基于多模态特征的无监督领域自适应多级对抗语义分割网络

王泽宇[1]，布树辉[2]，黄伟[1]，郑远攀[1]，吴庆岗[1]，常化文[1]，张旭[1]
（1. 郑州轻工业大学计算机与通信工程学院，河南 郑州 450000；2. 西北工业大学航空学院，陕西 西安 710072）

摘 要：为了解决领域自适应中存在领域间视觉、空间以及语义特征分布差异的问题，提出了基于多模态特征的无监督领域自适应多级对抗语义分割网络。首先，设计 3 层结构的注意力融合语义分割网络来分别从源域和目标域学习上述三类特征。然后，在单级对抗学习中引入联合分布置信度和语义置信度的自监督学习方法，从而在领域间所学特征的分布距离最小化过程中实现更多目标域像素的分布对齐。最后，通过基于多模态特征的多级对抗学习方法对 3 路对抗分支与 3 个自适应子网进行联合优化，从而能够有效学习各子网所提取特征的域间不变表示。实验结果表明，与当前先进方法相比，所提网络在 GTA5 到 Cityscapes、SYNTHIA 到 Cityscapes 和 SUN-RGBD 到 NYUD-v2 的数据集上分别取得最优的平均交并比 62.2%、66.9%和 59.7%。

关键词：无监督领域自适应；语义分割；多模态特征；注意力融合；多级对抗学习；自监督学习
中图分类号：TP391
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2022212

Unsupervised domain adaptation multi-level adversarial network for semantic segmentation based on multi-modal features

WANG Zeyu[1], BU Shuhui[2], HUANG Wei[1], ZHENG Yuanpan[1], WU Qinggang[1], CHANG Huawen[1], ZHANG Xu[1]
1. College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450000, China
2. School of Aeronautics, Northwestern Polytechnical University, Xi'an 710072, China

Abstract: In order to solve the problem of the distribution differences of visual, spatial, and semantic features between domains in domain adaptation, an unsupervised domain adaptation multi-level adversarial network for semantic segmentation based on multi-modal features was proposed. Firstly, an attentive fusion semantic segmentation network with three-layer structure was designed to learn the above three types of features from the source domain and target domain, respectively. Secondly, a self-supervised learning method jointing distribution confidence and semantic confidence was introduced into the single-level adversarial learning, so as to achieve the distribution alignment of more target domain pixels in the process of minimizing the distribution distance of the learnt features between domains. Finally, three adversarial branches and three adaptive sub-networks were jointly optimized by the multi-level adversarial learning method based on multi-modal features, which could effectively learn the invariant representation between domains for the features extracted from each sub-network. The experimental results show that compared with existing state-of-the-art methods, on the datasets of GTA5 to Cityscapes, SYNTHIA to Cityscapes, and SUN-RGBD to NYUD-v2 the proposed network achieves the best mean intersection over union of 62.2%, 66.9%, and 59.7%, respectively.

Keywords: unsupervised domain adaptation, semantic segmentation, multi-modal features, attentive fusion, multi-level adversarial learning, self-supervised learning

收稿日期：2022−06−24；**修回日期：**2022−09−20
基金项目：河南省科技攻关基金资助项目（No.222102210021）；河南省高等学校重点科研项目计划基金资助项目（No.21A520049, No.23A520004）
Foundation Items: The Science and Technology Project of Henan Province (No.222102210021), The Plan Support for Key Scientific Research Project of Higher Education in Henan Province (No.21A520049, No.23A520004)

0 引言

语义分割[1]作为计算机视觉的基础工作，它的核心问题是如何准确地对图像中每个像素进行分类。高精度的语义分割对有效实现机器人任务规划[2]、车辆自动驾驶[3]以及语义 SLAM（simultaneous localization and mapping）[4]等智能视觉任务起到至关重要的作用。因此，基于深度学习的语义分割被广泛研究。卷积神经网络（CNN, convolutional neural network）在语义分割的局部对象视觉特征提取中取得成功[1,5]，但是，由于卷积核感知域较小，因此提取的视觉特征一般缺少全局上下文信息，从而影响分割准确率。为弥补 CNN 空间结构化学习能力的不足，长短期记忆网络（LSTM, long short-term memory network）联合 CNN 的混合网络应运而生，文献[6]通过 LSTM 逐像素地遍历图像视觉特征来学习对象间的依赖关系，从而显式地推理全局场景的空间特征。为进一步提升分割精度，基于注意力机制的局部和全局特征融合被应用于语义分割，文献[7]根据对象视觉特征和所处全局场景空间特征的相关性自适应地聚合有用上下文信息并屏蔽噪声上下文信息，从而生成高质量的综合语义特征。

虽然上述有监督训练语义分割网络取得成功，但是，在有标签源域到无标签目标域的无监督领域自适应中，尽管领域间具有较高的语义相似性，由于目标域没有标签，不能直接优化网络参数，仅基于源域训练的网络无法理想地分割目标域场景，而人工制作目标域标签又必然提升成本。为能够利用无标签目标域间接调优网络参数，对抗学习[8]被广泛应用于领域自适应中。文献[9]首次通过对抗学习和附加类别约束减小领域间特征分布差异。基于此，文献[10]提出多级对抗学习，通过设置的多个判别器与语义分割网络的不同层次进行对抗，从而对齐不同抽象级别特征的分布。但是，由于领域间存在对象纹理不同和外界环境变化（季节、天气以及光照等）引起的视觉风格差异，从而出现误识别相兼容语义类别的问题。

为此，循环一致性[11]和域内风格不变表示（ISR, intra-domain style-invariant representation）[12]等图像风格转换方法被应用于减小领域间视觉风格差异。在此基础上，双向学习（BDL, bidirectional learning）[13]、标签驱动重建（LDR, label driven reconstruction）[14]和双路径学习（DPL, dual path learning）[15]等方法均提出一种双向学习框架，通过图像风格转换网络和语义分割网络相互促进，从而在确保语义内容不变的情况下实现源域到目标域图像的视觉风格转换，进而降低特征分布的错误对齐。进一步地，文献[16]提出零风格损失来分离图像的语义内容和视觉风格，从而使用去除风格差异的源域和目标域图像进行有监督训练。但是，引入图像风格转换方法会增大网络结构的复杂度，同时降低网络的训练效率。

为了利用无标签目标域有监督训练语义分割网络，自监督学习被用来为目标域图像生成标签，并基于多分类交叉熵损失调优网络，从而直接拉近领域间的特征分布差异[17-19]。两阶段目标域标签密集化（TPLD, two-phase pseudo label densification）生成策略[17]解决了目标域标签过于稀疏而导致的特征分布距离无法有效拉近问题。无监督域内自适应（UIA, unsupervised intra-domain adaptation）学习方法[18]首先按照同源域的分布接近程度对目标域进行划分，然后按照分布差异由小到大的顺序逐次对齐分布。文献[19]通过不确定学习策略迭代自动纠正目标域生成的错误标签，从而不断提升所生成标签的正确率。但是，上述自监督学习方法无法同时确保选定目标域子集的稠密性和目标域子集所生成标签的正确性，从而导致目标域中出现较多未充分对齐或错误对齐的像素。

值得注意的是，由于上述领域自适应方法所训练 ResNet-101（101-layer residual network）[5]的空间结构化学习能力有限，因此，虽然对抗学习、图像风格转换学习和自监督学习在对齐领域间局部对象视觉特征分布上取得成功，但是上述方法无法有效减小全局场景空间特征的分布差异，从而由于缺少目标域场景的全局上下文信息而影响综合语义特征的生成质量。为此，CDA（context-aware domain adaptation）[20]提出跨域的空间和通道注意力模块，用来学习领域间共享的上下文信息，并基于对抗学习减小上下文信息的分布差异。另外，文献[21]通过采样和聚类的方法显式学习领域间的上下文依赖关系，并同样基于对抗学习对齐结构化特征的分布。但是，上述方法未能全面地减小领域间视觉和空间特征的分布差异，同时没有考虑融合视觉和空间信息的综合语义特征的分布对齐。

综上，由于领域间不仅存在局部对象的颜色、形状以及纹理等视觉外观差异，而且存在全局场景

的环境、布局以及对象间边界等空间结构不同，因此，领域自适应不仅需要减小局部对象的视觉特征分布差异，而且需要减小全局场景的空间特征分布差异，同时需要对齐融合视觉和空间信息的综合语义特征分布。但是，现有方法[9-21]均未考虑全面减小上述三类特征的分布差异，从而导致无法在目标域场景有效生成融合对象视觉和空间信息的综合语义特征，这不仅会影响易混淆类别的区分，而且会出现尺寸较小对象的误分割，因此，如何全面最小化领域间视觉、空间以及语义等三类特征的分布距离成为领域自适应需要解决的核心问题。为此，本文提出基于多模态特征的无监督领域自适应多级对抗网络（UDAMAN-MF, unsupervised domain adaptation multi-level adversarial network based on multi-modal features），首先，设计 3 层结构语义分割网络分别从源域和目标域学习视觉、空间以及语义特征，从而为领域间上述三类特征的分布对齐奠定网络结构基础；然后，在单级对抗学习中引入改进的自监督学习，从而在特征分布距离最小化过程中实现更大目标域子集的分布对齐；最后，基于多级对抗学习全面对齐 3 层网络所学三类特征的分布，从而有效学习各类特征的域间不变表示。主要贡献如下。

1) 提出基于 3 层结构的注意力融合语义分割网络。所提网络由特征提取层、结构化学习层和特征融合层组成，3 层子网能够从源域和目标域分别学习局部对象的多维视觉特征（HVF, hierarchical visual feature）、全局场景的空间结构化特征（SSF, spatial structural features）以及包含综合语义的多模态混合特征（MHF, multi-modal hybrid features），为领域间视觉、空间以及语义特征的分布对齐奠定基础。

2) 联合分布置信度和语义置信度的自监督学习。为特征分布接近源域并且语义分类概率较高的目标域子集生成标签，以同时确保选定子集的稠密性和生成标签的正确性，从而能够通过有监督训练直接对齐接近源域的有标签目标域子集的分布，进而有助于无监督对抗学习对齐远离源域的无标签目标域子集，以实现更大目标域子集的分布对齐。

