无监督领域自适应语义分割（UDA-SS）旨在训练一个模型在有标签的源领域（例如，合成数据）上，并在外观和分布发生显著变化的情况下在无标签的目标领域（例如，真实世界数据）中进行部署。这对于机器人视觉非常重要，因为这些模型在策划或模拟的数据集上训练过，必须在没有额外标注的情况下推广到新环境。UDA-SS 方法主要分为基于对抗学习和自训练（ST）的方法。ST 方法通过解决类别级别的不对齐问题和改善训练稳定性，表现出更好的性能。使用教师-学生框架，其中教师生成伪标签来监督学生，ST 提供了更健壮的解决方案。随着 DAFormer 的出现，UDA-SS 通过基于转换器的自监督学习取得了突破。在 HRDA 中这一突破得到了进一步扩展，该方法整合了低分辨率和高分辨率特征，现在成为最新 UDA-SS 方法的骨干。

许多现有基于 HRDA 的方法虽然结合了空间和上下文细节，并利用对比学习来增强分割性能，但它们常常忽略了有效整合全局上下文和细粒度局部细节这一关键问题。高分辨率特征可以捕获细粒度的结构，但缺乏更广泛的上下文和类别级别的语义，而低分辨率 logits 提供了强大的全局先验和语义一致性。然而，大多数方法忽视了将这些低分辨率 logits 融入到高分辨率特征中，导致分割输出在空间上精确但在上下文上不一致。此外，边界信息往往没有被有效利用，导致对象边界的界定不清晰或不准确。

为了解决这个问题，我们建议通过引入边界信息以及全局和局部细节来增强基于 HRDA 的 UDA 方法。因此，我们提出一个自适应特征优化（AFR）模块，以便与基于 HRDA 的方法集成。AFR 模块通过利用来自低分辨率 logits 的全局信息来优化高分辨率特征，丰富上下文细节，同时从原始多尺度高分辨率特征中捕捉细粒度的局部空间细节。此外，AFR 整合了从高分辨率特征和低分辨率 logits 中提取的高频成分，增强边缘一致性。它通过不确定性驱动的注意力动态抑制噪声预测，并使用双重注意图自适应平衡局部和全局信息，确保稳健的边界预测。

虽然受到之前注意机制和边界优化方法的启发，我们的 AFR 模块引入了一种针对领域自适应的整合。虽然其各个组成部分是基于现有的研究成果，但据我们所知，它们在轻量级、即插即用的双注意力模块中的协调整合在 UDA 文献中是独一无二的。与仅在特征空间中运作的现有方法不同，AFR 首次将语义 logits 直接整合到细化过程中。它将语义 logits、不确定性和高频分量置于设计核心，而不是将这些元素视为辅助提示，从而在不使用显式边界头的情况下，实现语义一致和边界感知的高分辨率细化。AFR 的设计使现有的训练流程保持不变，而是专注于特征层面的模块化细化。这使得 AFR 能够在基于 HRDA 的 UDA 框架中作为即插即用的组件运作，而不干扰训练动态。这种模块化对于真实世界中的机器人部署尤为重要，因为保持训练的稳定性和简易性至关重要。我们在五个具有挑战性的数据集上验证了我们的方法，这些数据集涵盖了城市和越野环境。在标准的 GTA V [1] → Cityscapes [2] 和 SYNTHIA [3] → Cityscapes 设置中，AFRDA 在分割质量和细节保留方面优于先前的方法。更重要的是，我们将评估扩展到从 RUGD [4] 到我们内部的森林数据集（MESH）的越野适应，AFRDA 在包含植被、不平地面和自然障碍物的非结构化地形中保持了强大的性能——这一领域在 UDA-SS 研究中常被忽视。为了展示真实世界的适用性，我们在机器人上部署了 AFRDA，该机器人在户外环境中导航，我们的模型能够实现对可通行地面和障碍物的准确、稳定感知，支持可靠的自主导航。

I. 相关工作

A. 无监督领域适应

无监督领域自适应（UDA）弥合了标注源域和未标注目标域之间的领域差距，并使语义分割模型能够适应多样化的环境。已经提出了许多策略来提高 UDA 性能。其中，基于统计距离函数的方法利用熵最小化 [5] ，或 Wasserstein 距离 [6] 来减轻领域差异。另一类 UDA 方法利用对抗学习范式 [7] ，通过学习到的领域判别器全局对齐两个领域的特征。由于对抗学习忽略了类别级别的对齐，因此不能消除类别级别的偏移，并受到负迁移问题的困扰。尽管一些方法利用类别级别特征，目标标签的缺乏常常导致性能较弱和训练不稳定。基于自我训练的方法 [8] 已成为一种有前途的解决方案，采用教师-学生框架，其中教师模型为目标域生成伪标签。然而，噪声伪标签常常降低了这些方法的有效性和性能。最近的方法探索多裁剪一致性 [9] 、上下文线索 [10] 、对比学习 [11] ，以及辅助精化网络 [12] ，以创建可靠的伪标签并改善 UDA。

语义分割模型经常受到空间和上下文不一致性的影响，导致分类错误，尤其是在物体边界附近。为了解决这个问题，研究人员探索了特征优化技术，以提高空间精度和上下文理解。一个被广泛采用的策略是多尺度特征优化。方法如特征金字塔网络（FPN）、空洞空间金字塔池化（ASPP）和 HRNet 通过整合来自多种分辨率的特征来捕捉细粒度的细节和全局上下文。另一种有效的优化策略是利用注意力机制，专注于最相关的特征。

