Bilateral Cyclic Constraint and Adaptive Regularization for Unsupervised Monocular Depth Prediction

Alex Wong, Stefano Soatto
UCLA Vision Lab
University of California, Los Angeles, CA 90095
{alexw, soatto}@cs.ucla.edu

Abstract

Supervised learning methods to infer (hypothesize) depth of a scene from a single image require costly per-pixel ground-truth. We follow a geometric approach that exploits abundant stereo imagery to learn a model to hypothesize scene structure without direct supervision. Although we train a network with stereo pairs, we only require a single image at test time to hypothesize disparity or depth. We propose a novel objective function that exploits the bilateral cyclic relationship between the left and right disparities and we introduce an adaptive regularization scheme that allows the network to handle both the co-visible and occluded regions in a stereo pair. This process ultimately produces a model to generate hypotheses for the 3-dimensional structure of the scene as viewed in a single image. When used to generate a single (most probable) estimate of depth, our method outperforms state-of-the-art unsupervised monocular depth prediction methods on the KITTI benchmarks. We show that our method generalizes well by applying our models trained on KITTI to the Make3d dataset.

1. Introduction

Estimating the 3-dimensional geometry of a scene is a fundamental problem in machine perception with a wide range of applications, including autonomous driving [24], robotics [32, 43], pose-estimation [41], localization [18], and scene object composition [17, 26]. It is well-known that 3-d scene geometry can be recovered from multiple images of a scene taken from different viewpoints, including stereo, under suitable conditions. Under no conditions, however, is a single image sufficient to recover 3-d scene structure, unless prior knowledge is available on the shape of objects populating the scene. Even in such cases, metric information is lost in the projection, so at best we can use a single image to generate hypotheses, as opposed to estimates, of scene geometry.

Recent works [3, 6, 31, 33, 34, 50, 51] sought to exploit such strong scene priors by using pixel-level depth annotation captured with a range sensor (e.g. depth camera, lidar) to regress depth from the RGB image. Cognizant of the intrinsic limitations of this endeavor, we exploit stereo imagery to train a network without ground-truth supervision for generating depth hypotheses, to be used as a reference for 3-d reconstruction. We evaluate our method against ground-truth depths via two benchmarks from the KITTI dataset [13] and show that it generalizes well by applying models trained on KITTI to Make3d [40].

Rather than attempting to learn a prior by associating the raw-pixel values with depth, we recast depth estimation as an image reconstruction problem [12, 14] and exploit the epipolar geometry between images in a rectified stereo pair to train a deep fully convolutional network. Our network learns to predict the dense pixel correspondences (disparity field) between the stereo pair, despite only having seen one of them. Hence, our network implicitly learns the relative pose of the cameras used in training and hallucinates the existence of a second image taken from the same relative pose when given a single image during testing. From the disparity predictions, we can synthesize depth using the known focal length and baseline of the cameras used in training.

While [12, 14, 49] follow a similar training scheme, [49] does not scale to high resolution, and [12] uses a non-differentiable objectives. [14] proposed using two uni-directional edge-aware disparity gradients and left-right disparity consistency as regularizers. However, edge-awareness should inform bidirectionally and left-right consistency suffers from occlusions and dis-occlusions. Moreover, regularity should not only be data-driven, but also model-driven.

Our contributions are three-fold: (i) A model-driven adaptive weighting scheme that is both space- and training-time varying and can be applied generically to regularizers. (ii) A bilateral consistency constraint that enforces the cyclic application of left and right disparity to be the identity. (iii) A two-branch decoder that specifically learns the features necessary to maximize data fidelity and utilizes such features to refine an initial prediction by enforcing regularity. We formulate our contributions as an objective function that, when realized even by a generic encoder-decoder, achieves state-of-the-art performance on two KITTI [13] benchmarks and exhibits generalizability to Make3d [40].

2. Related Works

Supervised Monocular Depth Estimation. [39] proposed a patch-based model that combined local estimates with Markov random fields (MRF) to obtain the global depth. Similarly, [20, 25, 29, 40] exploited local monocular features to make global predictions. However, local methods lack the global context needed to generate accurate depth estimates. [34] instead employed a convolutional neural network (CNN). [30] further improved monocular methods by incorporating semantic cues into their model.

[5, 6] introduced a two scale network. [31] proposed a residual network with up-sampling modules to produce higher resolution depth maps. [3] learned depth using crowd-sourced annotations and [10] learned the ordinal relations using atrous spatial pyramid pooling. [38] used image patches with neural forests. [27, 50, 51] used conditional random fields (CRF) jointly with a CNN.

Unsupervised Monocular Depth Estimation. Recently, [9] introduced novel view synthesis by predicting pixel values based on interpolation from nearby images. [49] minimized an image reconstruction loss to hallucinate the existence of a right view of a stereo pair given the left by producing the distribution of disparities for each pixel.

[12] trained a network for monocular depth prediction by reconstructing the right image of a stereo pair with the left and synthesizing disparity as an intermediate step. Yet, their image formation model is not fully differentiable, making their objective function difficult to optimize. Unsupervised methods [14, 37, 57, 58] utilized a bilinear sampler modeled after the Spatial Transformer Network [23] to allow for a fully differentiable loss and end-to-end training of their respective networks. Specifically, [14] used SSIM [46] as a loss in addition to the image reconstruction loss. Also, [14] predicted both left and right disparities and used them for regularization via a left-right consistency check along with an edge-aware smoothness term. [2] trains a Generative Adversarial Network (GAN) [15] to constrain the output to reconstruct a realistic image to reduce the artifacts seen from stereo reconstruction. This class of method is also employed in depth completion [54].

Self-supervised methods [35, 44, 56, 59] used a pose network to learn ego-motion and depth from monocular videos, while [45, 52] leveraged visual odometry from off-the-shelf methods [7, 42] and [8] gravity as supervisors. [55] followed both unsupervised and self-supervised paradigms by using stereo video streams and proposed a feature reconstruction loss. While additional supervision and data are used to improve predictions, [14] still remains as the state-of-the-art in the unsupervised setting. Our method follows the unsupervised paradigm and we show that it not only outperforms [14], but also [55] who leveraged techniques from both unsupervised and self-supervised domains.

