智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于原型引导与自适应特征融合的域适应语义分割

杨宇宇, 杨霄, 潘在宇, 王军

引用本文:
杨宇宇, 杨霄, 潘在宇, 等. 基于原型引导与自适应特征融合的域适应语义分割[J]. 智能系统学报, 2025, 20(1): 150-161.
YANG Yuyu, YANG Xiao, PAN Zaiyu, et al. Domain adaptive semantic segmentation based on prototype-guided and adaptive feature fusion[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(1): 150-161.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202403010

DOI: 10.11992/tis.202403010
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20250103.0855.002

基于原型引导与自适应特征融合的域适应语义分割

杨宇宇, 杨霄, 潘在宇, 王军
(中国矿业大学 信息与控制工程学院, 江苏 徐州 221116)

摘　要: 无监督域自适应技术对于减少计算机视觉任务中的数据标注工作量具有重要意义, 尤其在像素级的语义分割中。然而, 目标域的特征分布离散和类别不平衡问题, 如模糊的类边界和某些类别的样本过少, 对无监督域自适应技术构成了挑战。针对上述挑战, 本文提出了一种原型引导的自适应特征融合模型。其中, 通过引入原型引导的双重注意力网络融合空间和通道注意力特征, 增强类内紧凑性。此外, 本文提出自适应特征融合模块, 灵活调整各特征的重要性, 使网络能够在不同的空间位置和通道上捕捉到更加具有类别区分性的特征, 进一步提升语义分割性能。在两个具有挑战性的合成−真实基准 GTA5-to-Cityscape 和 SYNTHIA-to-Cityscape 上的实验结果证明了本文方法的有效性, 展现出模型对复杂场景和不平衡数据的处理应对能力。

关键词: 深度学习; 无监督学习; 域适应; 语义分割; 注意力机制; 自训练学习; 自适应; 迁移学习; 原型引导
中图分类号: TP301　　**文献标志码**: A　　**文章编号**: 1673−4785(2025)01−0150−12

中文引用格式: 杨宇宇, 杨霄, 潘在宇, 等. 基于原型引导与自适应特征融合的域适应语义分割 [J]. 智能系统学报, 2025, 20(1): 150–161.
英文引用格式: YANG Yuyu, YANG Xiao, PAN Zaiyu, et al. Domain adaptive semantic segmentation based on prototype-guided and adaptive feature fusion[J]. CAAI transactions on intelligent systems, 2025, 20(1): 150–161.

Domain adaptive semantic segmentation based on prototype-guided and adaptive feature fusion

YANG Yuyu, YANG Xiao, PAN Zaiyu, WANG Jun
(School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: Unsupervised domain adaptation techniques are of significant importance to reducing the data annotation workload for computer vision tasks, particularly in pixel-level semantic segmentation. However, challenges such as the dispersed feature distribution and class imbalance in the target domain, such as blurred class boundaries and insufficient samples for certain categories, pose challenges to this technology. To address these challenges, this paper proposes a prototype-guided adaptive feature fusion model. It incorporates a dual attention network guided by prototypes to fuse spatial and channel attention features, enhancing class-wise compactness. Furthermore, this paper introduces an adaptive feature fusion module that flexibly adjusts the importance of each feature, enabling the network to capture more class-discriminative features across different spatial locations and channels, thereby further enhancing the performance of semantic segmentation. Experimental results on two challenging synthetic-to-real benchmarks of GTA5-to-Cityscape and SYNTHIA-to-Cityscape demonstrate the effectiveness of our method, showcasing the model's capability to handle complex scenes and imbalanced data.

Keywords: deep learning; unsupervised learning; domain adaptation; semantic segmentation; attention mechanism; self-training learning; self-adaptive; transfer learning; prototype guidance

在计算机视觉领域, 语义分割是一项关键任务, 长期以来在自动驾驶、医疗影像分析、遥感图像处理等应用中备受研究关注[1-2]。与简单的图

收稿日期: 2024−03−05.　**网络出版日期**: 2025−01−03.
基金项目: 新一代人工智能国家科技重大专项 (2020AAA0107300); 中央高校基本科研业务费专项 (2023QN1077).
通信作者: 王军. E-mail: jrobot@126.com.

像分类或目标检测不同, 语义分割旨在为图像中的每个像素分配语义标签, 为图像内容提供更深层次的理解。

随着深度学习和卷积神经网络(convolutional neural networks, CNN)的兴起, 语义分割取得了显著进展, 尤其是全卷积神经网络[3](fully convolutional networks, FCN)的引入, 使神经网络能够直接在像素级别进行训练。然而, 深度学习模型不仅需要大量标记数据来进行有效训练, 而且在处理真实场景时可能面临数据分布差异的挑战, 尤其是在合成数据(源域)到真实世界数据(目标域)的转换过程中。例如, 在自然影像分析中, 环境因素、成像条件和设备差异都可能会导致数据集的风格差异。

在上述背景下, 域适应技术[4-6]应运而生, 成为解决跨领域应用性能下降的有效策略。该技术的核心思想是通过使模型从一个领域(源域)学习并适应到另一个或多个领域(目标域), 从而显著提高模型的泛化能力和实用性。尤其是在目标域缺乏标记数据的情况下, 域适应技术为模型在未知环境中的应用提供了强大的支持。域适应技术在语义分割领域的研究主要集中在两个方面: 一是如何有效地对源域和目标域的数据分布进行对齐, 以减小它们之间的领域差异; 二是如何保持模型在源域性能的同时, 提高其在目标域的泛化能力。在这一基础上, 无监督域适应语义分割的研究逐渐演进为基于对抗学习的方法[7-10]、风格迁移技术[11-12]以及自训练学习[13-14]的策略。这些方向不仅拓展了域适应的应用范围, 也推动了语义分割领域的深入研究。

特别地, 自训练学习的域适应语义分割方法是一种简单且具有竞争力的方法。这些方法的关键原理是为目标图像生成一组伪标签, 作为真实标签的近似值, 然后利用伪标签的目标域数据来更新分割模型。这一方法特别适用于在目标域缺少标注数据的情况。在自训练学习的初始阶段, 尽管利用源域的大量标注数据训练的深度学习模型可以在源域上展现出优异的性能, 但由于源域与目标域之间存在的显著差异, 这些模型被直接应用于目标域时往往性能会下降。主要是因为深度学习模型在特征空间中受到域偏差的显著影响, 导致在目标域中生成的特征分布离散, 且类特征边界模糊不清。在这种情况下, 模型可能会对不确定的像素进行错误分类。此外, 由于自训练策略鼓励网络输出接近热编码的峰值分布, 可能会引发模型对频繁出现的类别过度拟合, 从而忽视较少见的类别。为解决上述问题, Zhang 等[15]提出了一种原型伪标签去噪的域适应策略, 其核心在于使用类别特征质心(原型)与各像素点特征的距离来估算伪标签的可信度, 并在训练过程中进行动态校正, 同时也对目标域中的特征分布进行紧凑化处理。然而, 此方法可能忽略了像素之间的上下文信息, 特别是对于那些大小、形状有较大变化的类别至关重要, 尤其是在解析复杂场景时。为此, 本文提出一种改进方案, 同时在像素位置层面和特征通道层面加强与原型的关联。这种多维度的联系增强了特征的判别力和同类特征的类内一致性。进一步地, 为了最大化的利用多维度特征, 本文还提出了一种自适应特征融合模块, 从而进一步提高域适应性能。本文主要贡献为

1)提出一种原型引导的自适应特征融合模型(prototype-guided adaptive feature fusion module, PG-AFFM), 通过融合多层次的目标特征, 实现类内的紧凑性和类间的可分性。

2)提出原型引导的双重注意力网络, 创新地整合原型伪标签去噪技术。其中, 原型引导的位置注意力模块(prototype-guided position attention module, PG-PAM)在空间位置上进行去噪和特征增强, 原型引导的通道注意力模块(prototype-guided channel attention module, PG-CAM)在通道层面强化特征表示能力。

