Machine Intelligence Research
www.mi-research.net

20(6), December 2023, 837-854
DOI: 10.1007/s11633-023-1458-0

DepthFormer: Exploiting Long-range Correlation and Local Information for Accurate Monocular Depth Estimation

Zhenyu Li[1] Zehui Chen[2] Xianming Liu[1] Junjun Jiang[1]

[1] Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China

[2] Department of Automation, University of Science and Technology of China, Hefei 230026, China

Abstract: This paper aims to address the problem of supervised monocular depth estimation. We start with a meticulous pilot study to demonstrate that the long-range correlation is essential for accurate depth estimation. Moreover, the Transformer and convolution are good at long-range and close-range depth estimation, respectively. Therefore, we propose to adopt a parallel encoder architecture consisting of a Transformer branch and a convolution branch. The former can model global context with the effective attention mechanism and the latter aims to preserve the local information as the Transformer lacks the spatial inductive bias in modeling such contents. However, independent branches lead to a shortage of connections between features. To bridge this gap, we design a hierarchical aggregation and heterogeneous interaction module to enhance the Transformer features and model the affinity between the heterogeneous features in a set-to-set translation manner. Due to the unbearable memory cost introduced by the global attention on high-resolution feature maps, we adopt the deformable scheme to reduce the complexity. Extensive experiments on the KITTI, NYU, and SUN RGB-D datasets demonstrate that our proposed model, termed DepthFormer, surpasses state-of-the-art monocular depth estimation methods with prominent margins. The effectiveness of each proposed module is elaborately evaluated through meticulous and intensive ablation studies.

Keywords: Autonomous driving, 3D reconstruction, monocular depth estimation, Transformer, convolution.

Citation: Z. Li, Z. Chen, X. Liu, 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. http://doi.org/10.1007/s11633-023-1458-0

1 Introduction

Monocular depth estimation plays a critical role in three dimensional reconstruction and perception. Since the groundbreaking work of [1], convolutional neural network (CNN) has dominated the primary workhorse for depth estimation, in which the encoder-decoder based architecture is designed[2–4]. Although there have been numerous work focusing on the decoder design[2, 4], recent studies suggest that the encoder is even more pivotal for accurate depth estimation[3, 5]. Due to the lack of depth cues, fully exploiting both the long-range correlation (i.e., distance relationship among objects) and the local information (i.e., consistency of the same object) are critical capabilities of an effective encoder[6]. Therefore, the potential bottleneck of current depth estimation methods may lie in the encoder where the convolution operators can scarcely model the long-range correlation with a limited receptive field.

In terms of CNN, there have been great efforts to overcome the above limitation, roughly grouped into two categories: manipulating the convolution operation and integrating the attention mechanism. The former applies advanced variations, including multi-scale fusion[7], atrous convolutions[8], and feature pyramids[9], to improve the effectiveness of convolution operators. The latter introduces the attention module[10] to model the global interactions of all pixels in the feature map. There are also several general approaches[2–4, 11] that explore the combination of both strategies. Though the performance is improved significantly, the dilemma persists.

Alternative to CNN, Vision Transformer (ViT)[12], which achieves tremendous success on image recognition, demonstrates the advantages of serving as the encoder for depth estimation. Benefiting from the attention mechanism, the Transformer is more expert at modeling the long-range correlation with a global receptive field. However, our pilot study (Section 3.1) indicates the ViT encoder cannot produce satisfactory performance due to the lack of spatial inductive bias in modeling the local information[13], leading to the poor performance on close-range depth estimation. In contrast, models with convolu-

Research Article
Manuscript received on March 5, 2023; accepted on May 26, 2023; published online on September 13, 2023
Recommended by Associate Editor Hao Dong
Colored figures are available in the online version at https://link.springer.com/journal/11633

tion encoders can better predict depth at these places. While a few papers propose to ensemble the operations[14–16], they focus on the classification task. For depth estimation, there is still a lack of research on combining them[13].

Therefore, we propose a novel monocular depth estimation framework, called DepthFormer (illustrated in Fig. 1), which boosts model performance by incorporating the advantages from both the Transformer and the CNN. The principle of DepthFormer lies in the fact that the Transformer branch models the long-range correlation while the convolution branch preserves the local information. We argue that the integration of parallel architecture can help achieve more accurate depth estimation. However, independent branches with simple late fusion lead to insufficient feature aggregation for the decoder. To bridge this gap, we design the hierarchical aggregation and heterogeneous interaction (HAHI) module to combine the best part of both branches. Specifically, it consists of a self-attention module to enhance the features among hierarchical layers of the Transformer branch via element-wise interaction and a cross-attention module to model the affinity between "heterogeneous" features (i.e., Transformer and CNN features) in a set-to-set translation manner. Since global attention on high-resolution feature maps leads to an unbearable memory cost, we propose to leverage the deformable scheme[17, 18] that only attends to a limited set of key sampling vectors in a learnable manner to alleviate this problem.

The main contributions are three-fold: 1) We design a parallel encoder architecture to exploit the long-range correlation and local information. 2) We design the HAHI to enhance features via element-wise interaction and model the affinity in a set-to-set translation manner. 3) Our proposed approach DepthFormer significantly outperforms state-of-the-arts with prominent margins on the KITTI[19], NYU[20] and SUN RGB-D[21] datasets. Furthermore, it achieves competitive result on the competitive KITTI depth estimation benchmark[1].

2 Related work

Estimating depth from RGB images is an ill-posed problem. Lack of cues, scale ambiguities, translucent or reflective materials all leads to ambiguous cases where appearance cannot infer the spatial construction. With the rapid development of deep learning, CNN has become a key component of mainstream methods to provide reasonable depth maps from a single RGB input. Meanwhile, depth estimation boosts other related fields such as the depth map super-resolution[22], high-quality[23], restoration[24], adversarial attack[25], etc.

Monocular depth estimation has drawn much attention in recent years. Among numerous effective methods, we consider dense prediction transformer (DPT)[5], Adabins[4] and TransDepth[13] as the most important three competitors.

DPT proposes to utilize ViT as the encoder and pretrain models on larger-scale depth estimation datasets. Adabins uses adaptive bins that dynamically change depending on representations of the input scene and proposes to embed the mini-ViT at a high resolution (after the decoder). TransDepth embeds ViT at the bottleneck to avoid the Transformer losing the local information and presents an attention gate decoder to fuse multi-level features. We focus on comparing these (and many other) methods in this paper.

Encoder-decoder is commonly used in monocular depth estimation[2–4, 11, 26–28]. In terms of the encoder, mainstream feature extractors, including EfficientNet[29], ResNet[1] and DenseNet[30], are adopted to learn representations. The decoder frequently consists of successive convolutions and upsampling operators to aggregate encoder features in a late fusion manner, recover the spatial resolution and estimate the depth. In this paper, we utilize the baseline decoder architecture in [31]. It allows us to more explicitly study the performance attribution of key contributions of this work, which are independent of the decoder.

Neck modules between the encoder and the decoder are proposed to enhance features. Many previous methods only focus on the bottleneck feature but ignore the lower-level ones, limiting the effectiveness[2, 3, 11, 13]. In this work, we propose the HAHI module to enhance all the multi-level hierarchical features. When another branch is available, it can model the affinity between the two-branch features as well, which benefits the decoder to aggregate the heterogeneous information.

Transformer networks are gaining greater interest in the computer vision community[12, 32–34]. Following the success of recent trends that apply the Transformer to solve computer vision tasks, we propose to leverage the Transformer as the encoder to model long-range correlations. In Section 3.1, we discuss our motivation and present differences between our method and several related works[4, 5, 13] that adopt the Transformer in monocular depth estimation.

3 Methodology

In this section, we present the motivation of this work and introduce the key components of DepthFormer: 1) an encoder consisting of a Transformer branch and a convolution branch and 2) the hierarchical aggregation and heterogeneous interaction (HAHI) module. Then, we introduce the simple decoder structure and supervision loss used in DepthFormer. An overview is shown in Fig. 2.

3.1 Motivation

To indicate the necessity of this work, we conduct a meticulous pilot study to investigate the limitations of ex-

[1] http://www.cvlibs.net/datasets/kitti/eval_depth.php?benchmark=depth_prediction

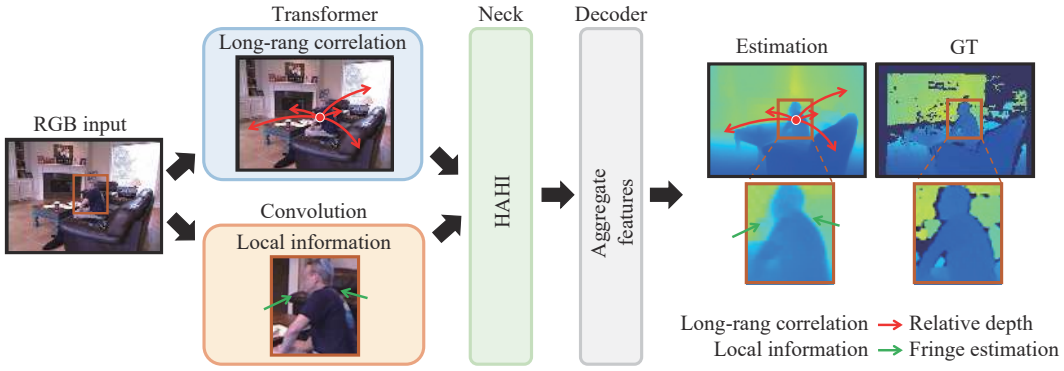


Fig. 1 Overview: We design a parallel encoder consisting of a Transformer branch to learn the long-range correlation and a convolution branch to extract the local information. To alleviate the lack of connections between the two branches, we propose the HAHI module to enhance features and model affinities.

isting methods that utilize pure CNN or ViT as the encoder in monocular depth estimation.

We first present several failure cases of the state-of-the-art CNN-based monocular depth estimation methods on the NYU dataset in Fig. 3. The depth results at the wall decorations and carpets are unexpectedly incorrect. Due to the pure convolutional encoder for feature extraction, it is hard for them to model the global context and capture the long-range distance relationship among objects through limited receptive fields. Such large-area counter-intuitive failures severely impair the model performance.