Adaptive Regularization. A number of computer vision problems can be formulated as energy minimization in a variational framework with a data fidelity term and a regularizer weighted by a fixed scalar. The solution found by the minimal energy involves a trade-off between data fidelity and regularization. Finding the optimal parameter for regularity is a long studied problem as [11] explored methods to determine the regularization parameter in image de-noising, while [36] used cross-validation as a selection criterion for the weight. [14, 47, 48] used image gradients as cues for a data-driven weighting scheme. [53] learned regularity conditioned on an image. Recently, [21, 22] proposed that regularity should not only be data-driven, but also model driven. The amount of regularity imposed should adapt to the fitness of the model in relation to the data rather than being constant throughout the training process.

We propose a novel objective function using bilateral cyclic consistency constraint along with a spatial and temporal varying regularization modulator. We show that despite using the fewer parameters than [14], we outperform [14] and other unsupervised methods. We detail our loss function with adaptive regularization, in Sec. 3, present a two-branch decoder architecture in Sec. 4, and specify hyper-parameters and data augmentation procedures used in Sec. 5. We evaluate our model on the KITTI 2015, KITTI Eigen Split, and Make3d benchmarks in Sec. 6. Lastly, we end with a discussion of our work in Sec. 7.

3. Method Formulation

We learn a model to hypothesize or "estimate" the disparity field d compatible with an image I^0 by exploiting the availability of stereo pairs (I^0, I^1) during training. We then synthesize the depth $z = FB/d$ of the scene using the focal length F and baseline B during test time. Given I^0, we estimate a function $d \in \mathbb{R}_+$ that represents the disparity of I^0, which we formulate as a loss function L (Eqn. 1), comprised of data terms and adaptive regularizers.

Our network, parameterized by ω, takes a single image I^0 as input and estimates a function $d = f(I^0; \omega)$, where d represents the disparity (which is monotonically related to inverse-depth) corresponding to I^0. We drive the training process with I^1, which is only used in the loss function, by a surrogate loss that minimizes the reprojection error of I^0 to I^1 and vice versa. We will refer to the disparity estimated by L as d^0 and d^1 for I^0 and I^1, respectively. Interested readers may refer to Supplementary Materials (Supp. Mat.) for more details on our formulation.

$$L = \underbrace{w_{ph}l_{ph} + w_{st}l_{st}}_{\text{data fidelity}} + \underbrace{w_{sm}l_{sm} + w_{bc}l_{bc}}_{\text{regularization}} \quad (1)$$

where each individual term l will be described in the next sections and their weights w in Sec. 5.

3.1. Data Fidelity

Our data fidelity terms seek to minimize the discrepancy between the observed stereo pair (I^0, I^1) and their reconstructions $(\hat{I}^0, \hat{I}^1)$. We generate each $\hat{I}$ term by applying a 1-d horizontal disparity shift to I at each position (x, y):

$$\hat{I}^0{}_{xy} = I^1_{xy-d^0_{xy}} \text{ and } \hat{I}^1{}_{xy} = I^0_{xy+d^1_{xy}} \tag{2}$$

We do so by using a 1-d horizontal bilinear sampler modeled after the image sampler from the Spatial Transformer Network [23] – instead of applying an affine transformation to activations, we warp an image to the domain of its stereo-counterpart using disparities. Our sampler is locally fully differentiable and each output pixel is the weighted sum of two (left and right) pixels. We propose to minimize the reprojection residuals as a two-part loss, which measures the standard color constancy (photometric) and the difference in illumination, contrast and image quality (structural).

Photometric loss. We model the image formation process via a photometric loss l_{ph}, which measures the $L1$ penalty of the reprojection residual for each I and $\hat{I}$ on each channel at every (x, y) position in the image space Ω:

$$l_{ph} = \sum_{(x,y)\in\Omega} |I^0_{xy} - \hat{I}^0_{xy}| + |I^1_{xy} - \hat{I}^1_{xy}| \tag{3}$$

Structural loss. In order to make inference invariant to local illumination changes, we use a perceptual metric (SSIM) that discounts such variability. We apply SSIM (ϕ) to image patches of size 3×3 at corresponding (x, y) in I and $\hat{I}$. Since two similar images give a SSIM score close to 1, we subtract 1 by the score to represent a distance:

$$l_{st} = \sum_{(x,y)\in\Omega} 2 - (\phi(I^0_{xy}, \hat{I}^0_{xy}) + \phi(I^1_{xy}, \hat{I}^1_{xy})) \tag{4}$$

3.2. Residual-Based Adaptive Weighting Scheme

A point estimate d can be obtained by maximizing the Bayesian criterion with a data fidelity term (energy) $\mathcal{D}(d)$ and a Bayesian or Tikhonov regularizer $\mathcal{R}(d)$ in the form:

$$\mathcal{D}(d) + \alpha\mathcal{R}(d) \tag{5}$$

where the weight α is a pre-defined positive scalar parameter that controls the regularity to impose on the model, leading to a trade-off between data fidelity and regularization.

The weight α modulates between data-fidelity and regularization, constraining the solution space. Yet, subjecting the entire solution, a dense disparity field, to the same regularity fails to address cases where the assumptions do not hold. Suppose one enforces a smoothness constraint to the output disparity field by simply taking the disparity gradient ∇d. This constraint would incorrectly penalize object boundaries (regions of high image gradients) and hence [14, 19] apply an edge-aware term to reduce the effects of regularization on edge regions. Although the edge-awareness term gives a data-driven approach on regularization, it is still static (the same image will always have the same weights) and independent of the performance of the model. Instead, we propose a space- and training-time varying weighting scheme based on the performance of our model measured by reprojection residuals.

