

Article

GAC-Net: A Geometric–Attention Fusion Network for Sparse Depth Completion from LiDAR and Image

Xingli Gan ^{1,*}, Kuang Zhu ¹ , Min Sun ¹ , Leyang Zhao ² and Canwei Lai ¹
¹ School of Computer Science and Technology, Zhejiang University of Science and Technology, Hangzhou 310023, China; 212308802004@zust.edu.cn (K.Z.); 222308855053@zust.edu.cn (M.S.); 222408855059@zust.edu.cn (C.L.)

² School of Mathematics and Statistics, University of New South Wales, Sydney, NSW 2052, Australia; z5167073@ad.unsw.edu.au

* Correspondence: 120070@zust.edu.cn

Abstract

Depth completion aims to reconstruct dense depth maps from sparse LiDAR measurements guided by RGB images. Although BPNet enhanced depth structure perception through a bilateral propagation module and achieved state-of-the-art performance at the time, there is still room for improvement in leveraging 3D geometric priors and adaptively fusing heterogeneous modalities. To this end, we proposed GAC-Net, a Geometric–Attention Fusion Network that enhances geometric representation and cross-modal fusion. Specifically, we designed a dual-branch PointNet++-S encoder, where two PointNet++ modules with different receptive fields are applied to extract scale-aware geometric features from the back-projected sparse point cloud. These features are then fused using a channel attention mechanism to form a robust global 3D representation. A Channel Attention-Based Feature Fusion Module (CAFFM) was further introduced to adaptively integrate this geometric prior with RGB and depth features. Experiments on the KITTI depth completion benchmark demonstrated the effectiveness of GAC-Net, achieving an RMSE of 680.82 mm, ranking first among all peer-reviewed methods at the time of submission.

Keywords: depth completion; sparse LiDAR; deep learning; computer vision; 3D geometric representation; point cloud representation; attention-based fusion



Received: 29 July 2025

Revised: 29 August 2025

Accepted: 30 August 2025

Published: 4 September 2025

Citation: Gan, X.; Zhu, K.; Sun, M.; Zhao, L.; Lai, C. GAC-Net: A Geometric–Attention Fusion Network for Sparse Depth Completion from LiDAR and Image. *Sensors* **2025**, *25*, 5495. <https://doi.org/10.3390/s25175495>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Depth completion [1] is a fundamental perception task that aims to reconstruct dense and accurate depth maps from sparse sensor inputs. It plays a critical role in various applications, such as autonomous driving [2–11] and sensor-based environmental understanding [12,13]. However, due to hardware constraints and occlusions, LiDAR sensors typically produce sparse and unevenly distributed depth maps. For example, a 64-line LiDAR provides valid measurements for only about 4% of image pixels [14–16], which poses significant challenges for downstream 3D perception tasks [17]. Although the KITTI dataset [18] remains the most widely used benchmark for depth completion, recent works such as RSUD20K [19] emphasize the importance of dataset diversity and robustness evaluation under diverse conditions, further underscoring the need for methods with strong generalization ability.

To overcome this challenge, mainstream approaches [20,21] typically adopt multi-modal fusion strategies by combining sparse depth maps with RGB images. Many existing methods [22–24] follow a two-stage architecture that performs early or late fusion of depth

and image features. However, these approaches struggle with distinguishing between valid and missing observations and often fail to effectively process the irregular distribution of sparse depth values [25]. A representative improvement, BPNet [26], introduces a three-stage framework that begins with a bilateral propagation module to generate an initial dense depth prior, followed by a multi-modal fusion stage and a final refinement stage. Nevertheless, its fusion module relies on a conventional U-Net backbone [27] and lacks explicit modeling of 3D geometric priors and adaptive integration of heterogeneous features.

To address these limitations, we propose GAC-Net, a Geometric–Attention Fusion Network that significantly improves depth completion performance by explicitly modeling 3D geometric priors and enhancing the multi-modal fusion mechanism. Specifically, GAC-Net incorporates two key modules. First, a dual-branch PointNet++-S encoder is designed with two parallel PointNet++ [28] branches of different receptive fields to extract scale-aware geometric features from the back-projected sparse point cloud. These features are fused using a channel attention mechanism [29] to form a robust global geometric representation. Second, a Channel Attention-Based Feature Fusion Module (CAFFM) is introduced to adaptively recalibrate and integrate the global 3D priors with RGB-depth features, improving the flexibility and accuracy of the fusion process.

The main contributions of this study are summarized as follows:

1. We propose a dual-branch PointNet++-S encoder to extract scale-aware geometric features from sparse point clouds and form a robust 3D representation;
2. We design a Channel Attention-Based Feature Fusion Module (CAFFM) to adaptively fuse geometric priors with RGB-depth features;
3. Extensive experiments on the KITTI depth completion benchmark demonstrate that GAC-Net outperforms BPNet and other published methods in both accuracy, robustness, and structural preservation, ranking first among peer-reviewed methods on the official leaderboard at the time of submission.

2. Related Work

2.1. Two-Dimensional-Based Depth Completion

Early deep learning methods for depth completion typically adopt a single-stage architecture [20,30,31], where sparse depth maps and RGB images are concatenated and processed by a shared convolutional encoder-decoder network. For example, the Sparse-to-Dense [31] framework proposed by Ma et al. fuses LiDAR inputs and monocular images through a unified backbone. Although these approaches are structurally simple and computationally efficient, they often produce overly smooth depth maps and fail to preserve fine geometric details around object boundaries.

To address these limitations, subsequent studies introduced two-stage frameworks [22–24,32–37], which first perform multi-modal feature fusion, followed by a refinement module for post-processing. CSPN [22] and CSPN++ [23], for instance, leverage convolutional spatial propagation networks to revisit the original sparse but accurate depth measurements and enhance edge sharpness. More recently, CFormer [32] extends this line of research by employing a CNN–Transformer hybrid encoder to extract and fuse RGB-depth features, followed by a refinement stage for fine-grained detail recovery. These methods gained popularity for their effectiveness in boundary preservation. However, directly applying convolution to sparse inputs compromises fusion quality, and this issue is also difficult to effectively resolve in the post-processing stage.

To further improve depth estimation quality, BPNet [26] proposes a three-stage architecture by adding a bilateral propagation module before fusion. This module generates an initial dense depth map from sparse input, providing structured geometric priors

that alleviate data sparsity and enhance the overall completion performance, achieving state-of-the-art results at the time.

Nevertheless, all the above methods are fundamentally 2D-guided and lack explicit modeling of 3D spatial structure, which limits their ability to fully exploit the geometric information inherently embedded in sparse depth measurements.

2.2. 2D–3D Joint Depth Completion: Geometric, Multi-View, and Transformer Approaches

To better exploit the 3D structural information embedded in sparse depth data, recent studies [38–48] have explored three major directions: geometry-based approaches, multi-view/BEV representations, and Transformer-based cross-modal reasoning, each contributing complementary strengths to depth completion performance.

Geometry-based approaches can be further divided into three categories. The first category involves estimating surface normals as intermediate cues to ensure local geometric coherence. For instance, DeepLiDAR [38] and Depth-Normal [39] leverage normal estimation to improve local consistency during the completion process. The second category employs graph-based representations to capture geometric and contextual relationships. Typical works, such as ACMNet [40] and GraphCSPN [41], use graph neural networks to model spatial affinities and guide the effective fusion of RGB and depth features. The third category adopts point-based modeling by treating sparse LiDAR measurements as 3D point clouds. Notable examples include FuseNet [42] and PointDC [43], which directly extract 3D geometric features from the reprojected point cloud branch and combine them with image features for joint depth estimation.

Beyond explicit geometry, multi-view or bird’s-eye-view (BEV) representations are leveraged to capture global 3D structural consistency. For example, BEV@DC [44] constructs a BEV representation of sparse depth measurements to encode global spatial context while preserving fine-grained local geometry. More recently, TPVD [45] decomposes sparse depth maps into multiple perspective views that are subsequently reprojected and fused in 2D space, thereby reconstructing dense depth maps with improved global consistency.

Another emerging line of research explores Transformer-based architectures for cross-modal reasoning. The Dual-Attention Fusion model [46] employs a CNN–Transformer hybrid encoder and introduces a Dual-Attention Fusion module that combines convolutional spatial/channel attention with Transformer-based cross-modal attention. DeCoTR [47] further advances this line by formulating depth completion purely as a cross-modal Transformer reasoning problem, leveraging global attention to integrate RGB and depth features. More recently, HTMNet [48] extends this paradigm by coupling Transformer with a Mamba-based state-space bottleneck, representing the first attempt to introduce state-space modeling into depth completion and demonstrating promising results in handling challenging scenarios such as transparent and reflective objects.

Despite these advancements, existing methods still face several challenges. Geometry-based approaches often struggle to balance local precision and global consistency, multi-view/BEV-based representations increase computational cost, and Transformer-based frameworks, while powerful in modeling cross-modal relationships, generally involve complex architectures and heavy computational overhead. In contrast, our method adopts a point-based geometric encoding to explicitly exploit 3D structure and integrates it with an SE-based attention fusion module, thereby achieving a better balance between accuracy and efficiency for robust depth completion.

3. Methods

GAC-Net is designed to fully exploit the rich texture cues from RGB images and the accurate yet incomplete geometric priors from sparse depth data to generate high-quality

dense depth maps. Let $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ be the input RGB image, $\mathbf{S} \in \mathbb{R}^{H \times W \times 1}$ denotes the sparse depth map, and $\mathbf{D} \in \mathbb{R}^{H \times W \times 1}$ denotes the predicted dense depth map.

