

城市级别的无人机占用网格感知

摘要

当前，占用网格预测在车辆自动驾驶领域发挥着重要作用，作为上游任务，为下游的决策规划提供了关键的可行性支持。目前，占用网格预测技术主要应用在车辆自动驾驶方面，然而，基于无人机的占用网格相关技术尚处于空白状态。为了填补这一空白，我们展开了深入的研究工作。首先，将 Urbanbis 仿真到 Unreal Engine 5 (Ue5) 上，接着使用 Airsim 进行数据采集。采集完成后，对数据进行细致的处理，使其转化为占用网格预测标签。在此基础上，我们精心开发出 Droneocc。与传统的 Surroundocc 相比，我们的 Droneocc 在性能上展现出显著优势，其交并比 (IOU) 比 Surroundocc 高了 2。这一成果表明，Droneocc 在无人机占用网格预测方面具有更高的准确性和可靠性。通过 Droneocc 的应用，有望为无人机自动驾驶领域带来新的突破和发展，为未来的空中交通和物流等领域提供更加安全、高效的解决方案。

关键词：占用网格预测；无人机自动驾驶；虚拟引擎；无人机

1 引言

随着自动驾驶技术的不断发展，环境感知系统已成为确保自动驾驶安全性的核心之一。尤其是在复杂的城市环境中，准确、实时地感知并理解周围的三维空间结构，对于自动驾驶系统的决策和规划至关重要。占用网格预测（Occupancy Grid Prediction）[\[6\]](#) [\[17\]](#)是环境感知中的重要任务之一，旨在通过为三维空间中的每个体素分配占用概率，构建详细的空间表示，进而推断物体、障碍物和空旷区域的分布。这一技术广泛应用于地面车辆的自动驾驶系统中，通过提供更高精度的环境建模，显著提高了系统的导航能力和安全性。

占用网格预测的传统方法主要集中于地面车辆的感知系统，通常依赖于激光雷达[\[4\]](#)、深度相机[\[16\]](#)等传感器获取周围环境的几何信息，并通过数据融合技术生成高精度的三维占用网格。然而，尽管激光雷达具有较高的精度，但其成本高昂且体积庞大，不适用于轻便、机动性要求高的无人机平台。因此，如何基于视觉传感器生成高效、低成本且高精度的占用网格预测，成为了无人驾驶领域中的一个研究难题。



图 1. 车辆驾驶中的占用网格预测

近来，随着虚拟仿真技术的发展，研究者们开始将城市级别的环境数据导入虚拟仿真平台，利用仿真环境生成占用网格数据。这一方法不仅能解决实际数据采集中的高成本问题，还能通过精确控制仿真环境中的各种参数，实现多样化的场景模拟和数据生成。

为了进一步填补现有无人驾驶数据集中的空白，本研究采用了 Urbanbis [18] 数据集，并将其仿真至 Unreal Engine 5 (Ue5) [12] 平台，利用 Airsim [13] 仿真无人机进行数据采集。通过这一仿真流程，我们成功生成了具有城市级别细节的无人机占用网格数据集，并开发了 Droneocc，一种专门针对无人机自动驾驶的占用网格预测模型。

尽管占用网格预测已在地面车辆自动驾驶中取得显著进展，但在无人机自动驾驶领域，尤其是在城市环境中的应用仍处于起步阶段。无人机的动态性和复杂的三维飞行环境使得其感知任务面临更多挑战。因此，发展专门针对无人机的占用网格预测模型，对于提升无人机在复杂城市环境中的感知能力具有重要意义。

本文的主要贡献包括：首先，通过将 Urbanbis 数据集导入虚拟仿真平台，构建了一个包含多种城市环境的无人机占用网格预测数据集；其次，我们提出了 DroneOcc 模型，并与传统的 SurroundOcc [17] 进行对比，证明了 DroneOcc 在准确性和鲁棒性上的优势。通过这一创新的占用网格预测方法，我们希望为无人机自动驾驶技术的进一步发展提供新的思路和解决方案，尤其是在未来城市空中交通和无人机物流等应用场景中，为提升飞行安全性和导航效率提供支持。



图 2. 无人机占用网格预测

2 相关工作

2.1 基于车辆驾驶的占用网格感知

占用网格预测在自动驾驶领域中最初的研究集中在地面车辆上。由于地面车辆的运动相对较为简单且常规感知传感器（如激光雷达和相机）能够较为精确地捕捉到环境信息，因

此相关的占用网格预测方法多用于车辆感知系统中。早期的研究工作主要依赖激光雷达 [10] [14] [2] [19] 或深度相机 [16] [5] 等单一传感器的数据来进行占用网格建模。然而，由于激光雷达的体积和成本问题，研究者们逐渐开始关注基于视觉的占用网格预测方法。

在视觉感知的背景下，许多基于深度学习的方法被提出以提升占用网格预测的精度和效率。例如，LiftSplat [11] 通过外积操作提升2D图像特征与其估计深度概率之间的上下文关系，从而生成更为精细的3D表示。这一方法被进一步扩展至目标检测领域，形成了BEVDet [5]，该方法通过生成鸟瞰图特征并结合注意力机制，显著提高了3D目标检测的效果。类似的，BEVDepth [8] 通过引入地面真值深度监督，对占用网格预测的深度信息进行进一步精细化。

随着注意力机制的发展，BEVFormer [9] 等方法采用了点可变形交叉注意力机制，将2D图像特征转化为鸟瞰图（BEV）视角的3D占用网格表示，这种方法能够更加灵活地处理动态场景中的物体。其他方法，如SparseOcc [15]，则利用稀疏体素解码器，有效地减少了空网格的计算量，进一步优化了模型的性能。

尽管这些方法在车辆自动驾驶中取得了显著进展，主要集中在复杂的城市街道和高速公路环境中，但它们通常假设感知目标相对固定且地面交通规则明确。因此，针对动态、三维变化复杂的空中交通环境，现有的占用网格预测技术仍然存在诸多挑战。

