A Global Occlusion-Aware Approach to Self-Supervised Monocular Visual Odometry

Yao Lu[1,2], **Xiaoli Xu**[1,2], **Mingyu Ding**[3], **Zhiwu Lu**[1,2*], **Tao Xiang**[4]

[1] Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China
[2] Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing 100872, China
[3] The University of Hong Kong, Pokfulam, Hong Kong, China
[4] University of Surrey, Guildford, Surrey GU2 7XH, United Kingdom
luyao777@ruc.edu.cn, luzhiwu@ruc.edu.cn, t.xiang@surrey.ac.uk

Abstract

Self-Supervised monocular visual odometry (VO) is often cast into a view synthesis problem based on depth and camera pose estimation. One of the key challenges is to accurately and robustly estimate depth with occlusions and moving objects in the scene. Existing methods simply detect and mask out regions of occlusions locally by several convolutional layers, and then perform only partial view synthesis in the rest of the image. However, occlusion and moving object detection is an unsolved problem itself which requires global layout information. Inaccurate detection inevitably results in incorrect depth as well as pose estimation. In this work, instead of locally detecting and masking out occlusions and moving objects, we propose to alleviate their negative effects on monocular VO implicitly but more effectively from two global perspectives. First, a multi-scale non-local attention module, consisting of both intra-stage augmented attention and cascaded across-stage attention, is proposed for robust depth estimation given occlusions, alleviating the impacts of occlusions via global attention modeling. Second, adversarial learning is introduced in view synthesis for monocular VO. Unlike existing methods that use pixel-level losses on the quality of synthesized views, we enforce the synthetic view to be indistinguishable from the real one at the scene-level. Such a global constraint again helps cope with occluded and moving regions. Extensive experiments on the KITTI dataset show that our approach achieves new state-of-the-art in both pose estimation and depth recovery.

Introduction

Visual odometry (VO) aims to estimate the relative camera poses from image sequences. It has found applications in a variety of computer vision fields including autonomous driving (Engel, Stückler, and Cremers 2015; Engel, Schöps, and Cremers 2014; Mur-Artal, Montiel, and Tardos 2015), augmented reality (Yang et al. 2013; Davison et al. 2007), and interactive collaborative robotics (Geiger, Ziegler, and Stiller 2011). Early geometric-based approaches (Engel, Koltun, and Cremers 2018; Leutenegger et al. 2015; Liu et al. 2018; Engel, Sturm, and Cremers 2013) exploit artificially designed rigid transformation for VO. This results in sub-optimal performance when dealing with low texture, complex scene structures and occlusions. Moreover, they suffer from the problem of large-scale drift and the need for hand-crafted feature design. Recent methods (Xue et al. 2019; Henriques and Vedaldi 2018; Clark et al. 2017; Wang et al. 2017) utilize deep neural networks (DNNs) to solve the VO problem, because they are able to learn scale priors from large amounts of data and jointly learn the optimal feature representation. However, data annotation for VO is very costly, which limits the scalability of these supervised methods. As a result, self-supervised monocular VO (SSM-VO) has attracted increasing attention lately.

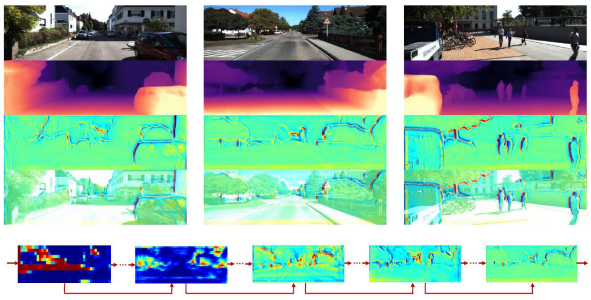

Figure 1: Example depth estimation obtained using our model. Top tow to bottom row: sample image, estimated depth, single-channel attention map, overlay of the sample image and attention map, attention maps extracted by our cascaded across-stage attention module at each stage of the depth decoder. Since the car is driving forward, most of the occlusions occur at the edge of the object.

SSM-VO is often cast into a view synthesis problem based on solving two closely intertwined problems: monocular depth estimation and relative camera pose regression. The key challenge faced by an SSM-VO method is the presence of occlusions and moving objects in the scene (Figure 1). Existing SSM-VO approaches (Bian et al. 2019; Luo et al. 2018) typically leverage multiple frames and additional models (e.g., optical flow models) to estimate an occlusion mask. With the mask, partial view synthesis is performed by excluding the masked regions. However, occlusion and moving object detection itself is an unsolved problem. Inaccurate detection inevitably results in incorrect depth as well as pose estimation. These methods thus choose to use more

*Corresponding author.

than two consecutive frames to improve the mask detection performance, giving rise to higher computational cost.

Humans often pay attention to the global layout of the scene first when detecting occlusions. Inspired by this, in this work, a novel SSM-VO approach is proposed (see Figure 2) which effectively alleviates the negative effect of occlusions and moving objects from two global perspectives. To this end, first, a new attention module is introduced for robust depth estimation in an unsupervised manner. Adopting a deep encoder-decoder architecture, most existing deep SSM-VO methods (Zhou et al. 2017; Wang et al. 2017; Bian et al. 2019; Luo et al. 2018) use skip connections to concatenate multi-level features computed over different layers. However, this ignores the correlation of attention information at different levels. In contrast, we propose a new cascaded across-stage attention module that contains both intra-stage augmented attention and cascaded across-stage attention for accurate perception and better feature extraction. With such an attention module, the depth estimation model is able to exploit global and multi-scale information to better estimate the depth in the occluded regions. Second, adversarial learning is introduced in view synthesis. Unlike existing methods that utilize pixel-level losses on the quality of synthesised views, we enforce the synthetic view to be indistinguishable from the real one at the scene-level. Such a global constraint again makes our proposed model more robust against occluded and moving regions as shown in Figure 1. With the improved capability for coping with occlusions and moving objects, our proposed model is able to estimate both depth and pose with two frames only instead of three or five required by most existing methods.

