

Article

Improved DPT-Hybrid for Monocular Depth Estimation with Geometry-Enhanced Encoding and Structure-Aware Gated Fusion

Wei Liu ^{1,2,*}, Shilei Hu ¹, Yi Qin ^{1,2}, Shengkai Hong ¹ and Dehua Zhang ³ ¹ School of Intelligent Manufacturing and Electrical Engineering, Nanyang Normal University, Nanyang 473061, China² Collaborative Innovation Center of Intelligent Explosion-proof Equipment, Nanyang Normal University, Nanyang 473061, China³ School of Artificial Intelligence, Henan University, Kaifeng 475004, China

* Correspondence: 20141010@nynu.edu.cn; Tel.: +86-18637716316

Abstract

Monocular depth estimation aims to recover dense 3D scene geometry from a single RGB image and plays an important role in autonomous driving, robotic perception, augmented reality, and 3D reconstruction. Although Transformer-based dense prediction models have achieved strong performance, existing DPT-Hybrid frameworks still suffer from three limitations: insufficient local geometric modeling in shallow stages, inadequate cross-scale fusion for preserving fine structures, and training objectives that only weakly constrain structural consistency. To address these issues, we propose a structure-aware enhanced DPT-Hybrid framework. First, a geometry-enhanced encoder introduces lightweight depth-wise separable convolution branches into shallow Transformer stages to better capture local edge and texture cues while preserving global contextual modeling. Second, a Structure-Aware Cross-Scale Gated Attention Fusion (S-GAF) module is proposed to improve decoder-side feature aggregation by jointly modeling channel-wise and spatial importance with an auxiliary RGB-gradient input. Third, joint structure–geometric consistency loss combines scale-invariant logarithmic loss, gradient consistency loss, and edge-focused loss to improve pixel-level accuracy, geometric plausibility, and boundary sharpness. Experiments on NYUv2 and KITTI demonstrate that the proposed method achieves lower single-run error metrics than the controlled DPT-Hybrid baseline under the evaluated settings. On NYUv2, our method achieves an absolute relative error (AbsRel) of 0.099 and an RMSE of 0.334. On KITTI, it achieves an AbsRel of 0.058 and an RMSE of 2.455. The proposed method introduces only modest additional complexity while producing more accurate and structurally sharper depth predictions.



Academic Editor: George A. Tsihrintzis

Received: 24 June 2026

Revised: 3 August 2026

Accepted: 4 August 2026

Published: 5 August 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: monocular depth estimation; dense prediction transformer; structure-aware fusion; gated attention; geometric consistency

1. Introduction

Monocular depth estimation aims to infer dense per-pixel scene depth from a single RGB image without relying on active sensors such as LiDAR or structured light cameras [1,2]. Because it avoids the hardware cost and deployment constraints associated with active sensing systems, it has become an important component in autonomous driving, robotic navigation, mobile augmented reality, 3D scene reconstruction, and other real-time

or resource-constrained perception systems [3–6]. However, the task remains inherently ill-posed because a 2D image does not explicitly encode the depth dimension of the observed 3D scene [7,8]. Accurate estimation therefore depends on effectively learning semantic context, geometric priors, and spatial layout regularities from visual data.

Progress in hierarchical representation learning has improved reconstruction, compression, and recognition across large-scale signal processing and computer vision tasks. For instance, high-order Dual Hahn polynomial methods enable efficient large-signal processing by retaining informative structures while reducing redundancy. This principle also applies to real-time dense prediction: models must preserve high-frequency geometric details without depending solely on larger backbones for autonomous navigation, augmented reality, and robotic perception. In monocular depth estimation, the corresponding challenge is to efficiently coordinate global semantic context with local boundaries, thin structures, and depth discontinuities.

Recent advances in Transformer architectures have substantially improved dense prediction tasks [9,10]. Among them, the Dense Prediction Transformer (DPT) has emerged as a strong framework for monocular depth estimation [3]. By using a Transformer-based encoder, the DPT can capture long-range dependencies and global scene semantics more effectively than conventional convolutional architectures. As a result, the DPT and its variants have achieved strong performance on widely used benchmarks such as the New York University Depth Dataset v2 (NYUv2) and the Karlsruhe Institute of Technology and Toyota Technological Institute Dataset (KITTI) [3,11,12].

Nevertheless, the vanilla DPT-Hybrid architecture still exhibits several limitations that affect prediction quality: First, shallow high-resolution stages rely heavily on Transformer representations and lack explicit local geometric modeling. This weakens the encoding of edges, textures, and fine structures, leading to blurred boundaries and degraded fine details in predicted depth maps [13,14]. Second, decoders commonly use simple fusion operations (e.g., concatenation or element-wise addition) for cross-scale aggregation. These operations fail to explicitly distinguish between fine spatial details carried by encoder features and progressively refined semantic context from decoder features [15,16]. Third, the commonly used scale-invariant depth loss mainly optimizes numerical regression accuracy and does not directly enforce structural sharpness or geometric consistency, making predictions prone to over-smoothing near depth discontinuities [1,17,18].

Several studies have attempted to improve DPT-style or dense prediction depth models. Some works strengthened local feature extraction through convolutional enhancement modules, whereas others incorporated attention-based fusion mechanisms or edge-aware supervision [19–24]. Although these efforts alleviated some of the above weaknesses, a unified design that jointly improves shallow local geometry modeling, structure-aware cross-scale fusion, and geometry-sensitive optimization remains underexplored.

To address these issues, we propose a structure-aware enhanced DPT-Hybrid framework that jointly improves the encoder, decoder fusion strategy, and training objective, achieving a better balance between global context modeling and local structure preservation. Specifically, the first two shallow stages of the DPT-Hybrid encoder are augmented with lightweight geometry-aware convolution branches; the decoder employs a Structure-Aware Cross-Scale Gated Attention Fusion (S-GAF) module guided by auxiliary image-gradient cues; and the training objective combines scale-invariant depth regression with gradient and edge consistency constraints.

The main contributions of this work are summarized as follows:

- We present a coordinated structure enhancement framework for DPT-Hybrid monocular depth estimation. The framework is motivated by the observation that local geometric

cues may be weakened during shallow encoding, diluted during encoder–decoder fusion, and insufficiently constrained by value-oriented regression losses.

- We implement this framework through three complementary mechanisms: residual depth-wise separable refinement in shallow high-resolution encoder stages, gradient-assisted adaptive gating for encoder–decoder fusion, and joint supervision of depth values, gradient transitions, and boundaries. These mechanisms align structural preservation across feature representation, feature fusion, and optimization.
- We conduct controlled component-level ablations on NYUv2 to quantify the individual and complementary effects of the proposed design. The complete framework obtains lower error metrics than the controlled DPT-Hybrid baseline on both NYUv2 and KITTI, with a moderate increase in the number of parameters and computational cost.

It should be emphasized that depth-wise separable convolution, channel–spatial gating, RGB-gradient extraction, and gradient-based supervision are not novel in isolation; the methodological contribution of this work lies in their task-specific coordination and placement within the DPT-Hybrid pipeline, rather than in proposing new primitive operators.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 presents the proposed method, including the geometry-enhanced encoder, the S-GAF module, and the joint loss design. Section 4 reports the experimental settings, quantitative comparisons, ablation studies, efficiency analysis, and qualitative results. Section 5 discusses the mechanistic interpretation, limitations, and future directions, and Section 6 concludes the paper.

2. Related Work

Monocular depth estimation is a fundamental yet ill-posed problem in computer vision. Early non-deep learning methods mainly relied on handcrafted cues and geometric assumptions, such as perspective, texture gradients, shading, occlusion, and piece-wise planar constraints [2,7]. Although these methods provided useful geometric intuition, their performance was often limited by weak generalization and strong dependence on scene-specific priors. With the development of deep learning, CNN-based methods enabled end-to-end depth prediction from RGB images and substantially improved estimation accuracy [1,8]. Representative works such as those of Eigen et al. and Laina et al. and DORN, BTS, and AdaBins progressively improved depth estimation through multi-scale learning, stronger backbones, and enhanced regression strategies [1,13,14,25]. More recently, Transformer-based architectures have further advanced dense prediction by modeling long-range dependencies more effectively [9,10]. Among them, the DPT is a representative framework that adapts Transformer features to dense prediction and has demonstrated strong performance on benchmarks such as NYUv2 and KITTI [3,11,12]. In particular, DPT-Hybrid combines convolutional and Transformer representations and provides a favorable balance between local inductive bias and global contextual reasoning [26]. However, its shallow feature extraction remains insufficiently sensitive to local geometric structures, and its feature fusion strategy remains relatively generic for recovering fine depth boundaries.

Feature fusion plays a central role in dense prediction because accurate estimation requires both high-level semantic context and low-level spatial details. Architectures such as U-Net, FPN, and RefineNet have demonstrated the effectiveness of top-down fusion and multi-scale feature aggregation [15,27]. In addition, lightweight attention modules such as SENet, CBAM, and ECA improve feature recalibration by emphasizing informative channels or spatial responses [19–21]. However, these designs are not specifically tailored to the structural requirements of monocular depth estimation. FPN-style modules offer a

general top-down propagation path, while CBAM and ECA only perform intra-feature recalibration rather than structure-guided cross-scale fusion between encoders and decoders. In contrast, our S-GAF module is tailored for depth estimation: it introduces auxiliary RGB-gradient cues into selective fusion to better preserve boundaries and thin structures.

Loss design is equally important for monocular depth estimation. Common objectives such as L_1 , L_2 , BerHu, ordinal regression, and SI-Log mainly focus on pixel-wise depth accuracy [1,3,13,14]. Additional supervision terms such as ranking loss, surface-normal loss, and virtual normal loss can improve geometric consistency [17,18,28]; however, they either increase sampling complexity or rely on accurate normal estimation, and such estimation is sensitive to annotation noise and boundary ambiguity. Since our goal is to improve DPT-Hybrid with limited overhead while strengthening local structure preservation, we instead adopt gradient consistency loss and edge-focused loss as complements to SI-Log loss. These terms more directly constrain local depth transitions and discontinuity sharpness, which is highly consistent with the design objectives of the geometry-enhanced encoder and S-GAF.

