2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

SuperDepth: Self-Supervised, Super-Resolved Monocular Depth Estimation

Sudeep Pillai, Rareș Ambruș, Adrien Gaidon
Toyota Research Institute (TRI)

***Abstract*— Recent techniques in self-supervised monocular depth estimation are approaching the performance of supervised methods, but operate in low resolution only. We show that high resolution is key towards high-fidelity self-supervised monocular depth prediction. Inspired by recent deep learning methods for Single-Image Super-Resolution, we propose a *sub-pixel convolutional layer extension* for depth super-resolution that accurately synthesizes high-resolution disparities from their corresponding low-resolution convolutional features. In addition, we introduce a *differentiable flip-augmentation layer* that accurately fuses predictions from the image and its horizontally flipped version, reducing the effect of left and right shadow regions generated in the disparity map due to occlusions. Both contributions provide significant performance gains over the state-of-the-art in self-supervised depth and pose estimation on the public KITTI benchmark. A video of our approach can be found at https://youtu.be/jKNgBeBMx0I.**

I. Introduction

Robots need the ability to simultaneously infer the 3D structure of a scene and estimate their ego-motion to enable autonomous operation. Recent advances in Convolutional Neural Networks (CNNs), especially for depth and pose estimation [1], [2], [3], [4] from a monocular camera have dramatically shifted the landscape of single-image 3D reconstruction. These methods cast monocular depth estimation as a supervised or semi-supervised regression problem, and require large volumes of ground truth depth and pose measurements that are sometimes difficult to obtain. On the other hand, self-supervised methods in depth and pose estimation [5], [6], [7] alleviate the need for ground truth labels and provide a mechanism to learn these latent variables by incorporating geometric and temporal constraints to effectively infer the structure of the 3D scene.

Recent works [6], [7], [8] in *self-supervised* depth estimation are limited to training in lower-resolution regimes due to the large memory requirements of the model and their corresponding self-supervised loss objective. High resolution depth prediction is, however, crucial for safe robot navigation, in particular for autonomous driving where high resolution enables robust long-term perception, prediction, and planning, especially at higher speeds. Furthermore, simply operating at higher image resolutions can be shown to improve overall disparity estimation accuracy (Section IV). We utilize this intuition and propose a deep architecture leveraging super-resolution techniques to improve monocular disparity estimation.

Contributions: We propose to use *subpixel-convolutional layers* to effectively and accurately super-resolve disparities

The authors are with the Toyota Research Institute (TRI) 4440 El Camino Real, Los Altos, CA 94022 USA and can be reached via email at firstname.lastname@tri.global

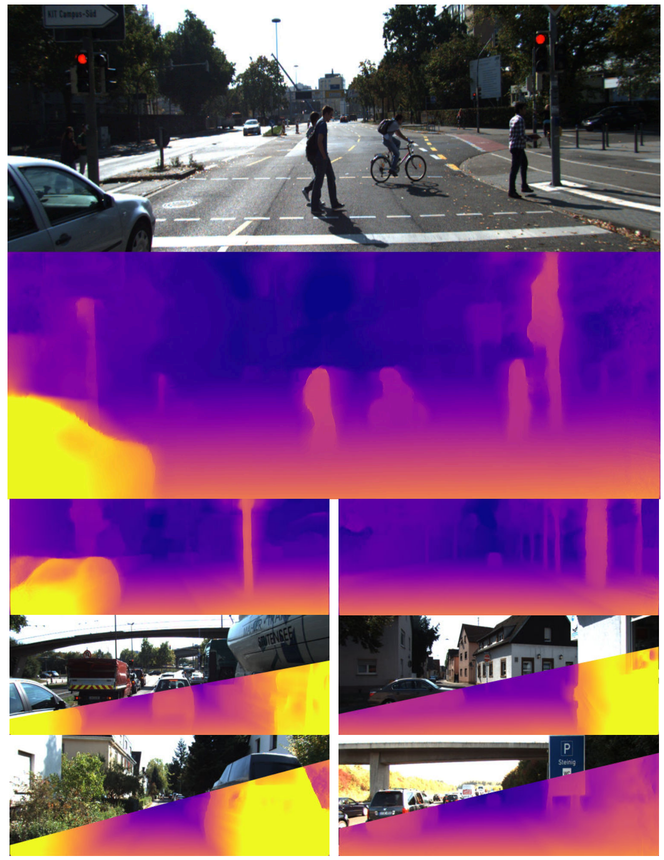

Fig. 1: Illustration of the accurate and crisp disparities produced by our method from a single monocular image. Our approach combines techniques from Single-Image Super-Resolution (SISR) [9] and spatial transformer networks (STN) [10] to estimate high-resolution, and accurate super-resolved disparity maps.

from their lower-resolution outputs, thereby replacing the deconvolution or resize-convolution [11] up-sampling layers typically used in the disparity decoder networks [6], [7]. Second, we introduce a *differentiable flip-augmentation layer* that allows the disparity model to learn an improved prior for disparities at image boundaries in an end-to-end fashion. This results in improved test-time depth predictions with reduced artifacts and occluded regions, effectively removing the need for additional post-processing steps typically used in other methods [6], [12]. We train our monocular disparity estimation network in a *self-supervised* manner using a synchronized stream of stereo imagery, relieving the need for ground truth depth labels. We show that our proposed layers provide significant performance gains to the overall monocular disparity estimation accuracy (Figure 1), especially at higher image resolutions as we detail in our experiments on the public KITTI benchmark.

