doi:10.12068/j.issn.1005－3026.2025.20230204

基于跨模态融合的玻璃类似物分割方法

万应才，房立金，赵乾坤
（东北大学 机器人科学与工程学院，辽宁 沈阳 110169）

摘 要：玻璃和镜子等物体因缺乏明显纹理和形状，使得传统语义分割方法难以有效识别，影响视觉任务准确性．为了解决这个问题提出了一种基于Transformer的RGBD跨模态融合方法，用于玻璃类似物的分割．该方法采用Transformer网络，通过跨模态融合模块提取RGB和深度特征的自注意力，并利用多层注意力机制（MLP）整合RGBD特征，实现3种注意力特征的融合．RGB和深度特征被反馈到各自分支，以增强网络的特征提取能力．最终，语义分割解码器结合4个阶段的融合特征输出玻璃类似物的分割结果．结果表明，本文方法与EBLNet方法相比在GDD，Trans10k和MSD数据集上的交并比分别提高1.64%，2.26%，7.38%，与PDNet方法比较在RGBD−Mirror数据集上交并比提高了9.49%，验证了其有效性．

关 键 词：注意力；语义分割；玻璃类似物；跨模态；深度估计
中图分类号：TP 753 **文献标志码**：A **文章编号**：1005−3026（2025）01−0001−08

Segmentation Method for Glass-like Object Based on Cross-Modal Fusion

WAN Ying-cai，*FANG Li-jin*，*ZHAO Qian-kun*
（School of Robot Science & Engineering，Northeastern University，Shenyang 110169，China. Corresponding author：FANG Li-jin，E-mail：ljfang@mail.neu.edu.cn）

Abstract：Due to the lack of distinct textures and shapes，objects such as glass and mirrors pose challenges to traditional semantic segmentation algorithms，compromising the accuracy of visual tasks. A Transformer-based RGBD cross-modal fusion method is proposed for segmenting glass-like objects. The method utilizes a Transformer network that extracts self-attention features of RGB and depth through a cross-modal fusion module and integrates RGBD features using a multi-layer perceptron（MLP）mechanism to achieve the fusion of three types of attention features. RGB and depth features are fed back to their respective branches to enhance the network's feature extraction capabilities. Finally，a semantic segmentation decoder combines the features from four stages to output the segmentation results of glass-like objects. Compared with the EBLNet method，the intersection-and-union ratio of the proposed method on the GDD，Trans10k and MSD datasets is improved by 1.64%，2.26%，and 7.38%，respectively. Compared with the PDNet method on the RGBD−Mirror dataset，the intersection-and-union ratio is improved by 9.49%，verifying its effectiveness.

Key words：attention；semantic segmentation；glass-like object（GLO）；cross-modal；depth estimation

室内玻璃类似物（glass-like object，GLO）包括镜面类物体和玻璃类物体．镜面类物体是指具有镜面反射特性的物体，它们能够反射周围场景．玻璃类物体是指透明的玻璃物体，它们通过透射将背后的场景投影出来．由于玻璃类似物的存在，许多计算机视觉任务可能会失败，例如语义分割、深度估计、目标检测、机器人导航、3D场景重建、Lidar测量等[1−4]．GLO没有特定的形状和视觉纹理特征，很难直接利用现有的语义分割方法对其进行检测和分割．因此研究高精度的GLO

收稿日期：2023−07−17
基金项目：国家自然科学基金资助项目（62273081）；辽宁省基础研究计划项目（2022JH2/101300202）．
作者简介：万应才（1990—），男，甘肃靖远人，东北大学博士研究生；房立金（1965—），男，辽宁沈阳人，东北大学教授，博士生导师．

分割方法对计算机视觉具有重要的意义[5].

近年来研究人员结合深度学习提出一些玻璃类似物分割方法,例如Yang等[6]构建了一个大规模镜像分割数据集(mirror segmentation dataset,MSD),并利用注意力模块生成多层次的纹理对比特征进行分割.在MirrorNet的基础上,Lin等[7-8]采用了关系语境对比局部模块(relational contextual contrasted local module,RCCLM)提取和比较镜面与上下文特征之间的关系,并采用边缘检测和融合模块提取多尺度的镜像边缘特征.Mei等[9]扩展了玻璃对象数据集,丰富了玻璃场景,通过大场景上下文特征融合模块实现了鲁棒的玻璃检测.在此基础上,He等[10]提出了一种新方法,利用边缘预测来指导GLO的分割结果.该方法利用了细化差分模块(refinement difference module,RDM)生成精确的边缘,并采用高效的基于点的图卷积网络模块(point-based graph convolution module,PGM)进行全局边缘特征学习.此外,一些研究人员探索了利用深度相机获取的深度信息来增强GLO分割的方法,如Mei等[11]提出了一种考虑图像和深度信息的镜像检测方法.Chang等[12]基于全景透明物体数据集提出了一种大视野可变形上下文特征(large-view deformable context feature,LDCF)来获取全景玻璃图像的宽视场和扭曲边界.