3) 基于多模态特征的多级对抗学习方法。通过 3 路对抗分支与 3 个自适应子网的联合对抗训练，以充分调优各子网的参数，从而全面减小低层子网所学视觉特征、中层子网所学空间特征以及整个网络所学语义特征的分布差异，进而有效学习上述三类特征的域间不变表示。

1　UDAMAN-MF 结构与学习方法

在领域自适应中，由于领域间不仅存在局部对象视觉外观特征的分布差异，而且存在全局场景空间结构化特征的分布差异，同时存在包含对象视觉和空间信息的综合语义特征的分布差异，因此，如何全面地减小上述三类特征的分布差异成为领域自适应研究的关键。为此，本文提出 UDAMAN-MF。UDAMAN-MF 由 2 个相互对抗的模块组成，即基于 3 层结构的注意力融合语义分割网络 G 和基于 3 路并行对抗分支的判别器 D，结构如图 1 所示。G 由特征提取层 G^{HVF}、结构化学习层 G^{SSF} 和特征融合层 G^{MHF} 组成，分别用来提取局部对象的多维视觉特征、推理全局场景的空间结构化特征以及融合生成包含对象综合语义的多模态混合特征，从而为领域间视觉、空间以及语义等三类特征的分布对齐奠定网络结构基础；D 由 3 路并行的对抗分支 D^{HVF}、D^{SSF} 和 D^{MHF} 构成，用来与低层子网 G^{HVF}、中层子网 G^{HVF}+G^{SSF} 以及整个网络 G^{HVF}+G^{SSF}+G^{MHF} 进行多级对抗训练，从而逐步减小领域间各子网所学特征的分布差异。

1.1　基于 3 层结构的注意力融合语义分割网络

为了分别从源域和目标域场景学习视觉、空间以及语义特征，本文提出基于 3 层结构的注意力融合语义分割网络，具体结构如图 2 所示，其中，前端的特征提取层 G^{HVF} 通过 ResNet-101 提取局部对象的多维视觉特征，中端的结构化学习层 G^{SSF} 采用 LSTM 推理全局场景的空间结构化特征，后端的特征融合层 G^{MHF} 基于注意力机制生成包含对象综合语义的多模态混合特征，从而为领域间上述三类特征的分布对齐奠定网络结构基础。

1.1.1　基于 ResNet-101 的特征提取层

在特征提取层，ResNet-101 提取局部对象的多维视觉特征。ResNet-101 共 5 层，设输入图像为 $\boldsymbol{I}$，则第 l 层的特征提取过程可以表示为

$$\boldsymbol{F}_l^{\mathrm{HVF}}=\begin{cases}\mathrm{maxpool}(\mathrm{conv}(\boldsymbol{I})) & ,l=1\\ \underbrace{\mathrm{resconv\_2}}_{x_l}(\mathrm{resconv\_1}(\boldsymbol{F}_{l-1}^{\mathrm{HVF}})) & ,2\leqslant l\leqslant 5\end{cases} \tag{1}$$

其中，$\boldsymbol{F}_l^{\mathrm{HVF}}$ 表示第 l 层的输出特征；函数 conv 和 maxpool 分别表示第一层中 7×7 卷积操作和 3×3 最大池化操作；函数 resconv_1 和 resconv_2 分别表示第 2 层～第 5 层中两类残差卷积操作；x_l 表示

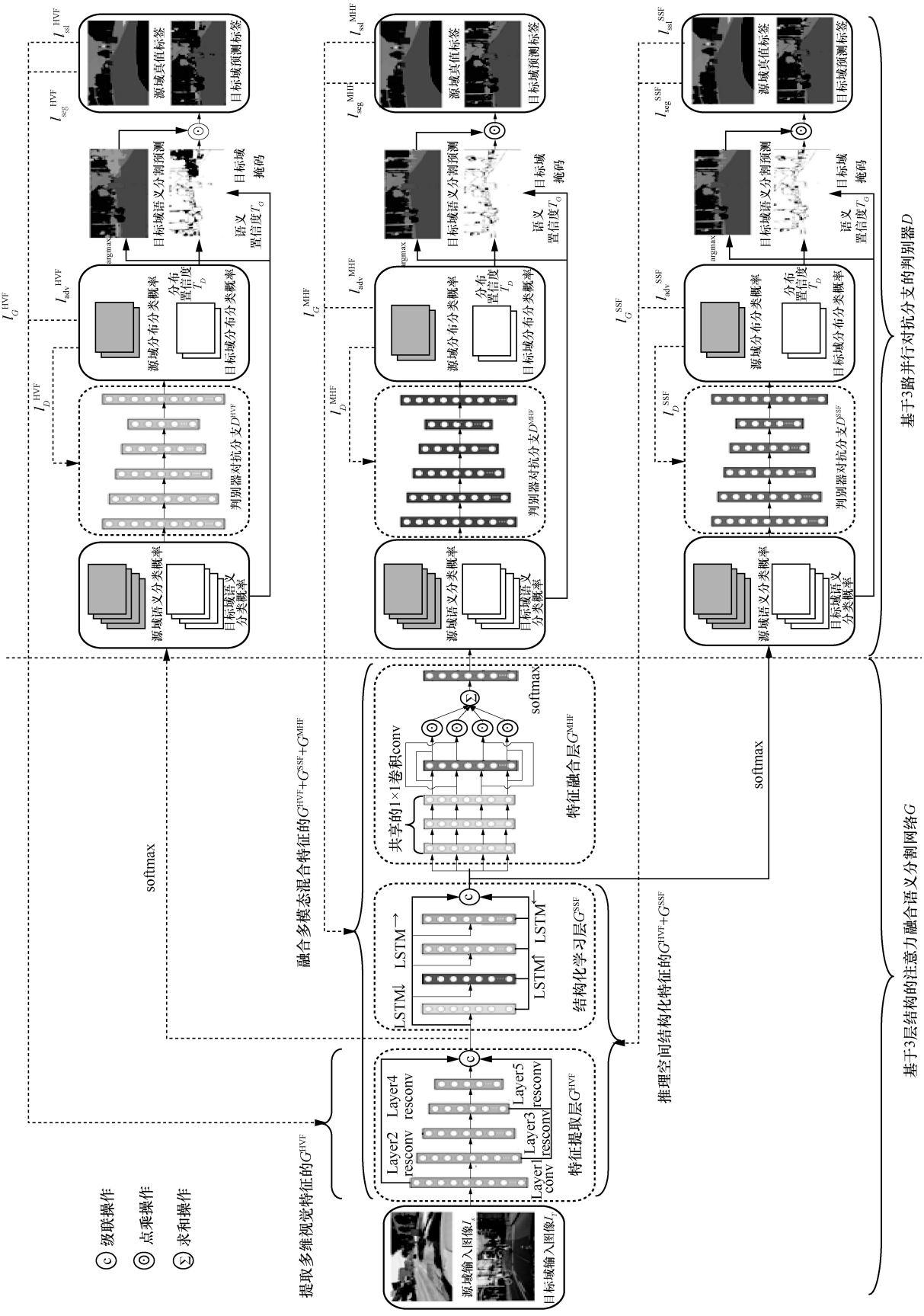


图 1　UDAMAN-MF 结构

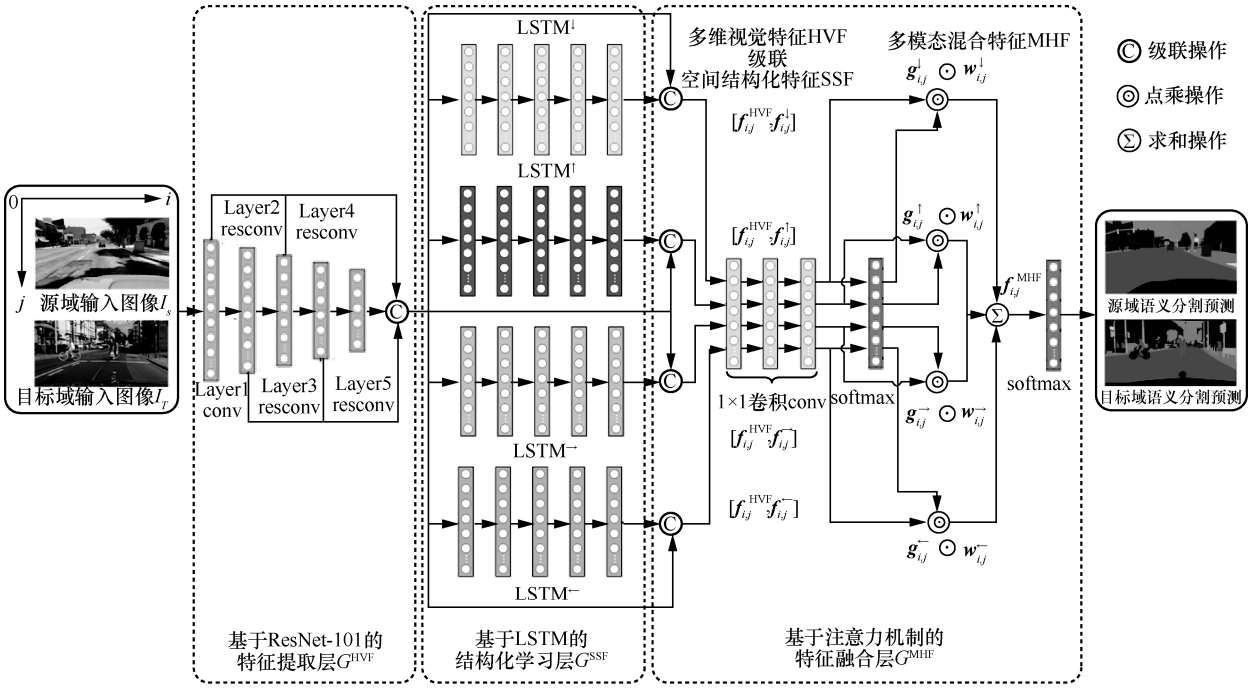


图 2　基于 3 层结构的注意力融合语义分割网络结构

第 l 层（$2 \leqslant l \leqslant 5$）中堆叠 resconv_2 的个数。

然后，级联 ResNet-101 各层上采样的特征，则图像中像素(i,j)的多维视觉特征可以表示为 $\boldsymbol{f}_{i,j}^{\mathrm{HVF}}$，如式(2)所示。

$$\begin{aligned}&\boldsymbol{F}^{\mathrm{HVF}}=\left[\mathrm{up}(\boldsymbol{F}_1^{\mathrm{HVF}}),\cdots,\mathrm{up}(\boldsymbol{F}_l^{\mathrm{HVF}}),\cdots,\mathrm{up}(\boldsymbol{F}_5^{\mathrm{HVF}})\right]\in\mathbb{R}^{n\times h\times w}\\&\boldsymbol{f}_{i,j}^{\mathrm{HVF}}=\boldsymbol{F}^{\mathrm{HVF}}(i,j)\in\mathbb{R}^n,1\leqslant i\leqslant w,1\leqslant j\leqslant h\end{aligned}\tag{2}$$

