TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214
DOI: 10.26599/TST.2025.9010036

Enhanced Monocular Depth Estimation Based on Uncertainty Edge Weighting Mask

Ye-Ji Kim, Byung-Gyu Kim*

Dept. of IT Engineering, Sookmyung Women's University, Seoul 04310, Republic of Korea;

Abstract: Monocular depth estimation is a critical component in understanding spatial relationships for various computer vision applications, including autonomous driving and augmented reality. However, accurate depth prediction remains challenging due to two primary factors: (1) the low pixel density of objects in distant regions and (2) the loss of essential features during the resolution reduction process in traditional encoder architectures. To address these challenges, this work introduces an innovative encoder-decoder architecture that incorporates uncertainty maps to improve feature extraction, particularly in long-distance regions. The proposed model utilizes auxiliary uncertainty networks to identify areas with high prediction difficulty, enabling the generation of more robust feature representations through hierarchical feature combinations. Additionally, the decoder architecture is designed to emphasize structural details by introducing an uncertainty edge weighting mask (UEWM) generation module, which further enhances depth prediction performance in challenging regions. Experimental results demonstrate that the proposed method significantly improves depth estimation accuracy in long-range scenarios, as evaluated on the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) and Dense Depth for Autonomous Driving (DDAD) datasets. These findings highlight the potential of this uncertainty-aware monocular depth estimation approach for practical applications, including autonomous driving and robotic perception.

Key words: monocular depth estimation, uncertainty, deep learning, uncertainty edge weighting

1 Introduction

Depth estimation, the process of predicting object distance from a given viewpoint, plays a crucial role in enabling spatial understanding across various applications. It is especially vital in autonomous driving, where precise depth information is required to detect obstacles, navigate environments, and ensure safe decision-making. Advanced technologies, such as Light Detection and Ranging (LiDAR), have been widely adopted for this purpose due to their ability to produce highly accurate depth maps. However, the high costs and operational complexities associated with LiDAR systems limit their scalability for mass-market deployment. In autonomous driving systems, depth estimation is integral to perceiving the environment and ensuring safety. Accurate depth maps enable vehicles to identify obstacles, maintain appropriate distances, and plan collision-free paths. Although LiDAR has proven effective, its reliance on specialized hardware introduces challenges such as device bulkiness, high maintenance costs, and potential reliability issues under certain conditions (e.g., adverse weather). To address these limitations, researchers have increasingly focused on camera-based depth estimation techniques [1],

• Ye-Ji Kim is with the Dept. of IT engineering, Sookmyung Women's University, Seoul 04310, Republic of Korea. E-mail: yj.kim@ivpl.sm.ac.kr

∗ Prof. Byung-Gyu Kim, Dept. of IT engineering, Sookmyung Women's University, E-mail: bg.kim@sookmyung.ac.kr.
Manuscript received: 2025-01-20; revised: 2025-02-28; accepted: 2025-03-18

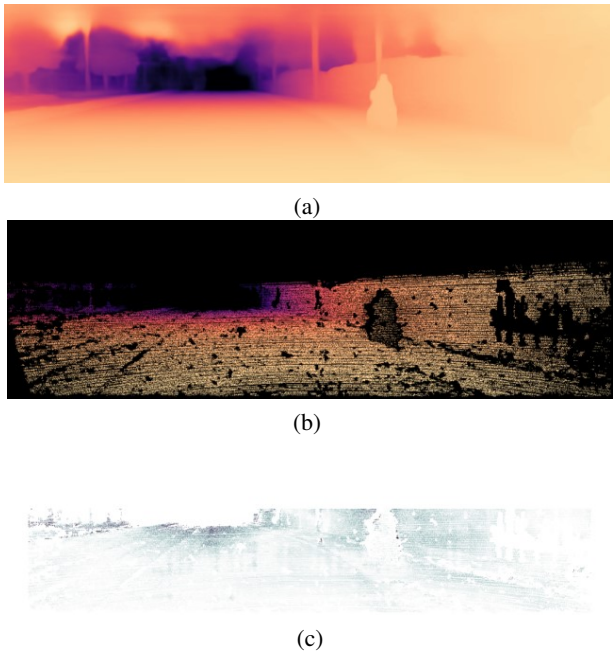

Fig. 1 The visualization of the the difference between the predicted depth map and ground truth depth map: (a) the depth map predicted by GEDepth [9], (b) the ground truth depth map, and (c) the relative difference in depth.

[2], [3]. Cameras, being compact, cost-effective, and widely available, offer an appealing alternative for depth prediction. Furthermore, advancements in computer vision and deep learning have enhanced the accuracy and efficiency of camera-based methods, demonstrating their viability for real-world autonomous driving applications [4], [5].

Conventional approaches to depth estimation primarily relied on disparity calculation using stereo image pairs. These methods aimed to achieve high-accuracy depth information by optimizing the correlation between two images through stereo matching techniques [6]-[14]. While effective, the reliance on stereo image pairs introduced complexities in hardware and system design. To address these challenges, recent research has focused on improving practicality and efficiency without compromising accuracy. Notably, monocular depth estimation techniques have emerged as a compelling alternative. These methods utilize a single image to predict depth, eliminating the need for stereo image pairs and the associated hardware complexity. Remarkably, monocular approaches have demonstrated performance comparable to, or even surpassing, traditional stereo-based methods, making them a promising solution for real-world applications.

Recent studies have advanced monocular depth estimation by categorizing it into self-supervised and supervised learning approaches. Self-supervised methods [15]-[22] aim to address the limitations of supervised learning by generating depth maps using pose networks, offering the advantage of reduced reliance on labeled data. However, their performance often falls short compared to supervised learning techniques [1]-[36]. However, supervised learning approaches continue to evolve, incorporating innovative techniques to achieve higher accuracy and significantly enhance the performance of monocular depth estimation.

Monocular depth estimation typically relies on an encoder-decoder architecture, where the encoder extracts features from the input image, and the decoder generates the corresponding depth map. However, this process faces a significant challenge: during iterative down-sampling in the encoder stage, spatial details are inevitably lost. This loss of detail adversely impacts prediction accuracy, particularly in long-range regions, where precise depth estimation is more challenging.

Figure 1 illustrates the error visualization example between the predicted depth map using the Ground Embedding for Monocular Depth Estimation (GEDepth) [9] model and the actual depth values. In this visualization, darker colors represent larger errors compared to the ground truth. The results reveal that the prediction errors are more pronounced in long-distance regions. This is primarily due to structural limitations, as only a small amount of pixel information is available for predicting depth values for distant objects. Additionally, when calculating the absolute depth difference for visualization, the absolute error tends to increase proportionally with the depth value, further amplifying the errors at greater distances. To compensate for this, the relative error is as follows:

$$\text{Relative Error} = \frac{\left|\hat{D}_i - D_i^{gt}\right|}{D_i^{gt}} \tag{1}$$

where $\hat{D}_i$ is the predicted depth map and D_i^{gt} is ground truth depth map.

This visualization of relative errors highlights the difficulty of the network in accurately predicting depth in long-range regions. These challenges stem from two primary factors: the limited amount of information available in distant areas and the loss of spatial detail caused by down-sampling during encoding. To address

these limitations, this paper introduces an encoder and decoder modeling technique designed to enhance prediction accuracy specifically in long-range regions. The contribution of this paper can be addressed as follows:

- We utilize an uncertainty map to identify some areas where more information is needed in the encode model. By providing more detailed feature information to areas with high uncertainty, this approach mitigates the errors associated with distant objects.
- A novel decoder learning method is proposed to improve the network's ability to effectively capture edge information in long-range regions, thereby boosting overall performance.

By providing more detailed feature information to areas with high uncertainty and designing a mechanism to give more attention for long-range regions, the network's ability can be effectively improved.

The rest of the paper is as follows. Section 2 presents the related work. Section 3 describes the proposed network framework in detail. Section 4 provides simulation results and performance evaluation. Finally, Section 5 concludes the paper and discusses future work.

2 Related Work

2.1 Encoder modeling for monocular depth estimation

Supervised learning-based monocular depth estimation models typically adopt an encoder-decoder architecture, with methods generally categorized into encoder modeling and decoder modeling. Encoder modeling focuses on extracting features that are well-suited for generating accurate depth maps.