综上所述,当前玻璃类似物分割研究主要集中在其上下文特征提取及边界信息提取两个方面.具体而言,玻璃类似物表面包括反射周围场景的镜像对象和透明玻璃中从后方投射周围场景.由于它的反射和透射特性,当前的玻璃类似物特征提取网络难以高效提取纹理上下文特征.此外,玻璃类似物的深度信息很难准确测量,但是观察深度估计结果可以发现在受到玻璃类似物的折射和反射时,镜面或者玻璃表面深度与周围边界深度相比会发生突变,这些玻璃类似物边界信息能够辅助分割网络对其边界进行定位.然而,现有方法在挖掘跨模态特征方面仍存在局限性,并且过于依赖深度传感器的深度图,限制了跨模态方法的应用范围.

针对上述问题,本文提出一种基于Transformer多层注意力机制的RGBD(red-green-blue-depth)跨模态融合玻璃类似物分割网络.该方法在玻璃类似物特征提取与边界提取两个方面都进行了改进.其中特征提取网络使用Transformer作为骨干网络分别对图像RGB(red-green-blue)与深度进行特征提取,并在每个Transformer解码层加入跨模态融合模块,然后网络解码器输入4层跨模态融合特征,输出玻璃类似物分割结果.特别地,本文通过跨模态融合模块引入深度信息增加网络感知空间信息的能力,以更加精确地定位玻璃类似物区域.实验结果表明,本文提出的方法在4个不同数据集上与其他先进方法相比均取得了领先.通过消融实验验证了本文提出的RGBD跨模态融合网络的有效性.

1 本文方法

1.1 网络结构

如图1所示,本文网络框架解码阶段分别由RGB与深度两个分支组成,每个分支分别包含4个阶段Transformer模块,并在不同阶段的Transformer模块之间加入特征融合模块,解码器融合不同阶段的融合特征,输出最终预测结果.图1左下方为Transformer模块之间的融合模块具体结构,上半部分表示RGB分支,下半部分表示深度图分支,两部分融合为RGBD分支.融合模块对RGB和深度分支的特征在空间与通道两个方向进行注意力操作,分别输出下一阶段RGB特征、下一阶段深度特征和融合特征.语义编码器结合各个阶段输出的融合特征,输出玻璃类似物分割预测结果.

1.2 玻璃类似物特征提取

由于玻璃类似物没有固定纹理信息,容易受到周围环境的影响,本文将视觉Transformer作为RGB与深度特征提取的骨干网络对玻璃类似物的各层级上下文信息特征进行提取.本文的Transformer结构采用Liu等[13]提出的Swin Transformer,它能通过自注意力机制实现全局信息的交互和依赖关系建模,具有处理任意大小图像和学习全局上下文信息的能力.此外,与其他融合深度传感器的方法不同,本文采用深度估计网络预测RGB图像对应的深度作为跨模态玻璃类似物检测的输入,其中深度估计网络为AdelaiDepth[14].

1.3 玻璃类似物跨模态融合模块

为了能够更好地利用RGB与深度模态的纹理、空间与结构等特征,本文设计了一种基于注意力机制的跨模态特征融合模块对玻璃类似物特征进行特征融合与特征交换.跨模态融合模块

分别加入到4个不同的RGB与深度特征提取阶段,如图1上半部分所示.每个跨模态融合模块分别包括RGB、深度和融合特征3个部分自注意力特征提取,最后输出3个部分自注意力特征的融合特征.其中特征融合模块的通道注意力与空间注意力具体结构如图2所示.