其中，n、h 和 w 分别表示多维视觉特征的维数、高度和宽度，函数 up 表示上采样操作。

1.1.2　基于长短期记忆网络的结构化学习层

结构化学习层由 4 路长短期记忆网络分支组成，各分支均由 5 层 LSTM 单元堆叠而成。4 路分支分别从 4 个方向（用↓↑→←表示）逐像素地遍历多维视觉特征，上述结构化学习过程可以表示为

$$\boldsymbol{h}_{l,i,j}^{\downarrow}=\begin{cases}\mathrm{LSTM}_l^{\downarrow}(\boldsymbol{h}_{l,i,j-1}^{\downarrow},\boldsymbol{f}_{i,j}^{\mathrm{HVF}})\in\mathbb{R}^{d_l},l=1,\\\mathrm{LSTM}_l^{\downarrow}(\boldsymbol{h}_{l,i,j-1}^{\downarrow},\boldsymbol{h}_{l-1,i,j})\in\mathbb{R}^{d_l},l>1,\end{cases}\quad\text{for } j=1,\cdots,h$$

$$\boldsymbol{h}_{l,i,j}^{\uparrow}=\begin{cases}\mathrm{LSTM}_l^{\uparrow}(\boldsymbol{h}_{l,i,j+1}^{\uparrow},\boldsymbol{f}_{i,j}^{\mathrm{HVF}})\in\mathbb{R}^{d_l},l=1,\\\mathrm{LSTM}_l^{\uparrow}(\boldsymbol{h}_{l,i,j+1}^{\uparrow},\boldsymbol{h}_{l-1,i,j})\in\mathbb{R}^{d_l},l>1,\end{cases}\quad\text{for } j=h,\cdots,1$$

$$\boldsymbol{h}_{l,i,j}^{\rightarrow}=\begin{cases}\mathrm{LSTM}_l^{\rightarrow}(\boldsymbol{h}_{l,i-1,j}^{\rightarrow},\boldsymbol{f}_{i,j}^{\mathrm{HVF}})\in\mathbb{R}^{d_l},l=1,\\\mathrm{LSTM}_l^{\rightarrow}(\boldsymbol{h}_{l,i-1,j}^{\rightarrow},\boldsymbol{h}_{l-1,i,j})\in\mathbb{R}^{d_l},l>1,\end{cases}\quad\text{for } i=1,\cdots,w$$

$$\boldsymbol{h}_{l,i,j}^{\leftarrow}=\begin{cases}\mathrm{LSTM}_l^{\leftarrow}(\boldsymbol{h}_{l,i+1,j}^{\leftarrow},\boldsymbol{f}_{i,j}^{\mathrm{HVF}})\in\mathbb{R}^{d_l},l=1,\\\mathrm{LSTM}_l^{\leftarrow}(\boldsymbol{h}_{l,i+1,j}^{\leftarrow},\boldsymbol{h}_{l-1,i,j})\in\mathbb{R}^{d_l},l>1,\end{cases}\quad\text{for } i=w,\cdots,1$$

$$\boldsymbol{h}_{l-1,i,j}=\mathrm{conv}([\boldsymbol{h}_{l-1,i,j}^{\downarrow},\boldsymbol{h}_{l-1,i,j}^{\uparrow},\boldsymbol{h}_{l-1,i,j}^{\rightarrow},\boldsymbol{h}_{l-1,i,j}^{\leftarrow}])\in\mathbb{R}^{d_{l-1}},l>1\tag{3}$$

其中，$\boldsymbol{h}_{l,i,j}{}^{\downarrow}$、$\boldsymbol{h}_{l,i,j}{}^{\uparrow}$、$\boldsymbol{h}_{l,i,j}{}^{\rightarrow}$和 $\boldsymbol{h}_{l,i,j}{}^{\leftarrow}$分别表示第 l（$1 \leqslant l \leqslant 5$）层 $\mathrm{LSTM}_l{}^{\downarrow}$、$\mathrm{LSTM}_l{}^{\uparrow}$、$\mathrm{LSTM}_l{}^{\rightarrow}$和 $\mathrm{LSTM}_l{}^{\leftarrow}$单元的隐藏层状态，$d^l$ 表示第 l 层各 LSTM 单元隐藏层状态的维数，函数 conv 表示 1×1 卷积操作。

多维视觉特征经过 4 路 LSTM 分支在 4 个不同方向上的逐像素遍历，推理的空间结构化特征可以定义为

$$\begin{aligned}\boldsymbol{F}^{\mathrm{SSF}}=&\{[\boldsymbol{f}_{i,j}^{\downarrow},\boldsymbol{f}_{i,j}^{\uparrow},\boldsymbol{f}_{i,j}^{\rightarrow},\boldsymbol{f}_{i,j}^{\leftarrow}]\}_{i=1,j=1}^{i=w,j=h}=\\&\{[\boldsymbol{h}_{5,i,j}^{\downarrow},\boldsymbol{h}_{5,i,j}^{\uparrow},\boldsymbol{h}_{5,i,j}^{\rightarrow},\boldsymbol{h}_{5,i,j}^{\leftarrow}]\}_{i=1,j=1}^{i=w,j=h}\in\mathbb{R}^{4d_5\times h\times w}\end{aligned}\tag{4}$$

其中，$\boldsymbol{f}_{i,j}{}^{\downarrow}$、$\boldsymbol{f}_{i,j}{}^{\uparrow}$、$\boldsymbol{f}_{i,j}{}^{\rightarrow}$和 $\boldsymbol{f}_{i,j}{}^{\leftarrow}$分别表示像素$(i,j)$在上、下、左和右 4 个不同方向上场景区域的全局上下文信息，$4d_5$ 表示空间结构化特征的维数。

1.1.3　基于注意力机制的特征融合层

为自适应融合多维视觉特征 HVF 和空间结构化特征 SSF，对于图像中每个像素(i, j)，首先，将其 HVF 分别与 4 个方向上的 SSF 级联，并通过卷积操作对 4 个方向上的级联特征依次降维，上述过程可以表示为

$$\boldsymbol{g}_{i,j}^{q}=\mathrm{conv}([\boldsymbol{f}_{i,j}^{\mathrm{HVF}},\boldsymbol{f}_{i,j}^{q}])\in\mathbb{R}^{e},q\in\{\downarrow,\uparrow,\rightarrow,\leftarrow\}\tag{5}$$

其中，conv 表示 3 层 1×1 卷积操作，$\boldsymbol{g}_{i,j}{}^{q}$ 表示 q 方向上的降维特征，e 表示 $\boldsymbol{g}_{i,j}{}^{q}$ 的维数。

然后，利用 softmax 函数分别计算 4 个方向上降维特征的注意力权重 $\boldsymbol{w}_{i,j}{}^{q}$。

$$\boldsymbol{w}_{i,j}^{q}=\mathrm{softmax}(\boldsymbol{g}_{i,j}^{q})\in[0,1]^{e},q\in\{\downarrow,\uparrow,\rightarrow,\leftarrow\} \quad (6)$$

其中，e 表示注意力权重 $\boldsymbol{w}_{i,j}{}^{q}$ 的维数。

接着，基于注意力机制对 4 个方向上的降维特征 $\boldsymbol{g}_{i,j}{}^{q}$ 分别赋予对应的权重 $\boldsymbol{w}_{i,j}{}^{q}$，从而生成多模态混合特征 $\boldsymbol{f}_{i,j}^{\mathrm{MHF}}$，表示为

$$\begin{aligned}&\boldsymbol{f}_{i,j}^{\mathrm{MHF}}=\sum_{q}\boldsymbol{w}_{i,j}^{q}\odot\boldsymbol{g}_{i,j}^{q}\in\mathbb{R}^{e},q\in\{\downarrow,\uparrow,\rightarrow,\leftarrow\}\\&\boldsymbol{F}^{\mathrm{MHF}}=\{\boldsymbol{f}_{i,j}^{\mathrm{MHF}}\}_{i=1,j=1}^{i=w,j=h}\in\mathbb{R}^{e\times h\times w}\end{aligned} \quad (7)$$

其中，⊙表示点乘操作，e、h 和 w 分别表示多模态混合特征的维数、高度和宽度。

最后，利用 softmax 分类器并根据多模态混合特征逐像素地标注图像的语义类别。

1.2 面向 3 层结构语义分割网络的多级对抗学习方法

为了全面对齐领域间网络 G 所学多维视觉特征 HVF、空间结构化特征 SSF 以及多模态混合特征 MHF 的分布，本文提出基于多模态特征的多级对抗学习方法。通过判别器 D 中设置的 3 路并行分支 D^{HVF}、D^{SSF} 和 D^{MHF} 与网络 G 中各自适应子网 G^{HVF}、G^{HVF}+G^{SSF} 和 G^{HVF}+G^{SSF}+G^{MHF} 分别进行单级对抗训练，从而逐步减小领域间各层子网所学模态特征的分布差异，进而有效学习上述三类特征的域间不变表示。

