MonoIndoor++: Towards Better Practice of Self-Supervised Monocular Depth Estimation for Indoor Environments

Runze Li, *Member, IEEE*, Pan Ji, Yi Xu, and Bir Bhanu, *Life Fellow, IEEE*

***Abstract*— Self-supervised monocular depth estimation has seen significant progress in recent years, especially in outdoor environments, *i.e.*, autonomous driving scenes. However, depth prediction results are not satisfying in indoor scenes where most of the existing data are captured with hand-held devices. As compared to outdoor environments, estimating depth of monocular videos for indoor environments, using self-supervised methods, results in two additional challenges: (i) the depth range of indoor video sequences varies a lot across different frames, making it difficult for the depth network to induce consistent depth cues for training, whereas the maximum distance in outdoor scenes mostly stays the same as the camera usually sees the sky; (ii) the indoor sequences recorded with handheld devices often contain much more rotational motions, which cause difficulties for the pose network to predict accurate relative camera poses, while the motions of outdoor sequences are pre-dominantly translational, especially for street-scene driving datasets such as KITTI. In this work, we propose a novel framework-*MonoIndoor++* by giving special considerations to those challenges and consolidating a set of good practices for improving the performance of self-supervised monocular depth estimation for indoor environments. First, a depth factorization module with transformer-based scale regression network is proposed to estimate a global depth scale factor explicitly, and the predicted scale factor can indicate the maximum depth values. Second, rather than using a single-stage pose estimation strategy as in previous methods, we propose to utilize a residual pose estimation module to estimate relative camera poses across consecutive frames iteratively. Third, to incorporate extensive coordinates guidance for our residual pose estimation module, we propose to perform coordinate convolutional encoding directly over the inputs to pose networks. The proposed method is validated on a variety of benchmark indoor datasets, *i.e.*, EuRoC MAV, NYUv2, ScanNet and 7-Scenes, demonstrating the state-of-the-art performance. In addition, the effectiveness of each module is shown through a carefully conducted ablation study and the good generalization and universality of our trained model is also demonstrated, specifically on ScanNet and 7-Scenes datasets.**

***Index Terms*— Monocular depth prediction, self-supervised learning.**

Manuscript received 23 June 2022; revised 25 August 2022; accepted 4 September 2022. Date of publication 15 September 2022; date of current version 6 February 2023. This work was supported in part by Bourns Endowment Funds. This article was recommended by Associate Editor J. Chen. *(Corresponding author: Runze Li.)*

Runze Li is with the Department of Computer Science, University of California at Riverside, Riverside, CA 92521 USA (e-mail: rli047@ucr.edu).

Pan Ji was with the OPPO US Research Center, InnoPeak Technology, Inc., Palo Alto, CA 94303 USA. He is now with the XR Vision Labs, Tencent, Shanghai 201900, China (e-mail: peterji1990@gmail.com).

Yi Xu is with the OPPO US Research Center, InnoPeak Technology, Inc., Palo Alto, CA 94303 USA (e-mail: yi.xu@oppo.com).

Bir Bhanu is with the Department of Electrical and Computer Engineering, University of California at Riverside, Riverside, CA 92521 USA (e-mail: bhanu@ece.ucr.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2022.3207105.

Digital Object Identifier 10.1109/TCSVT.2022.3207105

I. Introduction

MONOCULAR depth estimation has been applied in a variety of 3D perceptual tasks, including autonomous driving, virtual reality (VR), and augmented reality (AR). Estimating the depth map plays an essential role in these applications, in helping to understand environments, plan agents' motions, reconstruct 3D scenes, etc. Existing supervised depth methods [1], [2] can achieve high performance, but they require the ground-truth depth data during the training which is often expensive and time-consuming to obtain by using depth sensors (*e.g.*, LiDAR). To this end, a number of recent work [3], [4], [5] have been focused on predicting the depth map from a single image using self-supervised manners and they have shown advantages in scenarios where obtaining the ground-truth is not possible. In these self-supervised frameworks, photometric consistency between multiple views from monocular video sequences has been utilized as the main supervision for training models. Specifically, the recent work [5] has achieved significant success in estimating depth that is comparable to that by the supervised methods [2], [6]. For instance, on the KITTI dataset [7], the Monodepth2, proposed by Godard *et al.* [5], achieves an absolute relative depth error (AbsRel) of 10.6%, which is not far from the AbsRel of 7.2% by the DORN which is a supervised model proposed by Fu *et al.* [2]. However, most of these self-supervised depth prediction methods [3], [4], [5] are only evaluated on datasets of outdoor scenes such as KITTI, leaving their performance opaque for indoor environments. There are certainly ongoing efforts [8], [9], [10] which consider self-supervised monocular depth estimation for indoor environments, but their performance still trail far behind the one on the outdoor datasets by methods such as [3], [4], and [5] or the supervised counterparts [2], [11] on indoor datasets. In this paper, we concentrate on estimating the depth map from a single image for indoor environments in a self-supervised manner which only requires monocular video sequences for training.

This paper investigates the performance discrepancies between the indoor and outdoor scenes and takes a step towards examining what makes indoor depth prediction more challenging than the outdoor case. We *first* identify that the

1051-8215 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

scene depth range of indoor video sequences varies a lot more than in the outdoor and conjecture that this posits more difficulties for the depth network in inducing consistent depth cues across images from monocular videos, resulting in the worse performance on indoor datasets.

Our *second* observation is that the pose network, which is commonly used in self-supervised methods [4], [5], tends to have large errors in predicting rotational parts of relative camera poses. A similar finding has been presented in [12] where predicted poses have much higher rotational errors (*e.g.*, 10 times larger) than geometric SLAM [13] even when they use a recurrent neural network as the backbone to model the long-term dependency for pose estimation. We argue that this problem is not prominent on outdoor datasets, *i.e.*, KITTI, because the camera motions therein are mostly translational. However, frequent cameras rotations are inevitable in indoor monocular videos [14], [15] as these datasets are often captured by hand-held cameras or Micro Aerial Vehicles (MAVs). Thus, the inaccurate rotation prediction becomes detrimental to the self-supervised training of a depth model for indoor environments.

Our *third* conjecture is that the pose network in existing self-supervised methods is potentially suffering from insufficient cues to estimate relative cameras poses between color image pairs in different views. We argue that, rather than simply inducing camera poses based on color information of image pairs, encoding coordinates information can further improve the reliability of pose network in inferring geometric relations among changing views.

We propose **MonoIndoor++**, a self-supervised monocular depth estimation method tailored for indoor environments, giving special considerations for above problems. Our MonoIndoor++ consists of three novel modules: a *depth factorization* module, a *residual pose estimation* module, and a *coordinates convolutional encoding* module. In the depth factorization module, we factorize the depth map into a global depth scale (for the target image of the current view) and a relative depth map. The depth scale factor is separately predicted by an extra module (named as transformer-based scale regression network) in parallel with the depth network which predicts a relative depth map. In such a way, the depth network has more model plasticity to adapt to the depth scale changes during training. We leverage the recent advances of transformer [16] in designing the scale regression network to predict the depth scale factor. In the residual pose estimation module, we mitigate the issue of inaccurate camera rotation prediction by performing residual pose estimation in addition to an initial large pose prediction. Such a residual approach leads to more accurate computation of the photometric loss [5], which in turn improves model training for the depth prediction. In the coordinates convolutional encoding module, we encode the coordinates information (x, y) explicitly and incorporate them with color information in the residual pose estimation module, expecting to provide additional cues for pose predictions, which further consolidates residual pose estimation model during training.

It should be mentioned that this paper is an extended version of our previous conference paper [17], where we propose an unsupervised learning framework for monocular depth estimation in indoor environments. In this paper, we **i)** add more technical details of our proposed method; **ii)** present *coordinate convolutional encoding module* in the framework for improved performance of monocular depth prediction; **iii)** make a more clear explanation of our proposed *depth factorization module* with *transformer-based scale network*; **iv)** conduct extensive experiments and ablation studies on public benchmark datasets, *i.e.*, EuRoC MAV [18], NYUv2 [14], ScanNet [19] and 7-Scenes [20], and perform detailed analysis to demonstrate the effectiveness and good generalizability of our proposed framework in this journal paper.

In summary, our contributions are:

- We propose a novel depth factorization module with a transformer-based scale regression network to estimate a global depth scale factor, which helps the depth network adapt to the rapid scale changes for indoor environments during model training.
- We propose a novel residual pose estimation module that mitigates the inaccurate camera rotation prediction issue in the pose network and in turn significantly improves monocular depth estimation performance.
- We incorporate coordinates convolutional encoding in the proposed residual pose estimation module to leverage coordinates cues in inducing relative camera poses.
- We demonstrate the state-of-the-art performance of self-supervised monocular depth prediction on a wide-variety of publicly available indoor datasets, *i.e.*, NYUv2 [14], EuRoC MAV [18], ScanNet [19] and 7-Scenes [20].

The paper is organized as follows: Section II summarizes related published works in the field; then in Section III, we explain our proposed approach for monocular depth estimation in indoor environments, the proposed approach consisting of a depth factorization, a residual pose and a coordinates convolutional encoding modules; and in Section IV, we present experimental results and ablation studies on a variety of benchmark indoor datasets; and lastly a conclusion of our work is discussed in Section V.

II. Related Work

Much effort has been expended for the depth estimation in various environments. This paper addresses the self-supervised monocular depth estimation for indoor environments. In this section, we discuss the relevant work of depth estimation using both supervised and self-supervised methods.

A. Supervised Monocular Depth Estimation

The depth estimation problem was mostly solved by using supervised methods in early research. Saxena *et al.* [21] proposed the method to regress the depth from a single image by extracting superpixel features and using a Markov Random Field (MRF). Schönberger *et al.* [22] presented a system for the joint estimation of depth and normal information with photometric and geometric priors. These methods employ traditional geometry-based methods. Eigen *et al.* [23] proposed the first deep-learning based method for monocular depth estimation using a multi-scale convolutional neural network

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

(CNN). Later on, deep-learning based methods have shown significant progress on monocular depth estimation, specifically with massive ground-truth data during training the networks. One line of following work improves the performance of depth prediction by better network architecture design. Laina *et al.* [24] proposed an end-to-end fully convolutional architecture by encompassing the residual learning to predict accurate single-view depth maps given monocular images. Bhat *et al.* [25] proposed a transformer-based architecture block to adaptively estimate depth maps using a number of bins. Another line of work achieves improved depth estimation results by integrating more sophisticated training losses [2], [11], [26], [27], [28], [29]. Besides, a few methods [30], [31] proposed to use two networks, one for depth prediction and the other for motion, to mimic geometric Structure-from-Motion (SfM) or Simultaneous Localization and Mapping (SLAM) in a supervised framework. However, ground-truth depth maps with images are used to train these methods and obtaining ground-truth data is often expensive and time-consuming to capture. Some other methods then resorted to remedy this problem by generating pseudo ground-truth depth labels with traditional 3D reconstruction methods [32], [33], such as SfM [15] and SLAM [13], [34], or 3D movies [35]. Such methods have better capacity of generalization across different datasets, but cannot necessarily achieve the best performance for the dataset at hand. Some other ongoing efforts explore to improve robustness of supervised monocular depth estimation for zero-shot cross-dataset transfer. Ranftl *et al.* [35] proposed robust scale-and shift-invariant losses for training the model using data from mixed dataset and testing on zero-shot datasets, and improved it further by integrating vision transformer in network design [36].