However, convolutional neural network (CNN)-based encoders face challenges in capturing the global context of an image due to their inherently limited receptive field. To address this limitation, models such as From big to small (BTS) [10], Deep ordinal regression network (DORN) [2], multiscale feature fusion depth prediction (MFFDP) [11], and Global Understanding module (GUM) [12] have incorporated techniques such as atrous spatial pyramid pooling (ASPP) [13] further improves performance by leveraging channel and spatial attention modules to better utilize spatial information. Despite these advancements, the ability of CNN-based encoders to fully capture spatial details remains limited, leaving room for further improvement.

Transformer-based encoders have been introduced to address the limitations of CNN-based approaches. For example, the Dense Prediction Transformer (DPT) [34] model replaces CNNs with Vision Transformers and utilizes patch embedding resampling to generate dense and consistent depth maps in depth estimation tasks. Similarly, the bilateral residual depth network (BRNet) [35] enhances the accuracy of the depth map by incorporating transformer-based branches that effectively capture global contextual information. Although Transformers excel in modeling global contexts, they face challenges in capturing local patterns within images. Moreover, the patch-based embedding process can disrupt spatial continuity, which may hinder the network's ability to maintain consistent depth predictions across the image. As a result, Transformer-based encoders may struggle with accurate depth estimation, particularly in long-range regions where precision is critical.

DepthFormer [36] by Li *et al*. combines the strengths of CNNs and Transformers to enhance depth map accuracy, using modules that integrate features between encoders and decoders for improved performance across both close and far distances.

The use of attention mechanisms in Transformers allows the model to capture global information, ensuring consistent depth predictions across the same object within a single image. This approach has led to performance improvements in monocular depth estimation. However, it still has limitations in capturing fine-grained features in challenging regions, particularly in long-range areas. While Attention-based methods excel at incorporating global contextual information, they may suffer from information loss in distant regions, where fine-grained structural learning at the pixel level is crucial.

To address this issue, we propose an uncertainty edge weighting mask (UEWM) method. The proposed method is designed to effectively learn structural features in highly uncertain long-range regions, thereby improving depth prediction accuracy in areas where existing models struggle.

2.2 Monocular depth estimation based on uncertainty

Various studies have explored the use of uncertainty to improve the accuracy of monocular depth estimation. Poggi *et al*. [15] investigated the intrinsic causes of uncertainty in self-supervised methods, highlighting their limitations. Nie *et al*. [17] proposed a method to iteratively improve model performance by incorporating self-improving uncertainty during training. Xiang *et al*. [18] enhanced prediction reliability by enabling models to quantify and understand uncertainty. Zhou *et al*. [19] improved self-supervised depth estimation using self-distillation and uncertainty boosting techniques, enhancing overall reliability.

Hornauer *et al*. [23] introduced the Gradient-based Uncertaining method to quantify uncertainty in depth estimation. Asai *et al*. [24] proposed a multi-task learning approach that jointly optimizes monocular depth estimation and semantic segmentation, enhancing both depth prediction accuracy and predictive uncertainty. Additionally, significant efforts have been made to enhance monocular depth estimation models using self-supervised learning. Kendall *et al*. [25] also integrated Bayesian techniques into deep learning models, providing a robust framework for quantifying uncertainty and improving the reliability of depth estimation systems. Franci *et al*. [26] developed a scalable deterministic uncertainty method (DUM) that mitigates the Lipschitz constraint, facilitating efficient uncertainty estimation.

However, these studies primarily focus on overall depth estimation accuracy and overlook specialized improvements for distant regions. To address this gap, we propose an encoder-decoder modeling approach specifically designed to improve depth estimation accuracy in long-range regions. This approach complements previous works by optimizing depth predictions where existing models often struggle, ensuring high accuracy at far distances.

In addition, studies using vanishing points [43],[44] were conducted to improve the prediction accuracy of depth estimation. However, extracting vanishing points accurately is very difficult problem in various environments such as road driving, and securing a dataset containing vanishing point information is also practically limited. To overcome this limitation, we propose a method of dynamically calculating the area where uncertainty is concentrated instead of directly utilizing vanishing points and generating a weight mask based on them. Nearby vanishing points typically correspond to distant regions, and hence weighting around vanishing points enables effective distant region learning. The proposed method induces the network to learn structural features at long distances more effectively by weighting around regions with high uncertainty.

3 PROPOSED METHOD

This section highlights two key strategies for enhancing accuracy in high-uncertainty depth estimation: (1) hierarchical feature enhancement in the encoder and (2) uncertainty-edge-weighted depth regression in the decoder. The entire architecture of the proposed model is presented in Figure 2. The proposed network consists of three parts: Uncertainty Network, Encoder, and Decoder. Specifically, Uncertainty Networks play an important role for encoders and decoders. In the encoder stage, the uncertainty network provides uncertainty-aware features, enabling them to learn more robust feature representations. This allows the network to extract more robust features, even in areas where predictions are difficult, which are covered in detail in Section 3.1. In the decoder stage, the uncertainty network is leveraged to generate an uncertainty map, and based on this, a mask that highlights the critical edges of the long-distance regions is generated. This mask helps the network learn the long-distance depth information more precisely, which is discussed in Section 3.2.

3.1 Hierarchical Feature Enhancement Modeling in the Encoder

We propose a novel encoder structure that leverages uncertainty information to enhance depth estimation accuracy in high-uncertainty regions for monocular depth estimation. The encoder extracts compact features, capturing local and global relationships from input RGB images, to address challenging depth estimation areas. This design integrates an uncertainty-auxiliary network, CNN block, transformer block, and Hierarchical Aggregation and Heterogeneous Interaction (HAHI) module [36]. Specifically, the Uncertainty-Auxiliary Network (Θ_u) generates the uncertainty map (U) and hierarchical features $\mathbf{F}_u = \{F_u^i \in \mathbb{R}^{H/2^i \times W/2^i \times C}, i = 1, 2, 3, 4\}$ from the input

