SC-DepthV3: Robust Self-Supervised Monocular Depth Estimation for Dynamic Scenes

Libo Sun, Jia-Wang Bian, Huangying Zhan, Wei Yin, Ian Reid, and Chunhua Shen

Abstract—Self-supervised monocular depth estimation has shown impressive results in static scenes. It relies on the multi-view consistency assumption for training networks, however, that is violated in dynamic object regions and occlusions. Consequently, existing methods show poor accuracy in dynamic scenes, and the estimated depth map is blurred at object boundaries because they are usually occluded in other training views. In this paper, we propose SC-DepthV3 for addressing the challenges. Specifically, we introduce an external pretrained monocular depth estimation model for generating single-image depth prior, namely pseudo-depth, based on which we propose novel losses to boost self-supervised training. As a result, our model can predict sharp and accurate depth maps, even when training from monocular videos of highly dynamic scenes. We demonstrate the significantly superior performance of our method over previous methods on six challenging datasets, and we provide detailed ablation studies for the proposed terms.

Index Terms—Monocular depth estimation, unsupervised learning, self-supervised learning, knowledge distillation.

I. INTRODUCTION

MONOCULAR depth estimation [2] has attracted great attention in computer vision. It provides valuable cues for various downstream tasks, such as semantic image segmentation [3], salient object detection [4], 3D reconstruction [5], novel view synthesis [6], and visual odometry [1], [7]. Early work [2], [8] solves the monocular depth estimation problem by using supervised learning. However, these methods rely on ground-truth depth labels that are not always available in real-world scenes. To address this limitation, self-supervised monocular depth estimation methods were proposed and showed that a depth network could be trained from stereo image pairs [9] or monocular videos with ego-motions [10] without the need for ground-truth depth labels. We focus on self-supervised learning of monocular depth from videos since only a single camera is required to collect training data in this setup, which has great potential for advancing real-world applications.

Self-supervised methods typically rely on the multi-view consistency assumption for training networks, e.g., the photometric loss [10] and geometry consistency loss [1] that were used in previous methods. This assumption provides effective constraints for learning scene geometry, while it is violated at regions with occlusion (e.g., object boundaries) and moving objects. Therefore, existing methods often show only excellent results in (almost) static scenes such as KITTI [11] and NYUv2 [12] datasets. When training on more challenging dynamic datasets that have an amount of fast-moving objects, previous state-of-the-art methods [1], [13], [14] show poor accuracy. Moreover, the estimated depth map is blurred at object boundaries because they are usually occluded in other training views. We illustrate several examples of qualitative monocular depth results in Fig. 1.

To address the issues caused by moving objects and occlusions, existing approaches usually detect these bad regions and then exclude them from training. The methods can be categorized into four classes according to how they detect dynamic regions, involving in the prediction-based [10], semantic-based [15], [16], [17], [18], flow-based [19], [20], [21], [22], and geometry-based [1]. These methods can reduce corruption from noisy losses during training and generally improves overall accuracy, however, it leads to poor results on dynamic regions at inference time because these regions are not sufficiently regularized in training. There are also more sophisticated approaches [23], [24] that model the velocity of each moving object in multiple views, but they rely on solving a challenging problem in themselves.

We propose SC-DepthV3 in this paper, which addresses the above-mentioned issues by leveraging external single-image constraints. Specifically, we leverage an off-the-shelf monocular depth estimation model [25] to generate the single-image depth prior, which we term pseudo-depth. Based on it, we propose effective losses to constrain the depth estimation network in self-supervised learning. Here, we use LeReS [25] for generating pseudo-depth, which is trained in large-scale datasets with supervised learning and enables zero-shot generalization in previously unseen data. The excellent qualitative results have been demonstrated in [25], while we find that pseudo-depth may show low quantitative accuracy. Fig. 3 gives an example, where we visualize the error map of pseudo-depth by comparing it with
We conduct comprehensive experiments and ablation studies to validate our deeper understanding of the issues in self-supervised depth estimation. The results demonstrate the efficacy of our proposed methods.

Our contributions are as follows:

- We propose SC-DepthV3 for robust self-supervised learning of monocular depth in highly dynamic scenes, which allows for predicting accurate and sharp depth maps.
- We propose Dynamic Region Refinement (DRR) and Local Structure Refinement (LSR) modules, which are based on pseudo-depth to boost self-supervised learning.
- We conduct comprehensive experiments and ablation studies on six challenging datasets. The results demonstrate the efficacy of our proposed methods.

II. RELATED WORK

**Self-Supervised Monocular Depth Estimation:** Garg et al. [9] proposed to train monocular depth estimation models on stereo image pairs by using the photometric loss. Zhou et al. [10] proposed to train the depth estimation model on videos by jointly training a pose estimation model. Following them, many advanced techniques [1, 13, 14, 17, 19, 22, 27, 28, 29, 30] were proposed to boost the performance. However, multi-view ambiguities make the self-supervised method hard to handle dynamic objects and object boundaries. Previous methods either excluded these regions from training [1, 15, 16, 19] or modeled the object motions [23, 24], but both solutions have their drawbacks. More specifically, simply excluding dynamic regions would result in poor accuracy on these regions at the inference time, and modeling each object’s motion is ill-posed and may not be robust in dynamic scenes.

The key to solving the dynamic region issue is the proposed Dynamic Region Refinement (DRR) module. The method is inspired by an observation, i.e., we find that pseudo-depth maintains excellent depth ordinal (the further/nearer relations) between any two objects or pixels. To capitalize on these findings, we propose to extract the “ground-truth” depth ordinal information between dynamic and static regions (from pseudo-depth) and use it to regularize the self-supervised depth estimation in dynamic regions. Specifically, we sample point pairs between two regions and apply depth ranking loss [26]. This is effective because the static backgrounds have already been well-supervised by multi-view losses, and the dynamic regions could be uniquely localized by sampling sufficient point pairs between dynamic and static regions. Our method is also based on the fact that the depth ordinal in pseudo-depth is sufficiently accurate [25]. Furthermore, to segment dynamic regions from static backgrounds, we use the self-discovered mask that was proposed in SC-Depth [1] and generated by computing forward-backward depth inconsistency in self-supervised training, so the external segmentation networks are not required. Fig. 4 illustrates the proposed DRR module.

