联合组间对抗数据混合与变换器学习的协同显著性检测

吴 泱[1)] 宋慧慧[1)] 张开华[2)] 陈 虎[3)] 刘青山[2)]

[1)](南京信息工程大学自动化学院 南京 210044)

[2)](南京信息工程大学计算机与软件学院数字取证教育部工程研究中心 南京 210044)

[3)](四川大学视觉合成图形图像技术国家级重点实验室 成都 610041)

摘 要 协同显著性检测旨在发现并分割出一组图像中相同语义类别的前景显著目标.当前基于深度学习的协同显著性检测方法主要存在两方面局限:(1) 训练数据中仅含有单一显著目标,无法为模型训练提供对抗样本,导致其泛化性受限,难以有效应对未知类别目标、干扰显著目标、嘈杂背景等挑战;(2) 现有方法通常利用卷积神经网络提取特征,其感受野受限,无法建模长程依赖关系,限制了所学特征的表征力.为此,本文提出了一种新颖的基于组间对抗数据混合的协同显著性检测变换器,旨在通过纯视觉变换器构建序列到序列的协同显著性检测网络,并使用组间混合后的数据进行对抗训练,以提升模型的泛化性.所设计的网络结构包含数据混合子网络和协同显著性检测变换器两部分.具体而言,在数据混合子网络中,本文设计了目标细化模块,输入类激活图,引导网络以无监督的方式从一组图像中分割出边缘平滑的显著目标作为对抗对象,并通过设计调距模块将对抗对象以最小化重叠的方式混合至另一组图像之中,生成混合训练数据;在协同显著性检测变换器中,本文从序列建模的角度,设计了任务注入器,将组信息图符与显著性信息图符注入序列特征之中,并利用自注意力机制充分捕获特征之间的全局上下文信息.最后,将获得的组特征和显著性特征通过自注意力机制进行充分混合交互,以进一步增强特征的表征力,生成精确的协同显著性检测结果.本文在包含Cosal2015、CoCA和CoSOD3k等三个基准数据集上做了充分的实验评估,与多个领先方法的对比结果充分证明了本方法的优越性能.

关键词 数据混合;变换器;协同显著性检测;大数据

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2023.01838

Inter-Group Adversarial Mixup and Transformer Learning for Co-Saliency Detection

WU Yang[1)] SONG Hui-Hui[1)] ZHANG Kai-Hua[2)] CHEN Hu[3)] LIU Qing-Shan[2)]

[1)](*School of Automation, Nanjing University of Information Science and Technology, Nanjing* 210044)

[2)](*School of Computer and Software, Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing* 210044)

[3)](*Key Laboratory of Fundamental Science for National Defense on Vision Synthetization and Graphic Image, Sichuan University, Chengdu* 610041)

Abstract Co-saliency detection targets at segmenting the common salient objects in a group of relevant images. The current co-salient object detection methods based on deep learning have two limitations: (1) There is only a single target in training images, which can not provide adversari-

收稿日期:2022-06-15;在线发布日期:2023-01-10.本课题得到科技创新2030-"新一代人工智能"重大项目(No.2018AAA0100400)、国家自然科学基金项目(No.61876088,61872189,62276141,U20B2065,61532009)、江苏省333工程人才项目(No. BRA2020291)、视觉合成图形图像技术国家级重点实验室开放研究项目(No.2021SCUVS001)资助.**吴 泱**,硕士研究生,主要研究领域为协同显著性检测. E-mail: wuy98419@163.com. **宋慧慧(通信作者)**,博士,教授,主要研究领域为视频目标分割、图像超分. E-mail: songhuihui@nuist.edu.cn. **张开华**,博士,教授,中国计算机学会(CCF)会员(42089M),主要研究领域为协同显著性检测、视觉跟踪. **陈 虎**,博士,副教授,主要研究领域为计算机视觉、医学成像及医学图像处理. **刘青山**,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为视频内容分析与理解.

al samples for the model, making the model have poor generalization performance. When facing the interference of unknown class targets, similar salient objects, noisy background environments and so on, the model is greatly limited; (2) The existing methods usually use convolution neural networks (CNNs) to extract features. However, the CNNs can not obtain a large receptive field which makes the model unable to fully model the long-range dependencies, resulting in poor discriminative capability of the model. To this end, we propose a co-saliency detection transformer guided by intra-group adversarial mixup. Aiming at building the co-saliency detection network from a perspective of sequence-to-sequence and training the model on mixup adversarial data, making the model more generic. Our network mainly contains two parts, a mixup subnetwork and a co-saliency detection transformer. Specifically, in the mixup sub-network, we propose an object refinement module: we set input class activation maps(CAMs) as guidance to segment salient objects with smooth edges as the adversarial objects in an unsupervised way; a distance adjusting module: the adversarial objects are mixed into another group of images with the minimum overlap, constructing the mixed training data. In the co-saliency detection transformer, we construct the model from sequence-to-sequence. In this part, we design a task injector, which can inject group information and saliency information into the feature sequence, and we adopt self-attention to fully capture global information between features. Finally, we mix the group information and saliency information by self-attention, further enhancing the discriminative capability of the feature and generating the Precise results of co-saliency detection. Extensive experiments are carried out on three benchmark datasets including Cosal2015, CoCA, and CoSOD3k, demonstrating superiority of our method to state-of-the-art methods.

Keywords　mixup; transformer; co-salient object detection; big-data

1　引　言

协同显著性检测(Co-saliency Detection)旨在发现并分割出一组图片中语义类别相同的前景显著目标[1].相较于只关注于分割单个目标的显著目标检测任务[2],协同显著性检测更具挑战性,因为它需要在存在其他分散注意力物体的干扰下,区分出多幅图像中同时出现的显著物体.尽管如此,随着深度学习的发展,这项任务的研究已取得了长足进步,并被成功应用于一系列计算机视觉任务,如目标分割[3]、图像检索[4]、视频显著性检测[5]等领域.

随着卷积神经网络(CNNs,Convolution Neural Networks)[6]研究的快速发展,涌现出大量相关工作并不断刷新最佳性能[7-10].这类方法通过一系列创新性设计,如组信息融合机制[7]、梯度引导机制[9]、图像匹配技术[10]等,来学习更加鲁棒的特征表达,以应对传统方法难以提取高级语义特征,导致模型不能有效处理复杂场景中协同显著目标的大尺度表观变化挑战.尽管取得了不错的效果,但是这类基于卷积神经网络的工作存在两方面局限:

(1)现有主流方法都是基于经验风险最小化原则[11],利用神经网络强大的数据拟合能力在训练过程中追求对训练数据的平均误差最小化.这意味着神经网络可轻易过拟合训练数据,但在面对未知类别目标、相似显著目标、嘈杂背景环境等挑战时,泛化性较差,从而导致严重误检(如图 1 所示,实际场景中存在训练过程中类别未出现的目标,即未知类别目标(骰子组中,骰子为未知类别目标)、干扰显著目标(礼物盒组中,花朵、人物、圣诞树和人为干扰显著目标)、嘈杂背景(怀表组中怀表所处环境背景嘈杂)等挑战,基于卷积神经网络的方法难以有效处理这些挑战).文献[14]亦指出:即使增加训练数据与模型参数,或者采用强正则化等措施,在经验风险最小化原则下,模型仍更倾向于记忆训练数据而非提升泛化性.这些都极大降低了模型的实际应用价值;

(2)现有主流方法通常利用卷积神经网络提取特征,其感受野位于局部滑动窗口之中,导致所提取的特征存在固有局限性,难以捕获关键的全局线索[15].尽管最近提出了一些措施来弥补这方面缺

点，比如采用全连接层[16]、全局池化层[17]、非局部模块[18]等策略融入全局信息.但是，这些操作只局限于某些层中，而整体的卷积神经网络架构不变，导致模型的判别力仍受限.

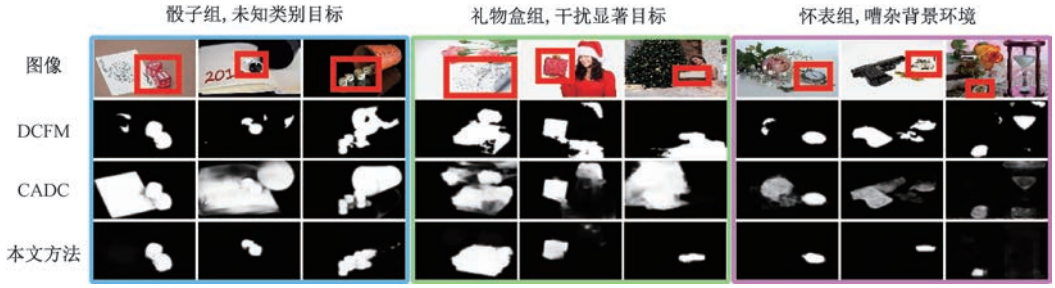

图1 相比于先进方法DCFM[12]和CADC[13]，本文方法在一系列挑战场景中的表现

为了突破第一类局限，文献[14]提出了数据混合(Mixup)增强策略，以提升网络的泛化性.该数据混合策略通过线性加权给定的一对输入图像，生成一幅新的混合图像作为模型输入.在此基础上，文献[19-20]对该数据混合策略进行了改进，在隐藏特征空间中混合输入图像或者以基于局部统计的方式进行数据混合.尽管这类数据增强策略在相关任务上取得了不错的效果，但是，这类简单的数据加权方法并未专门考虑图像中的显著目标区域，导致不能充分提取图像中的显著信息，从而在一定程度上影响了数据混合的有效性.为此，文献[21]提出了一种基于显著性和局部统计的数据融合方法，以充分利用显著信息来提升数据混合的有效性.文献[22]提出了一种新的优化策略，在最大化混合示例显著性度量的同时，增大数据之间的差异，以混合更多输入数据.然而，以上两种方法仍存在显著目标遮挡、背景噪声引入过多的问题，难以保证在混合显著目标的同时，最大程度保留图像中的主体背景信息.为此，本文针对协同显著性检测中训练数据存在的问题，即显著目标单一，缺少对抗样本而导致模型泛化性弱，设计了一种新的数据混合策略，以增强模型的泛化性.

