On-Site Adaptation for Monocular Depth Estimation with a Static Camera

Huan Li
huan.li3@unibo.it
Matteo Poggi
m.poggi@unibo.it
Fabio Tosi
fabio.tosi5@unibo.it
Stefano Mattoccia
stefano.mattoccia@unibo.it

Department of Computer Science and Engineering (DISI),
University of Bologna,
Italy

Abstract

We introduce a novel technique for easing the deployment of an off-the-shelf monocular depth estimation network in unseen environments. Specifically, we target a very diffused setting with a fixed camera mounted higher over the ground to monitor an environment and highlight the limitations of state-of-the-art monocular networks deployed in such a setup. Purposely, we develop an on-site adaptation technique capable of 1) improving the accuracy of estimated depth maps in the presence of moving subjects, such as pedestrians, cars, and others; 2) refining the overall structure of the predicted depth map, to make it more consistent with the real 3D structure of the scene; 3) recovering absolute metric depth, usually lost by state-of-the-art solutions. Experiments on synthetic and real datasets confirm the effectiveness of our proposal.

1 Introduction

Estimating the depth of a single image [38] represents a fascinating challenge in computer vision. In addition to the scientific charm, such an approach is desirable from a practical point of view, allowing for unconstrained depth sensing in almost any scenario without requiring cumbersome and expensive active sensors such as LiDARs [14] nor multiple, synchronized [34] cameras / a single, moving one [39]. Indeed, a single color camera represents the most common setup for several practical applications such as video surveillance [35], road traffic monitoring [33], or, more recently, social distancing [2, 31]. More importantly, these applications raise some crucial privacy concerns, and the possibility of estimating (accurate) depth maps in a surveillance setting also has the potential to improve this aspect. For instance, depth maps could be computed on edge and sent to the cloud to be processed by higher-level applications; under this assumption, the edge node would avoid sharing the color images containing more sensitive information. However, this comes at the cost of dealing with a highly ill-posed problem since the absence of triangulation cues from multiple views prevents the computation of a unique solution explaining the 3D geometry out of a single image.

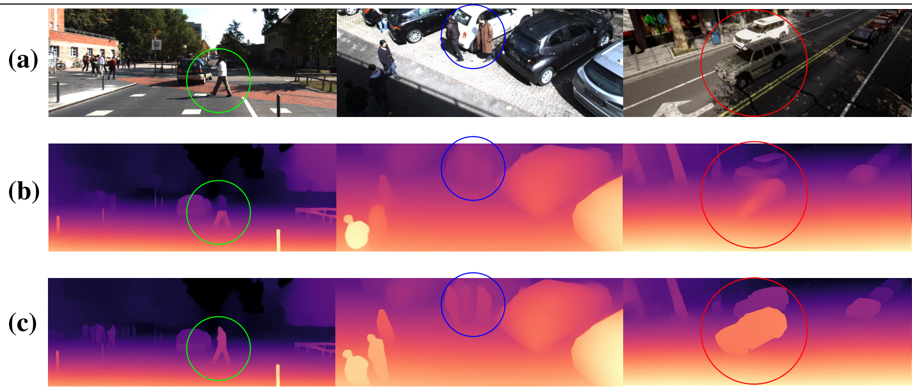


Figure 1: **Monocular depth estimation – before and after adaptation.** On fixed-camera settings (a), state-of-the-art depth estimation models [36] might fail in unseen environments or camera settings (b). Our adaptation scheme allows us to improve their reliability (c).

The advent of deep learning enabled the development of the first-ever solutions [10, 21, 29] making it possible to face such a problem, thanks to the increasing availability of images annotated with depth labels [14, 32] to be used for training, or to the introduction of alternative, self-supervised regimes replacing such annotations with synchronized stereo pairs [13, 15] or monocular video sequences [51]. These approaches, deploying Convolutional Neural Networks (CNNs) or, more recently, Transformers, learn to infer the distance from the camera for objects in the scene based on visual cues [7] such as shadows, perspective, vanishing lines, and more. As proof of this, by acting on some of these cues – e.g., manually shifting the height of the horizon in the image or simulating camera tilting with respect to the ground plane [7] – estimated depth for the very same scene might be sensibly altered.

