© 2025 IEEE。允许将本材料用于个人用途，对于任何其他用途，包括为广告或宣传目的重新印刷/再版本材料、创作新的汇编作品、转售或再分发至服务器或列表，或在其他作品中重复使用本作品中任何受版权保护的组成部分，均须获得 IEEE 的许可。数字对象标识符:10.1109/IROS60139.2025.11245809

arXiv:2601.00359v1 [cs.CV] 1 Jan 2026

通过蒸馏和 RGB-D Transformer 高效预测稠密视觉嵌入

S¨ohnke Benedikt Fischedick、Daniel Seichter、Benedict Stephan、Robin Schmidt 和 Horst-Michael Gross

摘要—在家庭环境中，机器人需要全面理解周围环境，才能与未经训练的人类进行有效且直观的交互。本文提出 DVEFormer——一种高效的基于 RGB-D Transformer 的方法，通过知识蒸馏预测与文本对齐的稠密视觉嵌入（DVE）。我们的方法不直接使用固定的预定义类别执行经典语义分割，而是利用来自 Alpha-CLIP 的教师嵌入，引导高效的学生模型 DVEFormer 学习细粒度的逐像素嵌入。尽管该方法仍支持经典语义分割（例如通过线性探测），但还进一步实现了灵活的基于文本的查询以及其他应用，例如创建全面的三维地图。在常用室内数据集上的评估表明，我们的方法在满足实时要求的同时实现了具有竞争力的性能：完整模型在 NVIDIA Jetson AGX Orin 上达到 26.3 FPS，较小变体达到 77.0 FPS。此外，我们展示了定性结果，突出了该方法在真实世界应用中的有效性和潜在用例。总体而言，我们的方法可作为传统分割方法的即插即用替代方案，同时支持灵活的自然语言查询，并能无缝集成到移动机器人技术的三维建图流程中。

I. 引言

在复杂室内环境中自主运行的移动机器人需要详细了解其运行所在的环境。这种场景知识应当全面且具有普适性，即应捕捉场景中相关对象的细节，而不局限于预先定义的固定类别范围（通常称为封闭集），从而适用于各种环境。

在我们正在开展的 CO-HUMANICS 研究项目中 [1–3], , 一台具备远程临场功能的移动机器人正在一位老年人的公寓中运行，家人和朋友可以远程访问并控制它，以拉近人与人之间的物理距离。在这种场景下，操作者往往未经培训，因此，机器人必须尽可能易于直观操控。例如，指定机器人的目标位置时，最好能使用自然语言标识符，而不是局限于预先定义的类别范围。图 1 通过命令 “驶向扶手椅” 展示了这一点。该命令要求机器人不仅能将扶手椅识别为一个独立物体，还能将其与其他类型的椅子区分开来。

作者隶属于神经信息学与认知机器人实验室，伊尔梅瑙工业大学，德国伊尔梅瑙，邮编98693。联系方式：soehnke.fischedick@tu-ilmenau.de，开放研究者与贡献者身份识别码：0000-0001-8447-0584。本研究获得卡尔·蔡司基金会对项目老年人与交互式陪伴者的共在 （CO-HUMANICS）的资助，项目执行期为2021年6月至2026年5月。http://www.co-humanics.de。

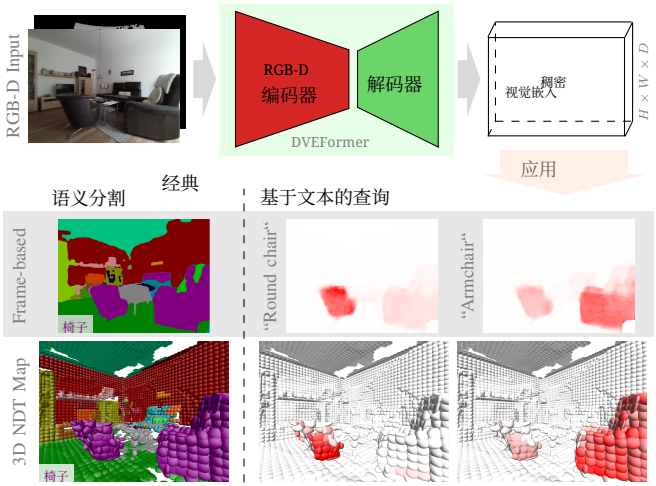


图1. 我们提出的方法 DVEFormer 在一个应用场景 [1]中，对此前未见过的数据预测稠密视觉嵌入（DVE）的真实示例。我们的方法不仅支持经典语义分割（固定的预定义类别——封闭集），还支持通过基于文本的查询指定超出固定预定义类别范围的感兴趣对象。此外，如底部所示，预测得到的嵌入可以轻松集成到现有的映射方法中，例如 [4]。语义颜色的选取遵循 [5]，并且是 NYUv2 [6]的默认颜色。图3展示了所描绘语义颜色的子集。基于文本的查询结果显示余弦相似度（红色：高，白色：低）。

然而，图像分割 [5, 7 – 15] 和三维建图方法 [4, 16–21] 通常仅适用于固定的类别范围，这对于许多现实世界的应用来说可能并不足够。这是因为它们通常依赖于常见室内分割数据集中的类别范围，例如 NYUv2 [6], 、SUN RGB-D [22], 或 ScanNet [23], ; 对于上述示例，这些数据集仅提供 椅子 和 沙发 这两个类别。尽管已有开放词汇分割方法 [24–33], , 但它们往往不适用于计算资源有限的移动应用。这些方法日益复杂，需要多个大型骨干网络，或并非专为室内环境设计。

为克服这些局限，我们提出了稠密视觉嵌入变换器（DVEFormer）——一种知识蒸馏方法 [34] ，采用高效的基于 Transformer 的 [35]架构，实现文本对齐视觉嵌入的密集预测。我们不依赖所用室内数据集预先定义的语义类别，而仅将数据集提供的掩码标注（不含语义标签）作为输入，送入 Alpha-CLIP [36]提取教师嵌入。我们的学生模型 DVEFormer 基于 EMSAFormer 架构 [12],, 采用经过改进的基于 Swin-Transformer 的 [37]RGB-D编码器和轻量级、类似 SegFormer 的 [38] 解码器来预测稠密分割，同时

针对移动机器人上的高效推理进行了优化。我们不预测预先定义的语义类别，而是训练模型直接输出从教师模型蒸馏得到的稠密视觉嵌入。这样一来，预测的嵌入便与文本对齐，并且无需重新训练模型本身，即可在运行时通过与输入的文本嵌入进行比较来查询。为进一步展现该方法的实际应用，我们展示了将估计得到的嵌入直接集成到三维建图流水线 [4, 16, 39], 中的定性结果，从而形成全面的三维场景表示。随后，未经训练的操作人员也可以通过文本查询该表示，例如指定机器人导航的目标。