To solve the above issue, ViT can serve as a proper alternative that is superior in modeling the long-range correlation with a global receptive field. Therefore, we experiment to analyze the performance of ViT-based and CNN-based methods on the KITTI dataset. Specifically, we adopt flatten ViT-Base[12] and ResNet-50[1] as the encoder to extract features, respectively. The results shown in Table 1 prove that the models applying ViT as encoder outperform those using ResNet-50 on distant object depth estimation. However, opposite results appear on the near objects. Since the depth values exhibit a long tail distribution and there are much more near objects (pixels) in scenes[35], the overall results of the models applying ViT are significantly inferior.

In general, it is tougher to estimate the depth of distant objects directly. Benefiting from modeling the long-range correlation, the ViT-based model can be more reliable to accomplish it via reference pixels in a global context. The knowledge of distance relationships among objects results in better performance on distant object depth estimations. As for the inferior near object depth estimation result, there are many potential explanations.

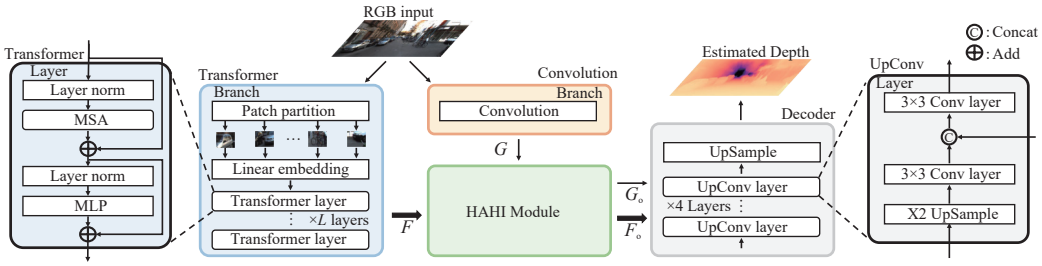


Fig. 2 An overview of DepthFormer. It comprises three major components: a parallel encoder consisting of a Transformer branch and a convolution branch, a hierarchical aggregation and heterogeneous interaction (HAHI) module, and a standard decoder. The HAHI enhances the Transformer features F and models the affinity between the Transformer and the convolution features G.

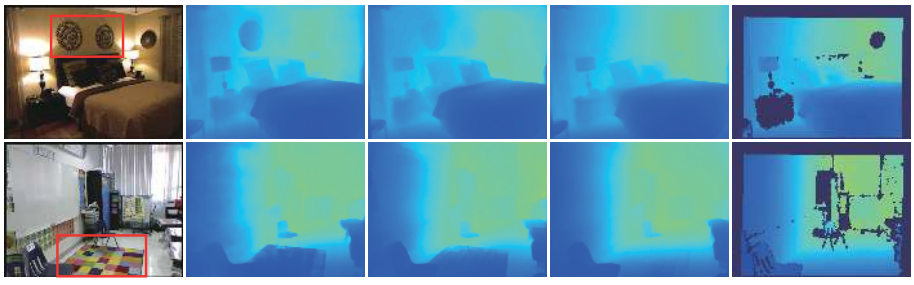


Fig. 3 Failure cases of previous methods on NYU dataset caused by a limited receptive field of the convolution operator

Table 1 Pilot study results on the KITTI dataset. Overall means the measurements are made from 0 m to 80 m. HF, Conv, LAtt., and GAtt. represent hierarchical features, convolution operations, local attention, and global attention, respectively.

Backbone	Conv	GAtt.	HF	LAtt.	Range	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_2 \uparrow$	REL↓	RMS↓
ResNet-50[1]	√	×	√	×	0–20 m	0.973	0.998	1	0.054	0.985
					60–80 m	0.600	0.900	0.972	0.188	14.30
					Overall	0.952	0.994	0.999	0.065	2.596
Flatten ViT[12]	×	√	×	×	0–20 m	0.955	0.995	0.999	0.071	1.275
					60–80 m	0.727	0.936	0.985	0.150	11.86
					Overall	0.938	0.992	0.999	0.080	2.695
PvT[14]	×	√	√	×	0–20 m	0.960	0.995	0.999	0.062	1.161
					60–80 m	0.678	0.927	0.983	0.161	12.64
					Overall	0.945	0.992	0.999	0.071	2.552
Swin Transformer[13]	×	√	√	√	0–20 m	0.972	0.998	1	0.050	0.948
					60–80 m	0.729	0.941	0.984	0.150	11.85
					Overall	0.961	0.993	0.999	0.062	2.40

We highlight two major concerns: 1) The Transformer lacks spatial inductive bias in modeling the local information[13]. As for depth estimation, the local information is reflected in the detailed context that is crucial for consistent and sharp estimation results. However, these detailed content tends to be lost during the patch-wise interaction of the Transformer. Since objects appearing nearer are larger with higher texture quality[6], the Transformer will lose more details at these locations, which severely deteriorates the model performance at a near range and leads to unsatisfying results. 2) Visual elements vary substantially in scale[32]. In general, a U-Net[7] shape architecture is applied for depth estimation, where the multi-scale skip connections are pivotal for exploiting multi-level information. Since the tokens in ViT are all of a fixed scale, the consecutive non-hierarchical forward propagation makes the multi-scale property ambiguous, which may also limit the performance.

Based on these analyses, we conduct more pilot studies to equip the flatten ViT with hierarchical features[36] and local attention[32].

As shown in the results, adopting the multi-scale architecture[36] can bring certain improvements. However, since ViT only utilizes global attention and the perspective field is invariant among the hierarchical features, the performance limitation exists. Furthermore, we install the local attention[32] into the ViT backbone with hierarchical architecture. Interestingly, the performance significantly boosts where the model achieves similar close-range depth estimation results compared to the one with ResNet, and simultaneously, the long-range depth estimation keeps on par with the model with vanilla ViT. Since the local attention essentially resembles the convolution operation, all these experiments enlighten us to combine the Transformer and convolution for monocular depth estimation.

In this paper, we propose to leverage an encoder consisting of parallel Transformer and convolution branches to exploit both the long-range correlation and local information. To highlight discrepancies with previous depth estimation work adopting Transformer, we present detailed comparisons in Fig. 4. We highlight the superiority of our parallel design as follows. 1) The parallel structure can independently extract the most useful information for both long-range and close-range depth estimation. Based on this, we design an effective fusion module to aggregate the best of Transformer and convolution branches. It makes sense and is well supported by our pilot studies. 2) Our design is pluggable and extendable. Based on a double-branch encoder, we can simply switch and scale up the Transformer encoder with different variants, such as ViT and Swin Transformer. With the fast development of Transformer architecture design, the performance of our method can also be improved without bells and whistles. Moreover, following a macro design, there is no inner modification for the Transformer encoder, which means we can easily adapt any pre-trained parameters for the Transformer encoder. 3) We provide a new perspective to enhance the model for depth estimation. Instead of stacking tremendous modules in a single-stream architecture, one can construct parallel branches and then aggregate the information via fusion strategies. This strategy can effectively combine the best of each branch and brings significant improvement to model performance.

3.2 Transformer and CNN feature extraction

We propose to extract image features via an encoder consisting of a Transformer branch and a light-weight convolution branch, thus fully exploiting the long-range correlation and the local information.

Transformer branch first splits the input image I

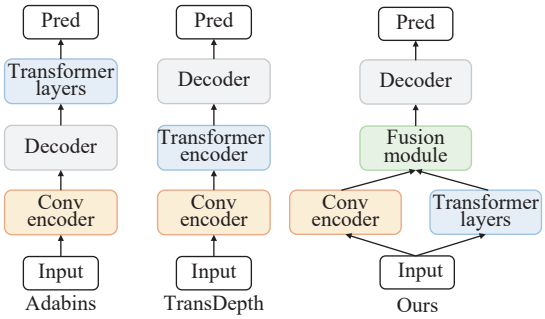


Fig. 4 Comparisons of model architecture with previous methods adopting Transformer layers. Our DepthFormer utilizes a parallel structure, better preserving the local and global information.

into non-overlapping patches by a patch partition module. The initial feature representation of each patch is set as a concatenation of the pixel RGB values. After that, a linear embedding layer is applied to project the initial feature representation to an arbitrary dimension, which is served as the input of the first Transformer layer and denoted as z^0. After that, L Transformer layers are applied to extract features. In general, each layer consists of a multi-head self-attention (MSA) module, followed by a multi-layer perceptron (MLP). A LayerNorm (LN) is applied before the MSA and the MLP, and a residual connection is utilized for each module. Therefore, the process of layer l is formulated as

$$\begin{aligned} \hat{z}^l &= \mathrm{MSA}\left(\mathrm{LN}\left(z^{l-1}\right)\right) + z^{l-1} \\ z^l &= \mathrm{MLP}\left(\mathrm{LN}\left(\hat{z}^l\right)\right) + \hat{z}^l \end{aligned} \tag{1}$$

where $\hat{z}^l$ and z^l denote the output features of the MSA module and the MLP module for layer l, respectively. The structure of a Transformer layer is illustrated in the left part of Fig. 2. Following DPT[5], we sample and reassemble N feature maps from the N selected Transformer layers as the output of the Transformer branch and symbolize them as $F = \{f^n\}_{n=1}^N$, where $f^n \in \mathbf{R}^{C_n \times H_n \times W_n}$ indicates the n-th reassembled feature map.

Notably, our framework is compatible with a variety of Transformer structures. In this paper, we prefer to utilize Swin Transformer[32] to provide hierarchical representations and reduce the computational complexity. The main differences from the standard Transformer layers lie in the local attention mechanism, the shifted window scheme, and the patch merging strategy.

Convolution branch contains a standard ResNet encoder to extract the local information, which is commonly used in depth estimation methods. Only the first block of the ResNet is used here to exploit the local information, which avoids the low-level features being washed out by consecutive multiplications[13] and greatly reduces the computational time. The output feature map with C_g channels is denoted as $G \in \mathbf{R}^{C_g \times H_g \times W_g}$.

Upon acquiring Transformer features F and convolution features G, we feed them to the HAHI module for further processing. Compared to TransDepth[13], we adopt an additional convolution branch to preserve the local information. It avoids the discarding of crucial information by CNN and enables us to predict sharper depth maps without artifacts, as shown in Fig. 5.

