ROIFormer: Semantic-Aware Region of Interest Transformer for Efficient Self-Supervised Monocular Depth Estimation

Daitao Xing[1*], Jinglin Shen[2], Chiuman Ho[2], Anthony Tzes[3]
[1] New York University, USA
[2] OPPO US Research Center, USA
[3] New York University Abu Dhabi and Center for Artificial Intelligence and Robotics, UAE
daitao.xing@nyu.edu, jinglin.shen, chiuman.ho@oppo.com, anthony.tzes@nyu.edu

Abstract

The exploration of mutual-benefit cross-domains has shown great potential toward accurate self-supervised depth estimation. In this work, we revisit feature fusion between depth and semantic information and propose an efficient local adaptive attention method for geometric aware representation enhancement. Instead of building global connections or deforming attention across the feature space without restraint, we bound the spatial interaction within a learnable region of interest. In particular, we leverage geometric cues from semantic information to learn local adaptive bounding boxes to guide unsupervised feature aggregation. The local areas preclude most irrelevant reference points from attention space, yielding more selective feature learning and faster convergence. We naturally extend the paradigm into a multi-head and hierarchic way to enable the information distillation in different semantic levels and improve the feature discriminative ability for fine-grained depth estimation. Extensive experiments on the KITTI dataset show that our proposed method establishes a new state-of-the-art in self-supervised monocular depth estimation task, demonstrating the effectiveness of our approach over former Transformer variants.

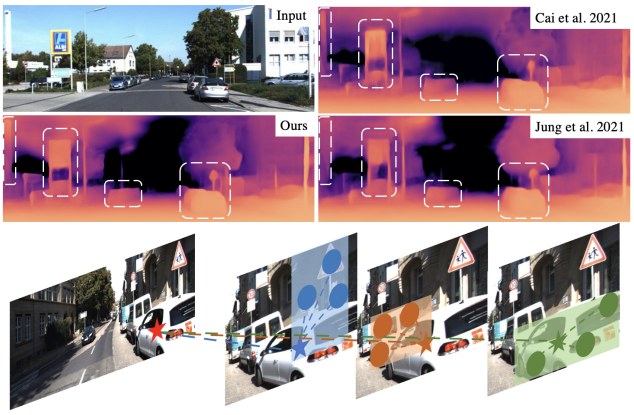

Figure 1: Depth prediction (top two rows) from a single image on KITTI with semantic guidance shows geometry consistency preservation in local uncertainty areas. ROI Former overview (bottom row) shows that the depth of the red star requires attention from various semantic regions.

Introduction

Accurate depth estimation is critical for many applications in computer vision and robotics fields such as perception, navigation, and path planning. The advancements in deep learning have brought significant breakthroughs in the accuracy of depth estimation methods in recent years. While supervised learning-based methods like (Ranftl, Bochkovskiy, and Koltun 2021a) achieved remarkable performance on pixel-wise dense predictions of monocular images, the requirement of a large number of dense labels for training and the excessive spending for acquiring those datasets using LiDAR constraint their usage in real-world applications. Instead, self-supervised depth estimation methods, which learn the depth values using only monocular or stereo image sequences, have become more popular.

Monocular self-supervised depth estimation utilizes the photometric loss and smoothness constraints of consecutive frames in image sequences to simultaneously learn the depth and pose networks. Despite the notable achievement, self-supervised methods, which only rely on similarity constraints, still have large performance gaps with the supervised methods. (Lyu et al. 2021) shows that the bottlenecks come from the inaccurate depth estimation, especially at object boundaries due to the moving objects, ambiguity in low-texture regions, reflective surfaces, occlusion, and the uncertainty of pose estimation. However, the sole consistency constraints of RGB images are insufficient to reduce those effects, which instead require external modalities to provide stronger geometric information.

Recent works, instead, leverage semantic information to improve monocular depth predictions via incorporating geometric guidance. While most of those works incorporate the semantics information explicitly, fewer works focus on designing cross-domain feature aggregation strategies to optimize the intermediate depth representations. Recently, Transformer based attention methods demonstrated their superiority over traditional CNNs in many vision tasks. (Jung, Park, and Yoo 2021) proposes a cross-modality attention module to refine depth features progressively on multi-scales and obtains significant improvement. However,

*Corresponding author.The work was done when Xing was an intern with OPPO

the interaction is restricted within corresponding features to avoid the computational overhead in the classic Transformer methods. (Zhu et al. 2020b) provides an alternative method to Transformer with significantly reduced parameters and faster convergence speed. However, we found that the performance of deformable attention drops dramatically when high-resolution images are used as inputs. We argue that this is due to the attention module collapse and failure to locate relative information as the feature space becomes too large.

The aforementioned analysis related to semantics-guided depth estimation and the success of attention-based feature aggregation indicates that these are mainly determined by two aspects: (1) to provide a dynamic attention region assigned to each reference point, which covers the intact local semantic information and excludes irrelevant points regardless of the spatial size of the feature maps, (2) locate the semantic positions within search areas to update the reference point. Subsequently, we propose a Region-of-Interests (ROIs) guided deformable attention module, named ROIFormer, which performs deformable attention within learnable adaptive local region proposals. Inspired by (Yang et al. 2018) and (Sun et al. 2021), those object-aware proposals can be inferred directly from semantically feature maps via lightweight networks. Unlike (Yang et al. 2018), which creates proposals with deformation on meta-anchors, we merge the proposal generation inside multi-head attention modules and define attention areas implicitly according to information from different semantic levels. The deformable attention is then performed within constrained regions to find the most relative semantic features. With the search space restrained into object-aware local areas, the search complexity is dramatically reduced, resulting in robust feature enhancement and fast convergence.