3)提出自适应特征融合模块, 能够高效地结合位置注意特征和通道注意特征, 使网络获得更多维的特征且更易捕捉到罕见类别特征, 从而增强类间的可分性, 以便更准确地区分不同类别。

1　相关工作

1.1　语义分割

在深度学习方法流行之前, 语义分割领域主要采用传统机器学习分类器的方法, 如支持向量机(support vector machine, SVM)和随机森林。然而, 深度学习的兴起极大地提升了语义分割算法的精度。全卷积网络(FCN)[3]的出现, 实现了对整个图像的端到端像素级预测, 无需将图像分割成小块或使用滑动窗口的方式。FCN 通过将传统的 CNN 中的全连接层转换为卷积层, 实现了对图像的任意尺寸输入和对应的密集预测输出, 这一创新极大地提高了语义分割的效率和精度。此后, 语义分割领域的研究迅速发展, 涌现出了许多创新的方法和模型。如 PSPNet(pyramid scene parsing network)[16]引入了金字塔池化模块, 有效

地捕获了不同尺度的上下文信息，DeepLab 系列[17-18]通过空洞卷积扩大感受野，并引入了条件随机场来优化分割边缘。此外，随着 Transformer 系列[19-21]在自然语言处理（natural language processing, NLP）领域的成功，其也被引入到图像语义分割中，通过全局自注意力机制捕捉长距离依赖，进一步提升了模型对复杂场景的理解能力。

1.2　无监督域适应

无监督域适应，旨在利用源域的大量标注数据来改善模型在目标域上的泛化能力，其中目标域仅含未标注数据。这一技术面临的主要挑战在于源域和目标域之间由于视觉风格、环境条件或数据采集设备的不同存在的数据分布差异。这些差异会导致在源域上训练的模型在目标域上性能下降，因为它们未能充分适应目标域的偏移特征。事实上，它们被认为在某种程度上是相互关联的，且它们的相关性越高，数据处理任务就越容易，从而可以在测试数据上获得较好的性能，但即使只有很小的视觉域偏移存在，也会导致较差的性能。因此，研究者们提出了众多基于距离的方法，包括最大平均差异[22]（maximum mean discrepancy, MMD）使源域和目标域之间的分布距离最小化。同时，随着生成对抗网络[23]（generative adversarial networks, GAN）的发展，对抗学习的方法开始流行，用于对齐源域和目标域之间的边缘或条件特征分布。

1.3　域适应语义分割

在语义分割任务中，模型的目标是将图像中的每个像素分配到其相应的语义类别，例如道路、汽车、行人等。然而，由于源域和目标域之间可能存在不同的数据分布，直接将在源域上训练的语义分割模型应用于目标域时，往往会导致性能下降。

为了克服这一挑战，Hoffman 等[24]在无监督语义分割领域的开创性工作为后续研究奠定了基础。他们提出的方法结合了全局和类别层面的域适应技术，利用对抗学习在分割网络上实现全局的域适应。继而，基于对抗学习的域自适应方法被广泛研究，核心是通过对抗性网络减少源域与目标域的特征差异，实现不同层面的域对齐，尤其是在图像级、特征级和像素级对齐不同的领域。图像级自适应（如 CycleGAN[25]）通过风格转换减少源目标域视觉差异，特征级对齐（如 AdaptSegNet[26]）使语义分割网络能探寻源域和目标域之间的共享特征，像素级的对抗方法[27]，采用熵最小化技术与对抗性训练相结合的方法来实现像素级域适应。

随着技术的不断发展，自训练学习（self-supervised learning, SSL）在域适应语义分割中显示出巨大的潜力。例如，Jiang 等[28]提出了一种原型对比自适应的方法（prototypical contrast adaptation, ProCA），通过自训练学习强调类内紧凑和类间分离。Hoyer 等[29]提出了一个掩码图像一致性模块，通过自训练学习目标域的空间上下文关系作为鲁棒视觉识别的额外线索来增强无监督域适应语义分割。最近，一种统一的像素和分片自训练学习框架[30]提出利用域内图像的固有结构，鼓励学习具有类内紧凑性和类间可分离性的区分像素特征，以及激励针对不同上下文或波动的相同区域的鲁棒特征学习。

2　原型引导的自适应特征融合模型

2.1　模型结构概览

本文提出了一种原型引导的自适应特征融合模型（PG-AFFM），其包含 3 个核心组件，获取深度特征和浅层特征的特征提取器、原型引导的双重注意力网络（PG-DAN）和自特征适应融合模块，如图 1 所示。

2.2　原型引导的双重注意力网络

为了避免原型对目标特征的片面引导并且有效加强目标域各类特征簇的紧凑性，为深度模型提供更准确的目标特征表征，本文提出了原型引导的双重注意力网络（PG-DAN），该网络结构创新性地结合了原型伪标签去噪技术和双重注意力机制，以更精确地捕捉目标域的特征分布。原型引导的通道注意力模块在通道层面上强化了特征表达。具体操作为：

首先，将源域预训练好的语义分割模型应用于目标域，对目标图像 X_t 的预测，得到伪标签 Y_t。根据伪标签计算各类特征的均值向量，定义这些向量为原型向量。这些原型向量是从目标域数据中提取的代表性样本，能够有效捕捉数据的关键特征，并被用来进一步引导自适应特征融合模型的训练。原型向量的初始化公式为

$$\eta^k=\frac{\sum\limits_{X_t\in D_T}\sum\limits_i f_{t,d}(X_t)^i\times\prod\left(Y_t^{(i,k)}==1\right)}{\sum\limits_{X_t\in D_T}\sum\limits_i\prod\left(Y_t^{(i,k)}==1\right)}$$

式中：D_T 是目标域数据集；$f_{t,d}(X_t)^i$ 表示 X_t 在像素点 i 处的深层特征；$\prod$ 为指示函数，当伪标签 Y_t 在像素点 i 处属于 k 类的热编码值为 1 时，$\prod(\cdot)=1$。

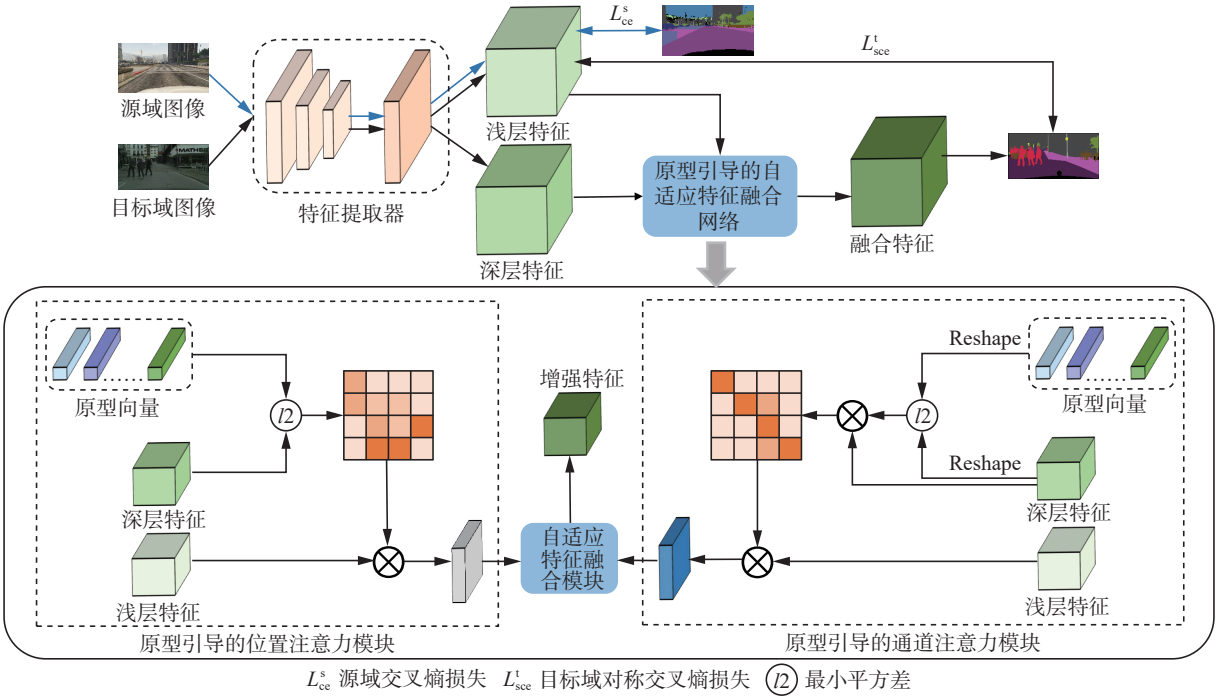


图 1　原型引导的自适应特征融合模型

Fig. 1　Prototype-guided adaptive feature fusion network

为了防范初始伪标签预测误差对原型向量精确度的潜在影响，本文在目标域参与训练时，通过原型引导的自适应特征融合模型持续生成伪标签，并据此动态更新原型向量。具体将原型估计视为特征提取器，计算得出小批量簇质心的移动平均值，从而能够精准地追踪原型的细微变化，确保原型向量的准确性与实时性：