2.2 基于无人机的占用网格预测

尽管占用网格预测在自动驾驶领域，尤其是地面车辆的感知任务中取得了显著进展，但在无人机领域，尤其是在城市级别的环境中，相关研究仍面临诸多挑战。无人机的感知任务不仅要处理三维环境中的障碍物、空旷区域和复杂的动态变化，还需要应对飞行高度、速度和方向等因素的影响。因此，无人机的感知任务往往比地面车辆更为复杂。

目前，针对无人机的感知任务，大多数研究仍集中在2D任务，如城市图像街景分割和目标检测等。主要原因在于3D感知任务对数据的需求较大，而现有的无人机数据集在城市级别的3D标注数据上存在严重不足。大多数公开的数据集，如 nuScenes [1]、KITTI [3] 等，虽然为地面车辆提供了丰富的3D标注，但对于无人机环境，尤其是在复杂城市环境中的3D数据集仍显稀缺。这种数据匮乏导致了3D感知任务，特别是占用网格预测在无人机领域的应用受到了极大的限制。

在这种背景下，许多研究将重点放在2D [?] [20]任务上。例如YOLO-based Detection [7]（基于YOLO的目标检测）等方法已广泛应用于无人机图像处理任务中。这些方法通过分析城市街景的2D图像，提供了基础的物体分割、目标检测和障碍物识别等能力。尽管这些方法能够处理无人机的实时视觉输入，但它们通常只提供二维平面上的信息，难以全面捕捉场景中的深度信息和三维结构。

尽管如此，随着虚拟仿真技术的不断发展，一些研究者尝试通过仿真环境生成数据并结合深度学习模型进行3D占用网格预测。例如，Airsim和Unreal Engine 5平台为无人机提供了可控且高效的仿真环境，使得研究者能够生成高质量的3D感知数据。然而，由于城市环境的复杂性和飞行的动态性，如何从有限的2D数据中推导出可靠的三维占用网格仍然是一个亟待解决的问题。

本研究通过将Urbanbis数据集转化为适用于无人机的仿真数据，提出了Droneocc，并采用了先进的视觉感知技术来生成无人机的占用网格预测。与传统的2D图像处理方法不同，我

们的方法尝试跨越从2D到3D感知的鸿沟，推动无人机占用网格预测技术的研究进展，尽管目前3D数据稀缺，我们的目标是通过创新的模型设计和数据生成方式，为未来无人机感知任务提供更为可靠的解决方案。

3 本文方法

3.1 数据整合

3.1.1 模型仿真

为了实现准确的占用网格预测，我们首先将 Urbanbis 数据集中的城市环境模型仿真到虚拟引擎 Unreal Engine 5 (Ue5) 中。Urbanbis 数据集包含了丰富城市级别的三维环境数据，但这些数据原本并不直接适用于无人机感知任务。因此，我们通过虚拟仿真环境对这些数据进行了适配，使其能够模拟真实世界中的城市飞行环境。虚拟引擎为我们提供了一个灵活且可控制的测试平台，能够在此平台上生成复杂的城市环境，包括道路、建筑、交通工具以及各种动态障碍物。

接下来，我们使用 Airsim (由微软提供的无人机仿真平台) 进行数据采集。Airsim能够模拟无人机的飞行，支持包括相机、激光雷达和GPS等传感器的集成。在这一过程中，我们特别注重模拟无人机的实际飞行视角，通过设定无人机的飞行轨迹，采集不同高度、不同角度的图像数据以及其他相关传感器数据（如位姿、速度等）。这些数据将为后续的占用网格预测任务提供丰富的输入信息。

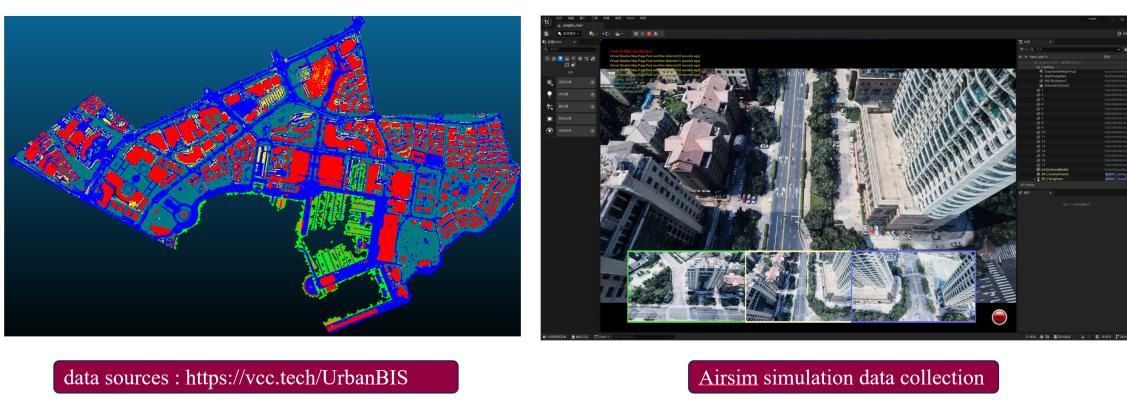


图 3. 将Urbanbis仿真到虚拟引擎中

关于无人机的设置，如图 4 所示，在本文中，将五个摄像头安装在仿真无人机上。其中一个摄像机正对着无人机下方，另外四个摄像机分别对应着无人机的前后左右四个视角。这四个视角均向下倾斜 60 度。每个相机的视场角 (FOV) 为 90 度，采集的图片大小为 1500×900。



图 4. 无人机设置

3.1.2 坐标系换算

在数据采集之后，下一步是将收集到的多源数据转换为统一的占用网格标签。由于 Urbanbis 数据集的坐标系与 Airsim 仿真环境中的坐标系存在一定差异，因此必须进行坐标系转换，以确保数据的准确性和一致性。