Indeed, obtaining a network capable of predicting accurate depth maps in any environment remains challenging, even in the availability of a vast amount of annotated data – e.g., *millions* of images [36, 37] collected from very different datasets. Moreover, the *scale ambiguity* intrinsic in single images also plays a role, making even state-of-the-art monocular depth estimation networks capable of predicting accurate relative depth, yet up to an unknown scale factor. As a consequence, despite the recent progress in cross-dataset generalization [36, 37], these models are still subject to failures in specific settings that are under-represented in the training data, e.g., on ambiguous objects such as mirrors [49] or, more commonly, when dealing with images taken from a perspective rarely – or *never* – observed during the training process [7]. Among them, we report an example in Fig. 1, showing a widespread surveillance setting (a), with the camera positioned high over the ground and slanted with respect to it. Although this configuration represents the perfect ground for deploying single image depth estimation – i.e., because of the lack of camera motion or multiple synchronized devices – existing approaches are not ready for unconstrained use there (b), often failing at properly estimating depth for common agents such as pedestrians and cars. We argue that *adaptation* techniques [4, 18] could attenuate this problem. In particular, *online* techniques [41] allow the monocular network to improve its accuracy in a new environment right at deployment time, without any prior assumption on it or requiring any sample beforehand. However, existing online strategies suited for monocular networks rely on the assumption that the camera is moving in the scene [17, 19, 43, 50] to exploit the same principles over which self-supervised approaches build upon [51], and thus cannot be exploited in the static-camera setting mentioned above. Moreover, using stereo images [17] to adapt a monocular network would have little practical sense – i.e., stereo networks [34] would be used instead.

To cope with these limitations, we propose a novel technique for the on-site adaption of an off-the-shelf, monocular depth estimation network when deployed in unseen environments. In contrast to the approaches mentioned above, requiring the camera to move and any other objects in the scene to be static, we exploit the opposite behavior. As we aim at running adaptation on fixed-camera installations, we identify any agents in the scene and use their motion to detect the ground plane over which they move. From these basic cues, we can extract pseudo depth labels for the agents themselves that can be used for a lightweight fine-tuning of the original depth network and exploit the detected ground plane for a test-time refinement step to better align the predicted depth with the actual 3D structure of the environment. Furthermore, by having access to simple priors about the specific camera installation – i.e., the camera height over the ground – we can recover the metric scale for depth maps predicted by the monocular network. To validate our proposal, we run experiments on a subset of the KITTI dataset featuring static camera sequences and two novel datasets composed of synthetic frames rendered through CARLA [8] and real images. In the latter case, the dataset frames indoor and outdoor scenes.

The main contributions of this work are:

- A novel, on-site adaptation scheme for monocular depth estimation networks working in fixed-camera setups, consisting of 1) a lightweight fine-tuning procedure aimed at correcting gross errors on moving agents, 2) a scene alignment step enabled by the moving agents in the scene identifying the ground plane, and 3) metric scale recovery by simply knowing the camera height over the ground.
- Two novel datasets with dense, ground-truth depth labels used to validate the effectiveness of our proposal, available at `https://sites.google.com/view/staticdepth-dataset`.

2 Related Work

Monocular Depth Estimation. After early attempts at learning for monocular depth estimation with classical machine learning [20, 38], this task attracted increasing interest with the rise of deep learning. Eigen *et al.* [9, 11] proposed a pivotal multi-stage, coarse-to-fine network for single image depth prediction, Laina et al. [21] developed a fully convolutional architecture with skip connections. DORN [12] deploys a densely connected backbone and casts depth prediction through ordinal regression, while BTS [22] uses local planar guidance to improve accuracy. More recent approaches estimate depth by predicting probability distributions and discrete bins [5, 27, 40], or by introducing self-attention mechanisms [1, 3, 23, 26, 36]. Among the latest works, NeWCRFs [48] and VA-DepthNet [28], respectively, resumed CRFs within the multi-head mechanism of transformers and first-order variational constraints to set the current state-of-the-art. However, any of the previous frameworks always focus on single domains – i.e., training and evaluating over indoor (NYU v2 [32]) and outdoor (KITTI [42]) data separately. A parallel research trend consists of training a single model for generalizing across different domains. MegaDepth [25] represents the first attempt in this direction, followed by MiDaS [37] and DPT [36]. Nonetheless, these models

