Self-supervised Monocular Trained Depth Estimation using Self-attention and Discrete Disparity Volume

Adrian Johnston Gustavo Carneiro
Australian Institute for Machine Learning
School of Computer Science, University of Adelaide
{adrian.johnston, gustavo.carneiro}@adelaide.edu.au

Abstract

Monocular depth estimation has become one of the most studied applications in computer vision, where the most accurate approaches are based on fully supervised learning models. However, the acquisition of accurate and large ground truth data sets to model these fully supervised methods is a major challenge for the further development of the area. Self-supervised methods trained with monocular videos constitute one the most promising approaches to mitigate the challenge mentioned above due to the wide-spread availability of training data. Consequently, they have been intensively studied, where the main ideas explored consist of different types of model architectures, loss functions, and occlusion masks to address non-rigid motion. In this paper, we propose two new ideas to improve self-supervised monocular trained depth estimation: 1) self-attention, and 2) discrete disparity prediction. Compared with the usual localised convolution operation, self-attention can explore a more general contextual information that allows the inference of similar disparity values at non-contiguous regions of the image. Discrete disparity prediction has been shown by fully supervised methods to provide a more robust and sharper depth estimation than the more common continuous disparity prediction, besides enabling the estimation of depth uncertainty. We show that the extension of the state-of-the-art self-supervised monocular trained depth estimator Monodepth2 with these two ideas allows us to design a model that produces the best results in the field in KITTI 2015 and Make3D, closing the gap with respect self-supervised stereo training and fully supervised approaches.

1. Introduction

Perception of the 3D world is one of the main tasks in computer/robotic vision. Accurate perception, localisation, mapping and planning capabilities are predicated on having access to correct depth information. Range finding sensors such as LiDAR or stereo/multi-camera rigs are often deployed to estimate depth for use in robotics and autonomous systems, due to their accuracy and robustness. However, in many cases it might be unfeasible to have, or rely solely on such expensive or complex sensors. This has led to the development of learning-based methods [49, 50, 20], where the most successful approaches rely on fully supervised convolutional neural networks (CNNs) [9, 8, 10, 15, 35].

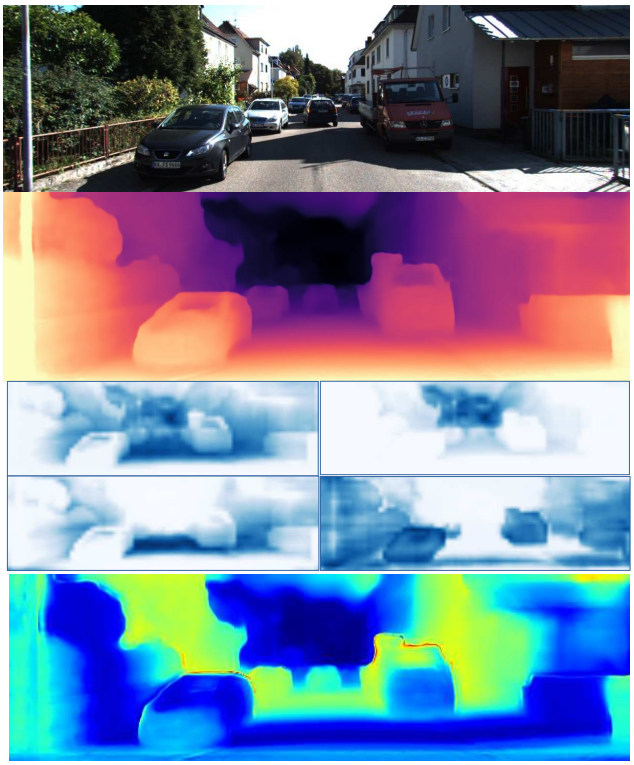

Figure 1. **Self-supervised Monocular Trained Depth Estimation using Self-attention and Discrete Disparity Volume**. Our self-supervised monocular trained model uses self-attention to improve contextual reasoning and discrete disparity estimation to produce accurate and sharp depth predictions and depth uncertainties. *Top: input image*; *Middle Top: estimated disparity*; *Middle Bottom: samples of the attention maps produced by our system (blue indicates common attention regions)*; *Bottom: pixel-wise depth uncertainty (blue: low uncertainty; green/red: high/highest uncertainty).*