对于网络 G 所学每类特征，为了充分对齐目标域中各像素的该类特征表示，在单级对抗学习中引入联合分布置信度和语义置信度的自监督学习方法。一方面，改进自监督学习为分布接近源域并且语义分类概率较高的目标域子集生成标签，以同时确保选定目标域子集的稠密性和所生成标签的正确性，从而能够基于多分类交叉熵损失直接对齐有标签目标域子集的分布；另一方面，对抗学习通过对抗分支与对应子网的竞争，从而间接拉近远离源域的无标签目标域子集与源域间的分布差异。两者相互结合，从而在领域间所学特征分布距离最小化中实现更多目标域像素的分布对齐。

1.2.1 基于改进自监督学习的单级对抗学习方法

设标签源域 S={($\boldsymbol{I}_S$,$\boldsymbol{Y}_S$)}，$\boldsymbol{Y}_S$ 表示源域图像 $\boldsymbol{I}_S$ 的标签，无标签目标域 T={$\boldsymbol{I}_T$}，$\boldsymbol{I}_T$ 表示目标域图像。在从源域 S 到目标域 T 的领域自适应中，一般对语义分割网络 G 和判别器 D 进行单级对抗训练[13]，其中，D 训练的目标是能够准确地区分源域和目标域语义分类概率间的分布差异；G 训练的目标是使目标域语义分类概率能够不断接近源域语义分类概率的分布，从而达到成功欺骗 D 的目的。两者相互对抗，从而逐步减小领域间网络 G 所学特征的分布差异。

1) 判别器 D 的训练

为训练判别器 D 区分源域和目标域语义分类概率的能力，目标函数定义为二分类交叉熵损失 l_D，即

$$\begin{aligned}&l_D(\boldsymbol{I}_S,\boldsymbol{I}_T,z)=\\&-\sum_{h,w}\left[(1-z)\ln(D(G(\boldsymbol{I}_T))^{(h,w,0)}+z\ln(D(G(\boldsymbol{I}_S))^{(h,w,1)}\right]\end{aligned} \quad (8)$$

其中，当 z = 1 时，判别器 D 的输入为源域语义分类概率 $G(\boldsymbol{I}_S)$；当 z = 0 时，判别器 D 的输入为目标域语义分类概率 $G(\boldsymbol{I}_T)$。

2) 语义分割网络 G 的训练

为对齐目标域中更多像素的分布，网络 G 的训练分为 3 个过程：①基于多分类交叉熵损失，使用有标签源域 S 对网络 G 进行有监督训练；②基于改进自监督学习损失为分布接近源域并且语义分类概率较高的目标域子集生成标签，并利用包含标签的目标域子集优化网络 G；③基于对抗学习损失，通过判别器 D 的竞争对抗再次调优网络 G。较不包含自监督学习的对抗训练，包含改进自监督学习的对抗训练能够对齐更大目标域子集的分布，如图 3(a)和图 3(b)所示。

首先，使用有标签源域 S 训练网络 G，该目标函数定义为多分类交叉熵损失 l_{seg}，计算式为

$$l_{\mathrm{seg}}(\boldsymbol{I}_S,\boldsymbol{Y}_S)=-\sum_{h,w}\sum_{c\in C}\boldsymbol{Y}_S^{(h,w,c)}\ln(G(\boldsymbol{I}_S)^{(h,w,c)}) \quad (9)$$

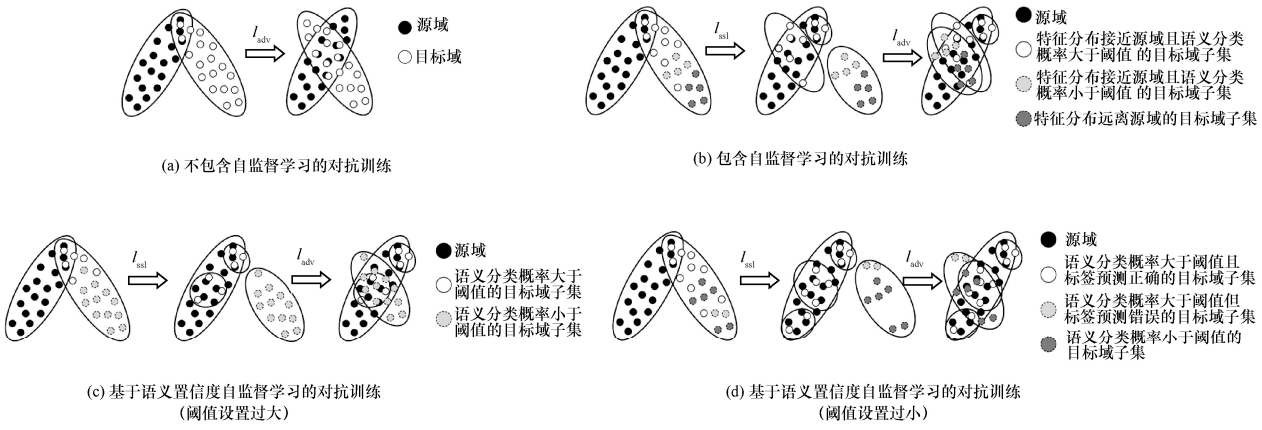
其中，$\boldsymbol{Y}_S$ 表示源域图像 $\boldsymbol{I}_S$ 的标签，$G(\boldsymbol{I}_S)$表示源域语义分类概率，C 表示语义类别数。

然后，为了利用无标签目标域 T 有监督训练网络 G，先通过基于源域 S 预训练的网络 G 为目标域图像 $\boldsymbol{I}_T$ 生成伪标签 $\boldsymbol{Y}_T$，计算式为

$$\boldsymbol{Y}_T^{(h,w,c)}=\begin{cases}1,c=\mathrm{argmax}(G(\boldsymbol{I}_T)^{(h,w,c)})\\0,c\neq\mathrm{argmax}(G(\boldsymbol{I}_T)^{(h,w,c)})\end{cases} \quad (10)$$

其中，函数 argmax 用来选择目标域语义分类概率 $G(\boldsymbol{I}_T)$中最大值对应的通道作为图像 $\boldsymbol{I}_T$ 的伪标签。

为选择伪标签中高可信的部分作为自监督学习的真值标签，一般基于语义置信度选择语义分类概率大于阈值的目标域子集生成标签并进行有监督训练[13]。但是，若语义置信度阈值设置过大，则无法保证包含标签的目标域子集的稠密性，从



(a) 不包含自监督学习的对抗训练　　(b) 包含自监督学习的对抗训练

(c) 基于语义置信度自监督学习的对抗训练（阈值设置过大）　　(d) 基于语义置信度自监督学习的对抗训练（阈值设置过小）

图 3　基于不同自监督学习方法的领域间特征分布对齐

而导致对抗训练无法充分对齐剩余较大无标签目标域子集的分布，如图 3(c)所示；若阈值设置过小，则无法保证选定目标域子集所生成标签的正确性，从而导致部分有标签目标域子集出现错误的分布对齐，如图 3(d)所示。但是，若先求得分布接近源域的目标域子集，再从接近源域的子集中选择语义分类概率大于阈值的像素生成标签，则不需要设置较大的阈值便可保证其生成标签的正确性，同时，迭代的自监督学习和对抗学习能够不断增大分布接近源域的目标域子集尺寸，进而可确保包含标签的目标域子集的稠密性。因此，为特征分布接近源域并且语义分类概率较高的目标域子集生成标签，并基于改进多分类交叉熵损失函数优化网络 G，该目标函数定义为自监督学习损失 l_{ssl}，计算式为

$$
\begin{aligned}
&l_{\text{ssl}}(\boldsymbol{I}_T)=-\sum_{h,w}\boldsymbol{M}_D^{(h,w,1)}\sum_{c\in C}\boldsymbol{M}_G^{(h,w,c)}\boldsymbol{Y}_T^{(h,w,c)}\ln(G(\boldsymbol{I}_T)^{(h,w,c)})\\
&\boldsymbol{M}_D^{(h,w,1)}=\begin{cases}1, D(G(\boldsymbol{I}_T))^{(h,w,1)}>T_D\\ 0, D(G(\boldsymbol{I}_T))^{(h,w,1)}\leqslant T_D\end{cases}\\
&\boldsymbol{M}_G^{(h,w,c)}=\begin{cases}1, G(\boldsymbol{I}_T)^{(h,w,c)}>T_G\\ 0, G(\boldsymbol{I}_T)^{(h,w,c)}\leqslant T_G\end{cases}
\end{aligned}
\tag{11}
$$

其中，$\boldsymbol{M}_D$ 表示伪标签 $\boldsymbol{Y}_T$ 的分布置信度掩码，用来选择分布接近源域的目标域子集；T_D 表示分布置信度阈值，当分布分类概率 $D(G(\boldsymbol{I}_T))>T_D$ 时，掩码置 1，否则置 0；$\boldsymbol{M}_G$ 表示伪标签 $\boldsymbol{Y}_T$ 的语义置信度掩码，用来选择语义分类概率较高的目标域子集；T_G 表示语义置信度阈值，当语义分类概率 $G(\boldsymbol{I}_T)>T_G$ 时，掩码置 1，否则置 0。

最后，利用判别器 D 与网络 G 进行对抗，从而达到目标域语义分类概率 $G(\boldsymbol{I}_T)$成功欺骗判别器 D 的目的，该目标函数定义为对抗学习损失 l_{adv}，计算式为

$$
l_{\text{adv}}(\boldsymbol{I}_T)=-\sum_{h,w}\ln(D(G(\boldsymbol{I}_T))^{(h,w,1)}) \tag{12}
$$

综上，为使网络 G 生成的目标域语义分类概率 $G(\boldsymbol{I}_T)$不断接近源域语义分类概率 $G(\boldsymbol{I}_S)$的分布，定义网络 G 的单级领域自适应损失 l_G 为

$$
\begin{aligned}
&l_G(\boldsymbol{I}_S,\boldsymbol{Y}_S,\boldsymbol{I}_T)=\\
&\lambda_{\text{seg}}l_{\text{seg}}(\boldsymbol{I}_S,\boldsymbol{Y}_S)+\lambda_{\text{ssl}}l_{\text{ssl}}(\boldsymbol{I}_T)+\lambda_{\text{adv}}l_{\text{adv}}(\boldsymbol{I}_T)
\end{aligned}
\tag{13}
$$

