2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Multimodal Scale Consistency and Awareness for Monocular Self-Supervised Depth Estimation

Hemang Chawla*, Arnav Varma*, Elahe Arani, and Bahram Zonooz

Abstract— **Dense depth estimation is essential to scene-understanding for autonomous driving. However, recent self-supervised approaches on monocular videos suffer from scale-inconsistency across long sequences. Utilizing data from the ubiquitously copresent global positioning systems (GPS), we tackle this challenge by proposing a dynamically-weighted *GPS-to-Scale (g2s)* loss to complement the appearance-based losses. We emphasize that the GPS is needed only during the multimodal training, and not at inference. The relative distance between frames captured through the GPS provides a scale signal that is independent of the camera setup and scene distribution, resulting in richer learned feature representations. Through extensive evaluation on multiple datasets, we demonstrate scale-consistent and -aware depth estimation during inference, improving the performance even when training with low-frequency GPS data.**

I. Introduction

Robots and autonomous driving systems require scene-understanding for planning and navigation. Therefore, spatial perception through depth estimation is essential for enabling complex behaviors in unconstrained environments. Even though sensors such as LiDARs can perceive depth at metric-scale [1], their output is sparse and they are expensive to use. In contrast, *monocular* color cameras are compact, low-cost, and consume less energy. While traditional camera-based approaches rely upon hand-crafted features from multiple views [2], deep learning based approaches can predict depth from a single image. Among these, self-supervised methods that predict the ego-motion and depth simultaneously by view-synthesis of adjacent frames [3]–[5] are preferred over supervised methods that require accurate ground truth labels for training [6]–[8].

However, monocular vision inherently suffers from scale ambiguity. Additionally, the self-supervised approaches introduce scale-inconsistency in estimated depth across different video snippets [9]. Consequently, most of the existing methods scale the estimated relative depth using the LiDAR ground truth during evaluation. Recent methods tackling this problem utilize additional 3D geometric constraints to introduce scale-consistency [9], [10], but require at least some depth or stereo supervision to predict at metric-scale [11], [12]. Nevertheless, obtaining metric scale predictions at low cost is necessary for practical deployment.

Since self-supervised learning allows training on large and varied data including crowdsourced data [13], [14], the ubiquitous GPS copresent with videos can be employed for multimodal training. Taking cues from how cross-modal learning leads to richer learned feature representations [15], [16], we hypothesize that the relative distance between image frames captured from the GPS can provide a scale signal that complements commonly used appearance-based losses to predict scale-consistent and -aware improved estimates, without requiring any ground truth depth annotations.

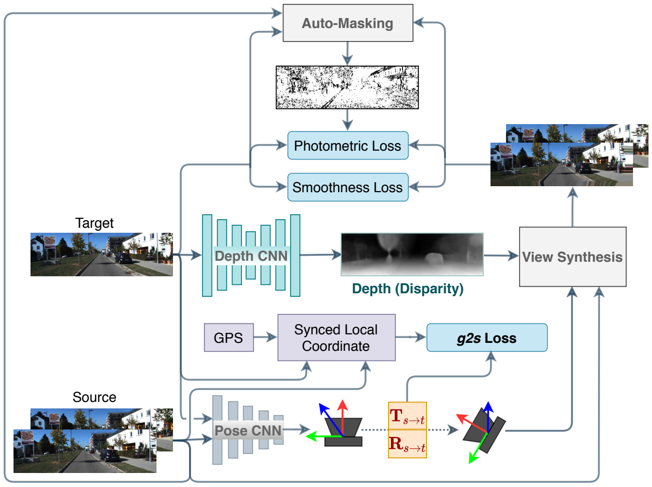

Fig. 1. A schematic of our proposed multimodal self-supervised depth and ego-motion prediction network for monocular videos. We introduce a *GPS-to-Scale (g2s)* loss that leads to scale-consistent and -aware estimates during inference.

In this work, we propose a *GPS-to-Scale (g2s)* loss that utilizes the ratio of magnitudes of the relative translation measured by the GPS and the relative translation predicted by the pose network to enforce scale-consistency and -awareness on the depth predictions, linked together via the perspective projection model [5]. Scale consistency implies that the standard deviation of the depth scale factors across the video is low. Scale awareness implies that the mean scale factor is close to 1. Note that this GPS information is only used during the training, while the inference is directly performed on the unlabeled monocular videos. Furthermore, we compare different weighting strategies for the proposed loss and demonstrate that exponentially increasing the weight on *g2s* over the epochs leads to the best performance. Experiments on the KITTI raw [17] Eigen [18] split as well as the improved KITTI depth benchmark [19] show that adding the *g2s* loss improves performance and scale-consistency over state-of-the-art-methods, even with low-frequency planar GPS (without altitude). Finally, with experiments on out-of-distribution Make3D [20] and Cityscapes [21] datasets, we

*Equal Contribution. All authors are with the Advanced Research Lab, Navinfo Europe, The Netherlands {hemang.chawla, arnav.varma, elahe.arani, b.yoosefizonooz}@navinfo.eu

978-1-7281-9077-8/21/$31.00 ©2021 IEEE

show that the introduced scale-consistency and -awareness is present across domains in comparison with other methods.

