Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

UP-Fuse: Uncertainty-guided LiDAR-Camera Fusion for 3D Panoptic Segmentation

Rohit Mohan*, Florian Drews†, Yakov Miron†‡, Daniele Cattaneo* and Abhinav Valada*
* University of Freiburg, Germany
Email: {mohan,valada,cattaneo}@cs.uni-freiburg.de
†Robert Bosch GmbH, Germany
Email: {florian.drews,yakov.miron}@de.bosch.com
‡University of Haifa

***Abstract*—LiDAR-camera fusion enhances 3D panoptic segmentation by leveraging camera images to complement sparse LiDAR scans, but it also introduces a critical failure mode. Under adverse conditions, degradation or failure of the camera sensor can significantly compromise the reliability of the perception system. To address this problem, we introduce UP-Fuse, a novel uncertainty-aware fusion framework in the 2D range-view that remains robust under camera sensor degradation, calibration drift, and sensor failure. Raw LiDAR data is first projected into the range-view and encoded by a LiDAR encoder, while camera features are simultaneously extracted and projected into the same shared space. At its core, UP-Fuse employs an uncertainty-guided fusion module that dynamically modulates cross-modal interaction using predicted uncertainty maps. These maps are learned by quantifying representational divergence under diverse visual degradations, ensuring that only reliable visual cues influence the fused representation. The fused range-view features are decoded by a novel hybrid 2D-3D transformer that mitigates spatial ambiguities inherent to the 2D projection and directly predicts 3D panoptic segmentation masks. Extensive experiments on Panoptic nuScenes, SemanticKITTI, and our introduced Panoptic Waymo benchmark demonstrate the efficacy and robustness of UP-Fuse, which maintains strong performance even under severe visual corruption or misalignment, making it well suited for robotic perception in safety-critical settings. We make the code and models publicly available at http://upfuse.cs.uni-freiburg.de.**

I. Introduction

3D panoptic segmentation [10, 23] unifies semantic and instance understanding of complex scenes, making it central for robotic perception, including autonomous driving [28]. LiDAR offers precise geometric measurements, but its sparsity and lack of appearance cues hinder reliable segmentation of small, distant, or geometrically similar objects. Incorporating dense, high-resolution camera data can mitigate these limitations by providing complementary texture and color information [39, 33, 40]. However, achieving effective fusion remains challenging, as the model has to learn not only *where* and *what* information to fuse, but also *when* to trust each modality. This reliability awareness is crucial under adverse conditions such as sensor failure, calibration errors, or visual domain shift, where camera inputs may become unreliable, as illustrated in Fig. 1.

Existing fusion approaches for 3D panoptic segmentation are ill-equipped to handle such scenarios and thus degrade sharply when a modality fails. A crucial gap lies in developing a fusion paradigm that can discern not only which features are relevant but also whether they are valid, enabling adaptive and context-dependent integration across modalities. To address these challenges, we propose UP-Fuse (**U**ncertainty-aware **P**anoptic **Fus**ion), an uncertainty-aware multi-modal fusion framework that leverages the range-view projection space. Our core contribution is an Uncertainty-Aware Fusion Module that jointly learns to evaluate cross-modal relevance and visual reliability. The module integrates deformable attention to identify informative visual features with an uncertainty head that estimates feature-level confidence. The fusion module is trained to associate feature patterns induced by camera sensor degradations (e.g., dropouts, occlusions, or out-of-domain distortions) with their impact on representational fidelity. The predicted uncertainty dynamically modulates the contribution of image features, enabling the network to attenuate unreliable cues while retaining informative ones. The fused representation is then processed by a Hybrid 2D-3D Panoptic Decoder that directly predicts 3D panoptic masks while alleviating projection ambiguities and boundary discontinuities inherent to the 360° range-view representation.

We extensively evaluate UP-Fuse on the Panoptic nuScenes, SemanticKITTI, and Waymo Open Dataset [38] benchmarks.

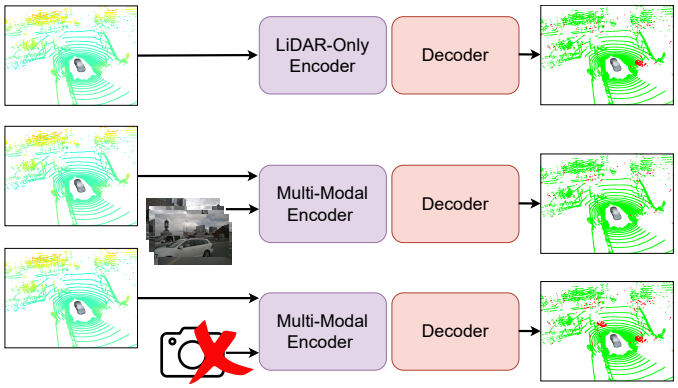


Fig. 1: Visualization of 3D panoptic segmentation: green indicates correct and red denotes errors. LiDAR-camera fusion significantly enhances segmentation over LiDAR-only methods, accurately detecting previously missed vehicles. However, camera sensor failure scenarios reveal a critical vulnerability, with fusion-based performance falling below LiDAR-only baselines and failing to detect previously identified objects. This highlights the crucial need for both relevance and reliability in multi-modal perception.

Since Waymo does not provide panoptic annotations, we generate them using the publicly available semantic segmentation and 3D bounding box labels, and establish Panoptic Waymo, a new multi-modal 3D panoptic benchmark with several baselines. The main contributions of this work are summarized as follows: (1) the UP-Fuse framework for uncertainty-aware multi-modal 3D panoptic segmentation, (2) an uncertainty-guided fusion module that enables reliability-aware integration of LiDAR and camera features, (3) a hybrid 2D-3D panoptic decoder that alleviates projection ambiguities in the 360° range-view representation, (4) a new multi-modal 3D panoptic benchmark for the Waymo Open Dataset with derived annotations and strong baselines, and (5) publicly released code and models upon acceptance.

II. Related Work

LiDAR Panoptic Segmentation: LiDAR Panoptic segmentation methods can be broadly categorized as top-down, bottom-up, or transformer-based approaches. Top-down methods [34, 43, 45] typically decompose the task into separate sub-tasks, using dedicated heads for instance and semantic segmentation. The instance head predicts object-level representations such as 2D instance masks from range-view (RV) projections [34] or 3D bounding boxes and instance masks in voxelized space [43, 45], while the semantic head produces dense per-point or per-pixel class predictions. The outputs from both heads are then fused through heuristic-based modules [10] to generate the final panoptic output. In contrast, bottom-up methods [8, 15, 31, 42, 47] jointly predict semantic labels and instance-related cues for all points within a single network. They typically regress offsets from each *thing* point toward its object center [8, 15, 27, 47] or learn affinities between points [31], followed by clustering or grouping to form complete instances [25].

Transformer-based approaches [6, 20, 41, 37] unify semantic and instance segmentation through a shared mask-classification framework, using learned queries to predict panoptic masks in an end-to-end manner. These methods differ in input representation, including point-based [37], sparse voxel [20], and RV projections [6]. The latter offers computational efficiency by transforming sparse 3D points into dense 2D grids, enabling the use of mature 2D segmentation architectures. However, lifting 2D predictions back into 3D heuristically, often introduces artifacts and struggles with occlusion [34, 6]. Our approach follows the RV transformer paradigm but incorporates a novel 2D-3D hybrid panoptic decoder, allowing more accurate and context-aware 3D panoptic segmentation.