尽管诸如 CBAM [13] 和 SE-Net [14] 的注意力模块通过动态优先考虑重要的空间和通道信息来细化特征，从而提高复杂场景中的分割精度，但它们没有结合语义 logits 或不确定性，而这正是我们 AFR 设计的核心。此外，边缘和边界引导的细化也被广泛探索以解决对象边缘周围的错误预测问题。像 Gated-SCNN [15] 和 STDC-Seg [16] 这样的技术通过使用额外的边界头来改善对象描绘，结合显式边界监督，而一些方法设计了边界感知损失函数，以提高对象边缘附近的特征质量。相比之下，我们的 AFR 模块通过高频残差隐式地捕获边界信息，从而无需额外的标签或专门的边界头。此外，最近的不确定性感知域适应方法如 UPA [17] 主要在无源时代通过像素级不确定性来过滤不可靠的伪标签，而不是像我们的 AFR 模块那样将不确定性集成到注意力引导的多分辨率特征细化中。我们的 AFR 模块独特地结合了全局语义 logits、增强边界的高频信号和双重不确定性图，用于自适应高分辨率特征细化，这在结构和功能上都与这些先前的方法不同。

B. 使用语义分割的视觉导航

语义分割对于视觉导航很重要，因为它将地形分类为可通行、不可通行、禁止等。传统的基于 RGB 的分割模型可以生成区分可通行和不可通行区域的 2D 地图 [18] 。然而，这些方法通常无法感知地表坡度和高度，尤其是在越野导航时，并且难以区分技术上可通行但可能不适合通行的地形类别（例如，小草与高草）。它们甚至可能陷入局部极小值。结合 LiDAR 点云的几何信息和分割地图，可以通过生成 2.5D 占据网格地图 [19] 来解决这些问题，让运动规划器创建语义感知的轨迹。最近，将本体感知（即线性和角速度、力、扭矩）与几何和视觉线索结合变得广受欢迎，因为它能够创建考虑表面颠簸和粗糙度的稳健 2D 和 3D 成本地图 [20] ，从而实现上下文感知导航。在我们提出的工作中，我们专注于利用语义分割来增强复杂环境中的视觉导航。

首先，在第 I-C 节中，我们提供了 UDA 方法的初步知识和我们所提出的框架 AFRDA 的概述。接着，在第 ?? 节中，我们介绍了注意力特征细化（AFR）模块，该模块增强了分割精度和边界稳定性。最后，在第 I-D 节中，我们讨论了用于在机器人车辆中部署我们框架的视觉规划。

C. UDA-SS 的预备知识

在 UDA 中，我们考虑一个源域 S，其中包含标记的图像（$X^S = \{x_i^S\}_{i=1}^{N_s}$，$Y^S = \{y_i^S\}_{i=1}^{N_s}$），以及一个目标域，该域仅包含原始图像 $X^T = \{x_i^T\}_{i=1}^{N_t}$，且没有任何真实标签 Y^T。一个由 θ 参数化的神经网络 h 在源域图像 X^S 上进行了训练，并在目标域图像 X^T 上进行了适配。然而，仅在源域数据上使用分类交叉熵损失

$$L_i^S = -\sum_{j=1}^{H\times W}\sum_{c=1}^{C} y_i^S \log h_\theta(x_i^S) \tag{1}$$

进行训练，其中 C 表示语义类别的数量，不能很好地推广到目标域图像，导致目标域预测效果不佳。为了应对这种领域偏移，最近的 UDA 方法如 DAFormer [8] 和 HRDA [9] 采用了自训练策略。这些方法利用教师模型 $h_{\bar{\theta}}$ 生成伪标签 $\bar{Y}^T = argmax\ h_{\bar{\theta}}(X^T)$，而不进行任何梯度的反向传播。教师模型 $h_{\bar{\theta}}$ 的权重基于学生模型 h_θ 的权重，在每次训练迭代 t 后使用指数移动平均（EMA）进行更新

$$\theta_{t+1} \leftarrow \alpha\theta_t + (1-\alpha)\theta_t. \tag{2}$$

生成的伪标签可能会有噪声，因此基于像素中最大 softmax 概率超过阈值 τ 的比例来生成质量估计 q^T。数学上，

稍后，学生模型 h_θ 使用伪标签及其质量估计再次在目标上进行训练，以优化目标域损失

然而，最近的一些方法如 DAFormer 加入了 ClassMix [21] 和其他混合技术 [22]，通过混合源域和目标域的图像、标签和伪标签生成混合对（$X^m = \{x_i^m\}_{i=1}^{N_m}$，$Y^m = \{y_i^m\}_{i=1}^{N_m}$）。学生模型并不是直接训练在伪标签上，而是在这些混合图像上进行训练，以减轻确认偏差，增强鲁棒性，并提高跨域的适应性。因此，目标域损失变为

这种方法提高了对目标领域的适应性。然而，由于仅使用低分辨率输入，它在准确分割小物体和保留细节方面仍然面临挑战，这限制了捕捉细粒度结构的能力。HRDA 通过引入一种多分辨率框架解决了这一问题，该框架结合了一个用于长距离依赖的大范围低分辨率（LR）上下文裁剪和一个用于精细分割的小范围高分辨率（HR）细节裁剪。

在 HRDA 框架中，我们采用自适应特征优化（AFR）模块来增强不同类别（见图 ??）的特征表示。AFR 模块通过利用低分辨率逻辑回归的全局信息来优化高分辨率特征，在捕捉原始多尺度高分辨率特征的细粒度空间细节的同时丰富上下文细节。此外，为了利用上下文线索，从目标图像中随机遮盖掉部分区域，在 UDA 方法中生成一组遮盖的目标图像 $X^M = \{x_i^M\}_{i=1}^{N_t}$。最后，通过目标域伪标签 $\bar{y}_i^T$ 监督，在遮盖的图像上训练学生模型以优化被遮盖的损失。

