This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Do More With What You Have: Transferring Depth-Scale from Labeled to Unlabeled Domains

Alexandra Dana Nadav Carmel Amit Shomer Ofer Manela Tomer Peleg

Samsung Israel R&D Center
Tel Aviv, Israel

alex.dana@samsung.com

Abstract

Transferring the absolute depth prediction capabilities of an estimator to a new domain is a task with significant real-world applications. This task is specifically challenging when images from the new domain are collected without ground-truth depth measurements, and possibly with sensors of different intrinsics. To overcome such limitations, a recent zero-shot solution was trained on an extensive training dataset and encoded the various camera intrinsics. Other solutions generated synthetic data with depth labels that matched the intrinsics of the new target data to enable depth-scale transfer between the domains.

In this work we present an alternative solution that can utilize any existing synthetic or real dataset, that has a small number of images annotated with ground truth depth labels. Specifically, we show that self-supervised depth estimators result in up-to-scale predictions that are linearly correlated to their absolute depth values across the domain, a property that we model in this work using a single scalar. In addition, aligning the field-of-view of two datasets prior to training, results in a common linear relationship for both domains. We use this observed property to transfer the depth-scale from source datasets that have absolute depth labels to new target datasets that lack these measurements, enabling absolute depth predictions in the target domain.

The suggested method was successfully demonstrated on the KITTI, DDAD and nuScenes datasets, while using other existing real or synthetic source datasets, that have a different field-of-view, other image style or structural content, achieving comparable or better accuracy than other existing methods that do not use target ground-truth depths.

1. Introduction

Monocular depth estimation is a fundamental problem in computer vision with numerous scene understanding applications, such as autonomous driving and navigation [16, 19, 39, 48], robotics [14, 34, 41] and augmented reality [13, 32]. Current methods for training Monocular Depth Estimators (MDE) include two major methodologies: the first uses full-supervision [5, 17, 25, 36] or mixed-supervision [2, 4, 24, 35], where ground-truth (GT) depths are measured directly by LiDARs or reconstructed using a stereo setting [11, 33, 40, 45], achieving *absolute* depth predictions. However, fine-tuning such models on new scenes or on images collected using different sensors, requires collecting their corresponding depth measurements, complicating the acquisition setup with additional depth sensors or cameras, increasing setup complexity and costs [14].

Thus, to enable training or fine-tuning MDEs on a new domain using only images, multiple efforts [10, 21–23, 27, 46, 53] were invested to improve the self-supervised regime.

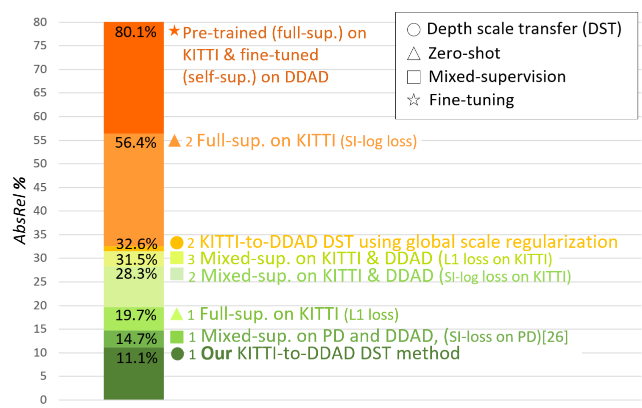

Figure 1. Demonstrating various methods for achieving absolute depth predictions on the DDAD target dataset (front camera), when no GT depth labels from DDAD are available for training. The *AbsRel* metric (lower is better) was measured on the DDAD validation dataset for each presented method. Mixed-supervised models (□) were trained using self-supervision on both DDAD and another source dataset (KITTI or Parallel Domain (PD)) and with full-supervision on source GT depths using the mentioned loss.

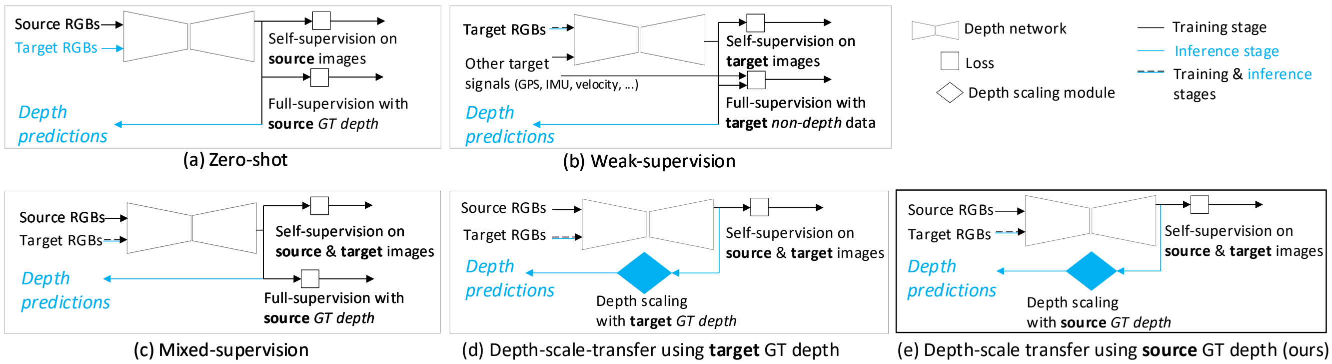

Figure 2. Methods for inferring *absolute* depth from single target RGB images, when no target GT *depth* measurements are available.

In this approach, two images acquired at different times are used to predict their depth using projective geometry [31]. Due to the nature of this concept, the scale of the predicted depth is ambiguous [31]. In this work we refer to predicted depths that lack real-world scale units as *up-to-scale* depths.