978-1-5386-6026-3/19/$31.00 ©2019 IEEE

II. Related Work

The problem of depth estimation from a single RGB image is an ill-posed inverse problem. Many 3D scenes can indeed correspond to the same 2D image, for instance because of scale ambiguities. Therefore, solving this problem requires the use of strong priors, in the form of geometric [6], [7], [13], ordinal [14], or temporal constraints [8], [7], [15]. Another effective form of strong prior knowledge is statistical in nature: powerful representations learned by deep neural networks trained on large scale data. CNNs have indeed shown consistent progress towards robust scene depth and 3D reconstruction [13], [16], [17]. State-of-the-art approaches in leveraging both data and structural constraints mostly differ by the type of data and supervision used.

Supervised Depth Estimation Saxena et al. [18] proposed one of the first monocular depth estimation techniques, learning patch-based regression and relative depth interactions using Markov Random Fields trained on ground truth laser scans. Eigen et al. [4] proposed a multiscale CNN architecture trained on ground truth depth maps by minimizing a scale-invariant loss. Fully supervised deep learning-based approaches have since then continuously advanced the state of the art through various architecture and loss improvements [19], [20], [1], [21]. Semi-supervised methods [3], [13] can, in theory, alleviate part of the labeling cost. However, so far they have only been evaluated when using similar amounts of labeled data, reporting significant improvements nonetheless.

Self-supervised Depth Estimation Procuring large amounts of ground truth depth maps is expensive, often requiring significant additional effort. *Self-supervised* learning methods have recently proven to be a promising direction to circumvent this major limitation. Recent advancements, for instance Spatial Transformer Networks [10], have opened the door to a variety of differentiable geometric constraints used as learning objectives capturing key scene properties characterizing optical flow [12], [22], depth [6], [5], [23], [15], and camera pose [7], [23], [24].

Self-supervised approaches thus typically focus on engineering the learning objective, for instance by treating view-synthesis as a proxy task [25], [26], [6], [8], [23], [27], as shown in Fig. 2, and further detailed in Section III-B. Our contributions rely on changing fundamental building blocks of the depth prediction CNN architecture using ideas developed initially for super-resolution [9], or transforming post-processing heuristics into trainable parts of our model.

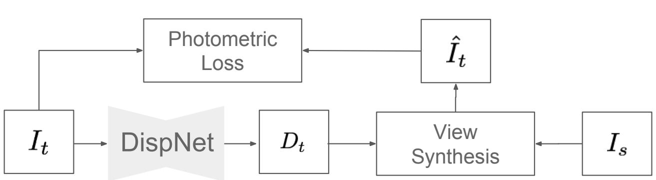

Fig. 2: Self-Supervised Disparity Estimation: The disparity network is self-supervised via a photometric loss defined between the target image I_t and the synthesized target image $\hat{I}_t$. This synthesized target image is generated via Spatial Transformer Networks [10] in the view-synthesis module, given the predicted target disparity D_t and source image I_s.

III. Self-supervised, Super-Resolved Monocular Depth Estimation

The task of monocular depth estimation entails the recovery of a function $f_z : I \to D_z$, that predicts the depth $\hat{z} = f_z(I(p))$ for every pixel p in the given input image I. In this work, we learn to recover instead the disparity estimation function $f_d : I \to D$, in a *self-supervised* manner from a synchronized stereo camera (Section III-A). Given f_d, the metric depth $\hat{z}$ can be trivially estimated for every pixel p in the image I via $\hat{z} = \frac{FB}{f_d(I(p))}$, where F is the camera focal length and B is the stereo baseline between the left and right camera pairs.

A. Monocular Depth Network Architecture

Our baseline disparity estimation model builds upon the popular DispNet [19] architecture. Following [6], we make similar modifications to the encoder-decoder network with skip connections [28] between the encoder's activation blocks. However, unlike the left-right (LR) disparity architecture proposed in [6], the model outputs a single disparity channel. In this work, we propose two key components to this base architecture that we detail in the following sections.