因此，结合 UDA 的基础和我们提出的 AFR 模块，我们框架的整体适应目标变为

为了识别一个对象，模型需要利用高分辨率和低分辨率表示。高分辨率特征捕获精细的空间细节，而低分辨率特征提供全局上下文。许多 UDA 架构整合了高分辨率和低分辨率特征，以平衡局部精度与大规模上下文推理。虽然特征融合可以增强空间和上下文信息的表示，但它忽略了低分辨率 logits 提供的类感知信息，这种信息可以为高分辨率特征提供结构上的改进信号。

与优化为内部表示的编码器特征不同，编码器特征可能包含冗余或噪声激活，低分辨率的 logits 是监督输出，明确地编码了类别级语义。结果，它们更易于解释，语义上对齐，并且更接近最终预测输出。此外，经过 softmax 归一化的 logits 能实现置信度估计，我们通过基于不确定性的细化来利用这一点。这使得 logits 更适合指导高分辨率特征，确保在细化过程中更好的类别级一致性和计算效率。因此，我们建议利用低分辨率 logits 的全局类别分布和不确定性来指导高分辨率特征的细化，确保它们同时结合空间细节和语义一致性。为了促进由低分辨率 logits 指导的高分辨率特征的细化，我们引入了注意力特征细化（AFR）模块，可以轻松集成到各种现有的基于 HRDA 的 UDA 方法中。尽管 AFR 不直接与伪标签过滤、自适应加权或掩码校正交互，其不确定性感知的细化通过抑制噪声特征和稳定预测，间接地补充了这些过程。这导致一个更强的学生模型，并且随着时间推移，通过基于 EMA 的教师更新，伪标签质量更高、更干净。特征细化过程 AFR 如图 ?? 所示，并在下文中描述。

AFR 通过结合两种互补的注意机制：类别感知逻辑注意（CALA）和不确定性抑制 HR 特征注意（UHFA），保留空间细节并整合语义一致性。

1)

基于类感知化的对数注意力 类别感知逻辑注意从低分辨率的逻辑中提取类别感知信息，同时结合来自 HR 特征的不确定性图。不确定性图由于其计算的简便性以及从分割输出中直接可用性，从 softmax 概率中估计出来，这不同于基于熵或校准的方法。该模块首先将 LR 逻辑 $L_{\text{LR}} \in \mathbb{R}^{B\times C\times H\times W}$ 作为输入，并通过一个 1×1 卷积将多通道表示简化为单通道注意力图

$$L^{\text{attn}} = \sigma(conv^{1\times 1}(L_{\text{LR}})), \tag{3}$$

其中 σ 表示 sigmoid 激活函数。在 CALA 模块的后半部分，HR 特征 $U_{\text{HR}} \in \mathbb{R}^{B\times 1\times H\times W}$ 的不确定性图也通过 sigmoid 激活函数处理

$$U_{\text{HR}}^{\text{attn}} = \sigma(U_{\text{HR}}). \tag{4}$$

然后这两个注意力图 L^{attn} 和 $U_{\text{HR}}^{\text{attn}}$ 通过逐元素相乘结合以获得调制后的注意力图，

$$A_{\text{logits}} = L^{\text{attn}} \odot U_{\text{HR}}^{\text{attn}} \tag{5}$$

这确保了在不确定的 HR 区域中强调低分辨率全局信息，同时在 HR 特征可靠的地方保持高分辨率空间精度。在获得调制后的注意力图后，CALA 模块提取 LR 逻辑的高频分量

$$L_{\text{LR}}^{\text{hf}} = L_{\text{LR}} - G_\gamma^{2D} \otimes L_{\text{LR}} \tag{6}$$

使用高斯平滑滤波器 G_γ^{2D}，并从原始 logits 中减去平滑版本。高斯平滑产生平滑的、可导的过渡，适合语义分割。与像 Sobel 这样的边缘检测器产生的锐利、二值化且不可导的边缘不同，高斯滤波捕捉软边界变化，并在重叠的类别边界之间保持语义连续性。其可导性支持基于梯度的优化，允许无缝集成进入端到端的训练中，而不需要额外的边界预测头。高斯平滑滤波器 G_γ^{2D} 通过计算两个一维高斯函数的外积构建。数学上，

$$G_\gamma^{2D}(i,j) = \frac{1}{2\pi\gamma^2}\exp\left(-\frac{i^2+j^2}{2\gamma^2}\right), \text{for } i,j \in \left[-\frac{k}{2},\frac{k}{2}\right] \tag{7}$$

其中，γ 是标准差，k 是核尺寸。最后，通过结合调制后的 logits 和原始 LR logits 的高频分量来计算由 logit 引导的注意力图

$$A_1 = \sigma(A_{\text{logits}} + L_{\text{LR}}^{\text{hf}}) \tag{8}$$

，它结合了边界信息以及基于全局类别分布和不确定性进行的细化。

2)

