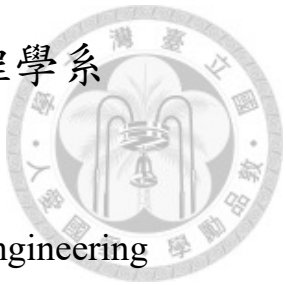


國立臺灣大學電機資訊學院資訊工程學系

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

運用消失點引導全景訊息預測

Vanishing-point Guided Semantic Scene Completion

楊盛評

Sheng-Ping Yang

指導教授: 鄭文皇 博士

Advisor: Wen-Huang Cheng Ph.D.

中華民國 114 年 8 月

August, 2025

doi:10.6342/NTU202503382

國立臺灣大學碩士學位論文
口試委員會審定書

MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

運用消失點引導全景訊息預測

Vanishing-point Guided Semantic Scene Completion

本論文係楊盛評君（學號 R12922163）在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 114 年 07 月 30 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Department of Computer Science and Information Engineering on 30 July 2025 have examined a Master's thesis entitled above presented by YANG, SHENG-PING (student ID: R12922163) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

莊永裕		
黃敬群	耿宏穎	簡韶逸

（指導教授 Advisor）

系主任/所長 Director: 陳祝嵩

Acknowledgements

在完成這篇碩士論文的此刻，我最想感謝的是鄭文皇教授。感謝老師在這兩年來提供一個自由且充實的學習環境，不僅引導我們在學術上建立紮實的基礎，也教會我們如何做人與做研究。老師的信任與支持，讓我能夠勇敢地嘗試與探索，是我這段碩士旅程中最珍貴的收穫。接著，我要感謝兩位在研究與生活中都給予我極大幫助的學長——志強老哥與昱文老哥。謝謝你們在我卡關與低潮的時候，總是不吝給予建議與鼓勵，尤其是志強老哥，除了學業上的協助，更在生活大小事上提供我很多幫助，讓我能安心前進。同時也要感謝實驗室的夥伴們：趙容、星逸、德易、康洋、儒怡、鄒玲、孟聰、靖家。我們在研究過程中互相討論、互相扶持，讓我在這段求學路上不孤單，並增添了許多溫暖與回憶。

謝謝我的女友，包容我在忙碌時的情緒與白目，也總是耐心陪我出門放鬆、轉換心情，更時時刻刻提醒我做事之餘也不要忘記她。她的理解與陪伴，讓我在壓力中依然感到幸福。感謝國中麻吉們，在我心煩的時候總能找到人聊聊天、分享近況，讓我暫時抽離焦慮，重新找回前進的力量。最後，最深的感謝獻給我的爸媽。謝謝您們一路以來無條件的支持與陪伴，是您們給了我最堅實的後盾，讓我能無所畏懼地追求理想，走到今天。

doi:10.6342/NTU202503382

摘要

本論文旨在解決自動駕駛領域中，僅使用單目相機進行三維語義場景補全（Semantic Scene Completion, SSC）時所面臨的關鍵挑戰，特別是對遠距離與微小物件感知準確度不足的問題。現有方法在處理因透視投影而在影像中變得微小、特徵模糊的遠方物體時，常因注意力分散而導致性能下降，進而對行車安全構成潛在威脅。為解決此問題，本研究提出一個名為「消失點聚合器」（Vanishing Point Aggregator, VPA）的創新架構。該方法的核心觀察在於：於駕駛場景影像中，消失點周圍自然聚集了來自遠距離場景的重要視覺資訊。VPA 引入一種新型的「消失點查詢」，專門用以強化此關鍵區域的特徵提取；並透過跨來源注意力融合機制，將富含遠場細節的 VPQ 與擷取全域物件語義的「標準實例查詢」進行整合，進而構建出更具完整性與辨識力的場景特徵表徵。

本研究於兩個具代表性的公開資料集——SemanticKITTI 與 SSCBench-KITTI-360 上進行系統性實驗與分析。實驗結果顯示，所提出的 VPA 模型在多項指標上皆達成目前最佳水準，尤其在遠距離區域與如行人、交通號誌等安全關鍵的微小物件類別上，顯著提升預測準確率。上述成果證實了本方法在提升單目 SSC 任務中遠場感知能力方面的有效性，對強化自動駕駛系統的環境感知穩定性與整體安全性具備實質貢獻。

關鍵字：自動駕駛、語義場景補全、消失點、小物件偵測

doi:10.6342/NTU202503382

Abstract

Semantic Scene Completion (SSC) aims to jointly predict semantic categories and 3D occupancy of a scene from coarse inputs, which is crucial for providing reliable perception in autonomous driving. In this paper, we enhance existing SSC models by unveiling the vanishing point region, specifically addressing challenges posed by tiny objects and voxels distant from the monocular camera. At the core of our method, we propose the Vanishing Point Aggregator (VPA) to prioritize features in high-density central areas. The proposed VPA seamlessly integrates the Vanishing Point Query (VPQ) with the vanilla instance query via a cross-attention fusion mechanism to refine feature representation. To evaluate the effectiveness of our method, we conduct comprehensive experiments on two standard SSC benchmarks and demonstrate that our method achieves SOTA performance. Our approach significantly improves the performance across various semantic classes, including a notable gain of 0.37 mIoU on SemanticKITTI and 0.5 mIoU on SSCBench-KITTI-360 for tiny objects. Ablation studies further validate the efficacy of our innovative query

doi:10.6342/NTU202503382

fusion strategy, showcasing its capability in long-range predictions for SSC tasks.

Keywords: Autonomous Driving, Semantic Scene Completion, Vanishing Point, Tiny Object Detection

doi:10.6342/NTU202503382

Contents

Page

Verification Letter from the Oral Examination Committee **i**

Acknowledgements **ii**

摘要 **iii**

Abstract **iv**

Contents **vi**

List of Figures **ix**

List of Tables **xii**

Chapter 1 Introduction **1**

1.1 Publication 5

Chapter 2 Related Works **6**

2.1 3D Semantic Scene Completion 6

2.2 Deformable Attention 7

2.3 Two-Stage Architecture: VoxFormer 10

2.4 Query-Centric Method: Symphonies 13

2.5 Vanishing Point-related Approaches 17

Chapter 3 Method **19**

3.1 Problem Formulation 19

doi:10.6342/NTU202503382

3.2 Motivations and Observations 20
3.3 Overview 22
3.4 Proposed Method 25
3.4.1 Vanishing Point Query (VPQ) Initialization 25
3.4.2 Vanishing Point Query Update with Regional Features 27
3.4.3 Cross-Source Query Update with Initial Voxel Features 29
3.4.4 Initial Scene Voxel Feature Generation 30
3.4.5 Iterative Feature Co-Refinement Module 33
3.4.6 Segmentation Head 34
3.4.7 Loss Functions 35
Chapter 4 Experiments **37**
4.1 Datasets 37
4.2 Evaluation Metrics 38
4.3 Implementation Details 39
4.4 Baseline Methods 40
4.5 Quantitative Results 40
4.5.1 Results on SemanticKITTI Validation Set 40
4.5.2 Results on SSCBench-KITTI-360 Test Set 44
4.6 Ablation Study 45
4.6.1 Impact of Query Design and Interaction 45
4.6.2 Performance Analysis across Distance Ranges 46
4.6.3 Effect of Vanishing Point Region Size 48

doi:10.6342/NTU202503382

Chapter 5 Conclusion **50**

5.1 Conclusion 50

5.2 Future Work 51

References **52**

doi:10.6342/NTU202503382

List of Figures

1.1 **Visualization of Semantic Scene Completion (SSC) output from the SemanticKITTI dataset.** The upper row displays input road scene images captured by a vehicle-mounted camera. The lower row illustrates the corresponding semantic scene completion outputs. 4

2.1 **Two-stage output of VoxFormer. Left (Stage 1):** The model predicts a class-agnostic occupancy map that reconstructs visible 3D geometry and generates sparse query proposals. **Right (Stage 2):** Those proposals are densified and semantically labeled via deformable cross-attention and self-attention, each color denotes a different scene class in the completed 3D voxel grid. 11

2.2 **Contrasting (a) traditional voxel-wise modeling with (b) the Symphonies framework.** While (a) directly maps pixels to voxels, potentially causing ambiguity, Symphonies (b) uses instance queries derived from images. These queries then mediate between 2D and 3D through ”Instance to Scene” (2) and ”Scene to Instance” (3) pathways, allowing Symphonies to utilize instance semantics for richer 3D scene comprehension. 14

doi:10.6342/NTU202503382

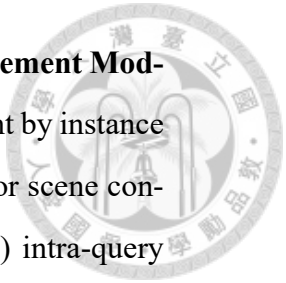
2.3 **Overview of Symphonies.** The Symphonies architecture highlights its iterative Decoder Layer. Multi-scale image features and instance queries are extracted, and initial voxels are proposed. These are fed into the core Symphonies Decoder, which is iterated N times. Within each iteration, a sequence of deformable cross-attention and self-attention operations interact sequentially—from 2D features to instance queries, from instance queries to 3D voxels, and then from 3D voxels back to instance queries. This procedure, effectively leveraging both deformable and standard mechanisms, facilitates comprehensive image-instance-scene feature interaction and refinement. A final Segmentation Head produces the 3D semantic segmentation. 15

3.1 Illustration of the Vanishing Point and the associated Information-rich Region in a typical driving scene. The VP, indicated by the red circle, represents the convergence of parallel lines. The surrounding orange-bounded area highlights the Information-rich Region, where features from distant objects are concentrated due to perspective projection, forming an area of high information density crucial for far-field perception. 21

3.2 **Overview of proposed model.** The pipeline of our method consists of three procedures: (a) initializing VPQ and instance queries from the encoded image features F. (b) Vanishing Point Aggregator (VPA), which dynamically aggregates pivot information from the VP region to enhance the VPQ. (c) VP-instance queries aggregation, where VPQ and instance queries are integrated using cross-attention, followed by deformable attention to refine 3D voxel features. 24

3.3 **Visualization of Reference and Vanishing Points.** The red marker denotes the reference points generated during the initialization steps, and the yellow marker indicates the vanishing point detected in this image. 27

doi:10.6342/NTU202503382



3.4 **Detailed view of the decoder's Iterative Feature Co-Refinement Module.** This image shows the decoder steps (1) voxel enrichment by instance and vanishing-point queries, (2) deformable self-attention for scene contextualization, (3) cross-attention query refinement, and (4) intra-query self-attention—all co-refining 3D voxel features and queries over M iterations before the final segmentation head. 32

4.1 **Results visualize on SemanticKITTI.** The regions marked with red boxes in the figure demonstrate that our method is capable of accurately detecting and locating small objects, even in challenging scenarios such as long distance prediction. 41

doi:10.6342/NTU202503382

List of Tables

2.1 Performance evaluation under different occupancy input conditions, showing the mean intersection over union (mIoU) for each scenario. 12

4.2 Ablation study on different query designs and their interaction mechanisms on the SemanticKITTI `val` set. "Instance Query Only" serves as our baseline, akin to Symphonies. "Fusion Query" refers to our proposed method with cross-source cross-attention. 47

4.3 Comparison of model performance on different distance settings. Far covers depth range from 38.4m to 51.2m and Medium covers range from 25.6m to 51.2m. 47

4.4 Ablation study on the size of the VP region, defined by $H_f/d \times W_f/d$. Performance is measured by mIoU (%). $d = 0$ disables VPA, and $d = 1$ uses the full image. 49

doi:10.6342/NTU202503382

Chapter 1 Introduction

With the rapid evolution of autonomous vehicles, a comprehensive and nuanced understanding of the operational environment is important, extending beyond isolated tasks such as vehicle detection [1], pedestrian detection [7], and rudimentary scene segmentation [2]. In this pursuit, 3D Semantic Scene Completion (SSC) [4, 16, 19] has garnered notable attention. SSC's core contribution lies in its ambitious goal: to predict not only the complete 3D geometry but also the semantic labels for every part of a scene, even from partial or occluded observations. This exhaustive, dense voxelized representation of the environment offers a profound advantage for autonomous systems. It moves beyond reactive perception to enable proactive path planning, sophisticated risk assessment, and robust decision-making, especially in complex, dynamic scenarios. Ultimately, by providing a holistic understanding, SSC plays a vital role in mitigating potential accidents and significantly enhancing the safety and reliability of autonomous driving systems.

Existing SSC methods can be broadly categorized into LiDAR-based [19, 21] and camera-based [4, 12, 16] approaches, differentiated by their primary sensor modality. LiDAR-based systems leverage precise geometric cues inherent to active sensing but often depend on expensive sensor hardware, which can limit their widespread adoption in cost-sensitive daily applications. In contrast, camera-based frameworks present a more accessible and economically viable alternative, endeavoring to infer the complete 3D scene

doi:10.6342/NTU202503382

structure and semantics solely from monocular RGB images. However, these camera-centric approaches face a significant inherent challenge: the accurate perception of distant objects. Due to the projective nature of camera imaging, distant entities occupy fewer pixels and possess less distinct features, rendering them difficult to discern and susceptible to misinterpretation. This limitation is particularly critical because these small or far-away objects frequently correspond to pedestrians, cyclists, or crucial traffic signs—elements for which perception systems have an extremely low tolerance for error due to their direct impact on safety. Consequently, the diminished performance in distant predictions directly compromises the system's ability to anticipate and react to potential hazards effectively. To address these critical issues, particularly the robust perception of distant, safety-critical elements, we propose a novel framework that introduces the Vanishing Point Aggregator (VPA).