3.3 HAHI module

To alleviate the limitation of insufficient aggregation, we introduce the HAHI module to enhance the Transformer features and further model the affinity of the Transformer and the CNN features in a set-to-set translation manner. It is motivated by Deform-DETR[18] and attempt to apply attention modules to solve the fusion of heterogeneous features.

We consider a set of hierarchical features $F = \{f^n\}_{n=1}^N$ as the inputs for feature enhancement. Since we use the Swin Transformer layers to extract the features, the reassembled feature maps will exhibit different sizes and channels, as shown in Fig. 6. Many previous works have to downsample the multi-level features to the resolution of the bottleneck feature and can only enhance the bottleneck feature with simple concatenation, or latent kernel schemes[3, 13, 34]. Oppositely, we aim to enhance all the features without downsampling operators that may lead to information loss.

Specifically, we first utilize 1×1 convolutions to

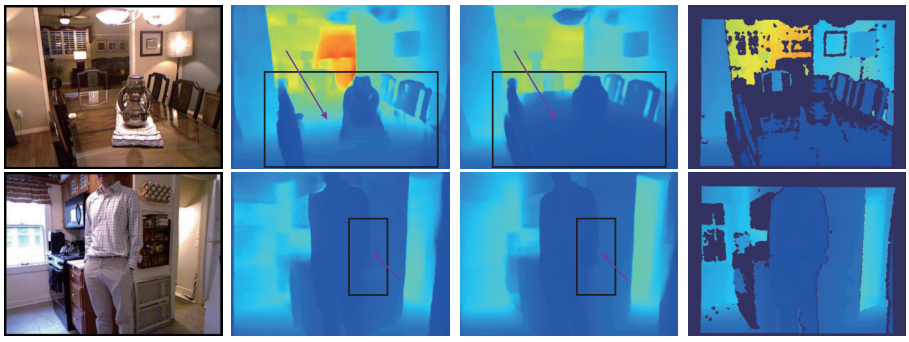


Fig. 5 Demonstration of artifacts and the lost of local information. Our method provides consistent and sharp depth estimation.

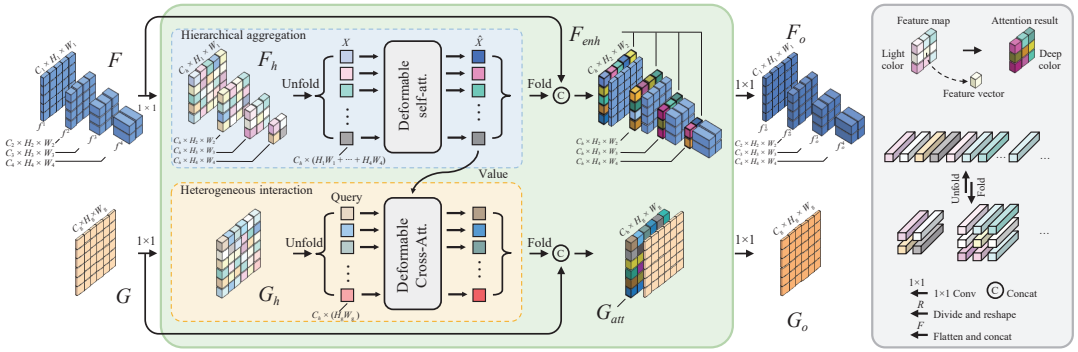


Fig. 6 Illustration of our proposed HAHI. The deformable self-attention module enhances the input Transformer features F. The deformable cross-attention module models the affinity between the Transformer features F and the convolution features G in a set-to-set translation manner. The output of the HAHI, F_o and G_o, are sent to the decoder for the final aggregation. Due to the adoption of Swin Transformer layers to extract the hierarchical features, F exhibits different sizes and channels.

project all the hierarchical features to the same channel C_h, denoting as $F_h = \{f_h^n\}_{n=1}^N$. Then, we unfold (i.e., flatten and concatenate) the feature maps to a two-dimensional matrix X, where each row is a C_h-dimensional feature vector of one pixel from the hierarchical features. After that, we compute the Q (query), K (key) and V (value) by linear projections of X as

$$Q = XP_Q, \quad K = XP_K, \quad V = XP_V \tag{2}$$

where P_Q, P_K and P_V are linear projections, respectively. We attempt to apply the self-attention module to enhance the features. However, extremely numerous feature vectors lead to an unbearable memory cost. To alleviate this issue, we propose to adopt a deformable version that only attends to a limited set of key sampling vectors in a learnable manner. It is reasonable for the depth estimation task since several key points that indicate the scene structure are enough for feature enhancement. Let q and v index a element with representation feature $x_q \in Q$ and $x_v \in V$, respectively. p_q represents the location of the query vector x_q. The processing can be formulated as

$$\mathrm{DAttn}(x_q, x_v, p_q) = \sum_{k \in \Omega_k} A_{qk} x_v \left(p_q + \Delta p_{qk}\right) \tag{3}$$

where the attention weight A_{qk} and the sampling offset Δp_{qk} of the k-th sampling point are obtained via linear projection over the query feature x_q. A_{qk} are normalized as $\sum_{k \in \Omega_k} A_{qk} = 1$. As $p_q + \Delta p_{qk}$ is fractional, bilinear interpolation is applied as in [17] in computing $x_v(p_q + \Delta p_{qk})$. We also add a hierarchical embedding to identify which feature level each query pixel lies in. The output denoted as $\hat{X}$ is folded (i.e., split and reshaped) back to the original resolutions to get the hierarchical enhanced features F_{enh}. After fusing F_{enh} and F via channel-wise concatenatetions followed by 1×1 convolutions, we obtain the output $F_o = \{f_o^n\}_{n=1}^N$ and achieve the feature enhancement.

When the additional convolution branch is available, we consider a feature map G as the second input of the HAHI for affinity modeling. Similar to the first input F, G can be any other type of representation. We utilize a 1×1 convolution to project G to G_h with a channel dimension C_h and then flatten G_h to a two-dimensional query matrix Q. Applying $\hat{X}$ as K and V, we calculate the cross-attention to model the affinity. Similarly, the unbearable memory cost still persists. We apply the deformable attention module in (3) to alleviate this issue, where the reference point locations p_q are dynamically predicted from the affinity query embedding via a learnable linear projection followed by a sigmoid function. After reshaping the result to the original resolution to form the attentive representation G_{att}, we fuse G_{att} and G by a channel-wise concatenation and a 1×1 convolution, getting another output of HAHI, denoted as G_o. This process achieves the affinity modeling and the feature interaction between the Transformer and the CNN branches.

It intuitively makes sense that we aim to combine both the local-aware convolutional and global-aware Transformer features. Given the heterogeneous features, our proposed HAHI module achieves the feature alignment in a dynamic learnable manner by considering the correlation between a convolutional feature and its corresponding Transformer features. Specifically, following the DETR-like structure[18, 33], the former self-attention aggregates and enhances the Transformer features, while the later cross-attention can efficiently model the affinities between heterogeneous features and aligns the heterogeneous features for simple concatenation-manner fusion in the decoder. The convolutional features serve as dense queries instead of the sparse object-level queries in [18, 33], which fully consider spatial information for the depth estimation task. In implementation, we only adopt single-layer cross-attention to save resources, which is much more efficient compared with the DETR-like decoders that adopt more than six attention layers in object detection tasks.

3.4 Decoder design

All the outputs of the HAHI (i.e., F_o and G_o) are sent

to the baseline decoder[31, 37] for depth estimation, which consists of several consecutive UpConv layers with skip connections from the encoder as illustrated in Fig. 7. In the decoder, the resolution of feature maps are progressively recovered via successive UpConv layers, and finally, we adopt a 3×3 convolution to regress the final depth prediction.

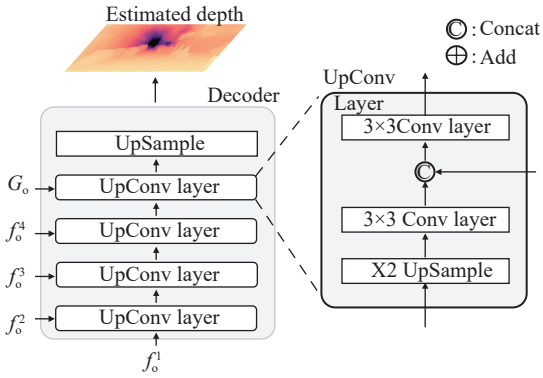


Fig. 7 Standard decoder adopted following previous work[4, 31] for fair comparisons. There are totally four decoder layers. Each 3×3 Conv layer consists of one convolution operation and one ReLU function.

Given the predicted depth estimation maps, the network optimization loss updated from[26] is

$$\mathcal{L}_{pixel} = \alpha \sqrt{\frac{1}{T}\sum_i h_i^2 - \frac{\lambda}{T^2}\left(\sum_i h_i\right)^2} \tag{4}$$

where $h_i = \log \tilde{d}_i - \log d_i$ with the ground truth depth d_i and predicted depth $\tilde{d}_i$. T denotes the number of pixels having valid ground truth values. Following [4], we use $\lambda = 0.85$ and $\alpha = 10$ for all experiments.

4 Experiment results

4.1 Datasets

KITTI is a dataset that provides stereo images and corresponding 3D laser scans of outdoor scenes captured by equipment mounted on a moving vehicle[19]. The RGB images have a resolution of around 1 241 × 376, while the corresponding ground truth depth maps are of low density. Following the standard Eigen training/testing split[26], we use around 26 K images from the left view for training and 697 frames for testing. When evaluation, we use the crop as defined by Garg et al.[38] and upsample the prediction to the ground truth resolution. For the online KITTI depth prediction, we use the official benchmark split[39], which contains around 72 K training data, 6 K selected validation data and 500 test data without the ground truth.