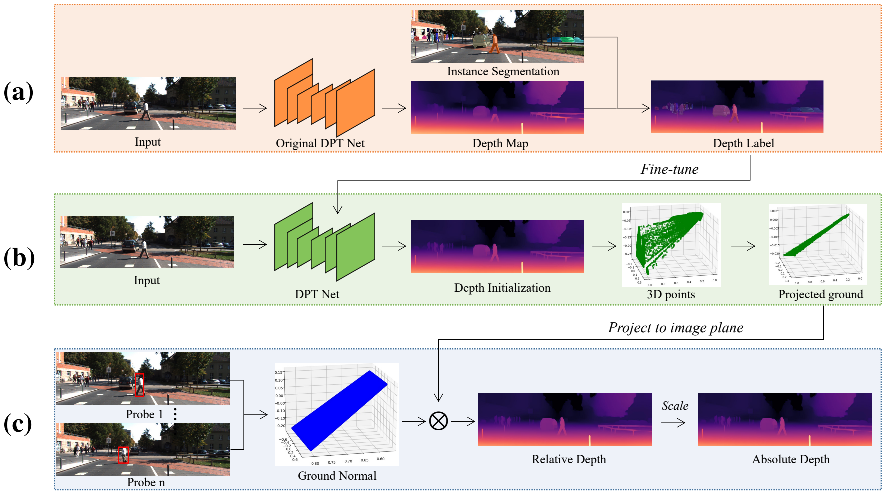


Figure 2: **Overview of our adaptation scheme.** First, (a) we rectify depth for agents in the scene by producing pseudo labels and running a lightweight fine-tuning of the original depth model; then, (b) the ground plane is extracted according to agents' motion, and used to align the overall structure of the depth maps predicted by the model. Eventually, (c) metric scale can be recovered from ground normal vectors by knowing camera height.

still fail when processing images from uncommon viewpoints or yield blurred predictions missing some of the agents in the scene, as already highlighted in Fig. 1.

Domain Adaptation for Monocular Depth Estimation. This research topic arose to deal with the inherent difficulties of obtaining a monocular depth estimation network capable of generalizing. At first, the focus has been on synthetic to real adaptation, exploiting image style transfer [4] or GANs [18] for the purpose, yet needing some samples from the target domain to be available beforehand, and focusing exclusively on outdoor [4] or indoor [18] environments – whereas state-of-the-art solutions [36] are nowadays capable of good generalization across the two. A more practical solution consists of directly adapting the model online during deployment [17, 19, 43, 50]. They build on the image reprojection principle at the core of self-supervised monocular depth estimation approaches [16] by exploiting consecutive frames acquired over time [51]. However, this strategy requires the camera to move constantly, with still objects, to obtain reliable self-supervision.

In contrast, our proposal aims to deal with the opposite setting, where the camera placement is fixed and moving agents appear in the scene.

3 Method

We now introduce our adaptation strategy to address three issues encountered when deploying monocular solutions [36, 37] in the wild: 1) incorrect/blurred predictions for some subjects in the scene, e.g., pedestrians, cars, etc. 2) inaccurate global structure, being not properly aligned with the real scene and 3) the predictions being up to an unknown scale factor. The three are dealt with by different steps in our pipeline, as spotlighted in Fig. 2.

3.1 Lightweight Fine-tuning with Pseudo Labels

Although modern monocular depth networks exhibit powerful generalization capabilities [36, 37], they sometimes fail when used in ever-seen environments. In particular, in the case of a fixed-camera installation with a viewpoint substantially different from those observed during training, these solutions might often miss the presence of agents in the scene, such as pedestrians or cars, as shown in Fig. 1. Purposely, we design a lightweight fine-tuning procedure to improve the perception of such agents by the monocular network by relying on pseudo labels obtained in two steps.

Pseudo labels initialization. We initialize the pseudo labels with the depth values estimated by monocular depth network [36, 37] considering its excellent generalization performance except for the weaknesses mentioned earlier.

Agents rectification and fine-tuning. In order to recover the miss-estimation of subjects in the scene or the blurred estimates, we distill proper depth labels. Purposely, we first detect possibly moving agents in the scene, for instance, through an instance segmentation network such as MaskRCNN from the Detectron2 framework [46]. Then, by assuming each agent is standing or moving over the ground plane, we generate pseudo labels by replacing the depth of each instance with the depth value of the lowest pixel in the instance itself – i.e., the contact point with the ground. For complex agents, such as bicycles or motorcycles, we approximate the riders' depth to that of the vehicles. For bags and hand-holding items, e.g. umbrellas, the depth will match that of the closest pedestrian. This process ignores other semantic classes. At deployment time, this allows for rapidly collecting a small set of samples for fine-tuning the original model, considerably improving the depth accuracy for such subjects, as shown in Fig. 1.

3.2 Ground Plane Estimation and Scene Alignment

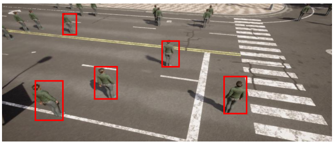
After correcting the depth for agents in the scene, there are still evident errors between scene structures reconstructed by the fine-tuned depth model and the sensed environment. As illustrated by Fig. 3 on the right, the ground plane reconstructed from the predicted depth map (red) is not properly aligned with the real 3D plane in the scene (green). This behaviour is probably a consequence of the very different viewpoints in training images [7], yielding a degradation of the predicted relative depth and scale recovery process (when feasible). To solve this issue, we aim to estimate the real ground plane in the scene and use it to restore the proper structure of the scene in the predicted depth map.