Our main contributions are: (1) A novel attention module, consisting of both intra-stage augmented attention and cascaded across-stage attention, is proposed to learn a better global feature representation for multi-scale depth estimation as well as dealing with the problem of occlusion. (2) We also introduce an adversarial-learning based framework to enforce a global constraint on view synthesis so that the detrimental effect of occlusion and moving objects can be mitigated effectively. (3) Extensive experiments demonstrate that our proposed approach achieves new state-of-the-art in both unsupervised depth estimation and pose estimation on the KITTI dataset (Geiger et al. 2013).

Related Work

Self-Supervised VO

Existing self-supervised monocular VO (SSM-VO) methods differ mostly in how the correlation between consecutive frames in the video sequence can be utilized as a supervision signal. Zhou et al. (Zhou et al. 2017) leveraged the geometric correlation of monocular depth and camera pose based on warping nearby views to the target. GeoNet (Yin and Shi 2018) further proposed an adaptive geometric consistency check to improve the robustness of depth estimation. GANVO (Feng and Gu 2019) formulated unsupervised VO as a generative learning problem (Goodfellow et al. 2014) with a depth generator and a discriminator. Based on GANVO (Feng and Gu 2019), (Li et al. 2019) further proposed to learn a compact representation of frame-to-frame correlation, which is updated by incorporating sequential information (with LSTM). Similar geometric constraints were also used in (Li et al. 2018; Almalioglu et al. 2019; Iyer et al. 2018). However, the above methods simply estimate and mask out regions of occlusions locally by several convolutional layers without any global information, resulting in sub-optimal performance. This problem is exacerbated by using only pixel-level losses. Both issues are resolved in our model from two global perspectives.

Unsupervised Monocular Depth Estimation

Many existing SSM-VO methods (Zhou et al. 2017; Zhan et al. 2018; Yin and Shi 2018; Feng and Gu 2019; Li et al. 2019) rely on accurate depth estimation, which sometimes is studied as a standalone problem (Garg et al. 2016; Godard, Mac Aodha, and Brostow 2017; Godard et al. 2019). (Garg et al. 2016) is the first end-to-end unsupervised depth estimation model based on a photometric consistency constraint. MonoDepth (Godard, Mac Aodha, and Brostow 2017) replaced the use of explicit depth data during training with easier-to-obtain binocular stereo footage. It exploited epipolar geometry constraints and generated disparity images by training the network with an image reconstruction loss. Improved upon (Godard, Mac Aodha, and Brostow 2017), Godard et al. (Godard et al. 2019) proposed to upsample depth predictions at different scales into the input resolution and then minimize the photometric reprojection errors to reduce visual artifacts, significantly improving the quality of depth prediction. Recently, (Johnston and Carneiro 2020) proposed a self-attention module to explore non-contiguous region features for better depth estimation. However, all existing methods use skip connections to concatenate multi-level features, ignoring the correlation of attention information at different levels. Importantly, without modeling multi-level attention, they lack the tool to deal with difficult regions for depth estimation due to occlusions and moving objects. In our work, a novel cascaded across-stage attention mechanism is introduced for better feature extraction and more robust depth estimation.

Methodology

Robust Depth Estimation

As shown in Figure 2, our proposed model consists of a depth generator, an ego-motion generator, and a discriminator. The depth generator is used to estimate the depth map of a target frame. The performance of pose prediction relies on accurate depth estimation.

We adopt a multi-stage encoder-decoder architecture in our depth generator G_d to generate multi-scale depth prediction. We use ResNet (He et al. 2016) as our encoder E_d, and a convolutional neural network with 5 layers as our decoder D_d. We denote the feature maps from encoder as $\mathbf{f}_e$ and the feature maps from decoder as $\mathbf{f}_d$, and they have the same spatial scale on the corresponding stage. The depth encoder E_d encodes the input image I_t at time t and outputs a depth latent feature $\mathbf{z}_d$, i.e., $E_d(I_t) = \mathbf{z}_d$. After the encoding phase, the depth latent feature $\mathbf{z}_d$ is decoded into a depth

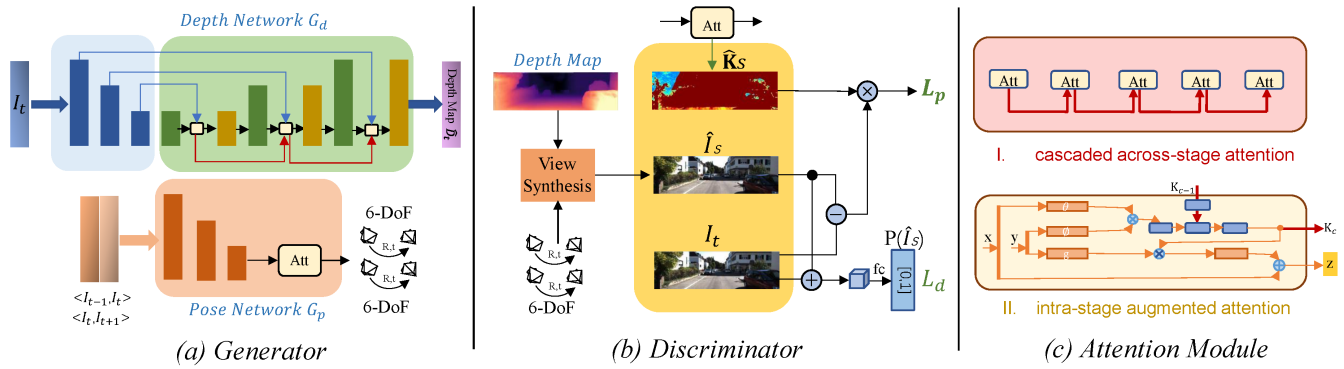

Figure 2: Illustration of our framework for self-supervised monocular VO. The depth generator (G_d) extracts the features of the target image with a convolutional neural network (CNN) and decodes the features with an attention module to generate the multi-scale depth predictions. The pose generator (G_p) estimates the camera pose and combines it with the multi-scale depth maps for view synthesis, which is guided by multi-scale soft occlusion maps $\hat{\mathbf{K}}_s$ estimated correspondingly from our non-local attention module. The synthesised image and real target image are then fed into the discriminator (D) for authenticity evaluating. Our attention module includes: (I) cascaded across-stage attention; (II) intra-stage augmented attention.

