Adaptive Self-supervised Depth Estimation in Monocular Videos

Julio Mendoza and Helio Pedrini(✉)

Institute of Computing, University of Campinas, Campinas, SP 13083-852, Brazil
helio@ic.unicamp.br

Abstract. In this work, we develop and evaluate two adaptive strategies to self-supervised depth estimation methods based on view reconstruction. First, we propose an adaptive consistency loss that extends the usage of minimum re-projection to enforce consistency on the pixel intensities, structure, and feature maps. Moreover, we evaluate two approaches to use uncertainty to weigh the error contribution in the input frames. Finally, we improve our model with a composite visibility mask. The results show that the adaptive consistency loss can effectively combine photometric, structure and feature consistency terms. Moreover, weighting the error contribution using uncertainty can improve the performance of a simpler version of the model, but cannot improve them model when all improvements are considered. Finally, our combined model achieves competitive results when compared to state-of-the-art methods.

Keywords: Depth estimation · View synthesis · Monocular videos

1 Introduction

Dense depth maps are useful representations of a scene that have been used in several computer vision applications, such as 3D reconstruction, virtual and augmented reality, robot navigation, scene interaction and autonomous driving. Depth maps can be obtained with sensors. However, in some scenarios, it is unfeasible to rely solely on them. Such demand has increased the interest in the development of effective methods, and recently it has motivated the development of approaches based on deep learning.

The existing data sets for depth estimation have enabled training deep models in a supervised approach. However, the size, quality and availability of labeled data sets are becoming a barrier for supervised approaches. Researchers use complex and costly procedures to collect ground truth and the available data sets are smaller than the ones used in other computer vision tasks.

In recent years, various self-supervised approaches have been proposed to learn depth maps from monocular videos. These methods rely on appearance and geometric consistency among nearby frames on videos, to reconstruct a reference frame with the intensities of another frame and to use the reconstruction error as a supervisory signal. Thus, these methods can learn dense depth maps without

Y. Peng et al. (Eds.): ICIG 2021, LNCS 12890, pp. 687–699, 2021.
https://doi.org/10.1007/978-3-030-87361-5_56

labeled data sets and can take advantage of the vast amount and rich variability of video data available.

One of the main challenges of self-supervised approaches based on reconstruction is that some pixels in frame cannot be explained from other frames because of occlusion, specular reflection, textureless regions among other reasons. Several approaches deal with these challenges excluding or attenuating the influence of pixels based on priors or adaptive approaches that leverage the availability of multiple frames neighboring a target frame to explain their pixels.

In this work, we develop and evaluate two adaptive strategies to improve the robustness of self-supervised depth estimation approaches with pixels that violate the assumptions of view reconstruction. Initially, we develop an adaptive consistency loss that extends the usage of minimum re-projection to enforce consistency on 3D structure and feature maps, in addition to the photometric consistency. Moreover, we evaluate the usage of uncertainty as loss attenuation mechanism, where the uncertainty is learned by modeling predictions as Laplacian, smooth-L1 or Cauchy probability distributions. Finally, we improve our model with a composite visibility mask.

2 Related Work

View Reconstruction Based Depth Estimation. Deep learning approaches based on view reconstruction leverage the correspondence between the pixels of two views of the same scene. This correspondence could be computed with the relative pose between cameras that captures both views, and a depth map of a single view. This principle was used by Garg et al. [5], where a stereo pair provides the views, the parameters of the device give the relative pose, and a depth network predicts the depth map. Similarly, Zhou et al. [28] proposed a method where the relative pose and the depth map are estimated with deep networks.

These approaches have been improved, for example, to deal with occluded regions that cannot be reconstructed using geometric priors [15,17], to deal with moving objects, which violate the static assumption of view reconstruction, using optical flow [9,26] or segmenting and estimating the motion of moving objects [2,14,23], and to improve the learning signal by enforcing consistency between several representations of the scene [3,17,26]. Our approach improves the learning signal enforcing consistency between 3D coordinates, feature maps, and color information of the views.