Ground Plane Estimation. According to the perspective principle, we can model a 3D plane as in the left part of Eq.1:

$$\begin{cases} a_g X + b_g Y + c_g Z = d_g \\ X = \frac{Zx}{f}, \quad Y = \frac{Zy}{f}, \quad H = \frac{Zh}{f} \end{cases} \quad \Rightarrow \quad a_g x + b_g y + c_g f = \frac{d_g f}{Z} = \frac{d_g h}{H} \tag{1}$$

where (x,y) and (X,Y) denote, respectively, the pixel coordinates and projected 3D coordinates, Z the depth, (a_g, b_g, c_g, d_g) the ground plane parameters, f the camera focal length, H and h the actual height and pixel height of an object in the scene. From it, we can derive (right side of Eq.1) the plane equation as a function of the 2D coordinates and height of a known object.

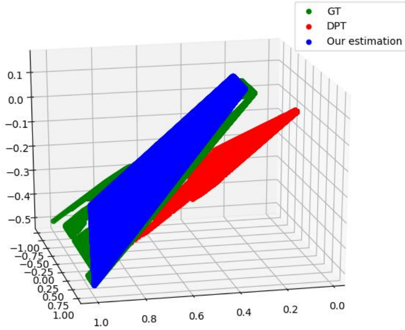
Inspired by [44], we use the moving agents previously detected as probes in the scene to estimate plane coefficients (a_g, b_g, c_g). Specifically, by detecting a single agent in more than



(a) Plane Estimation(people probes)

Ground normal	a	b	c
GT plane	-0.0002	0.7752	0.6316
DPT normal	-0.0041	0.8910	0.4540
Estimated normal	-0.0266	0.7321	0.6807

(b) Plane normal comparison



(c) Visulization of ground normal

Figure 3: **Structural misalignment between predicted depth and real scene.** For a single scene with pedestrians walking around (a), we report ground normals (a,b,c) obtained from predicted depth (red), our ground plane estimation method (blue), and ground truth (green). On the right, we visualize the misalignment between predicted and real planes.

three frames, we record their corresponding pixel height h and standing point coordinates (x,y). Then, assuming no co-linear positions, coefficients (a_g,b_g,c_g) can be estimated using the least squares method, with c and the actual height H of the agent regarded as constants. Once (a_g,b_g,c_g) is given, the relative depth Z of any ground point can be estimated as a function of $d_g H$ – i.e., up to an unknown scale factor. Despite this, this cue is enough to proceed with the alignment step discussed next.

Scene Alignment. We can exploit the estimated ground plane to align the predicted depth map, making it more congruent with the 3D structure of the scene. Given the predicted depth, a common practice for aligning it to a set of known depth values in metric scale consists of using the least squares algorithm [6, 36, 37] or a non-linear model, i.e. a CNN [45]. However, this would not fit with the priors we derive from the moving agents since we can estimate the ground plane model without precisely segmenting it from the rest of the scene. Hence, we introduce an alternative approach to align the ground plane in the predicted depth map with the one modeled by our method.

Any 3D point (X,Y,Z) in the estimated depth map can be projected onto the ground plane according to the ground parameters (a_p,b_p,c_p,d_p), as formulated in Eq.2.

$$X_g = \frac{a_p d_p - a_p c_p Z - a_p b_p Y + b_p^2 X}{a_p^2 + b_p^2} \qquad Y_g = \frac{b_p d_p - b_p c_p Z - a_p b_p X + a_p^2 Y}{a_p^2 + b_p^2} \qquad Z_g = Z \tag{2}$$

Thus, we can align the projections we obtain according to the ground plane model extracted from predicted depth maps with those yielded by the model obtained according to the moving agents' motion. To calculate the surface normal from depth predictions, we can fit $a_p X + b_p Y + c_p Z = d_p$ using the least square algorithm. It requires identifying a portion of the scene belonging to the ground plane by manually annotating a single image captured after installation or directly during deployment according to agents' motion.

After projecting all the 3D points onto the plane, i.e. (X_g,Y_g,Z), the 3D plane will be re-projected into the image plane according to the camera intrinsic parameters, yielding new pixel coordinates (x',y'). Finally, we can adjust the prediction to fit the ground plane model estimated through agents' motion substituting (x',y') in Eq. 1.

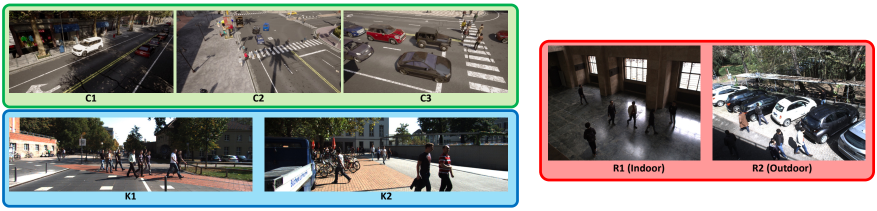


Figure 5: **Evaluation dataset.** We show a sample for each of the seven scenes from CARLA (green), KITTI (blue), or our acquisitions (red) used in our experiments.