Model-driven adaptive weight. We propose an adaptive weight α_{xy} that varies in space and training time for every position (x, y) of the solution based on the local residual $\rho_{xy} = |I_{xy} - \hat{I}_{xy}|$ and the global residual, represented by the average per-pixel residual, $\sigma = \dfrac{1}{\frac{1}{|\Omega|}\sum\limits_{(x,y)\in\Omega} |I_{xy} - \hat{I}_{xy}|}$:

$$\alpha_{xy} = \exp\left(-\frac{c\rho_{xy}}{\sigma}\right) \tag{6}$$

α is controlled by the local residual between an image I and its reprojection $\hat{I}$ at each position while taking into account of the global residual σ, which correlates to the training time step and decreases over time. c is a scale factor for the range of α. α is naturally small when residuals are large and tends to 1 as training converges.

Local adaptation. Consider a pair of poorly matched pixels, $(I_{xy}, \hat{I}_{xy})$, where the residual $|I_{xy} - \hat{I}_{xy}|$ is large. By reducing the regularity on the solution d_{xy}, we effectively allow for exploration in the solution space to find a better match and hence a d_{xy} that minimizes the data fidelity terms. Alternatively, consider a pair of perfectly matched pixels, $(I_{xy}, \hat{I}_{xy})$, where $|I_{xy} - \hat{I}_{xy}| = 0$. We should apply regularization to decrease the scope of the solution space such that we can allow for convergence and propagate the solution. Hence, a spatially adaptive α_{xy} must vary inversely to the local residual ρ_{xy} such that we impose regularization when the residual is small and reduce it when the residual is large.

Global adaptation. Consider a solution d_{xy} proposed at the first training time step $t = 1$. Imposing regularity effectively reduces the solution space based on an assumption about d_{xy} and biases the final solution. We propose that a weighting scheme $\alpha_{xy} \to 1$ as $t \to \infty$. However, if α_{xy} is directly dependent on the t, then α_{xy} will change if we continue to train even after convergence – causing the model to be unstable. Instead, let α_{xy} be inversely proportional to the global residual σ such that α_{xy} is small when the σ is large (generally corresponding to early time steps) and $\alpha_{xy} \to 1$ as $\sigma \to 0$. When training converges (i.e. the global residual has stabilized), α_{xy} likewise will be stable. This naturally lends to an annealing schedule where $\alpha_{xy} \to 1$ as time progresses in training steps.

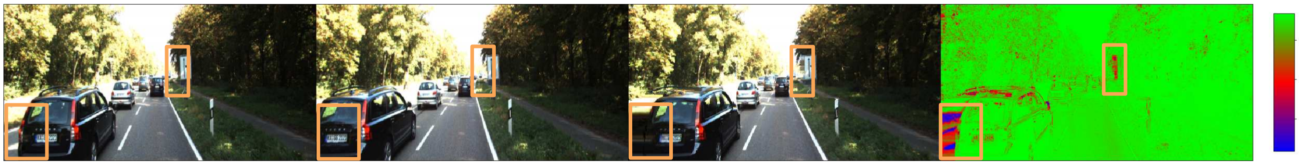


Figure 1: Left to right: left image, right image, left reconstruction, adaptive weights. The adaptive weights reduce regularization at regions of high residual; hence, they discount dis-occlusions and occlusions as in the highlighted regions.

3.3. Adaptive Regularization

Our regularizers assume local smoothness and consistency between the left and right disparities estimated. We propose to minimize the disparity gradient (smoothness) and the disparity reprojection error (bilateral cyclic consistency) while adaptively weighting both with α (Sec. 3.2).

Smoothness loss. We encourage the predicted disparities to be locally smooth by applying an $L1$ penalty to the disparity gradients in the x (∂_X) and y (∂_Y) directions. However, such an assumption does not hold at object boundaries, which generally correspond to regions of high changes in pixel intensities; hence, we include an edge-aware term λ to allow for discontinuities in the disparity gradient. We also weigh this term adaptively with α:

$$l_{sm} = \sum_{(x,y)\in\Omega} \alpha^0_{xy}(\lambda^0_{xy}|\partial_X d^0_{xy}| + \lambda^0_{xy}|\partial_Y d^0_{xy}|) + \alpha^1_{xy}(\lambda^1_{xy}|\partial_X d^1_{xy}| + \lambda^1_{xy}|\partial_Y d^1_{xy}|) \tag{7}$$

where $\lambda_{xy} = e^{-|\nabla^2 I_{xy}|}$ and the ∇^2 operator denotes the image Laplacian. We use the image Laplacian over the first order image gradients because it allows the disparity gradients to be aware of intensity changes in both directions. However, we regularize the disparity field using the disparity gradient so that we can allow for independent movement in each direction. Prior to computing the image Laplacian for λ, we smooth the image with a Gaussian kernel to reduce noise.

Bilateral cyclic consistency loss. A common regularization technique in stereo-vision is to maintain the consistency between the left (d^0) and right (d^1) disparities by reconstructing each disparity through projecting its counterpart with its disparity shifts:

$$d^{0p}_{xy} = d^1_{xy-d^0_{xy}} \text{ and } d^{1p}_{xy} = d^0_{xy+d^1_{xy}} \tag{8}$$

However, in doing so, the projected disparities suffer from the unresolved correspondences of both the disparity ramps, occlusions and dis-occlusions. We, propose a bilateral cyclic consistency check that is designed to specifically reason about occlusions while removing the effects of stereo dis-occlusions. We follow the intuition that the disparities d should have an identity mapping when projected to the domain of its stereo-counterpart and back-projected to the original domain as a reconstruction $\hat{d}$ so reconstruction of dis-occlusion is ignored.

$$\hat{d}^0_{xy} = d^0_{xy+d^1_{xy}-d^0_{xy}} \text{ and } \hat{d}^1_{xy} = d^1_{xy-d^0_{xy}+d^1_{xy}} \tag{9}$$

By applying an $L1$ penalty on the disparity field and its reconstruction, we are constraining that the cyclic transformations should be the identity transform, which keeps d^0 and d^1 consistent with each other in co-visible regions. If there exists an occluded region, the region in the reconstruction would be inconsistent with the original – yielding reprojection error. To avoid penalizing a model for an unresolvable correspondence due to the nature of the data, we propose to adaptively regularize the bilateral cyclic constraint using our residual-based weighting scheme (Eqn. 6). Unsurprisingly, local regions of high reprojection residual often correspond to occluded regions.