Consistency Constraints. The availability of a correspondence between the pixels on the source and target views allows supervision by enforcing consistency on representations, in addition to the pixel intensities. For example, we can enforce consistency between forward and backward optical flows [16,26], predicted and projected depth maps [7,16], 3D coordinates [17], and feature maps [20,27]. However, we cannot enforce consistency in the entire image because some regions do not have valid correspondences, for example, occluded regions produced by the motion of the camera or objects, or regions with specular reflection where the color is inconsistent with the structure of the scene, and also due to multiple correspondences for single pixels at homogeneous regions do not provide supervision. Approaches that exclude or attenuate the error contribution of these regions have been proposed in the literature. For example, learning an explainability mask [28], excluding pixels that are projected out of the field of view [17], excluding pixels with high inconsistencies on optical flows or depth maps [26], excluding stationary pixels [8], excluding occluded pixels using to geometric cues [9], attenuating the error using similar criteria [18]. Another approach leverage the availability of correspondences from multiple source frames [8] or estimated from different models [3], considering only the correspondences with minimum photometric error. Our approach extends the minimum re-projection error on other consistency constraints in addition to photometric consistency.

Adaptive Losses Based on Uncertainty. The importance of quantifying the uncertainty on predictions has motivated research endeavors in several problems on computer vision, such as robust regression [1], representation learning [22], object detection [11], image de-raining [25], optical flow [12] and depth estimation [13,19,21,24]. Researchers have explored approaches that leverage uncertainty information for depth estimation, for instance, a method that leverages existing uncertainty estimation techniques [21] and an approach that predicts the uncertainty using a neural network [13,19,24]. A recent work explored approaches to estimate epistemic uncertainty and aleatoric uncertainty on an unsupervised monocular setting [19]. In this work, we explore several probability functions to predict aleatoric uncertainty to improve depth estimation.

3 Method

Figure 1 illustrates the main components of our method. In this section, we provide an overview of our baseline system. Moreover, we introduce two adaptive strategies to improve the robustness of our approach. Finally, we explore additional constraints.

3.1 Preliminaries

Approaches that use view reconstruction as main supervisory signal require to find correspondences between pixel coordinates on frames that represent views of the same scene. These correspondences can be computed using multi-view geometry. Given a pixel coordinate x_t in a target frame $\mathbf{I_t}$, we can obtain its coordinate x_s in a source frame $\mathbf{I_s}$ by back-projecting x_t to the camera coordinate system of the $\mathbf{I_t}$ using its depth value $\mathbf{D_t}(x_t)$, and the inverse of its intrinsic matrix $\mathbf{K}^{-1}$. Then, the relative motion transformation $\mathbf{T_{t \rightarrow s}}$ is applied to project the coordinates form the coordinate system of the $\mathbf{I_t}$ to the coordinate system of $\mathbf{I_s}$. Finally, the coordinates are projected onto the image plane in the source