As illustrated in Figures 1 and 2, GAC-Net builds upon the multi-scale, three-stage architecture proposed in BPNet [26], which performs coarse-to-fine depth prediction over six hierarchical scales (from scale-5 to scale-0). At each scale, depth prediction is performed in three sequential stages:

- (1) Pre-processing stage: A bilateral propagation module generates an initial dense depth map from the sparse input, providing a more structured estimate for subsequent fusion;
- (2) Enhanced multi-modal fusion: a U-Net backbone [27] extracts multi-scale 2D features, while the proposed dual-branch PointNet++-S encodes density-adaptive local and contextual 3D priors. These features are adaptively fused through the proposed Channel Attention-Based Feature Fusion Module (CAFFM), which reweights channels to enhance cross-modal consistency;
- (3) Refinement stage: Finally, residual learning [49] and CSPN++ [23] are employed to iteratively refine the dense depth, enforcing local consistency and recovering fine-grained structures.

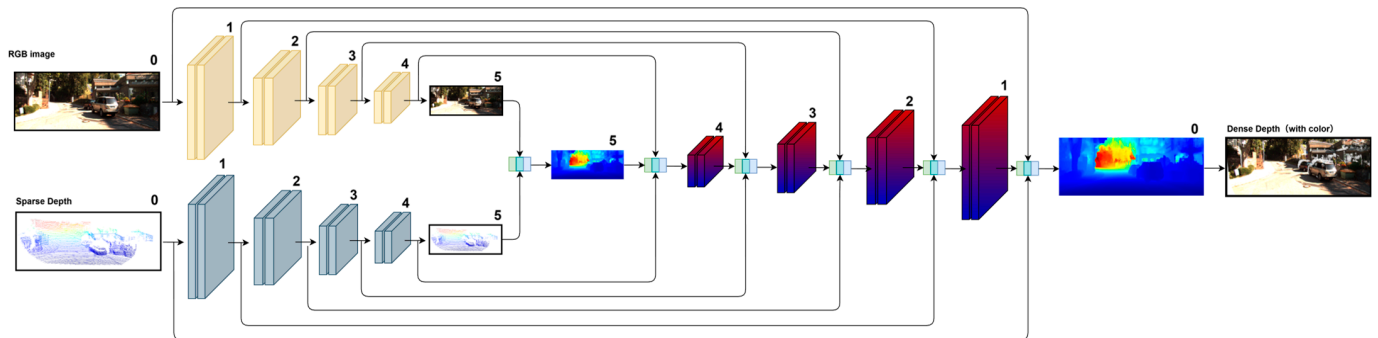


Figure 1. Overview of the multi-scale processing mechanism in GAC-Net.

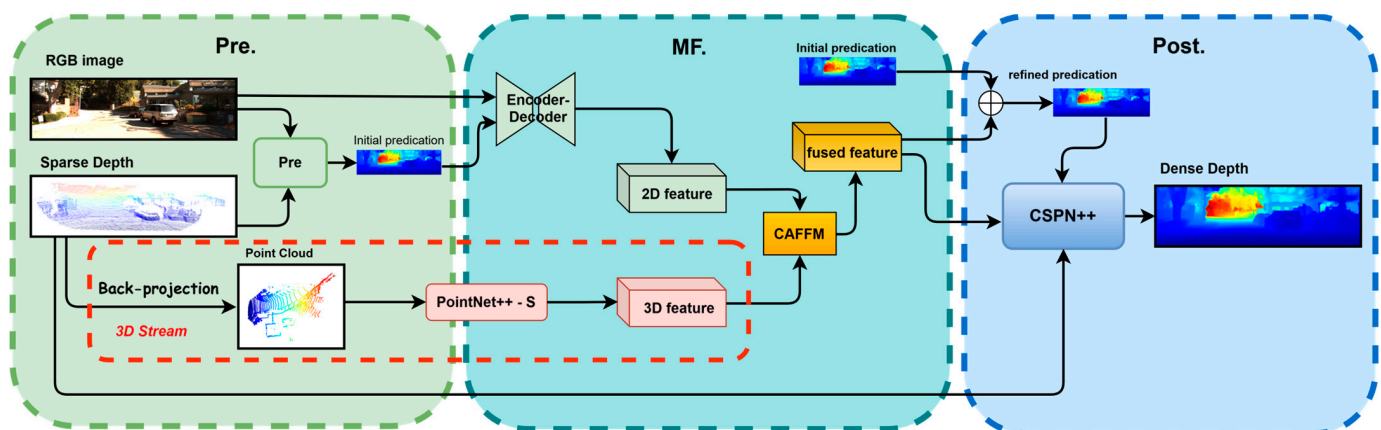


Figure 2. Overview of the three-stage network.

Figure 1 illustrates the progressive coarse-to-fine improvement of the depth map across scales, while Figure 2 shows the internal three-stage architecture within each resolution. The following sections present each module in detail (Sections 3.1–3.3) and describe the multi-scale training loss (Section 3.4).

3.1. Preprocessing via Bilateral Propagation

To generate a structured initial dense depth map from sparse input, we adopt the bilateral propagation module originally proposed by BPNet [26]. This module propagates valid depth measurements by dynamically aggregating neighboring pixels based on spatial proximity and radiometric similarity, effectively mitigating ambiguity caused by sparse sampling.

Given the input sparse depth map $\mathbf{S}$, the initial dense depth $\mathbf{D}^0$ can be computed as follows:

$$\mathbf{D}_i^0 = \sum_{j \in \mathcal{N}(i)} \omega_{ij} \mathbf{S}_j, \quad \sum_j \omega_{ij} = 1 \quad (1)$$

Here, $\mathcal{N}(i)$ denotes the set of neighboring pixels for pixel i , and ω_{ij} are adaptive weights reflecting both spatial distance and color similarity.

This propagation provides a spatially structured initial estimate that serves as a reliable basis for the subsequent U-Net [27] encoding and the proposed PointNet++-S-based geometry extraction. It should be noted that, consistent with BPNet [26], the multi-scale coarse-to-fine strategy and the refinement stage remain unchanged, while our main contribution focuses on the 3D geometric modeling and enhanced multi-modal fusion modules described in Sections 3.2 and 3.3.

3.2. Enhanced Multi-Modal Feature Fusion

To further enhance the robustness and accuracy of depth completion, GAC-Net improves the conventional multi-modal fusion design by explicitly incorporating global 3D geometric priors and a channel attention mechanism for adaptive cross-modal integration. Specifically, the proposed module combines a U-Net [27] backbone for multi-scale 2D feature extraction, a dual-branch PointNet++-S encoder for multi-scale 3D geometry representation, and a Channel Attention-Based Feature Fusion Module (CAFFM) to adaptively fuse the 2D and 3D features. The detailed structure of these components and their interaction is shown in Figure 3.

3.2.1. U-Net Backbone for 2D Feature Fusion

In the multi-modal fusion stage, the RGB image $\mathbf{I}$ and the initial dense depth map $\mathbf{D}^0$ generated by the bilateral propagation module are concatenated to form the 2D input tensor:

$$\mathbf{X}_{\text{in}} = \text{Concat}(\mathbf{I}, \mathbf{D}^0) \quad (2)$$

A U-Net [27] encoder-decoder backbone (encoder-decoder network) is then employed to extract hierarchical multi-scale 2D features $\mathbf{F}_{2D}$:

$$\mathbf{F}_{2D} = \text{U-Net}(\mathbf{X}_{\text{in}}) \quad (3)$$

This architecture progressively captures local texture, edge, and structural context through down-sampling and restores spatial resolution with skip connections during up-sampling. The resulting feature maps provide rich 2D cues for robust multi-modal fusion.

3.2.2. Dual-Branch PointNet++-S Encoder for 3D Geometry Representation

To capture expressive geometric features and maintain robustness under varying sparsity conditions, we propose a dual-branch PointNet++-S encoder, where S stands for Scale-aware, indicating its ability to extract multi-scale 3D representations from sparse depth inputs.

