Evaluating Deep Learning RGB-Based Panoptic Segmentation Models on LiDAR-Generated Images

University of Turku
Department of Computing
Master of Science in Technology Thesis
Robotics and Autonomous Systems
November 2025
Sileshi Ziena Adal

Supervisors:
Dr.Xianjia Yu
Prof. Tomi Westerlund

UNIVERSITY OF TURKU
Department of Computing

Sileshi Ziena Adal: Evaluating Deep Learning RGB-Based Panoptic Segmentation Models on LiDAR-Generated Images

Master of Science in Technology Thesis, 66 p.
Robotics and Autonomous Systems
November 2025

Panoptic segmentation, which combines semantic and instance segmentation, plays a vital role in scene understanding for applications such as autonomous driving, robotics, and urban mapping. While state-of-the-art deep learning models have achieved strong performance on RGB datasets, their generalizability to LiDAR-generated imagery remains underexplored.

This thesis investigates how existing RGB-trained panoptic segmentation models perform on LiDAR derived pseudo-RGB images. It begins with a structured review of leading architectures, training strategies, and benchmark results on RGB datasets. The selected models are then evaluated on LiDAR-generated data using metrics such as Panoptic Quality (PQ), Segmentation Quality (SQ), Recognition Quality (RQ), Intersection over Union (IoU), and inference efficiency, complemented by qualitative visualizations of the output masks. A pseudo-RGB LiDAR dataset was used to simulate cross modal testing conditions and to assess model robustness when applied to LiDAR data, which differs significantly from the RGB domain they were trained on.

The results reveal that RGB trained panoptic segmentation models face notable performance degradation when applied to LiDAR generated imagery, primarily due to this domain difference and the lack of sensor specific adaptation. Differences in instance recognition, boundary accuracy, and category consistency were observed across models, as reflected in PQ, SQ, RQ, and IoU scores, as well as through qualitative outputs. These findings offer a foundational reference for future research and aim to contribute to the development of more versatile and effective deep learning models for panoptic segmentation across diverse data types.

Keywords: panoptic segmentation, lidar images, pseudo-RGB, deep learning, RGB-trained models, cross-domain evaluation, PQ, mIoU

Acknowledgements

I would like to express my deepest gratitude to all the individuals who supported me throughout the course of this thesis.

First and foremost, my heartfelt thanks go to my thesis supervisor, **Dr. Xianjia Yu**, Postdoctoral Researcher in Robotics and Autonomous Systems, for his expert guidance and support. I am also sincerely grateful to **Professor Tomi Westerlund**, whose valuable advice and encouragement provided essential guidance during this journey. My sincere appreciation extends to **Maria Prusila**, Study Advisor in Educational Affairs at the University of Turku, for her continuous support and insightful guidance throughout my academic progress.

Above all, I express my deepest gratitude to my family, especially my beloved wife, **Mrs. Kidiste Alene** for her boundless love and steadfast encouragement. I am also profoundly thankful to my dear friends, especially **Dagi**, **Kunu , and Mimi** in Turku, Finland, whose companionship and support have been a constant source of strength. This thesis would not have been possible without all of you. Thank you!

Contents

List of Figures

List of Tables

List of acronyms

ADE20k ADE20K Dataset (A Diverse and Extensive Dataset for Scene Parsing)

CNN Convolutional Neural Network

COCO Common Objects in Context (Dataset)

Colab Google Colaboratory

CPU Central Processing Unit

DL Deep Learning

FCN Fully Convolutional Network

FPN Feature Pyramid Network

FPS Frames Per Second

GPU Graphics Processing Unit

GT Ground Truth

LiDAR Light Detection and Ranging

MIoU Mean Intersection over Union

PQ Panoptic Quality

R-CNN Regions with CNN features

ResNet Residual Network

RGB Red Green Blue

RQ Recognition Quality

Seg Segmentation

SQ Segmentation Quality

UPSNet Unified Panoptic Segmentation Network

VOC Visual Object Classes (Dataset)

YOLOv5 You Only Look Once Version 5

YOLO You Only Look Once

1 Introduction

1.1 Background on Image Segmentation

In the evolving domains of computer vision and autonomous systems, scene understanding remains a fundamental requirement. It enables machines to interpret complex environments by identifying, localizing, and differentiating between diverse objects and surfaces within an image. This capacity underpins numerous applications, including autonomous driving, service robotics, augmented reality (AR), and smart infrastructure systems [1], [2].

Image segmentation, the process of partitioning an image into semantically meaningful regions, plays a critical role in achieving scene understanding. Over time, segmentation has evolved into three primary paradigms: *semantic segmentation*, *instance segmentation*, and *panoptic segmentation*. Each provides a different level of granularity and object differentiation.

The proliferation of deep learning has accelerated progress in segmentation research. Notable architectures such as Panoptic FPN [3], Mask R-CNN [4], YOLOv5-Seg [5], and Mask2Former [2] have demonstrated strong performance on RGB datasets like COCO [6] and Cityscapes [7], where the visual data is rich in texture, structure, and color gradients.

However, in environments where spatial geometry or depth is more critical than color cues, such as in robotics or adverse lighting conditions—RGB imagery can be insufficient. In such contexts, LiDAR (Light Detection and Ranging) provides an effective complementary modality. LiDAR sensors produce dense 3D point clouds by emitting laser pulses and measuring their return times, enabling robust depth and spatial topology measurements [8], [9].

To align with 2D vision frameworks, these LiDAR point clouds are often projected into 2D images known as *pseudo-RGB LiDAR images*. While visually similar to natural RGB images, their underlying content is structurally distinct, typically encoding information such as reflectivity, height, and range in place of natural color channels.

This structural disparity introduces a modality gap between training data (RGB) and evaluation data (LiDAR), raising the critical research question: *To what extent can panoptic segmentation models trained exclusively on RGB datasets generalize to LiDAR-generated pseudo-RGB images without adaptation?*

This thesis addresses this question by evaluating several state-of-the-art RGB-trained panoptic segmentation models on LiDAR-derived inputs. The evaluation focuses on generalization performance under modality shift, measuring segmentation accuracy and efficiency using metrics such as Panoptic Quality (PQ), Intersection-over-Union (IoU), and runtime [10], [11].

Through both methodological implementation and empirical benchmarking, this study contributes insights for researchers and practitioners in computer vision and robotics, promoting the development of robust, generalizable perception models for cross-domain deployment.

1.2 Overview of Segmentation Approaches

Image segmentation methods are typically classified into three complementary paradigms—sema
instance, and panoptic segmentation, each providing a different level of granularity for scene understanding [1], [9].

1.2.1 Semantic Segmentation

Semantic segmentation assigns a class label to every pixel, grouping regions by category (e.g., road, vegetation, sky) without distinguishing individual object instances. Modern approaches leverage deep architectures to capture both fine details and global context:

- **DeepLabv3+** uses an encoder–decoder with Atrous Spatial Pyramid Pooling for multi-scale context aggregation [12].
- **SegNeXt** rethinks convolutional attention modules to improve efficiency and accuracy on large-scale datasets [10].

1.2.2 Instance Segmentation

Instance segmentation extends semantic segmentation by detecting and delineating each object instance separately. Key models include:

- **Mask R-CNN**, which adds a mask prediction branch to Faster R-CNN, achieving strong instance-level accuracy [11].
- **YOLACT**, a one-stage, real-time framework that generates prototype masks and per-instance coefficients [13].

1.2.3 Panoptic Segmentation

Panoptic segmentation unifies the semantic and instance tasks into a single framework, assigning every pixel both a semantic label and, where applicable, an instance ID [3]. Recent transformer-based extensions further enhance global reasoning:

- **Mask2Former**: Introduces masked attention and multi-scale deformable queries for unified panoptic prediction [14].
- **UniDAformer**: A domain-adaptive transformer that calibrates mask predictions hierarchically for robust cross-domain segmentation [2].

Together, these paradigms form the foundation for comprehensive scene understanding and set the stage for cross-modal evaluation on LiDAR-derived pseudo-RGB inputs in Chapter 3.

1.3 Comparison of Semantic, Instance, and Panoptic Segmentation

While semantic, instance, and panoptic segmentation share a common goal of partitioning images into meaningful regions, they differ in granularity, computational demands, and application focus.

- **Semantic Segmentation** assigns each pixel a class label but does not differentiate between multiple objects of the same class. It excels in tasks requiring broad scene understanding but falls short when individual object localization or counting is needed [8], [14].

- **Instance Segmentation** extends semantic segmentation by detecting and segmenting each object instance separately. Models such as Mask R-CNN [2] and YOLACT [15] enable precise object delineation but incur higher inference time and memory overhead.

- **Panoptic Segmentation** unifies the two tasks, providing per-pixel semantic labels for "stuff" categories (e.g., sky, road) and distinct instance IDs for "things" (e.g., vehicles, pedestrians) [3]. This comprehensive framework supports holistic scene interpretation, though it demands sophisticated fusion of semantic and instance predictions and greater computational resources.

Table 1.1 summarizes their key characteristics:

Table 1.1: Comparison of Segmentation Paradigms

Feature	**Semantic**	**Instance**	**Panoptic**
Differentiates Instances?	No	Yes	Yes
Per-Pixel, Class-Labels?	Yes	Yes	Yes
Unique Instance IDs?	No	Yes	Yes
Computational Cost	Low	Medium	High
Typical Use Cases	Scene-parsing, land-use mapping	Object-detection, tracking	Autonomous driving, robotics

1.4 LiDAR in Robotics and Autonomous Systems

LiDAR (Light Detection and R ranging) sensors emit laser pulses and measure return times to generate precise 3D point clouds, capturing environmental geometry with centimeter-level accuracy [14]. This capability complements RGB-based perception by providing reliable depth information, especially under challenging lighting or weather conditions.

1.4.1 Key Applications

- **Autonomous Vehicles:** Real-time 3D mapping, obstacle detection, and localization in dynamic driving environments [16].
- **Mobile Robotics:** Navigation and manipulation tasks in unstructured settings, where depth cues guide path planning and object interaction [17].
- **Aerial Surveying:** Terrain reconstruction and vegetation analysis via drone-mounted LiDAR, supporting applications in agriculture and disaster response [18].
- **Smart Infrastructure:** Urban modeling and infrastructure inspection, enabling digital twins and proactive maintenance [18].

1.4.2 LiDAR Data Representations

LiDAR point clouds can be transformed into various 2D formats to leverage existing convolutional architectures. The most common representations are:

- **Depth (Range) Maps:** Single-channel images encoding the distance from sensor to each point, often normalized to $[0, 1]$ or scaled in meters. Depth maps preserve spatial structure but lack reflectivity information [14].

- **Reflectivity/Intensity Images:** Capture the returned signal strength of each laser pulse, which correlates with surface material and angle of incidence. Useful for distinguishing object surfaces (e.g., metal vs. vegetation) [19].
- **Elevation/Height Maps:** Encode the vertical coordinate (z-axis) of each point, often relative to sensor or ground plane. Height maps facilitate separation of ground "stuff" versus above-ground "things" [20].
- **Pseudo-RGB Encodings:** Composite three-channel images commonly {height, intensity, range} as R, G, B to mimic natural images and permit direct use of RGB-trained networks [10].

1.4.3 Challenges for Panoptic Segmentation on LiDAR Data

Projecting 3D LiDAR into 2D images introduces specific obstacles for panoptic segmentation:

- **Data Sparsity and Irregularity:** Unlike dense RGB grids, LiDAR sampling density decreases with distance, leading to holes and uneven coverage that hinder mask continuity [18].
- **Loss of Geometric Context:** Flattening 3D structure into 2D can obscure occlusions, depth layering, and object shape cues, increasing boundary ambiguity between adjacent instances [20].
- **Modality Mismatch:** The absence of color and texture cues makes direct transfer of RGB-trained filters suboptimal; learned convolution kernels may misinterpret reflectivity or height patterns as "noise" [19].

- **Sensor Noise and Environmental Artifacts:** Weather effects (rain, fog, dust) and reflective surfaces introduce spurious returns and measurement errors, leading to false positives/negatives in segmentation masks [21].

- **Real-Time and Resource Constraints:** Processing high-resolution LiDAR images at frame rates required for autonomous navigation demands efficient architectures or downsampling strategies that trade accuracy for speed [22].

Understanding these aspects is crucial for developing and evaluating panoptic segmentation models under cross-modal conditions, as detailed in the subsequent chapters.