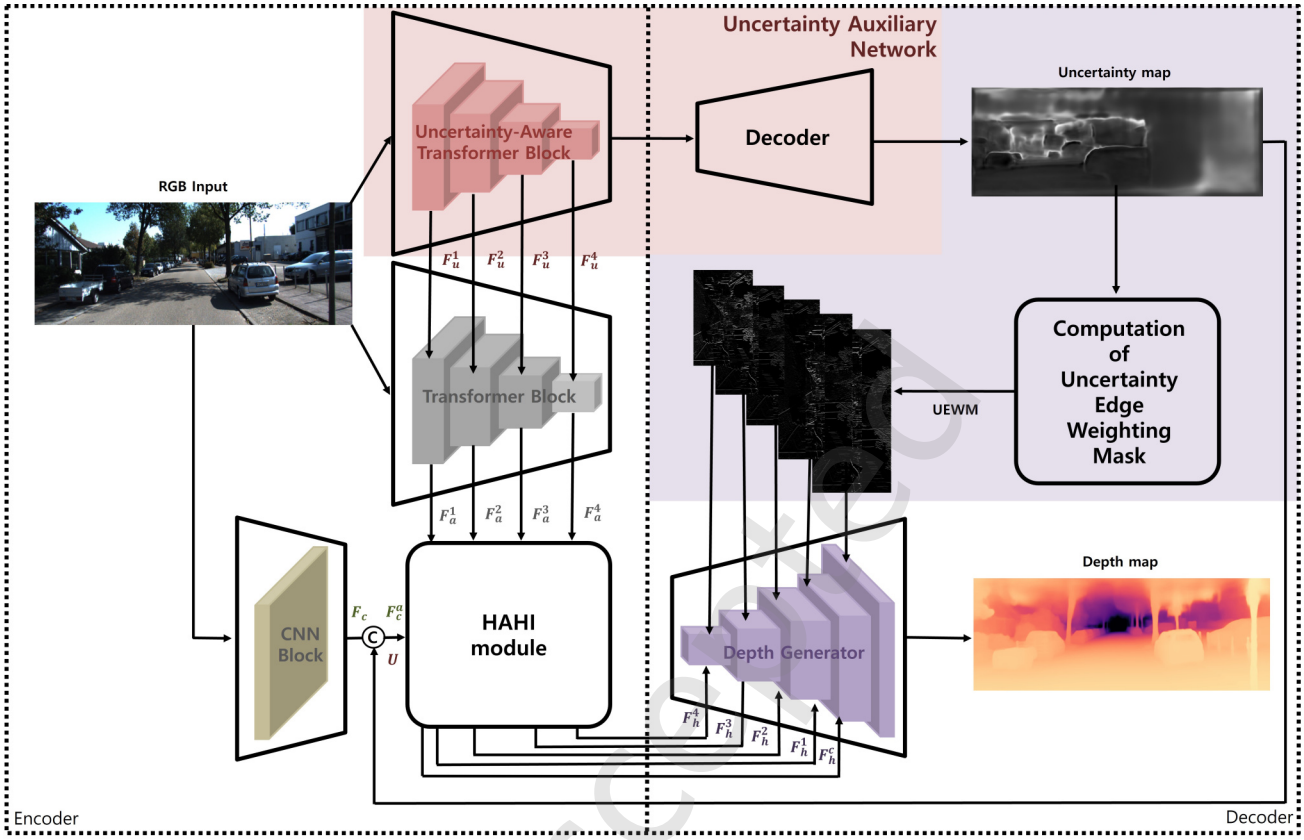

Fig. 2 The architecture of the proposed network model.

RGB image, $I = \{I \in \mathbb{R}^{H\times W\times 3}\}$.

$$F_u^1, F_u^2, F_u^3, F_u^4, U = \Theta_u(I), \tag{2}$$

where U is the predicted uncertainty map, F_u^i is the uncertainty-based feature, H is the height, W is the width, and C is the number of channels. The CNN block (Θ_c) learns the input image locally to extract the feature (F_c).

$$F_c = \Theta_c(I). \tag{3}$$

The Transformer block (Θ_t) combines the RGB image I and the uncertainty-encoded features (F_u), generating the feature $\mathbf{F}_a = \{F_a^i \in \mathbb{R}^{H/2^i\times W/2^i\times C}, i = 1,2,3,4\}$, which captures both the global dependencies and the uncertainty information.

$$F_a^1, F_a^2, F_a^3, F_a^4 = \Theta_t(I, F_u^1, F_u^2, F_u^3, F_u^4). \tag{4}$$

In the hierarchical feature fusion process, the F_u^i and F_a^i from each layer are concatenated along the channel dimension, and enhanced features F_a^i are generated through a 1×1 convolution, normalization, and ReLU activation.

$$\hat{z}_i = \text{MSA}(\text{LN}(z_{i-1})) + z_{i-1}, \tag{5}$$

$$z_i = \text{MLP}(\text{LN}(\hat{z}_i)) + \hat{z}_i, \tag{6}$$

$$F_a^i = \text{ReLU}(\text{Norm}(\text{Conv}(\text{cat}(z_i, F_u^i)))), \tag{7}$$

where i is a layer index, MSA is a Multi-Head Self-Attention, MLP is a Multi-Layer Perceptron, LN is Layer Normalization, and z is a feature extracted from transformer block. The uncertainty map (U) is added to F_c extracted from the CNN block to generate the uncertainty-aware feature (F_c^a).

$$F_c^a = F_c + U. \tag{8}$$

Finally, the HAHI module [36] (Θ_h) takes $\mathbf{F}_a$ and F_c^a as inputs, incorporates regional and global relationships, and preserves and reinforces uncertainty information. This eventually delivers the encoded feature ($\mathbf{F}_{enh}$) to the decoder (depth generator).

$$\mathbf{F}_{enh} : F_h^1, F_h^2, F_h^3, F_h^4, F_h^c = \Theta_h(F_a^1, F_a^2, F_a^3, F_a^4, F_c^a), \tag{9}$$

where $F_h^i (i = 1,2,3,4,c)$ is the output feature of the HAHI module [36].

The generated uncertainty perception feature accurately identifies high-uncertainty regions and provides representations closely matching their correct depth values. This approach addresses a key challenge in monocular depth estimation by improving depth estimation accuracy in such areas. The encoder structure is illustrated in Figure 3.

3.2 Depth Map Regression Based on UEWM in the Decoder

The uncertainty network generates an uncertainty map, where higher values indicate lower confidence in depth prediction, often corresponding to distant objects. We propose an uncertainty edge weighting mask (UEWM) to emphasize structural details in high-uncertainty regions. This mask enhances the network's ability to accurately predict depth in these areas.

3.2.1 Computation of UEWM

First, we apply a 3×3 Sobel filter mask to extract the edge of the object from the input RGB image (I), which we define as the edge map (S):

$$S = G_{x,y}(I). \tag{10}$$

Here, $G_{x,y}$ denotes the Sobel filter mask, and S represents the filtered edge map.

In regression-based Monocular Depth Estimation, it is challenging to directly interpret the uncertainty of depth predictions. Based on Xiang *et al.* [18], we apply Softmax to transform the model's predictions into a probabilistic perspective and use entropy to quantify the uncertainty. Based on this approach, an uncertainty map (U) is generated, and the uncertainty for each pixel is defined as follows:

$$P = Softmax(z), \tag{11}$$

$$U = \alpha \times \mathbb{H}(D') = -\alpha \sum_{i=1}^{n} P_i \times \log(P_i), \tag{12}$$

where z is the latent feature produced by the model before regressing the final depth map, i denotes the index of a depth range, and n represents the total number of depth ranges. P indicates the probability that a pixel belongs to each depth range. $\mathbb{H}$ is the entropy function, and D' is a random variable to represent the depth prediction. α is a learnable parameter that adjusts uncertainty values for stable training.

Based on the uncertainty map (U), we generate a weight mask to highlight distant regions by setting a threshold (U_{th}) as the average pixel value of the uncertainty map:

$$U_{th} = \frac{\sum_{p\in\Omega} U(p)}{N} \tag{13}$$

where Ω represents the entire set of pixels in the image, p represents individual pixels, and N represents the total number of pixels.

Then, the center point of the region is calculated by selecting the pixels with uncertainty values greater than or equal to U_{th}. The center points (V_x, V_y) represent the center positions of the region with high uncertainty.

$$V_x = \frac{\sum_{U(p)\geq U_{th}} x_p}{N_{high}}, \tag{14}$$

$$V_y = \frac{\sum_{U(p)\geq U_{th}} y_p}{N_{high}}, \tag{15}$$

where x_p and y_p are the location of the extracted uncertainty pixels, N_{high} means the number of pixels with high uncertainty. By calculating the Euclidean distance between the center points (V_x, V_y) and each pixel (x, y), a distance-based weight map W_{dist} is generated. Based on this, we subtract W_{dist} from the

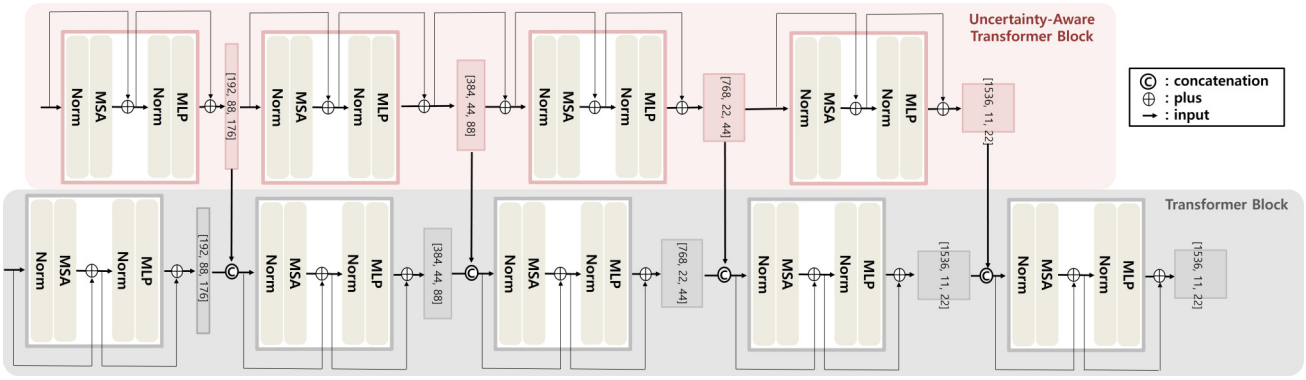

Fig. 3 The details of encoder architecture.