首先，我们通过 Airsim 采集的城市图像及位姿信息，记录下每一时刻无人机的飞行位置和姿态。然后，利用三点奇异值分解（SVD）方法，确定刚性矩阵，将无人机的位姿从仿真环境坐标系转换到 Urbanbis 模型的坐标系中。这一过程的目的是确保我们的数据能够与 Urbanbis 中的环境模型对齐，从而生成对应的占用网格标签。

接着，采用相似的坐标转换方法，将 Urbanbis 模型中的位姿转换到点云坐标系中。这一过程中，我们提取了与当前时刻无人机飞行位置相关的点云数据，并通过几何变换将点云从原始坐标系转换到新的坐标系下。此时，点云数据能够准确反映出当前飞行位置的三维空间结构，并且与相应的图像数据保持一致。

接下来，我们将点云坐标系转换到伪局部坐标系中。由于在实验过程中，接下来，我们将点云坐标系转换到伪局部坐标系中。由于在实验过程中，Urbanbis 坐标系与局部坐标系之间存在一定的固定偏差，因此我们通过利用出发点零位移时刻的坐标系与 Urbanbis 坐标系进行三点奇异值分解，构建伪局部坐标系，以补偿这一偏差。这样处理后，我们能够更加精准地确定每个体素的占用状态，并将其转化为占用网格标签。

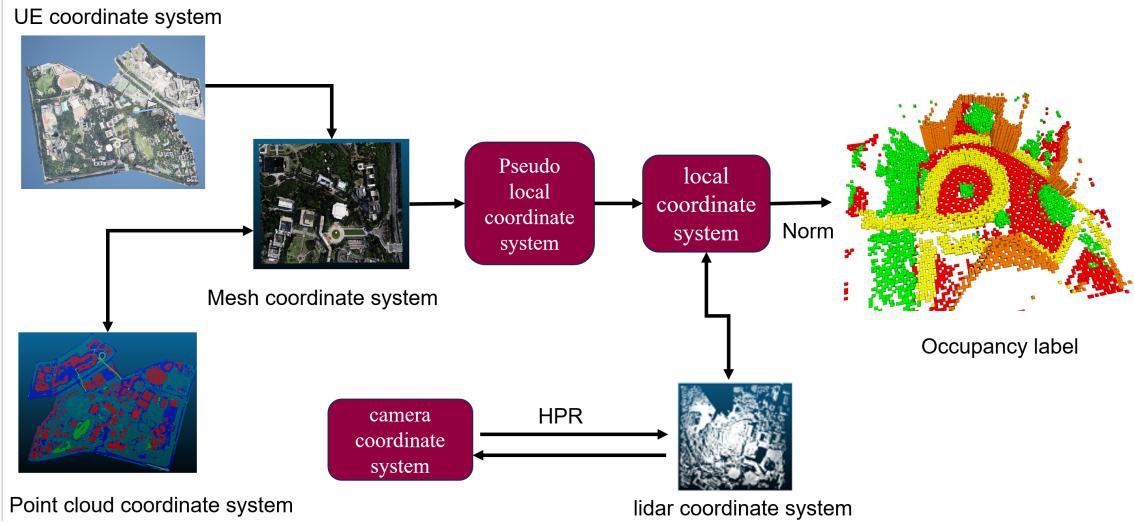


图 5. 坐标系换算

3.1.3 数据整理

经过上述步骤后，我们得到的每一帧数据包括了图像数据、点云数据以及对应的占用网格标签。这些数据由不同视角的相机提供（包括正下方视角和四个倾斜的视角），确保我们能够从多个角度捕捉到环境中的重要特征。每组数据都包含了飞行时的位姿信息，确保了数据的时间同步和空间一致性。

为了进一步提高模型训练的质量和多样性，我们针对 Urbanbis 中的六个城市（分别为 Yuehai、Yingrenshi、Lihu、Wuhu、Longhua 和 Qingdao）展开了全面的数据采集。最终，我们共采集了 55499 组数据，每组数据均涵盖五个视角的信息，确保了训练数据的丰富性和代表性。

Total: 55499 groups each with 5 perspectives							
Name	UE_map	Label_map	Number	Name	UE_map	Label_map	Number
Yuehai			11119	Wuhu			14854
Yingrenshi			798	Longhua			9706
Lihu			9741	Qingdao			9281

图 6. Airsim采集的数据统计

如图7所示，对于每组数据我们采集了如下信息。如图7展示了从仿真环境中采集的无人机数据的详细属性信息，这些属性涵盖了无人机的基础信息（如飞行位置和时间戳）、姿态信息（四元数和欧拉角）、图像与坐标系转换关系（相机内参、深度图和坐标系转换矩阵）

以及时间序列信息（前后帧时间戳）等。此外，还包括占用网格标签和深度图等关键感知数据。通过这些多维信息的整合，无人机的飞行位置、姿态及感知结果能够被准确地捕获，为占用网格预测任务提供全面且高质量的输入。

表 1. 可获取的属性信息

variable	analysis
Vehicle_Name	Drone name
Time_stamp	Time stamp
POS_X, POS_Y, POS_Z	flight position
Q_W, Q_X, Q_Y, Q_Z	Quadruple
Images	Images path
Roll, Pitch, Yaw	Euler angle
cam2pix	Camera coordinate system to depth map
internal	Camera internal reference
C2W	Convert camera coordinate system to world coordinate system
previous_time	The timestamp of the previous frame
next_time	The timestamp of the next frame
Occ_GT	Occupancy label
Depth	Depth