我们在 GitHub 上分享了数据准备流程、训练细节以及模型权重：`https://github.com/TUI-NICR/DVEFormer`。

II. 相关工作

下面我们将总结相关工作。首先回顾密集 RGB-D语义分割方法。接着讨论用于对齐视觉和文本模态的视觉-语言模型，重点介绍它们生成全局表示的能力，同时指出它们在细粒度任务和移动平台上的局限性。最后，简要讨论知识蒸馏，说明如何将特征从大型教师模型迁移到更高效的学生模型。

*A. RGB-D*语义分割

密集语义分割可以通过多种方式实现。大多数方法采用编码器–解码器方法进行逐像素预测 [5, 7, 10, 12–15], , 而近期方法（如 [40–43], ）则通过基于Transformer的解码器预测更大的分割区域，并将其分配给固定类别。尽管这些逐掩码方法表现出色，但它们往往较为复杂，且很少针对移动应用进行优化，因此我们专注于逐像素预测。

结合 RGB 和深度数据是提升分割性能的有效策略，因为这两种模态提供了互补信息 [44–46]。RGB 图像提供颜色和纹理等语义线索，而深度图像则提供几何信息。为结合 RGB 和深度，一些方法 [5, 44, 47, 48] 使用两个或更多编码器分别处理这两种模态，并在不同阶段融合特征。虽然这一策略有效，但需要精心设计融合机制，而且通常会增加计算开销，具体取决于所使用的骨干网络和融合机制。另一种方法是使用单个编码器，将 RGB 和深度作为联合输入送入模型。早期基于卷积神经网络的方法通过采用特殊的深度感知卷积来实现这一点 [8, 49, 50]。然而，这类操作通常不适用于移动平台。除了这些方法之外，研究者还提出了基于Transformer的方法。其中一些方法采用视觉Transformer [51], 在不使用金字塔结构的单个Transformer中处理这两种模态 [10], , 而另一些方法则使用基于Transformer的编码器，其采用卷积神经网络中常见的金字塔结构，因此适合作为现有模型的直接替代方案 [11, 12, 15]。

在我们的工作中，我们以 EMSAFormer [12], 为基础，它使用单个经过修改的 Swin-V2-T 编码器联合处理 RGB 和深度数据。与原始 Swin-V2-T 骨干网络相比，仅针对深度集成进行了少量修改，并未引入任何可能不受 NVIDIA TensorRT 等推理引擎支持或不适用于这些引擎的特殊操作。

B. 用于密集预测的视觉-语言模型

视觉-语言模型备受重视，因为它们能够在共享嵌入空间中对齐视觉和文本模态。CLIP [52] 在这一领域取得了突破，它通过对比学习联合训练视觉编码器和文本编码器，在各种图像分类任务上实现了出色的零样本性能。然而，CLIP 仅为整幅图像生成单个全局嵌入，因此在需要细粒度局部嵌入的任务中能力有限。

为克服这一限制，人们提出了多种方法来生成更细致的嵌入。一种策略是通过围绕感兴趣对象裁剪图像来提取局部嵌入 [53, 54]。MaskCLIP [24]和 CAT-Seg [27] 等方法通过将局部视觉嵌入直接与文本嵌入进行比较，或加入用于密集预测的额外层，在闭集和开放词汇设置下执行语义分割。其他方法 [25, 30, 32, 33] 也引入了额外机制，以从全局 CLIP 嵌入中提取局部嵌入。然而，这些方法通常需要同时使用两个大型视觉编码器，因此会增加计算复杂度。相比之下，Alpha-CLIP [36] 将额外的阿尔法通道作为输入，以引导模型为特定区域生成嵌入，同时不改变原始图像内容。尽管当关注不同掩码时，这种方法需要对每张图像进行多次推理，但它保留了宝贵的上下文信息，并维持了原始 CLIP 模型的泛化性能。为进一步提升分割性能，一些方法 [24, 31, 33, 54]采用知识蒸馏或伪标签策略。例如，MaskCLIP+ [24] 使用 MaskCLIP 的预测作为真实标签，指导分割专用模型的训练。这种方法克服了直接使用 CLIP 图像编码器所带来的架构僵化问题，且性能优于 MaskCLIP。

在我们的工作中，我们利用 Alpha-CLIP 生成教师嵌入，作为知识蒸馏框架中的目标，使学生模型 DVEFormer 能够预测与 Alpha-CLIP 和 CLIP 的视觉-语言空间对齐的密集逐像素嵌入。

C. 稠密视觉嵌入的知识蒸馏

知识蒸馏 [34],即将大型、计算密集型模型中的知识迁移到更小且

更高效的模型中，这种方法已在许多领域得到验证。例如，在自然语言处理中，大语言模型通常会被蒸馏为紧凑型模型，同时保留教师的大部分能力 [55, 56]。蒸馏可以通过多种方法实现，例如，最小化比较教师和学生模型输出的损失 [24, 54–56]。在计算机视觉中，训练模型来预测 CLIP 所提供的那类稳健视觉嵌入，需要大量数据和可观的计算资源，这对于密集的逐像素任务而言可能并不切实际。

我们利用提供二值分割掩码的现有数据集，并使用 Alpha-CLIP 生成教师嵌入，作为伪真实标签。我们的方法使用这些教师嵌入来监督学生模型，与 MaskCLIP+ 的做法类似，使其能够预测稠密的逐像素嵌入，这些嵌入与视觉-语言空间高度对齐，同时显著降低推理时的计算开销。

III. 稠密视觉嵌入的高效预测

我们的方法如图 2 所示，首先按照第 III-A 节的说明，使用 Alpha-CLIP [36]，提取教师嵌入。提取的嵌入用于训练我们的学生模型——稠密视觉嵌入变换器（DVEFormer）。我们在第 III-B 节介绍模型架构和蒸馏训练。DVEFormer 预测密集逐像素嵌入，其中编码了适用于各种下游任务的全面信息。例如，正如第 III-C 节所述，这些嵌入可以与对齐的文本编码器配合，用于文本引导检索，也可以用于分类任务。

A. 分割级视觉嵌入提取

我们的方法需要以像素级视觉嵌入作为离线知识蒸馏的教师信息。为提取这些嵌入，对于每个数据集样本，将 RGB图像 $\mathbf{I}_{\text{RGB}}$与对应的每个二值分割掩码 $\mathbf{M}_i \in \mathbb{M}$ 配对。这些二值分割掩码由扩展版数据集 [5, 12],提供的全景标注生成，其中包含不同的物体（例如椅子或桌子）以及无法计数的区域（例如墙或地板）。每一对 $(\mathbf{I}_{\text{RGB}}, \mathbf{M}_i)$ 都由 Alpha-CLIP 处理，以生成一个嵌入 $\mathbf{e}^{\mathbf{M}_i} \in \mathbb{R}^D$，其维度为 $D = 768$。虽然该维度可能高于我们的应用所需，但我们保留完整维度，以确保与 Alpha-CLIP 使用的现有文本编码器保持对齐，避免降维可能引入的伪影。