B. Self-Supervised Monocular Depth Estimation

Recently, significant progress has been made in self-supervised depth estimation as it does not require training with the ground-truth data. Garg *et al.* [3] proposed the first self-supervised method to train a CNN-based model for monocular depth estimation by using color consistency loss between stereo images. Zhou *et al.* [4] employed a depth network for depth estimation and a pose network to estimate relative camera poses between temporal frames, and used outputs to construct the photometric loss across temporal frames to train the model. Many follow-up methods then tried to propose new training loss terms to improve self-supervision for training models. Godard *et al.* [37] incorporated a left-right depth consistency loss for the stereo training. Bian *et al.* [38] put forth a temporal depth consistency loss to ensure predicted depth maps of neighbouring frames are consistent. Wang *et al.* [39] first observed the diminishing issue of the depth model during training and proposed a normalization method to counter this effect. Yin and Shi [40] and Zou *et al.* [41] trained three networks (*i.e.*, one depth network, one pose network, and one extra flow network) jointly by enforcing cross-task consistency between optical flow and dense depth. Wang *et al.* [42] and Zou *et al.* [12] explored techniques to improve the performance of pose network and/or the depth network by leveraging recurrent neural networks, such as LSTMs, to model long-term dependency. Tiwari *et al.* [43] designed a self-improving loop with monocular SLAM and a self-supervised depth model [5] to improve the performance of each one. Among these recent advances, Monodepth2 [5] significantly improved the performance over previous methods via a set of techniques: a per-pixel minimum photometric loss to handle occlusions, an auto-masking method to mask out static pixels, and a multi-scale depth estimation strategy to mitigate the texture-copying issue in depth. Watson *et al.* [44] proposed to use cost volume in the deep model and a new consistency loss calculated between a teacher and a student model for self-supervision training. Unlike Monodepth2, this method showed its advantages in using multiple frames during the testing. We implement our self-supervised depth estimation framework based on Monodepth2, but make important changes in designing both the depth and the pose networks.

Most of the aforementioned methods were only evaluated on outdoor datasets such as KITTI. Recent ongoing efforts [8], [9], [45] focus on self-supervised depth estimation for indoor environments. Zhou *et al.* [8] first observed existing large rotations on most existing indoor datasets, and then used a pre-processing step to handle large rotational motions by removing all the image pairs with "pure rotation" and designed an optical-flow based training paradigm using the processed data. Zhao *et al.* [9] adopted a geometry-augmented strategy that solved for the depth via two-view triangulation and then used the triangulated depth as supervision for model training. Bian *et al.* [10], [45] theoretically studied the reasons behind the unsatisfying deep estimation performance in indoor environments and argued that "the rotation behaves as noise during training". They proposed a rectification step during the data pre-processing to remove the rotation between consecutive frames and designed an auto-rectify network. We have an observation similar to [8], [45], and [10] that large rotations cause difficulties for training the network. However, we take a different strategy. Instead of removing rotations from training data during the data pre-processing, we progressively estimate camera poses in rotations and translations via a novel residual pose module in an end-to-end manner, and we validate the effectiveness of the proposed method in predicting improved depth on a variety of indoor benchmark datasets.

C. Transformer

We leverage the transformer in designing our scale regression network inspired by the recent advances [16], [46], [47] of the attention mechanism. Self-attention in the transformer was first used successfully in natural language processing (NLP) to model long-term dependencies. Wang *et al.* [46] proposed a non-local operations for computer vision tasks. Recently, self-attention and its variants have been widely used in transformer networks for high-level visions tasks such as image classification [16] and semantic segmentation [48], [49].

D. Coordinates Encoding

Convolutional neural networks (CNNs) have achieved significant success at many tasks, and it can be complemented

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

with specialized layers for certain usage. For instance, detection models like Faster R-CNN [50] make use of layers to compute coordinate transforms. Jaderberg *et al.* [51] proposed a spatial Transformer module that can be included into a standard CNN model to provide spatial transformation capabilities. Qi *et al.* [52], [53] designed the PointNet which took a set of 3D points represented as (x, y, z) coordinates as well as extra color features for 3D classification and segmentation. Recently, coordinates encoding has been widely used in vision transformers [16], [48], [49] and neural radiance fields (NERF) representations [54], [55], [56]. Vision transformers [16] take 2D images as the input, reshape the image into a sequence of flattened 2D patches and then employ self-attention blocks for image classification, detection and segmentation. Position embeddings are added to the patch embeddings as a standard processing step to retain positional information. Mildenhall *et al.* [54] proposed a method which took a 3D location (x, y, z) and 2D viewing direction (θ, ϕ) as the input for scene synthesis. Unlike in vision transformers where positional encoding is utilized to provide discrete positions of tokens in the sequence, in NERF, positional functions are used to map continuous input coordinates into a higher dimensional space for high frequency approximations. Liu *et al.* [57] defined the *CoordConv* operation to provide extra coordinates information as part of input channels to convolutional filters for the convolutional neural networks. Most of pose networks in monocular depth estimation pipelines [5], [37] simply take two consecutive frames as the input and outputs relative camera poses. We argue such designs infer rotational and translational relations by only focusing on photometric cues, but ignoring explicit coordinates cues. In our work, we leverage coordinates encoding in the proposed residual pose network.

E. Monocular Depth Estimation for the Circuits and Systems for Video Technology

Depth estimation is one of the most fundamental tasks in computer vision for the circuits and systems for video technology, and it has made great progress in recent years. Using deep learning techniques, efforts have been made to estimate dense depth maps, given input images, in a supervised manner. Cao *et al.* [29] formulated the estimation of depth as a pixelwise classification problem with a fully convolutional depth residual network. Song *et al.* [27] incorporated the idea of the Laplacian pyramid into the depth decoder. During the training, depth encoder features are fed into different streams which are predefined by the decomposition of the Laplacian pyramid for outputing the final depth map. Rather than employing a fully supervised approach for monocular depth estimation, Tian *et al.* [58] used quadtree constraint for calculating the photometric and depth loss during the training of a depth model. It leveraged the sparse depth information as a part of the input during semi-supervised training. To enable fully self-supervised training based on the standard framework designed by Zhou *et al.* [4], Chen *et al.* [59] incorporated additional losses derived from SURF features and mapped point clouds. However, different from [27], [29] which concentrated on supervised depth estimation, Tian *et al.* [58] and Chen *et al.* [59] leveraged sparse depth information from the visual odometry system and explored additional supervisions for self-supervised monocular training but they still suffered from unsatisfactory performance (AbsRel of 16.5% on NYUv2). In this paper, we proposed a novel framework, **MonoIndoor++**, with three new modules, a depth factorization module, a residual pose estimation module, and a coordinates convolutional encoding module, which target on solving existing problems, rapid scale changes in indoor environments, inaccurate camera rotation prediction issue and missing coordinates cues in inducing relative camera poses, for self-supervised monocular depth estimation in indoor environments. Our model can be trained with standard photometric loss derived from self-supervision and has established state-of-the-art (SOTA) performance on a wide-range of challenging benchmark indoor datasets.

III. METHOD

In this section, we present detailed descriptions of performing self-supervised depth estimation using the proposed **MonoIndoor++**. Specifically, we first give an overview of the standard framework for the self-supervised depth estimation. Then, we describe three core components including depth factorization, residual pose and coordinates convolution modules, respectively.

A. Self-Supervised Monocular Depth Estimation

Self-supervised monocular depth estimation is considered as a novel view-synthesis problem which is defined in [4], [5] and [12]. This key idea is to train a model to predict the target image from different viewpoints of source images. The image synthesis is achieved by using the depth map as the bridging variable between the depth network and pose network. Both the depth map of the target image and the estimated relative camera pose between a pair of target and source images are required to train such systems. Specifically, the depth network predicts a dense depth map D_t given a target image I_t as the input. The pose network takes a target image I_t and a source image $I_{t'}$ from another view and estimates a relative camera pose $T_{t\to t'}$ from the target to the source. The depth network and pose network are optimized jointly with the photometric reprojection loss which can then be constructed as follows:

$$\mathcal{L}_A = \sum_{t'} \rho(I_t, I_{t'\to t}), \tag{1}$$

and

$$I_{t'\to t} = I_{t'}\langle proj(D_t, T_{t\to t'}, K)\rangle, \tag{2}$$

where ρ denotes the photometric reconstruction error [4], [5]. It is a weighted combination of the L1 and Structured SIMilarity (SSIM) loss defined as

$$\rho(I_t, I_{t'\to t}) = \frac{\alpha}{2}\big(1 - \texttt{SSIM}(I_t, I_{t'\to t})\big) + (1-\alpha)\|I_t, I_{t'\to t}\|_1. \tag{3}$$

$I_{t'\to t}$ is the source image warped to the target coordinate frame based on the depth of the target image which is the output

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

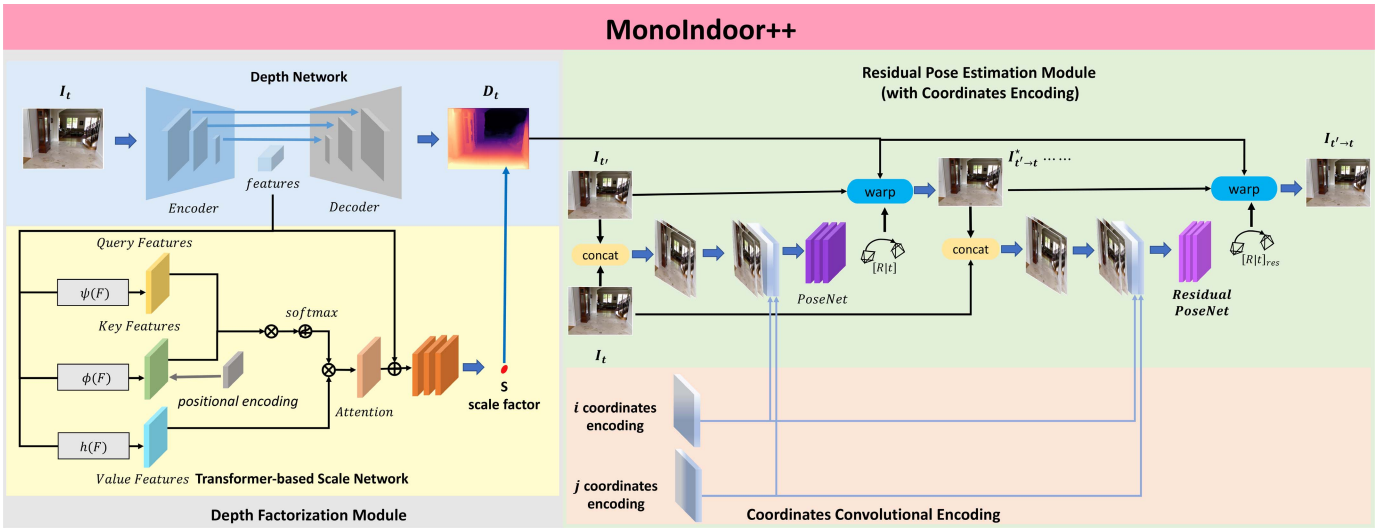


Fig. 1. Overview of the proposed **MonoIndoor++**. **Depth Factorization Module**: We use an encoder-decoder based depth network to predict a relative depth map and a transformer-based scale network to estimate a global scale factor. **Residual Pose Estimation Module**: We use a pose network to predict an initial camera pose of a pair of frames and residual pose network to iteratively predict residual camera poses based on the predicted initial pose. **Coordinates Convolutional Encoding**: We encode coordinates information along with the concatenated color image pairs as the input to the pose network and residual pose network for predicting relative camera poses.

from the depth network. $proj()$ is the transformation function to map image coordinated p_t from the target image to its $p_{t'}$ on the source image following

$$p_{t'} \sim K T_{t\to t'} D_t(p_t) K^{-1} p_t, \tag{4}$$

and $\langle\cdot\rangle$ is the bilinear sampling operator which is locally sub-differentiable.

