Digging Into Self-Supervised Monocular Depth Estimation

Shashank Kapoor, Syed Rizwan Ali Quadri, Shubham Patwar
School of computing and Electrical Electronics
Indian Institute of Technology Mandi,
Kamand 175-005
Email: S22022@students.iitmandi.ac.in
Email: T22113@students.iitmandi.ac.in
Email: T22108@students.iitmandi.ac.in

1. Problem Statement

Monocular Depth Estimation is the task of estimating the depth value (distance relative to the camera) of each pixel given a single (monocular) RGB image. Generally State-of-the-art methods usually fall into one of two categories: designing a complex network that is powerful enough to directly regress the depth map, or splitting the input into bins or windows to reduce computational complexity. The most popular benchmark is the KITTI datasets. Models are typically evaluated using RMSE or absolute relative error.

Per-pixel ground-truth depth data is challenging to acquire at scale. To overcome this limitation, self-supervised learning has emerged as a promising alternative for training models to perform monocular depth estimation. In this, we propose a set of improvements, which together result in both quantitatively and qualitatively improved depth maps compared to competing self-supervised methods.

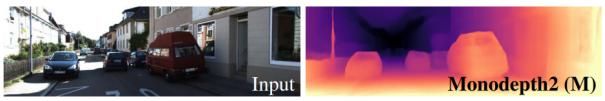

Figure 1. Depth from a single image.

Research on self-supervised monocular training usually explores increasingly complex architectures, loss functions, and image formation models, all of which have recently helped to close the gap with fully-supervised methods. We show that a surprisingly simple model, and associated design choices, lead to superior predictions. In particular, we propose:

(i) a minimum reprojection loss, designed to robustly handle occlusions

(ii) a full-resolution multi-scale sampling method that reduces visual artifacts

(iii) an auto-masking loss to ignore training pixels that violate camera motion assumptions.

We demonstrate the effectiveness of each component in isolation, and show high quality, state-of-the-art results on the KITTI benchmark.The goal is to find a dense depth map that minimizes the photometric error.

2. Literature Review

The paper discusses the challenge of inferring depth from a single image and proposes a self-supervised approach using monocular video or stereo pairs. The authors propose three innovations to improve monocular depth estimation:

- A novel appearance matching loss to address occluded pixels
- A simple auto-masking approach to ignore pixels with no relative camera motion
- A multi-scale appearance matching loss to reduce depth artifacts.

The proposed approach yields state-of-the-art results on the KITTI dataset and simplifies existing top-performing models.

The problem of estimating depth from a single image is difficult because there are multiple possible depths that an image can project to. Learning-based methods have been successful in addressing this problem by using predictive models that learn the relationship between color images and depth. Some of these methods require ground truth depth during training, which can be challenging to obtain in real-world settings. As a result, there is a growing body of work that exploits weakly supervised training data, such as known object sizes or synthetic depth data.

Recently, conventional structure-from-motion (SfM) pipelines have been used to generate sparse training signals for camera pose and depth. This approach has been shown to improve depth predictions when combined with traditional stereo algorithms.

3. Proposed Methods

The author discusses various methods for training depth estimation models in the absence of ground truth depth information. Self-supervised learning using stereo pairs or