图 3a 展示了从 NYUv2 数据集 [6]中提取的分割嵌入$\mathbf{e}^{\mathbf{M}_i}$ 的 t-SNE 可视化结果 [57] 。结果表明，这些嵌入往往按场景类别聚类，而非按分割的语义类别聚类。这种现象[1] 可以理解为一种取决于应用的权衡：侧重局部物体还是全局场景。 然而，我们的应用更依赖于区分每个分割的具体属性（例如物体类别、颜色或材质属性），而非全局场景线索。为此，我们通过抑制场景上下文来优化$\mathbf{e}^{\mathbf{M}_i}$ 。具体而言，我们将整张图像视为一个大型分割，使用 Alpha-CLIP 计算全局图像嵌入$\mathbf{e}^I$ ，并对每个掩码嵌入进行如下调整：

$$\tilde{\mathbf{e}}^{\mathbf{M}_i} = \frac{\mathbf{e}^{\mathbf{M}_i}}{\|\mathbf{e}^{\mathbf{M}_i}\|} - \alpha \cdot \frac{\mathbf{e}^I}{\|\mathbf{e}^I\|} \tag{1}$$

[1]另请参阅 Alpha-CLIP GitHub 议题：https://github.com/SunzeY/AlphaCLIP/issues/67。

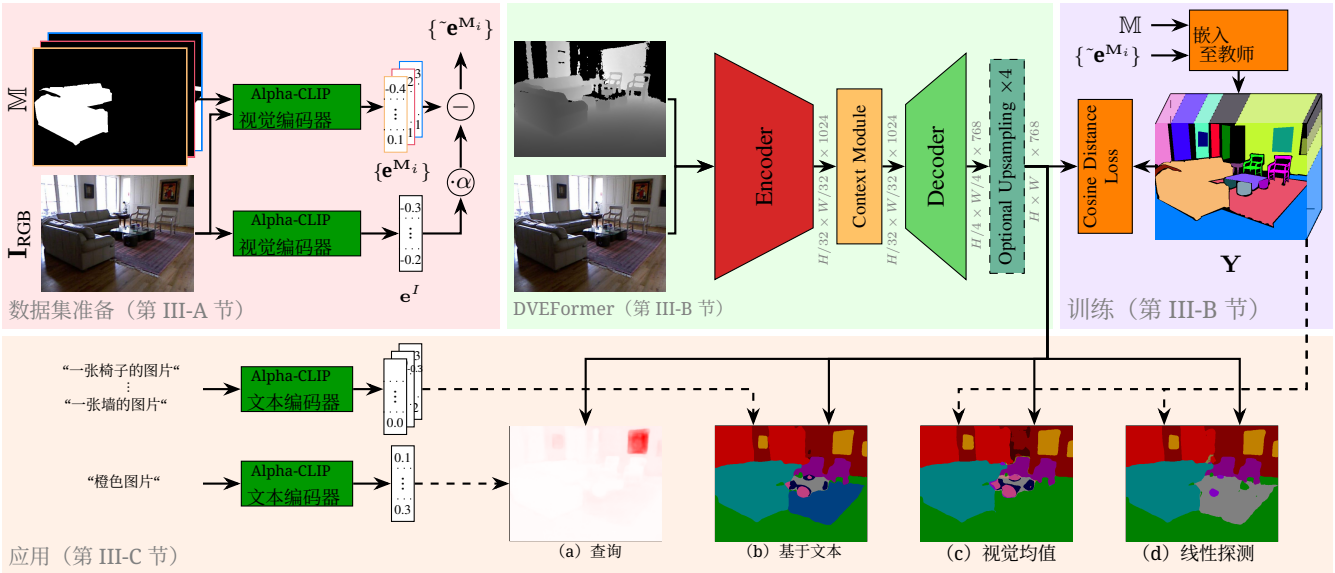
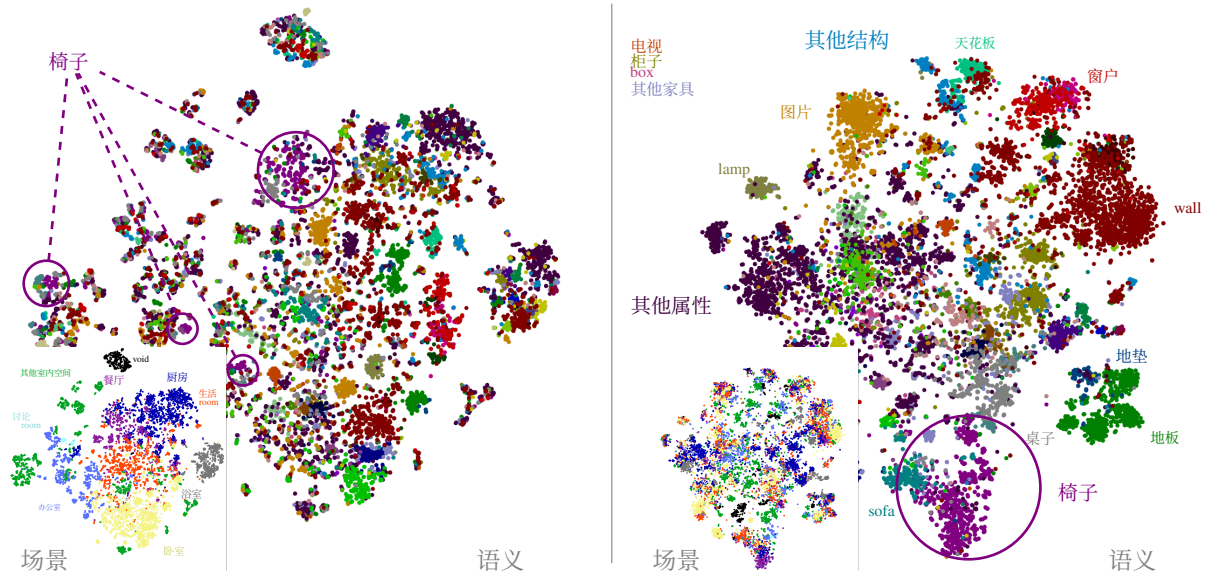


图 2. 我们提出的方法在 SUN RGB-D [22] 的示例图像上的概览。Alpha-CLIP 在离线状态下处理 RGB 图像和一组二值分割掩码，以提取一组教师嵌入。这些嵌入引导我们高效的稠密视觉嵌入变换器（DVEFormer）通过知识蒸馏学习稠密的逐像素视觉嵌入。对于下游应用，生成的嵌入可用于基于文本的检索（a），也可用于具有预定义类别的经典语义分割（b）-（d）或其他机器人应用，从而支持灵活的场景理解任务。图中所示语义颜色子集请参见图 3。