map space $\hat{D}_t$ using $D_d(\mathbf{z}_d) = \hat{D}_t$. The entire process of generating a depth map in the depth generator is given by

$$G_d(I_t) = D_d(\mathbf{z}_d) = D_d(E_d(I_t)) = \hat{D}_t. \quad (1)$$

Intra-Stage Augmented Attention. Inspired by the non-local networks (Wang et al. 2018b; Buades, Coll, and Morel 2005), we design our intra-stage augmented attention in our depth generator G_d, as shown in Figure 3. The intra-stage augmented attention bridges high-level and low-level features in each stage and captures long-range dependencies between the feature maps extracted from E_d and D_d, to provide context for dealing with occlusion and misalignment. Concretely, in our intra-stage attention, a non-local operation is defined on decoder feature maps $\mathbf{f}_d$ and encoder feature maps $\mathbf{f}_e$ as follows:

$$k_i = \frac{1}{\mathcal{C}(x,y)} \sum_{\forall j} f(x_i, y_j) g(y_j), \quad (2)$$

where x_i is the element of $\mathbf{f}_d$, and k_i is the element of our reconstructed feature maps (of the same size as $\mathbf{f}_d$), whose response is to be computed with all possible y_j in $\mathbf{f}_e$. The pairwise function $f(\cdot,\cdot)$ computes a scalar between x_i and y_j, which represents the attention score between position i in decoder feature maps $\mathbf{f}_d$ and position j in encoder feature maps $\mathbf{f}_e$. The unary function $g(\cdot)$ computes a representation in an embedded space of the encoder feature maps $\mathbf{f}_e$ at the position j. The response is normalized by a factor $\mathcal{C}(x,y)$.

There are many functions that can be used to define f, including Gaussian, embedded Gaussian, and dot product. In this work, we choose the embedded Gaussian for f

$$f(x_i, y_j) = e^{\theta(x_i)^T \phi(y_j)}, \quad (3)$$

where $\theta(x_i) = W_\theta x_i$ and $\phi(y_j) = W_\phi y_j$ are two embeddings in our module. The first is for our decoder feature maps $\mathbf{f}_d$, and the second is for our encoder feature maps $\mathbf{f}_e$. We set $\mathcal{C}$ as a softmax operation and have the self-attention:

$$k_i = \sum_{\forall j} \frac{e^{\theta(x_i)^T \phi(y_j)}}{\sum_{\forall i} e^{\theta(x_i)^T \phi(y_j)}} g(y_j). \quad (4)$$

The simplified representation for the above self-attention is:

$$\mathbf{k} = \text{softmax}(\mathbf{x}^T W_\theta^T W_\phi \mathbf{y}) g(\mathbf{y}) = \mathbf{K} g(\mathbf{y}). \quad (5)$$

Our intra-stage augmented attention block is defined as:

$$z_i = W_z k_i + x_i, \quad (6)$$

where k_i is given by Eq. (2), W_z is a weight matrix to be learned, and '$+x_i$' denotes a residual connection.

Cascaded Across-Stage Attention. Now an intra-stage non-local block has been introduced to each stage of the decoder, but there is still no connection among multiple non-local attention blocks. To overcome this drawback, we thus design cross-stage attention, as shown in Figure 2. We define the previous self-attention map (in the former stage) as $\mathbf{K}_{c-1}$. According to Eq. (5), the intermediate self-attention map of the current block is

$$\mathbf{K} = \text{softmax}(\mathbf{x}^T W_\theta^T W_\phi \mathbf{y}). \quad (7)$$

For our cascaded across-stage attention module, we combine the previous self-attention maps $\mathbf{K}_{c-1}$ and the intermediate self-attention maps $\mathbf{K}$ to produce the current self-attention maps $\mathbf{K}_c$. We first upsample the $\mathbf{K}_{c-1}$ to the same size as the $\mathbf{K}$ and concatenate them on channels. Then, we perform dimensionality reduction through a fully-connected (FC) layer with sigmoid as follows:

$$\mathbf{K}_c = f_{sig}([\mathbf{K}, u(\mathbf{K}_{c-1})]), \quad (8)$$

where u is the upsampling operation, $[\cdot,\cdot]$ denotes the concatenation, and f_{sig} is the FC layer with sigmoid.

Robust Pose Estimation

As the image sequence $I =< I_{t-1}, I_t, I_{t+1} >$ is given to the pose generator as input, we choose I_t as the target view and $I_s =< I_{t-1}, I_{t+1} >$ as the source view. The pose generator is used to regress the relative pose $\mathbf{p} \in \mathbf{SE}(3)$ which is introduced by motion and temporal dynamics across frames. The image sequence is split into two pairs $< I_{t-1}, I_t >$ and