Moreover, we observe that pseudo-depth shows smooth local structures and clean object boundaries. This motivates us to propose a Local Structure Refinement (LSR) module to improve the self-supervised depth estimation w.r.t. depth details. The proposed module contains two parts. On the one hand, we extract the surface normal from both pseudo-depth and network-predicted depth, and we constrain them to be consistent by applying a normal matching loss. This improves the overall depth significantly. On the other hand, we constrain depth estimation at object boundary areas by applying our proposed relative normal angle loss. More specifically, we sample point pairs around image edges and enforce their relative normal angles to be consistent between pseudo-depth and self-supervised depth. As a result, our method improves qualitative depth estimation results significantly, particularly at object boundaries. Fig. 1 shows several examples of the qualitative depth estimation results.
scenes. Compared with them, our method leverages pretrained single-image prior for resolving multi-view ambiguities, leading to a SOTA self-supervised depth estimation method. More recent methods include [31], [32], [33].

**SC-Depth Series Methods:** This paper is the third version of the SC-Depth series methods. In the SC-Depth [1], we addressed the scale inconsistency issue, so our method enables scale-consistent depth estimation over the video, which is beneficial to video-based tasks such as Visual SLAM. In the SC-DepthV2 [30], we analyzed the rotation issue in videos that are captured by handheld cameras, and we proposed an auto-rectify network to handle the large rotation. The V1 and V2 have shown great accuracy in both indoor and outdoor scenes. However, their predicted depth maps are blurred at object boundaries, and they suffer in highly dynamic scenes. In this paper, we propose SC-DepthV3 address the issue of dynamic objects and blurred object boundaries.

**Zero-Shot Monocular Depth Estimation:** Many existing methods leverage large-scale datasets and supervised training [25], [34], [35], [36], [37], [38], [39], [40] to train monocular depth estimation models towards zero-shot generalization on unseen data. For example, [34], [35], [36], [37], [38], [39] collect stereo images/videos from the internet and use geometric reconstruction tools [41], [42] to generate dense ground-truth depth labels. [40] export perfect ground-truth depths from the synthetic 3D movies [43]. Recently, LeReS [25] and DPT [44] achieve the state-of-the-art performance. However, note that their predicted depths are scale-shift-invariant, due to the high diversity of different scenes, which show low quantitative accuracy in out-of-distribution data and cannot be used for 3D reconstruction. Nevertheless, we find that their predicted depths carry good attributes that could be leveraged for boosting self-supervised learning of monocular depth estimation. Compared with these methods, our method enables consistent and accurate depth estimation for video-based tasks such as Visual SLAM, which has been demonstrated in SC-Depth [1], thanks to the scale-consistency constraints.

**Knowledge Transfer:** Our method is also related to knowledge transfer approaches, because the proposed method can be regarded as transferring the knowledge of pretrained monocular depth estimation models [25] to our self-supervised trained models. However, we argue that our method is very different from previous knowledge transfer or distillation methods. On the one hand, knowledge is often transferred by finetuning pretrained models in new datasets [45], which is not our case and cannot solve the challenges in our problem. The main issue in our problem is the imperfect self-supervised loss, so even if we finetune pretrained models (i.e., it provides a good initialization), the model would become worse and worse with training due to the deficient self-supervised loss functions. On the other hand, knowledge transfer could also be achieved by conducting semi-supervised learning on mixed datasets. Specifically, we can train models on both previous large-scale datasets with ground-truth labels and new datasets without annotations, and then we apply supervised loss in the former and self-supervised loss in the latter. However, this involves new challenges of mix-data training, long training time, and the maintenance of large-scale previous data. In contrast, our method is more elegant than semi-supervised training, since the teacher model is trained only once on large-scale datasets and can be used as an off-the-shelf tool to generate pseudo-depth in new scenes. Moreover, our student model shows significantly higher accuracy than the teacher model, which is rare in the field of knowledge distillation.

## III. Method

Fig. 2 illustrates an overview of the proposed method. First, our method is based on SC-Depth [1] for basic self-supervised training, which we describe in detail in Section III-A. Second, we discuss the single-image depth prior in Section III-B that is generated by using the off-the-shelf monocular depth estimation methods and used in our method for generating auxiliary supervision signals. Finally, we describe the Dynamic Region Refinement (DRR) in Section III-C and Local Structure Refinement (LSR) modules in Section III-D, respectively, which are the proposed terms to boost self-supervised training.

### A. Self-Supervised Depth Learning (SC-Depth)

In the self-supervised learning framework, a monocular depth estimation network (DepthNet) and a relative 6-DoF camera pose estimation network (PoseNet) are jointly trained on a large number of monocular videos. First, given a consecutive image pair $(I_a, I_b)$ randomly sampled from a training video, we predict their depths $(D_a, D_b)$ by forwarding the DepthNet and estimate their relative 6-DoF camera pose $P_{ab}$ by forwarding the PoseNet. Then, we generate the warping flow between two images using the predicted depth and pose, followed by synthesizing the $I'_b$ using the flow and $I_b$ via bi-linear interpolation. Finally, we penalize the color inconsistencies between $I_b$ and $I'_b$, and we also constrain the geometry consistency between $D_a$ and $D_b$, which back-propagates the gradients to the networks. The objective function is described below.

