中图分类号: 学校代码: 10055
UDC: 密级: 公开

南開大學

硕 士 学 位 论 文

视觉问题中弱监督信号的探索与利用

Exploring and Utilizing Weak Supervision in Computer Vision Problems

论文作者 张宇 指导教师 王恺 副教授
申请学位 工学硕士 培养单位 计算机学院
学科专业 计算机科学与技术 研究方向 计算机视觉
答辩委员会主席 杨巨峰 教授 评 阅 人 匿名评审

南开大学研究生院
二〇二一年五月

南开大学学位论文使用授权书

本人完全了解《南开大学关于研究生学位论文收藏和利用管理办法》关于南开大学 (简称“学校”) 研究生学位论文收藏和利用的管理规定，同意向南开大学提交本人的学位论文电子版及相应的纸质本。

本人了解南开大学拥有在《中华人民共和国著作权法》规定范围内的学位论文使用权，同意在以下几方面向学校授权。即：

1. 学校将学位论文编入《南开大学博硕士学位论文全文数据库》，并作为资料在学校图书馆等场所提供阅览，在校园网上提供论文目录检索、文摘及前 16 页的浏览等信息服务；

2. 学校可以采用影印、缩印或其他复制手段保存学位论文；学校根据规定向教育部指定的收藏和存档单位提交学位论文；

3. 非公开学位论文在解密后的使用权同公开论文。

本人承诺：本人的学位论文是在南开大学学习期间创作完成的作品，并已通过论文答辩；提交的学位论文电子版与纸质本论文的内容一致，如因不同造成不良后果由本人自负。

本人签署本授权书一份（此授权书为论文中一页），交图书馆留存。

学位论文作者暨授权人 (亲笔) 签字：_____

20　　年　　月　　日

南开大学研究生学位论文作者信息

论 文 题 目	视觉问题中弱监督信号的探索与利用				
姓　　名	张宇	学号	2120180486	答辩日期	2021 年 05 月 12 日
论 文 类 别	博士 ☐　学历硕士 ☑　专业学位硕士 ☐　同等学力硕士 ☐　划☑选择				
学院（单位）	计算机学院		学科/专业 (专业学位) 名称		计算机科学与技术
联 系 电 话	15619271627		电子邮箱	zhangyuygss@gmail.com	
通讯地址 (邮编)：300071					
非公开论文编号			备注		

注：本授权书适用我校授予的所有博士、硕士的学位论文。如已批准为非公开学位论文，须向图书馆提供批准通过的《南开大学研究生申请非公开学位论文审批表》复印件和“非公开学位论文标注说明”页原件。

南开大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下进行研究工作所取得的研究成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：_____ 年 月 日

非公开学位论文标注说明

(本页表中填写内容须打印)

根据南开大学有关规定，非公开学位论文须经指导教师同意、作者本人申请和相关部门批准方能标注。未经批准的均为公开学位论文，公开学位论文本说明为空白。

论文题目			
申请密级	☐ 限制 (⩽2 年) ☐ 秘密 (⩽10 年) ☐ 机密 (⩽20 年)		
保密期限	20 年 月 日至 20 年 月 日		
审批表编号		**批准日期**	20 年 月 日

南开大学学位评定委员会办公室盖章 (有效)

注：限制 ★2 年 (可少于 2 年); 秘密 ★10 年 (可少于 10 年); 机密 ★20 年 (可少于 20 年)

摘要

弱监督学习技术旨在利用比较弱的监督信号构建机器学习模型。计算机视觉中的弱监督学习的研究起源于人们对节省标注成本的迫切需求：细粒度的视觉任务中需要的强监督标注成本非常高昂，如何利用低成本的数据标注来训练细粒度的视觉任务成为近年来的研究热点。除了能够减少对标注的需求，弱监督学习技术更重要的意义还在于，从弱标注甚至无标注的开放数据中，挖掘图像或视频中的视觉信息，并基于这种视觉信息的组合与推理完成复杂的认知任务。这对于让机器模拟人类认知过程，形成更加通用的智能系统至关重要。本文关注计算机视觉中的弱监督技术，在图像和视频两个不同的场景中分别研究了如何利用弱监督的数据标注挖掘像素和物体两个层级的视觉信息，并利用从弱监督信号中挖掘的视觉信息进行更高层的认知任务。

在视频场景中，存在视频问答中的视频和句子的配对这一弱监督信息学习，针对视频问答任务中难以进行物体级别场景分析的问题，本文提出利用这种弱监督信息学习物体级别的视频区域和物体词语的对应关系。利用这种物体级别的视觉信息，本文设计了一种基于物体的注意力图生成模块，该模块通过问题中的物体词语引导，关注视频中的相关区域。为了关注到对回答起关键作用的物体区域，本文还提出一个注意力控制模块，该模块通过将注意力在问题中不同的词语之间转移实现视频关注区域的转移。在公开数据集上的实验表明了本文从粗糙的监督中挖掘到的物体级别视觉信息的有效性。

在图像场景中，本文研究了用户个性化图像的语义分割问题。个性化图像是指来自同一用户的图像集合。针对现有方法无法利用用户图像的个性化特点的问题，本文将图像属于同一用户这一弱监督信息具体化为用户图片之间的相互关联性，并将用户的图片聚类成组内图像高度相关的图像分组。本文提出一个组上下文语义辅助模块，在对组内图像进行分割时，通过组内图片之间的上下文语义补充，为图片中每一个像素的类别判定提供辅助。本文还收集了一个包含众多用户图片的个性化图像数据集。在数据集上的实验表明，本文的方法能够有效地利用用户个性化这一弱监督信息，提高分割性能。

关键词： 弱监督；视觉信息挖掘；视频问答；个性化；语义分割

Abstract

Weakly supervised learning aims to construct machine learning models from coarse and incomplete data annotations. The research origin of weakly supervised learning in computer vision is the urgent demand for saving labeling costs: The cost of labeling required by fine-grained vision tasks is very high. How to use low-cost data annotation to train fine-grained vision tasks has become a popular research area in recent years. In addition to reducing the need for labeling, the significance of weakly-supervised learning is to mine the visual information in images or videos from weakly labelled or even unlabeled data, and complete complex cognitive tasks based on the combination and inference of the learned visual information. This is essential for the machine to simulate the human cognitive process and form more general intelligent systems. This paper focuses on the weakly-supervised learning in computer vision. We study how to use coarse data annotation to mine both the pixel and object-level visual information in image segmentation and video understanding.

For video understanding, in view of the difficulty of object-level analysis in video question answering task, we propose to use the coarse pair between video and sentence to learn the correspondence between the object-level video areas and the object words. By using this object-level visual-language alignment, we designed an object-based attention map generation module, which is guided by the object words in the question and pays attention to the relevant areas in the video. In order to pay attention to the object area that is decisive for the question, we further propose an attention control module. This module enables the attention shift among different video area. Experiments on public dataset show the effectiveness of the object-level visual information we mined from coarse supervision.

For image segmentation, we study the semantic segmentation of user's personalized images. Personalized images are collections of images from the same user. Aiming at the problem that the existing methods cannot make use of the personalized characteristics of user's images, this paper embodies the weakly supervised information that

the images belong to the same user as the interrelationship between user's images. We cluster the personalized images into image groups so that images within one group are highly correlated. We propose a group context module, which provides assistance for determining the category of each pixel by supplementing the contextual semantics between images from the same group. We collected a personalized image dataset containing different user's personalized images. Experiments on the dataset show that our method can effectively utilize the weak supervision of image personalization and facilitate segmentation.

Key Words: weak supervision; visual information mining; video question answering; personalization; semantic segmentation

目录

插 图 目 录

表 格 目 录

第一章 绪论

计算机视觉中狭义的弱监督学习通常定义在具体的视觉任务中，例如弱监督语义分割是指利用图像级别的类别标签进行像素级别的语义分割任务的训练。本文探索更广泛的弱监督学习，即利用较弱的数据标注（不仅限于图像类别）挖掘图像或视频数据中的视觉信息，并利用这种视觉信息为不同的下游任务服务。

第一节 研究背景和动机

近年来，深度神经网络（Deep Neural Network）在计算机视觉和自然语言处理的研究中起到了重要作用。针对不同的具体任务，人们设计不同的网络结构，并利用对应的训练数据训练网络。如图 1.1 所示：在图像语义分割（Semantic Segmentation）任务中，研究者们需要标注图像每一个像素的类别[1]，并以此训练得到能够预测图像中每一个像素类别的分割网络；在显著性目标检测（Salient Object Detection）任务中，需要标注图像中每一个像素是前景还是背景[2]；在长视频的动作检测（Temporal Action Localization）任务中，则需要标注视频序列中发生动作的起止时间和动作类别[3]；在场景图生成（Scene Graph Generation）任务中，训练数据不仅需要标注图像中物体的位置，还需要标注这些物体之间的相互关系[4]。大量精细标注的数据，使得深度神经网络的训练变得实际可行，从而在各个任务上相比于传统算法取得了极大的性能提升。这种性能提升让计算机视觉算法能够在诸如摄影、复杂场景理解、多媒体内容审核、工业品缺陷检测等各种应用场景中发挥实际的作用，极大的减少了对于人工的需求。

虽然精细的标注能为神经网络提供有效的训练数据，然而这些标注的获取却需要消耗大量的资源。如何在不需要或仅需要少量粗糙的数据标注的情况下设计模型，并在这些任务上获得与精细标注相似的效果，从而进一步减少对人工的需求，这成为近年来的热点研究方向。一般将这种利用比较粗糙的数据标注进行训练，得到更精细的任务的模型的方法称为该任务的弱监督方法。例如在语义分割中，全监督训练需要提供图片中每个像素的类别标签作为训练数据。而在弱监督语义分割（Weakly Supervised Semantic Segmentation）中，训练数据中只包含图片中存在哪些类别的物体，但图片中特定像素是属于哪个类别的物

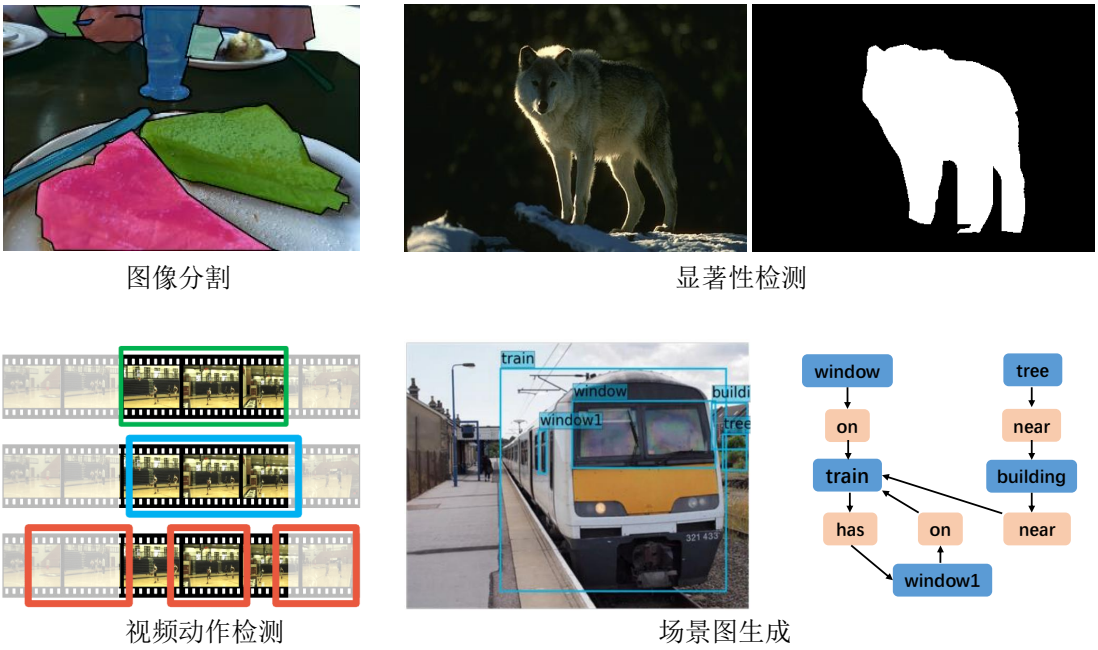

图 1.1 不同的任务需要标注相应的细粒度数据集，这种标注成本往往是高昂的。

体则是未知的。研究者们需要利用这种弱标注的数据去训练一个能够预测图片中每个像素的类别的模型。从广义上来说，弱监督条件下的视觉内容理解研究的核心就是利用更为简单易得（如图像包含物体类别、视频与对应的描述）的数据标注，去挖掘图像或者视频中的更为复杂、精细（如图像中每个像素的类别、视频中物体出现的区域）的视觉信息。基于这种更精细的视觉信息的挖掘，研究者们能够研究诸如图像场景分析、视频内容理解等更加抽象复杂的任务。

本文在视频理解和用户个性化图像分割两个场景中研究如何利用弱监督信号来挖掘图像或视频中的细粒度的视觉信息，并利用这种信息解决下游的视觉任务。

视频内容理解。图像和视频是计算机视觉研究中最重要的两个媒介。相比于图像，视频内容不仅有空间上的多样性，也增加了时序上的多样性。这一方面带来了额外的信息，另一方面也为视频内容理解带来了额外的挑战。早期的视频内容理解的相关任务主要关注在简单的动作视频的分类, 通常这种视频的场景比较简单，视频中通常是简单明确的动作以及和动作主体区分明显的背景。例如体育动作的分类[5] 中视频的画面都是人的单一运动。随着深度学习的发展，直接利用神经网络将这些视频的每一帧提取成一个特征，再在特征上进行简单的分类就能很好地解决这些问题。近年来研究者们将注意力更多地集

中在较为复杂的视频内容理解的问题中。比如动作比较复杂的视频分类数据集Something-Something[6] 中的动作类别不仅涉及空间上物体内容识别，还涉及时间上的简单逻辑推理。而在视频问答任务[7] 中，不仅仅需要模型做到视频整体内容的理解，还涉及到视频中具体物体的属性理解。在解决这些较为复杂的视频理解的问题的时候，如果还像之前的方法那样直接将视频的帧作为一个整体往往不能得到很好的结果。究其原因，还是因为用直接池化的特征来表征一帧的内容对于场景复杂的视频来说过于模糊，无法涵盖帧中包含的信息。为了解决这个问题，需要更加细致地分析场景中包含的物体以及它们的属性和关系。然而在视频问答任务的数据集中，训练集通常只包含视频以及关于视频的问题与答案，并不包含视频画面或者问题中出现的物体的详细标注。如何才能在仅有的粗糙标注下更加细致地分析视频中的内容从而帮助理解视频内容呢？我们通过粗糙的视频级别的弱监督标签挖掘视频中的物体级别的视觉信息，并利用物体级别的信息对视频进行分析。

用户个性化图像分割。图像语义分割的目标是预测一个图像中每一个像素的类别标签。通常需要在完全标注的数据集上训练网络，然后应用到需要预测分割图的图像上。然而当我们把这样训练的模型在用户的图像上实际应用时，却通常得不到理想的结果。这是因为公开的标注数据集和用户的个性化数据有很大的分布差异，在公开数据集上训练的模型会因为过拟合而无法在用户的数据取得好的泛化效果。尽管研究者们设计了诸如 drop out[8], 数据增强等不同的方法来减少过拟合效应，但是数据分布差异的问题依然无法得到妥善解决。为此，当然可以简单地在用户数据上进行像素级别的标注，然后利用这种标注数据再次训练模型。但是用户图像各有各的风格，数据分布自然也不尽相同，在不同的用户数据上标注显然成本巨大。另一方面，由于隐私问题，这种用户数据上的标注有时即便投入了成本也是无法得到的。那么能不能在没有用户的强标注数据，而只知道图片属于某个用户这种弱的数据标签的情况下，更好地学习到用户的数据特征，从而获得更好的分割结果呢？本文尝试利用这种弱标签数据挖掘图像中的像素级别语义信息，研究弱监督场景下的用户个性化图片的语义分割问题。

第二节 相关研究

在这一节中，我们简单介绍本文涉及到的相关领域的研究工作以及我们的思考。主要包含弱监督学习、视频内容理解与视频问答、个性化图像分割等相关内容的研究背景。

1.2.1 弱监督学习

深度学习的发展极大地推动了监督学习在各个领域的发展。然而监督学习通常需要大量的标注数据，这种数据标注非常耗费资源，并且在很多情况下很难甚至无法得到。为此急需发展弱监督学习技术，在较粗糙的监督信息下训练模型。根据周等人在[9] 中的梳理，弱监督学习（Weakly Supervised Learning）技术可以分为三种典型类型：不完整监督、不精确监督以及不准确监督。

不完整监督（Incomplete Supervision）[10–13] 是指大量的训练数据中一部分数据有标注，一部分没有标注，即标注不完全。在不人为添加数据标注的情况下，这种类型的技术又可以称为半监督学习（Semi-Supervised Learning）。半监督学习近年来在计算机视觉领域吸引了很多研究。魏等人[14] 研究了空洞卷积（Dilated Convolution）技术在半监督语义分割上的应用。Yan 等人[15] 通过为没有标注的部分数据生成伪标签，再利用这种伪标签训练模型，在视频显著性目标检测上取得了很好的效果。Kipf 等人[16] 在图结构的数据上直接利用卷积神经网络解决半监督问题，吸引了后续大量关于图神经网络的研究。主流的半监督方法都基于这样的假设: 数据的分布遵循聚类假设（Cluster Assumption）。即相同类别的数据在某个特征空间上是聚集在一起的，而不同的数据则在该空间上距离较远。实际上这个假设也被利用在诸如迁移学习、域自适应等其他机器学习技术中。视觉任务中常见的伪标签技术同样也是基于这样的假设: 在特征空间中相近的数据在输出空间上也有相同的标签，那么没有标注的数据也可以通过与已有标注的数据计算相似度来预测其输出，得到的输出作为伪标签就可以当作是额外的监督数据。

不精确监督（Inexact Supervision）指的是尽管数据有标注，但这种标注并不够准确。这种标注通常较为粗粒度，不足以直接训练需要的细粒度任务模型。比如，在视频分类中，可能会包含不同层级的标签，低层级的类别标签包含在较高层级的标签中。不精确监督技术就需要研究如何在只有高层级视频标签的情况下训练模型，得到能预测视频低层级标签的网络。[17] 等人研究了如何

利用多层级的标签帮助视频内容理解算法取得更好的效果。又如，在图像分割任务[14, 18, 19] 中，通常需要利用像素级别的标注数据进行训练。为了减少数据标注上的成本，弱监督的分割技术关注不精确监督场景下的分割技术：仅利用包含图片类别标签的数据进行训练。主流的弱监督图片分割方法都遵循这一设定，近年来很多方法在这个任务上取得了很好的进展。魏等人在 STC[20] 中提出利用自底向上的显著性检测算法预测较为简单图片的分割图，然后利用这个分割图作为监督逐步训练较复杂的模型。姜等人[18] 利用注意力累积机制，在图像分类的网络中挖掘注意区域作为物体区域，并生成伪标签，这个伪标签进而被用来训练分割网络。Wang 等人[19] 通过为注意力累积过程增加像素级别的约束进一步解决注意力累积图的不准确问题。不精确监督作为弱监督技术中的一个典型类型，在各个任务和各种实际应用中都有很多需求。本文处理的问题也属于不精确监督的范畴，即在粗粒度标注的场景下挖掘图像与视频中的语义信息。不同于上述狭义的语义分割弱监督方法，都通过图片级别的标签学习像素级别的图像分割图。我们考虑更加广泛的弱监督场景下的不同层级视觉信息的挖掘，并利用挖掘到的视觉信息为更上层的视觉理解问题服务。

不准确监督（Inaccurate Supervision）[21–23] 指的是用于监督用的数据不总是正确。也就是说，部分数据虽然有标注，但是这种标注是错误的。它不仅不能提供准确的监督，反而会给训练过程带来噪声，影响整体的模型训练效果。什么情况下我们会面对这样的不准确数据场景呢？一种情况是我们自动从网络或社交媒体上搜集大量包含标签的数据，这种标签是在互联网用户积累下自动产生的。由于缺少专家审核处理，所以这些数据的标签并不总是可靠，而我们又希望利用这种数据的海量性来提升算法的通用性能。这种时候就需要合适的方法来应对数据中的标注不准确问题。另一种有意思的情况是，尽管目前研究者们提出的标注数据集是经过培训的人工进行标注的，然而这种标注同样并不总是可靠。这是因为标注者的水平参差不齐，很多任务中的标签决定又存在一定的主观性。这导致用于监督的数据本身也不可靠，这种不准确监督数据训练的模型存在一个比理想的标签训练的模型低的性能上限。如何解决这种标签不准确的问题呢？一种直接的想法就是找到并去除这些错误的标签。侯等人在[24] 中提出所谓的“webly-supervised”问题。这可以看作是一种比常见的弱监督语义分割更加弱的问题。[24] 将类别标签当作关键词，在互联网上搜索相匹配的图片，并认为这个关键词是这些图片的类别标签。这样的标签显然是无法保证准确性

的。为此，他们设计了一个噪音去除网络用于去除数据中的不可靠部分。在训练过程中，他们将同一个关键词搜索到的一批（mini-batch）的数据作为一个整体，然后通过学习这批数据的图片间关系来找出与其他图片语义不一致的图片，这样的图片就被认为是与关键词不匹配的噪音图片被舍弃掉。留下的数据则被认为是正确标注的数据进行传统的弱监督语义分割模型训练。这种做法背后的想法是很朴素并且在传统机器学习问题中被经常使用的：即通过查询相邻的数据对当前数据进行修正或者去噪。计算机视觉中的图片去噪的一个简单的传统算法就是通过周围的像素进行平滑处理，达到去噪效果。本文中在处理个性化用户图片的分割问题时，同样也利用了相似的想法：通过咨询相关像素的语义来确定某个像素的类别。

1.2.2 视频内容理解与视频问答

视频内容理解一直是计算机视觉的一个活跃领域。由于计算资源有限，以往的工作主要关注相对简单视频中的动作识别[25, 26]，这些工作通常使用人工设计的能反应某种视频特性的特征。随着深度神经网络（DNN）的发展，研究人员在大规模视频分类数据集上取得了一些进展[5, 27, 28]。近年来，研究的重点已经从简单的动作识别转向复杂的场景分析[6, 29]，对动作的推理和归纳能力提出了更高的要求[30]。为了对复杂场景进行推理，需要对视频中的不同对象进行识别和定位。Baradel 等人[31] 和 Wang 等人[32] 从视频帧中抽取物体，并通过学习物体间的交互进行视频分类。Zhou 等人[33] 基于视频中的物体学习视频描述。通过将整个视频的网络输入缩小到物体边界框区域，取得了比以前更好的性能。然而，在这些模型中提取对象需要像 faster-RCNN[34] 这样的物体检测模型，这在实践中可能是有问题的：在推断时为每个视频提取对象边界框非常耗时；同时，目标检测模型从图像数据集到视频数据集[1, 35, 36] 的泛化能力得不到保证，这可能导致许多错误或遗漏的检测。总的来说，在这些以视频分类为目标的方法中，研究者们主要关注两个研究方向：一个是更好的基础网络，即通过研究不同的基础网络架构来提取视频特征，期望这种提取出的特征能够更好地表征视频输入，在映射空间上更加容易分类。无论是较为早期的双通道网络[37]，还是后来的 C3D[38]，I3D[39]，S3D[28] 等 3D 卷积的模型，还是最新的在视频上抛弃卷积，完全使用 Transformer[40] 的尝试[41]，这些都是在视频理解基础网络上的探索与尝试。除了更好的基础网络，另一个方向也备受关注，那就是在已有基础网络提取的视频特征的基础上，对这些表征进行组合与推理，期望能够通过模型设