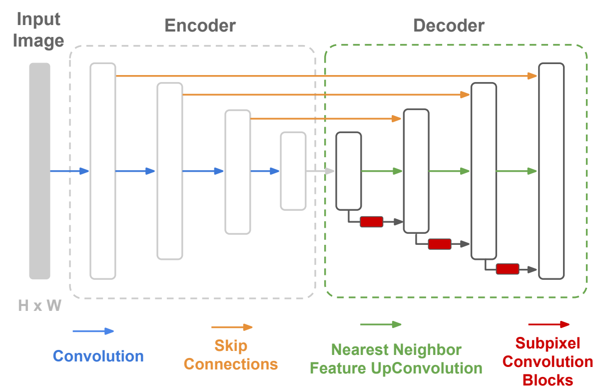

Fig. 3: Illustrated above is the proposed sub-pixel convolutional extension to the DispNet architecture [19]. The sub-pixel convolutional block architecture is further described in Table I.

1) Sub-pixel Convolution for Depth Super-Resolution: Most recent methods in multi-scale disparity estimation [6], [29] utilize de-convolutions [30], resize-convolutions [11] or naive interpolation operators (for e.g. bilinear, nearest-neighbor) to up-sample the lower-resolution disparities to their target image resolution. However, these methods fail to capture the fine details in images, and are limited in their representational capacity for high-quality disparity estimation. Inspired by recent Single-Image-Super-Resolution (SISR) [9] methods, we introduce a sub-pixel convolutional layer based on ESPCN [9] for depth super-resolution that accurately and efficiently synthesizes the high-resolution disparities from their corresponding low-resolution intermediate disparity decoder outputs (see Fig. 3). This replaces the disparity upsampling layers with relevant convolutional operators that can perform high-quality disparity synthesis from the low-resolution features. We swap the resize-convolution branches from each of the 4 pyramid scales in the disparity network

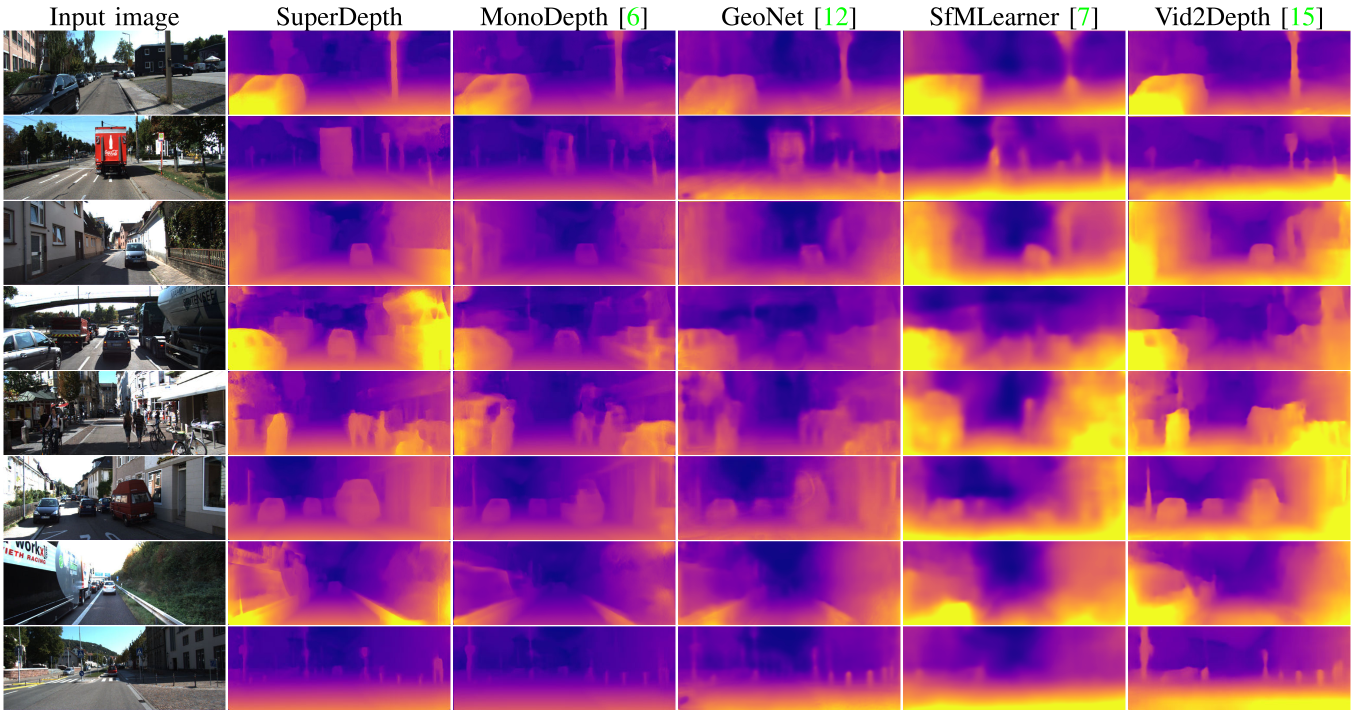

Fig. 4: Illustrated above are qualitative comparisons of our proposed self-supervised, super-resolved monocular depth estimation method with previous state-of-the-art methods. We show that our approach produces qualitatively better depth estimates with crisp boundaries. Our method also correctly reconstructs thin and far-off objects reliably compared to previous methods that tend to only estimate shadow artifacts in such regions.

with the sub-pixel convolutional branch consisting of a sequential block of 4 2D convolutional layers described in Table I. The final convolutional output is then re-mapped to the target depth resolution via a pixel re-arrange operation, resulting in an effective super-resolution operation as described in [9].