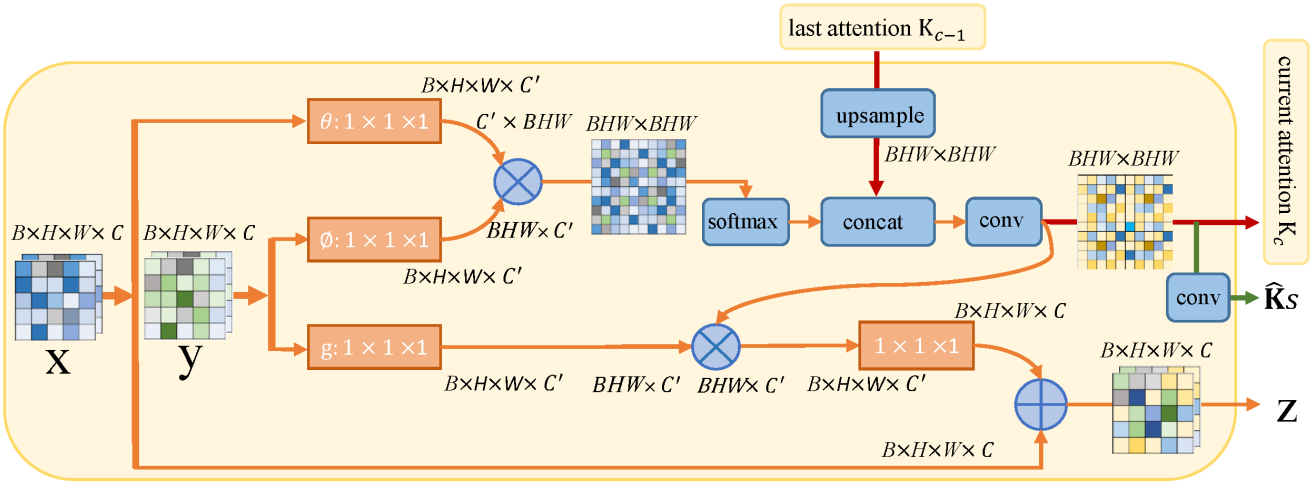

Figure 3: Details of our intra-stage augmented attention. **x**/**y** denotes the decoding/encoding feature map $\mathbf{f}_d$/ $\mathbf{f}_e$, and **z** denotes the reconstructed feature map. The feature maps are shown as the shape of their tensors, e.g., $B \times H \times W \times C$ for C channels (proper reshaping is performed when noted) in batch size B. "$\otimes$" denotes matrix multiplication, and "$\oplus$" denotes element-wise sum. The softmax operation is performed on each row. The orange boxes denote $1 \times 1 \times 1$ convolutions.

$< I_t, I_{t+1} >$, and we denote them as I_p. The pose generator takes the image pair I_p as input, and outputs 6-DoF pose values consisting of translation and rotation parameters. Following (Godard et al. 2019), we use a separate network to encode the image sequences for pose estimation. Given the pose encoder E_p, we can obtain the pose latent feature vector $\mathbf{z}_p$ as $\mathbf{z}_p = E_p(I)$. The relative camera pose $\hat{T}_{t\to s}$ is generated according to the pose latent feature vector $\mathbf{z}_p$ using the pose decoder D_p as

$$G_p(I_p) = D_p(\mathbf{z}_p) = D_p(E_p(I_p)) = \hat{T}_{t\to s}. \quad (9)$$

In our pose generator, a vanilla self-attention module is also used to select the effective feature, and the re-weighted pose feature vector is used for pose estimation:

$$\hat{T}_{t\to s} = f_{sig}(\mathbf{att}(\mathbf{z}_p)) = f_{sig}(\mathbf{z}_p \odot f_{sig}(\mathbf{z}_p)), \quad (10)$$

where $\mathbf{att}$ means the vanilla self-attention module and $\odot$ denotes the dot product.

Adversarial Learning

To simplify the notations, we use G to represent the two decoders, depth decoder D_d and pose decoder D_p, and use $\mathbf{z}$ to represent the two features, depth latent feature $\mathbf{z}_d$ and pose latent feature $\mathbf{z}_p$. The view synthesis process $S(\cdot)$ can be defined as follows:

$$S(\mathbf{z}) = S(G(\mathbf{z})) = S(D_d(\mathbf{z}_d), D_p(\mathbf{z}_p)). \quad (11)$$

The rendering of the reconstructed view $\hat{I}_s$ is based on the estimated depth map $\hat{D}_t$ from the depth generator, the 4×4 camera transformation matrix $\hat{T}_{t\to s}$ from the pose generator, and the source view I_s as in (Fehn 2004). Let the homogeneous coordinates of a pixel in the target view be p_t, and the camera intrinsic matrix be K. The coordinates of p_t are projected onto the source view p_s as

$$p_s \sim K\hat{T}_{t\to s}\hat{D}_t(p_t)K^{-1}p_t, \quad (12)$$

where the coordinates of p_s take continuous values. We utilize differentiable bilinear interpolation (Zhou et al. 2016) that has 4-pixel neighbours of p_s to approximate $I_s(p_s)$:

$$\hat{I}_s(p_t) = I_s(p_s) = \sum_{i\in\{top,bottom\},j\in\{left,right\}} w^{ij} I_s(p_s^{ij}), \quad (13)$$

where w_{ij} is the proximity value between the projected pixel p_s and its neighbouring pixels p_s^{ij} ($\sum_{i,j} w_{ij} = 1$).

After view reconstruction, the view discriminator D is trained to discriminate the reconstructed image from the real image. Following (Arjovsky, Chintala, and Bottou 2017) for easy convergence, we use the Wasserstein distance in the adversarial learning. Moreover, we simply remove the sigmoid in the view discriminator. After the weights w are updated, they are clipped into a stable range. For the discriminator, the network is trained by optimizing the loss function L_d:

$$L_d = \mathbf{E}_{\mathbf{I}\sim p_{data}(\mathbf{I})}[D(\mathbf{I})] - \mathbf{E}_{\mathbf{z}\sim p(\mathbf{z})}[D(G(\mathbf{z})], \quad (14)$$

where $\mathbf{I}$ is sampled from the data distribution p_{data}.

