(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116206133 B
(45)授权公告日 2023.09.05

(21)申请号 202310450234.5

(22)申请日 2023.04.25

(65)同一申请的已公布的文献号
申请公布号 CN 116206133 A

(43)申请公布日 2023.06.02

(73)专利权人 山东科技大学
地址 266590 山东省青岛市黄岛区前湾港路579号

(72)发明人 东野长磊 贾兴朝 赵文秀 彭延军

(74)专利代理机构 青岛智地领创专利代理有限公司 37252
专利代理师 王鸣鹤

(51)Int.Cl.
G06V 10/46(2022.01)
G06V 10/764(2022.01)
G06V 10/80(2022.01)
G06V 10/82(2022.01)
G06V 10/56(2022.01)
G06N 3/0455(2023.01)
G06N 3/0464(2023.01)
G06N 3/09(2023.01)

(56)对比文件
CN 113763422 A,2021.12.07
CN 115908789 A,2023.04.04
WO 2022166361 A1,2022.08.11
WO 2021088300 A1,2021.05.14
CN 113486865 A,2021.10.08
CN 111582316 A,2020.08.25
US 2019147318 A1,2019.05.16
US 2012113133 A1,2012.05.10
CN 115410046 A,2022.11.29
CN 113935433 A,2022.01.14
Nian Liu 等.Visual Saliency Transformer.《2021 IEEE/CVF International Conference on Computer Vision (ICCV)》.2022,第2022年卷摘要、第3节、图1.

审查员 杜学惠

权利要求书2页 说明书7页 附图3页

(54)发明名称

一种RGB-D显著性目标检测方法

(57)摘要

本发明提供了一种RGB-D显著性目标检测方法,涉及图像处理领域,具体包括如下步骤:输入RGB图像和深度图像;对RGB图像和深度图像进行特征提取,分别获取RGB图像和深度图像不同层级的显著性特征;融合深层次RGB特征和深度特征之间的互补语义信息,生成跨模态联合特征;利用密集连接增强的跨模态密集协作聚合模块实现两种不同模态的特征融合,逐级融合不同尺度上的深度特征和RGB特征,输入到显著性目标检测部分;按照预测的显著性图分辨率由小到大排序,并利用真值图对网络进行有监督学习,输出最终的显著性检测结果。本发明克服了现有技术中不能对跨模态特征进行有效融合,显著性目标检测精度不高的问题。

CN 116206133 B

1.一种RGB-D显著性目标检测方法,其特征在于,具体包括如下步骤:

S1,输入RGB图像和深度图像;

S2,利用基于T2T-ViT的transformer编码器对RGB图像进行特征提取,利用基于轻量级卷积网络MobileNet V2的编码器对深度图像进行特征提取,分别获取RGB图像和深度图像不同层级的显著性特征;

轻量级卷积网络MobileNet V2的编码器包括:MobileNet V2模块一、MobileNet V2模块二、MobileNet V2模块三和MobileNet V2模块四;

步骤S2的基于T2T-ViT的transformer编码器中的T2T操作包括:重组和软拆分,所述重组是将一个token序列$Tok \in \mathbb{R}^{l \times c}$重建成一个3D张量$I \in \mathbb{R}^{h \times w \times c}$,其中,l是token序列Tok的长度,c是token序列Tok和3D张量I的通道数,h,w分别代表I的高度和宽度,且l=h×w;

所述软拆分是通过展开操作将I软分割成k×k大小的块,$I \in R^{h \times w \times c}$经所述软拆分后得到token序列,其长度$l_0$可以表示为:

$$l_0 = \left\lfloor \frac{h+2p-k}{k-s}+1 \right\rfloor \times \left\lfloor \frac{w+2p-k}{k-s}+1 \right\rfloor,$$

其中,s代表块之间像素重叠个数,p代表块之间像素填充个数,k-s代表卷积运算中的步长,当s<k-1时,tokens序列的长度便可降低;

原始RGB图像$I_{input} \in \mathbb{R}^{H \times W \times C}$,其中H,W,C分别代表$I_{input}$的高度、宽度和通道数,经过重组得到的token序列$T \in \mathbb{R}^{h \times w \times c}$经过三轮Transformer转换和两轮T2T操作得到了多级tokens序列T′,T_1,T'_1,T_2,T'_2,这一过程可以表示为:

T′=Transformer(T),

T_1=Unfold(Reshape(T′)),

T_1'=Transformer(T_1),

T_2=Unfold(Reshape(T_1')),

T_2'=Transformer(T_2)

S3,利用基于跨模态Transformer融合模块,融合深层次RGB特征和深度特征之间的互补语义信息,生成跨模态联合特征;