表1对比了传统车辆与无人机在占用网格数据中的显著差异。传统车辆的占用网格更适用于动态场景，主要关注二维平面信息（如行人和车辆），其体素规模为 $200 \times 200 \times 16$ ，分辨率为 0.5m/体素0.5m/体素0.5m/体素。而无人机由于需要适应空中三维环境的复杂感知，采用了 $96 \times 64 \times 96$ 的体素规模，分辨率为 2m/体素2m/体素2m/体素。同时，无人机的占用网格更注重三维空间（X×Y×Z）的信息捕捉，能够支持其在复杂城市环境中的精准导航和路径规划。相比传统车辆，无人机的占用网格预测具备更高的维度复杂性和更广的感知范围。

表 2. 传统车辆与无人机的占用数据区别

	Traditional occupancy	drone occupancy
scene	Dynamic Scene	Static scene
Voxel size	$200 \times 200 \times 16$	$96 \times 64 \times 96$
Real size	$100 \times 100 \times 8$ (0.5m/voxel)	$192 \times 96 \times 192$ (2m/voxel)
Mainly considering size	$X * Y$	$X * Y * Z$

3.2 模型设计

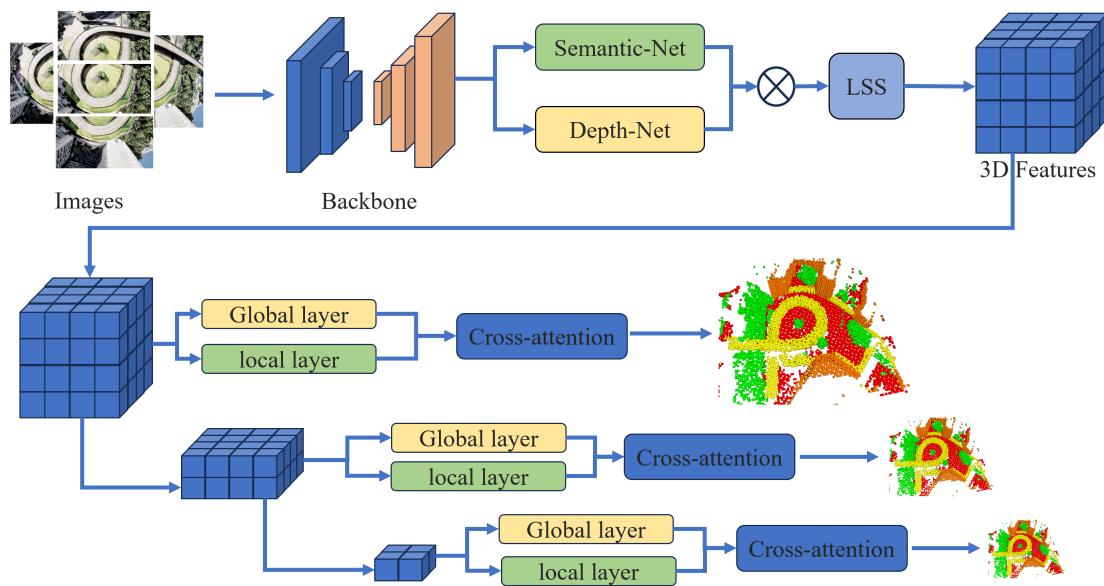


图 7. 模型设计

DroneOcc的输入为无人机拍摄的五张环视图像（分辨率为 1500×900 ），覆盖飞行场景的不同视角，包含丰富的环境信息。输入图像首先通过采用预训练的ResNet-50主干网络（Backbone）进行特征提取，生成形状为 $C \times H \times W$ 的高维度视觉特征，其中 C 是通道数， H 和 W 分别为特征图的高度和宽度。随后，这些特征被送入两个独立的子网络：语义网络（Semantic-Net）和深度网络（Depth-Net），分别提取语义特征和深度特征。语义网络基于语义分割任务框架，通过卷积操作提取多尺度特征，最终输出形状为 $C_s \times H \times W \times D$ 的语义特征图，其中 C_s 表示语义类别的数量， D 是深度维度，语义特征反映每个体素属于不同语义类别（如建筑物、道路）的概率分布。深度网络则对初始特征图进行扩展，生成形状为 $C_d \times H \times W \times D$ 的深度特征，其中 C_d 表示深度通道数，用于刻画空间中不同深度范围的信息分布。随后，语义特征和深度特征通过 LSS（Lift, Splat, Shoot）方法进行融合，利用点乘操作将二维特征提升为三维特征，生成最终形状为 $C_s \times H \times W \times D$ 的三维占用特征表示（Occupancy）。这种设计能够联合描述三维空间中的语义类别和深度分布，为后续解码阶段的多尺度预测提供高质量的三维特征输入。