First, we use the geometry consistency loss $L_G$ [1] to encourage the predicted depths $(D_a, D_b)$ to be consistent with each other in 3D space. Formally,

$$L_G = \frac{1}{|V|} \sum_{p \in V} D_{\text{diff}}(p),$$

where $V$ stands for valid points that are projected inside the image. $D_{\text{diff}}$ stands for the pixel-wise depth inconsistency between $D_a$ and $D_b$, which is detailed explained in [1]. With it, we can obtain the self-discovered mask:

$$M_s = 1 - D_{\text{diff}},$$

which assigns lower weights to dynamics and occlusions than static regions, since the former is geometrically inconsistent across multiple views. We use this mask in our proposed DRR module (Section III-C) to localize dynamic regions.

Second, we use the weighted photometric loss $L_P^M$ to constrain the warping flow between $I_a$ and $I_b$ that is generated by the $D_a$ and $P_{ab}$. Formally,

$$L_P^M = \frac{1}{|V|} \sum_{p \in V} (M_s(p) \cdot L_P(p)),$$

$$L_P = \frac{1}{|V|} \sum_{p \in V} \left( \lambda \| I_a(p) - I'_a(p) \|_1 + (1 - \lambda) \frac{1 - \text{SSIM}_{ab}(p)}{2} \right),$$
Fig. 2. Method overview. First, given a training sample (i.e., \( I_a \) and \( I_b \) two images), we follow SC-Depth [1] to compute self-supervised losses \( L_{Self} \), which is described in Section III-A. Second, we generate pseudo-depth \( PD_a \) using a pretrained depth estimation network, which is discussed in Section III-B. Finally, we propose DRR and LSR modules to constrain the network prediction \( D_a \) by using \( PD_a \), which are presented in Sections III-C and III-D, respectively.

where \( I'_a \) is synthesized from \( I_b \) using the warping flow, and SSIM [46] is a widely-used metric to measure image similarity. We set \( \lambda \) to 0.15 as in [1].

Third, we use the edge-aware smoothness loss to regularize the predicted depth map. Formally,
\[
L_S = \sum_p (e^{-\nabla I_a(p) \cdot \nabla D_a(p)})^2,
\]
where \( \nabla \) is the first derivative along spatial directions, which guides smoothness by image edges.

Overall, our objective function is formulated as follows:
\[
L_{Self} = \alpha L_P^M + \beta L_G + \gamma L_S.
\]
We set \( \alpha = 1 \), \( \beta = 0.5 \), and \( \gamma = 0.1 \) as in [1]. Note that we will replace \( L_S \) with the proposed normal loss \( L_N \) in Section III-D. Moreover, we also use the auto-masking and per-pixel minimum reprojection loss that are proposed in [13] to filter stationary and non-best points during training.

B. Single-Image Depth Prior

Our idea is to leverage the pretrained monocular depth estimation network for generating single-image depth prior, which is then used to boost self-supervised learning. Here we use LeReS [25] to generate pseudo-depth, which is trained on large-scale datasets with ground-truth depth labels. Thanks to the supervised training on large-scale data, it shows excellent zero-shot generalization performance on unseen scenes. Note that LeReS was not trained on datasets that we use in this paper for evaluation. An example of LeReS outputs is shown in Fig. 3, where it shows plausible visual results on the DDAD dataset but poor quantitative accuracy.

These phenomena echo our motivation, i.e., pseudo-depth is not accurate enough but has potential that can be leveraged for boosting self-supervised learning. More specifically, we find that the good visual results are contributed to several aspects, including (a) correct depth ordinal (nearer/further relation) between objects; (b) excellent smoothness in predicted depth; (c) sharp depth prediction at object boundaries.

Therefore, based on the above observation, we propose two modules to extract effective supervision signals from pseudo-depth. First, we propose a Dynamic Region Refinement (DDR) module that regularizes self-supervised training with a depth ranking constraint, particularly boosting depth prediction on moving objects. Second, we propose a Local Structure Refinement (LSR) module that constrains the smoothness and object boundaries of the predicted depth. The proposed two modules are presented in Sections III-C and III-D, respectively.
C. Dynamic Region Refinement

The key to our proposed dynamic region refinement (DRR) module is constraining depth estimation on dynamic regions by enforcing the nearest/further relation w.r.t. the predicted depths on static regions. Specifically, this is based on two assumptions including i) The accurate depth ranking relations between any two pixels can be extracted from pseudo-depth; ii) depth prediction on static regions is sufficiently accurate thanks to the self-supervised losses. The assumptions are valid, as demonstrated in prior work [13], so our idea is generally effective. In the proposed DRR module, we first sample point pairs between dynamic and static regions, where the segmentation of images is obtained in a self-supervised manner (2). Then we compute depth ranking loss on sampled point pairs to regularize the predicted depth map. The proposed sampling method and loss function are presented in the following paragraphs, and Fig. 4 illustrates a training example.

Dynamic-Focused Sampling: To sample point pairs between static and dynamic regions, we need the segmentation of training images. This could be achieved with the use of pretrained semantic segmentation networks, e.g., we can assume objects of certain classes such as vehicles as moving objects and others as static backgrounds. However, it involves extra data preprocessing and pretrained networks. Moreover, the dynamic of an object does not necessarily rely on the semantic classes, e.g., a chair can be dynamic while a person is moving it around. Instead, we derive dynamic/static segmentation from the self-discovered mask (2) that is computed based on the geometric consistency. It is a soft weight mask and assigns smaller values for depth-inconsistent regions (dynamics or occlusions) than others (static regions). To obtain binary segmentation, we propose to rank weights and pick the lowest 20% as potential dynamic regions, rather than doing hard thresholding. Here we assume that the ratio of moving objects pixels is around 20% or less, which is true in most real-world scenes. Then for each point in the dynamic regions, we pair it with a point that is randomly sampled from static regions. Moreover, other than constructing dynamic-static pairs as discussed above, we also sample point pairs randomly from the whole image, which serves as an additional global regularization.