	Layer Description	Output Tensor Dim.
1	5x5 conv, 32 planes, stride 2, padding 1, ReLu act.	(B, 32, H, W)
2	3x3 conv, 32 planes, stride 1, padding 1, ReLu act.	(B, 32, H, W)
3	3x3 conv, 16 planes, stride 1, padding 1, ReLu act.	(B, 16, H, W)
4	3x3 conv, 4 planes, stride 1, padding 1	(B, 4, H, W)
5	PixelShuffle, 1 plane	(B, 1, 2H, 2W)

TABLE I: Description of the sub-pixel convolutional block used to upsample the intermediate disparity outputs in DispNet to a higher image resolution.

2) Differentiable Flip Augmentation: In stereopsis, due to the occluding scene points in the left boundary of the right image, the disparity model will learn a poor prior on boundary-pixels. To circumvent this behavior, previous methods [6], [8] incorporate a post-processing step that alpha-blends the disparity images from the image and its horizontally flipped version. While this significantly reduces visual artifacts around the image boundary and improves overall accuracy, it decouples the final disparity estimation from the training. To this end, we replace this step with a *differentiable flip-augmentation layer* within the disparity estimation model itself, allowing us to fine-tune disparities in an end-to-end fashion. By leveraging the differentiable image-rendering in [10] to revert the flipped disparity, the model performs the forward pass with the identical model on both the original and horizontally flipped images. The outputs are fused in a differentiable manner with a pixel-wise mean operation while handling the borders similar to [6].

B. Self-supervising Depth with Stereopsis

We formulate the disparity estimation as a photometric error minimization problem across multiple camera views, following [6], [5]. We define D_t as the disparity image for the corresponding target image I_t, and cast the disparity estimation as an image synthesis task of a new source image I_s, as shown in Fig. 2. The photometric error is then re-written as the minimization of pixel-intensity difference between the target image I_t, and the synthesized target image re-projected from the source image's view $\hat{I}_t = I_s(p_s)$ [10]. Here, $p_s \sim K\mathbf{x}_{t\to s}\hat{D}_t(p_t)K^{-1}p_t$ is the source pixel derived from re-projecting the target pixel p_t in the source image's view $\mathbf{x}_s$, with $\mathbf{x}_{t\to s}$ describing the relative transformation between the target image view pose $\mathbf{x}_t$ and source image view pose $\mathbf{x}_s$. The disparity estimation model f_d parametrized by θ_d is defined as:

$$\hat{\theta_D} = \arg\min_{\theta_D} \sum_{s\in S} \mathcal{L}_D(I_t, \hat{I}_t; \theta_D) \tag{1}$$

where $s \in S$ are all the disparate views available for synthesizing the target image I^t. In the case of stereo cameras with a fixed baseline B, $\mathbf{x}_{s\to t}$ in Equation 1 is known a-priori, and directly incorporated as a constant within the overall minimization objective.

The overall loss $\mathcal{L}_d$ comprises of 3 terms (Equation 2), elaborated below:

$$\mathcal{L}_D(I_t, \hat{I}_t; \alpha_1) = \mathcal{L}_p(I_t, \hat{I}_t; \alpha_1) + \lambda_1\ \mathcal{L}_s(I_t) + \lambda_2\ \mathcal{L}_o(I_t) \tag{2}$$

Appearance Matching Loss We define an appearance matching loss term that captures the pixel-level similarity between the target image I_t and the view-synthesized target image $\hat{I}_t$. Following [6], [5], we use the Structural Similarity (SSIM) [31] term combined with an L1 pixel-wise reconstruction loss, inducing an overall appearance loss given by Equation 3 below.

$$\mathcal{L}_p(I_t, \hat{I}_t; \alpha_1) = \alpha_1 \frac{1 - \text{SSIM}(I_t, \hat{I}_t)}{2} + (1 - \alpha_1) \ |I_t - \hat{I}_t| \tag{3}$$

The appearance loss term is averaged per-pixel, pyramid-scale and image batch during training.

Disparity Smoothness Loss In order to regularize the disparities in textureless low-image gradient regions, we incorporate an edge-aware term (Equation 4), similar to [6], [13], [23]. The effect of each of the pyramid-levels is decayed by a factor of 2 on downsampling, starting with a weight of 1 for the 0^{th} pyramid level.

$$\mathcal{L}_s(I_t) = |\delta_x d_t| e^{-|\delta_x I_t|} + |\delta_y d_t| e^{-|\delta_y I_t|} \tag{4}$$

Occlusion Regularization Loss Additionally, we incorporate an occlusion regularization term [13] to minimize the shadow areas generated in the disparity map, especially across high gradient disparity regions.

$$\mathcal{L}_o(I_t) = |d_t| \tag{5}$$

In this term (Eq. 5), we induce a loss over the total sum of the absolute disparity estimate, encouraging lower disparity values (occluded pixel depths) over larger disparity values (occluding pixel depths).