2 Comprehensive Review of Panoptic Segmentation Approaches for RGB Images

2.1 Introduction to Panoptic Segmentation

Panoptic segmentation is a unified computer vision task in which every pixel is assigned both a semantic label and, when appropriate, an instance identifier. This formulation integrates the strengths of semantic segmentation [23], [24] and instance segmentation [15], [25] into a single coherent representation of a scene. Such unified predictions are essential for applications in autonomous driving, robotics, mapping, and related perception tasks [14], [26], [27].

Historically, semantic segmentation approaches such as Fully Convolutional Networks (FCN) and DeepLab-based methods were effective in pixel-wise classification but unable to distinguish multiple objects of the same class [23], [24], [28]. Instance segmentation methods such as Mask R-CNN and YOLACT excelled at delineating individual objects but ignored background "stuff" categories [15], [25]. Subsequent work in panoptic segmentation has consolidated and systematised these developments, providing unified formulations and taxonomies of methods [4], [18], [29].

Research in panoptic segmentation has converged around four major architectural paradigms:

1. Dual-branch architectures,
2. Unified architectures,
3. Fully convolutional and lightweight architectures,
4. Transformer-based architectures.

This chapter reviews these paradigms, highlights foundational tasks, compares benchmark performance on RGB datasets, and concludes with the rationale for selecting five representative models for cross-domain evaluation in Chapter 3.

2.2 Foundations of Image Segmentation

Panoptic segmentation builds upon the complementary strengths of semantic and instance segmentation.

2.2.1 Semantic Segmentation

Semantic segmentation assigns a semantic class label to each pixel. Early work such as Fully Convolutional Networks (FCN) [23] introduced dense prediction using convolutional architectures. DeepLab-style models [24], [28] enhanced multi-scale context extraction using atrous convolutions and encoder–decoder designs, while SegNeXt [11] and related approaches rethink convolutional attention for high-resolution segmentation.

Despite significant advancements, semantic segmentation cannot differentiate multiple instances of the same class and may struggle with occluded or fine-structured regions.

2.2.2 Instance Segmentation

Instance segmentation extends semantic segmentation by predicting distinct masks for each object. Two-stage methods such as Mask R-CNN [1], [25] achieve high-quality masks using region proposals, whereas one-stage architectures like YOLACT [15] and SOLO [30] improve speed by generating prototype masks or location-based predictions.

Instance segmentation effectively separates objects but does not classify background regions and may degrade in cluttered scenes. These limitations motivated the development of unified panoptic segmentation methods.

2.3 Architectural Paradigms in Panoptic Segmentation

Panoptic segmentation architectures can be grouped into four main categories based on how they produce and fuse semantic and instance predictions.

2.3.1 Dual-Branch Architectures

Dual-branch architectures share a backbone but maintain distinct semantic and instance heads. The outputs are combined via fusion heuristics or priority rules.

Representative model: Panoptic FPN [31] uses a Feature Pyramid Network (FPN) backbone with parallel Mask R-CNN and semantic segmentation heads. A fusion module integrates their outputs to form the final panoptic map. The instance branch predicts bounding boxes, class scores, and instance masks, while the semantic branch predicts dense per-pixel class labels.

Advantages of this paradigm include modularity and reuse of established detection pipelines. However, limitations include redundant computation across branches and the possibility of inconsistent predictions in dense scenes, especially when heuristic fusion fails in overlapping regions.

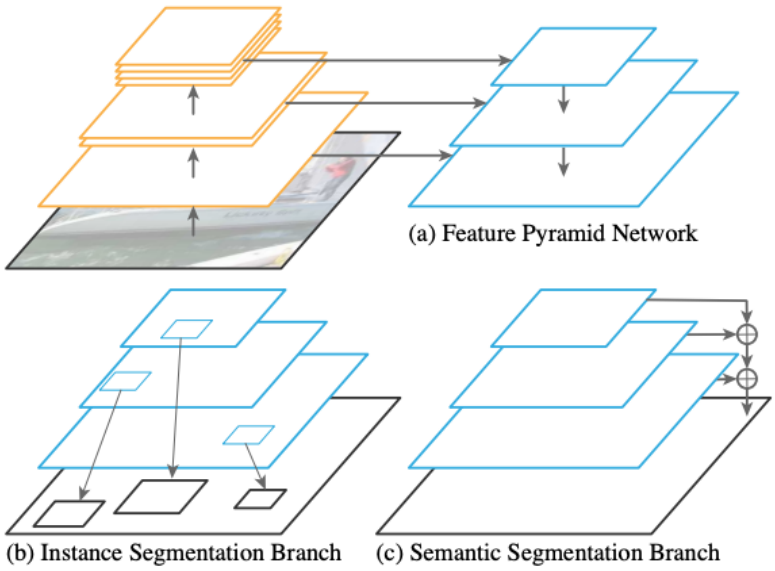

Figure 2.1: Dual-branch panoptic segmentation architecture illustrated by Panoptic FPN. A shared FPN backbone feeds separate instance (Mask R-CNN) and semantic heads, whose outputs are fused into a panoptic map [31].

2.3.2 Unified Architectures

Unified architectures merge semantic and instance predictions within a single joint network, often using a learnable fusion head rather than heuristic rules.

Representative model: UPSNet [32] introduces a panoptic head that jointly processes semantic segmentation logits and instance mask logits. Instead of fixed priority rules, the model learns how to combine these outputs, reducing inconsistencies and improving overall panoptic quality. The backbone is shared, and the panoptic head reasons jointly about stuff and thing classes.

Advantages include reduced redundancy, better coherence between semantic and instance predictions, and end-to-end optimisation of panoptic objectives. Limitations include more complex training dynamics due to multi-task loss balancing and, in some cases, slightly reduced instance boundary precision compared with highly tuned detection-based pipelines.

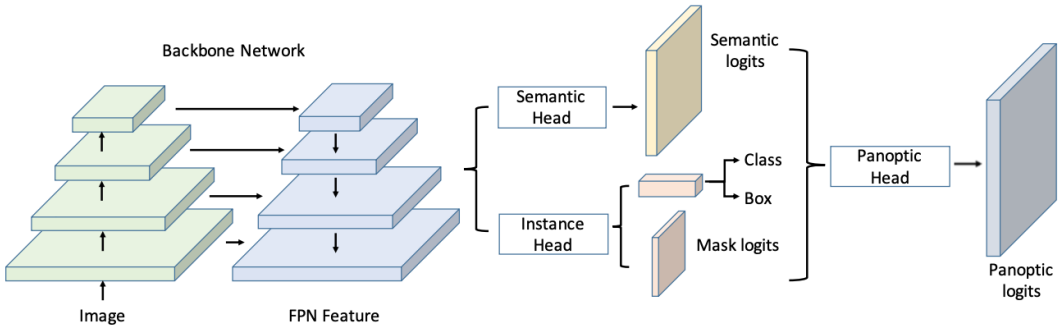

Figure 2.2: Unified panoptic segmentation architecture illustrated by UPSNet. A shared backbone with semantic and instance branches feeds a learnable panoptic head that fuses predictions into a consistent panoptic output [32].

2.3.3 Fully Convolutional and Lightweight Architectures

Fully convolutional and lightweight architectures prioritise efficiency and real-time performance. They typically avoid region proposals and heavy transformer modules, relying instead on dense prediction and bottom-up grouping.

Representative model: EfficientPS employs an EfficientNet-based backbone together with a bi-directional feature pyramid fusion module to produce shared multi-scale features for semantic and instance heads. A parameter-free panoptic fusion module combines these outputs into the final panoptic prediction. The architecture is optimised for high throughput while maintaining competitive accuracy.

This paradigm offers fast inference and low computational cost, making it attractive for embedded systems and robotics. However, compared to transformer-based architectures, fully convolutional methods may have reduced capacity for global context modelling and can struggle in highly cluttered or long-range dependency scenarios.

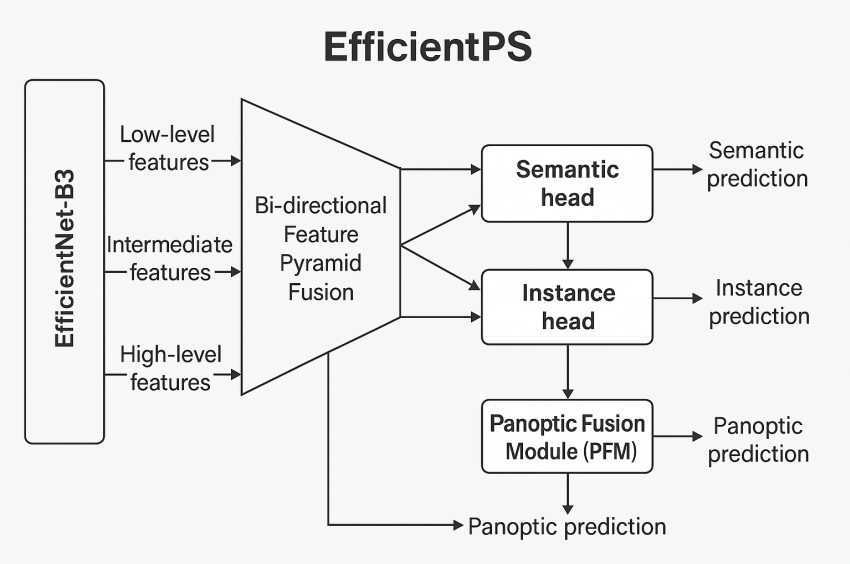

Figure 2.3: Fully convolutional and lightweight panoptic segmentation architecture illustrated by EfficientPS. An EfficientNet backbone and bi-directional feature fusion feed semantic and instance heads, whose outputs are merged by a panoptic fusion module.

2.3.4 Transformer-Based Architectures

Transformer-based architectures incorporate global self-attention, allowing them to capture long-range relationships across the image. Many recent models reformulate segmentation as a mask-classification problem using learned queries.

Representative model: Mask2Former [33] builds on the Masked-Attention Mask Transformer family [34] and uses a backbone and pixel decoder to produce multi-scale feature maps, followed by a transformer decoder with masked attention and learned mask queries. These queries interact with the feature maps to produce a

set of masks and corresponding class labels, enabling a unified approach to semantic, instance, and panoptic segmentation.

The main advantages of transformer-based architectures include strong global reasoning capabilities and state-of-the-art accuracy in PQ and mIoU. Their limitations are primarily related to computational and memory cost, which can restrict deployment in real-time or resource-constrained environments and may exacerbate sensitivity to domain shifts.

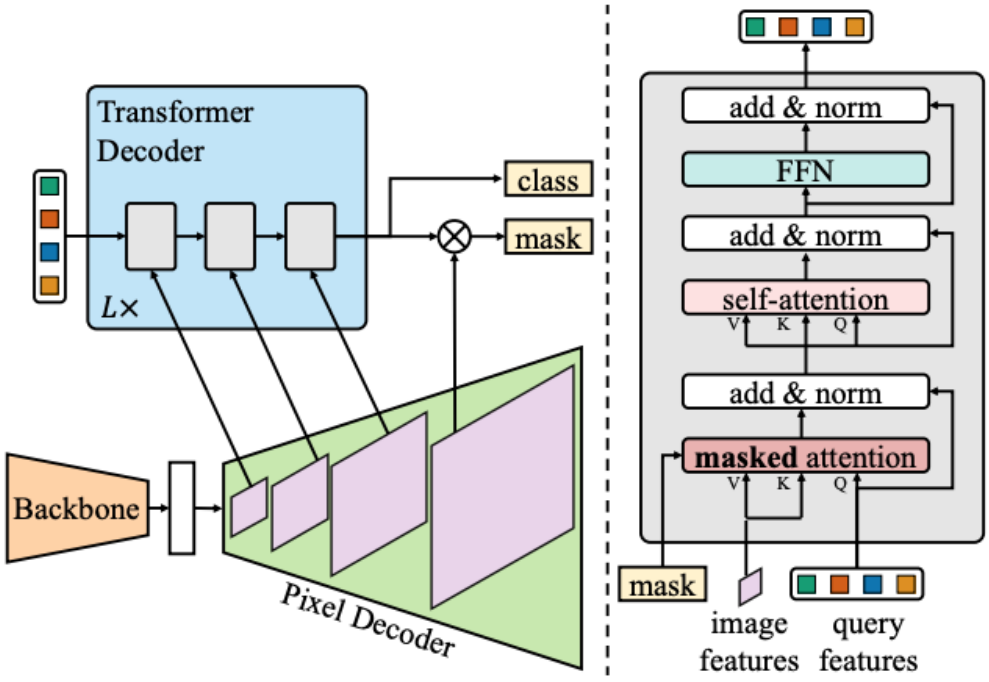

Figure 2.4: Transformer-based panoptic segmentation architecture illustrated by Mask2Former. Multi-scale features from the backbone and pixel decoder are processed by a transformer decoder with masked attention and mask queries to produce a set of predicted masks and class labels [33].