In addition, an edge-ware smoothness term is normally employed during training which can be written as

$$\mathcal{L}_s = |\partial_x d_t^*| e^{-|\partial_x I_t|} + |\partial_y d_t^*| e^{-|\partial_y I_t|}, \tag{5}$$

where $d_t^* = d/\bar{d}_t$ is the mean-normalized inverse depth from [39].

Further, inspired by [38], we incorporate the depth consistency loss to enforce the predicted depth maps across the target frame and neighbouring source frames to be consistent during the training. We first warp the predicted depth map $D_{t'}$ of the source image $I_{t'}$ by Equation (2) to generate $D_{t'\to t}$, which is a corresponding depth map in the coordinate system of the source image. We then transform $D_{t'\to t}$ to the coordinate system of the target image via Equation (4) to produce a synthesized target depth map $\widetilde{D}_{t'\to t}$. The depth consistency loss can be written as

$$\mathcal{L}_d = \frac{|D_t - \widetilde{D}_{t'\to t}|}{D_t + \widetilde{D}_{t'\to t}}. \tag{6}$$

Thus, the overall objective to train the model is

$$\mathcal{L} = \mathcal{L}_A + \tau \mathcal{L}_s + \gamma \mathcal{L}_d, \tag{7}$$

where τ and γ are the weights for the edge-aware smoothness loss and the depth consistency loss respectively.

As discussed in Section I, existing self-supervised monocular depth estimation models have been used widely in producing competitive depth maps on datasets collected in outdoor environments, *e.g.*, autonomous driving scenes. However, simply using these methods [5] still suffer from worse performance in indoor environments, especially compared with fully-supervised depth prediction methods. We argue that the main challenges in indoor environments come from the fact that i) the depth range changes a lot and ii) indoor sequences captured in existing public indoor datasets, *e.g.*, EuRoC MAV [18] and NYUv2 [14], contain regular rotational motions which are difficult to predict. To handle these issues, we propose **MonoIndoor++**, a self-supervised monocular depth estimation framework, as shown in Figure 1, to enable improved predicted depth quality in indoor environments. The framework takes as input a single color image and outputs a depth map via our MonoIndoor++ which consists of two core parts: a depth factorization module with a transformer-based scale regression network and a residual pose estimation module. In addition, when designing the residual pose estimation, we incorporate coordinates convolutional operations to encode coordinates information along with color information as input channels explicitly. The details of our main contributions are presented in the following sections.

B. Depth Factorization Module

Our depth factorization module consists of a depth prediction network and a transformer-based scale regression network.

1) Depth Prediction Network: The backbone model of our depth prediction network is based on Monodepth2 [5], which employs an auto-encoder structure with skip connections between the encoder and the decoder. The depth encoder learns a feature representation given a color image I as input. The decoder takes features from the encoder as the input and outputs relative depth map prediction. In the decoder, a sigmoid activation function is used to process features from the last convolutional layers and a linear scaling function is utilized to obtain the final up-to-the-scale depth prediction,

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

which can be written as follows,

$$d = 1/(a\sigma + b), \tag{8}$$

where σ is the outputs after the sigmoid function, a and b are specified to constrain the depth map D within a certain depth range. a and b are pre-defined as a minimum depth value and a maximum depth value empirically according to a known environment. For instance, on the KITTI dataset [7] which is collected in outdoor scenes, a is chosen as 0.1 and b as 100. The reason for setting a and b as these fixed values is that the depth range is consistent across the video sequences when the camera always sees the sky at the far point. However, it is observed that this setting is not valid for most indoor environments. For instance, on the NYUv2 dataset [14] which include various indoor scenes, *e.g.*, office, kitchen, *etc.*, the depth range varies significantly as scene changes. Specifically, the depth range in a bathroom (*e.g.*, 0.1m∼3m) can be very different from the one in a lobby (*e.g.*, 0.1m∼10m). We argue that pre-setting depth range will act as an inaccurate guidance that is harmful for the model to capture accurate depth scales in training models. This is especially true when there are rapid scale changes, which are commonly observed on datasets [14], [18], [19] in indoor scenes. Therefore, to mitigate this problem, our depth factorization module learns a disentangled representation in the form of a relative depth map and a global scale factor. The relative depth map is obtained by the depth prediction network aforementioned and a global scale factor is outputted by a transformer-based scale regression network which is introduced in the next subsection.

2) Transformer-Based Scale Regression Network: We propose a transformer-based scale regression network (see Figure 1) as a new branch which takes as input a color image and outputs its corresponding global scale factor. Our intuition is that the global scale factor can be informed by certain areas (*e.g.*, the far point) in the images, and we propose to use a transformer block to learn the global scale factor. Our expectation is that the network can be guided to pay more attention to a certain area which is informative to induce the depth scale factor of the target image of the current view in a scene.

The proposed transformer-based scale regression network takes the feature representations $\mathcal{F} \in \mathbb{R}^{D\times H\times W}$ learnt from the input image as the input and outputs the corresponding global scale factor, where D is dimension, H and W are the height and width of the feature map. Specifically, we project input features $\mathcal{F} \in \mathbb{R}^{D\times H\times W}$ to the query, the key and the value output, which are defined as

$$\begin{aligned} \psi(\mathcal{F}) &= \mathbf{W}_{\psi}\mathcal{F}, \\ \phi(\mathcal{F}) &= \mathbf{W}_{\phi}\mathcal{F}, \\ h(\mathcal{F}) &= \mathbf{W}_{h}\mathcal{F}, \end{aligned} \tag{9}$$

where $\mathbf{W}_{\psi}$, $\mathbf{W}_{\phi}$ and $\mathbf{W}_{h}$ are parameters to be learnt. The query and key values are then combined using the function $\mathcal{G}_{\mathcal{F}} = \texttt{softmax}(\mathcal{F}^{T}\mathbf{W}_{\psi}^{T}\mathbf{W}_{\phi}\mathcal{F})h(\mathcal{F})$, giving the learnt self-attentions as $\mathcal{G}_{\mathcal{F}}$. Finally, the $\mathcal{G}_{\mathcal{F}}$ and the input feature representation $\mathcal{F}$ jointly contribute to the output $\mathcal{S}_{\mathcal{F}}$ by using

$$\mathcal{S}_{\mathcal{F}} = \mathbf{W}_{\mathcal{S}_{\mathcal{F}}}\mathcal{G}_{\mathcal{F}} + \mathcal{F}. \tag{10}$$

Once we obtain $\mathcal{S}_{\mathcal{F}}$, we apply three residual blocks including two convolutional layers in each, followed by three fully-connected layers with dropout layers in-between, to output the global scale factor S for the target image of current view. We also use a 2D relative positional encoding [60] in calculating attentions with considerations of relative positional information of key features.

3) Probabilistic Scale Regression Head: The proposed transformer-based scale regression network is designed is to predict a single positive number given the input high-dimensional feature map $\mathcal{F} \in \mathbb{R}^{D\times H\times W}$. Inspired by the stereo matching work [61], we propose to use a probabilistic scale regression head to estimate the continuous value for scale factor. Specifically, given a maximum bound that the global scale factor is within, instead of outputting a single number directly, we first output a number of scale values $\widetilde{S}$ as the predictions of each scale s and then calculate the probability of s via the softmax operation $\texttt{softmax}(\cdot)$. Finally, the predicted global scale S is calculated as the sum of each scale s weighted by its probability of predicted values as

$$S = \sum_{s=0}^{D_{max}} s \times \texttt{softmax}(\widetilde{S}). \tag{11}$$

Thus, the probabilistic scale regression head enables us to resolve regression problem smoothly with a probabilistic classification-based strategy (see Section IV-E2 for ablation results).

C. Residual Pose Estimation

The principle of self-supervised monocular depth estimation is built upon the novel view synthesis, which requires both accurate depth maps from the depth network and camera poses from the pose network. Estimating accurate relative camera poses is important for calculating photometric reprojection loss to train the model because inaccurate camera poses might lead to wrong correspondences between the pixels in the target and source images, positing problems in predicting accurate depth maps. A standalone "PoseNet" is widely used in existing methods [5] to take two images as the input and to estimate the 6 Degrees-of-Freedom (DoF) relative camera poses. On datasets in outdoor environments (*e.g.*, autonomous driving scenes like KITTI), we argue that the relative camera poses are fairly simple because the cars which are used to collect video data are mostly moving forward with large translations but minor rotations. This means that pose estimation is normally less challenging for the pose network. In contrast, in indoor environments, the video sequences in widely-used datasets [14] are typically recorded with hand-held devices (*e.g.*, Kinect), so there are more complicated ego-motions involved as well as much larger rotational motions. Thus, it is relatively more difficult for the pose network to learn to predict accurate relative camera poses.

To better mitigate the aforementioned issues, existing methods [8], [45] concentrate on "removing" or "reducing" rotational components in camera poses during data preprocessing and train their models using preprocessed data. In this work, we argue these preprocessing techniques are not flexible in

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

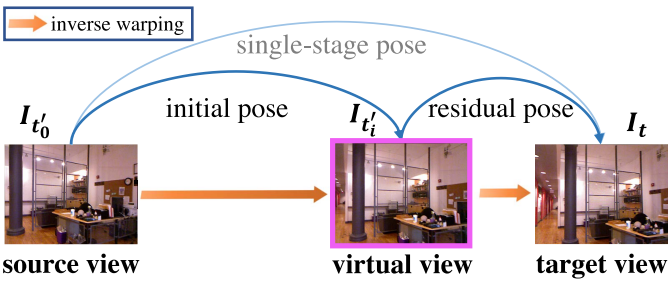


Fig. 2. Residual pose estimation. A single-stage pose can be decomposed into an *initial pose* and a *residual pose* by virtual view synthesis.

end-to-end training pipelines, instead, we propose a residual pose estimation module to learn the relative camera pose between the target and source images from different views in an iterative manner (see Figure 2 for core ideas).

Our residual pose module consists of a standard pose network and a residual pose network. In the first stage, the pose network takes a target image I_t and a source image $I_{t'_0}$ as input and predicts an initial camera pose $T_{t'_0 \to t}$, where the subscript 0 in t'_0 indicates that no transformation is applied over the source image yet. Then Equation (2) is used to bilinearly sample from the source image, reconstructing a warped target image $I_{t'_0 \to t}$ of a virtual view which is expected to be the same as the target image I_t if the correspondences are solved accurately. However, it will not be the case due to inaccurate pose prediction. The transformation for this warping operation is defined as

$$I_{t'_0 \to t} = I_{t'} \langle proj(D_t, T^{-1}_{t'_0 \to t}, K) \rangle. \tag{12}$$

Next, we propose a residual pose network (see ***Residual-PoseNet*** in Figure 1) which takes the target image and the synthesized target image of a virtual view ($I_{t'_0 \to t}$) as input and outputs a residual camera pose $T^{res}_{(t'_0 \to t) \to t}$, representing the camera pose of the synthesized image $I_{t'_0 \to t}$ with respect to the target image I_t. Then, we bilinearly sample from the synthesized image as

$$I_{(t'_0 \to t) \to t} = I_{t'_0 \to t} \langle proj(D_t, T^{res\ -1}_{(t'_0 \to t) \to t}, K) \rangle. \tag{13}$$

Once a new synthesized image of a virtual view is obtained, we can continue to estimate the residual camera poses for next view synthesis operation.

We define the general form of Equation (13) as

$$I_{t'_i \to t} = I_{t'_i} \langle proj(D_t, T^{res-1}_{t'_i \to t}, K) \rangle, \quad i = 0, 1, \cdots. \tag{14}$$

by replacing the subscript $t'_0 \to t$ with t'_1 to indicate that one warping transformation is applied, and similarly for the i^{th} transformation.