不确定性抑制的 HR 特征注意 不确定性抑制的 HR 特征注意力通过结合原始 HR 特征和 LR logits 中的不确定性信息来增强 HR 特征。UHFA 模块首先将 HR 特征图 $F_{\mathrm{HR}} = \{F_{\mathrm{HR}}^{(n)}\}_{n=1}^{N}$，$F_{\mathrm{HR}}^{(n)} \in \mathbb{R}^{B\times C_n\times H_n\times W_n}$ 作为输入，并在通道维度上应用全局平均池化（GAP）操作以计算全局特征表示 $F_{\mathrm{HR}}^{\mathrm{global}} \in \mathbb{R}^{B\times 1\times H\times W}$。随后，UHFA 模块通过使用高斯平滑从 $F_{\mathrm{HR}}^{\mathrm{global}}$ 中提取高频分量。数学上，

$$F_{\mathrm{HR}}^{\mathrm{hf}} = F_{\mathrm{HR}}^{\mathrm{global}} - G_{\gamma}^{2D} \otimes F_{\mathrm{HR}}^{\mathrm{global}}, \tag{9}$$

，其中，G_{γ}^{2D} 通过使用方程 (6) 计算。之后，全局池化的 HR 特征 $F_{\mathrm{HR}}^{\mathrm{global}}$ 及其高频对应部分 $F_{\mathrm{HR}}^{\mathrm{hf}}$ 被融合以强化边界结构，再通过基于 3×3 卷积注意力层的空间注意力。选择性地增强重要的空间区域，同时抑制信息量较少的区域，确保边界结构和模糊分类区域获得更高的关注以提高特征的区分度。在 UHFA 的下一步中，A_{HR} 与从 LR logits 获得的不确定性图的指数形式 $U_{\mathrm{LR}} \in \mathbb{R}^{B\times 1\times H\times W}$ 进行逐元素相乘，然后通过 sigmoid 激活函数获得不确定性抑制的 HR 特征注意力图

$$A2 = \sigma(A_{\mathrm{HR}} \otimes \exp(-U_{\mathrm{LR}})). \tag{10}$$

。这个注意力图优先考虑 LR logits 自信的区域中的 HR 特征，确保在分割边界中保留细粒度的空间细节。相反，它抑制不确定 LR 区域中 HR 特征的影响，以防止过拟合可能缺乏全局类别先验的高分辨率细节。在获得注意力图 A1 和 A2 之后，AFR 模块自适应地组合它们以生成最终的注意力图

$$A_{\mathrm{final}} = \alpha A1 + (1-\alpha)A2 \tag{11}$$

，其中 α 是一个可学习的参数。最后，AFR 模块通过对原始 HR 特征与最终注意力图进行逐元素相乘并通过残差连接，计算得到精炼的 HR 特征

$$F_{\mathrm{HR}}^{\mathrm{refined}} = (F_{\mathrm{HR}} \odot A_{\mathrm{final}}) + F_{\mathrm{HR}}. \tag{12}$$

。因此，AFR 模块生成的精炼 HR 特征在空间上连贯、边界感知和语义一致，同时也保留了原始 HR 特征以增强鲁棒性。

D. 视觉规划

我们将提出的 AFRDA 模块与我们最近开发的 POVNav 视觉规划器结合，以实现高效的基于视觉的导航。AFRDA 通过将像素分类为可导航（Ω_N）和不可导航（Ω_{NN}）区域来生成可导航图像 I_t^n，而 POVNav 应用视觉地平线概念以确保明确分离。3D 目标被投影到图像边界作为周边光学目标（POG），通过平衡偏离目标方向和前进的进展，选择一个称为地平线光学目标（HOG）的帕累托最优子目标。规划到 HOG 的无碰撞视觉路径，并提取两个特征——接近度（λ），测量到地平线的距离，以及对准度（ϕ），量化路径偏离。控制误差 $e(t) = [\lambda, \phi] - [\lambda_0, 0]$ 通过视觉伺服来最小化，其中 λ 调节前进速度而 ϕ 调整角速度。这使得导航按目标平稳、自适应地进行。

在这项工作中，我们使用五个数据集来评估我们的框架。在这些数据集中，GTA V 是一个从游戏环境中收集的合成数据集，包含 24,966 张高分辨率图像，并标注有 33 个类别。与 GTA V 类似，Synthia 也是一个模拟数据集，包含 9400 张图像，并与 GTA V 数据集共享 16 个语义类别。我们工作中使用的另一个城市数据集是 Cityscapes 数据集，具有 2975 个训练和 500 个验证图像。在森林数据集中，其中一个是 RUGD 数据集，一个越野数据集，包含 7453 张图像、24 个语义类别和 8 种独特的地形类型。MESH 数据集是在我们实验室前面收集的，包括 4415 张训练图像和 827 张验证图像。

我们在现有的 SOTA 方法 MIC [10] 上开发了我们的 AFRDA 模型，由于 GPU 内存限制，使用了批量大小为 2 和裁剪大小为 952。我们在 Cityscapes 和 MESH 数据集的验证集上评估模型，其中 Cityscapes 数据集使用平均交并比（mIoU）评估城市环境，而 MESH 数据集由于没有真实值，使用定性结果评估森林适应性。我们将我们提出的 AFRDA 与基线方法 MIC [10]，以及其他 SOTA 方法进行了定量和定性比较。在城市环境中，我们进行了两个领域适应任务：从 GTA V → Cityscapes 和 Synthia → Cityscapes，以及一个从 RUGD → MESH 的森林适应任务。首先，我们在表 I 中展示了 GTA V → Cityscapes 适应的定量结果。