To compensate for the lack of GT depth measurements when collecting new data, one might suggest using already existing datasets collected with GT depths that do have real world-scale (*source* dataset), and transferring from them the depth-scale property to the newly collected images (*target* dataset). To show the motivation behind our work and the non-triviality of such depth-scale transfer between datasets collected using various sensors, we evaluated different approaches that might seem intuitive. To demonstrate a real-world scenario, we selected for the target domain the relatively new DDAD [23] dataset and for the source domain we selected the well-known KITTI [20] dataset, which was collected using a sensor with a different field-of-view (FOV).

Our various early attempts included zero-shot estimation (see Figure 2a), fine-tuning using self-supervision on the target domain, and mixed-supervision on both domains (see Figure 2c). More details can be found in Sections 2 and 4.3. However, Figure 1 shows that such methods result in poor accuracy. A recent work [26] applied mixed-supervision on the target DDAD images (without using any DDAD GT depths) and another source dataset, but did achieve high absolute depth accuracy on DDAD (see Figure 1$\square_1$).

Thus, what impeded all our early attempts? An additional investigation indicated that the authors [26] used for the source domain a synthetic dataset that was specifically generated with the same intrinsics as the target dataset [1]. However, this effort is costly and needs to be repeated for each sensor with a different FOV. Thus, in this work we study an *alternative* for synthetic data generation for this purpose, by reusing existing synthetic or even *real* datasets with GT depth labels, collected using *variate* FOVs.

In order to use arbitrary datasets as source domains, we first studied fundamental aspects of the depth-scale, as predicted by self-supervised models trained using reprojected geometry. We show that although such models can predict only up-to-scale depths, these are *linearly* correlated with their respective GT depths, not only per a single image, but also across multiple images, displaying linear correlation characteristics per dataset, a property which we refer to in this work as *linear* depth ranking. Moreover, we show that when *adjusting* images from two different domains to a single FOV, under the assumption of similar camera heights, training the MDE on images from both domains results in a shared depth ranking scale, regardless of possible domain gaps [30, 51].

In this work we propose to transfer the depth-scale between domains by first training the MDE using self-supervision on images from both the source and target domain. Then, we model the relationship between the *source* up-to-scale depths and their GT depth values using a *single* scaling factor. Finally, we use this factor to scale the *target* up-to-scale depth predictions, achieving real-world absolute depth predictions on the new domain.

In summary, our contributions are as follows:

- We propose a novel depth-scale transfer method that uses existing source data with GT labels, that does not increase the computations of the used MDE.
- The suggested method can reuse any existing datasets with GT labels, real or synthetic.
- Our method achieves comparable or better accuracy on KITTI, DDAD and nuScenes with respect to existing depth-scale transfer methods, even when using a vanilla lite MDE, as long as local motion is filtered out efficiently.

2. Related work

Over the years, various solutions were suggested to overcome the lack of target GT depth measurements for training MDEs to predict absolute depth from target images. Here we cover the main approaches, that are also presented by category in Figure 2. The first approach is implemented as zero-shot [44] (see Figure 2a), where a model is trained on source datasets, and used to infer depth on target images, in the hope of generalizing well on the new domain. A recent zero-shot model [29] successfully overcame the geometrical domain gap between the source and the target domain by training a transformer-based architecture on a variety of source datasets (containing more than 700,000 training images with GT) that were further augmented to support various focal lengths. In addition, the camera parameters were embedded to enable zero-shot capabilities on various target datasets. In our work, we show an alternative solution to close the geometrical domain gap that uses only few annotated *source* samples (validation/test splits, less than 3,000 images) with a significantly lighter model (x50 less parameters). In addition, since our solution also uses target domain images, it could be re-adjusted to the new domain.

Another line of work used weak-supervision (see Figure 2b) to train the model directly on target data. Up-to-scale depth values were scaled during training with non-depth measurements such as the car velocity [23], its GPS location [9] or IMU measurements [50] through the relative translation regularization between frames in absolute distance units. However, such solutions also require signals acquisition with dedicated and costly setups. Other works [28, 47] trained the MDE using self-supervision on target images from multiple cameras and regularized the estimated translation and rotation between the cameras using the measured pose translation and rotation between them, to indirectly regularize for the depth-scale. However, such solutions require the use of at least two cameras for collecting training data and rely on their calibration accuracy.

A variety of works used mixed-supervision (see Figure 2c) to transfer depth-scale from source datasets with GT depth labels to new target domains that were acquired without it. To overcome difficulties arising from training the MDE on source and target images collected by different sensors types, these solutions used for the source domain synthetic data that was specifically created with the same intrinsics, extrinsics and even similar geometrical structures as in the target domain. Thus, a large number of works were limited to demonstrating their method only on the target KITTI dataset, while the source data was taken from the synthetic vKITTI [18] and vKITTI2 [7] that were specifically tailored to match KITTI. Since real and synthetic data differ in style, initial attempts focused on closing these gaps by incorporating style-transfer elements in the mix-supervision [3, 37, 51, 52], achieving absolute depth predictions. A recent work [26] incorporated unsupervised domain adaptation principles in the mixed-supervision, successfully demonstrating depth-scale transfer to DDAD. To achieve this, the authors generated a synthetic source dataset [1] with the same intrinsics of the target domain. In our work we show an alternative for this step.

Recently, two approaches decoupled the up-to-scale depth ranking problem of the self-supervised regularization from the depth-scale estimation. In the first approach, the MDE was directly trained using self-supervision on target images. Then, depth-scaling was separately estimated per target *image*, by using the known camera height and estimation of the road plane [38, 49] (see Figure 2d). However, this method requires sufficient visible free road during inference to estimate its plane, which is not always possible in traffic jams or turns. Our solution does not limit the target data by such condition.