$$\eta^k \leftarrow \lambda\,\eta^k + (1-\lambda)\,\eta^{k'}$$

式中：$\eta^{k'}$ 为特征提取器在当前训练批内计算出的 k 类的平均特征；λ 为动量系数，设置为 0.999 9。

获取类别原型后，PG-CAM 需要评估每个通道特征与类别原型之间的距离，图 2 给出了获取通道权重的示意。

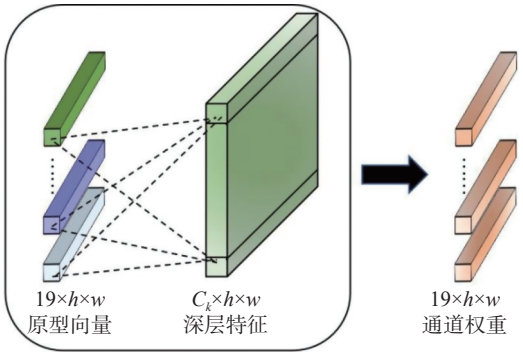


图 2　获取通道权重示意

Fig. 2　Schematic diagram of obtaining channel weights

每个通道特征与类别原型之间的距离转化为通道关联权重，通道与原型的相似度越高，权重越大。具体计算公式为

$$\omega_{\mathrm{c}}^{(i,k)} = \frac{\exp\left(-\| f_{\mathrm{t,d}}(X_{\mathrm{t}})^i - \hat{\eta}^k \| / \tau\right)}{\sum_{k'} \exp\left(-\| f_{\mathrm{t,d}}(X_{\mathrm{t}})^i - \hat{\eta}^{k'} \| / \tau\right)} \tag{1}$$

式中：τ 是 softmax 温度经验值，设置为 $\tau = 1$；$\hat{\eta}^k$ 是原型通过卷积操作为与深层特征同样的通道数的扩张原型：

$$\hat{\eta}^k = \mathrm{conv}1\times 1(\eta^k)$$

式中 conv 是卷积操作。接下来，使用矩阵乘法将得到的通道与原型之间的关联权重映射到深层特征上，通过 softmax 函数激活以确保每个通道的权重和为 1，关联权重越高的通道将得到与原型之间更高的通道注意力权重：

$$\hat{\omega}_{\mathrm{c}}^{(i,k)} = \mathrm{softmax}\left(\omega_{\mathrm{c}}^{(i,k)} \otimes f_{\mathrm{t,d}}(X_{\mathrm{t}})\right)$$

最后，使用通道注意力权重加权浅层特征，获得最终的通道注意力特征：

$$\hat{f}_{\mathrm{t,c}}^{(i,k)} = \hat{\omega}_{\mathrm{c}}^{(i,k)} f_{\mathrm{t,s}}^{(i,k)}$$

式中 $f_{\mathrm{t,s}}^{(i,k)}$ 是特征提取器提取的浅层特征。

原型引导的位置注意力模块（PG-PAM）与原型伪标签去噪策略一致，通过利用类别特征质心（即原型）来指导空间注意力的分配，从而在空间维度上进行有效的去噪和特征增强。该模块提高了网络对空间位置的敏感度，使其能够强调图像中与特定类别原型密切相关的像素，从而获得位置注意力特征，具体计算公式为

$$\hat{f}_{\mathrm{t,p}}^{(i,k)} = \omega_{\mathrm{p}}^{(i,k)} f_{\mathrm{t,s}}^{(i,k)}$$

式中 $\omega_{\mathrm{p}}^{(i,k)}$ 是通过式（1）计算各像素点与各类别原

型之间的关联权重(即位置注意力权重)得到的。

2.3　自适应特征融合模块

自适应特征融合模块旨在动态地学习结合位置和通道注意力特征的最优方式，消除对手动权重调整的依赖。该模块保证了网络在维持类别空间定位准确性的同时，还能够增强对少数关键类别的识别能力，缓解了类别不平衡的问题，进一步提高了模型的整体性能。自适应特征融合模块的具体结构如图 3 所示。

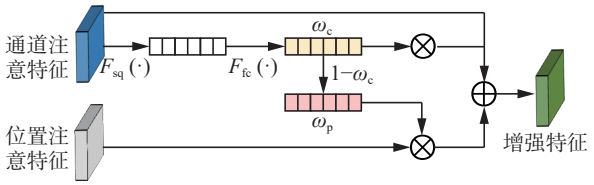


图 3　自适应特征融合模块示意

Fig. 3　Schematic diagram of adaptive feature fusion module

在自适应融合过程中，首先通过全局平均池化操作(表示为$F_{sq}(\cdot)$)，将通道注意力特征的全局空间信息聚合，这一步骤能够捕捉到整个特征图的全局信息。紧接着，通过一个全连接层(表示为$F_{fc}(\cdot)$)，学习通道注意力特征的权重ω_c，用于强调对当前任务更为重要的通道特征。具体计算公式为

$$\hat{f}_{sq}^{(i,k)} = F_{sq}(\hat{f}_{t,c}^{(i,k)}) = \frac{1}{h\times w}\sum_{i=1}^{h\times w}\hat{f}_{t,c}^{(i,k)}(i)$$

$$\omega_c = F_{fc}\left(\hat{f}_{sq}^{(i,k)}\right)$$

式中：$h\times w$代表通道注意力特征的总像素数，$\hat{f}_{sq}^{(i,k)}$代表压缩后的特征。

其次，将得到通道注意力特征的权重ω_c与通道注意力特征相乘。为了确保位置注意力特征与通道注意力特征的权重和为 1，位置注意力特征的权重ω_p设置为$\omega_p = 1-\omega_c$，将其同样地与位置注意力特征相乘。这种设置平衡了特征融合时位置注意力和通道注意力的权重，以免任何一方过度主导最终的特征表示。最后，通过加权融合这两种特征，得到包含更多有用信息的增强特征$\hat{f}_{strong}^{(i,k)}$：

$$\hat{f}_{strong}^{(i,k)} = \omega_c\hat{f}_{t,c}^{(i,k)} + \omega_p\hat{f}_{t,p}^{(i,k)}$$

最后，经过自适应特征融合模块进行特征融合，得到具有多维信息的增强特征$\hat{f}_{strong}^{(i,k)}$，在训练中进一步生成了改进的目标域硬伪标签$\hat{y}_t$，被用于监督目标域的自训练过程。

在训练的过程中，源域通过标准的交叉熵损失函数进行有监督训练，具体公式表示为

$$L_{ce}^{s} = -\sum_{i=1}^{h\times w}\sum_{k=1}^{K} y_s^{(i,k)}\log\left(p_s^{(i,k)}\right)$$