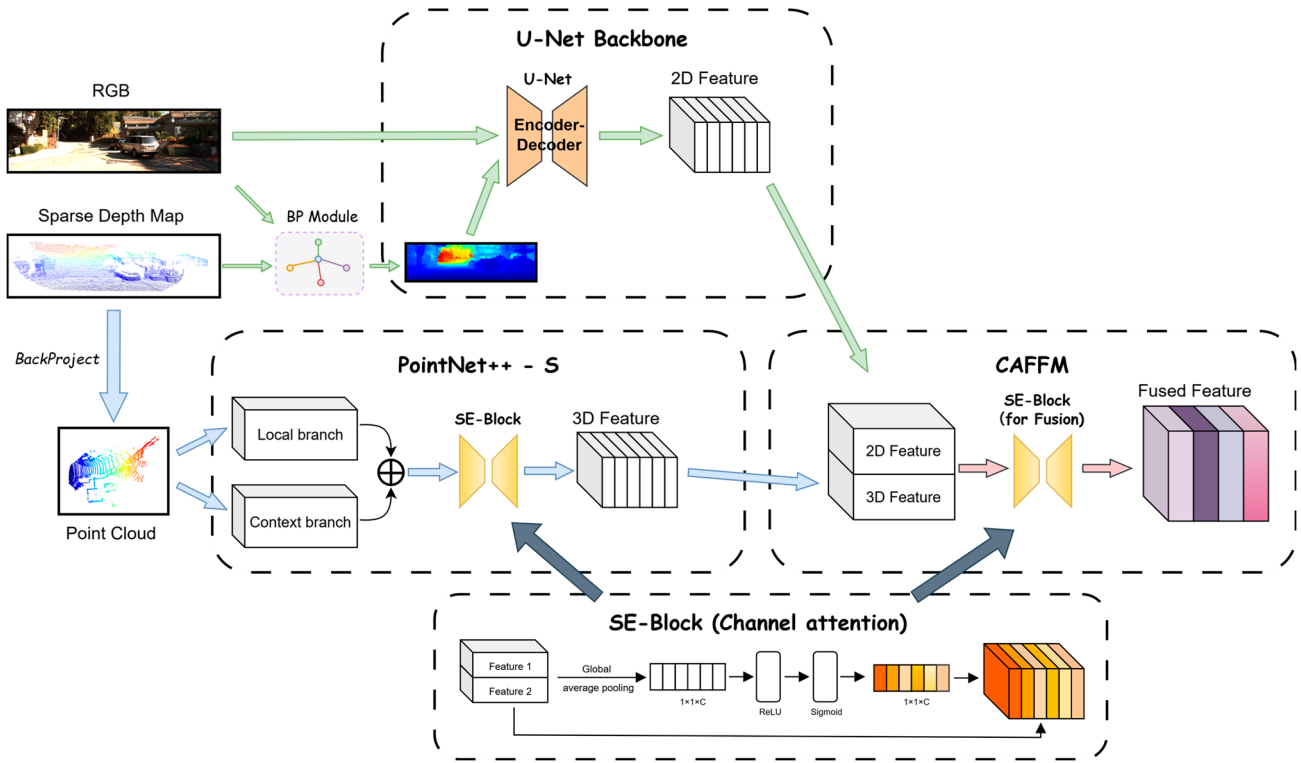


Figure 3. The architecture of the proposed 2D–3D feature extraction and fusion pipeline. The pipeline consists of a U-Net backbone for multi-scale 2D feature extraction, a dual-branch PointNet++-S module for density-adaptive 3D geometric modeling from back-projected point clouds, and a Channel Attention-Based Feature Fusion Module (CAFFM) for adaptive integration of 2D and 3D features. SE-Blocks are employed in both 3D encoding and feature fusion stages to enhance channel-wise interactions.

Let $\mathbf{P} = \{\mathbf{p}_i \in \mathbb{R}^3 | i = 1, \dots, N\}$ denote the 3D point cloud obtained by back-projecting valid sparse depth pixels using camera intrinsics. The encoder consists of two parallel PointNet++ [28] branches with distinct receptive field sizes, designed to extract both local and contextual geometric features.

The local branch aggregates fine-grained neighborhood information via two Set Abstraction (SA) layers:

$$\mathbf{F}_{\text{local}}^{(1)} = SA_{L_1}(\mathbf{P}; 384, 0.2m, 32) \quad (4)$$

$$\mathbf{F}_{\text{local}}^{(2)} = SA_{L_2}(\mathbf{F}_{\text{local}}^{(1)}; 128, 0.4m, 64) \quad (5)$$

where each SA layer samples a fixed number of neighboring points within a predefined radius.

Specifically, each Set Abstraction (SA) layer is defined by a tuple (n, r, k) , where the following is true:

- n is the number of center points sampled from the point cloud using farthest point sampling (FPS);
- r is the ball query radius used to group neighboring points around each center;
- k is the number of neighboring points selected within radius r for local feature aggregation.

These parameters control the receptive field and granularity of geometric information captured at each SA layer.

The context branch uses larger receptive fields to model broader scene-level geometry:

$$\mathbf{F}_{\text{context}}^{(1)} = SA_{C_1}(\mathbf{P}; 256, 0.4m, 32) \quad (6)$$

$$\mathbf{F}_{\text{context}}^{(2)} = \text{SAC}_2\left(\mathbf{F}_{\text{context}}^{(1)}; 64, 0.6m, 64\right) \quad (7)$$

The outputs from both branches are concatenated to form a unified multi-scale 3D feature:

$$\mathbf{F}_{3D} = \text{Concat}\left(\mathbf{F}_{\text{local}}^{(2)}, \mathbf{F}_{\text{context}}^{(2)}\right) \in \mathbb{R}^{256} \quad (8)$$

To emphasize informative channels, we apply a squeeze-and-excitation (SE) block [29] for channel-wise recalibration. The SE block first performs global average pooling:

$$\mathbf{z}_c = \frac{1}{N} \sum_{i=1}^N \mathbf{F}_{3D}^{(c)}(i), \quad c = 1, \dots, 256 \quad (9)$$

This channel descriptor is passed through two fully connected layers with ReLU and Sigmoid activations to compute attention weights:

$$\mathbf{s} = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})) \quad (10)$$

The recalibrated global 3D feature is then obtained via channel-wise scaling:

$$\mathbf{G}_{3D} = \mathbf{s} \odot \mathbf{F}_{3D} \quad (11)$$

where $\odot$ denotes element-wise multiplication. Finally, $\mathbf{G}_{3D}$ is broadcasted spatially to match the dimensions of the 2D feature map for cross-modal fusion in the next stage.

3.2.3. Channel Attention-Based Feature Fusion Module (CAFFM)

To integrate complementary local 2D texture features and global 3D geometric priors, we propose a Channel Attention-Based Feature Fusion Module (CAFFM). Let $\mathbf{F}_{2D} \in \mathbb{R}^{H \times W \times c_2}$ denote the multi-scale 2D feature map extracted by the U-Net [27] backbone, and $\mathbf{G}_{3D} \in \mathbb{R}^{c_3}$ represent the recalibrated global 3D feature vector obtained from the PointNet++-S encoder.

To enable effective fusion, the 3D feature is first broadcast spatially and concatenated with the 2D feature along the channel dimension:

$$\mathbf{F}_{\text{fusion-in}} = \text{Concat}\left(\mathbf{F}_{2D}, \tilde{\mathbf{G}}_{3D}\right), \quad \mathbf{F}_{\text{fusion-in}} \in \mathbb{R}^{H \times W \times (c_2 + c_3)} \quad (12)$$

where $\tilde{\mathbf{G}}_{3D}$ denotes the spatially broadcasted version of $\mathbf{G}_{3D}$.

A squeeze-and-excitation (SE) block [29] is then applied to adaptively learn the fusion attention. First, global average pooling compresses spatial information across all channels:

$$\mathbf{z}_c = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W \mathbf{F}_{\text{fusion-in}}^{(c)}(i, j), \quad c = 1, \dots, C \quad (13)$$

where $C = c_2 + c_3$.

This channel descriptor is passed through two fully connected layers with ReLU and Sigmoid activations to produce the channel attention weights:

$$\mathbf{s} = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})) \quad (14)$$

The final fused feature map is recalibrated via channel-wise scaling:

$$\mathbf{F}_{\text{fusion}} = \mathbf{s} \odot \mathbf{F}_{\text{fusion-in}} \quad (15)$$

where $\odot$ denotes element-wise multiplication across channels.

This adaptively fused representation is subsequently passed to the refinement stage to generate the final high-quality dense depth map.

3.2.4. Pseudocode for Enhanced Multi-Modal Feature Fusion

To provide a concise overview of the proposed fusion mechanism, we summarize the workflow in Algorithm 1. At a given scale s , the inputs include the RGB image $\mathbf{I}$, the sparse depth $\mathbf{D}$, and the pre-processed dense depth $\mathbf{D}_s^{(\text{pre})}$ obtained from the bilateral propagation (Section 3.1). The process first extracts 3D features from the back-projected point cloud using a dual-branch PointNet++-S with SE-based recalibration, while 2D features are encoded from the concatenated RGB-depth input using a U-Net encoder. Next, the 3D features are spatially broadcast to the image grid and fused with the 2D features through the CAFFM module with channel-attention weighting, producing the enhanced multi-modal feature $\mathbf{F}_s$. This procedure corresponds to Equations (2)–(15), and its pseudocode implementation is provided in Algorithm 1.

3.3. Depth Refinement

Although the fused depth map captures most scene structures, it may still suffer from local inconsistencies and blurred boundaries. To address this, we adopt a combination of residual learning [49] and CSPN++ [23] in the refinement stage.

Specifically, residual learning [49] corrects global prediction errors using the fused feature $\mathbf{F}_{\text{fused}}$:

$$\mathbf{D}_{\text{res}} = \mathbf{D}^* + R(\mathbf{F}_{\text{fused}}) \quad (16)$$

where $R(\cdot)$ denotes the residual regressor. Subsequently, CSPN++ [23] refines local details by iteratively propagating depth values:

$$\mathbf{D}_i^{t+1} = \sum_{j \in \mathcal{N}(i)} \kappa_{ij} \mathbf{D}_j^t + \left(1 - \sum_j \kappa_{ij}\right) \mathbf{s}_i \quad (17)$$

where κ_{ij} are learned affinity weights.

This refinement process preserves clear depth boundaries, improves the structural integrity of ambiguous regions, and enhances the overall accuracy of the predicted depth map, thereby supporting more precise estimation in the subsequent stages of the multi-scale framework.

3.4. Loss Function

To supervise the training of GAC-Net in an end-to-end manner, we adopt a multi-scale hybrid loss that combines the mean squared error (L2) and the mean absolute error (L1) terms, providing robust guidance for both global consistency and local structural accuracy. Specifically, for each predicted depth map at scale s , we apply a bilinear up-sampling operator to match the ground truth resolution and compute the loss within valid pixel regions. The total training loss is defined as follows:

$$L = \sum_{s=0}^5 \lambda_s \sum_{i \in \mathcal{P}} \left(\left\| \mathbf{D}_i^g - U_s(\mathbf{D}^*) \right\|_2^2 + \left\| \mathbf{D}_i^g - U_s(\mathbf{D}^s) \right\|_1 \right) \quad (18)$$

where $\mathbf{D}^*$ denotes the ground truth dense depth map, $\mathbf{D}^s$ is the predicted depth at scale s , $U_s(\cdot)$ is the bilinear up-sampling operator, $\mathcal{P}$ is the set of valid pixels in $\mathbf{D}^g$, and λ_s is a scale-dependent weight set to 4^{-s} , following the setting in BPNet [26].