In the second approach, the depth-scale is estimated by a module that only uses source data (see Figure 2e). A previous work [6] suggested to train the MDE on the source data with self-supervision and an additional regularization loss to align the depth-scale of all predictions to a single arbitrary value. Then, the real depth-scale factor was separately estimated and corrected using GT depth measurements from a few source images. We also evaluated this approach when applying self-supervision on *both* source and target domains using this regularization loss, and estimated the depth-scale from the source data, but did not achieve satisfactory results (see Figure 1$\bigcirc_2$ and Section 4.3).

In a recent work [42], the authors trained an MDE using self-supervision on datasets from both source and target domains. Then, they trained a separate CNN module with full-supervision using source GT depths to predict the depth-scale of up-to-scale depth maps. However, also this solution was demonstrated only on target datasets that have specially tailored synthetic source datasets.

In our work we also propose to transfer the depth-scale from another source domain (see Figure 2e). However, contrary to previous solutions, we reuse any *existing* source data type (real or synthetic), without limiting it to the intrinsics (FOV) of the target data. In addition, our depth-scale modality is implemented using a single scalar, thus does not increase the computations of the existing MDE.

3. Method

3.1. Depth prediction architecture

To demonstrate the concept behind this work, we selected for the depth Φ and pose Ψ networks similar designs as used for Monodepth2 [21], consisting of 4.7M parameters. The MDE was trained using self-supervision with only a photometric loss [21], after filtering local motion (see Supplementary material). Such architecture is sufficient for quantifying the effectiveness of the depth-scale transfer between domains. Additional depth accuracy improvements can be achieved using more complex architectures [5, 23] or regularization losses [26], but are beyond the scope of this study.

3.2. Adjusting the FOV of the source data to the target domain

Training an MDE on images collected using different camera FOVs (real or synthetic) introduces significant geometrical differences that a naïve self-supervised training regime cannot easily compensate for. To enable fine-tuning or

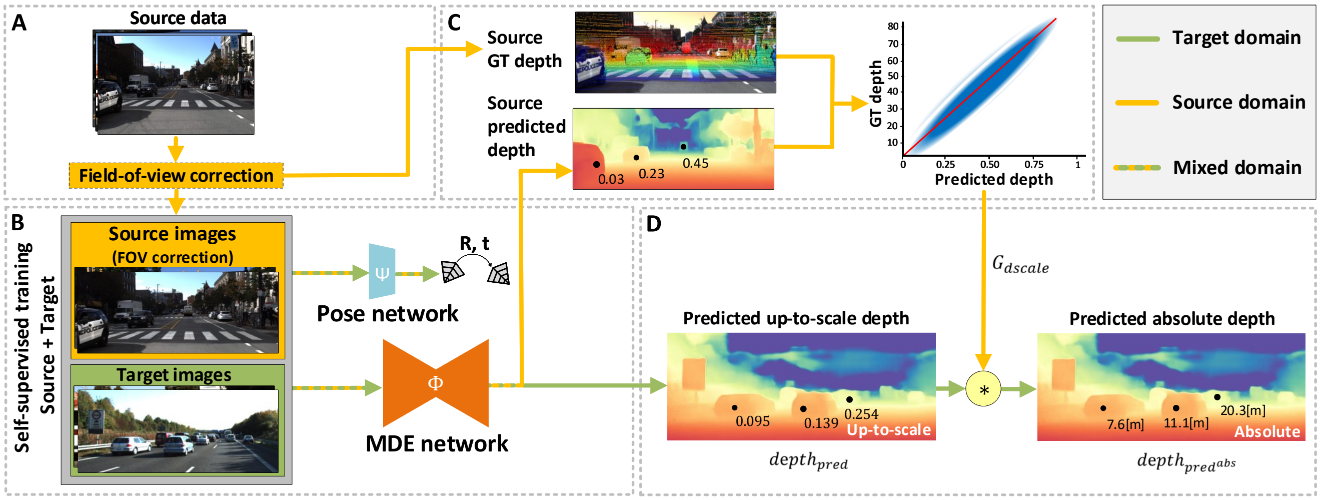

Figure 3. Overview of our solution. (A) The FOV of source data is adjusted to the target FOV. (B) The depth and pose networks are trained using self-supervision on both source and target training images (mixed batches). (C) Data from the source domain is used to generate the GT *vs*. predicted up-to-scale depths mapping. We apply the linear Theil-Sen regressor to calculate the G_{dscale} depth-scaling factor for the trained MDE. (D) Estimated target up-to-scale depths are multiplied by the G_{dscale} factor, resulting in absolute depth predictions.

training using self-supervision on images collected with sensors of different FOVs, without breaking the scene geometrical consistency, we propose to adjust the FOV of the source images to the FOV of the target images, resulting in training data with homogeneous FOV and aspect ratio. The FOV of a camera is calculated as:

$$\angle FOV = 2 \cdot atan\left(\frac{w}{2f}\right) \tag{1}$$

F)) where f is the focal length of the camera and w is the width of the image, both in pixel units. Let us denote the camera focal length and the image width in the target domain by f_T and w_T respectively, and the camera focal length in the source domain by f_S. To adjust the FOV of the source images to the target FOV, we center cropped the source image using the adjusted width w_S to $w_{S\to T}$, as follows (assuming $f_S < f_T$):

$$\angle FOV_T = 2 \cdot atan\left(\frac{w_T}{2f_T}\right) = 2 \cdot atan\left(\frac{w_{S\to T}}{2f_S}\right) \tag{2}$$

resulting in:

$$w_{S\to T} = w_T \frac{f_S}{f_T} \tag{3}$$

The adjusted crop height was similarly calculated $h_{S\to T} = h_T \frac{f_S}{f_T}$. Finally, the image crop was resized to the target image size to enable training on mixed batches. The same process was applied to the source GT map. In case where $h_{S\to T}$ was higher than h_T or f_S was larger than f_T the crop was padded, as detailed in the Supplementary material.