步骤S3中的跨模态Transformer融合模块CMTFM包括:跨模态交互注意力模块和Transformer层,所述跨模态交互注意力模块,用于对RGB图和深度图之间的远程跨模态依赖进行建模,整合RGB数据和深度数据之间的互补信息;来自RGB块标记的信息流T'_2和深度块标记的信息流C_4通过4次跨模态交互注意力模块来进行跨模态信息交互后,再经过一个4层Transformer层进行强化得到了token序列T_3;

S4,利用密集连接增强的跨模态密集协作聚合模块实现两种不同模态的特征融合,逐级融合不同尺度上的深度特征和RGB特征,输入到显著性目标检测部分;

步骤S4中,来自T2T-ViT的经重组后的RGB信息T′,T_1',T_3和来自MobileNet V2的深度信息C_1,C_2,C_3,C_4被输入到密集连接增强后的解码器,密集连接用于将不同尺度上的深度特征和RGB特征进行融合;其中MobileNet V2模块一、MobileNet V2模块二、MobileNet V2模块三和MobileNet V2模块四分别输出深度信息C_1、C_2、C_3、C_4;

S5,按照预测的显著图分辨率由小到大排序,并利用真值图对网络进行有监督学习,输出最终的显著性检测结果;

步骤S4的跨模态密集协作聚合模块包括:三个特征聚合模块和一个双重倒残差模块,跨模态密集协作聚合模块用于将低分辨率编码器特征扩展到与输入图像分辨率大小一致,所述特征聚合模块用于聚合特征和融合跨模态信息;

所述特征聚合模块包括:一个CBAM和两个倒残差结构IRB,还包含了两个元素相乘和一个元素相加操作;基于特征聚合模块的特征聚合和融合跨模态信息过程包括如下步骤:

S4.1,RGB特征T_R和深度特征T_D进行相乘,再经过一个IRB进行卷积后得到过渡的RGB-D特征图D,此过程表示为:

$D=IRB(T_R\times T_D)$;

其中,T_R包括:T′、T_1′和T_3,T_D包括:C_2、C_3、C_4;

S4.2,深度特征T_D经CBAM增强后的特征记为T''_D,此过程表示为:

$T'_D=Channel(T_D)\times T_D$;

$T''_D=Spatial(T'_D)\times T'_D$

S4.3,D再与深度特征T''_D再次相乘强化语义特征后得到D′,此过程表示为:

$D'=D\times T''_D$;

S4.4,D′与RGB特征T_R相加以重新增强显著特征,同时引入较低层次的输出特征T_{DC}进行元素相加,然后使用IRB,得到跨模态融合后的RGB-D特征D″,此过程表示为:

$T_R'=T_R+D'$

$D''=IRB(T_R'+T_{DC})$。

2.根据权利要求1所述的一种RGB-D显著性目标检测方法,其特征在于,步骤S2中,基于轻量级卷积网络MobileNet V2的编码器包括倒残差结构。

3.根据权利要求2所述的一种RGB-D显著性目标检测方法,其特征在于,经跨模态Transformer融合模块,得到跨模态交互信息的公式,表示为:

$$Attention(Q_R,K_D,V_D)=softmax(Q_RK_D^T/\sqrt{d_k})V_D$$
$$Attention(Q_D,K_R,V_R)=softmax(Q_DK_R^T/\sqrt{d_k})V_R,$$

其中,Q_R、Q_D分别为两种模态的查询,K_R、K_D分别为两种模态的键,V_R、V_D分别为两种模态的值。

4.根据权利要求1所述的一种RGB-D显著性目标检测方法,其特征在于,步骤S5中,预测的显著图由调整相应大小后的真值图进行监督,将这一阶段产生的四个损失表示为$L_d^i, i=1,2,3,4$,总的损失函数L_{total}计算公式如下:

$$L_{total}=\sum_{i=1}^{4}\lambda_i BCE(P_i,G_i),$$

其中,λ_i表示每个损失的权重,按照分辨率由小到大的顺序将四个显著性预测图依次记为P_i(i=1,2,3,4),G_i表示来自真值图的监督,其分辨率与P_i对应,BCE()表示交叉熵损失函数。

一种RGB-D显著性目标检测方法

技术领域

[0001] 本发明涉及图像处理领域,具体涉及一种RGB-D显著性目标检测方法。

背景技术

[0002] 在视觉场景中,人类能够快速地将注意力转移到最重要区域。计算机视觉中的显著性目标检测便是由计算机模拟人眼视觉来识别场景中最显著目标,显著目标检测作为计算机视觉应用中重要的预处理任务,已广泛应用于图像理解、图像检索、语义分割、图像修复和物体识别中。随着Kinect和RealSense等深度相机的发展,各种场景的深度图的获取变得更加容易,深度信息可以和RGB图像进行信息互补,有利于提高显著性检测的能力。因此,基于RGB-D的显著性目标检测得到了研究人员的关注。