Multi-Modal LiDAR Panoptic Segmentation: Panoptic-FusionNet [36] integrates camera features into a 3D backbone through multi-scale point-voxel-pixel correspondence tables, demonstrating the benefit of incorporating visual cues into LiDAR-based panoptic segmentation. Building on this idea, LCPS [46] introduces a voxel-space fusion pipeline that combines asynchronous temporal compensation, semantic-aware region alignment, and point-to-voxel propagation to enable content-aware cross-modal fusion. More recently, IAL [29] proposes a geometry-aware fusion architecture that performs cross-modal interaction through token-level attention between LiDAR and image representations. By incorporating modality-synchronized data augmentation during training, IAL encourages stronger geometric alignment between modalities. While these methods demonstrate the value of multi-modal fusion, Panoptic-FusionNet relies on static geometric associations [36] that fuse features without considering their contextual relevance. Both LCPS [46] and IAL [29] emphasize relevance-based fusion through content-aware feature selection or geometry-aware token interaction but lack an explicit mechanism to assess the reliability of visual features under adverse conditions. The reliability of camera features is closely related to aleatoric uncertainty, which captures irreducible noise arising from factors such as sensor imperfections, adverse lighting, or environmental conditions [9]. Existing fusion methods do not explicitly model this uncertainty, and traditional uncertainty estimation techniques such as Bayesian neural networks [2] or deep ensembles [13] are computationally expensive and primarily target epistemic uncertainty. In our work, we instead model aleatoric uncertainty at the feature level, enabling the network to quantify the reliability of camera-encoded features. UP-Fuse introduces an uncertainty-guided fusion module that integrates both relevance- and reliability-based fusion. By coupling cross-modal interaction with uncertainty awareness, our method adaptively balances the contributions of each modality. This design aligns with recent efforts to evaluate robustness under sensor degradation and failure [26, 35, 44].

III. UP-Fuse Architecture

Our proposed UP-Fuse architecture is illustrated in Fig. 2. The network first projects both LiDAR and camera data into a unified 2D range-view (RV) representation, enabling dense, pixel-aligned feature extraction (Sec. III-A1 and Sec. III-A2). These multi-scale, aligned features are then processed by our Uncertainty-Aware Fusion Module, which adaptively combines the modalities based on both cross-modal relevance and visual reliability (Sec. III-B). Finally, the resulting fused features are fed into our novel Hybrid 2D-3D Panoptic Decoder. This decoder generates direct 3D panoptic predictions for the original point cloud, while addressing projection ambiguities and 360° wrap-around discontinuities (Sec. III-C). We detail these components in the following sections.

A. Range-View Feature Representation

1) LiDAR Range-View Projection and Encoding: **Range-View Projection**: The raw LiDAR point cloud $\mathbf{P} \in \mathbb{R}^{N\times 4}$ consists of N points, where each point $\mathbf{p}_j = (x_j, y_j, z_j, i_j)$ contains the Cartesian coordinates (x_j, y_j, z_j) and the intensity value i_j of the reflected laser beam. The 3D point cloud is projected into a dense 2D spherical representation, referred to as the range-view image $\mathbf{L}_{\text{RV}}$, with spatial resolution $H \times W$. Each 3D point $\mathbf{p}_j$ is mapped to pixel coordinates (u_j, v_j) by computing its horizontal azimuth angle θ_j and vertical elevation angle ϕ_j. The range is defined as $r_j = \sqrt{x_j^2 + y_j^2 + z_j^2}$, and

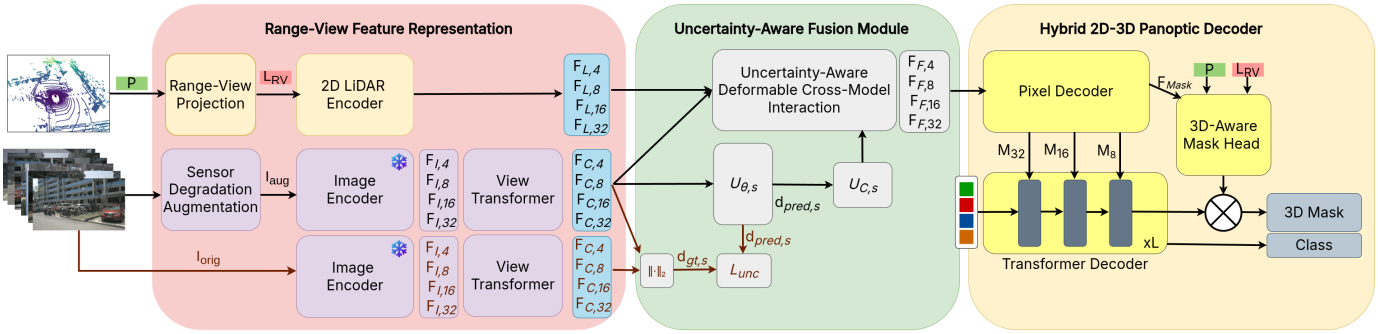


Fig. 2: Illustration of the proposed **UP-Fuse** architecture. LiDAR and multi-view camera images are fused onto a shared space of range-view feature representations. The Uncertainty-Aware Fusion Module adaptively integrates modalities via uncertainty-weighted deformable cross-modal interaction to attenuate unreliable visual cues. Finally, a Hybrid 2D-3D Panoptic Decoder generates 3D predictions. Paths and blocks shown in brown are used only during training.

the angular components are given by:

$$\theta_j = -\arctan(y_j, x_j), \quad \phi_j = \arcsin\left(\frac{z_j}{r_j}\right). \tag{1}$$

The angles are normalized according to the LiDAR sensor's field of view. Here, f_{up} and f_{down} denote the vertical angular limits above and below the horizontal plane, while f_{left} and f_{right} represent the horizontal limits to the left and right of the forward axis. The total spans are $f_h = |f_{\text{left}}| + |f_{\text{right}}|$ and $f_v = |f_{\text{up}}| + |f_{\text{down}}|$, which are used to scale the angles to the image dimensions:

$$\begin{bmatrix} u_j \\ v_j \end{bmatrix} = \begin{bmatrix} \left(\frac{\theta_j + |f_{\text{left}}|}{f_h}\right) W \\ \left(1 - \frac{\phi_j + |f_{\text{down}}|}{f_v}\right) H \end{bmatrix}. \tag{2}$$

The resulting range-view image $\mathbf{L}_{\text{RV}}$ is constructed by assigning to each pixel (u_j, v_j) the corresponding point features [7]:

$$\mathbf{L}_{\text{RV}}(u_j, v_j) = [r_j, z_j, i_j]. \tag{3}$$

To handle occlusions where multiple points map to the same pixel, all points are sorted by decreasing range and projected sequentially, ensuring that only the nearest point (with minimum r_j) is retained to maintain visibility consistency. This projection defines a mapping $\mathcal{P}_{\text{3D}\to\text{RV}} : \mathbf{p}_j \to (u_j, v_j)$, and we store its inverse $\mathcal{P}_{\text{RV}\to\text{3D}}$, which is later used in the 2D-3D hybrid decoder (see Sec. III-C) to lift 2D features back into 3D space.