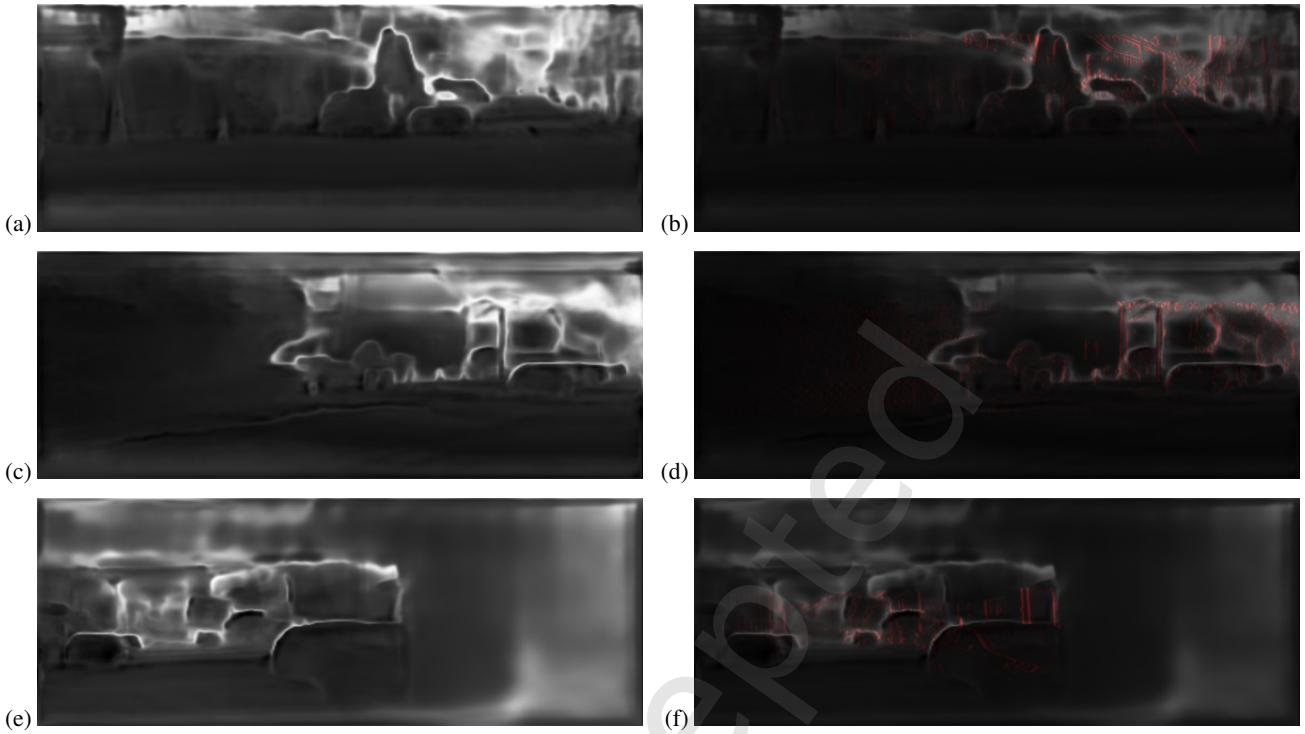

Fig. 4 Left column: Predicted uncertainty maps, right column: the edge-aware weighting mask based on this uncertainty map by the proposed method.

maximum value of the entire pixel value to give a larger weight to the pixels close to the center point (V_x, V_y), which is normalized to generate the final weight map $\hat{W}_{dist}$:

$$W_{dist}(p; x, y) = \sqrt{(V_x - x)^2 + (V_y - y)^2}, \quad (16)$$

$$W'_{dist}(p; x, y) = \max(W_{dist}(p; x, y)) - W_{dist}(p; x, y), \quad (17)$$

$$\hat{W}_{dist}(p; x, y) = \frac{W'_{dist}(p; x, y)}{\sqrt{H \times W}}, \quad (18)$$

Subsequently, we combine normalized distance-based weights with the uncertainty map (U) to generate the uncertainty weight map (E_{uncert}):

$$E_{uncert} = W_{dist}(p; x, y) \times U(p). \quad (19)$$

Finally, we multiply the edge map (S) with the uncertainty weight map (E_{uncert}) to generate the final uncertainty edge weight mask (E_w):

$$E_w = E_{uncert} \times S. \quad (20)$$

Figure 4 compares the predicted uncertainty map $(U(p))$ with the proposed UEWM (E_w). The left image displays the original $U(p)$, while the right image shows E_w, created by combining $U(p)$ and the edge map (S). Red areas represent E_w, clearly highlighting edges in high-uncertainty regions. This demonstrates that structural information in uncertain areas, including distant objects, is effectively emphasized.

3.2.2 Hierarchical Fusion of Uncertainty-Aware Features for Depth Map Regression

During decoding, the encoded feature F_n $(n = 1, 2, 3, 4, c)$ is combined with an uncertainty edge weighting map of the same resolution in each decoder layer. This hierarchical structure up-samples the features five times to the resolution of the next layer. The uncertainty edge weighting map provides shape information for distant objects, enabling the model to capture their details more effectively.

$$F_d^{i+1} = LeakyReLU(Norm(Conv(F_u^I(+)F_u^{i+1}))), \quad (21)$$

$$\hat{F}_d^{i+1} = F_d^{i+1} + E_w, \quad (22)$$

$$D = \theta_d(F_h^1, F_h^2, F_h^3, F_h^4, F_h^c, E_w^1, E_w^2, E_w^3, E_w^4, E_w^5). \quad (23)$$

Here, F_d^i $(i = 1, 2, 3, 4)$ represents features from the i-

th decoder layer, F^dc is the feature combined with the uncertainty edge weighting map, and D is the regressed depth map. The structure outputs the decoded feature F_d^5 to generate the depth map. The detailed decoder structure is shown in Figure 5. This structure guides the network to identify high-uncertainty regions, providing valuable information for more accurate predictions.

4 Experimental Results

4.1 Datasets

This study evaluates methods to enhance long-distance depth estimation performance using outdoor datasets collected in road driving environments. Specifically, the KITTI [37] and DDAD [38] datasets were used to compare and analyze monocular depth estimation performance in autonomous driving systems. These datasets, with their unique characteristics, serve as key benchmarks for validating depth estimation across diverse driving scenarios.

(1) KITTI: The KITTI [37] dataset is one of the most widely used benchmark datasets in monocular depth estimation research, in which RGB images with a resolution of 1241 × 376 pixels are used as input data and depth maps obtained by LiDAR sensors are used as ground truth. Furthermore, the KITTI [37] dataset provides calibration information between the sensor and the camera to help accurately apply intrinsic and extrinsic parameters. In this study, we follow the standard Eigen [39] training/test split method of the KITTI [37] dataset, using 23,158 images as training data and 697 images as test data.

(2) DDAD: Compared to the standardized KITTI [37] dataset, the DDAD dataset provides high-resolution and high-density depth information that supports a wide range of tasks required for autonomous systems. Specifically, the KITTI [37] dataset provides a depth range from 0 to 80 m, while the DDAD dataset [38] supports a wider prediction range from 0 to 250 m, providing a more challenging benchmark. In this work, we utilized the DDAD [38] dataset to use 12,650 images (150 scenes) for training and 3,950 images (50 scenes) for testing, and all experiments were conducted at four time points: forward, backward, left, and right, as defined in [40].

4.2 Evaluation Metrics

To demonstrate the performance of Monocular Depth Estimation, all results are evaluated with a total of seven metrics. These metrics include Absolute Relative Error (Abs Rel), Squared Relative Error (Sq Rel), Log root mean square error (RMSE log), and Threshold-based accuracy ($\delta < 1.25$, 1.25^2, and 1.25^3). Threshold-based accuracy refers to the probability that the ratio of predicted depth values to ground truth is less than certain thresholds (1.25, 1.25^2, 1.25^3), and so this probability increases as the threshold increases. The Abs Rel and RMSE metrics show higher performance with smaller values, and threshold-based accuracy shows higher performance with larger values.