In summary, existing monocular depth estimation frameworks suffer from three interconnected limitations: First, shallow feature processing often attenuates high-frequency geometric cues such as edges and fine structures, and balancing local details with global context remains challenging. Second, conventional top-down decoders rely on simple addition or concatenation for cross-scale fusion, without explicitly distinguishing the complementary roles of encoder details and decoder semantics. Third, standard pixel-wise regression losses prioritize numerical accuracy and provide weak supervision at depth discontinuities, frequently leading to over-smoothed boundaries. To address these limitations, our method introduces lightweight residual depth-wise separable convolutions into the shallow encoder stages to enhance local geometric representation. It further employs gradient-assisted cross-scale gated fusion (S-GAF) to adaptively regulate the aggregation of encoder and decoder features using boundary-related cues. Finally, joint structure–geometric consistency loss incorporates spatial gradient and edge supervision to improve structural consistency in the predicted depth maps. Table 1 provides a structured comparison between representative existing paradigms and the proposed method.

Table 1. Summary of representative existing methodologies, their primary limitations, and our proposed solutions. Note that the listed limitations represent common tendencies rather than universal characteristics of all methods within each category.

Method Category	Representative Works	Core Mechanism	Primary Limitations Drawbacks	Our Proposed
Multi-Scale CNNs	Eigen et al. [1], Laina et al. [13], DORN [14]	Multi-scale convolutional feature extraction and ordinal regression.	Limited receptive field; struggles to capture long-range global scene semantics.	Adopt DPT-Hybrid backbone to capture robust long-range global contextual dependencies.
Dense Transformers	DPT [3], Swin [10], NeW CRFs [29]	Self-attention mechanisms for dense prediction and windowed CRFs.	Shallow stages lose high-frequency local geometric details due to patchification.	Geometry-enhanced (GE) encoder: Inject residual DWConv branches into shallow Stages 1 and 2.
Attention Fusion	FPN [15], CBAM [19], ECA [20], AdaBins [25]	Top-down feature pyramids and intra-feature channel–spatial recalibration.	Lacks explicit cross-scale gating between encoder details and decoder semantics.	S-GAF module: Gradient-assisted cross-scale channel–spatial soft-selection gating.
Geometric Losses	Virtual Normal [17], Ranking Loss [28]	Surface-normal constraints and pair-wise ordinal ranking supervision.	High sampling complexity; sensitive to annotation noise and normal estimation errors.	Joint consistency loss: Direct spatial gradient consistency and Sobel edge-focused supervision.

Structure-sensitive representations have also been explored in computer-assisted medical image analysis. For example, quaternion orthogonal Rademacher–Fourier moments have been combined with deep learning for cancer disease classification [30]. Although that

study addressed image-level medical classification rather than dense depth regression, it illustrates the potential value of integrating explicit structural descriptors with learned representations. In contrast, our method learns local geometric and boundary-related cues end to end through the geometry-enhanced shallow encoder and the gradient-assisted fusion mechanism.

Beyond dense visual prediction, computationally efficient stacked recurrent architectures have also been investigated for sentiment classification [31]. Although that study addressed a different task and modality, it reflects the broader interest in balancing representation capacity and computational efficiency. Because its methodology was not designed for cross-scale visual feature fusion, it is cited here only as a general example of efficient machine learning architecture design.

Overall, following the development trajectory from general monocular depth estimation methods to DPT-based dense prediction, our work ultimately focuses on addressing the specific weaknesses of DPT-Hybrid through a coordinated set of improvements in shallow geometric feature extraction, structure-aware cross-scale fusion, and structure-sensitive optimization.

3. Method

3.1. Overall Network Architecture

The proposed method follows an encoder–decoder architecture for monocular depth estimation. It consists of three main components: a geometry-enhanced DPT encoder for global and local feature extraction, a decoder equipped with Structure-Aware Cross-Scale Gated Attention Fusion (S-GAF) modules for adaptive multi-scale aggregation, and joint structure–geometric consistency loss for training. The overall pipeline is shown in Figure 1.

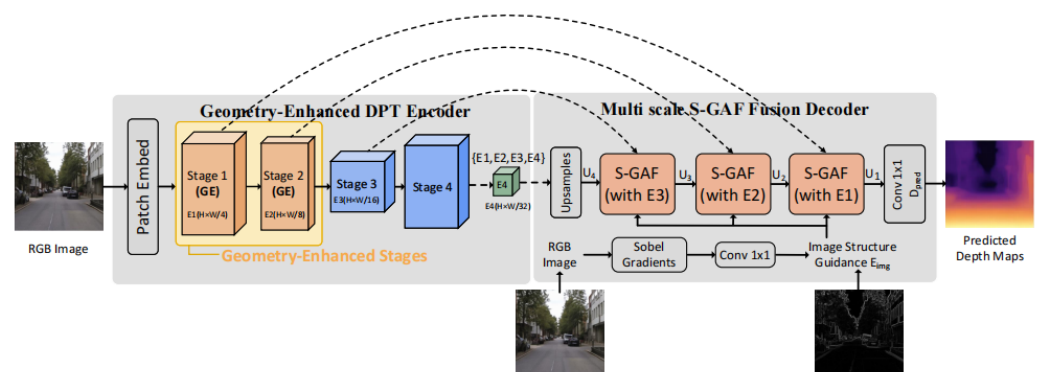


Figure 1. Overall architecture of the proposed structure-aware DPT-Hybrid framework. The network consists of a geometry-enhanced encoder, a top-down decoder with S-GAF modules, and joint structure–geometric consistency loss for training.

Given an input RGB image $I \in \mathbb{R}^{3 \times H \times W}$, the encoder extracts four hierarchical feature maps:

$$\{E_1, E_2, E_3, E_4\} = \mathcal{F}_{\text{enc}}(I), \quad (1)$$

where $\mathcal{F}_{\text{enc}}(\cdot)$ denotes the geometry-enhanced encoder. The first two shallow stages are augmented with lightweight convolutional branches to improve local structure modeling, whereas the deeper stages retain the original Transformer-dominant design for global semantic reasoning [3,10].

The decoder adopts a top-down progressive upsampling strategy. Starting from the deepest feature map E_4 , the decoder successively upsamples the current feature and fuses it with the corresponding encoder feature from the previous scale:

$$U_i = \mathcal{F}_{\text{S-GAF}}(\uparrow U_{i+1}, E_i), \quad i \in \{1, 2, 3\}, \quad (2)$$

where $\uparrow$ denotes bilinear upsampling and $\mathcal{F}_{\text{S-GAF}}(\cdot)$ denotes the structure-aware gated fusion function. Compared with simple concatenation or addition, the proposed fusion module adaptively weights semantic and detail features under an auxiliary RGB-gradient cue [15,19,20]. The final decoder feature U_1 is projected to a one-channel depth map by a 1×1 convolution:

$$d = \sigma(\mathcal{F}_{\text{conv}}(U_1)), \quad (3)$$

where $\sigma(\cdot)$ denotes the Softplus activation function, $\sigma(x) = \log(1 + \exp(x))$, which is applied to the final output to ensure a non-negative depth prediction. During training, the predicted depth map d is supervised by the proposed joint loss:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{si}} + \lambda_2 \mathcal{L}_{\text{grad}} + \lambda_3 \mathcal{L}_{\text{edge}}. \quad (4)$$

3.2. Geometry-Enhanced DPT Encoder

Although DPT-Hybrid provides strong global context modeling, its shallow stages are less effective at capturing fine local geometric details such as edges, textures, and thin structures [3,13]. To address this issue, we augment the first two shallow stages with lightweight geometry-aware enhancement branches. The design principle is straightforward: improve shallow local representation while preserving the efficiency and global modeling capability of the original backbone.

We specifically select a residual depth-wise separable convolution (DWConv) branch over alternative localized designs (such as local window attention or standard convolutions) for two compelling reasons: First, unlike local window attention, which computes data-dependent dot-product similarities within patches, DWConv offers explicit shift-invariant spatial filtering priors. By applying local kernels directly to high-resolution features, DWConv introduces a strong spatial inductive bias and learns filters sensitive to sharp boundaries and fine textures without requiring heavy data-driven pattern memorization. Second, compared to standard 3×3 convolutions that perform dense spatial and channel mixing simultaneously, DWConv decouples spatial filtering from channel mixing. As shown in Section 4.5.3, the depth-wise separable variant obtains lower reported errors than the tested standard-convolution variant under the evaluated setting. In addition, depth-wise separable convolution is theoretically more parameter-efficient than dense convolution for the same input/output channel dimensions and kernel size. The adopted residual branch therefore provides a practical local refinement mechanism while preserving the original backbone pathway.

The structure of the proposed geometry-enhanced shallow stage is illustrated in Figure 2.

Let E'_k denote the output feature of the k -th shallow Transformer stage, where $k \in \{1, 2\}$. We refine it using a depth-wise separable convolution branch:

$$E_k = E'_k + \text{DWConv}_{3 \times 3}(\text{LN}(E'_k)), \quad k = 1, 2, \quad (5)$$

where $\text{LN}(\cdot)$ denotes layer normalization and depth-wise separable convolution 3×3 (DWConv) denotes a 3×3 depth-wise separable convolution. This residual enhancement allows local edge and texture cues to be injected into the shallow Transformer features with limited overhead [32,33].

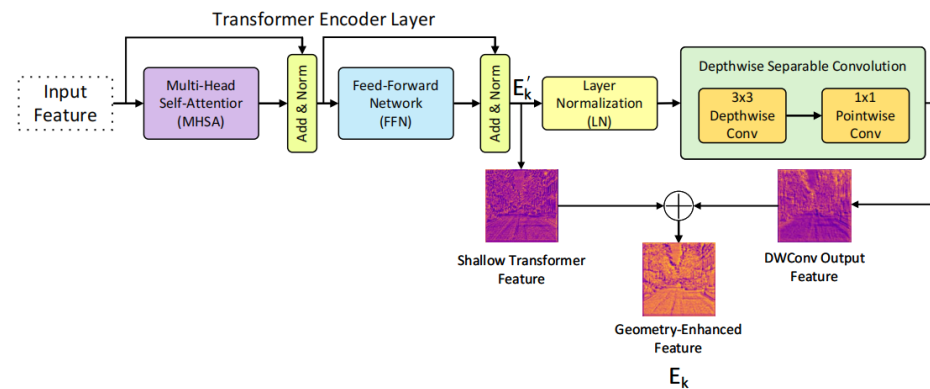


Figure 2. Structure of the geometry-enhanced shallow stage. A lightweight depth-wise separable convolution branch is used to enhance local geometric representation and is fused with the Transformer feature through a residual connection.