LiDAR Feature Encoding: The projected range-view image $\mathbf{L}_{\text{RV}}$ is processed by a 2D LiDAR encoder to extract hierarchical feature representations. We employ a Swin Transformer [18] backbone that effectively captures both local geometric structures and global spatial dependencies. The encoder produces feature maps at four resolution levels, denoted as $\mathbf{F}_{L,4}$, $\mathbf{F}_{L,8}$, $\mathbf{F}_{L,16}$, and $\mathbf{F}_{L,32}$, corresponding to output strides of 4, 8, 16, and 32 relative to the input resolution. These multi-scale features are subsequently fed to the Uncertainty-Aware Fusion Module (see Sec. III-B).

2) Image Encoding and View Transformation: **Image Encoding**: Each of the M calibrated camera views with spatial dimensions $C_H \times C_W$ is independently processed by an image encoder to capture rich visual cues. We adopt a Swin Transformer [18] backbone that outputs multi-scale feature maps at four hierarchical stages, denoted as $\mathbf{F}_{I,4}$, $\mathbf{F}_{I,8}$, $\mathbf{F}_{I,16}$, and $\mathbf{F}_{I,32}$, corresponding to feature strides of 4, 8, 16, and 32 relative to the original image resolution. The backbone is pre-trained and kept frozen during training. The extracted features from all camera views are then passed to the view transformation module for geometric alignment with the LiDAR range-view representation.

View Transformation: To enable spatial alignment between image and LiDAR representations, we establish a dense correspondence between the multi-view image planes and the LiDAR range-view image. This is accomplished by constructing a pseudo point cloud from the camera views and then projecting it into the LiDAR range-view space using the same mapping described in Sec. III-A1. Specifically, the LiDAR point cloud $\mathbf{P}$ is first projected into each of the M camera views to generate M sparse depth maps $\mathcal{D}_{\text{sparse}}$. These are densified using the depth completion method of [12], resulting in dense depth maps $\mathcal{D}_{\text{dense}}$. Each pixel (i, j) in view m is then back-projected into a 3D point expressed in the LiDAR coordinate frame:

$$\mathbf{p} = \mathbf{T}_m^{-1} \begin{bmatrix} \mathcal{D}_{\text{dense}}(i,j) \cdot \mathbf{I}_m^{-1} \cdot [i, j, 1]^T \\ 1 \end{bmatrix}, \tag{4}$$

where $\mathbf{I}_m$ is the 3×3 intrinsic matrix of camera m, and $\mathbf{T}_m$ is the 4×4 extrinsic matrix transforming from the camera to LiDAR frame. This produces a dense, camera-derived point cloud $\mathbf{P}_{\text{cam}}$ that aggregates information from all M views. We then project $\mathbf{P}_{\text{cam}}$ into the $H \times W$ range-view using the same projection formulation as in eq. (1) and eq. (2), yielding a dense mapping $\mathcal{M} : (m, i, j) \to (u, v)$ from each camera pixel to a corresponding range-view pixel. This mapping enables warping of the multi-scale image features $\mathbf{F}_{I,s}$ into the range-view domain to obtain RV-aligned image features $\mathbf{F}_{C,s}$ at each scale $s \in \{4, 8, 16, 32\}$. If multiple pixels project to the same RV location, their features are averaged. The aligned features serve as multi-modal counterparts to the LiDAR features $\mathbf{F}_{L,s}$ and are forwarded to the Uncertainty-Aware Fusion Module (see Sec. III-B).

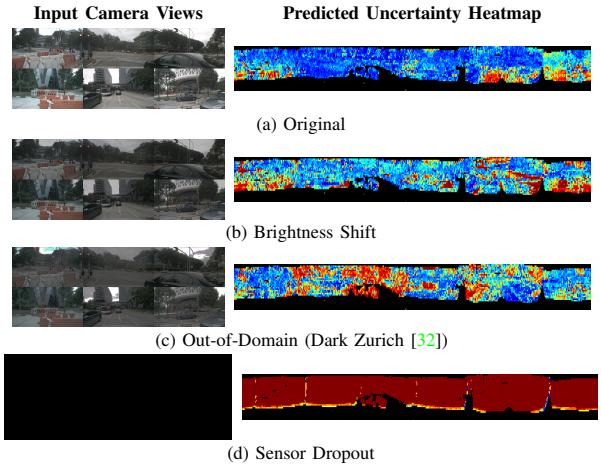


Fig. 3: Illustration of our uncertainty module on a Panoptic nuScenes sample. The right column shows predicted uncertainty under increasing synthetic degradations. The jet colormap marks low uncertainty in **blue** and high uncertainty in **red**. Mild distortions (b) keep uncertainty low, while strong distortions (c) and sensor dropout (d) produce high uncertainty. Black regions indicate areas without a camera to range-view (RV) mapping.

B. Uncertainty-Aware Fusion Module

At each scale $s \in \{4, 8, 16, 32\}$, the fusion module combines range-view (RV) aligned camera features $\mathbf{F}_{C,s}$ and LiDAR features $\mathbf{F}_{L,s}$ through uncertainty-guided cross-modal interaction to produce fused features $\mathbf{F}_{F,s}$.

Uncertainty Quantification of Camera Features: We model aleatoric uncertainty in camera features as instability under input degradations. Reliable features remain consistent across mild corruptions, whereas features from degraded inputs (e.g., underexposed or overexposed images) exhibit higher variability. To learn this relationship, we train a lightweight 3-layer MLP, $\mathcal{U}_{\theta,s}$, to regress feature instability. During training, for each image $\mathbf{I}_{\text{orig}}$, we sample a corrupted counterpart $\mathbf{I}_{\text{aug}}$ using a diverse set of non-spatial augmentations such as brightness and contrast shifts, sensor dropout, and cross-domain histogram matching. The latter adjusts image color statistics to match those from Cityscapes [4], COCO [17], and Dark Zurich [32], thereby introducing out-of-domain photometric variations. With probability of 0.5, we set $\mathbf{I}_{\text{aug}} = \mathbf{I}_{\text{orig}}$ to provide stable zero-uncertainty samples. Both $\mathbf{I}_{\text{orig}}$ and $\mathbf{I}_{\text{aug}}$ are processed by the frozen encoder and view transformation (Sec. III-A2) to obtain multi-scale, RV-aligned camera features $\mathbf{F}^{\text{orig}}_{C,s}$ and $\mathbf{F}^{\text{aug}}_{C,s}$.

We define the ground-truth instability $\mathbf{d}_{\text{gt},s}$ as the L2 norm between original and augmented features at each spatial location (u, v):

$$\mathbf{d}_{\text{gt},s}(u,v) = \left\| \mathbf{F}^{\text{orig}}_{C,s}(u,v) - \mathbf{F}^{\text{aug}}_{C,s}(u,v) \right\|_2 . \tag{5}$$

The MLP $\mathcal{U}_{\theta,s}$ is trained to predict this instability from the augmented features:

$$\mathbf{d}_{\text{pred},s} = \mathcal{U}_{\theta,s}(\mathbf{F}^{\text{aug}}_{C,s}). \tag{6}$$

It learns to associate corrupted feature patterns with their corresponding magnitude of instability. The network is optimized using a Huber loss, which provides robustness to outliers:

$$\mathcal{L}_{\text{unc}} = \frac{1}{N} \sum_{(u,v),s} \mathcal{L}_\delta(\mathbf{d}_{\text{pred},s}(u,v) - \mathbf{d}_{\text{gt},s}(u,v)) , \tag{7}$$

where N is the number of spatial locations and $\mathcal{L}_\delta$ is the Huber loss function with threshold δ, set to 1.0:

$$\mathcal{L}_\delta(a) = \begin{cases} 0.5a^2 & \text{if } |a| \le \delta, \\ \delta(|a| - 0.5\delta) & \text{otherwise.} \end{cases} \tag{8}$$

Finally, the predicted instability $\mathbf{d}_{\text{pred},s}$ is transformed into a probabilistic aleatoric uncertainty score $\mathbf{U}_{C,s} \in [0, 1]$:

$$\mathbf{U}_{C,s} = 1 - \exp(-\mathbf{d}_{\text{pred},s}). \tag{9}$$

This uncertainty score modulates the cross-modal interaction within the fusion module. Examples of augmentations and corresponding uncertainty maps for s=4 are shown in Fig. 3 and further discussed in the supplementary material.