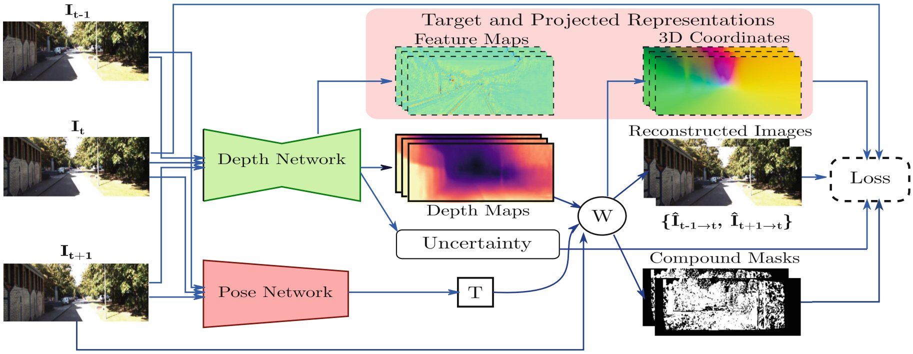

Fig. 1. Overview of our method. The depth network is used to predict the depth maps for the target $\mathbf{I_t}$ and source images $\mathbf{I_s} \in \{\mathbf{I_{t-1}}, \mathbf{I_{t+1}}\}$. The pose network predicts the Euclidean transformation between the target and source camera coordinate systems $\mathbf{T_{t\to s}}$.

frame. We express this correspondence in Eq. 1. We refer the reader to [28] for a detailed explanation.

$$x_s \sim \mathbf{K}\mathbf{T_{t\to s}}\mathbf{D_t}(x_t)\mathbf{K}^{-1}x_t \tag{1}$$

Once we know the projected coordinates and, therefore, the pixel intensities in the source image plane for each pixel x_t in the target image, we reconstruct the target frame $\hat{\mathbf{I}}_{\mathbf{s\to t}}(x_t) = \mathbf{I_s}(x_s)$. This process is known as image warping. This approach requires the dense depth map $\mathbf{D_t}$ of the target image, which we aim to reconstruct, the Euclidean transformation $\mathbf{T_{t\to s}}$, and camera intrinsics $\mathbf{K}$. Our model predicts the depth maps and the Euclidean transformation using convolutional neural networks and assumes that the camera intrinsics are given. The networks are trained using the reconstruction error as supervisory signal.

3.2 Adaptive Consistency Loss

Consistency could be enforced on representations of the scene such as 3D structure and feature maps. We propose an *adaptive consistency loss* that, in addition to the photometric consistency, also considers 3D structure and feature consistency constraints. This idea leverages the robustness of the min-reprojection error to pixels with high reconstruction error that could potentially be outliers. The adaptive consistency loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{ac} = \sum_{x_t \in \mathbf{I_t}} \min_{\mathbf{I_s}} \Big(M_o(x_t) \Big(\rho_{pc}\big(\mathbf{I_t}(x_t), \hat{\mathbf{I}}_{\mathbf{s\to t}}(x_s)\big) \\ + \lambda_{sc}\rho_{sc}\big(\mathbf{C_{s\to t}}(x_t), \hat{\mathbf{C}}_{\mathbf{s\to t}}(x_s)\big) + \lambda_{fc}\rho_{fc}\big(\mathbf{\Phi_t}(x_t), \hat{\mathbf{\Phi}}_{\mathbf{s\to t}}(x_s)\big)\Big)\Big)\end{aligned} \tag{2}$$

where ρ_{pc} measures the photometric consistency between pixels on the original $\mathbf{I_t}$ and reconstructed images $\hat{\mathbf{I}}_{\mathbf{s\to t}}$, ρ_{sc} measures the structure consistency between the 3D-coordinates of the target image projected to the camera coordinate system of the source image $\mathbf{C_{s\to t}}$, and the 3D-coordinates of the source image in its own camera coordinate system $\hat{\mathbf{C}}_{\mathbf{s\to t}}$, and ρ_{fc} measure the feature dissimilarity between the feature vectors for all pixels, and obtained from the target $\mathbf{\Phi_t}$, and the warped source feature maps $\hat{\mathbf{\Phi}}_{\mathbf{s\to t}}$. The feature maps are extracted from the decoder part of the depth network. M_o is a visibility mask that excludes pixels that lie out the field-of-view on the source frame [17].