For the deeper stages, we keep the original Transformer blocks unchanged:

$$E_3 = E'_3, \quad E_4 = E'_4. \quad (6)$$

In this way, the encoder outputs hierarchical features that combine local geometric detail with high-level contextual information. Because the enhancement is restricted to shallow high-resolution stages, the overall increase in the number of parameters and computation remains modest.

3.3. Structure-Aware Cross-Scale Gated Attention Fusion (S-GAF)

The decoder in the vanilla DPT-Hybrid uses relatively simple cross-scale fusion operations, which may not sufficiently distinguish the roles of encoder semantics and decoder details [3,15]. To improve structure preservation during decoding, we propose the Structure-Aware Cross-Scale Gated Attention Fusion (S-GAF) module.

The overall architecture of the proposed S-GAF module is shown in Figure 3.

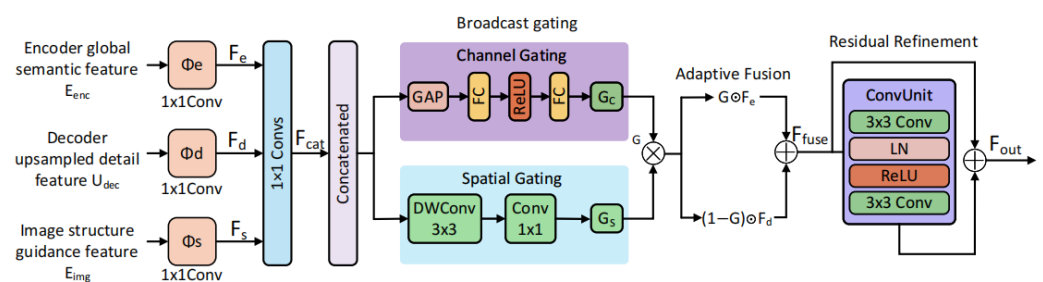


Figure 3. Structure of the proposed S-GAF module. The module takes an encoder feature, an upsampled decoder feature, and an auxiliary RGB-gradient feature as inputs, and performs adaptive channel–spatial gated fusion followed by residual refinement.

The module takes three inputs: the encoder feature E_{enc} , the upsampled decoder feature U_{dec} , and an auxiliary structure representation feature E_{img} extracted from RGB image gradients. We emphasize that the RGB-gradient cue is integrated via gradient-assisted fusion rather than acting as a hard top-down geometric prior. RGB gradients respond to both genuine depth discontinuities and non-depth visual changes (e.g., textures, illumination shifts and shadows). Accordingly, E_{img} is projected and concatenated as an equal structural input to the gating mechanism. This allows the learned channel and spatial gates to dynamically assess the reliability of local gradient cues, amplifying true boundary features while suppressing texture-induced noise during cross-scale feature aggregation [17,28].

To align the feature dimensions, we first project all three inputs to the same channel dimension:

$$F_e = \Phi_e(E_{\text{enc}}), \quad F_d = \Phi_d(U_{\text{dec}}), \quad F_s = \Phi_s(E_{\text{img}}), \quad (7)$$

where Φ_e , Φ_d , and Φ_s are 1×1 convolutions. The aligned features are concatenated along the channel dimension:

$$F_{\text{cat}} = \text{cat}(F_e, F_d, F_s). \quad (8)$$

A channel gate is then computed by global average pooling followed by two fully connected layers:

$$G_c = \sigma(\text{FC}_2(\text{ReLU}(\text{FC}_1(\text{GAP}(F_{\text{cat}}))))), \quad (9)$$

where $\text{GAP}(\cdot)$ denotes global average pooling and $\sigma(\cdot)$ is the Sigmoid activation function. In parallel, a spatial gate is computed as

$$G_s = \sigma(\text{Conv}_{1 \times 1}(\text{DWConv}_{3 \times 3}(F_{\text{cat}}))). \quad (10)$$

The final joint gate is obtained by broadcasted element-wise multiplication:

$$G = G_c \odot G_s. \quad (11)$$

Using the joint gate, the encoder and decoder features are fused adaptively:

$$F_{\text{fuse}} = G \odot F_e + (1 - G) \odot F_d, \quad (12)$$

where $\odot$ denotes element-wise multiplication with broadcasting where necessary. In this formulation, the module can preserve encoder-side spatial details near structural boundaries while relying more strongly on decoder-side semantic context in smooth or visually ambiguous regions. A residual refinement block is then applied:

$$F_{\text{out}} = F_{\text{fuse}} + \text{ConvUnit}(F_{\text{fuse}}), \quad (13)$$

In our implementation, $\text{ConvUnit}(\cdot)$ is instantiated as a residual convolutional unit containing two 3×3 convolutions, each with stride 1 and padding 1. A ReLU activation (`inplace=False`) is applied before each convolution, and the output of the second convolution is added to the unit input through a residual connection. For the depth model, no normalization layer and no dropout layer are used in this unit.

Overall, the S-GAF module improves cross-scale aggregation by jointly considering semantics, details, and auxiliary structural cues. This helps preserve clear boundaries and thin structures while maintaining computational efficiency [19–21].

Despite its structure-preserving capability, S-GAF has several methodological limitations: First, the auxiliary RGB-gradient feature is an appearance-based structural cue rather than a direct measurement of scene geometry. Consequently, strong texture patterns, shadows, specular reflections, and illumination changes may generate image gradients that do not correspond to actual depth discontinuities. Although learned gates partially suppress this interference, they cannot fully separate appearance edges from geometric boundaries. Second, the multi-branch channel–spatial gating introduces extra feature projection, concatenation, and sequential kernel execution. This causes moderate run-time overhead, even though the increase in theoretical FLOPs is small. Third, the current fusion mechanism operates deterministically and does not explicitly estimate the uncertainty or reliability of the structural guidance. As a result, its performance may degrade in textureless regions, under severe motion blur or poor illumination, and on transparent, reflective, or non-Lambertian surfaces. These weaknesses are examined further in Section 5.2, while corresponding methodological extensions are discussed in Section 5.3.

3.4. Joint Structure–Geometric Consistency Loss

To improve both numerical accuracy and structural quality, we use a joint loss that combines pixel-level depth regression, geometric gradient consistency, and edge preservation [1,17,18]. Figure 4 illustrates the composition of the proposed joint structure-geometric consistency loss, which consists of three complementary optimization terms. All loss terms are computed only on valid pixels. For gradient- and Sobel-based terms, pixels whose local convolution windows contain invalid ground-truth values are excluded from loss computation to avoid contamination from invalid regions.

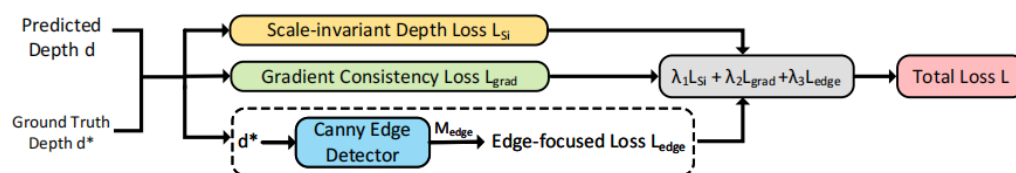


Figure 4. Composition of the proposed joint structure–geometric consistency loss. The joint loss consists of three complementary terms: scale-invariant logarithmic loss (SI-Log loss) for pixel-level accuracy constraints, gradient consistency loss for geometric smoothness constraints, and edge-focused loss for boundary sharpness constraints, which optimize the network from the perspectives of depth prediction accuracy, geometric plausibility, and structural clarity.

We first employ the scale-invariant logarithmic loss:

$$\mathcal{L}_{\text{si}} = \frac{1}{N} \sum_{i=1}^N (\log d_i - \log d_i^*)^2 - \frac{1}{N^2} \left(\sum_{i=1}^N (\log d_i - \log d_i^*) \right)^2, \quad (14)$$

where d_i and d_i^* denote the predicted and ground-truth depth values of the i -th valid pixel, respectively, and N is the number of valid pixels. This loss is widely used in monocular depth estimation because it reduces sensitivity to scale ambiguity [1,3].

To encourage dense local geometric consistency across the entire depth map, we further use a gradient consistency term:

$$\mathcal{L}_{\text{grad}} = \frac{1}{|M_g|} \sum_{i \in M_g} \|\nabla d_i - \nabla d_i^*\|_1, \quad (15)$$

where ∇ denotes the spatial gradient operator and M_g denotes the set of valid pixels whose local gradient computation windows contain only valid ground-truth depth values. This term encourages the predicted depth map to match the local geometric transitions of the ground truth [17,34].

To place stronger emphasis on salient depth discontinuities and boundary sharpness, we additionally introduce an edge-focused loss:

$$\mathcal{L}_{\text{edge}} = \frac{1}{|M_e|} \sum_{i \in M_e} \|\epsilon(d_i) - \epsilon(d_i^*)\|_1, \quad (16)$$

where $\epsilon(\cdot)$ denotes a Sobel-based edge extraction operator and M_e denotes the set of valid pixels whose local Sobel convolution windows contain only valid ground-truth values. Specifically,

$$\epsilon(d) = |G_x(d)| + |G_y(d)|, \quad (17)$$

where G_x and G_y are the horizontal and vertical Sobel responses, respectively.

The final training objective is defined as

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{si}} + \lambda_2 \mathcal{L}_{\text{grad}} + \lambda_3 \mathcal{L}_{\text{edge}}, \quad (18)$$

where $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.5$ in our implementation. The three terms are complementary: $\mathcal{L}_{\text{si}}$ improves depth regression accuracy, $\mathcal{L}_{\text{grad}}$ encourages geometric consistency, and $\mathcal{L}_{\text{edge}}$ helps preserve structural boundaries. In particular, $\mathcal{L}_{\text{grad}}$ and $\mathcal{L}_{\text{edge}}$ play different roles in structure supervision: the gradient consistency loss constrains local depth variations across all valid pixels by matching spatial derivatives between predictions and the ground truth. By contrast, the edge-focused loss targets high-response discontinuity regions extracted via Sobel filtering, which explicitly enhances boundary sharpness. Therefore, $\mathcal{L}_{\text{grad}}$ mainly improves overall local geometric transition consistency, while $\mathcal{L}_{\text{edge}}$ provides more explicit supervision for salient boundary regions. Their combination helps the model produce depth maps that are both geometrically smooth in local transitions and sharp near object boundaries.