II. Related Work

Estimating scene depth is a long-standing problem in computer vision. Traditional approaches solve this by utilizing disparity across multiple views within a non-linear optimization framework [2], [22]. Supervised methods that produce high-quality estimates have also been proposed [6]–[8], but necessitate the availability of accurate ground truth and cross-calibration of sensors for training. Instead, using view-synthesis as a signal, self-supervised methods produce accurate depth maps from stereo image pairs [23], [24] or monocular video snippets [3]–[5]. We focus on methods employing purely monocular setups, as they are more pervasive and do not depend upon prior knowledge of relative rotation and translation of the stereo camera pairs. However, most existing monocular approaches utilize only appearance-based losses with the assumption of brightness consistency that limits training on small video subsequences without any long sequence constraints. Hence, the depth and ego-motion estimates from these methods suffer from scale-inconsistency along with the global scale-ambiguity present in monocular vision. Therefore, ground truth LiDAR depth maps [4] or camera height [25] are used during inference to recover per-image scale.

Methods addressing this problem add 3D-geometry-based losses to introduce scale-consistency [9], [10], yet utilize at least some depth or stereo supervision to introduce scale-awareness [11], [12]. Recently [26] introduced a similar instantaneous velocity based multi-modal supervision. However, access to instantaneous velocity may require the use of inertial measurement units (IMU) that are less ubiquitous. In contrast, GPS is often copresent, such as in dashboard cameras albeit with lower frequency, allowing training on more data. In this work, we introduce a *GPS-to-scale (g2s)* loss that produces improved scale-consistent and -aware results even with low-frequency planar GPS without altitude.

III. Method

Our objective is to simultaneously train depth and ego-motion prediction networks that produce scale-consistent and -aware estimates from only a monocular color camera during inference. Here we describe the baseline network and appearance-based losses for self-supervised learning, followed by the motivation and description of our proposed dynamically-weighted *GPS-to-Scale (g2s)* loss.

A. Overview

Given a set of n images from a video sequence, and m loosely corresponding GPS coordinates, the inputs to the networks are a sequence of temporally consecutive RGB image triplets $\{I_{\text{-}1}, I_0, I_1\} \in \mathbb{R}^{H\times W\times 3}$ and the the synced GPS coordinates $\{G_{\text{-}1}, G_0, G_1\} \in \mathbb{R}^3$, when available. The depth network, $f_D : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H\times W}$, outputs dense depth (or disparity) for each pixel coordinate p of a single image. Simultaneously, the ego-motion network, $f_E : \mathbb{R}^{2\times H\times W\times 3} \to \mathbb{R}^6$, outputs relative translation (t_x, t_y, t_z) and rotation (r_x, r_y, r_z) forming the affine transformation $\left[\begin{smallmatrix}\hat{R} & \hat{T}\\ 0 & 1\end{smallmatrix}\right] \in$ SE(3) between a pair of adjacent images. The predicted depth $\hat{D}$ and ego-motion $\hat{T}$ are linked together via the perspective projection model [5], that warps the source (s) images $I_s \in \{I_{\text{-}1}, I_1\}$ to the target (t) image $I_t \in \{I_0\}$, given the camera intrinsics K.

We establish a strong baseline by following the best practices of appearance-based learning from Monodepth2 [4]. The networks are trained using the appearance-based *photometric* loss between the real and synthesized target images, as well as a *smoothness* loss for depth regularization in low texture scenes [4]. Following [4], [27], we use auto-masking (M) to disregard the temporally stationary pixels in the image triplets. The total appearance-based loss is calculated by upscaling the predicted depths from intermediate decoder layers to the input resolution.

Additionally, we introduce the dynamically-weighted *g2s* loss that enforces scale-consistency and -awareness using the ratio of the measured and estimated translation magnitudes. Fig. 1 illustrates the complete architecture that uses the proposed method.

B. GPS-to-Scale (g2s) Loss

Appearance-based losses provide supervisory signals on short monocular subsequences. This leads to scale-inconsistency of the predictions across long videos. Approaches addressing this problem through 3D-geometry-based losses provide a signal that depends upon the camera setup and the scene distribution [9], [10]. Therefore, we introduce the *GPS-to-Scale (g2s)* loss that provides an independent cross-modal signal leading to scale-consistent and -aware estimates.