(a) Alpha-CLIP 预测的原始嵌入　　(b) 精炼嵌入

图 3. Alpha-CLIP（见第 III-A 节）为 NYUv2 训练集划分生成的分割嵌入可视化：(a) 展示移除全局场景上下文前的嵌入，(b) 展示抑制上下文后的嵌入（$\alpha = 0.65$）。在两幅图中，主视图里的点按真实语义类别着色，子图则展示相同的点按其场景类别着色（定义见 [5]）。请注意，在 (a) 中，簇主要按场景类别形成；而在 (b) 中，由于场景上下文受到抑制，簇的组织方式更清晰地体现了语义类别。嵌入使用主成分分析和 t-SNE 进行降维，距离度量采用余弦相似度。

参数 α 控制全局场景上下文的去除程度，使分割嵌入能够更加以分割为中心，同时也可按需保留部分场景信息。如图 3b 所示，这种调整使嵌入按语义类别而非场景类别形成簇。计算完成后，这些调整后的分割嵌入 $\tilde{\mathbf{e}}^{\mathbf{M}_i}$ 会被存储起来，之后用于训练。

B. 模型架构与知识蒸馏

对于 DVEFormer，我们以 EMSAFormer [12] 为基础——这是一种最初为多种移动机器人任务设计的 RGB-D 编码器-解码器模型，例如全景分割（结合语义分割与实例分割）、实例方向估计和场景分类。本文仅关注语义分割分支，并在其架构基础上构建模型，以预测稠密的逐像素视觉嵌入。EMSAFormer 使用经过修改的 Swin-V2-T 主干网络 [37], , 并在 [12] 中对其进行了扩展，使其能够在单个编码器中联合处理 RGB 和深度输入。针对语义分割，EMSAFormer 探索了两种不同的解码器：轻量级、类似 SegFormer [38] 的基于 MLP 的解码器，以及 ESANet [9] 的解码器。鉴于基于 MLP 的解码器效率略高，且在大规模室内数据集上表现出色，我们以其为基础进行构建。与 EMSAFormer 的标准输出（生成具有 40 个通道的分割图，对应 40 个 NYUv2 语义类别）不同，我们的目标是为每个像素生成一个 768 维嵌入。为适应这一更高维度的输出，我们将解码器瓶颈的通道数从 128 增至 512，以确保具备足够的容量来捕获细粒度视觉细节。有关架构的更多详情，请参阅我们的 GitHub 仓库。

为了训练 DVEFormer 预测稠密的逐像素视觉嵌入，我们采用知识蒸馏策略。对于每个数据集样本，我们使用 Alpha-CLIP 离线提取教师嵌入 $\{\tilde{\mathbf{e}}^{\mathbf{M}_i}\}$（见 III-A 节）。随后，根据相应的分割掩码 $\{\mathbf{M}_i\}$（由全景标注得出）组合这些嵌入，形成教师体 $\mathbf{Y} \in \mathbb{R}^{H\times W\times D}$。由于分割掩码彼此唯一且不重叠，因此每个像素的教师目标都是明确的。鉴于基于 CLIP 的模型通常通过余弦相似度对齐视觉特征与文本特征，我们决定采用余弦距离作为损失。具体而言，将学生模型预测的嵌入 $\mathbf{Y} \in \mathbb{R}^{H\times W\times D}$ 与其对应的教师嵌入 $\mathbf{Y}$ 逐像素进行比较：

$$\mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{(x,y)\in\mathcal{P}} \left(1 - \frac{\mathbf{y}_{y,x} \cdot \hat{\mathbf{y}}_{y,x}}{||\mathbf{y}_{y,x}|| \cdot ||\hat{\mathbf{y}}_{y,x}||}\right) \tag{2}$$

其中，$\mathcal{P}$ 表示任意二值分割掩码所覆盖的像素集合，而 $\mathbf{y}_{y,\,x} \in \mathbb{R}^D$ 表示像素位置（x，y）处 $\mathbf{Y}$ 中的嵌入。以这种方式将学生模型的稠密输出与教师嵌入对齐，有助于保留与文本编码器输出的对齐关系，从而支持下游文本引导任务——不过，这取决于是否使用足够大的数据集进行训练。

教师模型每次处理一张 RGB 图像和一个阿尔法掩码，以生成一个分割区域嵌入；若要处理多个分割区域，则需要反复进行推理。而我们的学生模型从 RGB 和深度输入中生成密集逐像素嵌入，因此必须在无法获取掩码的情况下，隐式地学会识别分割区域。这种差异造成了领域差距，因为学生模型需要同时完成分割和嵌入。使用额外的

深度输入可以提供几何线索——例如物体边界——从而有助于弥合这一差距，并改进学生模型预测。

C. 稠密视觉嵌入的应用

我们的模型可预测稠密嵌入，从而编码可用于各种下游应用的视觉信息。例如，由于我们的训练目标使预测嵌入与 Alpha-CLIP 的视觉-文本联合空间对齐，因此可以利用提供的文本编码器将描述感兴趣对象的提示词（或多个提示词）嵌入。随后，可通过余弦相似度将每个像素的嵌入与文本嵌入进行比较，相似度较高的像素表示与提示词更为匹配（见图 2a）。进一步对相似度进行阈值处理，即可得到精细的分割掩码。无需额外训练，也可以将预测嵌入转换为闭集预测，以实现经典语义分割。一种方法是为每个预定义的语义类别生成具有代表性的嵌入。具体而言，可以采用基于提示词的策略（例如，“一张{语义类别名称}的图片”）来获得类别嵌入（见图 2b），也可以计算训练集上属于同一语义类别的所有分割片段的均值嵌入（称为视觉均值，见图 2c）。在这两种情况下，逐像素分配都是通过选择余弦相似度最高的类别（argmax）来完成的。对于需要在特定数据集上取得更强性能的应用，可以使用线性探测策略 [52]（见图 2d）。在这种方法中，会额外训练一个线性层，同时冻结模型的其余部分，将单个像素的 768 维嵌入直接映射到预定义类别集合，从而使表示适应目标数据集。尽管这需要一些额外训练，但无需重新训练整个模型，因此是一种快速且轻量的适配方法。

除了基于单张图像的分类任务之外，密集嵌入还可直接用于其他机器人应用，例如三维建图和导航。在这些应用中，相较于离散化为选定的预定义类别，保留场景的连续表示更具优势。