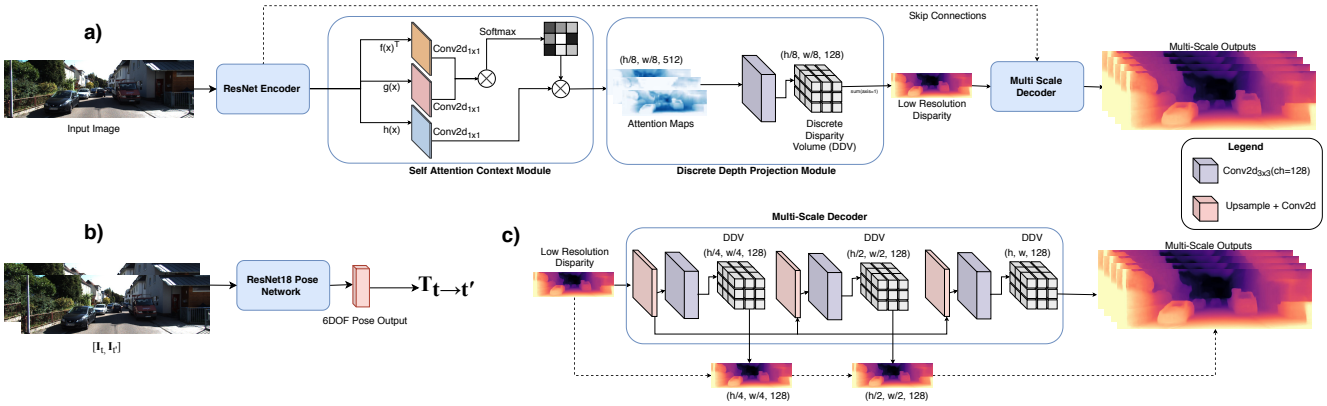

Figure 2. **Overall Architecture** The image encoding processes is highlighted in part *a)*. The input monocular image is encoded using a ResNet encoder and then passed through the Self-Attention Context Module. The computed attention maps are then convolved with a 2D convolution with the number of output channels equal to the number dimensions for the Discrete Disparity Volume (DDV). The DDV is then projected into a 2D depth map by performing a *softargmax* across the disparity dimension resulting in the lowest resolution disparity estimation (Eq. 4). In part *b)* the pose estimator is shown, and part *c)* shows more details of the Multi-Scale decoder. The low resolution disparity map is passed through successive blocks of UpConv (nearest upsample + convolution). The DDV projection is performed at each scale, in the same way as in the initial encoding stage. Finally, each of the outputs are upsampled to input resolution to compute the photometric reprojection loss.

While supervised learning methods have produced outstanding monocular depth estimation results, ground truth RGB-D data is still limited in variety and abundance when compared with the RGB image and video data sets available in the field. Furthermore, collecting accurate and large ground truth data sets is a difficult task due to sensor noise and limited operating capabilities (due to weather conditions, lighting, etc.).

Recent studies have shown that it is instead possible to train a depth estimator in a self-supervised manner using synchronised stereo image pairs [11, 13] or monocular video [62]. While monocular video offers an attractive alternative to stereo based learning due to wide-spread availability of training sequences, it poses many challenges. Unlike stereo based methods, which have a known camera pose that can be computed offline, self-supervised monocular trained depth estimators need to jointly estimate depth and ego-motion to minimise the photometric reprojection loss function [11, 13]. Any noise introduced by the pose estimator model can degrade the performance of a model trained on monocular sequences, resulting in large depth estimation errors. Furthermore, self-supervised monocular training makes the assumption of a moving camera in a static (i.e., rigid) scene, which causes monocular models to estimate 'holes' for pixels associated with moving visual objects, such as cars and people (i.e., non-rigid motion). To deal with these issues, many works focus on the development of new specialised architectures [62], masking strategies [62, 14, 52, 32], and loss functions [13, 14]. Even with all of these developments, self-supervised monocular trained depth estimators are less accurate than their stereo trained counterparts and significantly less accurate than fully supervised methods.

In this paper, we propose two new ideas to improve self-supervised monocular trained depth estimation: 1) self-attention [54, 51], and 2) discrete disparity volume [22]. Our proposed self-attention module explores non-contiguous (i.e., global) image regions as a context for estimating similar depth at those regions. Such approach contrasts with the currently used local 2D and 3D convolutions that are unable to explore such global context. The proposed discrete disparity volume enables the estimation of more robust and sharper depth estimates, as previously demonstrated by fully supervised depth estimation approaches [22, 29]. Sharper depth estimates are important to improving accuracy, and increased robustness is desirable to allow self-supervised monocular trained depth estimation to address common mistakes made by the method, such as incorrect pose estimation and matching failures because of uniform textural details. We also show that our method can estimate pixel-wise depth uncertainties with the proposed discrete disparity volume [22]. Depth uncertainty estimation is important for refining depth estimation [10], and in safety critical systems [21], allowing an agent to identify unknowns in an environment in order to reach optimal decisions. As a secondary contribution of this paper, we leverage recent advances in semantic segmentation network architectures that allow us to train larger models on a single GPU machine. Experimental results show that our novel approach produces the best self-supervised monocular depth estimation results for KITTI 2015 and Make3D. We also show in the experiments that our method is able to close the gap with self-supervised stereo trained and fully supervised depth estimators.

2. Related Work

Many computer vision and robotic systems that are used in navigation, localization and mapping rely on accurately understanding the 3D world around them [37, 16, 7, 1]. Active sensors such as LiDAR, Time of Flight cameras, or Stereo/Multi camera rigs are often deployed in robotic and autonomous systems to estimate the depth of an image for understanding the agent's environment [7, 1]. Despite their wipe-spread adoption [45], these systems have several drawbacks [7], including limited range, sensor noise, power consumption and cost. Instead of relying on these active sensor systems, recent advances leveraging fully supervised deep learning methods [9, 8, 10, 15, 35] have made it possible to learn to predict depth from monocular RGB cameras [9, 8]. However, ground truth RGB-D data for supervised learning can be difficult to obtain, especially for every possible environment we wish our robotic agents to operate. To alleviate this requirement, many recent works have focused on developing self-supervised techniques to train monocular depth estimators using synchronised stereo image pairs [11, 13, 41], monocular video [62, 14] or binocular video[60, 14, 32].

2.1. Monocular Depth Estimation

Depth estimation from a monocular image is an inherently ill-posed problem as pixels in the image can have multiple plausible depths. Nevertheless, methods based on supervised learning have been shown to mitigate this challenge and correctly estimate depth from colour input images [50]. Eigen *et al*. [9] proposed the first method based on Deep Learning, which applies a multi-scale convolution neural network and a scale-invariant loss function to model local and global features within an image. Since then, fully supervised deep learning based methods have been continuously improved [10, 15, 35]. However these methods are limited by the availability of training data, which can be costly to obtain. While such issues can be mitigated with the use of synthetic training data [35], simulated environments need to be modelled by human artists, limiting the amount of variation in the data set. To overcome fully supervised training set constraint, Garg *et al*. [11] propose a self-supervised framework, where instead of supervising using ground truth depth, a stereo photometric reprojection warping loss is used to implicitly learn depth. This loss function is a pixel-based reconstruction loss that uses stereo pairs, where the right image of the pair is warped into the left using a differentiable image sampler [19]. This loss function allows the deep learning model to implicitly recover the underlying depth for the input image. Expanding on this method, Godard *et al*. [13] add a left-right consistency loss term which helps to ensure consistency between the predicted depths from the left and right images of the stereo pair. While capable of training monocular depth estimators, these methods still rely on stereo-based training data which can still be difficult to acquire. This has motivated the development of self-supervised monocular trained depth estimators [62] which relax the requirement of synchronized stereo image pairs by jointly learning to predict depth and ego-motion with two separate networks, enabling the training of a monocular depth estimator using monocular video. To achieve this, the scene is assumed to be static (i.e., rigid), while the only motion is that of the camera. However, this causes degenerate behaviour in the depth estimator when this assumption is broken. To deal with this issue, the paper [62] includes a predictive masking which learns to ignore regions that violates the rigidity assumptions. Vijayanarasimhan *et al*. [52] propose a more complex motion model based on multiple motion masks, and GeoNet model [58] decomposes depth and optical flow to account for object motion within the image sequence. Self-supervised monocular trained methods have been further improved by constraining predicted depths to be consistent with surface normals [57], using pre-computed instance-level segmentation masks [3] and increasing the resolution of the input images [41]. Godard *et al*. [14] further close the performance gap between monocular and stereo-trained self-supervision with Monodepth2 which uses multi-scale estimation and a per-pixel minimum re-projection loss that better handles occlusions. We extend Monodepth2 with our proposed ideas, namely self-attention and discrete disparity volume.