[0003] 传统的RGB-D显著性目标检测方法通过手工特征提取,然后融合RGB图像和深度图。例如,Lang等人利用高斯混合模型来模拟深度诱导的显著性的分布。Ciptadi等人从深度测量中提取了三维布局和形状特征,利用不同区域之间的深度差异来测量深度对比度。尽管传统RGB-D检测方法很有效,但所提取的低级特征限制了模型的泛化能力,而且不适用于复杂场景。

[0004] 显著性目标检测的一个需求是有效融合跨模态信息,在对RGB图和RGB-D图进行编码后,还需要将学习到的两种模态特征融合起来。基于卷积神经网络(CNN)的显著性目标检测方法取得了许多令人印象深刻的结果。现有基于卷积神经网络的显著性检测方法,存在卷积感受野的限制,在学习全局远程依赖方面存在严重不足。其次,现有技术采用的早期或者后期融合策略,难以捕获RGB和深度图像之间的互补和交互作用信息。不能从两种模态中学习高层次的信息,挖掘出集成融合规则,从而不能有效地检测完整的显著性目标。

[0005] 因此,现需要一种能够对跨模态特征进行有效融合,有效提高显著性目标检测精度的方法。

发明内容

[0006] 本发明的主要目的在于提供一种RGB-D显著性目标侧方法,以解决现有技术中不能对跨模态特征进行有效融合,显著性目标检测精度不高的问题。

[0007] 为实现上述目的,本发明提供了一种RGB-D显著性目标检测方法,具体包括如下步骤:S1,输入RGB图像和深度图像;S2,利用基于T2T-ViT的transformer编码器对RGB图像进行特征提取,利用基于轻量级卷积网络MobileNet V2的编码器对深度图像进行特征提取,分别获取RGB图像和深度图像不同层级的显著性特征;S3,利用基于跨模态Transformer融合模块,融合深层次RGB特征和深度特征之间的互补语义信息,生成跨模态联合特征;S4,利用密集连接增强的跨模态密集协作聚合模块实现两种不同模态的特征融合,逐级融合不同尺度上的深度特征和RGB特征,输入到显著性目标检测部分;S5,按照预测的显著性图分辨率由小到大排序,并利用真值图对网络进行有监督学习,输出最终的显著性检测结果。

[0008] 进一步地,步骤S2的基于T2T-ViT的transformer编码器中的T2T操作包括:重组和

软拆分，重组是将一个token序列$Tok \in \mathbb{R}^{l \times c}$重建成一个3D张量$I \in \mathbb{R}^{h \times w \times c}$，其中，l是token序列Tok的长度，c是token序列Tok和3D张量I的通道数，h，w分别代表I的高度和宽度，且l＝h×w；

[0009] 软拆分是通过展开操作将I软分割成k×k大小的块，$I \in \mathbb{R}^{h \times w \times c}$经软拆分后得到token序列，其长度$l_0$可以表示为：

[0010] $$l_0 = \left\lfloor \frac{h+2p-k}{k-s}+1 \right\rfloor \times \left\lfloor \frac{w+2p-k}{k-s}+1 \right\rfloor$$

[0011] 其中，S代表块之间像素重叠个数，p代表块之间像素填充个数，k-S代表卷积运算中的步长，当s＜k-1时，tokens序列的长度便可降低。

[0012] 原始RGB图像$I_{input} \in \mathbb{R}^{H \times W \times C}$，其中H，W，C分别代表$I_{input}$的高度、宽度和通道数，经过重组得到的token序列$T \in \mathbb{R}^{h \times w \times c}$经过三轮Transformer转换和两轮T2T操作得到了多级tokens序列T′，T_1，T_1'，T_2，T'_2，这一过程可以表示为：

[0013] T′＝Transformer(T)，

[0014] T_1＝Unfold(Reshape(T′))，

[0015] T_1'＝Transformer(T_1)，

[0016] T_2＝Unfold(Reshape(T_1'))，

[0017] T'_2＝Transformer(T_2)。

[0018] 进一步地，步骤S2中，基于轻量级卷积网络MobileNet V2的编码器包括倒残差结构。

[0019] 进一步地，步骤S3中的跨模态Transformer融合模块CMTFM包括：跨模态交互注意力模块和Transformer层，跨模态交互注意力模块，用于对RGB图和深度图之间的远程跨模态依赖进行建模，整合RGB数据和深度数据之间的互补信息。