上述应用凸显了我们的方法在应用栈中的灵活性。某些应用可能更倾向于使用闭集预测，以确保一致性或与现有系统集成；而其他应用则可能受益于利用对齐的文本嵌入或视觉参考嵌入（例如视觉均值）进行检索。借助 DVEFormer，所有这些功能都可以基于同一密集嵌入输出无缝实现，且额外开销很小。

IV. 实验

我们的实验评估在多个广泛使用的室内分割数据集上进行，包括 NYUv2 [6], SUN RGB-D [22], 和 ScanNet [23]。我们首先聚焦于规模较小的 NYUv2 数据集，并将模型的输出仅与针对该数据集提取的教师嵌入对齐，以验证我们方法的可行性。我们在各种设置下开展经典语义分割的多项闭集实验，以确定表现良好的配置。在此基础上，考虑到我们预计模型能从更多数据中获益，我们按照 [4]中的训练方案扩大训练规模，将 NYUv2、SUN RGB-D、ScanNet 和 Hypersim [58], 这几个类别范围相似的数据集合并起来。此外，我们还利用 ADE20k 数据集 [59] 进一步增加训练数据量[2]。尽管该数据集的类别范围不同，也并非专为室内环境设计，但我们的方法不受固定语义类别的限制，因此可以轻松地将额外数据纳入训练集，这是相较于闭集方法的一项优势。我们在多个室内数据集上评估性能，并将我们的方法与其他最先进方法进行比较。最后，我们展示定性结果，以凸显我们方法的优势。

A. 实现细节

我们的实现基于 EMSAFormer [12]代码库构建。我们采用了与 EMSAFormer 相同的训练和数据增强流程。我们使用 ImageNet 预训练权重初始化编码器，并采用 640×480 的输入分辨率。训练在一块 NVIDIA A100（40 GB）上进行；不过，由于某些配置的内存需求较高，我们将批量大小降至 4。与 EMSAFormer 不同，我们使用 bfloat16 自动混合精度（AMP）来加速训练并降低内存消耗。我们使用 AdamW [61] 作为优化器，并对学习率$\{4\mathrm{e}{-5},\ 3\mathrm{e}{-5},\ 2\mathrm{e}{-5},\ 1\mathrm{e}{-5},\ 9\mathrm{e}{-6}\}$进行网格搜索，同时额外采用单周期学习率调度器。根据数据集设置，模型训练 500 或 250 个 epoch。在线性探测实验中，除输出端新增的线性层外，整个模型均被冻结；该线性层将 768 维输出嵌入投影到固定的语义类别集合。该层以 -3 的学习率训练初始训练总 epoch 数的 $1/10$，其余训练流程保持不变。为生成教师嵌入，我们使用了基于原始 CLIP-L/14@336 检查点 [52] 、并在 GRIT-20M [62] 上由 Alpha-CLIP 团队微调的 Alpha-CLIP [36]模型。对于涉及文本编码器的实验，我们采用了与所用 Alpha-CLIP 检查点关联的文本编码器。有关数据流程和训练参数的更多细节，请参阅我们的 GitHub 仓库。

B. 评估与指标

据我们所知，除了数据集的封闭集标注外，目前没有适用于室内应用的真实标注数据，可用于评估我们模型的各个方面。因此，我们通过报告平均

[2]由于 ADE20k 缺少深度图像，我们使用 Depth- Anything-V2 [60] 对其进行了估计。我们采用了基于 ViT-L 的模型，该模型使用来自 Hypersim 的合成图像进行训练，用于室内深度估计。

交并比（平均交并比），评估第 III-C 节中详述的三个互补场景下的性能。

基于文本： 根据余弦相似度，将预测嵌入与相应类别的文本嵌入（例如，“一张架子的图片”）匹配。这种方法尤其适用于事先未知分割类别，或类别可能取决于用户输入的情况。

视觉均值： 通过将预测嵌入与根据训练集得出的每个类别的均值嵌入进行比较，来完成像素分配。这种设置反映了感兴趣对象被多次观测到的情形。其均值嵌入可以存储起来，作为参考嵌入使用，而无需额外训练。

线性探测： 训练一个线性层，将密集输出映射到固定的语义类别集合，从而稳健地衡量闭集性能。

C. NYUv2 上的结果

我们首先在 NYUv2 上评估所提出的方法，并详细阐述该方法的关键方面。结果汇总于图 4。图中的真实标注（GT）性能指将目标嵌入$\mathbf{Y}$作为预测结果，然后计算平均交并比。我们报告这一结果以供参考，并将其作为理论上限。不过需要注意的是，GT 性能假设事先已获得完美的分割掩码；与评估预测的密集嵌入相比，这属于不同的评估范畴。结果表明，使用$\alpha = 0.65$ 移除全局场景上下文、聚焦于单个分割区域，与使用 $\alpha = 0$ 相比，能够持续提升平均交并比。此外，与 RGB 输入相比，RGB-D 输入带来了更高的性能，验证了将深度数据作为室内分割的额外输入所带来的益处。比较各种封闭集评估策略时，线性探测取得了最高的平均交并比，甚至超过了 EMSAFormer 的表现。使用根据训练数据计算的视觉均值嵌入，其表现始终优于文本提示，这表明文本描述在捕捉所需视觉嵌入方面可能不够理想，尽管仍能取得适当的结果。此外，如 SUN RGB-D 数据集 [22],中所做的那样，从 NYUv2 中排除三个模糊类别（其他结构、其他家具和其他属性），能够提升基于文本和视觉均值方法的平均交并比（见图 4）。这与图 3 的结果一致：这些类别未能形成明显的簇，表明它们的定义不够明确，且难以区分。

尽管预测 768 维的逐像素嵌入比 EMSAFormer 的 40 通道输出计算成本更高，但我们的模型在 NVIDIA Jetson AGX Orin 上仍满足实时性要求（$\geq$ 20 帧每秒）。此外，对于需要更高吞吐量的应用，我们的实验表明，将上采样比例降低至输入分辨率的 $1/4$ 可显著提高帧每秒数（即从 26.3 帧每秒提升至 77.0 帧每秒），而准确率仅略有下降，且性能仍优于 EMSAFormer。

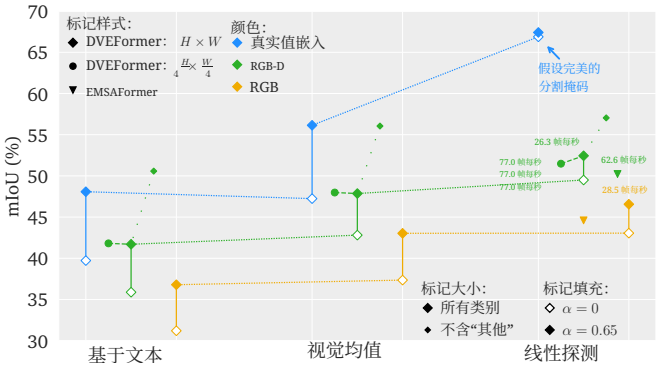


图 4. NYUv2 测试划分上不同配置的经典语义分割闭集结果（请参见所有图例和 x 轴）。请注意，计算 GT 性能时使用了真实标注掩码，因此假设分割是完美的。因此，学习模型与 GT 之间的差距并非完全由知识蒸馏过程造成。吞吐量是在 NVIDIA Jetson AGX Orin（JetPack 6.2、TensorRT 10.3、Float16）上测得的。关于指标和实验设置的更多细节见第 IV-C 节。