其中，l_{seg} 表示多分类交叉熵损失函数，l_{ssl} 表示自监督学习损失函数，l_{adv} 表示对抗学习损失函数，λ_{seg}、λ_{ssl} 和 λ_{adv} 分别表示上述三类损失的权重系数。

1.2.2　基于多模态特征的多级对抗学习方法

由于自适应子网 G^{HVF} 和 G^{HVF}+G^{SSF} 距离网络 G 输出端较远，低层和中层子网的参数无法通过单级对抗训练被充分调优，从而影响对应层次所学特征的分布对齐。因此，本文基于单级对抗学习，面向 3 层结构网络 G 提出基于多模态特征的多级对抗学习方法，通过判别器 D 中 3 路分支 D^{HVF}、D^{SSF} 和 D^{MHF} 与网络 G 中各子网 G^{HVF}、G^{HVF}+G^{SSF} 和 G^{HVF}+G^{SSF}+G^{MHF} 分别进行单级对抗训练，从而全面减小领域间所学视觉、空间以及语义等三类特征的分布差异。

为使单级领域自适应损失适用于基于多模态特征的多级对抗学习，扩展网络 G 的目标函数 l_G 为多级领域自适应损失，即

$$
\begin{aligned}
&l_G(\boldsymbol{I}_S,\boldsymbol{Y}_S,\boldsymbol{I}_T,\boldsymbol{F})=\sum_{i\in\boldsymbol{F}}l_G^i(\boldsymbol{I}_S,\boldsymbol{Y}_S,\boldsymbol{I}_T)=\\
&\sum_{i\in\boldsymbol{F}}\left[\lambda_{\text{seg}}^i l_{\text{seg}}^i(\boldsymbol{I}_S,\boldsymbol{Y}_S)+\lambda_{\text{ssl}}^i l_{\text{ssl}}^i(\boldsymbol{I}_T)+\lambda_{\text{adv}}^i l_{\text{adv}}^i(\boldsymbol{I}_T)\right]
\end{aligned}
\tag{14}
$$

其中，$\boldsymbol{F}$ = {HVF, SSF, MHF}表示网络 G 所学三类特征的集合；i 表示子网层次，当 i = HVF 时表示低层子网 G^{HVF}，当 i = SSF 时表示中层子网 G^{HVF}+G^{SSF}，当 i = MHF 时表示整个网络 G^{HVF}+G^{SSF}+G^{MHF}；$l_{\mathrm{seg}}{}^{i}$、$l_{\mathrm{ssl}}{}^{i}$ 和 $l_{\mathrm{adv}}{}^{i}$ 分别表示第 i 层子网的多分类交叉熵损失、自监督学习损失和对抗学习损失；$\lambda_{\mathrm{seg}}{}^{i}$、$\lambda_{\mathrm{ssl}}{}^{i}$ 和 $\lambda_{\mathrm{adv}}{}^{i}$ 分别表示第 i 层次子网的三类损失对应的权重系数。

同时，为与网络 G 中各子网进行对抗，在判别器 D 中设置 3 路对抗分支，并扩展判别器 D 的目标函数 l_D 为

$$l_D(\boldsymbol{I}_S,\boldsymbol{I}_T,z,\boldsymbol{F})=\sum_{i\in\boldsymbol{F}} l_D^i(\boldsymbol{I}_S,\boldsymbol{I}_T,z) \tag{15}$$

其中，i 表示判别器分支的层次，当 i = HVF 时表示低层子网的对抗分支 D^{HVF}，当 i = SSF 时表示中层子网的对抗分支 D^{SSF}，当 i = MHF 时表示整个网络的对抗分支 D^{MHF}。

为了清晰地说明语义分割网络 G 的参数调优，UDAMAN-MF 的多级对抗学习过程介绍如下。首先，基于多分类交叉熵损失 l_{seg}，使用有标签源域 S 对网络 G 迭代训练 100 次 epoch，从而初始化网络 G 的参数。然后，为保证网络 G 的参数在后续的多级对抗训练中较快收敛，对网络 G 和判别器分支 D^{MHF} 迭代单级对抗训练 200 次 epoch，在每次迭代中，一方面，基于二分类交叉熵损失 $l_D{}^{\mathrm{MHF}}$ 训练对抗分支 D^{MHF} 区分源域语义分类概率 $G(\boldsymbol{I}_S)$和目标域语义分类概率 $G(\boldsymbol{I}_T)$的能力；另一方面，基于单级领域自适应损失 $l_G{}^{\mathrm{MHF}}$ 训练网络 G，使网络 G 输出的目标域语义分类概率 $G(\boldsymbol{I}_T)$不断接近源域语义分类概率 $G(\boldsymbol{I}_S)$的分布。最后，对网络 G 中 3 个子网和判别器 D 中 3 路分支迭代多级对抗训练 200 次 epoch，在每次迭代中，自适应子网 G^{HVF}、G^{HVF}+G^{SSF} 和 G^{HVF}+G^{SSF}+G^{MHF} 分别与对应的判别器分支 D^{HVF}、D^{SSF} 和 D^{MHF} 依次进行单级对抗训练，从而逐步调优网络 G 中各子网的参数，进而全面对齐领域间所学三类特征的分布。

2 实验

2.1 训练数据集和性能评价标准

为了验证 UDAMAN-MF 的普适性，分别在室外和室内场景数据集上对所提网络进行训练和测试。

在室外场景的领域自适应中，选择合成的 GTA5[22]或 SYNTHIA 数据集[23]作为源域，同时选择真实的 Cityscapes 数据集[24]作为目标域。在训练阶段，使用有标签的 GTA5（SYNTHIA）数据集和无标签的 Cityscapes 训练数据集进行多级对抗训练；在测试阶段，使用 Cityscapes 验证数据集进行测试。

在从源域 SUN-RGBD 数据集[25]到目标域 NYUD-v2 数据集[26]的室内场景领域自适应中，由于 SUN-RGBD 数据集由 NYUD-v2、Berkeley B3BO、SUN3D 以及新制作的数据四部分组成，为满足领域间的差异性，选择去除 NYUD-v2 的 SUN-RGBD 数据集作为源域。在训练阶段，使用有标签的 SUN-RGBD 训练数据集和无标签的 NYUD-v2 训练数据集进行多级对抗训练，在测试阶段，使用 NYUD-v2 测试数据集进行验证。

另外，使用像素准确率（PA, pixel accuracy）、平均准确率（MA, mean accuracy）和平均交并比（mIoU, mean intersection over union）作为面向语义分割领域自适应网络的性能评价标准[1,13]。

2.2 实验环境和参数设置

基于开源的深度学习框架 PyTorch[27]编码实现 UDAMAN-MF，并在一台 2 个 2.4 GHz Intel Xeon Silver 4214R CPU（2×12 Cores）、24 GB NVIDIA GeForce GTX 3090 GPU 以及 128 GB 内存的计算机上进行训练和测试。

2.2.1 判别器的结构

判别器 D 由 3 路对抗分支 D^{HVF}、D^{SSF} 和 D^{MHF} 组成，每路分支均由 5 层核尺寸为 4×4、步长为 2 的卷积操作构成，各卷积层后均设置一个 leaky ReLU 激活函数，各卷积层输出特征的维数分别为 64、128、256、512 和 1。另外，为使判别器输出的分布分类概率图与输入图像尺寸相同，在最后一层卷积操作后添加一个上采样操作。

2.2.2 UDAMAN-MF 的训练

UDAMAN-MF 的训练共包括 3 个阶段：首先，基于多分类交叉熵损失训练网络 G；然后，对网络 G 和判别器 D 中的分支 D^{MHF} 进行单级对抗训练；最后，对判别器 D 中 3 路分支与网络 G 中 3 个子网进行多级对抗训练。

训练阶段 1，通过反向传播算法对网络 G 中各层联合优化。在特征提取层，首先，通过公用模型 resnet_v1_101_2016_08_28[5]初始化该层的参数；然后，上采样并级联 ResNet-101 各层输出特征，各层输出特征的维数分别为 64、256、512、1 024 和 2 048；

最后，将级联特征送入 3 层 1×1 卷积做降维，各层输出特征的维数分别为 2 048、1 024 和 512。在结构化学习层，首先，为 4 路遍历分支均设置 5 个 LSTM 单元，并设置各 LSTM 单元隐藏层状态的维数分别为 512、256、128、256 和 512；然后，在[–0.05, 0.05]的均匀分布下随机地初始化 4 路分支的参数。在特征融合层，首先通过 3 层 1×1 卷积将多维视觉特征依次与 4 个方向上的空间结构化特征做降维，各层输出特征的维数分别为 512、256 和 256；然后，利用 softmax 函数分别计算 4 个方向上降维特征的注意力权重，并将 4 个方向上的降维特征加权求和；最后，根据自适应聚合的多模态混合特征逐像素地标注语义类别；另外，在均值为 0、标准差为 0.05 的正态分布下初始化该层的参数。在完成 G 的网络参数设置后，设置 G 的训练参数如下：learning_rate = 5×10^{-4}、batch_size = 8、momentum = 0.9、weight_decay = 10^{-4} 以及 epoch = 100，并采用随机梯度下降算法优化 G 的参数。

训练阶段 2，对网络 G 和对抗分支 D^{MHF} 进行单级对抗训练。在每次迭代中，首先，基于多分类交叉熵损失微调网络 G；然后，基于二分类交叉熵损失训练对抗分支 D^{MHF}；接着，为目标域生成标签，并基于自监督学习损失优化网络 G；最后，固定对抗分支 D^{MHF} 的参数，并基于对抗学习损失调优网络 G。

训练阶段 3，对网络 G 和判别器 D 进行多级对抗训练，即判别器 D 中的 3 路分支 D^{HVF}、D^{SSF} 和 D^{MHF} 依次与网络 G 中的 3 个子网 G^{HVF}、G^{HVF}+G^{SSF} 和 G^{HVF}+G^{SSF}+G^{MHF} 进行重复的单级对抗训练。