Uncertainty-Guided Fusion: We propose an uncertainty-guided fusion mechanism that adaptively integrates LiDAR and camera features through cross-modal interaction. At each scale s, given LiDAR features $\mathbf{F}_{L,s}$ and camera features $\mathbf{F}_{C,s}$, fusion is achieved via a deformable attention [48] that aggregates spatially aligned visual information conditioned on LiDAR queries:

$$\mathbf{F}_{A,s} = \sum_{l=1}^{L} \sum_{p=1}^{P} A_{l,p}\, \tilde{\mathbf{F}}^{(l,p)}_{C,s}, \tag{10}$$

where $A_{l,p}$ are query-dependent attention weights measuring the relevance of each sampled camera feature $\mathbf{F}^{(l,p)}_{C,s}$ from the l-th level and p-th sampling point. We set L=1 attention level and P=4 sampling points per query. The uncertainty-modulated camera features are defined as

$$\tilde{\mathbf{F}}^{(l,p)}_{C,s} = (1 - \mathbf{U}_{C,s}) \odot \mathbf{F}^{(l,p)}_{C,s}, \tag{11}$$

where $\mathbf{U}_{C,s}$ denotes the predicted uncertainty map and $\odot$ represents element-wise multiplication. This formulation integrates deformable attention with uncertainty-driven weighting, allowing each LiDAR query to attend selectively to spatially relevant and reliable visual evidence. Regions with high uncertainty (e.g., overexposed or occluded areas) are attenuated, preventing them from dominating the fusion. The attended features are fused with the LiDAR representation as

$$\mathbf{F}_{F,s} = \mathbf{F}_{L,s} + \mathbf{F}_{A,s}. \tag{12}$$

This fusion preserves the geometric accuracy of LiDAR while enriching it with semantically reliable visual context. The resulting fused features $\mathbf{F}_{F,s}$ are subsequently passed to the 2D-3D hybrid decoder.

C. Hybrid 2D-3D Panoptic Decoder

The final component of our network is the hybrid 2D-3D panoptic decoder, which bridges 2D range-view features and 3D point-cloud outputs. It tackles two main challenges. First, directly predicting a 2D segmentation map and lifting it to 3D via $\mathcal{P}_{\text{RV}\to\text{3D}}$ is unreliable, since multiple 3D points project

to the same 2D pixel (u, v). Consequently, 2D predictions may propagate to occluded points, introducing label ambiguity. Second, the 360° LiDAR projection causes objects near the horizontal boundary (e.g., 0°/360°) to split across the 2D grid. A 2D-only decoder, unaware of this geometric continuity, would interpret them as separate instances, leading to fragmented predictions. To address these challenges, we propose a hybrid decoder that interleaves 2D feature processing with a 3D-aware unprojection mechanism, enabling explicit learning of 360° object continuity and resolving spatial ambiguity in 3D. Our approach follows the Mask2Former [3] paradigm, consisting of a pixel decoder and a transformer decoder.

2D Pixel and Transformer Decoder: The fused features $\mathbf{F}_{F,s}$ at scales $s \in \{4, 8, 16, 32\}$ are first processed by the pixel decoder, a multi-scale deformable attention-based feature pyramid [48]. It outputs (1) multi-scale feature maps $\mathbf{M}_s$ at $\{8, 16, 32\}$ resolutions and (2) a high-resolution mask feature map $\mathbf{F}_{\text{mask}}$ at scale 4 used for mask prediction. A Transformer decoder with a 3-layer block structure, repeated L times, then processes N_q learnable object queries, attending to $\mathbf{M}_s$ with each layer focusing on a different scale in alternation. Within each 3-layer block, the intermediate layers use the standard 2D mask head [3] for auxiliary supervision, while the block's final layer employs our 3D-aware mask head to generate predictions, capturing spatial continuity in the point cloud domain.

3D-Aware Mask Head: This head generates a per-point feature vector $\mathbf{f}_{\text{point},j}$ for each point $\mathbf{p}_j$ in the original point cloud $\mathbf{P}$. This is achieved through a learnable, geometry-aware feature aggregation process. The module takes three inputs: (1) the high-resolution feature map $\mathbf{F}_{\text{mask}}$, (2) the range channel of the 2D range image $\mathbf{L}_{\text{RV}}$ (from Sec. III-A1), which stores the minimum range $r_{(u,v)}$ at each pixel, and (3) the N 3D points, each with its true 3D range $r_{j,\text{true}}$ and its $\mathcal{P}_{\text{RV}\to\text{3D}}$ mapping from pixel (u_j, v_j). For each point $\mathbf{p}_j$, a $K \times K$ search window is defined in the range channel centered at (u_j, v_j). We then compute the absolute range difference $|r_{j,\text{true}} - r_{(u',v')}|$ between the point's true 3D range and the range value of every pixel (u', v') in the $K \times K$ window. The K pixels in this window with the smallest range difference are selected as 3D-consistent neighbors. The corresponding K features from $\mathbf{F}_{\text{mask}}$ are gathered, concatenated, and fused through a lightweight 2-layer MLP to yield a context-aware feature $\mathbf{f}_{\text{point},j} \in \mathbb{R}^D$, where D is the feature channel dimension. The resulting per-point feature map $\mathbf{F}_{\text{point}} \in \mathbb{R}^{D\times N}$ is thus a spatially consistent, 3D-aware representation, constructed from aggregation over all N points, that mitigates label bleeding and instance fragmentation.

Panoptic Prediction and Loss: The final transformer decoder layer produces the refined query embeddings $\mathbf{Q}_{\text{embed}}$ and the per-point feature map $\mathbf{F}_{\text{point}}$. Two lightweight heads generate the final outputs. The *class head* applies a linear projection on $\mathbf{Q}_{\text{embed}}$ to predict class logits $\mathbf{C} \in \mathbb{R}^{N_q\times(C_{\text{th}}+C_{\text{st}}+1)}$, where C_{th} and C_{st} denote the number of thing and stuff classes, respectively, and the additional class corresponds to the no-object label. The *mask head* transforms $\mathbf{Q}_{\text{embed}}$ into mask embeddings $\mathbf{E}_{\text{mask}} \in \mathbb{R}^{N_q\times D}$ through an MLP, and computes the 3D mask logits via a dot product with the per-point features:

$$\mathbf{M}_{\text{3D}}(q, j) = \mathbf{E}_{\text{mask}}(q, :) \cdot \mathbf{F}_{\text{point}}(:, j). \tag{13}$$

The decoder thus outputs the class logits $\mathbf{C}$ and per-point mask logits $\mathbf{M}_{\text{3D}} \in \mathbb{R}^{N_q\times N}$. Following [3], we employ a set-based loss computed on a subset of sampled points for efficiency. For each query, N_p points are selected using a mixture of importance sampling (points with highest uncertainty) and random sampling. Bipartite matching assigns ground-truth instances to queries, and the final loss is defined as

$$\mathcal{L}_{\text{panoptic}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}, \tag{14}$$

where $\mathcal{L}_{\text{cls}}$, $\mathcal{L}_{\text{mask}}$, and $\mathcal{L}_{\text{dice}}$ are the classification, mask, and Dice losses, respectively. This formulation enables supervision over dense 3D point predictions within the hybrid 2D-3D framework.