[0020] 进一步地，经跨模态Transformer融合模块，得到跨模态交互信息的公式，表示为：

[0021] $$Attention(Q_R, K_D, V_D) = softmax(Q_R K_D^T / \sqrt{d_k}) V_D$$

[0022] $$Attention(Q_D, K_R, V_R) = softmax(Q_D K_R^T / \sqrt{d_k}) V_R$$ 。

[0023] 其中，Q_R、Q_D分别为两种模态的查询，K_R、K_D分别为两种模态的键，V_R、V_D分别为两种模态的值。

[0024] 进一步地，步骤S4的跨模态密集协作聚合模块包括：三个特征聚合模块和一个双重倒残差模块，跨模态密集协作聚合模块用于将低分辨率编码器特征扩展到与输入图像分辨率大小一致，特征聚合模块用于聚合特征和融合跨模态信息。

[0025] 进一步地，特征聚合模块包括：一个CBAM和两个IRB，还包含了两个元素相乘和一个元素相加操作；基于特征聚合模块的特征聚合和融合跨模态信息过程包括如下步骤：

[0026] S4.1，RGB特征T_R和深度特征T_D进行相乘，再经过一个IRB进行卷积后得到过渡的RGB-D特征图D，此过程表示为：

[0027] D＝IRB(T_R×T_D)。

[0028] 其中，T_R包括：T′、T_1'和T_3，T_D包括：C_2、C_3、C_4。

[0029] S4.2,深度特征T_D经CBAM增强后的特征记为T''_D,此过程表示为:

[0030] $T_D' = Channel(T_D) \times T_D$

[0031] $T_D'' = Spatial(T_D') \times T_D'$。

[0032] S4.3,D再与深度特征T''_D再次相乘强化语义特征后得到D′,此过程表示为:

[0033] $D' = D \times T_D''$。

[0034] S4.4,D′与RGB特征T_R相加以重新增强显著特征,同时引入较低层次的输出特征T_{DC}进行元素相加,然后使用IRB,得到跨模态融合后的RGB-D特征D″,此过程表示为:

[0035] $T_R' = T_R + D'$

[0036] $D'' = IRB(T_R' + T_{DC})$;

[0037] 进一步地,步骤S4中,来自T2T-ViT的经重组后的RGB信息T′,T'_1,T_3和来自MobileNet V2的深度信息C_1,C_2,C_3,C_4被输入到密集连接增强后的解码器,密集连接用于将不同尺度上的深度特征和RGB特征进行融合。

[0038] 进一步地,步骤S5中,预测的显著图由调整相应大小后的真值图进行监督,将这一阶段产生的四个损失表示为$L_d^i, i = 1,2,3,4$,总的损失函数L_{total}计算公式如下:

[0039] $$L_{total} = \sum_{i=1}^{4} \lambda_i BCE(P_i, G_i)$$。

[0040] 其中,λ_i表示每个损失的权重,按照分辨率由小到大的顺序将四个显著性预测图依次记为P_i(i=1,2,3,4),G_i表示来自真值图的监督,其分辨率与P_i对应,BCE()表示交叉熵损失函数。

[0041] 本发明具有如下有益效果:

[0042] 1、本发明充分考虑到RGB图像和深度图像之间的不同。我们使用基于Transformer的T2T-ViT网络和轻量级MobileNet V2网络,分别实现对RGB信息和深度信息的提取。这种非对称双流学习网络设计使本发明相比其他显著性目标检测方法,降低了模型参数量,同时提高了显著性目标检测速度,并具有优秀的显著性目标检测性能。

[0043] 2、本发明所设计的解码器包括跨模态Transformer融合模块(CMTFM)和跨模态密集协作聚合模块(CMDCAM)。跨模态Transformer融合模块(CMTFM)作为解码器的块,可以建模RGB数据与深度数据之间的远程跨模态依赖,实现RGB数据与深度数据之间的跨模态信息交互。本发明采用密集连接来增强解码器,设计的跨模态密集协作聚合模块(CMDCAM),通过密集协作融合的方式聚合不同层次的特征,并有效地融合跨模态信息。本发明所设计的解码器有效地融合RGB图像信息和深度信息,提高了显著性目标的检测精度。

附图说明

[0044] 为了更清楚地说明本发明具体实施方式或现有技术中的技术方案,下面将对具体实施方式或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图是本发明的一些实施方式,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。在附图中:

[0045] 图1示出了本发明的一种RGB-D显著性目标检测方法的流程图。

[0046] 图2示出了本发明的一种RGB-D显著性目标检测方法的结构示意图。

[0047] 图3示出了图2的基于T2T-ViT的transformer编码器的结构示意图。

[0048] 图4示出了图2的解码器中的特征聚合模块FAM的结构示意图。

具体实施方式

[0049] 下面将结合附图对本发明的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0050] 如图1所示的一种RGB-D显著性目标检测方法，具体包括如下步骤：