Full Learning Objectives

To better mitigate the effect of occlusions and moving objects, we generate multi-scale pixel-level soft masks $\{\hat{\mathbf{K}}_s\}$ from our intra-stage augmented attention blocks, indicating the belief of the network in where the synthesized view and the real target view can match. In each stage of intra-stage augmented attention, $\hat{\mathbf{K}}_s$ is generated from $\mathbf{K}_c$ through a convolutional layer. Note that our occlusion map $\hat{\mathbf{K}}_s$, as shown in Figure 5, benefits from our non-local attention, which in turn helps to learn better attention at each stage of the network. The view synthesis objective is weighted correspondingly by minimizing the following photometric loss:

$$L_p = \sum_{<I_1,I_2,\ldots,I_N>} \sum_p \hat{\mathbf{K}}_s ||I_t(p) - \hat{I}_s(p)||_1, \quad (15)$$

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Zhou et al.* (Zhou et al. 2017)	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Yang et al. (Yang et al. 2018b)	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Mahjourian et al. (Mahjourian et al. 2018)	0.163	1.240	6.220	0.250	0.762	0.916	0.968
LEGO (Yang et al. 2018a)	0.162	1.352	6.276	0.252	-	-	-
DDVO (Wang et al. 2018a)	0.151	1.257	5.583	0.228	0.810	0.936	0.971
GANVO (Almalioglu et al. 2019)	0.150	1.141	5.448	0.216	0.808	0.939	0.975
DF-Net (Zou, Luo, and Huang 2018)	0.150	1.124	5.507	0.223	0.806	0.933	0.973
Li et al. (Li et al. 2019)	0.150	1.127	5.564	0.229	0.823	0.936	0.974
GeoNet* (Yin and Shi 2018)	0.149	1.060	5.567	0.226	0.796	0.935	0.975
Ranjan et al. (Ranjan et al. 2019)	0.148	1.149	5.464	0.226	0.815	0.935	0.973
EPC++ (Luo et al. 2018)	0.141	1.029	5.350	0.216	0.816	0.941	0.976
S2D'(M)' (Casser et al. 2019)	0.141	1.026	5.291	0.215	0.816	0.945	0.979
CC (Ranjan et al. 2019)	0.140	1.070	5.326	0.217	0.826	0.941	0.975
Bian et al. (Bian et al. 2019)	0.137	1.089	5.439	0.217	0.830	0.942	0.975
SGDepth (Klingner et al. 2020)	0.117	0.907	4.844	0.196	0.875	0.958	0.980
Monodepth2 (Godard et al. 2019)	0.115	0.903	4.863	0.193	0.877	0.959	0.981
DDV (ResNet18) (Johnston and Carneiro 2020)	0.111	0.941	**4.817**	0.189	**0.885**	0.961	0.981
Ours	**0.109**	**0.883**	4.827	**0.188**	0.881	**0.962**	**0.986**

Table 1: Comparative results for depth estimation on the KITTI raw dataset. * denotes the newer results obtained using the authors' updated implementations.

where $< I_1, I_2, ..., I_N >$ denotes the training image sequence, p is the pixel coordinate index, and $\hat{I}_s$ is the projected image of the source view I_s onto the target coordinate frame using relative pose and depth-based rendering. We choose the L_1 loss due to its robustness to outliers.

To encourage nonzero predictions for occlusion maps $\{\hat{\mathbf{K}}_s\}$ and prevent the network converge to a trivial solution, we define a regularization term $L_{reg}(\hat{\mathbf{K}}_s)$ by minimizing the cross-entropy loss with constant label 1 at each pixel location, as in (Zhou et al. 2017). The regularization term is:

$$L_{reg}(\hat{\mathbf{K}}_s) = \mathrm{BCE}(\hat{\mathbf{K}}_s, \mathbf{1}), \tag{16}$$

where $\mathbf{1}$ represents a tensor of all ones with the same dimension as $\hat{\mathbf{K}}_s$, and BCE is the binary cross-entropy between the soft mask $\hat{\mathbf{K}}_s$ and $\mathbf{1}$.

The overall appearance loss of the reconstructed image L_a is measured by both weighted photometric loss and structural similarity metric (SSIM) (Wang et al. 2004):

$$\begin{aligned} L_a =& L_{reg}(\hat{\mathbf{K}}_s) + (1-\alpha)L_p \\ &+ \frac{\alpha}{2}(1 - \mathrm{SSIM}(I_t(p), \hat{I}_s(p))), \end{aligned} \tag{17}$$

where the hyper-parameter α is set to 0.85 in this work.

As the photometric loss is not informative in the low-texture or homogeneous region of the scene, we incorporate an edge-aware smoothness prior (Yin and Shi 2018) to regularize the estimated depth map:

$$L_e = \sum_{p_t} |\nabla D(p_t)| \cdot (e^{|\nabla \mathbf{I}(p_t)|})^T, \tag{18}$$

where $|\cdot|$ denotes the element-wise absolute operator, ∇ is the vector differential operator, and T denotes the transpose of image gradient weighting.

The full learning objectives for self-supervised monocular VO (SSM-VO) are defined as follows:

$$L_{final} = \sum_l L_d^l + w_1 L_a^l + w_2 L_e^l, \tag{19}$$

where l indexes over different image scales and w_1, w_2 are the weighting hyper-parameters balancing these losses.

Experiments

Implementation Details

KITTI Dataset. For single-view depth estimation, we select the popular KITTI raw dataset (Geiger et al. 2013) with the Eigen (Eigen, Puhrsch, and Fergus 2014) split and the pre-processing method in (Zhou et al. 2017) to remove static frames, as in (Yin and Shi 2018; Zou, Luo, and Huang 2018; Ranjan et al. 2019). This provides 39,810 monocular triplets for training and 4,424 for the test. Moreover, for pose estimation, the KITTI odometry dataset (Geiger, Lenz, and Urtasun 2012) is used for performance evaluation. Following (Zhan et al. 2018), sequences 00-08 and 09-10 are used for training and test, respectively.

Network Architecture. The total deep learning framework is implemented on PyTorch (Paszke et al. 2019). For the depth generator network, we use ResNet18 (He et al. 2016) as the encoder backbone of the depth generator. The decoder of the depth generator has sigmoids at the output and ELU nonlinearities (Clevert, Unterthiner, and Hochreiter 2015) elsewhere. We also scale the sigmoid output $\hat{D}_t$ of the depth decoder between 0.1 and 100 units. In addition, for the pose generator network, the backbone ResNet18 is modified as in (Godard et al. 2019).