IV. Experimental Evaluation

In this section, we present a comprehensive evaluation of UP-Fuse. We first introduce the new Panoptic Waymo benchmark and the evaluation metrics in Sec. IV-A. We then detail our implementation in Sec. IV-B. Finally, we present comprehensive evaluation results, including benchmarking (Sec. IV-C), robustness analysis (Sec. IV-D), ablation studies (Sec. IV-E), and qualitative results (Sec. IV-F) .

A. Datasets and Evaluation Metrics

We evaluate UP-Fuse on Panoptic nuScenes [5], SemanticKITTI [1], and our newly introduced Panoptic Waymo benchmark derived from the Waymo Open Dataset [38].

Panoptic Waymo: While Panoptic nuScenes offers a 360° benchmark, its 32-beam LiDAR is comparatively sparse. To provide a more challenging, high-resolution benchmark for fine-grained fusion, we introduce panoptic annotations for the Waymo Open Dataset (WOD) [38]. WOD includes a dense 64-beam LiDAR and five high-resolution cameras, covering more than 180° of the scene, which makes it well-suited for multi-modal evaluation. The dataset officially provides 3D semantic segmentation and 3D bounding box labels. We generate our Panoptic Waymo ground truth by following the protocol of Panoptic nuScenes [5] to merge these annotations. We retain the original 798 training scenes and 202 validation scenes, resulting in a benchmark with 15 *stuff* classes and 6 *thing* classes. A detailed statistical overview is presented in the supplementary material. Since WOD is approximately 3× denser than nuScenes, we exclude thing instances with fewer than 50 points (15 in Panoptic nuScenes) to match the higher point cloud fidelity.

Evaluation Metrics: We evaluate our model using standard panoptic segmentation metrics [10]: Panoptic Quality (PQ), Segmentation Quality (SQ), and Recognition Quality (RQ). Furthermore, we report the modified Panoptic Quality metric ($\text{PQ}^{\dagger}$) introduced in [30]. All metrics are reported as averages over all classes, with additional breakdowns for *thing* (PQ^{th}) and *stuff* (PQ^{st}) categories.

TABLE I: Comparison of 3D panoptic segmentation performance on the Panoptic nuScenes validation set.

Method	Modality	PQ	PQ†	SQ	RQ	PQ^{th}	PQ^{st}	FPS
CPSeg HR [14]	L	71.1	75.6	82.5	85.5	71.5	70.6	—
Panoptic-FusionNet [36]	L	72.7	75.4	86.4	84.8	71.2	75.1	—
LCPS [46]	L	72.9	77.6	88.4	82.0	72.8	73.0	—
Panoptic-PHNet [15]	L	74.7	77.7	88.2	84.2	74.0	75.9	—
PUPS [37]	L	74.7	77.3	89.4	83.3	75.4	73.6	—
UP-Fuse (Ours)	L	74.9	78.2	87.8	85.1	75.1	74.5	—
CFNet [16]	L	75.1	78.0	88.8	84.6	74.8	76.6	—
P3Former [41]	L	75.9	78.9	89.7	84.7	76.9	75.4	—
CenterLPS [21]	L	76.4	79.2	86.2	88.0	77.5	74.6	—
IAL [29]	L	77.0	79.6	90.2	85.1	77.8	75.7	—
Panoptic-FusionNet [36]	LC	77.2	79.3	87.8	87.2	77.5	76.2	2.5
LCPS [46]	LC	79.8	84.0	89.8	88.5	82.3	75.6	1.7
IAL [29]	LC	80.3	82.8	91.0	87.9	—	—	0.9
UP-Fuse (Ours)	LC	80.7	84.3	90.3	89.0	83.0	77.0	**5.7**
IAL-PieAug [29]	LC	**82.3**	**84.7**	**91.5**	**89.7**	**85.3**	**77.3**	0.9

B. Implementation Details

Architecture and Pre-training: We use Swin-B [18] backbones for both LiDAR and camera. The camera encoder is pre-trained using a camera-only variant of our model, where LiDAR is used solely for view transformation. Its weights remain frozen during all multi-modal experiments. Our hybrid 2D-3D panoptic decoder operates with $N_q = 300$ queries and $L = 2$ blocks. The 3D-aware mask head aggregates features from $K = 5$ neighbors (see Sec. III-C). Further details are presented in the supplementary material.

Training Details: All models are optimized using AdamW [19] with a base learning rate of 10^{-4}. We apply standard spatial augmentations, including rotation, scaling, and horizontal flipping, to both modalities. For uncertainty training, we use diverse non-spatial camera corruptions such as photometric changes, sensor dropout, and out-of-domain histogram matching. The full list is detailed in the supplementary material. On Panoptic nuScenes, we use a range-view input of 32×1024 (resized to 256×2048), a batch size of 16, and train for 80 epochs with a $10\times$ learning rate decay applied at epochs 72 and 76. On Panoptic Waymo, we use a range-view input of 64×2560 (resized to 256×4096), a batch size of 8, and train for 48 epochs with learning rate decays at epochs 40 and 44. For SemanticKITTI, we use a range-view input of 64×2048 (resized to 256×4096), a batch size of 4, and train for 24 epochs with $10\times$ learning rate decays at epochs 16 and 20. The loss weights $(\lambda_{\text{cls}}, \lambda_{\text{dice}}, \lambda_{\text{mask}}, \lambda_{\text{unc}})$ are set to $(5, 5, 100, 1)$ for Panoptic nuScenes and $(2, 5, 50, 1)$ for both Panoptic Waymo and SemanticKITTI. During training, we sample $N_p = 12544$ points for Panoptic nuScenes and $N_p = 25088$ points for Panoptic Waymo and SemanticKITTI. Camera images are resized to 256×704 pixels for Panoptic nuScenes similar to [24] and Panoptic Waymo, and to 360×640 pixels for SemanticKITTI. Lastly, FPS is measured on a single NVIDIA A40 GPU with batch size 1 following the evaluation protocol used by IAL [29].

C. Benchmarking Results

We evaluate our UP-Fuse approach against published state-of-the-art methods. We first report results on the primary Panoptic nuScenes benchmark, followed by SemanticKITTI, and finally on our new Panoptic Waymo benchmark.

Results on Panoptic nuScenes: As shown in Tab. I, UP-Fuse (LC) reaches 80.7% PQ on the validation split, outperforming multi-modal baselines such as LCPS [46] by 0.9% in PQ and Panoptic-FusionNet [36] by 3.5% in PQ. It further yields a gain of 5.8% in PQ over our strong LiDAR-only (L) variant (74.9%), which removes the image encoder and uncertainty-driven fusion module. IAL [29], when trained with its proposed modality-synchronized augmentation (IAL-PieAug), achieves the highest absolute performance at 82.3% PQ. Without PieAug, the base IAL fusion architecture attains 80.3% PQ, 0.4% lower in PQ than UP-Fuse under the same augmentation protocol. Crucially, the competitive performance of UP-Fuse is achieved with substantially higher efficiency: it operates at 5.7 FPS, approximately $6\times$ faster than IAL (0.9 FPS), $3\times$ faster than LCPS, and over $2\times$ faster than Panoptic-FusionNet. This demonstrates that the previously unexplored unified 2D range-view space for multi-modal 3D panoptic segmentation enables an effective trade-off between accuracy and runtime efficiency. Beyond standard benchmarks, we further show in Sec. IV-D that our fusion framework exhibits improved robustness under sensor dropout, calibration drift, and visual domain shift.