Despite the depth consistency provision within each connected segment block using the semantic features, the instance level information is still missing, which results in uncertainty on the boundaries. Thus, we consider the spatial relationship between instance objects with the crowded areas (roads, sidewalks, and buildings) in 3D space. The points projected into the 3D space based on the estimation depth and intrinsic camera model should stay close to the nearby points within the same category and keep a reasonable distance to the reference crowded areas. Instead, the points far away from the reference points are masked as outliers which should not be used to calculate photometric similarity loss. Overall, our main contributions are summarized as follows:

- We provide a detailed comparison between different feature fusion strategies for efficient self-supervised depth estimation, indicating that search space complexity is critical for model convergence and performance improvement.
- We propose the ROIFormer, which guides the attention in local areas to most relative semantics information in an unsupervised and efficient way.
- The suggested self-supervised depth estimation with semantics guidance network achieves state-of-the-art performance on varies settings.

Related Work

Self-Supervised Monocular Depth Estimation

Significant improvement has been made since (Zhou et al. 2017) proposed the generalized framework, which enables supervision from consecutive frames via ego-motion. Later works focus on a more elegant loss design to filter unreliable propagation. (Yang et al. 2020) models the photometric uncertainties of pixels on input images. (Shu et al. 2020) proposed to feature metric loss to stabilize the loss landscape process. (Bian et al. 2019) upgraded the geometry consistency loss for scale-consistent predictions. In order to provide more geometry information for self-supervised learning, (Ranjan et al. 2019; Wang et al. 2019; Zhao et al. 2020; Petrovai and Nedevschi 2022; Zhu et al. 2020a) integrate optical flows and pseudo labels for extra constraints. (Guizilini et al. 2020a) and (Lyu et al. 2021) proposed optimized architectures for more efficient depth estimation.

Semantic-guidance for Depth Estimation

Semantic segmentation with strong geometry knowledge is widely used for the promotion of depth estimation. (Lee et al. 2021) improves the performance using an instance-aware geometric consistency loss. (Zhu, Brazil, and Liu 2020) explicitly measures the border consistency between segmentation and depth and minimizes it in a greedy manner. (Casser et al. 2019; Klingner et al. 2020) stabilize the photometric loss by removing moving dynamic-class objects. (Chen et al. 2019) performs region-aware depth estimation by enforcing semantics consistency, while (Guizilini et al. 2020b) uses pixel-adaptive convolutions to produce semantic-aware depth features via assigning weights to features within a local window. (Li et al. 2021) and (Jung, Park, and Yoo 2021) design a cross-task attention module to refine depth features progressively on multi-scales. (Tosi et al. 2020) and (Cai et al. 2021) apply knowledge distillation from semantic segmentation to depth estimation with a learnable domain transfer network.

Efficient Attention Network

Transformer (Carion et al. 2020) indicates a stronger performance over traditional CNNs. After that, lot of works, including (Dai et al. 2017; Xia et al. 2022; Xie et al. 2020; Chen et al. 2021; Yue et al. 2021; Liu et al. 2021; Wang et al. 2021) design efficient multi-scale attention for detection and classification tasks. (Xie et al. 2021) and (Ranftl, Bochkovskiy, and Koltun 2021b) customized Transformer with encoder and decoder frameworks for dense prediction tasks. More recently, (Bae, Moon, and Im 2022) and (Li et al. 2022) migrate Transformer into supervised depth estimation. (Johnston and Carneiro 2020; Zhao et al. 2022) and (Jung, Park, and Yoo 2021) integrate attention into self-supervised depth estimation and obtain significant improvement. (Nguyen et al. 2022) samples points over 3X3 transformed grids guided from ground truth boxes for efficient detection. However, those methods rely on either global dependence or local interactions. Instead, we select attention areas adaptively for optimal efficiency.

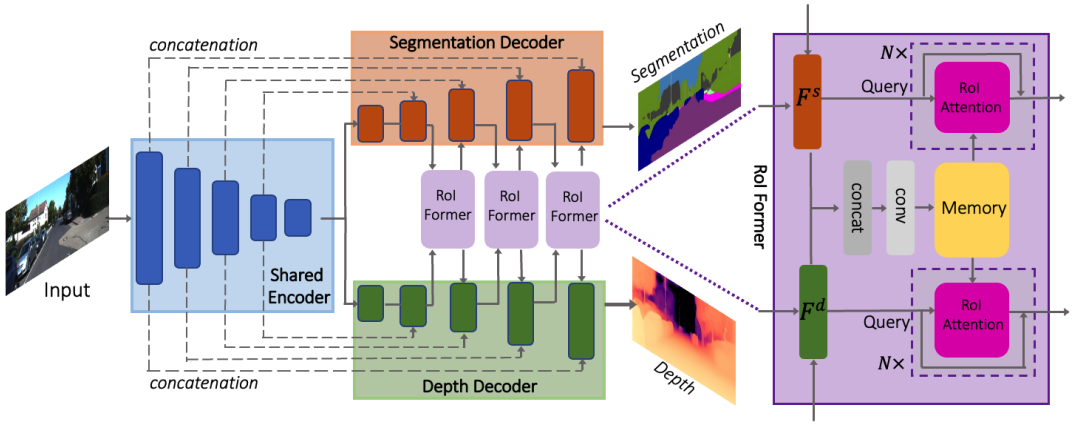

Figure 2: Proposed framework for self-supervised monocular depth estimation with semantic guidance. The segmentation and depth branches share the same backbone as the encoder, and capture mutual benefits information from other domains with adaptive local attention modules.