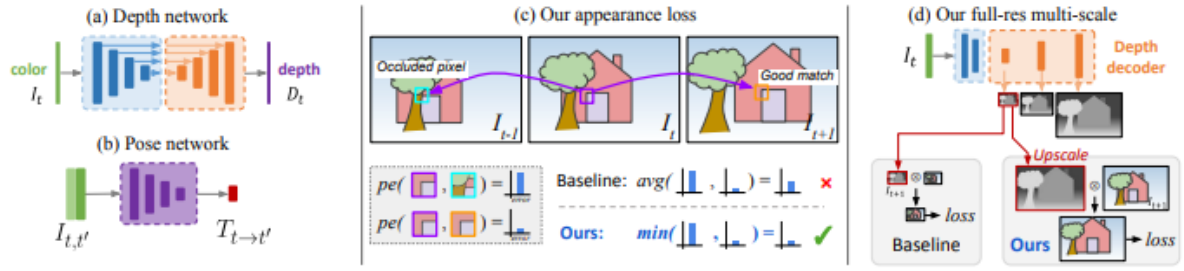

Figure 2. **Overview. (a) Depth network**: We use a standard, fully convolutional, U-Net to predict depth. **(b) Pose network**: Pose between a pair of frames is predicted with a separate pose network. **(c) Per-pixel minimum reprojection**: When correspondences are good, the reprojection loss should be low. However, occlusions and disocclusions result in pixels from the current time step not appearing in both the previous and next frames. The baseline average loss forces the network to match occluded pixels, whereas our minimum reprojection loss only matches each pixel to the view in which it is visible, leading to sharper results. (d) Full-resolution multi-scale: We upsample depth predictions at intermediate layers and compute all losses at the input resolution, reducing texture-copy artifacts.

monocular videos is a popular alternative, where the model is trained to minimize image reconstruction error by predicting depth or pixel disparities between frames.Appearance-based losses and careful resolution choices can improve performance, and recent approaches include geometry-based matching, instance segmentation, and explicit occlusion modeling. The article also proposes a novel minimum reprojection loss that improves the sharpness of depth predictions by only matching each pixel to the view in which it is visible.The training problem is formulated as the minimization of a photometric reprojection error, which measures the difference between the predicted image and the target image. The goal is to find a dense depth map that minimizes the photometric error.

The auto-masking method can also filter out pixels belonging to the object, but we found that this is a rare occurrence in practice. The auto-masking method can be seen as a form of data augmentation that helps the network learn to handle stationary scenes and moving objects.

Adaptive Depth Range Scaling, the predicted depths can vary between different scenes, making it difficult to compare results across dataset or even across different scenes within the same dataset. To address this issue, we propose a simple method to adaptively scale the predicted depth maps based on the median depth of each frame. Specifically, we compute the median depth of each frame and use it to scale the predicted depth map of that frame. This ensures that all frames have a similar scale, while still allowing for scene-specific depth ranges. We found that this simple method improves the accuracy of the predicted depth maps.

They also propose a multi-scale approach to depth estimation, where low-resolution depth maps are upsampled and the photometric error is computed at the high input resolution.

4. Results

This results in 39,810 monocular triplets for training and 4,424 for validation Here, we validate that:

(1) our reprojection loss helps with occluded pixels compared to existing pixel-averaging

(2) our auto-masking improves results, especially when training on scenes with static cameras

(3) our multi-scale appearance matching loss improves accuracy.

We evaluate our models, named Monodepth2, on the KITTI 2015 stereo dataset to allow comparison with previously published monocular methods.

We use the data split. Except in ablation experiments, for training which uses monocular sequences (i.e. monocular and monocular plus stereo) we follow pre-processing to remove static frames. This results in 39,810 monocular triplets for training and 4,424 for validation. We use the same intrinsics for all images, setting the principal point of the camera to the image center and the focal length to the average of all the focal lengths in KITTI. For stereo and mixed training (monocular plus stereo), we set the transformation between the two stereo frames to be a pure horizontal translation of fixed length. During evaluation, we cap depth to 80m per standard practice. For our monocular models, we report results using the per-image median ground truth scaling. for results we apply a single median scaling to the whole test set, instead of scaling each image independently. For results that use any stereo supervision we do not perform median scaling as scale can be inferred from the known camera baseline during training.

We compare the results of several variants of our model, trained with different types of self-supervision: monocular video only (M), stereo only (S), and both (MS). The results in Table 1 show that our monocular method outperforms all existing state-of-the-art self-supervised approaches. We also outperform recent methods that explicitly compute optical flow as well as motion masks. However, as with all image reconstruction based approaches to depth estimation, our model breaks when the scene contains objects that violate the Lambertian assumptions of our appearance loss .