To this end, after multiple residual poses are estimated, the camera pose of source image I'_t with respect to the target image I_t can be written as $T_{t \to t'} = T^{-1}_{t' \to t}$ where

$$T_{t' \to t} = \prod_i T_{t'_i \to t}, \quad i = \cdots, k, \cdots, 1, 0\ . \tag{15}$$

By iteratively estimating residual poses using a pose network and a residual pose network, we expect to obtain more accurate camera pose compared with the pose predicted from a single-stage pose network, so that a more accurate photometric reprojection loss can be built up for better depth prediction during the model training.

D. Coordinates Convolutional Encoding

For self-supervised monocular depth estimation, most of existing methods are designed to induce relative camera poses given a pair of color images. In this work, we propose to incorporate coordinates information as a part of input channels along with the color information explicitly to provide additional coordinates cues for pose estimation.

We extend standard convolutional layers to coordinates convolutional layers by initializing extra channels to process coordinates information which is concatenated channel-wise to the input representations (see **Coordinates Convolutional Encoding** in Figure 1). Given a pair of 2D images, we encode two coordinates x, y with color information (r, g, b), resulting in the 8-channels input as $(r_1, g_1, b_1, r_2, g_2, b_2, i, j)$ where (r_1, g_1, b_1) and (r_2, g_2, b_2) are rgb values of color images, respectively. The i coordinate channel is an $h \times w$ rank-1 matrix with its first row filled with 0's, its second row with 1's, its third with 2's, etc. The j coordinate channel is similar, but with columns filled in with constant values instead of rows. A linear scaling operation is applied over both i and j coordinate values to encode them in the range $[-1, 1]$. We adopt coordinates convolutional layers [57] in the residual pose estimation module to process 8-channels input for iterative pose estimation, and the pose estimation can be written as follows:

$$T_{t \to t'} = RPModule(\Omega; Concat(I_t, I_{t'}, i, j)) \tag{16}$$

where $RPModule$ is the proposed pose estimation module, Ω is the parameters of the module which are to be optimized.

IV. Experiments

A. Implementation Details

We implement our model using PyTorch [62]. In the depth factorization module, we use the same depth network as in Monodepth2 [5]; for the transformer-based scale regression network, we use a transformer module followed by two basic residual blocks and then three fully-connected layers with a dropout layer in-between. The dropout rate is empirically set to 0.5. In the residual pose module, we let the residual pose networks use a common architecture as in Monodepth2 [5] which consists of a shared pose encoder and an independent pose regressor. In the coordinates encoding module, 2D coordinates information (i, j) are directly concatenated with (r, g, b) channels of color images as the input and the convolutional layers are initialized with ImageNet-pretrained weights. Each experiment is trained for 40 epochs using the Adam [63] optimizer and the learning rate is set to 10^{-4} for the first 20 epochs and it drops to 10^{-5} for remaining epochs. The smoothness term τ is set as 0.001. The consistency term γ are set as 0.1 for EuRoC MAV dataset, 0.035 for NYUv2, ScanNet and 7-Scenes datasets, respectively.

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

B. Datasets

1) NYUv2 [14]: The NYUv2 depth dataset contains 464 indoor video sequences which are captured by a handheld Microsoft Kinect RGB-D camera. The dataset is widely used as a challenging benchmark for depth prediction. The resolution of videos is 640 × 480. Images are rectified with provided camera intrinsics to remove image distortion. We use the official training and validation splits which include 302 and 33 sequences, respectively. We use officially provided 654 images with dense labelled depth maps for testing. During training, images are resized to 320 × 256.

2) EuRoC MAV [18]: The EuRoC MAV Dataset contains 11 video sequences captured in two main scenes, a machine hall and a vicon room. Sequences are categorized as *easy*, *medium* and *difficult* according to the varying illumination and camera motions. For the training, we use three sequences of "Machine hall" (MH_01, MH_02, MH_04) and two sequences of "Vicon room" (V1_01 and V1_02). Images are rectified with provided camera intrinsics to remove image distortion. During training, images are resized to 512 × 256. We use the Vicon room sequence V1_03, V2_01, V2_02 and V2_03 for testing where the ground-truth depths are generated by projecting Vicon 3D scans onto the image planes. During training, images are resized to 512 × 256. In addition, we use V2_01 for ablation studies (see Section IV-E1 and Section IV-E2).

3) ScanNet [19]: The ScanNet dataset contains RGB-D videos of 1513 indoor scenes, which is captured by handheld devices. The dataset is annotated with 3D camera poses and instance-level semantic segmentations and is widely on several 3D scene understanding tasks, including 3D object classification, semantic voxel labeling, and CAD model retrieval. We use officially released train-validation-test splits. The resolution of color images is 1296 × 968. During training, images are resized to 320 × 256.

4) 7-Scenes [20]: 7-Scenes dataset contains a number of video sequences captured in 7 different indoor scenes, *i.e.*, *office*, *stairs*, etc. Each scene contains 500-1000 frames. All scenes are recorded using a handheld Kinect RGB-D camera at the resolution of 640 × 480. We use the official train-test split. During training, images are resized to 320 × 256.

C. Evaluation Metrics

We use both error metrics and accuracy metrics proposed in [23] for evaluation on all datasets, which include the mean absolute relative error (AbsRel), root mean squared error (RMS) and the accuracy under threshold ($\delta_i < 1.25^i$, $i = 1, 2, 3$). Following previous self-supervised depth estimation methods [5], [9], [10], we multiply the predicted depth maps by a scalar that matches the median with that of the ground-truth because self-supervised monocular methods cannot recover the metric scale. The predicted depths are capped at 10m in all indoor datasets except the EuRoC MAV dataset which one is set as 20m because it contains "Machine hall" scenes with observed large depth scale.

D. Experimental Results

1) Results on NYUv2 Depth Dataset: In this sub-section, we evaluate our **MonoIndoor**++ on the NYUv2 depth dataset [14]. Following [9], [10], the raw dataset is firstly downsampled 10 times along the temporal dimension to remove redundant frames, resulting in $\sim 20K$ images for training.

TABLE I

COMPARISON OF OUR METHOD WITH EXISTING SUPERVISED AND SELF-SUPERVISED METHODS ON NYUV2 [14]. BEST RESULTS AMONG SUPERVISED AND SELF-SUPERVISED METHODS ARE IN **BOLD**

Methods	Supervision	Error Metric ↓		Accuracy Metric ↑		
		AbsRel	RMS	δ_1	δ_2	δ_3
Make3D [21]	✓	0.349	1.214	0.447	0.745	0.897
Depth Transfer [64]	✓	0.349	1.210	-	-	-
Liu *et al.* [65]	✓	0.335	1.060	-	-	-
Ladicky *et al.* [66]	✓	-	-	0.542	0.829	0.941
Li *et al.* [67]	✓	0.232	0.821	0.621	0.886	0.968
Roy *et al.* [68]	✓	0.187	0.744	-	-	
Liu *et al.* [69]	✓	0.213	0.759	0.650	0.906	0.976
Wang *et al.* [70]	✓	0.220	0.745	0.605	0.890	0.970
Eigen *et al.* [1]	✓	0.158	0.641	0.769	0.950	0.988
Chakrabarti *et al.* [71]	✓	0.149	0.620	0.806	0.958	0.987
Laina *et al.* [24]	✓	0.127	0.573	0.811	0.953	0.988
Li *et al.* [26]	✓	0.143	0.635	0.788	0.958	0.991
DORN [2]	✓	0.115	0.509	0.828	0.965	0.992
Ranftl *et al.* [36]	✓	0.110	**0.357**	0.904	**0.988**	0.994
VNL [11]	✓	0.108	0.416	0.875	0.976	0.994
Bhat *et al.* [25]	✓	0.103	0.364	**0.903**	0.984	**0.997**
Fang *et al.* [72]	✓	**0.101**	0.412	0.868	0.958	0.986
Zhou *et al.* [8]	✗	0.208	0.712	0.674	0.900	0.968
Zhao *et al.* [9]	✗	0.189	0.686	0.701	0.912	0.978
Monodepth2 [5]	✗	0.160	0.601	0.767	0.949	0.988
SC-Depth [38]	✗	0.159	0.608	0.772	0.939	0.982
P^2Net (3-frame) [73]	✗	0.159	0.599	0.772	0.942	0.984
P^2Net (5-frame) [73]	✗	0.147	0.553	0.801	0.951	0.987
Bian *et al.* [45]	✗	0.147	0.536	0.804	0.950	0.986
Bian *et al.* [10]	✗	0.138	0.532	0.820	0.956	0.989
Monodepth2 [5] (Baseline)	✗	0.160	0.601	0.767	0.949	0.988
MonoIndoor [17] (Our ICCV21)	✗	0.134	0.526	0.823	0.958	0.989
MonoIndoor++ (Ours)	✗	**0.132**	**0.517**	**0.834**	**0.961**	**0.990**

a) Quantitative results: Table I presents the quantitative results of our model **MonoIndoor**++ and both SOTA supervised and self-supervised methods on NYUv2. It shows that our model outperforms all previous self-supervised SOTA methods [5], [10], [38], [73], reaching the best results across all metrics. Specifically, our method improves monocular depth prediction performance significantly by reducing AbsRel to 13.2% and increasing δ_1 to 83.4%. Besides, compared with the recent self-supervised methods by Bian *et al.* [10], [45] which concentrating on removing rotations via "rectification" as a data preprocessing step, our method gives the better performance without additional data preprocessing. It is noted that NYUv2 is very challenging and many previous self-supervised methods [40] fail to get satisfactory results. In addition to that, our model outperforms a group of supervised methods [1], [26], [65], [66], [71] and closes the performance gap between the self-supervised methods and fully-supervised methods [2], [24]. When compared with our preliminary work [17], our method can consistently improve depth estimation performance on all metrics, especially the δ_1, which is 83.4% and is better that these fully-supervised methods [2], [24]. Ablation studies for the effectiveness of each core module on NYUv2 are presented in Section IV-E1 and the ablation results of design choices for the coordinates convolutional encoding are shown in Section IV-E3.

b) Qualitative results: Figure 3 visualizes the predicted depth maps on NYUv2. Compared with the results from the