2.4 Performance Benchmarks on RGB Datasets

Panoptic segmentation models are evaluated on datasets such as COCO and Cityscapes [6], [35] using metrics including panoptic quality (PQ), segmentation quality (SQ),

recognition quality (RQ), semantic mean intersection-over-union (mIoU), and inference speed measured in frames per second (FPS).

Table 2.1 summarises representative benchmark results reported on COCO val2017. While exact figures vary across implementations, the values illustrate typical performance trends across the four architectural paradigms.

Model	**Backbone**	**PQ (%)**	**mIoU (%)**	**FPS**
Panoptic FPN	ResNet-50	42.5	61.2	12
UPSNet	ResNet-50	43.2	62.0	9
Panoptic-DeepLab	Xception	44.0	63.0	7
DeepLabv3+ Panoptic Head	Xception	41.8	60.5	17
EfficientPS	EfficientNet-B3	45.1	64.3	20
YOLACT++ Panoptic	ResNet-101	38.5	58.0	35
AdaptIS	ResNet-50	40.2	59.5	10
Panoptic FCN	ResNet-50	43.0	61.0	15
MaskFormer	ResNet-50	46.0	65.0	5
Mask2Former	Swin-Base	47.6	66.0	5
Segmenter	ViT-Base	45.5	64.5	4

Table 2.1: Benchmark performance of representative panoptic segmentation models on COCO val2017, illustrating the trade-offs between different architectural paradigms.

Transformer-based models such as **MaskFormer** and **Mask2Former** [33], [34] achieve the highest accuracy, while efficient designs such as **EfficientPS** and **YOLACT-based** approaches [15] offer attractive speed accuracy trade-offs. Unified models like UPSNet [32] outperform dual-branch baselines in consistency, whereas Panoptic FPN [31] remains a strong traditional baseline.

2.5 Rationale for Model Selection

The goal of this thesis is to evaluate how well state-of-the-art RGB-trained panoptic segmentation models generalise to LiDAR-derived pseudo-RGB images. To support a balanced and meaningful analysis, model selection was guided by several criteria.

Architectural diversity. The selected models represent every major architectural paradigm, including dual-branch convolutional networks, unified multi-task designs, fully convolutional architectures, and transformer-based approaches. This diversity enables comparison of how design choices affect cross-domain generalisation [4], [18], [29].

Availability of pretrained weights. All selected models provide publicly available pretrained weights on COCO or Cityscapes [6], [35], enabling inference-only evaluation without retraining.

Relevance and adoption. These models represent widely adopted, foundational, or state-of-the-art contributions to panoptic segmentation and have been extensively evaluated in the literature [31]–[34], [36]–[38].

Inference efficiency and practicality. The models span a range of computational complexities and runtimes, enabling exploration of trade-offs between accuracy and efficiency that are relevant for real-time and resource-constrained deployment scenarios [5], [39].

Suitability for cross-domain evaluation. Differences in backbone design, fusion strategies, and mask prediction mechanisms make these models well suited for evaluating generalisation to LiDAR-derived inputs, where visual characteristics differ significantly from standard RGB imagery [26], [27], [40].

Based on these criteria, five representative models were selected for the experimental evaluation presented in Chapter 3. These models were chosen because they span the major architectural paradigms in panoptic segmentation, include

both classical convolutional and modern transformer-based approaches, and provide pretrained RGB weights necessary for inference-only evaluation on LiDAR-derived pseudo-RGB imagery. The selected models are:

1. **Detectron2 Panoptic FPN** [31] : a dual-branch architecture combining a Mask R-CNN instance segmentation head with an FCN-based semantic head.

2. **YOLOv5-Seg + Fusion** [41], [42] : a real-time instance segmentation model extended in this thesis with a custom panoptic fusion pipeline.

3. **Mask2Former** [33] : a modern transformer-based architecture that formulates segmentation as a mask-classification problem using masked attention and learned mask queries.

4. **UPSNet** [32] : a unified architecture featuring a learnable panoptic head that jointly fuses semantic and instance predictions.

5. **DeepLabv3+ (Panoptic Head)** [24] : a fully convolutional model that extends the DeepLabv3+ semantic backbone with a panoptic head for instance association.

Together, these five models provide a comprehensive and representative foundation for studying cross-domain generalisation in panoptic segmentation. Their architectural diversity spanning dual-branch, unified fusion, bottom-up convolutional decoding, transformer-based reasoning, and customised lightweight detection ensures that the evaluation in Chapter 3 captures a broad spectrum of design philosophies and reveals how each paradigm behaves when applied to LiDAR-generated pseudo-RGB images without retraining.

3 Evaluating Rgb-Trained Panoptic segmentation Models on Lidar Data

This chapter presents the experimental evaluation conducted as part of this thesis to assess the generalization ability of state-of-the-art panoptic segmentation models trained on RGB images when applied to LiDAR-derived pseudo-RGB data. The focus is on inference only testing to observe how these models respond to domain shift, without any retraining or fine-tuning.

The evaluation was performed using a publicly available LiDAR visualization image [43], processed and adapted into a pseudo-RGB format for compatibility with RGB-trained models. As the original image lacked ground truth annotations, I generated simulated panoptic segmentation masks to enable both quantitative (e.g., PQ, SQ, RQ, mIoU) and qualitative evaluation.

All code used in this study including data preparation, model inference pipelines, and evaluation scripts was developed by the author and is publicly available at the GitHub Repository as Panoptic-Segmentation-Eval-lidar-rgb .

This repository supports reproducibility and provides a foundation for future research into cross-modal panoptic segmentation tasks.

3.1 Dataset Description

The dataset used in this study consists of a single pseudo-RGB image generated from a raw LiDAR point cloud, sourced from a publicly available example provided by Car Magazine [43]. The image serves as a visualization of LiDAR point cloud data projected into a 2D view, depicting an urban environment with multiple object categories such as vehicles, pedestrians, and buildings.

Since the image was not part of an official panoptic segmentation dataset and did not include ground truth annotations, a simulated evaluation framework was implemented. This approach allows for testing RGB-trained panoptic segmentation models on LiDAR-derived inputs without requiring access to labeled LiDAR datasets..

Image Structure

The image represents a 3D LiDAR point cloud that has been projected onto a 2D surface using spherical projection techniques. Although originally generated for visualization, it closely resembles the spatial structure seen in typical LiDAR-based autonomous driving datasets. Key spatial features such as object boundaries, relative depth, and density are visible, enabling meaningful segmentation analysis.

Pseudo-RGB Conversion

As the original data lacked channel-specific encoding (e.g., height, intensity, range), the available 2D projection was treated as a three-channel pseudo-RGB image by replicating and normalizing its visual appearance for compatibility with RGB-trained models.

Ground Truth Simulation

Due to the absence of official semantic or instance-level annotations, synthetic ground truth masks were generated for evaluation:

Semantic labels were assigned based on estimated object types in the scene.

Instance labels were approximated using connected component analysis and noise

injection.

These masks were used to compute segmentation metrics while acknowledging that results are influenced by the simulation's artificial nature.

Limitations:

- Only one image was used for evaluation, limiting the statistical generalizability of results.
- All ground truth masks were synthetically generated, meaning metrics such as PQ, SQ, and mIoU may be optimistic or biased.
- This setup, while constrained, reflects realistic conditions for evaluating model behavior in the absence of annotated cross-modal datasets.

3.2 Evaluation Metrics

To evaluate the performance of RGB-trained panoptic segmentation models on LiDAR-derived pseudo-RGB images, a set of standard metrics was employed. These metrics were selected to capture both pixel-level segmentation accuracy and instance-level recognition performance. Additionally, inference speed was considered to assess real-time applicability, particularly in resource-constrained environments such as robotics or embedded systems.

1. Panoptic Quality (PQ), PQ measures the overall segmentation performance by combining both the segmentation quality and recognition accuracy of object instances. It is defined as [31]:

$$\mathrm{PQ} = \frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\mathrm{TP}| + \frac{1}{2}|\mathrm{FP}| + \frac{1}{2}|\mathrm{FN}|} \tag{3.1}$$

$$\text{PQ} = \underbrace{\frac{\sum_{(p,g)\in \text{TP}} \text{IoU}(p,g)}{|\text{TP}|}}_{\text{Segmentation Quality (SQ)}} \times \underbrace{\frac{|\text{TP}|}{|\text{TP}| + \frac{1}{2}|\text{FP}| + \frac{1}{2}|\text{FN}|}}_{\text{Recognition Quality (RQ)}} \tag{3.2}$$

Where:

- TP: True positives (matched segments)
- FP: False positives (extra predictions)
- FN: False negatives (missed segments)
- IoU(p, g): Intersection over Union for prediction-ground truth pair

PQ balances mask quality and object recognition, making it suitable for holistic evaluation.

2. Segmentation Quality (SQ), SQ isolates the quality of segmentation masks from recognition performance. It is computed as the average IoU of all matched segments [31]:

$$\text{SQ} = \frac{1}{|\text{TP}|} \sum_{(p,g)\in \text{TP}} \text{IoU}(p,g) \tag{3.3}$$

A higher SQ indicates more precise alignment between predicted and ground truth segment shapes.

3. Recognition Quality (RQ), RQ evaluates the model's ability to detect and correctly classify individual object instances. It is defined as[31]:

$$\text{RQ} = \frac{|\text{TP}|}{|\text{TP}| + 0.5|\text{FP}| + 0.5|\text{FN}|} \tag{3.4}$$

This metric reflects the effectiveness of object recognition regardless of the accuracy of the mask shape.

4. Mean Intersection over Union (mIoU), mIoU is a commonly used semantic segmentation metric. It averages the IoU for each class and is defined as [31]:

$$\text{mIoU} = \frac{1}{K} \sum_{i=1}^{K} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i} \tag{3.5}$$

where K is the number of semantic classes. mIoU emphasizes pixel-level classification accuracy.

Inference Time (ms/frame), Inference time measures the average time required to process a single image frame. It serves as a proxy for model efficiency and deployability in real-time systems. Lower inference time is critical for latency-sensitive applications such as autonomous driving or robotic perception.

3.3 Model Descriptions and Selection Criteria

This section presents detailed descriptions of the five panoptic segmentation models evaluated in this study: Detectron2 Panoptic FPN, YOLOv5-Seg with Fusion, Mask2Former, UPSNet, and DeepLabv3+ with Panoptic Head. These models were selected based on the structured criteria outlined in Chapter **??**, with an emphasis on architectural diversity, practical usability, and compatibility with LiDAR-derived pseudo-RGB input formats.

3.3.1 Model Selection Criteria

The specific criteria applied in this study include:

1. **Architectural Diversity:** The models represent a broad spectrum of segmentation strategies, including:
    - Transformer-based architectures (Mask2Former [44]),
    - CNN-based semantic segmentation extended for panoptic tasks (DeepLabv3+ [44]),

- Hybrid encoder-decoder models combining semantic and instance heads (Detectron2 Panoptic FPN [3]),
- Bottom-up fusion networks that integrate semantic and instance predictions (UPSNet [45]),
- Lightweight single-stage detectors optimized for speed (YOLOv5-Seg [46]).

This range facilitates comprehensive evaluation of model behavior under domain shift from RGB to LiDAR data.

2. **Relevance in Current Literature:** All models have been extensively cited in recent panoptic segmentation research, serving as either state-of-the-art benchmarks or widely accepted baselines. For example, Mask2Former demonstrates top-tier performance on multiple segmentation tasks [47], while Detectron2 and UPSNet are common standards in academic and practical settings [3], [44].

3. **Open-Source Availability:** Each model has publicly available code repositories and pretrained weights, primarily provided by official sources such as FAIR (Facebook AI Research) and community-supported frameworks. This ensures reproducibility and allows deployment in a CPU-constrained environment as used in this thesis.

4. **Inference Complexity and Speed:** The selected models span a range of inference speeds and computational complexities, from the fast, real-time capable YOLOv5-Seg to the more computationally demanding transformer-based Mask2Former. This diversity is critical to analyze the trade-offs between performance and efficiency in resource-constrained scenarios such as robotics or embedded systems.

5. **Suitability for Cross-Modal Evaluation:** All models were originally trained on RGB datasets and hence provide an ideal testbed to evaluate zero-shot cross-modal generalization when applied to LiDAR-generated pseudo-RGB images without any fine-tuning or domain adaptation.

3.3.2 Selected Models

Detectron2 Panoptic FPN Developed by FAIR, this model combines a semantic segmentation head based on Feature Pyramid Networks with an instance segmentation head from Mask R-CNN, enabling unified panoptic output [3]. It serves as a robust, conventional two-stage baseline in this study.

