GlobalDepth: Global-Aware Attention Model for Unsupervised Monocular Depth Estimation

Huimin Yu, Ruoqi Li, Zhuoling Xiao*, Bo Yan
School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China
Email: {202122010225, ruoqi_li}@std.uestc.edu.cn, {zhuolingxiao, yanboyu}@uestc.edu.cn

Abstract—Monocular depth estimation is a significant task in computer vision, which can be widely used in Simultaneous Localization and Mapping (SLAM) and navigation. However, the current unsupervised approaches have limitations in global information perception, especially at distant objects and the boundaries of the objects. To overcome this weakness, we propose a global-aware attention model called GlobalDepth for depth estimation, which includes two essential modules: Global Feature Extraction (GFE) and Selective Feature Fusion (SFF). GFE considers the correlation among multiple channels and refines the encoder feature by extending the receptive field of the network. Furthermore, we restructure the skip connection by employing SFF between the low-level and the high-level features in element wise, rather than simply concatenation or addition at the feature level. Our model excavates the key information and enhances the ability of global perception to predict details of the scene. Extensive experimental results demonstrate that our method reduces the absolute relative error by 10.32% compared with other state-of-the-art models on KITTI datasets.

Index Terms—unsupervised monocular depth estimation, global feature extraction, selective feature fusion.

I. INTRODUCTION

Monocular depth estimation is the task of generating a dense depth map from a single RGB image, which has been studied extensively as a fundamental problem in computer vision. An accurate dense depth map estimated from a monocular camera is particularly useful for a low-cost system to understand the three-dimensional (3D) geometric information of a scene. However, the estimation of a depth map from a monocular image is usually ambiguous and ill-posed since a two-dimensional (2D) scene may be projected from an infinite number of real-world 3D scenes. Therefore, researchers apply deep learning to this problem to achieve monocular depth estimation in an end-to-end manner.

Several supervised depth estimation models [1]–[3] have achieved excellent results, but they require massive ground truth collected by sensors such as Light Detection And Ranging (LiDAR). As an alternative, unsupervised methods use geometrical constraints on stereo images [4], [5] or image sequences [6], [7] to minimize the photometric re-projection error as supervision. However, they are inaccurate in the prediction of the object edges. Recent works have attempted to overcome this problem by leveraging cross-domain knowledge, such as semantic information. Struct2Depth [8] used a trained instance segmentation model to split the moving objects, mitigating the impact of dynamic objects. Zhang et al. [9] proposed a bootstrapped self-supervised training method to improve semantic segmentation as well as depth estimation. Inspired by these works, we introduce semantic cues to help learn robust geometric features.

Current methods [6], [7] usually adopt the U-Net structure [10] without explicit geometric exploration of 3D scenes. Global and local details of complex scenes are ignored, which is the key to obtaining relative depth and sharp boundaries of objects [11]. To guide the model to perceive global and local information, attention mechanisms are introduced into many computer vision tasks. CBAM [12] extends SENet [13] with a spatial attention module, and performs adaptive feature selection from two dimensions of space and channel. Gao et al. [14] devised an attentional separation-and-aggregation network (ASANet) that can distinguish whether the pixels belong to static or dynamic scenes while estimating the camera’s ego-motion and the scene’s dynamic motion field. Johnston et al. [15] captured the context of similar disparity values at non-contiguous regions by exploring the feature similarity.

In this paper, a global-aware attention model for depth estimation is proposed, which aims to perceive global as well as detailed information. We embed semantic prior at first for data augmentation. Next, we leverage the global feature extraction module to refine the encoder feature by extending the receptive field of the network. Furthermore, the classical skip connection is redesigned by employing selective feature fusion in element wise. Finally, we construct the photometric re-projection loss as an internal constraint for unsupervised training. The main contributions of our work are as follows:

- A global feature extraction module is proposed to excavate the relations among different channels and obtain the global feature.
- For skip connection, we employ a tightly coupled method to fuse low-level and high-level features, enhancing the prediction of sharper edges.
- Extensive experiments show that our model achieves excellent performance both on the standard KITTI benchmark and the Make3D dataset.

II. GLOBAL-AWARE ATTENTION MODEL

In this section, we first present an overview of the proposed model in II-A and then introduce the global-aware attention model in II-B. Finally, we explain the training method for our depth estimation model in II-C.
A. Overview

An overview of the GlobalDepth is shown in Fig. 1. We embed semantic maps from the semantic segmentation network into original images. Then, the augmented data is sent to the attention-based depth network along with two essential modules called Global Feature Extraction and Selective Feature Fusion. A global-aware depth map is predicted through this attention-based depth model, which has a better estimation on boundaries and a global recognition of distant objects.