Single-View Depth Estimation. In the training phase, we use a snippet of three sequential video frames as a training sample, where we set the middle image as the reference

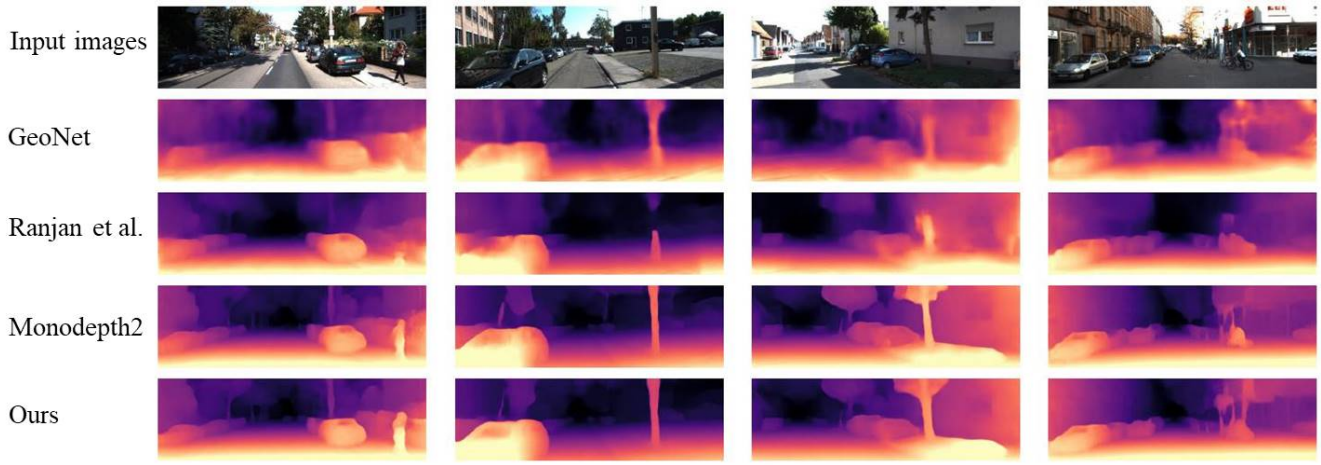

Figure 4: Qualitative results obtained by various methods for depth estimation on the KITTI Raw Dataset. It is evident that the estimated depth using our method preserves more accurate boundaries between objects with finer details.

Figure 5: Visualization for our occlusion maps. The ego-motion is stationary in the first two columns while dynamic in the last column. For stationary scenarios, only moving objects are detected. In dynamic scenarios, occlusions (e.g., caused by traffic signs), moving objects, and truncation are detected at the same time. Our occlusion map $\hat{\mathbf{K}}_s$ helps to remove ambiguous areas in view synthesis, thereby improving both the depth prediction and relative pose estimation.

frame to compute the loss, and the other as two for view synthesis. In the test phase, we cap the depth to 80m per standard practice (Godard, Mac Aodha, and Brostow 2017). We report the results using the per-image median ground-truth scaling (Zhou et al. 2017). For data augmentation, we select horizontal flips, random crop (at a random ratio of 0.7 to 1.0 of the original image size, ensuring the original aspect ratio for the training samples), and the following strategies (with 50% chance): random brightness, contrast, saturation, and hue jitter with respective ranges of ± 0.2, ± 0.2, ± 0.2, and ± 0.1. Importantly, the color augmentations are only applied to the images which are fed to the networks, not to those used to compute the photometric loss L_p. We set $w_1 = 0.1$ and $w_2 = 0.5$ in Eq. (19). Our model is trained with a Titan XP GPU for 60 epochs using the Adam optimizer. We take a learning rate of 10^{-4} for the first 35 epochs and reduce it to 10^{-5} for the remainder. We set the batch size to 12 and the input/output resolution to 640×192 unless otherwise specified. We employ Eigen et al.'s evaluation metrics (Eigen, Puhrsch, and Fergus 2014) for depth evaluation.

Pose Estimation. We trained our models on sequences 0-8 from the KITTI odometry split and tested on sequences 9 and 10. The performance of pose estimation is evaluated using Absolute Trajectory Error (ATE) for both translation and rotation. For trajectory evaluation, we choose the evaluation metrics as in (Zhou et al. 2017; Yin and Shi 2018).

Method	Frames	Sequence 09	Sequence 10
ORB-SLAM	-	0.014 ± 0.008	0.012 ± 0.011
(Zhou et al. 2017)*	5	0.016 ± 0.009	0.013 ± 0.015
GeoNet (Yin and Shi 2018)*	5	0.012 ± 0.007	0.012 ± 0.009
(Ranjan et al. 2019)	5	0.012 ± 0.007	0.012 ± 0.008
GANVO (Feng and Gu 2019)	5	0.009 ± 0.005	0.010 ± 0.013
DDVO (Wang et al. 2018a)	3	0.045 ± 0.018	0.033 ± 0.074
SGANVO (Feng and Gu 2019)	3	0.015 ± 0.006	0.014 ± 0.009
(Mahjourian et al. 2018)	3	0.013 ± 0.010	0.012 ± 0.011
EPC++ (Luo et al. 2018)	3	0.013 ± 0.007	0.012 ± 0.008
Monodepth2	2	0.017 ± 0.008	0.015 ± 0.010
Bian et al. (Bian et al. 2019)	2	0.016 ± 0.007	0.015 ± 0.015
Ours	**2**	$\mathbf{0.013 \pm 0.006}$	$\mathbf{0.012 \pm 0.006}$

Table 2: Comparative results on the KITTI odometry dataset. The absolute trajectory error (ATE) is used as the evaluation metric. * denotes the authors' updated results.