式中：$p_s^{(i,k)}$表示源域图像每个像素点i属于第k类的预测概率，$y_s^{(i,k)}$是对应的真实标签的热编码表示。

目标域的训练采用了对称交叉熵(symmetric cross entropy, SCE)[31]损失L_{sce}^{t}，以增强模型在面对噪声标签时的稳定性，具体公式表示为

$$L_{sce}^{t} = \alpha L_{ce}^{s}(p_t,\hat{y}_t) + \beta L_{ce}^{s}(\hat{y}_t,p_t)$$

式中：α和β是对称交叉熵的平衡系数，p_t表示目标图像的预测概率，$\hat{y}_t$表示原型引导的自适应特征融合模型持续生成的硬伪标签。

$$\hat{y}_t = \max_k \frac{\hat{f}_{strong}^{(i,k)}}{\sum_k \hat{f}_{strong}^{(i,k)}}$$

通过对增强特征$\hat{f}_{strong}^{(i,k)}$进行归一化并在每个像素点上取最大值，利用预设的各类阈值δ_k进行筛选得到伪标签$\tilde{y}_t$，由于硬伪标签是有噪声的，所以只有预测置信度高于给定阈值的硬伪标签才会被用于目标域的自训练过程，伪标签具体计算公式为

$$\tilde{y}_t = \begin{cases} \tilde{y}_t, \tilde{y}_t > \delta_k \\ \text{忽略}, \text{其他} \end{cases}$$

为了进一步增强特征分布的紧凑性，本文采用了原型去噪后的结构一致性损失L_{kl}^{t}[15]，通过计算原型与特征之间的相对距离，得到弱增强和强增强视图之间的 KL 散度，具体定义为

$$z_{\mathcal{T}}^{(i,k)} = \frac{\exp(-\| \mathcal{F}(\mathcal{T}(X_t))^i - \eta^k \| / \tau)}{\sum_{k'}\exp(-\| \mathcal{F}(\mathcal{T}(X_t))^i - \eta^k \| / \tau)}$$

$$L_{kl}^{t} = \mathrm{KL}(Z_{\mathcal{T}} \| Z_{\mathcal{T}'})$$

式中：默认τ为 1；$\mathcal{F}(\cdot)$表示特征提取；$\mathcal{T}(\cdot)$表示弱增强；$z_{\mathcal{T}}$是弱增强试图利用原型分配的相对特征距离；$Z_{\mathcal{T}'}$是强增强试图利用原型分配的相对特征距离，与$z_{\mathcal{T}}$计算一致。

为了保证训练过程中不会有任何类别的原型变为空，本文采用置信度正则化损失L_{reg}^{t}[32]，激励模型均匀地对所有类别分配预测概率：

$$L_{reg}^{t} = -\sum_{i=1}^{h\times w}\sum_{j=2}^{K}\log p_t^{(i,k)}$$

式中：$p_t^{(i,k)}$表示目标域图像每个像素点i属于第k类的预测概率。

因此，本文所提方法总体损失函数计算公式为

$$L_{total} = L_{ce}^{s} + L_{sce}^{t} + \lambda_1 L_{kl}^{t} + \lambda_2 L_{reg}^{t}$$

式中λ_1和λ_2分别设置为 10、0.1，代表不同损失平衡系数。

另外，本文采用学生−教师网络架构，其中学生网络的参数通过指数移动平均(exponential moving average, EMA)更新。EMA 方法有助于平

滑学生网络学习过程中的波动, 提高模型在面对多样化数据时的稳定性, 从而提高整体训练效果。具体更新方式为

$$\theta'_i = \delta\theta'_{i-1} + (1-\delta)\theta_i$$

式中: δ是 EMA 衰减率, θ'_i代表当前迭代教师模型的参数, θ'_{i-1}是上一迭代教师模型的参数, θ_i是当前迭代学生模型的参数。

3 实验与分析

3.1 数据集

本文选取 GTA5[33]、SYNTHIA[34] 和 Cityscapes[35]3 个主流数据集。实验分别在 GTA5-to-Cityscapes 和 SYNTHIA-to-Cityscapes 这 2 个经典域适应任务上进行训练和测试。

GTA5 数据集由 24 966 张图像构成, 分辨率为 1 914×1 052, 具有丰富多样的视觉场景和 33 种类别的精细像素级标注。其中包含与 Cityscapes 数据集 19 个公共类别相对应的精确语义标签, 是探索视觉域适应的理想选择。SYNTHIA 数据集是一个合成的城市景观数据集, 有 9 400 张像素为 1 280×760 的图片, 共包含 16 种与 Cityscapes 公共类别, 常用于进行 16 类和 13 类的分类任务。Cityscapes 数据集则专注于城市街道场景, 包含来自 50 个不同城市的实景图像。它提供了 5 000 张具有高质量像素级注释的图像, 其中包括 2 975 张用于训练、500 张用于验证以及 1 525 张用于测试, 共覆盖 19 个类别。

3.2 实验设置

本文的模型在以下硬件和软件环境中实现: CPU 为 12th Gen Intel(R) Core(TM) i9-12900K, GPU 为 NVIDIA GeForce RTX 3 090 Ti, 环境配置是 PyTorch 1.13.1、CUDA 11.6 以及 Python 3.9.16。本文使用 DeepLabv2[17] 与骨干 ResNet-101[36] 进行分割。之后利用 AdaptSegNet[26] 在分割输出上应用对抗性训练作为预热。本文使用 SGD 优化器, 初始学习率设置为 1×10^{-4}, 动量设置为 0.9, 输入图像批次为 2, 图像的大小随机裁剪为 896×512, 权重衰减系数为 0.000 2, EMA 的衰减率设置为 0.999, 硬伪标签的选择阈值设置为 0.95, 对称交叉熵的平衡系数α和β分别设置为 1.0 和 0.5, 迭代训练 100 个周期。

3.3 实验结果

本文提出的原型引导的自适应特征融合模型(PG-AFFM)与域适应语义分割领域的经典算法及最新的主流方法进行了综合比较, 这些对比方法包括 AdaptSegNet[26]、AdvEnt[27]、FDA(fourier domain adaptation)[37]、PLCA(pixel-level cycle association)[38]、SISC-PWL(spatially independent and semantically consistent-PWL)[39]、ASA(affinity space adaptation)[8]、CLAN(category-level adversarial adaptation)[40]、UDAclustering[41]、PixMatch[42]、ProDA[15]、DRSL(distribution regularized self-supervised learning)[43]、ARAS(adaptive refining-aggregation-separation)[44]、Multi OT[45]、HDL(hybrid domain learning)[46]、SAM[47]、PRLR[48]。

表 1 给出了 GTA5-to-Cityscapes 的域适应语义分割结果。本文的 PG-AFFM 方法取得了 54.7% 的最高平均交并比(mIoU)得分, 与仅在源域训练的非适应基线模型得分 36.6% 相比, PG-AFFM 实现了 18.1 百分点的性能提升, 比起预热模型 AdaptSegNet 的方法, 提高了 13.3 百分点。本文将 PG-AFFM 与 ProDA 蒸馏前的域适应语义分割结果进行对比。在 19 个类别中, PG- AFFM 在 11 个类别上获得了最佳成绩, 虽然在道路、建筑物等类别中不是最优结果, 但是在较难识别的类别(如交通灯、交通标志、行人等)上表现出显著的优势, 交通标志比其中的最低结果高出 36.8 百分点, 土地、摩托上也给出了最好的结果。PG-AFFM 性能的提升主要得益于原型先验知识的引入, 该方法充分利用了这一先验知识对目标域特征进行了全局特征增强, 成功捕捉到了每个类别特征的独特性, 从而能够有效缓解类别不平衡的问题。

表 1　GTA5-to-Cityscapes 域适应语义分割对比结果

Table 1　Comparison results of domain adaptive semantic segmentation from GTA5-to-Cityscapes　　%

方法	年份	具体类别																			mIoU
		道路	人行道	建筑物	墙	栅栏	杆	交通灯	交通标志	植物	土地	天空	行人	骑行者	汽车	卡车	巴士	火车	摩托	自行车	
仅源域	—	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
AdaptSegNet[26]	2018	85.6	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5	41.4
AdvEnt[27]	2019	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5