计来捕捉视频中的时序和空间信息。上面提到的利用物体检测技术先提取视频中的物体位置再在这些物体的特征上进行构建图模型分类的方法[32]，以及利用帧与帧之间关系进行时序推理的方法[30]，包括较近的利用不同采样速度的两个分支信息相互补充的 SlowFast 网络[42]，都是属于这一方向的探索。

与视频分类任务相比，视频问答（Video Question Answering, VQA）近几年才引起人们的注意。在视频分类中，输入是一个视频，模型需要输出它的类别。而在视频问答任务中，情况要相对复杂许多。给定一个视频以及一个关于这个视频的自然语言形式的问题，需要给出这个问题的答案。这个答案一般是在多个候选项中选择。在视频问答任务中，需要同时处理视觉（视频）和自然语言（问题）两方面的信息。在处理问题时，常见的方法将问题的每个词语替换成类似 BERT[43] 这种预训练模型输出的编码，然后将这个问题当作词语编码的序列输入到 LSTM[44] 网络中以获取表示这个问题的特征。在处理视频时，一般将视频的每一帧输入到卷积神经网络中得到对应的特征，然后在这一特征上处理得到的视频特征。在视频问答任务中，要回答问题中的关于物体、动作、场景等各种细致的问题。如果仅仅在上面提到的特征上进行简单的池化操作，无法应对这些关于场景细节的问题。为此，许多方法提出了利用注意力机制去关注视频中的特定部分，以提高回答的准确度。Na 等人[45] 和 Gao 等人[46] 两者都探索了利用所谓的记忆网络来动态解决问题。Jang 等人[7] 利用空间和时间的注意力来帮助提高模型性能。Mun 等人[47] 和 Xu 等人[48] 还利用注意力研究视频序列的时间关系。上述方法都可以归纳为一种软注意力机制。它们通过将提取出的问题特征作为一个先验，接着在处理视频特征时，再通过注意力机制利用这个先验，决定视频中的哪些区域应该获得更多关注。类似的，在处理问题时，也是将视频特征作为先验，然后用这个先验去决定问题中的哪些词语应该得到更多的关注。这种软注意力机制的方法的优点是简单易操作，不需要探究视频或问题中到底出现了哪些物体及其对应的词语。然而，这个优点同时也带来了缺点，那就是缺乏可解释性。网络只能隐式地赋予一些区域更大的权重，但无法理解它所关注的到底是什么。这样使用者也无法得知它到底关注了什么，无法探究模型在处理问答这种复杂问题时的推理过程。更重要的是，由于数据集的偏差，软注意力会倾向于关注出现次数较多的物体区域，而这很多情况下并不是解决问题所需的区域。视频问答任务需要网络对视频中的特定物体的属性与物体间的互动有很高的理解能力，这种需求是软注意力机制无法满足的。为了获得更

加可靠透明的问答模型，我们需要对视频中的各种物体进行显式的分析。而这种分析的前提就是能够识别视频中的物体区域，以及它们和问题中的词语的对应关系。Yu 等人在[49] 中利用几个 LSTM 模块检测视频中的概念词。然而，他们的方法只能检测到物体在整个视频中存在，无法在视频中定位物体。为了解决这一问题，我们利用视频问答数据中已知的视频和句子互相对应这一弱监督信号，建立了视频中的区域和问题中的相关词语的对应关系。

1.2.3 图像分割与个性化图像分割

图像语义分割是一个经典的计算机视觉问题。给定一个图像，它的目标是为图像中的每一个像素预测一个语义类别。最近 10 年来，基于深度学习的图像语义分割方法[50–52] 在这个领域取得了巨大的成功。我们考虑用户的个性化图像的分割问题：对于若干来自同一用户的无标注数据，如何利用现有的公开数据和模型在用户的图像上获得更好的语义分割效果。最简单的方法就是直接利用现有的模型在这些图像上进行预测，然而用于训练现有模型的数据集和用户数据之间往往存在很大的分布差异。尽管研究者们提出了很多关于语义分割的数据集[1, 35, 53–55]，并尽可能地使这些数据集覆盖尽可能多的场景，然而有限的数据和标注在面对新的数据时，依然会因为分布的差异面对泛化能力不强，分割效果不佳的问题。为了解决这个问题，可以通过在用户数据上进行标注来训练网络，使模型适应用户数据。然而用于语义分割的像素级别标注成本很高，对所有用户都进行数据标注是不现实的。我们需要能够直接将分割模型从现有的公开数据集上迁移到用户的个性化数据上，并适应个性化数据的分布。

关于语义分割的域自适应（Unsupervised Domain Adaptation for Semantic Segmentation）任务解决的问题和我们的情况类似，下面我们简称其为 UDASS。给定带有像素级标签的源数据集和未标注的目标数据集，UDASS 的目标是解决源数据集和目标数据集之间的分布不匹配问题，并使模型从源到目标具有更好的泛化性能。UDASS 的一个主要思路[56–60] 是使用基于对抗的方法来缩小源域和目标域数据在特征空间中的分布差距。另一个思路则更侧重于学习策略的设计：[61–64] 使用课程学习或者自监督训练来更新网络，通过伪标签在目标域由易到难地学习图像的语义。在这些域自适应的方法中，目标域的数据都被认为是相互独立分布的，这往往和现实情况不符。我们的用户个性化图像语义分割问题与 UDASS 之间的主要区别在于，用户个性化的图像并不是相互独立的。相反的，来自同一个人的图像是相互关联的，它们之间可能包含相同的物体，也可

能包含相同的背景。这种相互关联可以被看作是一种弱监督信息，同一用户的一个图像能够为其他图像提供有用的语义信息。如果能够利用用户个性化这一特点，挖掘图像之间的关联性，就能为图像的理解提供额外的帮助。

在计算机视觉和自然语言处理任务中利用用户的个性化信息，这一点在许多之前的工作中都有所讨论。MIRKIN 等人[65] 利用用户的个性特点来增强机器翻译系统。HORIGUCHI 等人[66] 提出了用于食品图像分类的个性化分类器。PARK 等人[67] 通过探索用户以前的帖子中的个性化特征来预测社交媒体图像中的标题和主题标签。KIM 等人[68] 根据问卷调查获取用户的偏好，再利用这种偏好来进行图像增强。这些方法的共同特点是，它们都从用户的数据中提取一个全局的个性化特征，然后在面对这个用户的新数据时将这个个性化特征用新数据的先验。在个性化图像分割的任务中，我们从更加广义的角度来利用用户个性化数据这一弱监督信息。除了用户全局的个性化特点外，我们认为用户的特点还局部地体现在不同的图像上。通过从关联的数据中获得互相补充的语义信息，也是利用数据的个性化特点的有效方法。在现有的图像分割方法中，已经存在一些利用相关数据来辅助分割的技术。这种技术主要出现在协同分割（co-segmentation）或协同显著性检测（co-saliency）的任务中。协同分割[69–72] 或协同显著性检测[73, 74] 任务的输入是一组包含共同前景物体的图像，它的目标是将图片中的这一共同前景物体分割出来。因为输入的一组图像中有一组共同的前景，因此可以通过从这组数据中提取出该前景的特征，然后利用这个特征来判断图像中的区域是否属于这个共同物体。Li 等人[71] 提出了一种递归网络架构，并通过该网络累积每张图像中的共同语义，形成一个共同前景表征。Zhang 等人[74] 利用分类网络关注的区域来发现共同物体类别的大致区域。和之前的个性化技术类似的，这些方法也要学习每个组的全局表征，然后作为对该组的图像进行分割时的先验使用。协同分割的方法能够成功的原因是需要分割的图片都包含同样的前景，因此对于每张图片都可以利用这样的前景特征来辅助区域的类别判定。然而和协同分割不同的是，用户个性化数据中不同的图像中包含不同类别的物体，而不是仅限于某种物体。如果对一个用户提取一个全局特征，这个特征会包含不同的类别的语义，导致其表征是信息模糊，无法为图像分割提供有意义的参考。为此，需要对现有的基于全局特征的技术进行改进，在利用图片之间语义信息的同时，为每张图片提供明确的语义补充。

第三节 本文研究内容与章节安排

本文围绕计算机视觉中弱监督信息的挖掘，分别在视频问答和个性化图像分割两个场景中，利用数据中的弱监督标注学习物体级别和像素级别的细粒度视觉信息。

在视频问答的场景中，训练数据中仅包含视频-问答这样的视频级别的弱标注的配对，而问题中则不仅包含视频级别的动作理解，也包含物体级别的属性与关系分析。为了回答这些细粒度的问题，我们必须在细粒度的物体区域级别分析视频中区域的属性、相互关系，以及这些视觉区域和自然语言的关系。为了能够在物体级别分析视频，我们将数据中视频和对应的问题分别解析成区域级别的视频节点和词语级别的物体词。并利用它们的配对学习物体级别的视频区域和物体词语的对应关系。针对先前方法的软注意力机制无法正确关注到正确的物体区域、缺乏可解释性的问题，我们利用从弱监督标注中学习的物体级别的区域-词语配对，通过问题中的物体词语在视频中生成物体区域的注意力图。这种基于物体词语的注意力图是透明可解释的，能够为模型分析提供帮助。视频问答任务中，自然语言的处理和理解对于回答问题也是至关重要的。在一个问题中，通常会涉及到多个不同的物体词语，而词语的选择会影响到我们方法中注意力图的生成。为了应对这个问题，关注问题中对回答起决定性作用的物体词语，我们设计了一个注意力控制模块。我们的注意力控制模块是一个能够级联使用的网络模组，通过级联多个注意力模块，网络能够在不同的步骤中选择不同的物体词语，从而使注意力在视频的不同区域上转移。这样的注意力转移能够模拟人类在解决问答任务时的推理过程。得益于我们基于物体区域的注意力机制的可解释性，我们能够可视化网络在不同的步骤关注的词语和区域，进而分析网络的行为。

在用户个性化图片分割的场景中，训练数据中仅包含来自同一用户的图片，我们尝试利用图片来自同一用户这一弱监督信息挖掘用户数据的个性化特点，并利用它来学习像素级别的语义分割图。图像语义分割的个性化研究是一个新的问题，我们首次在本文中研究这个问题和解决方法。为了促进这一问题的研究，我们收集了一个用户个性化图片数据集。这个数据集中包含来自 15 个不同的用户的个性化图片。为了在数据集上测试分割方法的性能，我们为这个数据集中每个用户 30% 的图片进行了像素级别的语义标注。个性化数据的最重要的特点就是同一个用户的不同图片之间相互关联，隐含了用户的个性化特点。为

了利用这一特点解决个性化图片语义分割的问题，我们尝试将同一个用户的图片进行分组。经过分组后，每一个组内的图片都包含相似的物体或者场景，也就是语义上相似度较高。在进行分割时，我们将每一组的图片当作一个共同的图片组处理，这些图片之间相互补充的上下文语义能够为组内图像中每一个像素的类别判定提供帮助。我们在提出的数据集上测试了基于相似图片语义辅助的个性化分割的方法，实验表明我们成功地利用了用户个性化这一弱监督信息，为像素级别的分割网络提供了有效的语义补充。

本文的后面章节的安排如下：在第二章中，我们介绍了视频问答场景下的弱监督信息的挖掘与利用。我们首先在第一节介绍视频问答的相关背景。接着在第二节中介绍了我们如何利用视频级别的标签学习物体级别的视觉-语言对应关系，进而利用这种关系解决视频问答任务。在第三节中我们报告了提出的方法在问答数据集上的表现，同时通过大量实验与可视化分析了提出的模型的不同模块的效果。在第三章中，我们介绍了个性化图像分割场景下的用户个性化信息的挖掘。我们同样在第一节介绍了问题的背景，在第二节中给出了我们提出的数据集的介绍。在第三节和第四节中，我们介绍了为了解决个性化分割问题的方法以及在提出的数据集上的实验结果与分析。最后，我们在第四章对本文内容进行总结，并对相关的问题未来的研究方向进行展望。

第二章 视频问答中的弱监督方法

在众多联合学习视觉和语言信息的任务中，视频问答（Video QA）是重要且有挑战性的一个。给定一个视频和关于它的问题，视频问答的目标是对问题作出正确回答。视频问答系统成功的一个重要因素是其是否能找到问题中提到的决定性视频区域，并根据这个区域的视觉信息回答问题，而这需要对视频进行物体级别的分析。为此，本章尝试通过粗粒度的视频与句子的配对这一弱监督信号学习物体级别的视频区域和词语的配对。基于这种配对，作者提出一个物体引导注意力（OGA）机制。使用显式的物体词语引导，告诉网络应该关注哪个区域。为了关注在适当的物体上，本章还设计了一个注意控制模块来动态选择问题中出现的物体词。作者在 TGIF-QA 数据集上评估了提出的模型。对于稀疏采样的视频帧，本章的模型在 TGIF-QA 数据集的不同任务上都优于现有的 SOTA 方法。

第一节 视频问答的背景介绍

传统视觉任务通常专注于从视频或图像中学习人工标注的有限的标签[75, 76]，例如图像分类需要学习固定数量的类别标签，语义分割则需要学习固定类别物体的区域。这些任务在很多现实场景中都得到了应用。然而基于强监督的学习面临许多瓶颈：例如类别的标注总是有限的，现有的模型无法处理没有标注的类别；另一方面在面对许多需要因果推理的较高层次的任务时，这种基于标签强监督学习的模型同样无法解决。为了解决上述问题，需要智能系统能够从开放获取的没有标注的信息中学习到有用的信息，并在这个信息的基础上进行推理与视觉理解。开放世界的信息自然是弱监督甚至是无监督的，但数据量却是海量的，它不仅包含视觉信息，也可能包含与之对应的语言信息。近年来，沟通视觉和语言之间关联的任务获得了大量的关注[77–80]。其目标是在自由形式、开放式的语言[78] 信号和视觉信号之间建立对应关系，并利用这一对应关系来解决更复杂的问题。这些任务为视觉系统提供了一个新的方向，那就是在没有明确的强人工标注的情况下，从视觉和自然语言的弱对应这种弱监督信号中联合学习。在这些任务中，视频问答（Video Question Answering）是一个富

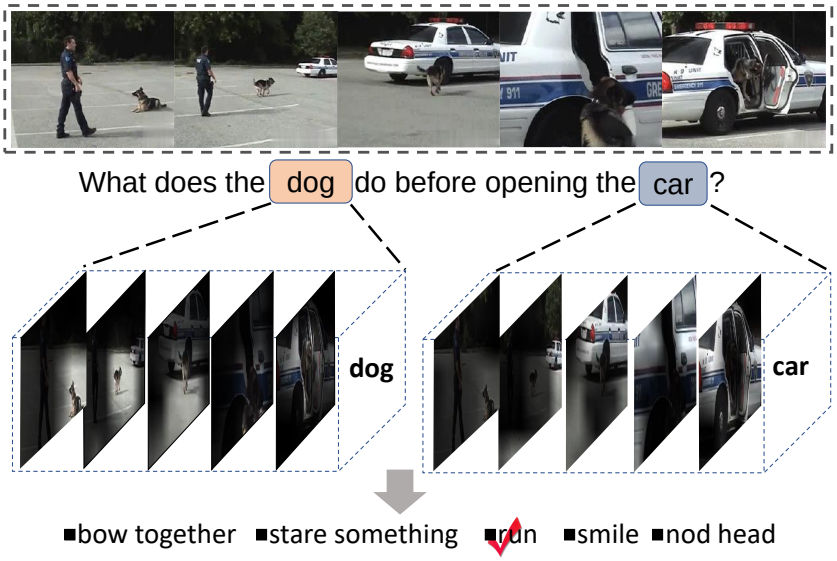

图 2.1 TGIF-QA 数据集上的一个例子。问题“狗在车门开前在做什么？”包含两个物体：“狗”和“车”，不同的物体对应视频中的不同区域。通过分析问题，我们知道包含“狗”的区域应被突出。我们通过视频-语句配对的弱监督信号来学习视频区域和物体名词之间的对应关系，进而在视频问答任务中通过物体词语关注到能够回答问题的视频时空区域。

有挑战性的问题，它需要视觉和语言的联合学习。给定一个视频和一个关于这个视频的问题，我们的模型需要分析问题，从视频中搜集需要的信息，然后给出答案。

在以前的视频理解任务（如动作识别）中，目标通常是对人类活动进行分类[5]。这种任务中的学习目标比较明确，那就是围绕人类的动作进行识别与分类。然而，在视频问答任务中，学习目标因问题而异。如图2.1所示，在视频的不同时空区域中会出现不同的物体：狗、人和车。为了正确地回答问题，模型必须基于问题关注到正确的时空区域。以往的工作在不同视频理解任务中都研究了注意力机制，从而让网络关注到合适的视频位置。Non-local 网络[81] 用自我注意力作为权值在不同的时空特征之间传递信息。注意力蒸馏方法[82] 使用光流来学习一个注意网络，该网络在测试时能抛弃光流直接生成关注动作部分的注意力图，用于视频分类。ST-VQA 方法[7] 结合视觉特征和语言特征在视频上生成软注意力图，用于视频问答。注意力蒸馏法[82] 可以利用光流的先验信息突出动作发生的区域，但是训练时的光流仍然是必不可少的，而光流的计算非常耗时。更重要的是，基于光流计算出来的注意力网络只能关注到视频中的运动区域，这导致其在视频问答中并不能满足要求，因为视频问答中关注的问题多种

多样，不仅有运动物体的动作，也可能有静态物体的性质。[7, 81] 都属于软注意力方法。通过软注意力机制虽然能提升视频理解任务的性能，但是它关注了哪些区域、为什么关注这些区域对于网络外部都是黑盒的，这就导致这些方法的可解释性较差。在一些分布不均衡、偏差较大的数据中，网络有可能仅仅是关注到多数问题中出现的物体区域，而非对当前问题有意义的区域。这一点我们能够在后文2.3.4的注意力图对比中看到。

为了解决上述问题，生成比软注意力方法更加可靠可解释的注意力图, 我们需要能够识别视频中和问题相关的物体区域，并显式地注意到这些区域。然而我们的视频问答数据中只有模糊的视频和语句的对应关系，没有物体级别的标注。为此，我们提出通过视频-问答对这一弱监督信号去学习物体级别的视频区域-物体词语对应关系。进一步的，我们提出了一种物体注意力引导机制，让网络利用学习到的 视频区域-物体词语配对注意到视频中的相关物体。相比于[82]，我们的方法仅利用原始的 RGB 数据突出物体区域，节省了计算光流的时间，并且能够同时关注到动态或静态的物体区域。我们首先利用视频问答对作为监督，建立视觉特征和物体词语之间的对应关系。接着对于给定的问题，我们会提取问题中关注的物体词语，并利用前面学习的对应关系查询视频中的相关区域，生成注意力图。通常，一个问题可能会提到数个物体词语，其中一些物体对预测答案具有决定性作用，一些物体可能仅仅是与决定性物体相关。为了明确地将注意力集中在对问题起决定性作用的物体上，我们设计了一个注意力控制模块，该模块动态地从问题中选择物体词语，并用这些物体词语在视频区域之间转移注意力。

本章的贡献可以概括为两个方面：

- 我们提出利用视频-问答对这一弱监督信号去学习物体级别的视频区域-物体词语对应关系，并在视频问答中利用这一关系引导网络关注到视频中的物体区域。与软注意方法相比，本章的注意力图能够显式地、准确地覆盖目标区域。该方法在 TGIF-QA[7] 数据集上实现了不同任务上的一致提升，并且计算量要小很多。
- 我们设计了一个注意力控制模块，从问题中选择物体词语。通过级联的注意力控制模块，我们的模型可以动态地转移注意力，选择能解决问题的视频区域。

第二节 基于视觉-语言匹配注意力机制的视频问答方法

在这一节中，我们将详细介绍如何利用视频级别的弱监督信号挖掘物体级别的视觉-语言信息，并用它来解决视频问答任务。

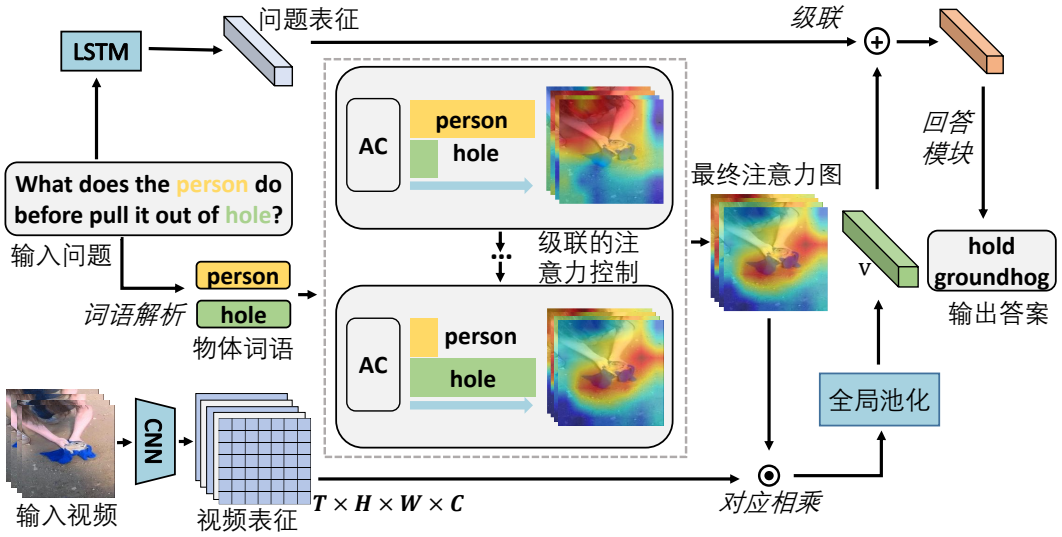

图 2.2 用于视频问答的物体词语引导的注意力模型的框架。输入是视频和问题，输出则是对应的答案。AC 表示我们的注意控制模块，它从问题中选择要关注的物体词语，然后利用注意力生成模块根据所选的词生成注意图。我们串联了数个注意力控制模块，当前 AC 模块产生的注意力图将传递给下一个 AC 模块，最后一个 AC 模块的注意力图用作最终的注意力图。由注意力图得到的最终视频表征 v 和由 LSTM 得到的问题表征融合并得到最终答案。

2.2.1 方法概述

这部分简要给出本章提出的利用物体词语引导注意力的视频问答方法的框架介绍。整体的框架将原始视频的 RGB 帧与其对应的问题作为输入，我们分别用 $X=\{x^1,x^2,...,x^T\}$ 和 $Q=\{qw^1,qw^2,...,qw^N\}$ 表示他们，这里 x^t 表示在 t 时刻的一帧的图像，$qw^n\in\mathbb{R}^d$ 表示问题的第 n 个词的编码。我们使用卷积神经网络将原始视频输入 X 转换为维度为 $T\times H\times W\times C$ 的中间视频特征 $V\in\mathbb{R}^{T\times H\times W\times C}$，这里 H 和 W 表示特征图的高度与宽度，T 表示帧数，C 为通道数。问题 Q 通过 LSTM[83] 网络处理为问题表征 q。如图2.2所示，通过多步的注意力控制，网络在目标词和视频区域之间动态转移注意力。最后一个 AC 模块的注意力图将用于获得最终的视频表征 v。通过将 q 和 v 输入给答案模块，就

能得到最终的答案预测。在整个过程中，有三部分起到了关键作用：弱监督视觉-语言匹配，注意力生成模块以及注意力控制模块。它们分别负责学习视觉区域与物体词语的匹配关系，利用匹配关系确定注意力区域，以及动态地决定问题中的哪个物体词语应该被关注。

弱监督视觉-语言匹配。这部分负责利用数据中的视频-问答弱监督信号，首先将视频和语句分别分解成时空区域和物体词语，然后建立它们之间的对应关系。中间特征图 $V \in \mathbb{R}^{T\times H\times W\times C}$ 可以被看作 $T \times H \times W$ 个维度为 C 的节点，那么其中每个节点都表示视频的一个时空区域，我们的目的是将每个节点和语句中的词语对应起来。通过这部分学习，我们将视频级别的弱标签转化为物体级别的匹配关系。通过利用这种较细粒度的匹配关系，我们的问答模型能够显式地在视频上运用注意力机制。