为了突破第二类局限，最近大热的视觉变换器(ViT，Vision Transformer)[23]作为一种新的模型范式，致力于解决卷积神经网络框架存在的固有缺陷.ViT[23]具有强大的捕获全局长程依赖关系的能力，便于建模不同区域之间的结构依赖关系，而这种能力对于本任务中建模组内或组间协同显著目标之间的关系至关重要.在此基础上，为了建模局部依赖关系，文献[24]通过分层递归相邻图符(tokens)建模相邻图符表征的局部结构.文献[25]采用滑动窗口的方式来兼顾局部信息.然而，这类变换器方法主要用于目标识别任务，其通过学习高级语义特征来预测目标类别标签.高级语义特征的分辨率较低，对大尺度目标表观变化的自适应性较好，适用于对类别标签的估计.与之不同，本任务为像素密集型预测任务，需要高分辨率的浅层特征来恢复空间细节信息.最近，文献[15]提出了一种基于纯视觉变换器的显著性检测方法，通过融合浅层特征与高级语义特征并改进了T2T上采样方式[24]来应对这个挑战.受此启发，本文通过设计一种新颖的任务注入器来将组信息图符和显著性图符融入特征序列，并通过跳跃连接将浅层特征与高级特征进一步融合，从而利用它们之间的互补特性，更精确地预测协同显著目标的掩模.

为此，本文提出了一种基于组间对抗数据混合的协同显著性检测变换器.其结构包含数据混合子网络和协同显著性检测变换器两部分.在数据混合子网络中，本文设计了目标细化模块：以类激活图(CAMs，Class Activation Maps)[26]为引导，通过无监督学习的方式从一组图中分割出显著目标作为对抗对象，并设计了一个调距模块将对抗对象以最小化重叠的方式混合至另一组图中，生成混合训练数据作为变换器的输入；接着，在变换器中，设计了一个任务注入器，将组信息图符与显著性信息图符注入序列特征之中，并利用自注意力机制充分捕获特征之间的全局上下文信息；最后，将获得的组特征和显著性特征通过注意力机制进行充分混合交互，生成精确的协同显著性目标掩模.在包含Cosal2015[27]、CoCA[9]、CoSOD3k[28]等三个基准数据集上的大量实验结果验证了本方法的有效性.

本文的主要贡献总结如下：

(1)据我们所知，本文首次提出了一种基于纯视觉变换器的协同显著性检测方法，在使用现有通用数据增强策略的情况下，即能在Cosal2015、CoCA、CoSOD3k等三个标准数据集上达到当前领先水平.

(2)本文提出了一种类激活图引导的数据混合方法.该方法能够以无监督的方式精细地分割出两组图像中的显著目标来作为对抗对象，并在每组图

像中通过替换对应背景区域混入对抗对象,以生成新的含对抗对象的训练数据.

(3)本文在纯视觉变换器中设计了一种新颖的任务注入器:通过学习组信息图符和显著性信息图符,将组信息与显著性信息通过注意力机制进行混合,引导变换器更好地关注于分割协同显著目标.

2　相关工作

2.1　协同显著性检测

早期的协同显著性检测方法[29]提取图像的低级特征,如 Gabor[30]或 SIFT[31],然后利用图像间这些特征的一致性信息来进行协同显著性检测.其方案主要包括通过流行排序来生成显著性图来捕获图像内的约束[29],或者使用聚类方法[32]与平移对齐方法[33]来生成全局关联信息.随后,一些工作[34]使用中级特征来处理本任务.所采用的中级特征包括显著性检测或者图像分割的结果.以上方法采用的都是手工提取的特征,难以有效处理真实场景下目标表观的大尺度变化.随着深度学习的兴起,大量相关工作[7,9,29,35-39]以端到端的方式直接从图像中学习出协同显著区域.其中,文献[7]为协同显著性检测设计了一种组协作学习框架,以组的方式探索一组图像特征的联合信息与单个图像特征的信息.文献[35]提出了一种分层设计的协同显著性检测框架,由卷积神经网络生成的协同显著性图经由标签平滑再处理.文献[9]提出了一种梯度引导的协同显著性检测模型,利用图像梯度信息使协同显著特征得到更多关注.文献[36]提出了一种图卷积框架来处理此任务.文献[37]提出了一种组与组之间协作学习的策略,以通过探索组与组之间的关系进行特征学习,这些工作都取得了不错的结果.

2.2　数据混合

为了防止深度神经网络对训练数据的过拟合,数据增强[40]被提出,并被广泛应用于网络模型的训练.传统方法[41]大都依赖于数据或任务的转换来生成新数据,缺乏对不同图像间关系的建模,限制了模型泛化力的提升.为此,文献[14]提出了数据混合策略,可独立应用于各种数据类型与任务,极大提升了模型的泛化力与鲁棒性.文献[14]在两个输入数据之间进行线性插值,并利用具有相应软标签的混合数据来训练模型.在此基础上,文献[19]和文献[20]分别在隐藏的特征流型空间中应用数据混合,以及通过剪切和拼接图像进行数据混合.文献[21]提出了一种基于显著性和局部信息统计的数据混合方法,能较好地保留显著目标区域.文献[22]提出了一种基于离散优化的数据增强方法,在所有输入数据中找到显著区域集合的最佳组合.

2.3　视觉变换器

文献[42]首次在机器翻译领域提出了基于变换器的编码和解码结构.最近,越来越多的工作将变换器引入计算机视觉任务并取得了优异的效果.文献[43]结合了卷积神经网络和变换器来处理目标检测任务.文献[44]也采用这种结构来处理全景分割任务.它们都采用卷积神经网络提取特征,再使用变换器捕获特征的长程依赖关系.ViT[23]首次在计算机视觉领域设计纯变换器结构,将输入图片裁剪为若干图符,从序列的角度处理图像分类任务.文献[45]提出了一种金字塔结构,将 ViT 调整为适应密集预测任务的结构.然而,变换器对局部信息的建模能力较差,为此,文献[24]使用一种 T2T 模块对局部特征结构进行建模,从而生成多尺度标记特征.文献[25]采用滑动窗口的方式来兼顾局部信息,以便更好地融合全局与局部特征信息.

3　本文方法

如图 2 所示,本文模型主要包含数据混合子网络和变换器两部分.在训练阶段,输入两组图像 $\mathcal{I}_1^o=\{I_{1,n}^o\in\mathbb{R}^{3\times H\times W}\}_{n=1}^N$ 和 $\mathcal{I}_2^o=\{I_{2,n}^o\in\mathbb{R}^{3\times H\times W}\}_{n=1}^N$ 其中,每组包含 N 张有相同前景显著目标的相关图像.本文的目标是学习一个前馈网络 f,来预测出两组图像中的协同显著目标掩膜 $\hat{\mathcal{O}}=\{\hat{\boldsymbol{O}}^n\in\{0,1\}^{1\times H\times W}\}_{n=1}^{2N}$:

$$\hat{\mathcal{O}}=f(\mathcal{I}_1^o,\mathcal{I}_2^o) \tag{1}$$

首先,本文设计了一个数据混合子网络 f_{mixup},$\mathcal{I}_1^o$ 和 $\mathcal{I}_2^o$ 同时输入该网络,生成混合图像组 $\mathcal{I}_1^m=\{I_{1,n}^m\in\mathbb{R}^{3\times H\times W}\}_{n=1}^N$,$\mathcal{I}_2^m=\{I_{2,n}^m\in\mathbb{R}^{3\times H\times W}\}_{n=1}^N$:

$$\{\mathcal{I}_1^m,\mathcal{I}_2^m\}=f_{mixup}(\mathcal{I}_1^o,\mathcal{I}_2^o) \tag{2}$$

具体如图 3 所示,f_{mixup} 由分类网络 f_{cls}、目标细化模块 f_{ref} 和调距模块 f_{adj} 等三部分组成.首先,$\mathcal{I}_1^o$ 和 $\mathcal{I}_2^o$ 同时输入文献[46]预训练过的 Densenet-169 网络,本文抛弃其最后的全连接层作为分类网络 f_{cls},经由 f_{cls} 得到对应的两组类激活图 $\mathcal{I}_1^{cam}=\{\boldsymbol{I}_{1,n}^{cam}\in(0,1)^{1\times H\times W}\}_{n=1}^N$ 以及 $\mathcal{I}_2^{cam}=\{\boldsymbol{I}_{2,n}^{cam}\in(0,1)^{1\times H\times W}\}_{n=1}^N$:

$$\{\mathcal{I}_1^{cam},\mathcal{I}_2^{cam}\}=f_{cls}(\mathcal{I}_1^o,\mathcal{I}_2^o) \tag{3}$$

然后,$\{\mathcal{I}_1^{cam},\mathcal{I}_2^{cam}\}$ 通过 1×1 卷积得到对应类别向量 $\hat{\boldsymbol{y}}_1\in\mathbb{R}^{1\times78}$ 和 $\hat{\boldsymbol{y}}_2\in\mathbb{R}^{1\times78}$.同时,将两组类

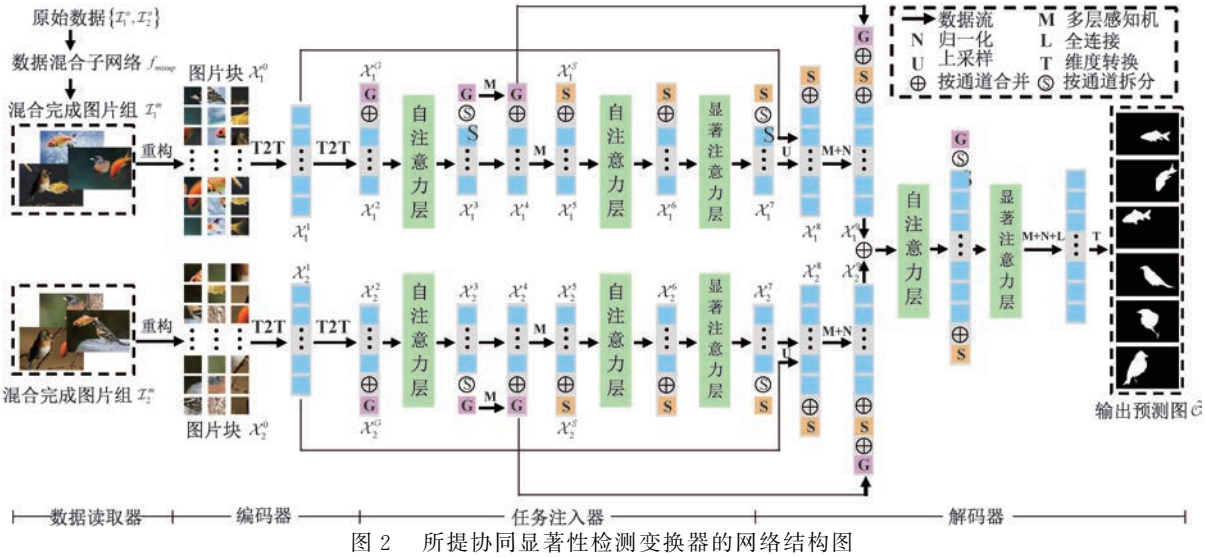

图 2 所提协同显著性检测变换器的网络结构图

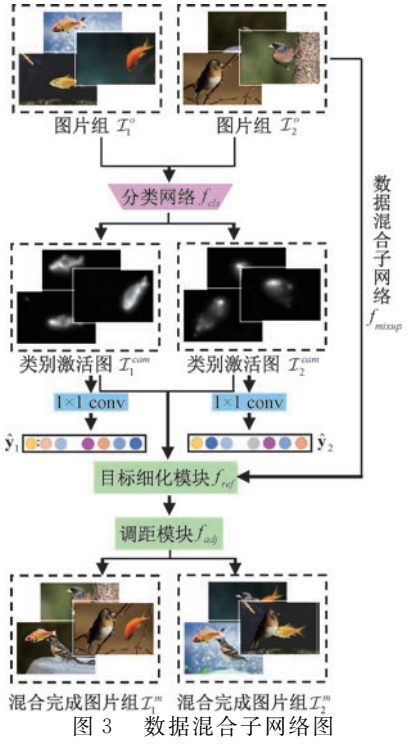

图 3 数据混合子网络图

激活图与原图一同送入目标细化模块 f_{ref}，生成待分割目标边界清晰的掩膜，再经由调距模块 f_{adj} 生成混合后的数据 $\mathcal{I}_1^m$ 和 $\mathcal{I}_2^m$：

$$\{\mathcal{I}_1^m,\mathcal{I}_2^m\}=f_{adj}(f_{ref}(\mathcal{I}_1^o,\mathcal{I}_2^o,\mathcal{I}_1^{cam},\mathcal{I}_2^{cam})) \quad (4)$$