Synced Local Coordinates: The GPS information, ubiquitously copresent with videos, consists of the latitude, longitude, and optionally the altitude of the vehicle. First, we convert these geodetic coordinates to local coordinates $\boldsymbol{G} = \{x_g, y_g, z_g\}$ using the Mercator projection such that,

$$x_g = \cos\left(\frac{\pi \cdot \text{lat}_0}{180}\right) r_e \log\left(\tan\frac{\pi \cdot (90 + \text{lat})}{360}\right) \quad (1)$$

$$y_g = \text{alt} \quad (2)$$

$$z_g = \cos\left(\frac{\pi \cdot \text{lat}_0}{180}\right) r_e \frac{\pi \cdot \text{lon}}{180} \quad (3)$$

where $r_e = 6\,378\,137\,\text{m}$ is taken as the radius of earth. Since the GPS frequency may be different from the frame-rate of the captured video, we additionally sync these local coordinates with the images using their respective timestamps.

Utilizing the ratio of the relative distance measured by the GPS and the relative distance predicted by the network, we additionally impose our proposed *g2s* loss given by,

$$\mathcal{L}_{g2s} = \sum_{s,t}\left(\frac{\|G_{s\to t}\|_2}{\|\hat{T}_{s\to t}\|_2} - 1\right)^2 \quad (4)$$

where $s \in \{-1, 1\}$ and $t \in \{0\}$. Note that these relative distance computations are unaffected by the rigid-body extrinsic calibration between the GPS and the camera. Following [5] we remove static frames while training, thereby allowing the *g2s* loss to be differentiable for all plausible inputs.

GPS noise and bias: By forming this loss upon the translation magnitude instead of the individual components (t_x, t_y, t_z), we account for any noise or systemic bias that may be present in the GPS measurements [28]. This loss encourages the ego-motion estimates to be closer to the common metric scale across the image triplets, thereby introducing the scale-consistency and -awareness which is extended to the depth estimates that are tied to the ego-motion via the perspective projection model.

Note that CNNs tend to learn surface statistical regularities by exploiting superficial clues (or shortcuts) specific to the distribution being trained on [29], [30]. Since the GPS signal does not depend upon the specific scene distribution or camera setup, we hypothesize that adding our proposed *g2s* loss in a multimodal context can help to disentangle intended higher-level abstractions [16] from the shortcut features to improve the estimates and help in generalizing scale-consistency to out-of-distribution (o.o.d.) datasets.

C. Dynamic Weighting Strategy

The networks learn to synthesize more plausible views of the target images I_t by improving their depth and ego-motion predictions over the training epochs. Thus, heavily penalizing the networks for the incorrect scales during the early training can interfere with the learning of individual translations, rotations, and pixel-wise depths. Hence, we dynamically weigh the *g2s* loss in an exponential manner to provide a scale signal that is low in the beginning and increases as the training progresses. The weight w to the *g2s* loss $\mathcal{L}_{g2s}$ is given by,

$$w = \exp\left(\text{epoch} - \text{epoch}_{\text{max}}\right). \tag{5}$$

D. Final Training Loss

The final loss combining the appearance-based losses [4], [5] with Eqs. 4 and 5 is given by,

$$\mathcal{L} = \mathcal{L}_{\text{appearance}} + w \cdot \mathcal{L}_{g2s}, \tag{6}$$

which is averaged over each batch.

IV. Experiments

For all our experiments, we follow the setup of Monodepth2 [4].

A. Depth Estimation

Following the established protocols, we compare our depth predictions on the Eigen Split [18] of KITTI [17] raw dataset as shown in Tables I and II. This contains $39,810$ training and 697 test images respectively. The depth is evaluated using metrics from [18] up to the fixed range of 80 m, unless specified otherwise. We also evaluate against the *Improved* ground truth depth [19] which contains 652 (93%) of the 697 *Original* test images. Best results for each metric are in bold. The second best results are underlined. * denotes results when trained on Cityscapes along with KITTI.

1) Performance and Scale-Consistency: For evaluating the performance and scale-consistency of depth estimation, we follow the standard procedure of scaling the per-image estimated depths $\hat{D}$ with individual scale factors given by the ratio of the median ground truth depths from LiDAR and the median predicted depths [5]. A lower standard deviation of the scale factors corresponds to a higher scale-consistency.

As shown in Table I, we outperform existing depth estimation methods on the KITTI *Original* as well as *Improved* ground truths for the Eigen split. This improvement can be attributed to the richer learned feature representations as explained in Sec III-B. Furthermore, Fig. 2 validates our results visually, and demonstrates that the learning of richer feature representation with our proposed multi-modal training leads to sharper depth estimates with improved structure preservation. As discussed earlier in Sec III-B, this can be explained by the disentangling of the intended higher-level abstractions from the shortcut features [16], [29], [30].

We also compare the variation of the scale factor for different methods as shown in Fig. 3. Note that the standard deviation of depth scale factors is the lowest for our method at 0.07. Unlike previous methods that measure scale-consistency by the standard deviation of the scales normalized by the median scale [4], [9], we report un-normalized standard deviation. This shows that the network is able to estimate scale-consistent depths with the use of our proposed *g2s* loss during training.