$$l_{bc} = \sum_{(x,y)\in\Omega} \alpha^0_{xy}|d^0_{xy} - \hat{d}^0_{xy}| + \alpha^1_{xy}|d^1_{xy} - \hat{d}^1_{xy}| \tag{10}$$

4. A Two-Branch Decoder

As our adaptive weighting scheme (Sec. 3.2) is function of the data fidelity residuals, we seek to ensure that the network learns a sufficient representation to minimize the data fidelity loss (Sec. 3.1). We propose a two-branch decoder (Fig. 2) with one branch (prefixed with 'i') dedicated to learning the features, `iconv`, necessary to make a prediction that minimizes data fidelity loss:

$$L^0 = w_{ph} l_{ph} + w_{st} l_{st} \tag{11}$$

using the reconstructed features via up-convolution and the corresponding `skip` connection from the encoder. We use a residual block [16] to learn the skip connection residual, `rskip`, necessary to minimize Eqn. 1 – both data fidelity and regularity loss. By concatenating `iconv` and `rskip` with the initial prediction (`idisp`) as features for the second branch (prefixed with 'r'), we have provided the decoder branch with a prediction that satisfies data fidelity along with features necessary to impose regularity. The branch can now utilize such information to refine the initial prediction by adaptively applying regularization based on the data fidelity residual. To maintain a similar network size and run-time, we reduce the depth of the network by 1 and added a single convolution as the first layer to enable a skip connection to the last layer. This, in fact, resulted

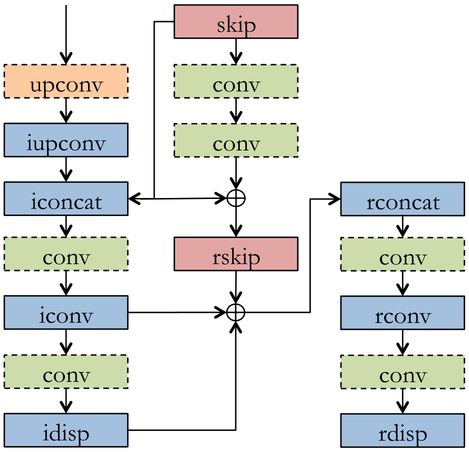


Figure 2: Two-branch decoder. `idisp` produces an initial prediction based only on the data terms and `rdisp` produces a refined prediction using the entire loss function (Eqn. 1). By minimizing just the data terms (Eqn. 11) in `idisp`, we force `iconv` to learn sufficient information for the reconstruction task such that `rdisp` can utilize such features along with the residual learned from the skip connection to refine a prediction that satisfies data fidelity by imposing regularity based on the data fidelity residual.

in our network having $\approx$ 10 million fewer parameters than [14]. We show qualitative results in Fig. 3 and 4 where we observe the benefits of learning the features that satisfy data fidelity as we recover more details about the scene geometry. Quantitatively, we show in Table 2 and 3 that this structure improves over the state-of-the-art performance on all metrics achieved by our generic encoder with a single branch decoder, where the final predictions of both decoders minimize our objective function (Eqn. 1).

5. Implementation Details

Our approach was implemented using TensorFlow [1]. There are $\approx$ 31 million trainable parameters in the generic encoder-decoder [14] and $\approx$ 21 million in our proposed structure (more details can be found in Supp. Mat. Table 2 and 3). Training takes $\approx$ 18 hours using an Nvidia GTX 1080Ti. Inference takes $\approx$ 32 ms per image. We used Adam [28] to optimize our network with a base learning rate of 1.8×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$. We then increase the learning rate to 2×10^{-4} after 1 epoch, decrease it by half after 46 epochs and by a quarter after 48 epochs for a total of 50 epochs. We use a batch size of 8 with a 512×256 resolution and 4 levels in our loss pyramid. We are able to achieve our results using the following set of weights for each term in our loss function: $w_{ph} = 0.15$, $w_{st} = 0.425$, $w_{sm} = 0.10$ and $w_{bc} = 1.05$. We choose the scale factor $c = 5.0$ for the adaptive weight α. For our smoothness term, we decrease it by a factor of 2^r for each r-th resolution in the loss pyramid where $r = 0$ refers to our highest resolution at 512×256 and $r = 3$ the lowest.

Metric	Definition
AbsRel	$\frac{1}{\lvert\Omega\rvert} \sum_{(x,y)\in\Omega} \frac{\lvert z_{xy} - z^{\text{gt}}_{xy}\rvert}{z^{\text{gt}}_{xy}}$
SqRel	$\frac{1}{\lvert\Omega\rvert} \sum_{(x,y)\in\Omega} \frac{\lvert z_{xy} - z^{\text{gt}}_{xy}\rvert^2}{z^{\text{gt}}_{xy}}$
RMS	$\sqrt{\frac{1}{\lvert\Omega\rvert} \sum_{(x,y)\in\Omega} \lvert z_{xy} - z^{\text{gt}}_{xy}\rvert^2}$
logRMS	$\sqrt{\frac{1}{\lvert\Omega\rvert} \sum_{(x,y)\in\Omega} \lvert \log z_{xy} - \log z^{\text{gt}}_{xy}\rvert^2}$
$\log_{10}$	$\frac{1}{\lvert\Omega\rvert} \sum_{(x,y)\in\Omega} \lvert \log z_{xy} - \log z^{\text{gt}}_{xy}\rvert$
Accuracy	% of z_{xy} s.t. $\delta \doteq \max\left(\frac{z_{xy}}{z^{\text{gt}}_{xy}}, \frac{z^{\text{gt}}_{xy}}{z_{xy}}\right) <$ threshold

Table 1: Error and accuracy metrics. z_{xy} is the predicted depth at $(x, y) \in \Omega$ and z^{gt}_{xy} is the corresponding ground truth. Three different thresholds $(1.25, 1.25^2$ and $1.25^3)$ are used in the accuracy metric as a convention in the literature.