3.3. Estimating the depth scaling factor using a linear estimator

We inferred the trained MDE Φ on the *source* test split and then analyzed the relationship between the predicted up-to-scale depths $depth_{pred}$ and their GT depth values. Since the training regime was designed to predict depths without any offset, we expected zero-depth predictions to match zero absolute depth. The GT *vs*. the predicted up-to-scale depth scatter plot indicated a linear relationship (see Section 4.1) across the evaluated data, therefore we chose to model this relationship using a linear fitter with G_{dscale} slope and zero intercept:

$$depth_{GT} \cong G_{dscale} \cdot depth_{pred} \tag{4}$$

G_{dscale} was estimated using the Theil-Sen regressor [12] (see Supplementary material), which was selected due to its high robustness to outliers. When the relationship was fitted per *image*, the slope was noted with I_{dscale}.

A previous work [42] chose to model the relationship between the source up-to-scale depth map and the depth-scale factor using a CNN-based network. For comparison purposes, we also implemented such solution. Additional details and analyses can be found in the Supplementary.

3.4. Overview of our solution

First, we adjusted the FOV of the source domain data (train and test splits) to match the FOV of the target domain, as described in Section 3.2 and Figure 3A. Training images from both source and target domains were randomly split into batches of four and used to train networks Φ and Ψ in a self-supervised manner (see Section 3.1 and Figure3B).

Next, we used the depth network Φ to infer the up-to-scale depths of images from the test split of the *source*. We fit the relationship between these predictions and their respective GT depths (see Section 3.3 and Figure 3C).

Finally, to estimate absolute depths on the $target$ domain, we inferred the up-to-scale depth of image i from the

Target	Source	Our Source-to-Target DST $\downarrow AbsRel$	$\downarrow AbsRel_{norm}$	$scale_{ratio}$	Full-sup. $\downarrow AbsRel$	Self-sup. $\downarrow AbsRel_{norm}$	Target-to-Target DST $\downarrow AbsRel$
DDAD1	KITTI	0.111	0.111	1.00±0.05	0.100	0.120	0.120
	vKITTI2	0.124	0.127	1.01±0.05			
KITTI	DDAD9	0.084	0.076	1.00±0.05	0.079	0.087	0.093
	vKITTI2	0.088	0.075	1.02±0.06			
nuScenes1	KITTI	0.165	0.19	0.90±0.05	0.059	0.332	0.425
	vKITTI2	0.141	0.19	0.94±0.04			

Table 1. Comparing the accuracy of our suggested depth-scale transfer (DST) method to other training regimes for the KITTI [43], DDAD [23] and nuScenes [8] target datasets. Our method is trained using self-supervision on a mix of images from the source and target datasets. In the last three columns show accuracy of models that are trained directly on the target train split using full-supervision, self-supervision and self-supervision and scaling using the G_{dscale} calculated on the target train split.

target test split using the depth network Φ and multiplied the resulting up-to-scale depths by G_{dscale}, obtaining target absolute depth predictions $depth^{i}_{pred^{abs}}$ (see Figure 3D):

$$depth^{i}_{pred^{abs}} = G_{dscale} \cdot depth^{i}_{pred} \tag{5}$$

For comparison, we also trained the MDE using full-supervision with a L1 loss and *target* GT depths to gauge the upper threshold accuracy of the used MDE architecture.

3.5. Datasets

KITTI. This dataset [18] is a common benchmark for depth evaluation. The front cameras have a FOV of 81° and are located 1.65 m above the ground.
DDAD. This dataset [23] was collected using six cameras. The front and rear cameras have a FOV of 47°/83° and are located 1.55 m above the ground. Let us denote the data collected using the front and rear cameras as DDAD1 and DDAD9 respectively, following the dataset camera numbering convention.
vKITTI2. This dataset [7] was recently released as a more photo-realistic version of vKITTI [18]. The camera has a FOV of 81° and is located 1.58 m above the ground.
nuScenes. In our experiments we used only images from the front camera [8], which has a FOV of 64.8°and located 1.51 m above the ground.

The KITTI, DDAD1 and nuScenes datasets were used as target domains. For the DDAD1 target domain we used KITTI and vKITTI2 as source domains. For the KITTI target domain we used vKITTI2, DDAD1 and DDAD9 as source domains (thus showing no limitation on the size of the source focal length) and for nuScenes1 we used KITTI and vKITTI2 as source domains. Additional details about these datasets can be found in the Supplementary material.

3.6. Depth evaluation metrics

To measure the accuracy of the predicted absolute depth values, we used the absolute relative depth accuracy (*AbsRel*) metric, without applying any normalization to the predicted depths [15]. To estimate the accuracy of the predicted up-to-scale depths, those were normalized first using the ratio between the medians of the predicted and the GT depth values (per image) [53], resulting in $AbsRel_{norm}$ (see Supplementary material).

To assess scaling similarity to GT depth measurements, we also calculated the median ratio between the GT and the predicted absolute depths per image, and then averaged this value across the entire test split, resulting in $scale_{ratio}$.

4. Results

Figure 1○$_1$ and Table 1 show that our method was able to transfer the depth-scale from various domain sources (real or synthetic) acquired with different FOVs to various target datasets, achieving an $AbsRel$ of 11.1% for the DDAD1 dataset, 8.4% for the KITTI dataset and 14.1% on the nuScenes1 dataset. The mean predicted depth deviated on average by less than 6% with respect to the GT depth (see $scale_{ratio}$ in the fifth column of Table 1), demonstrating the accuracy and effectiveness of our method.