Our photometric error function ρ_{pc} is a combination of an $L1$ distance and the structure similarity index metric (SSIM), with a trade-off parameter α. This function is shown in Eq. 3.

$$\rho_{pc}(p,q) = ||p-q||_1 + \alpha\frac{1-\mathrm{SSIM}(p,q)}{2} \tag{3}$$

Our structure error function ρ_{sc} is the average of a normalized absolute difference of the coordinates as follows:

$$\rho_{sc}(x,y) = \frac{1}{3}\sum_{i=1}^{3}\frac{|x_i-y_i|}{|x_i|+|y_i|} \tag{4}$$

Our feature dissimilarity function ρ_{sc} measures the squared L_2 distance of the L_2 normalized feature vectors $\hat{f}_s = f_s/||f_s||_2$ and $\hat{f}_t = f_t/||f_t||_2$, with $f_s = \hat{\mathbf{\Phi}}_{\mathbf{s\to t}}(x_s)$ and $f_t = \mathbf{\Phi_t}(x_t)$.

$$\rho_{fc}(f_s,f_t) = ||\hat{f}_s - \hat{f}_t||_2^2 \tag{5}$$

The total loss is the sum of the adaptive consistency loss and depth smoothness loss term [7] for the defined output scales $\mathcal{S}$.

$$\mathcal{L}_{total} = \sum_{i\in\mathcal{S}}\mathcal{L}_{ac}^{(i)} + \mathcal{L}_{ds}^{(i)} \tag{6}$$

3.3 Error Weighting Using Uncertainty

The adaptive consistency loss can handle cases in which at least one the source images can provide the information to reconstruct each pixel. However, several cases might break this condition, for instance, homogeneous regions and regions with specular reflection. Therefore, we aim to find other mechanisms to handle pixels with large error on these cases.

An approach is to allow the model to learn the uncertainty about the depth estimates, and leverage this information to attenuate the effect of pixels with large errors on the overall error. We can do that by placing a probability distribution function over the outputs of the model. The predicted depth values $\mathbf{D_t}(x_t)$ are modeled as corrupted with additive random noise sampled from a PDF with a scale parameter σ_{x_t} that is predicted by depth network. σ_{x_t} quantifies the uncertainty of the model on the predictions. The model is trained to minimize the negative log-likelihood.

First, we assume that noise comes from a Laplacian distribution, then the error function is the negative log-likelihood of this distribution. Equation 7 shows the error function.

$$\rho_{Laplacian}(p_t, p_s) = \frac{|\rho_{pc}(p_t, p_s)|}{\sigma_{x_t}} + \log(2\sigma_{x_t}) \tag{7}$$

where $p_t = \mathbf{I_t}(\mathbf{x_t})$, $p_s = \hat{\mathbf{I}}_{\mathbf{s}\to\mathbf{t}}(x_s)$, ρ_{pc} is the photometric error function, and σ_{x_t} is the predicted uncertainty for the pixel x_t.

We can observe that the first term in Eq. 7 attenuates the error when the uncertainty is high. Then, the second term discourages the model to predict high uncertainty values for all pixels. Thus, in order to minimize the function, the model is encouraged to predict high uncertainty values for pixels with large errors, attenuating the influence of large error in the overall error.

In order to explore the space of probability functions, we also evaluate our approach on the smooth-L1 functions and the Cauchy functions [1]. We define the probability distribution associated to the smooth-L1 function using the family of probability distributions defined in [1]. Equation 8 shows the negative log-likelihood associated to the smooth-L1 function.

$$\rho_{smooth\text{-}L1}(p_t, p_s) = \sqrt{\left(\frac{\rho_{pc}(p_t, p_s)}{\sigma_{x_t}}\right)^2 + 1} - 1 + \log(Z(1)) \tag{8}$$

where $Z(1)$ is a normalization factor for smooth-L1 function. We refer the reader to [1] for a detailed explanation. Finally, Eq. 9 shows the negative log-likelihood associated with the Cauchy distribution.

$$\rho_{Cauchy}(p_t, p_s) = \log\left(\frac{1}{2}\left(\frac{\rho_{pc}(p_t, p_s)}{\sigma_{x_t}}\right)^2 + 1\right) + \log(\sqrt{2}\pi\sigma_{x_t}) \tag{9}$$

Similarly, we propose to attenuate the error contribution in the scale of the images. This is a single uncertainty σ_t is predicted by each image. In the training process, the uncertainty is optimized to match to the distribution of errors for all the pixels of each image.