Abs Rel : $\frac{1}{N}\sum_i \frac{\left|\hat{D}_i - D_i^{gt}\right|}{D_i^{gt}}$,

Sq Rel : $\frac{1}{N}\sum_i \frac{\left|\hat{D}_i - D_i^{gt}\right|^2}{D_i^{gt}}$,

RMSE : $\sqrt{\frac{1}{N}\sum_i \left|\hat{D}_i - D_i^{gt}\right|^2}$,

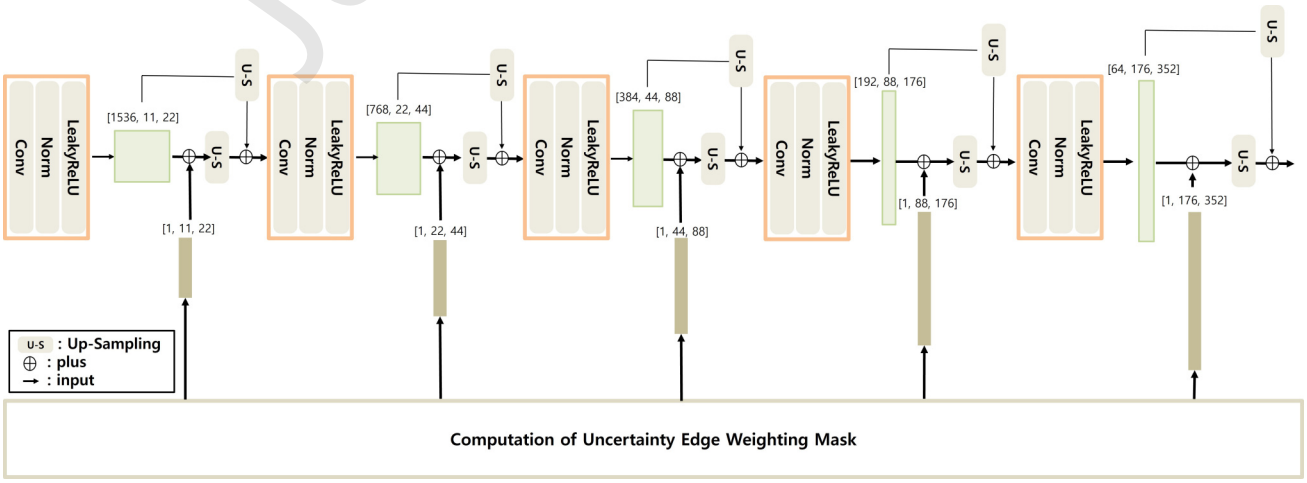

Fig. 5 The details of decoder architecture.

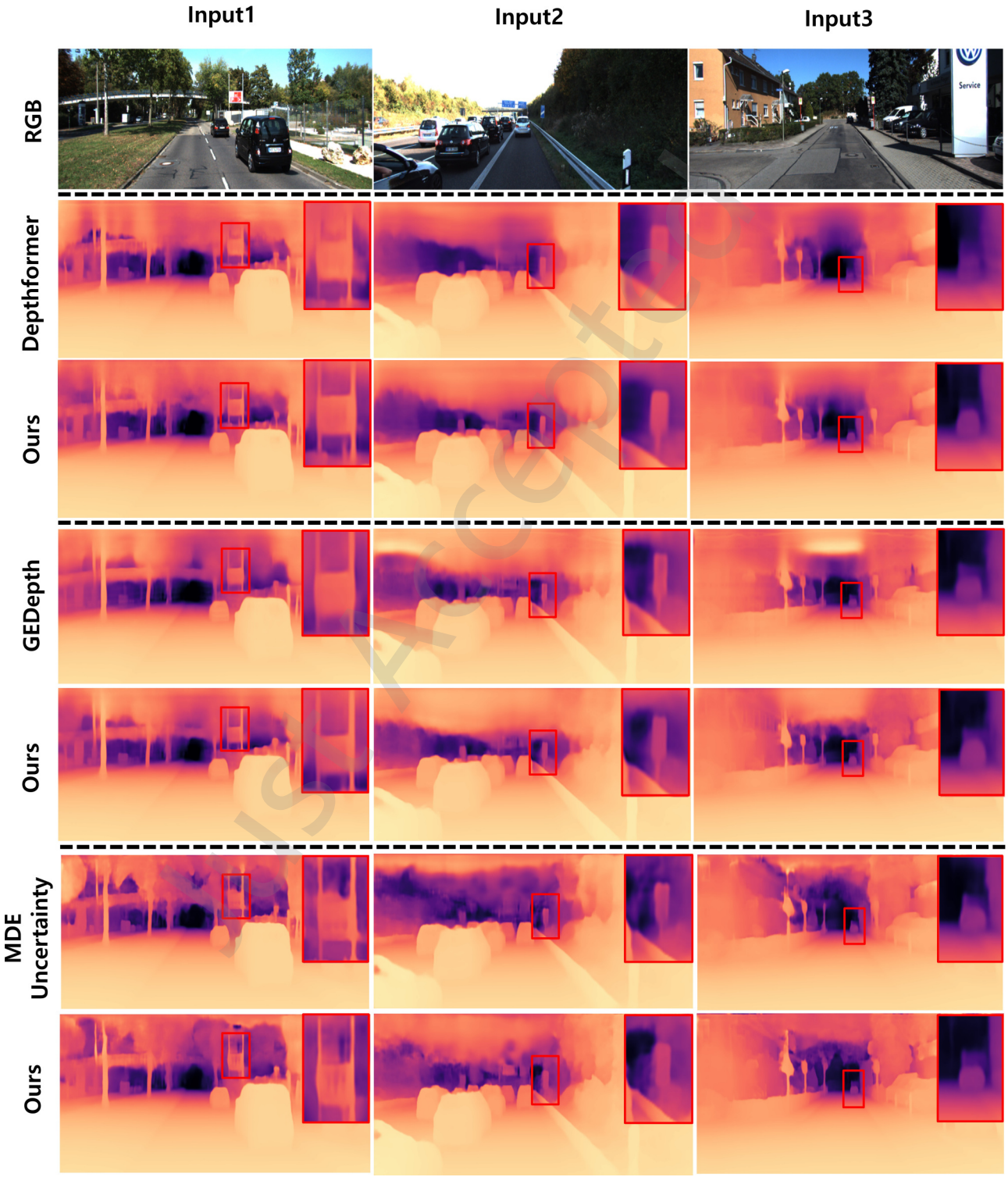

Fig. 6 Comparison of the depth prediction results by the other methods and our approach on three scenes of KITTI [37].

Table 1 Comparison of the model with the proposed method and monocular depth estimation model based on Swin-Transformer [41] on KITTI [37] dataset ([value] = amount of improvement, red color=best performance).