Proposed Methods

Self-Supervised Depth Estimation

Self-supervised monocular depth estimation utilizes source images I_{t-1} and I_{t+1} to build reference image $\hat{I}_{t'\to t}$ for target image I_t via geometric transformation. We learn a scale-ambiguous depth $\hat{D}_t$ map and a corresponding semantic segmentation map $\hat{S}_t$ from the multi-task network. With known intrinsic camera parameter matrix $K \in \mathbb{R}^{3\times 3}$, the pixels $p \in I_t$ with homogeneous coordinates u are projected to the 3D-space, resulting in a point cloud P_t with pixel value and semantic information. The pose network, instead, outputs the 6 DOF transformation $T_{t\to t'}|t' \in \{t-1, t+1\}$, rotates, shifts, and re-projects the point cloud to get pixel coordinates in image $I_{t'}$. The correspondence between I_t with I'_t can be summarized as:

$$p' = KT_{t\to t'}\hat{D}_t K^{-1}p. \tag{1}$$

The reference image $\hat{I}_{t'\to t}$ for I_t is obtained via the interpolation on I'_t according to the reprojected pixel coordinates p'. The objective is to minimize the similarity between $\hat{I}_{t'\to t}$ and I_t by calculating the Structural Similarity (SSIM) loss and L1 distance, as:

$$L_{p,t'} = \alpha \frac{1 - \text{SSIM}\left(I_t, \hat{I}_{t'\to t}\right)}{2} + (1-\alpha)\left|I_t - \hat{I}_{t'\to t}\right|. \tag{2}$$

We incorporate the minimum reprojection error by selecting the per-pixel minimum value cross all similarity losses, where $L_p = \min_{t'} L_{p,t'}$. Following (Godard et al. 2017), we include an edge-aware term to smooth the depth in low gradient areas, defined as:

$$L_s = |\partial_x D_t|\, e^{-|\partial_x I_t|} + |\partial_y D_t|\, e^{-|\partial_y I_t|}. \tag{3}$$

The smoothness loss reinforces the depth similarity between pixels with small differences in grey values.

It is proven that the multi-task joint training stimulates mutual benefits and results in significant improvements to both tasks. Therefore, we adopt pre-computed segmentation maps from (Jung, Park, and Yoo 2021) as ground truth and train the semantic segmentation branch with cross entropy loss.

Feature Enhancement with ROI Attention

The principle of jointly training depth estimation and segmentation is to distill the semantic and position information and obtain a more discriminative depth representation. Given a depth feature map $F^d \in \mathbb{R}^{H\times W\times C}$ and segmentation map $F^s \in \mathbb{R}^{H\times W\times C}$ from level l with feature dimension equals C and W, H are width and height of feature maps on level l. The enhanced geometric representations can be generalized as:

$$\text{Fusion}\langle f_i, F^s\rangle = \sum_{j\in\Omega(F^s)} A_{i,j}\mathbf{W}_{i,j}f_j, \tag{4}$$

where $f_i \in F^d$ is query feature from depth image at position i, A is assigned weight of j_{th} feature point f_j from sample space $\Omega(F^s)$ which is a subset of F^s based on sampling function Ω, and $\mathbf{W}_{i,j} \in \mathbb{R}^{C\times C}$ is the feature projection matrix. In (Guizilini et al. 2020b), the distribution of the weights is calculated as the correlation between the guiding features and the Gaussian kernel. The projection matrix $\mathbf{W}$ are convolutional weights with kernel size k and Ω is defined as a $k \times k$ convolutional window.

Transformer attention is designed to capture global dependencies to build spatial interactions over enlarged areas. Specifically, let the sampling space Ω include all feature points from F_s. The attention function is performed on query vector $\mathbf{W}_q f^d$, key vector $\mathbf{W}_k f^s$ and value vector $\mathbf{W}_v f^s$, where $\mathbf{W}_q, \mathbf{W}_k$ and $\mathbf{W}_v$ are three separate linear transform functions. The attention weight is expressed as $A_{i,j} \propto \exp\left\{\frac{f_i^T \mathbf{W}_q^T \mathbf{W}_k f_j}{\sqrt{C}}\right\}$, where $\sum_{j\in\Omega} A_{ij} = 1$. Since the attention weights are distributed over the entire feature space, the cross-attention module suffers from slow convergence and computational complexity overhead on feature

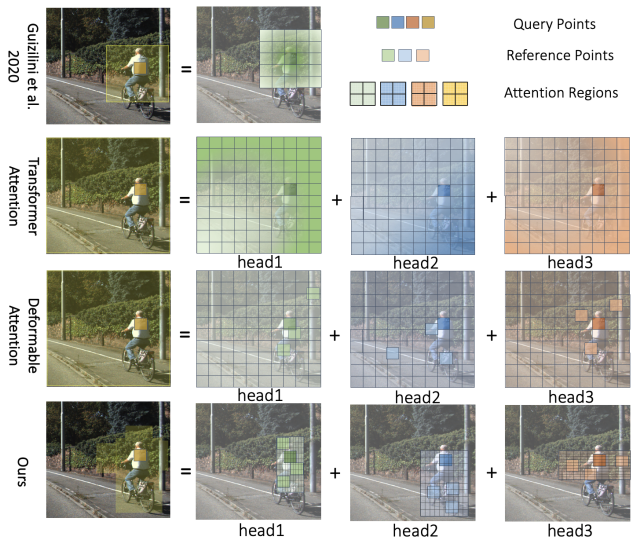

Figure 3: Comparison of feature fusion strategies for depth estimation: adaptive convolution (top row), global dense attention (second row), global sparse attention (third row) and our adaptive sparse attention (bottom row).