注意力生成模块。对于视频特征 V, 假设我们需要在这个视频的 $T \times H \times W$ 个区域中关注物体 w 对应的区域，w 表示物体词语的编码。我们设计了一个注意力生成（Attention Generation，AG）模块，以 V 和 w 为输入，注意力生成模块负责生成一个 V 上的注意力图，维度也是 $T \times H \times W$，它能突出显示物体 w 对应的节点。

注意力控制模块。假设问题中一共提到了 K 个物体 $W = \{w_1, w_2, ..., w_K\}$，在解决视频问答问题时，可能需要经过一个从 w_1 到 w_2，再到 w_k 的推理过程。例如对于问题“男人在上舞台前做了什么?”，人类的推理过程是先关注到舞台，然后关注上舞台的男人，最后观察他在上舞台前的动作再做出回答。为了让模型模拟这个过程，关注到正确的物体区域，我们设计了一个注意力控制（Attention Control，AC）模块。该模块能够利用问题表征 q 在 W 中选择物体词语。网络中通过级联多个注意力控制模块并将所选物体反馈给 AG 模块，以实现动态地将注意力转移到适当的物体区域的目的，从而对问题做出回答。

2.2.2 视觉-语言匹配

在这一部分，我们介绍如何利用粗粒度的视频-问答数据来建立物体级别的视频区域-物体词语对应关系。

视觉-语言匹配。构建视觉语言匹配的目标是将视频特征 V 中的节点和物体词汇表中的单词映射为同一特征空间的表征[84]，使得在这个统一的空间中互相匹配的视频区域和物体词语的距离接近，而不匹配的区域和词语之间的表征距离较远。考虑 V 中节点 V_i 和物体词语 w_+ 之间的对应关系。我们使用视觉编码

器 ϕ 和语言编码器 ψ 将 V_i 和 w_+ 映射到统一的 d 维空间 $\mathbb{R}^d$ 中。假设我们从物体词语汇表中再随机采样一个不匹配的物体词语 w_-。目标是训练编码器 ϕ 和 ψ，使得

$$s(\phi(V_i),\psi(w_+)) > s(\phi(V_i),\psi(w_-)), \tag{2.1}$$

其中函数 $s(\cdot,\cdot)$ 度量两个输入向量之间的相似度。遵循[85]，我们用余弦函数实现这个距离的度量。

匹配建立的策略。我们采用类似[84] 中的策略来训练这两个编码器。给定一个训练集的视频 V，我们将关于它的问题和答案中提到的物体视为 $w+$。我们使用 margin loss 来对齐视频节点 V_i 和 w_+，它可以表达为

$$\ell_{V_i} = |\delta + s(\psi(w_-),\phi(V_i)) - s(\psi(w_+),\phi(V_i))|_+, \tag{2.2}$$

这里 $|x|_+ = max(x,0)$, δ 是预先设定的区间幅度，这个幅度越大，就表示我们对不匹配表征之间的距离越敏感。这样我们就得到了视频区域 V_i 的匹配损失，一个视频中有 $T \times H \times W$ 个区域，将这些区域损失相加就能得到 $T \times H \times W$ 个区域损失。上面公式中的 w_- 是不匹配词语，或者称为负样本词语。我们用数据集中频繁出现的物体词语组成词库，然后在这个词库中出现频次最高的 150 个词中随机采样没有出现在问答候选词 W 中的词语作为 w_-。

值得注意的是，按照上面的规则，只要是在问答的语句中出现了的物体词语，我们就认为它对于视频中的所有区域都是匹配的，这与现实情况显然是不相符的。为了解决这个问题，我们还设置了一个和视频区域数一致的区域相关性图。这个相关性图的计算方法为

$$M_i = \frac{exp(s(\psi(w_+),\phi(V_i)))}{\sum_j exp(s(\psi(w_+),\phi(V_j)))}, \tag{2.3}$$

M 衡量节点 V_i 和物体词语 w_+ 的相关程度。得到相关图后，我们将它作为权重加到视频的区域损失上去，即：

$$\ell_{align} = \sum_{i \in T \times H \times W} M_i \cdot \ell_{V_i}. \tag{2.4}$$

也就是说，对于视频中的物体词语 w_+，我们在计算区域损失时给予和 w_+ 相匹配的区域更大的权重，从而使得区域损失专注在词语对应的区域，而不是视频中的不相关区域。在优化 ℓ_{align} 之后，我们得到了两个训练好的编码器 ϕ，ψ 作

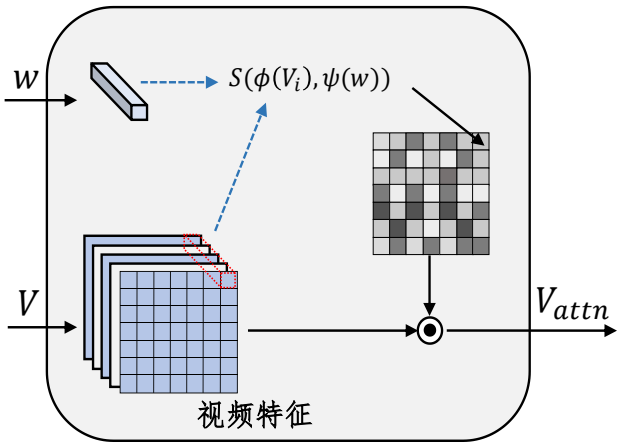

图 2.3 注意力生成模块示意图。为了简单起见，我们只画出帧数 $T=1$ 的情况。

为我们的视觉-语言对齐模块。后续在处理问答的时候，对于输入的视觉特征以及语言编码，我们首先将这些信息经过这两个编码器映射到统一的特征空间中，再进行注意力图的生成。

2.2.3 注意力生成模块

注意生成（AG）模块解决了在给定物体词语编码 w 的情况下在视频中应该注意哪个区域的问题。我们利用上面的视觉语言匹配来设计注意力生成模块。假设我们需要关注视频特征 V 中关于物体 w 的区域，通过采用上述编码器 ϕ 和 ψ，我们将 V 中的节点和物体词语编码 w 映射到统一的特征空间 $\mathbb{R}^d$。然后我们通过计算 $\mathbb{R}^d$ 空间中特征的余弦相似度来衡量节点 V_{thw} 与单词 w 的匹配程度，其中 t,h,w 表示视频中区域的时空位置。我们应用这个相似度来形成注意力图 $\mathcal{A}$，其中

$$\mathcal{A}_{t,h,w}=s(\phi(V_{t,h,w}),\psi(w)). \tag{2.5}$$

通过将注意力图复制拓展到和视频特征 V 形状相同的 $\mathcal{A}_{tile}$，我们就可以将特征 V 的不同区域加权为

$$V_{attn}=V\odot\mathcal{A}_{tile}, \tag{2.6}$$

这里 $\odot$ 表示点乘。我们在图2.3中展示了这个过程。

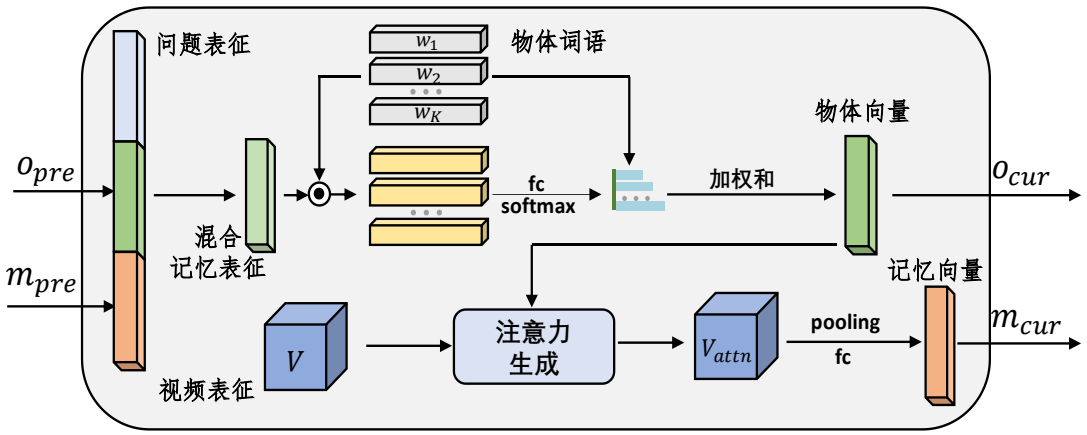

图 2.4 注意力控制模块（AC）。问题表征、视频表征和物体词是所有 AC 模块的全局信息。当前模块中生成的物体向量作为查询物体词语编码输入注意力生成模块，得到注意力图后生成 V_{attn}，池化后输出给下一个模块。

2.2.4 注意力控制模块

上一小节介绍了在给定查询物体词语 w 时如何在视频中关注相关的区域。然而，一个问题中可能会出现几个物体词，有的物体可能对回答起到关键作用，而有的则可能是推理答案的过程中涉及的物体，应该给哪些物体更多关注是一个需要解决的问题。在本节中，我们将介绍如何使用注意力控制（AC）模块在问题中的物体候选词语中选择要关注的物体。我们首先应用一个基于规则的语义解析器[86] 来解析问题并提取问题中的物体词语。对于由多个单词组成的物体，为了简单起见，我们使用最后一个单词来表示它们。物体词语被提取出来后，我们用 $\mathbb{R}^d$ 空间的编码来表征这些词，记为 $W=\{w_1,w_2,...,w_K\}$。将候选物体词语 W 和视频区域特征 V 看作知识库，注意力控制模块是一个可以级联使用的神经网络模块，通过级联数个 AC 模块，能够在词语知识库中动态地选择物体词语，然后在视觉知识库中注意不同的区域，提取视觉信息。在每个级联的模块中，我们都将问题表征 q 作为全局信息，并从 W 中选择目标词。选定的编码 w 将被用作2.2.3 中介绍的注意力生成模块的查询编码输入，从而生成关于该词语的注意力图，得到经过加权的视频特征。

受到以往的关于 QA 任务的工作[45, 46, 87] 的启发，我们还在 AC 模块中采用了记忆机制。串联的注意力控制模块之间，前一个模块的输出转化为记忆，作为后一个模块的输入。对于级联的多个 AC 模块，它们共享相同的问题表征 q、视频表征 V 和候选词 W 作为全局信息。

我们在图2.4中演示了注意力控制模块的内部结构。每个模块需要两个向量

作为输入。第一个向量 $o_{pre} \in \mathbb{R}^d$ 是物体向量，其表示由前面的 AC 模块选出的词语向量编码。另一个向量 $m_{pre} \in \mathbb{R}^d$ 是记忆向量，它代表了前一步的 AC 模块中经过注意力加权后的视频特征。在获取这些输入之后，我们首先通过下式计算一个混合的记忆向量 $m_{fuse} \in \mathbb{R}^d$：

$$m_{fuse} = W^{d\times 3d}[q, o_{pre}, m_{pre}] + b^d, \tag{2.7}$$

这里 $[\cdot,\cdot,\cdot]$ 表示级联操作，$W^{d\times 3d}$ 是线性变换的需要学习的参数，b^d 为偏置。m_{fuse} 包含了从之前的 AC 模块中关注的物体词语的表征以及这个词语对应的视频区域的视觉特征两方面的信息。得到 m_{fuse} 后，我们通过计算它与 W 中不同单词编码的相似性来决定当前的 AC 模块应该关注哪个物体词语。对于 K 个候选词 W，我们通过下式计算这些物体词语的相似度得分 $S \in \mathbb{R}^K$：

$$S^k = W^{1\times d}(m_{fuse} \odot w^k) + b^1, \tag{2.8}$$

这里 $\odot$ 表示逐元素乘，k 表示第 k 个单词，$W^{1\times d}$ 是线性变换的权重。随后，我们通过对候选物体词 W 的加权求和来获得当前步骤 o_{cur} 的物体向量：

$$o_{cur} = \sum_{w^k \in W} w^k \cdot softmax(S)^k. \tag{2.9}$$

这里，$softmax(S)^k$ 表示在 softmax 函数之后 S 的第 k 个值。在获得 o_{cur} 之后，我们将其提供给注意力生成模块，以获得当前步骤关注的视频区域特征 V_{attn}。最后，我们通过 V_{attn} 的线性变换来计算当前 AC 模块的记忆向量：

$$m_{cur} = W^{d\times C} pool(V_{attn}) + b^d, \tag{2.10}$$

$pool()$ 表示加权后的视频特征 V_{attn} 沿 $T \times H \times W$ 维度的池化操作，我们这里使用平均池化。综上所述，我们的注意力控制模块的目的是在候选物体词之间选择合适的词语，并在视频中突出这些词语对应区域的特征。我们将最后一个 AC 模块中的 V_{attn} 作为视频的最终表征。最终表征会作为回答模块的输入以得出问题的答案。

2.2.5 回答模块

除了上述的三个模块，视频问答网络中还需要一个回答模块，用以将上面提取的视频和问题表征映射到最后的答案上去。在 TGIF-QA[7] 数据集中，一共有

4 种不同的问答子任务。它们分别是动作（Action），状态转移（State Transition），帧属性 (Frame QA) 以及重复动作计数 (Repetition Counting)。动作提问中，问题是关于视频中物体的动作。答案会提供 5 个选项，需要模型在其中选择正确选项。状态转移提问中，问题同样是关于动作的，不过形式通常是 *某物在做某事之前/后做了什么*，需要识别多个动作，并且进行时间推理。状态转移任务同样需要模型在 5 个选项中选择一个争取的答案。帧属性提问的问题通常是关于视频中某些物体的属性，更注重空间上的视觉信息，这个子任务的答案没有选项，需要模型在开放的词库中选择正确的词语回答。重复动作计数中，要求模型对视频中发生的动作次数进行计数，答案是一个数字，需要模型回归得到。为了回答问题，我们需要一个同时包含问题和视频的最终表征 F，这里我们直接将前面的问题表征 q 和经过注意力突出的视频表征 V_{attn} 级联得到 $F = [q, pool(V_{attn})]$。遵循先前在 TGIF-QA 数据集上的工作[7, 46]。我们对不同的子任务使用不同的回答模块：动作和状态转移两个子任务都属于多项选择题，对于多项选择题，我们用线性变换 $s = W^{1\times d}F + b^1$ 将每个候选答案的最终表征 F 映射到实数域。分数越高，表明候选答案是正确答案的概率越高。我们使用 ranking loss 对模型进行优化。对于帧属性这一开放式问答，我们构建了一个包含 l 个词语，即所有答案的词汇表。然后我们把开放式问答看作是一个 l 个选项的多项选择问题：经过线性变换 $W^{l\times C}F + b^l$ 得到 l 个选项的分数，再利用交叉熵损失对模型进行优化。对于重复动作计数问题，我们用线性回归将 F 映射为 $W^{1\times C}F + b$，表示重复的次数，并用 $l2$ 损失对模型进行优化。在测试时用 $round(W^{1\times C}F + b)$ 将次数取整，其中 $round()$ 表示取整函数。

第三节 基于弱监督匹配的视频问答模型的实验与结果

2.3.1 实现细节

数据准备。我们使用采用 spaCy 库实现的 GloVe[88] 来初始化问题和答案中的单词。在训练时，我们把所有句子填充成 32 个单词的固定长度，不足的部分用 0 填充。对于多项选择任务，AC 模块中的问题表征为 q。而在问答模块前，我们会把每个候选答案和问题分别级联，再经过 LSTM 得到和答案数相同的问题-答案表征，再作为语言表征输入给回答模块。在训练时，我们将每个视频分成 8 个等长的片段，然后从每个片段中随机抽取一帧，得到 8 帧作为输入。测试时则对视频采用等间隔采样，同样也取 8 帧用于测试。对于重复动作计数任

务，因为需要在时序上计数，我们在每个视频上采样 12 帧。采样的视频帧我们统一调整为 224 × 224 的大小输入网络。

特征提取器。对于视频特征提取，我们采用 ResNet-152[89] 来提取视频帧的特征。“pool5”前的最后一层输出的特征被用作中间视频特征 V，我们的所有训练都是基于这个特征。视频特征提取器中的所有神经网络层在训练中不更新，也就是说 V 是固定不变的。我们使用一个两层的，隐藏维度为 512 的 LSTM 作为问题编码器，把问题 Q 转化为问题表征 q。

训练参数设置。我们在 TGIF-QA[7] 数据集的 4 个子任务上分别进行训练和测试。所有模型采用 Adam[90] 优化器进行优化。初始学习率设为 0.0001。除重复动作计数任务外，我们用最大值 20 对梯度进行范数裁剪，重复计数任务不进行梯度裁减。模型使用 Pytorch 实现[91]。

2.3.2 在 TGIF-QA 数据集上的实验

在这节中，我们报告提出的方法在 TGIF-QA 数据集上的结果。

数据集与评估。我们首先介绍 TGIF-QA 数据集和数据集的评估度量。TGIF-QA[7] 数据集是一个具有 165k 个 QA 对的大规模视频问答数据集。这个数据集中有四个不同的任务：动作、状态转换、帧属性和重复动作计数。前两项是多项选择题，我们需要从 5 个候选答案中选出一个答案。帧属性是一个开放式任务，答案是关于帧信息的开放式词汇。重复动作计数的重点是视频中动作的计数，这个任务的答案是 2 到 10 之间的数字（10+ 视为 10）。前三者都使用回答的准确率作为评估指标，重复动作计数则使用预测的次数和答案的均方误差作为评估指标。

与基准的对比。我们将在本节中比较我们的模型和系统基线在 TGIF-QA 上的性能。结果报告在表2.1中。我们在表中阐述了 3 种不同的实验设定：*Baseline* 表示系统基线，我们移除所有的注意力模块作为系统基准；*Ours w/o AC* 表示我们的模型去掉注意力控制模块，在这个设置中，我们从问题中随机选择一个目标词，然后将所选单词输入注意力生成模块以获取注意力图；*Ours full* 是我们的完整模型，网络使用注意力控制模块动态地选择物体词，然后利用注意力生成模块生成注意力图。从结果中，我们可以观察到随机选择问题中的物体词语的 *Ours w/o AC* 也在所有子任务上取得了性能提升，这表明了注意力生成模块的有效性。而之所以随机选择也能得到提升，我们推测是因为大部分问题中出现的物体词语数都比较少，导致很多问题随机选择也能够关注到合适的词语。通过

表 2.1 “Baseline”表示系统基线。“Ours w/o AC”表示去除 AC 模块的模型。“Ours full”表示本文的完整模型。除了重复动作计数，其他三个任务都是数字越大表示结果越好。

方法	动作	状态转移	帧属性	重复计数[1]
Baseline	66.3	72.6	53.1	4.45
Ours w/o AC	69.0	75.6	55.1	4.41
Ours full	**70.9**	**77.2**	**55.9**	**4.36**

表 2.2 与代表性方法的对比，“Random chance”表示随机从可能的候选中选择结果，“Most frequent words”代表出现次数最多的回答。

方法	动作	状态转移	帧属性	重复计数
Random chance	20.0	20.0	0.06	19.62
Most frequent words	31.4	30.1	17.5	7.78
VIS+LSTM(aggr)[92]	46.8	56.9	34.6	5.09
VIS+LSTM(avg)[92]	48.8	34.8	35.0	4.80
VQA-MCB(aggr)[93]	58.9	24.3	25.7	5.17
VQA-MCB(avg)[93]	29.1	33.0	15.5	5.54
CT-SAN[49]	56.1	64.0	39.6	5.13
ST-VQA(SP)[7]	57.3	63.7	45.5	4.28
ST-VQA(TP)[7]	60.8	67.1	49.3	4.40
Co-memory[46]	68.2	74.3	51.5	**4.10**
OGA (Ours)	**70.9**	**77.2**	**55.9**	4.36

添加注意力控制模块，我们的模型进一步在所有任务上都获得了性能改进，表明我们的注意控制模块能够有效地在问题中选择出合适的物体。

与 SOTA 的对比。我们在表2.2中报告了我们的方法与之前性能最好的模型在 TGIF-QA 数据集上的性能比较。*ST-VQA(SP)* 和 *ST-VQA(TP)* 分别表示使用空间注意力和时间注意力的 ST-VQA[7] 方法。ST-VQA 方法使用 ResNet-152[89] 与 C3D[38] 提取视频特征，并将两个网络提取的特征融合得到最终视频特征 V。Co-memory[46] 方法使用记忆机制来使用光流和 RGB 两个模态的输入分别建模运动和外观信息。与这些方法不同的是，我们的方法特别侧重于探索空间注意力的潜力。为了提高效率，在整个实验过程中，我们只在 RGB 帧上使用 ResNet-152 作为视频编码器。相比于要使用 3D 卷积的 ST-VQA 和使用光流的 Co-memory，我们的方法大幅减少了计算需求。与以前的方法相比，我们还使

[1]本文中所有的“重复动作计数”实验在获得最终的注意图后，使用了与其他任务不同的设置。我们将在小节2.3.5的第二部分和表2.6 中进行介绍。

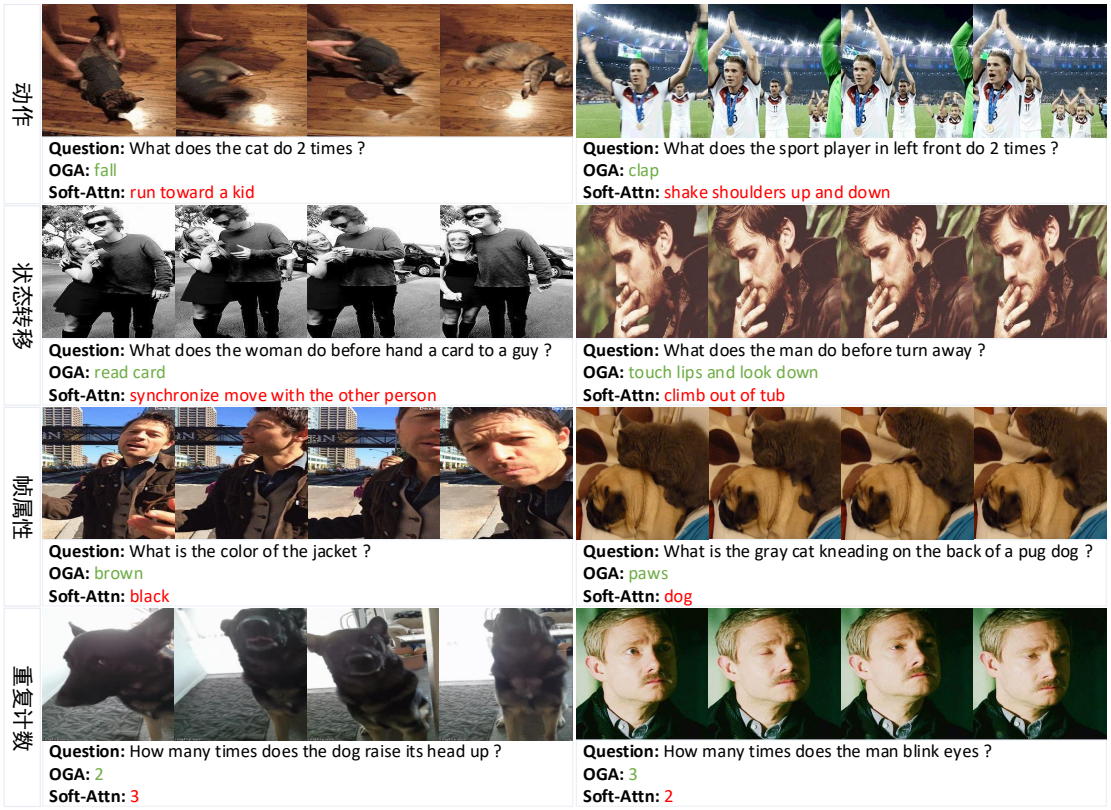

图 2.5 我们的物体引导注意模型在不同任务上的例子。OGA 表示我们的方法，Soft Attn 表示软注意力方法，如 ST-VQA[7]。正确答案是绿色，错误答案是红色。