The metric breakdown shows that the major contribution comes from *thing* classes, with UP-Fuse achieving a 7.9% boost in PQ^{th} compared to the LiDAR-only variant, with an additional 2.5% gain in PQ^{st} for *stuff* classes. This suggests that our uncertainty-guided fusion reliably exploits visual texture cues to disambiguate sparse and confusing *thing* instances in LiDAR data, while also better delineating *stuff* regions with similar geometry. A similar trend is observed on the test set (Tab. II), where our model achieves 81.1% in the PQ score.

Results on SemanticKITTI: We further validate our approach on the SemanticKITTI validation set (Tab. III), where the frontal camera setup limits visual overlap with LiDAR data. Under this constrained sensing setup, IAL-PieAug [29] achieves the highest PQ at 63.1%, while UP-Fuse (LC) attains 61.8% PQ, outperforming LCPS [46] by 2.8% in PQ. This indicates that UP-Fuse can effectively exploit complementary visual cues when available, even under limited camera coverage.

Results on Panoptic Waymo bridges the gap between Panoptic nuScenes, with full surround cameras and sparse LiDAR, and SemanticKITTI, with frontal cameras and denser LiDAR. As shown in Tab. IV, UP-Fuse (LC) achieves 60.9% PQ, a gain of 3.5% in PQ over its LiDAR-only baseline. This performance also surpasses other multi-modal methods, outperforming IAL [29] by 0.5% in PQ. The training protocols for all baselines are provided in the supplementary material.

D. Robustness Analysis

Real-world robotic deployment requires resilience to three failure modes [11]: complete sensor loss (dropout), mechanical misalignment (calibration drift), and environmental degradation (visual domain shift). We evaluate UP-Fuse under these conditions to assess its reliability for safety-critical robotic

TABLE II: Comparison of 3D panoptic segmentation results on the Panoptic nuScenes test set.

Method	Modality	PQ	PQ†	SQ	RQ	PQ^{th}	PQ^{st}
MaskPLS [20]	L	61.1	64.3	86.8	68.5	54.3	72.4
EfficientLPS [34]	L	62.4	66.0	83.7	74.1	57.2	71.1
Panopitc-PolarNet [47]	L	63.6	67.1	84.3	75.1	59.0	71.3
LCPS [46]	L	72.8	76.3	88.6	81.7	72.4	73.5
CPSeg [14]	L	73.2	76.3	88.1	82.7	72.9	74.0
Panoptic-PHNet [47]	L	80.1	82.8	91.1	87.6	82.1	76.6
LCPS [46]	LC	79.5	82.3	90.3	87.7	81.7	75.9
UP-Fuse (Ours)	LC	81.1	83.4	91.3	88.5	83.6	76.9
IAL-PieAug [29]	LC	**82.0**	**84.3**	**91.6**	**89.3**	**84.8**	**77.5**

TABLE III: Comparison of 3D panoptic segmentation results on the SemanticKITTI validation set.

Method	Modality	PQ	PQ†	SQ	RQ
LCPS [46]	L	55.7	65.2	65.8	74.0
EfficientLPS [34]	L	59.2	65.1	69.8	75.0
UP-Fuse (Ours)	L	60.4	67.8	74.9	71.7
Panoptic-PHNet [15]	L	61.7	—	—	—
IAL [29]	L	62.0	65.1	76.0	71.9
CenterLPS [21]	L	62.1	67.0	72.0	**80.7**
P3Former [41]	L	62.6	66.2	72.4	76.2
GP-S3Net [31]	L	63.3	**71.5**	81.4	75.9
PUPS [37]	L	**64.4**	68.6	**81.5**	74.1
LCPS [46]	LC	59.0	68.8	68.9	79.8
UP-Fuse (Ours)	LC	61.8	69.1	75.2	73.1
IAL-PieAug [29]	LC	63.1	66.3	81.4	72.9

perception. Additionally, Fig. 6 qualitatively compares IAL and UP-Fuse under these failure modes.

Robustness to Sensor Dropout: We evaluate robustness under complete camera failure, with results reported in Tab. V. To ensure a fair comparison, all multi-modal baselines are retrained from scratch using the same camera dropout augmentations applied in UP-Fuse. In the full modality setting (LC), all fusion-based methods achieve a substantial PQ improvement over their LiDAR-only counterpart (L), with LCPS [46] obtaining the largest gain of 6.4% in PQ, followed by 5.8% from our UP-Fuse. The critical evaluation setting for robustness is when the camera input is removed at inference (L*). In this regime, the limitations of fusion strategies that focus solely on relevance rather than reliability become evident. Panoptic-FusionNet [36], LCPS [46], and IAL [29] exhibit substantial degradation, with reductions of 5.0%, 4.2%, and 4.6% in PQ, respectively, all falling below their own L-only baselines. UP-Fuse demonstrates a markedly different response. Guided by its uncertainty-driven fusion module, the model suppresses corrupted image features and relies more heavily on the LiDAR representation. Its performance decreases by only 1.2% in PQ, remaining close to its strong L-only baseline. These findings indicate that robustness cannot be obtained through data augmentation alone and instead requires an architecture capable of adaptively modulating unreliable inputs.

TABLE IV: Comparison of 3D panoptic segmentation results on the Panoptic Waymo validation set.

Method	Modality	PQ	PQ†	SQ	RQ	PQ^{th}	PQ^{st}
Panoptic-FusionNet [36]	L	53.9	63.2	77.4	68.2	55.1	53.4
LCPS [46]	L	54.9	63.1	77.8	69.2	58.2	53.6
EfficientLPS [34]	L	56.3	64.5	78.3	70.6	59.7	54.9
UP-Fuse (Ours)	L	57.4	65.1	78.8	71.5	61.3	55.8
IAL [29]	L	58.2	66.1	79.2	72.1	62.5	56.4
P3Former [41]	L	59.3	67.4	79.3	73.5	63.8	57.5
Panoptic-FusionNet [36]	LC	56.2	64.8	78.5	70.9	58.7	55.2
LCPS [46]	LC	58.9	66.7	79.3	72.8	63.1	57.2
IAL [29]	LC	60.4	68.2	79.4	74.8	65.3	58.4
UP-Fuse (Ours)	LC	**60.9**	**69.0**	**79.6**	**75.4**	**66.7**	**58.6**

Robustness to Calibration Drift: We evaluate geometric robustness by simulating mechanical calibration drift through random rotational perturbations $\theta \in [0°, 5°]$ applied to the LiDAR-camera extrinsics at inference. As shown in Fig. 4, all multi-modal baselines exhibit increasing performance degradation as misalignment increases. While IAL [29] improves over Panoptic-FusionNet [36] and LCPS [46], it still incurs an 8.3% drop in PQ at 5°. In contrast, UP-Fuse limits the degradation to just 4.4%. These results confirm the adaptive behavior of UP-Fuse: as cross-modal alignment deteriorates, the uncertainty-driven fusion module progressively reduces the influence of unreliable image features.