IV. Experiments

A. Dataset

We use the KITTI [32] dataset for all our experiments. We compare with previous methods on the standard KITTI Disparity Estimation benchmark, and adopt the training protocols used in Eigen et al. [4]. More specifically, we used the KITTI *Eigen* splits described in [4] that contain 22600 training, 888 validation, and 697 test stereo image pairs. We evaluate the disparities estimated using the common metrics adopted by related methods (more details in [4]), and summarize our results in Table II.

B. Effect of High-Resolution in Disparity Estimation

As previously mentioned, the *self-supervised* photometric loss is limited by the image resolution and the corresponding disparities at which they operate. In their recent work [8], the authors discuss this limitation and up-sample the multi-scaled disparity outputs to their original input image resolution before computing the relevant photometric losses. In Figure 5, we show that with increasing input image resolutions of 1024 x 384, 1536 x 576, and 2048 x 768, the disparity estimation performance continues to improve for most metrics including Abs. Rel, Sq Rel. RMSE, and RMSE log. The performance of the baseline approach however saturates at the 1536 x 576 resolution since the original KITTI stereo images are captured at 1392 x 512 pixel resolution. It is however noteworthy that the fraction of the disparities within $\delta < z$ pixels show improvements with even higher input image resolutions indicating that the photometric losses are indeed limited by the disparity resolution.

Resolution	Abs Rel	Sq Rel	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
512 x 192	0.133	1.079	0.247	0.816	0.927	0.964
1024 x 384	0.116	0.935	0.210	0.842	0.945	**0.977**
1536 x 576	**0.114**	**0.869**	**0.209**	0.849	0.945	0.976
2048 x 768	0.116	1.055	**0.209**	**0.853**	**0.948**	**0.977**

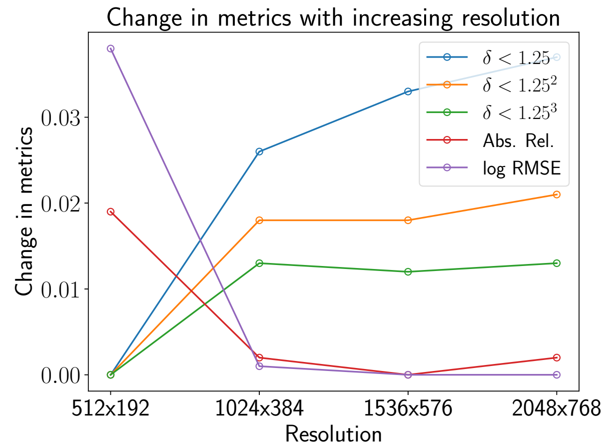

Fig. 5: **Effect of high-resolution**: The relative change in the disparity estimation metrics are plotted with increasing input image resolution. We show that by naively increasing the input image resolution, we are able to show considerable improvement (increase in $\delta < z$, and decrease in Abs Rel, Sq Rel, RMSE log metrics) without any changes to the underlying loss function. This motivates us to consider efficient and accurate methods in performing disparity estimation at much higher input image resolutions via sub-pixel convolutions.

C. Depth Estimation Performance

We re-implemented the modified DispNet with skip connections as described in Godard et al. [6] as *our* baseline (**Baseline**), and evaluate it with the proposed sub-pixel convolutional extension (**SuperDepth-SP**) and the differentiable flip-augmentation (**SuperDepth-FA**).

Improving Disparity Estimation with Sub-pixel Convolutions Using the insight of operating at high-resolution disparity regimes, we discuss the importance of super-resolving low-resolution disparities estimated within Encoder-Decoder-based disparity networks [19], [13], [1]. With **SuperDepth-SP**, we are able to achieve a considerable improvement in performance (0.112 abs. rel.) for the same input image resolution over our established baseline (0.116 abs. rel.). Furthermore, we notice that the Sq. Rel., RMSE, $\delta < z$ columns show equally consistent and improved performance over the baseline that utilizes resize-convolutions [11] instead for disparity up-sampling.