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

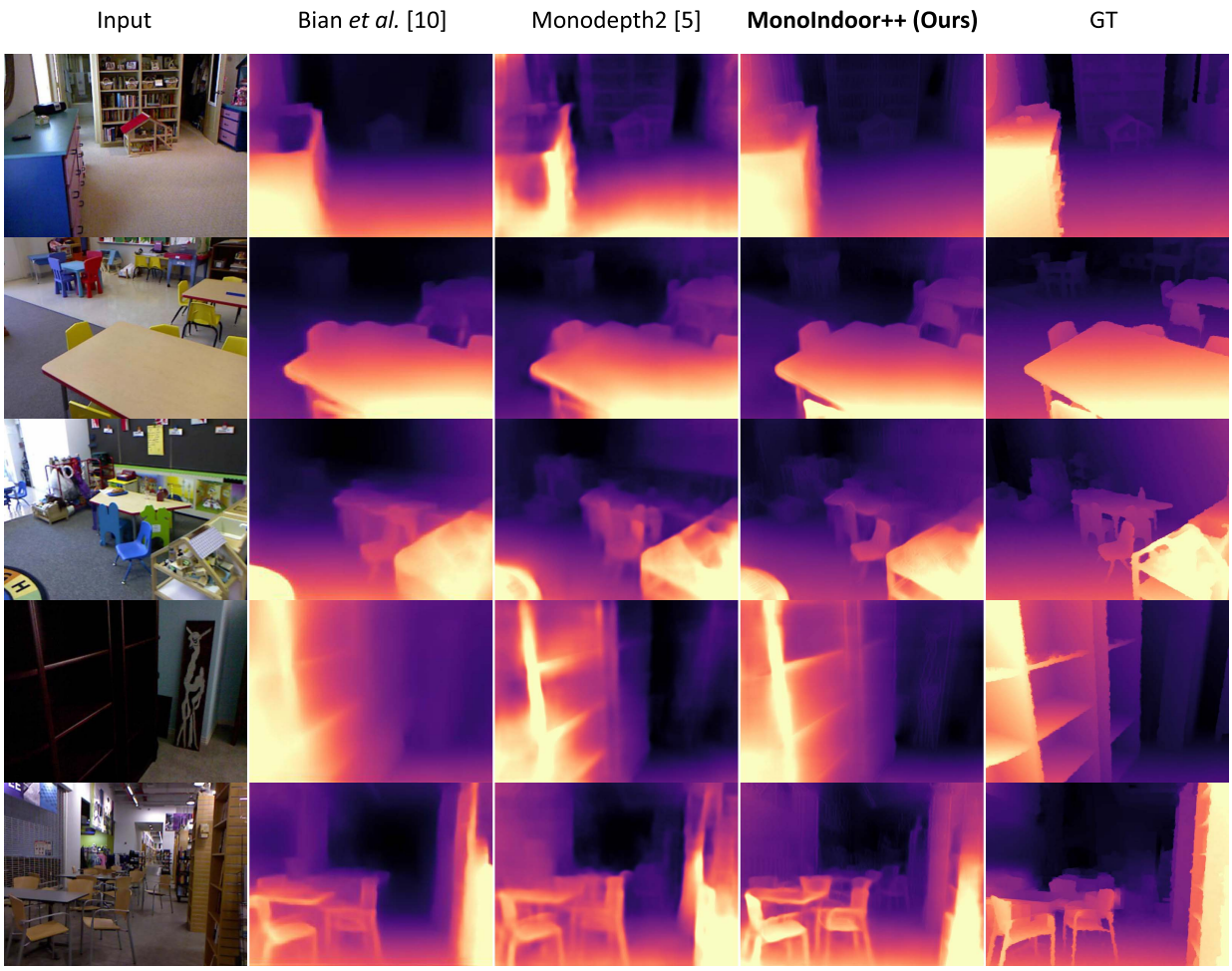


Fig. 3. Qualitative comparison on NYUv2 [14]. Images form the left to the right are: input, depth from [10] and [5], **MonoIndoor++(Ours)**, and ground-truth depth. Compared with both the baseline method Monodepth2 [5] and recent work [10], our model produces accurate depth maps that are closer to the ground-truth.

baseline method Monodepth2 [5] and recent work [10], depth maps predicted from our model (**MonoIndoor++**) are more precise and closer to the ground-truth. For instance, looking at the fourth column in the first row, the depth in the region of ***cabinet*** predicted from our model is much sharper and cleaner, being close to the ground-truth (the last column). These qualitative results are consistent with our quantitative results in Table I.

2) Results on EuRoC MAV Dataset: In this sub-section, we present evaluation results of self-supervised monocular depth estimation on the EuRoC MAV dataset [18]. As there are not many public results on the EuRoC MAV dataset, excepting for comparing between our **MonoIndoor++** and the baseline method Monodepth2 [5], we follow official implementations of Bian *et al.* [38],[1] P^{2}Net [73][2] and Bian *et al.* [10][3] to conduct experiments and make fair comparisons.

a) Quantitative results: We present quantitative results of our model **MonoIndoor++** and comparisons with other methods for the self-supervised monocular depth estimation on all Vicon room testing sequences in Table II. It can be observed that, when compared with recent self-supervised methods [10], [38], [73], our model achieves the best performance across all major evaluation metrics (AbsRel and δ_1) on various scenes including the "difficult" scene, *i.e.*, "Vicon room 203" (V2_03). Specifically, on the sequence V2_01, our model improves self-supervised monocular depth estimation performance significantly by reducing the AbsRel to 11.5% and increasing the δ_1 to 86.1%. Similar improvements can be observed on other test sequences. Besides, compared with our preliminary work [17], our method consistently and significantly improves depth estimation performance across all test sequences. In addition, ablation studies for the effectiveness of each core module are presented in Section IV-E1 and ablation experiments of the design choices for the scale network are shown in Section IV-E2.

b) Qualitative results: We present the qualitative results and comparisons of depth maps predicted by the baseline method Monodepth2 [5] and our **MonoIndoor++** in Figure 4. There are no ground-truth dense depth maps on the EuRoC MAV dataset. From Figure 4, it is clear that the depth maps

[1]https://github.com/JiawangBian/SC-SfMLearner-Release
[2]https://github.com/svip-lab/Indoor-SfMLearner
[3]https://github.com/JiawangBian/sc_depth_pl

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

TABLE II

QUANTITATIVE RESULTS AND COMPARISON BETWEEN OUR **MONOINDOOR++** WITH EXISTING SELF-SUPERVISED METHODS ON THE TEST SEQUENCES V1_03, V2_01 V2_02, V2_03 OF EUROC MAV [18]. BEST RESULTS ARE IN **BOLD**

Method	V1_03					V2_01				
	Error Metric ↓		Accuracy Metric ↑			Error Metric ↓		Accuracy Metric ↑		
	AbsRel	RMSE	δ_1	δ_2	δ_3	AbsRel	RMSE	δ_1	δ_2	δ_3
Bian *et al.* [38]	0.100	0.387	0.905	0.985	0.996	0.153	0.554	0.807	0.944	0.984
P^2Net [73]	0.104	0.387	0.905	0.986	0.997	0.155	0.557	0.780	0.953	0.989
Bian *et al.* [10]	0.094	0.360	0.925	0.985	0.995	0.148	0.536	0.800	0.950	0.987
Monodepth2 [5] (Baseline)	0.110	0.413	0.889	0.983	0.996	0.157	0.567	0.786	0.941	0.986
MonoIndoor [17] (Our ICCV21)	0.080	0.309	0.944	0.990	0.998	0.125	0.466	0.840	0.965	**0.993**
MonoIndoor++ (Ours)	**0.079**	**0.303**	**0.949**	**0.991**	**0.998**	**0.115**	**0.439**	**0.861**	**0.972**	0.992
Method	V2_02					V2_03				
	Error Metric ↓		Accuracy Metric ↑			Error Metric ↓		Accuracy Metric ↑		
	AbsRel	RMSE	δ_1	δ_2	δ_3	AbsRel	RMSE	δ_1	δ_2	δ_3
Bian *et al.* [38]	0.161	0.682	0.769	0.942	0.983	0.163	0.616	0.760	0.948	0.989
P^2Net [73]	0.150	0.604	0.800	0.955	0.989	0.152	0.541	0.792	0.954	0.991
Bian *et al.* [10]	0.154	0.637	0.783	0.948	0.987	0.149	0.534	0.792	0.962	0.992
Monodepth2 [5] (Baseline)	0.156	0.645	0.776	0.945	0.985	0.171	0.620	0.734	0.944	0.988
MonoIndoor [17] (Our ICCV21)	0.142	0.581	0.802	0.952	0.990	0.140	0.502	0.810	0.964	0.993
MonoIndoor++ (Ours)	**0.133**	**0.551**	**0.830**	**0.964**	**0.991**	**0.134**	**0.482**	**0.829**	**0.967**	**0.993**

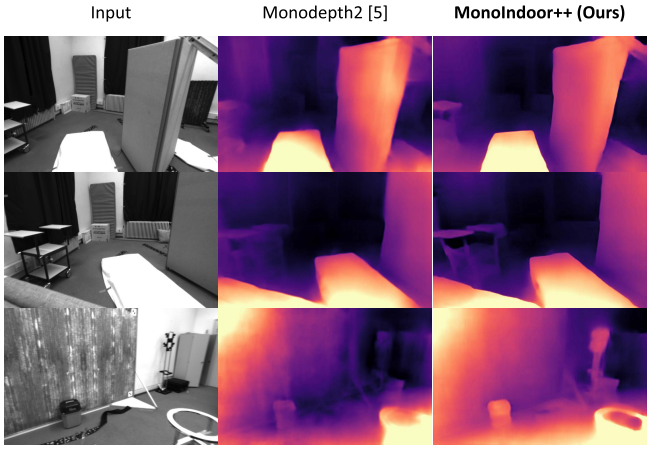


Fig. 4. Qualitative comparison of depth prediction on EuRoC MAV. Our **MonoIndoor**++ produces more accurate and cleaner depth maps.

TABLE III

RELATIVE POSE EVALUATION ON EUROC MAV [15]. RESULTS SHOW THE AVERAGE ABSOLUTE TRAJECTORY ERROR (ATE), AND THE RELATIVE POSE ERROR (RPE) IN METERS AND DEGREES, RESPECTIVELY. SCENE: TEST SEQUENCE NAME

Scene	Methods	ATE (m) ↓	RPE (m) ↓	RPE (°) ↓
V1_03	Monodepth2 [5] (Baseline)	0.0681	0.0686	1.3237
	MonoIndoor [17] (Our ICCV21)	0.0564	0.0638	0.7185
	MonoIndoor++ (Ours)	**0.0557**	**0.0542**	**0.5599**
V2_01	Monodepth2 [5] (Baseline)	0.0266	0.0199	1.1985
	MonoIndoor [17] (Our ICCV21)	0.0230	0.011	1.197
	MonoIndoor++ (Ours)	**0.0229**	**0.0050**	**1.1239**
V2_02	Monodepth2 [5] (Baseline)	0.0624	0.0481	**6.4135**
	MonoIndoor [17] (Our ICCV21)	0.0544	0.0360	7.2100
	MonoIndoor++ (Ours)	**0.0517**	**0.0350**	7.1928
V2_03	Monodepth2 [5] (Baseline)	0.0670	**0.0355**	5.3532
	MonoIndoor [17] (Our ICCV21)	0.0699	0.0748	5.304
	MonoIndoor++ (Ours)	**0.0644**	0.0676	**4.8559**

generated by our model are much better than the ones by Monodepth2. For instance, in the third row, our model can predict precise depths for the ***hole*** region at the right-bottom corner whereas such a hole structure in the depth map by Monodepth2 is missing. These observations are also consistent with the better quantitative results in Table II, proving the superiority of our model.

c) Relative pose evaluation: In Table III, we evaluate the proposed residual pose estimation module on all Vicon room test sequences V1_03, V2_01, V2_02 and V2_03 of the EuRoC MAV [18]. We follow [74] to evaluate relative camera poses estimated by our residual pose estimation module. We use the following evaluation metrics: absolute trajectory error (ATE) which measures the root-mean square error between predicted camera poses and ground-truth, and relative pose error (RPE) which measures frame-to-frame relative pose error in meters and degrees, respectively. As shown in Table III, compared with the baseline model Monodepth2 [5] which employs one-stage pose network, using our method leads to improved relative pose estimation across evaluation metrics on most test scenes. Specifically, on the scene V1_03, the ATE by our MonoIndoor++ is significantly decreased from 0.0681 meters to **0.0557** meters and RPE(°) is reduced from 1.3237° to **0.5599**°. Similar observations are made on the scene V2_02, where the ATE by our MonoIndoor++ is significantly decreased from 0.0624 meters to **0.0517** meters. When compared with our preliminary work [17], consistent improvements can also be observed across almost all testing

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

TABLE IV

COMPARISON OF OUR METHOD WITH EXISTING SUPERVISED AND SELF-SUPERVISED METHODS ON SCANNET [19]. BEST RESULTS AMONG SUPERVISED AND SELF-SUPERVISED METHODS ARE IN **BOLD**