训练阶段 2 和训练阶段 3，设置网络 G 的训练参数如下：optimizer(G) = SGD，learning_rate = 2.5×10^{-4}，batch_size = 4，decay_policy = Poly，decay_power = 0.9，momentum = 0.9，weight_decay = 5×10^{-4} 以及 epoch = 200。同时，在均值为 0、标准差为 0.05 的正态分布下初始化判别器 D 中各分支的网络参数，并设置其训练参数如下：optimizer(D) = SGD，learning_rate = 10^{-4}，batch_size = 4，momentum = 0.9，weight_decay = 5×10^{-4} 以及 epoch = 200。

在自监督学习损失的阈值设置中，为 3 个自适应子网对应的损失设置相同的阈值，其中，分布置信度阈值 T_D = 0.6，语义置信度阈值 T_G = 0.7。在多级领域自适应损失的权重系数设置中，为距离网络 G 输出端较远的子网对应的损失设置较小的权重系数，权重系数分别设置如下：$\lambda_{\mathrm{seg}}^{\mathrm{HVF}}$ = 0.3，$\lambda_{\mathrm{ssl}}^{\mathrm{HVF}}$ = 0.3，$\lambda_{\mathrm{adv}}^{\mathrm{HVF}}$ = 0.000 2，$\lambda_{\mathrm{seg}}^{\mathrm{SSF}}$ = 0.5，$\lambda_{\mathrm{ssl}}^{\mathrm{SSF}}$ = 0.5，$\lambda_{\mathrm{adv}}^{\mathrm{SSF}}$ = 0.000 6，$\lambda_{\mathrm{seg}}^{\mathrm{MHF}}$ = 1，$\lambda_{\mathrm{ssl}}^{\mathrm{MHF}}$ = 1 和 $\lambda_{\mathrm{adv}}^{\mathrm{MHF}}$ = 0.001。

2.3　实验结果与分析

2.3.1　GTA5 到 Cityscapes 的领域自适应

1) UDAMAN-MF 与先进方法的分割精度对比

源域 GTA5 到目标域 Cityscapes 上 UDAMAN-MF（基于 3 层结构语义分割网络）与先进方法（基于 ResNet-101）的训练方法与分割精度如表 1 所示，训练方法中，A 表示对抗学习方法，S 表示自监督学习方法，T 表示图像风格转换方法；mIoU 表示分割精度的评价标准。从总体上讲，所提网络取得最优的平均交并比 62.2%；从相兼容 19 种类别上看，相比其他方法，UDAMAN-MF 在 11 种类别上的交并比均有一定程度的提升。特别地，所提网络不仅在尺寸较小的类别上（如围栏、杆、信号灯和交通标识等）取得最优的交并比，而且在易混淆类别上（如行人和骑手、摩托车和自行车等）也取得更高的分割精度。UDAMAN-MF 获取成功的原因可归纳如下：第一，3 层结构语义分割网络不仅能够有效提取局部对象的多维视觉特征，而且可以准确推理全局场景的空间结构化特征，融合生成的多模态混合特征能够全面表达对象的综合语义；第二，改进的自监督学习方法能够确保选定目标域子集的稠密性和所生成标签的正确性，从而能够基于多分类交叉熵损失直接对齐接近源域的有标签目标域子集的分布，同时有助于对抗学习拉近远离源域的无标签目标域子集与源域间的分布差异，进而实现目标域中更多像素的分布对齐；第三，多级对抗学习方法能够充分调优 3 层网络中各子网的参数，从而全面减小领域间所学视觉、空间以及语义特征的分布差异，进而在目标域场景中生成融合对象视觉和空间信息的综合语义特征。

UDAMAN-MF 虽然在大多数类别上均取得较优的交并比，但是在少数类别上（如地面、卡车和火车等）的分割精度却较低，如表 1 所示。为分析分割精度不够理想的原因，图 4 列出语义分割混淆矩阵，其中，对角线表示各类别的像素准确率，非对角线表示行类别误预测为列类别的概率。从混淆矩阵中可发现：第一，“地面”易误分类为“马路”和“人行道”，这主要由于上述类别在外观或属性上存在较高的相似度；第二，“卡车”易误分类为

表 1　源域 GTA5 到目标域 Cityscapes 上 UDAMAN-MF 与先进方法的训练方法与分割精度

方法	训练方法	mIoU									
		马路	人行道	建筑	墙	围栏	杆	信号灯	交通标识	植物	地面
BDL[13]	AST	91.0%	44.7%	84.2%	34.6%	27.6%	30.2%	36.0%	36.0%	85.0%	43.6%
CDA[20]	A	91.3%	46.0%	84.5%	34.4%	29.7%	32.6%	35.8%	36.4%	84.5%	43.2%
TPLD[17]	AS	94.2%	60.5%	82.8%	36.6%	16.6%	39.3%	29.0%	25.5%	85.6%	44.9%
MetaCorrection[28]	S	92.8%	58.1%	86.2%	39.7%	33.1%	36.3%	42.0%	38.6%	85.5%	37.8%
ISR[12]	ST	93.0%	54.0%	86.6%	42.6%	34.7%	35.9%	40.8%	43.3%	86.0%	43.2%
DPL[15]	AST	92.8%	54.4%	86.2%	41.6%	32.7%	36.4%	49.0%	34.0%	85.8%	41.3%
SAC[29]	S	90.4%	53.9%	86.6%	42.4%	27.3%	45.1%	48.5%	42.7%	87.4%	40.1%
ProCA[30]	S	91.9%	48.4%	87.3%	41.5%	31.8%	41.9%	47.9%	36.7%	86.5%	42.3%
UCDA[31]	ST	92.6%	59.1%	88.5%	45.8%	40.5%	52.9%	53.6%	54.1%	88.0%	41.9%
UDAMAN-MF	AS	93.8%	61.7%	87.1%	48.9%	44.0%	54.4%	55.2%	56.5%	87.1%	43.3%

方法	训练方法	mIoU									
		天空	行人	骑手	汽车	卡车	公交车	火车	摩托车	自行车	平均
BDL[13]	AST	83.0%	58.6%	31.6%	83.3%	35.3%	49.7%	3.3%	28.8%	35.6%	48.5%
CDA[20]	A	83.0%	60.0%	32.2%	83.2%	35.0%	46.7%	0.0%	33.7%	42.2%	49.2%
TPLD[17]	AS	84.4%	60.6%	27.4%	84.1%	37.0%	47.0%	31.2%	36.1%	50.3%	51.2%
MetaCorrection[28]	S	87.6%	62.8%	31.7%	84.8%	35.7%	50.3%	2.0%	36.8%	48.0%	52.1%
ISR[12]	ST	85.4%	61.5%	34.4%	83.7%	29.2%	50.1%	4.0%	36.5%	50.9%	52.4%
DPL[15]	AST	86.0%	63.2%	34.2%	87.2%	39.3%	44.5%	18.7%	42.6%	43.1%	53.3%
SAC[29]	S	86.1%	67.5%	29.7%	88.5%	49.1%	54.6%	9.8%	26.6%	45.3%	53.8%
ProCA[30]	S	84.7%	68.4%	43.1%	88.1%	39.6%	48.8%	40.6%	43.6%	56.9%	56.3%
UCDA[31]	ST	86.0%	73.5%	44.1%	89.7%	39.3%	53.2%	26.8%	51.6%	61.8%	60.2%
UDAMAN-MF	AS	88.7%	76.2%	46.3%	88.1%	45.8%	53.9%	33.2%	54.1%	63.4%	62.2%

“汽车”，“火车”和“公交车”易相互混淆，除了视觉相似度较高外，主要与“卡车”和“火车”类别的出现频率较低有关，从而影响上述类别的充分学习。

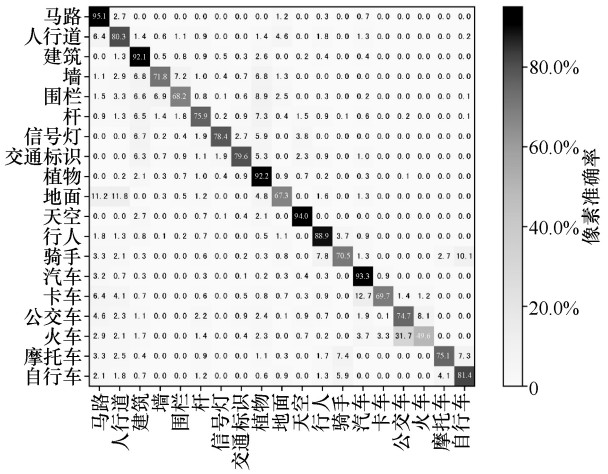


图 4　语义分割混淆矩阵

2) UDAMAN-MF 的消融学习

为研究 3 层结构语义分割网络、改进自监督学习方法以及多级对抗学习方法对于 UDAMAN-MF 的性能影响，在源域 GTA5 到目标域 Cityscapes 的领域自适应上进行消融学习，结果如表 2 所示。

表 2　源域 GTA5 到目标域 Cityscapes 上的消融学习结果

方法	mIoU
G^{HVF} (不包含自监督学习的单级对抗训练)	45.9%
G^{HVF}+G^{SSF} (不包含自监督学习的单级对抗训练)	49.2%
G^{HVF}+G^{SSF}+G^{MHF} (不包含自监督学习的单级对抗训练)	50.4%
G^{HVF}+G^{SSF}+G^{MHF} (包含改进自监督学习的单级对抗训练)	57.5%
UDAMAN-MF (包含改进自监督学习的多级对抗训练)	**62.2%**

首先，分别对低层子网 G^{HVF}、中层子网 G^{HVF}+G^{SSF} 和整个网络 G^{HVF}+G^{SSF}+G^{MHF} 进行不包含自监督学习的单级对抗训练。当结构化学习层 G^{SSF} 添加到特征提取层 G^{HVF} 末端时，mIoU 从