maps with large spatial sizes. An alternative way to cross-attention is deformable attention from (Zhu et al. 2020b), which only attends to a small set of key sampling points around the query point with:

$$\text{Fusion}\left\langle f_i, F^s\right\rangle = \sum_{j\in\Omega(F^s)} A_{i,j}\mathbf{W}_{i,j} f_{p_i+\Delta p_i}. \quad (5)$$

In specific, the Ω samples M pairs of key-value vectors from F^s. Those vectors are obtained from F^s via interpolation on position $p_i + \Delta p_i$, where p_i is the position coordinates of query point f_i; $\Delta p_i \in \mathbb{R}^2$ denotes the sampling offset with unconstrained range, which are learned from linear functions. While the computational cost, especially on large spatial resolutions, can be significantly reduced using sampled key-value pairs, the sample space is still the same size as F_s. The experimental results show that the sample space is too large for the linear function to find the most relative points, yielding divergence attention and worse performance in high-resolution settings, as shown in Figure 3.

To improve the attention efficiency and facilitate the location of relative information in limited iterations, we constraint the attention in a local region which is much smaller than the original sample space. Inspired by (Yang et al. 2018), we generate the Region of Interests (ROIs) for each query point. Thanks to the supervision from segmentation, the local semantic aware ROIs can be inferred from F^s efficiently. Specifically, we denote the ROI for a query point f_i as bin $b_i = [d_l, d_t, d_r, d_b] \in [0, W]\times[0, H]\times[0, W]\times[0, H]$ which represents the distances from f_i to the edges of the ROI box and $[w = d_l+d_r, h = d_t+d_b]$ are normalized width and height.In practice, we constraint the width and height of ROIs with a maximum value, r_{min} and r_{max}.Therefore, the feature fusion within a specific region of interest can be expressed as:

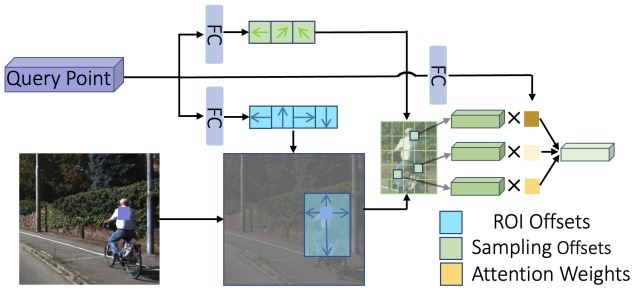

Figure 4: ROI attention overview: a local region of interest is first inferred from semantic cues, using sampling offsets and attention weights with two separate linear functions.The features are sampled via interpolation without cropping the ROI region explicitly.

$$\text{Fusion}\left\langle f_i, b_i\right\rangle = \sum_{j\in\Omega(b_i)} A_{i,j}\mathbf{W}_{i,j} f_{p_i+\Delta\frac{1}{2}\cdot p_i\cdot b_i^{wh}}, \quad (6)$$

where the key-value pairs are sampled from b_i via the interpolation on position $p_i+\Delta\frac{1}{2}\cdot p_i\cdot b_i^{wh}$, $\Delta p_i \in \mathbb{R}^2$ denotes the sampling offset with unconstrained range and p_i is the center position of ROI. ROI size b_i^{wh} guarantees the sampling offset is within $[-\frac{w}{2}, \frac{w}{2}]\times[-\frac{h}{2}, \frac{h}{2}]$. Without cropping and pooling explicitly, the sampling points can be obtained via interpolation over the neighboring positions inside the ROIs. We simplify the ROI generation with linear function which outputs a bounding box based on the hypothesis that most relevant information are distributed around the query point. Moreover, the deformable attention mechanism enables sampling in any position within ROIs. The attention module is shown in Figure 4. By introducing a lightweight module, our semantic bounded attention module prompts the attention efficiency and concentrates computational resources via narrowing down the search space.

ROI Former Module

Inspired by (Carion et al. 2020), the representative ability of ROI Attention can be boosted with a multi-head setting to capture guidance information from different semantic levels. Thus, we employ the multi-head structure to generate multiple proposals which provide separate ROIs for local attention. The multi-head ROI attention feature is:

$$f_i^d = \text{concat}(\text{Fusion}\left\langle f_i, b_i^1\right\rangle, \ldots, \text{Fusion}\left\langle f_i, b_i^M\right\rangle)\mathbf{W}_O \quad (7)$$

where Fusion $\left\langle f_i, b_i^m\right\rangle$ is the fusion feature after ROI attention on the m_{th} head. The features from the M head attention are merged by concatenation, following an output projection matrix $\mathbf{W}_O$. Beside the attention from depth to segmentation features, we apply the same multi-head ROI attention over depth feature to update semantic representation for mutual enhancement. To save the computational overhead when computing the guiding feature maps, we stack f^s and f^d at the beginning and use it as a shared guidance feature map after the convolutional layers, as shown in Figure 2 .