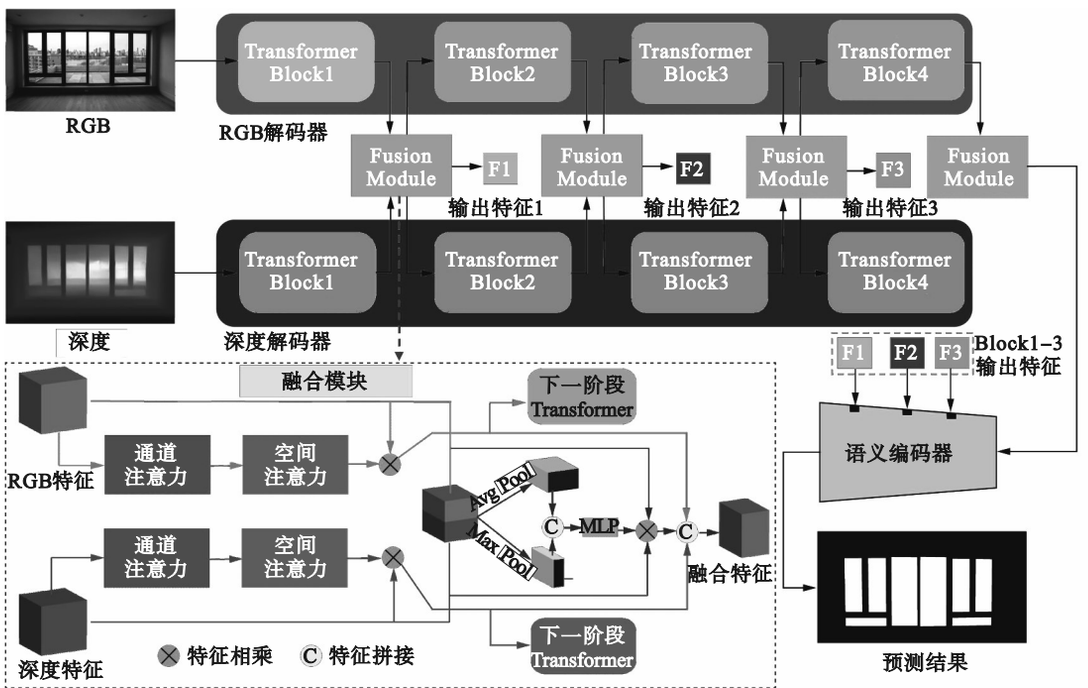


图1　RGBD跨模态融合玻璃类似物分割框图

Fig. 1　The framework of RGBD cross-modal fusion for glass-like object segmentation

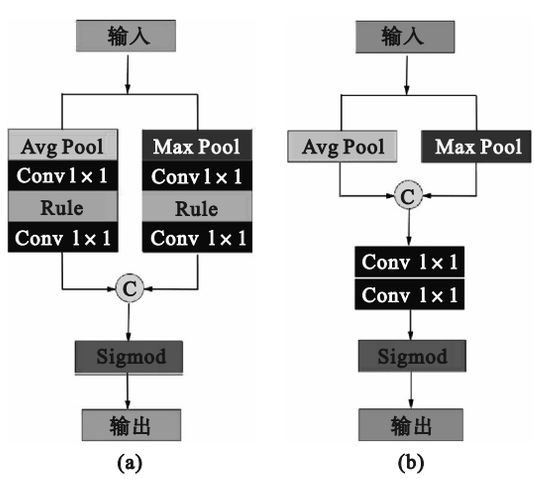


图2　玻璃类似物通道注意力与空间注意力特征提取结构

Fig. 2　Structure diagram of channel attention and space attention feature extraction of GLO

(a)—通道注意力;(b)—空间注意力.

1.3.1　RGB与深度通道方向注意力特征提取

RGB与深度通道方向注意力特征提取用于在融合模块中自适应地学习每个通道的重要性权重,以更好地提取特征.该通道注意力可以通过对每个通道进行加权平均来实现特征的加权融合,从而提高模型的性能和准确率.该通道注意力采用并行提取结构,对输入RGB与深度特征 F_{in}^{c} 分别进行通道注意力提取,其中平均池化的注意力 $F_{\text{avg}}^{\text{c}}$ 为

$$F_{\text{avg}}^{\text{c}}=C\left(R\left(C\left(O_{\text{p,avg}}\left(F_{\text{in}}^{\text{c}}\right)\right)\right)\right). \tag{1}$$

式中:$O_{\text{p,avg}}$ 为平均池化操作;C 为卷积核为1的卷积操作;R 为激活函数ReLU.

最大池化的注意力 $F_{\text{max}}^{\text{c}}$ 为

$$F_{\text{max}}^{\text{c}}=C\left(R\left(C\left(O_{\text{p,max}}\left(F_{\text{in}}^{\text{c}}\right)\right)\right)\right). \tag{2}$$

式中,$O_{\text{p,max}}$ 为最大池化操作.

然后合并两种不同的注意力操作并经Sigmoid激活函数,得到输出通道注意力特征 $F_{\text{out}}^{\text{c}}$:

$$F_{\text{out}}^{\text{c}}=S\left(f\left(F_{\text{avg}}^{\text{c}},F_{\text{max}}^{\text{c}}\right)\right). \tag{3}$$

式中:$f(\cdot)$ 为特征融合操作;$S(\cdot)$ 为Sigmoid激活函数.

1.3.2　RGB与深度空间方向注意力特征提取

RGB与深度空间方向注意力特征提取用于融合模块中自适应地学习每个空间位置权重,以提取玻璃类似物的空间特征.与1.3.1节中通道注意力类似,本文空间注意力也采用并行结构.输入特征 F_{in}^{s} 分别经过平均池化和最大池化操作得

到融合特征F_{AM}:

$$F_{\mathrm{AM}}=f\left(O_{\mathrm{p,avg}}\left(F_{\mathrm{in}}^{\mathrm{s}}\right),O_{\mathrm{p,max}}\left(F_{\mathrm{in}}^{\mathrm{s}}\right)\right). \tag{4}$$

F_{AM}经过两层卷积操作得到输出空间注意力特征$F_{\mathrm{out}}^{\mathrm{s}}$:

$$F_{\mathrm{out}}^{\mathrm{s}}=C\left(C\left(F_{\mathrm{AM}}\right)\right). \tag{5}$$