2.2. Self-attention

Self-attention has improved the performance of natural language processing (NLP) systems by allowing a better handling of long-range dependencies between words [51], when compared with recurrent neural networks (RNN) [47], long short term memory (LSTM) [18], and convolutional neural nets (CNN) [27]. This better performance can be explained by the fact that RNNs, LSTMs and CNNs can only process information in the local word neighbourhood, making these approaches insufficient for capturing long range dependencies in a sentence [51], which is essential in some tasks, like machine translation. Self-attention has been proposed in computer vision for improving Image Classification and Object Drection [2, 39]. Self-attention has also improved the performance of computer vision tasks such as semantic segmentation [59] by addressing more effectively the problem of segmenting visual classes in non-contiguous regions of the image, when compared with convolutional layers [4, 61, 6], which can only process information in the local pixel neighbourhood. In fact, many of the recent improvements in semantic segmentation performance stem from improved contextual aggregation strategies (i.e., strategies that can process spatially non-contiguous image regions) such as the Pyramid Pooling Module (PPM) in PSPNet [61], and the Atrous Spatial Pyramid Pooling [4]. In both of these methods, multiple scales of information are aggregated to improve the contextual representation by the network. Yuan *et al*. [59] further improve on this area with OCNet, which adds to a ResNet-101 [17] backbone a

self-attention module that learns to contextually represent groups of features with similar semantic similarity. Therefore, we hypothesise that such self-attention mechanisms can also improve depth prediction using monocular video because the correct context for the prediction of a pixel depth may be at a non-contiguous location that the standard convolutions cannot reach.

2.3. Discrete Disparity Volume

Kendall *et al*. [22] propose to learn stereo matching in a supervised manner, by using a shared CNN encoder with a cost volume that is refined using 3D convolutions. Liu *et al*. [29] investigate this idea further by training a model using monocular video with ground truth depth and poses. This paper [29] relies on a depth probability volume (DPV) and a Bayesian filtering framework that refines outliers based on the uncertainty computed from the DPV. Fu *et al*. [10] represent their ground-truth depth data as discrete bins, effectively forming a disparity volume for training. All methods above work in fully-supervised scenarios, showing advantages for depth estimation robustness and sharpness, allied with the possibility of estimating depth uncertainty. Such uncertainty estimation can be used by autonomous systems to improve decision making [21] or to refine depth estimation [10]. In this paper, we hypothesis that the extension of self-supervised monocular trained methods with a discrete disparity volume will provide the same advantages observed in fully-supervised models.

3. Methods

In the presentation of our proposed model for self-supervised monocular trained depth estimation, we focus on showing the importance of the main contributions of this paper, namely self-attention and discrete disparity volume. We use as baseline, the Monodepth2 model [14] based on a UNet architecture [44].

3.1. Model

We represent the RGB image with $\mathbf{I} : \Omega \rightarrow \mathbb{R}^3$, where Ω denotes the image lattice of height H and width W. The first stage of the model, depicted in Fig. 2, is the ResNet-101 encoder, which forms $\mathbf{X} = resnet_\theta(\mathbf{I}_t)$, with $\mathbf{X} : \Omega_{1/8} \rightarrow \mathbb{R}^M$, M denoting the number of channels at the output of the ResNet, and $\Omega_{1/8}$ representing the low-resolution lattice at $(1/8)^{th}$ of its initial size in Ω. The ResNet output is then used by the self-attention module [54], which first forms the query, key and value results, represented by:

$$\begin{aligned} f(\mathbf{X}(\omega)) &= \mathbf{W}_f \mathbf{X}(\omega), \\ g(\mathbf{X}(\omega)) &= \mathbf{W}_g \mathbf{X}(\omega), \\ h(\mathbf{X}(\omega)) &= \mathbf{W}_h \mathbf{X}(\omega), \end{aligned} \tag{1}$$

respectively, with $\mathbf{W}_f, \mathbf{W}_g, \mathbf{W}_h \in \mathbb{R}^{N \times M}$. The query and key values are then combined with

$$\mathbf{S}_\omega = softmax(f(\mathbf{X}(\omega))^T g(\mathbf{X})), \tag{2}$$

where $\mathbf{S}_\omega : \Omega_{1/8} \rightarrow [0, 1]$, and we abuse the notation by representing $g(\mathbf{X})$ as a tensor of size $N \times H/8 \times W/8$. The self-attention map is then built by the multiplication of value and $\mathbf{S}_\omega$ in (2), with:

$$\mathbf{A}(\omega) = \sum_{\tilde{\omega} \in \Omega_{1/8}} h(\mathbf{X}(\tilde{\omega})) \times \mathbf{S}_\omega(\tilde{\omega}), \tag{3}$$

with $\mathbf{A} : \Omega_{1/8} \rightarrow \mathbb{R}^N$.