Data augmentation is performed online during training. We perform a horizontal flip (with a swap to maintain correct relative positions) on the stereo pairs with 50% probability. Color augmentations on brightness, gamma and color shifts of each channel also occur with 50% chance. We uniformly sample from $[0.5, 1.5]$ for brightness, and $[0.8, 1.2]$ for gamma and each color channel separately.

6. Experiments and Results

We present our results on the KITTI dataset [13] under two different training and testing schemes, the KITTI 2015 split [14] and the KITTI Eigen split [6, 12]. The KITTI dataset contains 42,382 rectified stereo pairs from 61 scenes with approximate resolutions of 1242×375. We evaluate our method on the monocular depth estimation task on KITTI Eigen split and compare our approach with similar variants on a disparity error metric as an ablation study using the KITTI 2015 split. We show that our method outperforms state-of-the-art unsupervised monocular approaches and even supervised approaches on KITTI benchmarks, while generalizing to Make3d [40].

6.1. KITTI Eigen Split

We evaluate our method using the KITTI Eigen split [6], which has 697 test images from 29 scenes. The remaining 32 scenes contain 23,488 stereo pairs, of which 22,600 pairs are used for training and the rest for validation, following [12]. We project the velodyne points into the left input color camera frame to generate ground-truth depths. The ground-truth depth maps are sparse ($\approx$ 5% of the entire image) and prone to errors from rotation of the velodyne and motion of the vehicle and surrounding objects along with occlusions. As a result, we use the cropping scheme proposed by [12], which contains approximately 58% in height and 93% in width of the image dimensions.

We compare our approach with the recent monocular depth estimation methods at 80 and 50 meters caps in Ta-

			Error Metrics				Accuracy Metrics		
Method	Dataset	Cap	Abs Rel	Sq Rel	RMS	logRMS	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou et al. [56]	K	80m	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Mahjourian et al. [35]	K	80m	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Garg et al. [12]	K	80m	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Godard et al. [14]	K	80m	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Zhan et al. [55] (w/ video)	K	80m	0.144	1.391	5.869	0.241	0.803	0.928	**0.969**
Ours (Full Model)	K	80m	0.135	1.157	5.556	0.234	0.820	0.932	0.968
Ours (Full Model)*	K	80m	**0.133**	**1.126**	**5.515**	**0.231**	**0.826**	**0.934**	**0.969**
Zhou et al. [56]	CS+K	80m	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Mahjourian et al. [35]	CS+K	80m	0.159	1.231	5.912	0.243	0.784	0.923	0.970
Godard et al. [14]	CS+K	80m	0.124	1.076	5.311	0.219	0.847	0.942	0.973
Ours (Full Model)*	CS+K	80m	**0.118**	**0.996**	**5.134**	**0.215**	**0.849**	**0.945**	**0.975**
Zhou et al. [56]	K	50m	0.201	1.391	5.181	0.264	0.696	0.900	0.966
Garg et al. [12]	K	50m	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Godard et al. [14]	K	50m	0.140	0.976	4.471	0.232	0.818	0.931	0.969
Zhan et al. [55] (w/ video)	K	50m	0.135	0.905	4.366	0.225	0.818	0.937	**0.973**
Ours (Full Model)	K	50m	0.128	0.856	4.201	0.220	0.835	0.939	0.972
Ours (Full Model)*	K	50m	**0.126**	**0.832**	**4.172**	**0.217**	**0.840**	**0.941**	**0.973**

Table 2: Quantitative results[1] on the KITTI [13] Eigen split [6] benchmark. Depths are capped at 50 and 80 meters. K denotes training on KITTI. CS+K denotes pretraining on Cityscape [4] and fine-tuning on KITTI. Our full model using a generic encoder-decoder consistently outperforms other methods in all metrics across both depth caps with the exception of $\delta < 1.25^3$ where [55], which used temporal information (sequences of stereo-pairs), marginally beats our us by 0.1%. Our proposed decoder (*) improves over our encoder-decoder model across all metrics and is the state-of-the-art.

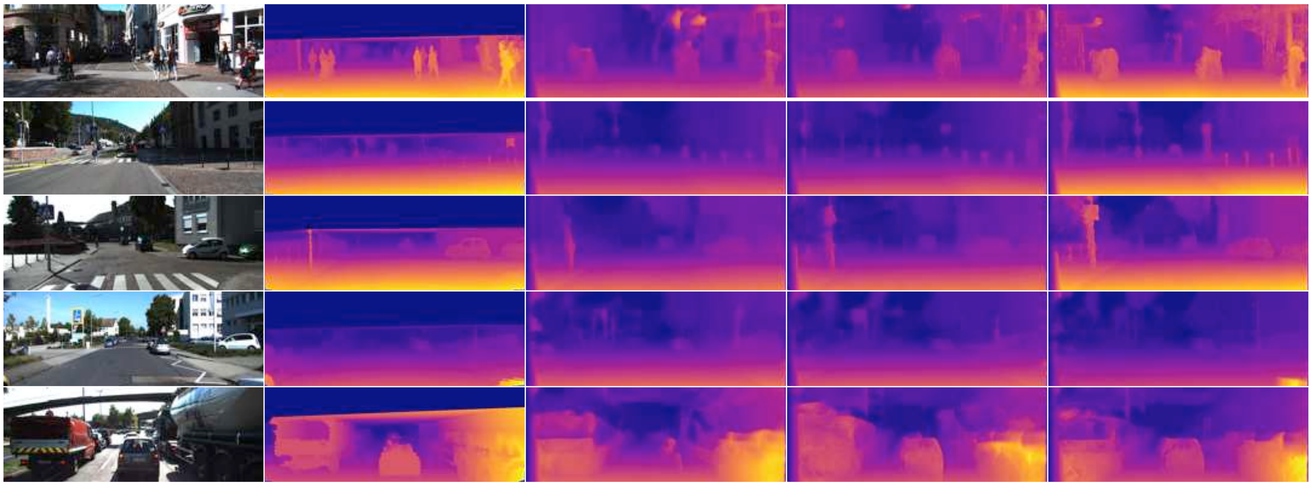


Figure 3: Qualitative results on KITTI Eigen split. From left to right: input images, ground-truth disparities, results of Godard et al. [14], our results with a generic decoder and our results with the proposed decoder. Our method under both decoders recovers more scene structures (row 2, 3: street signs, row 5: car in middle). Moreover, the predictions of the proposed two-branch structure are more realistic (row 1: pedestrian on right, row 4: tail of another car at bottom right corner, row 5: hollow trunk of truck on left, where both [14] and the generic decoder predicted as a surface).