B. Attention-Based Depth Network

1) Global Feature Extraction: We use HRNet-OCR network [16] as the backbone of the semantic segmentation network and ResNet18 [17] as the encoder of the depth network and pose network. The input of this model is the augmented data after embedding semantic cues. Then the encoder feature maps are fed to the hierarchical decoder for depth estimation. We introduce a channel attention to capture the vertical and horizontal correlation, as known as a loosely coupled scheme. This may cause artifacts in depth estimation, since the semantic gap between the encoder and decoder feature maps is too large. To tackle this problem, we adopt a tightly coupled approach, extracting features through two branches and selectively fusing the original features. Finally, after passing through a sigmoid function, it enables the network to conduct a weights allocation between the encoder and decoder feature maps.

2) Selective Feature Fusion: The SFF module is applied in skip connection to fuse the low-level and high-level features. The skip connection is the core of the U-Net structure [10] to prevent information lost in downsampling. In the previous work, the most common method was addition or concatenation, as known as a loosely coupled scheme. This may cause artifacts in depth estimation, since the semantic gap between the encoder and decoder feature maps is too large. To tackle this problem, we adopt a tightly coupled approach, extracting features through two branches and selectively fusing the original features. Finally, after passing through a sigmoid function, it enables the network to conduct a weights allocation between the encoder feature and the decoder feature in element wise. Through our insightful design, the fused features predict precise edges and improve the overall performance.

The process is shown in Fig. 3. Given two feature maps $L, H \in \mathbb{R}^{C \times H \times W}$ from encoder and decoder respectively, we generate the feature $F$ by element-wise summation:

$$F = L + H$$

In the upper branch, we sequeeze $F$ to a vector $\in \mathbb{R}^{C \times 1 \times 1}$ by global average pooling:

$$p(F) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F$$

Next, the channel attention $g(F) \in \mathbb{R}^{C \times 1 \times 1}$ is defined as:

$$g(F) = B(W_1 \otimes (\delta(B(W_2 \otimes p(F))))))$$

where $\otimes$ denotes $1 \times 1$ convolution with weights, $B$ refers to the batch normalization and $\delta$ is the ReLu activation function. Similarly, we get the spatial attention $l(F) \in \mathbb{R}^{C \times H \times W}$ :

$$l(F) = B(W_1 \otimes (\delta(B(W_2 \otimes F))))$$
Next, we utilize L1 and SSIM [18] to make the loss:

\[
T_p D
\]

where \( I \) the pixel in \{frames temporally adjacent to \( I \) \( I \) source image to target image is obtained by the pose network.

C. Constraints for Unsupervised Training

1) View Synthesis Loss: Similar to Monodepth2 [7], our model is trained in an unsupervised manner by minimizing the photometric reprojection error. The relative pose \( T_{t ightarrow s} \) from source image to target image is obtained by the pose network. \( I_t \) denotes the target frame \( I_t \) in the time \( t \). We use the two frames temporally adjacent to \( I_t \) as our source frames, i.e. \( I_s \in \{ I_{t-1}, I_{t+1} \} \). Besides, we also use \( I_s \) as the source frame \( I_s \) from the right camera to constrain the left-right consistency. In this case, \( T_{t ightarrow s} \) in Equation (9) is known from stereo cameras. Then, we can synthesize the target view by:

\[
p_s = KT_{t ightarrow s}D_tK^{-1}p_t
\]

\[
\hat{I}_t = warp(I_s, p_s)
\]

where \( D_t \) is the depth prediction result in \( I_t \) and \( K \) denotes the camera intrinsics. \( p_t \) is the homogeneous coordinates of the pixel in \( I_t \), \( p_s \) is the transformed coordinates of \( p_t \) by \( T_{t ightarrow s} \). After sampling operator \( warp() \), we get the synthesis view. Next, we utilize L1 and SSIM [18] to make the loss:

\[
pe(I_t, \hat{I}_t) = (1 - \alpha) \| I_t - \hat{I}_t \|_1 + \alpha \frac{1 - SSIM(I_t, \hat{I}_t)}{2}
\]

where \( \alpha \) is set to 0.85. We also apply a binary mask \( \mu \) followed [7] to mask the pixels that violate camera motion assumptions. The mask is produced by comparing the photometric error between \( I_s \) and \( I_t \) with that between \( \hat{I}_t \) and \( I_s \):

\[
\mu = [\min_s pe(I_t, \hat{I}_t) < \min_s (I_s, I_s)]
\]

The view synthesis loss is calculated by:

\[
L_p = \frac{1}{N} \sum_{n=1}^{N} \min(p(n) \times pe(I_t, I_t)(n))
\]

where \( N \) is the number of different scales.