The low-resolution discrete disparity volume (DDV) is denoted by $\mathbf{D}_{1/8}(\omega) = conv_{3\times3}(\mathbf{A}(\omega))$, with $\mathbf{D}_{1/8} : \Omega_{1/8} \rightarrow \mathbb{R}^K$ (K denotes the number of discretized disparity values), and $conv_{3\times3}(.)$ denoting a convolutional layer with filters of size 3×3. The low resolution disparity map is then computed with

$$\sigma(\mathbf{D}_{1/8}(\omega)) = \sum_{k=1}^{K} softmax(\mathbf{D}_{1/8}(\omega)[k]) \times disparity(k), \tag{4}$$

where $softmax(\mathbf{D}_{1/8}(\omega)[k])$ is the softmax result of the k^{th} output from $\mathbf{D}_{1/8}$, and $disparity(k)$ holds the disparity value for k. Given the ambiguous results produced by these low-resolution disparity maps, we follow the multi-scale strategy proposed by Godard *et al*. [14]. The low resolution map from (4) is the first step of the multi-scale decoder that consists of three additional stages of upconv operators (i.e., nearest upsample + convolution) that receive skip connections from the ResNet encoder for the respective resolutions, as shown in Fig. 2. These skip connections between encoding layers and associated decoding layers are known to retain high-level information in the final depth output. At each resolution, we form a new DDV, which is used to compute the disparity map at that particular resolution. The resolutions considered are (1/8), (1/4), (1/2), and (1/1) of the original resolution, respectively represented by $\sigma(\mathbf{D}_{1/8})$, $\sigma(\mathbf{D}_{1/4})$, $\sigma(\mathbf{D}_{1/2})$, and $\sigma(\mathbf{D}_{1/1})$.

Another essential part of our model is the pose estimator [62], which takes two images recorded at two different time steps, and returns the relative transformation, as in

$$\mathbf{T}_{t \rightarrow t'} = p_\phi(\mathbf{I}_t, \mathbf{I}_{t'}), \tag{5}$$

where $\mathbf{T}_{t \rightarrow t'}$ denotes the transformation matrix between images recorded at time steps t and t', and $p_\phi(.)$ is the pose estimator, consisting of a deep learning model parameterised by ϕ.

3.2. Training and Inference

The training is based on the minimum per-pixel photometric re-projection error [14] between the source image $\mathbf{I}_{t'}$ and the target image $\mathbf{I}_t$, using the relative pose $\mathbf{T}_{t \rightarrow t'}$ defined in (5). The pixel-wise error is defined by

$$\ell_p = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left(\min_{t'} \mu^{(s)} \times pe(\mathbf{I}_t, \mathbf{I}^{(s)}_{t \rightarrow t'}) \right), \tag{6}$$

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 [30]	D	0.201	1.584	6.471	0.273	0.680	0.898	0.967
Klodt [24]	D*M	0.166	1.490	5.998	-	0.778	0.919	0.966
AdaDepth [38]	D*	0.167	1.257	5.578	0.237	0.771	0.922	0.971
Kuznietsov [25]	DS	0.113	0.741	4.621	0.189	0.862	0.960	0.986
DVSO [55]	D*S	0.097	0.734	4.442	0.187	0.888	0.958	0.980
SVSM FT [33]	DS	0.094	0.626	4.252	0.177	0.891	0.965	0.984
Guo [15]	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 [62]†	M	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Yang [57]	M	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Mahjourian [34]	M	0.163	1.240	6.220	0.250	0.762	0.916	0.968
GeoNet [58]†	M	0.149	1.060	5.567	0.226	0.796	0.935	0.975
DDVO [53]	M	0.151	1.257	5.583	0.228	0.810	0.936	0.974
DF-Net [63]	M	0.150	1.124	5.507	0.223	0.806	0.933	0.973
LEGO [56]	M	0.162	1.352	6.276	0.252	-	-	-
Ranjan [43]	M	0.148	1.149	5.464	0.226	0.815	0.935	0.973
EPC++ [32]	M	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2depth '(M)' [3]	M	0.141	1.026	5.291	0.215	0.816	0.945	0.979
Monodepth2 [14]	M	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Monodepth2 (1024 × 320)[14]	M	0.115	0.882	4.701	0.190	0.879	0.961	**0.982**
Ours	M	**0.106**	**0.861**	**4.699**	**0.185**	**0.889**	**0.962**	**0.982**
Garg [11]†	S	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Monodepth R50 [13]†	S	0.133	1.142	5.533	0.230	0.830	0.936	0.970
StrAT [36]	S	0.128	1.019	5.403	0.227	0.827	0.935	0.971
3Net (R50) [42]	S	0.129	0.996	5.281	0.223	0.831	0.939	0.974
3Net (VGG) [42]	S	0.119	1.201	5.888	0.208	0.844	0.941	**0.978**
SuperDepth + pp [41] (1024 × 382)	S	0.112	0.875	4.958	0.207	0.852	0.947	0.977
Monodepth2 [14]	S	0.109	0.873	4.960	0.209	0.864	0.948	0.975
Monodepth2 (1024 × 320)[14]	S	**0.107**	**0.849**	**4.764**	**0.201**	**0.874**	**0.953**	0.977
UnDeepVO [28]	MS	0.183	1.730	6.57	0.268	-	-	-
Zhan FullNYU [60]	D*MS	0.135	1.132	5.585	0.229	0.820	0.933	0.971
EPC++ [32]	MS	0.128	0.935	5.011	0.209	0.831	0.945	0.979
Monodepth2[14]	MS	**0.106**	0.818	4.750	0.196	0.874	0.957	0.979
Monodepth2(1024 × 320)[14]	MS	**0.106**	**0.806**	**4.630**	**0.193**	**0.876**	**0.958**	**0.980**