3.3 Absolute Scale Recovery

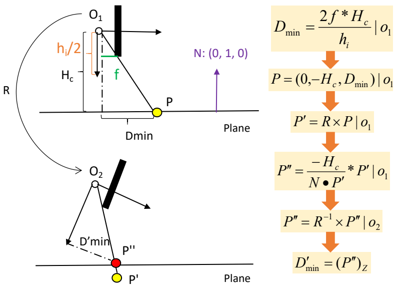


Figure 4: **Minimum depth estimation**. We exploit camera height and estimated ground plane orientation.

Although accurate in terms of relative depth, predictions of state-of-the-art models [36, 37] are often up to an unknown scale factor, whereas knowing the absolute depth is often necessary for practical applications. According to [30, 47], the missing scale can be restored by knowing the camera height over the ground. Unfortunately, they rely on the assumption that the Z axis of the camera is roughly parallel to the ground plane, a condition not always met in practice.

To remove this constraint, we design a custom method to deal with arbitrarily oriented cameras, by estimating the depth for the bottom-most pixel in the center of the image, assumed as anchor point. Fig.4 shows that in the O_1 setup, in which the Z-axis of the camera is parallel to the ground, the closest point laying on it can be represented as P, where the Z-axis projection D_{min} can be calculated from camera height H_c, image height h_i and focal f as they compose a similar triangle. In the O_2 setup, with an arbitrarily titled camera, point P is now positioned at P', and the closest point has changed to P'', which is the intersection between the origin-P' ray with the plane. Given the current ground normal and original normal $N(0,1,0)$, the pose matrix R from O_1 to O_2 can be estimated to achieve the coordinate transformation from P to P'. After transforming P'' to the system O_2 by multiplying with R^{-1}, the D'_{min} for the anchor point in the depth map will be estimated in metric scale. This latter will be used, together with predicted depth for this very same point, for restoring metric scale on the entire depth map, instead of using the median rescaling technique based on ground truth depth. This strategy assumes the anchor point is on the ground plane and not occluded by moving agents.

4 Experimental Results

4.1 Datasets

We run our experiments on a mixture of synthetic and real datasets, with a static camera mounted over the scene pointing toward roads, sidewalks, or pedestrian areas. A total of seven sequences are used – Fig. 5 shows an example for each – grouped into three categories:

Synthetic data (CARLA). We generate three sequences using CARLA simulator [8],

Scene	Method	SiLog↓	RMSE↓	Abs rel↓	Sq rel↓
C1	DPT [36]	0.051	4.098	0.184	0.707
	DPT-ft	**0.044**	**4.012**	**0.171**	**0.656**
	DPT-align	**0.018**	**3.421**	**0.103**	**0.399**
	DPT-ft-align	**0.026**	**3.677**	**0.131**	**0.512**
C2	DPT [36]	0.061	5.164	0.221	1.183
	DPT-ft	**0.029**	**3.476**	**0.144**	**0.591**
	DPT-align	**0.011**	**2.301**	**0.041**	**0.307**
	DPT-ft-align	**0.009**	**2.201**	**0.036**	**0.288**
C3	DPT [36]	0.056	2.355	0.214	0.521
	DPT-ft	**0.042**	**2.052**	**0.183**	**0.391**
	DPT-align	**0.018**	**1.282**	**0.112**	**0.154**
	DPT-ft-align	**0.014**	**1.103**	**0.098**	**0.115**
K1	DPT [36]	0.069	5.705	0.219	1.274
	DPT-ft	**0.024**	**3.824**	**0.121**	**0.541**
	DPT-align	**0.026**	**4.537**	**0.105**	**0.767**
	DPT-ft-align	**0.019**	**3.666**	**0.101**	**0.483**
K2	DPT [36]	0.149	4.718	0.337	2.289
	DPT-ft	**0.048**	**2.948**	**0.158**	**0.658**
	DPT-align	**0.057**	**3.686**	**0.235**	**1.441**
	DPT-ft-align	**0.045**	**2.876**	**0.152**	**0.552**
R1 (Indoor)	DPT [36]	0.052	1.626	0.151	0.546
	DPT-ft	**0.049**	**1.579**	**0.148**	**0.507**
	DPT-align	**0.033**	**0.999**	**0.098**	**0.293**
	DPT-ft-align	**0.036**	**1.075**	**0.107**	**0.337**
R2 (Outdoor)	DPT [36]	0.054	3.786	0.198	0.748
	DPT-ft	**0.052**	**3.654**	**0.189**	**0.723**
	DPT-align	**0.051**	**3.604**	**0.167**	**0.671**
	DPT-ft-align	**0.041**	**3.136**	**0.159**	**0.542**

Table 1: **Quantitative results – on-site adaptation.** We report error metrics on the seven sequences, for original DPT [36], DPT after lightweight fine-tuning (DPT-ft) and with test-time scene alignment (DPT-ft-align). We highlight **first**, **second**, and **third** best results.

each consisting of 800 frames at 800 × 400 resolution, dubbed C1, C2, and C3. We simulate a realistic traffic environment, where pedestrians and vehicles move around, and urban infrastructures, such as trees and buildings. The simulator also allows obtaining semantic segmentation labels, camera pose, and ground truth depth.