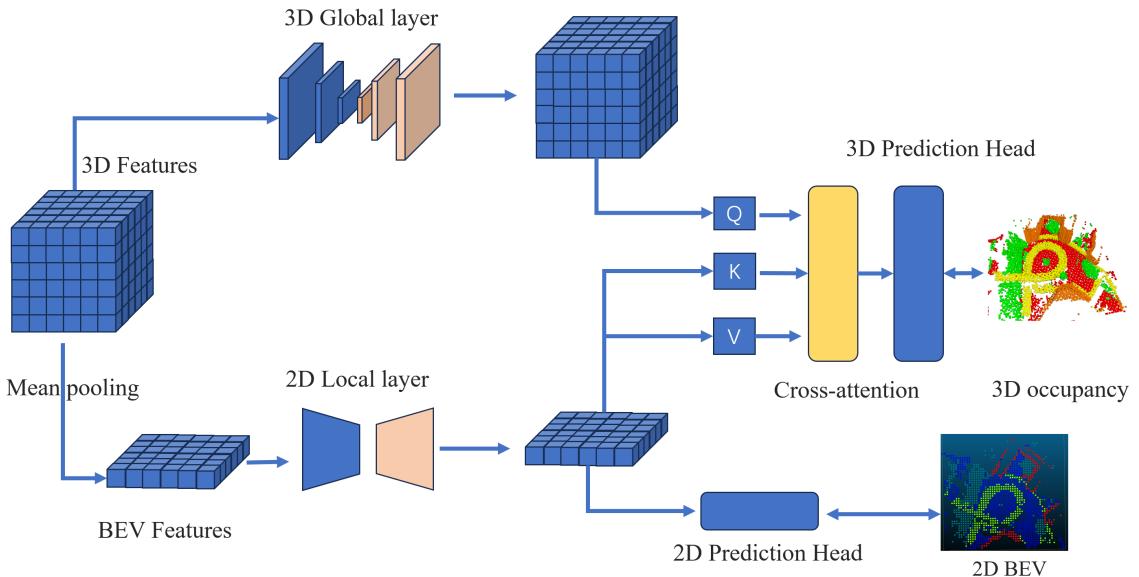


图 8. 全局特征与局部特征融合

在解码阶段，我们采用了局部解耦的思想，通过分离三维全局特征和二维局部特征的学习，有效提升了推理效率。输入的三维占用特征（Occupancy）具有形状为 $196 \times 16 \times 196$ ，其中 16 表示高度方向上的深度维度， 196×196 表示水平面的分辨率。首先，三维占用特征经过三维全局卷积层提取全局信息，生成形状为 $C3D \times 196 \times 16 \times 196$ 的三维全局特征（3D Occupancy）。与此同时，通过平均池化（Mean Pooling）操作，特征沿高度维度 16 进行池化，生成二维鸟瞰视角特征（BEV 特征），其形状为 $CBEV \times 196 \times 196$ ，用于学习水平面上的局部信息。接下来，BEV 特征被送入二维解码层进行解码，输出二维细粒度的特征表示。二维解码层的设计可以有效降低高度维度对模型的学习负担，使网络更专注于捕捉水平面的局部特征。

随后，为了融合三维全局信息和二维局部信息，模型通过 Cross-Attention[23] 模块实现特征的交互学习。在 Cross-Attention 模块中，三维全局特征（3D Occupancy）被用作查询（Query, Q），而二维 BEV 特征被用作键（Key, K）和值（Value, V）。具体而言，Cross-Attention 模块利用三维特征中的全局信息，关联并整合二维 BEV 特征中的细粒度信息，以提升对高度动态环境的感知能力。最终，经过预测头（Prediction Head），模型同时输出形状为 $196 \times 16 \times 196$ 的三维占用网格（3D Occupancy）和形状为 $CBEV \times 196 \times 196$ 的二维鸟瞰图（2D BEV）。这种局部解耦的设计通过将三维和二维特征分别解码，并在 Cross-Attention 中进行有效融合，使模型能够更加轻松地捕捉全局与局部的互补信息，同时显著提高预测的精度和效率。

3.3 损失函数设计

为了优化多尺度预测结果，本模型的总损失函数设计为多尺度加权组合的形式，定义如下：

$$L_{total} = L_1 + L_2 + L_3 \quad (1)$$

其中， L_1 、 L_2 、 L_3 分别对应三个不同尺度的预测输出，从低分辨率到高分辨率。每个尺度的损失函数 L_i 定义为二维鸟瞰图损失 (L_{BEV}) 和三维占用网格损失 ($L_{Occupancy}$) 的加和，

表达式如下：

$$L_i = L_{BEV} + L_{Occupancy} \quad (2)$$

二维鸟瞰图损失 L_{BEV} 用于约束二维BEV（Bird's Eye View）预测结果，目标是对水平平面上的特征进行细粒度优化。通过与地面真值（Ground Truth）对比，该损失引导网络学习二维水平平面的场景结构和语义信息。

三维占用网格损失 $L_{Occupancy}$ 聚焦于三维空间中每个体素的预测准确性，定义如下：

$$L_{Occupancy} = L_{ce} + L_{spatial} \quad (3)$$

其中， L_{ce} 表示语义损失（Semantic Loss），采用交叉熵损失函数（Cross-Entropy Loss）计算每个体素的预测语义概率与真实分布之间的差异，公式为：

$$L_{ce} = - \sum_{c=1}^C y_c \log(\hat{y}_c) \quad (4)$$

其中， C 为语义类别数， y_c 为真实语义标签， \hat{y}_c 为预测的语义概率。

$L_{spatial}$ 表示空间一致性损失（Spatial Consistency Loss），用于约束占用网格在三维空间中的几何一致性和结构化特征。通过对邻域体素间的平滑性进行约束， $L_{spatial}$ 能够提升占用网格预测的连贯性。

4 复现细节

4.1 实验环境搭建

为了验证本研究提出的模型，我们搭建了完整的实验环境，包括硬件和软件配置。硬件环境上，实验采用四张 NVIDIA A6000 GPU 和 Intel(R) Xeon(R) Gold 6430 CPU 的高性能计算服务器，GPU 拥有强大的计算能力，能够高效支持大规模数据的训练和推理。软件环境方面，模型训练基于 PyTorch 1.10.1 深度学习框架，支持 CUDA 11.3 加速，操作系统为 Ubuntu 22.04，保证了训练过程的高效性与稳定性。

在模型训练中，我们使用 AdamW 优化器，其具有良好的正则化效果，能够显著提升模型的泛化能力。训练总轮数设置为 24 epochs，通过批次更新优化模型参数，具体的超参数设置包括学习率衰减和权重初始化策略，确保模型在不同尺度上的特征学习达到最佳效。

4.2 创新点

提出了一种局部解耦的解码器结构，通过将三维全局特征与二维局部特征分别解码的方式，优化了特征学习效率。具体来说，三维全局特征通过全局卷积层学习三维空间的占用信息，而二维BEV特征通过平均池化降低高度维度的复杂度，并利用二维解码器重点捕捉水平面的细粒度特征。通过Cross-Attention模块在三维全局特征和二维局部特征之间进行有效融合，该设计不仅减轻了模型的计算负担，还显著提升了占用网格预测的精度。