用更少的视频帧作为模型的输入：ST-VQA 每四帧采样一次[7]；Co-memory 使用 32 帧图像作为视频输入，这是我们 8 帧输入的四倍。在视频输入时间分辨率为 1/4 的情况下，我们的模型在重复动作计数和状态转换任务上分别比 Co-memory 好 2.7%/2.9%。帧属性任务专注于物体级别的分析，有了明确的物体词语引导的注意力机制的帮助，我们在这项任务上有了 4.4% 的显著性能提升。不得不承认，我们在重复计数任务上落后于 Co-memory 方法，性能下降了 0.26。我们认为这种下降是合理的，因为重复计数任务严重依赖于输入帧的时间关系分析，在 Co-memory 中通过采用光流和更高的时间分辨率，可以更好地捕获输入帧。尽管如此，我们的模型在动作识别、帧属性和状态转换上的表现证明了从粗粒度弱监督中学习到的视觉-语言匹配的有效性以及基于匹配的物体引导注意力机制的优越性。这个结果还表明，虽然时间分析对视频理解很重要，但即便对于稀疏采样的 RGB 帧，物体级别的空间注意力方面仍有潜力可以探索。我们在图2.5中展示了不同任务上本文提出方法的一些结果示例，同时也和其他方法的结果进行了对比。

表 2.3 不同训练机制对应的结果。单独训练取得了较好的结果。

训练机制	动作	状态转移
单独训练	70.1	76.4
联合训练	70.9	77.2

表 2.4 用不同的训练数据子集训练匹配网络后，在动作识别和状态转移两个问答子任务上的实验结果。“帧属性 + 动作”表示这两个子集的合集。

测试任务 \ 训练子集	动作	状态转移	帧属性	重复计数	帧属性 + 动作
动作	66.9	70.1	70.9	65.6	67.1
状态转移	75.1	74.8	77.2	72.5	75.8

2.3.3 弱监督视觉-语言匹配训练

我们的方法中最重要的部分就是从弱监督的视频-句子对中挖掘物体级别的视频区域-物体词语信息并建立它们的配对，视频问答中的注意力机制都是基于这个区域匹配完成的。在这一部分，我们给出一些进行弱监督匹配训练时的不同选择以及得到的结果。

训练策略。首先我们研究匹配训练和 QA 训练两个网络的训练策略。弱监督匹配训练和视频问答的训练都有需要更新的网络参数。我们有两种训练策略可供选择：一种策略是同时训练弱监督匹配网络和视频问答网络，它们的参数同步更新，我们称这种策略为联合训练。我们用多目标学习的机制来进行联合训练，将匹配损失 ℓ_{align} 和问答损失 ℓ_{QA} 加在一起：

$$\ell_{joint} = \eta_a \ell_{align} + \eta_q \ell_{QA}, \tag{2.11}$$

η_a 和 η_q 表示两个部分损失的权重。在训练时，我们将前 3 轮（一轮表示整个训练集参与训练一次）的 η_q 设置为 0，随后将其设置为 1。η_a 按照经验设置为 10。另一种策略是先训练弱监督匹配的网络，得到视觉-语言匹配后，然后我们将这部分网络固定，后续训练问答网络的时候不再更新匹配网络，我们将这种策略称为单独训练。此时前三轮训练的方式和联合训练一致，不过后面的训练中将 η_a 设置为 0，不再更新匹配网络。我们在表2.3中报告了这两种训练机制得到的视频问答的结果。可以看到单独训练的性能要优于联合训练。

用不同的数据进行匹配训练。除了训练策略，使用不同的数据进行弱监督匹配训练也会影响最后的结果。TGIF-QA 数据集中有 4 个不同的子任务，它们

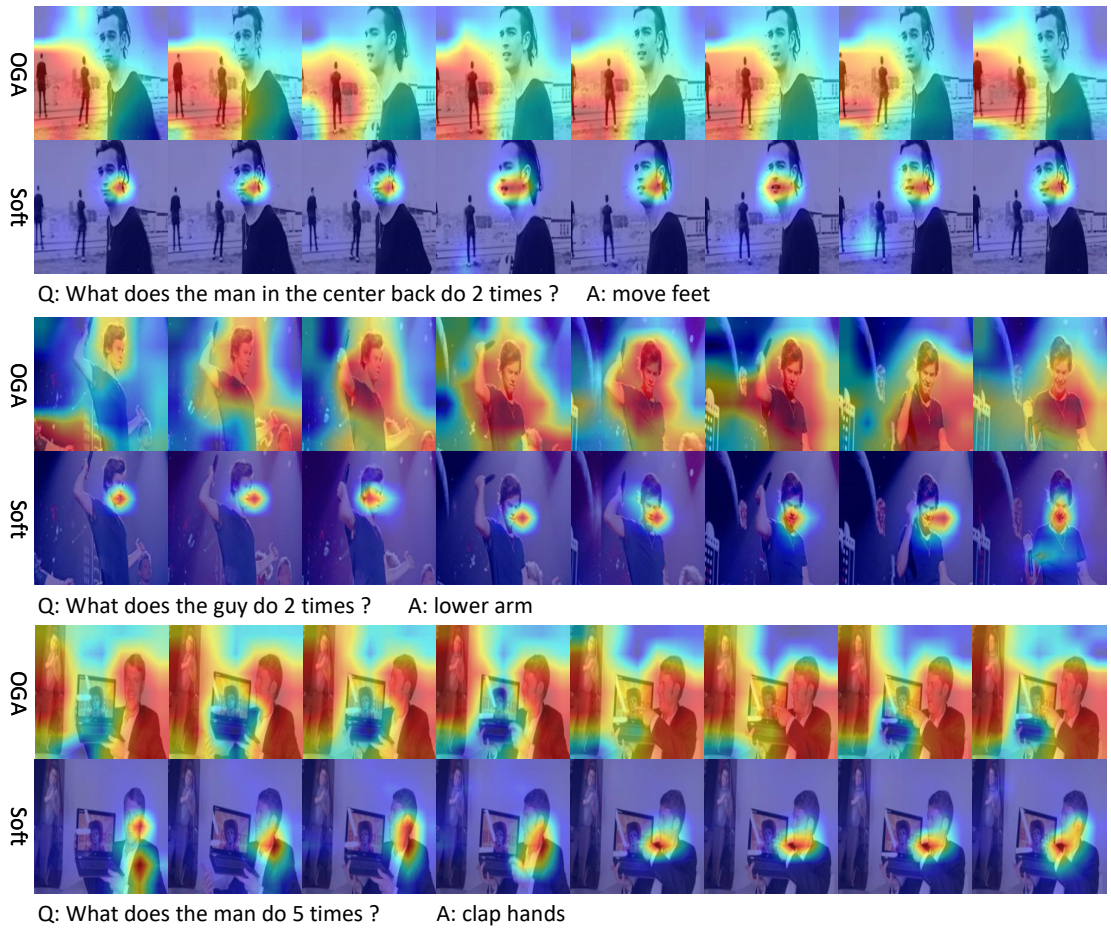

图 2.6 本文的方法（OGA）和软注意力方法（Soft）生成的注力意图比较。图中分别展示了三个视频问答上的 8 个视频帧上的注意力图。本文的注意力图更加平滑，能够覆盖需要关注的物体的大部分区域，而软注意力图只能注意到物体的很小的区域，并且很多情况下无法关注到正确的物体。软注意力图多数情况下关注的只是图片中的显著的物体，而非问题中的物体。

对应了 4 个不同的子数据集。这里我们分别用这 4 个子数据集中的视频-问答对来训练匹配网络。并用得到的视觉-语言匹配分别训练视频问答网络，观察各自的结果。我们在表2.4中报告了在动作识别和状态转移两个任务上的结果。从表中可以看到，不同的子集训练得到的匹配网络有很大的不同。在帧状态的子集上训练的匹配网络在两个任务上都取得了最好的效果。我们还将帧状态和动作识别两个数据集混合在一起训练匹配网络，得到的结果依然比帧状态上训练的网络差。帧状态这个任务特别关注物体以及物体的属性，而这正是我们训练的匹配网络的学习目标。我们认为这正是在帧状态子数据集上训练匹配网络效果最好的原因。

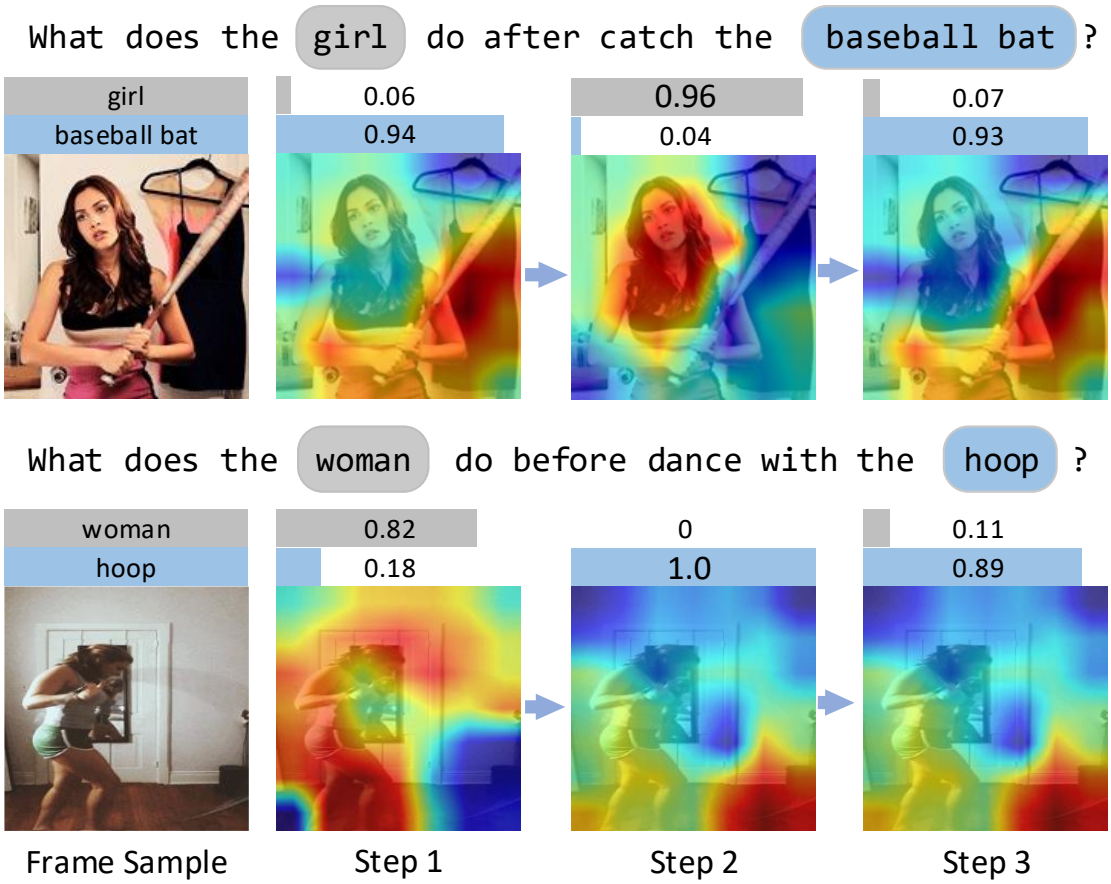

图 2.7 注意力控制模块的注意力转移展示。我们在这里展示了两个 QA 示例。我们为每个例子取样一帧，并在不同的步骤中显示注意力图。图中 *step n* 表示第 n 个注意力控制模块。我们同时展示物体词语的选择结果，不同的物体用不同的颜色表示。注意力图上方的条形图和数字表示每个步骤中的物体词语的得分。

2.3.4 模型可视化

注意力图可视化。为了更好地理解我们的注意力机制，我们演示了由我们的模型生成的注意力图并与 ST-VQA[7] 生成的软注意力图进行比较。在图2.6中，我们展示了三个视频片段和它们对应的问题以及答案。图中展示了每个视频用于测试的 8 帧，*OGA* 和 *Soft* 分别表示我们的模型注意力图和 ST-VQA 中的软注意力图。相比于软注意力图，我们的注意力图是在问题中目标词的明确引导下生成的，这为模型诊断带来了便利。因为我们的模型在生成这些注意图时“知道”它关注的是什么物体，因此具有更高的可解释性。观察图2.6，我们注意到，我们的注意力图更加平滑，往往涵盖了整个物体区域。与我们的模型不同的是，软注意力方法生成的注意力图倾向于关注一些物体的局部区域。如果视频中包

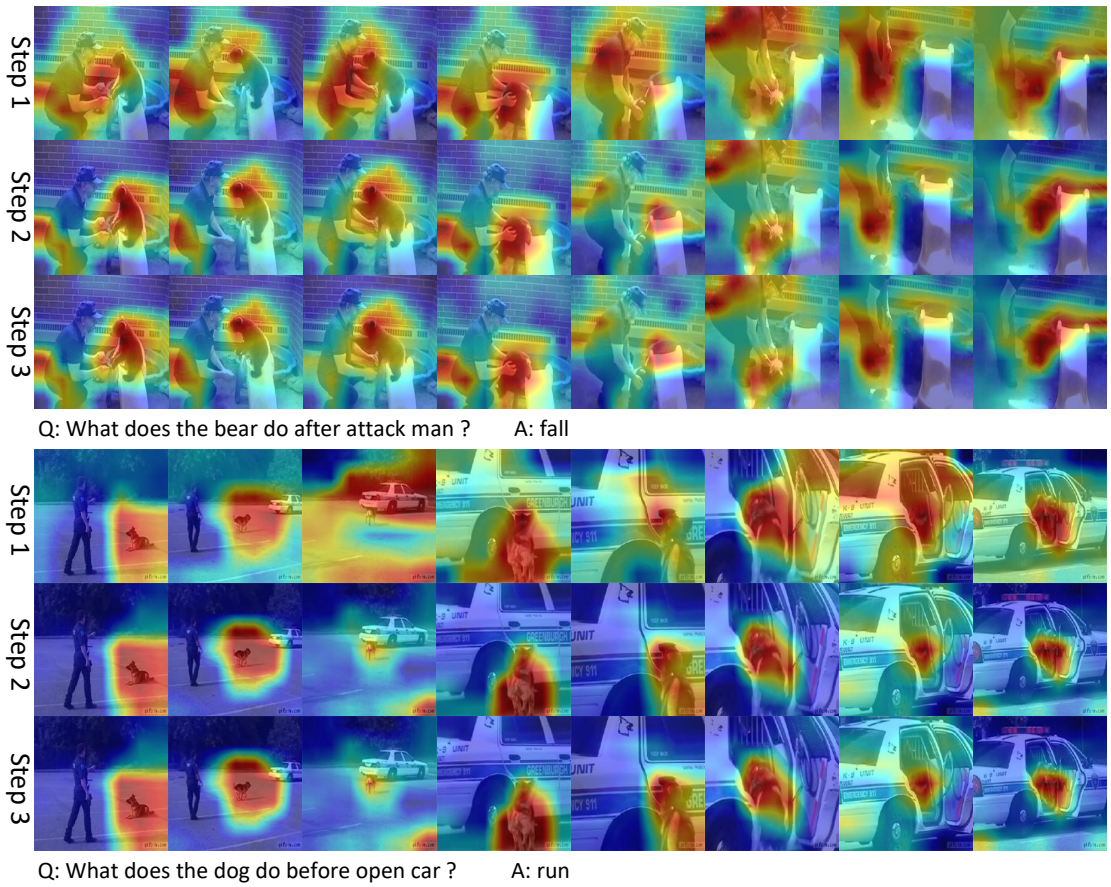

图 2.8 包含 8 帧的注意力图转移展示。本图展示了两个视频的区域注意力图。*step n* 表示第 n 个注意力控制模块。从图中可以观察到我们的网络成功地在物体之间转移注意区域，并最终关注到正确的物体区域。

含多个同一种类物体的实例，那么软注意力图往往只能关注到其中比较明显的那个实例的一部分区域，而不一定是问题中需要关注的区域。这是软注意力机制的模糊性导致的，软注意力模型的行为难以解释，产生的注意力图也可能受到数据集偏差的影响。例如图中的第一个视频例子，问题问的是后方中间的人在做什么，我们的方法成功注意到了后面的人，而软注意力方法则将注意力放在了前面的男人上。类似的，第三个例子中软注意力图只关注到了右边的显眼的男人的脸部，而我们的方法则关注到了两个男人的区域。

AC 模块中的注意力转移。我们的注意力控制（AC）模块可以选择问题中的不同物体词语，通过将数个 AC 模块级联，它们能够在不同的步骤中选择不同的物体词语，关注不同的视频区域，以模拟人类解决问答问题时的推理过程。得益于我们模型的可解释性，我们可以深入网络内部，看看注意力转移模块在每

一个步骤中选择了什么词语，关注了视频中哪个区域。我们在图2.7中通过可视化的例子来说明注意力转移过程的行为。图中一共展示了两个问答的例子，每个例子仅展示视频的一帧，我们的网络级联了三个注意力控制模块，因此这里我们也在每一行中展示这一帧在在三个模块中分别被关注的区域。从图中我们可以看到，在不同的步骤中，物体词语的得分会发生变化，从而导致视频区域中注意力的转移。进一步观察注意力区域的转移，我们发现模型在早期阶段倾向于关注人类区域，然后注意力通常转移到与人类相关的其他物体。由于人类是大多数视频中的共同物体，我们推测注意力转移模块倾向于将与人类相关的物体视为问题的决定性因素，并在学习过程中将注意力转移到这些区域。为了进一步展示注意力在所有帧中的的转移情况，我们在图2.8中展示了两个视频 8 个视频帧上的状态转移。从这两个例子上我们可以看到，网络的注意力在初始的 AC 模块中先关注到中间的物体，接着在最后的模块中转移到需要关注的物体的区域。

2.3.5 模型分析

注意力控制模块的数量对模型的影响。在本节中，我们将展示级联的注意力控制模块的数量 $step$ 如何影响模型在不同问答子任务上的性能。我们在表2.5中报告了 $step = 1,2,3,4$ 时的测试结果。重复动作和状态转换的最佳结果出现在 $step = 3$ 时，这表明注意力转移的机制促进了这两项任务的学习。而帧属性和重复动作计数则以较少的步数达到最佳效果，而改变步数对结果的影响较小。我们推测因为这两个模块关注的是动作发生的次数或者帧的属性，通常关注的是确定的某个物体的性质，并不像帧状态转移任务那样涉及物体之间的转移。因此注意力转移模块可以在这两个任务的早期步骤中聚焦到正确的物体。当 $step = 4$ 时，所有的结果都开始饱和，我们推测这是因为我们的问答任务中设计的物体词语数量并不多，因此 3 个级联的 AC 模块就足够完成注意力的转移。

空间注意力作为时间注意力。在这一节中，我们将演示基于物体词语的空间注意力如何提供时序上的信息，并判别出哪些帧对理解整个视频比较重要。具体来说，在获得注意力图 $\mathcal{A} \in \mathbb{R}^{T\times H\times W}$ 之后，我们沿着空间维度 HW 池化它，并通过 softmax 函数来获得时序上的重要性得分：$\mathcal{A}_{temp} = softmax(pool(\mathcal{A})) \in \mathbb{R}^{T}$。这个分数表明了视频中每一帧和目标物体词语的相关性。我们将 $\mathcal{A}_{temp}$ 作为时间注意力来进行问答实验：我们沿空间维度 H,W 平均池化视频表征 V 来得到 $V_T \in \mathbb{R}^{T\times C}$，然后使用 $V_T \odot \mathcal{A}_{temp}$ 计算时序上加权的特征，这个特征将用作最

表 2.5 不同任务上不同数量的注意力控制模块级联得到模型的效果对比。“Step”表示模块数量。

Step	动作	状态转移	帧属性	重复计数
1	66.8	75.4	55.2	4.36
2	69.0	76.1	55.9	4.39
3	70.9	77.2	55.5	4.36
4	70.6	76.0	55.7	4.36

表 2.6 空间注意力图的时序应用与结果。“SP2TP”表示池化空间注意力，将其转为时序注意力的结果；“SP”为本文原始的时间空间注意力设置。

方法	动作	状态转移	帧属性	重复计数
SP2TP	70.0	76.2	54.5	4.36
SP	70.9	77.2	55.5	4.43

终的视频表征使用。如表2.6所示，我们观察到，与我们原来的时空注意力设置相比，使用时序的注意力得到的结果是相当有竞争力的。实际上，我们报告的重复计数任务的最佳结果正是来自于此设置。计数任务在这个设置上得到最好的结果是符合预期的，因为计数任务主要侧重于时间的分析，这一点恰恰在本实验的设置中得到了强调。除了对帧进行加权以形成最终的视频表示之外，我们相信 $\mathcal{A}_{temp}$ 还可以用于其他领域，如长视频中时序上的动作的自动定位[94]。

第四节 本章小结

本章利用视频问答数据中的粗糙的视频-问答对的弱监督信息，通过匹配训练得到物体级别的视频区域-物体词语的对应关系。利用这种物体级别的对应关系，本章提出了一种简单有效的视频问答注意力机制。我们的方法利用问题中的目标词语来明确地引导网络关注对回答问题起决定性作用的视频区域。为了从问题中提到的多个物体词语中关注正确的物体，我们设计了一个注意力控制模块来选择词语，从而指导网络在视频的不同区域之间转移。得益于从弱监督信息中学习的物体级别视频区域-物体词语配对，我们的物体引导注意力（OGA）模型相较于软注意力方法可以生成可解释性更强的注意力图。TGIF-QA 数据集上的实验结果表明我们物体级别视觉-语言信息能够有效引导网络注意力，为问答任务提供帮助。

第三章 个性化图像语义分割中的弱监督方法

用户个性化图片指属于特定用户的图片。一般用于研究的数据集中图片之间是互相独立的，而个性化图片因为来自同一个用户，所以相互之间通常有语义或风格上的相关性，这种相关性体现了数据中用户的个性化特点。本章将无标注的用户个性化图片上的语义分割问题定义为 个性化图像分割问题。并将用户的个性化特点看作一种弱监督信号，探索如何利用图片间的相关性挖掘像素层级的语义信息，进而应用于个性化图片的语义分割任务中。

第一节 个性化图片分割背景介绍

语义分割是计算机视觉中一项吸引了大量研究的任务。该任务的目标是为给定图像的每个像素预测一个语义（类别）标签。与其他计算机视觉任务一样，研究者们利用深度学习强大的特征学习能力在语义分割任务上取得了极大的进展[50–52, 95–98]。这些利用深度学习的新技术主要集中在可公开获取的数据集上，如 Pascal VOC[1]，ADE20K[99]，CityScapes[53] 等。这些数据集都假定其中的图像是独立的。但是，这种假设在现实世界中并不成立。例如，在手机摄影中，用户可以通过拍照来记录他/她自己的生活，并组建起个性化的图像集。个性化的图像数据为视觉模型带来了挑战和机遇，如图3.1所示：一方面，个性化数据与公共数据集的分布不完全相同，从而导致在公开数据集上训练好的分割模型应用在个性化数据上时出现泛化问题；另一方面，图像来自同一用户是一种额外的弱监督信息，由于用户的个性化特点，来自同一用户的图像往往是相关的，如何利用这个信息去挖掘图片间的语义信息，并利用它来服务语义分割任务，这是一个潜在的研究方向。

个性化图像分割是在以前的研究中未曾讨论过的问题。该问题的困难主要集中在以下两个方面：(i) 首先，公开可用的数据集和用户自己的个性化数据之间存在很大的分布差距。尽管已经存在一些致力于从一个数据域向另一个分布不同的域进行域适应的图像分割方法，然而这些方法都集中在道路场景图像的分割上，学术界缺少一个聚焦于用户个性化图像的公开数据集。(ii) 另外，来自同一用户的个性化图像通常具有一些个人特征。如何在语义分割中适当地利用