Confident Depth Ranking Loss: We compute the depth ranking loss on the sampled point pairs in training. The original loss function was proposed in [26]. Formally, for a pair of points with predicted depth values \( p_0, p_1 \), the loss is

\[
\phi(p_0, p_1) = \begin{cases} 
\log(1 + \exp(-\ell(p_0 - p_1))), & \ell \neq 0 \\
 - \frac{1}{2}t, & \ell = 0
\end{cases}
\]

where \( t \) is the ground truth ordinal label, which can be induced by a ground truth depth map:

\[
\ell = \begin{cases} 
+1, & p_0^*/p_1^* \geq 1 + \tau, \\
-1, & p_0^*/p_1^* \leq \frac{1}{1+\tau}, \\
0, & \text{otherwise}.
\end{cases}
\]

Here \( \tau \) is a threshold, which is 0.03 in previous work [47], and \( p^\ast \) denotes pseudo-depth.

We empirically find that (7) is sub-optimal in our method since pseudo-depth is not as accurate as the ground-truth depth. Therefore, we have to take the confidence of pseudo-depth ordinals into consideration. Specifically, we observe that the ordinal is often sufficiently reliable when two points have sufficiently different depth values, i.e., when \( p_0^*/p_1^* \gg 1 \) or \( p_0^*/p_1^* \ll 1 \), and otherwise, it may be unreliable when two depth values are very close, i.e., when \( p_0^*/p_1^* \approx 1 \).

Based on the above observation, we propose to (a) increase \( \tau \) from 0.03 to 0.15 for higher tolerance; and (b) ignore point pairs that have \( \ell = 0 \). Formally, we reformulate (7) as

\[
\phi(p_0, p_1) = \log(1 + \exp(-\ell(p_0 - p_1))),
\]

Therefore, our Confident Depth Ranking Loss is defined as:

\[
L_{CDR} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \phi(p),
\]

where \( \Omega \) stands for the sampled point pairs that have \( \ell \neq 0 \).

D. Local Structure Refinement

As mentioned above in Section III-B, pseudo-depth provides excellent depth smoothness. In this section, we propose to leverage such attributes to regularize the self-supervised depth. Our idea is to i) constrain the surface normals that are derived from predicted depths and pseudo-depths to be matched; and ii) constrain two depth maps to be consistent w.r.t. relative normal angles of sampled point pairs around edges. Here, the first step is focused on refining the overall depth structures, and the second step is focused on improving object boundary regions. The details are provided below.

Normal Matching Loss: The edge-aware depth smoothness loss (5) is often used in self-supervised monocular depth estimation, and it is also used in our baseline. Here, we propose to replace it with the normal matching loss:

\[
L_N = \frac{1}{N} \sum_{i=1}^{N} ||n_i - n_i^\ast||_1,
\]

where \( n_i \) is the surface normal derived from the predicted depth, and \( n_i^\ast \) is the normal derived from pseudo-depth. \( N \) stands for the total number of pixels in the image. The pixel-wise loss function provides strong supervision for overall depth structures.

Edge-Aware Relative Normal Loss: Not only do overall structure refinement, but also we focus on object boundary areas. Specifically, we sample point pairs around image edges and constrain the relative normal angles of sampled point pairs to
be consistent with pseudo-depth. Here, we use edge-guided sampling that was proposed in [47] to construct point pairs \( \langle A, B \rangle \), and we define the Edge-aware Relative Normal Loss as:

\[
L_{ERN} = \frac{1}{N} \sum_{i=1}^{N} \| n_{A_i} \cdot n_{B_i} - n_{A_i}^* \cdot n_{B_i}^* \|_1,
\]  

(12)

where \( n_A \) denotes the normal of a sampled point from the predicted depth, and \( * \) denotes pseudo-depth. Combining the edge-guided sampling and relative normal loss, we can effectively constrain the depth estimation on object boundary regions.

The proposed \( L_{ERN} \) is similar to the pair-wise normal loss that is proposed in [25], while the latter samples point pairs from edges, planes, and whole images. In contrast, we sample points solely from edges because sampling from other regions requires high-quality ground-truth depth. In our case, pseudo-depth is not accurate enough to maintain high-quality global structures, hence we only constrain the local structure. We analyze this effect with a detailed ablation study in Section IV-C.

E. Training

**Losses:** Based on the proposed two refinement modules, we rewrite the overall loss function (6) of our baseline and obtain the new objective function as:

\[
L = \alpha L^M + \beta L_G + \gamma L_N + \delta L_{CDR} + \epsilon L_{ERN},
\]  

(13)

where we set \( \alpha = 1, \beta = 0.5, \) and \( \gamma = \delta = \epsilon = 0.1 \) in training based on empirical tuning.

**Networks:** Our depth and pose networks are the same as previous work [11, 13], where we use ResNet-18 [48] backbone for both depth and pose estimation networks. The depth network is a U-Net structure [49] with a DispNet [10] as the decoder. The activations are sigmoids at the output layer and ELU nonlinearities [50] elsewhere. We convert the sigmoid output \( x \) to depth with \( D = 1/(ax + b) \), where \( a \) and \( b \) are chosen to constrain \( D \) between 0.1 and 100 units. The pose network accepts two RGB frames as input and outputs the 6D relative pose. We modify the first layer of ResNet-18 to have six channels for accepting two-frame inputs, and features are decoded to 6-DoF parameters via four convolutional layers.

**Training Details:** We implement the proposed method using the PyTorch library [51]. Following [10, 21, 52], we use a snippet of three sequential video frames as a training sample. The images are augmented with random scaling, cropping, and horizontal flips during training. We use the Adam [53] optimizer and set the learning rate to be 10^{-4}. We initialize the encoder by using the pre-trained model on ImageNet [54]. We train our networks in 100 \( k \) iterations on each dataset.