表 I 显示，我们提出的 AFRDA 实现了 76.60 % mIoU，超过了基线 MIC 的 +1.05 mIoU。在 GTA V → Cityscapes 适应任务的 19 个类中，AFRDA 在 11 个类中表现领先，包括诸如 “围栏”、“杆”、“交通灯”、“交通标志” 和 “火车” 等小而稀有的类，这些类通常难以预测。此外，当 AFRDA 框架的核心组件 AFR 集成到 HRDA 框架中时，表现出良好的性能，导致 +0.76 mIoU 的提升。稍后，我们在表 II 中展示了 Synthia → Cityscapes 适应的定量结果。如表 II 所示，在 Synthia → Cityscapes 适应任务中，AFRDA 优于所有最新的先进方法（SOTA），并在基线 MIC 上超过了 +1.04 mIoU。此外，AFRDA 在此设置中预测具有挑战性的类时也展现了其效能。此外，将 AFR 与 HRDA 模块集成可使 HRDA 框架提高 +0.88 mIoU，显示 AFR 作为即插即用模块的可行性。我们还将我们的 AFR 模块接入到一个名为 ERF [25] 的最近的 UDA 方法上，并证明我们的 AFR 模块提升了 GTA V → Cityscapes 和 Synthia → Cityscapes 适应任务的整体结果。我们在图中展示了我们提出的 AFRDA 的定性结果。城市和森林环境适应任务分别见 1 和图 2 。如图 1 所示，在 GTA V → Cityscapes 和 Synthia → Cityscapes 适应方面，我们的模型比其他 SOTA 方法表现更好。结果表明，AFRDA 在墙、围栏、植被和人行道的分割上表现出色，生成了更清晰的边界、更光滑的边缘，并且在小类如交通标志和杆子的预测中更加准确。这是因为 AFR 模块明确利用边界细节并增强上下文感知，从而提高了模型区分人行道与道路以及围栏与墙壁的能力，同时准确地预测杆子和交通标志，从而在复杂的城市场景中实现更稳健和精确的分割。图 2 展示了我们提出的 AFRDA 框架在森林环境中的优越性。当所有其他方法都难以准确预测草地，特别是在草地显得较干燥或发黄时，我们的模型成功地高精度预测了 “草地”，“灌木”，“天空” 和其他元素。我们还在表 III 中报告了带有和不带 AFR 的各种 UDA 方法的吞吐量和 GPU 内存使用情况。将 AFR 添加到现有模型中导致最低限度的训练减速（例如，HRDA 仅为 7.6 % ），同时保持了相当或更快的推理速度。值得注意的是，AFR 稍微减少了 MIC 的训练 GPU 内存（从 20.58 → 减少到 20.51 GB），这可能是由于结构化注意力减少了中间特征冗余。当将其添加到 ERF 时，它也保持或增强了训练速度，突显其轻量化和高效的设计。总的来说，AFR 在几乎不增加计算成本的情况下提供了精度提升。