Model	Range	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
DORN [2]	Overall	0.072	0.307	2.727	0.120	-	-	-
DPT [34]	Overall	0.062	0.222	2.575	0.092	-	-	-
AdaBins [3]	Overall	0.058	0.190	2.360	0.088	-	-	-
NeWCRFs [1]	Overall	0.052	0.155	2.129	0.079	-	-	-
DepthFormer [36]	0-20m	0.039	0.104	2.087	0.057	0.995	0.998	0.999
	60-80m	0.198	0.948	12.981	0.826	0.588	0.957	0.981
	Overall	0.052	0.156	2.133	0.079	0.974	0.997	0.999
DepthFormer	0-20m	0.041	0.165	2.184	0.063	0.991	0.999	0.999
(+) Proposed	60-80m	0.190[-0.008]	0.828[-0.120]	9.653[-3.328]	0.400[-0.426]	0.782[+0.194]	0.942[+0.015]	0.983[+0.002]
method	Overall	0.050[-0.002]	0.144[-0.008]	2.105[-0.028]	0.074[-0.005]	0.975[+0.001]	0.997[-]	0.997[-]
GEDepth [9]	0-20m	0.034	0.089	1.912	0.049	0.996	0.998	1.000
	60-80m	0.155	0.877	10.854	0.692	0.618	0.888	0.925
	Overall	0.048	0.142	2.050	0.076	0.978	0.995	0.999
GEDepth	0-20m	0.039	0.103	1.995	0.051	0.996	0.997	1.000
(+) Proposed	60-80m	0.149[-0.006]	0.804[-0.073]	9.975[-0.879]	0.389[-0.303]	0.844[+0.226]	0.890[+0.010]	0.925[-]
method	Overall	0.047[-0.001]	0.138[-0.004]	2.033[-0.017]	0.069[-0.007]	0.979[+0.001]	0.996[-]	0.999[-]
MDEUncertainty [18]	0-20m	0.039	0.104	2.087	0.057	0.995	0.998	0.999
	60-80m	0.198	0.948	12.981	0.426	0.588	0.957	0.981
	Overall	0.052	0.156	2.133	0.079	0.974	0.997	0.999
MDEUncertainty	0-20m	0.039	0.103	1.995	0.051	0.996	0.997	1.000
(+) Proposed	60-80m	0.149[-0.006]	0.804[-0.073]	9.975[-0.879]	0.389[-0.303]	0.844[+0.226]	0.890[+0.010]	0.925[-]
method	Overall	0.047[-0.001]	0.138[-0.004]	2.033[-0.017]	0.069[-0.007]	0.979[+0.001]	0.996[-]	0.999[-]

RMSE log : $\sqrt{\frac{1}{N}\sum_i \left|\log \hat{D}_i - \log D_i^{gt}\right|^2}$,

Threshold-based accuracy :
$\%$ of d_i $s.t.$ $\max\left(\frac{\hat{D}_i}{D_i^{gt}}, \frac{D_i^{gt}}{\hat{D}_i}\right) = \delta < threshold.$

4.3 Results and Discussion

In this study, the performance of a monocular depth estimation model was evaluated separately for short and long distances. For the KITTI [37] dataset, within a depth range of [0m, 80m], short distances were defined as 0–20m and long distances as 60–80m. Similarly, for the DDAD [38] dataset, within a depth range of [0m, 250m], short distances were defined as 0–60m and long distances as 190–250m. This division was intended to focus on performance improvements in long-distance regions.

The proposed method, based on the Swin-Transformer backbone [41], is designed to be flexibly applicable to various depth estimation networks. Experimental results demonstrate significant improvement in depth prediction performance compared to existing Swin-Transformer-based monocular depth estimation models.

Notably, results on the KITTI [37] dataset (see Table 1) show substantial improvement in long-distance depth prediction, along with a modest overall performance enhancement. ↓ indicates that a lower value represents better performance, whereas ↑ indicates that a higher value represents better performance. We denoted the red color for the best performance in terms of the overall performance as shown in Table 1. These findings confirm the effectiveness of the proposed method in improving depth estimation for long-distance regions. Furthermore, compared to SOTAs such as DPT [34], AdaBins [3] and NeWCRF [1], the proposed scheme achieved better performance in terms of error metric and overall range.

A similar trend was observed with the DDAD [38] dataset as shown in Table 2, which covers a broader depth range than the KITTI [37] dataset. To compare performance, PackNet-SAN [42], BTS [10], PixelFormer [32], and BinsFormer [33] were employed. From the results, our proposed scheme showed the better performance in almost the performance measure. This led to even more significant performance improvements in distant regions, demonstrating the ability of the proposed model to achieve high accuracy

Table 2 Comparison of the model with the proposed method and monocular depth estimation model based on Swin-Transformer [41] on DDAD [38] dataset([value] = amount of improvement, red color=best performance).

Model	Range	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
PackNet-SAN [42]	Overall	0.187	2.776	11.936	0.276	-	-	-
BTS [10]	Overall	0.162	2.492	11.466	0.259	-	-	-
PixelFormer [32]	Overall	0.151	2.140	10.920	0.242	-	-	-
BinsFormer [33]	Overall	0.149	2.142	10.866	0.244	-	-	-
DepthFormer [36]	0-60m	0.114	1.986	10.003	0.213	0.962	0.975	0.992
	190-250m	3.728	8.452	92.213	3.842	0.378	0.751	0.978
	Overall	0.152	2.231	11.051	0.246	0.929	0.946	0.912
DepthFormer	0-60m	0.117	2.024	10.624	0.242	0.962	0.975	0.992
(+) Proposed	190-250m	3.116[-0.612]	8.043[-0.409]	89.872[-2.341]	3.587[-0.255]	0.378[-]	0.751[-]	0.978[-]
method	Overall	0.146[-0.006]	2.215[-0.016]	10.873[-0.178]	0.240[-0.006]	0.929[-]	0.946[-]	0.912[-]
GEDepth [9]	0-60m	0.098	1.787	9.882	0.197	0.975	0.980	0.992
	190-250m	3.445	8.026	84.432	3.519	0.417	0.782	0.979
	Overall	0.147	2.119	10.597	0.238	0.936	0.953	0.915
GEDepth	0-60m	0.102	1.804	9.911	0.230	0.975	0.980	0.992
(+) Proposed	190-250m	3.028[-0.417]	7.784[-0.242]	81.754[-2.678]	3.186[-0.333]	0.417[-]	0.782[-]	0.979[-]
method	Overall	0.143[-0.004]	2.093[-0.026]	10.432[-0.165]	0.235[-0.003]	0.975[-]	0.980[-]	0.992[-]

in environments requiring long-range depth prediction.

Qualitative evaluation compared depth maps predicted by the proposed model with those from existing models. As shown in Figure 6, the proposed method more accurately captured contours and depth information for long-distance objects. The images of the larger objects revealed clear distinctions between the objects and their backgrounds, with well-defined boundary lines. These results confirm that the proposed model excels in long-distance depth recognition, even in complex scenarios, and effectively performs monocular depth estimation.

4.4 Ablation Study

In the ablation study, we used the KITTI [37] dataset to evaluate the impact of each component of the proposed method on performance. To assess the effectiveness of the feature extraction approach, we compared and visualized features from baseline models and models using the proposed method. This analysis quantitatively and visually demonstrates the proposed method's contribution to improving depth estimation performance.

4.4.1 Evaluation of Key Components

The proposed encoder structure consists of two configurations: combining the uncertainty map with CNN-extracted features and integrating uncertainty-aware features with the transformer layer. Table 3 presents experimental results for applying the uncertainty map to the convolution layer, uncertainty-aware features to the transformer layer, and a combination of all modules. The results highlight that combining uncertainty-aware features with the transformer layer yields the most significant performance improvement, demonstrating their importance in enhancing feature extraction. Performance also improved when uncertainty maps were combined with the convolution layer. Applying only the uncertainty edge calculation module in the decoder resulted in modest improvements, attributed to residual errors outside edges in long-distance regions. However, when all proposed modules were applied together, all evaluation metrics showed significant performance gains. This confirms the individual effectiveness of each module and demonstrates their synergistic impact when combined.

4.4.2 Visual validation of feature extraction improvements for the proposed method in the encoder

We analyzed the limitations of existing depth network encoders and validated the improvements achieved with the proposed method. To assess its effectiveness, we compared and visualized features extracted from the first transducer layer, highlighting differences from existing methods. The feature resolution, 192×88×176, is a quarter of the input image size (352×712×3). Features were visualized by calculating the mean along the channel axis. We compared features extracted by