Recognizing that the vanishing point region in driving scenes typically contains a high concentration of relevant information, especially for distant objects aligned with the direction of travel, VPA is designed to prioritize and emphasize features from this crucial central area. The proposed VPA dynamically aggregates information from this vital region, significantly enhancing the model's ability to capture subtle semantic details and geometric structures that are often lost for distant entities. Furthermore, our VPA seamlessly integrates a novel Vanishing Point Query (VPQ) with established instance queries, thereby enriching the model's understanding by broadening both its global contextual awareness (via VPQ focusing on the far-field) and its local object-specific detail (via instance queries). We conduct rigorous experiments on two well-established benchmarks, i.e., SemanticKITTI and KITTI-360, achieving state-of-the-art (SOTA) performance that demonstrates the efficacy of our proposed method. By strategically focusing on these

doi:10.6342/NTU202503382

critical far-field areas, our model significantly improves the detection and semantic understanding of small objects appearing in distant regions—a persistent challenge for existing methods due to scale variations, feature degradation, and occlusion. This enhanced capability is indispensable for autonomous driving, where the accurate and timely detection of distant pedestrians, vehicles, and regulatory road signs is absolutely crucial for proactive hazard avoidance and accident prevention. In summary, our contributions are as follows:

1. We introduce a novel Vanish Point Aggregator (VPA) designed to explore and prioritize pivotal information concentrated within the vanishing point region of driving scenes. The proposed VPA employs a unique Vanish Point Query (VPQ), coupled with instance queries, to effectively address the challenges of Semantic Scene Completion, particularly for distant scene elements.

2. Our model significantly enhances perception accuracy for challenging categories, specifically tiny and long-distance objects, which are often safety-critical. This addresses longstanding challenges in SSC tasks through the synergistic application of VPQ and advanced fusion mechanisms for different query types.

3. We conduct abundant experiments and comprehensive ablation studies to rigorously verify the effectiveness of our presented model, demonstrating that it exceeds the performance of existing camera-based SSC methods, especially in scenarios demanding accurate far-field perception.

doi:10.6342/NTU202503382

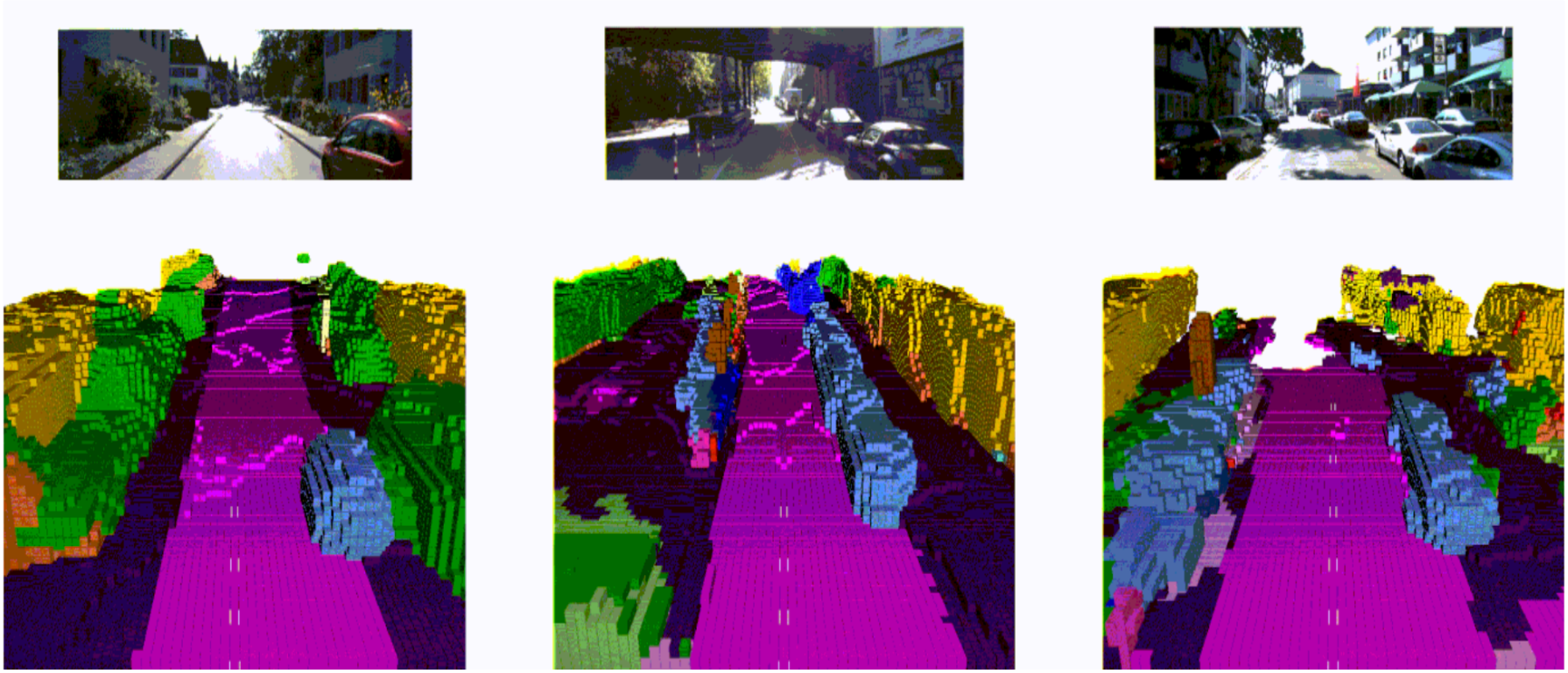


Figure 1.1: **Visualization of Semantic Scene Completion (SSC) output from the SemanticKITTI dataset.** The upper row displays input road scene images captured by a vehicle-mounted camera. The lower row illustrates the corresponding semantic scene completion outputs.

doi:10.6342/NTU202503382

1.1 Publication

The core of this thesis builds upon the works submitted to and accepted by the 2025 IEEE International Conference on Image Processing (IEEE ICIP 2025) in the following peer-reviewed publication:

Sheng-Ping Yang, Yu-Wen Tseng, Yung-Chieh Yang, I-Bin Liao, Chi-En Huang, Shen-Hsuan Liu, Yung-Hui Li, Jhih-Ciang Wu, Hong-Han Shuai, and Wen-Huang Cheng, "**Unraveling Vanishing Point and Calibrating Tiny Objects for Semantic Scene Completion.**", in Proc. of the IEEE ICIP 2025.

doi:10.6342/NTU202503382

Chapter 2 Related Works

2.1 3D Semantic Scene Completion

3D SSC aims to simultaneously infer complete 3D geometry and semantic labels from partial observations. Early approaches primarily relied on geometric priors [19, 21], utilizing point clouds and depth-maps obtained from LiDAR or other depth sensors to reconstruct missing structures and predict semantic labels. SSCNet [19] introduced a depth-map-based method, inferring both occupancy and semantic labels for each voxel within the view frustum. While depth sensors are costly and less portable, researchers have increasingly explored camera-based solutions [4, 12, 16]. MonoScene [4] was the first to introduce camera-based SSC, leveraging 2D-3D feature projections and a sequential combination of UNets. VoxFormer [16] employs a two-stage Transformer-based framework to sparsely query relevant 3D regions, facilitating more precise and effective interactions between 2D and 3D features. More recently, Symphonies [12] introduced instance queries as a bridge between 2D and 3D features, enhancing contextual awareness and semantic understanding. Despite these advances, objects in distant regions remain disproportionately underrepresented. To tackle this challenge, we propose the Vanishing Point Aggregator, which enhances feature prioritization in high-density central regions far from the observing camera. Leveraging fine-grained VP features, our proposed model becomes more

doi:10.6342/NTU202503382

perceptive of small objects, achieving a more balanced and comprehensive prediction.

2.2 Deformable Attention

Semantic Scene Completion (SSC) requires a network to reason about both observed and unobserved regions of a scene, predicting occupancy and semantic labels for every voxel within a 3D grid. Achieving this goal necessitates effective feature aggregation across multiple spatial domains, particularly between 2D image features and 3D voxel space. A central challenge in this process lies in how to efficiently propagate and fuse relevant information, especially when the input data is sparse, irregular, or partially occluded.

Traditional attention mechanisms, such as the full self-attention used in Transformer architectures, suffer from quadratic computational complexity with respect to the number of spatial positions. When directly applied to 3D voxel grids—which often contain millions of elements—this approach quickly becomes computationally prohibitive. Moreover, not all voxels contribute equally to the completion task. In outdoor driving scenarios, where the majority of the 3D volume consists of unoccupied free space, this uniform attention allocation leads to inefficient use of computational resources. Spending attention on these regions not only increases computational cost but also weakens the learning signal by diluting focus away from the truly informative areas. Furthermore, standard dense attention mechanisms lack the flexibility to dynamically adapt their receptive field according to object scale or contextual importance. This issue is particularly severe when dealing with small or distant objects—such as poles, bicycles, or traffic signs—which may occupy only a few voxels and are easily overshadowed by larger, dominant scene features.

doi:10.6342/NTU202503382

To address these challenges, *deformable attention* introduces a sparse and adaptive attention strategy. Instead of attending to the entire feature space uniformly, deformable attention allows each query to focus on a limited set of sampled key points located around learnable reference positions. Both the sampling offsets and the attention weights are learned during training, enabling the network to dynamically determine where and how to gather information. This flexibility allows the model to adaptively concentrate on spatially distributed and semantically meaningful regions, even when these regions are sparsely populated or irregularly shaped.

The deformable attention operation can be formally expressed as:

$$\mathrm{DA}(q, p, F) = \sum_{s=1}^{N_s} A_s W_s F(p + \Delta p_s), \tag{2.1}$$

where q denotes the query vector, p is the reference point, and F represents the input features. The summation is performed over N_s sampled points around the reference position p. For each sampled point, $W_s \in \mathbb{R}^{d \times d}$ is the learnable weight matrix for value transformation, $A_s \in [0, 1]$ is the attention weight, and Δp_s is the learnable sampling offset relative to p. The feature at the location $p + \Delta p_s$ is retrieved via bilinear interpolation in 2D feature space or trilinear interpolation in the 3D voxel grid.

This formulation provides an adaptive mechanism that adjusts the receptive field of each query according to the semantic complexity of the scene. In SSC, where objects may appear at varying scales and may be partially occluded, such adaptive sampling is essential for accurate geometry reconstruction and semantic reasoning.

Within the SSC pipeline, deformable attention plays a vital role across two key stages of the feature interaction process. First, *voxel-to-image deformable cross-attention* en-

doi:10.6342/NTU202503382

ables each 3D voxel query to selectively sample features from the projected 2D image regions using learned offsets. This approach prevents the ambiguity that often arises in dense projection methods, where features from foreground objects may incorrectly leak into background voxels. Second, *voxel-to-voxel deformable self-attention* facilitates selective information exchange among voxels after their initial feature encoding, allowing efficient propagation of contextual information across the 3D scene while maintaining spatial efficiency.

Beyond architectural efficiency, deformable attention directly addresses three fundamental challenges inherent in SSC tasks:

- **Occlusion and Partial Observability**: By focusing sampling on regions likely to contain informative features, deformable attention enables reasoning about occluded structures based on the visible context.
- **Scale Variation of Objects**: The use of learnable sampling offsets allows adaptive adjustment of the attention range, providing fine-grained focus on small objects while maintaining broad coverage for larger ones.
- **Spatial Imbalance**: Rather than allocating equal resources across occupied and unoccupied space, deformable attention prioritizes semantically rich regions, improving feature aggregation efficiency and learning effectiveness.

Together, these capabilities position deformable attention not merely as a computational optimization, but as a critical enabler of robust and reliable voxel representation learning—particularly important in monocular or camera-based SSC scenarios where depth cues are often incomplete or noisy.

doi:10.6342/NTU202503382

2.3 Two-Stage Architecture: VoxFormer

VoxFormer introduces a novel sparse query-based framework for camera-based Semantic Scene Completion (SSC), addressing key limitations of traditional dense projection approaches. Rather than relying on direct 2D-to-3D feature projection, which often leads to feature leakage and ambiguity—particularly in occluded or empty regions—VoxFormer proposes a principled two-stage design that separates geometry reconstruction from semantic completion.

At the core of VoxFormer's architecture lies the principle of reconstruction-before-hallucination. The first stage focuses on reconstructing the 3D geometry of visible regions by predicting an occupancy map using monocular or stereo depth estimation. This occupancy prediction serves as a class-agnostic filter that identifies potentially occupied voxels, substantially reducing the number of active queries by excluding large volumes of empty space. These voxel proposals are then used to guide feature aggregation and semantic reasoning in the subsequent stage.