Methods	Supervision	AbsRel ↓	SqRel ↓	RMS ↓	RMS_{log} ↓
Photometric BA [75]	✓	0.268	0.427	0.788	0.330
DeMoN [30]	✓	0.231	0.520	0.761	0.289
BANet [76]	✓	0.161	0.092	0.346	0.214
DeepV2D [31]	✓	0.069	**0.018**	0.196	0.099
NeuralRecon [77]	✓	**0.047**	0.024	**0.164**	**0.093**
Bian *et al.* [38]	✗	0.177	0.238	0.552	0.220
P^2Net [73]	✗	0.218	0.190	0.531	0.256
Bian *et al.* [10]	✗	0.163	0.096	0.428	0.188
Gu *et al.* [78]	✗	0.140	0.127	0.496	0.212
Monodepth2 [5] (Baseline)	✗	0.189	0.111	0.426	0.225
MonoIndoor [17] (Our ICCV21)	✗	0.126	0.057	0.329	0.163
MonoIndoor++ (Ours)	✗	**0.113**	**0.048**	**0.302**	**0.148**

TABLE V

ZERO-SHOT GENERALIZATION OF OUR METHOD FOR SELF-SUPERVISED DEPTH ESTIMATION ON SCANNET [19]. BEST RESULTS ARE IN **BOLD**

Methods	Supervision	Error Metric ↓		Accuracy Metric ↑		
		AbsRel	RMS	δ_1	δ_2	δ_3
Latina *et al.* [24]	✓	0.141	0.339	0.811	.958	0.990
VNL [11]	✓	**0.123**	**0.306**	**0.848**	**0.964**	**0.991**
Zhou *et al.* [8]	✗	0.212	0.483	0.650	0.905	0.976
Zhao *et al.* [9]	✗	0.179	0.415	0.726	0.927	0.980
Bian *et al.* [38]	✗	0.169	0.392	0.749	0.938	0.983
P^2Net [73]	✗	0.175	0.420	0.740	0.932	0.982
Bian *et al.* [10]	✗	0.156	0.361	0.781	0.947	0.987
Monodepth2 [5] (Baseline)	✗	0.170	0.401	0.730	0.948	0.991
MonoIndoor [17] (Our ICCV21)	✗	0.154	0.373	0.779	0.951	0.988
MonoIndoor++ (Ours)	✗	**0.138**	**0.347**	**0.810**	**0.967**	**0.993**

TABLE VI

ZERO-SHOT GENERALIZATION OF OUR METHOD FOR RELATIVE POSE ESTIMATION ON SCANNET [19]. BEST RESULTS ARE IN **BOLD**. ROT: ROTATIONAL ERROR OF THE RELATIVE POSE. TR: TRANSLATIONAL ERROR OF THE RELATIVE POSE

Methods	rot (deg) ↓	tr (deg) ↓	tr (cm) ↓
Zhou *et al.* [8]	1.96	39.17	1.4
P^2Net [73]	1.86	35.11	0.89
Bian *et al.* [10]	1.82	39.41	0.55
Monodepth2 [5] (Baseline)	2.03	41.12	0.83
MonoIndoor [17] (Our ICCV21)	1.36	23.42	1.04
MonoIndoor++ (Ours)	**1.19**	**21.33**	**0.27**

sequences, which can further validate the superiority of our model for self-supervised monocular depth estimation.

3) Results on ScanNet Dataset: In this sub-section, we evaluate our **MonoIndoor**++ and compare its performance with recent SOTA methods on the ScanNet dataset [19]. Referring to [76], the raw dataset is firstly downsampled 10 times along the temporal dimension and then $\sim 100K$ images are randomly selected for training. During testing, $\sim 4K$ are sampled from 100 different testing scenes to evaluate the trained model. It should be mentioned that we have observed that rarely research work have conducted thorough experiments on ScanNet for self-supervised monocular depth estimation. Instead, previous work [10], [73] simply conduct zero-shot generalization experiments. In this paper, we *first* present self-supervised depth estimation evaluation results, and *second*, we show evaluations of the zero-shot generalization on depth and relative pose estimation. As introduced in Section IV-D2, we follow official implementations of Bian *et al.* [38], P^2Net [73] and Bian *et al.* [10] to conduct experiments and make fair comparisons.

a) Self-supervised depth estimation evaluation: Table IV presents the quantitative results of our model **MonoIndoor**++ and both SOTA supervised and self-supervised methods on ScanNet. It shows that our **MonoIndoor**++ outperforms the previous self-supervised methods [5], [10], [73], [78] in depth estimation, reaching the best results across all metrics. For instance, our model gives 11.3% of the AbsRel, which is exceptionally competitive in indoor environments. When compared with our preliminary work [17], our MonoIndoor++ consistently improves depth estimation performance on this challenging dataset. In addition to that, our model outperforms a group of supervised methods [30], [76]. Ablation studies for the effectiveness of each core module are presented in Section IV-E1.

b) Zero-shot generalization: We present the zero-shot generalization results of self-supervised depth estimation on ScanNet [19] in Table V, where we evaluate the proposed **MonoIndoor**++ pretrained on NYUv2 dataset. From Table V, it is observed that our NYUv2 pretrained model generalizes better than other recent methods to new dataset. Besides, we show the zero-shot generalization results of relative pose estimation on ScanNet in Table VI. We follow [31], [73], and [10] to use 2000 image pairs selected from diverse indoor scenes for pose evaluation. It can be observed that our method outperforms other self-supervised methods. Specifically, compared to Bian *et al.* [10], our method significantly reduces translational error (tr (cm)) from 0.55 centimeters to **0.27** centimeters and decreases camera rotational error (rot (deg)) from 1.82 to **1.19**. When compared with our preliminary work [17], consistent improvements are observed on depth and relative pose evaluation results. Both depth and pose results validate the good zero-shot generalizability and capability of our method.

4) Results on RGB-D 7-Scenes Dataset: In this sub-section, we evaluate our **MonoIndoor**++ on the RGB-D 7-Scenes dataset [20] under two settings, the zero-shot generalization and the fine-tuning strategy, respectively. Following [45] and [10], we extract one image from every 30 frames in each video sequence. For fine-tuning, we first pre-train our model on the NYUv2 dataset, and then fine-tune the pre-trained model on each scene of 7-Scenes dataset.

Table VII presents the quantitative results and comparisons of our model **MonoIndoor**++ and latest SOTA self-supervised methods on 7-Scenes dataset. It can be observed that our model outperforms the baseline method

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
COMPARISON OF OUR METHOD TO LATEST SELF-SUPERVISED METHODS UNDER ZERO-SHOT GENERALIZATION AND FINE-TUNING SETTINGS ON RGB-D 7-SCENES [20]. BEST RESULTS ARE IN **BOLD**

Scenes	Zero-shot Generalization							
	Bian *et al.* [10]		Bian *et al.* [45]		Monodepth2 [5]		**MonoIndoor++ (Ours)**	
	AbsRel ↓	Acc (δ_1) ↑	AbsRel ↓	Acc (δ_1) ↑	AbsRel ↓	Acc (δ_1) ↑	AbsRel ↓	Acc (δ_1) ↑
Chess	0.179	0.689	0.169	0.719	0.193	0.654	**0.157**	**0.750**
Fire	0.163	0.751	0.158	0.758	0.190	0.670	**0.150**	**0.768**
Heads	0.171	0.746	**0.162**	**0.749**	0.206	0.661	0.171	0.727
Office	0.146	0.799	0.132	0.833	0.168	0.748	**0.130**	**0.837**
Pumpkin	0.120	0.841	0.117	0.857	0.135	0.816	**0.102**	**0.895**
RedKitchen	0.136	0.822	0.151	0.780	0.168	0.733	**0.144**	**0.795**
Stairs	**0.143**	0.794	0.162	0.765	0.146	**0.806**	0.155	0.753
Scenes	After Fine-tuning							
	Bian *et al.* [10]		Bian *et al.* [45]		Monodepth2 [5]		**MonoIndoor++ (Ours)**	
	AbsRel ↓	Acc (δ_1) ↑	AbsRel ↓	Acc (δ_1) ↑	AbsRel ↓	Acc (δ_1) ↑	AbsRel ↓	Acc (δ_1) ↑
Chess	0.150	0.780	0.103	0.880	0.123	0.853	**0.097**	**0.888**
Fire	0.105	0.918	0.089	0.916	0.091	0.927	**0.077**	**0.939**
Heads	0.143	0.833	0.124	0.862	0.130	0.855	**0.106**	**0.889**
Office	0.128	0.855	0.096	0.912	0.105	0.897	**0.083**	**0.934**
Pumpkin	0.097	0.922	0.083	**0.946**	0.116	0.877	**0.078**	0.945
RedKitchen	0.124	0.853	0.101	0.896	0.108	0.884	**0.094**	**0.915**
Stairs	0.134	0.823	0.106	0.855	0.127	0.825	**0.104**	**0.857**

Monodepth2 [5] significantly on each scene. Further, compared to the model [10], [45], our method achieve the best performance on most scenes before and after fine-tuning using NYUv2 pretrained models, which demonstrates better generalizability and capability of our model. Moreover, the results show that our method can perform well in a variety of different scenes.

E. Ablation Studies

1) Effects of Each Proposed Module in MonoIndoor++: In this sub-section, we perform ablation studies of each core module in our proposed **MonoIndoor**++ on NYUv2 [14], ScanNet [19] and EuRoC MAV [18] datasets in Table VIII.

Specifically, We first perform ablation study for the residual pose estimation module. In Table VIII, from methods of the "Monodepth2 [5] (Baseline)" and "**MonoIndoor**++ **(Ours)**" with the "Residual Pose" column checked, improved performance can be observed by using the proposed residual pose estimation module. For instance, on NYUv2, the AbsRel is decreased from 16% to 14.2% and δ_1 is increased from 76.7% to 81.3%; on ScanNet, the AbsRel is decreased from 18.9% to 13.6% and the δ_1 is increased from 70.9% to 83.3%; on EuRoC MAV V2_01, the AbsRel is decreased from 15.7% to 14.1% and the δ_1 is increased from 78.6% to 81.5% and similar observations can be made on other test sequences as well.

Next, we experiment to validate the effectiveness of the depth factorization module. Comparing with Monodepth2 which predicts depth without any guidance of global scales, by adding the depth factorization module with a separate scale network in our MonoIndoor++ (see "**MonoIndoor**++ **(Ours)**" with the "Residual Pose" and "Depth Factorization" columns checked), we further observe improved performance on all datasets. For instance, on NYUv2, the AbsRel is decreased from 14.2% to 13.4% and δ_1 is increased from 81.3% to 82.3%; on ScanNet, the AbsRel is decreased from 13.6% to 12.6% and the δ_1 is increased from 83.3% to 83.9%; on EuRoC MAV V2_01, the AbsRel is decreased from 14.1% to 12.5% and the δ_1 is increased from 81.5% to 84.0% and similar observations can be made on other test sequences.