4. Experiments

4.1. Datasets

We evaluated the proposed method on two standard monocular depth estimation benchmarks: NYUv2 for indoor scenes and KITTI for outdoor driving scenes.

For indoor evaluation, we used NYUv2 [11], which contains 1449 labeled RGB-D image pairs captured by a Kinect sensor. Following the commonly used split adopted in prior work [1,3,13], 795 samples were used for training and 654 for testing. The original image resolution was 640×480 . We used the official filled depth annotations and ignored invalid pixels during training and evaluation. Depth values were clipped to the range of 0–10 m. During training, images were randomly cropped and resized to 384×384 .

For outdoor evaluation, we used the KITTI benchmark [12] with the Eigen split [1], consisting of 23,488 training images and 697 test images. Depth values were clipped to the range of 10^{-3} –80 m, and invalid pixels were masked out. Input images were resized to 384×1216 while preserving the original aspect ratio as much as possible.

4.2. Implementation Details

All experiments were implemented using PyTorch 1.13 and CUDA 11.7 on a single NVIDIA GeForce RTX 3090 GPU with 24 GB of VRAM [35]. The backbone was initialized with the official ImageNet-pretrained DPT-Hybrid weights. In the shallow GE stages (E_1 and E_2), each DWConv branch follows the following order: layer normalization (LN), 3×3 depth-wise convolution (stride = 1, padding = 1), 1×1 point-wise convolution, and GELU activation. In the S-GAF module, all dimensional alignment projections (Φ_e , Φ_d , and Φ_s) are implemented as 1×1 standard convolutions followed by Batch Normalization (BN) and ReLU activation. The channel-gating branch comprises two fully connected layers with a reduction ratio of $r = 16$, an intermediate ReLU activation, and a final Sigmoid (σ) activation. The spatial-gating branch consists of a 3×3 depth-wise convolution followed by a 1×1 convolution and Sigmoid activation. A dropout rate of $p = 0.1$ is applied within the Transformer encoder blocks to reduce overfitting.

We employed the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 1×10^{-4} [36]. The initial learning rate was set to 1×10^{-5} for the pretrained DPT-Hybrid backbone and 1×10^{-4} for the newly introduced GE and S-GAF modules. A cosine-annealing learning rate scheduler was used to gradually decay the learning rate to a minimum of 1×10^{-6} over the training process [37]. During training, RGB images and their corresponding ground-truth depth maps were randomly cropped and resized to 384×384 for NYUv2 and 384×1216 for KITTI. The model was trained with a batch size of 4 for 60 epochs on NYUv2 and 80 epochs on KITTI. Standard data augmentation included random horizontal flipping with a probability of 0.5 and random color jittering of the RGB images with brightness, contrast, and saturation adjusted by ± 0.2 .

The results of the models trained in this study were obtained from a single training run for each configuration. Therefore, the reported differences should be interpreted as single-run numerical observations rather than statistically verified improvements.

4.3. Evaluation Metrics

We evaluated depth estimation performance using standard metrics widely adopted in the literature, including the absolute relative error (AbsRel), squared relative error (SqRel), root mean squared error (RMSE), logarithmic RMSE (RMSE-log), and threshold accuracies $\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$ [1,3]. Lower values of AbsRel, SqRel, RMSE, and RMSE-log indicate better performance, while higher values of the threshold metrics indicate better accuracy.

4.4. Quantitative Comparison

We compared the proposed method with representative monocular depth estimation approaches, including AdaBins, GLPDepth/GLPN, NeW CRFs, and Depth Pro [25,29,38], as well as several DPT-based variants. Unless otherwise stated, the results of the competing state-of-the-art methods reported in Tables 2 and 3 are quoted directly from their original publications under their standard benchmark settings. Absolute metrics are affected by differences in preprocessing, cropping protocols, training schedules, and hardware settings across studies. Thus, these quoted values serve only as a broad reference. For rigorous comparison, we focused primarily on controlled experiments against the DPT-Hybrid baseline. Both models were implemented in the same codebase and trained with identical data augmentation, optimization, and evaluation protocols.

Table 2 reports the results on NYUv2. Compared with the DPT-Hybrid baseline, our method improves all major metrics, reducing AbsRel from 0.107 to 0.099 and RMSE from 0.357 to 0.334. These improvements indicate that the proposed design enhances both numerical depth accuracy and structural prediction quality.

Figure 5 provides qualitative comparisons on representative indoor scenes. Compared with the baseline DPT-Hybrid, the proposed method produces depth maps with clearer object contours, more coherent planar structures, and fewer artifacts around thin objects and occlusion boundaries.

Table 2. Performance comparison on the NYUv2 dataset. Results of competing methods are quoted from their published papers when available.

Method	AbsRel ↓	SqRel ↓	RMSE ↓	RMSE-log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
AdaBins [25]	0.105	0.070	0.350	0.142	0.908	0.989	0.998
NeW CRFs [29]	0.103	0.068	0.347	0.140	0.910	0.990	0.998
GLPN [38]	0.101	0.066	0.343	0.137	0.913	0.991	0.999
Depth Pro [39]	0.099	0.065	0.335	0.133	0.919	0.992	0.999
DPT-Hybrid	0.107	0.072	0.357	0.145	0.904	0.988	0.998
DPT + CBAM	0.102	0.069	0.345	0.138	0.912	0.990	0.998
DPT + ECA	0.100	0.067	0.340	0.135	0.915	0.991	0.999
Ours	0.099	0.066	0.334	0.133	0.918	0.991	0.999

Table 3 reports the results on KITTI. Under the evaluated setting, our model reduces AbsRel from 0.062 to 0.058 and RMSE from 2.573 to 2.455 relative to the controlled DPT-Hybrid baseline. Together with the NYUv2 results, these single-run observations indicate improvements over the baseline on both an indoor benchmark and an outdoor driving benchmark. Their statistical robustness, however, requires repeated experiments with different random seeds.



Figure 5. Qualitative comparison on NYUv2. From left to right: input RGB image, DPT-Hybrid, AdaBins, NeW CRFs, GLPN, Depth Pro, and the proposed method. The proposed model produces sharper boundaries and better structural continuity in challenging regions.

Figure 6 presents qualitative depth prediction examples on the KITTI dataset. The first row displays the original RGB image and corresponding annotation reference, and the second row compares the prediction results of multiple methods. The proposed method produces sharper vehicle contours and more accurate depth discontinuities, which is consistent with the quantitative performance improvement.

Table 3. Performance comparison on the KITTI dataset (Eigen split). Results of competing methods are quoted from their published papers when available.

Method	AbsRel ↓	SqRel ↓	RMSE ↓	RMSE-log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
AdaBins [25]	0.061	0.360	2.550	0.091	0.960	0.995	0.999
NeW CRFs [29]	0.060	0.355	2.520	0.090	0.961	0.995	0.999
GLPN [38]	0.059	0.350	2.495	0.089	0.962	0.996	0.999
Depth Pro [39]	0.057	0.345	2.440	0.088	0.965	0.996	0.999
DPT-Hybrid	0.062	0.365	2.573	0.092	0.959	0.995	0.998
DPT-Large	0.058	0.352	2.487	0.089	0.963	0.996	0.999
Ours	0.058	0.348	2.455	0.088	0.964	0.996	0.999

Note that the specific baseline variants included in Tables 2 and 3 were selected to reflect the distinct challenges of indoor and outdoor environments. For NYUv2 (Table 2), which features fine object boundaries and heavy occlusions, we included attention-enhanced DPT variants as references for structure-related feature recalibration in indoor scenes. For the outdoor KITTI dataset (Table 3), where capturing vast depth ranges (0–80 m) often relies on high model capacity, we included DPT-Large as a capacity-oriented DPT reference to provide additional context for the accuracy–complexity trade-off.

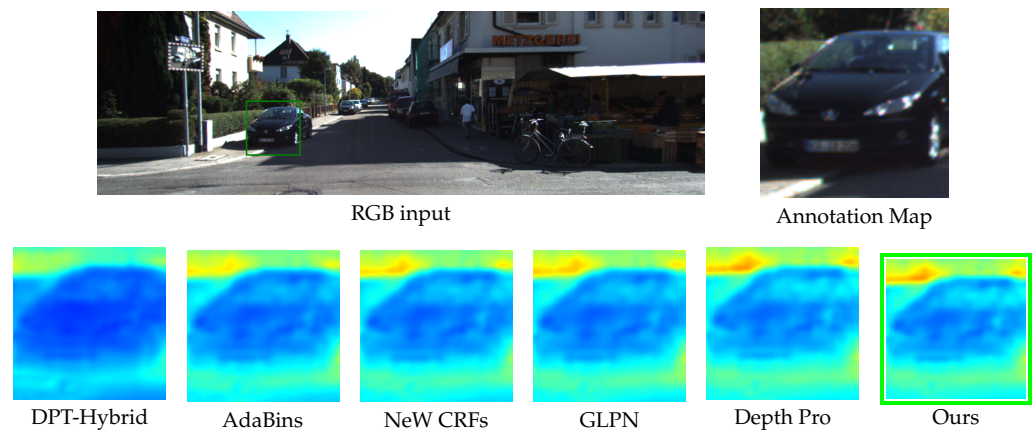


Figure 6. Qualitative comparison on the KITTI dataset. The first row displays the original RGB image and corresponding annotation reference. The second row presents depth predictions generated by different competing methods. From left to right: DPT-Hybrid, AdaBins, NeW CRFs, GLPN, Depth Pro, and the proposed method. The proposed model produces sharper vehicle contours.

4.5. Ablation Studies

To verify the effectiveness of each design choice, we conducted a series of ablation experiments on NYUv2 under the same training and evaluation protocol. The ablation analysis is organized into four parts: We first present an overall component-level ablation to evaluate the contribution of the three main modules. We then investigate the internal behavior of the proposed S-GAF module, the geometry-enhanced (GE) encoder, and the joint loss separately. This organization follows common scientific writing practice by progressing from global module validation to more fine-grained design analysis.

