//Proceedings of the 25th IEEE International Workshop on Multimedia Signal Processing (MMSP). Poitiers, France: IEEE: 1-5 [DOI: 10.1109/MMSP59012.2023.10337695]

Zhang Y, Hamidouche W and Deforges O. 2022. Channel-spatial mutual attention network for 360° salient object detection//Proceedings of the 26th International Conference on Pattern Recognition (ICPR). Montreal, Canada: IEEE: 3436-3442 [DOI: 10.1109/ICPR56361.2022.9956354]

Zhao J X, Liu J J, Fan D P, Cao Y, Yang J F and Cheng M M. 2019. EGNet: edge guidance network for salient object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 8778-8787 [DOI: 10.1109/ICCV.2019.00887]

Zhao R, Oyang W and Wang X G. 2017. Person re-identification by saliency learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(2): 356-370 [DOI: 10.1109/TPAMI.2016.2544310]

Zhao X Q, Pang Y W, Zhang L H, Lu H C and Zhang L. 2020. Suppress and balance: a simple gated network for salient object detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 35-51 [DOI: 10.1007/978-3-030-58536-5_3]

Zhao Y J, Zhao L X, Yu Q, Sheng L, Zhang J and Xu D. 2023. Distortion-aware transformer in 360° salient object detection//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 499-508 [DOI: 10.1145/3581783.3612025]

Zhou X M, Zhang Y, Li N, Wang X, Zhou Y and Ho Y S. 2021. Projection invariant feature and visual saliency-based stereoscopic omnidirectional image quality assessment. IEEE Transactions on Broadcasting, 67(2): 512-523 [DOI: 10.1109/TBC.2021.3056231]

作者简介

陈晓雷,男,教授,硕士生导师,主要研究方向为人工智能与计算机视觉。E-mail:chenxl703@lut.edu.cn

杜泽龙,男,硕士研究生,主要研究方向为全景图像显著目标检测。E-mail:1132911812@qq.com

张学功,男,硕士研究生,主要研究方向为全景图像显著目标检测。E-mail:864613727@qq.com

王兴,男,硕士研究生,主要研究方向为全景图像轻量级显著目标检测。E-mail:19312938573@163.com