3.4 Exploring Visibility Masks

We combine several strategies to filter out pixels that are likely to be outliers. We mask the pixels on the target image that lie out of the field-of-view on the source image, also known as principled mask [26], the pixels that belong to homogeneous regions and do not change their appearance, even when the camera is moving [8], and the target pixels that are occluded in the source view [9]. The resulting composite mask is applied to our adaptive consistency loss at each scale.

3.5 Implementation Details

The depth network is a convolutional encoder-decoder network with skip connections. We used a ResNet18 as backbone for the encoder part of the depth network. The decoder network is composed of deconvolutional layers that upsample the bottleneck representation in order to upscale the feature maps to the input resolution. For uncertainty estimation, we add a channel on the output of the depth network. In order to predict uncertainty pixel values, the extra channel is used as uncertainty map. On the other hand, when we aim to predict a single uncertainty value for image, we use spatial average pooling over the uncertainty map. The motion network predicts the relative motion between two input frames. The relative camera motion has a 6-DoF representation that corresponds to 3 rotation angles and the translation vector. The motion network is composed of the first five layers of the ResNet18 architecture, followed by a spatial average pooling and four 1×1 convolutional layers.

4 Experiments

In this section, we show the experiments conducted to evaluate each component of our system separately, as well the complete system with the proposed components.

4.1 Experiments Setup

Dataset. We use the KITTI benchmark [6]. It was created to reduce the bias and to complement available benchmarks with real-world data. It is composed of video sequences with 93 thousand images acquired through high-quality RGB cameras captured by driving on rural areas and highways of a city. We used the Eigen split [4] with 45023 images for training and 687 for testing. Moreover, we partitioned the training set on 40441 for training, 4582 for validation. For result evaluation, we used the standard metrics [4].

Training. Our networks are trained using ADAM optimization algorithm with a learning rate of $2e-5$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. We used the batch size of 12 snippets. Each snippet is a 3-frame sequence. The frames are resized to a resolution of 416×128 pixels.

4.2 Adaptive Consistency Loss

Table 1 shows the performance of the baseline model improved by considering the spatial and feature consistency loss terms individually, as well as combined using average and minimum re-projection. The first rows shows the results of our baseline model that only considers the photometric consistency and depth smoothness loss terms. Then, we evaluate the performance of the model including structure and feature consistency terms individually and jointly by using average or minimum re-projection. As other works in the literature [7,16,17,20,27], we show that including structure and feature consistency terms is beneficial. The results indicate that our implementations of structure and feature consistency can improve the performance of the model individually, in most of the metrics. Furthermore, we show that both terms are complementary and, together, can improve the performance with average and minimum re-projection losses. We obtained better results with minimum re-projection error.

Table 1. Ablation study. We evaluate the performance of structure and feature consistency terms with average re-projection, and the adaptive consistency loss, which uses minimum re-projection error.

				↓ Lower is better				↑ Higher is better		
Avg.	Min.	SC	FC	Abs Rel	Sq Rel	RMSE	Log RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
				0.1116	0.8905	4.7177	0.1840	0.8717	0.9564	0.9817
✓		✓		0.1113	1.0024	4.6312	0.1807	0.8797	0.9595	0.9823
✓			✓	0.1104	0.8747	4.6005	0.1800	0.8785	0.9587	0.9825
✓		✓	✓	0.1096	1.0134	4.5476	0.1776	**0.8838**	0.9611	0.9828
	✓	✓	✓	**0.1059**	**0.7520**	**4.4537**	**0.1737**	0.8834	**0.9620**	**0.9848**

4.3 Error Weighting Using Uncertainty

We evaluate the usage of uncertainty to weigh the error contribution when the uncertainty values are predicted by pixel and by image.