图 3.1 在将分割模型从公开数据集迁移到个性化图片时，往往面临数据分布不同导致的效果不佳问题。同时，用户的个性化图片间往往成组地有明显的关联性（包含相似的物体或背景等）。我们研究了如何利用特定用户图片间相互关联这个弱监督信息来挖掘语义信息，辅助个性化图片的语义分割。

这些个性化特征仍然是一个尚未解决的问题。尽管存在上述困难，但实际应用中对个性化图像分割有很多的需求。例如，相机应用需要为用户的图像生成高质量的分割图，为后续的图像优化、合成等复杂操作提供基础。为了解决上述问题并促进有关用户个性化图像分割的研究，我们收集了一个称为 PIS（personalized image segmentation）的个性化图片数据集。PIS 数据集包含 15 个个人用户的个性化图像，总计有 10080 张图像。对于每个用户的个性化数据，我们随机选择大约 30% 的图像并标注其像素级的语义分割图以方便进行模型效果验证。为了让我们的个性化数据和现有的公开数据集对齐，我们考虑了 PASCAL VOC[1] 数据集中的 20 种常见物体的类别进行标注。就我们所知，PIS 数据集是第一个关注用户个性化图像的分割问题的数据集。通过使用这个数据集，研究者们能够更好地探索如何利用好用户个性化特点这一弱监督信号，为语义分割任务提供帮助。

在视觉或自然语言处理的任务中利用数据的个性化特征这种粗粒度的弱监督信息，这种想法在现有的工作中已经有所探索。以前关于视觉任务的个性化问题的研究[66–68] 通常从个人数据中提取所谓的全局记忆，用以表示用户的偏好或个性，这个全局记忆通常用一个多维空间中的向量表征。在后续的任务中，这

些方法通常将这个全局表征用作先验，为特定的下游任务提供一个代表该用户偏好的参考。与这些方法不同的是，在我们的问题中，我们从更广泛的角度考虑如何利用个性化这个弱监督信号。我们认为，无需将每个用户图像的个性化特点提取成一个全局特征。实际上，3.4.6中的实验表明一个全局特征并不能处理好我们的用户图像分割问题。究其原因，是因为图像分割问题是一个底层的视觉问题。不同于图片分类或者视频分类中需要预测一个全局的类别标签，分割任务需要预测出图片中每一个像素的类别标签。尽管特定用户的图片中包含了该用户的一些特点，但是用户图片中有各种不同的物体与场景。对图片中所有的像素都使用一个全局记忆做先验难免会过于模糊。为了利用好用户数据的个性化信息，同时避免全局表征的模糊问题。我们提出在相互关联的个性化图片间利用它们的语义相关性来辅助每个像素的预测。这种关联性的利用是局部的、特定的，而不是所有的像素都利用一个全局的表征。

在本文的个性化图像分割目标中，用户的个性化图像是没有标签的。我们已知的监督信息只有图像是属于某个特定用户的。为了让模型预测像素级别的语义信息，本文使用域适应技术（Domain Adaptation）技术，将已有的公开数据集作为源域，将用户个性化数据作为目标域。实现从公开数据集上学习分割，并在无标签的用户个性化数据上学习个性化特点，增强分割效果的目标。在已有的语义分割域适应方法中[56, 60, 100], 都将目标域的图片看作是独立分布的，即相互不关联。这个设定与我们的个性化数据特点不符，且无法捕捉并利用数据个性化这一监督信号。在我们提出的方法中，我们在域适应转换的过程中引入一个基于相似图片的语义辅助模块。这个模块能够在域适应的同时利用数据的个性化特征，提供更好的分割效果。

本章的贡献主要有两个方面：

- 我们首先提出个性化图像分割这一问题，并收集了一个称为 PIS 的个性化图像数据集，其中包含 15 个不同的用户数据。
- 我们选择了一些与此问题相关的最新工作，并在我们的数据集上报告了它们的表现。此外，我们提出了一种基线方法，旨在利用数据个性化这一弱监督信号，挖掘图片之间的像素级语义关联。该方法在提出的个性化数据集上实现了最优秀的性能。

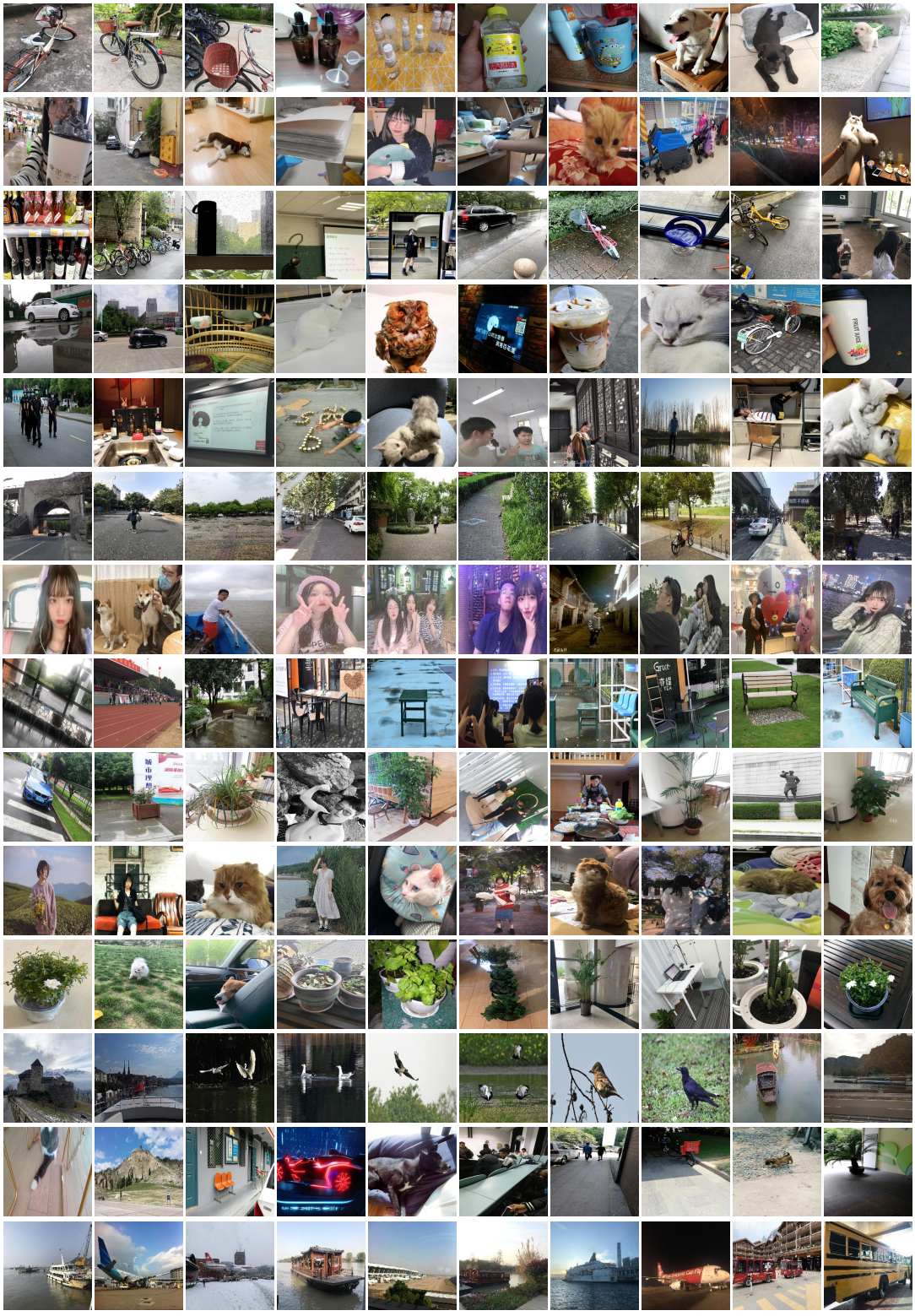

图 3.2 收集到的用户个性化图片展示。每一行的图片都随机取自一个用户。

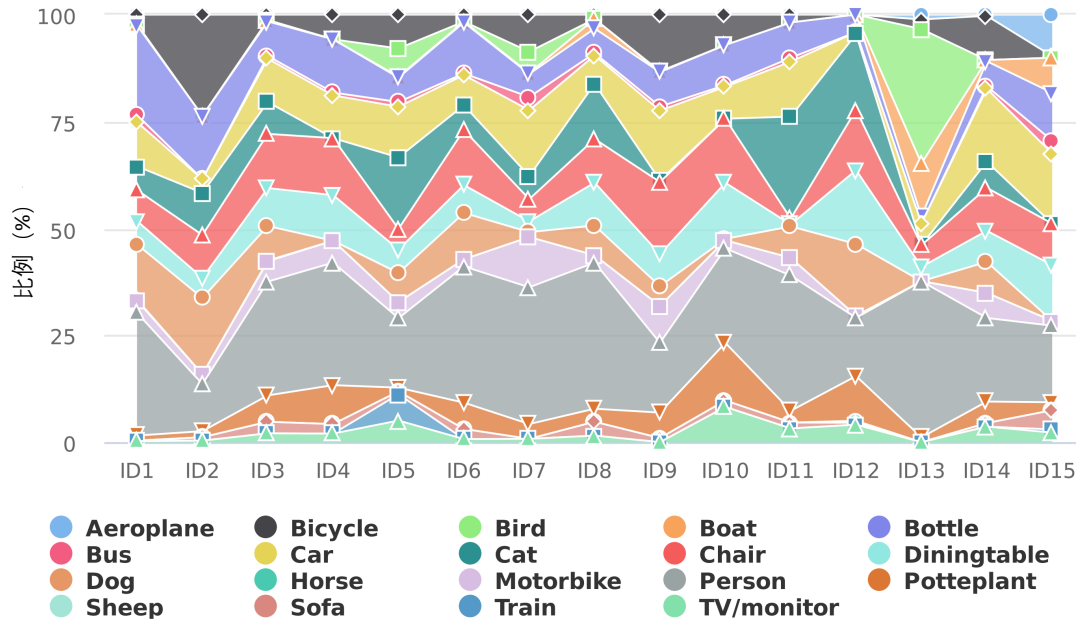

图 3.3 每个用户的个性化数据的不同物体类别的图片比例。

第二节 用户个性化图片数据集

本节从数据收集、数据标注和数据特征三个方面介绍我们收集的个性化图片数据集的相关信息。

3.2.1 数据收集

为了模拟现实世界中的个性化图片的数据分布，我们直接从不同的志愿者那里收集了属于他们的个性化图片数据集。每个志愿者都被要求在他/她的手机或照相机中导出图像以形成他/她的个人数据。为了保护隐私，我们要求志愿者浏览图像并过滤掉他/她不愿公开的图像。为了更好地从源数据集向个性化数据集进行域适应转换，我们的数据集关注 PASCAL VOC[1] 中的 20 个类别。最终，我们获得了包含 15 个用户个性化数据的 10080 张图像的大规模数据集。每个用户的个性化数据都具有与其他用户数据不同的分布。这些用户数据中的高层级的语义特点或低层级的统计特点都可能被用来帮助训练个性化的语义分割模型。在图3.2中，我们对部分用户的个性化图片进行采样，并展示这些图片。

3.2.2 数据标注

我们要求多位经过培训的专家对收集到的个性化数据进行标注。最终得到了所有图片的图像级标注（即图片中出现的物体的类别）以及测试集中图片的

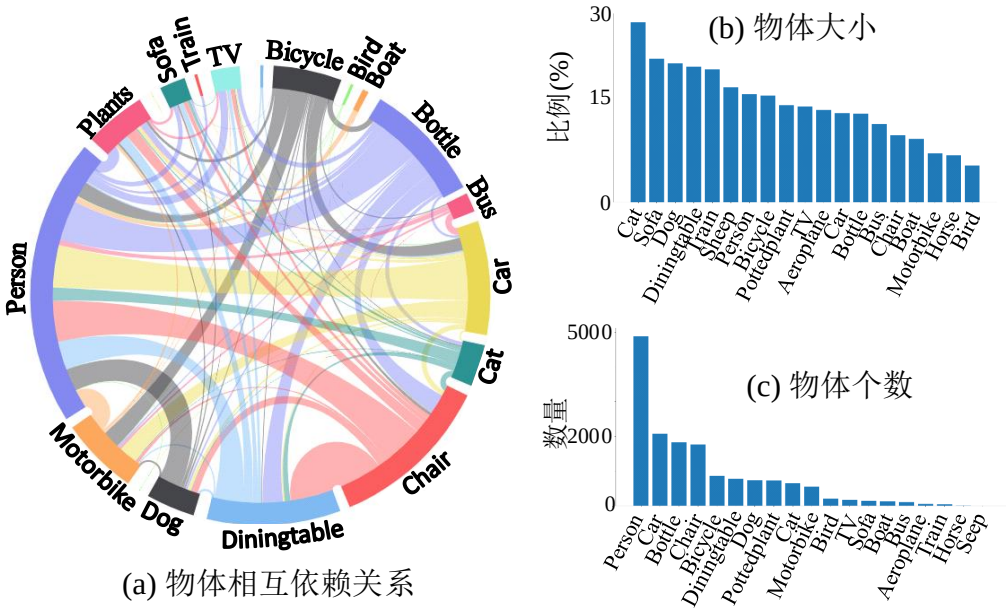

图 3.4 PIS 数据集的统计信息。(a) 数据集上不同类别物体的相互依存关系，相连的两个物体出现在用一张图片中。(b) 不同类别物体的平均占图片比例的大小。(c) 每个类别物体的实例数。

像素级标注。

图像级标注。图像级的标注主要用于数据集的分析以及未来可能的研究，本文提出的方法并不使用该标注。我们的数据集中的所有图片都被标注了与 PASCAL VOC 数据集[1] 一致的图像类别标签。我们在图3.3 中展示了每个用户的个性化数据的不同物体类别的占比。从图中可以看到大部分用户的照片中都含有很大比例的“person”，这符合普通人拍照的习惯。除此之外，其他类别物体的出现则各有多少，这反映了数据集的个性化这一特点，即不同的人会关注不同的物体。图3.4(a) 中展示了数据集中不同类别的物体的相互关系，两个类别有连接就表示这两个物体出现在了同一张图片中。通过分析这种依赖关系，可以学习到不同用户的拍照习惯等信息。很明显的，我们可以看到几乎所有的类别都大量地和“person”这一类别相关。这表明人们在拍照时一般都是以人为中心，然后记录一些和人有关的物体。

像素级标注。个性化图像分割问题的挑战是如何利用图片归属某个用户这一弱监督信息生成图像的像素级别分割图。为了让不同的方法在我们的数据集上评估效果，我们为每个用户的约 30% 的图像提供了像素级别的标注。和 PASCAL VOC[1] 一样，20 个类别的物体区域被标注成前景物体, 其他区域则被

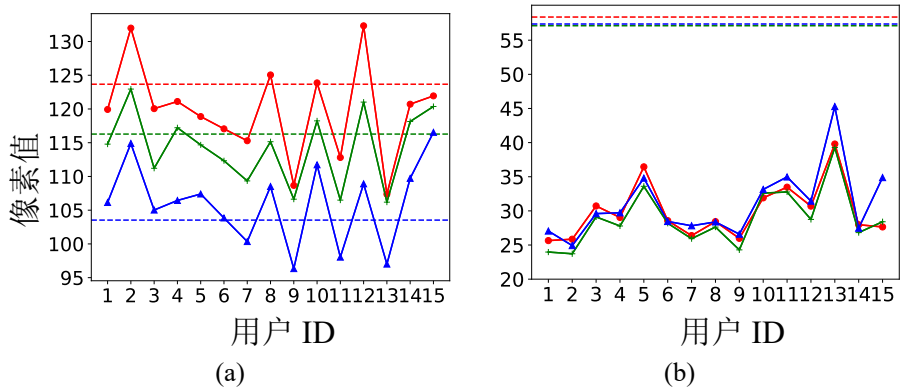

图 3.5 不同用户图片的平均像素值（a）以及其标准差（b）。图中红，绿和蓝 三种颜色分别表示三个颜色通道。虚线表示 ImageNet[101] 数据集的对应数值。

标注成背景。不同的类别区域在标注图中用不同的数值填充，从而形成一个像素级分割图，指示图像中每个像素的类别。我们在图3.4(b) 和图3.4(c) 中显示了不同类别物体的平均大小和不同类别物体在数据集中出现的次数。从图3.4(b) 中可以看到不同的类别在图片中所占的像素数量差别并不是很大, 这或许表明用户在拍摄不同物体时倾向于把它们缩放成占取景框比例类似的大小。图3.4(c) 则很明确地展示了用户图像中不同的类别数量有很大差异。

3.2.3 数据集特征

个性化的数据。我们的数据集最重要的特征是个性化。同一用户的图片间有一致的个性化特点，利用这一点可以辅助像素级的语义分割任务。另一方面，不同的用户的图片在底层细粒度属性（例如光照条件，图像质量）和上层语义属性（图像内容，背景）两个方面都可能有所不同。我们在图3.5（a）和图3.5（b）分别统计了每一个用户图片的不同颜色通道的灰度值均值和标准差，并和 ImageNet[101] 数据集（虚线表示）对比。从均值图中我们可以看到每一个用户都和 ImageNet 数据集有明显差异，而且不同的用户之间也各不相同。这表明了不同用户之间图片在底层细粒度属性上的差异。而用户间上层的语义属性差异可以从图3.3中观察到。不同用户的数据差异要求分割模型能够针对不同的用户进行相应的调整。通过观察图3.5（b），我们可以看到用户个性化图片的标准差与 ImageNet 有很大差异：各不相同且普遍偏小。这从另一角度说明我们的用户个性化内部的图片差异相对较小，也就是互相之间关联性强。

反映真实情况的数据。我们的个性化数据集非常接近现实情况。这体现在三个方面。首先，我们的数据集是直接从不同的用户那里收集的。这些图像忠

实地反映了他们在日常生活中关心并拍摄的事物，这意味着我们的数据集的结果可以反映出不同的方法在真实的用户数据上的有效性。其次，数据集中每个人的图像数量约为 600 到 900，这与一般的域适应方法中使用的其他数据集比起来相对较小。这种相对较小的数量是用户数据的真实特征，对于分割模型来说却是一种挑战：如何从较少的数据中学习到用户的个性化特点并应用到分割任务中。最后，我们数据集中的物体种类是长尾分布的，正如图3.4(c) 中展示的那样，不同的物体的数量是相差很大的。有些物体更有可能被拍摄，而另一些则没有。例如，在大多数图像中可能都有“person”，但只有很少的图像中包含“boat”。如何解决类别不平衡带来的问题也是一个有趣的研究方向。

第三节 基于弱监督上下文语义协同的个性化图片分割方法

在本节中，我们将介绍针对个性化图像分割而提出的方法。该方法的关键就是利用用户个性化这一弱监督信息挖掘用户图片间像素级的语义关联，进而为分割提供帮助。

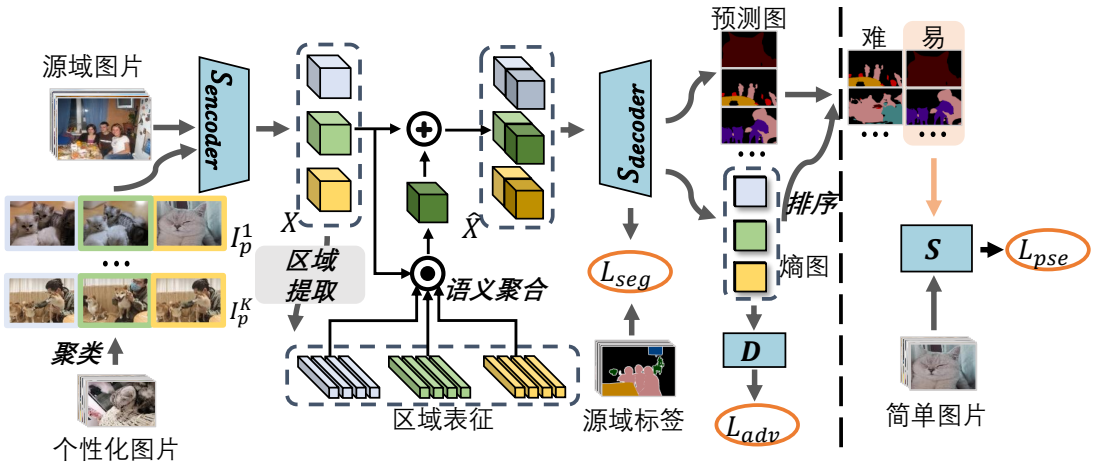

图 3.6 我们用于个性化图像分割方法的框架。我们的模型包含两个步骤：第一步是域自适应步骤，如 (a) 所示；第二步是伪标签微调步骤，如 (b) 所示。在 (a) 中，我们首先将个性化数据聚类为 K 组。然后在每个组中，我们使用组上下文语义辅助模块增强图像表征 X 以获得 $\hat{X}$。为简单起见，我们仅为每个组显示三张图像，并且仅显示标记为绿色的图像的组上下文语义辅助的过程。

本部分简要概述我们提出的个性化图像分割方法的流程。考虑有标注的公开数据集（源域）$\{I_s \subset \mathbb{R}^{3\times H\times W}\}$，它的 C-类分割标注图 $\{Y_s \subset \mathbb{R}^{C\times H\times W}\}$ 和无标注的个性化数据（目标域）$\{I_p \subset \mathbb{R}^{3\times H\times W}\}$。我们的网络需要从源域上学

习通用的分割能力，然后在个性化目标域上进行域适应学习。不同于先前的域适应方法中将目标域的数据看作是独立分布的，我们的用户个性化图像间互相关联，有用户的个性化特点，这为域适应过程提供了额外的弱监督信号。我们方法的关键思想是通过来自同一用户的其他图像的上下文语义来利用个性化图像之间的关联特性。图3.6中展示了我们方法的架构。我们的个性化图像分割框架有两个主要步骤：域适应步骤和随后的伪标签优化步骤。第一步，我们使用基于对抗的域适应框架，从源数据向个性化数据进行域适应学习。在域适应过程中，我们设计了一个组上下文语义辅助模块，利用用户图片之间的语义关联性提取出组上下文表征，辅助图片中每一个像素的分类。在第二步中，和其他弱监督图像分割方法类似，我们利用用户图像的伪标签来进一步微调分割网络。实际训练时，我们通过模型预测的分割图的熵来选择个性化数据中较为简单的图片，然后将这些简单图片的伪标签利用到网络微调过程中。

3.3.1 基于对抗的域适应技术

这部分简要介绍在框架第一步中使用的基于对抗的领域适应技术。记分割网络为 S，它接受图像 I_s 作为输出，并输出软预测图 $P_s = S(I_s) \in \mathbb{R}^{C\times H\times W}$，其中每个值 $P_s^{(c,h,w)}$ 表示像素 $I_s^{(h,w)}$ 属于类别 c 的概率。给定 I_s 的标签（即标注的分割图）Y_s，使用交叉熵损失函数

$$\mathcal{L}_{seg} = -\sum_{h,w}\sum_{c} Y_s^{c,h,w} log(P_s^{(c,h,w)}) \tag{3.1}$$

来训练分割网络。我们称损失 $\mathcal{L}_{seg}$ 为分割损失，它用于在有标签的源域图像上约束神经网络，使其能够在公开的标注数据集上学习到通用的分割能力。