续表 1

方法	年份	具体类别																			mIoU
		道路	人行道	建筑物	墙	栅栏	杆	交通灯	交通标志	植物	土地	天空	行人	骑行者	汽车	卡车	巴士	火车	摩托	自行车	
FDA[37]	2020	92.5	53.3	82.3	26.5	27.6	36.4	40.5	38.8	82.2	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.8	27.7	46.4	50.5
PLCA[38]	2020	84.0	30.4	82.4	35.3	24.8	32.2	36.8	24.5	85.5	37.2	78.6	66.9	32.8	85.5	40.4	48.0	8.8	29.8	41.8	47.7
SISC-PWL[39]	2020	89.0	45.2	78.2	22.9	27.3	37.4	46.1	43.8	82.9	18.6	61.2	60.4	26.7	85.4	35.9	44.9	**36.4**	37.2	49.3	49.0
ASA[8]	2020	89.2	27.8	81.3	25.3	22.7	28.7	36.5	19.6	83.8	31.4	77.1	59.2	29.8	84.3	33.2	45.6	16.9	34.5	30.8	45.1
CLAN[40]	2021	88.7	35.5	80.3	27.5	25.0	29.3	36.4	28.1	84.5	37.0	76.6	58.4	29.7	81.2	38.8	40.9	5.6	32.9	28.8	45.5
UDAclustering[41]	2021	89.4	30.7	82.1	23.0	22.0	29.2	37.6	31.7	83.9	37.9	78.3	60.7	27.4	84.6	37.6	44.7	7.3	26.0	38.9	45.9
PixMatch[42]	2021	91.6	51.2	84.7	37.3	29.1	24.6	31.3	37.2	86.5	44.3	85.3	62.8	22.6	87.6	38.9	52.3	0.7	37.2	50.0	50.3
ProDA[15]	2021	90.4	54.2	82.1	40.7	34.2	43.0	44.4	52.7	86.5	41.7	82.7	65.0	9.4	86.3	37.8	46.3	0.0	41.1	**50.8**	52.1
DRSL[43]	2022	**92.6**	55.9	82.4	29.0	24.6	42.7	38.3	35.7	85.5	39.5	77.0	64.2	26.2	83.9	19.5	31.6	9.3	27.1	42.5	47.8
ARAS[44]	2023	91.9	45.2	81.8	21.9	25.6	35.5	41.5	33.4	85.1	34.8	73.8	62.5	31.6	85.9	33.8	42.5	7.3	33.8	42.8	47.9
Multi OT[45]	2023	87.8	31.5	80.5	24.7	23.0	26.1	33.8	15.9	84.2	33.6	74.4	57.6	27.7	83.0	41.2	41.5	8.4	27.5	39.0	44.3
HDL[46]	2023	91.5	46.8	86.0	33.6	32.6	37.0	43.6	39.0	86.5	43.4	87.9	64.5	**36.6**	**87.8**	50.5	47.7	0.0	26.7	48.5	52.1
SAM[47]	2023	90.8	47.2	**86.8**	41.5	29.4	35.7	42.4	37.4	86.0	42.1	**88.3**	63.7	35.6	85.1	43.8	54.6	0.0	33.6	47.8	52.2
RPLR[48]	2022	92.3	52.3	84.8	34.7	29.7	32.6	36.7	32.7	83.2	42.5	81.5	60.6	33.3	85.0	**44.2**	48.0	3.8	35.7	37.3	50.1
本文算法	—	83.8	**57.9**	74.1	**44.1**	**38.1**	**45.0**	**51.4**	**52.7**	**88.6**	**47.7**	80.1	**67.8**	30.3	87.4	38.0	**60.2**	1.4	**44.2**	47.2	**54.7**

注：加粗表示在该列中最优。

表 2 给出了在 SYNTHIA-to-Cityscapes 的域适应语义分割任务上，PG-AFFM 方法取得了优异的成绩。表中 mIoU 和 mIoU*分别表示在 Cityscapes 上的 16 个和 13 个类别评估指数，PG-AFFM 分别实现了 53.3% 和 61.6% 的 mIoU 得分，这一方法超越了仅使用源域数据训练的非适应基线模型，后者在相同的类别评估中分别只达到了 34.9% 和 40.3% 的得分，表明了该方法的有效性。特别是在一些难以区分的类别上，如交通标志、行人、巴士等类别表现也较为突出，获得了最高的 mIoU 得分。这些成绩不仅凸显了 PG-AFFM 在域适应语义分割领域的优越性能，也强调了其在处理难度较大类别时相较于其他主流方法的优势。

表 2　SYNTHIA-to-Cityscapes 域适应语义分割对比结果

Table 2　Comparison results of domain adaptive semantic segmentation from SYNTHIA-to-Cityscapes　%

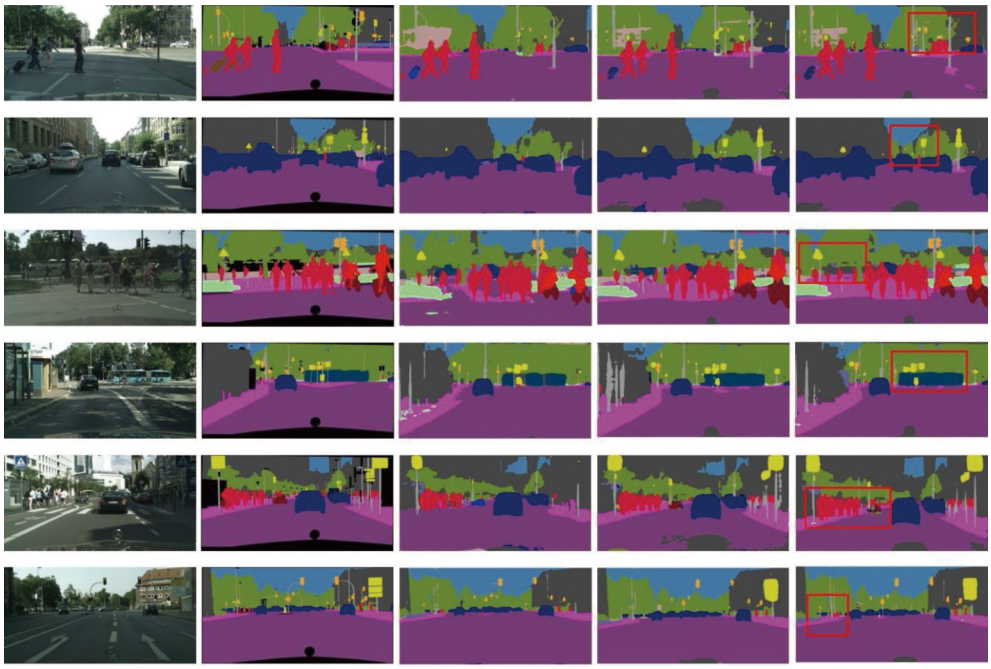
方法	年限	具体类别																mIoU	mIoU*
		道路	人行道	建筑物	墙	栅栏	杆	交通灯	交通标志	植物	天空	行人	骑行者	汽车	巴士	摩托	自行车		
仅源域	—	64.3	21.3	73.1	2.4	1.1	31.4	7.0	27.7	63.1	67.6	42.2	19.9	73.1	15.3	10.5	38.9	34.9	40.3
AdaptSegNet[26]	2018	79.2	37.2	78.8	—	—	—	9.9	10.5	78.2	80.5	53.5	19.6	67.0	29.5	21.6	31.3	—	45.9
AdvEnt[27]	2019	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	41.2	48.0
FDA[35]	2020	73.9	35.0	73.2	—	—	—	19.9	24.0	61.7	82.6	61.4	31.1	83.9	40.8	38.4	51.1	—	52.5
PLCA[36]	2020	82.6	29.0	81.0	11.2	0.2	33.6	24.9	18.3	82.8	82.3	62.1	26.5	85.6	48.9	26.8	52.2	46.8	54.0
SISC-PWL[37]	2020	59.2	30.2	68.5	22.9	1.0	36.2	32.7	28.3	86.2	75.4	68.6	27.7	82.7	26.3	24.3	52.7	45.2	51.0
ASA[8]	2020	91.2	48.5	80.4	3.7	0.3	21.7	5.5	5.2	79.5	83.6	56.4	21.9	80.3	36.2	20.0	32.9	41.7	49.3
CLAN[38]	2021	82.7	37.2	81.5	—	—	—	17.1	13.1	81.2	83.3	55.5	22.1	76.6	30.1	23.5	30.7	—	48.8
UDAclustering[39]	2021	88.3	42.2	79.1	7.1	0.2	24.4	16.8	16.5	80.0	84.3	56.2	15.0	83.5	27.2	6.3	30.7	41.4	48.2
PixMatch[40]	2021	**92.5**	**54.6**	79.8	4.8	0.1	24.1	22.8	17.8	79.4	76.5	60.8	24.7	85.7	33.5	26.4	54.4	46.1	54.5
ProDA[15]	2021	86.9	43.7	84.1	8.0	0.0	**41.9**	34.7	33.1	88.0	84.6	69.0	32.2	88.1	47.6	35.9	50.6	51.8	59.9
DRSL[43]	2022	82.8	40.1	81.3	13.0	1.6	41.6	19.8	33.1	85.3	84.3	59.5	30.1	78.6	25.3	19.8	51.7	46.7	53.2
RPLR[48]	2022	81.5	36.7	78.6	1.3	0.9	32.2	20.7	23.6	79.1	83.4	57.6	30.4	78.5	38.3	24.7	48.4	44.7	52.4
ARAS[44]	2023	85.6	39.2	79.9	15.5	0.3	32.2	19.3	23.9	79.1	81.7	61.1	19.3	82.9	25.7	10.6	51.9	44.3	50.8