The second stage performs semantic scene completion by progressively densifying the sparse voxel representation generated in stage one. Voxel queries interact with 2D image features via deformable cross-attention, fusing multi-view visual information into the 3D space. For voxels not selected during the query proposal phase, learnable mask tokens are introduced, and deformable self-attention is applied to propagate contextual information across the voxel grid. This sparse-to-dense mechanism allows the model to complete the semantic labeling process efficiently while maintaining high spatial resolution.

A key advantage of this query-based framework is its ability to mitigate feature con-

doi:10.6342/NTU202503382

tamination, which is a common issue in dense projection methods where overlapping projections can cause foreground features to incorrectly populate background voxels. By leveraging guided voxel queries and attention-based aggregation, VoxFormer significantly improves the quality of voxel representations, enhancing both geometric completion and semantic segmentation performance. Additionally, VoxFormer's modular design supports flexible integration of various depth estimation techniques and input modalities (monocular or stereo), making the framework adaptable to different SSC settings.

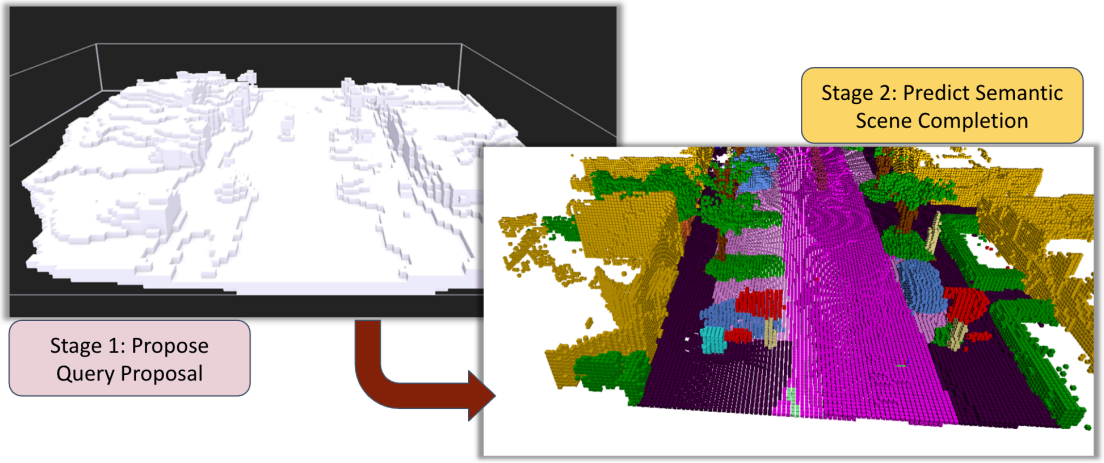


Figure 2.1: **Two-stage output of VoxFormer. Left (Stage 1):** The model predicts a class-agnostic occupancy map that reconstructs visible 3D geometry and generates sparse query proposals. **Right (Stage 2):** Those proposals are densified and semantically labeled via deformable cross-attention and self-attention, each color denotes a different scene class in the completed 3D voxel grid.

Despite these architectural strengths, VoxFormer' s two-stage design introduces an inherent limitation due to its sequential treatment of geometry and semantics. Specifically, the framework assumes that accurately predicted geometry provides a sufficient basis for semantic reasoning. However, this assumption implies a tight coupling between occupancy quality and semantic completion, which may not fully hold in practice.

To investigate this potential information gap between geometry and semantics, we designed an experiment where the stage-1 output (predicted occupancy) was replaced by

doi:10.6342/NTU202503382

ground truth occupancy maps. This setup isolates the influence of geometry prediction quality on the semantic segmentation outcome. To further analyze the relationship between occupancy density and semantic performance, we introduced random dropout to the ground truth occupancy with dropout rates of 50% and 90%, simulating varying levels of geometric completeness.

The experimental results, summarized in Table 2.1, reveal that even with 90% of the ground truth occupancy randomly dropped, the model still achieves better semantic performance than when using the original stage-1 output. This finding clearly indicates the existence of a significant information gap between the geometry prediction and semantic completion stages.

Scenario	**mIoU**
Original stage 1 output	12.04
Ground truth occupancy	19.71
Ground truth occupancy with 50% random dropout	16.38
Ground truth occupancy with 90 random dropout	12.44

Table 2.1: Performance evaluation under different occupancy input conditions, showing the mean intersection over union (mIoU) for each scenario.

These results suggest that the semantic completion stage has the capacity to leverage rich contextual cues even when geometric input is sparse, highlighting a disconnect between the two stages of the current design. This observation challenges the assumption that semantic reasoning should be strictly conditioned on geometry alone and points toward the need for more integrated or jointly optimized approaches.

While VoxFormer achieves notable efficiency and accuracy through its sparse query-based framework, our analysis reveals a fundamental limitation of its two-stage architecture. Addressing this geometry-semantic gap remains an important direction for further

doi:10.6342/NTU202503382

improving SSC models and advancing the robustness of camera-based 3D scene understanding systems.

2.4 Query-Centric Method: Symphonies

Symphonies introduces a new perspective on camera-based 3D Semantic Scene Completion (SSC) by reimagining the interaction between 2D and 3D representations through the lens of instance queries. While earlier methods primarily relied on dense projection from image features to voxel grids or on occupancy-first paradigms, Symphonies places instance-level queries at the center of its architecture. These queries serve not only as dynamic information bridges but also as semantic interpreters that navigate and refine multi-modal representations. Through this design, Symphonies achieves a balance of efficiency, scalability, and fine-grained semantic understanding that is particularly well-suited for real-world autonomous driving and robotics applications.

The backbone of the model begins with an image encoder built upon a standard Transformer architecture. This encoder extracts multi-scale 2D features from RGB inputs, capturing both low-level textures and high-level semantic representations. Compared to traditional convolutional networks, the Transformer-based encoder provides a larger receptive field and enhanced capacity for modeling long-range dependencies, which is essential for understanding complex spatial arrangements in urban scenes. The resulting feature maps preserve spatial hierarchies and semantic richness, enabling the model to represent critical elements such as pedestrians, vehicles, and buildings with high fidelity. These encoded features form the basis for subsequent cross-modal interactions with the 3D voxel space and serve as a semantically grounded foundation for scene-level reasoning.

doi:10.6342/NTU202503382

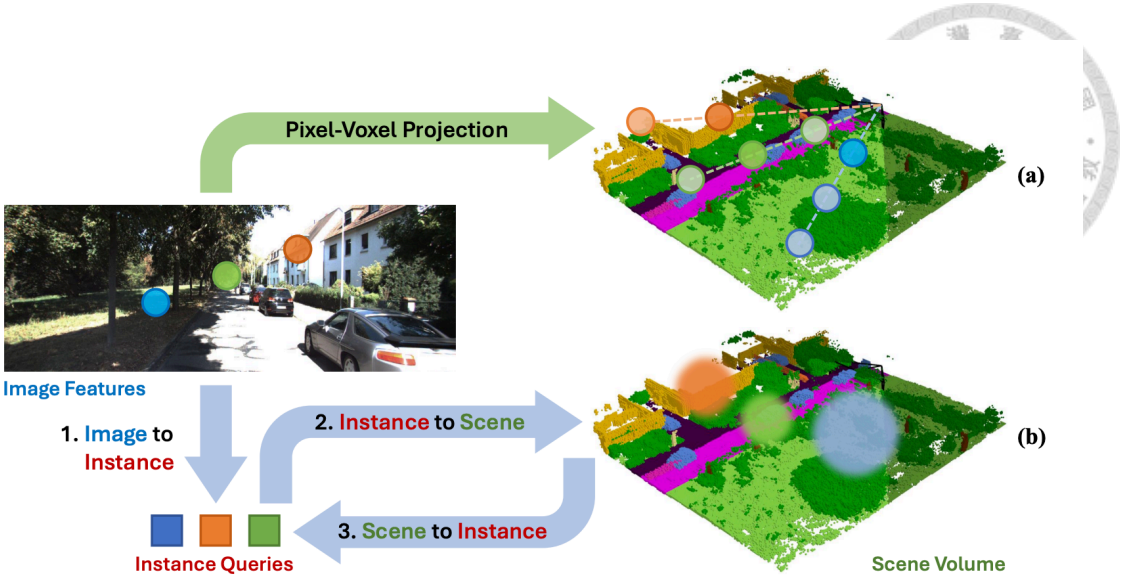


Figure 2.2: **Contrasting (a) traditional voxel-wise modeling with (b) the Symphonies framework.** While (a) directly maps pixels to voxels, potentially causing ambiguity, Symphonies (b) uses instance queries derived from images. These queries then mediate between 2D and 3D through "Instance to Scene" (2) and "Scene to Instance" (3) pathways, allowing Symphonies to utilize instance semantics for richer 3D scene comprehension.

The projected transition from 2D to 3D is performed via a depth-rectified Voxel Proposal Layer (VPL). Using either monocular or stereo-derived depth, the model lifts 2D features onto a coarse volumetric grid, generating voxel proposals that likely correspond to occupied areas of the scene. This early-stage filtering offers a computational advantage by discarding vast empty regions that dominate outdoor environments, especially in autonomous driving datasets. More importantly, it helps reduce the noise that commonly results from projecting uncertain 2D features into 3D, establishing a more robust base upon which semantic reasoning can be built.

At the core of Symphonies lies its decoder, where the architectural innovation is most evident. Here, the model employs a Serial Instance-Propagated Attention mechanism, which governs how learned instance queries interact with both image and voxel spaces. These queries are initialized to capture semantic categories or spatial instances and then refined over multiple attention layers. The decoder facilitates three critical streams of in-

doi:10.6342/NTU202503382

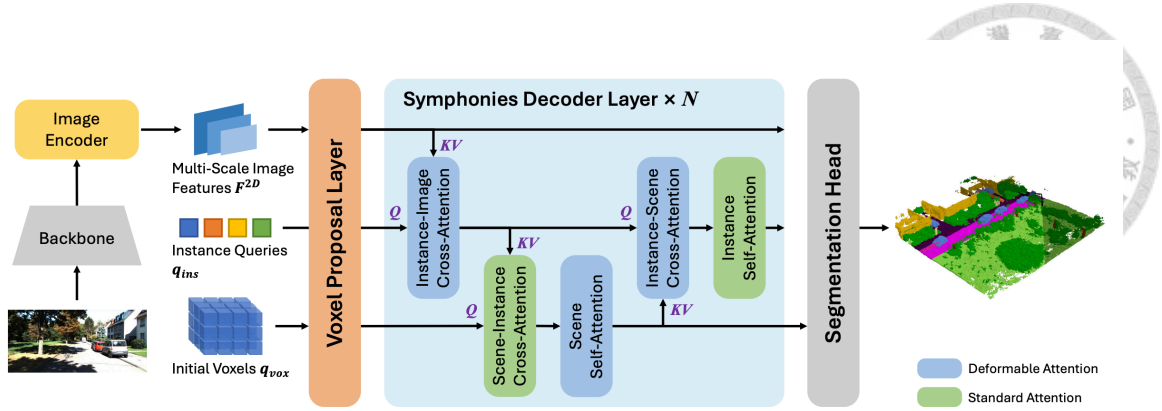


Figure 2.3: **Overview of Symphonies.** The Symphonies architecture highlights its iterative Decoder Layer. Multi-scale image features and instance queries are extracted, and initial voxels are proposed. These are fed into the core Symphonies Decoder, which is iterated N times. Within each iteration, a sequence of deformable cross-attention and self-attention operations interact sequentially—from 2D features to instance queries, from instance queries to 3D voxels, and then from 3D voxels back to instance queries. This procedure, effectively leveraging both deformable and standard mechanisms, facilitates comprehensive image-instance-scene feature interaction and refinement. A final Segmentation Head produces the 3D semantic segmentation.

teraction: from image features to queries , from queries to voxel grids, and from voxel to each other. In the first case, queries extract spatial visual cues from the image that help localize and define the structure of objects. In the second, they project this knowledge into 3D space, selectively aggregating voxel features based on geometric plausibility and semantic relevance. Finally, through intra-voxel communication, the model builds consistency across different entities, allowing for relational reasoning between, for instance, a pedestrian and a nearby vehicle.

One of the standout characteristics of Symphonies is its deliberate departure from the standard dense attention or projection-based paradigms. Instead of letting each voxel compete for global attention or passively inherit features from image pixels, Symphonies elevates instance queries to a central role. These queries actively guide the flow of information, effectively deciding what to extract, where to attend, and how to relate different spatial representations. This query-centric fusion mechanism provides several advantages. First, it reduces feature redundancy and leakage, common in dense projection methods

doi:10.6342/NTU202503382

where overlapping or occluded regions blur semantic boundaries. Second, it offers adaptive granularity, as the attention mechanism can adjust focus depending on the object's size or scene context. Third, it allows for modular extension—new queries or types of queries can be added for instance segmentation, panoptic reasoning, or even multi-frame temporal modeling.

The final output of the decoder is passed into a 3D segmentation head, which predicts semantic labels for each voxel. To enhance spatial resolution and multi-scale understanding, this module incorporates Atrous Spatial Pyramid Pooling (ASPP), which captures contextual cues at different dilation rates. The resulting voxel-wise semantic map reflects not only pixel-aligned texture but also object-level coherence and spatial awareness, all inferred from purely RGB inputs without explicit geometry supervision.