45.9%提升到 49.2%；当特征融合层 G^{MHF} 添加到结构化学习层 G^{SSF} 末端时，mIoU 从 49.2%提升到 50.4%。分割精度提升的原因如下。第一，结构化学习层能够准确推理对象邻近 4 个区域的全局上下文信息，同时，单级对抗学习可以减小领域间所学空间特征的分布差异，从而使网络 G 更加准确地推理目标域场景的全局上下文信息，进而能够基于空间结构化特征调优分类结果和避免分类错误。例如，虽然“行人”和“骑手”具有相似的视觉外观，但是能够根据邻近场景的空间结构化特性区分上述易混淆类别。第二，单级对抗训练的特征融合层能够基于注意力机制实现目标域场景中视觉和空间特征的有机融合，即根据对象邻近 4 个区域的全局上下文信息与其自身视觉信息的相关性进行加权求和，从而自适应聚合有用上下文信息和避免噪声上下文信息，进而显著提升目标域场景所学多模态混合特征的质量。例如，对于“杆”“信号灯”和“交通标识”等尺寸较小的类别，基于注意力机制的自适应聚合可屏蔽背景噪声的引入，以避免尺寸较小对象的视觉信息遭到破坏。

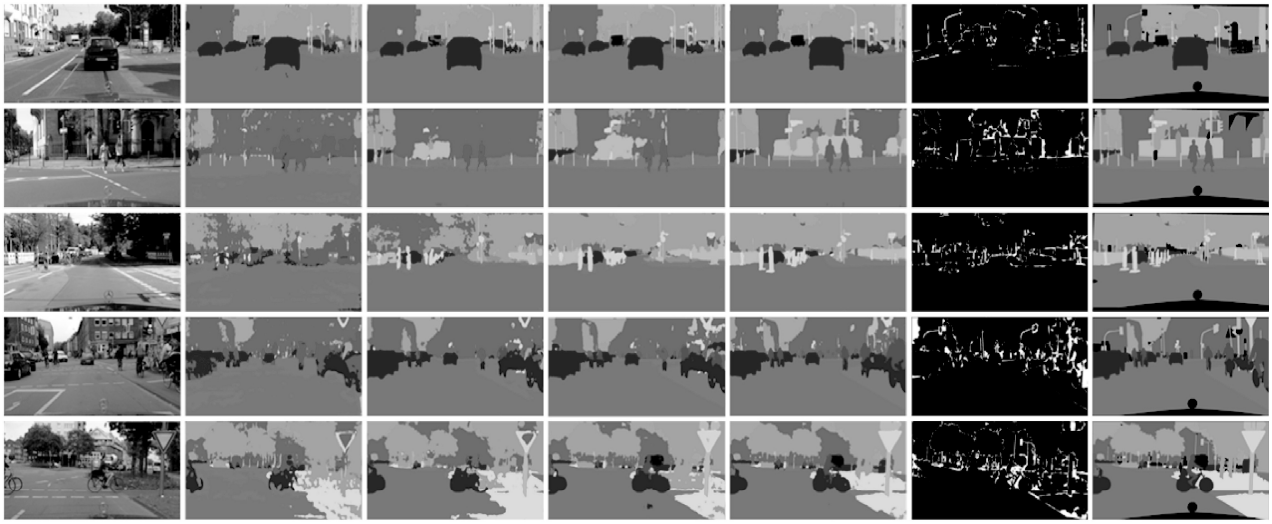
然后，对整个网络 G^{HVF}+G^{SSF}+G^{MHF} 分别进行不包含和包含改进自监督学习的单级对抗训练，实验结果表明包含自监督学习的对抗训练使分割精度提升了 7.1%。这说明联合分布置信度和语义置信度的自监督学习在减小领域间分布差异上起到重要作用，该方法为分布接近源域并且语义分类概率较高的目标域子集生成标签，从而可以直接对齐选中目标域子集的分布，并大幅减小尚未对齐的目标域子集尺寸，进而有助于对抗学习对齐远离源域的无标签目标域子集的分布，以实现更大目标域子集的分布对齐。

最后，对整个网络分别进行单级和多级对抗训练，实验结果表明，与单级对抗训练相比，多级对抗训练使分割精度提升了 4.7%。这说明多级对抗训练能够充分调优距离网络 G 输出端较远的低层子网 G^{HVF} 和中层子网 G^{HVF}+G^{SSF} 的参数，从而全面减小领域间所学视觉、空间以及语义特征的分布差异，进而有效学习上述三类特征的域间不变表示，融合生成的多模态混合特征能够更准确地表达对象的综合语义特征。

3) UDAMAN-MF 的语义分割视觉效果

源域 GTA5 到目标域 Cityscapes 领域自适应的分割视觉效果如图 5 所示。从图 5 可发现，首先，与图 5(b)相比，图 5(c)的误分类像素大量减少，从而证明 3 层结构网络有效提取多维视觉特征、显式推理空间结构化特征以及自适应融合多模态混合特征的能力；然后，与图 5(c)相比，图 5(d)对象轮廓更平滑，从而证明对抗学习联合改进自监督学习具有充分对齐目标域中各像素特征分布的能力；最后，与图 5(d)相比，图 5(e)中形状复杂对象的分割轮廓更精细，从而证明多级对抗学习具有全面拉近领域间视觉、空间以及语义等三类特征分布差异的能力。

4) UDAMAN-MF 与先进方法的稳健性对比



(a) 目标域输入图像　(b) 单级对抗训练G^{HVF}（不包含自监督学习）　(c) 单级对抗训练G（不包含自监督学习）　(d) 单级对抗训练G（包含自监督学习）　(e) 多级对抗训练G（UDAMAN-MF）　(f) 图5(e)与真值标签间误差　(g) 真值标签

图 5　源域 GTA5 到目标域 Cityscapes 领域自适应的分割视觉效果

为测试 UDAMAN-MF 的稳健性，使用基于快速梯度标签算法生成的 3 组扰动测试数据集评估所提网络抗噪声攻击的能力，3 组数据集分别由干净的 Cityscapes 验证数据集和不同扰动幅度 ε（分别为 0.1、0.25 和 0.5）的噪声输入 PSPNet（pyramid scene parsing network）生成[32]。源域 GTA5 到目标域 Cityscapes 的抗攻击能力如表 3 所示，其中，mIoU 表示扰动测试数据集上的平均交并比，mIoU* 表示干净测试数据集上的平均交并比，mIoU drop 表示 mIoU 较 mIoU*的下降值，若 mIoU 越高且 mIoU drop 越低，则表明网络的稳健性越强。

表 3 源域 GTA5 到目标域 Cityscapes 的抗攻击能力

方法	ε	抗攻击训练	mIoU	mIoU drop	mIoU*
BDL[13]		—	36.2%	12.3%	48.5%
DPL[15]		—	41.6%	11.7%	53.3%
ProCA[30]	0.1	—	39.9%	16.4%	56.3%
ASSUDA[32]		✓	**43.3**%	**0.6**%	43.9%
UDAMAN-MF		—	**52.0**%	**10.2**%	62.2%
BDL[13]		—	19.9%	28.6%	48.5%
DPL[15]		—	26.4%	26.9%	53.3%
ProCA[30]	0.25	—	25.1%	31.2%	56.3%
ASSUDA[32]		✓	**39.0**%	**4.9**%	43.9%
UDAMAN-MF		—	**37.7**%	**24.5**%	62.2%
BDL[13]		—	6.5%	42.0%	48.5%
DPL[15]		—	12.4%	40.9%	53.3%
ProCA[30]	0.5	—	11.6%	44.7%	56.3%
ASSUDA[32]		✓	**27.4**%	**16.5**%	43.9%
UDAMAN-MF		—	**23.6**%	**38.6**%	62.2%

从表 3 中可发现，第一，与没有抗攻击训练的方法相比，UDAMAN-MF 不仅在 3 组扰动数据集上的 mIoU 均最优，而且 mIoU drop 也均最低，这说明所提网络具有较强的稳健性，从而证明全面减小领域间所学视觉、空间以及语义等三类特征的分布差异能够有效对抗噪声扰动的攻击，进而降低噪声扰动对所生成多模态混合特征质量的影响；第二，较抗攻击训练的 ASSUDA（adversarial self-supervision unsupervised domain adaptation）网络[32]，虽然所提网络在 3 组数据集上的 mIoU 优于或接近 ASSUDA，但是 mIoU drop 却均高于 ASSUDA，这说明所提网络的稳健性逊于 ASSUDA，导致稳健性不强的原因在于噪声扰动会再次拉大领域间的特征分布差异，从而破坏所学特征的域间不变表示。

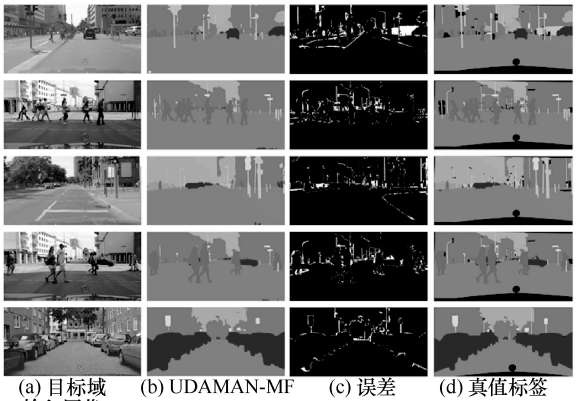
2.3.2 SYNTHIA 到 Cityscapes 的领域自适应

源域 SYNTHIA 到目标域 Cityscapes 的分割精度如表 4 所示，mIoU 为分割精度的评价标准。在相兼容 13 种领域自适应中，UDAMAN-MF 在总体上将分割精度从 63.1%提升到 66.9%；同时，在相兼容 16 种领域自适应中，UDAMAN-MF 在总体上也取得最优的平均交并比 58.8%。

表 4 源域 SYNTHIA 到目标域 Cityscapes 的分割精度

方法	训练方法	mIoU(13)	mIoU*(16)
BDL[13]	AST	51.4%	—
LDR[14]	AST	53.1%	—
DPL[15]	AST	54.2%	47.0%
TPLD[17]	AS	55.7%	48.1%
ISR[12]	ST	57.1%	49.0%
ProCA[30]	S	59.6%	53.0%
UCDA[31]	ST	63.1%	56.5%
UDAMAN-MF	AS	**66.9**%	**58.8**%

但是，对于墙、围栏以及杆等类别，UDAMAN-MF 分割效果不够理想，如图 6 所示。这主要由于上述类别出现的频率不高，从而影响其有效学习。此外，由于源域 SYNTHIA 和目标域 Cityscapes 间存在较大的视觉风格差异（如光照和物体纹理），从而在一定程度上影响上述类别的分布差异拉近。因此，后续工作将在多级对抗学习中引入多级图像风格转换方法，从而尽量降低视觉风格差异对特征分布对齐的影响。