除了用于学习通用分割能力的分割损失外，还需要将通用的分割能力迁移到目标数据域上，也就是我们的用户个性化数据域。为此，我们采用对抗训练的方法来缩小源数据 $\{I_s\}$ 和个性化数据 $\{I_p\}$ 之间的分布差异。首先，对于源图像和个性化图像的分割预测结果 P_s 和 P_p，我们通过下式计算它们的熵值图：

$$\begin{aligned} E_s^{h,w} &= \sum_{c} -P_s^{c,h,w} log(P_s^{c,h,w}), \\ E_p^{h,w} &= \sum_{c} -P_p^{c,h,w} log(P_p^{c,h,w}). \end{aligned} \tag{3.2}$$

熵值图反映了模型对于分割图的不确定性，$E^{h,w}$ 的值越大就表示在该像素的预测越不确定，也就是该像素对分割模型来说难度较高。对于同一张图像，不同

的模型会产生不同的熵值图；对于同一个模型，不同的输入图像也会产生不同的熵值图。在对抗训练中，我们通过训练一个判别器 D 来判断 E_s 和 E_p 的域标签，即判断熵值图是来自源域的图片还是来自目标域的图片。同时，我们通过训练分割网络 S，使得经过它的源和目标域图片产生的熵值图 E_s 和 E_p 特征相近，从而欺骗判别网络 D。在训练过程中，这两个网络依次更新，互相对抗。在这个对抗的过程中，我们就可以达到缩小源数据和个性化数据的预测图之间分布差距的目的，从而将分割能力从源域迁移到用户个性化数据域中。对抗损失可以表示为：

$$\mathcal{L}_{adv}(I_s, I_p) = -\sum_{h,w} log(1 - D(E_s^{h,w})) + log(D(E_p^{h,w})). \tag{3.3}$$

上面介绍的 $\mathcal{L}_{seg}$ 和 $\mathcal{L}_{adv}$ 两个损失函数分别负责分割网络 S 的通用分割能力的学习以及在用户个性化数据上的适应学习。同时进行两个损失的优化就是常见的图像分割域适应技术的常见做法。尽管这种对抗性训练可以一定程度上解决源数据和我们的个性化数据之间的分布不匹配问题。但是它不考虑个性化数据中的图像关联性，单独处理每个图像，从而无法考虑 $\{I_p\}$ 属于特定用户这一弱监督信号。为此，我们提出了一个组上下文语义辅助模块，以利用个性化数据的图像间的上下文信息。

3.3.2 组上下文语义辅助模块

为了在分割用户个性化图像时利用好用户个性化这一弱监督信号，我们设计了一个简单的组上下文语义辅助模块（Group Context Module）。该模块可以很方便地放置于分割网络后面，在进行域适应学习的同时利用用户互相关联的图像之间的上下文关系，辅助每个像素的类别判定。我们首先将每个用户的个性化数据聚类为多个组（group）。使得每个组中的图像包含相似语义特征，例如图像中的物体、背景等。在每个组中，我们为所有的图像提取区域语义表征。接着在预测每一个像素的类别时，网络会参考这些区域表征，利用这些表征作为先验进行预测。

图像聚类有很多可以选择的方法，这里我们选择最简单的 K 均值聚类。对于用户的个性化数据 $\{I_p\}$，我们把它们输入在 ImageNet[102] 上预训练的 ResNet-50[89] 网络，并在其全连接层之前获取特征 $\{F_p \in \mathbb{R}^{2048}\}$。之后，我们在 $\{F_p\}$ 上使用 K-means 聚类算法，获得 K 组图像，记为 $\{\{I_p^1\}, \{I_p^2\}, \cdots, \{I_p^K\}\}$。

将分割网络视作编码器 $S_{encoder}$ 和解码器 $S_{decoder}$ 的组合。编码器从组 K 中接受图像 I_p 作为输入并输出中间特征 $X = S_{encoder}(I_p) \in \mathbb{R}^{CH \times W \times H}$，其中 CH 和 W, H 分别表示 X 的通道数量和空间大小。我们的组上下文模块 F_{group} 通过使用组上下文学习增强后的特征 $\hat{X} = F_{group}(X) \in \mathbb{R}^{CH \times W \times H}$。组上下文模块可以分为两个步骤：区域表征提取和组内区域上下文聚合。

区域表征提取。受到[103]的启发，我们将图像 I_p 分割成 C 个软性的类别区域，并计算这些区域的表征。C 表示物体类别的数量，和个性化数据集中的数量一致。为了得到图像中的类别区域表征，我们需要利用一个辅助分割图。辅助分割图在图像分割方法中比较常见。在分割网络中，图像经过编码器变成中间特征 X，解码器则在这个中间特征的基础上学习并输出最终的图像分割结果。除此之外，一般还会在这个中间特征 X 前一层的输出 X' 后面再加一些卷积层并接上另一个解码器，这个解码器也输出一个分割结果，也就是辅助分割图。通过在辅助分割图上加分割损失，能够在较低层的特征上再加一层监督信息，使得网络有更好的多尺度性能。辅助分割图提供了一个较为粗糙的分割结果，可以用来参考每个像素的大致类别。我们将网络的辅助分割图记为 $P_p \in \mathbb{R}^{C \times W \times H}$，那么每个类别区域的表征可以用下式计算：

$$f_c = \sum_i r_{ci} X_i, \tag{3.4}$$

其中 i 表示空间位置，X_i 代表像素 i 的特征。r_{ci} 是像素 i 的权重，它是从该像素的辅助分割图的向量 $P_{pi} \in \mathbb{R}^C$ 在通道（即类别）这一维度上归一化计算得到的，计算方式为 $r_{ci} = softmax(P_{pi})_c$。对于有 N 张图的组来说，我们可以针对这个组提取 $N \times C$ 个区域表示。

组内区域上下文聚合。上面的区域表征提取过程为每张图片提取了不同类别区域的柔性表征，它能够作为图片中这个类别区域的代表。在后续分割种判断每个像素的类别时，我们将利用这些表征作为上下文的语义参考，也就是下面介绍的组内区域上下文聚合步骤。

在一些先前的考虑用户个性化问题的方法中，一般将个性化特征理解为某个用户的全局特性。因此它们一般对一个用户提取一个全局特征，然后用这个特征来处理所有的视觉区域。与之不同的是，在我们的个性化分割方法中，上下文并非从一个用户的全部图像中提取，而是从前面提到的图像分组中提取，也就是说个性化特征的提取是局部的。同时，我们也不会对所有的图像区域都使

用同样的上下文语义做参考，而是根据待分类像素和区域表征的相似度来决定哪些上下文语义能够作为像素分类的先验。具体来说，给定一个含有 N 张图像的用户图像分组的区域表征 $\{f_{i,j}|i \in [1,C], j \in [1,N]\}$，我们通过组内所有区域表征的加权得到 X 中的每个像素判别时所需的上下文语义表征：

$$c_{h,w} = \rho(\sum_{i,j} w_{(i,j),(h,w)} \sigma(f_{i,j})). \tag{3.5}$$

在这里，ρ 和 σ 是两个线性变换函数。权重 $w_{(\tilde{i},\tilde{j}),(h,w)}$ 是通过计算算像素 $X_{h,w}$ 和区域表征 $f_{\tilde{i},\tilde{j}}$ 之间的相似度而来的：

$$w_{(\tilde{i},\tilde{j}),(h,w)} = \frac{e^{s(X_{h,w},f_{\tilde{i},\tilde{j}})}}{\sum_{i\in[1,K],j\in[1,N]} e^{s(X_{h,w},f_{i,j})}}, \tag{3.6}$$

其中，$s(X_{h,w},f_{i,j})$ 是一个关系函数，可表示为 $s(X_{h,w},f_{i,j}) = \phi(X_{h,w})^T \varphi(f_{i,j})$，$\phi$ and φ 是由一个全连接层实现的两个变换函数。获得上下文语义后，我们将它和图像像素特征级联在一起，再接一个线性变换函数得到一个增强的像素特征，该特征不仅包含原有的视觉信息，还包含了来自其他图片的上下文语义信息：

$$\hat{X}_{h,w} = \psi([X_{h,w}, c_{h,w}]). \tag{3.7}$$

其中，$[*,*]$ 表示级联，ψ 是一个线性变换，由两个全连接层实现。这个增强的表征 $\hat{X}$ 将被输入解码器中并输出预测图：$\hat{P}_p = S_{decoder}(\hat{X})$。

对于 X 中的每一个像素，组上下文语义辅助模块将与像素特征相似度高的组内区域特征聚合在一起，为分割网络提供了额外的信息。通过将这样的模块添加在图像分割域适应的框架中，我们在从公开的源数据向无标注用户个性化数据的域适应学习的同时利用了个性化这一数据特点，利用图像之间相互关联这一弱监督信息，挖掘了像素级别的上下文语义关联，进而提高了个性化图片的分割性能。

3.3.3 使用伪标签优化

除了第一步域适应之外，用于图像语义分割的最新域适应方法[60, 104] 通常采用伪标签来进一步微调分割网络。这在利用图像标签的弱监督分割技术中也被广泛应用。在这些弱监督方法中有的[20] 尝试使用显著性图像分割的结果作为初始化伪标签并利用渐进学习的技术精调网络；有些则通过显示图像分类时网

络内部的注意力区域来生成区域注意力图[18]，并以此作为伪标签训练。我们的方法则利用第一步域适应训练出的模型来预测个性化图片的分割结果，并选择这些结果中比较可靠的部分作为伪标签。具体来说，先前介绍的熵图 E_p 是图像 I_p 的分割结果的不确定性指标，熵图中某个位置的值越大就表示分割网络对这个点的不确定性越高。而具有低不确定性的预测通常意味着输入的图像相对简单，并且结果具有较高的可靠性。因此，我们选择熵值较低的预测图作为伪标签。请注意，与 VOC 和 CityScapes 这样的数据集相比，每个人的个性化数据都相对较小，没有足够的数据来单独训练分割网络。因此，与[60] 不同，我们将伪标签添加到网络中，并增加额外的伪标签分割损失 $\mathcal{L}_{pse}$，而不是用伪标签替换源数据集的标签。

第四节　个性化图片分割模型的实验与分析

3.4.1　数据集和评估指标

我们的个性化数据集具有与 PASCAL VOC 相同的类别[1]。因此，在训练期间，我们将增强的 VOC 训练集用作源数据集，其中包含 10582 个带有 20 类物体的标记图像。在图像语义分割的研究中，一般使用 mIoU（mean Intersection over Union）作为定量评估指标。考虑一个预测分割图和它对应的真值图以及某种类别 c，预测图和真值图同时为该类别的区域记为 TP（True Positive），预测图预测为该类别而真值图不是该类别的区域记为 FP（False Positive）, 真值图为该类别而预测图不是该类别的区域记为 FN（False Negative）, 则该图像 i 上该类别的 IoU 计算为

$$IoU_i = \frac{TP}{TP + FP + FN},$$

所有图像在该类别下的 mIoU 则为

$$mIoU_c = \sum_{i}^{N} IoU_i,$$

N 表示图像的总数。然而，用户个性化数据通常是长尾分布的，这意味着不同类别的图像数量非常不平衡。如果按照上述的方法，分类别来计算 mIoU 并取所有类别的 mIoU 均值作为评测指标，就可能因为一些很少出现的类别的异常导致最终结果失真。因此我们使用了另一个指标 FIoU（Foreground Intersection over Union）来作为实验的主要衡量指标。FIoU 反映的是按照图像而不是按照类

别平均计算的 IoU。具体来说，考虑图像 i 的预测分割图和它的真值图，预测图和真值图标签一致且为前景的区域记为 T（True），预测图和真值图标签不一致且至少有一个为前景的区域记为 F（False）, 则该图像上的图像 IoU 记为

$$IoU_i = \frac{T}{T+F},$$

然后我们将所有图像上的 IoU 平均计算得到 FIoU:

$$FIoU = \sum_{i}^{N} IoU_i.$$

FIoU 相比于 mIoU 更能排除类别分布不均衡对最终数值的干扰。后面的实验结果也显示它在我们的数据集上表现更加稳定可靠。

3.4.2 实现细节

我们使用经过 ImageNet[102] 预训练后的 ResNet50[89] 网络作为分割网络的基础架构。分割网络配备了 PSP 模块[50]，与[56] 中类似。域适应训练过程中，输入是源域的图像，源域图像的标签以及聚类分成组的用户个性化图像。为了简化训练并节省计算量，我们不使用分组中的所有图像的区域来构建上下文语义。相反，我们将每个批次（mini-batch）采样的所有图像限制在同一组中，然后使用每个批次内的图像进行上下文语义的提取。我们采用随机裁剪进行图像增强，所有输入在训练过程中都被处理为 320×320 的大小。在伪标签微调的步骤中, 我们使用 $r = 0.5$ 的选择率来选择可靠的预测作为伪标签。本章将所有实验的批大小设置为 8。我们使用 SGD（Stochastic Gradient Descent）优化器进行网络优化，学习率设置为 2.5×10^{-4}，动量和权重衰减设置为 0.9 和 10^{-4}。实验代码由 PyTorch[105] 库实现。

3.4.3 和已有方法的性能对比

我们在用户个性化数据集上报告了一些已有的的域适应方法的性能，包括 AdaptSeg[58], MaxSquare[57], FDA[106], ADVENT[56] 和 MRNet[107]. 所有这些方法都单独处理目标图像，而不考虑个性化图像的相关属性。所有模型都以 VOC[1] 为源数据并以个性化数据为目标进行训练。像 MRNet[107] 这样的方法在伪标签微调步骤中仅使用目标域的伪标签来监督分割网络，而抛弃了源数据域的标签。由于我们的个性化数据每个用户的图像数量相对较少，导致难以训练好分割模型，所以我们为此类方法增加了 VOC[1] 标签的监督。实验结果由各种方法

表 3.1 个性化图像分割方法与 SOTA 方法 FIoU 指标的对比。这里报告使用 ResNet-50 和 VGG-16 两种不同 backbone 的结果。列号表示 15 个用户 ID。“Mean”列表示所有 Id 的平均。

Method	Backbone	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean
No-DA		45.39	51.99	48.95	47.60	58.03	48.15	56.86	62.45	48.23	45.14	62.37	51.68	48.56	48.13	41.57	51.01
AdaptSeg[58]		46.87	52.16	50.06	48.51	59.78	51.39	57.12	63.41	50.99	46.15	60.68	52.84	50.32	50.69	43.08	52.27
MaxSquare[57]		48.28	52.50	50.61	50.54	61.39	54.60	59.36	63.43	50.67	46.49	62.94	52.68	49.65	48.99	46.00	53.20
FDA[106]	ResNet-50	50.12	53.70	53.22	50.76	60.29	55.01	58.18	65.89	53.28	46.49	62.09	56.10	48.93	51.38	47.03	54.16
ADVENT[56]		53.39	57.33	52.42	52.51	64.63	55.04	60.61	61.69	55.34	49.18	66.05	57.83	56.04	54.38	52.34	56.59
MRNet[107]		**54.05**	58.62	54.29	53.17	61.72	57.24	62.20	66.46	56.75	50.27	66.76	54.20	53.87	54.38	51.38	57.02
OURS-S1		52.90	59.12	54.74	55.82	64.97	**60.38**	61.78	68.12	56.99	51.21	69.42	60.44	57.05	54.41	54.51	58.79
OURS-S2		53.28	**60.39**	**54.81**	**56.02**	**66.87**	60.11	**63.77**	**69.09**	**57.44**	**52.66**	**70.42**	**60.77**	**58.50**	**56.84**	**54.85**	**59.72**
No-DA		33.68	33.56	35.50	35.49	39.52	37.55	36.23	47.95	34.35	32.86	50.95	41.48	39.24	30.90	34.51	37.58
AdaptSeg[58]		32.70	37.65	37.16	33.54	40.55	41.11	43.17	52.12	36.95	31.83	49.04	40.97	33.54	31.49	34.06	38.39
MaxSquare[57]		36.17	32.99	38.81	37.36	42.64	42.03	49.88	50.06	37.99	35.93	51.33	41.98	36.27	36.35	37.13	40.46
FDA[106]	VGG-16	34.61	36.75	35.53	36.60	38.36	40.07	45.21	52.57	37.79	35.01	49.59	41.93	33.72	35.01	36.27	39.27
ADVENT[56]		39.89	44.39	39.88	40.01	49.89	44.24	47.99	54.59	43.84	38.29	53.00	43.07	42.83	40.02	41.36	44.22
MRNet[107]		34.40	41.18	36.67	32.18	44.63	38.12	41.99	46.78	39.51	36.54	39.39	44.17	35.93	37.17	38.35	39.13
OURS-S1		41.87	45.73	43.14	**44.04**	52.44	47.45	**52.32**	56.92	45.61	**42.67**	54.94	**48.38**	44.24	41.67	45.98	47.16
OURS-S2		**43.24**	**47.89**	**44.67**	44.00	**53.27**	**50.68**	52.18	**57.86**	**46.84**	42.34	**56.56**	46.28	**47.02**	**42.98**	**47.01**	**48.19**

表 3.2 个性化图像分割方法与 SOTA 方法 mIoU 指标的对比。这里报告使用 ResNet-50 和 VGG-16 两种不同 backbone 的结果。列号表示 15 个用户 ID。“Mean”列表示所有 Id 的平均。

Method	Backbone	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean
No-DA		28.05	29.18	30.78	33.05	42.52	31.31	35.85	28.63	39.60	36.99	33.15	38.51	29.78	32.75	31.85	33.47
AdaptSeg[58]		31.69	28.87	30.50	35.09	45.83	32.55	36.70	33.83	36.43	36.49	34.09	41.23	31.02	35.52	34.40	34.95
MaxSquare[57]		28.72	28.91	31.81	36.45	40.09	33.94	38.85	31.21	35.85	32.23	28.58	34.16	33.58	30.35	34.78	33.30
FDA[106]	ResNet-50	31.94	31.16	32.39	36.11	45.35	35.76	37.46	30.93	42.91	**43.28**	**37.51**	38.09	29.31	**37.25**	35.76	36.35
ADVENT[56]		36.04	34.04	36.98	39.98	43.76	40.52	41.59	29.69	36.26	39.19	33.46	39.05	38.17	33.43	37.44	37.31
MRNet[107]		**38.27**	**35.02**	36.98	36.54	43.99	**40.90**	40.22	36.26	32.35	33.10	36.26	31.78	37.77	35.89	32.24	36.51
OURS-S1		36.61	34.43	31.88	40.36	44.25	33.64	38.14	32.25	39.87	38.69	37.20	42.44	**39.60**	30.37	**42.18**	37.46
OURS-S2		33.85	33.38	**38.40**	**41.36**	**46.73**	37.58	**44.19**	**36.87**	**44.66**	42.03	37.42	**43.71**	35.12	34.18	37.89	**39.16**
No-DA		15.78	17.19	17.80	21.41	21.54	18.35	19.07	15.40	21.67	22.80	18.55	21.06	21.66	18.96	22.95	19.61
AdaptSeg[58]		16.59	19.04	18.96	23.81	21.43	23.12	25.47	16.26	23.33	22.60	19.08	20.20	22.12	20.10	24.30	21.09
MaxSquare[57]		18.46	18.19	18.46	22.29	26.01	23.88	25.36	17.07	25.01	25.37	19.99	20.56	24.52	20.82	25.90	22.13
FDA[106]	VGG-16	17.17	17.82	20.07	24.44	23.82	25.69	25.22	15.56	23.31	25.53	21.14	20.68	20.82	19.26	24.65	21.68
ADVENT[56]		27.32	21.95	**25.21**	25.46	35.50	23.75	28.18	**25.03**	**32.47**	29.73	24.76	25.00	26.74	24.76	26.31	26.81
MRNet[107]		20.30	16.46	20.92	27.62	29.70	25.46	27.74	22.30	30.64	26.46	17.64	27.12	23.43	23.86	26.08	24.38
OURS-S1		24.40	**24.93**	20.31	31.01	35.54	30.67	28.81	20.68	27.45	**32.06**	27.63	**31.44**	27.68	23.87	**33.53**	28.00
OURS-S2		**25.53**	24.40	22.62	**33.55**	**35.73**	**31.86**	**32.14**	21.84	28.62	30.57	**30.26**	24.97	**28.82**	**25.25**	31.91	**28.54**

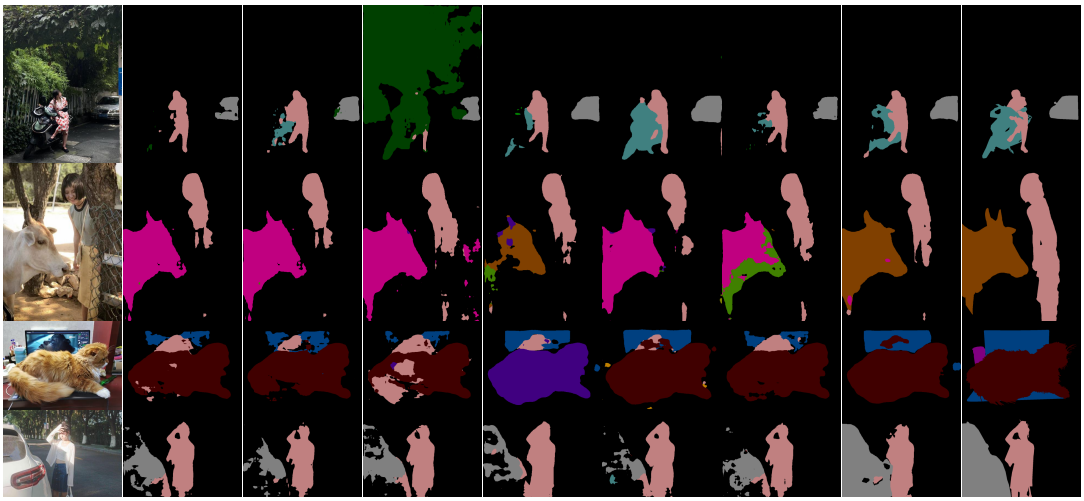

图 3.7 不同方法在测试集上的结果展示。Ours 表示我们的方法。GT 表示人工标注的分割图。

在个性化数据集中标注的验证集上进行测试得到。我们分别在表3.1 和表3.2中报告了 FIoU 和 mIoU。我们分别用 *OURS-S1* 和 *OURS-S2* 来表示没有加上第二步伪标签微调的结果和加上伪标签训练的最终结果。总的来说，使用 ResNet50 作为 backbone，*OURS-S1* 获得了 37.46 的 mIoU 和 58.79 的 FIoU。与基准方法 ADVENT 相比，该方法分别将性能提高了 0.15 和 2.20，这表明我们的组上下文模块非常有效。值得注意的是，与 FIoU 相比，mIoU 的 0.15 改善相对较小。我们推测这是由于个性化数据的长尾属性引起的：由于组上下文模块结合了其他图像的上下文以帮助学习，因此它倾向于在包含许多图像的类上表现更好，而可能会损坏稀有类的结果。在进行 mIoU 的评估时，最终的结果可能会受到这些稀有类别的影响。通过使用伪标签，*OURS-S2*，获得了 39.16 的 mIoU 和 59.72 的 FIoU，分别进一步提升了 1.7 和 0.93 的性能。我们在图3.7中展示了我们的用户个性化图像分割的结果，并与一些之前的方法的结果进行对比。从图中我们可以看到，我们的方法相比其他方法能够获得更加准确的分割图。

3.4.4 按照类别的 mIoU 结果

上面一节我们提到我们的数据是长尾分布的，类别间的数量不平衡会影响平均 mIoU 指标的结果。在这一部分，我们详细报告我们的方法在不同类别上的 mIoU 并与部分现有方法对比。我们在表3.3中报告了用户 $ID1 - ID10$ 上的

表 3.3 不同的方法在用户图像上按照不同类别评测的 mIoU 结果。对于每个用户，“Number”这一行表示测试图像中包含某类别的图像的数量。“Baseline”表示本文方法的基线，“MRNet”是最近的代表方法，“OURS-S1”和“OURS-S2”分别表示本文方法的第一步训练结果和最终结果。由于空间有限，我们仅展示 $ID1-10$ 的结果。