Symphonies offers several practical advantages over previous SSC methods. In contrast to VoxFormer, which separates geometry prediction from semantic reasoning and introduces a potential gap between the two stages, Symphonies embraces a joint reasoning framework guided by semantic queries. There is no intermediate occupancy mask that constrains or biases the semantic completion stage. Instead, the entire reasoning process is conditioned on learnable queries that can attend to both visible evidence and infer hidden structures. Compared to methods like MonoScene, which rely on heavy projection networks and multi-stage decoders, Symphonies remains lightweight and more interpretable, as its query dynamics can be visualized and analyzed directly.

Furthermore, by avoiding explicit dependency on depth supervision, Symphonies becomes more flexible in environments where accurate geometry is unavailable or unreliable. It performs competitively even with monocular images, and its query-based at-

doi:10.6342/NTU202503382

tention design ensures robustness under viewpoint changes, occlusions, and sensor noise. This makes it particularly suitable for deployment in resource-constrained or partially observed environments where depth data might be sparse or corrupted.

In conclusion, Symphonies introduces a powerful and elegant baseline for SSC by repositioning instance queries as the primary interface for 2D – 3D semantic fusion. It unifies vision and geometry within a single-stage architecture, where interaction is guided not by rigid projections or dense maps but by semantically meaningful, learnable entities. This model not only delivers strong performance on standard benchmarks but also offers a versatile framework that can be extended to a wide range of 3D perception tasks. As camera-based SSC continues to evolve, Symphonies sets a new baseline in how semantic understanding, geometric reasoning, and visual abstraction can be harmoniously integrated through the language of queries.

2.5 Vanishing Point-related Approaches

The vanishing point (VP), where parallel lines in 3D space converge, serves as a crucial spatial cue for object localization and motion estimation in vision-based perception tasks. A key challenge associated with VP is vanishing point detection. For instance, VPDETR [6] introduces an end-to-end Transformer-based framework for VP detection, eliminating the need for explicit line extraction and candidate sampling, thereby improving efficiency and accuracy. Beyond detection, VPs have been increasingly utilized as auxiliary cues in various visual tasks. VP-guided techniques [9, 13] exploit perspective geometry to enhance semantic understanding and improve spatial reasoning. VPSeg [9] refines feature extraction for small-scale objects near the vanishing point, effectively ad-

doi:10.6342/NTU202503382

dressing challenges in distant object segmentation and rapid motion scenarios, particularly relevant to autonomous driving applications. Building on VP-guided methods, we exploit the knowledge in VP region, which tackles depth-induced density variations in driving images by dynamically prioritizing information-rich regions. By leveraging VP cues, our VPA enhances long-range predictions and tiny object detection within the SSC framework, improving scene completion performance and 3D environment understanding.

doi:10.6342/NTU202503382

Chapter 3 Method

3.1 Problem Formulation

The primary goal of SSC is to generate a dense 3D voxelized representation from a single 2D image captured by a camera mounted on a vehicle. This voxelized representation integrates geometric and semantic information, where each voxel in the 3D space encodes whether it is occupied and, if so, its semantic class. Formally, given an input RGB image $I \in \mathbb{R}^{h \times w \times 3}$, where h and w denote the height and width of the image, the objective is to predict voxel grid $\mathcal{V} \in \mathbb{R}^{H \times W \times D \times (N+1)}$ and $\mathcal{V}(x, y, z) \in \{c_0, c_1, ..., c_N\}$. Here, c_0 corresponds to the "empty space", while the remaining ($c_1, \ldots, c_N$) represent the semantic classes of interests, such as road, building, or fence.

We utilize a deep neural network, denoted by Θ, to perform this prediction task. The model Θ takes the input image I and outputs the 3D voxel grid Y, such that:

$$Y = \Theta(I), \tag{3.1}$$

where each voxel $Y(i, j, k)$ contains a one hot label from the classes in C, providing both geometry and semantic information.

doi:10.6342/NTU202503382

3.2 Motivations and Observations

Recent advancements in camera-based Semantic Scene Completion (SSC), exemplified by frameworks like Symphonies [12], have demonstrated the efficacy of employing instance queries as intermediaries. These queries aim to bridge the inherent gap between 2D image features and their corresponding 3D scene representations. Symphonies, in particular, leverages deformable cross-attention mechanisms to facilitate dynamic interactions between each instance query (Q_{ins}) and the entirety of the 2D image features. This global interaction strategy is designed to allow instance queries to capture a broad contextual understanding of the scene from comprehensive image features.

However, while this global approach is beneficial for general scene understanding and capturing large, salient objects, its uniform treatment of the image space can inadvertently lead to a dilution of focus. When tackling the nuanced challenge of perceiving small or distant objects—such as faraway pedestrians, cyclists, or crucial traffic signs—such globally-oriented interactions often sacrifice the requisite specificity. The attention mechanism, when spread across the entire image for every query, may not allocate sufficient focus or processing power to the subtle cues indicative of these critical, yet diminutive, entities. This can result in their features being underrepresented or overshadowed by more dominant elements in the scene, consequently degrading performance on these low-tolerance, safety-critical object categories that demand specialized attention to maintain accuracy.

The vanishing point in driving scenes is a geometrically significant locus, representing the convergence of parallel lines and typically aligning with the vehicle's direction of motion. Crucially, we observe that due to the inherent characteristics of perspective

doi:10.6342/NTU202503382

projection—where objects appear smaller as their distance from the camera increases, as known as ”big in, small out” phenomenon relative to the 3D world versus the 2D image plane—the region surrounding the vanishing point inherently aggregates a disproportionately large amount of information from the far-field. This projective geometry effectively concentrates features from numerous distant objects and scene elements into a relatively compact area on the image, forming a region of high information density. Consequently, we posit that it is imperative to effectively leverage this intrinsic property of imaging systems.

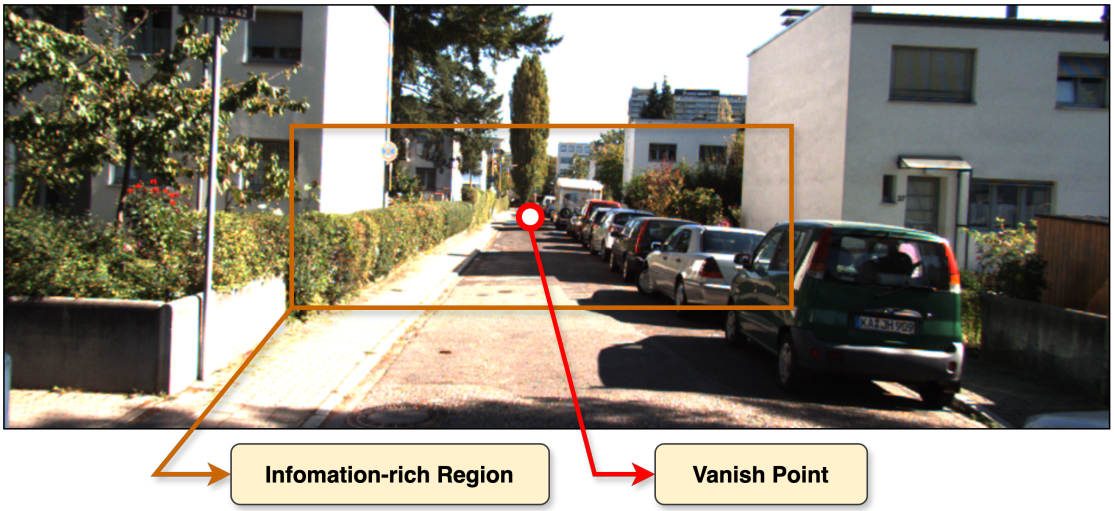


Figure 3.1: Illustration of the Vanishing Point and the associated Information-rich Region in a typical driving scene. The VP, indicated by the red circle, represents the convergence of parallel lines. The surrounding orange-bounded area highlights the Information-rich Region, where features from distant objects are concentrated due to perspective projection, forming an area of high information density crucial for far-field perception.

To address these limitations and capitalize on the observed information density in the VP region, we propose to augment the instance query mechanism by incorporating explicit spatial guidance through a novel Vanishing Point Query (VPQ), denoted as Q_{vp}. Unlike Symphonies, which relies on instance queries learning to attend to relevant global image features implicitly, our VPQ is explicitly designed to flexibly yet intensively focus on this high-density VP region Ω. By directing attention specifically to Ω, the VPQ aims

doi:10.6342/NTU202503382

to extract richer, more discriminative features for objects situated at a distance, which inherently project to this area. This targeted attention mechanism is hypothesized to provide more robust and accurate feature representations for small or distant objects, mitigating the feature dilution effect seen in purely global interaction strategies and thereby significantly improving performance on these vital categories.

3.3 Overview

Building upon the motivation to enhance the perception of distant and small objects critical for autonomous navigation, we propose an innovative model architecture that strategically integrates the Vanishing Point Query (VPQ) with conventional instance queries (Q_{ins}). The overarching goal of this integration is to effectively capture and prioritize both the distinct spatial and rich semantic information emanating from the region surrounding the vanishing point in the input image. A conceptual illustration of our approach, highlighting its key components and their interplay, is presented in Fig. 3.2.

Our architecture meticulously orchestrates this integration through three principal stages, each specifically designed to progressively enhance the feature representation derived from, or guided by, the VP region Ω:

1. **VPQ Initialization:** The Vanishing Point Queries (Q_{vp}) are initialized by first identifying the image's vanishing point (ρ). Reference points are then strategically sampled, prioritizing locations near ρ based on their distance, to ensure focus on this information-rich region. Local image features are extracted around these sampled points and subsequently formatted into the initial Q_{vp} set. This procedure grounds the queries in the geometry of the vanishing point and seeds them with pertinent

doi:10.6342/NTU202503382

local visual information from its surrounding area.

2. **VPQ Updating with Regional Features:** Following initialization, the Q_{vp} undergoes an update process where it directly interacts with and aggregates features specifically extracted from the pre-defined VP region Ω of the image features. This targeted interaction ensures that the VPQ becomes a rich embedding of the most relevant information contained within this crucial far-field zone, capturing details that might be overlooked by queries interacting with the entire image indiscriminately.

3. **VP-Instance Query Aggregation:** In the final stage, the updated, contextually rich Q_{vp} is synergistically aggregated with the standard instance queries (Q_{ins}). This aggregation process is designed to allow the instance queries to benefit from the specialized spatial and semantic insights captured by the VPQ. By fusing these query types, the model can leverage both the general object-level understanding of Q_{ins} and the distant-object-centric, perspective-aware information from Q_{vp}, leading to a more comprehensive and accurate feature representation, particularly for elements residing in or near Ω.

Each of these components is carefully designed to ensure that the enhanced feature representations accurately reflect the geometric and semantic nuances of the critical VP region. The detailed mechanisms and formulations underlying these stages will be elaborated upon in the subsequent sections.

doi:10.6342/NTU202503382

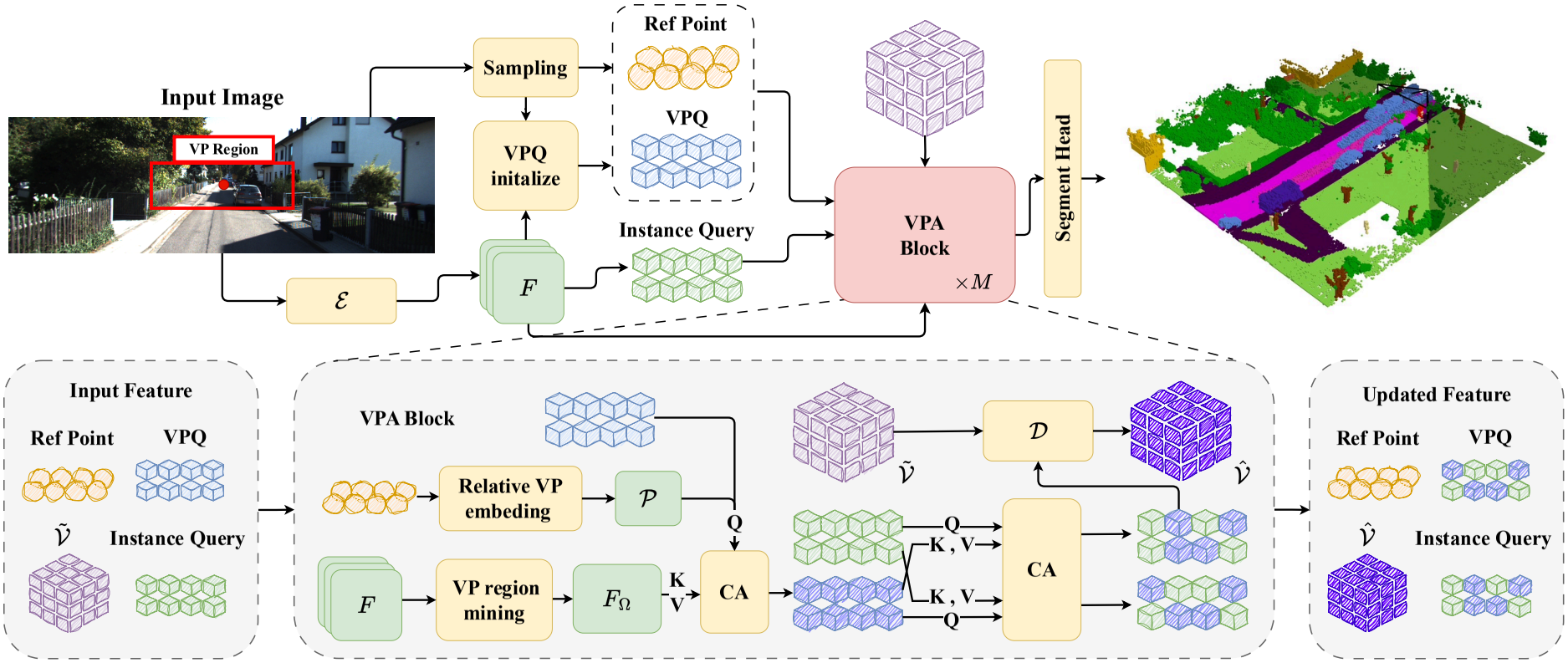


Figure 3.2: **Overview of proposed model.** The pipeline of our method consists of three procedures: (a) initializing VPQ and instance queries from the encoded image features F. (b) Vanishing Point Aggregator (VPA), which dynamically aggregates pivot information from the VP region to enhance the VPQ. (c) VP-instance queries aggregation, where VPQ and instance queries are integrated using cross-attention, followed by deformable attention to refine 3D voxel features.

doi:10.6342/NTU202503382

3.4 Proposed Method

In this section, we present the main components of our approach. We begin by explaining how Vanishing Point Queries (VPQ) are defined and initialized from the encoded image features. Next, we describe the Vanishing Point Aggregator (VPA) that refines these queries by collecting geometric cues from the VP region. Finally, we show how the enhanced VPQ are fused with instance queries via cross-attention and deformable attention to produce accurate 3D voxel features.