ble 2. Fig. 3 provides a qualitative comparison between our method and the baseline. We note that [55] trained two networks using stereo video streams (as opposed to a single network with stereo pairs like ours and [14]), which allows their networks to learn a depth prior in both spatial and temporal domains. Using the network of [14] (generic encoder with a single branch decoder), we outperform all competing methods in all metrics under both depth caps except for $\delta < 1.25^3$ where we are comparable to [55]. We improve consistently over [14] and [55] by an average of

Method	Error Metrics					Accuracy Metrics		
	Abs Rel	Sq Rel	RMS	logRMS	D1-all	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
[14] w/ Deep3D [49]	0.412	16.37	13.693	0.512	66.850	0.690	0.833	0.891
[14] w/ Deep3Ds [49]	0.151	1.312	6.344	0.239	59.640	0.781	0.931	0.976
$ph + st + \lambda^G sm$ ([14] w/o Left-Right Consistency)	0.123	1.417	6.315	0.220	30.318	0.841	0.937	0.973
$ph + st + \lambda^G sm + lr$ [14]	0.124	1.388	6.125	0.217	30.272	0.841	0.936	0.975
$ph + st + \alpha\lambda^G sm + \alpha lr$ ([14] w/ Our Adaptive Regularization)	0.120	1.367	6.013	0.211	30.132	0.849	0.942	0.975
Aleotti et al. [2]	0.119	1.239	5.998	0.212	29.864	0.846	0.940	0.976
$ph + st + \lambda^L sm + bc$ (Ours w/o Adaptive Regularization)	0.117	1.264	5.874	0.207	29.793	0.851	0.944	0.977
$ph + st + \alpha\lambda^L sm + \alpha lr$ (Ours w/o Bilateral Cyclic Consistency)	0.117	1.251	5.876	0.206	29.536	0.851	0.944	0.977
$ph + st + \alpha\lambda^G sm + \alpha bc$ (Ours w/o Bidirectional Edge-Awareness)	0.115	1.211	5.743	0.203	28.942	0.852	0.945	0.977
$ph + st + \alpha\lambda^L sm + \alpha bc$ (Ours Full Model)	0.114	1.172	5.651	0.202	28.142	0.855	**0.947**	0.979
$ph + st + \alpha\lambda^L sm + \alpha bc$ * (Ours Full Model w/ 2 Branch Decoder)	**0.110**	**1.119**	**5.576**	**0.200**	**27.149**	**0.856**	**0.947**	**0.980**

Table 3: Quantitative comparison[1] amongst variants of our model on KITTI 2015 split proposed by [14]. Each variant is named according to its loss function. ph and st denote data terms, sm local smoothness, α our adaptive weights, λ^G image gradients [14], λ^L image Laplacian, lr left-right consistency [14], and bc our bilateral cyclic consistency. We show the effectiveness of our adaptive regularization (Sec. 3.3) by applying it to [14] and improving their model. Our full model using a generic encoder-decoder outperforms all variants on every metric, including [2] which predicts disparities that generate photo-realistic images. Our full model using our proposed two-branch decoder (*) further improves the state-of-the-art.

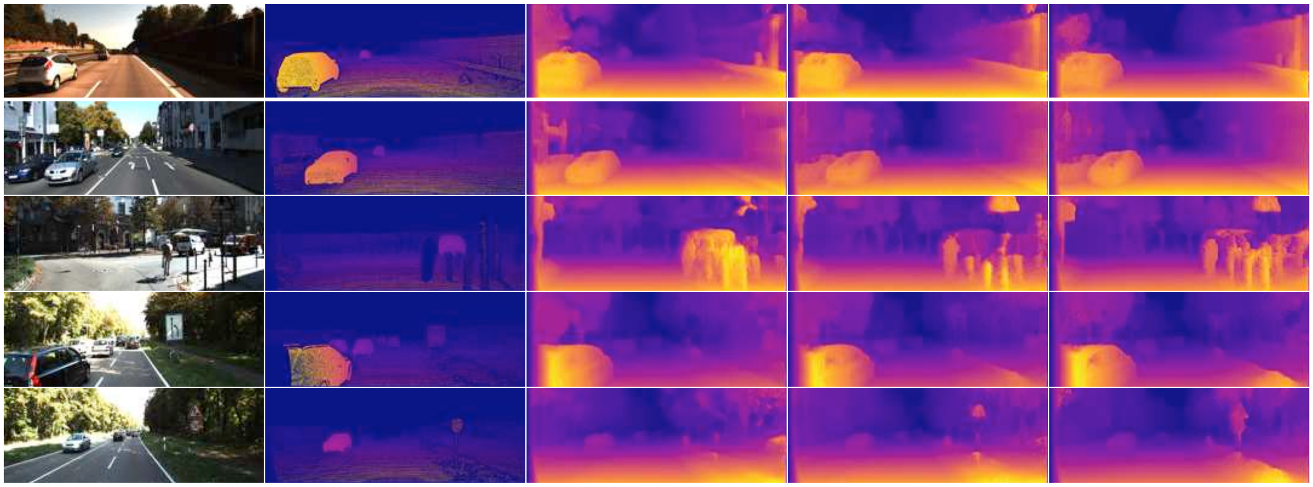
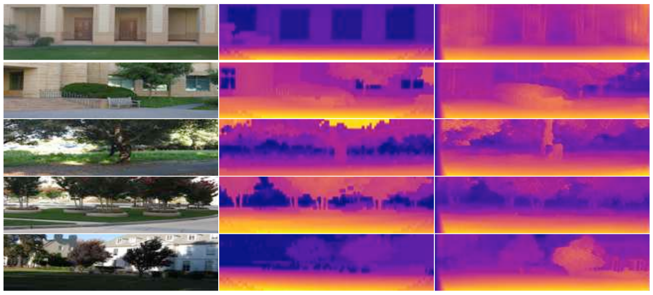