KITTI static sequences. Among the many samples provided by the KITTI raw dataset [14], a small amount of short, static sequences – mainly concentrated in the *Campus* category – are suitable for our experiments. We obtain two main sequences by grouping frames from *2011_09_28_drive_0016* + *2011_09_28_drive_0021* and *2011_09_28_drive_0039* + *2011_09_28_drive_0043*, dubbed K1 and K2, and counting 395 and 506 samples.

Real sequences. To further stress the flexibility of our approach, we collect two real sequences in indoor and outdoor environments, dubbed R1 and R2, counting 3024 and 2952 images. For both, we mounted a camera tilted toward the ground plane at about five meters over it. We collect images with a 27cm baseline stereo camera and use CREStereo [24] to estimate disparity maps and triangulate them into depth to obtain ground truth labels. Although imperfect, we consider these annotations accurate enough for our purposes.

4.2 Implementation details

Given its outstanding generalization performance, we adopt DPT [36] as the baseline monocular depth estimation network in our experiments, on a single 3090 GPU. We adapt it for each of the seven sequences through the pipeline introduced earlier. Concerning the lightweight fine-tuning process, we select the first 342, 463, and 406 frames from the three synthetic sequences, the first 273 and 107 frames from the KITTI sequences, and the first 684 and 679 frames from the real scenes. On top of them, we generate pseudo labels and fine-tune DPT for 30 epochs. This strategy simulates an on-site adaptation carried out on the first

		C1	C2	C3	K1	K2	R1	R2
Ground truth	Depth	6.281	9.799	5.256	5.992	5.895	6.125	6.675
[30, 47]	Depth	10.751	19.431	10.686	6.439	6.439	24.065	25.651
	Error (m)	4.470	9.632	5.430	0.447	**0.544**	17.940	18.976
Ours	Depth	6.102	9.789	5.105	5.772	5.285	6.304	6.421
	Error (m)	**0.179**	**0.010**	**0.151**	**0.220**	0.610	**0.179**	**0.254**

Table 2: **Scale recovery evaluation – anchor point.** From top to bottom: ground truth depth for the anchor point, average depth and its error according to [30, 47] and our method.

Scene	Method	Rescale	SiLog	RMSE	Abs rel	Sq rel
C1	Monodepth2 (S) [16]	Stereo	**0.014**	**4.376**	**0.065**	**0.304**
	DPT-ft-align	Ours	0.038	5.628	0.124	0.727
C2	Monodepth2 (S) [16]	Stereo	**0.011**	**8.231**	**0.051**	**0.383**
	DPT-ft-align	Ours	0.021	8.426	0.096	0.651
C3	Monodepth2 (S) [16]	Stereo	0.031	2.653	0.077	0.256
	DPT-ft-align	Ours	**0.012**	**2.101**	**0.068**	**0.095**
K1	Monodepth2 (S) [16]	Stereo	**0.005**	**2.307**	**0.042**	**0.163**
	DPT-ft-align	Ours	0.019	3.901	0.085	0.504
K2	Monodepth2 (S) [16]	Stereo	0.081	3.181	0.116	0.511
	DPT-ft-align	Ours	**0.042**	**3.026**	**0.139**	**0.553**
R1 (Indoor)	Monodepth2 (S) [16]	Stereo	0.041	1.233	0.101	0.381
	DPT-ft-align	Ours	**0.028**	**1.102**	**0.106**	**0.265**
R2 (Outdoor)	Monodepth2 (S) [16]	Stereo	**0.032**	**3.112**	**0.104**	**0.461**
	DPT-ft-align	Ours	0.081	5.076	0.141	1.112

Table 3: **Scale recovery evaluation – comparison with stereo self-supervision.** We report error metrics by MonoDepth2 trained with stereo self-supervision and our method.

frames collected after installation. Then, we evaluate the effectiveness of our pipeline on the remaining frames by enabling test-time scene alignment as well. We compute standard metrics concerning the monocular depth estimation task [11, 42], such as SiLog, RMSE, Abs Rel, and Sq Rel. The evaluation is performed both by rescaling predictions using ground truth depth itself [36], as well as by restoring absolute scale following our approach.