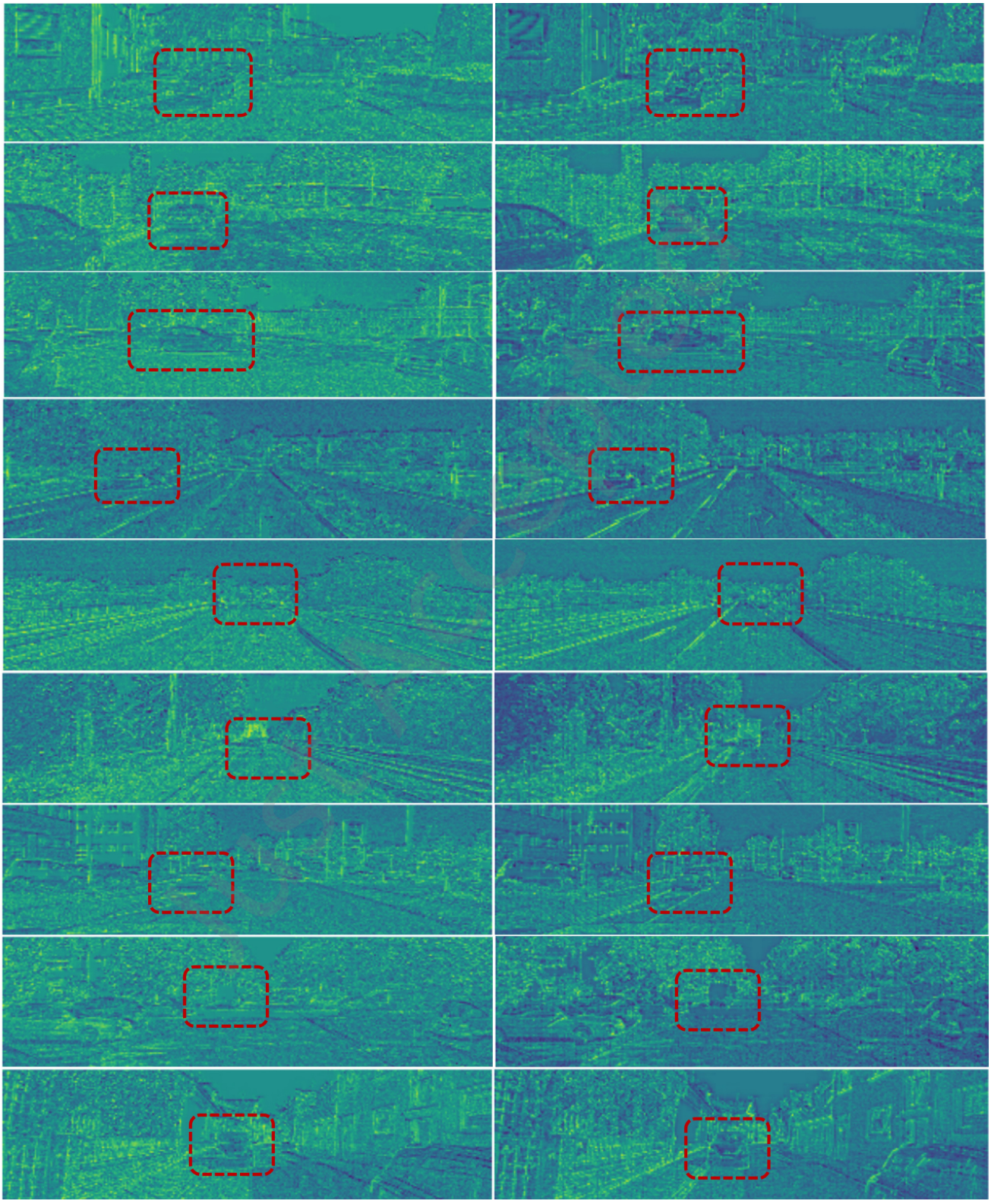

Fig. 7 Comparison between feature extracted from the encoder of GEDepth [9] (left column) and the proposed encoder modeling methods (right column) on KITTI [37] dataset.

Fig. 8 Comparison between feature extracted from the encoder of GEDepth [9] (left column) and the proposed encoder modeling methods (right column) on DDAD [38] dataset.

Table 3 Performance analysis for each component of the proposed method on KITTI [37] dataset.

	Component	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
	w/o proposed methods	0.048	0.142	2.050	0.076	0.978	0.995	0.999
Encoder	(a) CNN + uncertainty map	0.048	0.136[-0.006]	2.176	0.075[-0.001]	0.978	0.995	0.999
	(b) Transformer + uncertainty-aware feature	0.046[-0.002]	0.134[-0.008]	2.188	0.073[-0.003]	0.979[+0.001]	0.995	0.999
	(a) + (b)	0.047[-0.001]	0.139[-0.003]	2.187	0.073[-0.002]	0.978	0.995	0.999
Decoder	uncertainty edge weighting method	0.054	0.147	1.892[-0.175]	0.065[-0.011]	0.978	0.995	0.999
	All	0.047[-0.001]	0.138[-0.002]	2.033[-0.017]	0.069[-0.007]	0.978[+0.001]	0.995	0.999

the proposed encoder with those from the same layer of the existing GEDepth [9] model. Figures 7 and 8 illustrate these comparisons, with each row aligned by image type. Features from GEDepth [9] appear on the left, while those from the proposed method are on the right. The proposed method produced features that more clearly and precisely distinguished objects, particularly in long-distance regions. Figures 7 and 8 show that the proposed method better defines object boundaries and effectively identifies vehicles and lanes.

In contrast, features from GEDepth [9] showed unclear shapes and ambiguous boundaries. These findings demonstrate that the proposed feature enhancement modeling improves clarity and object distinction in high-uncertainty areas. Additionally, it mitigates information loss in distant regions caused by resolution reduction during feature extraction. This confirms that the proposed encoder modeling method is more efficient and provides superior performance compared to existing methods.

4.4.3 Inference Time

The inference speed of the proposed model remains similar to that of GEDepth [9], averaging 185 ms and 1029 ms per image on KITTI [37] and DDAD [38] datasets, respectively. This implies that long-distance depth estimation performance can be significantly improved without additional computational costs, especially in applications requiring high accuracy and real-time processing, such as autonomous driving.

5 Conclusions

This study has focused on enhancing monocular depth estimation, particularly in long-distance regions, using road-driving data. In autonomous driving scenarios, low pixel density in distant regions often leads to information loss during encoder resolution reduction. To address this, we proposed a novel encoder feature extraction technique utilizing uncertainty maps, enabling the identification and enhancement of challenging long-distance regions through hierarchical feature integration. Additionally, a decoder module was designed to generate edge weighting maps, effectively highlighting structural details and improving long-distance depth estimation. Experimental results demonstrated that the proposed technique significantly enhanced depth recognition in complex scenarios, delivering robust performance in monocular depth estimation. This highlights its practical applicability to real-world tasks like autonomous driving.

The proposed method is effective in improving depth estimation performance in long-range regions, but it is likely to show limitations in scenes containing low-light conditions or dynamic objects. In low-light environments, the signal-to-noise ratio (SNR) in the image may be lowered, which may reduce the reliability of the uncertainty map, and in the case of dynamic objects, the depth prediction performance may be degraded due to the occlusion problem. To overcome these limitations, future studies need to explore learning reliability-aware correction techniques and time-series information for dynamic objects considering low-light environments. Furthermore, it is necessary to focus on further improving long-distance accuracy while maintaining short-range performance and exploring multi-modal integration to enhance generalization across diverse environmental conditions. These advancements will pave the way for more reliable monocular depth estimation solutions.

Acknowledgment

References

References

[1] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan, New crfs: Neural window fully-connected crfs for monocular depth estimation, *arXiv preprint arXiv:2203.01502*, 2022

[2] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, Deep ordinal regression network for monocular depth

estimation, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2002-2011, 2018.

[3] S. F. Bhat, I. Alhashim, and P. Wonka, Adabins: Depth estimation using adaptive bins, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4009-4018, 2021.

[4] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, and Z. Li, Bevdepth: Acquisition of reliable depth for multi-view 3d object detection, *AAAI Conf. Artif. Intell.*, vol. 37, no. 2, pp. 1477–1485, Jun. 2023.

[5] J. Philion, and S. Fidler, Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d, in European Conference on Computer Vision (ECCV), pp. 194-210, 2020.

[6] Imran, S., Khan, M. U. K., Mukarram, S. B., and Kyung, C. M, Increased-Range Unsupervised Monocular Depth Estimation, arXiv preprint arXiv:2006.12791, 2020.

[7] T. Li and Y. Zhang, A Contour-Aware Monocular Depth Estimation Network using Swin Transformer and Cascaded Multi-scale Fusion, *IEEE Sensors Journal*, vol. 24, pp. 13620-13628, 2024.

[8] S. Shao, Z. Pei, W. Chen, X. Wu, and Z. Li, Nddepth: Normal-distance assisted monocular depth estimation, in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7931-7940, 2023.