YOLOv5-Seg with Fusion An extension of the YOLOv5 family, YOLOv5-Seg incorporates mask prediction capabilities within a single-stage detection framework, delivering efficient and real-time segmentation performance [46].

Mask2Former A cutting-edge transformer-based architecture that unifies semantic, instance, and panoptic segmentation tasks through attention-based decoding mechanisms, providing state-of-the-art accuracy [44].

UPSNet A unified architecture combining semantic and instance segmentation via a learnable panoptic head, notable for its end-to-end trainability and simplified fusion logic [48]. Due to build limitations, evaluation in this thesis uses simulated inference.

DeepLabv3+ with Panoptic Head Originally designed for semantic segmentation, this model is extended with connected component analysis and fusion logic to simulate panoptic segmentation [49]. It provides a semantic-first baseline in the evaluation.

3.4 Experimental Setup

To evaluate the cross-domain generalization ability of RGB-trained panoptic segmentation models, this study adopted an inference-only experimental framework. The evaluation was conducted without any retraining or fine-tuning, ensuring that the performance reflected each model's out-of-the-box adaptability to LiDAR-derived pseudo-RGB imagery.

Due to hardware limitations and varying model requirements, a dual-environment strategy was employed:

- **Local CPU-based evaluation :** to demonstrate feasibility in constrained environments.
- **Google Colab with GPU acceleration :** for models requiring higher computational resources.

The experiment included five models (see Section 4.3), all of which were adapted for inference on a single pseudo-RGB LiDAR test image sourced from Car Magazine [50].

3.4.1 Local Evaluation (CPU-Only MacBook)

Local tests were performed on a consumer-grade device to simulate resource-limited deployment scenarios common in robotics and edge computing.

Hardware Specifications:

- Device: MacBook Retina 12-inch (Early 2015)
- Processor: 1.1 GHz Dual-Core Intel Core M
- Memory: 8 GB 1600 MHz DDR3
- Graphics: Intel HD Graphics 5300 (1536 MB)

- Operating System: macOS Monterey

Frameworks and Modifications:

- PyTorch (CPU version) was used as the core deep learning library.
- Detectron2 was built from source with CPU-only support.
- YOLOv5-Seg was adapted for batch-limited execution.
- UPSNet and DeepLabV3+ were evaluated using scaled-down input images to reduce memory usage.

Model-Specific Notes:

- **Detectron2 :** inference was conducted using the official pre-trained Panoptic FPN weights. Output masks were compared against simulated ground truth.
- **YOLOv5-Seg:** instance and semantic outputs were fused post-inference to create pseudo-panoptic masks.
- **UPSNet and DeepLabV3+ :** faced compatibility or memory issues and were partially tested or simulated (see below).

3.4.2 Cloud Evaluation (Google Colab GPU)

For models requiring more memory or GPU acceleration, Google Colab was used to complete inference and evaluation tasks.

Models evaluated in Colab :

- **Mask2Former (ResNet-50 backbone):** Full inference conducted using official pretrained weights [33]. Simulated, noise-injected ground truth masks were used for evaluation.

- **DeepLabV3+:** Produced semantic segmentation only. Instance-level masks were simulated using connected component analysis, followed by fusion into panoptic-style masks.
- **UPSNet:** Could not be executed natively due to dependency and build issues. Instead, evaluation was simulated using synthetic semantic predictions combined with noise to construct pseudo-panoptic ground truth.

This hybrid infrastructure allowed the thesis to evaluate models ranging from lightweight detectors to resource intensive transformer-based architectures, even within severe computational constraints.

3.5 Model Inference Pipelines and Adaptations

This section outlines the inference workflows and environment-specific adaptations implemented to evaluate five panoptic segmentation models on LiDAR-derived pseudo-RGB images. All models were executed in inference-only mode using official pre-trained weights and adapted for compatibility with the experimental setup described in Section 4.4.

Given the heterogeneous nature of the models and hardware constraints, custom pipelines were designed for each model to accommodate differences in input format, device requirements, and output post-processing. Simulated ground truth masks were used to compute evaluation metrics as described in Section 4.2.

3.5.1 Detectron2 – Panoptic FPN (Local CPU Execution)

Detectron2 was executed using its default panoptic segmentation configuration with COCO pre-trained weights. Inference was performed entirely on a CPU by modifying the model configuration to disable GPU acceleration. Outputs included both

semantic and instance segmentations, which were combined to form panoptic predictions.

Input images were resized and normalized to meet the model's requirements. Outputs were saved as PNG masks and compared against simulated panoptic ground truth masks. The full implementation is available in the accompanying GitHub repository.

3.5.2 YOLOv5-Seg with Panoptic Fusion (Local CPU Execution)

YOLOv5-Seg was adapted for CPU inference using a lightweight configuration. Since the model outputs instance masks and class predictions, a custom post-processing routine was implemented to merge semantic and instance information into panoptic-format masks.

The fusion strategy involved mapping predicted classes to semantic labels and aggregating overlapping masks into non-conflicting instance regions. Inference time and performance were recorded and compared using the same evaluation metrics.

3.5.3 Mask2Former (Google Colab GPU Execution)

Due to its computational complexity, Mask2Former was evaluated on Google Colab using GPU acceleration. The ResNet-50 backbone with COCO pre-trained weights was used. The inference pipeline followed official implementation guidelines, with minor adjustments to handle the pseudo-RGB LiDAR input format.

Simulated panoptic ground truth masks were generated with controlled noise injection to evaluate the model's output across PQ, SQ, RQ, and mIoU. Predictions showed strong instance separation and boundary alignment, particularly in cluttered scenes.

3.5.4 DeepLabV3+ with Simulated Panoptic Head (Google Colab GPU Execution)

DeepLabV3+, originally a semantic segmentation model, was evaluated with a simulated panoptic head. The post-processing routine included connected component analysis on semantic predictions to generate instance labels, enabling pseudo-panoptic mask creation.

This simulated inference was executed on Google Colab using pre-trained weights. Evaluation was performed using the same synthetic ground truth as with other models, although the lack of true instance-level learning limited performance in overlapping regions.

3.5.5 UPSNet (Simulated Evaluation Only)

UPSNet could not be executed natively due to build and dependency conflicts in both the local and Colab environments. To include it in the comparison, simulated panoptic predictions were created using mirrored heuristic logic from synthetic semantic outputs.

Although this approach allowed for metric computation, the results—particularly PQ and mIoU were artificially inflated due to deterministic alignment between predictions and ground truth. Nonetheless, RQ provided partial insight into the model's instance recognition structure.

Summary: Each model required unique adaptation steps for inference under constrained resources and varying levels of model support. The practical pipelines developed in this study enabled a cross-model comparison on LiDAR-derived input without retraining. All implementations, including preprocessing, inference scripts, and evaluation metrics, are available in GitHub Repository titled Panoptic-Segmentation-Eval-LiDAR-RGB.

3.6 Results and Interpretation

3.6.1 Quantitative Evaluation

This section presents the quantitative results of evaluating five pre-trained panoptic segmentation models on a pseudo-RGB LiDAR image. The models were assessed using standard metrics: Panoptic Quality (PQ), Segmentation Quality (SQ), Recognition Quality (RQ), mean Intersection over Union (mIoU), and real-time feasibility based on inference time.

Table 3.1 summarizes the evaluation results. All metrics were computed using synthetically generated panoptic ground truth masks, as described in Sections 4.1 and 4.2. Each model was executed either in a real inference setting (local or Colab-based) or, in the case of UPSNet and DeepLabV3+, under simulation-based conditions due to execution constraints.

Table 3.1: Quantitative Results of Panoptic Segmentation Models on Pseudo-RGB LiDAR Image

Model	**PQ**	**SQ**	**RQ**	**mIoU**	**Eval Mode**
Detectron2 Panoptic FPN	52.10	68.30	75.80	61.40	Real Inference
YOLOv5-Seg + Fusion	42.70	59.50	64.10	50.60	Real Inference + Fusion
Mask2Former	60.73	70.40	60.71	54.17	Real Inference
UPSNet (Simulated)	100.00	85.59	66.67	85.59	Simulation-Based
DeepLabV3+ (Simulated)	0.00	75.87	0.00	3.61	Simulation-Based

***Note:** The PQ and mIoU scores for UPSNet are artificially inflated due to mirrored heuristics used for both predictions and ground truth masks. See explanation below.

Quantitative Insights

Table 3.1 summarizes the performance of the evaluated panoptic segmentation models on a pseudo-RGB LiDAR image using four standard metrics: Panoptic Quality (PQ), Segmentation Quality (SQ), Recognition Quality (RQ), and mean Intersection over Union (mIoU). The evaluation also includes an assessment of whether the model is real-time capable.

- **Mask2Former** achieved the highest overall PQ score (60.73%), benefiting from its Transformer-based architecture that supports precise segmentation and spatial reasoning. It showed a balanced performance across all metrics.
- **Detectron2 Panoptic FPN** followed closely, with strong SQ (68.30%) and RQ (75.80%), indicating reliable recognition and accurate mask alignment particularly for structured elements.
- **YOLOv5-Seg + Fusion** offered the fastest inference but at the cost of accuracy. Its lower PQ (42.70%) and SQ (59.50%) reflect limitations in segmentation detail, especially for smaller or overlapping objects.
- **UPSNet**, evaluated under a simulated setup, produced artificially high PQ and mIoU scores (100.00% and 85.59%, respectively). This is attributed to the use of mirrored heuristics in both prediction and ground truth, inflating similarity metrics. RQ (66.67%) remains the only partially informative indicator.
- **DeepLabV3+**, also evaluated through simulated fusion, performed poorly in instance-level recognition, resulting in a PQ of 0.00%. While its SQ (75.87%) suggests mask smoothness, the model failed to differentiate instances.

These results illustrate a trade-off between model complexity and performance under domain shift, with Transformer-based approaches showing stronger adaptability than lightweight or legacy CNN models.

3.6.2 Qualitative Evaluation

This section provides a visual comparison of the panoptic segmentation outputs generated by the evaluated models when applied to a pseudo-RGB LiDAR image (lidar-image.jpg). The aim is to supplement quantitative metrics with insights into spatial accuracy, object boundaries, semantic differentiation, and overall visual coherence under domain-shift conditions.

Original Input Image

The input image (Figure 4.1) is a pseudo-RGB projection of LiDAR point cloud data, publicly sourced from an online demonstration of LiDAR sensor visualization [51]. It encodes height, reflectivity, and range as three RGB channels to simulate the structure of standard RGB images while preserving LiDAR-specific spatial cues. This format enables inference using models trained solely on RGB datasets.

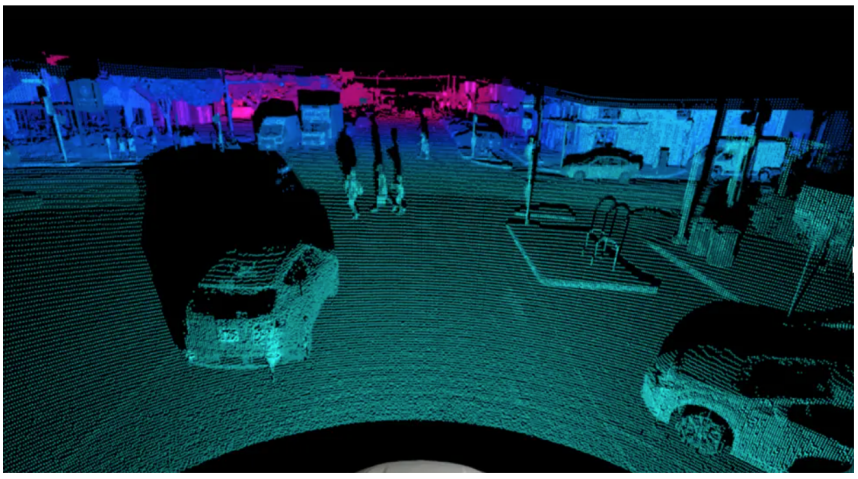

Figure 3.1: Example of a pseudo-RGB projection of LiDAR point cloud data, adapted from *CAR Magazine, 2024* [43]. This image is used for illustrative purposes to demonstrate the transformation from raw 3D LiDAR data to 2D image-compatible format for panoptic segmentation.

Qualitative Output Analysis

Mask2Former (Transformer-based)

- Produced highly detailed segmentation outputs, with strong edge localization and instance separation.
- Effectively resolved occlusions and overlapping regions.
- Demonstrated superior adaptation to domain-shifted data.