IV. EXPERIMENT

A. Datasets and Evaluation Metrics

The proposed method focuses on boosting self-supervised monocular depth estimation in challenging dynamic scenes, so we mainly evaluate our methods on three dynamic datasets, including DDAD driving dataset [14], BONN dynamic dataset [55], and TUM dataset [56] (dynamic object split). Note that these datasets contain fast-moving objects, which are much more challenging than the widely-used KITTI [11] and NYUv2 [12] datasets. We assume that the latter two datasets are almost static in this paper, and we also report results on them. All the mentioned self-supervised methods are trained on each dataset individually for a fair comparison. Moreover, following previous methods, we analyze the depth results at object boundaries and plane regions in the IBims-1 dataset [57]. In the following paragraphs, the details of each dataset are described.

**DDAD:** The dataset contains 200 driving videos that are captured in urban scenes. The LiDAR scanned point clouds are provided, which we use to generate sparse ground-truth depths for evaluation. In this dataset, almost vehicles are moving on the road, and there are fewer stopping cars than KITTI, making it more challenging to train self-supervised models. We use the standard training/testing split, which has 150 training scenes (12650 images) and 50 validation scenes (3950 images). We use the validation scenes for evaluation. Depth ranges are capped to at most 200 meters, and images are resized to the resolution of 640 \( \times \) 384 for training depth and pose networks.

**BONN:** The dataset contains 26 dynamic indoor videos that have fast-moving people or other objects. The Kinect captured depth maps are provided as the ground truth for evaluation. We manually find 4 challenging video sequences with fast-moving people (1785 images) for testing, and we use the remaining videos for training. Depth ranges are capped at 10 meters, and images are resized to the resolution of 320 \( \times \) 256 for training networks.

**KITTI:** The dataset provides a collection of indoor videos with Kinect-captured depth maps as the ground truth. We choose only videos that belong to the Dynamic Objects category, making sure that the model is trained in dynamic scenes. There are in total 11 sequences, and we use the last two sequences that contain moving people (1375 images) for testing. The remaining 9 dynamic videos are used for training, and images are resized to the resolution of 320 \( \times \) 256 for feeding to networks.

**NYUv2:** The dataset provides a large collection of indoor videos, and it is widely-used in the computer vision community. There are 654 testing images of static scenes for depth evaluation, and we use the remaining videos that do not contain testing images for training neural networks. Images are resized to the resolution of 320 \( \times \) 256 before feeding to the network. Note that this dataset contains almost-static scenes.

**IBims-1:** The dataset provides 100 accurate and dense ground truths for analyzing depth details, including object boundaries and planes. Images are collected in different kinds of indoor environments, and it does not provide a training set. For a fair comparison with previous work, we use the model trained on the NYUv2 dataset for all methods.

**Depth Evaluation Metrics:** We use standard depth evaluation metrics, including mean absolute relative error (AbsRel),
TABLE I
SELF-SUPERVISED MONOCULAR DEPTH ESTIMATION RESULTS ON THE DDAD DRIVING DATASET [14]

| Methods         | Full Image |             |             | Dynamic |             |             | Static |             |
|-----------------|------------|-------------|-------------|---------|-------------|-------------|--------|-------------|
|                 | AbsRel     | SqRel       | RMS         | RMSLog  | δ₁          | δ₂          | δ₃     | AbsRel      | δ₁          | AbsRel      | δ₁          |
| Monodepth2 [13] | 0.239      | 12.547      | 18.392      | 0.316   | 0.752       | 0.899       | 0.949  | 0.747       | 0.432       | 0.188       | 0.771       |
| PackNet [14]    | 0.182      | 7.945       | 15.021      | 0.259   | **0.828**   | **0.925**   | 0.961  | 0.564       | 0.520       | **0.137**   | **0.843**   |
| SGDDepth [18]   | 0.200      | 7.944       | 17.149      | 0.289   | 0.769       | 0.911       | 0.957  | 0.619       | 0.446       | 0.170       | 0.786       |
| DynamicDepth [58]| **0.156** | **3.305**   | 15.612      | **0.258**| 0.785       | 0.914       | **0.962**| **0.258**   | **0.612**   | 0.149       | 0.792       |
| SC-Depth [1]    | 0.169      | 3.877       | 16.290      | 0.280   | 0.773       | 0.905       | 0.951  | 0.345       | 0.546       | 0.155       | 0.783       |
| Ours w/o DRR    | 0.153      | 3.124       | **15.237**  | 0.252   | 0.799       | 0.920       | 0.963  | 0.259       | 0.612       | 0.146       | 0.806       |
| Ours w/o LSR    | 0.149      | 3.094       | 16.198      | 0.262   | 0.794       | 0.913       | 0.956  | 0.210       | 0.666       | 0.146       | 0.799       |
| Ours            | **0.142** | **3.031**   | **15.868**  | **0.248**| **0.813**   | **0.922**   | **0.963**| **0.199**   | **0.697**   | **0.140**   | **0.813**   |

We segment vehicles and pedestrians as dynamic objects and consider the remaining regions as static backgrounds. This dataset is more challenging than KITTI due to more complex scenes, fewer stopping cars, and longer depth ranges (200 m versus 80 m). Note that DynamicDepth [58] uses two frames for depth estimation.