4.5.1. Overall Component Ablation

We first evaluated the contribution of the three main components: the geometry-enhanced encoder (GE), the S-GAF module, and the joint loss (JL). Starting from the DPT-Hybrid baseline, we incrementally added each component and report the results in Table 4. Representative visual comparisons are shown in Figure 7.

Table 4. Overall ablation study on NYUv2. GE denotes the geometry-enhanced encoder, S-GAF denotes the proposed Structure-Aware Cross-Scale Gated Attention Fusion module, and JL denotes the joint structure–geometric consistency loss.

Model	Components			Metrics		
	GE	S-GAF	JL	AbsRel ↓	RMSE ↓	$\delta < 1.25$ ↑
DPT-Hybrid	×	×	×	0.107	0.357	0.904
DPT-Hybrid + S-GAF	×	✓	×	0.104	0.351	0.907
DPT-Hybrid + JL	×	×	✓	0.105	0.353	0.908
DPT-Hybrid + GE	✓	×	×	0.103	0.348	0.910
DPT-Hybrid + GE + S-GAF	✓	✓	×	0.100	0.340	0.916
Ours	✓	✓	✓	0.099	0.334	0.918

The results in Table 4 show that all three components contribute positively to the final performance. Among the single-module variants, GE provides the largest individual improvement, reducing AbsRel from 0.107 to 0.103 and RMSE from 0.357 to 0.348. This observation is consistent with the role of shallow features in DPT-Hybrid. While Transformer backbones excel at modeling long-range dependencies and global scene layouts, local structures—edges, corners, textures and thin regions—are primarily encoded by high-resolution

shallow features. Once such fine-grained geometric information is weakened during the early stages, it becomes difficult to recover it completely during decoding.

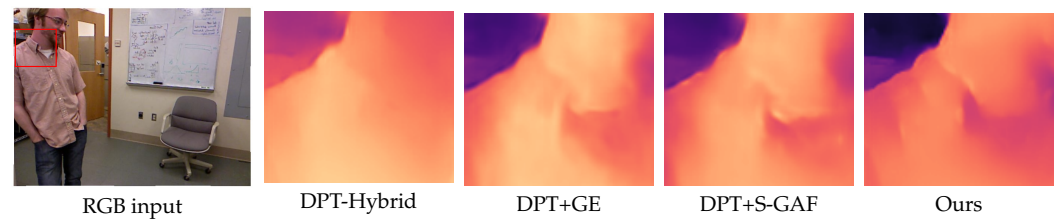


Figure 7. Qualitative overall ablation results on NYUv2. The complete model produces sharper boundaries and more coherent local geometry than the baseline and partial variants.

The S-GAF-only variant also improves upon the baseline, indicating that decoder-side fusion is another important source of error in the original DPT-Hybrid. Simple cross-scale addition or concatenation does not explicitly distinguish between semantically rich encoder features and progressively refined decoder features. The proposed gated fusion enables the network to adaptively select between these feature sources according to the local structure, which is particularly beneficial around depth discontinuities and thin structures.

The JL-only variant provides a moderate but consistent improvement. This suggests that loss design alone can improve local depth transitions and boundary sharpness, although its effect is more limited when the underlying feature representations are not sufficiently structure-aware. When GE and S-GAF are used together, the improvement becomes more pronounced, and the best result is achieved when all three components are combined. This indicates that the three designs are complementary rather than redundant. In other words, improved shallow representations, more effective cross-scale fusion, and structure-sensitive supervision reinforce one another.

The visual results in Figure 7 further support the quantitative findings. Compared with the baseline, the GE-enhanced variant preserves clearer object contours, the S-GAF variant produces more stable local depth transitions, and the full model achieves the best overall balance between boundary sharpness and regional consistency.

4.5.2. Ablation on the S-GAF Module

To further understand the contribution of the proposed fusion strategy, we conducted an internal ablation study on S-GAF. The purpose of this experiment was to disentangle the effects of gated fusion, channel–spatial attention, and auxiliary structure guidance. Since accurate monocular depth estimation requires both high-level semantic understanding and low-level detail recovery, the quality of decoder fusion directly affects the preservation of object boundaries and thin structures.

To rigorously isolate the contribution of each structural design choice within the S-GAF module and distinguish it from conventional attention mechanisms, Table 5 presents an expanded internal ablation study on the NYUv2 dataset.

From a methodological perspective, S-GAF differs from generic feature fusion modules, such as FPN-style addition, and lightweight attention blocks, such as CBAM or ECA. FPN mainly defines a top-down information propagation path, while CBAM and ECA primarily recalibrate an already fused feature tensor. By contrast, S-GAF is explicitly designed for cross-scale fusion between encoder and decoder features in monocular depth estimation. Its gating mechanism does not impose a fixed preference for either feature source; instead, it learns spatially and channel-wise adaptive weights conditioned on both feature content and auxiliary RGB-gradient cues.

Table 5. Expanded internal ablation study isolating each design choice of the S-GAF module on NYUv2. Starting from simple concatenation, we progressively introduced CBAM-style recalibration, cross-scale gating, channel gating, spatial gating, and an auxiliary RGB-gradient input.

Variant/ Fusion Strategy	Channel Gate	Spatial Gate	Cross-Scale Gating ($G \odot F_e + (1 - G) \odot F_d$)	Auxiliary RGB-Gradient Input (F_s)	AbsRel ↓	RMSE ↓
Simple Concatenation	×	×	×	×	0.107	0.357
			(Direct Concat)			
CBAM-Style Recalibration	✓	✓	×	×	0.106	0.354
			(Self-Attention on Concat)			
S-GAF (Channel Only)	✓	×	✓	×	0.106	0.355
S-GAF (Spatial Only)	×	✓	✓	×	0.106	0.355
S-GAF (Channel + Spatial)	✓	✓	✓	×	0.105	0.353
Full S-GAF (Ours)	✓	✓	✓	✓	0.104	0.351

This design is particularly important because encoder features usually retain stronger local details, whereas decoder features carry progressively integrated semantic context. Around object boundaries and geometrically ambiguous regions, these two sources are not equally reliable. A learned gate can therefore improve fusion quality by preserving details where necessary and emphasizing semantic stability where local cues are weak or noisy.

The quantitative results in Table 5 support this interpretation. As shown in the table, applying standard CBAM-style self-recalibration to concatenated features yields only minor improvements (0.354 RMSE). Relative to simple concatenation, both channel-only and spatial-only cross-scale gating variants provide measurable but limited improvements, reducing AbsRel from 0.107 to 0.106 and RMSE from 0.357 to 0.355. When channel and spatial gates are combined under our cross-scale arbitration mechanism, the performance further improves to an AbsRel of 0.105 and an RMSE of 0.353. Under the evaluated setting, the combined cross-scale soft-selection gate achieves lower reported errors than the tested CBAM-style self-recalibration and single-gate variants. This indicates that explicit cross-scale arbitration is better suited for encoder–decoder fusion than intra-feature self-recalibration. The two gate types are complementary: channel gating reweights features globally, while spatial gating highlights locally important regions.

The full S-GAF configuration achieves the best result after introducing auxiliary RGB-gradient modulation, with AbsRel and RMSE being further reduced to 0.104 and 0.351, respectively. This final gain supports the incremental contribution of the auxiliary RGB-gradient input under the evaluated setting. The relatively small numerical gain also indicates that the gradient cue serves as a complementary input rather than the dominant source of improvement.

The visual examples in Figure 8 are consistent with the quantitative trends. Compared with simpler fusion strategies, the full S-GAF variant produces sharper transitions near object boundaries and more coherent predictions in structurally complex regions. These observations suggest that the gain mainly comes from more appropriate cross-scale aggregation rather than merely from adding extra parameters.

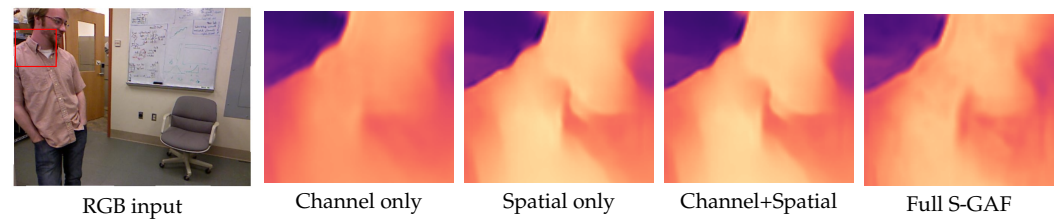


Figure 8. Qualitative comparison of internal S-GAF variants on NYUv2. Introducing channel–spatial gating and auxiliary RGB-gradient guidance progressively improves boundary clarity and local structural continuity.

4.5.3. Ablation on the Geometry-Enhanced Encoder

We next analyzed the internal design of the geometry-enhanced encoder. The goal of GE is to improve shallow local geometric modeling while preserving the strong global reasoning capability of the original DPT-Hybrid backbone. Since the first two stages operate at a relatively high spatial resolution, they are the most suitable locations for injecting local inductive bias through lightweight convolutional enhancement.

The overall ablation has already shown that GE provides the largest individual improvement among the three main components. This result is intuitively reasonable. Although Transformer-based backbones are strong in long-range contextual reasoning, they do not always explicitly emphasize local edge and texture patterns in shallow stages to the same extent as convolutional operators. For monocular depth estimation, however, local geometric structures are critical because they are closely related to depth discontinuities, object contours, and thin regions.

The quantitative results in Table 6 further clarify this design choice. Without GE, the model built on DPT-Hybrid + S-GAF obtains an AbsRel of 0.104 and an RMSE of 0.351. Adding GE only to Stage 1 improves the result to 0.102/0.344, whereas adding it only to Stage 2 yields 0.103/0.345. This indicates that both shallow stages are useful, but Stage 1 contributes slightly more, likely because it preserves the highest-resolution local cues.

Table 6. Ablation study of the geometry-enhanced encoder design on NYUv2. All variants are built on DPT-Hybrid + S-GAF.

Variant	GE Stage 1	GE Stage 2	Conv. Type	AbsRel ↓	RMSE ↓
No GE	×	×	-	0.104	0.351
GE only Stage 1	✓	×	DW Conv	0.102	0.344
GE only Stage 2	×	✓	DW Conv	0.103	0.345
GE on Stage 1 + Stage 2	✓	✓	DW Conv	0.100	0.340
GE (standard conv)	✓	✓	Standard Conv	0.102	0.343

When GE is inserted into both Stage 1 and Stage 2, the performance improves further to an AbsRel of 0.100 and an RMSE of 0.340, which is the best configuration among the tested variants. This confirms that the two shallow stages provide complementary information for local structure recovery. By contrast, replacing the depth-wise separable convolution with a standard convolution degrades the result to 0.102/0.343, even though GE is still applied to both stages. This indicates that the depth-wise separable design is not only lightweight, but also better suited for its purpose: injecting an efficient local inductive bias without disrupting the original Transformer representations.