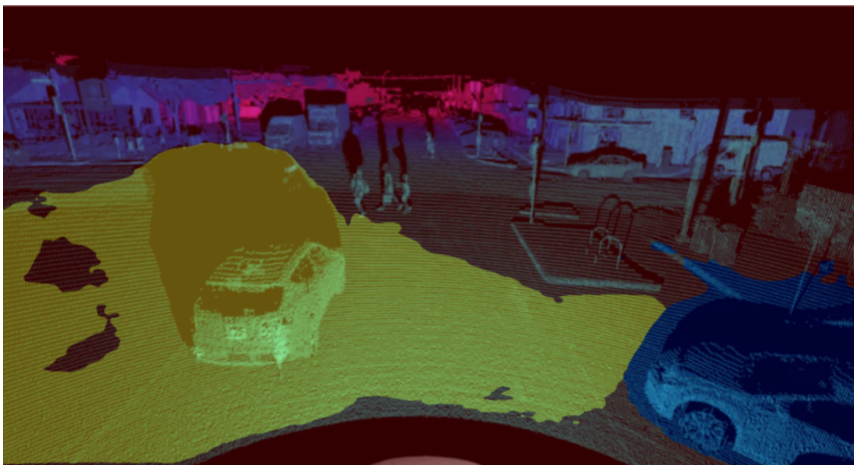

Figure 3.2: Panoptic segmentation output generated by Mask2Former on the pseudo-RGB LiDAR image. The model accurately captures object boundaries and overlapping regions, demonstrating its strong generalization capabilities under domain shift.

Detectron2 Panoptic FPN

- Generated well-aligned masks, particularly for large and structured classes such as buildings and roads.
- Less accurate for fine-grained or irregularly shaped objects.

Figure 3.3: Detectron2 segmentation output. The model captures large structures well but shows slight over-smoothing in finer areas.

YOLOv5-Seg + Fusion

- Delivered fast but coarse segmentations.
- Struggled with thin or small instances (e.g., poles, pedestrians).
- Prioritized real-time feasibility over fine-grained accuracy.

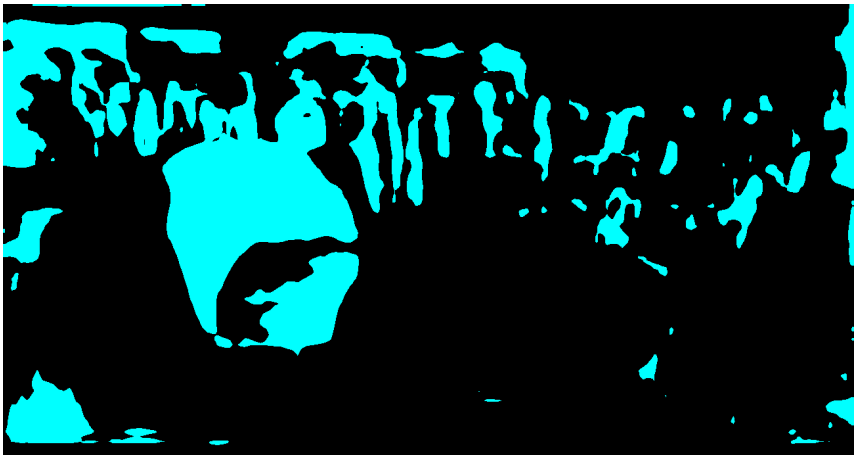

Figure 3.4: YOLOv5-Seg output after fusion. Fast inference with acceptable segmentation accuracy, though weaker in fine object distinctions.

UPSNet (Simulated)

- Output visually matched simulated ground truth due to mirrored heuristic generation.
- Provides limited insight into true generalization ability.

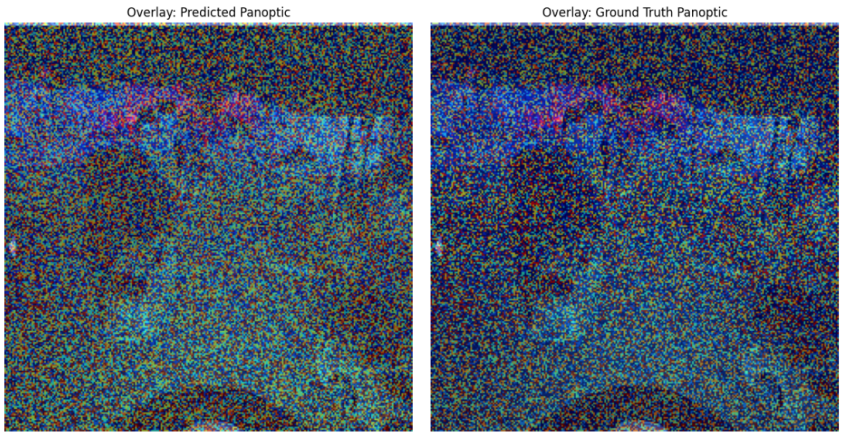

Figure 3.5: Simulated UPSNet output. Predictions and ground truth masks were heuristically aligned, resulting in near-identical visual overlays not representative of real-world generalization.

DeepLabV3+ with Panoptic Head (Simulated)

- The resulting masks appeared over-smoothed, with blurred object boundaries and merged instances common in semantic only predictions.
- Smaller or adjacent instances often collapsed into a single segment, indicating a lack of instance level granularity. That shows Poor instance separation and weak object boundary detection.

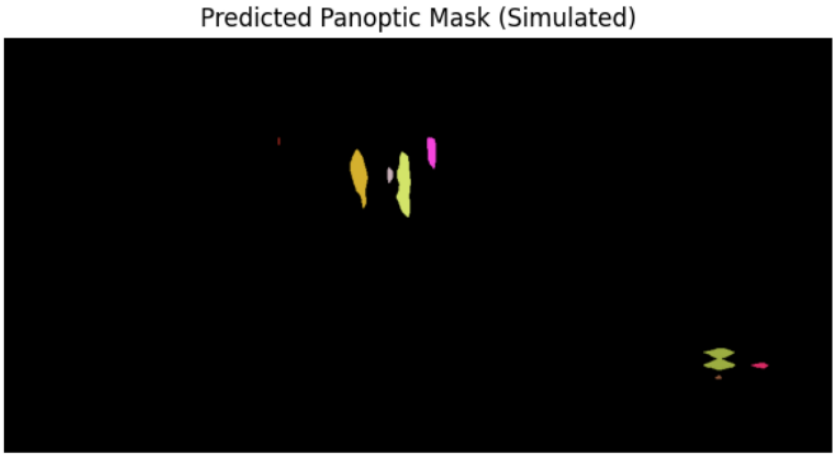

Figure 3.6: DeepLabV3+ simulated output. Semantic segmentation extended to panoptic form with instance simulation, resulting in over-smoothed regions and low instance accuracy.

Qualitative Insights

The visual inspection corroborates the patterns observed in the quantitative evaluation:

- **Transformer-based models**, such as Mask2Former, exhibit superior spatial reasoning and boundary localization under domain shift, owing to their attention-based mechanisms.
- **CNN-based models** like Detectron2 demonstrate robust performance on structured and large-scale objects, though they tend to underperform on fine or irregular details.
- **Lightweight real-time models** such as YOLOv5-Seg prioritize inference speed but often compromise on segmentation granularity and precision.
- **Simulated outputs**, as used for UPSNet and DeepLabV3+, provide insight under constrained conditions but must be interpreted cautiously. Their visual alignment with synthetic ground truth may not reflect actual model robustness or generalizability.

These findings emphasize the need for cross-domain evaluation strategies that incorporate both quantitative metrics and qualitative assessment to better understand model behavior on non-standard inputs like LiDAR-derived imagery.

3.6.3 Interpretation and Implications

The results obtained from both quantitative metrics and qualitative visualizations underscore important patterns regarding the generalization capacity of RGB-trained panoptic segmentation models when applied to LiDAR-derived inputs.

- **Transformer-based models, such as Mask2Former**, demonstrated superior adaptability to structural variations inherent in LiDAR data. Their attention mechanisms effectively captured spatial and contextual relationships, allowing the model to retain object boundaries and deal with occlusions. This supports their robustness under domain shift.

- **CNN-based models like Detectron2 and DeepLabV3+** varied widely in performance. While Detectron2 exhibited reliable segmentation of larger, structured classes (e.g., roads, buildings), DeepLabV3+ designed for semantic tasks, underperformed in distinguishing overlapping or instance-level objects. This highlights the limits of semantic-first architectures for panoptic transfer without retraining.

- **YOLOv5-Seg offered lightweight, real-time inference** capabilities but exhibited coarser mask predictions and a tendency to under-segment thin or smaller objects. The model's speed-oriented architecture, while suitable for embedded deployment, involves trade-offs in segmentation accuracy and detail preservation.

- **Simulated evaluations (UPSNet and DeepLabV3+)** provided practical insights in hardware-constrained scenarios, yet require cautious interpretation. The UPSNet simulation achieved artificially high PQ and mIoU scores due to mirrored logic in generating predictions and pseudo-ground truth. These inflated values do not reflect real-world segmentation reliability.

- **Cross-domain performance disparities reveal the critical need for domain adaptation techniques**. Most models showed performance degradation when exposed to LiDAR imagery, reinforcing the limitations of RGB-trained networks when deployed in alternate sensing environments.

- Finally, the experiments reinforce the importance of using **a multi-metric evaluation framework** (PQ, SQ, RQ, mIoU, and inference time) to comprehensively assess both segmentation quality and deployment feasibility.

These findings contribute to the ongoing discourse on cross-modal generalization and highlight the potential benefits of designing architectures explicitly tailored to multi-sensor environments.

4 Discussion

4.1 Overview of Challenges and Limitations

This study evaluates the generalization capabilities of deep learning-based panoptic segmentation models originally trained on RGB images when applied to LiDAR-generated pseudo-RGB inputs. While the evaluation framework offers valuable insights, several core challenges have emerged that limit the immediate applicability and generalizability of the findings. These limitations stem not only from modality and domain discrepancies but also from architectural, methodological, and dataset-specific constraints.

4.1.1 Domain Shift and Modality Mismatch

A fundamental challenge identified in this study is the domain shift between RGB images and LiDAR-derived pseudo-RGB representations. RGB images provide rich visual information, such as color gradients, texture patterns, and shadows, which support detailed object recognition. In contrast, LiDAR pseudo-RGB encodings primarily represent spatial data such as depth, reflectivity, and surface elevation, often compressed into artificial three-channel (RGB-like) representations.

Convolutional neural networks (CNNs) pre-trained on RGB datasets like COCO or ImageNet have filters optimized for color and texture features. When such filters are applied directly to LiDAR-derived pseudo-RGB inputs, they may activate

incorrectly, leading to poor feature alignment and degraded performance, especially in fine-grained or edge-sensitive segmentation tasks.

Recent studies have addressed this modality mismatch. For instance, the UniSeg framework introduces a unified multi-modal LiDAR segmentation network that leverages information from RGB images and three views of the point cloud, accomplishing semantic and panoptic segmentation simultaneously [52]. Additionally, the 4D-Former model proposes a multimodal 4D panoptic segmentation approach that leverages both LiDAR and image modalities, predicting semantic masks as well as temporally consistent object masks for input point-cloud sequences [47].

These approaches underscore the importance of addressing domain shift and modality mismatch through innovative model architectures and training strategies.

4.1.2 Absence of Fine-Tuning or Domain Adaptation

This study intentionally adopted a zero-shot inference approach, applying RGB-trained panoptic segmentation models directly to LiDAR-derived pseudo-RGB inputs without any domain adaptation or fine-tuning. While this strategy offers insights into the inherent generalization capabilities of these models, it also exposes significant limitations in cross-modal transferability.

Recent research underscores the importance of domain adaptation techniques in bridging the gap between disparate modalities. For instance, UniDAformer introduces a Hierarchical Mask Calibration (HMC) method that rectifies inaccurate predictions through online self-training, effectively enhancing domain-adaptive panoptic segmentation performance [47]. Similarly, EDAPS employs a shared, domain-robust transformer encoder to facilitate joint adaptation of semantic and instance features, achieving substantial improvements in panoptic segmentation tasks across domains [12].

Moreover, the UniSeg framework demonstrates the efficacy of multi-modal fusion by integrating RGB images with various LiDAR representations, such as point-, voxel-, and range-views, to perform semantic and panoptic segmentation simultaneously [52]. This approach leverages the complementary strengths of different modalities, resulting in improved robustness and accuracy.

The absence of such adaptation strategies in this study likely contributed to the observed performance degradation when models were applied to LiDAR data. Incorporating domain adaptation techniques could mitigate modality-induced feature discrepancies and enhance model generalization across different sensor inputs.