Error Weighting by Pixel. Table 2 shows that predicting uncertainty to weight the error contribution by pixel improves the performance of the baseline model using smooth-L1 probability function. However, the variants of the model that use the Laplacian and Cauchy distribution degrade the results.

We observe that the model predicts incorrect depth values on regions where the pixel intensities vary. This variation occurs because the predictive uncertainty is formulated on the photo-metric consistency term (Eq. 7).

Table 2. Using uncertainty to weigh the error contribution by pixel.

Method	↓ Lower is better				↑ Higher is better		
	Abs Rel	Sq Rel	RMSE	Log RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Baseline-L1	0.1894	4.1497	5.9739	0.2433	**0.8111**	0.9228	0.9599
Laplacian	0.1987	4.4033	6.0675	0.2481	0.8034	0.9216	0.9593
Smooth-L1	**0.1810**	**3.0795**	**5.6726**	**0.2386**	0.8027	**0.9245**	**0.9634**
Cauchy	0.1968	3.2513	5.9439	0.2540	0.7836	0.9147	0.9565

Error Weighting by Image. Table 3 shows the effect of predictive uncertainty by image to weight the error contribution of images using the Laplacian, Smooth-L1, and Cauchy probability functions. The first row shows our baseline, which use an L1 distance between pixel intensities to measure photometric consistency and depth smoothness.

Our results indicate that the Smooth-L1 function improves the performance of the baseline and outperforms the approaches that assume other distributions. However, using uncertainties predicted through Laplacian and Cauchy functions does not improve the performance. Qualitative results are illustrated in Fig. 2.

Table 3. Using uncertainty to weigh the error contribution by image.

Method	↓ Lower is better				↑ Higher is better		
	Abs Rel	Sq Rel	RMSE	Log RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Baseline-L1	0.1894	4.1497	5.9739	0.2433	0.8111	0.9228	0.9599
Laplacian	0.1928	4.4074	5.9921	0.2472	**0.8153**	0.9234	0.9598
Smooth-L1	**0.1561**	**1.3712**	**5.3931**	**0.2239**	0.8018	**0.9286**	**0.9683**
Cauchy	0.1976	3.3892	6.0628	0.2530	0.7846	0.9160	0.9600

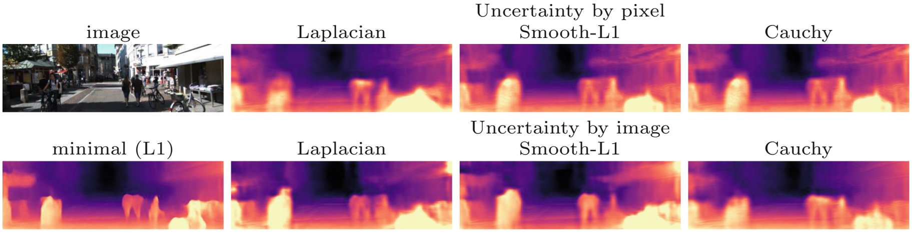

Fig. 2. Qualitative results of error weighting approach with uncertainty. The first column shows a target image and its depth maps predicted with the minimal model. The remaining columns show the results for the error weighting approaches for the PDF associated to Laplacian, Smooth-L1 and Cauchy functions. For each function, the first and second rows show the result of considering an uncertainty value by pixel and by image, respectively.

4.4 Visibility Masks

We performed ablation studies with visibility masks to filter out inconsistent pixels. We used model trained with the adaptive consistency loss as baseline. Table 4 shows that every mask improves the error metrics, as well as the thresholded accuracy metrics. Moreover, the model trained with all mask formulations achieved better results. Qualitative results are illustrated in Fig. 3.

Table 4. Ablation study of additional masks. We considered the Field-of-View masks (FOV), Auto mask (AM), Geometric mask (GM).