(a) 目标域输入图像 (b) UDAMAN-MF (c) 误差 (d) 真值标签

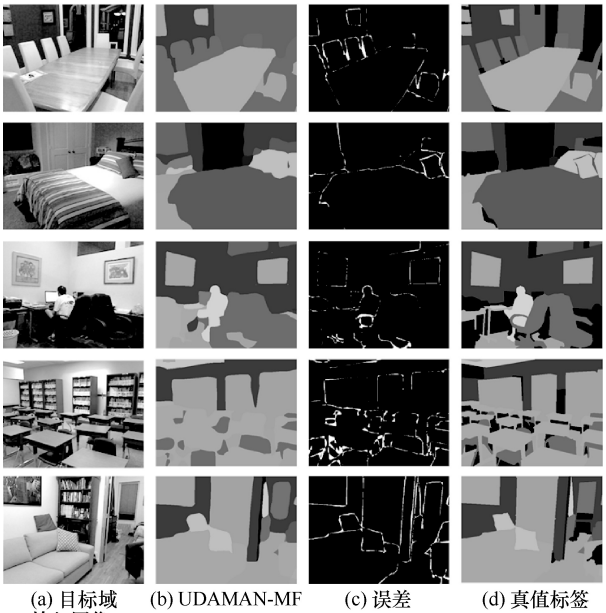
图 6 源域 SYNTHIA 到目标域 Cityscapes 领域自适应的分割视觉效果

2.3.3 SUN-RGBD 到 NYUD-v2 的领域自适应

为验证所提网络的普适性，本节在源域SUN-RGBD到目标域NYUD-v2的室内场景进行领域自适应学习，分割精度如表 5 所示，像素准确率、平均准确率以及平均交并比为分割精度的评价标准。虽然相兼容类别多达 37 种，但是所提网络在上述评价标准上仍取得最优的成绩，其 PA、MA 和 mIoU 分别为 84.9%、74.6%和 59.7%，较当前先进方法中的最优结果，上述 3 项评价标准分别提升了 3.6%、4.9%和 4.5%。源域 SUN-RGBD 到目标域 NYUD-v2 领域的自适应分割视觉效果如图 7 所示。从图 7 可以看出，所提网络不仅能够准确地分割易识别类别（如墙面、地面、床、沙发和天花板等），而且可以较理想地解析形状复杂类别（如椅子和人等）。所提网络在室内场景领域自适应中取得的成绩主要归功于精心设计的3层结构语义分割网络和基于改进自监督学习的多级对抗学习方法，从而能够在目标域场景有效地生成融合对象视觉和空间信息的综合语义特征。

表 5　源域 SUN-RGBD 到目标域 NYUD-v2 的分割精度

方法	训练方法	PA	MA	mIoU
UIA[18]	AS	71.4%	57.8%	43.6%
BDL[13]	AST	74.2%	61.0%	46.3%
DPL[15]	AST	77.5%	63.2%	49.8%
SAC[29]	S	79.6%	66.1%	52.7%
ProCA[30]	S	81.3%	69.7%	55.2%
UDAMAN-MF	AS	**84.9%**	**74.6%**	**59.7%**



(a) 目标域输入图像　(b) UDAMAN-MF　(c) 误差　(d) 真值标签

图 7　源域 SUN-RGBD 到目标域 NYUD-v2 领域自适应的分割视觉效果

3　结束语

本文面向语义分割提出基于多模态特征的无监督领域自适应多级对抗语义分割网络。首先，所提 3 层结构的语义分割网络能够分别从两域学习视觉、空间以及语义特征，从而为特征分布对齐奠定基础。然后，改进自监督学习能够确保选定目标域子集的稠密性和所生成标签的正确性，从而可以直接对齐有标签目标域子集的分布，与对抗学习相互结合，能够实现更大目标域子集的对齐。最后，多级对抗学习对 3 路对抗分支与 3 个子网分别进行单级对抗训练，从而有效学习各子网输出特征的域间不变表示。实验结果表明，在室外和室内场景的 3 个数据集上，UDAMAN-MF 均取得最优的分割精度，证明其不仅具有全面对齐领域间视觉、空间以及语义特征分布的能力，而且具有良好的普适性。但是，当目标域数据遭受噪声扰动攻击时，所提网络无法理想地对齐特征分布，因此，后续工作将在多级对抗学习中引入抗攻击训练，从而提升网络的稳健性，以满足机器人任务规划和车辆自动驾驶等智能视觉任务对安全性的要求。

参考文献：

[1] 徐英姿, 刘原, 时梦然, 等. 语义在通信中的应用综述[J]. 电信科学, 2022, 38(Z1): 43-59.
XU Y Z, LIU Y, SHI M R, et al. A survey of semantic applications in communications[J]. Telecommunications Science, 2022, 38(Z1): 43-59.

[2] AGIA C, JATAVALLABHULA K M, KHODEIR M, et al. Taskography: evaluating robot task planning over large 3D scene graphs[C]//Proceedings of Conference on Robot Learning. Cambridge: JMLR, 2022: 46-58.

[3] CAESAR H, BANKITI V, LANG A H, et al. nuScenes: a multimodal dataset for autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 11621-11631.

[4] YU C, LIU Z X, LIU X J, et al. DS-SLAM: a semantic visual SLAM towards dynamic environments[C]//Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE Press, 2018: 1168-1174.

[5] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[6] LI Z, GAN Y, LIANG X, et al. LSTM-CF: unifying context modeling and fusion with LSTMS for RGB-D scene labeling[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2016: 541-557.

[7] YUAN Y H, CHEN X L, WANG J D. Object-contextual representations for semantic segmentation[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2020: 173-190.

[8] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv Preprint, arXiv: 1511.06434, 2015.

[9] HOFFMAN J, WANG D, YU F, et al. FCNs in the wild: Pixel-level adversarial and constraint-based adaptation[J]. arXiv Preprint, arXiv: 1612.02649, 2016.

[10] TSAI Y H, HUNG W C, SCHULTER S, et al. Learning to adapt structured output space for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7472-7481.

[11] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2223-2232.

[12] LI Z Y, TOGO R, OGAWA T, et al. Learning intra-domain style-invariant representation for unsupervised domain adaptation of semantic segmentation[J]. Pattern Recognition, 2022, 132(12): 108911.

[13] LI Y S, YUAN L, VASCONCELOS N. Bidirectional learning for domain adaptation of semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 6936-6945.

[14] YANG J, AN W, WANG S, et al. Label-driven reconstruction for domain adaptation in semantic segmentation[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2020: 480-498.

[15] CHENG Y, WEI F, BAO J, et al. Dual path learning for domain adaptation of semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 9082-9091.

[16] LEE S, HYUN J, SEONG H, et al. Unsupervised domain adaptation for semantic segmentation by content transfer[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 8306-8315.

[17] SHIN I, WOO S, PAN F, et al. Two-phase pseudo label densification for self-training based domain adaptation[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2020: 532-548.

[18] PAN F, SHIN I, RAMEAU F, et al. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 3764-3773.

[19] PENG C L, MA J Y. Domain adaptive semantic segmentation via entropy-ranking and uncertain learning-based self-training[J]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(8): 1524-1527.

[20] YANG J, AN W, YAN C, et al. Context-aware domain adaptation in semantic segmentation[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2021: 514-524.

[21] HUANG J, LU S, GUAN D, et al. Contextual-relation consistent domain adaptation for semantic segmentation[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2020: 705-722.

[22] RICHTER S R, VINEET V, ROTH S, et al. Playing for data: ground truth from computer games[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2016: 102-118.

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

[24] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 3213-3223.

[25] SONG S R, LICHTENBERG S P, XIAO J X. SUN RGB-D: a RGB-D scene understanding benchmark suite[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 567-576.

[26] SILBERMAN N, HOIEM D, KOHLI P, et al. Indoor segmentation and support inference from RGBD images[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2012: 746-760.

[27] PASZKE A, GROSS S, MASSA F, et al. PyTorch: an imperative style, high-performance deep learning library[J]. Advances in Neural Information Processing Systems, 2019, 32(12): 8024-8035.

[28] GUO X Q, YANG C, LI B P, et al. MetaCorrection: domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 3926-3935.

[29] ARASLANOV N, ROTH S. Self-supervised augmentation consistency for adapting semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 15384-15394.

[30] JIANG Z K, LI Y X, YANG C Y, et al. Prototypical contrast adaptation for domain adaptive semantic segmentation[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer, 2022: 36-54.

[31] ZHANG F, KOLTUN V, TORR P, et al. Unsupervised contrastive domain adaptation for semantic segmentation[J]. arXiv Preprint, arXiv:

2204.08399, 2022.

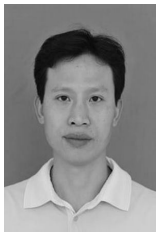
[32] YANG J Y, LI C Y, AN W Z, et al. Exploring robustness of unsupervised domain adaptation in semantic segmentation[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 9174-9183.

[作者简介]

王泽宇（1989− ），男，河南郑州人，博士，郑州轻工业大学讲师，主要研究方向为计算机视觉、图像处理、深度学习等。

布树辉（1978− ），男，河南洛阳人，博士，西北工业大学教授、博士生导师，主要研究方向为计算机视觉、图像处理、机器学习等。

黄伟（1982− ），男，河南郑州人，博士，郑州轻工业大学副教授、硕士生导师，主要研究方向为遥感图像处理、深度学习等。



郑远攀（1983− ），男，河南郑州人，博士，郑州轻工业大学副教授、硕士生导师，主要研究方向为图像处理、智慧应急等。

吴庆岗（1984− ），男，河南濮阳人，博士，郑州轻工业大学副教授、硕士生导师，主要研究方向为计算机视觉、遥感图像处理、深度学习等。

常化文（1980− ），男，河南郑州人，博士，郑州轻工业大学讲师、硕士生导师，主要研究方向为图像质量评价、计算机视觉等。

张旭（1979− ），女，河南南阳人，郑州轻工业大学讲师，主要研究方向为图像处理、模型检测等。