Figure 4: Qualitative results on KITTI 2015 split. From left to right: input images, ground-truth depths, results of Godard et al.[14], our results using a generic decoder and our results the proposed decoder. Our approach generates more consistent depths (row 1: walls on right, row 2: building on left) and recovers more detailed structures (row 3: biker and poles on right, rows 4, 5: street signs), with the two-branch decoder recovering the most.

8.7% and 5.75% in AbsRel, 13.1% and 10.5% in SqRel and even 5.25% and 2.55% in logRMS, respectively. Furthermore, we score significantly higher in $\delta < 1.25$ (the hardest accuracy metric), which suggests that our model produces more correct and realistically detailed depths than all competing methods. In addition, our two-branch decoder improves over the said results across all metrics and depth caps and is the current state-of-the-art. Table 2 shows that our model also beats [14] when pretraining on Cityscape [4] and fine-tuning on KITTI. An ablation study on Eigen Split examining the effects of each of our contributions (Sec. 3.3) can be found in our Supp. Mat.

6.2. KITTI 2015 Split

We evaluate our method on 200 high quality disparity maps provided as part of the official KITTI training set [13]. These 200 stereo pairs cover 28 of the total 61 scenes. From 30,159 stereo pairs covering the remaining 33 scenes, we choose 29,000 for training and the rest for validation. While typical training and evaluation schemes project velodyne laser values to depth, we choose to use the provided disparity maps as they are less erroneous than velodyne data points. In addition, we also use the official KITTI dispar-



(a)

Method	Supervised	Error Metrics			
		AbsRel	Sq Rel	RMS	$\log_{10}$
Karsch et al. [25]	Yes	0.417	4.894	8.172	0.144
Liu et al. [34]	Yes	0.462	6.625	9.972	0.161
Laina et al. [31]	Yes	**0.198**	**1.665**	**5.461**	**0.082**
Godard et al. [14]	No	0.468	9.236	12.525	0.165
Ours	No	0.454	8.470	12.211	0.163
Ours*	No	**0.427**	**8.183**	**11.781**	**0.156**

(b)

Figure 5: Qualitative (a) and quantitative (b) results[1] on Make3d [40] with maximum depth of 70 meters. In (a), top to bottom: input images, ground-truth disparities, our results. In (b), unsupervised methods listed are all trained on KITTI Eigen split. Despite being trained on KITTI, we perform comparably to a number of supervised methods trained on Make3d.

ity metric of end-point-error (D1-all) to measure our performance as it is a more appropriate metric on our class of approach that outputs disparity and synthesizes depth from the output using camera focal length and baseline.

We show qualitative comparisons in Fig. 4 and quantitative comparisons in Table 3. Table 3 also serves as an ablation study on variants belonging to the stereo unsupervised paradigm using different image formation model and regularization terms. We show that by simply applying our adaptive regularization to [14], we achieve improvement over their model. We also study the effects of substituting our bilateral cyclic consistency with the left-right consistency regularizer [14]. We also substitute image Laplacian with image gradients for edge-aware weights. In addition, we find that adaptive regularization and bilateral cyclic consistency contribute similarly to the improvements of the models. However, when combined they achieve significantly improvements over the baseline method (and all variants) in every metric. Furthermore, when using our proposed decoder, we again surpass all variants on every metric. We additionally outperform [2], who uses a GAN to constrain the output disparities to produce photo-realistic images during reconstruction. This result aligns with our performance on accuracy metrics – our method produces accurate and realistic depths.

6.3. Generalizing to Different Datasets: Make3d

To show that our model generalizes, we present our qualitative and quantitative results in Fig. 5 on the Make3d dataset [40] containing 134 test images with 2272 × 1707 resolution. Make3d provides range maps (resolution of 305 × 55) for ground-truth depths, which must be rescaled and interpolated. We use the central cropping proposed by [14] where we generate a 852 × 1707 crop centered on the image. We use the standard $C1$ evaluation metrics[1] proposed for Make3d and limit the maximum depth to 70 meters. The results of the supervised methods are taken from [14]. Because Make3d does not provide stereo pairs, we are unable to train on it. However, we find that despite having trained our model on KITTI Eigen split, our performance is comparable to that of supervised methods trained on Make3d and is better than the baseline across all metrics.

7. Discussion

In this work, we proposed an adaptive weighting scheme (Sec. 3.3) that is both spatially and time varying, allowing for not only a data-driven, but also model-driven approach to regularization. Moreover, we introduce a bilateral cyclic consistency constraint that not only enforces consistency between the left and right disparities, but also removes stereo dis-occlusions while discounting unresolved occlusions when combined with our weighting scheme. Finally, we propose a two-branch decoder that achieves the state-of-the-art by learning features to improve data residual for imposing our adaptive regularity. We achieve state-of-the-art performance on two KITTI benchmarks and show that our method generalizes to Make3d. Our two-branch decoder further improves over those results. Our experiments (Table 2 and 3) show that our approach produces depth maps with more details while maintaining global correctness.

For future work, we plan to improve robustness to specular and transparent surfaces as these regions tend to produce inconsistent depths. We are also exploring more sophisticated regularizers in place of the simple disparity gradient. Finally, we believe that the task should drive the network architecture. Rather than using a generic network, finding a better architectural fit could prove to be ground-breaking and further push the state-of-the-art.