D. 最先进方法比较

表 I 展示了我们的模型在组合数据集上训练后（参见第 IV 节引言），在最常用的真实室内分割数据集上的评估表现。尽管我们的主要目标是预测密集文本对齐视觉嵌入，而非最大化闭集性能，但我们的线性探测方法取得了有竞争力的结果，始终优于 EMSAFormer，并达到其他同等规模模型的水平。值得注意的是，除表 I 最后一行的 EMSANet 外，其他竞争方法均在目标数据集上进行了端到端微调，而我们只在冻结模型之上训练了额外的线性探测层。这不仅简化了训练过程，也表明我们的蒸馏嵌入只需极少的进一步优化即可支持分割。此外，基于文本和视觉均值的平均交并比（表 I 中以灰色标示）略低于线性探测所得结果，但考虑到它们所提供的灵活性，表现依然不错。我们的实验还表明，减少输出端上采样、从而加快推理速度的模型，在所有数据集上均能保持强劲表现，与全分辨率版本相比仅有小幅下降。总的来说，这些结果表明，我们的方法在多种室内数据集上实现了有竞争力的分割性能，同时兼具灵活性和效率，适用于真实应用场景。

V. DVEFORMER 在三维建图中的应用

为进一步展示我们方法的适用性，我们将预测得到的稠密视觉嵌入集成到现有的三维正态分布变换（NDT）建图框架中 [4],，该框架最初设计于 [16] ，用于融合语义分割。图5展示了一个室内场景及其对应的三维NDT地图，其中每个单元存储的是来自多次观测的累积视觉嵌入，而非逐类别直方图计数。该地图可用于两种应用。对于经典语义分割，

表 I：各种数据集上闭集语义分割的最先进方法比较。对于 DVEFormer，我们以黑色报告线性探测结果，以灰色报告基于文本的/视觉均值结果。对于 NYUv2 和 SUN RGB-D，我们报告测试划分上的指标。对于 ScanNet，我们纳入验证集和隐藏测试集的结果。吞吐量是在 NVIDIA Jetson AGX Orin（JetPack 6.2、TensorRT 10.3、Float16）上测得的，输入分辨率为 640×480。图例： *–采用测试时增强的结果； ▽–预计会更低； **–输入分辨率为 1024×768，与指定分辨率不同。

模型	骨干网络	额外信息 训练数据	NYUv2 Test 平均交并比↑	SUN RGB-D Test 平均交并比↑	ScanNet 验证集 平均交并比↑	ScanNet（20 个类别） 验证集 平均交并比↑	ScanNet（20 个类别） Test 平均交并比↑	FPS↑
ShapeConv [8]	ResNet-101		47.4	47.6	-	-	-	- ▽
	ResNext-101		50.2	-	-	-	-	- ▽
Omnivore [11]	Swin-T	✓	49.7	-	-	-	-	- ▽
	Swin-B	✓	54.0	-	-	-	-	- ▽
MultiMAE [15, 33]	ViT-B	✓	56.0	51.1	-	-	-	- ▽
CMX [13]	2x MiT-B2		54.1	49.7*	-	-	61.3	33.37
	2x MiT-B4		56.0	52.1*	-	-	-	18.76
CMNeXt [14, 15]	2x MiT-B4		56.9	51.9	-	-	-	14.84
DFormer [15]	DFormer-S	✓	53.6	50.0	-	-	-	55.03
	DFormer-B	✓	55.6	51.2	-	-	-	35.11
EMSAFormer [12]	SwinV2-T-128		50.23	48.61	-	64.75	56.4	62.63
EMSANet [5]	2x ResNet34-NBt1D		51.15	48.39	-	-	-	78.19
EMSANet [4, 5]	2x ResNet34-NBt1D	✓	56.55	49.31	50.84	66.96	60.00	38.57**
DVEFormer $(H/4 \times W/4)$ (我们的方法)	SwinV2-T-128	✓	56.25 43.45 / 50.02	50.11 44.01 / 45.66	48.75 33.58 / 39.59	67.55 50.32 / 56.75	- -	77.0
DVEFormer（我们的方法）	SwinV2-T-128	✓	57.7 44.07 / 50.31	50.99 44.56 / 46.25	49.11 33.77 / 39.59	67.72 50.16 / 56.57	62.6 -	26.30

首先对每个单元的嵌入进行归一化，然后使用我们实验中的线性探测权重进行分类，并赋予其置信度最高的类别。此外，文本查询通过计算每个单元的嵌入与相应文本嵌入之间的余弦相似度，检索超出固定类别范围的物体。总体而言，这些结果表明，我们按图像预测的视觉嵌入能够无缝集成到三维建图等下游应用中，从而提升整体灵活性。

有关真实应用场景的更多展示和示例，请参阅我们的 GitHub 仓库。

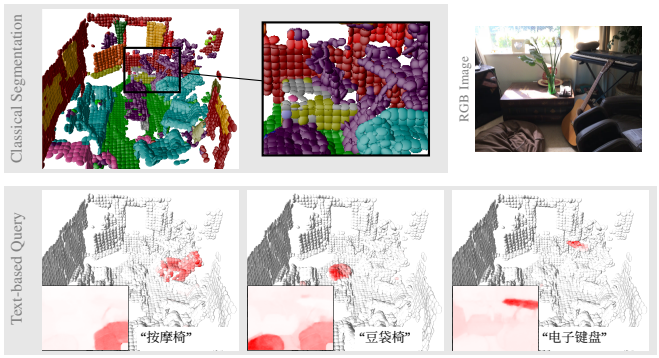


图5. ScanNet 验证场景（有效场景*0608*）的三维NDT地图建图结果。该地图展示了经典语义分割（通过线性探测）以及基于文本的查询（余弦相似度：红色表示高，白色表示低）。此外，还展示了同一文本查询下单个位姿对应的RGB图像。关于所示语义颜色的子集，请参见图3。