[0051] S1，输入RGB图像和深度图像。

[0052] S2，利用基于T2T-ViT的transformer编码器对RGB图像进行特征提取，利用基于轻量级卷积网络MobileNet V2的编码器对深度图像进行特征提取，分别获取RGB图像和深度图像不同层级的显著性特征。

[0053] 如图2所示，轻量级卷积网络MobileNet V2的编码器包括：MobileNet V2模块一、MobileNet V2模块二、MobileNet V2模块三和MobileNet V2模块四。其中MobileNet V2模块一、MobileNet V2模块二、MobileNet V2模块三和MobileNet V2模块四分别输出深度信息C_1、C_2、C_3、C_4；

[0054] T2T-ViT网络是对ViT网络的改进，在ViT的基础上增加了T2T操作，相当于卷积神经网络中的下采样，用于同时建模图像的局部结构信息与全局相关性。T2T能将相邻的tokens聚合成一个新的token，从而减少token的长度。

[0055] 具体地，步骤S2的基于T2T-ViT的transformer编码器中的T2T操作包括：重组和软拆分，重组是将一个token序列$Tok \in \mathbb{R}^{l \times c}$重建成一个3D张量$I \in \mathbb{R}^{h \times w \times c}$，其中，l是token序列Tok的长度，c是token序列Tok和3D张量I的通道数，h，w分别代表I的高度和宽度，且l＝h×w。

[0056] 软拆分是通过展开操作将I软分割成k×k大小的块，$I \in \mathbb{R}^{h \times w \times c}$经软拆分后得到token序列，其长度$l_0$可以表示为：

[0057] $$l_0 = \left\lfloor \frac{h+2p-k}{k-s}+1 \right\rfloor \times \left\lfloor \frac{w+2p-k}{k-s}+1 \right\rfloor$$

[0058] 其中，S代表块之间像素重叠个数，p代表块之间像素填充个数，k-s代表卷积运算中的步长，当s＜k-1时，tokens序列的长度便可降低。

[0059] 原始RGB图像$I_{input} \in \mathbb{R}^{H \times W \times C}$，其中H，W，C分别代表$I_{input}$的高度、宽度和通道数，经过重组得到的token序列$T \in \mathbb{R}^{h \times w \times c}$经过三轮Transformer转换和两轮T2T操作得到了多级tokens序列T′，T_1，T_1'，T_2，T'_2，这一过程可以表示为：

[0060] T′＝Transformer(T)，

[0061] T_1＝Unfold(Reshape(T′))，

[0062] T_1'＝Transformer(T_1)，

[0063] T_2＝Unfold(Reshape(T_1'))，

[0064] T'_2＝Transformer(T_2)。

[0065] 具体地，步骤S2中，基于轻量级卷积网络MobileNet V2的编码器包括倒残差结构。语义信息主要存在于RGB图像中，深度图传达了没有对象细节的信息。深度图中所含信息相对于RGB较单一，且量少，而且往往深度图中颜色最深的部位便是显著性目标检测任务所要寻找的显著目标。所以本发明采用轻量级的MobileNet V2网络便能很好地提取深度图的信息。MobileNet V2是对MobileNet V1的改进，提出了倒残差结构(Inverted Residual Block，IRB)结构。倒残差结构与残差结构中维度先缩减再扩增正好相反，更有利于特征的学习。如图2所示，将MobileNet V2侧输出的4级深度特征图标注为C_1、C_2、C_3、C_4。

[0066] S3，利用基于跨模态Transformer融合模块，融合深层次RGB特征和深度特征之间的互补语义信息，生成跨模态联合特征。

[0067] 具体地，步骤S3中的跨模态Transformer融合模块(CMTFM，Cross-modality Transformer Fusion Module)包括：跨模态交互注意力模块和Transformer层，跨模态交互注意力模块，用于对RGB图和深度图之间的远程跨模态依赖进行建模，整合RGB数据和深度数据之间的互补信息，从而提高显著性预测的准确性。CMTFM基于视觉显著性转换器(Visual Saliency Transformer，VST)中的RGB-D转化器，为了节省参数和计算资源，我们去掉了RGB-D转化器中的自注意力部分。

[0068] 具体地，如图2所示，在CMTFM中，融合T'_2和C_4以整合RGB和深度数据之间的互补信息。通过三个线性投影操作将T'_2转化生成查询Q_R，键K_R，值V_R。类似地，用另外三个线性投影操作将C_4转化成查询Q_D，键K_D，值V_D。由Transformer层中的多头注意力中的“缩放点积注意力”公式可以得到跨模态交互信息的公式，表示为：