Table 1. **Quantitative results.** Comparison of existing methods to our own on the KITTI 2015 [12] using the Eigen split [8]. The Best results are presented in **bold** for each category, with second best results underlined. The supervision level for each method is presented in the *Train* column with; D – Depth Supervision, D* – Auxiliary depth supervision, S – Self-supervised stereo supervision, M – Self-supervised mono supervision. Results are presented without any post-processing [13], unless marked with – + pp. If newer results are available on github, these are marked with – †. Non-Standard resolutions are documented along with the method name. Metrics indicated by red: *lower is better*, Metrics indicated by blue: *higher is better*

where $pe(.)$ denotes the photometric reconstruction error, $\mathcal{S} = \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, \frac{1}{1}\}$ is the set of the resolutions available for the disparity map, defined in (4), $t' \in \{t-1, t+1\}$, indicating that we use two frames that are temporally adjacent to $\mathbf{I}_t$ as its source frames [14], and $\mu^{(s)}$ is a binary mask that filters out stationary points (see more details below in Eq.10) [14]. The re-projected image in (6) is defined by

$$\mathbf{I}^{(s)}_{t\to t'} = \mathbf{I}_{t'}\big\langle proj(\sigma(\mathbf{D}^{(s)}_t), \mathbf{T}_{t\to t'}, \mathbf{K})\big\rangle, \tag{7}$$

where $proj(.)$ represents the 2D coordinates of the projected depths $\mathbf{D}_t$ in $\mathbf{I}_{t'}$, $\langle . \rangle$ is the sampling operator, and $\sigma(\mathbf{D}^{(s)}_t)$ is defined in (4). Similarly to [14], the pre-computed intrinsics $\mathbf{K}$ of all images are identical, and we use bi-linear sampling to sample the source images and

$$pe(\mathbf{I}_t, \mathbf{I}^{(s)}_{t'}) = \frac{\alpha}{2}(1-\text{SSIM}(\mathbf{I}_t, \mathbf{I}^{(s)}_{t'}))+(1-\alpha)\|\mathbf{I}_t-\mathbf{I}^{(s)}_{t'}\|_1, \tag{8}$$

where $\alpha = 0.85$. Following [13] we use an edge-aware smoothness regularisation term to improve the predictions around object boundaries:

$$\ell_s = |\partial_x d^*_t|\, e^{-|\partial_x \mathbf{I}_t|} + |\partial_y d^*_t|\, e^{-|\partial_y \mathbf{I}_t|}, \tag{9}$$

where $d^*_t = d_t/\overline{d_t}$ is the mean-normalized inverse depth from [53] to discourage shrinking of the estimated depth. The auto-masking of stationary points [14] in (6) is necessary because the assumptions of a moving camera and a static scene are not always met in self-supervised monocular trained depth estimation methods [14]. This masking filters out pixels that remain with the same appearance between two frames in a sequence, and is achieved with a binary mask defined as

$$\mu^{(s)} = \big[\min_{t'} pe(\mathbf{I}_t, \mathbf{I}^{(s)}_{t'\to t}) < \min_{t'} pe(\mathbf{I}_t, \mathbf{I}_{t'})\big], \tag{10}$$

where $[.]$ represents the Iverson bracket. The binary mask μ in (10) masks the loss in (6) to only include the pixels where the re-projection error of $\mathbf{I}^{(s)}_{t'\to t}$ is lower than the error of the un-warped image $\mathbf{I}_{t'}$, indicating that the visual object is moving relative to the camera. The final loss is computed as the weighted sum of the per-pixel minimum reprojection loss in (6) and smoothness term in (9),

$$\ell = \ell_p + \lambda \ell_s \tag{11}$$

where λ is the weighting for the smoothness regularisation term. Both the pose model and depth model are trained jointly using this photometric reprojection error. Inference is achieved by taking a test image at the input of the model and producing the high-resolution disparity map $\sigma(\mathbf{D}_{1/1})$.

4. Experiments

We train and evaluate our method using the KITTI 2015 stereo data set [12]. We also evaluate our method on the Make3D data set [50] using our model trained on KITTI 2015. We use the split and evaluation of Eigen *et al*. [8], and following previous works [62, 14], we remove static frames before training and only evaluate depths up to a fixed range of 80m [8, 11, 13, 14]. As with [14], this results in 39,810 monocular training sequences, consisting of sequences of three frames, with 4,424 validation sequences. As our baseline model, we use Monodepth2 [14], but we replace the original ResNet-18 by a ResNet-101 that has higher capacity, but requires more memory. To address this memory issue, we use the inplace activated batch normalisation [46], which fuses the batch normalization layer and the activation functions to reach up to 50% memory savings. As self-supervised monocular trained depth estimators do not contain scale information, we use the per-image median ground truth scaling [62, 14]. Following architecture best practices from the Semantic Segmentation community, we adopt the *atrous convolution* [5], also known as the *dilated convolution*, in the last two convolutional blocks of the ResNet-101 encoder [61, 59, 5, 6] with dilation rates of 2 and 4, respectively. This has been shown to significantly improve multi-scale encoding by increasing the models field-of-view [5]. The results for the quantitative analysis are shown in Sec. 4.2. We also present an ablation study comparing the effects of the our different contributions in Sec. 4.4. Final models are selected using the lowest absolute relative error metric on the validation set.

4.1. Implementation Details

Our system is trained using the PyTorch library [40], with models trained on a single Nvidia 2080Ti for 20 epochs. We jointly optimize both our pose and depth networks with the Adam Optimizer [23] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of $1e^{-4}$. We use a single learning rate decay to $lr = 1e^{-5}$ after 15 epochs. As with previous papers [14], our ResNet encoders use pre-trained ImageNet [48] weights as this has been show to reduce training time and improve overall accuracy of the predicted depths. All models are trained using the following data augmentations with 50% probability; Horizontal flips, random contrast (± 0.2), saturation (± 0.2), hue jitter (± 0.1) and brightness (± 0.2). Crucially, augmentations are only performed on the images input into the depth and pose network and the loss in (11) is computed using the original ground truth images, with the smoothness term set to $\lambda = 1e^{-3}$. Image resolution is set to 640×192 pixels.