NYU-Depth-v2 provides images and depth maps for different indoor scenes captured at a pixel resolution of 640 × 480[20]. Following previous works, we train our network on a 50 K RGB-Depth pairs subset. The predicted depth maps of DepthFormer have a resolution of 320 × 240 and an upper bound of 10 meters. We upsample them by 2× to match the ground truth resolution during both training and testing. We evaluate the results on the pre-defined center cropping by Eigen et al.[26]

SUN RGB-D is an indoor dataset consisting of around 10 K images with high scene diversity collected with four different sensors[21, 40, 41]. We apply this dataset for generalization evaluation. Specifically, we cross-evaluate our NYU pre-trained models on the official test set of 5 050 images without further fine-tuning. The depth upper bound is set to 10 meters. Note that this dataset is only for evaluation. We do not train on this dataset.

4.2 Evaluation metrics

In our experiments, we follow the standard evaluation protocol of the prior work[26] to confirm the effectiveness of DepthFormer in experiments. For the NYU, KITTI Eigen split and SUN RGB-D dataset, we utilize the accuracy under the threshold ($\delta_i < 1.25^i, i = 1, 2, 3$), mean absolute relative error (AbsRel), mean squared relative error (SqRel), root mean squared error (RMSE), root mean squared log error (RMSElog), and mean log10 error (log10) to evaluate our methods. In terms of the online KITTI benchmark[39], we use the scale-invariant logarithmic error (SILog), percentage of AbsRel and SqRel (absErrorRel, sqErrorRel), and root mean squared error of the inverse depth (iRMSE).

4.3 Implementation details

Since we find there is no commonly used codebase for the monocular depth estimation task, we develop a unified benchmark based on the MMSegmentation[42]. We believe it can further boost the development of this field and achieve fair comparisons. We train the entire network with the batch size 2, learning rate 1×10^{-4} for 38.4 K iterations on a single node with 8 NVIDIA V100 32 GB GPUs, which takes around 5 hours. The linear learning rate warm-up strategy is applied for the first 30% iterations following [4]. The cosine annealing learning rate strategy is adopted for the learning rate decay. Following [5, 32], we sample $N = 4$ results from the transformer features as the output of the transformer branch. The number of reference points in deformable attention modules and C_h is experientially set to 8 and the median value of the channel dimension of F, respectively. Following [18], we adopt 8 deformable attention heads. The default patch size of ViT-Base and window size of Swin Transformer are 16 and 7, respectively. Following previous works, our encoders are pre-trained on ImageNet dataset[43] and then fine-tuned on depth datasets. The de-

coder is trained from scratch.

As for the pilot study, the baseline model consists of an encoder and a decoder. We adopt the decoder in [31] as a default setting and mainly focus on the influence of encoder choices. In terms of the convolution encoder, we utilize the standard ResNet-50[1]. For the Transformer encoder, we adopt the ViT-B[12] following the design of the DPT[5], PvT-T[36], and Swin-T[32]. During training, we adopt the AdamW optimizer. The weight decay is set to 0.01. We experientially use the 1-cycle policy with the learning rate $lr = 6 \times 10^{-5}$ for the Transformer-based model and $lr = 1 \times 10^{-4}$ for the ResNet-based model, respectively. We also apply a linear warm-up scheduler for the first 500 iterations. The cosine annealing learning rate strategy is adopted for the learning rate decay. When evaluation, we divide the depth range to 0–20 m, 20–60 m and 60–80 m. The results of 0–20 m and 60–80 m can indicate the model performance predicting the depth of near and distant objects, respectively.

4.4 Comparison to state-of-the-arts

We compare the proposed methods with the leading monocular depth estimation models. Primarily, we choose the Adabins[4] as our main competitor, which is a solid counterpart and achieved state-of-the-art on all of the datasets we consider. We reproduce the codes of Adabins and load the pre-trained models provided by the authors to get the resulting depth images. Other results are from their official codes.

NYU-Depth-v2. Table 2 lists the performance comparison results on the NYU-Depth-v2 dataset. While the performance of the state-of-the-art models tends to approach saturation, DepthFormer outperforms all the competitors with prominent margins in all metrics. It indicates the effectiveness of our proposed methods. Qualitative comparisons can be seen in Fig. 8. DepthFormer achieves more accurate and sharper depth estimation results. We combine camera parameters and predicted depth maps to inv-project the 2D images into the 3D world. As shown in Fig. 9, our reconstructed scenes are satisfying with sharp boundaries of objects and reasonable depth estimations.

Since previous work applies different encoders, we conduct more fair performance comparisons with them by aligning the encoder architecture to the Swin Transformer and then providing the quantitative metrics, runtime, and the number of parameters of each method. As shown in Table 3, DepthFormer outperforms previous methods with significant margins meanwhile introducing acceptable overhead on frames per second (FPS) and parameters, which indicates the effectiveness of our method.

KITTI. We evaluate on the Eigen split[26] and report the results on Table 4. DepthFormer significantly outperforms all the leading methods. Qualitative comparisons can be seen in Fig. 10. We then train our model on the training set of the standard KITTI benchmark split and submit the prediction results of the testing set to the online website. We report the results in Table 5. While a saturation phenomenon persists in sqErrorRel, DepthFormer still achieves 16% improvement on this metric and achieves the most competitive result on the highly competitive benchmark as the submission time of Nov. 16th, 2021. We report some qualitative comparison results in Fig. 11.

SUN RGB-D. Following Adabins[4], we conduct a cross-dataset evaluation by training our models on the NYU-Depth-v2 dataset and evaluating them on the test set of the SUN RGB-D dataset without any fine-tuning. As shown in Table 6, significant improvements in all the metrics indicate an outstanding generalization perform-

Table 2 Comparison of performances on the NYU-Depth-v2 dataset. The reported numbers are from the corresponding original papers. Sup and Unsup represent supervised training and unsupervised training, respectively. Best / Second best results are marked bold / underlined.

Method	Train	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL ↓	RMS ↓	$log_{10} \downarrow$
StructDepth[44]	Unsup	0.817	0.955	0.988	0.140	0.534	0.060
MonoIndoor[45]	Unsup	0.823	0.958	0.989	0.134	0.526	–
Eigen et al.[26]	Sup	0.769	0.950	0.988	0.158	0.641	–
Laina et al.[46]	Sup	0.811	0.953	0.988	0.127	0.573	0.055
DORN[2]	Sup	0.828	0.965	0.992	0.115	0.509	0.051
BTS[3]	Sup	0.885	0.978	0.994	0.110	0.392	0.047
DAV[11]	Sup	0.882	0.980	0.996	0.108	0.412	–
TransDepth[15]	Sup	0.900	0.983	0.996	0.106	0.365	0.045
DPT[5]	Sup	<u>0.904</u>	0.988	0.998	0.110	0.357	0.045
AdaBins[4]	Sup	0.903	0.984	<u>0.997</u>	0.103	0.364	<u>0.044</u>
NeWCRFs[47]	Sup	**0.923**	**0.992**	**0.998**	<u>0.095</u>	<u>0.331</u>	<u>0.044</u>
DepthFormer	Sup	**0.923**	<u>0.989</u>	<u>0.997</u>	**0.094**	**0.329**	**0.040**

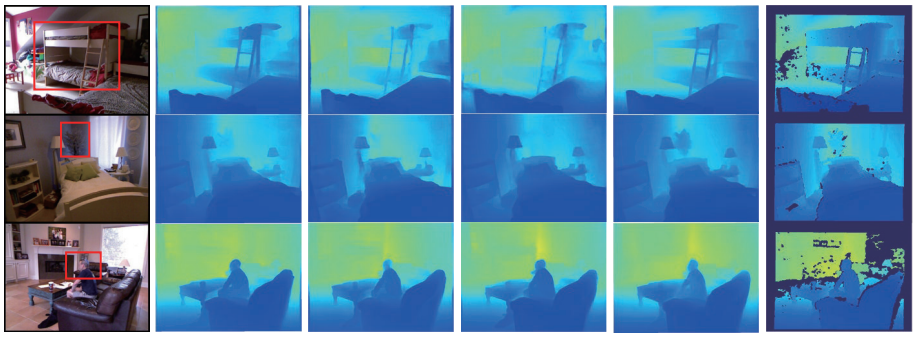


Fig. 8 Qualitative comparison on the NYU-Depth-v2 dataset

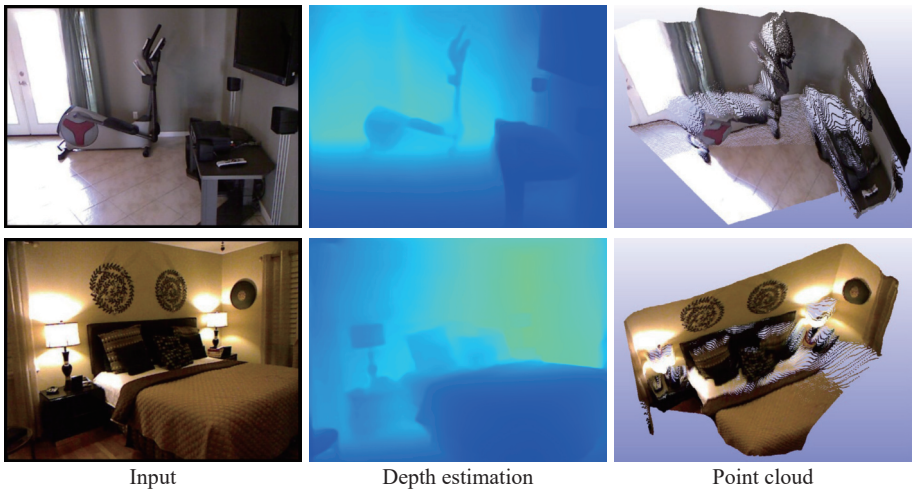


Fig. 9 Visualization of reconstructed 3D scenes. We first estimate the depth maps based on the input RGB images, and then inv-project image pixels into the real 3D world with estimated depth maps and camera intrinsic parameters.

Table 3 Comparisons with other depth estimation counterparts

Method	Encoder	$\delta_1 \uparrow$	REL ↓	FPS ↑	#param ↓
Adabins[4]	ResNet-50	0.850	0.136	**13.79**	83.5 M
Adabins[4]	EfficientNetB5	0.903	0.103	11.48	**77 M**
Adabins[4]	Swin-L-22K	0.917	0.102	7.63	210 M
DPT[5]	Swin-L-22K	0.914	0.099	7.51	225 M
DepthFormer (Ours)	Swin-L-22K	**0.923**	**0.094**	5.48	273 M

ance of DepthFormer. Qualitative results are shown in Fig. 12. It is engaging that DepthFormer presents a strong generalization performing. Especially, our method can predict accurate depth estimation for extremely dark areas which are extremely hard to handle without training on the corresponding dataset.