Comparison with the State-of-the-Art

Depth Estimation Results on KITTI Raw Dataset. Table 1 compares our proposed model with the existing self-supervised depth estimation methods. Note that most of the compared methods are based on deep learning and some results are updated according to the authors' updated implementation. The comparative results in Table 1 show that our depth generator outperforms all self-supervised methods and achieves new state-of-the-art on this dataset. Particularly, with the same backbone, our model beats the latest and strongest competitor (Johnston and Carneiro 2020) on five out of seven metrics. The qualitative results for various depth estimation methods on the KITTI Eigen split are shown in Figure 4. It is evident that the estimated depth using our method preserves more accurate boundaries between

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Baseline (BL)	0.163	1.253	6.232	0.263	0.774	0.923	0.965
BL+ISAA (inner)	0.153	1.248	5.593	0.237	0.792	0.939	0.970
BL+ISAA (stage-wise)	0.147	1.142	5.412	0.223	0.816	0.941	0.975
BL+ISAA (non-local)	0.140	1.072	5.339	0.216	0.827	0.946	0.975
BL+ISAA (ours)	0.134	1.065	5.152	0.211	0.834	0.948	0.976
BL+ISAA+CASA	0.119	0.925	4.963	0.197	0.862	0.955	0.981
BL+ISAA+CASA+GC	**0.109**	**0.883**	**4.827**	**0.188**	**0.881**	**0.962**	**0.986**

Table 3: Ablation study for our proposed model. Notations: ISAA – intra-stage augmented attention; CASA – cascaded across-stage attention; GC – global constraint. ISAA (inner), ISAA (stage-wise), and ISAA (non-local) are alternatives of ISAA.

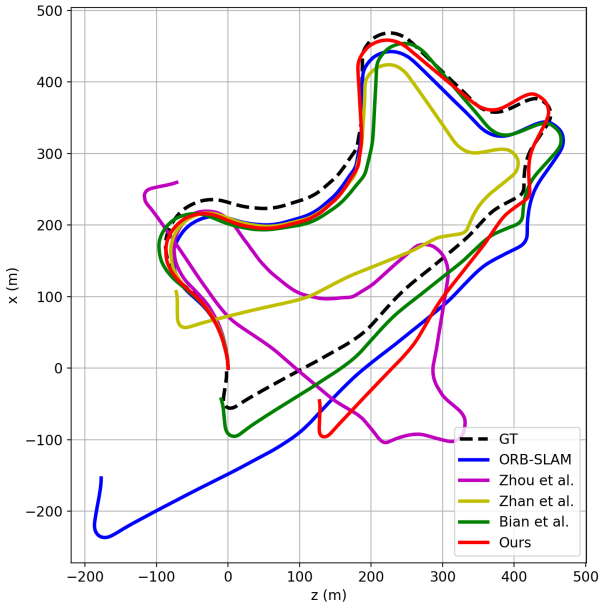

Figure 6: Qualitative results obtained by our proposed model for visual odometry on Sequence 09.

objects with finer details. This is mainly due to the proposed attention module which enables our model to better cope with occlusion and moving objects (see the visualization of our occlusion maps in Figure 5). Note that although these enhanced edges/details are relatively small in size and thus contribute little to various depth estimation metrics in Table 1, visual inspection on Figure 4 suggests that the improvement in depth estimation quality is significant.

Pose Estimation Results on KITTI Odometry Dataset. We compare our proposed approach with recent SSM-VO methods. We also report the results of ORB-SLAM system (without loop closing) (Mur-Artal, Montiel, and Tardos 2015) as a reference. It is worth noting that our simple frame-to-frame pose estimation framework is not expected to beat a visual SLAM system, which has a strong back-end optimization system (i.e., bundle adjustment) for improving the performance. Also note that our pose generator takes only two frames as input whilst most other methods take more frames. The odometry results on the KITTI odometry dataset are shown in Table 2. Our proposed approach clearly outperforms other methods using the same 2 input frames and achieves similar results w.r.t. the methods using more input frames (3 or 5 frames). Some qualitative results (sequence 09) for visual odometry are shown in Figure 6.

Ablation Study

Ablation Study for Our Full Model. We conduct ablation studies for our full model, which consists of three main components: intra-stage augmented attention (ISAA), cascaded across-stage attention (CASA) and global constraint (GC). Specifically, ISAA bridges high-level and low-level features in each sub-network and makes it focus more on the areas that are difficult to estimate from a global perspective; CASA enables our network to provide gradually refined attention and balance the attention map from different scales; GC is induced by adversarial learning and enforces the synthetic view to be indistinguishable from the real one to help cope with occluded and moving regions at the scene-level. Table 3 clearly demonstrates the contribution of each component to the overall performance.

Ablation Study for Our ISAA Module. Note that ISAA is crucial for feature selection during the whole train process. We compare our ISAA with three alternative implementations: (a) ISAA (inner): The self-attention, like the vanilla self-attention module in pose generator, implemented on depth latent feature $\mathbf{z}_d$ existed in the inner of the depth generator. (b) ISAA (stage-wise): The self-attention implemented on the feature maps $\mathbf{f}_d$ at each stage of the depth decoder. (c) ISAA (non-local): Traditional implementation of non-local self-attention on feature maps $\mathbf{f}_d$ at each stage of the depth decoder. From Table 3, we can see that our ISAA performs better than alternative implementations. More ablative results can be found in the suppl. material.

Conclusion

We have proposed an adversarial-learning based framework for monocular VO which learns the depth estimation and view synthesis from a global perspective. We have also introduced a novel non-local attention module, consisting of both intra-stage augmented attention and cascaded across-stage attention, to learn a better feature representation for depth estimation and deal with the problem of moving objects and occlusions. With the proposed module, our global occlusion-aware approach to monocular VO achieves new state-of-the-art on the KITTI dataset.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (61976220 and 61832017), and Beijing Outstanding Young Scientist Program (BJJWZYJH012019100020098). Note that the first three authors (Yao Lu, Xiaoli Xu, and Mingyu Ding) are of equal contribution to this work.

References

Almalioglu, Y.; Saputra, M. R. U.; de Gusmao, P. P.; Markham, A.; and Trigoni, N. 2019. GANVO: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, 5474–5480.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN. *arXiv preprint arXiv:1701.07875* .

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. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 35–45.

Buades, A.; Coll, B.; and Morel, J.-M. 2005. A non-local algorithm for image denoising. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 60–65.

Casser, V.; Pirk, S.; Mahjourian, R.; and Angelova, A. 2019. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *AAAI Conference on Artificial Intelligence (AAAI)*, 8001–8008.