4.3 Quantitative results

Lightweight fine-tuning and scene alignment. We start by evaluating the effectiveness of the first two steps in our pipeline. Table 1 gathers the results achieved by the original DPT model and those obtained after performing the lightweight fine-tuning and scene alignment steps. In this experiment, absolute scale is restored according to median rescaling [51]. Starting from the former, we refer to DPT-ft for the model fine-tuned on pseudo labels without alignment. We can notice how the lightweight fine-tuning improves the accuracy compared to the original model. In particular, by correcting several errors in correspondence of the moving pedestrians or vehicles. By performing scene alignment on the predictions by the original DPT model (DPT-align) we improve the accuracy as well, often with a major gain with respect to what is achieved by the light-weight finetuning – since this latter only acts on moving agents, representing a minority of the pixels in the scene. Finally, combining fine-tuning and alignment – DPT-ft-align entries – further decreases the error in most cases, except on C1 and R1 in which the ground plane covers the largest portion of the scene compared to other sequences.

Absolute scale recovery. To conclude, we evaluate the effectiveness of our scale recovery strategy. For this purpose, we first compare our approach with alternative techniques

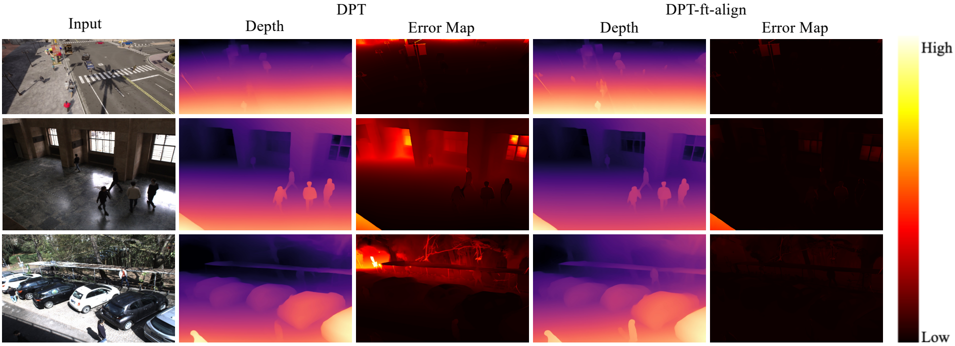


Figure 6: **Qualitative results – error maps.** We show error maps by DPT and DPT-ft-align.

[30, 47] exploiting knowledge of the camera height but assuming the camera to be parallel to the ground plane. Table 2 reports the average depth value for the anchor points estimated by the different techniques, followed by their error compared to ground truth depth. For KITTI, the anchor is replaced by the closest pixel with available ground truth. We can notice how our approach consistently restores the absolute scale more accurately. Although the two are substantially equivalent on KITTI, where the camera is almost parallel to the ground, our solution results superior in the real datasets featuring a significant camera tilt.

Finally, we compare the accuracy of depth maps predicted by DPT after having recovered scale through our technique with the results obtained by a model trained directly on-place with supervision from a stereo camera in the scene, i.e., with knowledge about the metric scale of the scene. We assume this configuration sets an upper bound at the performance a monocular network could achieve by having perfect knowledge of the scale during training. Table 3 shows how the outcome of this experiment, with MonoDepth2 [16] used for the comparison. After our scale recovery step, DPT-ft-align often yields results close to those by MonoDepth2 – outperforming it a few times – confirming the proposal's effectiveness.

4.4 Qualitative Results

To conclude, Fig. 6 shows how our whole framework dramatically reduces the error being the primary source of failure – i.e., in the presence of agents such as pedestrians and cars, or in the farthest parts of the scene, where the ground plane misalignment between predicted and real depth becomes more prominent. After processing, the error remains slightly higher on objects farther from the ground plane – e.g., structures on the sidewalk (left) or the obstacle in the very foreground (center) – where no optimization is performed by our method.

5 Conclusion

This paper proposes a novel pipeline for the on-site adaptation of a monocular depth network specifically designed to deal with fixed-camera installations. Our method allows for a lightweight fine-tuning of the model and a test-time scene alignment of the predicted depth maps by leveraging the presence of agents moving freely in the scene, as well as for recovering the metric scale of the scene by only knowing the height at which the camera is. Our experiments show how a few images collected just after deployment allow for improving the results achieved by the DPT network thanks to our solution.

Acknowledgment. We sincerely thank the scholarship supported by China Scholarship Council (CSC).

References

[1] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention. *arXiv preprint arXiv:2210.09071*, 2022.

[2] Maya Aghaei, Matteo Bustreo, Yiming Wang, Gianluca Bailo, Pietro Morerio, and Alessio Del Bue. Single image human proxemics estimation for visual social distancing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2785–2795, 2021.

[3] Shubhra Aich, Jean Marie Uwabeza Vianney, Md Amirul Islam, and Mannat Kaur Bingbing Liu. Bidirectional attention network for monocular depth estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11746–11752. IEEE, 2021.