This hybrid loss ensures that GAC-Net effectively minimizes both global depth estimation errors and local fine-grained residuals across multiple resolutions, producing high-quality dense depth maps even under sparse and noisy input conditions.

Algorithm 1: Enhanced Multi-Modal Feature Fusion at scale s (Equations (2)–(15))**Input:**RGB image $I \in R^{H \times W \times 3}$ Sparse depth map $D \in R^{H \times W \times 1}$ Pre-processed dense depth (from BP) $D_s^{(pre)} \in R^{H \times W}$ Camera intrinsics $K \in R^{3 \times 3}$ **Output:**Fused feature map $F_s \in R^{H \times W \times C}$ **Notion:** $\oplus$: channel concatenation; $\odot$: channel-wise multiplication;

GAP: global average pooling;

 $\delta(\cdot)$: ReLU; $\sigma(\cdot)$: Sigmoid $D_s^{(pre)}$: dense depth pre-processed via Bilateral Propagation**Procedure:**

- 1: Step 1: Back-projection to 3D
- 2: Construct sparse point cloud $P = \text{BackProject}(S, K)$.
- 3: Step 2: 2D feature encoding
- 4: Form 2D input $X = [I \oplus D_s^{(pre)}]$
- 5: $F_{2D} = U - \text{Net}(X)$
- 6: Step 3: Dual-branch PointNet++-S
- 7: For each branch $b \in \{local, context\}$:
- 8: For each SA layer with config (n_l, r_l, k_l) :
- 9: $g_l^{(b)} \leftarrow \text{PointNetSA}(P; n_l, r_l, k_l)$
- 10: End for
- 11: End for
- 12: Aggregate to multi-scale 3D feature $F_{3D} = \text{Concat}(F_{local}^{(2)}, F_{context}^{(2)})$
- 13: Step 4: Channel recalibration on 3D
- 14: $z = \text{GAP}(F_{3D})$;
- 15: $s = \sigma(W_2 \cdot \delta(W_1 z))$;
- 16: $G_{3D} = s \odot F_{3D}$.
- 17: Step 5: Spatial broadcasting of 3D feature
- 18: $\tilde{G}_{3D} = \text{Broadcast}(G_{3D} \rightarrow H \times W)$
- 19: Step 6: CAFFM: channel-attention fusion
- 20: $F_{in} = \text{Concat}(F_{2D}, \tilde{G}_{3D})$;
- 21: $z = \text{GAP}(F_{in})$;
- 22: $\alpha = \sigma(W_2 \cdot \delta(W_1 z))$;
- 23: $F_s = \alpha \odot F_{in}$.
- 24: **return** F_s .

4. Experiment

4.1. Experiment Setting

4.1.1. Datasets

All experiments are conducted on the KITTI Depth Completion (DC) benchmark [18], which is widely used to evaluate outdoor depth completion performance. The dataset is collected from an autonomous driving platform, with the ground truth depth generated by temporally registering multiple LiDAR scans and refined using stereo image pairs. It

consists of 86,898 training frames, 1000 validation frames, and 1000 test frames. Following common practice, we crop the upper 100-pixel region (where LiDAR data is typically missing) from the original resolution of 1216×352 to 1216×256 during training. In inference, the full resolution is used for evaluation.

4.1.2. Training Details

We implement GAC-Net using PyTorch and train it on a workstation equipped with four NVIDIA RTX 4090 GPUs. The sparse depth map is first back-projected and processed with the proposed dual-branch PointNet++-S encoder to extract a robust global geometric feature.

Following common practice in recent depth completion works [26,50], we use the AdamW optimizer with a weight decay of 0.05 and apply gradient clipping with a maximum L2 norm of 0.1. The learning rate is scheduled using a warm-up and cosine annealing strategy: it starts at $1/40$ of the maximum value (2.5×10^{-4}), linearly increases to the maximum in the first 10% of iterations, and then decays to 10% of the maximum following a cosine curve for the remaining 90% of training. The batch size is set to 2 per GPU. The model is trained for 30 epochs. Random brightness and contrast augmentation is applied to RGB images during training.

To further illustrate the convergence process, Figure 4 presents the normalized training curves of the per-scale losses (Scale-0 to Scale-5), which are plotted together for clarity. All curves consistently decrease and converge to a stable plateau after sufficient iterations. Figure 5 shows the validation curve of the main evaluation metric (RMSE), which decreases smoothly and stabilizes after approximately 20 epochs.

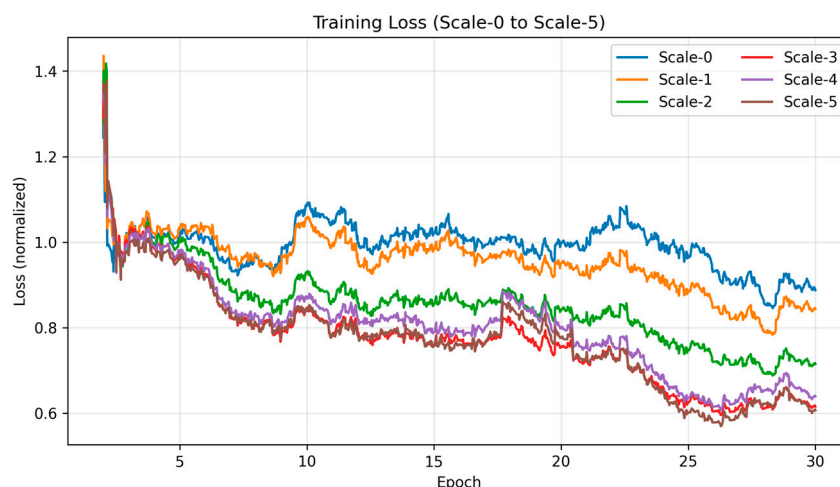


Figure 4. Normalized training curves of the per-scale losses (Scale-0 to Scale-5) during training. All curves decrease consistently and converge to a stable plateau after sufficient iterations.

4.1.3. Metrics

Four standard metrics are used to evaluate performance: root mean squared error (RMSE, in mm), mean absolute error (MAE, in mm), root mean squared error of inverse depth (iRMSE, in $1/\text{km}$), and mean absolute error of inverse depth (iMAE, in $1/\text{km}$). Among these, RMSE is the primary metric used for ranking on the official KITTI benchmark.

The following formula shows the detailed definition of each indicator:

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - y_{gt})^2} \quad (19)$$

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^m |y_i - y_{gt}| \quad (20)$$

$$\text{iRMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m \left(\frac{1}{y_i} - \frac{1}{y_{gt}} \right)^2} \quad (21)$$

$$\text{iMAE} = \frac{1}{m} \sum_{i=1}^m \left| \frac{1}{y_i} - \frac{1}{y_{gt}} \right| \quad (22)$$

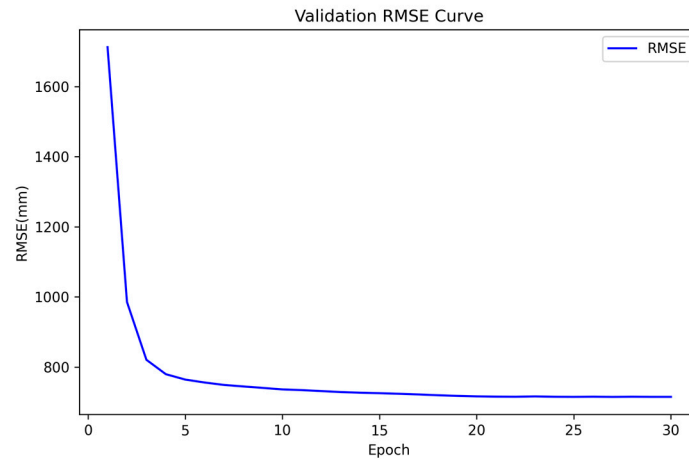


Figure 5. Validation curve of the main evaluation metric, RMSE. The curve decreases smoothly and reaches a stable plateau after approximately 20 epochs.

4.2. Comparison with SoTA Methods

To comprehensively validate the superiority of the proposed GAC-Net, we conduct comparison experiments from both quantitative and qualitative perspectives.

4.2.1. Quantitative Comparison

Table 1 summarizes the performance of GAC-Net and representative state-of-the-art (SOTA) methods. The upper section lists purely 2D-based approaches, while the lower section includes 2D–3D joint methods that incorporate geometric priors from LiDAR point clouds.

Our GAC-Net achieves the lowest RMSE of 680.82 mm, surpassing BPNet [26] (684.90 mm), TPVD [45] (693.97 mm), and other recent methods. This result demonstrates the effectiveness of our enhanced multi-modal fusion framework, particularly the integration of scale-aware 3D geometry via PointNet++-S and adaptive channel-wise fusion through CAFFM.

In addition to RMSE, GAC-Net also achieves competitive performance on other metrics, including MAE, iRMSE, and iMAE, indicating both global accuracy and robustness in local structures. Notably, GAC-Net ranks first among all peer-reviewed methods on the official KITTI leaderboard at the time of submission.