随后，$\{\mathcal{I}_1^m,\mathcal{I}_2^m\}$ 被送入协同显著性检测变换器 f_{trans} 以进行后续处理：第一步，通过编码器将输入数据 $\{\mathcal{I}_1^m,\mathcal{I}_2^m\}$ 裁剪为图片块，输入预训练过的变换器骨干网络[24]，获取特征序列 $\{\mathcal{X}_1^1,\mathcal{X}_2^1\}$. 第二步，通过任务注入器中的组信息图符 $\mathcal{X}_G$ 和显著性信息图符 $\mathcal{X}_G$，分别学习输入图片组中的组共性特征与显著性特征，经自注意力层和显著注意力层捕获全局信息并进行特征融合. 第三步，在解码器阶段，特征序列经过上采样得到 $\{\mathcal{X}_1^9,\mathcal{X}_2^9\}$. 最后，将 $\{\mathcal{X}_1^9,\mathcal{X}_2^9\}$ 按通道维度叠加，再依次通过自注意力层和显著注意力层交互信息，预测出协同显著图 $\hat{\mathcal{O}}$.

在测试阶段，输入一组图像 $\mathcal{I}=\{I_n^o\}_{n=1}^N$，不需要经过数据混合，直接经过训练好的 f_{trans}，得到对应的一组协同显著性图：

$$\hat{\mathcal{O}}=\{\hat{\boldsymbol{O}}^n\}_{n=1}^N=f_{trans}(\mathcal{I}) \quad (5)$$

本文的创新点在于所设计的类激活图引导无监督学习的数据混合和针对协同显著性检测的纯视觉变换器，在接下来的章节中将详细介绍这两部分.

3.1 类激活图引导无监督学习的数据混合子网络

当前协同显著性检测的训练图像存在协同显著目标单一、对抗目标数量少的特点. 模型在这种图象上训练很容易过拟合，导致泛化性差. 鉴于此，本文设计了一种能生成组间对抗样本的数据混合子网络. 如图 3 所示，本文将不同类别的两组图片输入分类网络，生成对应的类激活图. 尽管类激活图具有丰富的位置信息和精确的语义信息，但是分割结果粗糙，难以恢复出显著目标的精细边缘. 受文献[47]的启发，本文从像素分类的角度，以无监督学习的方式分割显著目标，将拥有相同外观属性的像素归于同一类别. 根据协同显著性检测的任务特性，仅需将像素归为前景与背景两类. 在文献[47]和像素自适应卷积(PAC)[48]的基础上，本文设计了目标细化模块，运用像素级相似性核(Pixel-level affinity kernel)迭代更新每一个像素，以细化类激活图. 同时，本文也设计了调距模块来调整一组图中待混合显著目标的位置，以尽可能地避免图中显著目标之间互相遮挡.

3.1.1　目标细化模块

为了更加精确地分割边缘,本文利用像素自适应卷积.根据其中的双边滤波思想,不仅考虑空间域中的像素点位置,亦考虑像素域内的像素值差异,并将像素点投影到高维空间,在高维空间中减轻滤波后边缘模糊的影响.如图 4 所示,本文将图像 I 对应的类激活图 $\boldsymbol{I}^{cam}$ 定义为其初始显著性掩膜 $\boldsymbol{M}^0 \in (0,1)^{1\times h\times w}$,滤波器定义为 k.在第 t 个迭代周期,$\boldsymbol{M}$ 在位置(i,j)处像素的更新公式为

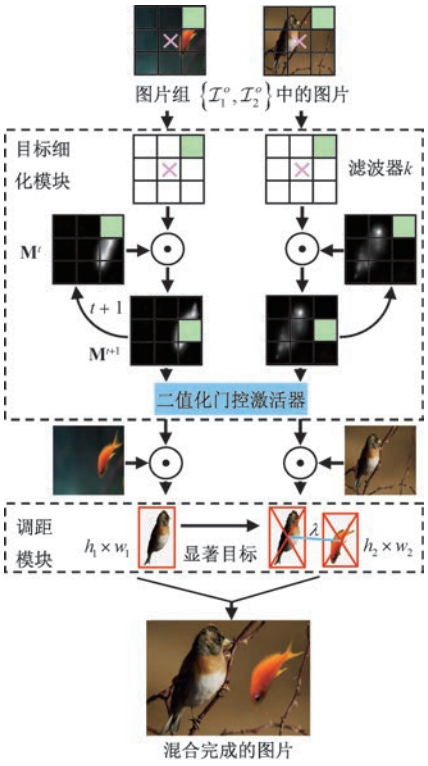

图 4　数据混合具体步骤

$$\boldsymbol{M}_{i,j}^t=\sum_{(l,n)\in\mathcal{N}(i,j)}\alpha_{i,j,l,n}\cdot\boldsymbol{M}_{l,n}^{t-1} \tag{6}$$

其中,$\mathcal{N}(i,j)$表示位置(i,j)的邻域,(l,n)表示邻域内各点.滤波器 k 定义为

$$k(I_{i,j},I_{l,n})=-\frac{|I_{i,j}-I_{l,n}|}{\sigma_{i,j}^2} \tag{7}$$

其中,$I_{i,j}$ 表示 I 在位置(i,j)处的像素值,$\sigma_{i,j}$ 为原图像素值的标准差.在此基础上,用归一化指数函数(softmax)得到(i,j)与其邻域内各点(l,n)的最终亲和值 $\alpha_{i,j,l,n}$:

$$\alpha_{i,j,l,n}=\frac{e^{\bar{k}(I_{i,j},I_{l,n})}}{\sum\limits_{(l,n)\in\mathcal{N}(i,j)}e^{\bar{k}(I_{i,j},I_{l,n})}} \tag{8}$$

其中,$\bar{k}$ 为图像在 RGB 三个通道上的平均亲和值.经过迭代细化的显著性掩膜已具有较为清晰的边缘信息,但其值分布于 0～1.为了将其二值化以更好地分割原图像,本文设计了一个二值化门控激活器.当原掩膜满足一定条件时,将其像素置为 1,否则,置为 0,公式为

$$\boldsymbol{B}_{i,j}=\begin{cases}1, & \boldsymbol{M}_{i,j}>\alpha\dfrac{\sum_{i=1,j=1}^{i=h,j=w}\boldsymbol{M}_{i,j}}{h\times w}\\ 0, & \text{otherwise}\end{cases} \tag{9}$$

其中,超参数 α 用来调节二值化门控激活器对像素的激活程度.根据表 7 中的实验效果,本文将 α 设置为 1.1,将迭代周期 t 设置为 8.

图 5 展示了目标细化模块中各阶段的可视化效果.不难发现,类激活图具备丰富的位置信息.但是,边缘信息较差.本文通过迭代细化较好地恢复出掩膜的边缘.但是,其存在目标内部响应值过低的问题,不利于分割原图中的显著目标.而掩膜进一步通过二值化门控激活器后,能在激活目标内部响应的同时,进一步滤除部分噪声,更好地二值化掩膜,以精确分割出显著目标.

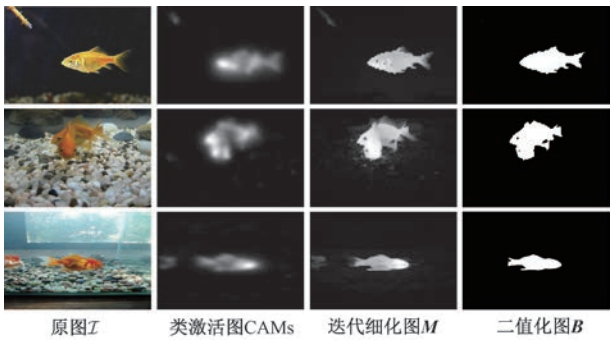

图 5　目标细化模块各阶段可视化效果图

3.1.2　调距模块

在模型得到二值掩膜 $\boldsymbol{B}$ 后,将其与原图逐通道按对应元素相乘,即可分割出显著目标.为尽可能地避免一对图中的显著目标互相遮挡,本文以目标框的形式定位显著目标,对二值化掩膜 $\boldsymbol{B}$ 分别进行从上至下和从左至右的遍历,得到上下左右四个方向上最边缘的像素位置,并将其分别往上下左右四个方向外扩 2 个像素单位定位其上下左右四条边线,即可得到显著目标的定位框,定义目标框尺寸分别为 $h_1\times w_1$,$h_2\times w_2$,定义 λ 为两个目标框中心的距离,当 $\lambda>0.8\dfrac{(h_1+h_2)^2+(w_1+w_2)^2}{2}$ 时予以拼接.对两组图片成对数据混合,最终得到混合完成数据组 $\mathcal{I}_1^m$ 和 $\mathcal{I}_2^m$,并将其输入协同显著性检测变换器 f_{trans},进行进一步的操作.

在数据混合子网络中,本文采用分类损失函数 $\mathcal{L}_{cls}$ 监督学习类别向量 $\hat{\boldsymbol{y}}_1$ 和 $\hat{\boldsymbol{y}}_2$,以更新其网络参数,

引导模型适应本任务的数据特性. $\mathcal{L}_{cls}$ 由交叉熵损失 ℓ_{ce} 组成：

$$\mathcal{L}_{cls}=\ell_{ce}(\hat{\boldsymbol{y}}_1,\boldsymbol{y}_{gt})+\ell_{ce}(\hat{\boldsymbol{y}}_2,\boldsymbol{y}_{gt}) \tag{10}$$

其中，$\boldsymbol{y}_{gt}\in\mathbb{R}^{1\times 78}$ 为类别标签，交叉熵损失 ℓ_{ce} 定义为

$$\ell_{ce}(\hat{\boldsymbol{y}}_n,\boldsymbol{y}_{gt})=-\sum_{n=1}^{N}\boldsymbol{y}_{gt}\cdot\log(\hat{\boldsymbol{y}}_n) \tag{11}$$

3.2 协同显著性检测变换器

图 2 中展示的本文变换器的整体结构主要包括四个阶段：数据混合基础上的数据读取器、编码器、任务注入器、解码器. 接下来将具体介绍后面三个结构.