4.2. KITTI Results

The results for the experiment are presented in Table 1. When comparing our method (grayed row in Table 1) on the KITTI 2015 data set [12] (using Eigen [8] split), we observe that we outperform all existing self-supervised monocular trained methods by a significant margin. Compared to other methods that rely on stronger supervision signals (e.g., stereo supervision and mono+stereo supervision), our approach is competitive, producing comparable results to the current state of the art method Monodepth2. As can be seen in Figure 3 our method shows sharper results on thinner structures such as poles than the baseline Monodepth2. In general, Monodepth2 (Mono and Mono+Stereo) struggles with thin structures that overlap with foliage, while our method is able to accurately estimate the depth of these smaller details. We attribute this to the combination of the dilated convolutions and the contextual information from the self-attention module. As can be seen in car windows, Monodepth2 and our method struggle to predict the depth on glassy reflective surfaces. However, this is a common issue observed in self-supervised methods because they cannot accurately predict depth for transparent surfaces since the photometric reprojection/warping error is ill-defined for such materials/surfaces. For instance, in the example of car windows, the correct depth that would minimise the photometric reprojection loss is actually the depth from the car interior, instead of the glass depth, as would be recorded by the ground truth LiDAR. When comparing our method against some specific error cases for Monodepth2 [14] (Figure 4), we can see that our method succeeds in estimating depth of the highly reflective car roof (*left*) and successfully disentangles the street sign from the background (*right*). This can be explained by the extra context and receptive field afforded by the self-attention context module as well as the regularisation provided by the discrete disparity volume.

4.3. Make3D Results

Table 3 presents the quantitative results for the Make3D data set [50] using our model trained on KITTI2015. We follow the same testing protocol as Monodepth2 [14] and methods are compared using the evaluation criteria outline in [13]. It can be seen in Table 3 that our method produces superior results compared with previous methods that also rely on self-supervision.

4.4. Ablation Study

Table 2 shows an ablation study of our method, where we start from the baseline Monodepth2 [14] (row 1). Then, by first adding DDV (row 2) and both self attention and DDV (row 3), we observe a steady improvement in almost all evaluation measures. We then switch the underlying encoding model ResNet-18 to ResNet-101 with dilated convolutions in row 4. Rows 5 and 6 show the addition of DDV and then both self-attention and DDV, respectively, again with

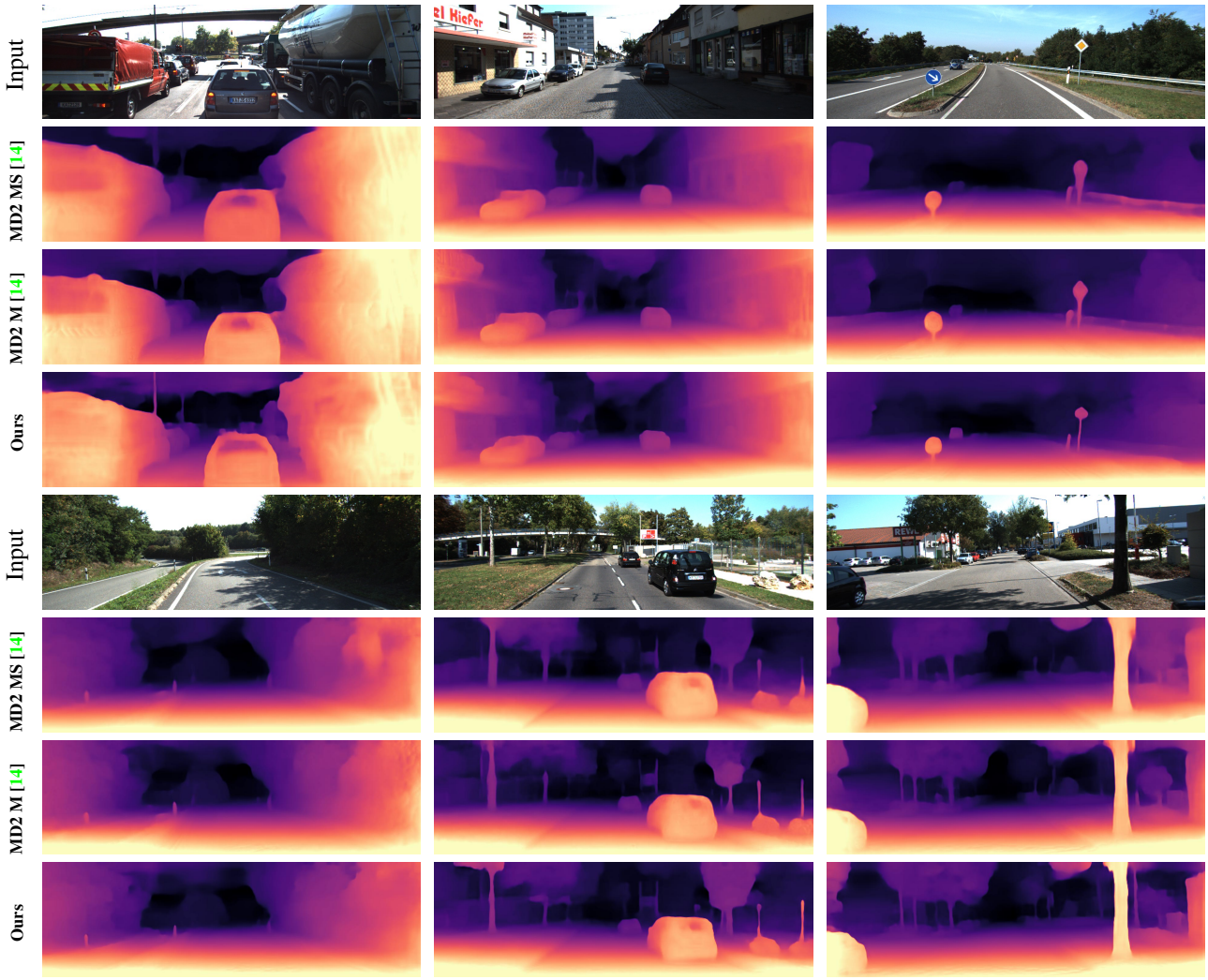

Figure 3. **Qualitative results on the KITTI Eigen split [8] test set.** Our models perform better on thinner objects such as trees, signs and bollards, as well as being better at delineating difficult object boundaries.