Acknowledgements. Byung-Woo Hong (*hong@cau.ac.kr*; Chung-Ang University, Korea) provided mentorship and helpful discussions, which we believe warrant authorship. Here, we recognize him as an author (as listed in other versions of this paper) and thank him for his contributions. This work was supported by NRF-2017R1A2B4006023, NRF-2018R1A4A1059731, ONR N00014-17-1-2072, ARO W911NF-17-1-0304.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016. 5

[2] F. Aleotti, F. Tosi, M. Poggi, and S. Mattoccia. Generative adversarial networks for unsupervised monocular depth prediction. In *15th European Conference on Computer Vision (ECCV) Workshops*, 2018. 2, 7, 8

[3] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. In *Advances in Neural Information Processing Systems*, pages 730–738, 2016. 1, 2

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6, 7

[5] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015. 2

[6] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 1, 2, 5, 6

[7] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2018. 2

[8] X. Fei, A. Wong, and S. Soatto. Geo-supervised visual depth prediction. *arXiv preprint arXiv:1807.11130*, 2018. 2

[9] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2016. 2

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

[11] N. P. Galatsanos and A. K. Katsaggelos. Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation. *IEEE Transactions on image processing*, 1(3):322–336, 1992. 2

[12] R. Garg, V. K. BG, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016. 1, 2, 5, 6

[13] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361. IEEE, 2012. 1, 2, 5, 6, 7

[14] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017. 1, 2, 3, 5, 6, 7, 8

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[17] T. He, H. Huang, L. Yi, Y. Zhou, C. Wu, J. Wang, and S. Soatto. Geonet: Deep geodesic networks for point cloud analysis. *arXiv preprint arXiv:1901.00680*, 2019. 1

[18] T. He and S. Soatto. Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors. *arXiv preprint arXiv:1901.03446*, 2019. 1

[19] P. Heise, S. Klose, B. Jensen, and A. Knoll. Pm-huber: Patchmatch with huber regularization for stereo matching. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2360–2367. IEEE, 2013. 3

[20] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007. 2

[21] B.-W. Hong, J.-K. Koo, M. Burger, and S. Soatto. Adaptive regularization of some inverse problems in image analysis. *arXiv preprint arXiv:1705.03350*, 2017. 2

[22] B.-W. Hong, J.-K. Koo, H. Dirks, and M. Burger. Adaptive regularization in convex composite optimization for variational imaging problems. In *German Conference on Pattern Recognition*, pages 268–280. Springer, 2017. 2

[23] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 2, 3

[24] J. Janai, F. Güney, A. Behl, and A. Geiger. Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. *arXiv preprint arXiv:1704.05519*, 2017. 1

[25] K. Karsch, C. Liu, and S. B. Kang. Depth extraction from video using non-parametric sampling. In *European Conference on Computer Vision*, pages 775–788. Springer, 2012. 2, 8

[26] K. Karsch, K. Sunkavalli, S. Hadap, N. Carr, H. Jin, R. Fonte, M. Sittig, and D. Forsyth. Automatic scene inference for 3d object compositing. *ACM Transactions on Graphics (TOG)*, 33(3):32, 2014. 1

[27] S. Kim, K. Park, K. Sohn, and S. Lin. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In *European Conference on Computer Vision*, pages 143–159. Springer, 2016. 2

[28] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[29] J. Konrad, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee. Learning-based, automatic 2d-to-3d image and video conversion. *IEEE Transactions on Image Processing*, 22(9):3485–3496, 2013. 2

[30] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–96, 2014. 2

[31] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016. 1, 2, 8

[32] I. Lenz, H. Lee, and A. Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015. 1

[33] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5170, 2015. 1

[34] F. Liu, C. Shen, G. Lin, and I. Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2016. 1, 2, 8

[35] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, 2018. 2, 6

[36] N. Nguyen, P. Milanfar, and G. Golub. Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement. *IEEE Transactions on image processing*, 10(9):1299–1308, 2001. 2

[37] V. Patraucean, A. Handa, and R. Cipolla. Spatio-temporal video autoencoder with differentiable memory. *arXiv preprint arXiv:1511.06309*, 2015. 2

[38] A. Roy and S. Todorovic. Monocular depth estimation using neural regression forest. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5506–5514, 2016. 2

[39] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168, 2006. 2

[40] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2009. 1, 2, 5, 8

[41] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013. 1

[42] F. Steinbrücker, J. Sturm, and D. Cremers. Real-time visual odometry from dense rgb-d images. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 719–722. IEEE, 2011. 2

[43] D. Stoyanov, M. V. Scarzanella, P. Pratt, and G.-Z. Yang. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 275–282. Springer, 2010. 1

[44] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. Demon: Depth and motion network for learning monocular stereo. In *IEEE Conference on computer vision and pattern recognition (CVPR)*, volume 5, page 6, 2017. 2

[45] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, 2018. 2

[46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2

[47] A. Wedel, D. Cremers, T. Pock, and H. Bischof. Structure- and motion-adaptive regularization for high accuracy optic flow. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1663–1668. IEEE, 2009. 2

[48] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic huber-l1 optical flow. In *BMVC*, volume 1, page 3, 2009. 2

[49] J. Xie, R. Girshick, and A. Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision*, pages 842–857. Springer, 2016. 1, 2, 7

[50] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *Proceedings of CVPR*, volume 1, 2017. 1, 2

[51] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3917–3925, 2018. 1, 2

[52] N. Yang, R. Wang, J. Stückler, and D. Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *European Conference on Computer Vision*, pages 835–852. Springer, 2018. 2

[53] Y. Yang and S. Soatto. Conditional prior networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 271–287, 2018. 2

[54] Y. Yang, A. Wong, and S. Soatto. Dense depth posterior (ddp) from single image and sparse range. *arXiv preprint arXiv:1901.10034*, 2019. 2

[55] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 340–349, 2018. 2, 6

[56] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7, 2017. 2, 6

[57] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 117–126, 2016. 2

[58] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *European conference on computer vision*, pages 286–301. Springer, 2016. 2

[59] Y. Zou, Z. Luo, and J.-B. Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. *arXiv preprint arXiv:1809.01649*, 2018. 2