In addition, by using the residual pose estimation module with both the proposed depth factorization module and coordinates convolutional encoding module (see "**MonoIndoor**++ **(Ours)**" with all columns checked, the performance can be improved consistently. For instance, on NYUv2, the AbsRel is decreased to 13.2% and δ_1 is increased to 83.4%; on ScanNet, the AbsRel is decreased to 11.3% and the δ_1 is increased to 87.3%; on EuRoC MAV V2_01, the AbsRel is

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII

ABLATION RESULTS ON EACH CORE MODULE OF OUR **MONOINDOOR++** AND COMPARISON WITH THE BASELINE METHOD ON THE NYUV2 [14], SCANNET [19] AND EUROC MAV [18] DATASETS. BEST RESULTS ARE IN **BOLD**. RESIDUAL POSE: OUR RESIDUAL POSE ESTIMATION MODULE. DEPTH FACTORIZATION: OUR DEPTH FACTORIZATION MODULE WITH SCALE NETWORK. COORDINATES CONV. ENCODING: OUR COORDINATES CONVOLUTIONAL ENCODING MODULE

Method	Residual Pose	Depth Factorization	Coordinates Conv. Encoding	NYUv2					ScanNet				
				Error Metric ↓		Accuracy Metric ↑			Error Metric ↓		Accuracy Metric ↑		
				AbsRel	RMSE	δ_1	δ_2	δ_3	AbsRel	RMSE	δ_1	δ_2	δ_3
Monodepth2 [5] (Baseline)	✗	✗	✗	0.16	0.601	0.767	0.949	0.988	0.189	0.426	0.709	0.929	0.984
MonoIndoor++ (Ours)	✓	✗	✗	0.142	0.553	0.813	0.958	0.988	0.136	0.345	0.833	0.968	0.995
MonoIndoor++ (Ours)	✓	✓	✗	0.134	0.526	0.823	0.958	0.989	0.126	0.329	0.839	0.973	0.995
MonoIndoor++ (Ours)	✓	✓	✓	**0.132**	**0.517**	**0.834**	**0.961**	**0.990**	**0.113**	**0.302**	**0.873**	**0.979**	**0.996**

Method	Residual Pose	Depth Factorization	Coordinates Conv. Encoding	EuRoC MAV V1_03					EuRoC MAV V2_01				
				Error Metric ↓		Accuracy Metric ↑			Error Metric ↓		Accuracy Metric ↑		
				AbsRel	RMSE	δ_1	δ_2	δ_3	AbsRel	RMSE	δ_1	δ_2	δ_3
Monodepth2 [5] (Baseline)	✗	✗	✗	0.110	0.413	0.889	0.983	0.996	0.157	0.567	0.786	0.941	0.986
MonoIndoor++ (Ours)	✓	✗	✗	0.100	0.379	0.913	0.987	0.997	0.141	0.518	0.815	0.961	0.991
MonoIndoor++ (Ours)	✓	✓	✗	0.080	0.309	0.944	0.990	0.998	0.125	0.466	0.840	0.965	**0.993**
MonoIndoor++ (Ours)	✓	✓	✓	**0.079**	**0.303**	**0.949**	**0.991**	**0.998**	**0.115**	**0.439**	**0.861**	**0.972**	0.992

Method	Residual Pose	Depth Factorization	Coordinates Conv. Encoding	EuRoC MAV V2_02					EuRoC MAV V2_03				
				Error Metric ↓		Accuracy Metric ↑			Error Metric ↓		Accuracy Metric ↑		
				AbsRel	RMSE	δ_1	δ_2	δ_3	AbsRel	RMSE	δ_1	δ_2	δ_3
Monodepth2 [5] (Baseline)	✗	✗	✗	0.156	0.645	0.776	0.945	0.985	0.171	0.620	0.734	0.944	0.988
MonoIndoor++ (Ours)	✓	✗	✗	0.150	0.619	0.792	0.950	0.988	0.147	0.538	0.806	0.963	0.989
MonoIndoor++ (Ours)	✓	✓	✗	0.142	0.581	0.802	0.952	0.990	0.140	0.502	0.810	0.964	0.993
MonoIndoor++ (Ours)	✓	✓	✓	**0.133**	**0.551**	**0.830**	**0.964**	**0.991**	**0.134**	**0.482**	**0.829**	**0.967**	**0.993**

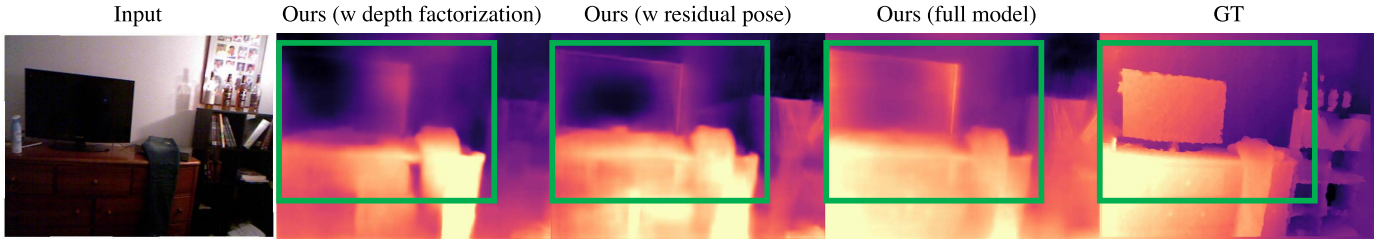


Fig. 5. Qualitative ablation comparisons of depth prediction on NYUv2. Our full model with both depth factorization and residual pose modules produce better depth maps.

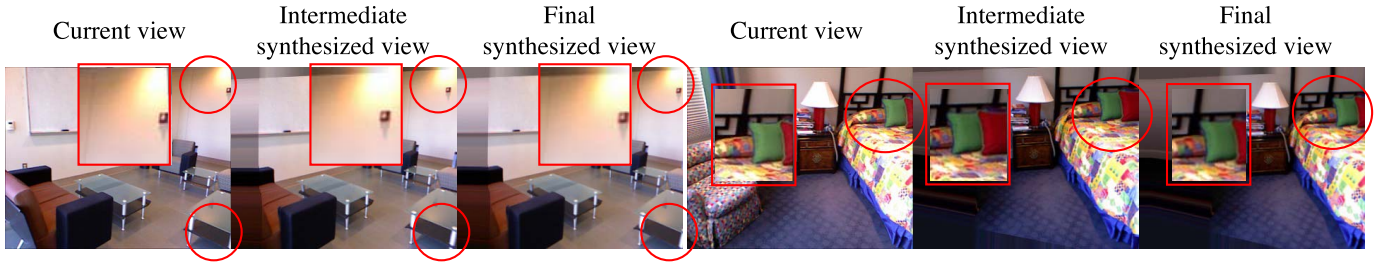


Fig. 6. Intermediate synthesized views on NYUv2.

decreased to 11.5% and the δ_1 is increased to 86.1% and similar observations can be made on other test sequences as well. Our full model achieves the best performance by giving competitive depth estimation results on these challenging datasets. We argue that these ablation results clearly prove the effectiveness of the each proposed module in our model, **MonoIndoor++**.

We also present the exemplar depth visualizations by our proposed modules on NYUv2 dataset in Figure 5. In addition, we visualize intermediate and final synthesized

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

TABLE IX

ABLATION RESULTS OF DESIGN CHOICES AND THE EFFECTIVENESS OF COMPONENTS IN THE TRANSFORMER-BASED SCALE REGRESSION NETWORK OF OUR MODEL (**MONOINDOOR++**) ON EUROC MAV V2_01 [18]. PORB. REG.: THE PROBABILISTIC SCALE REGRESSION BLOCK. NOTE: WE ONLY USE THE RESIDUAL POSE ESTIMATION MODULE WHEN EXPERIMENTING WITH DIFFERENT NETWORK DESIGNS FOR THE DEPTH FACTORIZATION MODULE

Network Design	Attention	Prob. Reg.	Error Metric ↓		Accuracy Metric ↑		
			AbsRel	RMSE	δ_1	δ_2	δ_3
I. ScaleCNN	✓	✓	0.140	0.518	0.821	0.956	0.985
II. ScaleNet	✓	✓	0.141	0.519	0.817	0.959	0.988
III. ScaleRegressor	✗	✗	0.139	0.508	0.817	0.960	0.987
III. ScaleRegressor	✓	✗	0.135	0.501	0.825	0.964	0.989
III. ScaleRegressor	✓	✓	**0.125**	**0.466**	**0.840**	**0.965**	**0.993**

views compared with the current view on NYUv2 in the Figure 6. Highlighted regions show that final synthesized views are better than the intermediate synthesized views and closer to the current view.

2) Effects of Network Design for Transformer-Based Scale Regression Network: We perform ablation studies for our network design choices for the transformer-based scale regression network in depth factorization module on the test sequence V2_01 of the EuRoC MAV dataset [18]. Firstly, we consider the following designs as the backbone of our scale regression network: I) a pre-trained ResNet-18 [79] followed by a group of Convolutional-BN-ReLU layers; II) a pre-trained ResNet-18 [79] followed by two residual blocks; III) a lightweight network with two residual blocks which shares the feature maps from the depth encoder as input. These three choices are referred to as the ScaleCNN, ScaleNet and ScaleRegressor, respectively in Table IX. Next, we validate the effectiveness of adding new components into our backbone design. As described in Section III-B, we mainly integrate two sub-modules: i) a transformer module and ii) a probabilistic scale regression block.

As shown in Table IX, the best performance is achieved by ScaleRegressor that uses transformer module and probabilistic scale regression. It proves that sharing features with the depth encoder is beneficial to scale estimation. Comparing the results of three ScaleRegressor variants, the performance gradually improves as we add more components (*i.e.*., attention and probabilistic scale regression (Prob. Reg.)). Specifically, adding the transformer module improves the overall performance over the baseline backbone; adding the probabilistic regression block leads to a further improvement, which validates the effectiveness of our proposed sub-modules.

3) Ablation Results of Coordinates Convolutional Encoding: We present ablation studies for the encoding position of the coordinates convolutional encoding module on the NYUv2 [14] and V2_01 of the EuRoC MAV [18] datasets in Table X. It should be mentioned that, to fully explore the effectiveness of using coordinates encoding technique, we only run our **MonoIndoor++** with the residual pose estimation module. We perform coordinates convolutional encoding with the following choices. Specifically, we first encode coordinates information with the color image pairs and extend coordinates convolutional layers to process combined input data. Second, we perform coordinates encoding operations with the feature representations outputted from the pose encoder and the processed features are taken as the input to the pose decoder. Third, we incorporate coordinates encoding operations with both input and features from pose encoder for pose estimation.

TABLE X

ABLATION RESULTS OF ENCODING POSITION FOR COORDINATES CONVOLUTIONAL WITH OUR **MONOINDOOR++** ON NYUV2. INIT.: INITIALIZATION OF WEIGHTS. NOTE: WE ONLY USE THE RESIDUAL POSE ESTIMATION MODULE WHEN EXPERIMENTING WITH DIFFERENT NETWORK DESIGNS FOR THE COORDINATES CONVOLUTIONAL ENCODING MODULE

Model	Encoding Position	NYUv2				
		Error Metric ↓		Accuracy Metric ↑		
		AbsRel	RMS	δ_1	δ_2	δ_3
MonoIndoor	✗	0.142	0.553	0.813	0.958	0.988
MonoIndoor++ (Random Init.)	Input	0.140	**0.543**	0.817	0.959	0.989
MonoIndoor++ (ImageNet Init.)	Input	**0.139**	0.545	**0.821**	**0.958**	**0.989**
MonoIndoor++ (ImageNet Init.)	Encoder Features	0.145	0.565	0.806	0.954	0.988
MonoIndoor++ (ImageNet Init.)	Input & Encoder Features	0.141	0.554	0.815	0.957	0.989
Model	Encoding Position	EuRoC MAV V2_01				
		Error Metric ↓		Accuracy Metric ↑		
		AbsRel	RMS	δ_1	δ_2	δ_3
MonoIndoor	✗	0.141	0.518	0.786	0.941	0.986
MonoIndoor++	Input	**0.130**	**0.492**	**0.840**	**0.965**	**0.992**

From the Table X, it can be observed that, by using coordinates convolutional encoding in residual pose estimation module, performance can be improved. For instance, the AbsRel is decreased to **13.9%** from 14.2% and the δ_1 is improved from 81.7% to **82.1%**. Besides, comparing with encoding coordinates information with feature representations after the pose encoder, applying the coordinates convolutional encoding operation over the input image pairs directly gives the best performance. Further, we test two different initialization methods for coordinates convolutional layers which are with random initializations or ImageNet-pretrained [80] initialization, respectively. The coordinates convolutional encoding layers which are initialized with ImageNet-pretrained weights give slightly improved performance compared to ones with random weights. Given the above observations, we further perform experiments under the same settings on EuRoC MAV V2_01 sequence, significant improvements have been observed for self-supervised monocular depth estimation by using our residual pose estimation module with coordinates convolutional encoding module, which can further validate the effectiveness of the coordinates convolutional encoding module.