Backbone	Self-Attn	DDV	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Baseline (MD2 ResNet18)	✗	✗	0.115	0.903	4.863	0.193	0.877	0.959	0.981
ResNet18	✗	✓	0.112	0.838	4.795	0.191	0.877	0.960	0.981
ResNet18	✓	✗	0.112	0.845	4.769	0.19	0.877	0.96	**0.982**
ResNet18	✓	✓	0.111	0.941	4.817	0.189	0.885	0.961	0.981
ResNet101 w/ Dilated Conv	✗	✗	0.110	0.876	4.853	0.189	0.879	0.961	**0.982**
ResNet101 w/ Dilated Conv	✗	✓	0.110	0.840	4.765	0.189	0.882	0.961	**0.982**
ResNet101 w/ Dilated Conv	✓	✗	0.108	**0.808**	4.754	**0.185**	0.885	**0.962**	**0.982**
ResNet101 w/ Dilated Conv	✓	✓	**0.106**	0.861	**4.699**	**0.185**	**0.889**	**0.962**	**0.982**

Table 2. **Ablation Study.** Results for different versions of our model with comparison to our baseline model Monodepth2 [14](MD2 ResNet18). We evaluate the impact of the Discrete Disparity Volume (DDV), Self-Attention Context module and the larger network architecture. All models were trained with Monocular self-supervision. Metrics indicated by red: *lower is better*, Metrics indicated by blue: *higher is better*

a steady improvement of evaluation results in almost all evaluation measures. The DDV on the smaller ResNet-18 model provides a large improvement over the baseline in the *absolute relative* and *squared relative* measures. However, ResNet-101 shows only a small improvement over the baseline when using the DDV. The Self-Attention mechanism drastically improves the close range accuracy ($\delta < 1.25$) for both backbone models. The significantly larger improvement of the self-attention module in the ResNet-101 model (row 6), is likely because of the large receptive field produced by the dilated convolutions, which increases the amount of contextual information that can be computed by the self-attention operation.

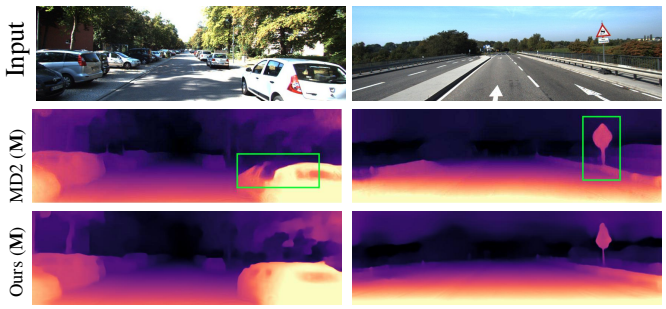

Figure 4. **Monodepth2 Failure cases.** Although trained on the same loss function as the monocular trained (**M**) Monodepth2 [14], our method succeeds in estimating depth for the reflective car roof *(Left)* and the difficult to delineate street sign *(Right)*.

	Type	Abs Rel	Sq Rel	RMSE	$\log_{10}$
Karsch [20]	D	0.428	5.079	8.389	0.149
Liu [31]	D	0.475	6.562	10.05	0.165
Laina [26]	D	**0.204**	**1.840**	**5.683**	**0.084**
Monodepth [13]	S	0.544	10.94	11.760	0.193
Zhou [62]	M	0.383	5.321	10.470	0.478
DDVO [53]	M	0.387	4.720	8.090	0.204
Monodepth2 [14]	M	0.322	3.589	7.417	0.163
Ours	M	**0.297**	**2.902**	**7.013**	**0.158**

Table 3. **Make3D results.** All self-supervised mono (M) models use median scaling.

4.5. Self-attention and Depth Uncertainty

While the self-attention module and DDV together provide significant quantitative and qualitative improvements, they also provide secondary functions. The attention maps (Eq. 3) from the self-attention module can be visualized to interrogate the relationships between objects and disparity learnt by the model. The attention maps highlight non-contiguous image regions (Fig. 5), focusing on either foreground, midground or background regions. The maps also tend to highlight either distant objects or stationary visual objects, like cars. Moreover, as the DDV encodes a probability over a disparity ray, using discretized bins, it is possible to compute the uncertainty for each ray by measuring the variance of the probability distribution. Figure 6 shows a trend where uncertainty increases with distance, up until the background image regions, which are estimated as near-infinite to infinite depth with very low uncertainty. This has also been observed in supervised models that are capable of estimating uncertainty [29]. Areas of high foliage and high shadow (row 2) show very high uncertainty, likely attributed to the low contrast and lack of textural detail in these regions.

5. Conclusion

In this paper we have presented a method to address the challenge of learning to predict accurate disparities solely from monocular video. By incorporating a self-attention mechanism to improve the contextual information available to the model, we have achieved state of the art results for monocular trained self-supervised depth estimation on the KITTI 2015 [12] dataset. Additionally, we regularised the training of the model by using a discrete disparity volume, which allows us to produce more robust and sharper depth estimates and to compute pixel-wise depth uncertainties. In the future, we plan to investigate the benefits of incorporating self-attention in the pose model as well as using the estimated uncertainties for outlier filtering and volumetric fusion.

Figure 5. **Attention maps from our network**. Subset of the attention maps produced by our method. Blue indicates region of attention.