In contrast to early 2D-based methods such as CSPN++ [23] and PENet [21], which primarily rely on image-guided convolutions, GAC-Net leverages explicit 3D geometric encoding to address the irregularity and sparsity of LiDAR inputs. While 2D methods can effectively propagate depth in high-texture regions, they often fail to capture reliable geometric structures in textureless or occluded areas, leading to blurred or distorted reconstructions. By integrating point-based representations through the PointNet++-S encoder, our model significantly improves structural completeness and robustness in such challenging scenarios.

Table 1. Quantitative result on KITTI dataset.

Method	2D	3D	RMSE↓	MAE↓	iRMSE↓	iMAE↓	Publication
CPSN [22]	✓		1019.64	279.46	2.93	1.15	ECCV 2018
TWISE [51]	✓		840.20	195.58	2.08	0.82	CVPR 2021
CSPN++ [23]	✓		743.69	209.28	2.07	0.90	AAAI 2020
PENet [21]	✓		730.08	210.5	2.17	0.94	ICRA 2021
RigNet [35]	✓		712.66	203.25	2.08	0.90	ECCV 2022
LRRU [52]	✓		696.51	189.96	1.87	0.81	ICCV 2023
BP-Net [26]	✓		684.90	194.69	1.82	0.84	CVPR 2024
FuseNet [42]	✓	✓	752.88	221.19	2.34	1.14	ICCV 2019
ACMNet [40]	✓	✓	744.91	206.09	2.08	0.90	T-IP 2021
PointFusion [53]	✓	✓	741.90	201.10	1.97	0.85	ICCV 2021
GraphCSPN [41]	✓	✓	738.41	199.31	1.96	0.84	ECCV 2022
PointDC [43]	✓	✓	736.07	201.87	1.97	0.87	ICCV 2023
DeCoTR [47]	✓	✓	717.07	195.30	1.92	0.84	CVPR 2024
TPVD [45]	✓	✓	693.97	188.60	1.82	0.81	CVPR 2024
GAC-Net(ours)	✓	✓	680.82	193.85	1.81	0.84	-

↓ indicates that a lower value is better. The optimal numerical results are highlighted in bold.

Compared with recent 2D–3D joint methods such as TPVD [45] and DeCoTR [47], GAC-Net adopts a more modular and geometry-driven design. TPVD [45] decomposes the sparse input into multiple perspective views to enhance structural awareness, but it also introduces repeated projection steps, which tend to increase error accumulation and processing complexity. DeCoTR [47], on the other hand, integrates 2D and 3D features via global attention, facilitating long-range interactions across modalities. However, both approaches involve complex processing pipelines. In contrast, GAC-Net focuses on scale-aware point-based geometry encoding and adaptive channel-wise fusion, yielding lower RMSE while maintaining architectural clarity and generalizability across varying input sparsity levels.

4.2.2. Qualitative Comparison

We further conduct visual comparisons to assess the perceptual quality of the predicted depth maps. As shown in Figures 6–8, GAC-Net is compared with four representative state-of-the-art methods: LRRU [52], TPVD [45], CFormer [32], and ImprovingDC [50], as well as its baseline model BPNet [26].

Figure 6 focuses on the reconstruction of large-scale foreground objects, particularly a complete vehicle. The zoom-in region highlights GAC-Net’s ability to recover the rear windshield and its boundary with higher completeness and clarity, while other methods either produce incomplete structures (e.g., LRRU [52]) or blurred and collapsed outlines (e.g., TPVD [45] and CFormer [32]). This illustrates GAC-Net’s superior capability in handling spatially sparse but structurally significant areas.

Figure 7 compares the preservation of fine structures such as thin poles and distant pedestrians. In the zoomed views, GAC-Net demonstrates sharper edges and better geometric integrity, whereas other methods show common issues such as edge blurring, structural breakage, or background fusion. For instance, TPVD [45] exhibits strong edge diffusion, and ImprovingDC [50] fails to recover the basic shape of the pedestrian’s head. These results indicate that GAC-Net can effectively retain high-frequency details even under highly sparse input conditions.

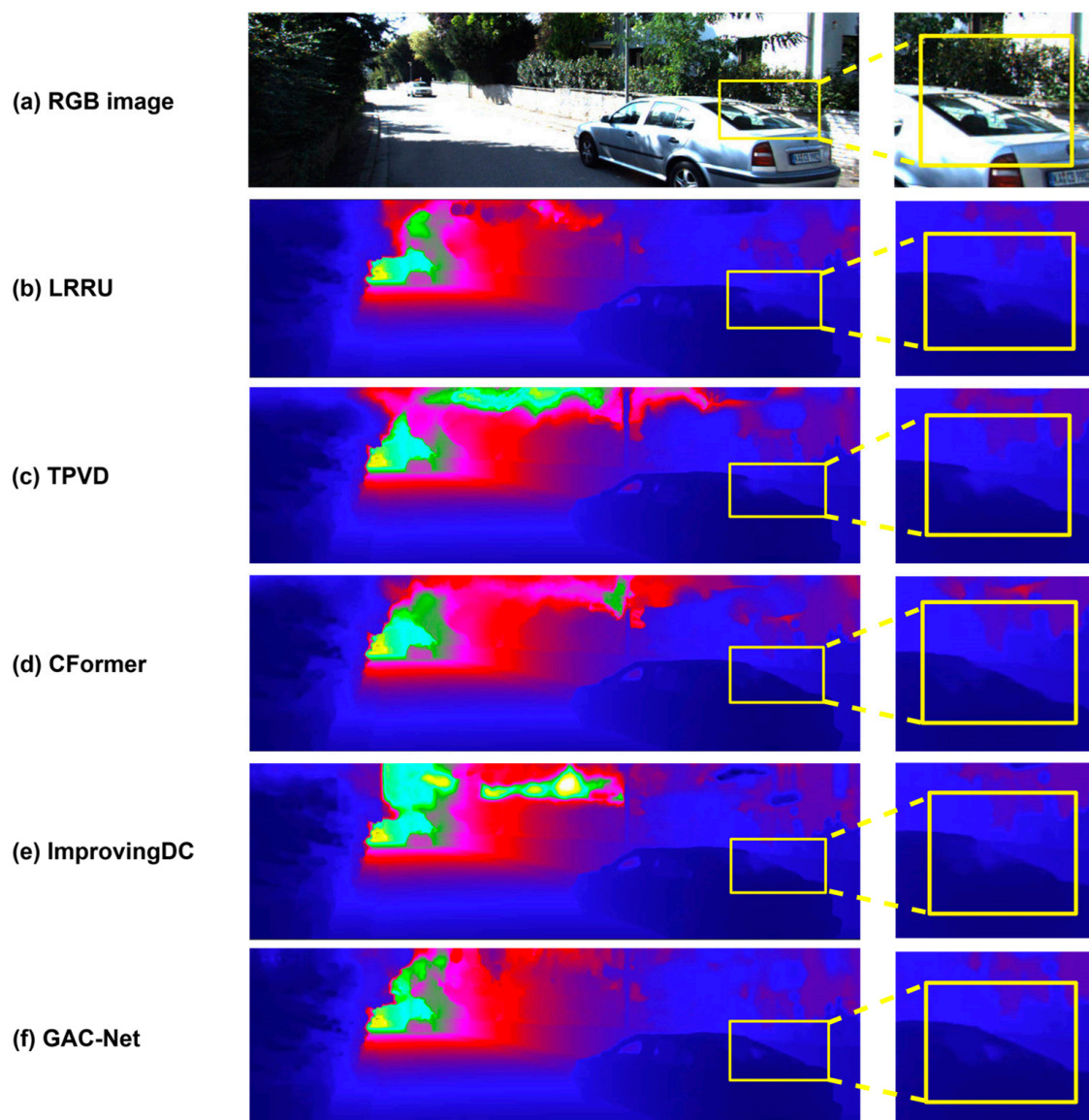


Figure 6. Qualitative comparison with state-of-the-art methods (LRRU [52], TPVD [45], CFormer [32] and ImprovingDC [50]) on preserving the structural completeness of objects, highlighting the reconstruction of full vehicles (e.g., rear windshield and boundary contours) under sparse input.

Figure 8 provides a direct comparison between GAC-Net and BPNet [26], verifying the practical improvements brought by our proposed modules. In the zoom-in regions, GAC-Net exhibits sharper, more continuous, and more complete reconstructions of edge structures such as street signs and tree trunks. By contrast, BPNet [26] suffers from elongated contours and edge spreading. These differences confirm that the incorporation of 3D geometric encoding and channel-aware fusion in GAC-Net significantly enhances its ability to reconstruct complex scene elements.

In summary, the differences observed in the zoom-in regions clearly demonstrate that GAC-Net achieves more complete, sharper, and geometrically consistent depth estimation results across various object types, including large foreground objects (e.g., vehicles), mid-scale structures (e.g., poles), and fine, distant elements (e.g., pedestrians).