4.5 Ablation studies

For our ablation study, we conduct evaluations with each component of DepthFormer to prove the effectiveness of our method on the NYU and KITTI dataset.

Effectiveness of key components. We first validate the effectiveness of the key components of DepthFormer. From the baseline network (i.e., ResNet-50, Swin-T), we reinforce the network with our proposed methods and evaluate the improvement of the model performance. The results are reported in the Table 7. As the additional convolution branch and the HAHI are adopted, the overall performance is significantly improved, which demonstrates the effectiveness of our methods. Moreover, following previous methods[5, 13], we utilize larger-scale dataset (i.e., ImageNet-22K) to pre-train our encoder. The results (+LP) indicate that the Transformer encoder can better benefit from the larger model capacity and the larger-scale pre-training dataset compared with the CNN encoder.

Since our design is pluggable and extendable, we can

Table 4 Comparison of performances on the KITTI validation dataset. The reported numbers are from the corresponding papers. Measurements are made for the depth range from 0 m to 80 m.

Method	Train	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL ↓	Sq ↓	RMS ↓	log ↓
Godard et al.[48]	Unsup	0.861	0.949	0.976	0.114	0.898	4.935	0.206
Johnston and Garneiro[49]	Unsup	0.889	0.962	0.982	0.106	0.861	4.699	0.185
Gan et al.[50]	Sup	0.890	0.964	0.985	0.098	0.666	3.933	0.173
DORN[2]	Sup	0.932	0.984	0.994	0.072	0.307	2.727	0.120
Yin et al.[51]	Sup	0.938	0.990	0.998	0.072	–	3.258	0.117
PGA-Net[52]	Sup	0.952	0.992	0.998	0.063	0.267	2.634	0.101
BTS[3]	Sup	0.956	0.993	0.998	0.059	0.245	2.756	0.096
TransDepth[15]	Sup	0.956	0.994	**0.999**	0.064	0.252	2.755	0.098
DPT-Hybrid[5]	Sup	0.959	0.995	**0.999**	0.062	–	2.573	0.092
AdaBins[4]	Sup	0.964	0.995	**0.999**	0.058	0.190	2.360	0.088
NeWCRFs[47]	Sup	0.974	**0.997**	**0.999**	**0.052**	**0.155**	**2.129**	**0.079**
DepthFormer	Sup	**0.975**	**0.997**	**0.999**	**0.052**	0.158	2.143	**0.079**

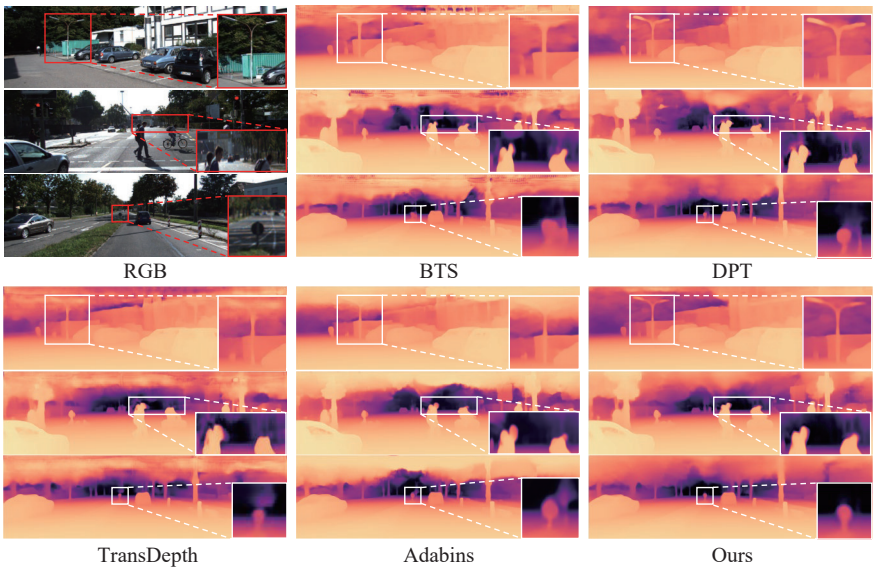


Fig. 10 Qualitative comparison on the KITTI validation dataset

Table 5 Comparison of performances on the KITTI depth estimation benchmark test set

Method	SILog ↓	sqErrorRel ↓	absErrorRel ↓	iRMSE ↓
DORN[2]	11.77	2.23	8.78	12.98
BTS[3]	11.67	2.21	9.04	12.23
BANet[53]	11.55	2.31	9.34	12.17
PWA[54]	11.45	2.30	9.05	12.32
ViP-DeepLab[55]	10.80	2.19	8.94	11.77
NeWCRFs[47]	**10.39**	1.83	**8.37**	**11.03**
Ours	10.46	**1.82**	8.54	11.17

simply switch and scale up the Transformer encoder with different variants. In this ablation, we conduct experiments with different Transformer backbones on NYU dataset. As shown in Table 8, our DepthFormer can bring improvements to models with various Transformer backbones. It is possible that the performance of our method can also be improved without bells and whistles with the fast development of Transformer architecture design.

Fine-grained evaluation on convolution branch. The standard CNN encoder can be divided into several sequential blocks. We further scrutinize the influence of different level convolution features on the model performance. Following the default setting, we adopt ResNet-50 as the additional convolution branch. Results are shown in Fig. 13. Interestingly, the model achieves the best performance with only one convolutional block and then downgrades if more blocks are added. A possible explanation for this might be that the consecutive convolutions

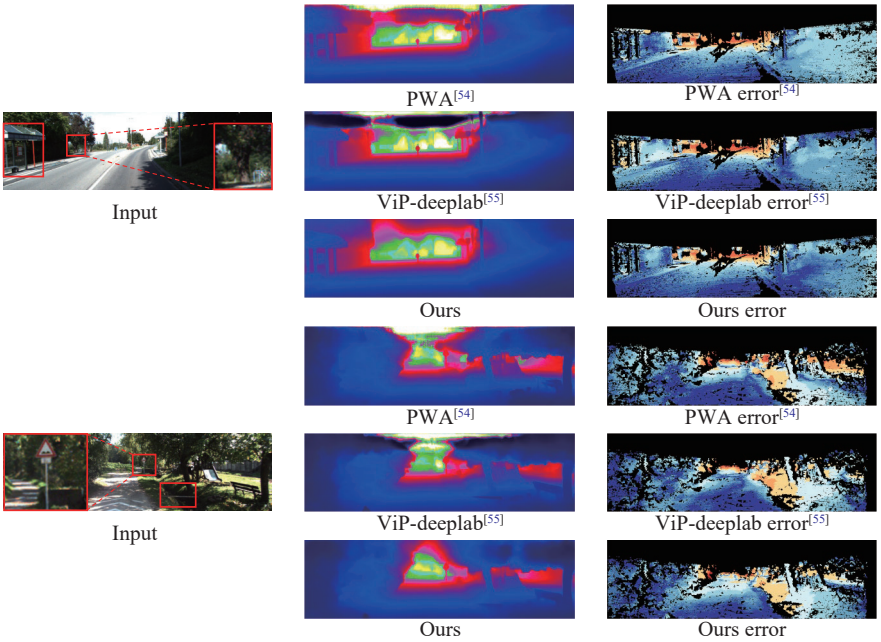


Fig. 11 Qualitative comparison with the state-of-the-art on the KITTI benchmark, better viewed by zooming on screen. Deeper red pixels in the error maps indicate higher errors. Deeper blue means lower errors. The figures are from the official KITTI benchmark website.

Table 6 Results of models trained on the NYU-Depth-v2 dataset and tested on the SUN RGB-D dataset[23] without fine-tuning. The reported numbers are from [4].

Method	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL ↓	RMS ↓	$log_{10} \downarrow$
Chen et al.[56]	0.757	0.943	0.984	0.166	0.494	0.071
Yin et al.[51]	0.696	0.912	0.973	0.183	0.541	0.082
BTS[3]	0.740	0.933	0.980	0.172	0.515	0.075
Adabins[4]	0.771	0.944	0.983	0.159	0.476	0.068
NeWCRFs[47]	0.799	0.959	0.986	0.150	0.429	0.064
Ours	**0.815**	**0.970**	**0.993**	**0.137**	**0.408**	**0.059**

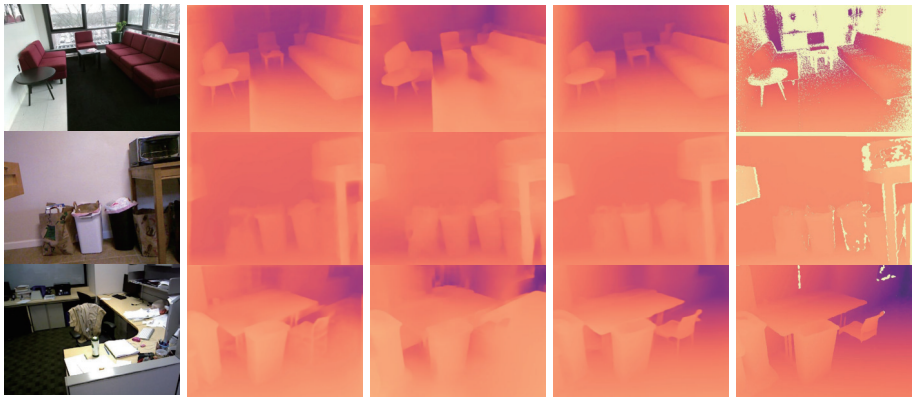


Fig. 12 Qualitative comparison on the SUN RGB-D dataset

wash out low-level features, and the gradually reducing spatial resolution discards the fine-grained information[13]. Adopting the first block achieves a win-win scenario: It optimizes accuracy by preserving crucial local information while reducing complexity. This can reduce the training time by 2.5× or more and likewise decrease memory consumption, enabling us to easily scale our Transformer branch to large models.