			↓ Lower is better				↑ Higher is better		
FOV	AM	GM	Abs Rel	Sq Rel	RMSE	Log RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
✓			0.1059	**0.7520**	4.4537	0.1737	0.8834	0.9620	**0.9848**
✓	✓		0.1063	0.8071	4.5570	0.1779	0.8829	0.9612	0.9831
✓		✓	0.1073	0.9355	**4.4135**	0.1734	0.8877	**0.9629**	0.9840
✓	✓	✓	**0.1015**	0.7692	4.4297	**0.1719**	**0.8890**	0.9622	0.9839

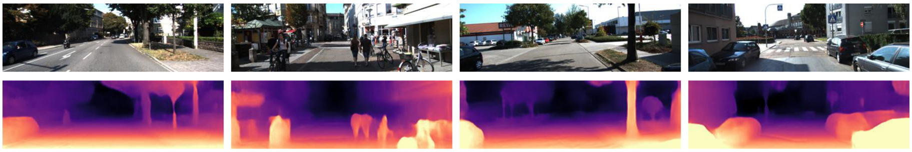

Fig. 3. Qualitative results. Depth prediction using our final model.

4.5 Comparison with the State of the Art

Table 5 shows that our method achieved competitive results when compared to state-of-the-art methods. Moreover, our approach is compatible and it could be improved with advanced strategies such as inference-time refinement [2,3], joint depth and optical flow estimation [3], and effective architecture designs [10].

Table 5. Results of depth estimation on the Eigen split of the KITTI dataset. We compared our results against several methods of the literature. In order to allow a fair comparison, we report the results of competitive methods trained with a resolution of 416×128 pixels. (*) indicates newly results obtained from an official repository. (-ref.) indicates that the online refinement component is disabled.

Method	↓ Lower is better				↑ Higher is better		
	Abs Rel	Sq Rel	RMSE	Log RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou et al. [28]*	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Mahjourian et al. [17]	0.163	1.240	6.220	0.250	0.762	0.916	0.967
Ying et al. [26]*	0.149	1.060	5.567	0.226	0.796	0.935	0.975
Casser et al. [2] (-ref.)	0.141	1.026	5.290	0.215	0.816	0.945	0.979
Chen et al. [3] (-ref.)	0.135	1.070	5.230	0.210	0.841	0.948	**0.980**
Gordon et al. [9]	0.129	**0.959**	5.230	0.213	0.840	0.945	0.976
Ours	0.131	1.037	5.173	**0.204**	0.846	0.952	**0.980**
Godard et al. [8]	**0.128**	1.087	**5.171**	**0.204**	**0.855**	**0.953**	0.978

5 Conclusions

In this work, we show that minimum re-projection can be used to jointly enforce consistency on photometric, 3D structure, and feature representations of frames. This approach reduces the influence of pixels without valid correspondences on other consistency constraints, in addition to photometric consistency.

Moreover, our results suggest that the error weighting approaches based on predictive uncertainty at pixel and image levels can be beneficial when the model is minimal, when the model does not implement additional strategies to handle invalid correspondences and when the outputs are assumed to follow the probability distribution derived from the smooth-L1 function. Further exploration could be done to leverage uncertainty to improve the performance of self-supervised depth estimation methods that consider several priors to handle invalid correspondences.

Acknowledgments. The authors are thankful to National Council for Scientific and Technological Development (CNPq grant #309330/2018-1) and São Paulo Research Foundation (FAPESP grants #17/12646-3 and #2018/00031-7) for their financial support.