To estimate an upper bound accuracy of the used MDE architecture, we also directly trained the MDE with full-supervision using target GT depths. This training regime achieved for DDAD1 and KITTI an $AbsRel$ of 10.0% and 7.9% respectively (see Table 1, sixth column). This shows that our depth-scale transfer method is competitive with fully-supervised methods that directly use target GT depth measurements, when controlling for the same MDE architecture. Additional depth-accuracy metrics, as well as visual examples of the estimated depth maps can be found in the Supplementary material.

The $AbsRel_{norm}$ metric is agnostic to the depth-scale estimation quality, thus reflects errors related to other factors such as the limited capacity of the network, insufficient loss regularization, poor image quality and domain gaps due to limited training data, *etc*. The results in Table 1 (third *vs*. fourth columns) indicate of a small difference between the $AbsRel$ and the $AbsRel_{norm}$, providing an additional metric to quantify our depth-scale related error.

Target	All predictions			Predictions with $AbsRel_{norm}$ < 15%			
	I_{dscale}	G_{dscale}	↑Pearson coefficient	Remaining pixels	I_{dscale}	G_{dscale}	↑Pearson coefficient
KITTI	85.7±9.3	84.4	0.93	83.6%	84.6±8.3	83.1	0.98
vKITTI2	105.2±16.1	107.5	0.79	61.1%	115.2±11.1	113.4	0.98
DDAD1	124.2±14.1	122.5	0.90	77.6%	125.9±12.3	124.4	0.98
DDAD9	115.7±14.6	118.0	0.85	59.9%	120.9±13.9	121.7	0.97
nuScenes1	70.8±19.2	75.8	0.76	53.0%	70.4±18.0	78.9	0.97

Table 2. Relationship between the GT and the predicted up-to-scale depths of our MDE, when separately trained on various datasets using self-supervision. Evaluations were done on their test split, using all predictions or predictions with $AbsRel_{norm}$ < 15% (remaining pixels % after the filtering is mentioned in the fifth column). For each data slicing the relationship was separately fitted *per image* using the linear model described in Section 3.3. Its mean and standard deviation across all test images is reported in the I_{dscale} column. The relationship was also fitted using the same model directly on *all* test images and depicted by the G_{dscale} factor and the Pearson correlation coefficient.

To evaluate the impact of training the MDE on data from an additional (source) domain on the depth estimation accuracy of the target domain, we also trained the MDE using self-supervision directly on the target domain. The results in Table 1 (fourth *vs*. seventh column) indicate that training on additional data did not deteriorate the accuracy of the model. This was specifically prominent on the nuScenes1 dataset, where training on an additional source of data significantly improved the $AbsRel_{norm}$ metric from 33.3% to 19%, which was also previously reported as significantly poor [29]. We refer to the reasons behind this significant improvement in the Supplementary material.

For comparison, we also scaled the self-supervised predictions using the estimated global G_{dscale} calculated on the target data, achieving similar results to the scaling per image using GT target measurements (Table 1, seventh *vs*. eight columns). In the next sub-sections we will present the analyses that laid the foundations for our suggested inter-domain depth-scale transfer method.

4.1. Studying per-domain the relationship between GT and predicted up-to-scale depths

We started our analysis by separately training the MDE on datasets from various domains. For each trained model the relationship between the GT and the predicted up-to-scale depths was analyzed on its respective test split. An MDE trained using projective geometry [15, 31] is expected to linearly rank the predicted depth per *image* [53]. During training, no additional bias corrections were applied to the network output, thus we expected GT values close to zero to be mapped to predicted up-to-scale depths close to zero (*i.e*. zero offset) (see Figure 4A). For each evaluated test split we fit *per image* the GT to up-to-scale depths using the linear model described in Section 3.3 (I_{dscale} in Table 2 and Figure 4C).

A more thorough analysis of the GT *vs*. the predicted up-to-scale scatter plot in Figure 4A indicated of a common linear trend across all test images, achieving a Pearson correlation coefficient bigger than 0.76 (see Table 2). Measurements were fitted using the linear model described in Section 3.3 and the calculated G_{dscale} factor is reported in Table 2. The scatter plots in Figure 4A revealed that some measurements are outside of the main linear trend. We hypothesized that these outliers could result from poor up-to-scale predictions of the MDE. To validate this assumption, we filtered out predicted depths with $AbsRel_{norm}$ >15%. As seen in Figure 4B, the filtering indeed removed the majority of outliers, increasing Pearson correlation coefficient to above 0.97. Although the obtained slope might slightly vary across images (see Figure 4C), their mean is comparable to the slope G_{dscale} calculated across the entire split (see Table 2), suggesting that the training process of the MDE converges the depth ranking per-image into a *common* linear ranking scale for all depths in the scenes of the domain it was trained on.

We hypothesize that the observed I_{dscale} variability could be explained by imperfect convergence due to non-optimal training losses, architecture capacity, image quality, generalization gaps and other factors. Conducting experiments to measure the impact of each factor on the slope variability across images is beyond the scope of this work; however, Table 2 shows that when removing poor up-to-scale predictions ($AbsRel_{norm}$ > 15%), the variability decreased, supporting this assumption.

To reinforce that the observed linearity results from the projective geometry based loss, regardless of the network architecture, we also analyzed up-to-scale predictions of a significantly different architecture, PackNet [23]. This network uses 3D convolutions and a different encoder-decoder architecture, but was trained with a similar loss. As shown in the Supplementary material, the predicted up-to-scale depths of this self-supervised MDE are also linearly correlated to the GT depths.

The obtained mean of the slopes per individual image is similar to G_{dscale} and their variability is substantially smaller than their mean. Thus, in this work we evaluated how well the single depth-scaling factor G_{dscale} per dataset could be used to model the relationship between the GT and the predicted up-to scale depths.