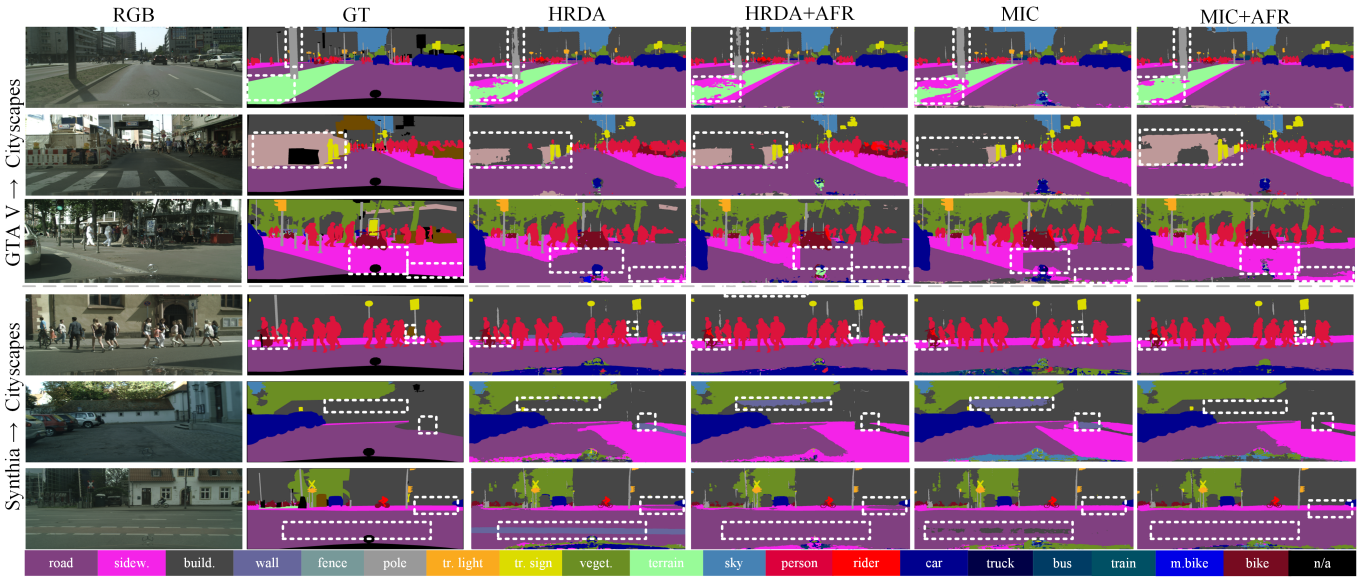

Figure 1: 适配 GTA V → Cityscapes 和 Synthia → Cityscapes 的定性结果。

Table I: 从 GTA V → 到 Cityscapes 数据集适应的定量比较。

Method	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
ADVENT [5]	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
DACS [23]	89.9	39.7	87.9	30.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1
ProDA [24]	87.8	56.0	79.7	46.3	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
DAFormer [23]	95.7	70.2	89.4	53.5	48.1	49.6	55.8	59.4	89.9	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.3
HRDA [9]	97.12	78.03	90.83	60.49	47.42	57.63	59.73	70.38	91.17	44.02	93.43	79.04	54.46	94.62	85.26	85.01	67.97	63.16	68.10	73.05
HRDA + AFR	96.97	77.11	90.87	63.06	50.28	58.98	63.87	71.39	91.39	46.62	94.09	78.15	53.04	94.40	83.17	85.70	74.07	60.62	68.63	73.81 ↑ 0.76 %
MIC [10]	97.37	80.37	91.31	60.85	51.83	59.33	61.92	73.19	91.82	50.16	94.33	80.17	55.53	94.60	86.04	90.33	81.63	65.49	69.13	75.55
MIC+AFR (Ours)	97.66	81.65	91.43	61.99	56.97	61.98	64.53	74.62	91.66	51.29	93.86	81.12	58.30	94.86	85.03	90.74	82.41	64.87	70.33	76.60 ↑ 1.05 %
ERF [25]	97.01	78.16	91.31	60.18	56.61	59.22	63.62	68.41	91.67	52.95	94.5	79.71	55.27	94.65	86.1	90.14	81.55	65.2	70.58	75.62
ERF+AFR	97.38	80.11	91.16	60.6	58.28	60.43	64.5	73.88	91.75	52.52	94.77	80.11	54.47	94.53	85.5	90.17	81.76	64.48	70.23	76.14 ↑ 0.52 %

E. 消融研究

我们通过对 GTA V → Cityscapes 适应任务进行消融研究来评估我们提出的 AFR 模块的有效性和设计，如表 IV 所示。

为了展示我们 AFR 模块的可解释性和效果，我们在图 **??** 中可视化了关键的注意力组件。CALA 注意力图突出显示了道路、天空和建筑物外立面等大型且自信的语义区域，并具有较高的注意力值，这与其提供语义感知的全局指导的角色一致。虽然一些较小的物体如电杆和交通标志是可见的，但它们获得的注意力相对平滑且细节较少，表明对细微结构的敏感性有限。相比之下，在一个高分辨率裁剪中显示的 UHFA 注意力图则侧重于高频空间信号，并增强了边缘和边界，尤其是在电杆、建筑物、交通标志和行人周围。最终的注意力图通过可学习的融合整合了这两种来源。为了可视化 UHFA 的具体效果，我们展示了最终注意力和 CALA 之间的差异，这揭示了结构边界、类别过渡和物体轮廓周围的强烈激活。在这两个示例中，电杆、交通标志和建筑边界的轮廓清晰可见。这些较高的值确认了 UHFA 在边界敏感区域提高注意力的作用，同时保留了来自 CALA 的更广泛的语义结构。虽然由于阈值变动，我们没有提供二值化的 UHFA 掩码，但可视化表明 UHFA 抑制了平坦区域并保留了高频边界内容，支持了我们的说法，即 AFR 增强了沿类别边界和物体边缘的高分辨率特征细化。我们进行了一系列实验，通过逐个移除每个模块来评估基于类别感知对数的注意力（CALA）和抑制不确定性高分辨率特征注意力（UHFA）模块的贡献。结果显示，禁用 CALA 导致性能有小幅下降（76.04 % mIoU），而禁用 UHFA 导致的下降略大（75.86 % mIoU）。这些结果表明，两个模块都对 AFR 的有效性具有贡献，但 UHFA 的影响稍微强一些。尽管每个模块单独运行时都能超越基线提高分割效果，但它们的结合产生了最显著的增益，这加强了 AFR 通过 CALA 和 UHFA 之间的互补交互来增强局部和全局表示的观点。

1) 不确定性估计的影响

我们进行了实验，分别从 CALA 中的高分辨率特征和 UHFA 中的低分辨率 logits 中去除不确定性估计。结果表明，去除高分辨率不确定性导致观察到的性能最大降幅（75.17 % mIoU），而去除低分辨率 logit 不确定性则实现