[0069] $$Attention(Q_R,K_D,V_D)=softmax(Q_RK_D^T/\sqrt{d_k})V_D$$

[0070] $$Attention(Q_D,K_R,V_R)=softmax(Q_DK_R^T/\sqrt{d_k})V_R。$$

[0071] 这样来自RGB块标记的信息流T'_2和深度块标记的信息流C_4通过4次跨模态交互注意力模块来进行跨模态信息交互后，再经过一个4层Transformer层进行强化得到了token序列T_3。

[0072] 来自编码器的RGB和深度序列必须通过线性投影层，以将其嵌入维度从384转换为64，以减少计算和参数。

[0073] S4，利用密集连接卷积神经网络增强特征融合模块，逐级融合不同尺度上的深度特征和RGB特征，输入到显著性目标检测部分。

[0074] 具体地，步骤S4的跨模态密集协作聚合模块(CMDCAM，Cross-modal dense cooperative Aggregation Module)包括：三个特征聚合模块(FAM，Feature Aggregation Module)和一个双重倒残差模块，跨模态密集协作聚合模块用于将低分辨率编码器特征扩展到与输入图像分辨率大小一致，以便进行像素级分类。特征聚合模块既能作为解码器网络的组成，承担起聚合特征的作用，也能有效地融合跨模态信息。

[0075] 具体地，如图4所示，特征聚合模块包括：一个CBAM和两个IRB，还包含了两个元素相乘和一个元素相加操作；深度图仅传达了一个先验区域，缺乏对像细节。因此，我们先通过两次乘法增强了RGB的语义特征。基于特征聚合模块的特征聚合和融合跨模态信息过程包括如下步骤：

[0076] S4.1，RGB特征T_R和深度特征T_D进行相乘，再经过一个IRB进行卷积后得到过渡的

RGB-D特征图D,此过程表示为:

[0077] $D=IRB(T_R\times T_D)$。

[0078] 其中,T_R包括:T′、T_1′和T_3,T_D包括:C_2、C_3、C_4。S4.2,深度特征T_D经CBAM增强后的特征记为T''_D,此过程表示为:

[0079] $T_D'=Channel(T_D)\times T_D$

[0080] $T_D''=Spatial(T_D')\times T_D'$。

[0081] S4.3,D再与深度特征T''_D再次相乘强化语义特征后得到D′,此过程表示为:

[0082] $D'=D\times T_D''$。

[0083] S4.4,D′与RGB特征T_R相加以重新增强显著特征,同时引入较低层次的输出特征T_{DC}进行元素相加,然后使用IRB,得到跨模态融合后的RGB-D特征D″,此过程表示为:

[0084] $T_R'=T_R+D'$

[0085] $D''=IRB(T_R'+T_{DC})$。

[0086] 具体地,步骤S4中,来自T2T-ViT的经重组后的RGB信息T′,T_1′,T_3和来自MobileNet V2的深度信息C_1,C_2,C_3,C_4被输入到密集连接增强后的解码器,密集连接用于将不同尺度上的深度特征和RGB特征进行融合。

[0087] S5,按照预测的显著性图分辨率由小到大排序,并利用真值图对网络进行有监督学习,输出最终的显著性检测结果。

[0088] 具体地,如图1所示,步骤S5中,通过在每个解码器模块的输出中依次添加1×1单通道卷积和Sigmoid激活函数来进行显著性映射。在训练期间,预测的显著图由调整相应大小后的真值图进行监督,将这一阶段产生的四个损失表示为$L_d^i, i=1,2,3,4$,总的损失函数L_{total}计算公式如下:

[0089] $$L_{total}=\sum_{i=1}^{4}\lambda_i BCE(P_i,G_i)$$。

[0090] 其中,λ_i表示每个损失的权重,按照分辨率由小到大的顺序将四个显著性预测图依次记为P_i(i=1,2,3,4),G_i表示来自真值图的监督,其分辨率与P_i对应,BCE()表示交叉熵损失函数。

[0091] 按照分辨率由小到大的顺序将四个显著性预测图依次记为P_i(i=1,2,3,4)。G_i表示来自GT的监督,其分辨率与P_i对应。利用交叉熵损失函数(BCE)公式,则可计算总的损失函数L_{total},计算公式如下:

[0092] $$L_{total}=\sum_{i=1}^{4}\lambda_i BCE(P_i,G_i)$$。

[0093] 其中,λ_i表示每个损失的权重。

[0094] 在显著性目标检测方法中,使用基于图像分类的经预训练的模型作为主干网,有助于训练过程中损失收敛,从而能够有效的提高显著目标检测的精度。本发明使用了经过预训练的基于T2T-ViT的transformer编码器和基于轻量级卷积网络MobileNet V2的编码器来作为主干网提取特征。

[0095] 本发明设计了跨模态密集协作聚合模块(CMDCAM),该模块基于倒残差模块,具有计算参数量和计算量小的优点。该模块不但可以融合RGB信息和深度信息两种模态信息,而