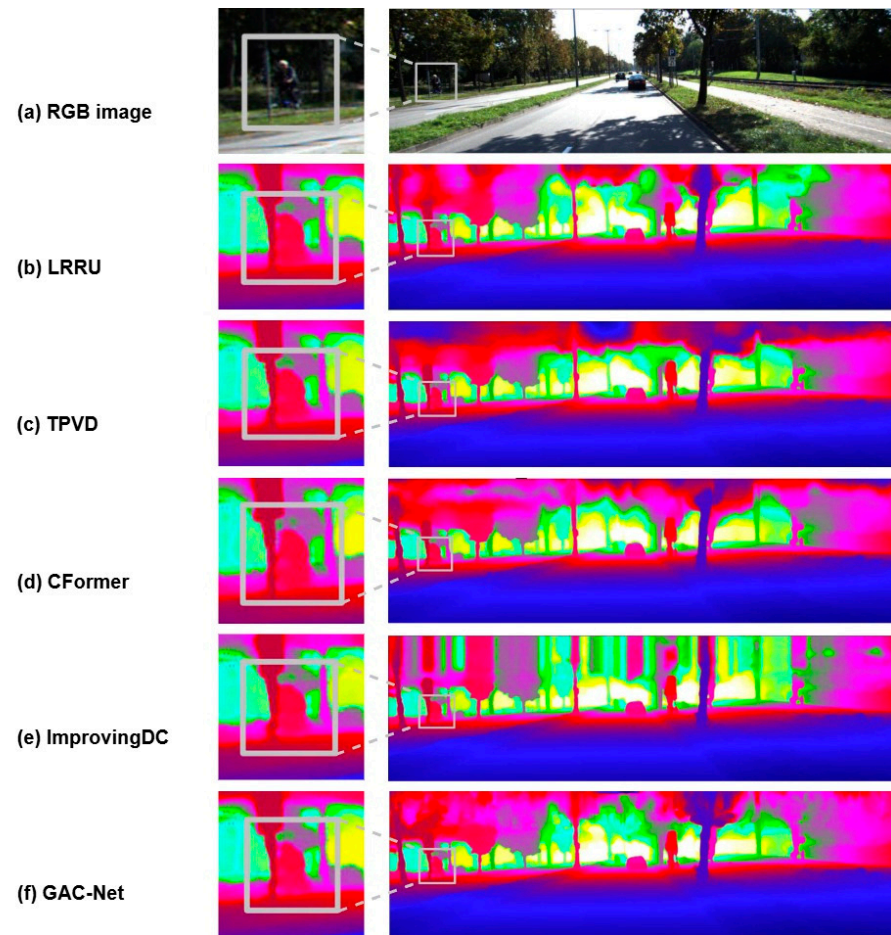


Figure 7. Visual comparison of local structural details among state-of-the-art methods (LRRU [52], TPVD [45], CFormer [32] and ImprovingDC [50]), with a focus on edge sharpness and boundary preservation for objects such as traffic signs and cyclists.

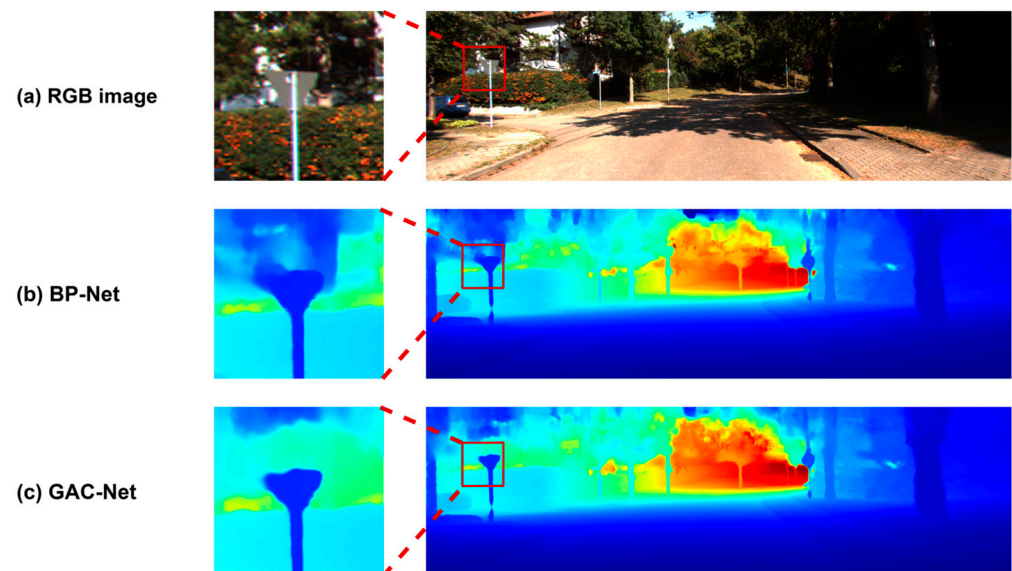


Figure 8. Visual comparison between BPNet [26] and our GAC-Net.

4.3. Ablation Studies

To comprehensively evaluate the contributions of each core module in GAC-Net, we conduct ablation studies on the KITTI validation set [18]. Specifically, we analyze the impact

of the geometric encoding module (PointNet++-S) and the channel-aware feature fusion module (CAFFM). The experimental results are summarized in Tables 2 and 3, respectively.

Table 2. Ablation studies on PointNet++-S.

GAC-Net	Without	PointNet++	PointNet++-S	RMSE	MAE
i	✓			719.08	204.36
ii		✓		714.34	200.12
iii			✓	709.4	195.63

Table 3. Ablation studies on CAFFM.

GAC-Net	Add	Concat	CAFFM	RMSE	MAE
i	✓			716.69	202.54
ii		✓		715.06	201.36
iii			✓	709.4	195.63

4.3.1. Effectiveness of PointNet++-S

Table 2 shows the performance gain brought by incorporating the PointNet++-S module. Starting from a baseline model (i) that only adopts the original multi-modal fusion pipeline without any 3D encoder or attention mechanism, we observe an RMSE of 719.08 mm and an MAE of 204.36 mm. By adding a standard PointNet++ encoder (ii), the performance improves to 714.34 mm RMSE and 200.12 mm MAE, confirming the benefit of introducing explicit 3D geometric priors. Furthermore, replacing the vanilla encoder with our proposed dual-branch PointNet++-S module (iii) leads to the best performance (RMSE: 711.40 mm, MAE: 197.63 mm), demonstrating the effectiveness of the scale-aware geometric extraction strategy.

4.3.2. Effectiveness of CAFFM

As shown in Table 3, we investigate different feature fusion strategies to validate the advantage of our CAFFM module. The baseline model (i) uses element-wise addition for combining local and contextual features, yielding 716.69 mm RMSE and 202.54 mm MAE. Switching to channel-wise concatenation (ii) slightly improves the performance to 715.06 mm RMSE and 201.36 mm MAE. Finally, applying our proposed channel-aware feature fusion module (CAFFM) (iii) further reduces the RMSE to 711.40 mm and MAE to 197.63 mm. This confirms that CAFFM enables richer feature interactions and enhances overall depth completion accuracy.

4.3.3. Visual Comparison of Ablation Results

To complement the quantitative results in Tables 2 and 3, Figure 9 provides visual comparisons of ablation experiments on representative KITTI samples. As shown on the left (A–D), the baseline model without PointNet++-S and CAFFM produces blurry boundaries, where the traffic sign is merged with the background. Adding PointNet++-S improves geometric consistency, but without CAFFM, the edges remain less distinct. The complete GAC-Net successfully separates the sign from the background, yielding sharper and clearer boundaries. On the right (E–H), a larger-scale scene further demonstrates that both PointNet++-S and CAFFM contribute to reconstructing the complete structure of the vehicle, with the full GAC-Net achieving the most accurate and visually coherent depth maps. These qualitative results are consistent with the quantitative improvements reported in the ablation study.

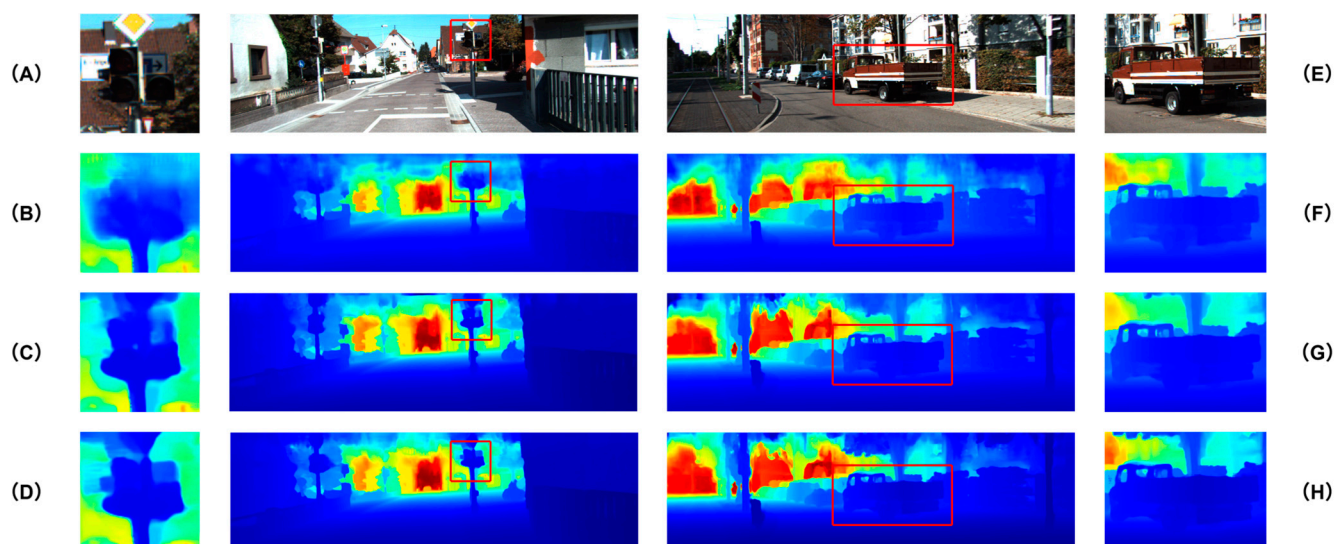


Figure 9. Visual comparison of ablation experiments on the KITTI validation set [18]. (A–D) show a local region around a traffic sign and its background, while (E–H) present a large-scale scene containing a vehicle. From top to bottom: RGB input image (A,E), baseline model without PointNet++-S and CAFFM (B,F), variant with PointNet++-S but simple concatenation instead of CAFFM (C,G), and the full GAC-Net with both PointNet++-S and CAFFM (D,H).