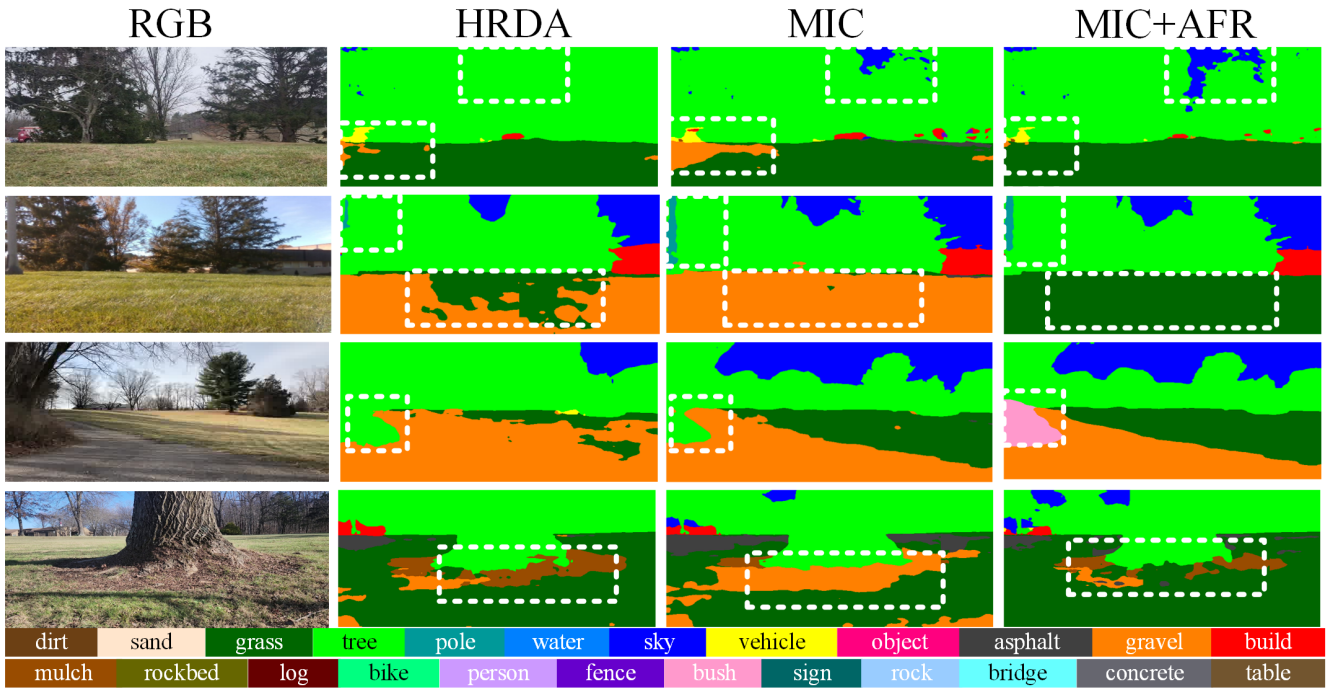

Figure 2: RUGD → MESH 适应性的定性结果。

Table II: 从 Synthia → Cityscapes 数据集适应性的定量比较。

Method	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
ADVENT [5]	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
DACS [23]	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	–	90.8	67.6	38.3	82.9	–	38.9	–	28.5	47.6	48.3
ProDA [24]	87.8	45.7	84.6	37.1	0.6	44.0	54.6	37.0	88.1	–	84.4	74.2	24.3	88.2	–	51.1	–	40.5	45.6	55.5
DAFormer [23]	84.5	40.7	88.4	41.5	6.5	50.0	55.0	54.6	86.0	–	89.8	73.2	48.2	87.2	–	53.2	–	53.9	61.7	60.9
HRDA [9]	84.89	45.73	89.13	45.32	8.21	56.26	66.05	64.26	85.9	-	93.33	78.99	53.42	87.8	-	63.86	-	62.17	64.63	65.62
HRDA + AFR	88.74	49.77	89.37	49.05	2.87	58.26	65.9	61.76	89.28	-	93.57	80.13	51.87	89.88	-	67.43	-	63.85	62.22	66.50 ↑ 0.88 %
MIC [10]	88.39	51.74	89.19	48.05	7.85	56.02	66.69	62.09	88.04	-	94.02	81.01	58.02	90.84	-	67.02	-	63.07	67.04	67.26
MIC+AFR (Ours)	88.50	54.83	88.89	48.42	8.74	59.33	66.43	63.67	87.41	-	93.86	81.77	59.22	91.18	-	66.76	-	68.36	65.44	68.30 ↑ 1.04 %
ERF [25]	89.67	55.87	88.44	46.25	1.28	58.41	66.76	62.73	87.17	-	94.64	81.5	59.13	94.65	-	62.37	-	66.93	63.69	67.46
ERF+AFR	89.89	55.9	89.23	47.24	8.09	59.29	67.62	62.52	87.73	-	94.26	81.25	57.78	90.67	-	65.62	-	67.67	65.72	68.16 ↑ 0.70 %

Table III: 在 RTX 4090 上进行训练和推理时的运行时间和内存消耗。

Method	Training		Inference	
	Throughput	GPU Memory	Throughput	GPU Memory
HRDA [9]	0.92 it/s	20.47 GB	2.02 img/s	9.45 GB
HRDA+AFR	0.85 it/s	20.47 GB	1.88 img/s	9.85 GB
MIC [10]	0.73 it/s	20.58 GB	1.92 img/s	9.91 GB
MIC+AFR	0.70 it/s	20.51 GB	1.91 img/s	9.91 GB
ERF [25]	0.75 it/s	18.78 GB	2.24 img/s	8.21 GB
ERF+AFR	0.79 it/s	18.83 GB	2.21 img/s	8.21 GB

了 76 % 的 mIoU。这表明在高分辨率下抑制不确定性对于 AFR 的有效性至关重要。由于高分辨率不确定性与低分辨率 logits 相乘，其去除扰乱了基于置信度的细化，降低了全局类别先验的有效性并增加了空间不一致性。尽管高分辨率特征仍然通过注意力和高频提取进行处理，它们失去了适当加权的低分辨率指导。与维持稳定特征流的基线不同，没有高分辨率不确定性的 AFR 使全局上下文与局部细节不对齐，导致分割错误，而其他组件无法弥补。

我们还进行了消融研究，通过分别和同时从 CALA 和 UHFA 中移除高频组件。从 CALA 中仅移除边界优化导致性能下降更大 (75.20 %) 相比于从两个模块中移除 (75.58 %)，而仅从 UHFA 中移除则结果是 75.65 % 。这表明，基于类别的边界优化确保了类别先验与空间细节的正确对齐，直接影响了低分辨率的逻辑值如何与高分辨率特征相互作用。没有这种优化，模型依赖于未对齐的边界提示，从而增加了分割错误。从两个模块中移除边界优化减少了这种未对齐现象，因为模型适应使用其他可用特征，如不

Table IV: 消融研究结果显示不同组件对 mIoU 性能的影响。

Model Variation	mIoU (%)	
	Absolute	δ_{AFR}
Baseline (No AFR)	75.55	-1.05
AFR w/o CALA	76.04	-0.56
AFR w/o UHFA	75.86	-0.74
AFR w/o HR Uncertainty (CALA)	75.17	-1.43
AFR w/o LR Logits Uncertainty (UHFA)	76.00	-0.60
AFR w/o Boundary (CALA & UHFA)	75.58	-1.02
AFR w/o Boundary in CALA	75.20	-1.4
AFR w/o Boundary in UHFA	75.65	-0.95
Full AFR	76.60	0