构建了包含三层尺度的多任务损失函数 $L_{total} = L1 + L2 + L3$ ，针对不同分辨率的预测结果分别进行优化。在每个尺度中，损失函数由二维BEV损失 L_{BEV} 和三维占用网格损失 $L_{Occupancy}$ 组

成。三维占用网格损失进一步结合了语义交叉熵损失 L_{ce} 和空间一致性损失 $L_{spatial}$ ，从语义分类和几何一致性两方面优化模型性能。这种多尺度损失设计，保证了模型在不同分辨率下都能生成高质量的预测结果。

基于Urbanbaits数据集，利用Unreal Engine 5 (UE5)和Airsim平台模拟无人机飞行，生成了城市级别的高质量三维占用网格数据集。该数据集包含多视角图像、深度信息和精细化的占用网格标签，填补了无人机占用网格预测领域数据稀缺的空白，为后续研究提供了重要支撑。

5 实验结果分析

5.1 结果分析

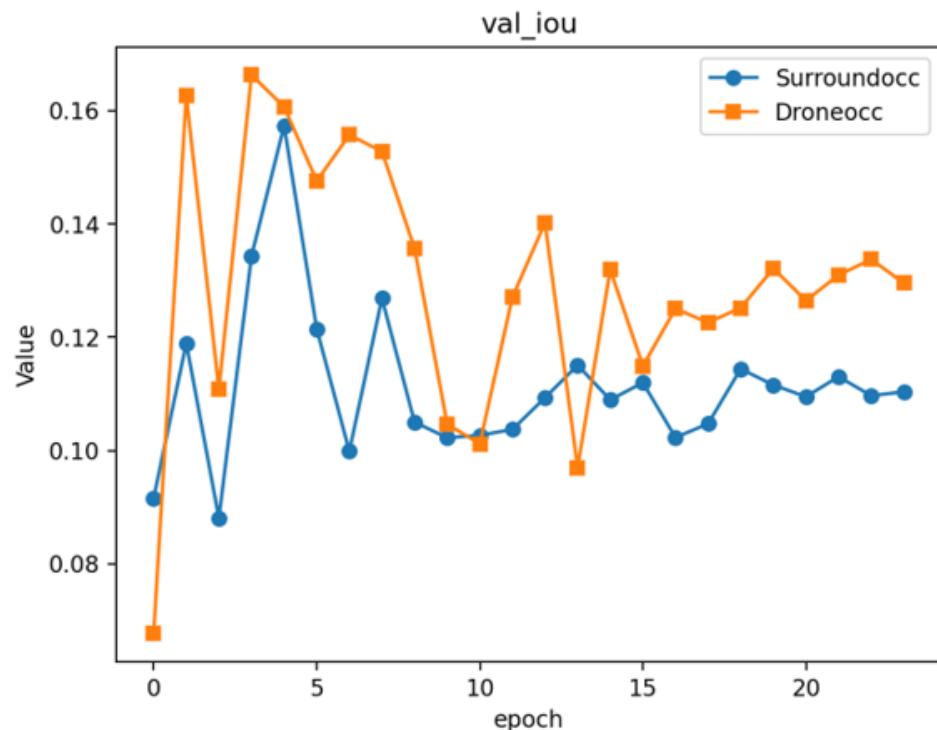


图 9. 训练时IOU

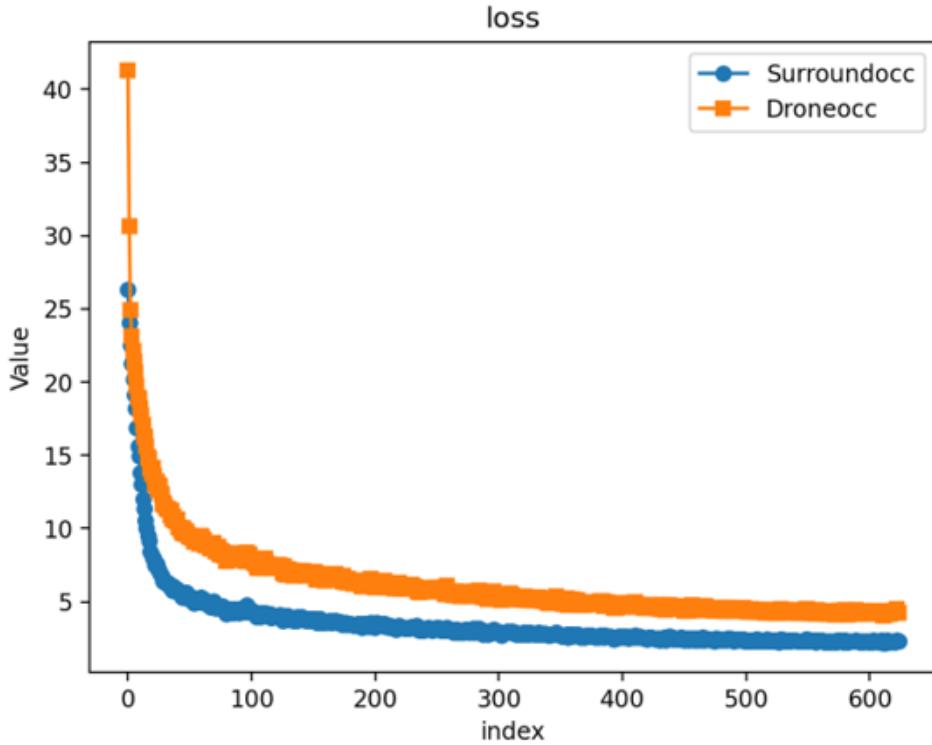


图 10. 训练损失函数

表 3. 测试结果

model	IOU	Terrain	vegetation	water	Bridge	Vehicle	Boat	Building
Surroundoc	0.1681	0.0576	0.1565	0.0003	0.0064	0	0	0.09434
Dronocc	0.1828	0.0527	0.1611	0.0014	0.01138	0.0014	0	0.10727