1.3.3 RGB与深度融合特征提取

在分别提取RGB与深度融合特征提取之后,本文采用多层感知机(multi-layer perceptron,MLP)对RGB与深度融合的特征进行特征提取.MLP是一种注意力机制,它使用多层感知机来计算每个位置的权重,以更好地提取特征[15].具体而言,将输入的RGB与深度融合为RGBD的特征向量,通过MLP进行注意力特征提取,然后使用Softmax函数将变换后的向量转换为概率分布,最后将概率分布与输入RGBD特征进行加权平均,得到注意力加权后的融合RGBD特征.具体过程如下:RGB与深度拼接的特征F_{RGBD}经过多层感知机进行特征提取得到特征$F_{\mathrm{RGBD}}^{\mathrm{MLP}}$:

$$F_{\mathrm{RGBD}}^{\mathrm{MLP}}=f_{\mathrm{MLP}}\left(F_{\mathrm{RGBD}}\right). \tag{6}$$

式中,$f_{\mathrm{MLP}}(\cdot)$为MLP操作.

1.3.4 跨模态特征融合

利用本文的跨模态注意力融合模块对每个阶段的输入RGB特征F_{RGB}和深度特征F_{d}分别进行通道注意力、空间注意力和MLP特征提取得到RGB特征F_{RGB}、深度特征F_{d}和MLP特征$F_{\mathrm{RGBD}}^{\mathrm{MLP}}$.其中$F_{\mathrm{RGB}}$和$F_{\mathrm{d}}$分别作为下一阶段RGB和深度Transformer的输入,然后与$F_{\mathrm{RGBD}}^{\mathrm{MLP}}$组成新的跨模态融合特征$F_{\mathrm{fusion}}$:

$$F_{\mathrm{fusion}}=F_{\mathrm{RGB}}+F_{\mathrm{d}}+F_{\mathrm{RGBD}}^{\mathrm{MLP}}. \tag{7}$$

1.4 损失函数

本文采用交叉熵二值(binary cross-entropy,BCE)损失函数作为监督训练损失函数.BCE损失函数是一种用于二分类问题的常见损失函数,其基本思想是将模型的输出(通常是一个概率值)与真实标签进行比较,从而计算模型预测错误的程度.训练数据中真实标签为G_{t},解码器输出分割预测结果为P_{s},则输出损失为

$$L_{\mathrm{BCE}}\left(P_{\mathrm{s}},G_{\mathrm{t}}\right)=-\frac{1}{N}\sum_{i=1}^{N}[G_{\mathrm{t},i}\ln\left(P_{\mathrm{s},i}\right)+(1-G_{\mathrm{t},i})\ln(1-P_{\mathrm{s},i})]. \tag{8}$$

式中:L_{BCE}为BCE损失函数的输出损失;N表示样本数量;$G_{\mathrm{t},i}$为第i个样本的真实标签(0或1);$P_{\mathrm{s},i}$为模型对第i个样本预测为正类(1)的概率.

2 实验结果与分析

2.1 实现细节

本文方法在PyTorch框架上实现,并使用随机梯度下降(stochastic gradient descent,SGD)优化网络,动量为0.9,权重衰减为5×10^{-4}.本文实验将批次大小设置为14.在NVIDIA RTX 3090Ti显卡、8核AMD 5800X 3.8 GHz CPU和64 GB RAM上进行200轮训练.在训练和测试时,将输入图像的大小调整为416像素×416像素,并通过随机水平翻转进行增强.

2.2 数据集和评估指标

2.2.1 数据集

1)GDD数据集[9]为镜子分割任务创建的,该数据集是从室内外场景中选取的大量类似玻璃的数据集,其中包含3 916张镜子图像和标签.在实验过程中,将数据集分为2 980个训练图像和936个测试图像.

2)RGBD-Mirror数据集[11]是第一个RGB-D镜面数据集,其中包含3 094张RGB图像和深度图.RGBD-Mirror是一个综合性的数据集,它从4个流行的室内数据集(Matterport 3D,SUNRGBD,ScanNet和2D3DS)中选择包含镜子的图像、对应的深度图及镜面的真实标签.实验中,选取2 000个进行训练,其余1 046个用于测试.

3)MSD[6]是第一个可用的大型镜面数据集,包括4 018张图像及其对应的标签.在数据集的分割中,将3 063张图片用于训练,其余955张图片用于测试.

4)Trans10k是一个大规模的透明物体数据集,包含10 428张图像和对应的深度图.Trans10k包括两种类别的透明物体,即房间图像和物体.在本文的实验中,使用5 000,1 000和4 428个图像与标签组成的图像对进行训练、验证和测试.

2.2.2 评估指标

本文采用平均交并比R_{mIoU}和F_{β}评估分割性能,其中R_{mIoU}衡量预测与真实分割的重叠程度,F_{β}结合精确度和召回率进行评估.此外,使用MAE(mean absolute error)和平衡误差率R_{be}评估预测结果与真实标签的差异,使用模型区分GLO与非GLO区域的平衡性.