3.4.1 Vanishing Point Query (VPQ) Initialization

The initialization of Vanishing Point Queries (Q_{vp}) is a multi-step process designed to generate queries inherently focused on the VP region and seeded with relevant local image features. This procedure, illustrated conceptually in Fig. 3.3, unfolds as follows:

1. **Vanishing Point Detection:** First, the vanishing point (ρ) of the input image is detected. This can be achieved using established computer vision techniques, for this work, I adapted an off-the-shelf algorithm — Hough transform-based edge detector that identifies converging lines indicative of perspective depth.

2. **Reference Point Sampling:** Subsequently, a set of N_{vp} reference points (r_i, where $i = 1, \ldots, N_{vp}$) are sampled from the 2D image plane. To ensure that these points are concentrated around the area of interest, the sampling probability $P(r_i)$ for each point r_i is made inversely proportional to the square of its Euclidean distance $d(r_i, \rho)$

doi:10.6342/NTU202503382

from the detected vanishing point ρ:

$$P(r_i) \propto \frac{1}{d(r_i, \rho)^2}. \tag{3.2}$$

This probabilistic sampling strategy biases the selection towards regions closer to the vanishing point, effectively anchoring the subsequent queries in this information-rich area.

3. **Local Feature Extraction:** For each sampled reference point position r_i, local image features are captured from the multi-scale image feature maps F^{2D}, which obtained from DINO image backbone feature extractor. This is done by applying a set of small convolutional kernels (e.g., with varying receptive field sizes such as 3×3, 5×5, 7×7) centered at the projected locations of r_i onto the feature maps. This step embeds local contextual visual information from the vicinity of the reference points into preliminary feature vectors.

4. **Query Formatting:** Finally, the N_{vp} extracted feature vectors are processed and reshaped to form the initial set of N_{vp} Vanishing Point Queries, $Q_{\text{vp}} \in \mathbb{R}^{N_{vp} \times C}$, where N_{vp} is the number of VP queries equal to the number of instance queries N_{ins} and C is the feature dimension. This processing involve a sequence of linear projections to match the required query size and format of our downstream iterative refinement module

The rationale behind this initialization process is to establish VP-related queries that are inherently tied to the geometry of the vanishing point and are initialized using discriminative local image features sourced directly from its surrounding high-density region Ω. Standard instance queries, $Q_{\text{ins}} \in \mathbb{R}^{N_{\text{ins}} \times C}$, are typically initialized as learnable embed-

doi:10.6342/NTU202503382

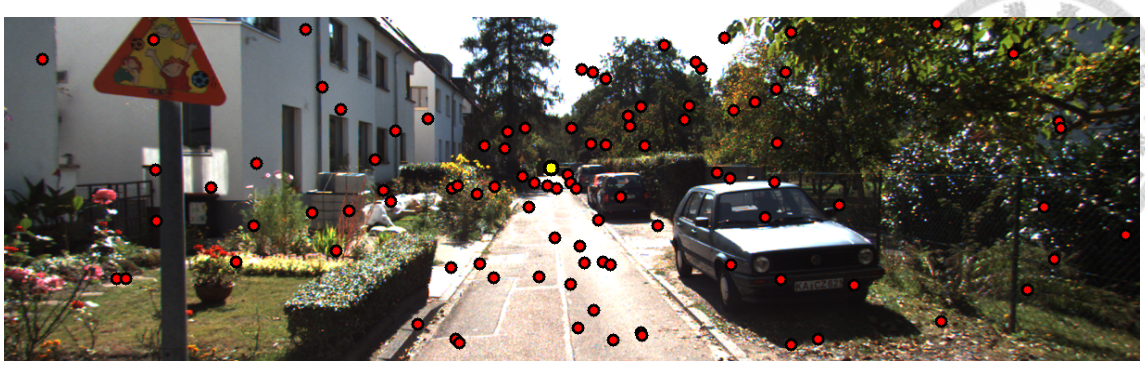


Figure 3.3: **Visualization of Reference and Vanishing Points.** The red marker denotes the reference points generated during the initialization steps, and the yellow marker indicates the vanishing point detected in this image.

dings, similar to common practices in transformer-based object detectors.

3.4.2 Vanishing Point Query Update with Regional Features

After their initial formation, the Vanishing Point Queries (Q_{vp}) undergo a dedicated update step to further enrich them with aggregated information specifically from the Vanishing Point Region, denoted as Ω. The goal is to refine each Q_{vp} by enabling it to attend to a consolidated representation of this critical image area. This process involves the following operations:

1. **Defining the VP Region :** A specific sub-region of the input image's feature map is designated as the VP feature region, Ω. This region is defined as a central area of size $H/d \times W/d$ of the high-resolution image feature map, where H and W are the height and width of input image and d is the hyperparameter to determine the size of VP region. This region is expected to encapsulate the vital visual cues converging towards the vanishing point.

2. **Extracting VP Region Features (F_Ω):** To obtain a compact yet informative representation of Ω, a low-stride multi-layer Convolutional Neural Network (CNN)

doi:10.6342/NTU202503382

is applied exclusively to this defined region, primarily leveraging high-resolution image features. This multi-layer CNN acts as a feature sampler and aggregator, processing the visual information within Ω to produce a set of condensed VP region features, denoted as F_Ω.

3. **Positional Encoding for VP Queries:** To provide spatial context to the Q_{vp} relative to the center of the VP region or the detected vanishing point itself, a positional embedding (PE_{vp}) is computed for each query. This embedding can be based on the normalized distance of each query's conceptual anchor point from the detected vanishing point. We reuse the distance function in (3.2) to encode the positional information from the r_i to ρ. The positional embedding could be formulate as

$$PE_{\text{vp}} = \phi(d(r_i, \rho)), \tag{3.3}$$

where ϕ is the relative embedding function. This PE_{vp} is then added to the existing Q_{vp} features.

4. **Cross-Attention with Regional Features:** Finally, the VP queries are updated via a cross-attention mechanism. The positionally-aware VP queries ($Q_{\text{vp}} + PE_{\text{vp}}$) serve as the *Query* (Q) sequence. The aggregated VP region features (F_Ω) extracted in previous step serve as both the *Key* (K) and *Value* (V) sequences.

$$Q'_{\text{vp}} = \text{CrossAttn}(Q_{\text{vp}} + PE_{\text{vp}}, F_\Omega, F_\Omega). \tag{3.4}$$

This attention step allows each VP query to selectively draw information from the most relevant parts of the consolidated VP region features F_Ω, resulting in updated queries Q'_{vp} that are more attuned to the holistic context of the far-field information.

doi:10.6342/NTU202503382

These updated Q'_{vp} are then utilized in the subsequent iterative refinement module.

3.4.3 Cross-Source Query Update with Initial Voxel Features

Once the Vanishing Point Query (Q'_{vp}) has been enriched with features from the VP region (see Sec. 3.4.2), we perform a cross-source cross-attention to fuse information between these specialized queries and the standard instance queries (Q_{ins}). Concretely, we simultaneously update both sets of queries:

First, the instance queries incorporate far-field context by attending to the enriched VP queries:

$$\hat{Q}_{\text{ins}} = \text{CrossAttn}\big(Q_{\text{ins}},\, Q'_{\text{vp}},\, Q'_{\text{vp}}\big). \tag{3.5}$$

At the same time, the VP queries assimilate object-centric semantics by attending to the instance queries:

$$\hat{Q}'_{\text{vp}} = \text{CrossAttn}\big(Q'_{\text{vp}},\, Q_{\text{ins}},\, Q_{\text{ins}}\big). \tag{3.6}$$

Here, $\hat{Q}_{\text{ins}}$ and $\hat{Q}'_{\text{vp}}$ denote the co-refined instance and VP queries after this bidirectional fusion.

By enabling this cross-source co-refinement, our model retains its ability to capture global scene information while enhancing representations of small or distant objects. These enriched queries then serve as more informative inputs for subsequent interactions with the 3D scene voxel features.

doi:10.6342/NTU202503382

3.4.4 Initial Scene Voxel Feature Generation

Parallel to the preparation of the specialized Vanishing Point Queries and instance queries , an initial 3D volumetric representation of the scene is established. This process, commonly referred to as a Voxel Proposal Layer (VPL), aims to generate features for a sparse set of proposed voxels v_{p} that are estimated to correspond to occupied surfaces within the camera's field of view. This provides a geometrically grounded, albeit coarse, foundation for subsequent 3D scene understanding and refinement. The generation of these initial scene voxel features involves the following key operations:

1. **Depth Map Prediction:** A monocular depth estimation network processes the input RGB image I to predict a corresponding dense depth map D. This depth map provides an estimate of the distance z_c from the camera center to each visible point in the scene.

2. **3D World Point Generation from Depth:** Using the predicted depth $z_c = D(x^I)$ for each pixel x^I (represented in homogeneous coordinates), along with the camera's intrinsic matrix K and extrinsic parameters (rotation R, translation T), each pixel is projected into a 3D point x^W in the world coordinate system. This transformation is typically performed in two steps: first to camera coordinates x^C, then to world coordinates x^W:

$$x^C = K^{-1} \cdot (z_c x^I) \tag{3.7}$$

$$x^W = [R, T]^{-1} \cdot x^C. \tag{3.8}$$

This effectively project the 3D world points representing the visible surfaces from

doi:10.6342/NTU202503382

the camera's perspective.

3. **Voxel Grid Association:** The generated 3D world points x^W are then associated with a predefined discrete 3D voxel grid that spans the target scene volume $\mathcal{V}$. Points x^W falling into the same voxel cell contribute to the potential occupancy of that voxel. This step identifies which voxels are likely to be occupied based on the projected depth information, forming a set of candidate voxel locations V_p .

4. **Voxel Feature Initialization and Refinement via Deformable Attention:** Let $q_{\text{vox}} \in \mathbb{R}^{N_{\text{vox}} \times C}$ represent an initial feature embedding for all voxels within the entire scene volume $\mathcal{V}$,which is learnable embeddings features from a coarse 3D representation. We select the features corresponding to the set of candidate voxel locations V_p. Let these selected initial features be q_p. These selected features q_p are then further refined by aggregating information from the multi-scale 2D image features F^{2D} using a deformable attention mechanism. The 3D positions V_p of these voxel proposals are projected onto the image plane to obtain corresponding 2D reference points P_I. The deformable attention allows each selected voxel proposal q_p (acting as the query) to dynamically sample relevant visual information from F^{2D} (acting as keys and values) around its projected location P_I:

$$q'_p = \text{DeformAttn}(q_p, P_I, F^{\text{2D}}). \tag{3.9}$$

The resulting refined features q'_p will constitute the set of active initial voxel features $q_{\text{vox}} \in \mathbb{R}^{N_{\text{vox}} \times C}$ that will be used in subsequent processing, where N_{vox} is the number of voxel in 3D coordination.