续表 2

方法	年限	具体类别																mIoU	mIoU*
		道路	人行道	建筑物	墙	栅栏	杆	交通灯	交通标志	植物	天空	行人	骑行者	汽车	巴士	摩托	自行车		
Multi OT[45]	2023	87.6	43.8	80.6	—	—	—	11.2	12.1	81.1	81.2	56.7	20.1	74.8	33.7	16.8	34.2	—	48.8
HDL[46]	2023	90.8	53.0	83.3	21.2	**3.4**	33.9	**36.9**	24.5	84.2	85.1	63.9	29.9	84.6	51.8	28.3	55.4	51.9	59.3
SAM[47]	2023	77.5	32.3	82.6	**25.5**	1.9	34.6	33.6	32.4	81.7	85.1	63.8	31.8	82.3	35.2	31.9	**54.6**	49.2	55.7
本文算法		83.7	44.3	**84.3**	10.3	0.0	41.4	36.5	**35.3**	**88.3**	**86.8**	**69.8**	**36.3**	**88.1**	**54.8**	**40.0**	48.6	**53.3**	**61.6**

注: 加粗表示在该列中最优。

图 4 给出了 AdaptSegNet、PLCA、UDAclustering、ProDA 算法在 GTA5-to-Cityscapes 任务上的语义分割效果对比。观察红框内区域, 可以发现本文模型在小物体识别上表现出更高的精度, 特别是在细节化地预测人群、交通信号灯、路标和杆状物体等方面。结果表明, 本文方法能够更精确地捕获目标域中的类别特征信息, 并有效地缓解了类别不平衡问题, 从而取得了较好的分割性能。



(a) 原图　(b) 真值　(c) AdaptSegNet　(d) ProDA　(e) 本文算法

图 4　GTA5-to-Cityscapes 域适应语义分割可视化对比结果

Fig. 4　Visual comparison results of GTA5-to-Cityscapes domain adaptation semantic segmentation

3.4　消融实验

本文通过对比完整模型与去除特定模块后的模型性能, 可以更深入地展现每个模块的作用及其在整体模型中的重要性。表 3 为消融实验的结果, 揭示了各个模块对模型最终性能的影响。

表 3　消融各模块的域适应语义分割结果

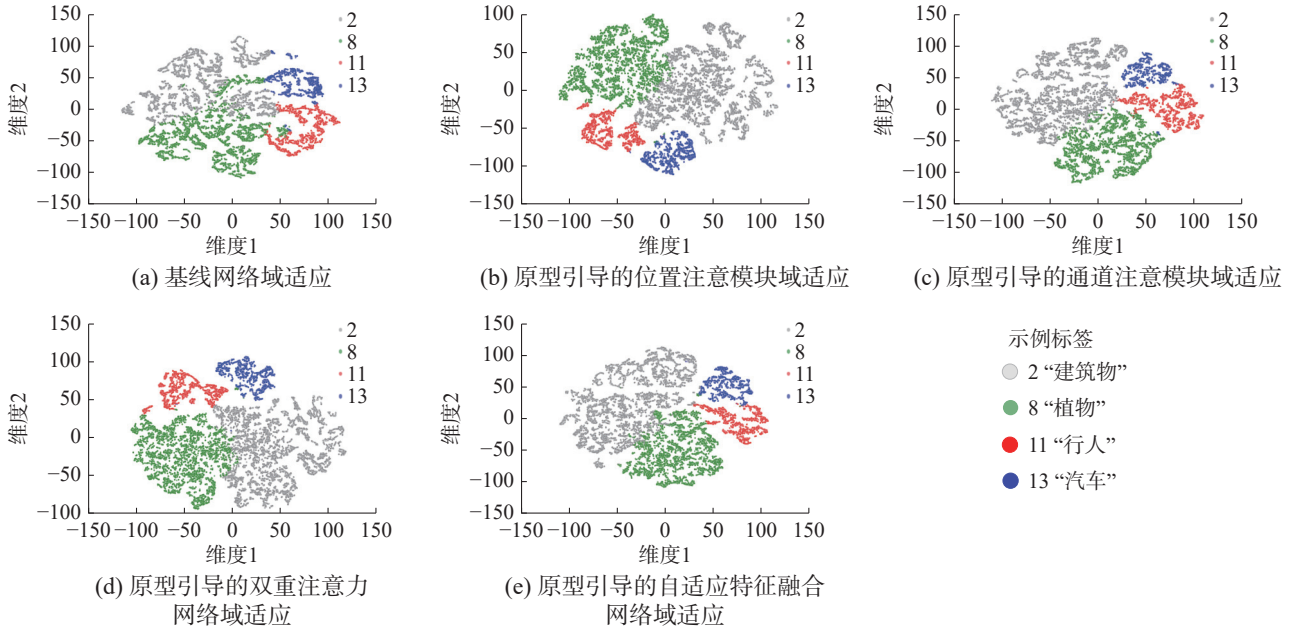
Table 3　Domain adaptive semantic segmentation results for each ablation module　%

原型引导的位置注意模块	原型引导的通道注意模块	自适应特征融合模块	mIoU	增量
√			52.0	+0.0
	√		52.1	+0.1
√	√		52.5	+0.5
√	√	√	54.7	+2.7

训练的初始阶段以 ProDA 模型为基础, 本文设立了一个 52.0% 的 mIoU 基准得分, 这一步骤为后续的实验奠定了基础。接着, 为了提升模型对不同特征通道之间差异的敏感性, 本文引入了原型引导的通道注意力模块。该模块的引入使得分提升至 52.1%。进一步地, 将位置注意力与通道注意力结合, 平等地利用空间和通道信息。这种融合方式使得 mIoU 得分进一步提升至 52.5%, 这一结果表明, 通过综合考虑空间位置和通道特征的不同注意力模型, 可以有效地对数据进行增强。在实验的最后阶段, 本文引入了自适应特征融合模块, 这一模块的加入将 mIoU 得分提高至 54.7%, 表明了特征自适应融合模块在整合关键特