3.2.1 编码器

为了节约训练开销，本文采用预训练好的 T2T-ViT[24] 模型作为主干网络. 输入一个图符序列，T2T-ViT 迭代应用 T2T 模块对序列编码. 如图 6 所示，T2T 模块由重构和展开两部分组成，以充分交互输入数据 $\mathcal{X}^0$ 的局部信息. 首先，输入数据 $\mathcal{X}^0$ 经过多头注意力（MHA，Multi-head Attention）和多层感知机（MLP，Multi-layer Perceptron）层[42]，得到序列 $\mathcal{X}^{0'}$. 然后，将 $\mathcal{X}^{0'}$ 以 s 个重叠区、p 个 0 填充（padding）的形式重构成 $k\times k$ 个特征图符 $\mathcal{X}^0_{2d}$ 以交互局部信息. 随后，将 $\mathcal{X}^0_{2d}$ 展开成序列的形式 $\mathcal{X}^0_{un}$，再经过 MHA 和 MLP 层，形成新的序列 $\mathcal{X}^1$.

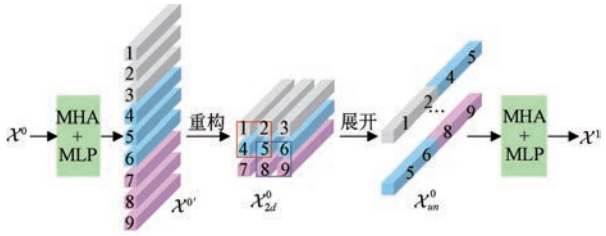

图 6 T2T 模块

T2T-ViT 中的重构和展开操作对相邻特征图符之间的局部关系进行建模，克服了 ViT 忽视充分利用局部信息的缺陷，并充分利用了空间先验信息来提升分辨率. T2T 模块可迭代多次，每次都将先前的图符序列转换为新的序列，从而充分建模所有图符中的长程依赖关系. 本文参考了文献[15]的设计：将输入图片以块状形式输入 T2T 模块，并迭代两次. 将三次裁剪的尺寸设置为 $k=[7,3,3]$，重叠区个数设置为 $s=[3,1,1]$，填充尺寸设置为 $p=[2,1,1]$. 由此，可获得多尺寸序列 $\mathcal{X}^0\in\mathbb{R}^{b\times N\times l_0\times c}$，$\mathcal{X}^1\in\mathbb{R}^{b\times N\times l_1\times c}$，$\mathcal{X}^2\in\mathbb{R}^{b\times N\times l_2\times c}$. 其中，$b$ 为组的数量，N 为每个图片组包含的图片的数量，$l_0=\frac{H}{4}\times\frac{W}{4}$，$l_1=\frac{H}{8}\times\frac{W}{8}$，$l_2=\frac{H}{16}\times\frac{W}{16}$，$c$ 为特征通道数，序列的长度不断缩短. 此外，本文参照文献[42]将余弦位置嵌入到多尺寸序列之中，以编码位置信息.

3.2.2 任务注入器

本文针对协同显著性检测任务的特性设计了组信息图符 $\mathcal{X}^G\in\mathbb{R}^{b\times N\times 1\times d}$ 和显著性图符 $\mathcal{X}^S\in\mathbb{R}^{b\times N\times 1\times d}$，其中 d 为特征通道数，设置为 384. 本任务需要发现并分割出一组图片中语义类别相同的前景显著目标，因此，如图 2 所示，本文首先将组信息图符 $\mathcal{X}^G\in\mathbb{R}^{b\times N\times 1\times d}$ 嵌入 $\mathcal{X}^2$，并利用自注意力层融合带有全局信息的组信息图符. 在此基础上，本文将 $\mathcal{X}^G$ 输入 MLP 层，以进一步交互组信息，并将交互完成的 $\mathcal{X}^G$ 嵌入回图符序列 $\mathcal{X}^4$ 以避免模型在早期阶段就丢失了对同组信息的关注. $\mathcal{X}^S$ 也经过相同的步骤处理并得到 $\mathcal{X}^6$，以引导模型在关注组信息的前提下，进一步将显著性信息融入图符序列.

图 7 所示为显著注意力层的结构图. 不同于自注意力层[42]，输入的特征序列 $\mathcal{X}_{in}$ 将被按通道维度拆分为 $\mathcal{X}_{in\_s}$ 和 $\mathcal{X}_{in\_f}$，$\mathcal{X}_{in\_s}=\mathcal{X}_{in}(:,0)$，$\mathcal{X}_{in\_f}$ 即为 $\mathcal{X}_{in}$ 余下部分. 显著性注意力层将经过全连接层处理后的 $\mathcal{X}_{in\_s}$ 作为查询（*query*）$\mathcal{Q}$，$\mathcal{X}_{in\_f}$ 经全连接层处理后作为键（key）$\mathcal{K}$ 和值（value）$\mathcal{V}$，以增强对显著性信息的建模.

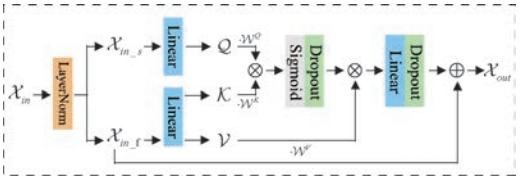

图 7 显著注意力层

图 8 为展示任务注入器有效性的示例. 在“手风琴组”中可以观察到，未注入显著性图符时，由于没有显著性信息的引导，模型对目标边缘的分割并不精确，会部分地将显著目标周围的无关区域分割出来. 未注入组信息图符时，“闹钟组”和“手风琴组”均受到了显著但非协同的对抗目标的干扰，错误地定位了协同目标，分割效果较差. 当组信息图符和显著性图符均未注入时，模型的表现最差，出现了协同目标定位错误、边缘分割模糊的问题，这表明在协同显著目标外观变化巨大、背景嘈杂等情况下，本文设计的任务注入器能很好地帮助模型定位与分割出协同显著目标.

3.2.3 解码器

由于 $\mathcal{X}^7$ 的序列较短，难以恢复出高质量的协同显著性预测结果. 一系列基于卷积神经网络的模型[10,36-37,49]都采用双线性插值的方式进行上采样，以恢复特征图. 为此，本文采用文献[15]设计的一种

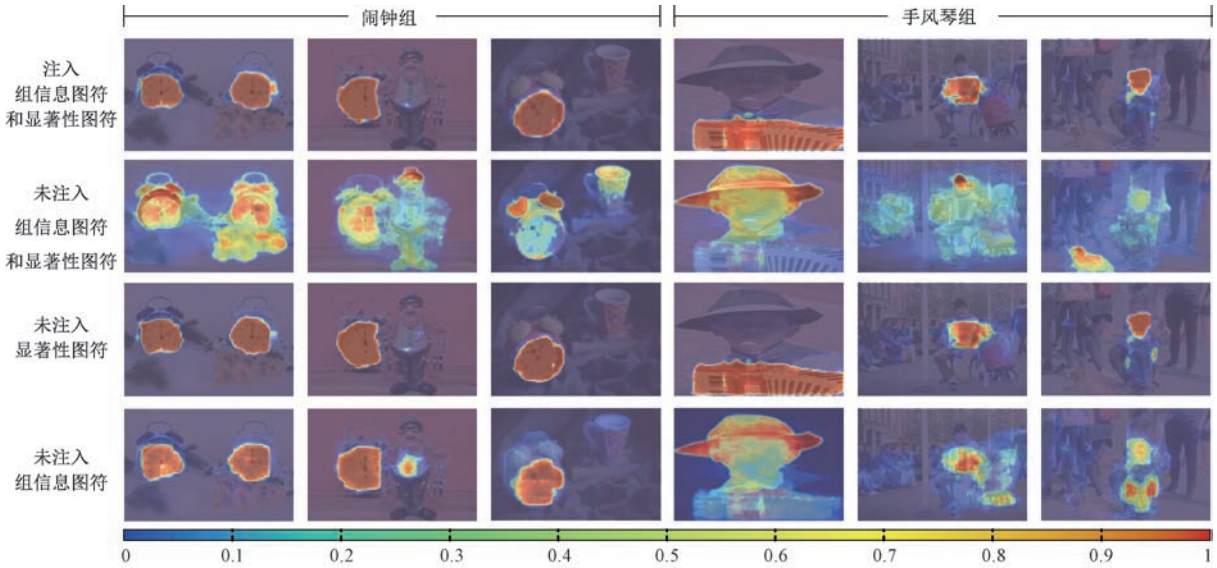

图 8　任务注入器效果图

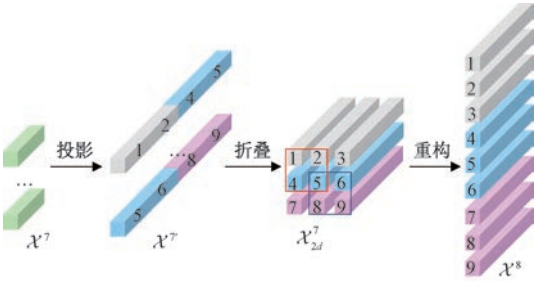

图 9　RT2T 模块

在变换器架构下的上采样方式. 图 9 展示了文献[15]在文献[24]的基础上设计的逆 T2T 模块(RT2T). RT2T 通过将每个图符序列扩展为多个子图符序列来实现上采样. 首先,将输入图符的维度从 $d=384$ 投影到 $c=64$. 然后,使用额外的全连接层将维度从 c 投影到 ck^2,之后与 T2T 模块相同,将图符视为 $k\times k$ 个特征块,交叠个数为 s. 接着,本文用 p 维 0 填充来扩充原特征序列至 $\mathcal{X}^8$. 本文借鉴了卷积神经网络中的多层特征融合机制[15],在上采样阶段融合了浅层更具结构信息的特征,以帮助网络恢复细节信息. 接着,本方法再次融合显著性图符 $\mathcal{X}^S$,输入 MLP 层并进行归一化. 此后,将从两组图像中提取的特征图符、显著性图符、组信息图符融合,经过自注意力层混合组间信息,以更好地提升模型的泛化性. 最终,在分解出显著性图符和组信息图符后,经过显著注意力层和维度转换,得到模型的输出结果 $\hat{\mathcal{O}}$.