在验证集 IoU 曲线中，DroneOcc 的 IoU 值在整个训练过程中始终高于 SurroundOcc，并呈现出更稳定的提升趋势。这表明 DroneOcc 模型在多尺度解码和损失函数设计的帮助下，能够更好地学习并泛化三维场景的语义和空间信息。最终，DroneOcc 的验证集 IoU 达到了 0.1828，比 SurroundOcc 的 0.1681 提高了约 8.8%，充分证明了模型设计的优势。

从损失值曲线可以看出，DroneOcc 和 SurroundOcc 的损失值均随着训练迭代逐渐下降。然而，DroneOcc 的初始损失值较高，这可能是由于更复杂的特征融合过程。但在后续训练中，DroneOcc 的损失下降速度较快，并最终趋于收敛，这表明模型的优化过程更加高效。

从各类别的 IoU 指标对比可以看到，DroneOcc 在大多数语义类别上均优于 SurroundOcc。尤其是在关键类别（如建筑物）上，DroneOcc 的 IoU 为 0.10727，相比 SurroundOcc 提高了约 1.4%。这表明，本研究的局部解耦设计和多尺度融合策略能够有效提升不同场景语义信息的捕捉能力。

5.2 局限

尽管本研究提出的 DroneOcc 模型在 IoU 指标上仅比 SurroundOcc 稍有提升，但从推理效果的可视化结果来看，模型在三维占用网格的预测上已能够接近真实标签，展现出较为优异

的性能。以图中的两幅图为例，红色表示模型预测的建筑物，棕色表示模型预测的地面上；黄色为标签中的建筑物，蓝色为标签中的地面。可以看出，模型的预测结果能够大致识别建筑物和地面的分布位置，并在三维空间中提供占用区域的轮廓信息。

然而，值得注意的是，模型的预测效果仍然难以达到雷达级别的精准度。与激光雷达不同，本研究基于纯图像输入进行占用网格预测，这一过程中受限于输入的稀疏性和无人机场景的复杂性，使得模型在捕捉精细边界和复杂几何信息时存在一定误差。例如，在建筑物的边界区域以及地面与建筑的过渡区域，模型的预测往往存在较大不确定性。

这一限制主要归因于以下两点：首先，纯视觉输入在深度信息获取上存在天然的劣势；其次，无人机场景规模庞大，数据分布复杂，导致模型难以达到高精度预测。这也提示我们未来的改进方向：通过引入多模态数据（如结合激光雷达或深度相机的输入）以及更高分辨率的标签数据，进一步提升预测的精度和鲁棒性。