As expected, the combination of M and S training data increases accuracy, which is especially noticeable on metrics that are sensitive to large depth errors e.g. RMSE. Despite

Method	Train	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen [9]	D	0.203	1.548	6.307	0.282	0.702	0.890	0.890
Liu [36]	D	0.201	1.584	6.471	0.273	0.680	0.898	0.967
Klodt [28]	D*M	0.166	1.490	5.998	-	0.778	0.919	0.966
AdaDepth [45]	D*	0.167	1.257	5.578	0.237	0.771	0.922	0.971
Kuznietsov [30]	DS	0.113	0.741	4.621	0.189	0.862	0.960	0.986
DVSO [68]	D*S	0.097	0.734	4.442	0.187	0.888	0.958	0.980
SVSM FT [39]	DS	0.094	0.626	4.252	0.177	0.891	0.965	0.984
Guo [16]	DS	0.096	0.641	4.095	0.168	0.892	0.967	0.986
DORN [10]	D	**0.072**	**0.307**	**2.727**	**0.120**	**0.932**	**0.984**	**0.994**
Zhou [76]†	M	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Yang [70]	M	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Mahjourian [40]	M	0.163	1.240	6.220	0.250	0.762	0.916	0.968
GeoNet [71]†	M	0.149	1.060	5.567	0.226	0.796	0.935	0.975
DDVO [62]	M	0.151	1.257	5.583	0.228	0.810	0.936	0.974
DF-Net [78]	M	0.150	1.124	5.507	0.223	0.806	0.933	0.973
LEGO [69]	M	0.162	1.352	6.276	0.252	-	-	-
Ranjan [51]	M	0.148	1.149	5.464	0.226	0.815	0.935	0.973
EPC++ [38]	M	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2depth '(M)' [5]	M	0.141	1.026	5.291	0.215	0.816	0.945	0.979
Monodepth2 w/o pretraining	M	0.132	1.044	5.142	0.210	0.845	0.948	0.977
Monodepth2	M	**0.115**	**0.903**	**4.863**	**0.193**	**0.877**	**0.959**	**0.981**
Monodepth2 (1024 × 320)	M	**0.115**	**0.882**	**4.701**	**0.190**	**0.879**	**0.961**	**0.982**
Garg [12]†	S	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Monodepth R50 [15]†	S	0.133	1.142	5.533	0.230	0.830	0.936	0.970
StrAT [43]	S	0.128	1.019	5.403	0.227	0.827	0.935	0.971
3Net (R50) [50]	S	0.129	0.996	5.281	0.223	0.831	0.939	0.974
3Net (VGG) [50]	S	0.119	1.201	5.888	0.208	0.844	0.941	**0.978**
SuperDepth + pp [47] (1024 × 382)	S	0.112	0.875	**4.958**	**0.207**	0.852	0.947	0.977
Monodepth2 w/o pretraining	S	0.130	1.144	5.485	0.232	0.831	0.932	0.968
Monodepth2	S	**0.109**	**0.873**	4.960	0.209	**0.864**	**0.948**	0.975
Monodepth2 (1024 × 320)	S	**0.107**	**0.849**	**4.764**	**0.201**	**0.874**	**0.953**	0.977
UnDeepVO [33]	MS	0.183	1.730	6.57	0.268	-	-	-
Zhan FullNYU [73]	D*MS	0.135	1.132	5.585	0.229	0.820	0.933	0.971
EPC++ [38]	MS	0.128	0.935	5.011	0.209	0.831	0.945	**0.979**
Monodepth2 w/o pretraining	MS	0.127	1.031	5.266	0.221	0.836	0.943	0.974
Monodepth2	MS	**0.106**	**0.818**	**4.750**	**0.196**	**0.874**	**0.957**	**0.979**
Monodepth2 (1024 × 320)	MS	**0.106**	**0.806**	**4.630**	**0.193**	**0.876**	**0.958**	**0.980**

Figure 3. **Quantitative results.** Comparison of our method to existing methods on KITTI 2015 using the Eigen split. Best results in each category are in bold; second best are underlined.

our contributions being designed around monocular training, we find that the in the stereo-only case we still perform well. We achieve high accuracy despite using a lower resolution than 1024 × 384, with substantially less training time (20 vs. 200 epochs) and no use of post processing.

5. Comparison

All results here are presented without post-processing. While our contributions are designed for monocular training, we still gain high accuracy in the stereo-only category.

We additionally show we can get higher scores at a larger 1024 × 320 resolution. These high resolution numbers are bolded if they beat all other models, including our low-res versions.