The shared guidance feature map is then fed into two attention blocks with N stacked multi-head ROI attention layers to update depth and semantic feature separately. The outputs are fed into the upper level after upsampling to refine features on high resolution layers.

Network Architecture

Our segmentation and depth estimation networks have the same architecture as in(Godard et al. 2019), i.e., a U-Net with skip connections, except for the depth fusion module. The feature maps of encoder $P = C6, C5, C4, C3, C2$ from 1/32 to 1/2 are extracted and fed into different levels of the encoder for dense prediction. Those feature maps are projected into $P = P6, P5, P4, P3, P2$ of dimension $C = 256, 128, 64, 32, 16$ with five separate convolutional layers. The decoder consists of five upsample stages where, in each stage, the feature from the last decoding level is pre-processed with a convolutional layer and then upsampled and concatenated with features in the current level. The concatenated features are fed into another convolutional layer. The same operations are applied to the segmentation branch. Finally, the semantic features and depth features are fed into the ROIFormer module to obtain the fusion features, as shown in Figure 2. Within ROIFormer, the depth feature and semantic feature are first concatenated into a common feature map. For efficiency purposes, we stack the feature maps from the segmentation branch and depth branch directly and use them as common attention memory.

Semantic Guided Re-projection Mask

Similar to (Godard et al. 2019), we also predict depth maps $\hat{D}^L$ on intermediate layers to calculate projection loss on multiple scales. For the segmentation map, we only use the final predictions. Photometric loss calculation is less accurate near the object boundaries and instance segments connected regions due to the depth estimation uncertainties. To overcome the boundary contamination problem, we create a mask for each point with a penalty coefficient according to the distance from the instance points to the areas. To this end, we build a graph from instance points set $\mathcal{S}_{Ins}$ to reference points set $\mathcal{S}_{Ref}$ and apply K-nearest-neighbors to sample the K and get average relative distance $d_{t,i\rightarrow\mathcal{S}_{Ref}}$. Consequently, after re-projection into the consecutive images, we get the confidential mask as:

$$\mu_{t,i} = \begin{cases} \frac{1}{e^{\alpha d_{t,i}}}, & i \in \mathcal{S}_{ins} \\ 1, & i \notin \mathcal{S}_{ins} \end{cases} \tag{8}$$

We set 1 to each ith pixel position not belonging to instance points set $\mathcal{S}_{ins}$ and penalty weight to instance points obtained from an exponentially decreasing function with a scale factor α. As shown in Figure 5, compared with (Godard et al. 2019)'s automask, our semantic bounded mask concentrate on the border areas where the depth value varies dramatically or reflective areas like windows of cars. In practice, we only consider the salient objects as reference points. The loss function of our model is obtained as :

$$L = \mu * L_p + \beta \cdot L_s + \gamma \cdot L_{sem}. \tag{9}$$

where β and γ are weighted factors for smoothness loss and segmentation loss.

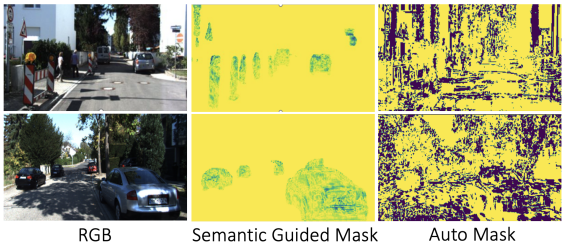

Figure 5: Visualization comparison between the semantic guided re-projection mask (second column) and auto mask (third column).

Experimental Results

Implementation Details

Our encoder is built using Resnet-18 and Resnet-50 backbones pre-trained on the ImageNet. The pose network is a pre-trained Resnet-18 model. The input image size is set to a medium resolution of 192×640 and a high resolution of 320×1024 following the same setting in the former methods for a fair comparison. We train our model with a batch size of 12 on a single NVIDIA Tesla V100 GPU. Following (Godard et al. 2019), all experiments are trained with 20 epochs with the learning rate of 10^{-4} and decayed by 10 on the 10th and 15th epochs. All experiments are trained with an ADAM optimizer. For supervised segmentation training, we employ the Cross-Entropy loss with a hyper-parameter $\gamma = 0.5$ to control the weights of segmentation loss. The α for photometric loss is set to 0.85 and $\beta = 1 \times 10^{-3}$. The ROIs range is set to $r_{min} = \{0.3W, 0.3H\}$ and $r_{max} = \{0.7W, 0.7H\}$ For evaluation, the input image size is the same as the training set, and the final output of the model is resized into the ground truth resolution (384×1280). We apply the ground truth scale technique to restore the absolute depth estimation, and the depth values are restricted to the range of 0 to 80 meters. No further post-processing is required.

The KITTI dataset is used for self-supervised depth evaluation using the same settings as in (Zhou et al. 2017) to remove the static frames from the data split of (Eigen and Fergus 2015), resulting in 39810 images used for training, 4424 for validation ,and 697 for final evaluation.

Ablation Study

ROIFormer Module ROIFormer is the core operation in the proposed method, which guarantees the efficient mutual information fusion between semantics and depth and the feature representation enhancement. We first investigate the

# of layers	AbsRel	SqRel	RMSE	RMSElog
0	1.037	0.692	4.428	0.179
1	1.011	0.679	4.323	0.175
2	1.002	0.667	4.320	0.174
3	1.015	0.617	4.359	0.176
4	1.027	0.678	4.364	0.177

Table 1: Ablation study on the number of attention layers. The number of attention layers are shared across P3 to P5.