This VPL thus furnishes an initial, sparse set of active 3D voxel features q_{vox}, which

doi:10.6342/NTU202503382

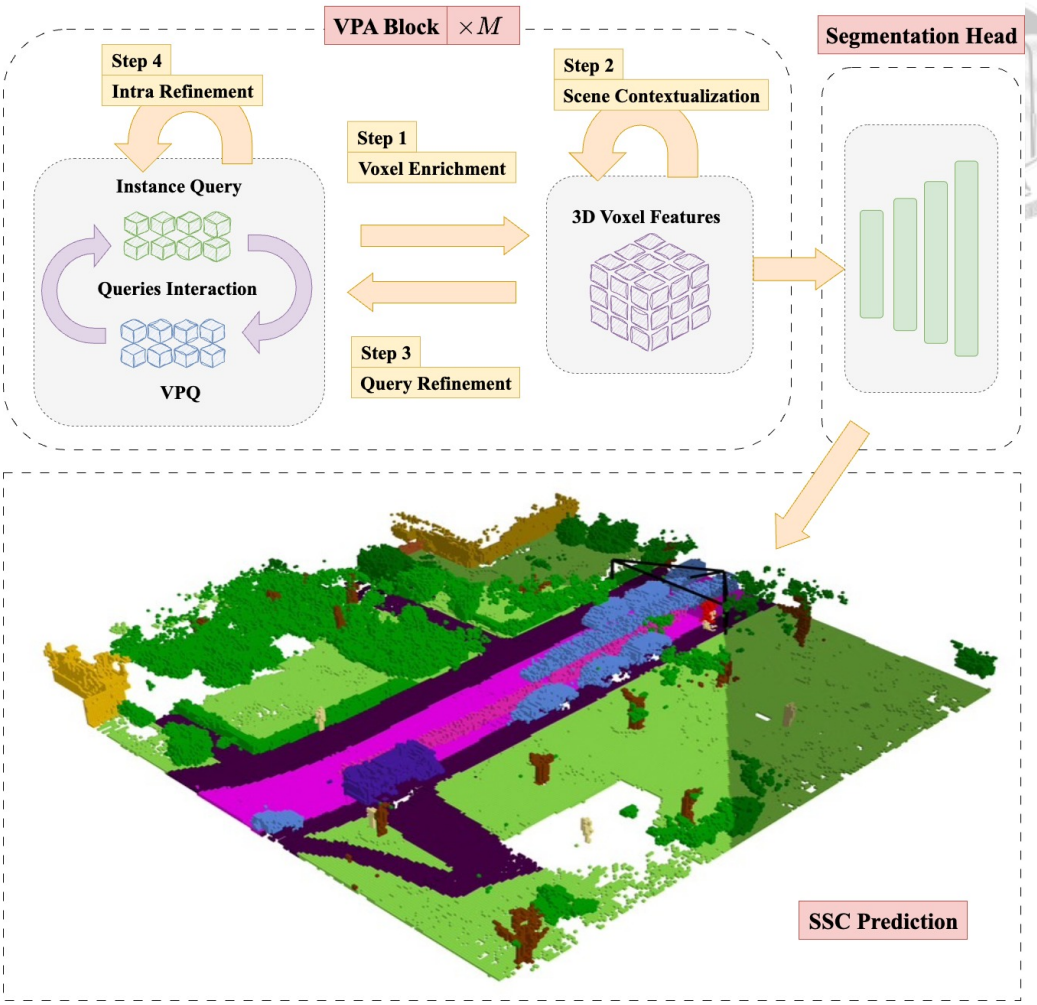


Figure 3.4: **Detailed view of the decoder's Iterative Feature Co-Refinement Module.** This image shows the decoder steps (1) voxel enrichment by instance and vanishing-point queries, (2) deformable self-attention for scene contextualization, (3) cross-attention query refinement, and (4) intra-query self-attention—all co-refining 3D voxel features and queries over M iterations before the final segmentation head.

are geometrically grounded to the estimated visible surfaces and contextually enriched by relevant image information. These initial scene voxel features q_{vox}, along with the co-refined Vanishing Point Queries and instance queries from Sec. 3.4.3, serve as primary intermediary, providing essential information for the interaction between 2D image feature F^{2D} and 3D voxel feature q_{vox}.

doi:10.6342/NTU202503382

3.4.5 Iterative Feature Co-Refinement Module

With initial scene voxel features (q_{vox}), co-refined Vanishing Point Queries ($\hat{Q}'_{\text{vp}}$), and instance queries ($\hat{Q}_{\text{ins}}$) prepared, our framework employs a core iterative refinement module. This module, illustrated in Fig. 3.4, is executed for N_{iter} iterations to progressively enhance all three representations through a synergistic exchange of information across image, instance, and 3D scene domains. Within each iteration, a sequence of attention-based interactions occurs:

1. **Voxel Enrichment by Queries:** The 3D voxel features (q_{vox}) are updated by aggregating information from the queries $\hat{Q}_{\text{ins}}$ through cross-attention, infusing q_{vox} with instance semantics and VP-guided far-field cues.

2. **Scene Contextualization:** A deformable self-attention mechanism is applied to q_{vox}, allowing voxel features to propagate and capture broader 3D contextual relationships within the scene volume.

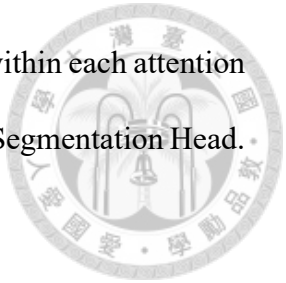
3. **Query Refinement by Scene Context:** Both $\hat{Q}_{\text{ins}}$ and $\hat{Q}'_{\text{vp}}$ are further updated by attending to the contextually enriched q_{vox} via deformable cross-attention, allowing them to incorporate 3D structural information.

4. **Intra-Query Refinement:** Finally, self-attention is applied independently to the sets of $\hat{Q}_{\text{ins}}$ and $\hat{Q}'_{\text{vp}}$, enabling them to refine their representations by capturing internal consistencies and relationships among queries of the same type.

This cycle of interactions allows the specialized $\hat{Q}'_{\text{vp}}$ to continuously guide scene understanding towards distant elements, while $\hat{Q}_{\text{ins}}$ hones in on object-specific details, all contributing to a progressively more accurate and complete 3D scene representation

doi:10.6342/NTU202503382

q_{vox}. Standard transformer block components (FFN, LN) are utilized within each attention operation. After N_{iter} iterations, the final refined q_{vox} are passed to the Segmentation Head.



3.4.6 Segmentation Head

Following N_{iter} iterations of the co-refinement module, the final enhanced scene voxel features, $q_{\text{vox}}^{(N_{\text{iter}})} \in \mathbb{R}^{N_{\text{vox}} \times C}$, encapsulate a rich, multi-scale understanding of the 3D scene. These features are then processed by a Segmentation Head to produce the ultimate per-voxel semantic predictions.

Our Segmentation Head is designed to effectively translate these dense features into class probabilities. It typically involves these operations:

1. **Feature Upsampling:** If the voxel features $q_{\text{vox}}^{(N_{\text{iter}})}$ are at a coarser resolution than the desired output grid (e.g., $256 \times 256 \times 32$), an upsampling mechanism is employed. This can be achieved using 3D transposed convolutions or trilinear interpolation to restore the features to the target spatial dimensions while preserving learned details.

2. **Atrous Spatial Pyramid Pooling (ASPP):** To capture contextual information at multiple scales robustly before the final classification, we incorporate an Atrous Spatial Pyramid Pooling (ASPP) module [5]. The ASPP module applies several parallel 3D dilated convolutions with different dilation rates to the voxel features. This allows the network to probe features at various receptive fields, effectively capturing object and scene context at multiple scales. An image pooling branch also be included. The outputs of these parallel branches are then concatenated and passed through a 1×1 bottleneck convolution to fuse the multi-scale features.

3. **Final Classification Layer:** Finally, the last of linear layers reduces the feature

doi:10.6342/NTU202503382

dimension to N_{classes}, where N_{classes} is the total number of semantic categories . This produces the final per-voxel class logits. A softmax activation is then applied to these logits during inference to obtain class probabilities.

3.4.7 Loss Functions

To train our comprehensive model, we employ a carefully designed compound loss function $\mathcal{L}_{\text{total}}$ that addresses both the geometric completion and semantic segmentation aspects of the SSC task, while also accounting for inherent class imbalances.

1. **Semantic Segmentation Loss (L_{sem}):** For the semantic labeling of each voxel, we primarily use a weighted cross-entropy loss. Given the significant class imbalance typical in SSC datasets, where classes like 'road' or 'vegetation' dominate over rare classes like 'cyclist', a weighting scheme is applied. The weight for each class is inversely proportional to its frequency in the training set:

$$L_{\text{sem}} = -\frac{1}{|\mathcal{V}_{\text{valid}}|} \sum_{v \in \mathcal{V}_{\text{valid}}} \sum_{c=1}^{N_{\text{classes}}} w_c \cdot y_{v,c} \log(p_{v,c}), \tag{3.10}$$

where $\mathcal{V}_{\text{valid}}$ are the voxels considered for loss computation, $y_{v,c}$ is the ground truth label (1 if voxel v belongs to class c, 0 otherwise), $p_{v,c}$ is the predicted probability, and w_c is the class weight.

2. **Geometric Completion Loss (L_{geo}):** To explicitly supervise the geometric structure and the relationship between semantic predictions and occupancy, we incorporate a Scene-Class Affinity Loss (L_{scal}), similar to that introduced in MonoScene [4]. This loss encourages consistency between semantic predictions and the predicted occupancy, penalizing semantically labeled voxels that are geometrically empty and

doi:10.6342/NTU202503382

vice-versa. The geometric completion loss could be formulated as:

$$L_{\text{geo}} = L_{\text{scal}}^{\text{geo}} + L_{\text{scal}}^{\text{sem}}, \tag{3.11}$$

where $L_{\text{scal}}^{\text{geo}}$ focuses on the binary occupancy prediction and $L_{\text{scal}}^{\text{sem}}$ ensures that semantic predictions are consistent within occupied regions.

3. **Total Loss and Auxiliary Supervision:** The total loss is a weighted sum of these components:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{sem}} L_{\text{sem}} + \lambda_{\text{geo}} L_{\text{geo}}, \tag{3.12}$$

where λ_{sem} and λ_{geo} are hyperparameter weights balancing the different loss terms. Furthermore, to encourage effective learning throughout the depth of our iterative co-refinement module, auxiliary losses are applied after each of the N_{iter} iterations. The voxel features from each intermediate iteration are passed through auxiliary segmentation heads to predict class logits. The total loss formulation (Eq. 3.12) is then applied to these intermediate predictions. This deep supervision strategy helps mitigate vanishing gradients and guides the intermediate layers to produce more meaningful representations.

This comprehensive loss structure ensures that our model learns to accurately predict both the complete 3D geometry and the semantic labels of the scene, with particular attention to handling class imbalance and fostering coherent scene understanding.

doi:10.6342/NTU202503382

Chapter 4 Experiments

4.1 Datasets

We rigorously evaluate our proposed method on two widely adopted public benchmarks for 3D semantic scene completion: SemanticKITTI [3] and SSCBench-KITTI-360 [15]. Both datasets are derived from the KITTI Odometry Benchmark [8], providing rich outdoor driving scenarios.

SemanticKITTI [3] offers dense, voxel-wise semantic annotations for 20 distinct classes across 21 LiDAR sequences, sequences 00-10 for training, 08 for validation, and 11-21 for testing. The dataset provides RGB images corresponding to the voxel label processed from LiDAR scans aggregation.

SSCBench-KITTI-360 [15] is a more recent benchmark built upon the KITTI-360 dataset [17], providing annotations for 19 semantic classes across 9 driving sequences. It features more diverse viewpoints and longer trajectories compared to the original KITTI Odometry sequences.

For both datasets, the standard evaluation protocol defines a 3D scene volume spanning $51.2\text{m} \times 51.2\text{m} \times 6.4\text{m}$ centered around the ego-vehicle. This volume is discretized into a $256 \times 256 \times 32$ voxel grid, resulting in a voxel size of $0.2\text{m} \times 0.2\text{m} \times 0.2\text{m}$. Our

doi:10.6342/NTU202503382

method, being camera-based, utilizes the provided RGB images as input.

4.2 Evaluation Metrics

To assess the performance of our approach, we employ standard evaluation metrics consistent with prior works in semantic scene completion [4, 11, 12, 16, 20, 22, 23]. This ensures fair and direct comparability with existing state-of-the-art methods. The primary metrics are:

- **Intersection over Union (IoU) for Scene Completion (SC):** This metric evaluates the accuracy of the geometric completion, i.e., the binary prediction of whether each voxel is occupied or empty, without regard to semantic labels. It is calculated as the ratio of the intersection to the union of the predicted and ground truth occupied voxels.

- **Mean Intersection over Union (mIoU) for Semantic Scene Completion (SSC):** This metric assesses the class-specific prediction accuracy. The IoU is calculated independently for each of the N_{classes} semantic categories, and the mIoU is the average of these per-class IoUs. This provides a comprehensive measure of how well the model identifies both the geometry and the correct semantic label of each occupied voxel.

- **Small Object IoU (sIoU):** To specifically evaluate performance on classes that are typically small, distant, or otherwise challenging to perceive accurately, we define and report the Small Object IoU (sIoU). This metric is calculated as the mean Intersection over Union (mIoU) exclusively over a predefined subset of these critical

doi:10.6342/NTU202503382

small object categories. For instance, on the KITTI-360 benchmark, these classes include car, bicycle, motorcycle, other-vehicle, person, fence, pole, traffic sign, and other-objects. The sIoU provides a focused measure of the model's efficacy in handling these safety-critical and perceptionally demanding elements.