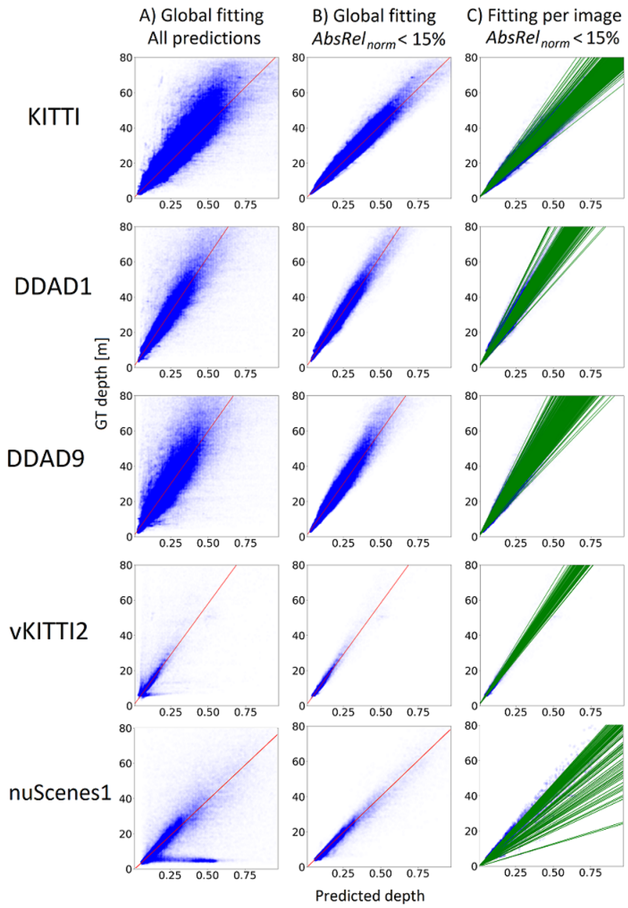

Figure 4. Scatter plots of the GT *vs*. the predicted up-to-scale depth values. Our MDE was separately trained on various datasets using self-supervision and inferred on test images from the same domain. (A) All test data. The red line depicts the linear fit applied on all data (see Eq. (4). (B) Predictions with $AbsRel_{norm}$ > 15% were filtered out and the fitting was recalculated. (C) The GT *vs*. the predicted depth relationship was similarly linearly fitted *per image* and depicted by different green lines.

4.2. Training the MDE on images from two domains

Modeling the GT to predicted up-to-scale depth relationship resulted in highly varying G_{dscale} values for different domains (see Table 3, third column). Thus, even if a self-supervised MDE is able to linearly rank depths, the modeled G_{dscale} is not transferable to other domains. To demonstrate that G_{dscale} differences cannot be explained only by different FOVs, an additional MDE was separately trained on source datasets after adjusting the source images to match the target FOV (see Table 3, fourth column). Applying the source G_{dscale} on the target, achieved an $AbsRel$ of 0.225/0.337/0.491/0.123 for each row in Table 3, showing that this step on its own is insufficient.

		G_{dscale} ($AbsRel_{norm}$ <15%)		
Test domain	**Type**	Separate trainings	Separate trainings $S \xrightarrow{FOV} T$	Joint training $S \xrightarrow{FOV} T$
KITTI	T	83.1	83.1	99.3
DDAD9	S	121.7	102.8	99.5
KITTI	T	83.1	83.1	100.7
vKITTI2	S	113.4	113.4	106.5
DDAD1	T	124.4	124.4	136.5
KITTI	S	83.1	63.9	136.1
DDAD1	T	124.4	124.4	127.3
vKITTI2	S	113.4	127.6	129.2

Table 3. G_{dscale} values estimated on test splits of various datasets (first column). Second column indicates which dataset was used as source (S) or target (T). Third column: two MDEs were trained separately on the source and the target training datasets; Fourth column: two MDEs were trained separately on the source and the target training datasets, the source images were adjusted to the target FOV ($S \xrightarrow{FOV} T$); Fifth column: a single MDE was trained on a mixture of training data from the target and the source domains after the source images were adjusted to the target FOV.

To overcome the G_{dscale} high differences between each pair of domains, we hypothesized that training the MDE using self-supervision jointly on both the source and the target datasets could potentially result in ranking the depths of both domains on a *common* scale, thus achieving a *de facto* inter-domain depth ranking. To enable training on two domains, the images from the source domain were adjusted to the FOV of the target domain, as described in Section 3.2. The fifth column in Table 3 shows that training the MDE on data from two domains resulted in similar G_{dscale} values for both the source and the target test splits, suggesting that the calculated G_{dscale} on the source data could be applied also on the target data. Finally, we also evaluated the effect of incorrect source image resizing to align the size of source and target image without considering the effect of FOV, we implemented two naïve alternatives. More details about this analysis can be found in the Supplementary material.

4.3. Straightforward attempts for depth-scale-transfer between domains

We report the accuracy of various early approaches that we implemented in attempt to compensate for the lack of GT target measurements, as mentioned in Section 1. We started with a zero-shot approach, where the MDE was trained using full-supervision with varying losses (L1, SI-log) on the KITTI dataset. The model was inferred on the DDAD1 dataset, but achieved non-satisfactory results (see Figure 1△), as expected when the geometry of the target scene is significantly different. We also evaluated the accuracy of a fully-supervised model on KITTI, that was fine-tuned using self-supervision on the DDAD1 target. However, Figure 1☆

shows poor accuracy, due to the loss of depth-scale.