2.3 玻璃类似物分割实验结果

本实验选择了包含玻璃类和镜子类物体的

4个数据集进行实验,并对实验结果进行定性与定量分析.

2.3.1　GDD和Trans10k数据集结果分析

本文方法在GDD和Trans10k数据集上与其他分割方法对比.在表1中,其他方法包括语义分割方法ICNet[1],DeepLabv3+[16];显著目标检测方法MINet-R[17],ITSD[18];玻璃类似物分割方法MirrorNet[6],TransLab[19],GDNet[9],PGSNet[8],GSD[7]和EBLNet[10].由表1可知,本文方法在GDD数据集[9]和Trans10k数据集[19]上,R_{mIoU}指标分别达到了89.61%和92.32%,领先EBLNet方法1.64%,2.26%.本文方法的MAE和R_{be}均取得了领先,证明了本文方法的有效性.

表1　在GDD和Trans10k数据集上与其他方法进行定量比较

Table 1　Quantitative comparison with other methods on the GDD and Trans10k datasets.

方法	骨干网络	GDD				Trans10k			
		R_{mIoU}/%	F_β	MAE	R_{be}/%	R_{mIoU}/%	F_β	MAE	R_{be}/%
ICNet[1]	ResNet-50	69.59	0.747	0.164	16.10	74.94	0.784	0.110	10.92
DeepLabv3+[16]	ResNet-50	69.95	0.767	0.147	15.49	51.52	0.602	0.229	23.80
MINet-R[17]	ResNet-50	82.03	0.847	0.092	8.55	85.88	0.881	0.060	6.03
ITSD[18]	ResNet-50	83.72	0.862	0.087	7.77	85.44	0.871	0.063	6.26
MirrorNet[6]	ResNeXt-101	85.07	0.866	0.083	7.67	88.30	0.907	0.047	4.95
TransLab[19]	ResNet-50	81.64	0.849	0.097	9.70	87.10	0.897	0.051	5.44
GDNet[9]	ResNeXt-101	87.63	0.898	0.063	5.62	88.68	0.907	0.046	4.72
GSD[7]	ResNeXt-101	88.07	0.932	0.059	5.71	89.16	0.937	0.043	4.50
PGSNet[8]	ResNeXt-101	87.81	0.901	0.062	5.56	89.79	0.917	0.042	4.39
EBLNet[10]	ResNeXt-101	88.16	0.939	0.059	5.58	90.28	0.947	0.048	4.14
本文方法	Swin-s	89.61	0.942	0.060	5.02	92.32	0.949	0.035	2.98

2.3.2　MSD数据集结果分析

在表2中,本文方法与ICNet[1],DeepLabv3+[16],MirrorNet[6],EBLNet[10]对比.结果显示本文方法表现较好.本文方法与EBLNet相比,R_{mIoU}与F_β分别提高了7.38%和2.94%,而MAE和R_{be}分别下降了8.16%和6.95%,在多个关键性能指标上显著优于EBLNet方法,显示出更准确和更稳定的性能.

表2　在MSD数据集上与其他方法进行定量比较

Table 2　Quantitative comparison with other methods on the MSD dataset

方法	R_{mIoU}/%	F_β	MAE	R_{be}/%
ICNet[1]	57.25	0.710	0.124	18.75
DeepLabv3+[16]	78.81	0.872	0.054	8.95
MirrorNet[6]	78.95	0.857	0.065	6.39
EBLNet[10]	80.33	0.883	0.049	8.63
本文方法	86.26	0.909	0.045	8.03

2.3.3　RGBD-Mirror数据集结果分析

由表3可知,本文方法在所有指标上都超过了其他方法.在R_{mIoU}指标上,本文方法达到了85.15%,而最接近的方法PDNet[11]为77.77%.

在表3中,与PDNet方法对比,本文方法的R_{mIoU}和F_β分别提高了9.49%和11.76%,MAE和R_{be}分别降低了11.90%和21.10%.由此可知,本文方法对于镜面物体分割任务具有更高的准确性,进一步验证了本文方法的有效性.