The proposed GE module introduces a lightweight depth-wise separable convolution branch in a residual manner. This design is advantageous for two reasons: First, it strengthens sensitivity to high-frequency local patterns with limited computational overhead. Second, the residual formulation preserves the original Transformer representation

and therefore acts as a complementary refinement rather than a complete architectural replacement. This makes the enhancement both stable and efficient.

The visual comparison in Figure 9 supports the numerical trends. The best GE configuration recovers clearer boundaries and produces more consistent local depth transitions, especially in cluttered regions where small geometric details are easily blurred in the baseline. These observations indicate that the gain is structurally meaningful rather than reflected only in global error metrics.

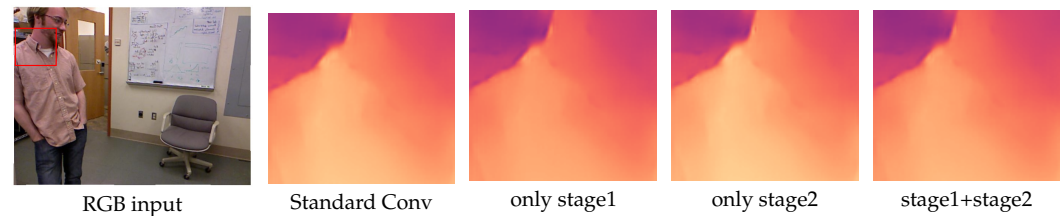


Figure 9. Qualitative comparison of geometry-enhanced encoder variants on NYUv2. Enhancing the first two shallow stages with depth-wise separable convolutions yields clearer contours and more stable local geometry.

4.5.4. Ablation on the Joint Loss

Finally, we investigated the effect of the proposed joint structure–geometric consistency loss. Although the SI-Log loss is a strong baseline objective for monocular depth estimation, it mainly emphasizes overall pixel-wise regression quality and scale-invariant consistency. When used alone, it may still lead to overly smooth predictions around boundaries and in regions with rapid depth variation.

As shown in Table 7, adding either the gradient loss or the edge loss improves the reported results compared with using the SI-Log loss alone. The gradient loss reduces AbsRel from 0.100 to 0.099 and RMSE from 0.340 to 0.335, while the edge loss achieves an AbsRel of 0.099 and an RMSE of 0.336. By combining both loss terms, the full joint loss achieves the lowest reported RMSE of 0.334 while maintaining an AbsRel of 0.099.

Table 7. Ablation study of the joint loss design on NYUv2. All variants are built on the full network architecture (DPT-Hybrid + S-GAF + GE) with the same training and evaluation settings and differ only in the included loss terms. Results were obtained from a single training run for each configuration.

Variant	SI-Log	Grad Loss	Edge Loss	AbsRel ↓	RMSE ↓
SI-Log only	✓	×	×	0.100	0.340
SI-Log + grad	✓	✓	×	0.099	0.335
SI-Log + edge	✓	×	✓	0.099	0.336
SI-Log + grad + edge (Ours)	✓	✓	✓	0.099	0.334

The reduction in AbsRel from 0.100 to 0.099 is marginal and should be interpreted as an observed numerical difference rather than a statistically verified improvement. The reduction in RMSE from 0.340 to 0.334 is comparatively more noticeable, suggesting that the joint loss may help reduce relatively large depth prediction errors. However, all results in Table 7 come from single training runs per configuration. Repeated experiments with different random seeds are needed to verify the statistical robustness of the observed improvements.

The visual examples in Figure 10 are consistent with the reported numerical results. Compared with SI-Log alone, the configurations with additional structural supervision produce clearer boundaries and more stable local transitions in the illustrated examples. The full joint loss visually provides a favorable balance between boundary sharpness and regional consistency, suggesting the potential benefit of combining regression accuracy with

structure-sensitive optimization. Since these are qualitative examples and the quantitative results are obtained from single runs, they should be interpreted as illustrative evidence rather than statistical validation of the observed improvements.

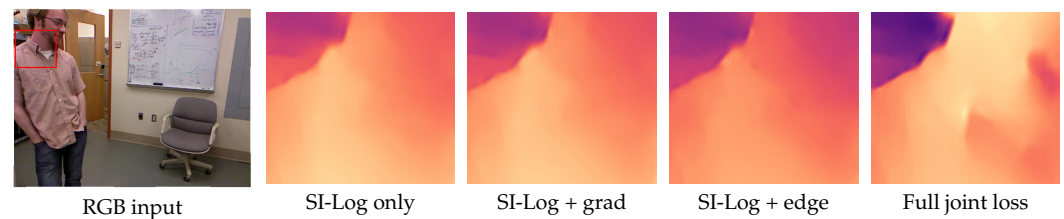


Figure 10. Qualitative comparison of different loss configurations. The examples compare the depth predictions obtained using the SI-Log loss alone, the SI-Log loss with the gradient loss, the SI-Log loss with the edge loss, and the full joint loss.

4.5.5. Sensitivity to Representative Loss Weight Settings

To evaluate the sensitivity of the selected loss weights in Equation (18), we conducted a sensitivity analysis on the NYUv2 dataset by fixing the primary regression weight $\lambda_1 = 1.0$ and systematically varying the gradient consistency weight λ_2 and edge-focused weight λ_3 . Table 8 summarizes the impact of different weight combinations on depth accuracy and boundary preservation.

Table 8. Sensitivity analysis of loss weights λ_2 (gradient loss) and λ_3 (edge loss) on NYUv2 ($\lambda_1 = 1.0$).

Configuration	λ_2 (Grad)	λ_3 (Edge)	AbsRel ↓	RMSE ↓	RMSE-log ↓	$\delta < 1.25$ ↑
Config A (Under-weighted)	0.2	0.2	0.101	0.341	0.136	0.914
Config B (Ours—Optimal)	0.5	0.5	0.099	0.334	0.133	0.918
Config C (Gradient-heavy)	0.8	0.2	0.100	0.336	0.134	0.916
Config D (Edge-heavy)	0.2	0.8	0.100	0.337	0.135	0.916
Config E (Over-weighted)	1.0	1.0	0.102	0.343	0.137	0.913

Among the five evaluated single-run configurations, the setting $(\lambda_2, \lambda_3) = (0.5, 0.5)$ obtains the lowest reported AbsRel and RMSE. Lower equal auxiliary weights provide weaker structural regularization, whereas the larger equal-weight setting yields higher reported errors. The gradient-heavy and edge-heavy configurations also perform slightly worse than the adopted setting. These observations validate the chosen weights within the tested configurations. They do not, however, confirm global optimality or statistical robustness across different random seeds.

4.6. Efficiency Analysis

In addition to prediction accuracy, computational efficiency is an important consideration for practical monocular depth estimation, especially in scenarios with limited computational resources or real-time requirements. We compared the baseline DPT-Hybrid and the proposed full model in terms of parameter count, floating-point operations (FLOPs), inference throughput (FPS), and single-frame latency. Both models were evaluated with a batch size of 1 on the same NVIDIA RTX 3090 GPU using an identical input resolution and testing protocol.

Table 9 summarizes the results. The parameter count reflects the model size, while FLOPs estimate the theoretical arithmetic cost of a single forward pass. FPS and latency characterize the actual run-time efficiency on the target hardware. The latency values are calculated from the measured throughput as $1000/\text{FPS}$ and provide a more intuitive indication of the processing time required for each frame.

Table 9. Efficiency comparison of DPT-Hybrid and the proposed full model. FLOPs are calculated for a single forward pass. FPS was measured with a batch size of 1 on an NVIDIA RTX 3090 under the same input resolution and testing protocol, and latency was calculated as 1000/FPS.

Model	Params (M)	FLOPs (G)	FPS ↑	Latency (ms/frame) ↓	NYUv2 AbsRel ↓	KITTI AbsRel ↓
DPT-Hybrid	123.15	109.96	13.04	76.68	0.107	0.062
Full model (Ours)	129.75	115.56	12.11	82.57	0.099	0.058

As shown in Table 9, the proposed model increases the number of parameters from 123.15 M to 129.75 M (+5.36%) and FLOPs from 109.96 G to 115.56 G (+5.09%). Its throughput decreases from 13.04 to 12.11 FPS, representing a reduction of 7.13%, while the corresponding single-frame latency increases from 76.68 to 82.57 ms, representing an increase of 7.68%.

The run-time reduction is slightly larger than the increase indicated by the FLOPs because FLOPs measure only arithmetic operations and do not fully represent hardware execution efficiency. In particular, the S-GAF module introduces multi-branch feature transformations, feature concatenation, pooling, gating, and element-wise modulation during multi-scale decoding. These operations add relatively few FLOPs, but they incur extra intermediate tensor reads/writes and additional GPU kernel launches. Their run-time cost is therefore influenced by memory traffic, kernel launch overhead, and synchronization rather than arithmetic complexity alone. Moreover, these operations are repeatedly applied to relatively high-resolution feature maps, causing their overhead to accumulate across decoder stages. This explains why the observed FPS reduction of 7.13% is moderately greater than the theoretical FLOPs increase of 5.09%.

To substantiate our claim of a lightweight architectural design, Table 10 provides a detailed component-wise breakdown of the computational complexity, parameter overhead, and single-frame latency introduced by each proposed module relative to the DPT-Hybrid baseline.

Table 10. Component-wise complexity breakdown of the proposed modules on an NVIDIA RTX 3090 GPU (384 × 384 input).