[9] X. Yang, Z. Ma, Z. Ji, and Z. Ren, Gedepth: Ground embedding for monocular depth estimation, in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, pp. 12719-12727, 2023.

[10] J. H. Lee, M. K. Han, D. W. Ko, and I. H. Suh, From big to small: Multi-scale local planar guidance for monocular depth estimation, *arXiv preprint arXiv:1907.10326*, 2019.

[11] X. Xu, Z. Chen, and F. Yin, Monocular depth estimation with multi-scale feature fusion, *IEEE Signal Processing Letters*, vol. 28, pp. 678-682, 2021.

[12] J. Xiao, L. Li, X. Su, and G. Tan, Multi-scale Monocular Depth Estimation Based on Global Understanding, *IEEE Access*, vol. 12, pp. 46930-46939, 2024.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834-848, 2017.

[14] Godard, C., Mac Aodha, O., and Brostow, G. J, Unsupervised monocular depth estimation with left-right consistency, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 270-279, 2017.

[15] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, On the uncertainty of self-supervised monocular depth estimation, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 3227-3237, 2020.

[16] N. Hirose, S. Taguchi, K. Kawano, and S. Koide, Variational monocular depth estimation for reliability prediction, in 2021 International Conference on 3D Vision (3DV), pp. 637–647, 2021.

[17] X. Nie, D. Shi, R. Li, Z. Liu, and X. Chen, Uncertainty-aware self-improving framework for depth estimation, IEEE Robotics and Automation Letters, vol. 7, no. 1, pp. 41–48, 2021.

[18] M. Xiang, J. Zhang, N. Barnes, and Y. Dai, Measuring and modeling uncertainty degree for monocular depth estimation, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 5716–5727, Jul. 2024.

[19] H. Zhou, S. Taylor, D. Greenwood, and M. Mackiewicz, Sub-depth: Self-distillation and uncertainty boosting self-supervised monocular depth estimation, *arXiv preprint arXiv:2111.09692*, 2021.

[20] H. Zhou, S. Taylor, and D. Greenwood, Sub-depth: Self-distillation and uncertainty boosting self-supervised monocular depth estimation, arXiv preprint arXiv:2111.09692, 2021.

[21] H. Choi, H. Lee, S. Kim, S. Kim, S. Kim, K. Sohn, and D. Min, Adaptive confidence thresholding for monocular depth estimation, in IEEE International Conference on Computer Vision (ICCV), pp. 12808–12818, 2021.

[22] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," arXiv preprint arXiv:1511.02680, 2015.

[23] J. Hornauer and V. Belagiannis, Gradient-based uncertainty for monocular depth estimation, in European Conference on Computer Vision (ECCV), pp. 613–630, 2022.

[24] A. Asai, D. Ikami, and K. Aizawa, Multi-task learning based on separable formulation of depth estimation and its uncertainty, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop, pp. 21–24, 2019.

[25] A. Kendall and Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, in Advances in Neural Information Processing Systems (NeurIPS), pp. 5580–5590, 2017.

[26] G. Franchi, X. Yu, A. Bursuc, E. Aldea, S. Dubuisson, and D. Filliat, Latent discriminant deterministic uncertainty, in European Conference on Computer Vision (ECCV), pp. 243–260, 2022.

[27] N. Fang, L. Qiu, S. Zhang, Z. Wang, Z. Zhou, and K. Hu, GSDC Transformer: An Efficient and Effective Cue Fusion for Monocular Multi-Frame Depth Estimation, *IEEE Robotics and Automation Letters*, vol. 9, pp. 2256-2263, 2024.

[28] Roy, A., and Todorovic, S, Monocular depth estimation using neural regression forest, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 5506-5514, 2016.

[29] Luo, Y., Ren, J., Lin, M., Pang, J., Sun, W., Li, H., and Lin, L, Single view stereo matching, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 155-163, 2018.

[30] Tosi, F., Aleotti, F., Poggi, M., and Mattoccia, S, Learning monocular depth estimation infusing traditional stereo knowledge, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 9799-9809, 2019.

[31] Cho, J., Min, D., Kim, Y., and Sohn, K, A large RGB-D dataset for semi-supervised monocular depth estimation, arXiv preprint arXiv:1904.10230, 2019.

[32] A. Agarwal and C. Arora, Attention attention everywhere: Monocular depth prediction with skip attention, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5861-5870, 2023.
[33] Z. Li, X. Wang, X. Liu and J. Jiang, Binsformer: Revisiting adaptive bins for monocular depth estimation, *IEEE Trans. Image Process.*, 2024.
[34] R. Ranftl, A. Bochkovskiy, and V. Koltun, Vision transformers for dense prediction, in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12179-12188, 2021.
[35] W. Han, J. Yin, X. Jin, X. Dai, and J. Shen, Brnet: Exploring comprehensive features for monocular depth estimation, in *Proc. European Conference on Computer Vision*, Springer, pp. 586-602, 2022.
[36] Z. Li, Z. Chen, X. Liu, and J. Jiang, Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation, *Machine Intelligence Research*, vol. 20, no. 6, pp. 837-854, 2023.
[37] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, Vision meets robotics: The KITTI dataset, *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231-1237, 2013.
[38] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, 3D packing for self-supervised monocular depth estimation, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2485-2494, 2020
[39] D. Eigen and R. Fergus, Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 2650-2658, 2015.
[40] V. Guizilini, R. Ambrus, W. Burgard, and A. Gaidon, "Sparse auxiliary networks for unified monocular depth prediction and completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11078-11088, 2021.
[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in *Proc. IEEE/CVF International Conference on Computer Vision (CVPR)*, pp. 10012-10022, 2021.
[42] V. Guizilini, R. Ambrus, W. Burgard, and A. Gaidon, Sparse auxiliary networks for unified monocular depth prediction and completion, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11078-11088, 2021.
[43] H. Ibrahem, A. Salem, and H. S. Kang, Seg2Depth: Semi-supervised Depth Estimation for Autonomous Vehicles Using Semantic Segmentation and Single Vanishing Point Fusion. *IEEE Trans. Intell. Vehicles*, 2024.
[44] B. Li and X. Han, Enhanced monocular depth estimation: A CNN integrating semantic segmentation embedding and vanishing point detection, in *Proc. 7th Int. Conf. Adv. Algorithms Control Eng. (ICAACE)*, Mar. 2024, pp. 363–367.

Ye-Ji Kim received the BS degree in IT Engineering from Sookmyung Women's University, Seoul, South Korea, in 2022. She received a Master's degree in IT Engineering from the Graduate School of Sookmyung Women's University in 2024. Her research interests include scene understanding, depth estimation, segmentation, and object detection.

Byung-Gyu Kim has received his BS degree from Pusan National University, Korea, in 1996 and an MS degree from Korea Advanced Institute of Science and Technology (KAIST) in 1998. In 2004, he received a PhD degree in the Department of Electrical Engineering and Computer Science from Korea Advanced Institute of Science and Technology (KAIST). In March 2004, he joined the real-time multimedia research team at the Electronics and Telecommunications Research Institute (ETRI), Korea, where he was a senior researcher. In ETRI, he developed so many real-time video signal processing algorithms and patents and received the Best Paper Award in 2007. From February 2009 to February 2016, he was an associate professor in the Division of Computer Science and Engineering at SunMoon University, Korea. In March 2016, he joined the Department of Information Technology (IT) Engineering at Sookmyung Women's University, Korea where he is currently a full professor. In 2007, he is serving or served as an associate editor of Circuits, Systems and Signal Processing (Springer), The Journal of Supercomputing (Springer), Discover Computing (Springer), The Journal of Real-Time Image Processing (Springer), Heliyon Journal (Elsevier), and IEEE Access (IEEE). From 2018 to 2024, he served as the Editor-in-Chief (EiC) of the Journal of Multimedia Information System. From 2022, he was honored as the World Top 2% Scientist in Image & Artificial Intelligence field by Scopus and Stanford University.
He has published over 280 international journal and conference papers, patents in his field. His research interests include deep learning-based object detection, image enhancement, depth estimation, emotion recognition, video coding techniques, and intelligent scene understanding. He is a senior member of IEEE and a professional member of ACM, and IEICE.