表3　在RGBD-Mirror数据集上与其他方法进行定量比较

Table 3　Quantitative comparison with other methods on the RGBD-Mirror dataset

方法	R_{mIoU}/%	F_β	MAE	R_{be}/%
F3Net[20]	65.15	0.707	0.069	14.25
MirrorNet[6]	68.37	0.723	0.062	8.66
PMD[8]	72.27	0.775	0.054	10.71
PDNet[11]	77.77	0.825	0.042	7.77
本文方法	85.15	0.922	0.037	6.13

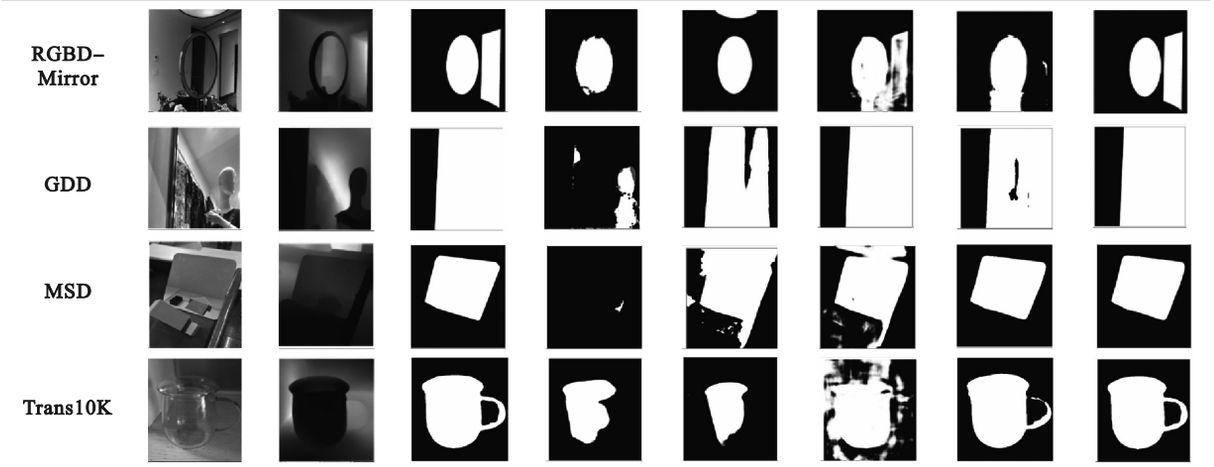
2.4　定性分析

从表4可知,本文方法在4个数据集上都能够获得令人满意的分割结果,证明该方法适用于不同类型的玻璃类似物的分割.

此外,本文方法在小物体和细节方面也表现出较好的分割结果.例如,在数据集Trans10k对透明玻璃杯子进行分割时,本文方法可以准确地分割出杯子把手.通过定性分析,验证了本文方法对于玻璃类似物的分割能力.

表4 本文方法与其他方法在RGBD-Mirror,GDD,MSD和Trans10k数据集结果对比

Table 4 Comparison with other methods on RGBD-Mirror,GDD,MSD,and Trans10k datasets



2.5 消融实验

2.5.1 深度影响

在RGBD跨模态学习中,本文选择网络预测深度作为跨模态深度输入.为了验证本文用网络估计深度替代RGBD相机采集的有效性,在带有深度相机采集深度数据集RGBD-Mirror上,分别对比相机采集深度和网络估计深度对于玻璃类似物分割的影响.

由表5可知,网络估计深度的分割结果明显高于使用相机采集深度的结果.本文选取的网络估计深度图相比相机采集深度滤除了深度噪声,网络估计深度图的边缘锐利,深度值相对较平滑,具有较高质量的深度图.由此可知,本文采用网络估计深度既提高了玻璃类似物分割的准确性又提高了方法的通用性,使其不依赖于深度传感器.

表5 不同类型深度消融实验结果

Table 5 Experimental results of different types of depth ablation

方法	R_{mIoU}/%	F_β	MAE	R_{be}/%
PDNet[11](相机采集深度)	77.77	0.825	0.042	7.77
PDNet(网络估计深度)	78.58	0.849	0.041	7.01
本文方法(相机采集深度)	84.15	0.908	0.042	6.50
本文方法(网络估计深度)	85.15	0.922	0.037	6.13

2.5.2 注意力模块及骨干网络的影响

为了验证本文提出基于Transformer的跨模态融合玻璃类似物分割方法的有效性.本文对骨干网络ResNet和Swin-s[11]及融合模块进行消融实验,如表6所示.

相比ResNet网络,本文采用了基于Transformer的Swin-s结构作为骨干网络,明显提升了玻璃类似物的分割精度,在此基础上特征融合模块进一步提升了平均交并比R_{mIoU},验证了特征融合模块的有效性.

表6 不同骨干网络及融合模块的平均交并比指标结果

Table 6 Mean intersection-over-union ratio results of different backbone networks and fusion modules %

ResNet101	Swin-s	ResNet101+融合模块	Swin-s+融合模块
79.56	86.02	83.32	89.61

在图3所示的注意力分布可视化图中,“RGB输入”表示只有RGB分支,“RGB+深度”表示直接融合RGB与深度作为网络输入,“RGB+深度+融合模块”表示跨模态融合模块融合RGB与深度.如图3所示,当只使用RGB分支时注意力无法集中在扶梯玻璃区域上,在引入深度后,注意力的分布得到了显著改善.特别地,跨模态融合模块融合RGB与深度,网络的注意力分布进一步提升,网络注意力效率显著提高,证明本文所提出的跨模态注意力融合机制是有效的.