VI. 结论

本文提出了一种利用 RGB-D Transformer 高效预测密集文本对齐视觉嵌入的方法，称为 DVEFormer。我们不再采用固定预定义类别进行经典语义分割，而是使用知识蒸馏策略，以来自 Alpha-CLIP 的教师嵌入引导模型学习逐像素视觉嵌入。我们在 NYUv2、SUN RGB-D 和 ScanNet 上的评估表明，该方法不仅通过线性探测实现了具有竞争力的闭集性能，还支持灵活的自然语言查询，因此适用于各种下游应用。尽管输出维度更高，我们的方法仍能满足计算资源有限的移动平台上的实时推理要求，例如在 NVIDIA Jetson AGX Orin 上，我们的较大型模型达到 26.3 帧每秒，较小型模型达到 77.0 帧每秒。在未来的工作中，我们计划探索减少预测嵌入所需显存的方法，例如降低嵌入维度，或为每个掩码预测一个嵌入向量，同时预测二值分割掩码。此外，我们还计划在更大型数据集上进行训练，更全面地评估实例感知和建图性能，并将该方法扩展至全景分割。

参考文献

[1] S. B. Fischedick、K. Richter、T. Wengefeld等，“通过协作式远程临场机器人跨越距离——CO-HUMANICS 项目进展报告”，载于*ISR* 会议论文集，2023年。[2] S. A. Arboleda、S. Fischedick、C. Diao等，“机器人控制程度的变化对机器人媒介交流中临场感影响的探索性研究”，载于*RO-MAN* 会议论文集，2024年，第83–88页。[3] M. Conde、V. Mikhailova 和 N. D¨oring，“‘我觉得那个人就在这里’：老年人对远程临场机器人的态度、使用意向和需求”，《国际社会机器人学期刊》，第16卷，第7期，第1619–1639页，2024年7月。[4] D. Seichter、B. Stephan、S. B. Fischedick、S. Mueller、L. Rabes 和 H.-M. Gross，“PanopticNDT：高效且鲁棒的全景映射”，载于*IROS* 会议论文集，2023年，第7233–7240页。[5] D. Seichter、S. B. Fischedick、M. K¨ohler 和 H.-M. Gross，“室内环境的高效多任务 RGB-D 场景分析”，载于*IJCNN* 会议论文集，2022年。[6] N. Silberman、D. Hoiem、P. Kohli 和 R. Fergus，“基于 RGBD 图像的室内分割与支撑关系推断”，载于*ECCV* 会议论文集，2012年。[7] L.-C. Chen、Y. Zhu、G. Papandreou、F. Schroff 和 H. Adam，“用于语义图像分割的编码器-解码器与空洞可分离卷积”，载于*ECCV* 会议论文集，2018年，第801–818页。[8] M. Cao、H. Leng 和 Y. Lischinski、Dani d Li，“ShapeConv：用于室内 RGB-D 语义分割的形状感知卷积层”，载于*ICCV* 会议论文集，2021年。

Classical segmentation “按摩椅”