Improving Disparity Estimation with Differentiable Flip-Augmentation Fine-Tuning In their previous works Godard et. al [6] use a hand-engineered post-processing step to fuse the disparity estimates of the left image and the horizontally flipped image. While this reduces the artifacts at the borders of the image, we show that this technique can be used in a differentiable manner to allow further fine-tuning

Method	Resolution	Dataset	Train	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Garg et al.[5] cap 50m	620 x 188	K	M	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Godard et al. [8]	640 x 192	K	M	0.129	1.112	5.180	0.205	0.851	0.952	0.978
SfMLearner [7] (w/o explainability)	416 x 128	K	M	0.221	2.226	7.527	0.294	0.676	0.885	0.954
SfMLearner [7]	416 x 128	K	M	0.208	1.768	6.856	0.283	0.678	0.885	0.957
SfMLearner [7]	416 x 128	CS+K	M	0.198	1.836	6.565	0.275	0.718	0.901	0.960
GeoNet [12]	416 x 128	K	M	0.155	1.296	5.857	0.233	0.793	0.931	0.973
GeoNet [12]	416 x 128	CS+K	M	0.153	1.328	5.737	0.232	0.802	0.934	0.972
Vid2Depth [15]	416 x 128	K	M	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Vid2Depth [15]	416 x 128	CS+K	M	0.159	1.231	5.912	0.243	0.784	0.923	0.970
UnDeepVO [23]	416 x 128	K	S	0.183	1.73	6.57	0.268	-	-	-
Godard et al. [6]	640 x 192	K	S	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Godard et al. [6]	640 x 192	CS+K	S	0.124	1.076	5.311	0.219	0.847	0.942	0.973
Godard et al. [8]	640 x 192	K	S	0.115	1.010	5.164	0.212	**0.858**	0.946	0.974
Baseline	1024 x 384	K	S	0.116	0.935	5.158	0.210	0.842	0.945	0.977
SuperDepth-SP	1024 x 384	K	S	**0.112**	0.880	4.959	**0.207**	0.850	0.947	0.977
SuperDepth-FA	1024 x 384	K	S	0.115	0.922	5.031	0.206	0.850	**0.948**	**0.978**
SuperDepth-SP+FA	1024 x 384	K	S	**0.112**	**0.875**	**4.958**	**0.207**	0.852	0.947	0.977

TABLE II: Single-view depth estimation results on the KITTI dataset [32] using the Eigen Split [4] for depths reported less than 80m, as indicated in [4]. The mode of self-supervision employed during training is reported under the **Train** column - Stereo (S), Mono (M). Above, we compare our baseline approach along with the proposed sub-pixel convolutions variant (**SuperDepth-SP**), the differentiable flip-augmentation (**SuperDepth-FA**), and the combined variants (**SuperDepth-SP+FA**). Training datasets used by previous methods are listed as either CS=Cityscapes [33], K=KITTI[32]. For Abs Rel, Sq Rel, RMSE, and RMSE log, lower is better. For $\delta < 1.25$, $\delta < 1.25^2$ and $\delta < 1.25^3$, higher is better.

of the disparity network in an end-to-end manner. With the differentiable flip-augmentation training, we improve the baseline and the sub-pixel variant (**SuperDepth-SP**) on all metrics except the Abs. Rel which remains unchanged. Finally, by training with the subpixel-variant (**SuperDepth-SP**) and fine-tuning with the flip-augmentation (**SuperDepth-FA**) we are able to achieve state-of-the-art performance on the KITTI Eigen split benchmark as listed in Table II.

Effects of fine-tuning and pre-training Many recent state-of-the-art results [8], [29] provide strong performance by either using pre-trained ImageNet weights [34] and fine-tuning or adapting the task domain from a model trained on an alternate dataset training. While we realize the implications of transferring well-conditioned model weights for warming up training, in this work we only consider the case of *self-supervised training from scratch*. Despite training *from scratch*, we show in Table II that the performance of our models (**Baseline**, **SuperDepth-SP**, **SuperDepth-SP+FA**) are competitive with those of recent state-of-the-art self-supervised disparity estimation methods [8], [29] that utilize ImageNet pre-trained weights.

Qualitative results We contrast the results of our method alongside related methods in Figure 4. We note that our method is able to capture finer details of objects as compared to previously established state-of-the-art. The effect of our *sub-pixel convolutions* is particularly noticeable around smaller objects (e.g. poles, traffic signs), where the super-resolved depths successfully recover the underlying geometry.

D. Self-Supervising Pose Estimation with SuperDepth

To further validate our contributions, we perform a second set of experiments where we use our disparity network trained on stereo data to train a pose network which estimates the 6 DoF pose between subsequent monocular frames. We bootstrap the self-supervised SuperDepth training process with a similar self-supervised pose estimation network that is able to recover scale-aware, long-term camera trajectories. For the pose network, we use the architecture described in [7] *without* the explainability mask layer. Formally, the pose network recovers a function $f_{\mathbf{x}} : (I_t, I_S) \rightarrow \mathbf{x}_{t\rightarrow s} = \left(\begin{smallmatrix} R & t \\ 0 & 1 \end{smallmatrix}\right) \in SE(3)$, for all $s \in S$, where $\mathbf{x}_{t\rightarrow s}$ is the 6 DoF transformation between image I_t and I_s. The pose network is fed with the concatenation of the target image I_t and a set of context images I_S, which are temporally adjacent to the target image. Following [35], [36], [37], we use the logarithm of a unit quaternion to parameterize the rotation in $\mathbb{R}^3$ and do not require an added normalization constraint unlike previous works [38].