P5	P4	P3	P2	AbsRel	SqRel	RMSE
✓				0.103	0.7049	4.437
✓	✓			0.1008	0.7004	4.392
✓	✓	✓		0.1003	0.6844	4.366
✓	✓	✓	✓	0.1002	0.6603	4.330

Table 2: Ablation study on the impact of layers for feature fusion. P2-inclusion improves the performance, while P3-P5 was only used in all other experiments.

impact of the number of attention layers, where the attention block is applied on P3, P4, and P5 levels with N attention layers stacked for both segmentation and depth queries. Thanks to the shared memory and local attention design, the memory cost is linear to the input feature size, making it possible to test more stacked layers. As shown in Table 1, removing the attention blocks degenerate the fusion layer into a simple concatenation operation. The interaction between two domains relies on stacked convolutional operations, which provides limited feature enhancement. Adding one attention layer yields significant improvement over all metrics. We found that using two stack layers achieves the best trade-off between accuracy and complexity while additional layers contribute less or even hinder the performance.

We further investigate the impact of attention layers on different pyramid levels and show results in Table 2. The experiments show that feature fusion only on deep layers with limited resolution leads to a minor improvement in depth precision. Shallow layers with fewer feature channels bring more fine-grained details, resulting in compelling results. We further test our attention module on P2 level with 1/2 input size resulting in similar results. Finally, we compare the two sampling strategies and the effects of attention points in the attention module. Similar to the aforementioned cases, the features in shallows are vital for final outputs. Thus, it is reasonable that assigning more attention points in shallow layers brings more performance benefits. In our experiments, the combination of [8,16,32] is the best setting considering the precision and efficiency trade-off.

Complexity Comparison The model efficiency of our proposed method is explored by comparing it with other attention-based feature fusion methods, i.e., transformer attention and deformable attention. As shown in Table3, the transformer performs better than deformable attention but exceeds the memory limit on HR inputs. The performance of

Type	#	AbsRel	SqRel	RMSE
Fixed	[8]	0.1036	0.7269	4.416
	[16]	0.1009	0.7027	4.395
	[32]	0.1012	0.7164	4.414
Total	[4,8,16]	0.1010	0.6972	4.386
	[8,16,32]	0.1003	0.6855	4.360
	[16,32,64]	0.0998	0.7137	4.415

Table 3: Ablation study on attention points sampling strategies and number of points. The features of pyramid layers are fixed, as well as the attention head numbers.

Type	Input	Net	Sq Rel	RMSE
Transformer	640	Res18	0.722	4.547
Transformer	1024	Res18	0.687	4.366
Transformer	640	Res50	0.675	4.393
Transformer	1024	Res50	×	×
Deformable	640	Res18	0.717	4.472
Deformable	1024	Res18	0.704	4.391
Deformable	640	Res50	0.696	4.399
Deformable	1024	Res50	0.738	4.517
ROIFormer	640	Res18	0.695	4.438
ROIFormer	1024	Res18	0.674	4.335
ROIFormer	640	Res50	0.673	4.351
ROIFormer	1024	Res50	0.616	4.148

Table 4: Comparison between different attention types

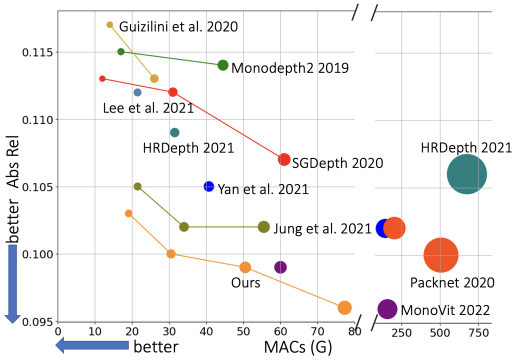

Figure 6: Speed-Accuracy trade-off curve.

deformable attention on HR inputs drops dramatically. We argue that the sample space is too large to select valid key-value pairs, resulting in divergence. ROIFormer, achieves the best performance across all settings, demonstrating its superiority over other attention methods. Figure. 7 shows the attention areas and sampled key values based on different attention variants. ROIFormer samples the relative features more efficiently. We further summarize the model complexity against the performance of all STOA methods and plot the trade-off curve in Figure.6. The model size is shown as areas of circles.

Impact of Individual Component. We summarize the impact on depth estimation accuracy of each main component in Table 5. For the baseline, we keep only self-supervised depth estimation modules including smoothness loss and photometric loss. Applying our attention module only to the intermediate depth features brings over 6% im-

M	Sem	ROIFormer	Mask	AbsRel	RMSE
✓				0.115	4.863
✓		✓		0.108	4.595
✓	✓	✓		0.1005	4.351
✓	✓	✓	✓	0.1002	4.356

Table 5: Ablation study for the contribution of three main components: monocular (M) training, semantics (Sem) information, ROIFormer and semantics guided mask loss.