[4] Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2800–2810, 2018.

[5] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021.

[6] Alexandra Dana, Nadav Carmel, Amit Shomer, Ofer Manela, and Tomer Peleg. One scalar is all you need - absolute depth estimation using monocular self-supervision. *ArXiv*, abs/2303.07662, 2023.

[7] Tom van Dijk and Guido de Croon. How do neural networks see depth in single images? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2183–2191, 2019.

[8] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[9] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

[10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[12] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep Ordinal Regression Network for Monocular Depth Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] Ravi Garg, Vijay Kumar BG, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision (ECCV)*, pages 740–756, Amsterdam, The Netherlands, 2016. Springer.

[14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237, 2013.

[15] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

[16] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[17] Muhammad Umar Karim Khan. Towards continual, online, self-supervised depth. *arXiv preprint arXiv:2103.00369*, 2021.

[18] Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, and R Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2656–2665, 2018.

[19] Yevhen Kuznietsov, Marc Proesmans, and Luc Van Gool. Comoda: Continuous monocular depth adaptation using past experiences. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2907–2917, January 2021.

[20] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *CVPR*, pages 89–96, 2014.

[21] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *Fourth International Conference on 3D Vision (3DV)*, pages 239–248, Stanford University, California, 2016. IEEE.

[22] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[23] Sihaeng Lee, Janghyeon Lee, Byungju Kim, Eojindl Yi, and Junmo Kim. Patch-wise attention network for monocular depth estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):1873–1881, May 2021. doi: 10.1609/aaai.v35i3.16282. URL https://ojs.aaai.org/index.php/AAAI/article/view/16282.

[24] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16263–16272, 2022.

[25] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[26] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022.

[27] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022.

[28] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. Va-depthnet: A variational approach to single image depth prediction. *CoRR*, abs/2302.06556, 2023. doi: 10.48550/arXiv.2302.06556. URL https://doi.org/10.48550/arXiv.2302.06556.

[29] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 38(10):2024–2039, 2016.

[30] Robert McCraith, Lukás Neumann, and Andrea Vedaldi. Calibrating self-supervised monocular depth estimation. *CoRR*, abs/2009.07714, 2020. URL https://arxiv.org/abs/2009.07714.

[31] Alessio Mingozzi, Andrea Conti, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Monitoring social distancing with single image depth estimation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(6):1290–1301, 2022.

[32] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[33] Valentino Peluso, Antonio Cipolletta, Andrea Calimera, Matteo Poggi, Fabio Tosi, Filippo Aleotti, and Stefano Mattoccia. Monocular depth perception on microcontrollers for edge applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1524–1536, 2021.

[34] Matteo Poggi, Fabio Tosi, Konstantinos Batsos, Philippos Mordohai, and Stefano Mattoccia. On the synergies between machine learning and binocular stereo for depth estimation from images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[35] Chakravartula Raghavachari, V. Aparna, S. Chithira, and Vidhya Balasubramanian. A comparative study of vision based human detection techniques in people counting applications. *Procedia Computer Science*, 58:461–469, 2015. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2015.08.064. Second International Symposium on Computer Vision and the Internet (VisionNet'15).

[36] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021.

[37] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022.

[38] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008.

[39] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[40] Fei Sheng, Feng Xue, Yicong Chang, Wenteng Liang, and Anlong Ming. Monocular depth distribution alignment with low computation. *arXiv preprint arXiv:2203.04538*, 2022.

[41] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[42] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017.

[43] Niclas Vödisch, Kürsat Petek, Wolfram Burgard, and Abhinav Valada. Codeps: Online continual learning for depth estimation and panoptic segmentation. *arXiv preprint arXiv:2303.10147*, 2023.

[44] Yifan Wang, Brian L. Curless, and Steven M. Seitz. People as scene probes. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 438–454, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58607-2.

[45] Diana Wofk, René Ranftl, Matthias Müller, and Vladlen Koltun. Monocular visual-inertial depth estimation, 2023.

[46] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[47] Feng Xue, Guirong Zhuo, Ziyuan Huang, Wufei Fu, Zhuoyue Wu, and Marcelo H. Ang Jr. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. *CoRR*, abs/2004.05560, 2020. URL https://arxiv.org/abs/2004.05560.

[48] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Newcrfs: Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[49] Pierluigi Zama Ramirez, Alex Costanzino, Fabio Tosi, Matteo Poggi, Samuele Salti, Luigi Di Stefano, and Stefano Mattoccia. Booster: a benchmark for depth from images of specular and transparent surfaces. *arXiv preprint arXiv:2301.08245*, 2023.

[50] Zhenyu Zhang, Stephane Lathuiliere, Elisa Ricci, Nicu Sebe, Yan Yan, and Jian Yang. Online depth learning against forgetting in monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[51] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.