方法	ID	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	table	dog	horse	m.bike	person	plant	sheep	sofa	train	tv	Mean
Number		0	19	0	2	177	14	87	47	61	46	109	2	24	242	10	0	0	1	2	
Baseline		-	56.06	-	0	47.16	46.68	53.94	66.23	19.9	31.49	74.5	29.24	19.9	52.34	15.91	-	-	27.28	0.01	36.04
MRNet[107]	1	-	54.11	-	0.27	51.99	38.16	56.04	56.31	22.2	25.98	67.58	54.39	25.18	56.57	22.63	-	-	38.76	3.92	38.27
OURS-S1		-	41.31	-	54.67	45.91	34.98	55.57	62.48	20.57	27.8	75.43	7.55	18.82	54.41	25.75	-	-	23.56	0.3	36.61
OURS-S2		-	42.86	-	4.68	42.81	43.51	56.87	54.29	17.63	33.6	69.17	8.58	8.33	56.48	19.44	-	-	49.5	0	33.85
Number		0	267	0	0	167	0	36	111	113	53	201	0	26	124	16	0	9	0	4	
Baseline		-	70.61	-	-	34.37	-	37.83	67.37	42.02	14.03	61.31	-	14.68	31.47	13.85	-	18.64	-	2.3	34.04
MRNet[107]	2	-	67.14	-	-	50.98	-	38.62	66.95	26.65	17.9	64.32	-	14.47	29.75	24.21	-	14.64	-	4.55	35.02
OURS-S1		-	66.14	-	-	43.46	-	32.19	67.01	39.9	15.61	70.58	-	8.85	35.71	18.33	-	7.95	-	7.47	34.43
OURS-S2		-	72.19	-	-	40.84	-	26.68	69.25	35.15	14.28	71.97	-	8.24	37.13	20.3	-	0.77	-	3.74	33.38
Number		0	8	0	2	51	4	64	47	80	56	51	1	30	168	39	0	17	0	13	
Baseline		-	19.14	-	46.84	21.77	2.47	53.79	68.44	22.15	28.97	59.16	75.85	11.29	63.26	65.14	-	11.02	-	5.34	36.98
MRNet[107]	3	-	16.13	-	20.14	27.36	17.71	60.81	73.04	25.58	30.75	59.8	60.22	24.16	65.34	60.65	-	9.75	-	3.19	36.98
OURS-S1		-	4	-	21.16	11.09	0.88	68.77	68.18	28.88	40.03	57.37	19.43	17.52	63.19	56.61	-	10.15	-	10.96	31.88
OURS-S2		-	16.33	-	51.35	17.34	11.18	66.17	70.17	25.31	39.87	58.7	66.95	15.2	62.8	60.54	-	6.95	-	7.21	38.40
Number		0	26	0	0	56	3	45	0	60	48	0	0	23	129	41	0	10	0	9	
Baseline		-	68.98	-	-	53.17	31.65	61.62	-	30.23	24.49	-	-	11.78	70.51	52.07	-	35.33	-	0	39.98
MRNet[107]	4	-	60.01	-	-	60.14	28.21	69.56	-	31.04	9.93	-	-	10.67	74.17	51.86	-	6.33	-	0	36.54
OURS-S1		-	66.28	-	-	56.95	24.89	73.41	-	33.62	23.62	-	-	6.13	75.21	53.54	-	30.36	-	0	40.36
OURS-S2		-	55.5	-	-	53.53	23.07	79.65	-	31.29	33.58	-	-	5.92	72.71	62.36	-	37.22	-	0.11	41.36
Number		0	36	30	1	26	5	54	76	22	25	31	0	17	74	4	0	3	28	23	
Baseline		-	69	72.52	0	52.03	75.36	81.65	78.82	2.92	21.65	59.49	-	15.38	36.36	4.24	-	0	75.65	55.1	43.76
MRNet[107]	5	-	62.29	70.63	0	47.89	73.75	80.82	80.36	0.78	7.85	65.37	-	27.9	48.78	0.86	-	20.71	65.48	50.41	43.99
OURS-S1		-	66.77	59.21	0	37.47	78.14	84.88	77.17	2.96	19.7	64.23	-	21.17	44.54	12.77	-	0	73.67	65.26	44.25
OURS-S2		-	71.56	69.59	0	41.27	82.77	80.12	79.03	1.28	35.9	68.19	-	22.24	54.55	6.34	-	0	75.91	58.86	46.72
Number		0	9	0	1	66	2	39	32	71	36	60	0	9	176	34	0	13	0	4	
Baseline		-	56.93	-	22.87	47.74	45.37	31.27	77.46	14.77	32.74	69.24	-	25.24	60.78	66.24	-	12.58	-	4.09	40.52
MRNet[107]	6	-	61.65	-	0	51.78	61.25	32.99	73.44	24.24	32.75	71.01	-	17.58	66.11	62.36	-	16.77	-	0.71	40.90
OURS-S1		-	6.77	-	19.94	46.33	7.11	36.45	78.14	21.04	30.67	68.52	-	9.09	65	65.15	-	7.64	-	9.09	33.64
OURS-S2		-	36.89	-	21.09	31.26	51.77	40.24	68.66	24.38	37.34	69.9	-	7.86	65.65	58.63	-	10.43	-	2.04	37.58
Number		0	45	26	0	28	16	80	28	27	13	6	0	61	166	18	0	0	0	4	
Baseline		-	40.05	64.25	-	0.61	26.8	77.09	81.7	33.16	19.19	22.13	-	58.27	77.13	40.35	-	-	-	0	41.59
MRNet[107]	7	-	49.83	54.09	-	1.75	26.31	76.96	88.05	20.58	16.42	8.78	-	58	80.32	41.74	-	-	-	0	40.22
OURS-S1		-	49.05	44.63	-	2.64	38.72	74.38	81.54	23.83	3.68	8.44	-	57.46	77.7	33.81	-	-	-	0	38.14
OURS-S2		-	54.21	62.57	-	4.7	50.44	79.03	87.74	39.68	10.2	6.23	-	59.55	78.59	41.49	-	-	-	0	44.19
Number		0	6	0	13	34	3	39	76	60	59	41	0	10	201	18	0	19	0	9	
Baseline		-	15.93	-	18.89	31.23	4.17	54.33	74.94	6.61	19.04	56.13	-	8.73	73.59	26.33	-	17.84	-	7.93	29.69
MRNet[107]	8	-	19.44	-	16.21	35.69	63.37	56.97	75.43	11.13	26.01	63.51	-	14.2	76.62	24.81	-	6.97	-	17.28	36.26
OURS-S1		-	9.25	-	15.21	44.85	17.84	58.45	77.29	9.95	26.09	59.67	-	3.47	80.78	21.29	-	17.79	-	9.63	32.25
OURS-S2		-	9.39	-	32.35	47.81	24.01	58.36	79.15	6.96	8.3	65.49	-	25.14	80.7	22.98	-	32.52	-	23.06	36.87
Number		0	69	1	0	42	5	85	2	88	39	24	0	45	85	31	0	5	0	0	
Baseline		-	62.81	22.83	-	42.13	0	87.76	5.32	24.81	8.28	78.95	-	9.23	65.4	10.22	-	53.65	-	-	36.26
MRNet[107]	9	-	64.65	0.7	-	53.46	4.05	85.41	1.98	37.43	3.79	67.21	-	12.15	58.87	19.68	-	11.2	-	-	32.35
OURS-S1		-	58.23	7.02	-	53.64	26.56	90.85	2.99	31.31	14.16	72.39	-	33.5	65.58	16.25	-	45.78	-	-	39.87
OURS-S2		-	64.74	28.45	-	55.6	64.46	92.29	2.29	27.02	12.86	68.07	-	43.29	63.89	12.35	-	45.21	-	-	44.66
Number		0	19	0	0	25	1	20	0	40	35	1	0	5	59	36	0	4	0	22	
Baseline		-	71.64	-	-	65.33	0	80.5	-	24.21	20.07	0	-	17.59	62.95	52.09	-	13	-	62.95	39.19
MRNet[107]	10	-	65.25	-	-	47.52	0	71.5	-	26.94	13.51	0	-	8.94	60.74	62.26	-	6.88	-	33.71	33.10
OURS-S1		-	68.72	-	-	66.3	0	72.79	-	32.85	23.02	0	-	15.7	66.74	62.45	-	0	-	55.66	38.69
OURS-S2		-	71.69	-	-	60.39	0	73.14	-	32.41	23.42	0	-	13.16	64.39	64.53	-	43.94	-	57.29	42.03

结果。所有的实验都使用 ResNet-50 作为基础网络。表中所列的 4 种方法分别是：*Baseline* 是没有利用用户个性化信息作辅助监督的方法；*MRNet*[107] 是一个最近提出的域适应分割方法；*OURS-S1* 和 *OURS-S2* 分别表示我们的方法只进行第一步训练和加上伪标签微调完整进行两步训练的结果。我们同时报告了每个用户验证集中包含不同类别物体的图像的数量，并列在 *Number* 一行中。我们使用的上下文语义模块能够通过相似图片组内的上下文语义来辅助分割，这个模块需要有内容相似的图片之间的语义相互补充。然而对于少见的物体类别，因为无法得到相似语义的图片的辅助，上下文模块则有可能无法发挥作用。而因为我们的聚类方法并不能保证包含这些图片的组内语义都相关，所以组上下文模块还有可能因为组内的无关图片引入干扰性的上下文语义，导致对这些图片效果不佳。以 *ID*6 为例，少见的物体类别如”bicycle”（9 张图片），”boat”（1 张图片）以及”bus”（两张图片），*OURS-S1* 的 mIoU 明显比 *Baseline* 要低，然而比较多见的类别如”car”（39 张图片），”cat”（32 张图片）和”person”（176 张图片），*OURS-S1* 超过了 *Baseline*。对所有类别平均，*OURS-S1* 仅达到了 33.64 mIoU，比 *Baseline* 的 40.52 要低。这是因为这些少见的物体类别上的数值较低导致的。如果观察 FIoU，就能发现 *OURS-S1* 达到了 60.38，比 *Baseline* 的 55.04 要高。我们认为通过改进聚类方法，或者减少上下文语义聚合过程中的无关语义的干扰，都能改善这个问题。

3.4.5 用户图片分组分析

为了对用户个性化数据的图像间的相关性以及我们的聚类分组效果有一个直观的了解，我们在图 3.8中展示了用户 *ID*12 上的一些分组的结果。我们直接使用 K 均值聚类算法在图片经过由 ImageNet 上训练的 ResNet50 网络提取的特征上进行聚类。可以预期的是，我们的聚类会将前景相似的图片分在同样的组中。从图3.8中，我们也可以观察到这一点：第一行的分组大都是宠物的图片，第二行主要是一些显示器，第三行是一些家具的图片，而最后一行的分组多是一些盆栽。另一方面，我们也可以观察到同一组中并不都是同样的物体的图片，例如第四行多是盆栽的分组中也出现了两张宠物的图片。这提醒我们简单的 K 均值聚类算法并不能够达到完美的分组效果，未来应该应该应用更加先进的分组方法来获取更可靠的分组，从而得到更好的组内上下文语义。

图3.8中只展示了用户 *ID*12 的个性化图片的 4 个分组，而一个用户的图片分组数量一般是多于 4 的。在 K 均值聚类算法中，组的数量是一个需要自定义

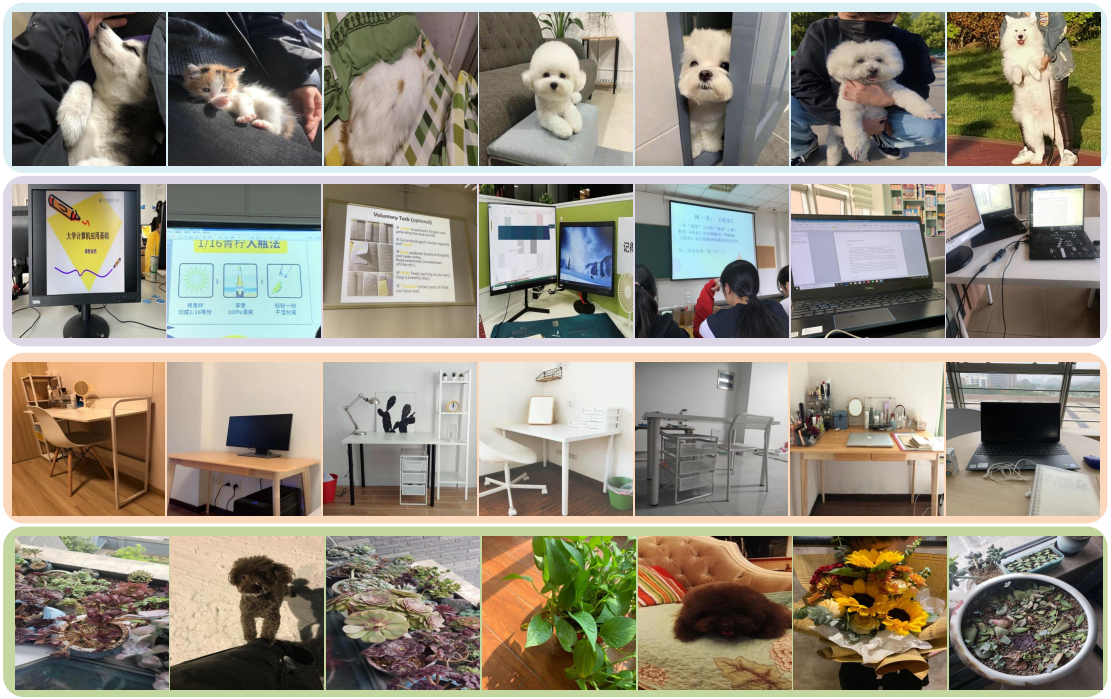

图 3.8 用户 *ID*12 上的图片聚类分组效果展示。图中包含随机选择的 4 个分组，每一行代表一个小组。为了简化显示，我们只展示每个组的 7 张图片。

的参数，下面我们将每个用户的个性化图片分为不同数量的组，并研究在推理时组的数量如何影响分割的性能。如表3.4中所示，不同的行表示不同数量的组。在 *Groups=1* 的情况下，一个用户的所有图像都被视为一组。这样在计算组区域上下文时，就相当于在所有用户图片中随机选择图像来计算上下文语义，从而可能会将不相关的图像考虑进来导致网络混乱。我们可以看到此时的平均 FIoU 为 46.64，相对较低。当 *Groups=200* 时，平均的 FIoU 46.66 也比较低。此时每个组中的图像数量太少，所以无法为组上下文模块提供足够的语义补充。虽然从平均来看，我们可以观察到中等数量的 80 组可以实现更好的性能。但是我们仍然可以注意到不同的用户的最佳 FIoU 出现在不同的分组数量上，我们认为这是因为不同用户的数据具有各自的分布，从而适合不同数量的分组。将来，我们将研究更灵活的方法来对个性化图像进行聚类，而不是使用固定的数量来对所有用户进行分组。

3.4.6 组上下文模块的有效性

在本节中，我们通过与两个基线进行比较来研究本章提出的组上下文模块的有效性，结果报告在表3.5中。其中 *None* 指的是直接将编码器的输出特征 X 输

表 3.4 不同的分组数量对不同用户个性化图像分割效果的影响。不同的列表示不同的用户 ID，“Mean”表示在所有用户上的平均。本表使用的指标是 FIoU。

数量	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean
1	42.39	46.11	42.95	43.79	52.45	46.15	51.23	55.88	45.57	42.51	55.30	47.89	43.66	39.54	44.12	46.64
10	42.49	45.52	42.75	44.07	52.64	46.54	50.85	56.52	45.81	42.25	54.75	46.86	44.62	41.18	45.36	46.81
80	41.87	45.73	43.14	44.04	52.44	47.45	52.32	56.92	45.61	42.67	54.94	48.38	44.24	41.67	45.98	47.16
200	42.11	45.66	43.14	43.33	52.08	45.85	52.05	56.51	45.78	42.13	53.75	46.73	43.80	41.59	45.35	46.66

入给解码器得到分割图，不使用组上下文模块。*Global* 表示使用一个全局的语义表征来增强所有的像素特征，即使用一个全局表征来代表一个用户的个性化特征，这类似于[71] 中的方法。本部分的实验中，分割网络的主干是 VGG-16[108]。如表3.5中报告的，*None* 的基线实现了 44.22 的 FIoU。*Global* 轻微提升了 0.34 的性能，这表明在我们的场景下，全局的用户表征不够有效。这个结果符合我们在前面的分析：一个用户的个性化图片之间虽然有相关性，但是这种关联是相对较弱的，并不是所有的图片都强相关。如果使用一个全局表征来代表一个用户，然后在这个用户的所有图片上作为先验，这个先验会过于模糊，难以起到明显的效果。*OURS* 表示在网络中加上组上下文模块的我们的方法，它的 FIoU 达到了 47.16，相比于 *None* 的基线提高了 2.94 的性能，这表明了我们提出的组上下文模块的有效性。

表 3.5 组上下文模块的效果验证。“None”和“Global”分别表示无上下文和全局上下文。“Mean”表示所有 ID 上的均值。

方法	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean
None	39.89	44.39	39.88	40.01	49.89	44.24	47.99	54.59	43.84	38.29	53.00	43.07	42.83	40.02	41.36	44.22
Global	38.89	42.13	42.13	38.96	50.88	45.09	51.62	55.67	44.23	38.51	49.80	45.70	43.20	39.84	41.71	44.56
OURS	41.87	45.73	43.14	44.04	52.44	47.45	52.32	56.92	45.61	42.67	54.94	48.38	44.24	41.67	45.98	47.16

3.4.7 个性化训练的价值

在本节中，我们混合来自所有用户的图像以形成一个混合图像集 *MixAll*。同时，我们从 *MixAll* 中随机采样一个 1/15 的子集，得到与单个用户图像数量类似的 *MixSample*。我们在这两个图像集上训练模型，然后在不同用户的数上据评估训练出的模型，结果在表3.6中展示。表中上面两行分别表示上述两个采样数据上进行训练的结果，*Personal* 则表示我们的方法，即在对应用户上训练并测试的结果。尽管 *MixAll* 具有大约 15 倍的图像用来进行域适应训练，它的所有用户的平均 FIoU 却只有 45.56，低于 *Personal* 的 47.16。这个结果显示了从个性化数据

表 3.6 将个性化数据集混合并进行实验。“MixAll”表示混合所有用户图像，“MixSample”对“MixAll”的图像进行 1/15 的采样，以使其和每个用户的个性化图像集具有相似的大小。

方法	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean
MixSample	40.92	43.18	40.73	40.55	49.84	46.18	50.42	57.22	42.92	39.39	54.26	45.83	43.67	38.74	44.90	45.25
MixAll	42.54	45.11	41.73	39.87	49.58	43.74	52.64	56.88	43.82	38.55	55.26	47.18	42.33	38.97	45.16	45.56
Personal	41.87	45.73	43.14	44.04	52.44	47.45	52.32	56.92	45.61	42.67	54.94	48.38	44.24	41.67	45.98	47.16

中学习的价值。

3.4.8 在道路场景数据上的实验

这一节中我们在常见的道路数据集上进行无监督语义分割域适应训练，测试我们的图像间语义辅助模块在这些道路场景数据上的效果如何。现有的一些无监督语义分割域适应方法一般都在 GTA5[54] 和 Cityscapes[53] 之间进行域适应训练，这里我们也遵循这样的设置。GTA5 包含 24966 张分辨率为 1914×1024 的虚拟图片，Cityscapes 则是实拍的街景图片，包含 2975 张训练图片和 500 张测试图片，分辨率为 2048×1024。我们进行从虚拟的 GTA5 向真实的 Cityscapes 上迁移的训练，并在后者的测试集上测试效果。跟随之前的方法[56, 109]，我们这里评测 19 个类别上的 mIoU，同时分割网络也使用 ResNet-101 为基础网络的 DeepLab v2[51] 模型。训练和测试时的图片大小我们也设置成和 ADVENT[56] 一致，批的大小设置为 4. 本实验的主要目的是测试我们的上下文语义辅助模块能否在道路场景数据集分割上起到作用，因此我们仅进行第一步的域适应训练，忽略第二步的伪标签微调部分，实验结果报告在表3.7中。从结果中我们可以看到，我们的方法超过基线方法 ADVENT 1.6%，达到 45.4 的 mIoU。这表示即便是在道路场景的数据集上，我们的上下文语义模块仍然能捕捉到图片间的相关语义，并对图像分割起到明显作用。应该注意到，道路场景中的图片其实都是关于类似道路、车辆和行人等物体的。而我们的分组方法是基于图像级特征聚类得到的，可能并不能很好区分这些图片。相信更加灵活的聚类方法能够得到更加理想的效果。

表 3.7 在道路场景数据上的分割实验结果，本实验包含 GTA5→Cityscapes 的域适应训练。注意我们仅进行第一步的域适应训练，没有加上伪标签微调的步骤。表中标注ST表示增加了伪标签微调的方法的结果。表中黑色数字下标表示发表年份。

GTA5 ⟶ Cityscapes																				
方法	road	side.	buil.	wall	fence	pole	t-light	t-sign	vege.	terr.	sky	pers.	rider	car	truck	bus	train	motor	bike	mIoU
Source-Only	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
Fully-Supervised	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.1
AdaptSeg$_{18}$[58]	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
DCAN$_{18}$[110]	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
DISE$_{19}$[111]	91.5	47.5	82.5	31.3	25.6	33.0	33.7	25.8	82.7	28.8	82.7	62.4	30.8	85.2	27.7	34.5	6.4	25.2	24.4	45.4
CLAN$_{19}$[112]	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
ADVENT$_{19}$[56]	89.9	36.5	81.6	29.2	25.2	28.5	32.3	22.4	83.9	34.0	77.1	57.4	27.9	83.7	29.4	39.1	1.5	28.4	23.3	43.8
SSF-DAN$_{19}$[113]	90.3	38.9	81.7	24.8	22.9	30.5	37.0	21.2	84.8	38.8	76.9	58.8	30.7	85.7	30.6	38.1	5.9	28.3	36.9	45.4
SIBAN$_{19}$[114]	88.5	35.4	79.5	26.3	24.3	28.5	32.5	18.3	81.2	40.0	76.5	58.1	25.8	82.6	30.3	34.4	3.4	21.6	21.5	42.6
LTIR$_{20}$ w/o ST[115]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	44.6
LTIR$_{20}$[115]	92.9	55.0	85.3	34.2	31.1	34.9	40.7	34.0	85.2	40.1	87.1	61.0	31.1	82.5	32.3	42.9	0.3	36.4	46.1	50.2
FDA$_{20}$ w/o ST[106]	90.0	40.5	79.4	25.3	26.7	30.6	31.9	29.3	79.4	28.8	76.5	56.4	27.5	81.7	27.7	45.1	17.0	23.8	29.6	44.6
Ours w/o ST	89.2	41.5	83.3	33.3	15.1	34.6	42.9	29.0	85.9	38.3	79.9	65.8	28.9	85.8	40.4	46.7	0.0	22.1	0.0	45.4

第五节 本章小结

在本章中，我们研究了如何利用用户个性化这一弱监督信息来挖掘图像之间的关联性，并将这种关联性应用到语义分割的任务中。我们首次提出图像语义分割中的个性化问题，并收集了包含 15 个用户数据的大型个性化图像数据集 PIS。我们的数据集可以作为用户个性化图像分割问题的一个研究起点。用户个性化这一弱监督信息是我们的数据集的最大特点，在为我们研究视觉任务提供一个先验知识的同时，如何利用这一弱监督信号也是研究个性化图片语义分割的一个难点。用户的图片之间有相互关联性，但是并不是所有图片都高度相关。在进行语义分割这些细粒度视觉问题时，也不是所有的语义都能对像素类别的判别起到正面的互补作用。如何能够利用好用户个性化这一弱监督信号，同时又能排除不相关图片的语义干扰，这是一个值得继续研究的问题。未来我们将在更好的图像分组，以及更好的上下文语义辅助模块这两个方面继续探索。