2) Depth Hints Loss: To help the depth network converge to a global optimum, we follow the method in [5] to incorporate depth hints loss. The hints \( D^h \) are generated from the Semi-Global Matching (SGM) algorithm [19]. It performs pixel-by-pixel matching of stereo images and uses mutual information as the cost volume. Only when the current depth prediction from the depth network is inaccurate, the depth hints loss acts as a penalty to guide the network to jump out of the local optimum. The depth hints loss is formulated for pixel \( i \) as:

\[
L_h = \begin{cases} 
\log(|h_i - d_i| + 1) & \text{if} \ L_p(I_t, \hat{I}_t)i < L_p(I_t, \hat{I}_t)i \\
0 & \text{otherwise}
\end{cases}
\]

where \( \hat{I}_t \) denotes the reconstructed view with hint \( h \). \( \hat{I} \) represents the reconstructed view with depth network’s output. The total loss is defined as follows:

\[
L_{total} = L_p + L_h
\]

III. Experiments and results

A. Datasets

Our model is trained on KITTI datasets [26], which has been widely used for depth estimation benchmarks. We adopt the Eigen split [27] of the dataset for distance to 80 meters and use pre-processing to remove static frames before training like [6]. Thus, 39810 and 4424 images are used for training and validation, and 697 images are used for evaluation. We also test our trained model in Make3D datasets [28] to evaluate the generalization of our model.

B. Implement Details

We implement our models on a single Nvidia Titan X GPU based on PyTorch. The global-aware attention depth network and pose network are jointly trained for 20 epochs using Adam Optimizer [29] with \( \beta_1 = 0.9, \beta_2 = 0.999 \). No pre-trained weights are used for the depth network. The learning rate is set to \( 10^{-4} \) and decay to \( 10^{-5} \) in the last 5 epochs. We resize the input images to \( 640 \times 192 \) and with a batch size of 12. The total parameters of GlobalDepth are 95.74M (25.36M except for the semantic segmentation network), with 73 frames per second (fps) during inference time.
The quantitative results on KITTI datasets compared with other recent models are presented in Table I. Best results are in bold and second best are underlined. For error evaluating metrics Absolute Relative error (Abs Rel), Squared Relative error (SqRel), Root Mean Squared Error (RMSE), Root Mean Squared logarithmic error (RMSElog), the lower they are, the better performance the model achieves. For accuracy evaluating metrics $\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$, higher is better. The results show that our model outperforms most unsupervised monocular depth estimation methods. The Abs Rel of the proposed method is 0.113, which reduced the absolute relative error by 10.32% compared to MonaDA [25]. To verify the effectiveness of the proposed model, the results of the ablation experiment are shown in Table II.

We also present some qualitative samples in Fig. 4. The results show that our network can predict the edge of the object accurately, since the SFF module realizes the tightly-coupled features in element wise. In addition, our method also behaves robustly in predicting far objects such as the sky and cars in the lane, effectively alleviating the phenomenon of depth artifacts.

### C. KITTI Results

![Qualitative results on the KITTI dataset. GlobalDepth performs better on the objects boundaries such as signs and trees. The depth of far objects including sky and cars is further improved by our model.](image)

Fig. 4. Qualitative results on the KITTI dataset. GlobalDepth performs better on the objects boundaries such as signs and trees. The depth of far objects including sky and cars is further improved by our model.

### D. Make3D Results

We evaluate on the Make3D datasets after trained on the KITTI datasets to test the generalization of our model. In Table III, The type named Depth means supervised methods, and No refers to the unsupervised methods. The results show our method is comparable to supervised methods and surpass all listed unsupervised methods.

### IV. Conclusion

In this paper, we propose a novel global-aware attention model for unsupervised depth estimation. The design of the Global Feature Extraction considers the correlation among different channels and aggregates different region responses. Thus, it helps the depth network to capture a wide range of depth relations through channel attention. The Selective Feature Fusion module fuses the high-level and low-level features to obtain the most useful features and achieve a better estimation of object edge. It enables the network to conduct a weights allocation between the encoder feature and the decoder feature in element wise. The experimental results show that our model solves the problem of artifacts of the current monocular depth estimation framework in the remote region and the edge of the object, and achieves SOTA performance on the KITTI and Make3D datasets.