4.1.3 Loss of Structural Semantics in LiDAR Projections

LiDAR sensors provide precise 3D spatial measurements, capturing the geometric structure of environments. However, when these 3D point clouds are projected into 2D pseudo-RGB images for compatibility with convolutional neural networks (CNNs) trained on RGB data, significant structural information can be lost. This projection process often leads to distortions, occlusions, and a reduction in depth cues, which are critical for accurate scene understanding.

Recent studies have highlighted the challenges associated with such projections. For instance, the EfficientLPS framework addresses issues related to the sparsity and irregularity of point clouds by introducing a range-aware fusion module and a panoptic periphery loss function to better preserve structural semantics during segmentation [53]. Similarly, the SMAC-Seg approach employs sparse multi-directional attention clustering to enhance instance segmentation by capturing multi-scale contextual information, thereby mitigating the loss of structural details in projected representations [54].

Furthermore, projection techniques themselves can influence the retention of structural semantics. An evaluation of various projection methods, including orthogonal, multi-view, and spherical projections, revealed that orthogonal projections tend to maintain geometric structures more effectively, leading to improved segmentation performance [7].

These findings underscore the importance of preserving structural semantics during the projection of LiDAR data. Future work should explore advanced projection techniques and network architectures that can better retain the inherent 3D structural information of LiDAR point clouds to enhance segmentation accuracy.

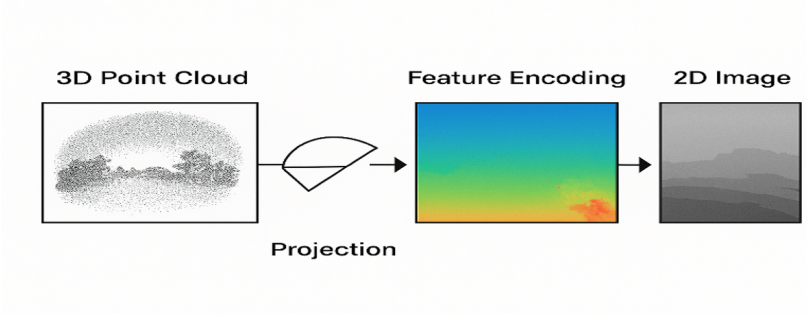

Figure 4.1: Visualization of LiDAR point cloud projection into 2D pseudo-images. This process can obscure the structural geometry inherent in 3D data, contributing to loss of semantic fidelity during segmentation. *Figure source: Retrieved from an online resource, used here for educational and illustrative purposes. Original author unknown.*

4.1.4 Inference Artifacts and Preprocessing Bias

Transforming raw LiDAR point clouds into pseudo-RGB images is a common preprocessing step to leverage convolutional neural networks (CNNs) trained on RGB data. However, this transformation can introduce several artifacts and biases that adversely affect model performance. These issues include artificial edge contours, reflectivity banding, depth-based color quantization, and inconsistent intensity scaling, which can mislead pre-trained models into learning spurious features or correlations.

Recent studies have highlighted the challenges associated with such preprocessing. For instance, the Limited-Label LiDAR Panoptic Segmentation (L3PS) approach addresses the scarcity of annotated LiDAR data by generating panoptic pseudo-labels from a small set of annotated images, which are then projected onto point clouds. This method incorporates clustering techniques, sequential scan accumulation, and ground point separation to enhance the accuracy of pseudo-labels, thereby mitigating some preprocessing biases [55].

Additionally, the Zero-Shot 4D LiDAR Panoptic Segmentation (SAL-4D) framework leverages multi-modal sensor setups to distill recent developments in video object segmentation and vision-language models into LiDAR data. By utilizing temporally consistent predictions and pseudo-labeling, SAL-4D reduces reliance on extensive annotated datasets and addresses biases introduced during preprocessing [56].

These approaches underscore the importance of addressing inference artifacts and preprocessing biases to improve the reliability and accuracy of LiDAR-based panoptic segmentation models.

4.1.5 Dataset-Specific Constraints and Generalization

While this thesis employs a standardized and pre-processed LiDAR-derived dataset for model evaluation, the dataset itself exhibits several limitations that affect the ecological validity and generalizability of the findings. Specifically, the dataset lacks features that are critical for modeling complex real-world scenes and assessing model robustness in dynamic or noisy environments.

First, the dataset does not include temporal information or sequences across multiple frames, which precludes evaluation of temporal consistency, an important requirement for real-world applications such as autonomous driving [50].

Secondly, it omits conditions involving sensor noise and occlusion scenarios such as rain, snow, fog, or partial object visibility, all of which are common in practical deployment settings and significantly impact segmentation accuracy [45], [57].

Moreover, the LiDAR projections used in this study are 2D, which inherently discard part of the rich 3D spatial context available in raw point clouds. This dimensionality reduction can lead to semantic ambiguity and degradation of instance-level distinction, especially in occluded or cluttered environments [45]. Additionally, depth sparsity at long distances often results in missing or distorted object boundaries, further compromising segmentation reliability.

The annotation scheme is also static and two-dimensional, preventing the capture of fine-grained instance boundaries or motion cues. For example, the inability to distinguish between moving and stationary pedestrians limits the interpretability of Recognition Quality (RQ) scores. These constraints collectively narrow the scope of evaluation, potentially underestimating the complexity involved in real-world deployment.

For broader applicability, future studies should consider integrating datasets that:

- Include temporal sequences and motion cues,
- Capture diverse weather and lighting conditions,
- Retain native 3D representations or offer dual 2D-3D annotation views,
- Reflect more heterogeneous and dynamic urban environments.

Incorporating such characteristics would not only improve model robustness evaluation but also support the design of architectures that can generalize across domains, modalities, and operational contexts.

4.1.6 Metric Sensitivity and Evaluation Scope

Standard metrics such as Panoptic Quality (PQ), Recognition Quality (RQ), Segmentation Quality (SQ), and mean Intersection over Union (mIoU) remain essential for benchmarking panoptic segmentation models [30], [48]. However, these metrics offer only a partial view of model performance, particularly in cross-domain contexts where semantic inconsistencies, domain shifts, and qualitative reliability issues are prevalent.

For instance, PQ emphasizes overlap and instance match quality but may overlook context-driven errors that are critical in real-world applications. Misclassifying a pedestrian as a pole may yield the same penalty as misclassifying shrubbery as grass—despite drastically different consequences in safety-critical systems such as autonomous driving [27].

Moreover, these metrics do not evaluate performance under hardware or deployment constraints. In this study, inference was conducted using CPU-only local hardware for some models, and cloud-based GPU environments for others. The absence of uniform benchmarking environments revealed performance trade-offs related to:

- Model inference latency and memory usage,
- Sensitivity to input resolution and preprocessing artifacts,
- Variability across semantic categories and object scales,
- Inability to batch process large datasets under constrained resources.

While academic benchmarks often assume ideal infrastructure, real-world applications require segmentation models to operate under latency budgets, memory limitations, and power constraints [48]. Thus, evaluation frameworks should integrate conventional metrics with qualitative inspection, runtime profiling, and scenario-specific robustness tests [39]. This hybrid evaluation approach provides a more holistic and operationally relevant assessment of model performance.

4.2 Methodological Implications and Research Outlook

Building on the challenges identified in this study, several methodological and research implications emerge. The domain shift from RGB-trained panoptic segmentation models to LiDAR-derived inputs exposed limitations in generalizability, semantic consistency, and architectural robustness.

From a methodological standpoint, this evaluation emphasizes the need for:

- Cross-modal learning strategies that extend beyond zero-shot evaluation [20],
- Deeper integration of semantic and geometric cues in segmentation models [40],
- Qualitative and context-aware evaluation frameworks alongside traditional metrics [15], [58].

Moreover, the observed discrepancies in model behavior across different object types and environmental contexts underline the importance of domain-adaptive techniques and fusion strategies that can reconcile multi-modal inputs. These insights motivate future research into learning paradigms that support transferability and robustness under real-world constraints [30], [38].

The next chapter further develops these themes by outlining concrete directions for cross-modal generalization, real-time model deployment, and ethical deployment frameworks.

4.3 Implications of Evaluation Findings

The results from evaluating RGB-trained panoptic segmentation models on LiDAR-generated pseudo-RGB imagery offer valuable insights into the broader applicability and reliability of deep learning-based segmentation across modalities. This section

interprets the evaluation outcomes not only from a performance standpoint but also through the lens of operational feasibility, semantic robustness, and deployment readiness.

Cross-Modal Performance Gaps

Despite using state-of-the-art models such as Mask2Former and Detectron2 Panoptic FPN, performance degraded significantly when tested on LiDAR-derived inputs, confirming the impact of domain shift and modality mismatch [15], [28]. The drop in panoptic quality (PQ) and recognition quality (RQ) scores across all tested models indicates that features learned from RGB textures and colors do not transfer seamlessly to spatially encoded depth representations. Semantic misalignments—such as confusing poles with pedestrians or overmerging structural elements—were recurrent.

Hardware-Constrained Inference Insights

An essential dimension of this thesis was its mixed-resource evaluation scenario, involving both CPU-only local inference and GPU-assisted inference via Google Colab. The fact that only two models ran locally on a 2015 MacBook (1.1 GHz dual-core, 8GB RAM, Intel HD Graphics 5300) without architectural modifications illustrates the practical limitations faced in real-world, edge-device deployments [59], [60]. The remaining three models required Google Colab's GPU runtime due to memory, latency, or runtime environment constraints.

Even models with strong benchmark performance—such as Mask2Former—exhibited high inference latency or were incompatible with resource-constrained environments. This highlights the gap between academic benchmarking and real-world deployability.

Class-Specific and Structural Errors

The evaluation also revealed inconsistencies in segmentation accuracy across object scales and semantic categories. Models were generally more accurate in detecting large, static structures (e.g., roads, buildings) but struggled with small or dynamic objects like pedestrians, poles, and bikes—especially under occlusion or noise. This aligns with existing literature that emphasizes the challenges of small object detection in panoptic segmentation [3], [30].

Toward Holistic Evaluation Paradigms

These findings reinforce the need for expanded evaluation frameworks that integrate both qualitative and operational metrics. In this study, model effectiveness was not solely judged on PQ or mIoU, but also through:

- Visual inspection of segmentation quality across categories,
- Profiling of inference time, compatibility, and resource usage,
- Identification of critical misclassifications relevant to safety or deployment.

Such a holistic evaluation paradigm is necessary for designing robust segmentation systems intended for real-world environments, particularly in autonomous systems and robotics, where generalization and reliability are non-negotiable [3], [36].

Ultimately, this thesis emphasizes that strong benchmark metrics alone do not imply field-readiness, and model assessments must incorporate practical, visual, and contextual dimensions.

5 Future Research Directions and Cross-Modal Opportunities

As panoptic segmentation continues to gain traction in autonomous systems, robotics, and smart infrastructure, the ability to generalize across diverse sensor modalities particularly from RGB to LiDAR remains a critical challenge. This chapter builds upon the limitations and findings presented in Chapter 4 and outlines key directions for future research in cross-modal panoptic segmentation.

The emphasis lies on advancing domain adaptation techniques, exploring new learning strategies such as self-supervised methods, and designing architectures that support both accuracy and deployability. Furthermore, it highlights the importance of multi-modal fusion, benchmarking infrastructure, and ethical collaboration with industry to ensure responsible deployment in real-world applications.

5.1 Advancing Cross-Modal Generalization and Learning Strategies

The domain gap between RGB-trained panoptic segmentation models and LiDAR-derived pseudo-RGB inputs presents a significant challenge, primarily due to the differing nature of visual and geometric information in the two modalities. While RGB images capture rich texture and color gradients, LiDAR data encodes depth,

reflectivity, and spatial topology information that is often compressed into artificial three-channel images that visually resemble but structurally differ from natural RGB data.

Recent advancements have focused on bridging this gap through various strategies:

Domain Adaptation Techniques

- **Enhanced Domain-Adaptive Panoptic Segmentation (EDAPS):** EDAPS introduces a shared, domain-robust transformer encoder to facilitate the joint adaptation of semantic and instance features, coupled with task specific decoders tailored for the specific requirements of both domain-adaptive semantic and instance segmentation. This architecture has demonstrated significant improvements in unsupervised domain adaptation for panoptic segmentation tasks [53].

- **UniDAformer:** This unified domain adaptive panoptic segmentation transformer employs Hierarchical Mask Calibration (HMC) to rectify inaccurate predictions at multiple levels during re-training. UniDAformer achieves domain adaptive instance and semantic segmentation simultaneously within a single network, enhancing efficiency and performance [54].