We report results based on the official evaluation scripts and protocols provided by each benchmark where available. Detailed per-class IoU scores and qualitative visualizations, particularly for the SemanticKITTI dataset, are provided in the following sections to offer a more granular insight into our method's performance characteristics.

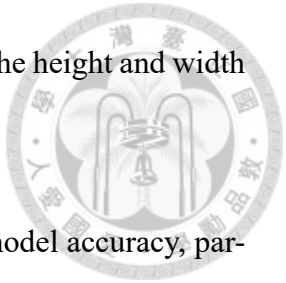
4.3 Implementation Details

All experiments are conducted on a single NVIDIA RTX 3090 GPU, highlighting the model's potential for computational efficiency. Our model is trained for a total of 30 epochs. We utilize the ResNet-50 [10] architecture as the image backbone, and the subsequent image encoder components are initialized with pre-trained weights from MaskDINO [14], which is recognized for its strong performance in various dense prediction tasks. For the optimization process, we employ the AdamW optimizer [18] with an initial learning rate of 2×10^{-4} and a weight decay of 1×10^{-4}. A MultiStepLR learning rate schedule is adopted to adjust the learning rate during training, reducing it at predefined epoch milestones.

Regarding our proposed Vanishing Point Aggregator (VPA), the number of VPA blocks , which corresponds to the number of iterations N_{iter} in Sec. 3.4.5, is set to 3. We use $N_{\text{vp}} = 100$ Vanishing Point Queries (VPQs) and an equal number of instance queries, $N_{\text{ins}} = 100$. The feature dimension C for both queries and voxel representations is set to 128. The VP region, crucial for guiding the VPQs, is defined as the central $H/3 \times W/3$

doi:10.6342/NTU202503382

area of the relevant high-resolution feature map, where H and W are the height and width of the input image.



This configuration aims to strike an effective balance between model accuracy, particularly for small and distant objects prioritized by our VPA, and overall computational efficiency.

4.4 Baseline Methods

To demonstrate the efficacy of our proposed approach, we conduct a comprehensive comparison against several current state-of-the-art camera-based Semantic Scene Completion methodologies. These include Symphonies [12], MonoScene [4], TPVFormer [11], VoxFormer [16], OccFormer [23], NDC-Scene [22], and H2GFormer [20]. The selected baselines represent leading advancements in leveraging monocular RGB imagery for dense 3D scene reconstruction and semantic understanding.

4.5 Quantitative Results

We present a comprehensive quantitative evaluation of our proposed method against state-of-the-art camera-based Semantic Scene Completion approaches. The performance is assessed on the SemanticKITTI validation set and the SSCBench-KITTI-360 test set.

4.5.1 Results on SemanticKITTI Validation Set

Table 4.1a details the comparative performance on the SemanticKITTI validation set. Our VPA method achieves a state-of-the-art mIoU of **15.26%**, surpassing all listed

doi:10.6342/NTU202503382

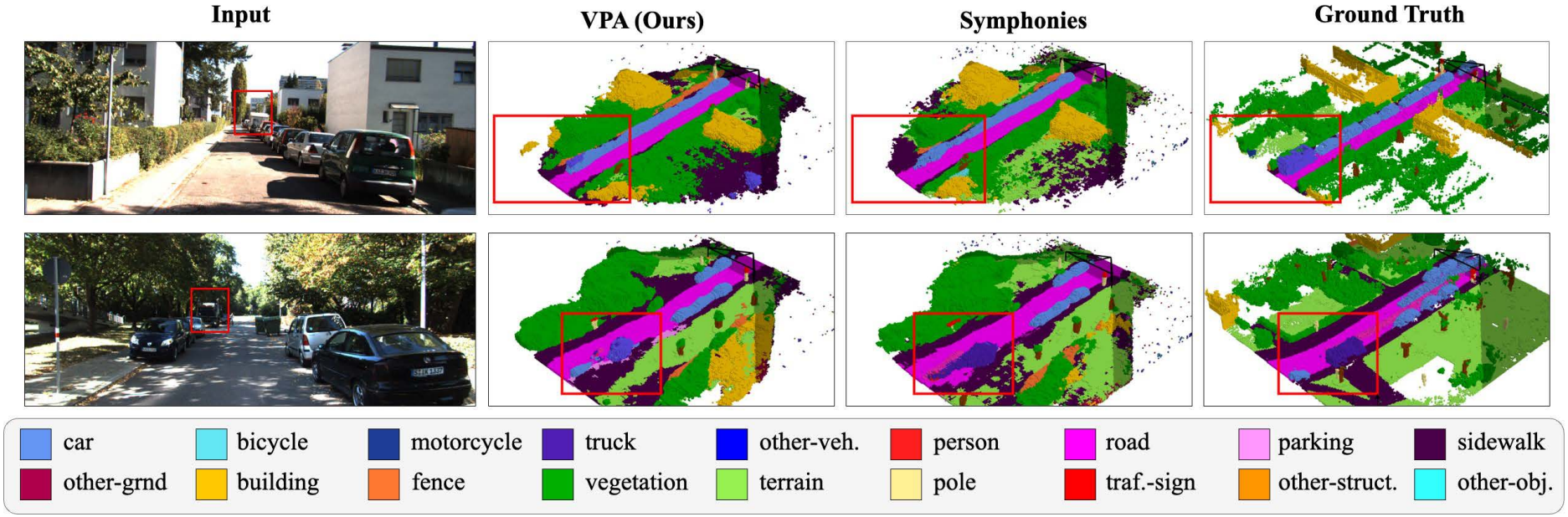


Figure 4.1: **Results visualize on SemanticKITTI.** The regions marked with red boxes in the figure demonstrate that our method is capable of accurately detecting and locating small objects, even in challenging scenarios such as long distance prediction.

doi:10.6342/NTU202503382

baseline methods. Notably, this represents an improvement of 0.37 mIoU points over the strong Symphonies [12] baseline with 14.89% mIoU, underscoring the efficacy of our VPQ-guided strategy in refining feature representations for enhanced scene understanding. The overall IoU for scene completion stands at a competitive 41.99%.

Analyzing per-class performance, VPA demonstrates its robustness by achieving the highest IoU scores in several key categories. For large, structured elements and dense regions, our method excels, securing top performance for **building** (22.01%), **car** (29.03%), and **vegetation** (25.39%). Additionally, VPA outperforms others in challenging categories, such as **truck** (19.27%) and **bicycle** (3.55%), where accurate long-range predictions are critical.

A critical advantage of our VPA is its enhanced capability in perceiving smaller, yet crucial, objects. This is evidenced by leading scores for **person** (4.07%) and **traffic sign** (5.91%). While categories with extremely sparse instances and small visual footprints, such as **motorcyclist** (0.00% for all methods) and **bicyclist** (2.40% for VPA, with VoxFormer-S at 3.32%), remain universally challenging, our method's competitive performance in **person** and superior results in **bicycle** and **traffic sign** highlight the benefits of the VPQ and instance query fusion for prioritizing these safety-critical elements. The overall strong performance, particularly the improvements in recognizing small and distant objects, validates the core contributions of our VPA in leveraging perspective cues for more accurate semantic predictions in dense urban environments.

doi:10.6342/NTU202503382

Method	IoU	mIoU	sIoU	road	sidewalk	parking	building	truck	vegetation	trunk	terrain	other-grnd.	car	bicycle	motorcycle	other-veh.	person	bicyclist	motorcyclist	fence	pole	traf.-sign
MonoScene	36.86	11.08	3.78	56.52	26.72	14.27	14.09	6.98	17.89	2.81	29.64	0.46	23.26	0.61	0.45	1.48	1.86	1.20	0.00	5.84	4.14	2.25
TPVFormer*	35.61	11.36	3.77	56.50	25.87	20.60	13.88	8.08	16.92	2.26	30.38	0.85	23.81	0.36	0.05	4.35	0.51	0.89	0.00	5.94	3.14	1.52
VoxFormer-S	44.02	12.35	5.04	54.76	26.35	15.50	17.65	5.63	24.39	5.08	29.96	0.70	25.79	0.59	0.51	3.77	1.78	**3.32**	0.00	7.64	7.11	4.18
OccFormer	36.50	13.46	4.93	58.85	26.88	19.61	14.40	**25.53**	19.63	3.93	32.62	0.31	25.09	0.81	1.19	8.52	2.78	2.82	0.00	5.61	4.26	2.86
NDC-Scene	37.24	12.70	5.45	**59.20**	28.24	**21.42**	14.94	14.75	19.09	3.51	31.04	**1.67**	26.26	1.67	2.37	7.73	3.60	2.74	0.00	6.65	4.53	2.73
H2GFormer-S	**44.57**	13.73	5.51	56.08	**29.12**	17.83	19.74	10.00	**26.25**	**7.80**	**34.42**	0.45	27.60	0.50	0.47	7.39	1.54	2.88	0.00	7.24	7.88	4.68
Symphonies	41.92	14.89	7.12	56.37	27.58	15.28	21.64	20.44	25.72	6.60	30.87	0.95	28.68	2.54	**2.82**	**13.89**	3.52	2.24	0.00	8.40	**9.57**	5.76
VPA (Ours)	41.99	**15.26**	**7.19**	58.40	27.26	21.21	**22.01**	19.27	25.39	6.20	31.37	0.52	**29.03**	**3.55**	2.55	12.52	**4.07**	2.40	0.00	**9.19**	9.39	**5.91**

(a) **Quantitative results on SemanticKITTI `val`.** * represents the reproduced results in [11]. The sIoU is the mean IoU over the subset of classes on the right side of the single vertical line following the main non-sIoU classes. The best results are in **bold**.

Method	IoU	mIoU	sIoU	truck	road	parking	sidewalk	other-grnd.	building	vegetation	terrain	other-struct.	car	bicycle	motorcycle	other-veh.	person	fence	pole	traf.-sign	other-obj.
MonoScene	37.87	12.31	4.72	8.02	48.35	11.38	28.13	3.32	32.89	26.15	16.75	4.20	19.34	0.43	0.58	2.03	0.86	3.53	6.92	5.67	3.09
TPVFormer*	40.22	13.64	5.53	8.06	52.99	11.99	31.07	3.78	34.83	30.08	17.52	5.48	21.56	1.09	1.37	2.57	2.38	4.80	7.46	5.86	2.70
VoxFormer	38.76	11.91	4.93	4.56	47.01	9.67	27.21	2.89	31.18	28.99	14.69	3.79	17.84	1.16	0.89	2.06	1.63	4.97	6.51	6.92	2.43
OccFormer	40.27	13.81	6.20	9.89	54.30	13.44	31.53	3.55	**36.42**	31.00	**19.51**	6.95	22.58	0.66	0.26	3.82	2.77	4.80	7.77	8.51	4.60
Symphonies	44.12	18.58	11.28	**25.07**	54.94	13.83	32.76	**6.93**	35.11	38.33	11.52	**14.44**	30.02	**1.85**	5.90	12.06	8.20	8.58	14.01	9.57	11.28
VPA (Ours)	**44.87**	**18.80**	**11.78**	21.32	**56.64**	**14.45**	**33.38**	6.83	35.00	**38.41**	11.00	13.88	**30.60**	1.19	**8.31**	**13.29**	**8.42**	**8.74**	**14.51**	**10.07**	**11.88**

(b) **Quantitative results on SSCBench-KITTI-360 test.** The best results are in **bold**. Our method surpasses all previous methods across multiple metrics and shows substantial improvements in vary small object classes.

doi:10.6342/NTU202503382

4.5.2 Results on SSCBench-KITTI-360 Test Set

Table 4.1b presents the quantitative evaluation of our VPA method on the challenging SSCBench-KITTI-360 test set. Our approach establishes a new state-of-the-art, achieving the highest overall mean Intersection over Union (mIoU) of **18.80%**. This result outperforms all contemporary camera-based SSC methods, including the strong Symphonies baseline (18.58% mIoU). Furthermore, VPA also leads in overall geometric IoU with **44.87%**. Critically, our method excels in the Small Object IoU (sIoU) metric, securing a top score of **11.78%**. This superior performance across mIoU, IoU, and particularly sIoU, highlightss the comprehensive capabilities of our VP-guided architecture in accurately modeling both broad scene context and the fine-grained details crucial for small or distant object perception in complex driving scenarios.

Further analysis of per-class performance on KITTI-360 (Table 4.1b) reveals the breadth of VPA's effectiveness. For prevalent, large-scale environmental classes, our method secures top IoU scores, including **road** (56.64%), **sidewalk** (33.38%), **parking** (14.45%), and **vegetation** (38.41%). It also achieves leading performance for the crucial class **car** (30.60%) and maintains highly competitive results for complex structures like **building** (35.00%, close to the top score). This strong performance across diverse, large object types and expansive scene layouts underscores VPA's robust feature extraction and contextual refinement capabilities.