[9] D. Seichter、M. K¨ohler、B. Lewandowski 等，“用于室内场景分析的高效 RGB-D 语义分割”，载于 ICRA 会议论文集，2021年。[10] R. Bachmann、D. Mizrahi、A. Atanov 和 A. Zamir，“MultiMAE：多模态多任务掩码自编码器”，载于 ECCV 会议论文集，2022年。[11] R. Girdhar、M. Singh、N. Ravi 等，“Omnivore：适用于多种视觉模态的单一模型”，载于 CVPR 会议论文集，2022年。[12] S. Fischedick、D. Seichter、R. Schmidt、L. Rabes 和 H.-M. Gross，“基于 RGB-D Transformer 的高效多任务场景分析”，载于 IJCNN 会议论文集，2023年，第1–10页。[13] J. Zhang、H. Liu、K. Yang、X. Hu、R. Liu 和 R. Stiefelhagen，“CMX：基于 Transformer 的 RGB-X 语义分割跨模态融合”，ITS，第24卷，第12期，第14 679–14 694页，2023年。[14] J. Zhang、R. Liu、H. Shi、K. Yang、S. Reiß、K. Peng、H. Fu、K. Wang 和 R. Stiefelhagen，“实现任意模态语义分割”，载于 CVPR，2023年。[15] B. Yin、X. Zhang、Z.-Y. Li、L. Liu、M.-M. Cheng 和 Q. Hou，“DFormer：重新思考用于语义分割的 RGBD 表征学习”，载于 ICLR 会议论文集，2024年。[16] D. Seichter、P. Langer、T. Wengefeld 等，“室内环境的高效鲁棒语义映射”，载于 ICRA 会议论文集，2022年。[17] G. Narita、T. Seno、T. Ishikawa 和 Y. Kaji，“PanopticFusion：在背景类与物体类层级上进行在线体积语义映射”，载于 IROS 会议论文集，2019年。[18] L. Schmid、J. Delmerico、J. Sch¨onberger 等，“全景多 TSDF：一种灵活的表示方法，用于在线多分辨率体积映射和长期动态场景一致性”，载于 ICRA 会议论文集，2022年。[19] L. Gan、R. Zhang、J. W. Grizzle 等，“用于可扩展稠密语义映射的贝叶斯空间核平滑”，RAL，2020年。[20] M. Kolodiazhnyi、A. Vorontsova、A. Konushin 和 D. Rukhovich，“Oneformer3d：用于统一点云分割的单一 Transformer”，载于 CVPR 会议论文集，2024年，第20 943–20 953页。[21] D. Robert、H. Raguet 和 L. Landrieu，“基于超点图聚类的可扩展 3D 全景分割”，3DV 会议论文集，2024年。[22] S. Song、S. P. Lichtenberg 和 J. Xiao，“SUN RGB-D：RGB-D 场景理解基准套件”，载于 CVPR 会议论文集，2015年。[23] A. Dai、A. X. Chang、M. Savva 等，“ScanNet：带有丰富标注的室内场景 3D 重建”，载于 CVPR 会议论文集，2017年。[24] C. Zhou、C. C. Loy 和 B. Dai，“从 CLIP 中提取免费的稠密标签”，载于 ECCV 会议论文集，2022年。[25] J. Ding、N. Xue、G.-S. Xia 和 D. Dai，“解耦零样本语义分割”，载于 CVPR 会议论文集，2022年，第11 583–11 592页。[26] B. Xie、J. Cao、J. Xie、F. S. Khan 和 Y. Pang，“SED：用于开放词汇语义分割的简易编码器-解码器”，载于 CVPR 会议论文集，2024年，第3426–3436页。[27] S. Cho、H. Shin、S. Hong、A. Arnab、P. H. Seo 和 S. Kim，“CAT-Seg：用于开放词汇语义分割的代价聚合”，载于 CVPR 会议论文集，2024年，第4113–4123页。[28] M. Bucher、T.-H. Vu、M. Cord 和 P. P´erez，“零样本语义分割”，NeurIPS 会议论文集，第32卷，2019年。[29] Y. Xian、S. Choudhury、Y. He、B. Schiele 和 Z. Akata，“用于零标签与少标签语义分割的语义投影网络”，载于 CVPR 会议论文集，2019年，第8256–8265页。[30] X. Chen、S. Li、S.-N. Lim、A. Torralba 和 H. Zhao，“基于嵌入调制的开放词汇全景分割”，载于 CVPR 会议论文集，2023年，第1141–1150页。[31] S. Jiao、Y. Wei、Y. Wang、Y. Zhao 和 H. Shi，“学习用于零样本分割的掩码感知 CLIP 表征”，NeurIPS 会议论文集，第36卷，第35 631–35 653页，2023年。[32] Y. Liu、S. Bai、G. Li、Y. Wang 和 Y. Tang，“基于语义辅助校准的开放词汇分割”，载于 CVPR 会议论文集，2024年，第3491–3500页。[33] J. Guo、Q. Wang、Y. Gao、X. Jiang、S. Lin 和 B. Zhang，“MVP-Seg：用于开放词汇语义分割的多视图提示学习”，载于中国模式识别与计算机视觉大会（PRCV）。Springer，2023年，第158–171页。[34] G. Hinton、O. Vinyals 和 J. Dean，“蒸馏神经网络中的知识”，载于 NeurIPS 会议论文集（研讨会），2015年。[35] A. Vaswani、N. Shazeer、N. Parmar 等，“注意力就是你所需要的一切”，NeurIPS 会议论文集，第30卷，2017年。[36] Z. Sun、Y. Fang、T. Wu、P. Zhang、Y. Zang、S. Kong、Y. Xiong、D. Lin 和 J. Wang，“Alpha-clixp：一款聚焦于任意目标的 CLIP 模型”，载于 CVPR 会议论文集，2024年，第13 019–13 029页。[37] Z. Liu, H. Hu, Y. Lin 等，《Swin Transformer V2：扩展容量与分辨率》，发表于 CVPR 会议论文集，2022年。[38] E. Xie, W. Wang 等，《SegFormer：采用 Transformer 进行语义分割的简单高效设计》，发表于 NeurIPS 会议论文集，2021年。[39] E. Einhorn 和 H.-M. Gross，《基于 NDT 地图的通用 2D/3D 同步定位与建图，适用于长期应用》，发表于 ECMR 会议论文集，2013年。[40] H. Wang、Y. Zhu、H. Adam、A. Yuille 和 L.-C. Chen，《MaX-DeepLab：基于掩码 Transformer 的端到端全景分割》，发表于 CVPR 会议论文集，2021年，第 5463–5474 页。[41] B. Cheng、A. G. Schwing 和 A. Kirillov，《语义分割并非只需逐像素分类》，发表于 NeurIPS 会议论文集，2021年。[42] B. Cheng、I. Misra、A. G. Schwing 等，《用于通用图像分割的掩码注意力掩码 Transformer》，发表于 CVPR 会议论文集，2022年。[43] Z. Li、W. Wang、E. Xie 等，《全景SegFormer：通过 Transformer 深入探索全景分割》，发表于 CVPR 会议论文集，2022年，第 1280–1289 页。[44] C. Hazirbas、L. Ma、C. Domokos 和 D. Cremers，《FuseNet：通过基于融合的 CNN 架构将深度信息融入语义分割》，发表于 ACCV 会议论文集，2016年，第 213–228 页。[45] J. Jiang、L. Zheng、F. Luo 和 Z. Zhang，《RedNet：用于室内 RGB-D 语义分割的残差编码器-解码器网络》，arXiv 预印本 arXiv:1806.01054，2018年。[46] X. Hu、K. Yang、L. Fei 和 K. Wang，《ACNet：利用互补特征进行 RGBD 语义分割的注意力网络》，发表于 ICIP 会议论文集，2019年。[47] A. Valada、R. Mohan 和 W. Burgard，《用于多模态语义分割的自监督模型适应》，IJCV，2019年。[48] X. Chen、K.-Y. Lin、J. Wang 等，《通过分离与聚合门实现 RGB-D 语义分割的双向跨模态特征传播》，发表于 ECCV 会议论文集，2020年，第 561–577 页。[49] W. Wang 和 U. Neumann，《用于 RGB-D 分割的深度感知 CNN》，发表于 ECCV 会议论文集，2018年，第 144–161 页。[50] Y. Xing、J. Wang 和 G. Zeng，《可变形2.5D卷积：沿深度轴学习 RGB-D 场景解析的感受野》，发表于 ECCV 会议论文集，2020年，第 1–17 页。[51] A. Dosovitskiy、L. Beyer、A. Kolesnikov 等，《一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer》，发表于 ICLR 会议论文集，2021年。[52] A. Radford、J. W. Kim、C. Hallacy 等，《通过自然语言监督学习可迁移的视觉模型》，发表于 ICML 会议论文集，PMLR，2021年，第 8748–8763 页。[53] S. Subramanian、W. Merrill、T. Darrell 等，《ReCLIP：用于指代表达理解的强力零样本基线》，发表于 ACL 会议论文集，2022年5月。[54] K. Chen、X. Jiang、Y. Hu 等，《OvarNet：迈向开放词汇对象属性识别》，发表于 CVPR 会议论文集，2023年6月，第 23 518–23 527 页。[55] V. Sanh、L. Debut、J. Chaumond 和 T. Wolf，《DistilBERT：BERT 的蒸馏版本，更小、更快、更便宜、更轻量》，发表于 NeurIPS 会议论文集（研讨会），2019年。[56] X. Jiao、Y. Yin、L. Shang 等，《TinyBERT：通过蒸馏 BERT 实现自然语言理解》，arXiv 预印本 arXiv:1909.10351，2019年。[57] L. van der Maaten 和 G. Hinton，《使用 t-SNE 可视化数据》，JMLR，第 9 卷，第 86 期，第 2579–2605 页，2008年。[58] M. Roberts、J. Ramapuram、A. Ranjan 等，《Hypersim：用于整体室内场景理解的照片级真实感合成数据集》，发表于 ICCV 会议论文集，2021年。[59] B. Zhou、H. Zhao、X. Puig 等，《通过 ADE20k 数据集实现对场景的语义理解》，IJCV，第 127 卷，第 3 期，第 302–321 页，2019年。[60] L. Yang、B. Kang、Z. Huang 等，《Depth Anything V2》，发表于 NeurIPS 会议论文集，第 37 卷，第 21 875–21 911 页，2025。[61] I. Loshchilov 和 F. Hutter，《解耦权重衰减正则化》，发表于 ICLR 会议论文集，2019年。[62] Z. Peng、W. Wang、L. Dong、Y. Hao、S. Huang、S. Ma 和 F. Wei，《Kosmos-2：将多模态大语言模型 grounding 到现实世界》，arXiv 预印本 arXiv:2306.14824，2023年。