且可以聚合不同层次的特征信息。该模型可以实现在降低检测方法计算量前提下，明显提高了显著性目标的检测性能，并提高了显著性目标的检测精度。

[0096] 当然，上述说明并非是对本发明的限制，本发明也并不仅限于上述举例，本技术领域的技术人员在本发明的实质范围内所做出的变化、改型、添加或替换，也应属于本发明的保护范围。


输入RGB图像
输入深度图像
RGB图像特征提取
深度图像特征提取
基于Transformer跨模态融合模块
密集连接增强特征融合模块
显著性目标检测

图 1

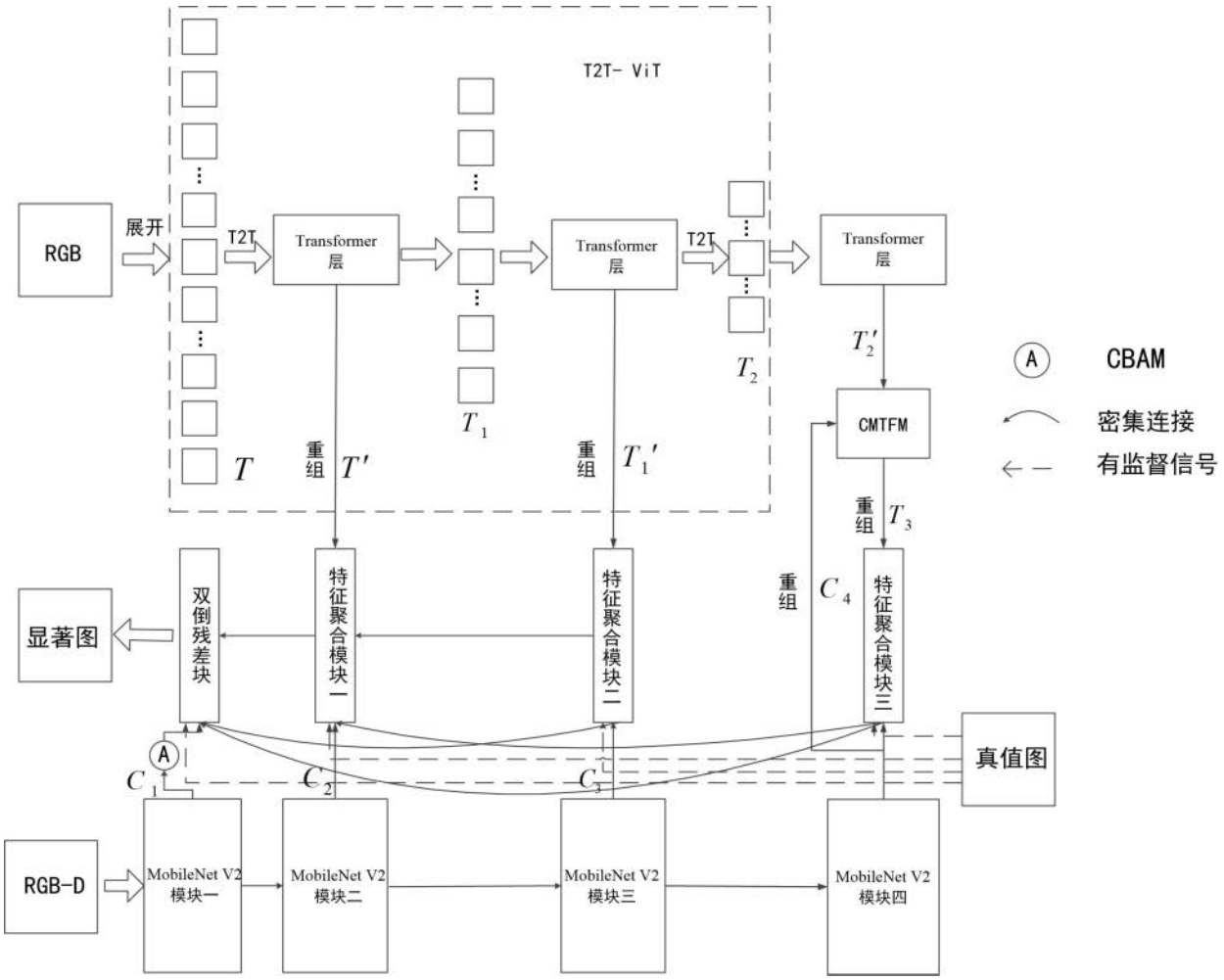
T2T- ViT
RGB
展开
T2T
Transformer 层
Transformer 层
T2T
Transformer 层
T_2'
T_2
T_1
T
重组
T'
重组
T_1'
CMTFM
重组
T_3
A
CBAM
密集连接
有监督信号
显著图
双倒残差块
特征聚合模块一
特征聚合模块二
重组
C_4
特征聚合模块三
真值图
A
C_1
C_2
C_3
RGB-D
MobileNet V2 模块一
MobileNet V2 模块二
MobileNet V2 模块三
MobileNet V2 模块四