对最终结果 $\hat{\mathcal{O}}$,本文采用显著性损失 $\mathcal{L}_{sal}$ 损失函数来监督训练过程. $\mathcal{L}_{sal}$ 由二值交叉熵损失 ℓ_{bce}[50] 和 $SIoU$ 损失 ℓ_{siou}[37] 组成:

$$\mathcal{L}_{sal}=\frac{1}{N}\sum_{n=1}^{N}\ell_{bce}(\mathcal{Y}(n,:),\hat{\mathcal{O}}(n,:))+\frac{\theta}{N}\sum_{n=1}^{N}\ell_{siou}(\mathcal{Y}(n,:),\hat{\mathcal{O}}(n,:)) \tag{12}$$

其中,$\mathcal{Y}$为图像标签集合,θ 为权重系数,根据实验结果本文将其设置为 0.3. ℓ_{bce} 损失函数定义为:

$$\ell_{bce}(\mathcal{Y}(n,:),\hat{\mathcal{O}}(n,:))=-(\mathcal{Y}(n,:))^{\top}\log(\hat{\mathcal{O}}(n,:))+(1-\mathcal{Y}(n,:))^{\top}\log(1-\hat{\mathcal{O}}(n,:))) \tag{13}$$

ℓ_{siou} 损失函数定义为

$$\ell_{siou}(\mathcal{Y}(n,:),\hat{\mathcal{O}}(n,:))=1-\frac{\mathcal{Y}(n,:)^{\top}\hat{\mathcal{O}}(n,:)}{\mathbf{1}^{\top}(\mathcal{Y}(n,:)+\hat{\mathcal{O}}(n,:))-\mathcal{Y}(n,:)^{\top}\hat{\mathcal{O}}(n,:)} \tag{14}$$

3.3　损失函数

本文联合端到端训练数据混合子网络和协同显著性变换器. 网络参数通过优化如下多任务损失函数$\mathcal{L}$得到:

$$\mathcal{L}=\lambda_1\mathcal{L}_{cls}+\lambda_2\mathcal{L}_{sal} \tag{15}$$

其中,λ_1 和 λ_2 为权重系数,根据实验效果,我们将其分别设置为 0.2 和 0.8.

4　实验结果分析

4.1　训练细节

本方法运行于 PyTorch 1.6.0[51] 框架下,使用的 CPU 是 Intel(R) Xeon(R) E5−2678 v3 @ 2.50 GHz,在一块 GeForce GTX 2080Ti GPU 上训练模型. 训

练集采用文献[52]提出的 COCO-SEG，包含 78 个类别的 20 万余张图像.每一张图都有二值化真值标注和类别标注.

在每个训练周期中，本文随机选取两组图片，调整大小至 $224\times224\times3$ 作为输入.根据训练设备的显存，本文将每组图片数量 N 设置为不大于 4 的数，N 可在设备允许的情况下设置为更大的数，组个数 b 则固定设置为 2，输入图像总数即为 $K=b\times N$.本文采用 Adam[53]优化器，动量的第一次和第二次衰减率分别设置为 0.9 和 0.99.初始学习率设置为 10e−4，在第 40 个训练周期衰减至 10e−5.方法共训练了 50 个周期，训练完成共需要 8.5 个小时.

4.2 测试集和评价指标

本文采用的基准测试数据集包括 Cosal2015[27]，CoCA[9]，以及 CoSOD3k[28].其中，Cosal2015 包含 50 类，共 2015 幅图像.这些图像在颜色、尺寸、背景、造型方面差异较大，导致 Cosal2015 具备一定挑战；CoCA 是最新的测试集，包含 80 类，共 1297 幅图片.这些图片含有极大的复杂背景干扰，导致 CoCA 是一个难度极大的数据集；CoSOD3k 包含来自 160 类的 3316 幅图片，是目前最大的基准测试集.它针对协同显著性检测任务的特性，设计了更具干扰性的对抗目标和复杂背景，使其颇具挑战性.

本文的测试指标为：用来衡量标签与预测图之间的平均像素误差的指标 MAE；基于局部像素信息差异和全局均值差异的、更注重全局信息和局部信息统一的指标 E_{ϕ}^{max}；用来衡量标签和预测图之间的结构相似性指标 S_{α}；描述测试值在准确率和召回率之间平衡程度的指标 F_{β}^{max}：$F_{\beta}=\dfrac{(1+\beta^2)Precision\times Recall}{\beta^2\times Precision+Recall}$，$\beta^2$ 一般设置为 0.3.

4.3 测试集和评价指标

本文采用文献[1]开源的测评代码与多个先进方法进行了比较.对比方法包括 CBCD[32]，DIM[54]，GW[7]，RCAN[8]，CSMG[35]，SSNM[55]，GCAGC[36]，GICD[9]，ICNet[56]，GCoNet[37]，DeepACG[10]，CADC[13]，DCFM[12].

4.3.1 可视化结果比较

图 10 展示了本文方法与其他方法在三个测试数据集上采样的可视化对比结果.对比方法包括三个先进方法 DeepACG[10]，CADC[13]，DCFM[12].本文方法面对背景杂乱、目标尺寸小、目标遮挡、目标外观差异巨大、存在对抗目标或训练分布外的目标等挑战时，展现出了优于其它方法的优异性能.

在 CoCA 数据集中的“手风琴组”和“蝴蝶组”皆为未知类别目标，且存在目标遮挡(“手风琴组”的第三列、第五列和第六列)、尺寸差异大(“手风琴组”的第一列、第二列与第三列、第四列、第五列)、目标与背景颜色相近(“手风琴组”的第四列、“蝴蝶组”的第二列)、存在干扰显著目标(“手风琴组”的第一列、第二列、第三列和“蝴蝶组”的全列)的问题，使得模型很难准确分割出协同显著目标.在应对目标存在遮挡时，CADC、DCFM 和 DeepACG 都不能将协同显著目标与遮挡物很好地分割出来(见“手风琴组”的第三列、第五列和第六列)；在协同显著目标尺寸差异大的情况下，其它方法都错误地分割出远大于显著目标的区域(见“手风琴组”的第一列和第二列)；在面对目标与背景颜色相近的情况时，DCFM 和 DeepACG 错误地将目标周围的区域分割出来(见“蝴蝶组”第二列)；当测试样本存在对抗目标时，其它方法都不能准确分割出协同显著目标：存在误检情况，会把对抗目标的全部或部分分割出来(见“手风琴组”的第一列、第二列、第三列和“蝴蝶组”的全列).

在 CoSOD3k 数据集中的“篮球组”和“蜜蜂组”皆为训练分布外的目标，同时有协同显著目标尺寸小(“篮球组”第二至第五列、“蜜蜂组”第三至第六列)、存在对抗目标(“篮球组”全列和“蜜蜂组”全列)、目标与背景颜色相近(“蜜蜂组”第三列和第五列)的挑战，对模型的性能提出了更高要求.当协同显著目标尺寸很小时，其它方法均不能准确分割出显著目标，而会将显著目标周围的无关信息一同错误分割出来(见“篮球组”全列)；当存在对抗目标时，其它方法会出现对协同显著目标的误判(见“篮球组”全列和“蜜蜂组”全列)；当目标与背景颜色很相近时，其它方法会错误地将显著区域扩展至背景区域(见“蜜蜂组”全列).另外，其它方法亦存在检测不到显著目标的情况(见“篮球组”CADC 结果的第二列和第三列，DeepACG 结果的第一列、第三列至第五列).

在 Cosal2015 数据集中的“咖啡杯组”和“青蛙组”，存在细节较多的协同显著目标(“咖啡杯组”全列)和伪装的协同显著目标(“青蛙组”全列)，同时还有部分目标存在干扰目标(“咖啡杯组”第一列、第六列).DCFM 和 DeepACG 在恢复协同显著目标的细节部分时效果较差，不能很好地分割出咖啡杯的杯把(见“咖啡杯组”第三列和第五列)；在面对伪装目

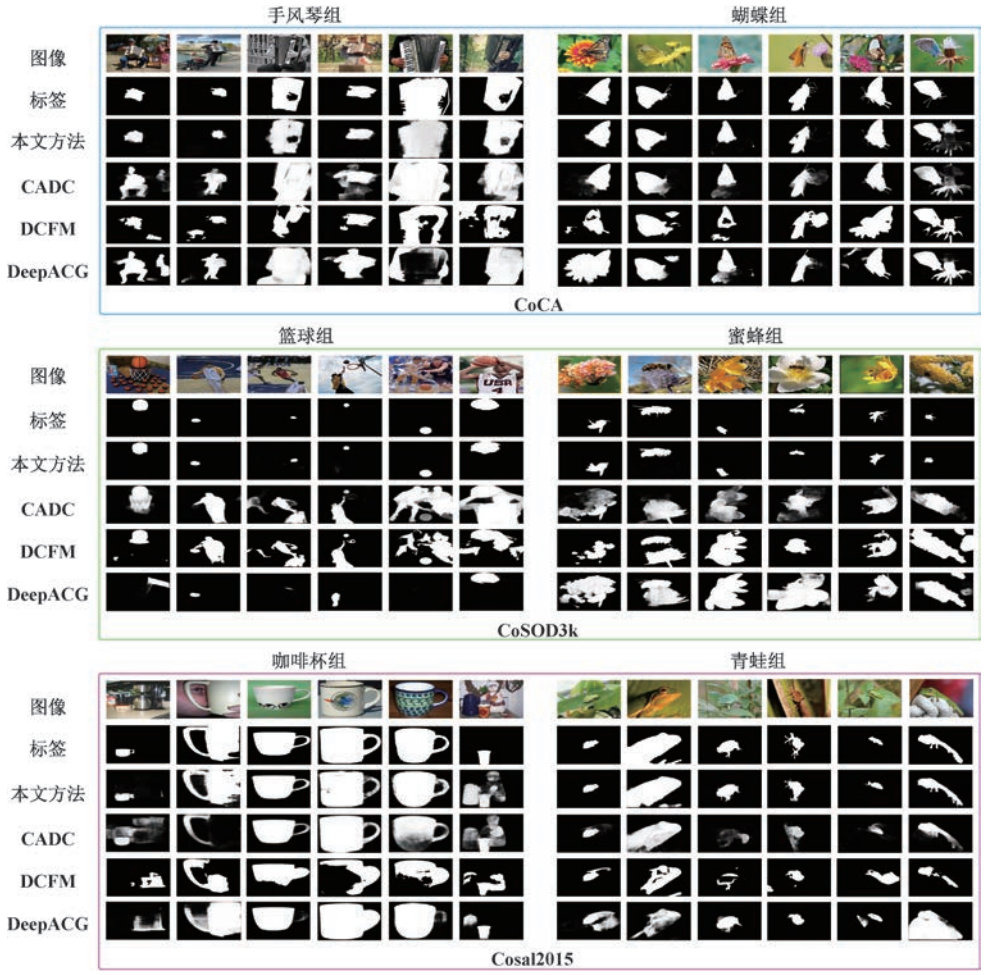

图 10　本文方法与其他先进方法的可视化结果比较

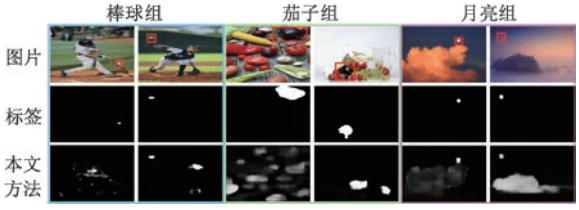

图 11　本文方法的失败案例

标时,CADC 存在丢失目标的情况(见“青蛙组”第三列);存在干扰目标时,其他方法都会错误地分割出干扰目标(见“咖啡杯组”第一列).