TABLE II
SELF-SUPERVISED MONOCULAR DEPTH ESTIMATION RESULTS ON THE BONN DYNAMIC DATASET [57]

| Methods         | Full Image |             |             | Dynamic |             |             | Static |             |
|-----------------|------------|-------------|-------------|---------|-------------|-------------|--------|-------------|
|                 | AbsRel     | RMS         | δ₁          | δ₂      | δ₃          | AbsRel      | δ₁     | AbsRel      |
| Monodepth2 [13] | 0.565      | 2.337       | 0.352       | 0.591   | 0.728       | 0.474       | 0.172  | 0.594       | 0.383       |
| SC-Depth [1]    | 0.272      | 0.733       | 0.623       | 0.858   | 0.948       | 0.704       | 0.166  | 0.180       | 0.714       |
| SC-Depth V2 [30]| **0.211** | **0.619**   | **0.714**   | **0.873**| **0.936**   | 0.488       | **0.247**| **0.152**   | **0.803**   |
| Ours w/o DRR    | 0.138      | 0.396       | 0.885       | 0.951   | 0.974       | 0.248       | 0.690  | 0.106       | 0.939       |
| Ours w/o LSR    | 0.130      | 0.382       | 0.874       | 0.951   | 0.977       | 0.274       | 0.613  | 0.097       | 0.937       |
| Ours            | **0.126** | **0.379**   | **0.889**   | **0.961**| **0.980**   | **0.220**   | **0.720**| **0.102**   | **0.931**   |

This dataset is super-challenging because all training and testing videos contain fast-moving objects, which occupy a large proportion of pixels.

B. Evaluation Results

Results on Dynamic Datasets: We use three dynamic datasets mentioned above to evaluate the proposed method, and the quantitative depth estimation results are reported in Tables I, II, and III, respectively. We show the qualitative comparison results in Fig. 5, and demo videos for depth estimation are in available online at github. A more detailed analysis is conducted below.

Table I shows the results on DDAD dataset, where we compare our method with previous state-of-the-art methods, including Monodepth2 [13], PackNet [14], and SC-Depth [1]. The results show that our method outperforms previous methods by a large margin, and particularly on dynamic regions. Note that our method outperforms PackNet [14], although the latter uses a significantly larger network backbone than ours. This demonstrates our main contribution in this paper, i.e., robust learning of monocular depth in dynamic scenes. Besides, we also report the result without our proposed DRR and LSR modules. Here our baseline method is a modified version of SC-Depth, and it incorporates the advantages of Monodepth2. The results show that the performance of these models is significantly lower than
that of our full model, which demonstrates the efficacy of our proposed losses.

Tables II and III show the depth estimation results on BONN and TUM datasets, respectively. These indoor datasets are more challenging than driving datasets such as DDAD since the ratio of dynamic regions to the full image of the former is significantly larger than that of the latter. Consequently, previous methods such as Monodepth2 [13] and SC-Depth [1] show poor accuracy in BONN and TUM datasets. Compared with these approaches, our method presents significantly better results. This is contributed to our proposed losses, which enables our method to learn depth estimation robustly from dynamic videos.

**Results on Static Datasets:** Although our main contribution in this paper is boosting self-supervised monocular depth in dynamic scenes, we show that our method is also working well in almost-static scenes. The results are reported in the widely-used KITTI driving dataset and NYUv2 indoor dataset. Sampled qualitative results are illustrated in Fig. 6.

Table IV shows the depth estimation results on KITTI, where our method is comparable but does not outperform the previous state-of-the-art methods. The reasons are two folds. First, the dataset contains a large number of stopping cars that help self-supervised methods learn depth on vehicles, so our method is hard to further improve the performance when previous methods have obtained good results on dynamic regions. Second, PackNet [14] uses a large network backbone, while other methods, including ours, use ResNet-18, which is much smaller than the former. Overall, we argue that the existing methods have reached a bottleneck in the KITTI dataset, and due to the low impact of dynamic objects on self-supervised learning here, our method is hard to further improve the performance. Moreover, we show qualitative results in Fig. 6(a), which shows that our method generates sharper depth maps than other methods.

**Depth Quality at Object Boundaries:** Table VI shows the detailed analysis of depth results on the IBims-1 dataset, where we compare our method with Monodepth2 [13] and SC-DepthV2 [30]. All models are trained on NYUv2 for a fair comparison. The AbsRel metric shows the overall accuracy of depth estimation results on full images, and other metrics reflect the detailed depth quality at object boundaries and plane regions. The results show that our method significantly outperforms previous methods. We also remove the proposed LSR from our full model for ablation study purposes, and the results in Table VI show that the performance is clearly degraded. This demonstrates the efficacy of our proposed LSR module.
Fig. 5. Qualitative depth estimation results. Existing methods show poor results in dynamic scenes (a–c) because they are hard to handle fast-moving objects during training. Even though they show good accuracy in (d), where models are trained in static scenes, the depth is blurred at object boundaries. By contrast, our method predicts sharp and accurate depth robustly.

qualitative depth estimation results on the IBims-1 dataset are illustrated in Fig. 5(d).

C. Ablation Studies

We have shown results with and without our proposed DRR and LSR in Tables I, II and III. The results demonstrate the efficacy of the proposed modules. In this section, we make a more detailed analysis of the proposed methods, and we also discuss the performance by using different methods to generate pseudo-depth.

Dynamic Region Refinement: The proposed DRR module consists of dynamic-focused sampling and confident depth ranking loss. We make ablation studies by comparing our method with random sampling (RS) and original ranking loss (RL) that are used in [26]. Table VII shows the evaluation results, which show that the performance is significantly degraded when replacing our proposed terms with the existing methods. This demonstrates the efficacy of our proposed methods.