We train the pose network through an additional *self-supervised* photometric loss, $\mathcal{L}_X$ between the target image I_t and image $\hat{I}_t$ inferred via the mapping $\mathbf{x}_{t\rightarrow s}$ from the context image I_s. We define $\mathcal{L}_X = \mathcal{L}_p(I_t, \hat{I}_t; \alpha_2)$ and we note here that, although similar to the $\mathcal{L}_p$ loss defined in Eq. 3, the multi-view photometric loss, $\mathcal{L}_X$, uses a different weight, α_2, to trade-off between the $L1$ and the $SSIM$ components. In all our pose estimation experiments, $\alpha_2 = 0.05$, so that the optimization favors the $L1$ component of the loss while training the pose network. This is important, as the $SSIM$ loss is better suited for images that are fronto-parallel (e.g. stereo camera images), an assumption which is often invalidated in images which are acquired sequentially as the camera is undergoing ego-motion. The total loss becomes $\mathcal{L} = \mathcal{L}_D + \mathcal{L}_X$, through which we train the pose estimating network f_x while ensuring that the network which estimates disparity, f_d, does not diverge during this optimization step; this is important for recovering trajectories that are metrically accurate.

For long term trajectory estimation we report Average Translational (t_{rel}) and Rotational (r_{rel}) RMS drift over trajectories of 100-800 meters. We use the KITTI odometry benchmark for evaluation, and specifically sequences 00 -

	SfMLearner [7]‡		UnDeepVO [23]		Ours	
Seq	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
00†	66.35	6.13	4.41	1.92	6.12	2.72
03†	10.78	3.92	5.00	6.17	7.90	4.30
04†	4.49	5.24	4.49	2.13	11.80	1.90
05†	18.67	4.10	3.40	1.50	4.58	1.67
07†	21.33	6.65	3.15	2.48	7.60	5.17
01*	35.17	2.74	69.07	1.60	13.48	1.97
02*	58.75	3.58	5.58	2.44	3.48	1.10
06*	25.88	4.80	6.20	1.98	1.81	0.78
08*	21.90	2.91	4.08	1.79	2.25	0.84
09*	18.77	3.21	7.01	3.61	3.74	1.19
10*	14.33	3.30	10.63	4.65	2.26	1.03
Train	29.26	4.45	11.70	2.75	**4.50**	**1.15**
Test	16.56	3.26	8.82	4.13	**7.60**	**3.15**
Avg	29.95	4.23	11.18	2.55	**5.91**	**2.06**

TABLE III: Long term trajectory results on the KITTI odometry benchmark. We report the following metrics: t_{rel} - average translational RMSE drift (%) on trajectories of length 100-800m, and r_{rel} - average rotational RMSE drift ($°/100m$) on trajectories of length 100-800m. $*$ and † represent train and respectively test seq. for our method. The methods of [7] and [23] are trained on seq. 00-08. ‡ The results of [7] were scaled using the scale from the ground truth depth. The last 3 rows show the Train and Test averages computed on the appropriate sequences for each method as well as the overall average computed across all the sequences.

10, for which ground truth is available. We note that in this case we still train our disparity and pose networks on the KITTI *Eigen* train split, with the mention that this data split includes all the images from sequences 01, 02, 06, 08, 09 and 10. We report our results on all sequences 00 - 10 in III, where we clearly mark the sequences that are used for training and testing, both for our method and the related work. We leave out model based methods (e.g. [39], [13]) and limit our quantitative comparison to similar self-supervised learning based methods. In all our experiments we use a context of size 3 (i.e. target frame plus 2 additional frames).

We compare against: (a) SfMLearner [7] which is trained using monocular video and thus we scale their depth predictions using the scale from the ground truth; and (b) UnDeepVO [23] which, like us, is trained on a combination of monocular and stereo imagery and returns metrically accurate depths and trajectories. We note that our quantitative results are superior to those of [23], which we attribute to the fact that our pose network is bootstrapped with much more accurate depth estimates. We further note that through the proposed combination of monocular and stereo losses our approach is able to overcome the scale ambiguity and recover metrically accurate trajectories which exhibit little drift over extended periods of time (see Table. III and Fig. 6).

E. Implementation

We follow the implementation of [6] closely, and implement our depth estimation network in PyTorch. The sub-pixel convolution and differentiable flip-augmentation take advantage of the native `PixelShuffle` and `index_select` operations in PyTorch, with the model and losses parallelized across 8 Titan V100s during training. We train the disparity network for 200 epochs using the Adam optimizer [40]. The learning rate and batch size are estimated via hyper-parameter search. In most cases, we use a batch size of 4 or 8, with an initial learning rate of 5e-4. As training proceeds, the learning rate is decayed every 40 epochs by a factor of 2. We set the following parameter values for all training runs: $\lambda_1 = 0.1$, $\lambda_2 = 0.01$, $\alpha_1 = 0.85$, $\alpha_2 = 0.05$. For fine-tuning with the differentiable flip-augmentation layer, we use a learning rate of 5e-5, batch size of 2, and only consider the first 2 pyramid scales for computing the loss as the lower-resolution pyramid scales tend to over-regularize the depth maps.