We also conduct experiments to investigate different convolution variants in Table 9. We first evaluate a

Table 7 Ablation study results on the NYU dataset. CB: Convolution branch. LP: Larger-scale pre-training dataset (22 K ImageNet) for boosting the model performance. For fair comparison, we utilize the 22 K-ImageNet pre-trained ResNet-50-x3 provided by [57] to get the results of R-50 (+HAHI+LP).

Backbone	Various	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL ↓	RMS ↓
R-50[1]	–	0.811	0.961	0.991	0.145	0.482
	+HAHI	0.866	0.976	0.994	0.122	0.411
	+HAHI+LP	0.865	0.976	0.995	0.124	0.406
Swin-T[13]	–	0.847	0.975	0.993	0.131	0.432
	+HAHI	0.866	0.977	0.995	0.125	0.409
	+CB	0.851	0.977	0.995	0.129	0.425
	+CB+HAHI	0.871	0.979	0.995	0.120	0.399
Swin-B[13]	+CB+HAHI	0.892	0.984	**0.997**	0.109	0.373
	+CB+HAHI+LP	0.910	0.987	**0.997**	0.101	0.348
Swin-L[13]	+CB+HAHI+LP	**0.923**	**0.989**	**0.997**	**0.094**	**0.329**

Table 8 Ablation study of the extendable property. We adopt various Transformer to validate the effectiveness of DepthFormer.

Transformer backbone	DepthFormer	$\delta_1 \uparrow$	REL ↓	RMS ↓
Flatten Transformer, ViT-B		0.846	0.141	0.437
	√	**0.882**	**0.121**	**0.390**
Pyramid Transformer, PVT-T		0.822	0.145	0.467
	√	**0.835**	**0.141**	**0.451**
Swin Transformer, Swin-T		0.847	0.131	0.432
	√	**0.871**	**0.120**	**0.399**

single-layer convolution with different kernel size. There are only slight performance discrepancies among them, indicating simple and various convolution operations can be effective enough for our purpose of providing local information. Moreover, we also compare the default ResNet architecture with other convolution backbones such as EfficientNet[29] and DenseNet[30]. The differences are also slight with a common performance drop when applying more CNN blocks. Recall our experiments have shown that a deeper convolution network cannot bring performance gains but leads to degradation. Since the results of models with EfficientNet and DenseNet backbones present a similar trend of a performance drop, it is unnecessary to further stack more layers in this ablation study. All these experimental results indicate that a simple single-layer convolution can work effectively in our proposed parallel architecture.

Fine-grained evaluation on HAHI. Since the HAHI consists of a deformable self-attention module (DSA) for hierarchical aggregation and a deformable cross-attention module (DCA) for heterogeneous interaction, we conduct more detailed ablation studies on both of these two modules. For fair comparison, we choose the Swin-T with CB as the default backbone. The results are reported in Table 10. We propose to apply the attention mechanism on all the hierarchical features (multi-level DSA) for sufficient aggregation. Compared with the one where only each single-layer feature is considered in the attention module, denotes as single-level DSA, the multi-level aggregation strategy obtains a 4.9% enhancement on RMS. It demonstrates that the multi-level aggregation strategy is much more effective. When DSA is added without the multi-level DSA, the model performance is seriously impaired. However, with the multi-level DSA, DCA achieves a 2.2% improvement on RMS, verifying

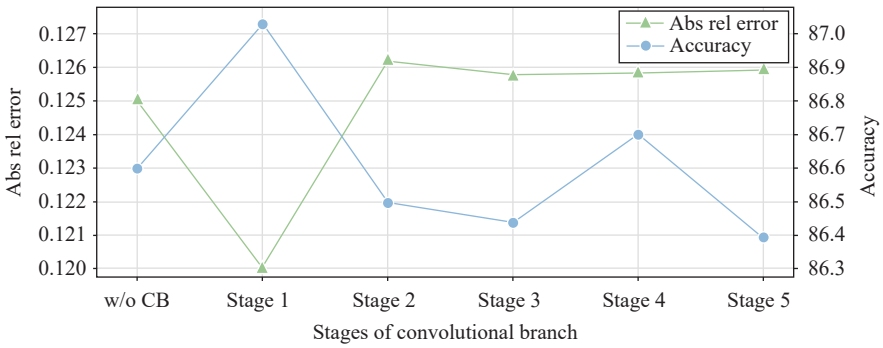


Fig. 13 Effect of the convolution branch block on NYU depth estimation performance. We can observe that behaviour decreases after the first block of R-50[1]. Notably, Stage 1 only consists of one 7×7 convolution, ReLU and BN.

Table 9 Comparisons of CNN variants

CNN variants	$\delta_1 \uparrow$	REL ↓	RMS ↓
Conv3×3+BN+ReLU	**0.871**	**0.120**	0.403
Conv5×5+BN+ReLU	**0.871**	**0.120**	0.401
Conv9×9+BN+ReLU	0.870	0.121	0.405
ResNet50-C1, 2[1]	0.865	0.126	0.410
EfficientNetB5-C1, C2[29]	0.866	0.125	0.408
DenseNet169-C1, C2[30]	0.866	0.124	0.409
Conv7×7+BN+ReLU (ResNet-C1)	**0.871**	**0.120**	**0.399**

Table 10 Ablation study of the HAHI module on NYU dataset. DSA, DCA: Deformable self-attention and cross-attention.

Aggregation (DSA)		Interaction (DCA)	$\delta_1 \uparrow$	REL ↓	RMS ↓
single-level	multi-level				
			0.847	0.131	0.432
√			0.850	0.131	0.427
	√		0.867	0.124	0.408
		√	0.828	0.143	0.443
√		√	0.831	0.144	0.449
	√	√	**0.871**	**0.120**	**0.399**

the importance of both the multi-level DSA and the DCA for heterogeneous interaction. We infer the reason that there are large discrepancies between the heterogeneous features. Multi-level DSA achieves the alignment of the features, which propels the affinity modeling. All the results demonstrate the effectiveness of our proposed HAHI module.

Since HAHI module is essentially a late fusion strategy, we further compare it with other effective late fusion strategies. As shown in Table 11, we not only compare our HAHI with widely-adopted fusion strategies proposed for classification tasks[58–60], but also compare recent multimodal and depth fusion strategies[13, 61, 62]. We highlight the best performance achieved by our HAHI with hierarchical aggregation and heterogeneous interaction. Compared with these counterparts, HAHI can dynamically align the heterogeneous features from the Transformer branch and convolution branch, yielding a more effective fusion strategy for depth estimation facing various scenes. Moreover, compared with other potential effective strategies such as AGD in [13], HAHI also saves tons of GPU memory. In the ablation study, we use a batch size of 8 as default for DepthFormer equipped with HAHI, but the AGD is out of memory even when we only use a batch size of 1. Here, we also adopt a two-direction version for the HAHI module, where an extra self-attention is applied for the convolution feature and an extra cross-attention is run in a reverse direction (regard the Transformer features as queries). There is a slight performance improvement, but it leads to more inference time. As a result, we use the single-direction described in Section 3.3 as default.

Table 11 Comparisons with late fusion strategies

Method	$\delta_1 \uparrow$	REL ↓	RMS ↓	Field
Concate[58]	0.847	0.131	0.432	–
CAM[59]	0.851	0.131	0.426	Classification
SAM[60]	0.852	0.130	0.424	Classification & Detection
CBAM[60]	0.853	0.130	0.423	Classification & Detection
AGD[15]	OOM even with bs=1			Depth estimation
MFA[61]	0.861	0.124	0.412	RGBD saliency detection
DFM[62]	0.860	0.125	0.414	RGBD saliency detection
HAHI (Ours)	**0.871**	**0.120**	**0.399**	Depth estimation

We then conduct experiments to investigate the influence of the number of attention heads in HAHI. As shown in Table 12, with the increase in the number of attention heads, the model performance improves significantly, indicating the representation capacity of HAHI is enhanced. However, the performance trends to saturation when the number of attention heads surpasses 8. Therefore, we choose 8 as a default setting, considering a tradeoff between performance and computational cost.

Table 12 Comparisons of HAHI with different numbers of heads. Convolution is followed by ReLU and BN.

#head of att.	$\delta_1 \uparrow$	REL ↓	RMS ↓
2	0.860	0.128	0.410
4	0.866	0.123	0.406
8	**0.871**	**0.120**	**0.399**
16	0.870	**0.120**	0.403

Details about pilot study results. We have discussed that the CNN branch can provide local information lost in the Transformer branch and the HAHI further promotes the depth estimation via feature enhancement and affinity modeling. They improve the model performance, especially on near object depth estimation. Table 13 demonstrates the effectiveness of our methods. Moreover, we draw more fine-grained results in Fig. 14. We present qualitative comparison results in Fig. 15. One can observe sharper and more accurate results can be achieved with our proposed CNN branch and HAHI module. The results show that our methods achieve a tradeoff between long and short-range estimation, improving the overall performance. It makes sense that adding the local details inevitably smears some global information, leading to such a slight performance drop on long-range depth estimation.