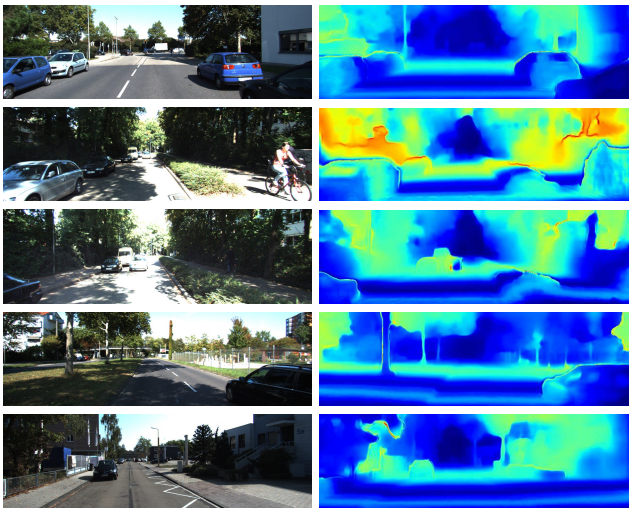

Figure 6. **Uncertainty from our network**. The Discrete Disparity Volume allows us to compute pixel-wise depth uncertainty. Blue indicates areas of low uncertainty, green/red regions indicate areas of high/highest uncertainty.

6. Acknowledgment

This research was in part supported by the Data to Decisions Cooperative Research Centre (A.J) and the Australian Research Council through grants DP180103232, CE140100016. G.C. acknowledges the support by the Alexander von Humboldt-Stiftung for the renewed research stay sponsorship.

References

[1] Markus Achtelik, Abraham Bachrach, Ruijie He, Samuel Prentice, and Nicholas Roy. Stereo vision and laser odometry for autonomous helicopters in gps-denied indoor environments. In *Unmanned Systems Technology XI*, volume 7332, page 733219. International Society for Optics and Photonics, 2009. 3

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3286–3295, 2019. 3

[3] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *AAAI*, 2019. 3, 5

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016. 3

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 6

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 3, 6

[7] Gregory Dudek and Michael Jenkin. *Computational principles of mobile robotics*. Cambridge university press, 2010. 3

[8] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015. 1, 3, 5, 6, 7

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 1, 3, 5

[10] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 1, 2, 3, 4, 5

[11] Ravi Garg, Vijay Kumar BG, and Ian Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016. 2, 3, 5, 6

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*, 2012. 5, 6, 8

[13] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 2, 3, 5, 6, 8

[14] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. *The International Conference on Computer Vision (ICCV)*, October 2019. 2, 3, 4, 5, 6, 7, 8

[15] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *ECCV*, 2018. 1, 3, 5

[16] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *European conference on computer vision*, pages 345–360. Springer, 2014. 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[19] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NeurIPS*, 2015. 3

[20] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *PAMI*, 2014. 1, 8

[21] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017. 2, 4

[22] 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 *ICCV*, 2017. 2, 4

[23] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 6

[24] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: learning SFM from SFM. In *ECCV*, 2018. 5

[25] Yevhen Kuznietsov, Jörg Stückler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *CVPR*, 2017. 5

[26] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016. 8

[27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. 3

[28] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. UnDeepVO: Monocular visual odometry through unsupervised deep learning. *arXiv*, 2017. 5

[29] Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G Narasimhan, and Jan Kautz. Neural rgb (r) d sensing: Depth and uncertainty from a video camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10986–10995, 2019. 2, 4, 8

[30] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *PAMI*, 2015. 5

[31] Miaomiao Liu, Mathieu Salzmann, and Xuming He. Discrete-continuous depth estimation from a single image. In *CVPR*, 2014. 8

[32] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3D holistic understanding. *arXiv*, 2018. 2, 3, 5

[33] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *CVPR*, 2018. 5

[34] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints. In *CVPR*, 2018. 5

[35] Nikolaus Mayer, Eddy Ilg, Philipp Fischer, Caner Hazirbas, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. What makes good synthetic training data for learning disparity and optical flow estimation? *IJCV*, 2018. 1, 3

[36] Ishit Mehta, Parikshit Sakurikar, and PJ Narayanan. Structured adversarial training for unsupervised monocular depth estimation. In *3DV*, 2018. 5

[37] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015. 3

[38] Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, and R. Venkatesh Babu. AdaDepth: Unsupervised content congruent adaptation for depth estimation. In *CVPR*, 2018. 5

[39] Niki Parmar, Prajit Ramachandran, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems*, pages 68–80, 2019. 3

[40] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NeurIPS-W*, 2017. 6

[41] Sudeep Pillai, Rares Ambrus, and Adrien Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. In *ICRA*, 2019. 3, 5

[42] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *3DV*, 2018. 5

[43] Anurag Ranjan, Varun Jampani, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *CVPR*, 2019. 5

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4

[45] Francisca Rosique, Pedro J Navarro, Carlos Fernández, and Antonio Padilla. A systematic review of perception system and simulators for autonomous vehicles research. *Sensors*, 19(3):648, 2019. 3

[46] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. In-place activated batchnorm for memory-optimized training of dnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5639–5647, 2018. 6

[47] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988. 3

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 6

[49] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168, 2006. 1

[50] Ashutosh Saxena, Min Sun, and Andrew Ng. Make3d: Learning 3d scene structure from a single still image. *PAMI*, 2009. 1, 3, 6

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 3

[52] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. SfM-Net: Learning of structure and motion from video. *arXiv*, 2017. 2, 3

[53] Chaoyang Wang, Jose Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, 2018. 5, 8

[54] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 2, 4

[55] Nan Yang, Rui Wang, Jörg Stückler, and Daniel Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *ECCV*, 2018. 5

[56] Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. LEGO: Learning edge with geometry all at once by watching videos. In *CVPR*, 2018. 5

[57] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry with edge-aware depth-normal consistency. In *AAAI*, 2018. 3, 5

[58] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018. 3, 5

[59] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018. 3, 6

[60] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, 2018. 3, 5

[61] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 3, 6

[62] Tinghui Zhou, Matthew Brown, Noah Snavely, and David Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 2, 3, 4, 5, 6, 8

[63] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency. In *ECCV*, 2018. 5