本文方法利用数据混合构造新的训练样本以增强模型处理对抗样本的能力,并使模型能更好地泛化至训练分布外的数据.同时,本方法采用变换器的架构以更好地利用全局信息,使组信息更好地融合到每一幅图像中,以突出协同目标.因此本文模型具有更好的鲁棒性、泛化性以及判别力,能取得最好的检测效果.

如图 11 所示,本文亦展示了在面对一些极困难样本时的失败案例.在面对协同显著目标特别小的情况时(见“棒球组”),由于目标过于细小,本文数据混合的方法较难针对性地提升模型面对该类样本时的性能,会出现失误,不能完全准确地将其分割;在面对干扰目标极多的情况时(见“茄子组”),由于目标与干扰目标存在交错、重叠,外观也极其相似,本文方法会在分割出协同显著目标的同时将部分干扰目标一同分割,不能做到完全准确.显著目标与背景或干扰物颜色十分接近或没有明确边缘时(见“月亮组”),本文的数据混合方法和任务注入器都很难把握主要目标,容易错把语义信息近似的干扰目标当成协同显著目标.尽管本文方法能取得不错的效果,但在一些困难样本上出现失误,存在改进空间.

4.3.2　测试指标比较

图 12 展示了本文方法和所有参与比较的方法在三个基准测试集上的 PR 曲线及 F-measure 曲线结果.从图中可以发现本文方法取得了最佳的性能.在所有的图中,本文方法的曲线都在最上端.

表 1 列出了所有参与比较的方法的指标.其中,加粗的为最优指标,加下划线的为次优指标.表 1 中的 CADC 和 DCFM 为最新提出的先进方法.本文

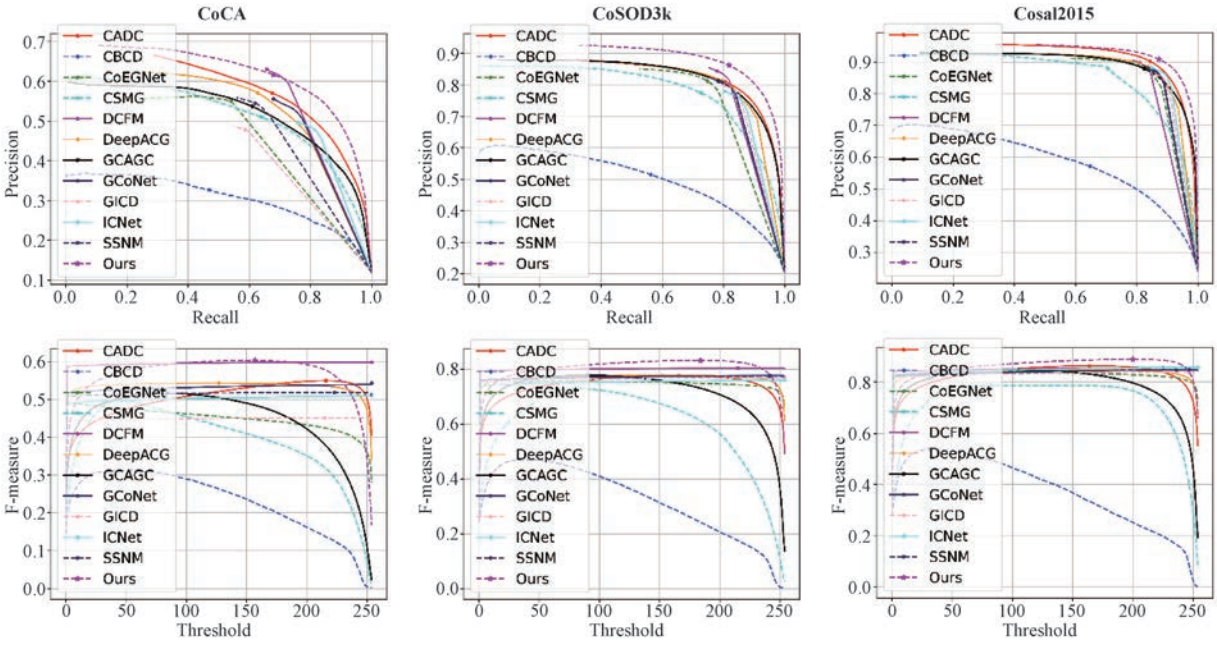

图 12 本文方法与其它先进方法在三个基准测试集上 PR 曲线和 F-measure 曲线的比较

在三个数据集的 11 个指标上取得了最优结果，在 1 个指标上取得了次优结果. 具体的，在 Cosal2015 数据集上，本文方法的 MAE，S_α，E_ϕ^{max}，F_β^{max} 分别为 0.049，0.890，0.938，0.890，大幅领先于次优方法 1.5%，2.4%，3.2%，2.8%；在 CoSOD3k 数据集上，本文方法的四个指标分别为 0.619，0.856，0.911，0.833，大幅领先于次优方法 0.5%，4.5%，3.7%，2.8%；在 CoCA 数据集上，本文方法的四个指标分别为 0.097，0.729，0.811，0.603. 其中，仅有 MAE 指标是次优结果，略低于最优结果 1.2%. 但是，其它指标均为最优结果且优于次优方法 1.9%，2.8%，0.5%. 这表明本方法有优异的性能. 为了对模型复杂度进行测评，我们将输入图片数设置为 4 对部分开源代码的模型的浮点运算数(FLOPs)、参数量、推理速度进行了测试与比较. 表 2 展示了本文方法与部分最新方法在算法复杂度上的差异及模型在 CoCA 数据集上的 F_β^{max} 指标. 从浮点运算数与参数量可以看出，相较于其他方法，本文方法的复杂度最低. 尽管本文的推理速度仅快于 CADC，但 20.4fps 足以满足实时需求，且本文方法能达到最佳性能.

4.4 消融实验

为了验证本方法各关键部分的有效性，本文在文献[15]的基础上加以改动，并将其作为基线(baseline)方法，在三个数据集上都针对数据混合子网络、任务注入器、数据混合子网络各部分、交叉熵损失函数以及部分参数做了消融实验.

4.4.1 主体消融实验

为了验证数据混合策略和任务注入器的有效性，本文在三个数据集上做了消融实验. 表 3 展示了本方法主要部分在三个数据集上的消融实验结果. 通过在 Cosal2015 上的实验可以看出，本文设计的数据混合方法对指标 S_α、F_β^{max} 和 E_ϕ^{max} 均有较大提升，仅有指标 MAE 略有下降. 这是因为数据混合为训练样本提供了对抗目标，在一定程度上会影响模型对协同显著目标细节的恢复，所以衡量平均像素误差的 MAE 会略有下降. 同时，衡量结构相似性的 S_α 和衡量准确率与召回率平衡程度的 F_β^{max} 分别得到了 1.5%和 1.9%的提升，这展示了数据混合策略能有效提升模型的泛化性能，能更好地处理训练分布外的数据；本文设计的任务注入器使 MAE 得到了 1.3%的大幅提升，且其余三个指标也得到了提升. 其中，衡量局部像素信息差异和全局像素均值差异的 E_ϕ^{max} 更是提升了 4.7%，这表明组信息图符能很好地将组特征注入显著性图符，增强了模型的判别力，使模型能够更加准确地分割协同显著目标. 通过在 CoSOD3k 上的实验，可以发现，本文的数据混合策略会在一定程度上影响 MAE，但对其余三个指标有积极作用. 尤其是 S_α 和 F_β^{max} 有大幅提升，分别提升了 1.7%和 3.3%. 由于数据集 CoSOD3k 相较于 Cosal2015 有更多的干扰目标和更为杂乱的背景，得益于数据混合，模型在泛化性和鲁棒性上的提升更为明显. 本文设计的任务注入器能有效提升模型对协同显著目标的细化能力和检测准确率，使

表 1　本文方法与其他先进方法在三个数据集上测试指标的比较

方法	Cosal2015				CoSOD3k				CoCA			
	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{\max}\uparrow$	$F_\beta^{\max}\uparrow$	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{\max}\uparrow$	$F_\beta^{\max}\uparrow$	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{\max}\uparrow$	$F_\beta^{\max}\uparrow$
CBCD (TIP2013)	0.233	0.544	0.656	0.503	0.228	0.528	0.589	0.363	0.172	0.526	0.659	0.313
DIM (TNNLS2016)	0.312	0.593	0.697	0.559	0.327	0.559	0.610	0.420	—	—	—	—
GW (IJCAI2019)	0.147	0.743	0.793	0.697	0.147	0.716	0.777	0.649	0.166	0.602	0.701	0.408
RCAN (IJCAI2019)	0.126	0.779	0.842	0.764	0.130	0.744	0.808	0.688	0.160	0.616	0.702	0.422
CSMG (CVPR2019)	0.130	0.774	0.818	0.777	0.157	0.711	0.723	0.645	0.124	0.632	0.734	0.503
SSNM (AAAI2020)	0.102	0.788	0843	0.794	0.120	0.726	0.756	0.675	0116	0.628	0.741	0.482
GCAGC (CVPR2020)	0.085	0.817	0.866	0.813	0.100	0.785	0.816	0.740	0.118	0.669	0.754	0.523
GICD (ECCV2020)	0.072	0.842	0.884	0.834	0.089	0.794	0.831	0.743	0.125	0.658	0.701	0.504
ICNet (NIPS2020)	0.058	0.857	0.900	0.858	0.089	0.794	0.845	0.762	0.147	0.654	0.705	0.514
CoEGNet (TPAMI2021)	0.077	0836	0.882	0.832	0.092	0.762	0.825	0.736	0.106	0.612	0.717	0.493
GCoNet (CVPR2021)	0.069	0.845	0.887	0.847	0.071	0.802	0.860	0.750	0.105	0.673	0.760	0.524
DeepACG (CVPR2021)	0.066	0.853	0.893	0.847	0.079	0.811	0.859	0.779	0.104	0.685	0.759	0.564
CADC (ICCV2021)	0.064	0.866	0.906	0.862	0.096	0.801	0.840	0.759	0.132	0.681	0.744	0.548
DCFM (CVPR2022)	0.067	0.838	0.892	0.856	0.067	0.810	0.874	0.805	**0.085**	0.710	0.783	0.598
本文方法	**0.049**	**0.890**	**0.938**	**0.890**	**0.062**	**0.856**	**0.911**	**0.833**	0.097	**0.728**	**0.811**	**0.603**

表 2　本文模型与其他模型复杂度比较

方法	FLOPs/G	参数量/M	推理速度/fps	$F_\beta^{\max}\uparrow$
GICD (ECCV2020)	364.7	278.8	40.8	0.504
GCoNet (CVPR2021)	259.9	142.0	116.2	0.524
CADC (ICCV2021)	330.0	392.8	18	0.548
DCFM (CVPR2022)	251.9	142.3	84.4	0.598
本文方法	234.0	58.3	20.4	0.603

MAE 得到了 1.4%的提升. 通过在 CoCA 上的测试可以发现,杂乱的背景和繁多的干扰目标使 CoCA 在三个数据集中最具挑战性. 本文的数据混合策略使模型能应对对抗目标的干扰, S_α 和 F_β^{max} 分别取得了 3.3%和 1.6%的提升. 在困难测试集上,数据混合策略的有效性进一步得到验证. 同时,任务注入器引导组信息与显著性特征融合,能有效改善模型性能,对四个指标分别提升了 3.4%,2.1%,3.6%,1.3%,证明了本文所提方法的有效性.