Interestingly, Swin Transformer based model achieves better performance compared with ResNet50 based ones

Table 13 More detailed ablation quantitative results on KITTI dataset

Backbone	Various	Range	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_2 \uparrow$	REL $\downarrow$	RMS $\downarrow$
ResNet-50[1]	–	0–20 m	0.973	0.998	1	0.054	0.985
		60–80 m	0.600	0.900	0.972	0.188	14.30
		Overall	0.952	0.994	0.999	0.065	2.596
ViT-Base[12]	–	0–20 m	0.955	0.995	0.999	0.071	1.275
		60–80 m	0.727	0.936	0.985	0.150	11.86
		Overall	0.938	0.992	0.999	0.080	2.695
ViT-Base[12]	+CB	0–20 m	0.960	0.996	0.999	0.067	1.223
		60–80 m	0.725	0.950	0.984	0.147	11.69
		Overall	0.942	0.994	0.999	0.076	2.644
ViT-Base[12]	+CB+HAHI	0–20 m	0.964	0.995	0.999	0.064	1.172
		60–80 m	0.712	0.946	0.984	0.150	11.90
		Overall	0.948	0.993	0.999	0.073	2.596
Swin-T[13]	–	0–20 m	0.972	0.998	1	0.050	0.948
		60–80 m	0.729	0.941	0.984	0.150	11.85
		Overall	0.961	0.993	0.999	0.062	2.402
Swin-T[13]	+CB	0–20 m	0.979	0.998	1	0.049	0.934
		60–80 m	0.726	0.945	0.984	0.149	11.76
		Overall	0.964	0.995	0.999	0.060	2.310
Swin-T[13]	+CB+HAHI	0–20 m	0.981	0.998	1	0.049	0.911
		60–80 m	0.744	0.939	0.981	0.146	11.61
		Overall	0.966	0.996	0.999	0.059	2.261

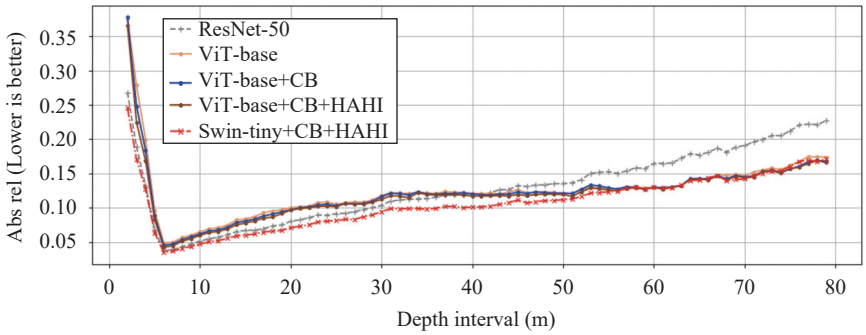


Fig. 14 Fine-grained quantitative results of our pilot study on KITTI dataset. We divide the depth range (0–80 m) into 80 intervals. Point (i, j) in the plot represents the abs rel of the model is j on depth interval $(i, i+1]$m. Our method achieves a tradeoff between long and short range estimation.

on near object depth estimation and also provides satisfactory results compared with ViT-based ones on long-range depth estimation. We infer that the hierarchical design and the local attention of the Swin Transformer benefits the extraction of the local information (See Table 1), and the cross-window attention mechanism can also successfully model the long-range correlation.

5 Conclusions

We have presented DepthFormer, a novel framework for accurate monocular depth estimation. Our method fully exploits the long-range correlations and the local information by an encoder consisting of a Transformer branch and a CNN branch. Since independent branches with late fusion lead to insufficient feature aggregation for the decoder, we propose the hierarchical aggregation and heterogeneous interaction module to enhance the multi-level features and further model the feature affinity. DepthFormer achieves significant improvements compared with state-of-the-arts in the most popular and challenging datasets. We hope our study can encourage more works applying the Transformer architecture in monocular depth estimation and enlighten the framework design

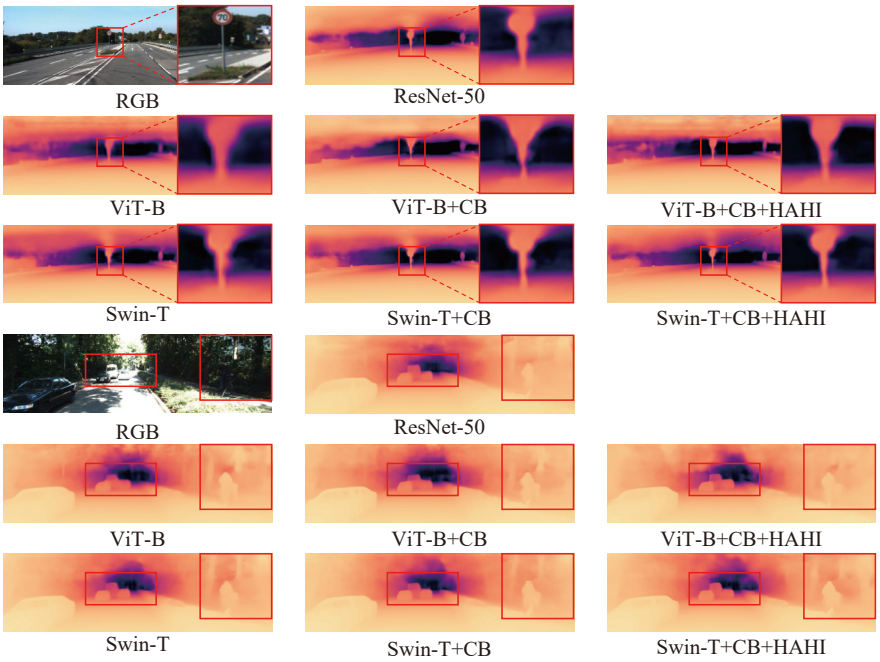


Fig. 15 Qualitative comparisons in our pilot study

of other tasks.

Potential impact: Beyond the direct application of our work for autonomous driving or spatial reconstruction, there are several venues that warrant future investigation. For example, the common dense global attention in Transformer might be sumptuous. In terms of depth estimation, several key points that indicate the scene structure could be enough to provide crucial long-range information. Designing a more dedicated attention mechanism would improve the effectiveness of the Transformer branch. Furthermore, the HAHI is input-agnostic, and including other modalities such as sparse LiDAR would enhance performance and generalization. Finally, due to the lack of theoretical guarantees, future work to improve the applicability of DepthFormer might consider challenges of explainability and transparency.

Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

Open Access

References

[1] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[2] H. Fu, M. M. Gong, C. H. Wang, K. Batmanghelich, D. C. Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 2002–2011, 2018. DOI: 10.1109/CVPR.2018.00214.

[3] J. H. Lee, M. K. Han, D. W. Ko, I. H. Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation, [Online], Available: https://arxiv.org.abs/1907.10326, 2019.

[4] S. F. Bhat, I. Alhashim, P. Wonka. AdaBins: Depth estimation using adaptive bins. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 4008–4017, 2021. DOI: 10.1109/CVPR46437.2021.00400.

[5] R. Ranftl, A. Bochkovskiy, V. Koltun. Vision transformers for dense prediction. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 12159–12168, 2021. DOI: 10.1109/ICCV48922.2021.01196.

[6] A. Saxena, S. H. Chung, A. Y. Ng. Learning depth from single monocular images. In *Proceedings of the 18th Inter-*

national Conference on Neural Information Processing Systems, Vancouver, Canada, pp. 1161–1168, 2005.

[7] O. Ronneberger, P. Fischer, T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Munich, Germany, pp. 234–241, 2015. DOI: 10.1007/978-3-319-24574-4_28.

[8] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018. DOI: 10.1109/TPAMI.2017.2699184.

[9] H. S. Zhao, J. P. Shi, X. J. Qi, X. G. Wang, J. Y. Jia. Pyramid scene parsing network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 6230–6239, 2017. DOI: 10.1109/CVPR.2017.660.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, USA, pp. 6000–6010, 2017.

[11] L. Huynh, P. Nguyen-Ha, J. Matas, E. Rahtu, J. Heikkilä. Guiding monocular depth estimation using depth-attention volume. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 581–597, 2020. DOI: 10.1007/978-3-030-58574-7_35.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[13] G. Yang, H. Tang, M. Ding, N. Sebe, E. Ricci. Transformers solve the limited receptive field for monocular depth prediction. In *Proceedings of International Confonference on Computer Vision*, 2021.

[14] K. Yuan, S. P.Guo, Z. W.Liu, A. J. Zhou, F. W. Yu, W. Wu. Incorporating convolution designs into visual transformers. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 559–568, 2021. DOI: 10.1109/ICCV48922.2021.00062.

[15] Z. H. Dai, H. X. Liu, Q. V. Le, M. X. Tan. Coatnet: Marrying convolution and attention for all data sizes. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pp. 3965–3977, 2021.

[16] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, R. B. Girshick. Early convolutions help transformers see better. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pp. 30392–30400, 2021.

[17] J. F. Dai, H. Z. Qi, Y. W. Xiong, Y. Li, G. D. Zhang, H. Hu, Y. C. Wei. Deformable convolutional networks. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 764–773, 2017. DOI: 10.1109/ICCV.2017.89.

[18] X. Z. Zhu, W. J. Su, L. W. Lu, B. Li, X. G. Wang, J. F. Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *Proceedings of 9th International Conference on Learning Representations*, 2021.

[19] A. Geiger, P. Lenz, C. Stiller, R. Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013. DOI: 10.1177/0278364913491297.

[20] N. Silberman, D. Hoiem, P. Kohli, R. Fergus. Indoor segmentation and support inference from RGBD images. In *Proceedings of the 12th European Conference on Computer Vision*, Springer, Florence, Italy, pp. 746–760, 2012. DOI: 10.1007/978-3-642-33715-4_54.

[21] S. R. Song, S. P. Lichtenberg, J. X. Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 567–576, 2015. DOI: 10.1109/CVPR.2015.7298655.

[22] T. W. Hui, C. C. Loy, X. O. Tang. Depth map super-resolution by deep multi-scale guidance. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 353–369, 2016. DOI: 10.1007/978-3-319-46487-9_22.

[23] J. Lee, Y. Kim, S. Lee, B. Kim, J. Noh. High-quality depth estimation using an exemplar 3D model for stereo conversion. *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 7, pp. 835–847, 2015. DOI: 10.1109/TVCG.2015.2398440.

[24] J. X. Dong, J. S. Pan, J. S. Ren, L. Lin, J. H. Tang, M. H. Yang. Learning spatially variant linear representation models for joint filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8355–8370, 2022. DOI: 10.1109/TPAMI.2021.3102575.

[25] Z. Q. Zhang, X. G. Zhu, Y. W. Li, X. Q. Chen, Y. Guo. Adversarial attacks on monocular depth estimation, [Online], Available: https://arxiv.org/abs/2003.10315, 2020.

[26] D. Eigen, C. Puhrsch, R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 2366–2374, 2014.

[27] J. J. Hu, M. Ozay, Y. Zhang, T. Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, IEEE, Waikoloa, USA, pp. 1043–1051, 2019. DOI: 10.1109/WACV.2019.00116.