Method	Net	Sem	Input	AbsRel	SqRel	RMSE	RMSElog	$\theta < 1.25$	$\theta < 1.25^2$	$\theta < 1.25^3$
Zou et al. 2020	Res18		MR	0.115	0.871	4.778	0.191	0.874	0.963	**0.984**
Johnston et al. 2020	Res18		MR	0.111	0.941	4.817	0.185	0.885	0.961	0.981
Hui et al. 2022	Res18		MR	0.108	0.71	4.513	0.183	0.884	0.964	0.983
Jung et al. 2021	Res18	✓	MR	0.105	0.722	4.547	0.182	0.886	0.964	**0.984**
Ours	Res18	✓	MR	**0.103**	**0.696**	**4.438**	**0.178**	**0.889**	**0.965**	**0.984**
Klingner et al. 2020	Res18	✓	HR	0.107	0.768	4.468	0.186	0.891	0.963	0.982
Choi et al. 2020	Res18		HR	0.106	0.743	4.489	0.181	0.884	0.965	**0.984**
Lyu et al. 2021	Res18		HR	0.106	0.755	4.472	0.181	0.892	0.966	**0.984**
Jung et al. 2021	Res18	✓	HR	0.102	0.687	4.366	0.178	0.895	**0.967**	**0.984**
Ours	Res18	✓	HR	**0.100**	**0.675**	**4.335**	**0.176**	**0.896**	**0.967**	**0.984**
Godard et al. 2019	Res50		MR	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Guizilini et al. 2020b	Res50	✓	MR	0.113	0.831	4.663	0.189	0.878	0.971	0.983
Kumar et al. 2021	Res50	✓	MR	0.109	0.718	4.516	0.18	**0.896**	**0.973**	**0.986**
Yan et al. 2021	Res50		MR	0.105	0.769	4.535	0.181	0.892	0.964	0.983
Li et al. 2021	Res50	✓	MR	0.103	0.709	4.471	0.18	0.892	0.966	0.984
Jung et al. 2021	Res50	✓	MR	0.102	0.675	4.393	0.178	0.893	0.966	0.984
Ours	Res50	✓	MR	**0.100**	**0.673**	**4.351**	**0.175**	**0.896**	0.967	0.985
Godard et al. 2019	Res50		HR	0.115	0.882	4.701	0.19	0.879	0.961	0.982
Shu et al. 2020	Res50		HR	0.104	0.729	4.481	0.179	0.893	0.965	0.984
Gurram et al. 2021	Res50		HR	0.104	0.721	4.396	0.185	0.88	0.962	0.983
Kumar et al. 2021	Res50	✓	HR	0.102	0.701	4.347	**0.166**	0.901	**0.98**	**0.99**
Cai et al. 2021	Res50	✓	HR	0.102	0.698	4.439	0.18	0.895	0.965	0.983
Chanduri et al. 2021	Res50		HR	0.102	0.723	4.374	0.178	0.898	0.966	0.983
Petrovai et al. 2022	Res50		HR	0.101	0.72	4.339	0.176	0.898	0.967	0.984
Ours	Res50	✓	HR	**0.096**	**0.616**	**4.148**	0.169	**0.905**	0.969	0.986

Table 6: Comparison with the state-of-the-art on KITTI Eigen test set.

Figure 7: Visualization of attention distribution.

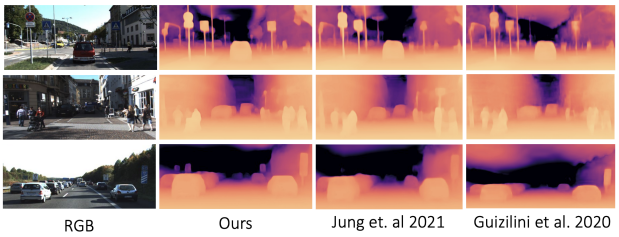

Figure 8: Visualization of depth estimation on KITTI.

provement on relative errors which imply the significant impact of attention on feature representation enhancement. Together with semantics information, the attention module is capable of getting the most of mutual benefits from segmentation and depth domains.

Comparison with State-Of-The-Art Methods

We evaluate ROIFormer on the KITTI dataset based on the metrics from (Eigen and Fergus 2015). As shown in Table 6, our proposed method outperforms all existing SOTA self-supervised monocular depth estimation methods, including approaches utilizing semantic information. ROIFormer enables flexible local feature interaction, resulting in a much higher performance under HR settings with more fine-grained details. We also compare the model complexity by calculating the MACs and model parameter numbers. The trade-off curve is shown in Figure 6. We achieve the same performance with only 30% MACs, compared to the recent SOTA method (Zhao et al. 2022) which utilizes the transformer as the backbone. With Resnet18, the ROIFormer runs at 51 fps on MR and 33 fps on HR. With Resnet50, it runs at 45 and 24 fps respectively. Finally, Figure 8 illustrates the performance comparison with other SOTA methods qualitatively.

Conclusion

In this paper, we propose a novel attention module to improve the self-supervised depth estimation accuracy. Our method enhances the representative feature ability by learning spatial dependencies from local semantic areas. We leverage geometric cues from segmentation feature maps to learn the regions of interest to guide feature aggregation cross domains. The attention module is employed in a multi-head and multi-scale schema to enable the feature learning from different semantic levels. Furthermore, we introduce a semantic aware projection mask to improve the model robustness on uncertain areas.We conducted extensive experiments on the KITTI dataset and achieved new SOTA performance, which demonstrated its effectiveness.

References

Bae, J.; Moon, S.; and Im, S. 2022. MonoFormer: Towards Generalization of self-supervised monocular depth estimation with Transformers. *arXiv preprint arXiv:2205.11083*.

Bian, J.; Li, Z.; Wang, N.; Zhan, H.; Shen, C.; Cheng, M.-M.; and Reid, I. 2019. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Advances in neural information processing systems*, 32.

Cai, H.; Matai, J.; Borse, S.; Zhang, Y.; Ansari, A.; and Porikli, F. 2021. X-distill: Improving self-supervised monocular depth via cross-task distillation. *arXiv preprint arXiv:2110.12516*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Casser, V.; Pirk, S.; Mahjourian, R.; and Angelova, A. 2019. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI conference on artificial intelligence*, 01, 8001–8008.