Benefits of auto-masking The full Eigen KITTI split contains several sequences where the camera does not move between frames e.g. where the data capture car was stopped at traffic lights. These 'no camera motion' sequences can cause problems for self-supervised monocular training, and as a result, they are typically excluded at training time using expensive to compute optical flow. We report monocular results trained on the full Eigen data split in, i.e. without removing frames. The baseline model trained on the full KITTI split performs worse than our full model.

Effect of ImageNet pretraining We follow previous work in initializing our encoders with weights pretrained on ImageNet. While some other monocular depth prediction works have elected not to use ImageNet pretraining, we show in Table 1 that even without pretraining, we still achieve state-of-the-art results. We train these 'w/o pretraining' models for 30 epochs to ensure convergence. Table 2 shows the benefit our contributions bring both to pretrained networks and those trained from scratch.

		Auto-masking	Min. reproj.	Full-res multi-scale	Pretrained	Full Eigen split	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
(a)	Baseline				✓		0.140	1.610	5.512	0.223	0.852	0.946	0.973
	Baseline + min reproj.		✓		✓		0.122	1.081	5.116	0.199	0.866	0.957	0.980
	Baseline + automasking	✓					0.124	0.936	5.010	0.206	0.858	0.952	0.977
	Baseline + full-res m.s.			✓	✓		0.124	1.170	5.249	0.203	0.865	0.953	0.978
	Monodepth2 w/o min reprojection	✓		✓	✓		0.117	0.878	4.846	0.196	0.870	0.957	0.980
	Monodepth2 w/o auto-masking		✓	✓	✓		0.120	1.097	5.074	0.197	0.872	0.956	0.979
	Monodepth2 w/o full-res m.s.	✓	✓		✓		0.117	**0.866**	4.864	0.196	0.871	0.957	**0.981**
	Monodepth2 with [76]'s mask		✓	✓	✓		0.123	1.177	5.210	0.200	0.869	0.955	0.978
	Monodepth2 smaller (416 × 128)	✓	✓	✓	✓		0.128	1.087	5.171	0.204	0.855	0.953	0.978
	Monodepth2 (full)	✓	✓	✓	✓		**0.115**	0.903	**4.863**	**0.193**	**0.877**	**0.959**	**0.981**
(b)	Baseline w/o pt						0.150	1.585	5.671	0.234	0.827	0.938	0.971
	Monodepth2 w/o pt	✓	✓	✓			**0.132**	**1.044**	**5.142**	**0.210**	**0.845**	**0.948**	**0.977**
(c)	Baseline (full Eigen dataset)				✓	✓	0.146	1.876	5.666	0.230	0.848	0.945	0.972
	Monodepth2 (full Eigen dataset)	✓	✓	✓	✓	✓	**0.116**	**0.918**	**4.872**	**0.193**	**0.874**	**0.959**	**0.981**

Figure 4. **Ablation.** Results for different variants of our model **(Monodepth2)** with monocular training on KITTI 2015 using the Eigen split. **(a)** The baseline model, with none of our contributions, performs poorly. The addition of our minimum reprojection, auto-masking and full-res multi-scale components, significantly improves performance. **(b)** Even without ImageNet pretrained weights, our much simpler model brings large improvements above the baseline – see also figure 3. **(c)** If we train with the full Eigen dataset (instead of the subset introduced for monocular training) our improvement over the baseline increases