Local Structure Refinement: The proposed LSR module consists of normal matching loss and edge-guided relative normal ranking loss. The ablation study results are summarized in Table VIII. We replace the normal matching loss with edge-aware depth smoothness loss (EDS), and we also add random sampling (RS) to the edge-guided sampling. These variants degenerate the depth accuracy, which demonstrates that our proposed methods are better than existing solutions.
Fig. 6. Qualitative monocular depth estimation results on static datasets. Our method allows for generating sharper depth maps than previous methods—See object boundaries.

| TABLE V |
| --- |
| MONOCULAR DEPTH ESTIMATION RESULTS ON THE NYUv2 [12] DATASET |

| Methods       | Error ↓ | Accuracy ↑ |
|---------------|---------|------------|
|               | AbsRel  | RMS        |
|               | δ1      | δ2         | δ3         |
| Make3D [59]   | 0.349   | 1.214      | 0.447      | 0.745 | 0.897 |
| DepthTransfer [60] | 0.349 | 1.210      | -          | -     | -    |
| Liu et al. [61] | 0.335  | 1.060      | -          | -     | -    |
| Ladicky et al. [62] | -     | -          | 0.542      | 0.829 | 0.941 |
| Li et al. [63] | 0.232   | 0.821      | 0.621      | 0.866 | 0.968 |
| Roy et al. [64] | 0.187   | 0.744      | -          | -     | -    |
| Wang et al. [65] | 0.220  | 0.745      | 0.605      | 0.890 | 0.970 |
| Eigen et al. [66] | 0.158  | 0.641      | 0.760      | 0.950 | 0.988 |
| Chakrabarti et al. [67] | 0.149  | 0.620      | 0.806      | 0.958 | 0.987 |
| Li et al. [68] | 0.143   | 0.635      | 0.788      | 0.958 | 0.991 |
| DORN [69]      | 0.115   | 0.509      | 0.828      | 0.960 | 0.992 |
| VNL [70]       | **0.108** | **0.416**  | **0.875**  | **0.976** | **0.994**  |
| Zhou et al. [71] | 0.208  | 0.712      | 0.674      | 0.900 | 0.968 |
| Zhao et al. [72] | 0.189  | 0.686      | 0.701      | 0.912 | 0.978 |
| Monodepth2 [13] | 0.169  | 0.614      | 0.745      | 0.946 | 0.987 |
| SC-Depth [1]   | 0.159   | 0.608      | 0.772      | 0.939 | 0.982 |
| P2Net [73]     | 0.150   | 0.561      | 0.796      | 0.948 | 0.986 |
| SC-DepthV2 [30] | 0.138  | 0.532      | 0.820      | 0.956 | 0.989 |
| Monolndoor [74] | 0.134  | 0.526      | 0.823      | 0.958 | 0.989 |
| Ours           | **0.123** | **0.486**  | **0.848**  | **0.963** | **0.991** |

Our method outperforms a majority of supervised methods (first row) and all the self-supervised methods (second row).

| TABLE VI |
| --- |
| EVALUATION OF DEPTH BOUNDARIES (DBE) AND PLANES (PE) ON iBims-1 [57] |

| Method                  | iBims-1 |
|-------------------------|---------|
|                         | e\text{acc} | e\text{comp} | e\text{Rec} | e\text{Prec} | AbsRel |
| Monodepth2 [13]         | 4.269    | 89.771       | 10.943      | 29.327      | 0.202 |
| SC-DepthV2 [30]         | 4.206    | 69.846       | 7.049       | 23.190      | 0.172 |
| Ours w/o LSR            | 3.138    | 65.692       | 3.684       | 14.696      | 0.152 |
| Ours                    | **3.001** | **48.047**   | **2.701**   | **13.372**  | **0.146** |

All models are trained on NYUv2.

| TABLE VII |
| --- |
| ABLATION STUDIES OF THE PROPOSED DRR ON DDAD DATASET |

| Methods                  | Full | Dynamic | Static |
|--------------------------|------|---------|--------|
|                         | AbsRel | δ1     | AbsRel | δ2     | AbsRel | δ1     |
| Baseline                 | 0.179  | 0.753  | 0.355  | 0.536  | 0.163  | 0.761  |
| B+DRR (Ours)             | **0.149** | **0.794** | **0.210** | **0.666** | **0.146** | **0.799** |
| DRR w/ RS                | 0.154  | 0.785  | 0.219  | 0.654  | 0.151  | 0.790  |
| DRR w/ RL                | 0.159  | 0.767  | 0.214  | 0.659  | 0.159  | 0.765  |

RS denotes random sampling used in [26], and rl denotes ranking loss used in [26]. The decreased performance demonstrates the effectiveness of our proposed methods.

**Pseudo-Depth:** We use LeReS [25] (ResNet-101) in this paper for generating pseudo-depth, while it is also possible to use other monocular depth estimation networks. Table IX shows the ablation study results on DDAD dataset, where we also include DPT [44] and ResNet-50 version of LeReS. The results show that the pseudo-depths that are generated by all three variants are not accurate in the DDAD dataset. However, when applying our proposed method that uses pseudo-depth for
We use pseudo-depth to boost self-supervised depth estimation results. This demonstrates that our proposed method is not limited to one specific method for generating pseudo-depth. The results also show that our method with LeReS (ResNet-101) outperforms other variants, including DPT. We hypothesize that the reason is that our method incorporates the normal information in training which is shared with LeReS but not with DPT.

**Discussion:** We use pseudo-depth to boost self-supervised monocular depth estimation, which somewhat degrades our claim of self-supervised learning. However, in practice, the monocular depth estimation models such as [25], [44] are only trained once in large-scale datasets and can be used as off-the-shelf tools in new unseen scenes, so our method has almost no extra cost compared to pure self-supervised depth estimation methods [1], [13].

### V. Conclusion

We propose SC-DepthV3 for robust self-supervised learning of monocular depth from challenging dynamic videos. The key to our method is that we use pseudo-depth, which is generated by a pretrained monocular depth estimation network, for addressing the challenges in self-supervised monocular depth estimation framework. More specifically, we address the issues of dynamic objects and blurred object boundaries. As a result, our proposed method can predict sharp and accurate depth maps, even when the model is trained from highly dynamic videos. We comprehensively evaluate our method on six challenging datasets, including both dynamic and static scenes. The results show that our method significantly outperforms previous alternatives, and the ablation study results demonstrate that the proposed modules are effective.