2) Scale-Awareness: We also compare the *unscaled* depth estimates in Table II (LR and HR denote methods trained on low and high resolution images respectively. pp [4] denotes post-processing during inference). As shown, most state-of-the-art monocular self-supervised methods produce poor estimates without the per-image scaling based on the LiDAR ground truth depths. However, our *unscaled* estimates are close to that from Table I. As shown in Fig. 3, our mean depth scale factor is ≈ 1 (specifically 1.03), establishing the scale-awareness introduced by our method for monocular depth estimation. We outperform Roussel et al. [12] which uses stereo pre-training on CityScapes to predict scale-aware monocular depth for KITTI. We also show comparable performance against Packnet-SfM [26] which uses a much heavier depth-estimation-dedicated architecture unlike the ResNet family based methods such as ours. Thus, while our method has an inference time of 40 ms on an NVIDIA 1080Ti GPU, [26] is slower with an inference time of 60 ms even on a Titan V100 GPU.

Hence, through experiments in Sec IV-A.1 and IV-A.2, we demonstrate that the *G2S* loss provides a scale-signal based on relative distance between image frames resulting in scale-consistent and -aware estimates.

B. KITTI Depth Prediction Benchmark

We also measure the performance of our method on the KITTI Depth Prediction Benchmark using the metrics from [19]. We train our method with a ResNet50 encoder on an image size of 1024×320 for 30 epochs, and evaluate it

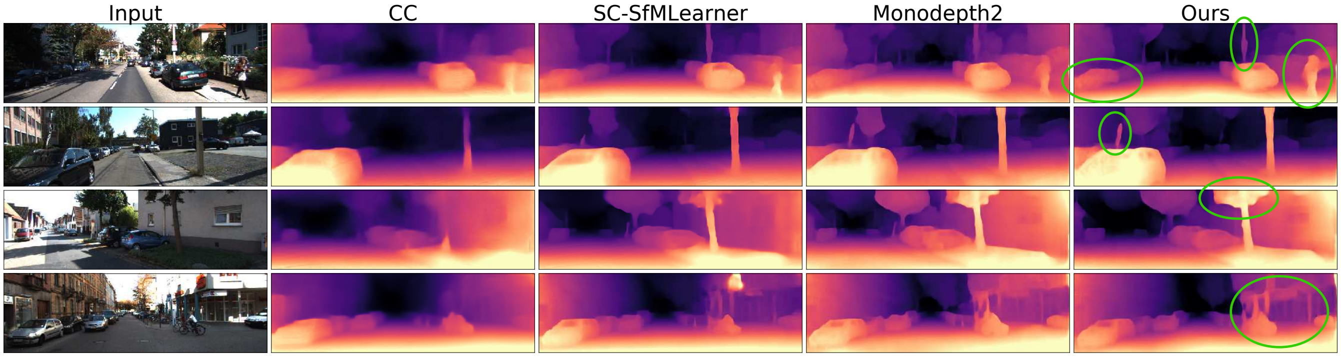

Fig. 2. Single-image depth estimates on the KITTI Eigen split. Our method produces sharper, high-quality predictions that preserve more structure when compared against existing methods.

TABLE I

Per-image scaled DENSE DEPTH PREDICTION (WITHOUT POST-PROCESSING) ON KITTI *Original* [17] AND *Improved* [19].

GT	Methods	Resolution	Error↓				Accuracy↑		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Original	SfMLearner [5]	416×128	0.208	1.768	6.856	0.283	0.678	0.885	0.957
	GeoNet [31]	416×128	0.155	1.296	5.857	0.233	0.793	0.931	0.973
	Vid2Depth [10]	416×128	0.163	1.240	6.220	0.250	0.762	0.916	0.968
	Struct2Depth [3]	416×128	0.141	1.026	5.291	0.215	0.816	0.945	0.979
	VITW [14]	416×128	0.128	0.959	5.230	0.212	0.845	0.947	0.976
	Roussel et al. [12]	416×128	0.179	1.545	6.765	0.268	0754	0916	0.966
	CC [32]	832×256	0.140	1.070	5.326	0.217	0.826	0.941	0.975
	SC-SfMLearner [9]	832×256	0.137	1.089	5.439	0.217	0.830	0.942	0.975
	Monodepth2 [4]	640×192	0.115	0.903	4.863	0.193	**0.877**	**0.959**	**0.981**
	SG Depth [33]	640×192	0.117	0.907	**4.844**	0.194	0.875	0.958	0.980
	Ours	640×192	**0.112**	**0.894**	4.852	**0.192**	**0.877**	0.958	**0.981**
Improved	SfMLearner* [5]	416×128	0.176	1.532	6.129	0.244	0.758	0.921	0.971
	Geonet* [31]	416×128	0.132	0.994	5.240	0.193	0.883	0.953	0.985
	Vid2Depth* [10]	416×128	0.134	0.983	5.501	0.203	0.827	0.944	0.981
	Monodepth2 [14]	640×192	0.090	**0.545**	**3.942**	**0.137**	**0.914**	**0.983**	**0.995**
	Ours	640×192	**0.088**	0.554	3.968	**0.137**	0.913	0.981	**0.995**

TABLE II

Unscaled DENSE DEPTH PREDICTION ON KITTI *Original* [17].