For the mixed-supervision implementation (see Figure 2c), we did not have access to the specially tailored synthetic data (PD [1]) that the authors [26] used as source data for the target DDAD1 dataset, thus we used as a source the KITTI dataset. However, this approach obtained unsatisfactory results (see Figure 1$\square_{2,3}$), since retrospectively, these datasets cannot be straightforwardly mixed together due to their different geometry (as supported by results in Table 3).

We also trained the MDE using self-supervision on both the source and target train splits, with an additional regularization loss that aligns the depth-scale of predictions in both domains [6], and then estimated the depth-scale factor using GT depths of a few source images (category (e) in Figure 2). However, this approach also failed to achieve satisfactory accuracy (see Figure 1$\bigcirc_{2}$). This low accuracy can be also explained by the inability to straightforwardly mix together datasets with different FOVs.

4.4. Accuracy comparison to other methods

Finally, we compared our method to other absolute depth predictors that were trained without any target GT depth measurements (see Table 4), using the $AbsRel$ metric on KITTI, DDAD1 and nuScenes1 target domains. These were described in Section 2 and Figure 2a-e and include zero-shot [29], weakly-supervised methods [9, 23, 28, 50], mixed-supervised methods [26, 37] and depth-scale transfer from target GT depth [38, 49] or source GT depth [42].

The zero-shot model [29] achieved better accuracy on two out of the three target domains, however it used a significantly bigger training dataset (700,000 images) and larger architecture (232.6M parameters *vs*. 4.7M in our model). Nonetheless, our suggested method achieved competitive or better accuracy than the rest of the methods, while using both existing real or synthetic source datasets, collected with different FOVs than the target data, without being limited by road visibility.

Target	Method	Ref.	Source	↓$AbsRel$		
				[15]	[23]	[8]
KITTI	Zero-shot	[29]	Multiple	0.100	-	-
	Weak-sup	[23]	-	0.107	-	-
		[50]	-	0.108	-	-
		[9]	-	0.109	-	-
	Mixed-sup	[37]	CS	0.119	-	-
		[37]	vKITTI	0.120	-	-
		[26]	vKITTI2	0.107	-	-
	DST-target	[38]	-	0.113	-	-
		[49]	-	0.118	-	-
	DST-source	[42]	vKITTI2	0.109	-	-
		Ours	vKITTI2	0.108	-	-
		Ours	DDAD9	0.110	-	-
		Ours	DDAD1	0.117	-	-
DDAD1	Zero-shot	[29]	Multiple	-	0.100	-
	Weak-sup	[28]	-	-	0.130	-
	Mixed-sup	[26]	PD	-	0.147	-
	DST-source	**Ours**	KITTI	-	0.111	-
		Ours	vKITTI2	-	0.124	-
nuSc1	Zero-shot	[29]	Multiple	-	-	0.150
	Weak-sup	[28]	-	-	-	0.186
	DST-source	**Ours**	KITTI	-	-	0.165
		Ours	vKITTI2	-	-	0.141

Table 4. Comparing absolute depth estimation accuracy of methods that do not have access to target GT depth measurements during training. nuSc1, CS and PD are abbreviations of nuScenes1, Cityscapes and Parallel Domain. References in the $AbsRel$ columns indicate the used evaluation dataset. Training regimes abbreviations: weak-supervision (Weak-sup), mixed-supervision (Mixed-sup), depth-scale transfer using source-data (DST-source), depth-scale transfer using target-data (DST-target).

5. Discussion

Training or fine-tuning absolute depth estimators on collected images from a different domain without using their GT depth measurements is an existing challenge with real-world significance. Current solutions include zero-shot models that are trained on a large number of labeled images, mixed-supervised solutions that require generating tailored synthetic data to resemble the scenes of the new domain or the target sensors specification, or solutions that estimate the depth-scale from the target domain using some apriori knowledge about the camera setup.

In this work we suggested an alternative method that reuses a relatively small number of existing labeled real or synthetic images, to transfer from them the depth-scale to new domains, emphasizing the insensitivity of this approach to style or structural domain gaps. Our method leveraged the observed depth ranking linearity of projective geometry self-supervision, resulting in a lite-weight solution that does not alter the MDE architecture.

Although we demonstrated the method using a self-supervised vanilla architecture and efficient local motion filtering, we achieved competitive or better results than other existing solutions. More advanced architectures and losses can independently improve the up-to-scale depth predictions, which in turn impact the accuracy of the absolute depth predictions, even when using an ideal depth-scaling factor. We postulate that more accurate self-supervised MDEs can also reduce the variability of the GT to up-to-scale relationship across images, thus enabling a better generalization of the single depth-scale correction factor.

Finally, the suggested solution is also highly applicative for continuous learning, where depth models require adjustments to new scenes, without the need to collect additional GT depth measurements.

References

[1] Parallel domain. https://paralleldomain.com/, 2021. 2, 3, 8

[2] Ali Jahani Amiri, Shing Yan Loo, and Hong Zhang. Semi-supervised monocular depth estimation with left-right consistency using deep neural network. In *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 602–607. IEEE, 2019. 1

[3] 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. 3

[4] Jongbeom Baek, Gyeongnyeon Kim, and Seungryong Kim. Semi-supervised learning with mutual distillation for monocular depth estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 4562–4569. IEEE, 2022. 1

[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. 1, 3

[6] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Advances in neural information processing systems*, 32, 2019. 3, 8

[7] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. 3, 5

[8] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 5, 8

[9] Hemang Chawla, Arnav Varma, Elahe Arani, and Bahram Zonooz. Multimodal scale consistency and awareness for monocular self-supervised depth estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5140–5146. IEEE, 2021. 3, 8

[10] Xingyu Chen, Thomas H Li, Ruonan Zhang, and Ge Li. Frequency-aware self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5808–5817, 2023. 1