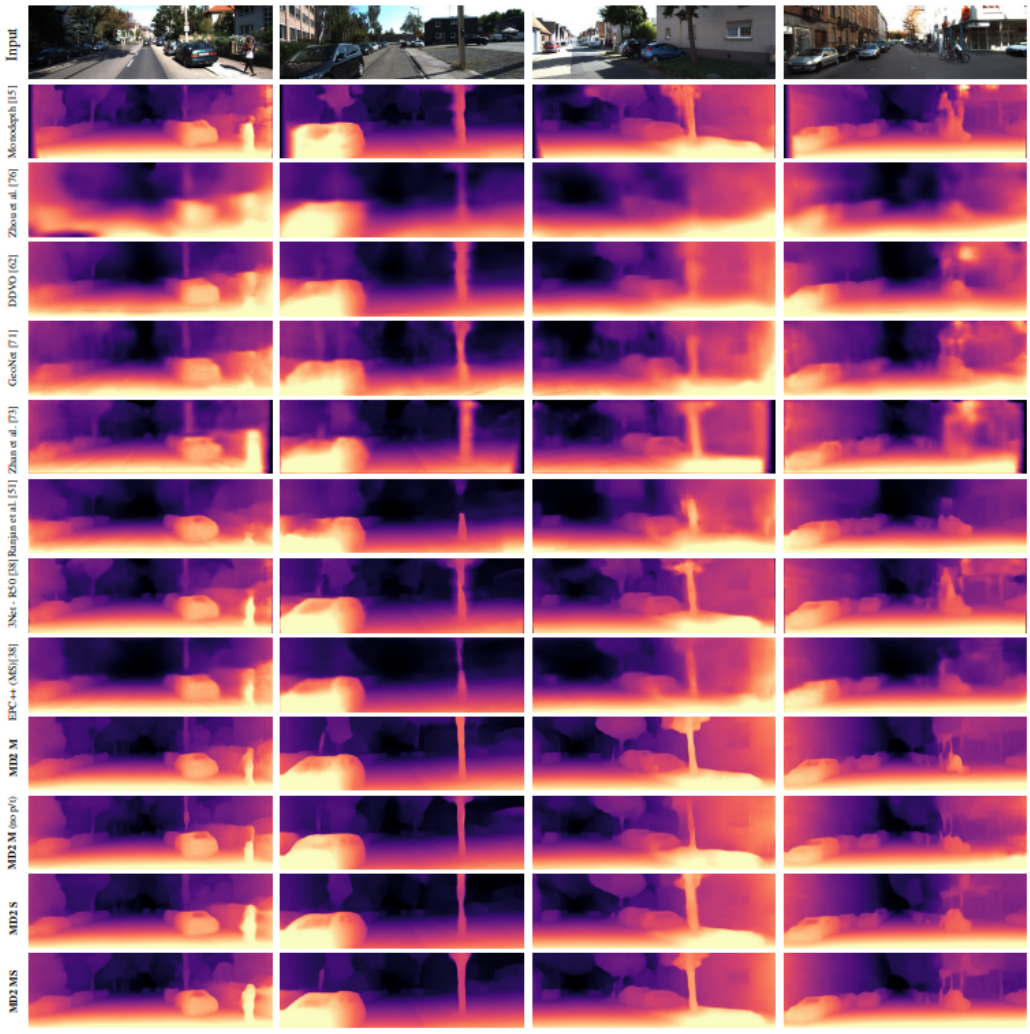

Figure 5. **Qualitative results on the KITTI Eigen split.** Our models (MD2) in the last four rows produce the sharpest depth maps, which are reflected in the superior quantitative results in Figure 1.

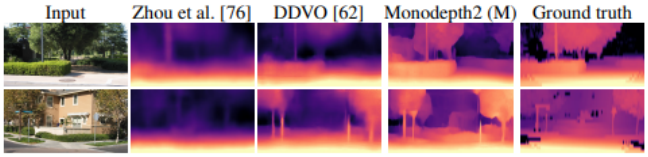

Figure 6. **Qualitative Make3D results.** All methods were trained on KITTI using monocular supervision.

6. Conclusion

We have presented a versatile model for self-supervised monocular depth estimation, achieving state-of-the-art depth predictions. We introduced three contributions: (i) a minimum reprojection loss, computed for each pixel, to deal with occlusions between frames in monocular video, (ii) an auto-masking loss to ignore confusing, stationary pixels, and (iii) a full-resolution multi-scale sampling method. We showed how together they give a simple and efficient model for depth estimation, which can be trained with monocular video data, stereo data, or mixed monocular and stereo data.

Acknowledgement Thanks to the course instructor Dr. Parimala Kancharla for discussing with us the latest topic on Monocular Depth estimation. we learned, by taking this as the project work. we looking forward for the more of the collaborative work in future. **Thank you Mam!**

References

[1] Clement Godard et. al. *Digging Into self-Supervised Monocular Dept Estimation*, CVPR, August 2019