	Methods	Resolution	Error↓				Accuracy↑		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
LR	SfMLearner [5]	416×128	0.977	15.161	19.189	3.832	0.0	0.0	0.0
	Roussel et al. [12]	416×128	0.175	1.585	6.901	0.281	0.751	0.905	0.959
	CC [32]	832×256	0.961	14.672	18.838	3.280	0.0	0.0	0.0
	SC-SfMLearner [9]	832×256	0.961	14.915	19.089	3.264	0.0	0.0	0.0
	Monodepth2 [4]	640×192	0.969	15.126	19.199	3.489	0.0	0.0	0.0
	Packnet-SfM [26]	640×192	0.111	**0.829**	**4.788**	0.199	0.864	**0.954**	**0.980**
	Ours	640×192	0.111	0.900	4.935	0.200	0.863	0.953	0.979
	Ours (pp)	640×192	**0.109**	0.860	4.855	**0.198**	**0.865**	**0.954**	**0.980**
HR	Packnet-SfM [26]	1280×384	**0.107**	**0.803**	**4.566**	0.197	**0.876**	0.957	0.979
	Ours	1024×384	0.109	0.844	4.774	0.194	0.869	0.958	0.981
	Ours (pp)	1024×384	0.109	0.809	4.705	**0.193**	0.869	**0.959**	**0.982**

using the online KITTI benchmark server.[1]

Results, ordered based on their rank, are shown in Table III (D, M, and S represent supervised training with ground truth depths, monocular sequences, and stereo pairs, respectively. Seg represents additional supervised semantic segmentation training. G represents the use of GPS for multi-modal self-supervision). We outperform all self-supervised methods while also performing better than many supervised methods which use ground truth depth maps during training.

C. Ablation Studies

To study the efficacy of the proposed *g2s* loss in detail, we perform ablation studies on the introduced weighting strategy, as well as the frequency and dimensionality of the GPS used in the multi-modal training.

1) Weighting Strategy: In Table V, we compare our proposed weighting strategy (Eq. 5) against the alternative constant and linearly increasing weights for the *g2s* loss. The mean and standard deviation of scale factors as well as

[1] http://www.cvlibs.net/datasets/kitti/eval_depth.php?benchmark=depth_prediction. See results under g2s

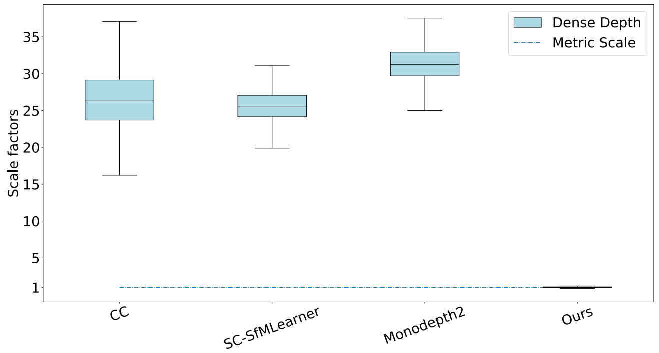

Fig. 3. Box-plot visualizing the mean and standard deviation of scale factors for per-image dense depth estimation on the test set of Eigen split [18]. Existing methods scaled the estimated depth using the per-image ground truth during inference. Our method is scale-consistent and -aware and does not need ground truth during inference.

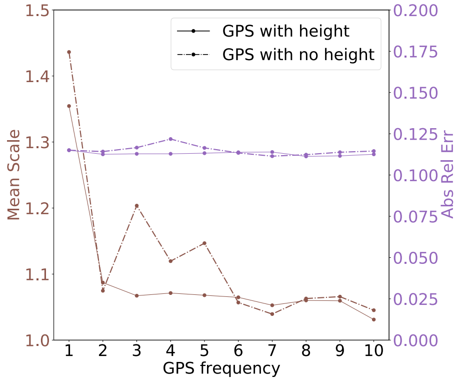

Fig. 4. Ablation study on different GPS frequencies and dimensionality. Mean scale factor and performance of depth estimation indicated by the Abs Rel Error [18] is shown.

the corresponding metrics on *scaled* predictions are shown. We confirm that utilizing an exponential weighting strategy effectively leverages the scale signal to produce scale-consistent and -aware depth estimates. As explained earlier in Section III-C, this is because penalizing the networks for the incorrect scales during the early training can interfere with the learning. Therefore, providing an increasing scale signal over the epochs, while allowing effective appearance-based learning in the early training, leads to the best results.