More significantly, our VPA demonstrates substantial advancements in the perception of smaller, often dynamic, and traditionally challenging object classes on this benchmark, which directly contribute to its leading sIoU score. As detailed in Table 4.1b, VPA achieves state-of-the-art IoU scores for a majority of these small object categories: **motor-**

doi:10.6342/NTU202503382

cycle (8.31%, a notable +2.41% improvement over Symphonies), **other-veh.** (13.29%), **person** (8.42%), **fence** (8.74%), **pole** (14.51%), **traffic sign** (10.07%), and **other-obj.** (11.88%). The marked improvements, particularly for vulnerable road users such as motorcycle and person, and critical navigational elements like traffic sign and pole, strongly affirm the benefits of our VPA's targeted attention to the vanishing point region for enhancing the perception of distant and small entities and highlighting its robustness in detecting small and spatially distributed objects. These comprehensive per-class results on KITTI-360 validate the precision, robustness, and generalization capabilities of our VP-guided approach in diverse urban driving scenarios.

4.6 Ablation Study

To dissect the contributions of our proposed query mechanisms and their interactions, we conduct a series of ablation studies on the SemanticKITTI validation set. The primary focus is to evaluate the effectiveness of integrating Vanishing Point Queries with standard Instance Queries.

4.6.1 Impact of Query Design and Interaction

To assess the effectiveness of Vanishing Point Queries, standard instance queries, and their integration, we perform an ablation study comparing how each query type bridges 2D and 3D information. Table 4.2 summarizes the performance achieved by different query configurations. Our baseline, employing **Instance Query Only**, yields an mIoU of 14.89%. This configuration, akin to the approach in Symphonies [12], demonstrates the inherent capability of instance queries to capture global contextual information from

doi:10.6342/NTU202503382

the scene. When utilizing **VP Query Only**, which interacts exclusively with the critical VP region, the model achieves an mIoU of 14.42%. While effective for its targeted area, this focused interaction limits its ability to capture non-VP-centric global information, resulting in a slightly lower overall mIoU compared to using instance queries alone.

A naive **Mix Query** approach, which directly sums the features of Instance Queries and VPQs one-by-one, results in an mIoU of 14.47%. This marginal change from ”VP Query Only” and slight degradation compared to ”Instance Query Only” indicates that simple additive combination is suboptimal for integrating these distinct query types and may introduce conflicting signals. In contrast, our proposed **Fusion Query** method, which employs a dedicated cross-source cross-attention mechanism to intelligently integrate the VPQs and Instance Queries, achieves the highest performance with an mIoU of **15.26%** and an IoU of **41.99%**. This result clearly demonstrates the superiority of the sophisticated fusion strategy. The significant mIoU gain of +0.37% over ”Instance Query Only” and +0.79% over ”Mix Query” validates that the cross-source cross-attention effectively synergizes the global contextual understanding from instance queries with the spatially prioritized, far-field insights from VPQs. This intelligent fusion mitigates the limitations of using either query type in isolation or combining them naively, leading to an enhanced overall scene comprehension and validating the core design principle of our VPA framework.

4.6.2 Performance Analysis across Distance Ranges

To specifically investigate the impact of our Vanishing Point Aggregator (VPA) on perceiving objects at varying distances, we evaluate its performance against the strong Symphonies baseline across different depth-based settings on both SemanticKITTI and

doi:10.6342/NTU202503382

Query Configuration	IoU	mIoU
Instance Query Only (Baseline)	41.92	14.89
VP Query Only	41.86	14.42
Mix Query	41.90	14.47
Fusion Query (Ours)	**41.99**	**15.26**

Table 4.2: Ablation study on different query designs and their interaction mechanisms on the SemanticKITTI `val` set. "Instance Query Only" serves as our baseline, akin to Symphonies. "Fusion Query" refers to our proposed method with cross-source cross-attention.

Dataset	Setting	Symphonies	VPA(Ours)
SemanticKitti	Far	5.17	**5.81**
	Medium	13.92	**14.52**
	Full	14.89	**15.26**
SSCBench-KITTI-360	Far	7.61	**8.23**
	Medium	17.33	**17.69**
	Full	18.58	**18.80**

Table 4.3: Comparison of model performance on different distance settings. Far covers depth range from 38.4m to 51.2m and Medium covers range from 25.6m to 51.2m.

SSCBench-KITTI-360. The results, presented in Table 4.3, categorize performance, measured relative to the front of the camera, into "Far" (38.4m to 51.2m), "Medium" (25.6m to 51.2m), and "Full" (entire range) settings based on object depth.

On SemanticKITTI, our VPA method demonstrates consistent improvements over Symphonies across all distance ranges. Most notably, in the challenging "Far" range, VPA achieves an mIoU of **5.81%**, outperforming Symphonies (5.17%) by a significant margin of +0.64%. This advantage is maintained in the "Medium" range (VPA: **14.52%** vs. Symphonies: 13.92%) and for the "Full" scene (VPA: **15.26%** vs. Symphonies: 14.89%).

A similar trend is observed on the SSCBench-KITTI-360 dataset. For "Far" distance predictions, VPA scores **8.23%** mIoU compared to Symphonies' 7.61%, an improvement

doi:10.6342/NTU202503382

of +0.62%. In the ”Medium” range, VPA achieves **17.69%** (vs. 17.33%), and for the ” Full” range, it records **18.80%** (vs. 18.58%).

These results compellingly demonstrate the effectiveness of VPA in enhancing long-distance predictions, a scenario where conventional methods often exhibit degraded performance. By explicitly prioritizing the information-rich VP region and effectively fusing these specialized cues with broader contextual understanding, our method achieves superior performance, particularly in challenging long-range scenarios. This capability is critical for robust perception in autonomous driving, where timely detection of distant entities is paramount for safety.

4.6.3 Effect of Vanishing Point Region Size

The definition of the Vanishing Point (VP) region, from which specialized features are extracted to guide the VP Queries, is a critical hyperparameter in our VPA framework. To determine an optimal configuration, we conducted an ablation study varying the size of this region, denoted by a divisor d. The VP region is defined as $H_f/d \times W_f/d$, where H_f and W_f are the height and width of the high-resolution feature map from which regional features are extracted. To ensure a fair comparison in terms of model parameters and feature density when varying d, the cropped VP region is resized to a consistent spatial dimension before features are processed by the subsequent low-stride CNN. We also include a setting where $d = 0$ (VPA disabled), which effectively reverts to our baseline ” Instance Query Only” model, and $d = 1$, which uses the full image feature map as the ” VP region.”

The mIoU results on both SemanticKITTI and SSCBench-KITTI-360 for different

doi:10.6342/NTU202503382

values of d are presented in Table 4.4. As observed, setting $d = 3$ (i.e., a VP region

VP Region Denominator (d)	SemanticKITTI	KITTI-360
$d = 0$ (VPA Disabled)	14.89	18.58
$d = 1$ (Full Image)	15.21	18.76
$d = 3$ (Proposed)	**15.26**	**18.80**
$d = 5$	15.09	18.68
$d = 7$	14.83	18.61

Table 4.4: Ablation study on the size of the VP region, defined by $H_f/d \times W_f/d$. Performance is measured by mIoU (%). $d = 0$ disables VPA, and $d = 1$ uses the full image.

of $H_f/3 \times W_f/3$) yields the best performance on both datasets, achieving an mIoU of **15.26%** on SemanticKITTI and **18.80%** on KITTI-360. This configuration surpasses the performance when VPA is disabled ($d = 0$), which scores 14.89% and 18.58% respectively, highlighting the overall benefit of incorporating VP-guided features.

When the VP region is made larger (e.g., $d = 1$, using the full image), performance slightly decreases to 15.21% on SemanticKITTI and 18.76% on KITTI-360 compared to $d = 3$. This suggests that while a broader view captures more context, it may also introduce an excessive amount of less relevant or potentially noisy features from peripheral areas, diluting the focused information from the true far-field region. Conversely, making the VP region progressively smaller also leads to a decline in performance. This indicates that an overly constrained region may not capture sufficient contextual information from the surroundings of the vanishing point, thereby limiting the VPQ's ability to effectively enhance long-range predictions.

Therefore, $d = 3$ appears to strike the most effective balance, providing a sufficiently focused yet adequately contextualized region for extracting potent far-field features. This empirically validates our choice of $H_f/3 \times W_f/3$ as the default VP region size for the VPA module in our main experiments.

doi:10.6342/NTU202503382

Chapter 5 Conclusion

5.1 Conclusion

In this paper, we introduced a novel Vanishing Point Aggregator (VPA) framework for 3D semantic scene completion, designed to address critical perception challenges in autonomous driving. By synergistically integrating specialized Vanishing Point Queries with conventional instance queries, our method effectively prioritizes and aggregates crucial spatial and semantic information from the information-dense region surrounding the vanishing point. This targeted approach significantly enhances the model's ability to perceive small, distant, and traditionally hard-to-detect objects, which are often overlooked by methods that rely on more uniform feature aggregation.

Our extensive experiments on the challenging SemanticKITTI and SSCBench-KITTI-360 benchmarks demonstrate that VPA achieves SOTA performance. Notably, our approach exhibits superior accuracy in long-distance predictions and for small object categories, highlighting its practical value for safety-critical autonomous driving applications where robust far-field understanding is paramount. This improvement can directly translate to more reliable early detection of distant pedestrians, cyclists, and obstacles—allowing vehicles to initiate braking or evasive maneuvers sooner, thereby reducing collision risk and enhancing overall roadway safety.

doi:10.6342/NTU202503382

5.2 Future Work

Building upon the successes of VPA, a promising direction for future research involves extending this VP-guided paradigm to **incorporate temporal consistency and motion cues**, **real-time efficiency**, and **zero-shot generalization**. First, by propagating vanishing-point queries across consecutive frames via a lightweight spatiotemporal transformer or recurrent module, the model can enforce smooth far-field predictions, adapt to dynamic scene changes, and virtually eliminate flicker. Second, to enable deployment on resource-constrained automotive hardware, we will explore adaptive query pruning driven by scene complexity, mixed-precision quantization of both backbone and query-fusion layers, and knowledge distillation into compact student networks that sustain high small-object IoU at >30 FPS on embedded SoCs. Finally, avenue toward zero-shot generalization in unseen real-world settings is to explore a self-supervised domain adaptation scheme. For example, one could apply contrastive consistency objectives to align feature distributions between synthetic and unlabeled real video streams, and introduce controlled perturbations of vanishing-point cues to diversify geometric guidance. By using these techniques to train the VPA block and fusion modules without manual labels, the framework may better adapt its semantic and geometric completion capabilities to entirely new environments.

doi:10.6342/NTU202503382

References

[1] S. Ardianto, H.-M. Hang, and W.-H. Cheng. Fast vehicle detection and tracking on fisheye traffic monitoring video using cnn and bounding box propagation. In ICIP, 2022.

[2] F. Barbato, E. Camuffo, S. Milani, and P. Zanuttigh. Continual road-scene semantic segmentation via feature-aligned symmetric multi-modal network. In ICIP, 2024.

[3] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In ICCV, 2019.

[4] A.-Q. Cao and R. De Charette. Monoscene: Monocular 3d semantic scene completion. In CVPR, 2022.

[5] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV), pages 801–818, 2018.

[6] T. Chen, X. Ying, J. Yang, R. Wang, R. Guo, B. Xing, and J. Shi. Vpdetr: End-to-end vanishing point detection transformers. In AAAI, 2024.

[7] A. Das, S. Das, G. Sistu, J. Horgan, U. Bhattacharya, E. Jones, M. Glavin, and C. Eis-

doi:10.6342/NTU202503382

ing. Revisiting modality imbalance in multimodal pedestrian detection. In ICIP, 2023.

[8] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In CVPR, 2012.

[9] D. Guo, D.-P. Fan, T. Lu, C. Sakaridis, and L. Van Gool. Vanishing-point-guided video semantic segmentation of driving scenes. In CVPR, 2024.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.

[11] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In CVPR, 2023.

[12] H. Jiang, T. Cheng, N. Gao, H. Zhang, T. Lin, W. Liu, and X. Wang. Symphonize 3d semantic scene completion with contextual instance queries. In CVPR, 2024.

[13] S. Lee, J. Kim, J. Shin Yoon, S. Shin, O. Bailo, N. Kim, T.-H. Lee, H. Seok Hong, S.-H. Han, and I. So Kweon. Vpgnet: Vanishing point guided network for lane and road marking detection and recognition. In ICCV, 2017.

[14] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In CVPR, 2023.

[15] Y. Li, S. Li, X. Liu, M. Gong, K. Li, N. Chen, Z. Wang, Z. Li, T. Jiang, F. Yu, et al. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. In IROS, 2024.

doi:10.6342/NTU202503382

[16] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In CVPR, 2023.

[17] Y. Liao, J. Xie, and A. Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3):3292–3310, 2022.

[18] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

[19] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. In CVPR, 2017.

[20] Y. Wang and C. Tong. H2gformer: Horizontal-to-global voxel transformer for 3d semantic scene completion. In AAAI, 2024.

[21] Z. Xia, Y. Liu, X. Li, X. Zhu, Y. Ma, Y. Li, Y. Hou, and Y. Qiao. Scpnet: Semantic scene completion on point cloud. In CVPR, 2023.

[22] J. Yao, C. Li, K. Sun, Y. Cai, H. Li, W. Ouyang, and H. Li. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In ICCV, 2023.

[23] Y. Zhang, Z. Zhu, and D. Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In ICCV, 2023.

doi:10.6342/NTU202503382