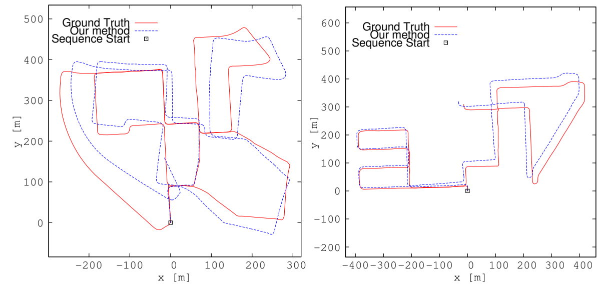

Fig. 6: Illustrations of our pose estimation results on the KITTI odometry benchmark, sequences 00 (left) and 08 (right). Our results are rendered in blue while the ground truth is rendered in red.

V. Conclusion

In this work, we propose two key extensions to self-supervised monocular disparity estimation that enables state-of-the-art performance on the KITTI disparity estimation benchmark. Inspired by the strong performance in monocular disparity estimation in high-resolution regimes, we incorporate the concept of sub-pixel convolutions within a disparity estimation network to enable super-resolved depths. The super-resolved depths operating at higher-resolutions tend to reduce ambiguities in the self-supervised photometric loss estimation (unlike their lower-resolution counterparts), thereby resulting in improved depth estimation. In addition to super-resolution, we introduce a differentiable flip-augmentation layer that further reduces artifacts and ambiguities while training the monodepth model. Through experiments, we show that both contributions provide significant performance gains to the proposed self-supervised technique, resulting in state-of-the-art performance in depth estimation on the public KITTI benchmark. As a consequence of improved disparity estimation, we study its relation to the strongly correlated problem of pose estimation and show strong quantitative and qualitative performance compared to previous self-supervised pose estimation methods.

In the accompanying video we show qualitative results and comparisons with other methods. To further illustrate the usefulness of our method, we show 3D reconstructions generated from RGB images as a vehicle navigates through an environment. These illustrations highlight one potential avenue for future work: using the estimated depth for tasks such as navigation or mapping using only RGB information.

Acknowledgments

We would like to thank John Leonard, Mike Garrison, and the whole TRI-ML team for their support. Special thanks to Vitor Guizilini for his help.

References

[1] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5038–5047.

[2] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, "Learning depth from monocular videos using direct methods," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2022–2030.

[3] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6647–6655.

[4] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374.

[5] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*. Springer, 2016, pp. 740–756.

[6] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[7] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1851–1858.

[8] C. Godard, O. Mac Aodha, and G. Brostow, "Digging into self-supervised monocular depth estimation," *arXiv preprint arXiv:1806.01260*, 2018.

[9] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

[10] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[11] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016.

[12] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1983–1992.

[13] N. Yang, R. Wang, J. Stuckler, and D. Cremers, "Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 817–833.

[14] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.

[15] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5667–5675.

[16] M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. J. Davison, "Codeslamlearning a compact, optimisable representation for dense visual slam," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2560–2568.

[17] H. Zhou, B. Ummenhofer, and T. Brox, "Deeptam: Deep tracking and mapping," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 822–838.

[18] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 824–840, 2009.

[19] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4040–4048.

[20] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1592–1599.

[21] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 66–75.

[22] S. Meister, J. Hur, and S. Roth, "Unflow: Unsupervised learning of optical flow with a bidirectional census loss," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[23] R. Li, S. Wang, Z. Long, and D. Gu, "Undeepvo: Monocular visual odometry through unsupervised deep learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7286–7291.

[24] S. Pillai and J. J. Leonard, "Towards visual ego-motion learning in robots," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5533–5540.

[25] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, "View synthesis by appearance flow," in *European conference on computer vision*. Springer, 2016, pp. 286–301.

[26] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "Deepstereo: Learning to predict new views from the world's imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5515–5524.

[27] X. Fei, S. Soatto, and A. Wong, "Geo-supervised visual depth prediction," *IEEE Robotics and Automation Letters*, 2019.

[28] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[29] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang, "Learning monocular depth by distilling cross-domain stereo networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 484–500.

[30] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[32] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[33] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz, "Geometry-aware learning of maps for camera localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2616–2625.

[36] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni, "Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[37] F. S. Grassia, "Practical parameterization of rotations using the exponential map," *Journal of graphics tools*, vol. 3, no. 3, pp. 29–48, 1998.

[38] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

[39] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.