4.4. Sparsity Level Analysis

To assess the robustness of our method under varying input sparsity levels, we simulate different degrees of depth sparsity by uniformly sampling valid pixels from the original sparse depth maps. Following standard practice [26,45], the sparsity ratio is varied from 0.4 to 1.0, where 1.0 corresponds to the original input sparsity.

The results are presented in Figure 10, where our GAC-Net variants are compared against three representative methods: BPNet [26], TPVD [45], RigNet [35], and ACM-Net [40]. As expected, all methods suffer a noticeable performance drop in RMSE as the input becomes sparser. However, both versions of GAC-Net consistently outperform other approaches across all sparsity levels.

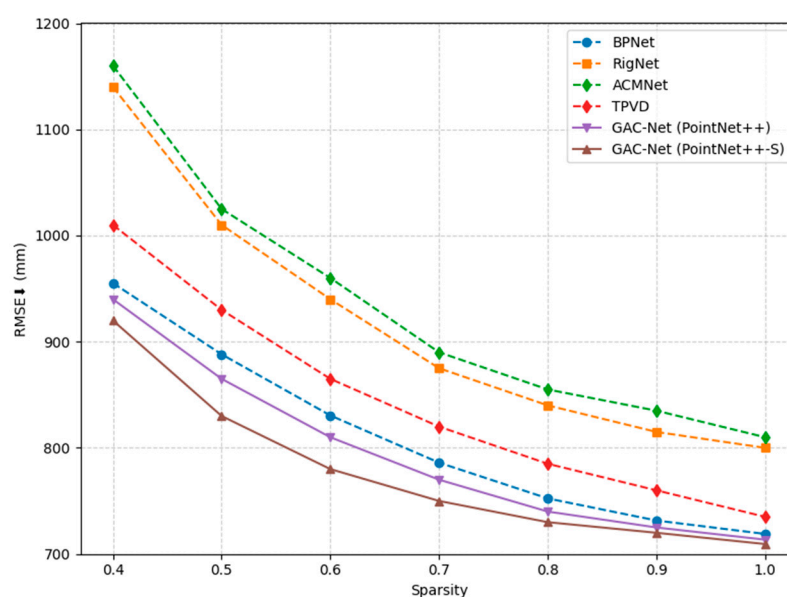


Figure 10. Comparison of RMSE (mm) across different input depth sparsity levels.

Notably, GAC-Net (PointNet++-S) achieves the lowest RMSE in every case, especially under more challenging sparsity conditions (e.g., 0.4–0.6). This performance validates the effectiveness of our scale-aware dual-branch PointNet++-S module, which captures both local geometric detail and global spatial structure. These results confirm that our method maintains higher accuracy and stronger generalization ability in sparse depth completion tasks under real-world constraints.

4.5. Complexity Analysis

To further evaluate the efficiency of our framework, we compare the model complexity of GAC-Net with BPNet [26] and the Transformer-based Dual-Attention Fusion model [46]. The results are summarized in Table 4. As shown, our method contains 91.07 M parameters, which is slightly higher than BPNet [26] (89.87 M) due to the additional PointNet++-S branch for 3D geometric feature extraction. This design, however, only introduces a marginal increase in inference time (0.17 s vs. 0.16 s), demonstrating that the added geometric modeling branch does not significantly impact efficiency.

Table 4. Comparison of model complexity.

Method	Publication	Params (M)	Inference Time (s)
BPNet [26]	CVPR 2024	89.87	0.160
Dual-Attention Fusion [46]	Sensors 2024	31.38	0.191
Ours	-	91.07	0.170

Compared with the Transformer-based Dual-Attention Fusion model, GAC-Net runs faster (0.17 s vs. 0.191 s), despite having more parameters (91.07 M vs. 31.38 M). This advantage mainly stems from the lightweight SE-based CAFFM fusion module, which effectively avoids the runtime overhead introduced by Transformer modules.

Overall, the results indicate that GAC-Net achieves a favorable balance between accuracy and efficiency. Although not strictly real-time, the design demonstrates that incorporating 3D modeling through the PointNet++-S branch and the SE-based CAFFM fusion module can be achieved with only marginal computational overhead.

5. Conclusions

This paper presented GAC-Net, a geometric attention fusion framework for image-guided depth completion. Building on the multi-scale, three-stage pipeline of BPNet, GAC-Net introduces two key modules: (i) a dual-branch PointNet++-S encoder that extracts scale-aware 3D geometric features from back-projected sparse point clouds and (ii) a Channel-Attention-Based Feature Fusion Module (CAFFM) that adaptively integrates geometric priors with RGB-depth features. These designs enhance explicit 3D modeling and cross-modal alignment, leading to improved preservation of both global structures and fine details. On the KITTI benchmark, GAC-Net demonstrates competitive performance among peer-reviewed methods, achieving accuracy improvements in both global structures and local details.

We further analyzed model complexity and showed that, while GAC-Net includes an additional 3D branch, the overall inference overhead remains marginal and indicates a favorable trade-off between accuracy and efficiency compared with Transformer-based baselines.

Limitations and future work:

- (1) Efficiency under deployment constraints. Although not strictly real-time, our results indicate that incorporating 3D modeling through the PointNet++-S branch

and the SE-based CAFFM fusion module can be achieved with only modest computational overhead. Future work will explore hardware-friendly acceleration (e.g., TensorRT, mixed precision), model compression (pruning/quantization/distillation), and lightweight design choices (e.g., streamlined multi-scale stages and operator fusion) to further reduce latency and memory footprint. We will also investigate efficient cross-modal reasoning modules (e.g., state-space models such as Mamba or linear-attention variants) under strict efficiency constraints.

- (2) Generalization beyond KITTI. To strengthen external validity, we plan to evaluate cross-dataset performance and domain transfer (e.g., additional outdoor/indoor benchmarks) and study robustness to sparsity, noise, transparency/reflectivity, and calibration shifts. In particular, recent benchmarks such as RSUD20K highlight the importance of dataset diversity and robustness evaluation under diverse environmental conditions, which will inspire our future exploration of generalization across challenging scenarios.
- (3) Temporal and system aspects. We will extend GAC-Net to sequential/streaming depth completion with temporal consistency and conduct system-level measurements (FLOPs and inference memory on diverse hardware, including embedded platforms), providing a more comprehensive view of computational costs for practical deployment.

Author Contributions: K.Z.: methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, visualization. C.L.: software. M.S. and L.Z.: writing—review and editing, supervision. X.G.: conceptualization, funding acquisition, project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Research and Development Program Project of Zhejiang Province (No. 2024C01109), the Scientific Research Innovation Fund for Postgraduates of Zhejiang University of Science and Technology (No. 2022yjskc08), and the General Scientific Research Project in Zhejiang Province (No. Y202249472).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xie, Z.; Yu, X.; Gao, X.; Li, K.; Shen, S. Recent Advances in Conventional and Deep Learning-Based Depth Completion: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 3395–3415. [\[CrossRef\]](#)
2. Huang, Z.; Lv, C.; Xing, Y.; Wu, J. Multi-Modal Sensor Fusion-Based Deep Neural Network for End-to-End Autonomous Driving With Scene Understanding. *IEEE Sens. J.* **2021**, *21*, 11781–11790. [\[CrossRef\]](#)
3. Song, Z.; Lu, J.; Yao, Y.; Zhang, J. Self-Supervised Depth Completion from Direct Visual-LiDAR Odometry in Autonomous Driving. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 11654–11665. [\[CrossRef\]](#)
4. Bai, L.; Zhao, Y.; Elhousni, M.; Huang, X. DepthNet: Real-Time LiDAR Point Cloud Depth Completion for Autonomous Vehicles. *IEEE Access* **2020**, *8*, 227825–227833. [\[CrossRef\]](#)
5. Gofer, E.; Praisler, S.; Gilboa, G. Adaptive LiDAR Sampling and Depth Completion Using Ensemble Variance. *IEEE Trans. Image Process.* **2021**, *30*, 8900–8912. [\[CrossRef\]](#)
6. Cui, Y.; Chen, R.; Chu, W.; Chen, L.; Tian, D.; Li, Y.; Cao, D. Deep Learning for Image and Point Cloud Fusion in Autonomous Driving: A Review. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 722–739. [\[CrossRef\]](#)
7. Su, S.; Wu, J. GeometryFormer: Semi-convolutional transformer integrated with geometric perception for depth completion in autonomous driving scenes. *Sensors* **2024**, *24*, 8066. [\[CrossRef\]](#)
8. Zou, N.; Xiang, Z.; Chen, Y.; Chen, S.; Qiao, C. Simultaneous Semantic Segmentation and Depth Completion with Constraint of Boundary. *Sensors* **2020**, *20*, 635. [\[CrossRef\]](#)