Component/Variant	Params (M)	Added Params	FLOPs (G)	Added FLOPs	Latency (ms)	Added Latency
DPT-Hybrid (Baseline)	123.15	–	109.96	–	76.68	–
+GE Encoder Only	123.55	+0.40 M (+0.32%)	111.12	+1.16 G (+1.05%)	77.82	+1.14 ms
+S-GAF Module Only	129.35	+6.20 M (+5.03%)	114.40	+4.44 G (+4.04%)	81.45	+4.77 ms
Full Model (GE + S-GAF)	129.75	+6.60 M (+5.36%)	115.56	+5.60 G (+5.09%)	82.57	+5.89 ms

As quantitatively demonstrated in Table 10, the geometry-enhanced (GE) encoder is exceptionally lightweight. Injecting residual depth-wise separable convolution branches into the first two shallow stages adds only 0.40 M parameters (+0.32%), 1.16 G FLOPs (+1.05%), and 1.14 ms of single-frame latency. This evidence substantiates our claim that the GE encoder introduces modest additional complexity while delivering substantial gains in shallow local geometric modeling.

Despite the additional computational cost, the proposed full model achieves higher depth estimation accuracy than the baseline DPT-Hybrid on both benchmarks. Specifically, the AbsRel metric improves from 0.107 to 0.099 on NYUv2 and from 0.062 to 0.058 on KITTI. These results demonstrate that the geometry-enhanced encoding and structure-aware gated fusion modules effectively improve prediction accuracy while introducing only a small increase in model size and FLOPs, at the cost of a moderate reduction in inference speed.

Overall, the proposed model provides a reasonable trade-off between accuracy and efficiency. The gated fusion design reduces inference speed relative to the vanilla DPT-

Hybrid. However, its improved depth accuracy makes the model well suited for accuracy-critical tasks, such as 3D reconstruction and offline robotic mapping, where geometric precision and boundary sharpness outweigh strict real-time requirements.

5. Discussion

5.1. Mechanistic Analysis and Comparison with Published Methods

The improvements achieved by the proposed framework can be attributed to the complementary roles of the geometry-enhanced (GE) encoder, the structure-guided adaptive fusion (S-GAF) module, and the joint structure–geometric consistency loss. Transformer-based depth models are effective at capturing long-range dependencies and global scene layouts, but patch projection and repeated semantic aggregation may weaken high-frequency details. The residual depth-wise separable convolution branches in the shallow encoder stages compensate for this limitation by preserving local edges and geometric transitions with limited computational overhead. Unlike increasing backbone capacity, as in larger DPT variants, GE improves the representation of local structures within the original DPT-Hybrid architecture. The controlled ablation results confirm that the resulting gains arise from shallow geometric enhancement rather than model scaling alone.

S-GAF further addresses the semantic gap between encoder and decoder features. Conventional addition or concatenation treats the two feature sources uniformly, while generic attention modules such as SENet, CBAM, and ECA mainly recalibrate a single feature tensor or an already merged representation [19–21]. In contrast, S-GAF learns a complementary gate that determines where encoder details should be retained and where decoder semantics should dominate. RGB gradients serve as an auxiliary input to the learned gate, not as rigid geometric priors. This allows the network to mitigate interference from textures and illumination variations. The ablation results show that this cross-scale gating strategy performs better than the tested CBAM-style fusion, while gradient modulation provides an additional improvement.

The proposed loss terms complement the network design by explicitly supervising local depth variations. SI-Log primarily optimizes scale-invariant numerical accuracy, whereas the gradient term encourages consistent first-order depth changes and the edge term emphasizes discontinuities. Their combination mainly reduces relatively large localized errors near object boundaries, which explains the clearer improvement in RMSE than in AbsRel. Unlike ranking-based or normal-based geometric constraints [17,18,28], our loss terms offer a simpler local supervision scheme that operates directly on valid depth neighborhoods.

The quantitative results position the proposed method favorably relative to representative published approaches such as AdaBins, NeW CRFs, GLPN, and DPT variants [3,25,29,38]. These methods improve depth estimation through adaptive depth discretization, conditional random fields, global–local pathways, or larger representations, whereas our method focuses on shallow geometric preservation and structure-aware cross-scale fusion. The underlying objective is also consistent with efficient structured signal representations, including high-order Dual Hahn polynomial methods for signal reconstruction, compression, and recognition: both seek to retain informative structures while controlling redundant computation. However, our approach employs learned local filtering rather than explicit polynomial transforms, making it more suitable for end-to-end dense prediction. Since training and evaluation protocols may vary across publications, the cross-paper results should be regarded as contextual comparisons; the controlled comparisons with DPT-Hybrid provide the strongest evidence for the effectiveness of the proposed components.

Nevertheless, the measured throughput decreases from 13.04 FPS for DPT-Hybrid to 12.11 FPS for the proposed model on an RTX 3090. Therefore, the current implementation offers an improved accuracy–complexity trade-off for offline reconstruction and moderate-rate robotic perception, but it does not yet satisfy strict high-frame-rate or embedded real-time requirements.

5.2. Limitations

Despite its consistent improvements on the evaluated benchmarks, the proposed method has several limitations: First, RGB gradients are appearance cues rather than direct measurements of scene geometry. Dense textures, shadows, reflections, and illumination changes can create gradients unrelated to depth discontinuities. Conversely, boundaries between similarly colored objects may produce only faint gradients. Although S-GAF learns to adaptively adjust the contribution of gradient-assisted structural cues, it cannot completely distinguish photometric edges from physical boundaries. Consequently, the method may introduce local depth fluctuations in highly textured regions or over-smooth low-contrast depth boundaries.

Figure 11 visualizes the pixel-wise depth error distribution of the vanilla DPT-Hybrid baseline and the proposed full model on NYUv2. Compared with the baseline that produces blurred object boundaries and uneven local depth transitions, our method generates sharper structural contours and reduces overall local depth errors. However, noticeable errors still persist in fine boundary regions and low-contrast areas, which corroborates the aforementioned limitation.

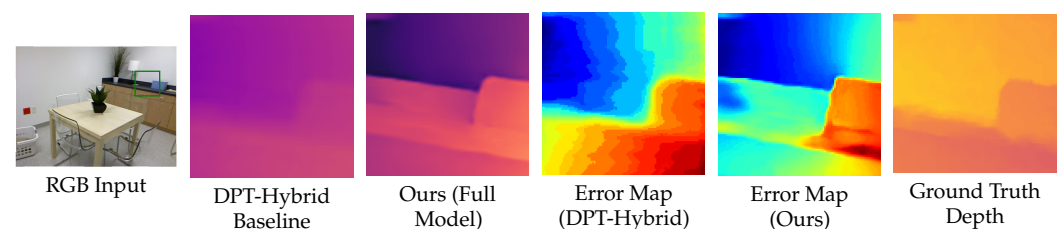


Figure 11. Qualitative comparison of DPT-Hybrid and the proposed full model on NYUv2. From left to right in the first row: input RGB image, depth prediction obtained by the vanilla DPT-Hybrid, and depth prediction of our full model. From left to right in the second row: pixel-wise depth error heatmap of DPT-Hybrid, pixel-wise depth error heatmap of our method, and ground-truth depth. The baseline produces blurred object boundaries and uneven local depth transitions. By combining geometry-enhanced encoding, gradient-assisted gated fusion, and structure-sensitive supervision, the full model produces sharper structural contours and reduces local depth errors relative to the baseline.

Second, transparent, reflective, and non-Lambertian surfaces remain challenging. For example, mirrors may contain strong virtual contours, while glass surfaces may reveal background structures instead of their actual geometry. These ambiguities may mislead both the structural guidance and the monocular depth predictor. Missing or noisy sensor annotations for such materials further reduce the available supervision, even though invalid measurements are excluded from the loss computation. Large textureless regions and distant outdoor areas remain challenging. GE and S-GAF can preserve existing structural cues but cannot reliably recover geometry when informative visual evidence is missing.

Third, the current model processes individual images and does not exploit temporal or multi-view consistency. Motion blur, adverse illumination, weather variations, and dynamic objects may therefore weaken the extracted structural cues or cause unstable frame-wise predictions. Moreover, although GE introduces only a small increase in the theoretical computation, the multi-scale feature projections and gating operations in S-GAF

incur additional memory access and kernel launch overhead. The measured throughput of 12.11 FPS on an RTX 3090 consequently limits direct deployment in high-frame-rate or resource-constrained applications.

Finally, the method relies on densely supervised depth data and has been evaluated only on NYUv2 and KITTI under their standard protocols. Although these datasets represent indoor and outdoor driving environments, respectively, results based on them are insufficient to establish broad cross-dataset or zero-shot generalization. Variations in camera intrinsics, scene composition, illumination, weather, depth range, and annotation modality may cause substantial domain shifts on other datasets. Accordingly, these results demonstrate controlled improvements over DPT-Hybrid on the tested benchmarks but do not prove universal cross-domain robustness. In addition, differences in preprocessing, network backbones, training settings, and evaluation protocols limit definitive comparisons with results reported by other studies. The reported results for the models trained in this study are single-run point estimates and characterize endpoint performance under the evaluated settings. Since complete epoch-wise records are not available for all controlled variants, the present results do not support a formal comparison of convergence speed or optimization stability. Future work will include repeated experiments with different random seeds and complete training and validation curves to evaluate both learning variance and convergence behavior.

5.3. Future Research Goals and Methodological Extensions

First, future work will prioritize standardized cross-dataset and zero-shot evaluation. Models trained on NYUv2 will be evaluated without target-domain fine-tuning on ScanNet, ETH3D, Matterport3D, and Hypersim, while models trained on KITTI will be evaluated on DDAD and other outdoor datasets. These experiments will examine robustness to changes in camera intrinsics, depth ranges, illumination conditions, weather, scene composition, and sensing or annotation modalities. The depth scale, valid-region mask, and evaluation protocol of each dataset will be carefully standardized to ensure fair comparisons. More robust structural priors will also be investigated to reduce the influence of texture, shadows, reflections, and weak-contrast boundaries.

Second, we will focus on lightweight deployment through structural re-parameterization, operator fusion, low-rank projection, quantization-aware training, and knowledge distillation. A multi-branch structure-aware module could be used during training and converted into a simpler inference-time operator, thereby reducing the memory access and kernel launch overhead of S-GAF. The resulting models will be evaluated on embedded platforms rather than only on desktop GPUs, with parameter count, FLOPs, throughput, latency, memory consumption, and energy consumption reported jointly.