2) GPS Frequency and Dimensionality: While GPS is ubiquitously copresent with driving video sequences, crowd-sourced data often consists of high frames-per-second (fps) videos but lower frequency GPS. Furthermore, while altitude can be trilaterated by the GPS receivers, it is often not measured by the low-cost setups. Therefore, we study the efficacy of the *g2s* loss over different GPS frequencies, and the impact of the lack of altitude/height measurements in two-dimensional GPS.

Note that the images in the KITTI dataset are captured at 10 fps. To simulate the GPS frequencies lower than $10\,\mathrm{Hz}$ we randomly select the GPS data for $f < 10$ frames for each 10-frame non-overlapping subsequences ($\approx 1\,\mathrm{s}$) in the training data. Thereafter, we apply our *g2s* loss as described in Eq. 4 on the adjacent image pairs that have corresponding GPS available. The results are visualized in Fig. 4.

TABLE III
QUANTITATIVE COMPARISON ON THE KITTI DEPTH PREDICTION BENCHMARK (ONLINE SERVER).

Method	Train	SILog	SqErrRel	AbsErrRel	iRMSE
DORN [6]	D	11.77	2.23	8.78	12.98
SORD [34]	D	12.39	2.49	10.10	13.48
VNL [35]	D	12.65	2.46	10.15	13.02
DS-SIDENet [36]	D	12.86	2.87	10.03	14.40
PAP [37]	D	13.08	2.72	10.27	13.95
Guo et al. [38]	D+S	13.41	2.86	10.60	15.06
Ours	M+G	14.16	3.65	11.40	15.53
Monodepth2 [4]	M+S	14.41	3.67	11.22	14.73
DABC [39]	D	14.49	4.08	12.72	15.53
SDNet [40]	D	14.68	3.90	12.31	15.96
APMoE [41]	D	14.74	3.88	11.74	15.63
CSWS [27]	D	14.85	348	11.84	16.38
HBC [42]	D	15.18	3.79	12.33	17.86
SGDepth [33]	M+Seg	15.30	5.00	13.29	15.80
DHGRL [43]	D	15.47	4.04	12.52	15.72
MultiDepth [44]	D	16.05	3.89	13.82	18.21
LSIM [45]	S	17.92	6.88	14.04	17.62
Monodepth [24]	S	22.02	20.58	17.79	21.84

We observe that our method is able to learn scale-aware depth estimation by using even the low-frequency GPS, thereby indicating the strength of the proposed *g2s* loss. Moreover, our method improves upon the baseline Monodepth2 [4] even with a low-frequency scale-signal. We also observe that our method performs equally well without the availability of the altitude information. Thus, we conclude that our method would be applicable in the case of datasets with 2-dimensional or sparse GPS.

D. Out-of-Distribution Performance

We also study the generalization capability of our method on o.o.d. [30] datasets - Make3D (M3D) [20] and Cityscapes (CS) [21]. We evaluate our method (trained on the KITTI Eigen split) on the 2 : 1 center crop of o.o.d. test images. Table IV shows the mean and standard deviation of the scale factors for the estimated depths, capped at $70\,\mathrm{m}$. The standard deviation on the depth scale factor is the lowest for our method, indicating scale-consistency. This has also been visualized for the Make3D and Cityscapes test sets in Figs. 5 and 6. Also note that the mean of depth scale factors is significantly closer to 1 than for other methods, even though metric-scale is no longer maintained. Finally, the qualitative results on the Make3D and Cityscapes dataset as shown in Figs. 7 and 8, demonstrate that the proposed multi-modal

TABLE IV
SCALE FACTORS ON OUT-OF-DISTRIBUTION DATASETS.

	Method	μ_{scale}	$\sigma_{\text{scale}} \downarrow$
M3D	SC-SfMLearner	40.62	17.24
	Monodepth2	76.02	24.40
	Ours	2.81	0.85
CS	SC-SfMLearner	60.99	22.44
	Monodepth2	118.61	36.49
	Ours	4.01	1.22

TABLE V

ABLATION STUDY OF DIFFERENT WEIGHTING STRATEGIES ON THE KITTI EIGEN SPLIT [18].

Weights	μ_{scale}	$\sigma_{\text{scale}} \downarrow$	**Error↓**				**Accuracy↑**		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Const. 1	0.776	0.126	1.280	53.454	21.915	0.934	0.217	0.427	0.604
Const. 10^{-3}	1.159	0.120	0.125	1.032	5.214	0.203	0.860	0.955	0.979
Linear	0.124	**0.020**	0.443	4.757	12.083	0.588	0.303	0.561	0.766
Ours (Eq. 5)	**1.031**	$\underline{0.073}$	**0.112**	**0.894**	**4.852**	**0.192**	**0.877**	**0.958**	**0.981**

training improves the delineation of different objects in the depth estimation even for new scenes. These results can be explained by the learning of richer transferable discriminative features due to the scene and camera-setup independence of the GPS scale signal as explained in Sec III-B.