确定性加权和上下文先验。当在 CALA 和 UHFA 中都有边界优化时，分割效果提升，因为类先验边界与空间一致的边界对齐；CALA 优化低分辨率的逻辑值，而 UHFA 锐化物体边缘并减少噪声，展示了 AFR 的有效设计。为了进一步验证高频提示对小物体分割的贡献，我们在表 **??** 中呈现了类别别的 IoU 结果。移除高频信息稳定地降低了所有小结构类别的性能。例如，与完整的 AFR 模型相比，Pole 的 IoU 下降了 1.14 % ，Traffic Light 下降了 1.35 % ，Rider 下降了 1.94 % 。这些结果证实了 UHFA 的高频优化通过提高边缘敏感性和保持细粒度的结构细节，提高了对小而细的类别的检测。

F. 导航任务

我们评估了基于 RUGD → MESH 设置训练的 AFRDA 模型，通过与 POVNav 规划器 [26] 集成，并在实验室附近的森林中部署于 Husky 机器人上（图 **??** ）。在 RTX 2060 GPU 上使用 640×480 图像时，AFRDA 的分割耗时为 0.72 秒，而整个管道耗时约 0.77 秒。机器人以 0.1 米/秒的速度沿约 10 米路径移动，成功避开不可导航区域以到达目标。

II. 结论

我们提出了一种名为 AFRDA 的模型，以改进无监督领域自适应语义分割。我们的方法基于一个自我训练框架，该框架同时探讨局部和全局细节以及边界信息。为此，我们开发了一个名为关注特征优化的模块，该模块基于低分辨率的 logits 优化高分辨率特征。AFR 利用低分辨率 logits 的语义意识和高分辨率特征的空间意识来改善语义分割。我们在基准数据集上测试了我们的方法，我们的模型显著超越了 SOTA。此外，我们在一个机器人车辆中部署了我们的框架，以便在非结构化环境中导航。

References

[1] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 102–118.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[3] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3234–3243.

[4] M. Wigness, S. Eum, J. G. Rogers, D. Han, and H. Kwon, "A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments," in *IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2019, pp. 5000–5007.

[5] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 2517–2526.

[6] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 285–10 295.

[7] R. Gong, W. Li, Y. Chen, D. Dai, and L. Van Gool, "Dlow: Domain flow and applications," *Int. J. Comput. Vis.*, vol. 129, no. 10, pp. 2865–2888, 2021.

[8] L. Hoyer, D. Dai, and L. Van Gool, "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9924–9935.

[9] ——, "Hrda: Context-aware high-resolution domain-adaptive semantic segmentation," in *European conference on computer vision*. Springer, 2022, pp. 372–391.

[10] L. Hoyer, D. Dai, H. Wang, and L. Van Gool, "Mic: Masked image consistency for context-enhanced domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11 721–11 732.

[11] J. Xiang, C. Wan, and Z. Cao, "Pseudolabel guided pixels contrast for domain adaptive semantic segmentation," *Scientific Reports*, vol. 14, no. 1, p. 31615, 2024.

[12] X. Zhao, N. C. Mithun, A. Rajvanshi, H.-P. Chiu, and S. Samarasekera, "Unsupervised domain adaptation for semantic segmentation with pseudo label self-refinement," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2024, pp. 2399–2409.

[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.

[15] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-scnn: Gated shape cnns for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 5229–5238.

[16] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei, "Rethinking bisenet for real-time semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 9716–9725.

[17] X. Chen, Y. Zheng, Y. Wei, and Y. Shen, "Uncertainty-aware pseudo-label filtering for source-free unsupervised domain adaptation," *Neurocomputing*, vol. 575, p. 127190, 2024.

[18] T. Guan, D. Kothandaraman, R. Chandra, A. J. Sathyamoorthy, K. Weerakoon, and D. Manocha, "Ga-nav: Efficient terrain segmentation for robot navigation in unstructured outdoor environments," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 8138–8145, 2022.

[19] T. Guan, Z. He, R. Song, D. Manocha, and L. Zhang, "Tns: Terrain traversability mapping and navigation system for autonomous excavators," *arXiv preprint arXiv:2109.06250*, 2021.

[20] M. Sivaprakasam, S. Triest, C. Ho, S. Aich, J. Lew, I. Adu, W. Wang, and S. Scherer, "Salon: Self-supervised adaptive learning for off-road navigation," *arXiv preprint arXiv:2412.07826*, 2024.

[21] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2021, pp. 1369–1378.

[22] Z. Chen, Z. Ding, J. M. Gregory, and L. Liu, "Ida: Informed domain adaptive semantic segmentation," in *IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*. IEEE, 2023, pp. 90–97.

[23] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "Dacs: Domain adaptation via cross-domain mixed sampling," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2021, pp. 1379–1389.

[24] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. Wen, "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 12 414–12 424.

[25] Y. Song, J. Shi, C. Liu, S. Bai, Y. Yuan, X. Shu, Q. Qian, D. Xu, and Y. Sun, "Extended receptive field uda semantic segmentation based on spatial alignment and knowledge distillation," *IEEE Trans. Autom. Sci. Eng.*, 2025.

[26] D. Pushp, Z. Chen, C. Luo, J. M. Gregory, and L. Liu, "Povnav: A pareto-optimal mapless visual navigator," in *Int. Symp. Exp. Robot. (ISER)*. Springer, 2023, pp. 250–263.