References

1. Barron, J.T.: A general and adaptive robust loss function. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4331–4339 (2019)
2. Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Depth prediction without the sensors: leveraging structure for unsupervised learning from monocular videos. In: AAAI Conference on Artificial Intelligence, vol. 33, pp. 8001–8008 (2019)
3. Chen, Y., Schmid, C., Sminchisescu, C.: Self-supervised learning with geometric constraints in monocular video: connecting flow, depth, and camera. In: IEEE International Conference on Computer Vision, pp. 7063–7072 (2019)
4. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in Neural Information Processing Systems, pp. 2366–2374 (2014)
5. Garg, R., B.G., V.K., Carneiro, G., Reid, I.: Unsupervised CNN for single view depth estimation: geometry to the rescue. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 740–756. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_45
6. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. Int. J. Robot. Res. **32**(11), 1231–1237 (2013)
7. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 270–279 (2017)
8. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth prediction. In: International Conference on Computer Vision (ICCV), October 2019
9. Gordon, A., Li, H., Jonschkowski, R., Angelova, A.: Depth from videos in the wild: unsupervised monocular depth learning from unknown cameras. arXiv preprint arXiv:1904.04998 (2019)

10. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3D packing for self-supervised monocular depth estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2485–2494 (2020)
11. He, Y., Zhu, C., Wang, J., Savvides, M., Zhang, X.: Bounding box regression with uncertainty for accurate object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2888–2897 (2019)
12. Ilg, E., et al.: Uncertainty estimates and multi-hypotheses networks for optical flow. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 677–693. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_40
13. Klodt, M., Vedaldi, A.: Supervising the new with the old: learning SFM from SFM. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 713–728. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_43
14. Lee, M., Fowlkes, C.C.: CeMNet: self-supervised learning for accurate continuous ego-motion estimation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–8 (2019)
15. Li, R., Wang, S., Long, Z., Gu, D.: UnDeepVO: monocular visual odometry through unsupervised deep learning. In: IEEE International Conference on Robotics and Automation, pp. 7286–7291. IEEE (2018)
16. Luo, C., et al.: Every pixel counts++: joint learning of geometry and motion with 3D holistic understanding. arXiv preprint arXiv:1810.06125 (2018)
17. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5667–5675 (2018)
18. Mendoza, J., Pedrini, H.: Self-supervised depth estimation based on feature sharing and consistency constraints. In: 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Valletta, Malta, pp. 134–141, February 2020
19. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the uncertainty of self-supervised monocular depth estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3227–3237 (2020)
20. Shi, Y., Zhu, J., Fang, Y., Lien, K., Gu, J.: Self-supervised learning of depth and ego-motion with differentiable bundle adjustment. arXiv preprint arXiv:1909.13163 (2019)
21. Tonioni, A., Poggi, M., Mattoccia, S., Di Stefano, L.: Unsupervised domain adaptation for depth prediction from images. IEEE Trans. Pattern Anal. Mach. Intell. **42**(10), 2396–2409 (2019)
22. Wiles, O., Sophia Koepke, A., Zisserman, A.: Self-supervised learning of class embeddings from video. In: IEEE/CVF International Conference on Computer Vision Workshops, pp. 1–8 (2019)
23. Xu, H., Zheng, J., Cai, J., Zhang, J.: Region deformer networks for unsupervised depth estimation from unconstrained monocular videos. arXiv preprint arXiv:1902.09907 (2019)
24. Yang, N., von Stumberg, L., Wang, R., Cremers, D.: D3VO: deep depth, deep pose and deep uncertainty for monocular visual odometry. arXiv preprint arXiv:2003.01060 (2020)
25. Yasarla, R., Patel, V.M.: Uncertainty guided multi-scale residual learning-using a cycle spinning CNN for single image de-raining. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8405–8414 (2019)

26. Yin, Z., Shi, J.: GeoNet: unsupervised learning of dense depth, optical flow and camera pose. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1983–1992 (2018)
27. Zhan, H., Garg, R., Saroj Weerasekera, C., Li, K., Agarwal, H., Reid, I.: Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: The IEEE Conference on Computer Vision and Pattern Recognition, pp. 340–349 (2018)
28. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1851–1858 (2017)