4.4.2　数据混合中关键步骤的有效性

为了验证本文数据混合方法中关键步骤的有效性,如表 4 所示,本文在三个数据集上针对迭代细化模块、二值化门控函数和调距模块做了消融实验. 结果显示,简单地应用迭代细化模块或者二值化门控函数都会损失模型的性能. 以在最具挑战性的 CoCA 上的结果为例,在单独运用迭代细化模块时,指标 MAE, E_ϕ^{max} 分别降低了 3.4%,2.2%. 在单独运用二值化门控函数时,四个指标分别降低了 2.8%,3.3%,3.2%,0.3%. 在同时应用三个策略中的两种时,都不能达到最佳性能,尤其是同时应用二值化门控激活器和调距模块时,四个指标距离最优性能分别体现出了 3.6%,4.4%,7.1%,7.2%的差距,在没有应用迭代细化操作时,训练数据会出现非常多的噪声,导致模型效果变差. 同时应用迭代细化操作和调距模块时, E_ϕ^{max} 达到了最优,其余三个指标距离最佳性能的差距为 0.9%,0.1%,0.6%. 同时应用迭代细化操作和二值化门控激活器时,四个指标与最佳性能的差距为 1.4%,0.9%,1.3%,0.6%.

表 3 本文方法与其他先进方法在三个数据集上测试指标的比较

策略		Cosal2015				CoSOD3k				CoCA			
①	②	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{\max}\uparrow$	$F_\beta^{\max}\uparrow$	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{\max}\uparrow$	$F_\beta^{\max}\uparrow$	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{\max}\uparrow$	$F_\beta^{\max}\uparrow$
		0.058	0.862	0.873	0.863	0.073	0.831	0.873	0.795	0.123	0.679	0.751	0.569
√		0.062	0.877	0.911	0.882	0.078	0.848	0.881	0.828	0.129	0.712	0.788	0.585
	√	0.045	0.867	0.920	0.871	0.059	0.840	0.894	0.811	0.089	0.700	0.787	0.582
√	√	0.049	**0.890**	**0.938**	**0.890**	0.062	**0.856**	**0.911**	**0.833**	0.097	**0.728**	**0.811**	**0.603**

注：策略①是数据混合，策略②是任务注入器

表 4 数据混合中关键步骤的有效性

策略			Cosal2015				CoSOD3k				CoCA			
①	②	③	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{\max}\uparrow$	$F_\beta^{\max}\uparrow$	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{\max}\uparrow$	$F_\beta^{\max}\uparrow$	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{\max}\uparrow$	$F_\beta^{\max}\uparrow$
			0.045	0.867	0.920	0.871	**0.059**	0.840	0.894	0.811	**0.089**	0.700	0.787	0.582
√			0.066	0.857	0.881	0.863	0.074	0.801	0.883	0.761	0.123	0.712	0.765	0.585
	√		0.073	0.849	0.879	0.855	0.096	0.823	0.874	0.777	0.117	0.667	0.755	0.579
		√	0.095	0.840	0.862	0.812	0.101	0.810	0.845	0.748	0.146	0.673	0.734	0.534
	√	√	0.087	0.841	0.867	0.833	0.078	0.813	0.850	0.751	0.133	0.685	0.740	0.531
√		√	0.057	0.878	0.915	0.860	0.070	0.837	0.904	0.821	0.106	0.728	**0.814**	0.591
√	√		0.052	0.885	0.911	0.864	0.061	0.849	0.900	0.827	0.111	0.720	0.801	0.597
√	√	√	0.049	**0.890**	**0.938**	**0.890**	0.062	**0.856**	**0.911**	**0.833**	0.097	**0.729**	0.811	**0.603**

注：策略①是迭代细化操作，策略②是二值化门控激活器，策略③是调距模块

可以看出，迭代细化操作能有效剔除噪声，二值化门控激活器则可以在此基础上进一步分割目标，有效提升模型性能. 在此基础上，调距模块可以最小化目标重叠，提升数据混合的有效性，进一步提升模型性能. 需要指出的是，数据混合会降低 MAE，因为混合后的数据在一定程度上使模型不只是完全关注于协同显著目标，在分割协同显著目标的细节上有所损失. 但模型的泛化性能得到了改善，在其余三个关键指标上取得了大幅提升.

4.4.3 交叉熵损失函数的有效性

为了验证本文数据混合子网络中交叉熵损失函数的有效性，本文将交叉熵损失函数权重 λ_1 固定为 0.2，对其做了消融实验. 如表 5 所示，使用交叉熵损失函数能使模型更好地在当前任务类型的数据上进行泛化，将四个指标分别提高了 1.2%、3.7%、2.0%、2.9%. 图 13 展示了没有交叉熵损失时模型关注区域的变化. 可以观察到，当不使用交叉熵损失函数监督训练过程时，模型会关注单幅图像上更大的区域，而不关注协同目标所在的区域.

图 13 交叉熵损失函数的效果图

4.4.4 参数分析

本文针对公式(6)中迭代周期 t 的取值对模型性能的影响，做了相应的消融实验，从表 6 可以观察到，迭代次数的增加能提升模型的性能，但当迭代次数达到 8 次之后，继续增加迭代次数对模型性能的提升十分有限，且会增加额外的计算开销，据此，我们将迭代次数 t 设定为 8 次.

表 5 交叉熵损失函数在 CoCA 上的实验

交叉熵损失函数	CoCA			
	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{\max}\uparrow$	$F_\beta^{\max}\uparrow$
	0.109	0.692	0.791	0.574
√	0.097	0.729	0.811	0.603

表 6 不同 t 值在 CoCA 上的消融实验

t 取值	CoCA			
	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{\max}\uparrow$	$F_\beta^{\max}\uparrow$
6	0.100	0.719	0.798	0.597
7	0.102	0.722	0.811	0.596
8	0.097	0.729	0.805	0.603
9	0.097	0.727	0.809	0.602
10	0.096	0.730	0.802	0.602

为了验证公式(9)中不同 α 取值对模型性能的影响，本文针对不同的 α 取值在 CoCA 上做了对比实验. 如表 7 所示，当 α 取得较小时，四个指标均有所下降，这是因为较小的 α 会导致更多的背景信息被错误分割，引入了更多的无效信息，使模型的训练效率下降. 当 α 取得较大时，会导致显著目标不能被完整分割，会使构造的新训练样本无法包含完整的

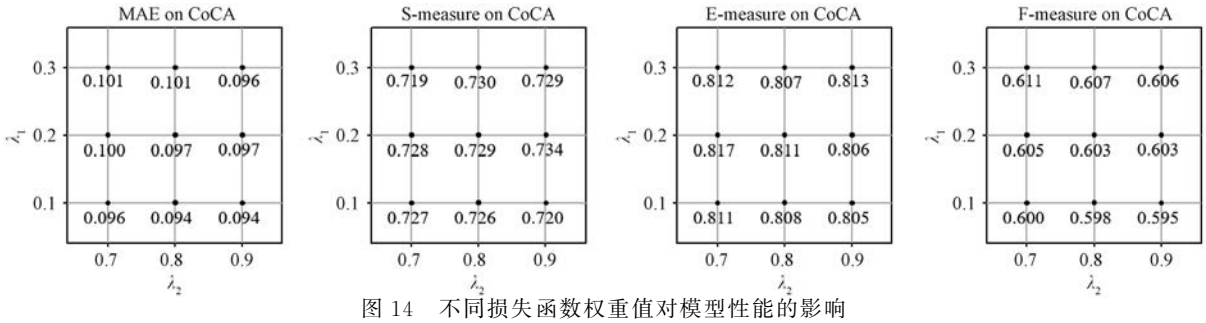

图 14　不同损失函数权重值对模型性能的影响

表 7　不同 α 值在 CoCA 上的消融实验

α 取值	CoCA			
	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{\max}\uparrow$	$F_\beta^{\max}\uparrow$
0.8	0.111	0.701	0.791	0.591
0.9	0.104	0.717	0.787	0.592
1.0	0.104	0.710	0.804	0.599
1.1	0.097	0.729	0.811	0.603
1.2	0.096	0.722	0.813	0.589
1.3	0.098	0.727	0.802	0.593

对抗目标.实验表明当 α 取 1.1 时,模型性能达到最佳.

图 14 展示了公式(15)中 λ_1 和 λ_2 的不同取值对模型性能的影响.根据对任务的侧重,本文将 λ_1 设置为 0.2,将 λ_2 设置为 0.8,并各额外取两组参数进行参数分析.从实验中可以观察到,四个指标的最大波动分别为 0.7%,1.4%,1.2%,1.6%,四个指标的变化趋势没有统一的规律且 λ_1 取 0.2,λ_2 取 0.8 时模型性能已十分出色,因此本文将 λ_1 定为 0.2,λ_2 定为 0.8.

在固定显著性损失函数 $\mathcal{L}_{sal}$ 的权重系数 λ_2 为 0.8 的前提下,本文针对公式(12)中不同的 θ 取值做了消融实验.从图 15 中可以观察到,θ 取 0.3 时,模型性能达到最佳,当 θ 取值过大时,模型会过度关注协同显著目标的位置准确性,而极大地忽视目标的边缘等细节准确性,会严重影响模型性能.