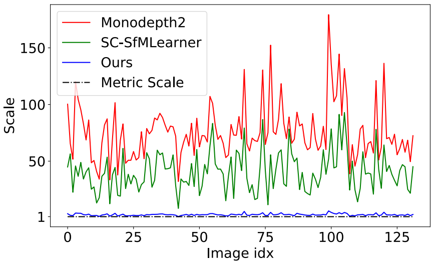

Fig. 5. Out-of-Distribution depth scale variation on the Make3D test set [20].

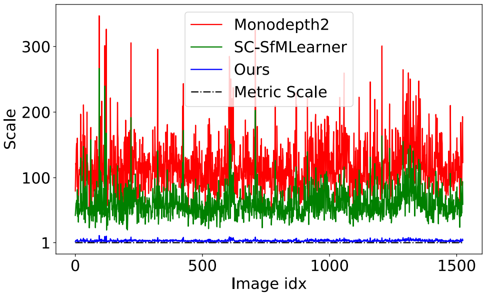

Fig. 6. Out-of-Distribution depth scale variation on the Cityscapes test set [21].

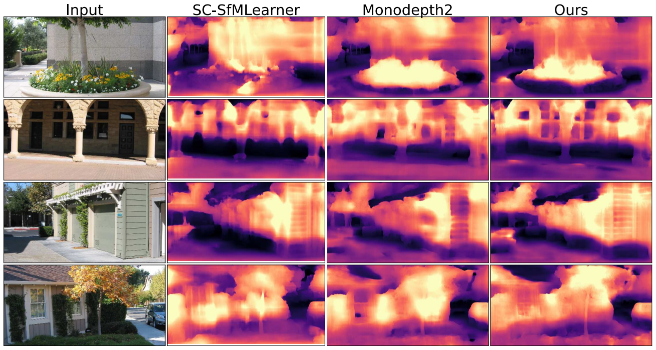

Fig. 7. Qualitative results on Make3D test set [20]. All methods were trained on the monocular sequences from the KITTI Eigen split [18]. Note that finer details are present in our predictions, such as building structures, silhouettes of flowers, and tree trunks.

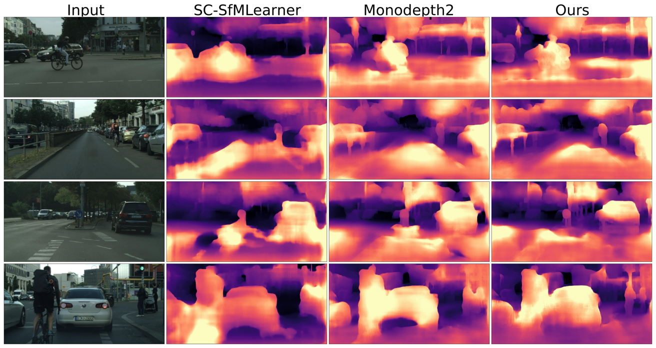

Fig. 8. Qualitative results on Cityscapes test set [21]. All methods were trained on the monocular sequences from the KITTI Eigen split. Note that finer details are present in our predictions, such as vehicle details, silhouettes of humans, and traffic signs.

V. CONCLUSION

This work addresses the problem of estimating scale-consistent and -aware monocular dense depths in a self-supervised setting, a feature essential for many practical autonomous vehicle applications. Previously, only appearance-based losses were used, and hence it was necessary to scale the predictions using the LiDAR ground truth. In contrast, by utilizing the camera-setup- and scene-independent GPS information, we propose an exponentially-weighted *GPS-to-Scale (g2s)* loss to predict metrically accurate single-image depths within a multimodal self-supervised learning framework, without requiring any ground truth depth annotations. Also, no GPS information is used during the inference. Validating our approach on the KITTI dataset, we improve upon existing methods to predict sharper depths with finer-delineation of objects at scale. Through ablation studies, we also demonstrate the efficacy of our proposed loss, even when training on low-frequency or sparse GPS without height information. Finally, we show that our method results in better scale-consistency and -awareness even on out-of-distribution datasets. We posit that these improved results are a consequence of learning richer representations within a multimodal self-supervised framework. In the future, it seems promising to also study the impact of such at framework on the adversarial robustness of monocular depth estimation.