第四章 总结展望

第一节 本文工作总结

从粗糙的弱监督信号中挖掘细粒度的视觉信息，并利用这种视觉信息的组合与推理完成更复杂的认知任务。这是一个非常有意义的研究方向。尽管目前各种不同任务上的细粒度标注的数据集层出不穷，然而这种数据集能包含的数据毕竟是有限的。如果通过弱监督技术的研究，使得机器能够从互联网的海量弱标注信息中挖掘细粒度的视觉信息，并在这些信息的基础上进行更加高层级的推理与复杂认知，将会极大地推动智能系统的通用性发展。本文在视频问答和个性化图像分割两个场景下研究如何挖掘弱监督数据中的物体级别和像素级别的视觉信息，并利用它来解决下游的任务。

在视频问答的场景中，本文通过将视频和语句分解为视频区域和物体词语，然后通过在统一的特征空间中学习区域和词语的对应关系来建立物体级别的视觉认知能力。基于这种物体级别的匹配关系的建立我们得以实现由问题中的物体词语引导的视频注意力机制，将网络的注意力集中到视频中的相关物体上，从而回答问题。同时，为了更精准地关注到对回答起关键作用的物体本文还设计了一个能够在视频的不同区域转移注意力的模块。公开数据集上的实验结果表明了本文的方法的优越性。相比于基于软注意力的方法，本文的基于视觉-语言配对的注意力机制在关注视频的不同区域的同时可以显式地输出它关注的物体词语，这种透明与可解释性为模型分析提供了基础。通过可视化网络的注意力转移过程，我们能够观察模型在解决问答这样的复杂问题时的推理过程，进而设计更加复杂、通用的认知系统。

在个性化图像分割的场景中，本文通过将个性化这一较难定义的弱监督信号具象化成用户图像之间的关联性，进而通过关联图片之间的语义互补性为图像分割提供额外的信息补充。为了促进个性化图像分割的研究，我们收集了一个包含 15 个用户的个性化图像的数据集。为了在利用用户图像之间的语义相关性的同时排除相关性较低的图像带来的干扰。本文提出在处理用户的图像时，首先将其聚类成若干个语义相似度高的分组。我们进一步设计了一个组上下文

语义辅助模块：在每一个分组内，我们提取图片的区域特征，在对具体像素进行类别判定时，组上下文模块会度量该像素和这些区域特征的相关性，并用相关性高的区域的语义来辅助像素类别的判定。在个性化数据集上的实验结果表明我们的组上下文模块能够有效地挖掘图像之间的关联性并应用到分割任务上。

第二节 未来工作展望

开放数据中弱监督信号的探索与利用是一个非常有挑战性的研究方向，从这些信号中挖掘基础的视觉信息的能力对于更通用的智能系统的发展不可或缺。本文在视频和图像两个场景下利用弱监督信号挖掘其中的物体和像素级别的视觉信息。尽管这些信息为下游的任务带来了帮助，但是本文的系统还是较为简单的设计，挖掘到的视觉信息也还较为粗糙。例如从视频中挖掘到的视频区域和物体词语的对应关系中，我们得到的只是网格状区域的视频内容和物体词语的对应关系。本文的系统暂时还不能得到更加精细的像素级别的物体区域，同时问题中的一些表达动作或者相互关系的词语我们还没有能够将其和视频中的时空区域对应起来。未来，我们将尝试更加细粒度的区域配对，同时考虑学习视频中的时空区域的组合和表达动作的词语的对应关系。个性化图像分割的场景中，我们虽然能够将个性化信息转换成图像之间关联，但是这种关联依然是较为模糊的基于特征向量相似度的特征增强。未来，我们希望能够探索用户图像之间高层级的物体关联性，比如实现图像间物体的追踪。总之，弱监督信息的探索与利用还有很多值得研究的具体问题，希望本文能够给读者一些关于这个方向的参考，推动弱标注的智能系统的开发。

参考文献

[1] EVERINGHAM M, VAN GOOL L, WILLIAMS C K, et al. The pascal visual object classes (voc) challenge. [J]. Int. J. Comput. Vis., 2010, 88 (2): 303–338.

[2] WANG J, JIANG H, YUAN Z, et al. Salient Object Detection: A Discriminative Regional Feature Integration Approach. [J]. International Journal of Computer Vision, 2017, 123 (2): 251–268. DOI: 10.1007/s11263-016-0977-3. ISSN: 1573-1405.

[3] IDREES H, ZAMIR A R, JIANG Y.-G, et al. The THUMOS challenge on action recognition for videos "in the wild". [J]. Computer Vision and Image Understanding, 2017, 155: 1–23.

[4] XU D, ZHU Y, CHOY C, et al. Scene Graph Generation by Iterative Message Passing. [C] // Computer Vision and Pattern Recognition (CVPR). 2017.

[5] SOOMRO K, ZAMIR A R, SHAH M, et al. UCF101: A dataset of 101 human actions classes from videos in the wild. [J]. CoRR, 2012: 2012.

[6] GOYAL R, KAHOU S E, MICHALSKI V, et al. The" Something Something" Video Database for Learning and Evaluating Visual Common Sense. [C] // Int. Conf. Comput. Vis. Vol. 1. 2017: 3.

[7] JANG Y, SONG Y, YU Y, et al. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017: 2758–2766.

[8] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. [J/OL]. Journal of Machine Learning Research, 2014, 15 (56): 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html.

[9] ZHOU Z.-H. A brief introduction to weakly supervised learning. [J]. National science review, 2018, 5 (1): 44–53.

[10] SETTLES B. Active learning literature survey. Computer sciences technical report 1648. [J]. University of Wisconsin-Madison, 2009.

[11] CHAPELLE O, SCHOLKOPF B, ZIEN A. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. [J]. IEEE Transactions on Neural Networks, 2009, 20 (3): 542–542.

[12] ZHU X J. Semi-supervised learning literature survey. [J]. 2005.

[13] ZHOU Z.-H, LI M. Semi-supervised learning by disagreement. [J]. Knowledge and Information Systems, 2010, 24 (3): 415–439.

[14] WEI Y, XIAO H, SHI H, et al. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7268–7277.

[15] YAN P, LI G, XIE Y, et al. Semi-supervised video salient object detection using pseudo-labels. [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7284–7293.

[16] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks. [J]. ArXiv preprint arXiv:1609.02907, 2016.

[17] NAUATA N, SMITH J, MORI G. Hierarchical label inference for video classification. [J]. ArXiv preprint arXiv:1706.05028, 2017.

[18] JIANG P.-T, HOU Q, CAO Y, et al. Integral object mining via online attention accumulation. [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 2070–2079.

[19] WANG Y, ZHANG J, KAN M, et al. Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.

[20] WEI Y, LIANG X, CHEN Y, et al. STC: A Simple to Complex Framework for Weakly-Supervised Semantic Segmentation. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (11): 2314–2320. DOI: 10.1109/TPAMI.2016.2636150.

[21] FRÉNAY B, VERLEYSEN M. Classification in the presence of label noise: a survey. [J]. IEEE transactions on neural networks and learning systems, 2013, 25 (5): 845–869.

[22] ANGLUIN D, LAIRD P. Learning from noisy examples. [J]. Machine Learning, 1988, 2 (4): 343–370.

[23] BLUM A, KALAI A, WASSERMAN H. Noise-tolerant learning, the parity problem, and the statistical query model. [J]. Journal of the ACM (JACM), 2003, 50 (4): 506–519.

[24] HOU Q, HAN L, LIU J, et al. Autonomous Learning of Semantic Segmentation from Internet Images. [J]. SCIENTIA SINICA Informationis,

[25] DALAL N, TRIGGS B. Histograms of Oriented Gradients for Human Detection. [C/OL] // IEEE Conf. Comput. Vis. Pattern Recog. Washington, DC, USA: IEEE Computer Society, 2005: 886–893. ISBN: 0-7695-2372-2. http://dx.doi.org/10.1109/CVPR.2005.177. DOI: 10.1109/CVPR.2005.177.

[26] WANG H, KLÄSER A, SCHMID C, et al. Action recognition by dense trajectories. [C] //. 2011.

[27] KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset. [J]. ArXiv preprint arXiv:1705.06950, 2017.

[28] ZHANG D, DAI X, WANG X, et al. S3D: single shot multi-span detector via fully 3D convolutional networks. [J]. ArXiv preprint arXiv:1807.08069, 2018.

[29] GU C, SUN C, ROSS D A, et al. AVA: A video dataset of spatio-temporally localized atomic visual actions. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2018: 6047–6056.

[30] ZHOU B, ANDONIAN A, OLIVA A, et al. Temporal relational reasoning in videos. [C] // Eur. Conf. Comput. Vis. 2018: 803–818.

[31] BARADEL F, NEVEROVA N, WOLF C, et al. Object level visual reasoning in videos. [C] // Eur. Conf. Comput. Vis. 2018: 105–121.

[32] WANG X, GUPTA A. Videos as space-time region graphs. [C] // Eur. Conf. Comput. Vis. 2018: 399–417.

[33] ZHOU L, KALANTIDIS Y, CHEN X, et al. Grounded video description. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 6578–6587.

[34] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. [C] //. 2015: 91–99.

[35] LIN T.-Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context. [C] // European conference on computer vision. Springer. 2014: 740–755.

[36] KRISHNA R, ZHU Y, GROTH O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. [J]. Int. J. Comput. Vis., 2017, 123 (1): 32–73.

[37] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos. [J]. ArXiv preprint arXiv:1406.2199, 2014.

[38] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks. [C] // Proceedings of the IEEE international conference on computer vision. 2015: 4489–4497.

[39] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset. [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299–6308.

[40] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need. [J]. ArXiv preprint arXiv:1706.03762, 2017.

[41] BERTASIUS G, WANG H, TORRESANI L. Is Space-Time Attention All You Need for Video Understanding? [J]. ArXiv preprint arXiv:2102.05095, 2021.

[42] FEICHTENHOFER C, FAN H, MALIK J, et al. Slowfast networks for video recognition. [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6202–6211.

[43] DEVLIN J, CHANG M.-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. [J]. ArXiv preprint arXiv:1810.04805, 2018.

[44] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging. [J]. ArXiv preprint arXiv:1508.01991, 2015.

[45] NA S, LEE S, KIM J, et al. A read-write memory network for movie story understanding. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017: 677–685.

[46] GAO J, GE R, CHEN K, et al. Motion-appearance co-memory networks for video question answering. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2018: 6576–6585.

[47] MUN J, HONGSUCK SEO P, JUNG I, et al. Marioqa: Answering questions by watching gameplay videos. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017: 2867–2875.

[48] XU D, ZHAO Z, XIAO J, et al. Video question answering via gradually refined attention over appearance and motion. [C] //. ACM. 2017: 1645–1653.

[49] YU Y, KO H, CHOI J, et al. End-to-end concept word detection for video captioning, retrieval, and question answering. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017: 3165–3173.

[50] ZHAO H, SHI J, QI X, et al. Pyramid Scene Parsing Network. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017.

[51] CHEN L.-C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2017, 40 (4): 834–848.

[52] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn. [C] // Int. Conf. Comput. Vis. 2017: 2961–2969.

[53] CORDTS M, OMRAN M, RAMOS S, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2016.

[54] RICHTER S R, VINEET V, ROTH S, et al. Playing for Data: Ground Truth from Computer Games. [C] // Eur. Conf. Comput. Vis. 2016: 102–118.

[55] ROS G, SELLART L, MATERZYNSKA J, et al. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2016: 3234–3243.

[56] VU T.-H, JAIN H, BUCHER M, et al. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 2517–2526.

[57] CHEN M, XUE H, CAI D. Domain adaptation for semantic segmentation with maximum squares loss. [C] // Int. Conf. Comput. Vis. 2019: 2090–2099.

[58] TSAI Y.-H, HUNG W.-C, SCHULTER S, et al. Learning to adapt structured output space for semantic segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2018: 7472–7481.

[59] HOFFMAN J, TZENG E, PARK T, et al. Cycada: Cycle-consistent adversarial domain adaptation. [C] // International Conference on Machine Learning. 2018: 1989–1998.

[60] PAN F, SHIN I, RAMEAU F, et al. Unsupervised Intra-domain Adaptation for Semantic Segmentation through Self-Supervision. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020: 3764–3773.

[61] ZHANG Y, DAVID P, FOROOSH H, et al. A curriculum domain adaptation approach to the semantic segmentation of urban scenes. [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2019.

[62] ZHANG Y, DAVID P, GONG B. Curriculum domain adaptation for semantic segmentation of urban scenes. [C] // Int. Conf. Comput. Vis. 2017: 2020–2030.

[63] ZOU Y, YU Z, VIJAYA KUMAR B, et al. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. [C] // Eur. Conf. Comput. Vis. 2018: 289–305.

[64] LIAN Q, LV F, DUAN L, et al. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. [C] // Int. Conf. Comput. Vis. 2019: 6758–6767.

[65] MIRKIN S, NOWSON S, BRUN C, et al. Motivating Personality-aware Machine Translation. [C] // Conf. Empir. Meth. Natur. Lang. Process. 2015: 1102–1108.

[66] HORIGUCHI S, AMANO S, OGAWA M, et al. Personalized Classifier for Food Image Recognition. [J]. IEEE Trans. Multimedia, 2018, 20 (10): 2836–2848.

[67] PARK C C, KIM B, KIM G. Attend to You: Personalized Image Captioning with Context Sequence Memory Networks. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017: 6432–6440.

[68] KIM H.-U, KOH Y J, KIM C.-S. PieNet: Personalized Image Enhancement. [C] // Eur. Conf. Comput. Vis. 2020.

[69] HSU K.-J, LIN Y.-Y, CHUANG Y.-Y. DeepCO3: Deep Instance Co-Segmentation by Co-Peak Search and Co-Saliency Detection. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019.

[70] ZHU C, XU K, CHAUDHURI S, et al. AdaCoSeg: Adaptive Shape Co-Segmentation With Group Consistency Loss. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020.

[71] LI B, SUN Z, LI Q, et al. Group-Wise Deep Object Co-Segmentation With Co-Attention Recurrent Neural Network. [C] // Int. Conf. Comput. Vis. 2019: 8519–8528.

[72] HAN J, QUAN R, ZHANG D, et al. Robust object co-segmentation using background prior. [J]. IEEE Trans. Image Process., 2017, 27 (4): 1639–1651.

[73] FAN D.-P, LIN Z, JI G.-P, et al. Taking a Deeper Look at Co-Salient Object Detection. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020.

[74] ZHANG Z, JIN W, XU J, et al. Gradient-Induced Co-Saliency Detection. [C] // Eur. Conf. Comput. Vis. 2020.

[75] DENG J, DONG W, SOCHER R, et al. ImageNet: A Large-Scale Hierarchical Image Database. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2009.

[76] SIMONYAN K, ZISSERMAN A. Two-Stream Convolutional Networks for Action Recognition in Videos. [G/OL] // Adv. Neural Inform. Process. Syst. Ed. by GHAHRAMANI Z, WELLING M, CORTES C, et al. Curran Associates, Inc., 2014: 568–576. http://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos.pdf.

[77] FANG H, GUPTA S, IANDOLA F, et al. From captions to visual concepts and back. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2015: 1473–1482.

[78] ANTOL S, AGRAWAL A, LU J, et al. VQA: Visual Question Answering. [C] // Int. Conf. Comput. Vis. 2015.

[79] FROME A, CORRADO G S, SHLENS J, et al. DeViSE: A Deep Visual-Semantic Embedding Model. [G/OL] // Adv. Neural Inform. Process. Syst. Ed. by BURGES C J C, BOTTOU L, WELLING M, et al. Curran Associates, Inc., 2013: 2121–2129. http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model.pdf.

[80] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator. [C/OL] // IEEE Conf. Comput. Vis. Pattern Recog. 2015. http://arxiv.org/abs/1411.4555.

[81] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2018: 7794–7803.

[82] LIU M, CHEN X, ZHANG Y, et al. Paying More Attention to Motion: Attention Distillation for Learning Video Representations. [J]. ArXiv preprint arXiv:1904.03249, 2019.

[83] HOCHREITER S, SCHMIDHUBER J. Long short-term memory. [J]. Neural computation, 1997, 9 (8): 1735–1780.

[84] WU H, MAO J, ZHANG Y, et al. Unified Visual-Semantic Embeddings: Bridging Vision and Language with Structured Meaning Representations. [J]. ArXiv preprint arXiv:1904.05521, 2019.

[85] FAGHRI F, FLEET D J, KIROS J R, et al. Vse++: Improving visual-semantic embeddings with hard negatives. [J]. ArXiv preprint arXiv:1707.05612, 2017.

[86] SCHUSTER S, KRISHNA R, CHANG A, et al. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. [C] //. 2015: 70–80.

[87] HUDSON D A, MANNING C D. Compositional attention networks for machine reasoning. [J]. ArXiv preprint arXiv:1803.03067, 2018.

[88] PENNINGTON J, SOCHER R, MANNING C. Glove: Global vectors for word representation. [C] // Conf. Empir. Meth. Natur. Lang. Process. 2014: 1532–1543.

[89] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2016: 770–778.

[90] KINGMA D P, BA J. Adam: A method for stochastic optimization. [J]. ArXiv preprint arXiv:1412.6980, 2014.

[91] PASZKE A, GROSS S, CHINTALA S, et al. Automatic Differentiation in PyTorch. [C] //. 2017.

[92] REN M, KIROS R, ZEMEL R. Exploring models and data for image question answering. [C] //. 2015: 2953–2961.

[93] FUKUI A, PARK D H, YANG D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. [J]. ArXiv preprint arXiv:1606.01847, 2016.

[94] YU G, YUAN J. Fast action proposals for human action detection and search. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2015: 1302–1311.

[95] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation. [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431–3440.

[96] NOH H, HONG S, HAN B. Learning deconvolution network for semantic segmentation. [C] // Proceedings of the IEEE international conference on computer vision. 2015: 1520–1528.

[97] LIN G, MILAN A, SHEN C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1925–1934.

[98] CHEN L.-C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. [C] // Proceedings of the European conference on computer vision (ECCV). 2018: 801–818.

[99] ZHOU B, ZHAO H, PUIG X, et al. Scene Parsing through ADE20K Dataset. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017.

[100] TSAI Y.-H, SOHN K, SCHULTER S, et al. Domain adaptation for structured output via discriminative patch representations. [C] // Int. Conf. Comput. Vis. 2019: 1456–1465.

[101] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2009: 248–255.

[102] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2009: 248–255.

[103] YUAN Y, CHEN X, WANG J. Object-contextual representations for semantic segmentation. [J]. ArXiv preprint arXiv:1909.11065, 2019.

[104] ZHENG Z, YANG Y. Unsupervised Scene Adaptation with Memory Regularization in vivo. [C] // International Joint Conference on Artificial Intelligence (IJCAI). 2019.

[105] PASZKE A, GROSS S, MASSA F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. [C] // Adv. Neural Inform. Process. Syst. 2019: 8024–8035.

[106] YANG Y, SOATTO S. Fda: Fourier domain adaptation for semantic segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020: 4085–4095.

[107] ZHENG Z, YANG Y. Rectifying Pseudo Label Learning via Uncertainty Estimation for Domain Adaptive Semantic Segmentation. [J]. Int. J. Comput. Vis., 2020.

[108] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. [C] // Int. Conf. Learn. Represent. 2015.

[109] KANG G, WEI Y, YANG Y, et al. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. [C] // Adv. Neural Inform. Process. Syst. 2020.

[110] WU Z, HAN X, LIN Y.-L, et al. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. [C] // Eur. Conf. Comput. Vis. 2018: 518–534.

[111] CHANG W.-L, WANG H.-P, PENG W.-H, et al. All about structure: Adapting structural information across domains for boosting semantic segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 1900–1909.

[112] LUO Y, ZHENG L, GUAN T, et al. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 2507–2516.

[113] DU L, TAN J, YANG H, et al. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. [C] // Int. Conf. Comput. Vis. 2019: 982–991.

[114] LUO Y, LIU P, GUAN T, et al. Significance-aware information bottleneck for domain adaptive semantic segmentation. [C] // Int. Conf. Comput. Vis. 2019: 6778–6787.

[115] KIM M, BYUN H. Learning texture invariant representation for domain adaptation of semantic segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020: 12975–12984.

致谢

逝者如斯夫，不舍昼夜。三年的硕士生涯即将结束，回顾这段时间的学习与生活，感慨万千。曾经的各种迷茫，无助，希望与失望仿佛历历在目。在此，向帮助过我的师长，鼓励我们朋友与亲人致以最诚挚的感谢。

首先要感谢的是我的导师程明明教授，程老师在科研上对我悉心教导，在生活上对我关心及时帮助。除此之外，程老师的一丝不苟的学术作风更是我现在以级今后的学习典范。师承程老师，我感到非常幸运，他毫无保留地支持我的研究，为我提供力所能及的帮助。从研究想法到实验过程，他都能耐心地帮我分析、推敲，我地所有研究工作都离不开他的汗水与辛劳。同时，我要感谢的的联合导师王恺副教授对我的指导和帮助，让我更加明确硕士生涯的努力方向，以实事求是的研究态度，做有价值的研究。感谢南开大学的所有老师们。特别感谢南开大学媒体计算实验室的任博老师，卢少平老师以及计算机学院的杨巨峰老师，他们是我学术研究上的榜样，让我树立了精益求精的学术态度与精神。

感谢媒体计算实验室的所有同学。特别感谢各位师兄对我的指导与帮助，姜鹏涛、侯淇彬、刘姜江等各位师兄对我的科研生活与职业规划提供了帮助；特别感谢同年的谭永强、高尚华、李炫毅、陈林卓、林铮、许刚、吴宇寰等同学，他们为我的学术与生活提供了不间断的帮助与理解；特别感谢张长彬、张钊等学弟，他们为我的论文与研究提供了许多帮助和意见。向在媒体计算实验室中共同奋斗与成长的各位同学表达深深的谢意和不舍。三年时间，我们一起分享过喜悦与压力，共同经历了挫折与成长，大家互相帮助，共同学习也共同放松的时间多么值得回味。

时逢新冠疫情期间，每个人都是防疫的一部分。而特别需要感谢地是抗疫战役中的医护人员，公安民警，军人，是他们的坚强与付出让我们能够获得一个平稳安全的生活环境。

最后，我要感谢我的家人，二十多年的人生中，是父母与亲人不间断的支持、鞭策与鼓励让我能够安心求学，努力研究。感谢我的女朋友赵思浓，她的陪伴、认可与支持为我提供了源源不断的前进动力。

个人简历

张宇，1996 年 7 月出生于安徽省合肥市。

2011 年 9 月至 2014 年 6 月就读于安徽省庐江中学。

2014 年 9 月至 2018 年 6 月就读于西北工业大学自动化专业，并获得工学学士学位。

于 2018 年 9 月至 2021 年 6 月在南开大学就读于计算机科学与技术专业研究生。

于 2020 年 6 月至 9 月在杭州阿里巴巴图灵安全实验室实习。

参与的项目

华为手机中基于显著性检测的手机自拍人像分割技术。

华为手机中个性化用户图像的分割技术。

专利

基于图像间语义辅助的个性化图像分割方法及系统。申请公布号：CN 112381831 A。程明明，张宇，姜鹏涛

研究生期间主要研究工作:

- Yu Zhang, Xuan-Yi Li, Ming-Ming Cheng, Bo Ren. Tell Me Where to Look: Object Guided Attention Machanism for Video Question Answering, In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), submitted.
- Yu Zhang, Chang-Bin Zhang, Peng-Tao Jiang, Feng Mao, Ming-Ming Cheng. Personalized Image Semantic Segmentation, In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021, submitted.