征，尤其是在小物体识别任务上的有效性，如行人、交通灯和交通标志等。

为了展示本文方法如何增强目标特征的类内紧凑性和类间边界的清晰度，并深入分析各模块在特征表示上的作用，本文利用 t-SNE 技术[49]对特征空间进行了可视化。图 5 给出了 4 个类别（建筑物、植物、人和汽车）的特征分布，分别用灰色、绿色、红色和蓝色标识。从图 5（a）可以看出，使用 AdaptSegNet 作为基线模型时，各类别的特征表示相对较为分散。从图 5（b）、（c）可看出，通过引入原型校正机制后，各类特征变得更加集中且密集，表明类内的紧凑性得到了增强。从图 5（d）中，可以明显看到不同类别间的特征开始显著分离，展示了类间边界的清晰化。最终，从图 5（e）可以看出本文提出的模型进一步减少了汽车与建筑物、建筑物与植物间的交叉点。这一系列可视化结果直观地揭示了本文方法在提升语义分割任务中目标特征表示的质量方面的优势。



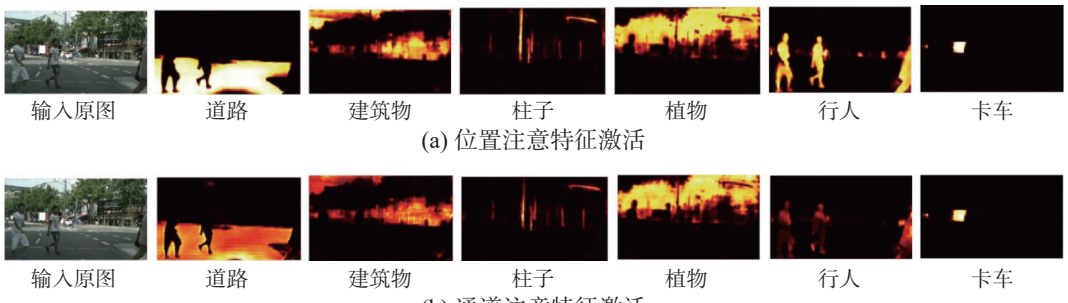
(a) 基线网络域适应　(b) 原型引导的位置注意模块域适应　(c) 原型引导的通道注意模块域适应

(d) 原型引导的双重注意力网络域适应　(e) 原型引导的自适应特征融合网络域适应

图 5　不同类别特征表示的 t-SNE 可视化图

Fig. 5　t-SNE visualization of feature representations across different categories

此外，本文深入分析了自适应特征融合模块对模型性能的必要性。图 6 通过位置和通道注意力模块生成的激活图揭示了这一点，这些图根据输入图像中存在的对象展示了 6 个不同通道的特征表征，深色区域表示较高的注意力，黑色区域表示背景信息。通过观察道路、建筑物、植物、行人等类别的激活情况，本文注意到位置注意力模块可能专注于捕捉边缘和纹理等基本特征，展示了模型层级化处理信息的能力，这对于逐步构建复杂物体的识别非常有用；同时，通道注意力模块增强了与类别判别相关的通道特征，关注哪些通道对于当前任务最为重要。特别是在分析行人特征时，位置和通道注意力的互补性得到了充分体现。因此，自适应特征融合模块的引入充分融合了位置特征和通道特征，形成一个更为全面和综合的特征表征，从而有助于模型更准确地分割和定位物体。



(a) 位置注意特征激活

(b) 通道注意特征激活

图 6　位置和通道注意特征激活

Fig. 6　Activation maps of position and channel attention features

本文在实验中采用的原型是基于目标域中通过伪标签预测得到的。尽管源域数据也可用于构建原型, 但源域和目标域之间存在的域差异可能导致这些原型之间的差异, 引入额外的噪声。为了验证这一点, 本文比较了基于目标域原型和源域原型的方法性能。当以目标域原型为类别参考原型时, 模型训练达到的 mIoU 是 54.7%; 当以源域原型为类别参考原型时, 模型训练达到的 mIoU 是 54%。由此可见, 目标域原型引导的方法得到了最佳的模型性能, 同时也表明了域间隙对域适应语义分割性能的影响。

在目标域的训练阶段, 本文采用了对称交叉熵损失来优化模型, 其中通过平衡交叉熵和逆交叉熵的系数α和β来调节损失。表 4 表明, 随着特征分布变得更加清晰, 适当提高α值有助于稳步提升模型性能。当α=0.5 和β=1 时, 模型展现了最优性能。

表 4　不同α和β设置下的参数敏感性实验

Table 4　Different α and β sensitivity experiment of parameters under setting

α	β			
	0.1	0.5	1	5
0.01	49.3	53.1	54.1	53.8
0.1	49.8	53.7	54.2	54.0
0.5	51.6	54.4	54.7	54.2
1	51.9	53.2	54.3	54.5

4　结束语

本文旨在解决域适应语义分割任务中的两个关键问题: 增强目标域特征的类内紧凑性和缓解数据集类别不平衡引起的类别过度拟合现象。本文提出了一种原型引导的自适应特征融合模型。引入了原型引导的双重注意力网络, 并通过自适应特征融合模块优化目标域特征表征, 确保少数类别的特征在融合过程中不会被那些多数类别所支配, 从而提升模型对于复杂视觉任务的处理能力。在 GTA5-to-Cityscapes 的任务中, 本文提出的 PG-AFFM 模型在 mIoU 评价指标上达到了 54.7%, 相较于本文的最高对比方法, 性能提升了 2.5 百分点。在 SYNTHIA-to-Cityscapes 任务中, 16 类和 13 类分类任务的 mIoU 分别达到了 53.3% 和 61.6%。对比实验结果和消融实验验证了本文方法的有效性, 可以看出其在跨域语义分割及未来自动驾驶视觉系统中的应用潜力。

未来, 我们将专注于将本研究方法应用于多源数据的域适应任务, 尤其是在复杂样本分布的场景中。我们将开发新策略以更好地处理多源数据间的差异, 提高模型在跨域应用中的泛化能力和鲁棒性。

参考文献:

[1] 景庄伟, 管海燕, 彭代峰, 等. 基于深度神经网络的图像语义分割研究综述[J]. 计算机工程, 2020, 46(10): 1−17.
JING Zhuangwei, GUAN Haiyan, PENG Daifeng, et al. Survey of research in image semantic segmentation based on deep neural network[J]. Computer engineering, 2020, 46(10): 1−17.

[2] 计梦予, 袭肖明, 于治楼. 基于深度学习的语义分割方法综述[J]. 信息技术与信息化, 2017(10): 137−140.
JI Mengyu, XI Xiaoming, YU Zhilou. A review of semantic segmentation based on deep learning[J]. Information technology and informatization, 2017(10): 137−140.

[3] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence. Boston: IEEE, 2017: 640−651.

[4] 范苍宁, 刘鹏, 肖婷, 等. 深度域适应综述: 一般情况与复杂情况[J]. 自动化学报, 2021, 47(3): 515−548.
FAN Cangning, LIU Peng, XIAO Ting, et al. A review of deep domain adaptation: general situation and complex situation[J]. Acta automatica sinica, 2021, 47(3): 515−548.

[5] 高德鹏. 基于跨域正则化模型的域适应方法研究[D]. 哈尔滨: 哈尔滨工业大学, 2020.
GAO Depeng. Research on domain adaptation method based on cross-domain regularization model[D]. Harbin: Harbin Institute of Technology, 2020.

[6] 王格格, 郭涛, 余游, 等. 基于生成对抗网络的无监督域适应分类模型[J]. 电子学报, 2020, 48(6): 1190−1197.
WANG Gege, GUO Tao, YU You, et al. Unsupervised domain adaptation classification model based on generative adversarial network[J]. Acta electronica sinica, 2020, 48(6): 1190−1197.

[7] BOUSMALIS K, SILBERMAN N, DOHAN D, et al. Unsupervised pixel-level domain adaptation with generative adversarial networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 95−104.

[8] ZHOU Wei, WANG Yukang, CHU Jiajia, et al. Affinity space adaptation for semantic segmentation across domains[J]. IEEE transactions on image processing, 2021, 30: 2549−2561.