Clark, R.; Wang, S.; Wen, H.; Markham, A.; and Trigoni, N. 2017. Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Clevert, D.-A.; Unterthiner, T.; and Hochreiter, S. 2015. Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv preprint arXiv:1511.07289* .

Davison, A. J.; Reid, I.; Molton, N. D.; and Stasse, O. 2007. MonoSLAM: Real-Time Single Camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(6): 1052–1067.

Eigen, D.; Puhrsch, C.; and Fergus, R. 2014. Depth map prediction from a single image using a multi-scale deep network. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2366–2374.

Engel, J.; Koltun, V.; and Cremers, D. 2018. Direct Sparse Odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(3): 611–625.

Engel, J.; Schöps, T.; and Cremers, D. 2014. LSD-SLAM: Large-scale direct monocular SLAM. In *European Conference on Computer Vision (ECCV)*, 834–849.

Engel, J.; Stückler, J.; and Cremers, D. 2015. Large-scale direct SLAM with stereo cameras. In *IROS*, 1935–1942.

Engel, J.; Sturm, J.; and Cremers, D. 2013. Semi-dense visual odometry for a monocular camera. In *IEEE International Conference on Computer Vision (ICCV)*, 1449–1456.

Fehn, C. 2004. Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. In *Stereoscopic Displays and Virtual Reality Systems XI*, 93–104.

Feng, T.; and Gu, D. 2019. SGANVO: Unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks. *IEEE Robotics and Automation Letters* 4(4): 4431–4437.

Garg, R.; BG, V. K.; Carneiro, G.; and Reid, I. 2016. Unsupervised CNNn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision (ECCV)*, 740–756.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research* 32(11): 1231–1237.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Geiger, A.; Ziegler, J.; and Stiller, C. 2011. StereoScan: Dense 3d Reconstruction in Real-time. In *IEEE Intelligent Vehicles Symposium*, 963–968.

Godard, C.; Mac Aodha, O.; and Brostow, G. J. 2017. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 270–279.

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

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2672–2680.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Henriques, J. F.; and Vedaldi, A. 2018. MapNet: An Allocentric Spatial Memory for Mapping Environments. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8476–8484.

Iyer, G.; Krishna Murthy, J.; Gupta, G.; Krishna, M.; and Paull, L. 2018. Geometric consistency for self-supervised end-to-end visual odometry. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 267–275.

Johnston, A.; and Carneiro, G. 2020. Self-supervised Monocular Trained Depth Estimation using Self-attention and Discrete Disparity Volume. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4756–4765.

Klingner, M.; Termöhlen, J.-A.; Mikolajczyk, J.; and Fingscheidt, T. 2020. Self-Supervised Monocular Depth Esti-

mation: Solving the Dynamic Object Problem by Semantic Guidance. *arXiv preprint arXiv:2007.06936* .

Leutenegger, S.; Lynen, S.; Bosse, M.; Siegwart, R.; and Furgale, P. 2015. Keyframe-based visual-inertial odometry using nonlinear optimization. *International Journal of Robotics Research* 34(3): 314–334.

Li, R.; Wang, S.; Long, Z.; and Gu, D. 2018. UnDeepVO: Monocular visual odometry through unsupervised deep learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 7286–7291.

Li, S.; Xue, F.; Wang, X.; Yan, Z.; and Zha, H. 2019. Sequential Adversarial Learning for Self-Supervised Deep Visual Odometry. In *IEEE International Conference on Computer Vision (ICCV)*, 2851–2860.

Liu, H.; Chen, M.; Zhang, G.; Bao, H.; and Bao, Y. 2018. ICE-BA: Incremental, Consistent and Efficient Bundle Adjustment for Visual-Inertial SLAM. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1974–1982.

Luo, C.; Yang, Z.; Wang, P.; Wang, Y.; Xu, W.; Nevatia, R.; and Yuille, A. 2018. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *arXiv preprint arXiv:1810.06125* .

Mur-Artal, R.; Montiel, J. M. M.; and Tardos, J. D. 2015. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics* 31(5): 1147–1163.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 8024–8035.

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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12240–12249.

Wang, C.; Miguel Buenaposada, J.; Zhu, R.; and Lucey, S. 2018a. Learning depth from monocular videos using direct methods. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022–2030.

Wang, S.; Clark, R.; Wen, H.; and Trigoni, N. 2017. DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2043–2050.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018b. Non-local neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7794–7803.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13(4): 600–612.

Xue, F.; Wang, X.; Li, S.; Wang, Q.; Wang, J.; and Zha, H. 2019. Beyond tracking: Selecting memory and refining poses for deep visual odometry. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8575–8583.

Yang, M.-D.; Chao, C.-F.; Huang, K.-S.; Lu, L.-Y.; and Chen, Y.-P. 2013. Image-based 3D scene reconstruction and exploration in augmented reality. *Automation in Construction* 33: 48–60.

Yang, Z.; Wang, P.; Wang, Y.; Xu, W.; and Nevatia, R. 2018a. LEGO: Learning edge with geometry all at once by watching videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 225–234.

Yang, Z.; Wang, P.; Xu, W.; Zhao, L.; and Nevatia, R. 2018b. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In *AAAI Conference on Artificial Intelligence (AAAI)*, 7493–7500.

Yin, Z.; and Shi, J. 2018. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1983–1992.

Zhan, H.; Garg, R.; Saroj Weerasekera, C.; Li, K.; Agarwal, H.; and Reid, I. 2018. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 340–349.

Zhou, T.; Brown, M.; Snavely, N.; and Lowe, D. G. 2017. Unsupervised learning of depth and ego-motion from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1851–1858.

Zhou, T.; Tulsiani, S.; Sun, W.; Malik, J.; and Efros, A. A. 2016. View synthesis by appearance flow. In *European Conference on Computer Vision (ECCV)*, 286–301.

Zou, Y.; Luo, Z.; and Huang, J.-B. 2018. DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency. In *European Conference on Computer Vision (ECCV)*, 36–53.