---

### ACKNOWLEDGMENTS

The authors would like to thank anonymous reviewers for their valuable suggestions.

### REFERENCES

[1] J.-W. Bian et al., “Unsupervised scale-consistent depth learning from video,” *Int. J. Comput. Vis.*, vol. 129, pp. 2548–2564, 2021.

[2] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2366–2374.

[3] S.-J. Park, K.-S. Hong, and S. Lee, “RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation,” in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 4980–4989.

[4] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, “RGB-D salient object detection: A survey,” *Comput. Vis. Media*, vol. 7, pp. 37–69, 2021.

[5] R. A. Newcombe et al., “KinectFusion: Real-time dense surface mapping and tracking,” in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, 2011, pp. 127–136.

[6] J.-W. Bian, H. Zhan, and I. Reid, “NVSS: High-quality novel view self-synthesis,” in *Proc. Int. Conf. 3D Vis.*, 2021, pp. 1085–1094.

[7] H. Zhan, C. S. Weerasekera, J. Bian, and I. Reid, “Visual odometry revisited: What should be learnt?”, 2019, arXiv:1909.09803.

[8] F. Liu, C. Shen, G. Lin, and I. Reid, “Learning depth from single monocular images using deep convolutional neural fields,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.

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

[10] 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.*, 2017, pp. 6612–6619.

[11] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets Robotics: The KITTI dataset,” *Int. J. Robot. Res.*, vol. 32, pp. 1231–1237, 2013.

[12] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from RGBD images,” in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.

[13] C. Godard, O. M. Aodha, M. Firman, and G. J. Brostow, “Digging into self-supervised monocular depth prediction,” in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 3828–3838.

[14] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, “3D packing for self-supervised monocular depth estimation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2482–2491.

[15] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, “Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8001–8008.

[16] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova, “Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras,” in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 8976–8985.

[17] V. Guizilini, R. Hou, J. Li, R. Ambrus, and A. Gaidon, “Semantically-guided representation learning for self-supervised monocular depth,” in *Proc. Int. Conf. Learn. Representation*, 2020.

[18] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, “Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance,” in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 582–600.

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

[20] Y. Zou, Z. Luo, and J.-B. Huang, “DF-Net: Unsupervised joint learning of optical flow and camera pose,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 38–53.

[21] A. Ranjan et al., “Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12232–12241.

[22] Y. Chen, C. Schmid, and C. Sminchisescu, “Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera,” in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 7063–7072.

[23] S. Liu, S. Im, S. Lin, and I. S. Kweon, “Learning monocular depth in dynamic scenes via instance-aware projection consistency,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1863–1872.

[24] H. Li, A. Gordon, H. Zhao, V. Casser, and A. Angelova, “Unsupervised monocular depth learning in dynamic scenes,” in *Proc. Conf. Robot Learn.*, 2021, pp. 1908–1917.
[25] W. Yin et al., “Learning to recover 3D scene shape from a single image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 204–213.

[26] W. Chen, Z. Fu, D. Yang, and J. Deng, “Single-image depth perception in the wild,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2016, pp. 730–738.

[27] C. Guo, D. M. Ashtiani, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6602–6611.

[28] H. Zhan, R. Garg, C. S. Weerasekara, K. Li, H. Agarwal, and I. Reid, “Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Web., 2018, pp. 340–349.

[29] R. Mahjourian, M. Wicke, and A. Angelova, “Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 5667–5675.

[30] J.-W. Bian, H. Zhan, N. Wang, T.-J. Chin, C. Shen, and I. Reid, “Auto-regnet network for unsupervised indoor depth estimation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, pp. 9802–9813, Dec. 2022.

[31] R. Wang, Z. Yu, and S. Gao, “PlaneDepth: Self-supervised depth estimation via orthogonal planes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2023, pp. 2145–2154.

[32] H. Si et al., “Fully self-supervised depth estimation from defocus clue,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2023, pp. 9140–9149.

[33] A. Bangunharcana, A. M. Aly, and K. Kim, “DualRefine: Self-supervised depth and pose estimation through iterative epipolar sampling and refinement toward equilibrium,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2023, pp. 726–738.

[34] K. Xian et al., “Monocular relative depth perception with web stereo data supervision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 311–320.

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

[36] Z. Li et al., “Learning the depths of moving people by watching frozen people,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 15–23.

[37] C. Wang, S. Lucey, F. Perazzi, and O. Wang, “Web stereo video supervision,” in Proc. Int. Conf. 3D Vis., 2019, pp. 348–357.

[38] W. Chen, S. Qian, and J. Deng, “Learning single-image depth from videos using quality assessment networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 5604–5613.

[39] W. Yin et al., “Learning to recover 3D scene shape from a single image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 204–213.

[40] 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. 742–756, 2022.

[41] H. Hirschmuller, “Accurate and efficient stereo processing by semi-global matching and mutual information,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2005, pp. 807–814.

[42] 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., 2012, pp. 611–625.

[43] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, “A naturalistic open source movie for optical flow evaluation,” in Proc. Eur. Conf. Comput. Vis., 2012, pp. 611–625.

[44] R. Ranftl, A. Bochkovskiy, and V. Koltun, “Transformer for dense correspondence,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 4524–4530.

[45] X. Luo, J.-B. Huang, R. Szlendak, K. Matzen, and J. Kopf, “Consistent video depth estimation,” ACM Trans. Graph., vol. 39, no. 20, Art. no. 71, 2020.

[46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image Quality Assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[47] K. Xian, J. Zhang, W. Ong, L. Mai, Z. Lin, and Z. Cao, “Structure-guided ranking loss for single image depth prediction,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2020, pp. 608–617.

[48] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[49] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Assist. Interv., Springer, 2015, pp. 234–241.

[50] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (ELUs),” 2015, arXiv:1511.07289.