Robustness to Visual Domain Shift: We further assess robustness under a visual domain shift from daytime to nighttime conditions. To this end, all models are retrained using the 630 daytime scenes from Panoptic nuScenes and evaluated on 15 held-out nighttime scenes. This setting represents a naturally occurring domain shift, where camera observations are substantially degraded while LiDAR data remain reliable. As shown in Tab. VI, all baseline fusion methods are highly sensitive to this shift. Having learned to rely on visual cues under daytime illumination, Panoptic-FusionNet [36], LCPS [46], and IAL [29] incorrectly fuse dark and uninformative image features, resulting in drops of 3.1%, 2.7% and 2.1% in PQ relative to their LiDAR-only baselines (L), respectively. In contrast, UP-Fuse demonstrates strong architectural robustness. Its uncertainty-driven fusion module identifies regions where visual features are unreliable and selectively incorporates only the complementary cues that remain informative. As a result, UP-Fuse maintains its performance and achieves a 0.1% improvement in PQ, effectively reverting to its robust L-only baseline and avoiding any significant degradation. We note that the night scene evaluation lacks four *thing* classes (bus, construction vehicle, trailer, and traffic cone).

E. Ablation Study

Architectural Ablation: As shown in Tab. VII, we perform a bottom-up analysis of our architecture. Our first baseline (M1) adapts a 2D decoder [3] to the range-view and uses a KNN-based post-processing [22] to lift 2D predictions to 3D. Replacing this with our Hybrid 2D-3D Panoptic Decoder (M2) provides a significant 2.1% gain in PQ (74.9% vs. 72.8% PQ), confirming it successfully alleviates projection ambiguities in the 360° range-view. We then introduce the camera modality with concatenation fusion. We find that using dense depth

TABLE V: Robust performance comparison on the Panoptic nuScenes validation set under missing camera modality conditions. ΔPQ denotes the change compared to the corresponding LiDAR-only variant.

Method	Modality	PQ	PQ^{th}	PQ^{st}	ΔPQ
Panoptic-FusionNet [36]	L	72.7	71.2	75.1	—
LCPS [46]	L	72.9	72.8	73.0	—
UP-Fuse (Ours)	L	74.9	75.1	74.5	—
IAL [29]	L	77.0	77.8	75.7	—
Panoptic-FusionNet [36]	LC	76.6	77.1	76.0	+3.9
LCPS [46]	LC	79.3	81.7	75.3	**+6.4**
IAL [29]	LC	80.1	82.2	76.5	+3.7
UP-Fuse (Ours)	LC	80.7	83.0	77.0	+5.8
Panoptic-FusionNet [36]	L*	67.7	65.3	71.7	-5.0
LCPS [46]	L*	68.7	67.9	70.1	-4.2
IAL [29]	L*	72.4	71.6	73.8	-4.6
UP-Fuse (Ours)	L*	73.7	73.6	74.1	**-1.2**

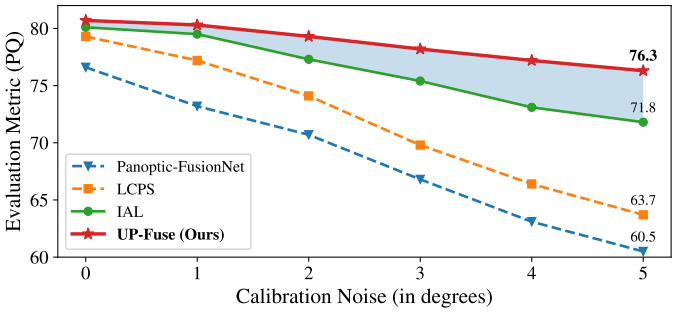


Fig. 4: Robust performance comparison on the Panoptic nuScenes validation set under increasing calibration drift between LiDAR and camera over rotation magnitudes from 0° to 5°. UP-Fuse (red) outperforms all baselines, dropping only 4.4% in PQ compared to $> 8\%$ for state-of-the-art methods. Refer to the supplementary material for visualization of the projection shifts.

(VT-$\mathcal{D}_{\text{dense}}$) provides a 0.8% gain in PQ over sparse depth (VT-$\mathcal{D}_{\text{sparse}}$) in view transformer. Next, we replace simple concatenation with Deformable Cross-Modal Interaction (D-CMI), which provides a substantial 2.4% boost in PQ, demonstrating the clear superiority of attention-based fusion. Adding our sensor degradation augmentations (Augs.) results in a slight drop to 78.9% PQ, as the model learns a robust but overly-cautious representation. The drop is mitigated by enabling our Uncertainty-Aware D-CMI, which elevates performance to 80.7% PQ. This highlights that uncertainty-awareness is essential for effectively leveraging the augmentations.

TABLE VI: Comparison of robustness to visual domain shift on the Panoptic nuScenes validation set. Models trained on day scenes are evaluated on the night split to assess performance under corrupted visual features.

Method	Modality	PQ	PQ^{th}	PQ^{st}	ΔPQ
Panoptic-FusionNet [36]	L	63.1	61.3	64.8	—
LCPS [46]	L	63.3	63.1	63.4	—
UP-Fuse (Ours)	L	64.2	63.9	64.4	—
IAL [29]	L	65.6	65.5	65.7	—
Panoptic-FusionNet [36]	LC	60.0	56.2	63.7	-3.1
LCPS [46]	LC	60.6	58.3	62.9	-2.7
IAL [29]	LC	63.5	63.1	64.2	-2.1
UP-Fuse (Ours)	LC	64.3	64.9	63.7	**+0.1**

TABLE VII: Ablation study of our UP-Fuse architecture on the Panoptic nuScenes validation set. VT denotes View-Transformation. D-CMI denotes Deformable Cross-Modal Interaction. Augs. denotes Augmentations.

Model Variant	Modality	PQ	PQ^{th}	PQ^{st}
M1 (2D-3D Post Processing [22])	L	72.8	72.2	73.8
M2 (Hybrid 2D-3D Decoder)	L	74.9	75.1	74.5
M2 + Concat Fusion (VT-$\mathcal{D}_{\text{sparse}}$)	LC	76.4	77.2	75.1
M2 + Concat Fusion (VT-$\mathcal{D}_{\text{dense}}$)	LC	77.2	78.4	75.2
M2 + D-CMI (VT-$\mathcal{D}_{\text{dense}}$)	LC	79.6	81.6	76.3
+ Sensor Degradation Augs.	LC	78.9	80.7	75.9
+ Uncertainty-Aware D-CMI	LC	**80.7**	**83.0**	**77.0**

TABLE VIII: Pearson correlation $\rho(d_{\text{gt}}, d_{\text{pred}})$ under camera sensor degradation on the Panoptic nuScenes validation set. The predicted instability d_{pred} is consistently positively correlated with the ground-truth feature instability d_{gt}, with the strongest correlation under sensor dropout.