REFERENCES

[1] I. A. Barsan, S. Wang, A. Pokrovsky, and R. Urtasun, "Learning to localize using a lidar intensity map." in *Conference on Robot Learning (CoRL)*, 2018, pp. 605–616. 1

[2] K. Karsch, C. Liu, and S. B. Kang, "Depth extraction from video using non-parametric sampling," in *European Conference on Computer Vision*. Springer, 2012, pp. 775–788. 1, 2

[3] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8001–8008. 1, 2, 4

[4] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 3828–3838. 1, 2, 3, 4, 5

[5] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," 2017. 1, 2, 3, 4

[6] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011. 1, 2, 5

[7] C. Tang and P. Tan, "Ba-net: Dense bundle adjustment network," *arXiv preprint arXiv:1806.04807*, 2018. 1, 2

[8] H. Zhou, B. Ummenhofer, and T. Brox, "Deeptam: Deep tracking and mapping," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 822–838. 1, 2

[9] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," in *Advances in Neural Information Processing Systems*, 2019, pp. 35–45. 1, 2, 3, 4

[10] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5667–5675. 1, 2, 4

[11] V. Guizilini, J. Li, R. Ambrus, S. Pillai, and A. Gaidon, "Robust semi-supervised monocular depth estimation with reprojected distances," in *Conference on Robot Learning*. PMLR, 2020, pp. 503–512. 1, 2

[12] T. Roussel, L. Van Eycken, and T. Tuytelaars, "Monocular depth estimation in new environments with absolute scale," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1735–1741. 1, 2, 3, 4

[13] H. Chawla, M. Jukola, T. Brouns, E. Arani, and B. Zonooz, "Crowdsourced 3d mapping: A combined multi-view geometry and self-supervised learning approach," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS)*. IEEE, 2020. 1

[14] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova, "Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras," 2019. 1, 4

[15] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011. 1

[16] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017. 1, 3

[17] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013. 1, 3, 4

[18] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374. 1, 3, 5, 6

[19] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *International Conference on 3D Vision (3DV)*, 2017. 1, 3, 4

[20] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 824–840, 2008. 1, 5, 6

[21] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223. 1, 5, 6

[22] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015. 2

[23] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*. Springer, 2016, pp. 740–756. 2

[24] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279. 2, 5

[25] F. Xue, G. Zhuo, Z. Huang, W. Fu, Z. Wu, and M. H. Ang Jr, "Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS)*. IEEE (in press), 2020. 2

[26] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2485–2494. 2, 3, 4

[27] B. Li, Y. Dai, and M. He, "Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference," *Pattern Recognition*, vol. 83, pp. 328–339, 2018. 2, 5

[28] A. Das and G. Dubbelman, "An experimental study on relative and absolute pose graph fusion for vehicle localization," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 630–635. 3

[29] J. Jo and Y. Bengio, "Measuring the tendency of cnns to learn surface statistical regularities," *arXiv preprint arXiv:1711.11561*, 2017. 3

[30] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *arXiv preprint arXiv:2004.07780*, 2020. 3, 5

[31] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1983–1992. 4

[32] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black, "Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 12 240–12 249. 4

[33] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, "Self-Supervised Monocular Depth Estimation: Solving the Dynamic Object Problem by Semantic Guidance," in *ECCV*, 2020. 4, 5

[34] R. Diaz and A. Marathe, "Soft labels for ordinal regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4738–4747. 5

[35] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5684–5693. 5

[36] H. Ren, M. El-Khamy, and J. Lee, "Deep robust single image depth estimation neural network using scene understanding." in *CVPR Workshops*, 2019, pp. 37–45. 5

[37] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, and J. Yang, "Pattern-affinitive propagation across depth, surface normal and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4106–4115. 5

[38] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang, "Learning monocular depth by distilling cross-domain stereo networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 484–500. 5

[39] R. Li, K. Xian, C. Shen, Z. Cao, H. Lu, and L. Hang, "Deep attention-based classification network for robust depth prediction," in *Asian Conference on Computer Vision (ACCV)*, 2018. 5

[40] M. Ochs, A. Kretz, and R. Mester, "Sdnet: Semantically guided depth estimation network," in *German Conference on Pattern Recognition*. Springer, 2019, pp. 288–302. 5

[41] S. Kong and C. Fowlkes, "Pixel-wise attentional gating for scene parsing," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1024–1033. 5

[42] H. Jiang and R. Huang, "Hierarchical binary classification for monocular depth estimation," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2019, pp. 1975–1980. 5

[43] Z. Zhang, C. Xu, J. Yang, Y. Tai, and L. Chen, "Deep hierarchical guidance and regularization learning for end-to-end depth estimation," *Pattern Recognition*, vol. 83, pp. 430–442, 2018. 5

[44] L. Liebel and M. Körner, "Multidepth: Single-image depth estimation via multi-task regression and classification," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1440–1447. 5

[45] M. Goldman, T. Hassner, and S. Avidan, "Learn stereo, infer mono: Siamese networks for self-supervised, monocular, depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0. 5