9. Jeong, Y.; Park, J.; Cho, D.; Hwang, Y.; Choi, S.B.; Kweon, I.S. Lightweight Depth Completion Network with Local Similarity-Preserving Knowledge Distillation. *Sensors* **2022**, *22*, 7388. [\[CrossRef\]](#)
10. Pan, J.; Zhong, S.; Yue, T.; Yin, Y.; Tang, Y. Multi-Task Foreground-Aware Network with Depth Completion for Enhanced RGB-D Fusion Object Detection Based on Transformer. *Sensors* **2024**, *24*, 2374. [\[CrossRef\]](#)
11. El-Yabroudi, M.Z.; Abdel-Qader, I.; Bazuin, B.J.; Abudayyeh, O.; Chabaan, R.C. Guided Depth Completion with Instance Segmentation Fusion in Autonomous Driving Applications. *Sensors* **2022**, *22*, 9578. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Zhai, D.-H.; Yu, S.; Wang, W.; Guan, Y.; Xia, Y. TCRNet: Transparent Object Depth Completion with Cascade Refinements. *IEEE Trans. Autom. Sci. Eng.* **2025**, *22*, 1893–1912. [\[CrossRef\]](#)
13. Wang, M.; Huang, R.; Liu, Y.; Li, Y.; Xie, W. suLPCC: A Novel LiDAR Point Cloud Compression Framework for Scene Understanding Tasks. *IEEE Trans. Ind. Inform.* **2025**, *21*, 3816–3827. [\[CrossRef\]](#)
14. Lu, H.; Xu, S.; Cao, S. SGTBN: Generating Dense Depth Maps from Single-Line LiDAR. *IEEE Sens. J.* **2021**, *21*, 19091–19100. [\[CrossRef\]](#)
15. Chen, Z.; Wang, H.; Wu, L.; Zhou, Y.; Wu, D. Spatiotemporal Guided Self-Supervised Depth Completion from LiDAR and Monocular Camera. In Proceedings of the 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), Shenzhen, China, 1–4 December 2020; pp. 54–57. [\[CrossRef\]](#)
16. Wang, Y.; Dai, Y.; Liu, Q.; Yang, P.; Sun, J.; Li, B. CU-Net: LiDAR Depth-Only Completion with Coupled U-Net. *IEEE Robot. Autom. Lett.* **2022**, *7*, 11476–11483. [\[CrossRef\]](#)
17. Fan, Y.-C.; Zheng, L.-J.; Liu, Y.-C. 3D Environment Measurement and Reconstruction Based on LiDAR. In Proceedings of the 2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Houston, TX, USA, 14–17 May 2018; pp. 1–4. [\[CrossRef\]](#)
18. Uhrig, J.; Schneider, N.; Schneider, L.; Franke, U.; Brox, T.; Geiger, A. Sparsity Invariant CNNs. In Proceedings of the 2017 International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; IEEE: New York, NY, USA, 2017; pp. 11–20. [\[CrossRef\]](#)
19. Zunair, H.; Khan, S.; Hamza, A.B. Rsud20K: A dataset for road scene understanding in autonomous driving. In Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 27–30 October 2024; IEEE: New York, NY, USA, 2024; pp. 708–714. [\[CrossRef\]](#)
20. Tang, J.; Tian, F.-P.; Feng, W.; Li, J.; Tan, P. Learning Guided Convolutional Network for Depth Completion. *IEEE Trans. Image Process.* **2021**, *30*, 1116–1129. [\[CrossRef\]](#)
21. Hu, M.; Wang, S.; Li, B.; Ning, S.; Fan, L.; Gong, X. PENet: Towards Precise and Efficient Image Guided Depth Completion. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 13656–13662. [\[CrossRef\]](#)
22. Cheng, X.; Wang, P.; Yang, R. Learning Depth with Convolutional Spatial Propagation Network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2361–2379. [\[CrossRef\]](#)
23. Cheng, X.; Wang, P.; Guan, C.; Yang, R. CSPN++: Learning Context and Resource Aware Convolutional Spatial Propagation Networks for Depth Completion. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10615–10622. [\[CrossRef\]](#)
24. Lin, Y.; Cheng, T.; Zhong, Q.; Zhou, W.; Yang, H. Dynamic Spatial Propagation Network for Depth Completion. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 1638–1646.
25. Liu, L.; Liao, Y.; Wang, Y.; Geiger, A.; Liu, Y. Learning Steering Kernels for Guided Depth Completion. *IEEE Trans. Image Process.* **2021**, *30*, 2850–2861. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Tang, J.; Tian, F.-P.; An, B.; Li, J.; Tan, P. Bilateral Propagation Network for Depth Completion. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 9763–9772. [\[CrossRef\]](#)
27. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
28. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 5099–5108.
29. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
30. Imran, S.; Long, Y.; Liu, X.; Morris, D. Depth Coefficients for Depth Completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 12438–12447. [\[CrossRef\]](#)
31. Ma, F.; Karaman, S. Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 4796–4803. [\[CrossRef\]](#)

32. Zhang, Y.; Guo, X.; Poggi, M.; Zhu, Z.; Huang, G.; Mattoccia, S. CompletionFormer: Depth Completion with Convolutions and Vision Transformers. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; IEEE: New York, NY, USA, 2023; pp. 18527–18536. [\[CrossRef\]](#)
33. Park, J.; Joo, K.; Hu, Z.; Liu, C.K.; So Kweon, I. Non-local Spatial Propagation Network for Depth Completion. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 120–136.
34. Chen, D.; Huang, T.; Song, Z.; Deng, S.; Jia, T. Agg-Net: Attention Guided Gated-Convolutional Network for Depth Image Completion. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; IEEE: New York, NY, USA, 2023; pp. 8853–8862. [\[CrossRef\]](#)
35. Yan, Z.; Wang, K.; Li, X.; Zhang, Z.; Li, J.; Yang, J. RigNet: Repetitive Image Guided Network for Depth Completion. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; Springer Nature: Cham, Switzerland, 2022; pp. 214–230.
36. Yan, Z.; Li, X.; Hui, L.; Zhang, Z.; Li, J.; Yang, J. Rignet++: Semantic Assisted Repetitive Image Guided Network for Depth Completion. *arXiv* **2023**, arXiv:2309.00655. [\[CrossRef\]](#)
37. Nazir, D.; Pagani, A.; Liwicki, M.; Stricker, D.; Afzal, M.Z. SemAttNet: Toward Attention-Based Semantic Aware Guided Depth Completion. *IEEE Access* **2022**, *10*, 120781–120791. [\[CrossRef\]](#)
38. Qiu, J.; Cui, Z.; Zhang, Y.; Zhang, X.; Liu, S.; Zeng, B.; Pollefeys, M. DeepLiDAR: Deep Surface Normal Guided Depth Prediction for Outdoor Scene from Sparse LiDAR Data and Single Color Image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3308–3317. [\[CrossRef\]](#)
39. Xu, Y.; Zhu, X.; Shi, J.; Zhang, G.; Bao, H.; Li, H. Depth Completion from Sparse LiDAR Data with Depth-Normal Constraints. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2811–2820. [\[CrossRef\]](#)
40. Zhao, S.; Gong, M.; Fu, H.; Tao, D. Adaptive Context-Aware Multi-Modal Network for Depth Completion. *IEEE Trans. Image Process.* **2021**, *30*, 5264–5276. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Liu, X.; Shao, X.; Wang, B.; Li, Y.; Wang, S. Graphcspn: Geometry-aware depth completion via dynamic GCNs. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer Nature: Cham, Switzerland, 2022; pp. 90–107.
42. Chen, Y.; Yang, B.; Liang, M.; Urtasun, R. Learning Joint 2D-3D Representations for Depth Completion. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10022–10031.
43. Yu, Z.; Sheng, Z.; Zhou, Z.; Luo, L.; Cao, S.-Y.; Gu, H.; Zhang, H.; Shen, H.-L. Aggregating Feature Point Cloud for Depth Completion. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 8698–8709.
44. Zhou, W.; Yan, X.; Liao, Y.; Lin, Y.; Huang, J.; Zhao, G.; Cui, S.; Li, Z. BEV@DC: Bird’s-Eye View Assisted Training for Depth Completion. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 9233–9242.
45. Yan, Z.; Lin, Y.; Wang, K.; Zheng, Y.; Wang, Y.; Zhang, Z.; Li, J.; Yang, J. Tri-Perspective View Decomposition for Geometry-Aware Depth Completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 4874–4884.
46. Wang, S.; Jiang, F.; Gong, X. A Transformer-Based Image-Guided Depth-Completion Model with Dual-Attention Fusion Module. *Sensors* **2024**, *24*, 6270. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Shi, Y.; Singh, M.K.; Cai, H.; Porikli, F. DeCoTR: Enhancing Depth Completion with 2D and 3D Attentions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 10736–10746.
48. Xie, G.; Zhang, Y.; Jiang, Z.; Liu, Y.; Xie, Z.; Cao, B.; Liu, H. HTMNet: A Hybrid Network with Transformer-Mamba Bottleneck Multimodal Fusion for Transparent and Reflective Objects Depth Completion. *arXiv* **2025**, arXiv:2505.20904. [\[CrossRef\]](#)
49. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
50. Wang, Y.; Zhang, G.; Wang, S.; Li, B.; Liu, Q.; Hui, L.; Dai, Y. Improving Depth Completion via Depth Feature Upsampling. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; IEEE: New York, NY, USA, 2024; pp. 21104–21113. [\[CrossRef\]](#)
51. Imran, S.; Liu, X.; Morris, D. Depth Completion with Twin Surface Extrapolation at Occlusion Boundaries. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; IEEE: New York, NY, USA, 2021; pp. 2583–2592. [\[CrossRef\]](#)

52. Wang, Y.; Li, B.; Zhang, G.; Liu, Q.; Gao, T.; Dai, Y. LRRU: Long-Short Range Recurrent Updating Networks for Depth Completion. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; IEEE: New York, NY, USA, 2023; pp. 9388–9398. [\[CrossRef\]](#)
53. Huynh, L.; Nguyen, P.; Matas, J.; Rahtu, E.; Heikkilä, J. Boosting Monocular Depth Estimation with Lightweight 3D Point Fusion. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 12747–12756. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.