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

V. Conclusion

In this work, a novel monocular self-supervised depth estimation framework, called the **MonoIndoor++**, has been proposed to predict depth map of a single image in indoor environments. The proposed model consists of three modules: (a) a novel *depth factorization module* with a transformer-based scale regression network which is designed to jointly learn a global depth scale factor and a relative depth map from an input image, (b) a novel *residual pose estimation module* which is proposed to estimate accurate relative camera poses for novel view synthesis of self-supervised training that decomposes a global pose into an initial pose and one or a few residual poses, which in turn improves the performance of the depth model, (c) a *coordinates convolutional encoding module* which is utilized to encode coordinates information explicitly to provide additional cues for the residual pose estimation module. Comprehensive evaluation results and ablation studies have been conducted on a wide-variety of indoor datasets, establishing the state-of-the-art performance and demonstrating the effectiveness and universality of our proposed methods.

References

[1] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2650–2658.

[2] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2002–2011.

[3] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 740–756.

[4] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1851–1858.

[5] C. Godard, O. M. Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3828–3838.

[6] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang, "Learning monocular depth by distilling cross-domain stereo networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 484–500.

[7] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[8] J. Zhou, Y. Wang, K. Qin, and W. Zeng, "Moving indoor: Unsupervised video depth learning in challenging environments," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8618–8627.

[9] W. Zhao, S. Liu, Y. Shu, and Y.-J. Liu, "Towards better generalization: Joint depth-pose learning without posenet," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 9151–9161.

[10] J. Bian, H. Zhan, N. Wang, T.-J. Chin, C. Shen, and I. Reid, "Auto-rectify network for unsupervised indoor depth estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 21, 2022, doi: 10.1109/TPAMI.2021.3136220.

[11] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5684–5693.

[12] Y. Zou, P. Ji, Q.-H. Tran, J.-B. Huang, and M. Chandraker, "Learning monocular visual odometry via self-supervised long-term modeling," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 710–727.

[13] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source slam system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[14] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGB-D images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.

[15] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2016, pp. 4104–4113.

[16] A. Dosovitskiy *et al.*, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–11.

[17] P. Ji, R. Li, B. Bhanu, and Y. Xu, "Monoindoor: Towards good practice of self-supervised monocular depth estimation for indoor environments," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12787–12796.

[18] M. Burri *et al.*, "The EuRoC micro aerial vehicle datasets," *Int. J. Robot. Res.*, vol. 35, no. 10, pp. 1157–1163, Sep. 2016.

[19] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5828–5839.

[20] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in RGB-D images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2930–2937.

[21] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2008.

[22] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 501–518.

[23] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," 2014, *arXiv:1406.2283*.

[24] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 239–248.

[25] S. F. Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 4009–4018.

[26] J. Li, R. Klein, and A. Yao, "A two-streamed network for estimating fine-scaled depth maps from single RGB images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3372–3380.

[27] M. Song, S. Lim, and W. Kim, "Monocular depth estimation using Laplacian pyramid-based depth residuals," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4381–4393, Nov. 2021.

[28] Y. Cao, T. Zhao, K. Xian, C. Shen, Z. Cao, and S. Xu, "Monocular depth estimation with augmented ordinal depth relationships," *IEEE T-CSVT*, vol. 30, no. 8, pp. 2674–2682, Oct. 2020.

[29] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3174–3182, Nov. 2018.

[30] B. Ummenhofer *et al.*, "DeMoN: Depth and motion network for learning monocular stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2017, pp. 5038–5047.

[31] Z. Teed and J. Deng, "DeepV2D: Video to depth with differentiable structure from motion," in *Proc. ICLR*, 2018, pp. 1–20.

[32] Z. Li and N. Snavely, "MegaDepth: Learning single-view depth prediction from internet photos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2041–2050.

[33] Z. Li *et al.*, "Learning the depths of moving people by watching Frozen people," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4521–4530.

[34] Z. Teed and J. Deng, "DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras," in *Proc. NeurIPS*, 2021, pp. 1–12.

[35] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, Mar. 2022.

[36] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12179–12188.

[37] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 270–279.

[38] J.-W. Bian *et al.*, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," in *Proc. NeurIPS*, 2019, pp. 1–11.

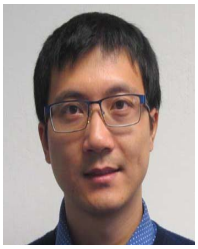
[39] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, "Learning depth from monocular videos using direct methods," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2022–2030.

[40] Z. Yin and J. Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1983–1992.

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

[41] Y. Zou, Z. Luo, and J.-B. Huang, "DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 36–53.
[42] R. Wang, S. M. Pizer, and J.-M. Frahm, "Recurrent neural network for (UN-) supervised learning of monocular video visual odometry and depth," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5555–5564.
[43] L. Tiwari, P. Ji, Q.-H. Tran, B. Zhuang, S. Anand, and M. Chandraker, "Pseudo RGB-D for self-improving monocular SLAM and depth prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 437–455.
[44] J. Watson, O. M. Aodha, V. Prisacariu, G. Brostow, and M. Firman, "The temporal opportunist: Self-supervised multi-frame monocular depth," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 1164–1174.
[45] J.-W. Bian, H. Zhan, N. Wang, T.-J. Chin, C. Shen, and I. Reid, "Unsupervised depth learning in challenging indoor video: Weak rectification to rescue," 2020, *arXiv:2006.02708*.
[46] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[47] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 1–11.
[48] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted Windows," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[49] Z. Liu *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 12009–12019.
[50] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NeurIPS*, 2015, pp. 1–9.
[51] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. NeurIPS*, 2015, pp. 1–9.
[52] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 652–660.
[53] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. NeurIPS*, 2017, pp. 1–10.
[54] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–25.
[55] K. Park *et al.*, "Nerfies: Deformable neural radiance fields," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5865–5874.
[56] M. Tancik *et al.*, "Block-NeRF: Scalable large scene neural view synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 8248–8258.
[57] R. Liu *et al.*, "An intriguing failing of convolutional neural networks and the CoordConv solution," in *Proc. NeurIPS*, 2018, pp. 1–12.
[58] F. Tian *et al.*, "Depth estimation using a self-supervised network based on cross-layer feature fusion and the quadtree constraint," *IEEE T-CSVT*, vol. 32, no. 4, pp. 1751–1766, Apr. 2022.
[59] S. Chen, Z. Pu, X. Fan, and B. Zou, "Fixing defect of photometric loss for self-supervised monocular depth estimation," *IEEE T-CSVT*, vol. 32, no. 3, pp. 1328–1338, Mar. 2022.
[60] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3286–3295.
[61] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2018, pp. 5410–5418.
[62] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, 2019, pp. 1–12.
[63] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic gradient descent," in *Proc. ICLR*, 2015, pp. 1–15.
[64] K. Karsch, C. Liu, and S. B. Kang, "Depthtransfer: Depth extraction from video using non-parametric sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2144–2158, Nov. 2014.
[65] M. Liu, M. Salzmann, and X. He, "Discrete-continuous depth estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2014, pp. 716–723.
[66] L. Ladicky, J. Shi, and M. Pollefeys, "Pulling things out of perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 89–96.
[67] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1119–1127.
[68] A. Roy and S. Todorovic, "Monocular depth estimation using neural regression forest," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5506–5514.
[69] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2015, pp. 5162–5170.
[70] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Towards unified depth and semantic prediction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2800–2809.
[71] A. Chakrabarti, J. Shao, and G. Shakhnarovich, "Depth from a single image by harmonizing overcomplete local network predictions," in *Proc. NeurIPS*, 2016, pp. 1–9.
[72] Z. Fang, X. Chen, Y. Chen, and L. V. Gool, "Towards good practice for CNN-based monocular depth estimation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1091–1100.
[73] Z. Yu, L. Jin, and S. Gao, "P^{2}Net: Patch-match and plane-regularization for unsupervised indoor depth estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 206–222.
[74] H. Zhan, C. S. Weerasekera, J.-W. Bian, and I. Reid, "Visual odometry revisited: What should be learnt?" in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2020, pp. 4203–4210.
[75] H. Alismail, B. Browning, and S. Lucey, "Photometric bundle adjustment for vision-based SLAM," 2016, *arXiv:1608.02026*.
[76] C. Tang and P. Tan, "BA-Net: Dense bundle adjustment network," 2018, *arXiv:1806.04807*.
[77] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao, "NeuralRecon: Real-time coherent 3D reconstruction from monocular video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 15598–15607.
[78] X. Gu, W. Yuan, Z. Dai, C. Tang, S. Zhu, and P. Tan, "DRO: Deep recurrent optimizer for structure-from-motion," 2021, *arXiv:2103.13201*.
[79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[80] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

Runze Li (Member, IEEE) received the B.E. degree in information engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2014, and the M.S. degree in information technology (distributed computing) from The University of Melbourne, VIC, Australia. He is currently pursuing the Ph.D. degree in computer science with the University of California at Riverside, Riverside, CA, USA.



Pan Ji received the Ph.D. degree from The Australian National University in 2016. He worked as a Researcher with NEC Labs America from February 2018 to September 2020. Before moving to USA in February 2018, he has been working as an ARC Senior Research Associate (Post-Doctoral Researcher) with The University of Adelaide since July 2016. From September 2020 to June 2022, he was a Senior Staff Research Engineer and the Manager of the OPPO US Research Center, InnoPeak Technology, Inc. He is currently the Director of visual perception with XR Vision Labs, Tencent. His research interests lie in computer vision (especially 3D vision), unsupervised learning (such as clustering), and various other aspects of machine learning. He received the Best Student Paper Award at the International Conference on Image Processing (ICIP) 2014.

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.

Yi Xu received the Ph.D. degree from Purdue University in 2010. He is currently the Director of XR Technology with the OPPO US Research Center, InnoPeak Technology, Inc. Before joining InnoPeak, he worked at various industrial labs, such as GE Research and JD.COM Silicon Valley Labs. His research interests lie in 3D computer graphics and computer vision, with a focus on extended reality.

Bir Bhanu (Life Fellow, IEEE) received the B.S. degree (Hons.) from IIT-BHU, the M.E. degree (Hons.) from BITS Pilani, the S.M. and E.E. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, and the M.B.A. degree from the University of California at Irvine, Irvine, CA, USA. He is currently the Bourns Endowed University of California Presidential Chair in engineering, a Distinguished Professor of electrical and computer engineering, and the Founding Director of the Interdisciplinary Center for Research in Intelligent Systems from 1998 to 2019, and the Visualization and Intelligent Systems Laboratory since 1991 at the University of California at Riverside (UCR), Riverside, CA, USA. He is the Founding Professor of electrical engineering with UCR and served as its first Chair from 1991 to 1994. He has been a Cooperative Professor of computer science and engineering since 1991, bioengineering since 2006, and mechanical engineering since 2008. Recently, he served as the Interim Chair for the Department of Bioengineering from 2014 to 2016. He also served as the Director for the National Science Foundation Graduate Research and Training Program in video bioinformatics with UCR. Prior to joining UCR in 1991, he was a Senior Honeywell Fellow with Honeywell Inc. He holds U.S. and international patents. He received the Faculty Research Lecturer Award from UCR in 2019 commencement. He has published extensively and received university and industry awards for research excellence, outstanding contributions, and team efforts; and journal/conference best paper awards. His research interests include computer vision; pattern recognition and data mining; machine learning; artificial intelligence; image processing; image and video database; graphics and visualization; robotics; human–computer interactions; and biological, medical, military, and intelligence applications. He is a fellow of AAAS, IAPR, SPIE, NAI, and AIMBE.

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on February 04,2023 at 07:47:54 UTC from IEEE Xplore. Restrictions apply.