Chen, P.-Y.; Liu, A. H.; Liu, Y.-C.; and Wang, Y.-C. F. 2019. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2624–2632.

Chen, Z.; Zhu, Y.; Zhao, C.; Hu, G.; Zeng, W.; Wang, J.; and Tang, M. 2021. Dpt: Deformable patch-based transformer for visual recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2899–2907.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Eigen, D.; and Fergus, R. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, 2650–2658.

Godard, C.; Mac Aodha, O.; Firman, M.; and Brostow, G. J. 2019. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3828–3838.

Guizilini, V.; Ambrus, R.; Pillai, S.; Raventos, A.; and Gaidon, A. 2020a. 3D Packing for Self-Supervised Monocular Depth Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guizilini, V.; Hou, R.; Li, J.; Ambrus, R.; and Gaidon, A. 2020b. Semantically-guided representation learning for self-supervised monocular depth. *arXiv preprint arXiv:2002.12319*.

Johnston, A.; and Carneiro, G. 2020. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 4756–4765.

Jung, H.; Park, E.; and Yoo, S. 2021. Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12642–12652.

Klingner, M.; Termöhlen, J.-A.; Mikolajczyk, J.; and Fingscheidt, T. 2020. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *European Conference on Computer Vision*, 582–600. Springer.

Lee, S.; Im, S.; Lin, S.; and Kweon, I. S. 2021. Learning monocular depth in dynamic scenes via instance-aware projection consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3, 1863–1872.

Li, R.; He, X.; Xue, D.; Su, S.; Mao, Q.; Zhu, Y.; Sun, J.; and Zhang, Y. 2021. Learning depth via leveraging semantics: Self-supervised monocular depth estimation with both implicit and explicit semantic guidance. *arXiv preprint arXiv:2102.06685*.

Li, Z.; Chen, Z.; Liu, X.; and Jiang, J. 2022. DepthFormer: Exploiting Long-Range Correlation and Local Information for Accurate Monocular Depth Estimation. *arXiv preprint arXiv:2203.14211*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Lyu, X.; Liu, L.; Wang, M.; Kong, X.; Liu, L.; Liu, Y.; Chen, X.; and Yuan, Y. 2021. Hr-depth: High resolution self-supervised monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3, 2294–2301.

Nguyen, D.-K.; Ju, J.; Booij, O.; Oswald, M. R.; and Snoek, C. G. 2022. BoxeR: Box-Attention for 2D and 3D Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4773–4782.

Petrovai, A.; and Nedevschi, S. 2022. Exploiting Pseudo Labels in a Self-Supervised Learning Framework for Improved Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1578–1588.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021a. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12179–12188.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021b. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12179–12188.

Ranjan, A.; Jampani, V.; Balles, L.; Kim, K.; Sun, D.; Wulff, J.; and Black, M. J. 2019. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12240–12249.

Shu, C.; Yu, K.; Duan, Z.; and Yang, K. 2020. Feature-metric loss for self-supervised learning of depth and ego-motion. In *European Conference on Computer Vision*, 572–588. Springer.

Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. 2021. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14454–14463.

Tosi, F.; Aleotti, F.; Ramirez, P. Z.; Poggi, M.; Salti, S.; Stefano, L. D.; and Mattoccia, S. 2020. Distilled semantics for comprehensive scene understanding from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4654–4665.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 568–578.

Wang, Y.; Wang, P.; Yang, Z.; Luo, C.; Yang, Y.; and Xu, W. 2019. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8071–8081.

Xia, Z.; Pan, X.; Song, S.; Li, L. E.; and Huang, G. 2022. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4794–4803.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Xie, Z.; Zhang, Z.; Zhu, X.; Huang, G.; and Lin, S. 2020. Spatially adaptive inference with stochastic feature sampling and interpolation. In *European conference on computer vision*, 531–548. Springer.

Yang, N.; Stumberg, L. v.; Wang, R.; and Cremers, D. 2020. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1281–1292.

Yang, T.; Zhang, X.; Li, Z.; Zhang, W.; and Sun, J. 2018. Metaanchor: Learning to detect objects with customized anchors. *Advances in neural information processing systems*, 31.

Yue, X.; Sun, S.; Kuang, Z.; Wei, M.; Torr, P. H.; Zhang, W.; and Lin, D. 2021. Vision transformer with progressive sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 387–396.

Zhao, C.; Zhang, Y.; Poggi, M.; Tosi, F.; Guo, X.; Zhu, Z.; Huang, G.; Tang, Y.; and Mattoccia, S. 2022. MonoViT: Self-Supervised Monocular Depth Estimation with a Vision Transformer. In *International Conference on 3D Vision*.

Zhao, W.; Liu, S.; Shu, Y.; and Liu, Y.-J. 2020. Towards better generalization: Joint depth-pose learning without posenet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9151–9161.

Zhou, T.; Brown, M.; Snavely, N.; and Lowe, D. G. 2017. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1851–1858.

Zhu, J.; Shi, Y.; Ren, M.; and Fang, Y. 2020a. MDA-Net: Memorable Domain Adaptation Network for Monocular Depth Estimation. In *BMVC*.

Zhu, S.; Brazil, G.; and Liu, X. 2020. The edge of depth: Explicit constraints between segmentation and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13116–13125.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020b. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.