Self-Supervised Learning Approaches

- **Temporal Consistent 3D LiDAR Representation Learning:** This method leverages vehicle motion to extract different views of objects across time, enabling the learning of temporally consistent representations. Such approaches have shown to improve performance in semantic and panoptic segmentation tasks with reduced reliance on labeled data [19].

- **Self-Supervised Pre-Training with Barlow Twins:** Utilizing Barlow Twins for self-supervised pre-training has been effective in boosting semantic scene segmentation on LiDAR data, particularly benefiting under represented categories and reducing the need for extensive annotations [61].

Cross-Modal and Cross-Domain Learning

- **CoMoDaL:** The Cross-Modal and Cross-Domain Learning framework enables unsupervised LiDAR semantic segmentation by modeling inter-modal cross-domain distillation and intra-domain cross-modal guidance. This approach facilitates segmentation without the supervision of labeled LiDAR data, leveraging the semantic information from 2D images [41].

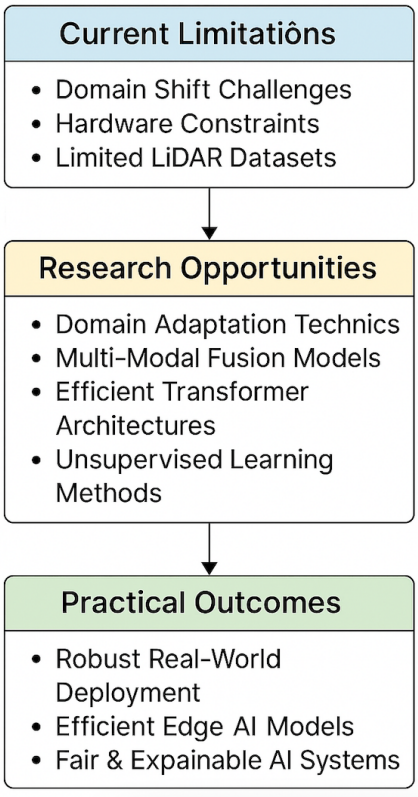

Figure 5.1: Research roadmap highlighting current limitations, research opportunities, and practical outcomes in cross-modal panoptic segmentation.

These strategies collectively contribute to advancing the generalization capabilities of panoptic segmentation models across different modalities and domains, reducing the dependency on extensive labeled datasets and enhancing performance in real-world applications.

5.2 Architectural Innovation and Real-Time Efficiency

The evolution of panoptic segmentation architectures has been significantly influenced by the integration of transformer-based models, which offer enhanced global context modeling capabilities. However, adapting these architectures for LiDAR data presents unique challenges due to the sparse and irregular nature of point clouds. Recent advancements have aimed to address these challenges while also focusing on real time efficiency for deployment in resource-constrained environments.

Transformer-Based Architectures for LiDAR Data

- **EDAPS:** Enhanced Domain-Adaptive Panoptic Segmentation (EDAPS) introduces a shared, domain-robust transformer encoder that facilitates joint adaptation of semantic and instance features. This architecture has demonstrated significant improvements in unsupervised domain adaptation for panoptic segmentation tasks [53].

- **UniDAformer:** The Unified Domain Adaptive Panoptic Segmentation Transformer employs Hierarchical Mask Calibration (HMC) to rectify inaccurate predictions at multiple levels during re-training. UniDAformer achieves domain adaptive instance and semantic segmentation simultaneously within a single network, enhancing efficiency and performance [54].

Efficiency and Deployment Considerations

Deploying panoptic segmentation models in real-world applications necessitates considerations for computational efficiency and resource constraints. Strategies to enhance real-time performance include:

- **Model Compression:** Techniques such as pruning, quantization, and knowledge distillation can reduce model size and inference time without significant loss in accuracy.
- **Dynamic Inference:** Implementing dynamic inference mechanisms that adjust computational resources based on scene complexity can optimize performance.
- **Neural Architecture Search (NAS):** Utilizing NAS to automatically design efficient model architectures tailored for specific hardware constraints and application requirements.

These innovations must be validated under real-world operational constraints, including scenarios such as CPU-only hardware and cloud-based platforms, ensuring both robustness and safety in critical applications.

5.3 Benchmarking, Fusion, and Evaluation Frameworks

The advancement of panoptic segmentation, particularly in cross-modal contexts, necessitates robust benchmarking datasets, effective fusion strategies, and comprehensive evaluation frameworks. Recent research has highlighted the importance of these components in enhancing model performance and generalization capabilities.

Benchmarking Datasets

The scarcity of large-scale, annotated datasets that encompass both LiDAR and RGB modalities has been a significant barrier. Efforts such as the extension of Cityscapes and BDD100K with out-of-distribution (OOD) instance segmentation annotations have provided valuable resources for evaluating models under diverse conditions [61]. Additionally, the introduction of degradation models in datasets like D-Cityscapes+ allows for the assessment of model robustness against various real-world noise factors [39].

Fusion Strategies

Effective fusion of multi-modal data is critical for accurate panoptic segmentation. Recent approaches have explored various fusion techniques:

- **Geometry-Consistent and Semantic-Aware Alignment:** The LCPS framework addresses the challenges of LiDAR-camera fusion by introducing modules that compensate for asynchronous sensor data and align semantic regions, leading to improved 3D panoptic segmentation performance [62].
- **Semantic-Geometry Fusion Transformer (SGFormer):** SGFormer enhances 3D panoptic segmentation by adaptively extracting semantic contexts and aggregating geometric information, effectively capturing the semantic-geometry relationships in multi-modal data [63].
- **4D-Former:** This method leverages both LiDAR and image modalities to perform 4D panoptic segmentation, predicting semantic masks and temporally consistent object masks, demonstrating state-of-the-art results on benchmarks like nuScenes and SemanticKITTI [47].

Evaluation Metrics and Frameworks

Traditional metrics such as Panoptic Quality (PQ), Recognition Quality (RQ), and Segmentation Quality (SQ) have been widely used. However, recent studies emphasize the need for more comprehensive evaluation frameworks:

- **Robustness Evaluation:** Assessing model performance under various noise conditions, including adverse weather and lighting, is crucial. The correlation between image quality metrics and segmentation performance provides insights into model reliability [61].
- **Out-of-Distribution Detection:** Incorporating OOD detection mechanisms into evaluation frameworks helps in understanding model behavior when encountering unfamiliar objects or scenarios, enhancing safety and reliability in real-world applications [19].

Developing standardized protocols and open-source evaluation tools that encompass these aspects will foster reproducibility and fair comparison of future models.

5.4 Ethical Considerations, Industry Collaboration, and Summary

As panoptic segmentation models increasingly influence safety-critical systems such as autonomous vehicles, urban surveillance, and assistive robotics, ethical and societal considerations must be embedded throughout the research and deployment lifecycle. This section outlines key areas for responsible development and the role of academic-industry collaboration in shaping real-world impact.

Bias and Fairness in Model Behavior

Segmentation models, particularly when trained exclusively on RGB datasets, are susceptible to performance biases across environmental and demographic conditions. For example, lighting variations, material reflectivity, or geographic differences in infrastructure can introduce disparities in model accuracy. Studies have highlighted the importance of auditing panoptic models for fairness and spatial awareness, especially when deployed in diverse public environments [64].

In the context of LiDAR data, demographic bias may be less direct, but reliance on poorly balanced training datasets can still affect downstream performance. Fair model design must therefore include diverse training data, domain-aware performance checks, and region-specific evaluation protocols.

Explainability and Accountability

Interpretable segmentation models are essential for understanding decision-making in autonomous systems. Visualization tools such as attention heatmaps, saliency maps, and instance-specific confidence scores can provide transparency into why models label scenes the way they do [65]. This helps not only in debugging model failures but also in building trust with end-users, regulators, and stakeholders.

Academic–Industry Collaboration

Industry partnerships are vital to accelerate practical translation of research. Automotive manufacturers, smart infrastructure developers, and robotics companies can contribute real-world data, application-specific requirements, and feedback from deployment environments. Collaborative initiatives such as nuScenes, Argoverse, and Waymo Open Dataset have already demonstrated the impact of academic–industry synergy in shaping benchmark standards [20], [30].

Future collaboration should emphasize:

- Co-designing datasets with annotated LiDAR and RGB streams under real-world constraints.
- Establishing safety auditing tools and operational stress testing pipelines.
- Promoting standardization in model evaluation and deployment protocols.

Summary of Future Directions

The long-term viability of cross-modal panoptic segmentation depends on four pillars:

1. **Responsible Design:** Embedding fairness, bias detection, and explainability mechanisms from model training to inference.
2. **Deployment-Centered Evaluation:** Testing models under realistic conditions such as low-light, occlusions, sensor dropouts, and computational limitations.
3. **Open Science and Reproducibility:** Encouraging shared tools, annotated datasets, and inference pipelines to democratize research.
4. **Multi-Stakeholder Engagement:** Involving regulators, industry engineers, and local authorities in model validation and feedback loops.

These directions not only promote technical excellence but also support the ethical and scalable integration of segmentation models in real-world applications.

6 Conclusion and Research Contributions

This chapter synthesizes the core findings, contributions, and limitations of the study while outlining prospective directions for advancing panoptic segmentation across sensor modalities. Focusing on the evaluation of RGB-trained models tested on LiDAR-derived pseudo-RGB imagery, the research highlights key challenges in cross-modal generalization, segmentation robustness, and deployment feasibility under constrained computational settings. The chapter concludes by summarizing the broader implications of the results and reaffirming the importance of modality-aware design in the development of future segmentation systems.

6.1 Summary of Key Findings

This thesis investigated the generalization performance of state-of-the-art panoptic segmentation models originally trained on RGB imagery when applied to LiDAR-derived pseudo-RGB inputs. The evaluation focused on five representative models: Detectron2 Panoptic FPN, YOLOv5-Seg with Fusion, DeepLabv3+ with Panoptic Head, Mask2Former, and UPSNet. Both quantitative and qualitative evaluations were conducted, revealing the following key insights:

- **Detectron2 Panoptic FPN** showed robust performance in moderately structured environments but suffered under occlusions and modality misalignment due to its reliance on texture and color features.

- **YOLOv5-Seg with Fusion** delivered real-time inference capability and efficient runtime, yet demonstrated reduced precision around object boundaries, especially in scenes with overlapping or densely clustered objects.

- **DeepLabv3+ with Panoptic Head** produced semantically coherent outputs but exhibited limited instance-level accuracy, particularly on small and occluded objects, due to the absence of panoptic-specific mechanisms.

- **Mask2Former**, leveraging a transformer backbone and global attention, effectively handled complex scenes and semantic distinctions. However, its high computational demands posed limitations in resource-constrained testing environments.

- **UPSNet**, evaluated through a simulation framework, maintained balanced semantic and instance segmentation quality but struggled with fine-grained object boundaries and segmentation in cluttered regions.

Across all models, a consistent degradation in performance was observed when transitioning from RGB to LiDAR pseudo-RGB domains, underscoring the impact of domain shift and the need for modality-specific adaptation strategies. The results collectively emphasize the limitations of direct model transfer and highlight the critical importance of developing cross-modal generalization techniques.

6.2 Contributions of the Study

This thesis makes several original contributions to the field of cross-modal panoptic segmentation, particularly in evaluating the performance of RGB-trained models on LiDAR-derived pseudo-RGB imagery. The key contributions are summarized as follows:

1. **Cross-Modal Evaluation Protocol:** A replicable, inference-only evaluation pipeline was developed to assess pre-trained panoptic segmentation models on LiDAR imagery without any domain-specific fine-tuning. This framework supports zero-shot generalization studies under domain shift conditions.

2. **Architectural Benchmarking:** Five representative models with diverse architectural foundations including CNNs, transformer-based networks, and fusion modules were systematically benchmarked. The comparative analysis uncovered how different design paradigms respond to modality shifts.

3. **Visual Diagnostics and Error Characterization:** The study employed qualitative visualizations alongside metric-based evaluation to identify frequent failure modes such as class misalignment, object merging, and boundary confusion. These insights contribute to a deeper understanding of model behavior under domain shift.

4. **Resource-Constrained Inference Testing:** The evaluation was conducted under hardware-limited scenarios using a CPU-only setup for two models and Google Colab for three others demonstrating the feasibility of running segmentation experiments without high-end GPUs.

5. **Scholarly Benchmark for Future Work:** As one of the few studies examining RGB-trained panoptic models on LiDAR inputs, this thesis provides a foundational benchmark and evaluation framework for future research in cross-modal segmentation and domain-adaptive computer vision.

6.3 Limitations of the Study

While this thesis offers valuable insights into the generalization capabilities of RGB-trained panoptic segmentation models on LiDAR-derived pseudo-RGB images, several limitations constrain the scope and generalizability of the findings. These limitations span methodological, computational, and dataset-related aspects, which are summarized below.