[28] X. B. Yang, L. Y. Zhou, H. Q. Jiang, Z. L. Tang, Y. B. Wang, H. J. Bao, G. F. Zhang. Mobile3DRecon: Real-time monocular 3D reconstruction on a mobile phone. *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 12, pp. 3446–3456, 2020. DOI: 10.1109/TVCG.2020.3023634.

[29] M. X. Tan, Q. V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, USA, pp. 6105–6114, 2019.

[30] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 2261–2269, 2017. DOI: 10.1109/CVPR.2017.243.

[31] I. Alhashim, P. Wonka. High quality monocular depth estimation via transfer learning, [Online], Available: https://arxiv.org/abs/1812.11941, 2018.

[32] Z. Liu, Y. T. Lin, Y. Cao, H. Hu, Y. X. Wei, Z. Zhang, S. Lin, B. N. Guo. Swin Transformer: Hierarchical vision

transformer using shifted windows. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 9992–10002, 2021. DOI: 10.1109/ICCV48922.2021.00986.

[33] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 213–229, 2020. DOI: 10.1007/978-3-030-58452-8_13.

[34] S. X. Zheng, J. C. Lu, H. S. Zhao, X. T. Zhu, Z. K. Luo, Y. B. Wang, Y. W. Fu, J. F. Feng, T. Xiang, P. H. S. Torr, L. Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, USA, pp. 6877–6886, 2021. DOI: 10.1109/CVPR46437.2021.00681.

[35] J. B. Jiao, Y. Cao, Y. B. Song, R. Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 55–71, 2018. DOI: 10.1007/978-3-030-01267-0_4.

[36] W. H.Wang, E. Z. Xie, X. Li, D. P. Fan, K. T. Song, D. Liang, T. Lu, P. Luo, L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 548–558, 2021. DOI: 10.1109/ICCV48922.2021.00061.

[37] Z. Y. Li, Z. H. Chen, A. Li, L. J. Fang, Q. H. Jiang, X. M. Liu, J. J. Jiang, B. L. Zhou, H. Zhao. SimIPU: Simple 2D image and 3D point cloud unsupervised pre-training for spatial-aware visual representations. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, Vancouver, Canada, pp. 1500–1508, 2022. DOI: 10.1609/aaai.v36i2.20040.

[38] R. Garg, V. K. B.G., G. Carneiro, I. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 740–756, 2016. DOI: 10.1007/978-3-319-46484-8_45.

[39] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, A. Geiger. Sparsity invariant CNNs. In *Proceedings of International Conference on 3D Vision*, IEEE, Qingdao, China, pp. 11–20, 2017. DOI: 10.1109/3DV.2017.00012.

[40] J. X. Xiao, A. Owens, A. Torralba. SUN3D: A database of big spaces reconstructed using SfM and object labels. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Sydney, Australia, pp. 1625–1632, 2013. DOI: 10.1109/ICCV.2013.458.

[41] A. Janoch, S. Karayev, Y. Q. Jia, J. T. Barron, M. Fritz, K. Saenko, T. Darrell. A category-level 3D object dataset: Putting the kinect to work. *Consumer Depth Cameras for Computer Vision*, A. Fossati, J. Gall, H. Grabner, X. F. Ren, K. Konolige, Eds., London, UK: Springer, pp. 141–165, 2013. DOI: 10.1007/978-1-4471-4640-7_8.

[42] M. Contributors. *MMsegmentation: Openmmlab semantic segmentation toolbox and benchmark*, [Online], Available: https://gitee.com/deadkany/mmsegmentation, 2020.

[43] A. Krizhevsky, I. Sutskever, G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Lake Tahoe, USA, pp. 1097–1105, 2012.

[44] B. Y. Li, Y. Huang, Z. Y. Liu, D. P. Zou, W. X. Yu. StructDepth: Leveraging the structural regularities for self-supervised indoor depth estimation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 12643–12653, 2021. DOI: 10.1109/ICCV48922.2021.01243.

[45] P. Ji, R. Z. Li, B. Bhanu, Y. Xu. Monoindoor: Towards good practice of self-supervised monocular depth estimation for indoor environments. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 12767–12776, 2021. DOI: 10.1109/ICCV48922.2021.01255.

[46] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, N. Navab. Deeper depth prediction with fully convolutional residual networks. In *Proceedings of the 4th International Conference on 3D Vision*, IEEE, Stanford, USA, pp. 239–248, 2016. DOI: 10.1109/3DV.2016.32.

[47] W. H. Yuan, X. D. Gu, Z. Z. Dai, S. Y. Zhu, P. Tan. Neural window fully-connected CRFs for monocular depth estimation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 3906–3915, 2022. DOI: 10.1109/CVPR52688.2022.00389.

[48] C. Godard, O. M. Aodha, G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 6602–6611, 2017. DOI: 10.1109/CVPR.2017.699.

[49] A. Johnston, G. Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 4755–4764, 2020. DOI: 10.1109/CVPR42600.2020.00481.

[50] Y. K. Gan, X. Y. Xu, W. X. Sun, L. Lin. Monocular depth estimation with affinity, vertical pooling, and label enhancement. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 232–247, 2018. DOI: 10.1007/978-3-030-01219-9_14.

[51] W. Yin, Y. F. Liu, C. H. Shen, Y. L. Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, pp. 5683–5692, 2019. DOI: 10.1109/ICCV.2019.00578.

[52] D. Xu, X. Alameda-Pineda, W. L. Ouyang, E. Ricci, X. G. Wang, N. Sebe. Probabilistic graph attention network with conditional kernels for pixel-wise prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2673–2688, 2022. DOI: 10.1109/TPAMI.2020.3043781.

[53] S. Aich, J. M. U. Vianney, M. A. Islam, M. K. B. Liu. Bidirectional attention network for monocular depth estimation. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Xi'an, China, pp. 11746–11752, 2020. DOI: 10.1109/ICRA48506.2021.9560885.

[54] S. Lee, J. Lee, B. Kim, E. Yi, J. Kim. Patch-wise attention network for monocular depth estimation. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pp. 1873–1881, 2021. DOI: 10.1609/aaai.v35i3.16282.

[55] S. Y. Qiao, Y. K. Zhu, H. Adam, A. Yuille, L. C. Chen. ViP-DeepLab: Learning visual perception with depth-aware video panoptic segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern*

Recognition, IEEE, Nashville, USA, pp. 3996–4007, 2021. DOI: 10.1109/CVPR46437.2021.00399.

[56] X. T. Chen, X. J. Chen, Z. J. Zha. Structure-aware residual pyramid network for monocular depth estimation. In *Proceedings of the 28th International Joint conference on Artificial Intelligence*, Macao, China, pp. 694–700, 2019.

[57] A. Kolesnikov, L. Beyer, X. H. Zhai, J. Puigcerver, J. Yung, S. Gelly, N. Houlsby. Big transfer (BiT): General visual representation learning. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 491–507, 2020. DOI: 10.1007/978-3-030-58558-7_29.

[58] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Columbus, USA, pp. 1725–1732. DOI: 10.1109/CVPR.2014.223.

[59] J. Hu, L. Shen, G. Sun. Squeeze-and-excitation networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 7132–7141, 2018. DOI: 10.1109/CVPR.2018.00745.

[60] S. Woo, J. Park, J. Y. Lee, I. S. Kweon. CBAM: Convolutional block attention module. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 3–19, 2018. DOI: 10.1007/978-3-030-01234-2_1.

[61] T. Zhou, H. Z. Fu, G. Chen, Y. Zhou, D. P. Fan, L. Shao. Specificity-preserving RGB-D saliency detection. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 4661–4671, 2021. DOI: 10.1109/ICCV48922.2021.00464.

[62] W. B. Zhang, G. P. Ji, Z. Wang, K. R. Fu, Q. J. Zhao. Depth quality-inspired feature manipulation for efficient RGB-D salient object detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, ACM, pp. 731–740, 2021. DOI: 10.1145/3474085.3475240.

Zhenyu Li received the B. Sc. degree in computer science from Harbin Institute of Technology, China in 2021. He is a master student in computer science from Harbin Institute of Technology, China.

His research interests include depth estimation and 3D object detection.

E-mail: zhenyuli17@hit.edu.cn

ORCID iD: 0000-0003-2932-9179

Zehui Chen received the B. Sc. degree in software engineering from Tongji University, China in 2020. He is currently a Ph. D. degree candidate with School of Information Science and Technology, University of Science and Technology of China (USTC), China.

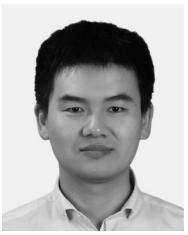
His research interests include object detection and multi-modal learning.

E-mail: lovesnow@mail.ustc.edu.cn

Xianming Liu received the B. Sc. M. Sc., and Ph. D. degrees in computer science from Harbin Institute of Technology (HIT), China in 2006, 2008 and 2012, respectively. In 2011, he spent half a year at Department of Electrical and Computer Engineering, McMaster University, Canada, as a visiting student, where he was a post-doctoral fellow from 2012 to 2013. He was a project researcher with National Institute of Informatics (NII), Japan from 2014 to 2017. He is currently a professor with School of Computer Science and Technology, HIT. He was a receipt of the IEEE ICME 2016 Best Student Paper Award.

His research interests include object detection and multi-modal learning.

E-mail: csxm@hit.edu.cn



Junjun Jiang received the B. Sc. degree in mathematics from Huaqiao University, China in 2009, and the Ph. D. degree in computer science from Wuhan University, China in 2014. From 2015 to 2018, he was an associate professor with School of Computer Science, China University of Geosciences, China. From 2016 to 2018, he was a project researcher with National Institute of Informatics (NII), Japan. He is currently a professor with School of Computer Science and Technology, Harbin Institute of Technology, China. He won the Best Student Paper Runner-up Award at MMM 2017, the Finalist of the World's FIRST 10K Best Paper Award at ICME 2017, and the Best Paper Award at IFTC 2018. He received the 2016 China Computer Federation (CCF) Outstanding Doctoral Dissertation Award and 2015 ACM Wuhan Doctoral Dissertation Award, China.

His research interests include image processing and computer vision.

E-mail: jiangjunjun@hit.edu.cn (Corresponding author)

ORCID iD: 0000-0002-5694-505X