图 2

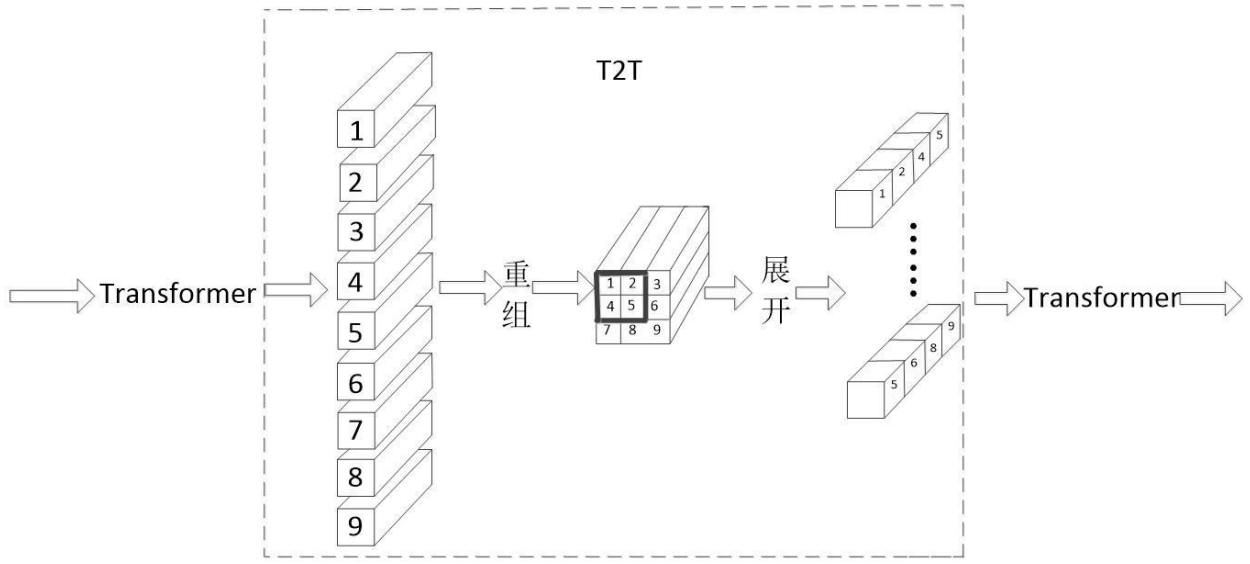
T2T
Transformer
1
2
3
4
5
6
7
8
9
重组
1 2 3
4 5 6
7 8 9
展开
Transformer

图 3

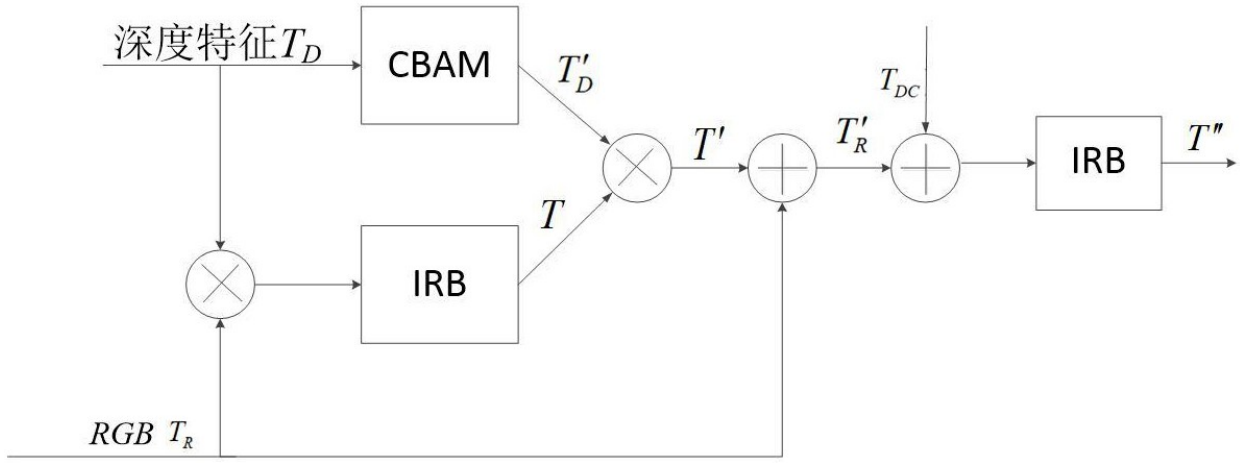

图 4