[11] Jaehoon Cho, Dongbo Min, Youngjung Kim, and Kwanghoon Sohn. Deep monocular depth estimation leveraging a large-scale outdoor stereo dataset. *Expert Systems with Applications*, 178:114877, 2021. 1

[12] Xin Dang, Hanxiang Peng, Xueqin Wang, and Heping Zhang. Theil-sen estimators in a multiple linear regression model. *Olemiss Edu*, 2008. 4

[13] Catherine Diaz, Michael Walker, Danielle Albers Szafir, and Daniel Szafir. Designing for depth perceptions in augmented reality. In *2017 IEEE international symposium on mixed and augmented reality (ISMAR)*, pages 111–122. IEEE, 2017. 1

[14] Xingshuai Dong, Matthew A Garratt, Sreenatha G Anavatti, and Hussein A Abbass. Towards real-time monocular depth estimation for robotics: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):16940–16961, 2022. 1

[15] 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. 5, 6, 8

[16] João M Faria and António HJ Moreira. Implementation of an autonomous ros-based mobile robot with ai depth estimation. In *IECON 2021–47th Annual Conference of the IEEE Industrial Electronics Society*, pages 1–6. IEEE, 2021. 1

[17] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 1

[18] A Gaidon, Q Wang, Y Cabon, and E Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. 3, 5

[19] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 740–756. Springer, 2016. 1

[20] 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. 2

[21] 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. 1, 3

[22] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8977–8986, 2019.

[23] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2485–2494, 2020. 1, 2, 3, 5, 6, 8

[24] Vitor Guizilini, Jie Li, Rares Ambrus, Sudeep Pillai, and Adrien Gaidon. Robust semi-supervised monocular depth estimation with reprojected distances. In *Conference on robot learning*, pages 503–512. PMLR, 2020. 1

[25] Vitor Guizilini, Rares Ambrus, Wolfram Burgard, and Adrien Gaidon. Sparse auxiliary networks for unified monocular depth prediction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11078–11088, 2021. 1

[26] Vitor Guizilini, Jie Li, Rareș Ambruș, and Adrien Gaidon. Geometric unsupervised domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8537–8547, 2021. 2, 3, 8

[27] Vitor Guizilini, Kuan-Hui Lee, Rareş Ambruş, and Adrien Gaidon. Learning optical flow, depth, and scene flow without real-world labels. *IEEE Robotics and Automation Letters*, 7 (2):3491–3498, 2022. 1

[28] Vitor Guizilini, Igor Vasiljevic, Rares Ambrus, Greg Shakhnarovich, and Adrien Gaidon. Full surround monodepth from multiple cameras. *IEEE Robotics and Automation Letters*, 7(2):5397–5404, 2022. 3, 8

[29] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9233–9243, 2023. 2, 6, 8

[30] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 484–500, 2018. 2

[31] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2, 6

[32] Masayuki Kanbara, Takashi Okuma, Haruo Takemura, and Naokazu Yokoya. A stereoscopic video see-through augmented reality system based on real-time vision-based registration. In *Proceedings IEEE Virtual Reality 2000 (Cat. No. 00CB37048)*, pages 255–262. IEEE, 2000. 1

[33] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017. 1

[34] Wan-Soo Kim, Dae-Hyun Lee, Yong-Joo Kim, Taehyeong Kim, Won-Suk Lee, and Chang-Hyun Choi. Stereo-vision-based crop height estimation for agricultural robots. *Computers and Electronics in Agriculture*, 181:105937, 2021. 1

[35] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6647–6655, 2017. 1

[36] 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. 1

[37] Shao-Yuan Lo, Wei Wang, Jim Thomas, Jingjing Zheng, Vishal M Patel, and Cheng-Hao Kuo. Learning feature decomposition for domain adaptive monocular depth estimation. *arXiv preprint arXiv:2208.00160*, 2022. 3, 8

[38] Robert McCraith, Lukas Neumann, and Andrea Vedaldi. Calibrating self-supervised monocular depth estimation. *arXiv preprint arXiv:2009.07714*, 2020. 3, 8

[39] Jaswanth Nidamanuri, Chinmayi Nibhanupudi, Rolf Assfalg, and Hrishikesh Venkataraman. A progressive review: Emerging technologies for adas driven solutions. *IEEE Transactions on Intelligent Vehicles*, 7(2):326–341, 2021. 1

[40] 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):1623–1637, 2020. 1

[41] Shuwei Shao, Zhongcai Pei, Weihai Chen, Baochang Zhang, Xingming Wu, Dianmin Sun, and David Doermann. Self-supervised learning for monocular depth estimation on minimally invasive surgery scenes. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7159–7165. IEEE, 2021. 1

[42] Kunal Swami, Amrit Muduli, Uttam Gurram, and Pankaj Bajpai. Do what you can, with what you have: Scale-aware and high quality monocular depth estimation without real world labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 988–997, 2022. 3, 4, 8

[43] 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. 5

[44] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2022–2030, 2018. 2

[45] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2162–2171, 2019. 1

[46] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1164–1174, 2021. 1

[47] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *Conference on Robot Learning*, pages 539–549. PMLR, 2023. 3

[48] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M López. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):537–547, 2020. 1

[49] Feng Xue, Guirong Zhuo, Ziyuan Huang, Wufei Fu, Zhuoyue Wu, and Marcelo H Ang. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2330–2337. IEEE, 2020. 3, 8

[50] Sen Zhang, Jing Zhang, and Dacheng Tao. Towards scale-aware, robust, and generalizable unsupervised monocular depth estimation by integrating imu motion dynamics. In *European Conference on Computer Vision*, pages 143–160. Springer, 2022. 3, 8

[51] Shanshan Zhao, Huan Fu, Mingming Gong, and Dacheng Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9788–9798, 2019. 2, 3

[52] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 3

[53] 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. 1, 5, 6