[9] 高子航, 刘兆英, 张婷, 等. 基于对抗域适应的红外舰船目标分割[J]. 数据采集与处理, 2023, 38(3): 598−607.
GAO Zihang, LIU Zhaoying, ZHANG Ting, et al. Infrared ship target segmentation based on adversarial domain adaptation[J]. Journal of data acquisition and pro-

cessing, 2023, 38(3): 598−607.

[10] 张桂梅, 鲁飞飞, 龙邦耀, 等. 结合自集成和对抗学习的域自适应城市场景语义分割[J]. 模式识别与人工智能, 2021, 34(1): 58−67.
ZHANG Guimei, LU Feifei, LONG Bangyao, et al. Domain adaptation semantic segmentation for urban scene combining self-ensembling and adversarial learning[J]. Pattern recognition and artificial intelligence, 2021, 34(1): 58−67.

[11] ZHAO Yihao, WU Ruihai, DONG Hao. Unpaired image-to-image translation using adversarial consistency loss[M]// Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 800−815.

[12] 李美丽, 杨传颖, 石宝. 基于语义分割的图像风格迁移技术研究[J]. 计算机工程与应用, 2020, 56(24): 207−213.
LI Meili, YANG Chuanying, SHI Bao. Research on image style transfer technology based on semantic segmentation[J]. Computer engineering and applications, 2020, 56(24): 207−213.

[13] 吕佳, 李婷婷. 半监督自训练方法综述[J]. 重庆师范大学学报 (自然科学版), 2021, 38(5): 98−106.
LYU Jia, LI Tingting. A summary of semi-supervised self-training methods[J]. Journal of Chongqing normal university (natural science edition), 2021, 38(5): 98−106.

[14] 张勋晖, 周勇, 赵佳琦, 等. 基于熵增强的无监督域适应遥感图像语义分割[J]. 计算机应用研究, 2021, 38(9): 2852−2856.
ZHANG Xunhui, ZHOU Yong, ZHAO Jiaqi, et al. Entropy enhanced unsupervised domain adaptive remote sensing image semantic segmentation[J]. Application research of computers, 2021, 38(9): 2852−2856.

[15] ZHANG Pan, ZHANG Bo, ZHANG Ting, et al. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 12409−12419.

[16] ZHAO Hengshuang, SHI Jianping, QI Xiaojuan, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6230−6239.

[17] 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 transactions on pattern analysis and machine intelligence, 2018, 40(4): 834−848.

[18] YANG Zhen, PENG Xiaobao, YIN Zhijian, et al. Deeplab_v3_plus-net for image semantic segmentation with channel compression[C]//2020 IEEE 20th International Conference on Communication Technology. Nanning: IEEE, 2020: 1320−1324.

[19] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]// 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 9992−10002.

[20] XIE Enze, WANG Wenhai, YU Zhiding, et al. SegFormer: simple and efficient design for semantic segmentation with transformers[J]. Advances in neural information processing systems, 2021, 34: 12077.

[21] LIU Ze, HU Han, LIN Yutong, et al. Swin transformer V2: scaling up capacity and resolution[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 11999−12009.

[22] SMOLA A J, GRETTON A, BORGWARDT K. Maximum mean discrepancy[C]//2006 ICONIP 13th International Conference on Neural Information Processing. Hong-Kong: Springer International Publishing, 2006: 3−6.

[23] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[EB/OL]. (2014−06−10)[2024−03−05]. https://arxiv.org/abs/1406.2661v1.

[24] HOFFMAN J, WANG Dequan, YU F, et al. FCNs in the wild: pixel-level adversarial and constraint-based adaptation[EB/OL]. (2016−12−08)[2024−02−15]. https://doi.org/10.48550/acxiv.

[25] ZHU Junyan, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2242−2251.

[26] TSAI Y H, HUNG W C, SCHULTER S, et al. Learning to adapt structured output space for semantic segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7472−7481.

[27] VU T H, JAIN H, BUCHER M, et al. ADVENT: adversarial entropy minimization for domain adaptation in semantic segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 2512−2521.

[28] JIANG Zhengkai, LI Yuxi, YANG Ceyuan, et al. Prototypical contrast adaptation for Domain adaptive semantic segmentation[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 36−54.

[29] HOYER L, DAI Dengxin, WANG Haoran, et al. MIC: masked image consistency for context-enhanced domain adaptation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 11721−11732.

[30] CHEN Mu, ZHENG Zhedong, YANG Yi, et al. PiPa: pixel- and patch-wise self-supervised learning for domain adaptative semantic segmentation[C]//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023: 1905−1914.

[31] WANG Yisen, MA Xingjun, CHEN Zaiyi, et al. Symmetric cross entropy for robust learning with noisy labels[C]//

2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 322−330.

[32] ZOU Yang, YU Zhiding, LIU Xiaofeng, et al. Confidence regularized self-training[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 5981−5990.

[33] RICHTER S R, VINEET V, ROTH S, et al. Playing for data: ground truth from computer games[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016: 102−118.

[34] ROS G, SELLART L, MATERZYNSKA J, et al. The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vega: IEEE, 2016: 3234−3243.

[35] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 3213−3223.

[36] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[37] YANG Yanchao, SOATTO S. FDA: fourier domain adaptation for semantic segmentation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 4084−4094.

[38] KANG Guoliang, WEI Yunchao, YANG Yi, et al. Pixel-level cycle association: a new perspective for domain adaptive semantic segmentation[J]. Advances in neural information processing systems, 2020, 33: 3569.

[39] IQBAL J, ALI M. MLSL: multi-level self-supervised learning for domain adaptation with spatially independent and semantically consistent labeling[C]//2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass: IEEE, 2020: 1853−1862.

[40] LUO Yawei, LIU Ping, ZHENG Liang, et al. Category-level adversarial adaptation for semantic segmentation using purified features[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 44(8): 3940−3956.

[41] TOLDO M, MICHIELI U, ZANUTTIGH P. Unsupervised domain adaptation in semantic segmentation via orthogonal and clustered embeddings[C]//2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021: 1357−1367.

[42] MELAS-KYRIAZI L, MANRAI A K. PixMatch: unsupervised domain adaptation via pixelwise consistency training[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 12430−12440.

[43] IQBAL J, RAWAL H, HAFIZ R, et al. Distribution regularized self-supervised learning for domain adaptation of semantic segmentation[J]. Image and vision computing, 2022, 124: 104504.

[44] CAO Yihong, ZHANG Hui, LU Xiao, et al. Adaptive refining-aggregation-separation framework for unsupervised domain adaptation semantic segmentation[J]. IEEE transactions on circuits and systems for video technology, 2023, 33(8): 3822−3832.

[45] GUO Yaqian, WANG Xin, LI Ce, et al. Domain adaptive semantic segmentation by optimal transport[J]. Fundamental research, 2024, 4(5): 981−991.

[46] ZHANG Yuhang, TIAN Shishun, LIAO Muxin, et al. A hybrid domain learning framework for unsupervised semantic segmentation[J]. Neurocomputing, 2023, 516: 133−145.

[47] CHUNG I, YOO J, KWAK N. Exploiting inter-pixel correlations in unsupervised domain adaptation for semantic segmentation[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. Waikoloa: IEEE, 2023: 12−21.

[48] LI Jing, ZHOU Kang, QIAN Shenhan, et al. Feature representation and reliable pseudo label retraining for cross-domain semantic segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2024, 46(3): 1682−1694.

[49] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11): 01301.

作者简介：

杨宇宇，硕士研究生，主要研究方向为深度学习、域适应语义分割。E-mail：yyb904yyy@163.com。

杨霄，博士研究生，主要研究方向为计算机视觉、多模态表征学习。E-mail：yangxiao523x@163.com。

王军，教授，博士生导师，主要研究方向为智能机器人与无人系统、生物特征识别、机器视觉。主持新一代人工智能国家科技重大专项。E-mail：jrobot@126.com。