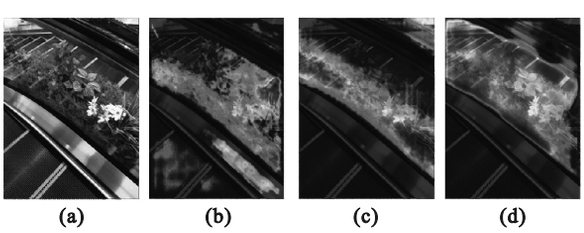


图3 注意力分布可视化图

Fig. 3 Visualization of attention distribution

(a)—原图像;(b)—RGB输入;(c)—RGB+深度;(d)—RGB+深度+融合模块.

2.5.3　跨模态融合模块消融实验

本文在Transformer解码阶段采用多阶段融合策略,为了研究不同融合阶段对分割结果的影响,实验过程中设计了不同融合阶段及其组合的消融实验.表7的第2至第5行展示单个融合模块被集成至Transformer编码第1至第4阶段,而第6至第8行则呈现多个融合模块在多个编码阶段的组合加入.实验结果如表7所示,在不同融合阶段下,R_{mIoU}随着融合阶段增加逐步提高,结果表明在4个Transformer解码阶段均加入融合模块能够最大限度提升分割结果.

表7　不同融合阶段平均交并比指标对比

Table7　Results of mean intersection-over-union ratio at different fusion stages

第1阶段	第2阶段	第3阶段	第4阶段	R_{mIoU}
				87.71
✓				87.31
	✓			87.29
		✓		87.89
			✓	88.69
✓	✓			89.48
✓	✓	✓		89.61
✓	✓	✓	✓	

注:"✓"代表加入融合模块.

2.5.4　真实场景玻璃类似物测试

为了验证本文提出方法的通用性,收集了日常环境中手机拍摄的含有玻璃窗户、玻璃门、眼镜和镜子的玻璃类似物数据,其中包括镜面和玻璃类物体.由图4可知,在实际场景中测试表明,本文方法能够准确地分割出玻璃类似物区域,从而验证了本文方法在实际应用中的有效性.特别是在周围复杂反射和干扰的场景,本文方法依然能从干扰物中分割出镜子.

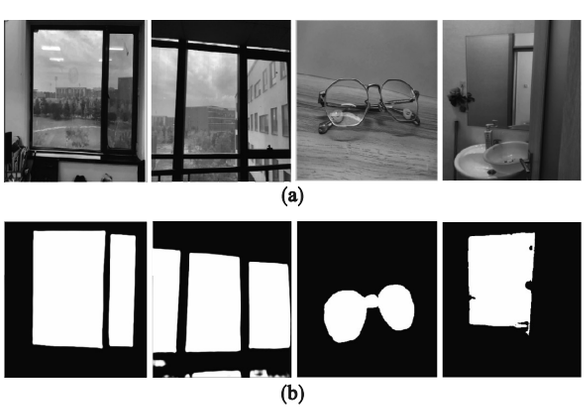


图4　真实场景玻璃类似物实验结果

Fig. 4　Results of glass-like objects in real scenes

(a)—实物;(b)—分割结果.

2.6　玻璃区域深度恢复

受到玻璃类似物透射或反射的影响,玻璃类似物区域的深度无法得到真实值,从而影响了3D重建和机器人导航.

玻璃类似物区域通常为平面且其表面深度与周围边界的深度相似,而本文方法能够分割得到玻璃类似物区和边界.为了进一步验证本文所提出的融合深度估计镜面分割的优势,设计了一种深度恢复方法,将边界10个像素宽的扩展区域的平均深度作为玻璃类似物区域的深度.

图5展示了如何利用玻璃类似物分割恢复镜面深度.如图5所示,首先,使用边缘来定位需要恢复的深度区域,然后使用边界(图5e)外的10个像素区域的平均深度作为玻璃区域深度的基准,调整玻璃区域的深度.调整后的深度图如图5f所示,借助镜面周围边界的深度,将镜面深度估计结果修正为边界附近深度,使其接近镜面真实物理深度.这种深度恢复的方法能够对镜面或玻璃平面区域的深度进行整体估计,为机器人导航、2D或3D语义分割和3D重建等提供深度参考.

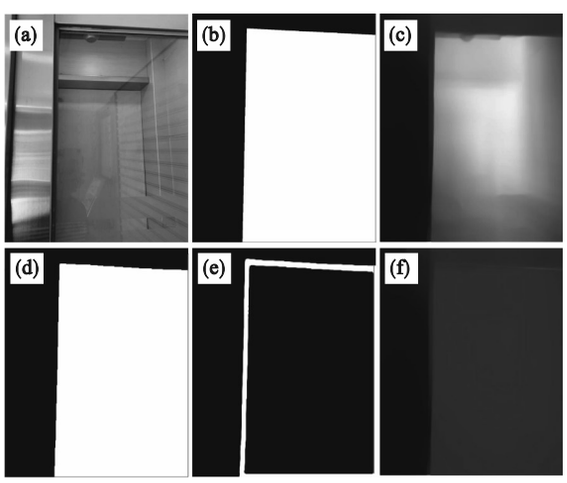


图5　深度恢复过程

Fig. 5　The process of depth recovery

(a)—输入图像;(b)—真实标签;(c)—深度图;(d)—玻璃区域分割结果;(e)—玻璃边界估计结果;(f)—调整后深度图.