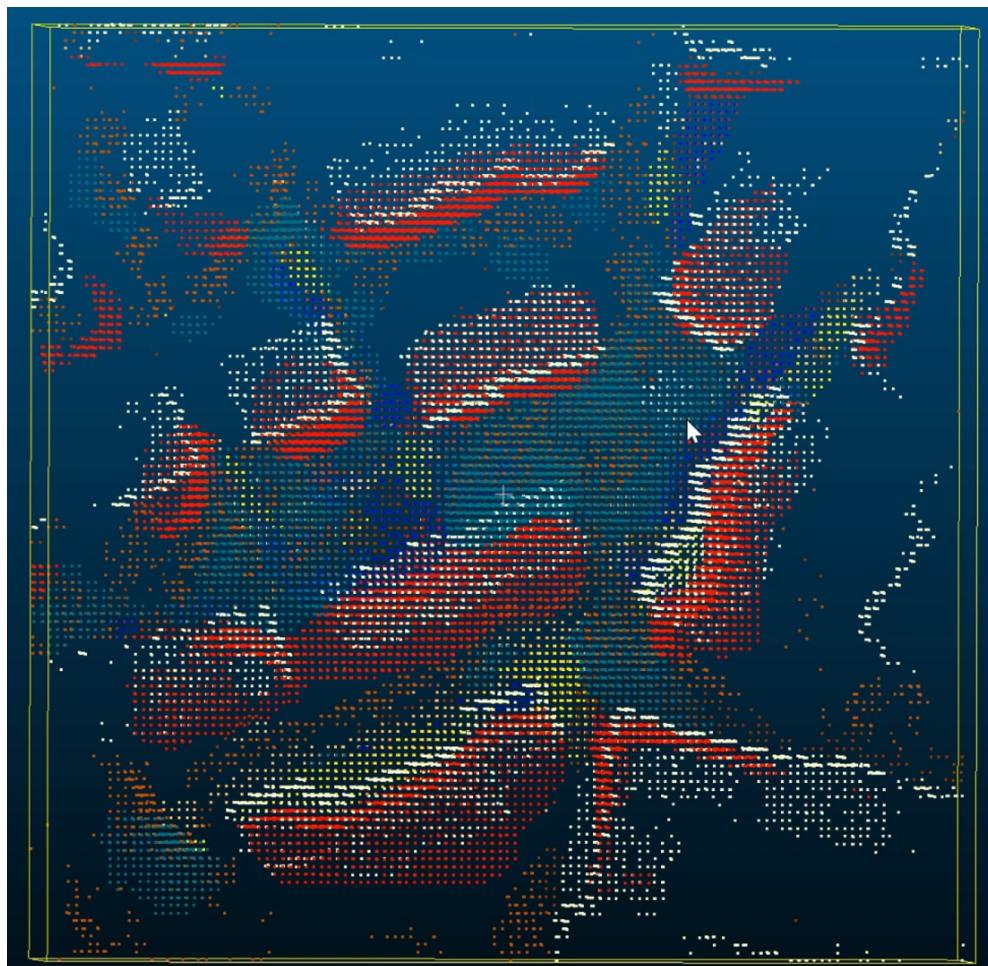


图 11. 预测与标签重合情况1

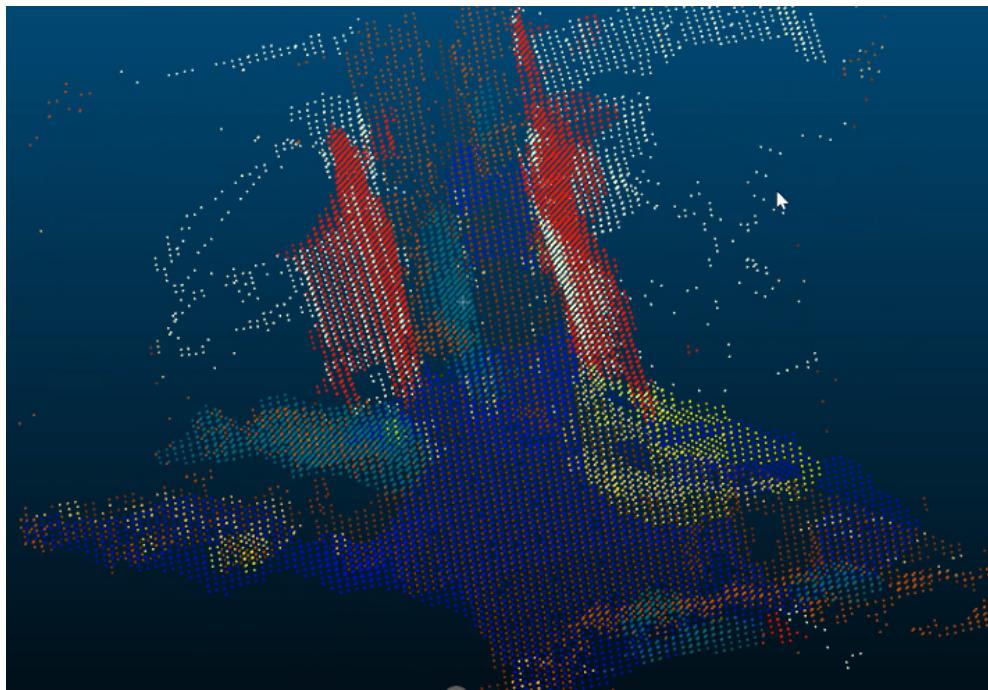


图 12. 预测与标签重合情况2

6 总结与展望

本研究提出了一种基于纯视觉输入的无人机三维占用网格预测方法 DroneOcc，通过局部解耦的多尺度解码器设计，实现了三维全局特征与二维局部特征的有效融合，显著提升了无人机复杂场景的三维感知能力。同时，结合多尺度损失函数，从语义分类和空间一致性两个方面优化了预测结果。实验表明，DroneOcc 在 IoU 和关键类别（如建筑物）的预测精度上相较基准方法 SurroundOcc 有显著提升，尽管整体分数尚未达到激光雷达级别的精准度，但已能够较好地捕捉目标的大致位置和类别，为无人机的环境感知和自主飞行提供了有力支持。未来工作将致力于进一步提升模型的性能，包括引入激光雷达或深度相机进行多模态数据融合，优化模型结构以提高对复杂场景的适应能力，扩展更大规模和多样化的真实数据集以提升泛化性，以及通过轻量化设计优化推理效率以满足实际应用需求。总之，本研究为无人机感知技术的未来发展提供了理论支持和实践参考，对城市空中交通、无人机物流等领域具有重要的应用潜力。

参考文献

- [1] H. Caesar, V. Bankiti, A. H. Lang, et al. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.
- [2] R. CHENG, R. RAZANI, E. TAGHAVI, et al. (af)2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. *Cornell University - arXiv, Cornell University - arXiv*, 2021.

- [3] A. Geiger, P. Lenz, C. Stiller, et al. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.
- [4] V. GUIZILINI, I. VASILJEVIC, R. AMBRUS, et al. Full surround monodepth from multiple cameras. *IEEE Robotics and Automation Letters*, pages 5397–5404, 2022.
- [5] J. Huang, G. Huang, Z. Zhu, et al. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.
- [6] Y. Huang, W. Zheng, Y. Zhang, et al. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023.
- [7] P. Jiang, D. Ergu, F. Liu, et al. A review of yolo algorithm developments. *Procedia computer science*, 199:1066–1073, 2022.
- [8] Y. Li, Z. Ge, G. Yu, et al. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023.
- [9] Z. Li, W. Wang, H. Li, et al. Bevformer: Learning bird’ s-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Cham: Springer Nature Switzerland, 2022.
- [10] V. LIONG, T. NGUYEN, S. WIDJAJA, et al. Amvnet: Assertion-based multi-view fusion network for lidar semantic segmentation. *arXiv: Computer Vision and Pattern Recognition, arXiv: Computer Vision and Pattern Recognition*, 2020.
- [11] J. Philion and S. Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer International Publishing, 2020.
- [12] W. Qiu and A. Yuille. Unrealcv: Connecting computer vision to unreal engine. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 909–916. Springer International Publishing, 2016.
- [13] S. Shah, D. Dey, C. Lovett, et al. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, pages 621–635. Springer International Publishing, 2018.
- [14] H. TANG, Z. LIU, S. ZHAO, et al. Searching efficient 3d architectures with sparse point-voxel convolution. In *Computer Vision – ECCV 2020, Lecture Notes in Computer Science*, pages 685–702, 2020.

- [15] P. Tang, Z. Wang, G. Wang, et al. Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15035–15044, 2024.
- [16] Y. WEI, L. ZHAO, W. ZHENG, et al. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation.
- [17] Y. Wei, L. Zhao, W. Zheng, et al. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023.
- [18] G. YANG, F. XUE, Q. ZHANG, et al. Urbanbis: a large-scale benchmark for fine-grained urban building instance segmentation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings*, 2023.
- [19] D. YE, Z. ZHOU, W. CHEN, et al. Lidarmultinet: Towards a unified multi-task network for lidar perception. 2022.
- [20] 何少林, 徐京华, and 张帅毅. 面向对象的多尺度无人机影像土地利用信息提取. *自然资源遥感*, (2):107–112, 2013.