Third, temporal consistency and self-supervised adaptation will be explored to reduce dependence on densely annotated depth data. Photometric reconstruction, occlusion-aware masking, teacher–student pseudo-labeling, and multi-frame geometric constraints may improve stability in textureless or motion-blurred regions and reduce frame-wise depth fluctuations in videos. Unsupervised domain adaptation and test-time adaptation will also be investigated to improve robustness to unseen cameras, illumination conditions, weather, and scene distributions. Future evaluation will therefore consider boundary accuracy, temporal consistency, uncertainty calibration, and deployment efficiency in addition to standard depth estimation metrics.

6. Conclusions

This paper presented a structure-aware enhanced DPT-Hybrid framework for monocular depth estimation. The proposed method improves the baseline architecture via three

aspects: a geometry-enhanced shallow encoder for stronger local structure modeling, a Structure-Aware Cross-Scale Gated Attention Fusion module for adaptive multi-scale aggregation, and joint structure–geometric consistency loss for improving both numerical accuracy and structural fidelity.

Under the evaluated indoor (NYUv2) and outdoor (KITTI) protocols, the proposed method achieves lower single-run error metrics than the controlled DPT-Hybrid baseline. The quantitative and qualitative results indicate improved preservation of object boundaries, thin structures, and local depth discontinuities. The proposed modules cause a moderate increase in the number of parameters and theoretical computation; however, the extra feature projections, gating operations, and memory access overhead lead to a measurable drop in practical inference throughput.

Nevertheless, the current experimental scope is limited to the standard NYUv2 and KITTI protocols and therefore does not establish broad cross-dataset or zero-shot generalization. Real-world scenarios such as mobile AR and unstructured-environment robot navigation may encounter severe domain shifts. These stem from varying camera parameters, extreme illumination, adverse weather, dynamic motion, and transparent or reflective surfaces. Thus, our conclusions are limited to the datasets and evaluation protocols tested in this work. Since the current results come from single training runs, future experiments will use multiple random seeds to evaluate variance and statistical robustness.

Future work will conduct standardized zero-shot and cross-dataset evaluations on more diverse indoor, outdoor, real-world, and synthetic datasets. We will also investigate more robust structural priors, lightweight decoder and fusion designs, temporal constraints, and semi-supervised, self-supervised, or domain-adaptive learning strategies. These extensions are expected to improve cross-domain robustness, inference efficiency, temporal stability, and applicability to resource-constrained real-world systems.

Author Contributions: Conceptualization, W.L. and S.H. (Shilei Hu); methodology, W.L. and S.H. (Shilei Hu); software, S.H. (Shilei Hu) and S.H. (Shengkai Hong); validation, Y.Q. and S.H. (Shengkai Hong); formal analysis, S.H. (Shilei Hu) and D.Z.; investigation, Y.Q. and S.H. (Shengkai Hong); resources, W.L.; data curation, S.H. (Shilei Hu) and Y.Q.; writing—original draft preparation, S.H. (Shilei Hu) and W.L.; writing—review and editing, W.L., Y.Q., and D.Z.; visualization, S.H. (Shilei Hu) and S.H. (Shengkai Hong); supervision, W.L.; project administration, W.L.; funding acquisition, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Henan Provincial University-Enterprise Collaborative Innovation Project (26BXQXT050), in part by the Science and Technology Research Project of the Henan Province under grant (262102211033), in part by the Commissioned Project of the Han Culture Discipline Cluster, Nanyang Normal University under grant (hwhxkq202510), and in part by the Nanyang Normal University Foundation of China under grant 2024PY011.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The NYUv2 dataset is publicly available at https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html (accessed 20 June 2026), and the KITTI dataset is publicly available at <http://www.cvlibs.net/datasets/kitti/> (accessed 20 June 2026). No new datasets were created in this study. The publicly available NYUv2 and KITTI datasets were analyzed as described above.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AbsRel	Absolute Relative Error
DPT	Dense Prediction Transformer
DWConv	Depth-Wise Separable Convolution
FLOPs	Floating-Point Operations
FPS	Frames Per Second
GE	Geometry-Enhanced
KITTI	Karlsruhe Institute of Technology and Toyota Technological Institute Dataset
NYUv2	New York University Depth Dataset v2
RMSE	Root Mean Square Error
S-GAF	Structure-Aware Cross-Scale Gated Attention Fusion
SI-Log	Scale-Invariant Logarithmic
SqRel	Squared Relative Error

References

1. Eigen, D.; Puhersch, C.; Fergus, R. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2014), Montreal, QC, Canada, 8–13 December 2014; pp. 2366–2374. [\[CrossRef\]](#)
2. Scharstein, D.; Szeliski, R. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *Int. J. Comput. Vis.* **2002**, *47*, 7–42. [\[CrossRef\]](#)
3. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision Transformers for Dense Prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021), Montreal, QC, Canada, 11–17 October 2021; pp. 12179–12188.
4. Azuma, R.T. A Survey of Augmented Reality. *Presence Teleoperators Virtual Environ.* **1997**, *6*, 355–385. [\[CrossRef\]](#)
5. Qi, C.R.; Liu, W.; Wu, C.; Su, H.; Guibas, L.J. Frustum PointNets for 3D Object Detection from RGB-D Data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 918–927. [\[CrossRef\]](#)
6. Mur-Artal, R.; Montiel, J.M.M.; Tardos, J.D. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Trans. Robot.* **2015**, *31*, 1147–1163. [\[CrossRef\]](#)
7. Saxena, A.; Sun, M.; Ng, A.Y. Make3D: Learning 3D Scene Structure from a Single Still Image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 824–840. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Liu, F.; Shen, C.; Lin, G. Deep Convolutional Neural Fields for Depth Estimation from a Single Image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), Boston, MA, USA, 7–12 June 2015; pp. 5162–5170. [\[CrossRef\]](#)
9. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR 2021), Virtual Event, 3–7 May 2021.
10. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021), Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022. [\[CrossRef\]](#)
11. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor Segmentation and Support Inference from RGBD Images. In Proceedings of the European Conference on Computer Vision (ECCV 2012), Florence, Italy, 7–13 October 2012; pp. 746–760. [\[CrossRef\]](#)
12. Geiger, A.; Lenz, P.; Urtasun, R. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012), Providence, RI, USA, 16–21 June 2012; pp. 3354–3361. [\[CrossRef\]](#)
13. Laina, I.; Ruppel, C.; Belagiannis, V.; Tombari, F.; Navab, N. Deeper Depth Prediction with Fully Convolutional Residual Networks. In Proceedings of the International Conference on 3D Vision (3DV 2016), Stanford, CA, USA, 25–28 October 2016; pp. 239–248. [\[CrossRef\]](#)
14. Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Tao, D. Deep Ordinal Regression Network for Monocular Depth Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2002–2011. [\[CrossRef\]](#) [\[PubMed\]](#)

15. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125. [\[CrossRef\]](#)
16. Lin, G.; Milan, A.; Shen, C.; Reid, I. RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
17. Yin, W.; Liu, Y.; Shen, C.; Yan, Y. Enforcing Geometric Constraints of Virtual Normal for Depth Prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5684–5693. [\[CrossRef\]](#)
18. Hu, J.; Ozay, M.; Zhang, Y.; Okatani, T. Revisiting Single Image Depth Estimation: Toward Higher Resolution Maps with Accurate Object Boundaries. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV 2019), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1043–1051. [\[CrossRef\]](#)
19. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018; pp. 3–19. [\[CrossRef\]](#)
20. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020), Virtual Event, 14–19 June 2020; pp. 11534–11542. [\[CrossRef\]](#)
21. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141. [\[CrossRef\]](#)
22. Godard, C.; Mac Aodha, O.; Brostow, G.J. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 270–279.
23. Zhou, L.; Shen, C.; van den Hengel, A. Edge-Guided Depth Estimation Network. In Proceedings of the Asian Conference on Computer Vision (ACCV 2018), Perth, Australia, 2–6 December 2018; pp. 253–268.
24. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Attention to Scale: Scale-Aware Semantic Image Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 3640–3649. [\[CrossRef\]](#)
25. Bhat, S.F.; Alhashim, I.; Wonka, P. AdaBins: Depth Estimation using Adaptive Bins. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021), Virtual Event, 19–25 June 2021; pp. 4009–4018.
26. Chen, Y.; Yin, Q.; Zhao, L.; Wang, J.; Zhou, S.; Tang, J. Enhancing long-range depth estimation via heterogeneous CNN-transformer encoding and cross-dimensional semantic fusion. *Sci. Rep.* **2026**, *16*, 9396. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Munich, Germany, 5–9 October 2015; pp. 234–241. [\[CrossRef\]](#)
28. Xian, K.; Zhang, J.; Wang, O.; Mai, L.; Xu, Z.; Cao, Z. Structure-Guided Ranking Loss for Single Image Depth Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 611–620.
29. Yuan, W.; Gu, X.; Dai, Z.; Zhu, S.; Tan, P. NeW CRFs: Neural Window Fully-Connected CRFs for Monocular Depth Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), New Orleans, LA, USA, 18–24 June 2022; pp. 3916–3925.
30. El Ogri, O.; El-Mekkaoui, J.; Hjouji, A. A computer-assisted medical diagnosis system for cancer diseases based on quaternion orthogonal Rademacher-Fourier moments and deep learning. *Biomed. Signal Process. Control* **2026**, *89*, 108744. [\[CrossRef\]](#)
31. Chan, K.H.; Im, S.K. Sentiment analysis by using Naïve-Bayes classifier with stacked CARU. *Electron. Lett.* **2022**, *58*, 411–413. [\[CrossRef\]](#)
32. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
33. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520. [\[CrossRef\]](#)
34. Wang, P.; Shen, X.; Lin, Z.; Cohen, S.; Price, B.; Yuille, A.L. Towards Unified Depth and Semantic Prediction from a Single Image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), Boston, MA, USA, 7–12 June 2015; pp. 2800–2809. [\[CrossRef\]](#)
35. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; pp. 8024–8035.

36. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA, 6–9 May 2019.
37. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the International Conference on Learning Representations (ICLR 2017), Toulon, France, 24–26 April 2017.
38. Kim, D.; Sohn, C. Global-Local Path Networks for Monocular Depth Estimation with Vertical CutDepth. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), New Orleans, LA, USA, 18–24 June 2022; pp. 2858–2868.
39. Bochkovskii, A.; Delaunoy, A.; Germain, H.; Santos, M.; Zhou, Y.; Richter, S.R.; Koltun, V. Depth Pro: Sharp Monocular Metric Depth In Less Than A Second. In Proceedings of the International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025.

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.