3　结　　语

针对玻璃类似物分割任务的挑战,本文提出了一个基于Transformer的RGBD跨模态融合分割方法.该方法结合了2个Transformer分支,通过1个跨模态融合模块来整合RGB和深度信息,并利用空间、通道及多层注意力机制优化特征提

取,增强对玻璃类似物纹理及深度空间结构的识别能力.实验结果表明,本文方法与EBLNet方法相比,在GDD,Trans10k和MSD数据集上交并比分别提高1.64%,2.26%,7.38%,与PDNet方法比较在RGBD-Mirror数据集上交并比提高了9.49%.消融实验进一步验证了本文方法对玻璃类似物区域的识别能力及设计的合理性;相比传统深度传感器,使用深度估计网络生成的深度图更为有效.未来工作计划将此技术应用于机器人导航、语义分割及3D重建等领域,以提高任务的精度和通用性.

参考文献:

[1] Zhao H S, Qi X J, Shen X Y, et al. ICNet for real-time semantic segmentation on high-resolution images [C]// Proceedings of the European Conference on Computer Vision (ECCV 2018). Munich: Springer International Publishing, 2018: 418-434.

[2] Wang D Q, Zhang T, Süsstrunk S. NEMTO: neural environment matting for novel view and relighting synthesis of transparent objects [C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE, 2023: 317-327.

[3] 王璐,王帅,张国峰,等.基于语义分割注意力与可见区域预测的行人检测方法[J].东北大学学报(自然科学版),2021,42(9):1261-1267.
(Wang Lu, Wang Shuai, Zhang Guo-feng, et al. Pedestrian detection based on semantic segmentation attention and visible region prediction [J]. *Journal of Northeastern University* (*Natural Science*), 2021, 42(9): 1261-1267.)

[4] 张之敏,乔建忠,林树宽,等.一种基于深度网络的视图重建方法[J].东北大学学报(自然科学版),2020,41(8):1065-1069.
(Zhang Zhi-min, Qiao Jian-zhong, Lin Shu-kuan, et al. A view reconstruction method based on deep network [J]. *Journal of Northeastern University* (*Natural Science*), 2020, 41(8): 1065-1069.)

[5] Wang Z Y, Li Y C, Cheng X N, et al. Key points trajectory and multi-level depth distinction based refinement for video mirror and glass segmentation [J]. *Multimedia Tools and Applications*, 2024, 83(39): 86513-86535.

[6] Yang X, Mei H Y, Xu K, et al. Where is my mirror? [C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 8808-8817.

[7] Lin J Y, He Z B, Lau R W H. Rich context aggregation with reflection prior for glass surface detection [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 13410-13419.

[8] Lin J Y, Wang G D, Lau R W H. Progressive mirror detection [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 3694-3702.

[9] Mei H Y, Yang X, Wang Y, et al. Don't hit me! glass detection in real-world scenes [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 3684-3693.

[10] He H, Li X T, Cheng G L, et al. Enhanced boundary learning for glass-like object segmentation [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021: 15839-15848.

[11] Mei H Y, Dong B, Dong W, et al. Depth-aware mirror segmentation [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 3043-3052.

[12] Chang Q L, Liao H H, Meng X F, et al. PanoGlassNet: glass detection with panoramic RGB and intensity images [J]. *IEEE Transactions on Instrumentation and Measurement*, 2024, 73: 5019015.

[13] 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). Montreal: IEEE, 2021: 9992-10002.

[14] Yin W, Zhang J M, Wang O, et al. Learning to recover 3D scene shape from a single image [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 204-213.

[15] Taud H, Mas J F. Multilayer perceptron (MLP) [M]// Cámacho O M T, Paegelow M, Mas J F, et al. Geomatic Approaches for Modeling Land Change Scenarios. Cham: Springer, 2018: 451-455.

[16] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network [C]//2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 6230-6239.

[17] Deng J J, Pan Y W, Yao T, et al. MINet: meta-learning instance identifiers for video object detection [J]. *IEEE Transactions on Image Processing*, 2021, 30: 6879-6891.

[18] Zhou H J, Xie X H, Lai J H, et al. Interactive two-stream decoder for accurate and fast saliency detection [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 9138-9147.

[19] Xie E Z, Wang W J, Wang W H, et al. Segmenting transparent objects in the wild [C]//Computer Vision and Pattern Recognition. Cham: Springer International Publishing, 2020: 696-711.

[20] Wei J, Wang S H, Huang Q M. F3Net: fusion, feedback and focus for salient object detection [C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: IEEE, 2020: 12321-12328.