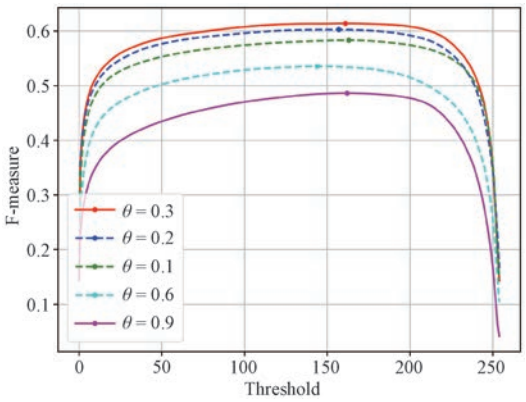

图 15　不同 Θ 取值的 F-measure 曲线比较

5　结　论

本文提出了一种新颖的基于组间对抗数据混合的协同显著性检测变换器,旨在通过视觉变换器构建序列到序列的协同显著性检测网络,并使用组间混合后的数据进行对抗训练,以增强模型的泛化性.本文在数据混合子网络中集成了目标细化模块和调距模块,前者以类激活图 CAMs 作为引导,以无监督的方式分割出平滑边缘的显著目标,以其作为对抗对象细化显著目标掩膜.后者将对抗对象以最小化重叠的方式混合至另一组图中,生成混合训练数据;在协同显著性检测变换器中,本文从序列角度建模,设计了任务注入器,将组信息图符与显著性信息图符注入至序列特征之中,并通过自注意力机制充分捕获特征间的全局上下文信息.最终将组特征与显著性特征经由自注意力机制混合,进一步增强特征的表征力,生成高质量协同显著性检测结果.本文在包含 Cosal2015,CoCA,以及 CoSOD3k 等三个标准测试集上的结果充分验证了本文所提算法的优越性.

参　考　文　献

[1] Fan D P, Li T, Lin Z, et al. Re-thinking co-salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(8): 4339-4354

[2] Liu J J, Hou Q, Liu Z A, et al. Poolnet+: Exploring the potential of pooling for salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 887-904

[3] Wang W, Shen J, Yang R, et al. Saliency-aware video object seg-mentation. IEEE Transactions on Pattern Analysis and Machine intelligence, 2017, 40(1): 20-33

[4] Cheng M M, Mitra N J, Huang X, et al. Salientshape: group saliency in image collections. The Visual Computer, 2014, 30(4):443-453

[5] Piao Y, Jiang Y, Zhang M, et al. Panet: Patch-aware network for light field salient object detection. IEEE Transactions

on Cybernetics, 2023, 53(1): 379-391

[6] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[7] Wei L, Zhao S, Bourahla O E F, et al. Group-wise deep co-saliency detection//International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 3041-3047

[8] Li B, Sun Z, Tang L, et al. Detecting robust co-saliency with recurrent co-attention neural network//International Joint Conference on Artificial Intelligence. Macao, China, 2019: 818-825

[9] Zhang Z, Jin W, Xu J, et al. Gradient-induced co-saliency detection//European Conference on Computer Vision. Glasgow, UK, 2020: 455-472

[10] Zhang K, Dong M, Liu B, et al. Deepacg: Co-saliency detection via semantic-aware contrast gromov-wasserstein distance//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13703-13712

[11] Vapnik V N. An overview of statistical learning theory. IEEE Transactions on Neural Networks, 1999, 10(5): 988-999

[12] Yu S, Xiao J, Zhang B, et al. Democracy does matter: Compre-hensive feature mining for co-salient object detection// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 979-988

[13] Zhang N, Han J, Liu N, et al. Summarize and search: Learning consensus-aware dynamic convolution for co-saliency detection// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 4167-4176

[14] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization//International Conference on Learning Representations. Vancouver, Canada, 2018: 1-13

[15] Liu N, Zhang N, Wan K, et al. Visual saliency transformer// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 4722-4732

[16] Han J, Chen H, Liu N, et al. Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. IEEE Transactions on Cybernetics, 2017, 48(11): 3171-3183

[17] Liu N, Han J, Yang M H. Picanet: Learning pixel-wise contextual attention for saliency detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 3089-3098

[18] Liu N, Zhang N, Han J. Learning selective self-mutual attention for rgb-d saliency detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 13756-13765

[19] Verma V, Lamb A, Beckham C, et al. Manifold mixup: Better representations by interpolating hidden states//International Conference on Machine Learning. Long Beach, USA, 2019: 6438-6447

[20] Yun S, Han D, Oh S J, et al. Cutmix: Regularization strategy to train strong classifiers with localizable features//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 6023-6032

[21] Kim J H, Choo W, Song H O. Puzzle mix: Exploiting saliency and local statistics for optimal mixup//International Conference on Machine Learning. Vienna, Austria, 2020: 5275-5285

[22] Kim J H, Choo W, Jeong H, et al. Co-mixup: Saliency guided joint mixup with supermodular diversity//International Conference on Learning Representations. Vienna, Austria, 2021: 1-21

[23] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale //International Conference on Learning Representations. Vienna, Austria, 2021: 1-21

[24] Yuan L, Chen Y, Wang T, et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 558-567

[25] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 10012-10022

[26] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2921-2929

[27] Song H, Liu Z, Xie Y, et al. Rgbd co-saliency detection via bagging-based clustering. IEEE Signal Processing Letters, 2016, 23(12): 1722-1726

[28] Fan D P, Lin Z, Ji G P, et al. Taking a deeper look at co-salient object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 2919-2929

[29] Li Y, Fu K, Liu Z, et al. Efficient saliency-model-guided visual co-saliency detection. IEEE Signal Processing Letters, 2014, 22(5): 588-592

[30] Gabor D. Theory of communication. part 1: The analysis of information. Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering, 1946, 93(26): 429-441

[31] Lowe D G. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91-110

[32] Fu H, Cao X, Tu Z. Cluster-based co-saliency detection. IEEE Transactions on Image Processing, 2013, 22(10): 3766-3778

[33] Jacobs D E, Goldman D B, Shechtman E. Cosaliency: Where people look when comparing images//Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology. New York, USA, 2010: 219-228

[34] Jiang B, Jiang X, Tang J, et al. Co-saliency detection via a general optimization model and adaptive graph learning. IEEE Transactions on Multimedia, 2020(23): 3193-3202

[35] Zhang K, Li T, Liu B, et al. Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3095-3104

[36] Zhang K, Li T, Shen S, et al. Adaptive graph convolutional network with attention graph clustering for co-saliency detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 9050-9059

[37] Fan Q, Fan D P, Fu H, et al. Group collaborative learning for co-salient object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12288-12298

[38] Zhang Dong-Ming, Jin Guo-Qing, et al. Salient Object Detection Based on Deep Fusion of Hand-Crafted Features. Chinese Journal of Computers, 2019, 42(09): 2076-2086. (in Chinese)
(张冬明,靳国庆,代锋,等. 基于深度融合的显著性目标检测算法. 计算机学报, 2019, 42(09):2076-2086.)

[39] Chen Bing-Cai, Tao Xin, Chen Hui, et al. Saliency detection via fusion of boundary connectivity and local. Chinese Journal of Computers, 2020, 43(1): 16-28. (in Chinese)
(陈炳才, 陶鑫, 陈慧, 等. 融合边界连通性与局部对比性的图像显著性检测. 计算机学报, 2020, 43(1): 16-28.)

[40] Bishop C M. Training with noise is equivalent to tikhonov regularization. Neural Computation, 1995, 7(1): 108-116

[41] Devries T, Taylor G W. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv: 1708.04552, 2017

[42] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017(30): 1-11

[43] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers//European Conference on Computer Vision. Glasgow, UK:Springer,2020: 213-229

[44] Wang H, Zhu Y, Adam H, et al. Max-deeplab: End-to-end panoptic segmentation with mask transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 5463-5474

[45] Wang W, Xie E, Li X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 568-578

[46] Piao Y, Wang J, Zhang M, et al. Mfnet: Multi-filter directive network for weakly supervised salient object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 4136-4145

[47] Araslanov N, Roth S. Single-stage semantic segmentation from image labels//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 4253-4262

[48] Su H, Jampani V, Sun D, et al. Pixel-adaptive convolutional neural networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 11166-11175

[49] Li T, Zhang K, Shen S, et al. Image co-saliency detection and instance co-segmentation using attention graph clustering based graph convolutional network. IEEE Transactions on Multimedia, 2022(24): 492-505

[50] Zhu J, Liao S, Yi D, et al. Multi-label cnn based pedestrian attribute learning for soft biometrics//2015 International Conference on Biometrics. Phuket, Thailand, 2015: 535-540

[51] Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 2019(32): 1-12

[52] Wang C, Zha Z J, Liu D, et al. Robust deep co-saliency detection with group semantic//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 8917-8924

[53] Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv: 1412.6980, 2014

[54] Zhang D, Han J, Han J, et al. Cosaliency detection based on intra saliency prior transfer and deep intersaliency mining. IEEE Transactions on Neural Networks and Learning Systems, 2015, 27(6): 1163-1176

[55] Zhang K, Chen J, Liu B, et al. Deep object co-segmentation via spatial-semantic network modulation//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 12813-12820

[56] Jin W D, Xu J, Cheng M M, et al. Icnet: Intra-saliency correlationnetwork for co-saliency detection. Advances in Neural Information Processing Systems, 2020(33): 18749-1875

WU Yang, master. His research interests include co-saliency detection.

SONG Hui-Hui, Ph. D., professor. Her research interests include video object segmentation and image super-resolution.

ZHANG Kai-Hua, Ph. D., professor. His research interests include co-saliency detection and visual tracking.

CHEN Hu, Ph. D., associate professor. His research interests include computer vision and medical image processing.

LIU Qing-Shan, Ph. D., professor. His research interests include video content analysis and understanding.

Background

Co-saliency detection aims to find and segment out the salient objects with the same semantic information in a group of images. In this paper, we propose the intra-group adversarial mixup for co-saliency detection transformer. As far as we know, our method is the first pure vision transformer design in co-saliency detection field. The network mainly has two innovative designs: a mixup sub-network to produce new training data and a co-salient transformer with the task injector. In the mixup sub-network, we design an object refinement module, using the CAMs as guidance to segment the salient objects with smoothing edges. We further design a distance adjusting module to mix the adversarial objects, producing new training samples. In the transformer, we design a task injector, which can inject group information and saliency information into the feature sequence, and we adopt self-attention to fully capture global information between features, enhancing the representational ability of the feature.

Experiments are carried out on three benchmark datasets, include Cosal2015, CoSOD3k and CoCA. The results show the excellent performance of our methods. The F_{β}^{max} is increased by 2.8%, 2.8%, 0.5% on three datasets compared with the second best-performing methods, showing the leading performance of our method.