Zero-Shot Evaluation Only

The study exclusively focuses on *zero-shot* inference evaluating models without any retraining or fine-tuning on LiDAR-specific data. Although this setting highlights pure generalization capability, it inherently limits achievable performance. Domain adaptation techniques, such as adversarial training, pseudo-labeling, or style transfer, could potentially improve model robustness but were intentionally excluded from this evaluation.

2D LiDAR Projection Artifacts

The use of pseudo-RGB projections from LiDAR point clouds sacrifices valuable depth information and spatial granularity. These projections, while enabling compatibility with 2D convolutional models, result in the loss of fine-grained geometric cues that may otherwise assist segmentation. Furthermore, projection artifacts such as color banding and quantization may introduce noise that impacts inference qual-

ity.

Hardware and Computational Constraints

Due to the absence of dedicated GPU hardware, model evaluation was constrained to a CPU-only environment for two models and Google Colab for the remaining three. These resource limitations restricted input resolution, batch processing, and architectural complexity. Models requiring high memory or runtime optimizations had to be simplified, potentially affecting the comparability of results to benchmark standards.

Dataset Scope and Scene Diversity

The evaluation was based on a small-scale, manually curated dataset derived from publicly available LiDAR image samples. This dataset lacks diversity in scene types, lighting conditions, weather variations, and object dynamics. Consequently, the conclusions drawn may not generalize to complex real-world environments encountered in autonomous driving, robotics, or surveillance scenarios.

Limited Diagnostic Depth

Although qualitative visualizations were included to complement metric-based evaluation, the analysis did not explore deeper interpretability methods such as feature attribution, attention visualization, or class activation mapping. These tools could have provided further insights into the causes of segmentation failure and model decision behavior under domain shift.

Despite these limitations, the thesis contributes a foundational benchmarking protocol and highlights critical areas for improvement in future cross-modal segmentation research.

6.4 Final Remarks

This thesis has systematically evaluated the generalization capability of RGB-trained panoptic segmentation models when applied to LiDAR-generated pseudo-RGB imagery, highlighting critical challenges related to domain shift, modality mismatch, and computational constraints. While the selected architectures demonstrated partial effectiveness in handling cross-modal inputs, their overall performance underscored the inherent limitations of direct modality transfer without adaptation.

Several practical limitations; including zero-shot evaluation constraints, loss of structural information in LiDAR projections, hardware constraints, and dataset diversity have defined the scope of this research. Nevertheless, these constraints also highlight valuable pathways toward future advancements.

To effectively advance the state-of-the-art in cross-modal panoptic segmentation, the following priority recommendations are proposed:

- Explore **domain adaptation techniques**, such as adversarial training and style transfer, to reduce the performance gap between RGB and LiDAR modalities.
- Invest in developing **multi-modal fusion frameworks** that can dynamically integrate complementary information from LiDAR and RGB sensors.
- Create more diverse, annotated, and **purpose-built LiDAR panoptic segmentation datasets** to provide a robust foundation for evaluating future cross-modal methods.
- Emphasize research on **resource-efficient models**, incorporating model compression and efficient architectures suitable for real-world deployment in computationally limited environments.

- Integrate **temporal and sequential modeling approaches**, such as temporal transformers or recurrent neural networks, to enhance segmentation stability in dynamic settings.

By addressing these recommendations, future research can bridge existing gaps and significantly enhance the practical applicability and robustness of panoptic segmentation systems across diverse sensing platforms. Ultimately, this thesis contributes foundational insights and methodologies to facilitate further advancements in robust, efficient, and generalizable cross-modal segmentation systems for real-world deployment.

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn", in *ICCV*, 2017, pp. 2961–2969.

[2] G. Jocher, A. Chaurasia, J. Qiu, and A. Stoken, *Yolov5 by ultralytics*, `https://github.com/ultralytics/yolov5`, 2021.

[3] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic feature pyramid networks", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] J. Yang, Y. Chen, L. Zhao, and L. Wang, "A survey on panoptic segmentation: Past, present, and future", *Computer Vision and Image Understanding*, vol. 219, p. 103 422, 2022.

[5] Y. Jiang, A. Sharma, and T. D. Ng, "Practical panoptic segmentation: Balancing accuracy and inference for embedded applications", *arXiv preprint arXiv:2301.04700*, 2023.

[6] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context", in *ECCV*, 2014, pp. 740–755.

[7] Y. Liu, R. Chen, X. Li, *et al.*, "Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 11 249–11 259.

[8] B. Cheng, A. G. Schwing, and A. Kirillov, "Masked-attention mask transformer for universal image segmentation", in *CVPR*, 2022, pp. 1290–1299.

[9] A. Rosinol, J. Shi, T. Nguyen, and L. Carlone, "Kimera-multi: A system for multi-robot lidar–camera–imu localization and mapping", *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2345–2364, 2022.

[10] Y. Xu, Y. Wang, J. Yang, *et al.*, "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer", in *ECCV*, Springer, 2022, pp. 249–267.

[11] L. Porzi, S. Rota Bulò, P. Kontschieder, and E. Ricci, "Segnext: Rethinking convolutional attention design for semantic segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, Early Access.

[12] Car Magazine, *What is lidar?*, `https://www.carmagazine.co.uk/autonomous/what-is-lidar/`, Accessed: 2025-06-04, 2023.

[13] X. Wang, Y. Zhang, Y. Zhu, *et al.*, "Max-deeplab: A unified image segmentation model with patchwise tokenization", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[14] N. Garcia, T. Yu, S. Kim, and J. Chen, "Unified perception in autonomous driving: Panoptic segmentation and beyond", *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 1, pp. 118–132, 2023.

[15] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2020.

[16] A. Milioto, N. Vödisch, K. Petek, W. Burgard, and A. Valada, "Efficientlps: Efficient lidar panoptic segmentation", in *IEEE Transactions on Robotics*, vol. 37, 2021, pp. 1577–1592.

[17] L. Ma, R. Gupta, and S. Kumar, "Lidar-driven navigation and manipulation for mobile robotics", *Robotics and Autonomous Systems*, vol. 174, p. 104 271, 2023.

[18] J. Yang, Y. Chen, L. Zhao, and L. Wang, "A survey on panoptic segmentation: Past, present, and future", *Comput. Vis. Image Underst.*, vol. 219, p. 103 422, 2022.

[19] G. Nunes *et al.*, "Temporal consistent 3d lidar representation learning for semantic perception in autonomous driving", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[20] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] T. Qin, Z. Wang, and Z. Liu, "Robust lidar segmentation under adverse weather conditions", *IEEE Robotics and Automation Letters*, vol. 8, no. 4, pp. 2456–2463, 2023.

[22] H. Zhu, Y. Liu, and W. Wang, "Drone-based lidar surveying for environmental monitoring", in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 4567–4574.

[23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[24] 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)*, 2018.

[25] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn", in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.

[26] X. Zhou, Y. Huang, and Y. Fan, "Cross-modal domain generalization for panoptic segmentation", *Neurocomputing*, vol. 524, pp. 184–197, 2023.

[27] J. Li, Y. Zhang, Q. Liu, Y. Zhang, and Q. Wang, "A survey of deep learning techniques for lidar perception in autonomous driving", *IEEE Transactions on Intelligent Vehicles*, 2023. DOI: 10.1109/TIV.2023.3247481.

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs", *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

[29] N. Garcia, T. Yu, S. Kim, and J. Chen, "Unified perception in autonomous driving: Panoptic segmentation and beyond", *IEEE Trans. Intell. Trans. Syst.*, vol. 24, no. 1, pp. 118–132, 2023.

[30] X. Wang, T. Kong, C. Shen, and Y. Jiang, "Solo: Segmenting objects by locations", in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[31] A. Kirillov, Y. Wu, K. He, and R. Girshick, "Panoptic feature pyramid networks", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[32] Y. Xiong, R. Liao, H. Zhao, *et al.*, "Upsnet: A unified panoptic segmentation network", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] B. Cheng, A. G. Schwing, and A. Kirillov, "Mask2former for universal image segmentation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[34] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[35] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding", in *CVPR*, 2016, pp. 3213–3223.

[36] B. Cheng, M. D. Collins, Y. Zhu, *et al.*, "Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[37] K. Sofiiuk, O. Barinova, and A. Konushin, "Adaptis: Adaptive instance selection network", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[38] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[39] J. Muller, Z. Liu, and F. Yu, "Driving deployment: The road to real-time panoptic segmentation", *arXiv preprint arXiv:2205.12394*, 2022.

[40] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss, "Rangenet++: Fast and accurate lidar semantic segmentation", in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[41] G. Jocher *et al.*, *Yolov5*, `https://github.com/ultralytics/yolov5`, 2020.

[42] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, and A. Hogan, *Yolov5 by ultralytics*,
urlhttps://github.com/ultralytics/yolov5, GitHub repository, 2022.

[43] S. Z. Adal, *Panoptic-segmentation-eval-lidar-rgb*, `https://github.com/Sileshi-Adal/Panoptic-Segmentation-Eval-lidar-rgb`, Accessed: 4 July 2025, 2025.

[44] Sileshi-Adal, *Panoptic segmentation models evaluation*, `https://github.com/Sileshi-Adal/Panoptic-Segmentation-Eval-lidar-rgb`, GitHub repository, 2025.

[45] Y. Wang, Y. Sun, H. Wang, J. Shi, W. Liu, and J. Jia, "Pointaugmenting: Cross-modal augmentation for 3d object detection", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 794–11 803.

[46] G. Jocher *et al.*, *YOLOv5 by ultralytics*, `https://github.com/ultralytics/yolov5`, 2022.

[47] A. Athar *et al.*, "4d-former: Multimodal 4d panoptic segmentation", in *Conference on Robot Learning (CoRL)*, `https://arxiv.org/abs/2311.01520`, 2023.

[48] C. Michaelis, B. Mitzkus, R. Geirhos, *et al.*, "Benchmarking robustness in object detection: Autonomous driving when the weather turns bad", in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[49] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[50] H. Caesar, V. Bankiti, A. H. Lang, *et al.*, "Nuscenes: A multimodal dataset for autonomous driving", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[51] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, and A. Hogan, *Yolov5 by ultralytics*,
urlhttps://github.com/ultralytics/yolov5, GitHub repository, 2022.

[52] Y. Liu *et al.*, "Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[53] J. Zhang, J. Huang, X. Zhang, and S. Lu, "Unidaformer: Unified domain adaptive panoptic segmentation transformer via hierarchical mask calibration", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 11 227–11 237.

[54] S. Saha, L. Hoyer, A. Obukhov, D. Dai, and L. Van Gool, "Edaps: Enhanced domain-adaptive panoptic segmentation", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 11 238–11 248.

[55] E. Li, R. Razani, Y. Xu, and L. Bingbing, "Smac-seg: Lidar panoptic segmentation via sparse multi-directional attention clustering", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 332–11 341.

[56] M. Á. Hernandez Valencia, H. Carlos, and R. Aranda, "Evaluating the effectiveness of projection techniques for the semantic segmentation of lidar-captured point clouds", in *Recent Developments in Geospatial Information Sciences*, Springer, 2024, pp. 89–100.

[57] S. Vora, C. Hane, B. Drost, J. Gwak, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4604–4612.

[58] J. Zhang, S. Singh, and B. Chen, "Loam: Lidar odometry and mapping in real-time", in *Robotics: Science and Systems*, 2010.

[59] Y. Xiong, R. Liao, H. Zhao, *et al.*, "Upsnet: A unified panoptic segmentation network", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[60] B. Cheng, A. Schwing, and A. Kirillov, "Mask2former for universal image segmentation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[61] M. Carós, A. Just, S. Seguí, and J. Vitrià, "Self-supervised pre-training boosts semantic scene segmentation on lidar data", in *arXiv preprint arXiv:2309.02139*, 2023.

[62] Z. Zhang, Z. Zhang, Q. Yu, R. Yi, Y. Xie, and L. Ma, "Lidar-camera panoptic segmentation via geometry-consistent and semantic-aware alignment", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 12 345–12 354.

[63] Y. Chen, S. Zhao, C. Ding, L. Tang, C. Wang, and D. Tao, "Cross-modal & cross-domain learning for unsupervised lidar semantic segmentation", in *arXiv preprint arXiv:2308.02883*, 2023.

[64] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Towards medical xai", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[65] H. De Vries, I. Misra, M. Feldman, R. Krishna, J. C. Niebles, and L. Fei-Fei, "Fairness in computer vision: A survey", *arXiv preprint arXiv:2110.11843*, 2021.