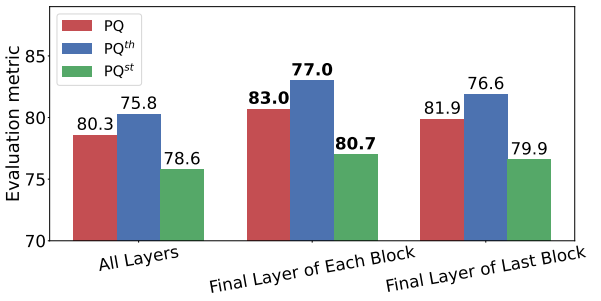
Degradation Type	$\rho(d_{\text{gt}}, d_{\text{pred}})$
Brightness Shift	0.6907
Color Jitter	0.6369
Out-of-Domain	0.7802
Fog	0.7258
Sensor Dropout	0.9595

Hybrid Decoder Design: In Fig. 5, we ablate the two key hyperparameters of our Hybrid 2D-3D Panoptic Decoder. First, in Fig. 5a, we analyze the placement of the 3D-aware mask head within the decoder layers. We find that performing the 3D-lifting operation in the final layer of each decoder block provides the best performance (80.7% PQ), as it allows the queries to first refine themselves in the 2D range-view before being lifted to 3D. Second, in Fig. 5b, we ablate the neighborhood size K for the 3D-aware feature aggregation, showing that $K = 5$ is the optimal choice for balancing context aggregation and feature specificity.

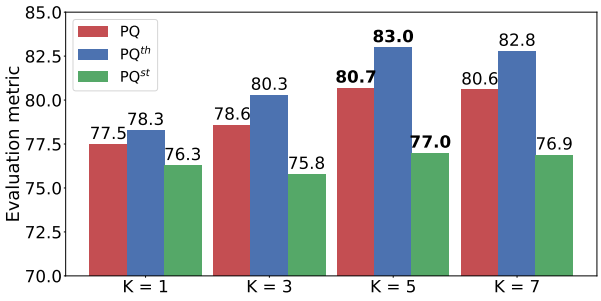
Uncertainty Quantification: Lastly, we evaluate whether our aleatoric uncertainty estimator captures visual feature instability under camera degradation. As shown in Tab. VIII, the predicted instability is consistently positively correlated with the ground-truth feature deviation across different corruptions. The correlation is highest under sensor dropout ($\rho = 0.9595$), where the degradation causes severe and localized feature collapse. Strong correlations under out-of-domain shifts ($\rho = 0.7802$) and fog ($\rho = 0.7258$) further indicate that the estimator generalizes beyond simple missing-sensor cases. These results support our design choice of using feature-level uncertainty to identify unreliable visual cues, enabling the Uncertainty-Aware D-CMI module to down-weight degraded camera features during fusion.

F. Qualitative Evaluation

In Fig. 6, we compare UP-Fuse with the strongest baseline, IAL, on the Panoptic nuScenes validation set. We show three challenging settings: sensor dropout, calibration drift, and nighttime domain shift. Under sensor dropout, IAL fails to resolve geometric ambiguity. It misclassifies concrete barriers as man-made structures. With calibration drift, IAL merges



(a) Analysis of Head Configuration



(b) Impact of Neighborhood Size K

Fig. 5: Ablation studies on the key hyperparameters of our Hybrid 2D-3D Panoptic Decoder.

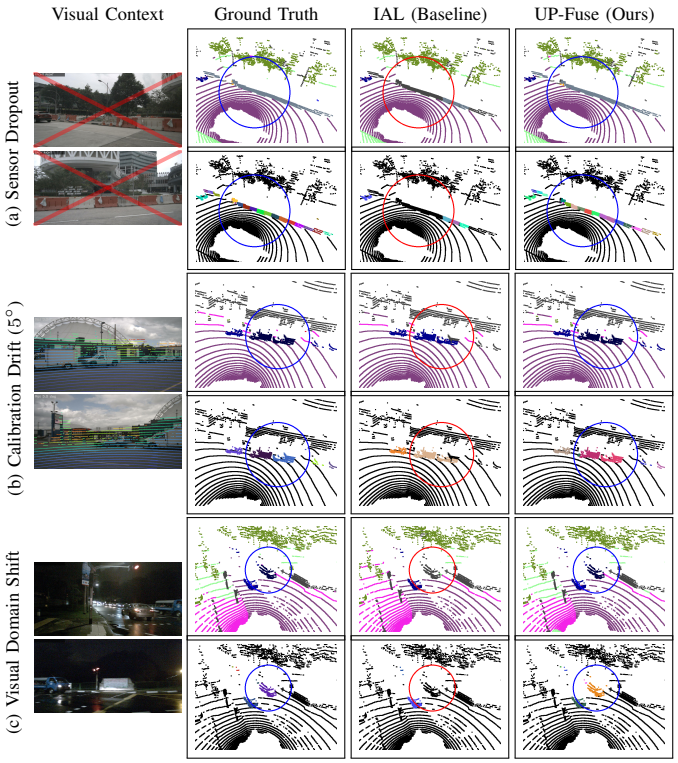


Fig. 6: Qualitative robustness comparison of 3D panoptic segmentation between UP-Fuse and the strongest baseline IAL on the Panoptic nuScenes validation set. (a) IAL fails to resolve geometric ambiguity, misclassifying concrete barriers as man-made structures. (b) Under calibration drift, IAL incorrectly merges two trucks into a single instance and mislabels vehicle tops as background. (c) Nighttime domain shift leads to similar errors. In contrast, UP-Fuse (right) suppresses unreliable visual cues and maintains correct semantic and instance predictions across all scenarios. **red**: incorrect prediction, **blue**: correct prediction (Best viewed at $4\times$ zoom).

two trucks into a single instance. It also mislabels vehicle tops as background. Similar errors appear at night, where unreliable visual cues degrade both semantic and instance predictions. In contrast, UP-Fuse suppresses unreliable visual information. It preserves correct semantic labels and instance separation across all scenarios. These results show that uncertainty-aware fusion improves robustness under degraded sensing conditions.

V. Limitations and Future Work

A limitation of our framework is its reliance on fixed camera parameters for view transformation. While our uncertainty module mitigates calibration drift by suppressing the visual stream, it does not explicitly correct the extrinsics. Consequently, in scenarios with severe misalignment, performance is bounded by the LiDAR-only baseline. Future work will explore joint extrinsic refinement to recover the benefits of fusion under significant sensor displacement.

Furthermore, while UP-Fuse successfully targets robustness against visual degradation, it does not currently model severe LiDAR corruption, such as ghost or missing returns. Extending this uncertainty-aware fusion to handle LiDAR degradation is an important future direction. This requires symmetric, bidirectional uncertainty modeling across both modalities. Currently, the main architectural bottleneck for this is our camera-to-range-view mapping, which explicitly depends on LiDAR depth. A more general solution would decouple this mapping from measured point clouds, for instance, through learned depth prediction.

VI. Conclusion

In this work, we presented UP-Fuse, a robust LiDAR-Camera fusion framework for 3D panoptic segmentation built on a unified range-view representation. UP-Fuse achieves strong performance on Panoptic nuScenes, SemanticKITTI, and our newly introduced Panoptic Waymo benchmark, while demonstrating consistent robustness across camera sensor degradation and failure. The proposed uncertainty-guided fusion mechanism jointly models cross-modal relevance and feature reliability, allowing the network to adaptively attenuate unreliable visual cues under degradation. The hybrid 2D-3D transformer decoder effectively resolves spatial ambiguities inherent to range-view projections. Together, these architectural components enable a practical balance between accuracy, efficiency, and robustness for multi-modal robotic perception.

Acknowledgments

This research was funded by Bosch Research as part of a collaboration between Bosch Research and the University of Freiburg on AI-based automated driving.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall.