Generative Adversarial Networks for unsupervised monocular depth prediction

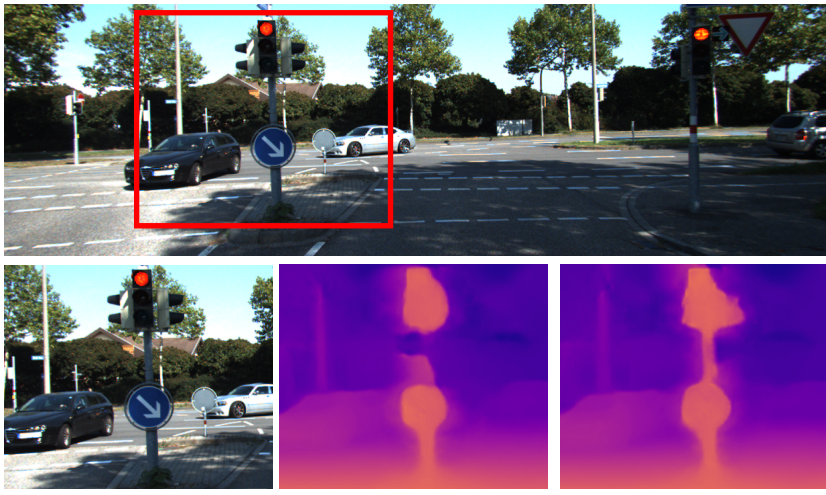
Filippo Aleotti, Fabio Tosi, Matteo Poggi, Stefano Mattoccia

University of Bologna, Viale del Risorgimento 2, Bologna, Italy
{filippo.aleotti}@studio.unibo.it
{fabio.tosi5,m.poggi,stefano.mattoccia}@unibo.it

Abstract. Estimating depth from a single image is a very challenging and exciting topic in computer vision with implications in several application domains. Recently proposed deep learning approaches achieve outstanding results by tackling it as an image reconstruction task and exploiting geometry constraints (e.g., epipolar geometry) to obtain supervisory signals for training. Inspired by these works and compelling results achieved by Generative Adversarial Network (GAN) on image reconstruction and generation tasks, in this paper we propose to cast unsupervised monocular depth estimation within a GAN paradigm. The generator network learns to infer depth from the reference image to generate a warped target image. At training time, the discriminator network learns to distinguish between fake images generated by the generator and target frames acquired with a stereo rig. To the best of our knowledge, our proposal is the first successful attempt to tackle monocular depth estimation with a GAN paradigm and the extensive evaluation on CityScapes and KITTI datasets confirm that it enables to improve state-of-the-art. Additionally, we highlight a major issue with data deployed by a standard evaluation protocol widely used in this field and fix this problem using a more reliable dataset recently made available by the KITTI evaluation benchmark.

1 Introduction

Accurate depth estimation is of paramount importance for many computer vision tasks and for this purpose active sensors, such as LIDARs or Time of Flight sensors, are being extensively deployed in most practical applications. Nonetheless, passive depth sensors based on conventional cameras have notable advantages compared to active sensors. Thus, a significant amount of literature aims at tackling depth estimation with standard imaging sensors. Most approaches leverage on multiple images acquired from different viewpoints to infer depth through binocular stereo, multi-view stereo, structure from motion and so on. Despite their effectiveness, all of them rely on the availability of multiple acquisitions of the sensed environment (e.g., binocular stereo requires two synchronized images).



(a) (b) (c)

Fig. 1. Estimated depth maps from single image. On top, frame from KITTI 2015 dataset, on bottom (a) detail from reference image (red rectangle), (b) depth predicted by Godard et al. [13] and (c) by our GAN architecture.

Monocular depth estimation represents an appealing alternative to overcome such constraint and recent works in this field achieved excellent results leveraging on machine learning [21,6,24,13]. Early works tackled this problem in a supervised manner [21,6,24] by training on a large amount of images with pixel-level depth labels. However, is well known that gathering labeled data is not trivial and particularly expensive when dealing with depth measurements [12,11,30,38]. More recent methods [53,13] aim to overcome this issue casting monocular depth estimation as an image reconstruction problem. In [53] inferring camera ego-motion in image sequences and in [13] leveraging on a stereo setup. In both cases, difficult to source labeled depth data are not required at all for training. The second method yields much better results outperforming even supervised methods [21,6,24] by a large margin.

Recently, Generative Adversarial Networks (GANs) [14] proved to be very effective when dealing with high-level tasks such as image synthesis, style transfer and more. In this framework, two architectures are trained to solve competitive tasks. The first one, referred to as *generator*, produces a new image from a given input (e.g., a synthetic frame from noise, an image with a transferred style, etc.) while the second one called *discriminator* is trained to distinguish between real images and those generated by the first network. The two models play a *min-max* game, with the generator trained to produce outputs good enough to fool the discriminator and this latter network trained to not being fooled by the generator.

Considering the methodology adopted by state-of-the-art methods for unsupervised monocular depth estimation and the intrinsic ability of GANs to detect

inconsistencies in images, in this paper we propose to infer depth from monocular images by means of a GAN architecture. Given a stereo pair, at training time, our generator learns to produce meaningful depth representations, with respect to left and right image, by exploiting the epipolar constraint to align the two images. The warped images and the real ones are then forwarded to the discriminator, trained to distinguish between the two cases. The rationale behind our idea is that a generator producing accurate depth maps will also lead to better reconstructed images, harder to be distinguished from original unwarped inputs. At the same time, for the discriminator will be harder to be fooled, pushing the generator to build more realistic warped images and thus more accurate depth predictions.

In this paper, we report extensive experimental results on the KITTI 2015 dataset, which provides a large amount of unlabeled stereo images and thus it is ideal for unsupervised training. Moreover, we highlight and fix inconsistencies in the commonly adapted split of Eigen [6], replacing Velodyne measurements with more accurate labels recently made available on KITTI [40]. Therefore, our contribution is threefold:

- Our framework represents, to the best of our knowledge, the first method to tackle monocular depth estimation within a GAN paradigm
- It outperforms the state-of-the-art methods
- We propose a more reliable evaluation protocol for the split of Eigen et al. [6]

2 Related Work

Depth estimation from images has a long history in computer vision. Most popular techniques rely on synchronized image pairs [39], multiple acquisitions from different viewpoints [9], at different time frames [35] or in presence of illumination changes [45]. Although certainly relevant to our work, these methods are not able to infer depth from a single image while recent methods casting depth prediction as a learning task and applications of GANs to other fields are strictly related to our proposal.

Learning-Based Stereo. Traditional binocular stereo algorithms perform a subset of steps as defined in [39]. The matching cost computation phase is common to all approaches, encoding an initial similarity score between pixels on reference image, typically the left, and matching candidates on the target. The seminal work by Zbontar and LeCun [49,50] computes matching costs using a CNN trained on image patches and deploys such strategy inside a well-established stereo pipeline [15] achieving outstanding results. In a follow-up work, Luo et al. [25] obtained more accurate matching representation casting the correspondence search as a multi-class classification problem. A significant departure from this strategy is represented by DispNet [29], a deep architecture aimed at regressing per-pixel disparity assignments after an end-to-end training. These latter methods require a large amount of labeled images (i.e., stereo pairs with ground-truth

disparity) for training [29]. Other works proposed novel CNN-based architectures inspired by traditional stereo pipeline as GC-Net [18] and CLR [31].

Supervised monocular depth estimation. Single image depth estimation is an ill-posed problem due to the lack of geometric constraints and thus it represents a much more challenging task compared to depth from stereo. Saxena et al. [37] proposed Make3D, a patch-based model estimating 3D location and orientation of local planes by means of a MRF framework. This technique suffers in presence of thin structures and lack of global context information often useful to obtain consistent depth estimations. Liu et al. [24] trained a CNN to tackle monocular depth estimation, while Ladicky et al. [21] exploited semantic information to obtain more accurate depth predictions. In [17] Karsch et al. achieved more consistent predictions at testing time by copying entire depth images from a training set. Eigen et al. [6] proposed a multi-scale CNN trained in supervised manner to infer depth from a single image. Differently from [24], whose network was trained to compute more robust data terms and pairwise terms, this approach directly infers the final depth map from the input image. Following [6] other works enabled more accurate estimations by means of CRF regularization [23], casting the problem as a classification task [2], designing more robust loss functions [22] or using scene priors for plane normals estimation [43]. Luo et al. [26] formulated monocular depth estimation as a stereo matching problem in which the right view is generated by a view-synthesis network based on Deep 3D [46]. Fu et al. [8] proposed a very effective depth discretization strategy and a novel ordinal regression loss achieving state-of-the-art results on different challenging benchmarks. Kumar et al. [4] demonstrated that recurrent neural networks (RNNs) can learn spatio-temporally accurate monocular depth prediction from video sequences. Atapour at al. [1] take advantage of style transfer and adversarial training on synthetic data to predict depth maps from real-world color images. Lastly, Ummenhofer et al. [41] proposed DeMoN, a deep model to infer both depth and ego-motion from a pair of subsequent frames acquired by a single camera. As for deep stereo models all these techniques require a large amount of labeled data at training time to learn meaningful depth representation from a single image.

Unsupervised monocular depth estimation. Pertinent to our proposal are some recent works concerned with view synthesis. Flynn et al. [7] proposed DeepStereo, a deep architecture trained on images acquired by multiple cameras in unsupervised manner to generate novel view points. Deep3D by Xie et al. [46] generates corresponding target view from an input reference image in the context of binocular stereo, by learning a distribution over all possible disparities for each pixel on the source frame and training their model with a reconstruction loss. Similarly, Garg et al. [10] trained a network for monocular depth estimation using a reconstruction loss over a stereo pair. To make their model fully differentiable they used Taylor approximation to make their loss linear, resulting in a more challenging objective to optimize. Godard et al. [13] overcome this problem by using a bilinear sampling [16] to generate images from depth prediction. At training time, this model learns to predict depth for both images

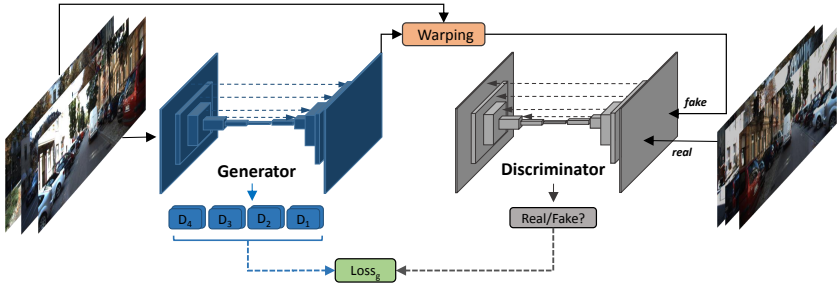


Fig. 2. Proposed adversarial model. Given a single input frame, depth maps are produced by a Generator (blue) and used to warp images. Discriminator (gray) process both raw and warped images, trying to classify the former as real and the latter as fake. The generator is pushed to improve depth prediction to provide a more realistic warping to fool the discriminator. At the same time the discriminator learns to improve its ability to perform this task.

in a stereo pair thus enabling to enforce a left-right consistency constraint as supervisory signal. A simple post-processing step allows to refine depth prediction. This approach was extended by including additional temporal information [51] and by training with semi-supervised data [20,48]. While previous method requires rectified stereo pairs for training, Zhou et al. [53] proposed to train a model to infer depth from unconstrained video sequences by computing a reconstruction loss between subsequent frames and predicting, at the same time, the relative pose between them. This strategy removes the requirement of stereo pairs for training but produces a less accurate depth estimation. Wang et al. [42] proposed a simple normalization strategy that circumvent problems in the scale sensitivity of the depth regularization terms employed during training and empirically demonstrated that the incorporation of a differentiable implementation of Direct Visual Odometry (DVO) improves previous monocular depth performance [53]. Mahjourian et al. [27] used a novel approximate ICP based loss to jointly learn depth and camera motion for rigid scenes, while Yin et al. [47] proposed a learning framework for jointly training monocular depth, optical flow and camera motion from video. [51]. Concurrently with our work, Poggi et al. [32] deployed a thin model for depth estimation on CPU and proposed a trinocular paradigm [33] to improve unsupervised approaches based on stereo supervision.

Generative Adversarial Networks. GANs [14] recently gained popularity by enabling to cast computer vision problems as a competitive task between two networks. Such methodology achieved impressive performance for image generalization [5,34], editing [54] and representation learning [34,28] tasks. More recent applications include text-to-image [36,52] and image-to-image [55] translations.

3 Method overview

In this section we describe our adversarial framework for unsupervised monocular depth estimation. State-of-the-art approaches rely on single network to accomplish this task. In contrast, at the core of our strategy there is a novel loss function based on a two players min-max game between two adversarial networks, as shown in Figure 2. This is done by using both a generative and a discriminative model competing on two different tasks, each one aimed at prevailing the other. This section discusses the geometry of the problem and how it is used to take advantages of 2D photometric constraints with a generative adversarial approach in a totally unsupervised manner. We refer to our framework as *MonoGAN*.

3.1 Generator model for monocular depth estimation

The main goal of our framework is to estimate an accurate depth map from a single image without relying on hard to find ground-truth depth labels for training. For this purpose, we can model this problem as a domain transfer task: given an input image x, we want to obtain a new representation $y = G(x)$ in the depth domain. In other contexts, GAN models have been successfully deployed for image-to-image translation [55]. For our purpose a generator network, depicted in blue in Figure 2, is trained to learn a transfer function $G : \mathcal{I} \rightarrow \mathcal{D}$ mapping an input image from $\mathcal{I}$ to $\mathcal{D}$, respectively, the RGB and the depth domain. To do so, it is common practice to train the generator with loss signals enforcing structure consistency across the two domains to preserve object shapes, spatial consistency, etc. Similarly, this can be done for our specific goal by exploiting view synthesis. That is, projecting RGB images into 3D domain according to estimated depth and then back-projecting to new synthesized view for which we need a real image to compare with. To make it possible, for each training sample at least two images from different points of view are required to enable the image reconstruction process described so far. In literature, this strategy is used by other unsupervised techniques for monocular depth estimation, exploiting both unconstrained sequences [53] or stereo imagery [13]. In this latter case, given two images i^l and i^r acquired by a stereo setup, the generator estimates inverse depth (i.e., disparity) d^l used to obtain a synthesized image $\tilde{i}^l$ by warping i^r with bilinear sampler function [16] being it fully differentiable and thus enabling end-to-end training. If d^l is accurate, shapes and structures are preserved after warping, while an inaccurate estimation would lead to distortion artifacts as shown on the right of Figure 3. This process is totally unsupervised with respect to the $\mathcal{D}$ domain and thus it does not require at all ground-truth labels at training time. Moreover, by estimating a second output d^r, representing the inverse mapping from i^l to i^r, allows to use additional supervisory signals by enforcing consistency in the $\mathcal{D}$ domain (i.e., Left-Right consistency constraint).

3.2 Discriminator model

To successfully accomplish domain transfer, GANs rely on a second network trained to distinguish images produced by the generator from those belonging to the target domain, respectively *fake* and *real* samples. We follow the same principle using the gray model in Figure 2, but acting differently from other approaches. In particular, to discriminate synthesized images from real ones we need a large amount of samples in the target domain. While for traditional domain transfer applications this does not represent an issue (requiring images without annotation), this becomes a limitation when depth is the target domain being ground-truth label difficult to source in this circumstance. To overcome this limitation, we train a discriminator to work on the RGB domain to tell original input images from synthesized ones. Indeed, if estimated disparity by the generator is not accurate, the warping process would reproduce distortion artifacts easily detectable by the discriminator. On the other hand, an accurate depth prediction would lead to a reprojected image harder to be recognized from a real one. Figure 3 shows, on the left, an example of real image and, on the right, a warped one synthesized according to an inaccurate depth estimation. For instance, by looking at the tree, we can easily tell the real image from the warped one. By training the discriminator on this task, the generator is constantly forced to produce more accurate depth maps thus leading to a more realistic reconstructed image in order to fool it. At the same time the discriminator is constantly pushed to improve its ability to tell real images from synthesized ones. Our proposal aims at such *virtuous behavior* by properly modeling the adversarial contribution of the two networks as described in detail in the next section.

4 Adversarial formulation

To train the framework outlined so far in end-to-end manner we define an objective function $\mathcal{L}(G, D)$ sum of two terms, a $\mathcal{L}_{GAN}$ expressing the min-max game between generator G and discriminator D:

$$\begin{aligned}\mathcal{L}_{GAN} = \min_G \max_D V(G, D) =& \mathbb{E}_{i_0 \sim \mathcal{I}}[\log(D(i_0))] \\ &+ \mathbb{E}_{i_1 \sim \tilde{\mathbb{I}}}[\log(1 - D(i_1))]\end{aligned} \tag{1}$$

with i_0 and i_1 belonging, respectively, to real images $\mathcal{I}$ and fake images $\tilde{\mathcal{I}}$ domains being the latter obtained by bilinear warping according to depth estimated by G and a data term $\mathcal{L}_{data}$ resulting in:

$$\mathcal{L}(G, D) = \mathcal{L}_{GAN} + \mathcal{L}_{data} \tag{2}$$

According to this formulation, generator G and discriminator D are trained to minimize loss functions $\mathcal{L}_G$ and $\mathcal{L}_D$:

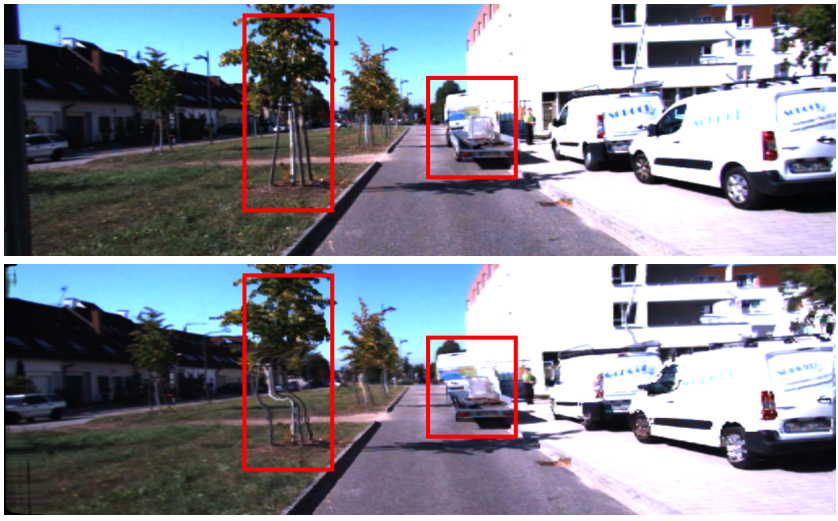


Fig. 3. Example of real (top) and warped (bottom) image according to an estimated depth. We can clearly notice how inaccurate predictions lead to warping artifacts on the reprojected frame (e.g., distorted trees) not perceivable elsewhere.

$$\mathcal{L}_G = \mathcal{L}_{data} + \alpha_{adv}\mathbb{E}_{i_0,i_1\sim\tilde{\mathbb{I}}}[\log\left(D(i_1)\right)] \tag{3}$$

$$\mathcal{L}_D = -\frac{1}{2}\mathbb{E}_{i_0\sim\mathcal{I}}\log(D(i_0)) - \frac{1}{2}\mathbb{E}_{i_1\sim\tilde{\mathcal{I}}}\log(1-D(i_1)) \tag{4}$$

To give an intuition, G is trained to minimize the loss from data term and the probability that D will classify a warped image $i_1 \sim \tilde{\mathcal{I}}$ as fake. This second contribution is weighted according to α_{adv} factor, hyper-parameter of our framework. Consistently, D is trained to classify a raw image $i_0 \sim \mathcal{I}$ as real and a warped one as fake. Despite our framework processes a transfer from $\mathcal{I}$ to depth domain $\mathcal{D}$, we highlight how in the proposed adversarial formulation the discriminator does not process any sample from domain $\mathcal{D}$, neither fake nor real. Thus it does not require any ground-truth depth map and perfectly fits with an unsupervised monocular depth estimation paradigm.

4.1 Data term loss

We define the data term $\mathcal{L}_{data}$ part of the generator loss function $\mathcal{L}_G$ as follows:

$$\mathcal{L}_{data} = \beta_{ap}(\mathcal{L}_{ap}) + \beta_{ds}(\mathcal{L}_{ds}) + \beta_{lr}(\mathcal{L}_{lr}) \tag{5}$$

where the loss consists in the weighted sum of three terms. The first one, namely *appearence* term, measures the reconstruction error between warped image $\tilde{I}$ and real one I by means of SSIM [44] and L1 difference of the two

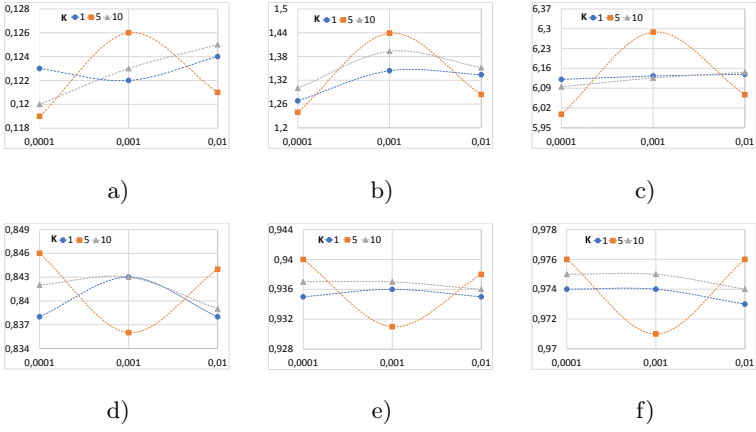


Fig. 4. Analysis of hyper-parameters α_{adv} and k of our GAN model, on x axis α_{adv}, on y axis an evaluation metric. a) Abs Rel, b) Sq Rel and c) RMSE metrics (lower is better). d) $\delta < 1.25$, e) $\delta < 1.25^2$, f) $\delta < 1.25^3$ metrics (higher is better). Interpolation is used for visualization purpose only. We can notice how our proposal using a weight α_{adv} of 0.0001 and a step k of 5achieves the best performance with all metrics.

$$\mathcal{L}_{ap} = \frac{1}{N}\sum_{i,j} \gamma \frac{1 - SSIM(I_{i,j}, \tilde{I}_{i,j})}{2} + (1-\gamma)||I_{i,j} - \tilde{I}_{i,j}||_1 \tag{6}$$

The second term is a *smoothness* constraint that penalizes large disparity differences between neighboring pixels along the x and y directions unless a strong intensity gradients in the reference image I occurs

$$\mathcal{L}_{ds} = \frac{1}{N}\sum_{i,j} |\delta_x d_{i,j}| e^{-||\delta_x I_{i,j}||} + |\delta_y d_{i,j}| e^{-||\delta_y I_{ij}||} \tag{7}$$

Finally, by building the generator to output a second disparity map d^r, we can add the term proposed in [13] as third supervision signal, enforcing left-right consistency between the predicted disparity maps, d^l and d^r, for left and right images:

$$\mathcal{L}^l_{lr} = \frac{1}{N}\sum_{i,j} |d^l_{i,j} - d^r_{i,j+d^l_{i,j}}| \tag{8}$$

Moreover, estimating d^r also enables to compute the three terms for both images in a training stereo pair.

5 Experimental results

In this section we assess the performance of our proposal with respect to state-of-the-art. Firstly, we describe implementation details of our model outlining

			Proposed method			Lower is better		Higher is better		
Exp.	Method	Dataset	Abs Rel	Sq Rel	RMSE	RMSE log	D1-all	δ <1.25	$\delta < 1.25^2$	$\delta < 1.25^3$
i)	Godard et al. [13]	K	0.124	1.388	6.125	0.217	30.272	0.841	0.936	0.975
	MonoGAN	K	**0.119**	**1.239**	**5.998**	**0.212**	**29.864**	**0.846**	**0.940**	**0.976**
ii)	Godard et al. [13]	CS	0.699	10.060	14.445	0.542	94.757	0.053	0.326	0.862
	MonoGAN	CS	**0.668**	**9.488**	**14.051**	**0.526**	**94.092**	**0.063**	**0.394**	**0.876**
iii)	Godard et al. [13]	CS+K	0.104	1.070	5.417	0.188	25.523	0.875	0.956	0.983
	MonoGAN	CS+K	**0.102**	**1.023**	**5.390**	**0.185**	**25.081**	**0.878**	**0.958**	**0.984**
iv)	Godard et al. [13] + pp	CS+K	0.100	0.934	**5.141**	0.178	25.077	0.878	**0.961**	**0.986**
	MonoGAN + pp	CS+K	**0.098**	**0.908**	5.164	**0.177**	**23.999**	**0.879**	**0.961**	**0.986**

Table 1. Results on KITTI stereo 2015 [30]. We compare MonoGAN with [13] using different training schedules, respectively only KITTI sequences (K), only CityScapes (CS) and both sequentially (CS+K). Adversarial contribution always improves the results. We indicate with pp results obtained after applying the final post-processing step proposed in [13].

the architecture of generator and discriminator networks. Then, we describe the training protocols followed during our experiments reporting an exhaustive comparison on KITTI 2015 stereo dataset [30] with state-of-the-art method [13]. This evaluation clearly highlights how the adversarial formulation proposed is beneficial when tackling this unsupervised monocular depth estimation. Moreover, we compare our proposal with other frameworks known in literature, both supervised and unsupervised, on the split of data used by Eigen et al. [6]. In this latter case we provide experimental results on the standard Eigen split as well as on a similar one made of more reliable data. This evaluation highlights once again the effectiveness of our proposal.

5.1 Implementation Details

For our GAN model, we deploy a VGG-based generator as in [13] counting 31 million parameters. We designed the discriminator in a similar way but, since the task of the discriminator is easier compared to the one tackled by the generator, we reduced the amount of feature maps extracted by each layer by a factor of two to obtain a less complex architecture. In fact, it counts about 8 million parameters, bringing the total number of variables of the overall framework to 39 million at training time. At test time, the discriminator is no longer required, restoring the same network configuration of [13] and thus the same computational efficiency.

For a fair comparison, we tune hyper-parameters such as learning rate or weights applied to loss terms to match those in [13], trained with a multi-scale data term while the adversarial contribution is computed at full resolution only. Being the task of D easier compared to depth estimation performed by G, we interleave the updates applied to the two. To this aim we introduce a further hyper-parameter k as the ratio between the number of training iterations performed on G and those on D, in addition to α_{adv}. In other words, discriminator weights are updated only every k updates of the generator. We will report evaluations for different values of parameter k. To jointly train both generator

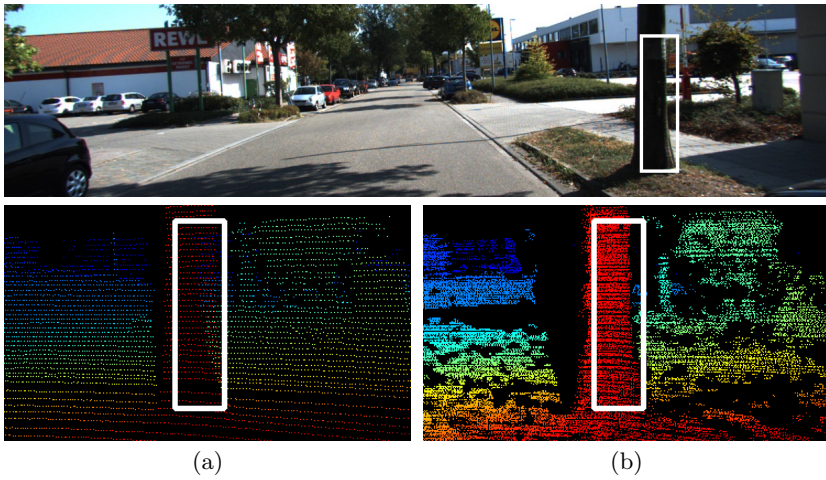
			Proposed method		Lower is better		Higher is better		
Method	cap	Dataset	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou et al. [53]	80 m	CS+K	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Mahjourian et al. [27]	80 m	CS+K	0.159	1.231	5.912	0.243	0.784	0.923	0.970
Yin et al. [47]	80 m	CS+K	0.153	1.328	5.737	0.232	0.802	0.934	0.972
Wang et al. [42]	80 m	CS+K	0.148	1.187	5.496	0.226	0.812	0.938	0.975
Poggi et al. [32] (200)	80 m	CS+K	0.146	1.291	5.907	0.245	0.801	0.926	0.967
Godard et al. [13]	80 m	CS+K	0.124	1.076	5.311	0.219	0.847	0.942	0.973
MonoGAN	80 m	CS+K	0.124	1.055	5.289	0.220	0.847	0.942	0.973
Godard et al. [13] + pp	80 m	CS+K	**0.118**	0.923	5.015	**0.210**	0.854	0.947	**0.976**
MonoGAN + pp	80 m	CS+K	**0.118**	**0.908**	**4.978**	**0.210**	**0.855**	**0.948**	**0.976**
Garg et al. [10]	50 m	K	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Zhou et al. [53]	50 m	CS+K	0.190	1.436	4.975	0.258	0.735	0.915	0.968
Mahjourian et al. [27]	50 m	CS+K	0.151	0.949	4.383	0.227	0.802	0.935	0.974
Poggi et al. [32] (200)	50 m	CS+K	0.138	0.937	4.488	0.230	0.815	0.934	0.972
Godard et al. [13]	50 m	CS+K	0.117	0.762	3.972	0.206	0.860	0.948	0.976
MonoGAN	50 m	CS+K	0.118	0.761	3.995	0.208	0.860	0.949	0.976
Godard et al. [13] + pp	50 m	CS+K	**0.112**	0.680	3.810	**0.198**	0.866	**0.953**	**0.979**
MonoGAN + pp	50 m	CS+K	**0.112**	**0.673**	**3.804**	**0.198**	**0.868**	**0.953**	**0.979**

Table 2. Results for unsupervised techniques on the original Eigen et al. [6] split based on raw Velodyne data.

and discriminator we use two instances of Adam optimizer [19], with $\beta_1 = 0.9$, $\beta_2 = 0.99$ and $\epsilon = 10^{-8}$. The learning rate is the same for both instances: it is set at $\lambda = 10^{-4}$ for the first 30 epochs and then halved each 10 epochs. Number of epochs is set to 50 as for [13]. Training data are extracted from both KITTI raw sequences [30] and CityScapes dataset [3] providing respectively about 29000 and 23000 stereo pairs, these latter samples are cropped to remove lower part of the image frames (depicting a portion of the car used for acquisition) as in [13]. Moreover, as in [13] we perform data augmentation by randomly flipping input images horizontally and applying the following transformations: random gamma correction in [0.8,1.2], additive brightness in [0.5,2.0], and color shifts in [0.8,1.2] for each channel separately. The same procedure is applied before forwarding images to both generator and discriminator.

5.2 Hyper-parameters analysis

As mentioned before, our GAN model introduces two additional hyper-parameters: the weight α_{adv} applied to the adversarial loss acting on the generator and the iteration interval k between subsequent updating applied to the discriminator. Figure 4 reports an analysis aimed at finding the best configuration (α_{adv}, k). On each plot, we report an evaluation metric used to measure accuracy in the field of monocular depth estimation (e.g., in [13]) as a function of both α_{adv} and k. Respectively, on top we report from left to right Abs Rel, Sq Rel and RMSE (lower scores are better), on bottom $\delta < 125$, $\delta < 125^2$ and $\delta < 125^3$ (higher scores are better). These results were obtained training MonoGAN on the 29000 KITTI stereo images [30], with α_{adv} set to 0.01, 0.001 and 0.0001 and k to 1, 5 and 10, for a total of 9 models trained and evaluated in Figure 4. We can notice how the configuration α_{adv}=0.0001 and k=5 achieves the best performance with all evaluated metrics. According to this analysis we use these hyper-parameters



(a) (b)

Fig. 5. Qualitative comparison between (a) reprojected raw Velodyne points as done in the original Eigen split for results reported in Table 2 and (b) reprojected ground-truth labels filtered according to [40], available on the KITTI website, deployed for our additional experiments reported in Table 3. Warmer colors encode closer points.

in the next experiments, unless otherwise stated. It is worth to note that despite the much smaller magnitude of α_{adv} compared to weights α_{ap}, α_{ds} and α_{lr} in data term (5), its contribution will affect significantly depth estimation accuracy as reported in the remainder.

5.3 Evaluation on KITTI dataset

Table 1 reports experimental results on the KITTI 2015 stereo dataset. For this evaluation, 200 images with provided ground-truth disparity from KITTI 2015 stereo dataset are used for validation, as proposed in [13]. We report results for different training schedules: running 50 epochs on data from KITTI only (K), from CityScapes only (CS) and 50 epochs on CityScapes followed by 50 on KITTI (CS+K). We compare our proposal to state-of-the-art method for unsupervised monocular depth estimation proposed by Godard et al. [13] reporting for this method the outcome of the evaluation available in the original paper. Table 1 is divided into four main sections, representing four different experiments. In particular, i) compares MonoGAN with [13] when both trained on K. We can observe how our framework significantly outperforms the competitor on all metrics. Experiment ii) concerns the two models trained on CityScapes data [3] and evaluated on KITTI stereo images, thus measuring the generalization capability across different environments. In particular, CityScapes and KITTI images differ not only in terms of scene contents but also for the camera setup. We can notice that MonoGAN better generalizes when dealing with different data. In iii), we train both models on CityScapes first and then on KITTI, showing that MonoGAN better benefits from using different datasets at training time compared

			Proposed method		Lower is better		Higher is better		
Method	cap	Dataset	Abs Rel	Sq Rel	RMSE	RMSE log	δ <1.25	$\delta < 1.25^2$	$\delta < 1.25^3$
Godard et al. [13]	80 m	CS+K	0.097	0.728	4.279	0.151	0.898	0.973	0.991
MonoGAN	80 m	CS+K	**0.096**	**0.699**	**4.236**	**0.150**	**0.899**	**0.974**	**0.992**
Godard et al. [13] + pp	80 m	CS+K	0.092	0.596	3.977	0.145	0.902	0.975	0.992
MonoGAN + pp	80 m	CS+K	**0.090**	**0.566**	**3.911**	**0.143**	**0.906**	**0.977**	**0.993**
Godard et al. [13]	50 m	CS+K	0.095	0.607	**4.100**	0.149	0.896	0.975	0.992
MonoGAN	50 m	CS+K	**0.094**	**0.600**	4.110	**0.148**	**0.897**	**0.976**	**0.993**
Godard et al. [13] + pp	50 m	CS+K	0.091	0.544	3.996	0.145	0.899	0.976	0.993
MonoGAN + pp	50 m	CS+K	**0.089**	**0.522**	**3.958**	**0.143**	**0.902**	**0.978**	**0.994**

Table 3. Results for MonoGAN and Godard et al. [13] on 93.5% of Eigen et al. [6] split using accurate ground-truth labels [40] recently made available by KITTI evaluation benchmark.

to [13] thus confirming the positive trend outlined in the previous experiments. Finally, in iv) we test the network trained in iii) refining the results with the same post-processing step described in [13]. It consists in predicting depth for both original and horizontally flipped input image, then taking 5% right-most pixels from the first and 5% left-most from the second, while averaging the two predictions for remaining pixels. With such post-processing, excluding one case out of 6 (i.e., with the RMSE metric) MonoGAN has better or equivalent performance compared to state-of-the-art. Overall, the evaluation on KITTI 2015 dataset highlights the effectiveness of the proposed GAN paradigm. In experiments iii) and iv), we exploited adversarial loss only during the second part of the training (i.e., on K) thus starting from the same model of [13] trained as in experiment ii), with the aim to assess how the discriminator improves the performance of a pre-trained model. Moreover, when fine-tuning we find beneficial to change the α_{adv} weight, similarly to traditional learning rate decimation techniques. In particular, we increased the adversarial weight α_{adv} from 0.0001 to 0.001 after 150k iterations (out of 181k total).

5.4 Evaluation on Eigen split

We report additional experiments conducted on the split of data proposed by Eigen et al. in [6]. This validation set is made of 697 depth maps obtained by projecting 3D points inferred by a Velodyne laser into the left image of the acquired stereo pairs in 29 out of 61 scenes. The remaining 32 scenes are used to extract 22600 training samples. We compare to other monocular depth estimation framework following the same protocol proposed in [13] using the same crop dimensions and parameters.

Table 2 reports a detailed comparison of unsupervised methods. On top, we evaluated depth maps up to a maximum distance of 80 meters. We can observe how MonoGAN performs on par or better than Godard et al. [13] outperforming it in terms of Sq Rel and RMSE errors and $\delta < 1.25$, $\delta < 1.25^2$ metrics. On the bottom of the table, we evaluate up to 50 meters maximum distance to compare with Garg et al. [10]. This evaluation substantially confirms the previous trend. As for experiments on KITTI 2015 stereo dataset, we find out that increasing by

a factor 10 the adversarial weight α_{adv} from 0.0001 to 0.001 after 150k iterations out of 181k total increases the accuracy of MonoGAN. Apparently, the margin between MonoGAN and [13] is much lower on this evaluation data. However, as already pointed out in [13] and [40], depth data obtained through Velodyne projection are affected by errors introduced by the rotation of the sensor, the motion of the vehicle and surrounding objects and also incorrect depth readings due to occlusion at object boundaries. Therefore, to better assess the performance of our proposal with respect to state-of-the-art we also considered the same split of images with more accurate ground-truth labels made available by Uhrig et al. [40] and now officially distributed as depth ground-truth maps by KITTI benchmark. These maps are obtained by filtering Velodyne data with disparity obtained by the Semi Global Matching algorithm [15] so as to remove outliers from the original measurements. Figure 5 shows a qualitative comparison between depth labels from raw Velodyne data reprojected into the left image, deployed in the original Eigen split, and labels provided by [40], deployed for our additional evaluation. Comparing (a) and (b) to the reference image at the top we can easily notice in (a) several outliers close to the tree trunk border not detectable in (b). Unfortunately, accurate ground-truth maps provided by [40] are not available for 45 images of the original Eigen split. Therefore, the number of testing images is reduced from from 697 to 652. However, at the expense of a very small reduction of validation samples (i.e., 6.5%) we get much more reliable ground-truth data according to [40]. With such accurate data, Table 3 reports a comparison between [13] and MonoGAN with and without post-processing, thresholding at 80 and 50 meters as for previous experiment on standard Eigen split. From Table 3 we can notice how with all metrics, excluding one case, MonoGAN on this more reliable dataset outperforms state-of-the-art [13] confirming the trend already reported in Table 1 on the accurate KITTI 2015 benchmark.

6 Conclusions

In this paper, we proposed to tackle monocular depth estimation as an image generation task by means of a Generative Adversarial Networks paradigm. Leveraging at training time on stereo images, the generator learns to infer depth from the reference image and from this data to generate a warped target image. The discriminator is trained to distinguish between real images and fake ones generated by the generator. Extensive experimental results confirm that our proposal outperforms state-of-the-art techniques for supervised and unsupervised monocular depth estimation.

Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

1. A. Atapour-Abarghouei and T. P. Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
2. Y. Cao, Z. Wu, and C. Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.
3. 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.
4. A. CS Kumar, S. M. Bhandarkar, and P. Mukta. Depthnet: A recurrent neural network architecture for monocular depth prediction. In *1st International Workshop on Deep Learning for Visual SLAM, (CVPR)*, 2018.
5. E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.
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.
7. 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.
8. H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
9. Y. Furukawa, C. Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015.
10. 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.
11. A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.
12. 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.
13. C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017.
14. 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.
15. H. Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 807–814. IEEE, 2005.
16. M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.
17. K. Karsch, C. Liu, and S. Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2144–2158, 2014.

18. A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry. End-to-end learning of geometry and context for deep stereo regression. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
19. D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
20. Y. Kuznietsov, J. Stuckler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
21. 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.
22. 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.
23. B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1119–1127, 2015.
24. 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.
25. W. Luo, A. G. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5695–5703, 2016.
26. Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin. Single view stereo matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
27. R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
28. M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, pages 5040–5048, 2016.
29. N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016.
30. M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
31. J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
32. M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia. Towards real-time unsupervised monocular depth estimation on cpu. In *IEEE/JRS Conference on Intelligent Robots and Systems (IROS)*, 2018.
33. M. Poggi, F. Tosi, and S. Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *6th International Conference on 3D Vision (3DV)*, 2018.
34. A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

35. R. Ranftl, V. Vineet, Q. Chen, and V. Koltun. Dense monocular depth estimation in complex dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4058–4066, 2016.
36. S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.
37. 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.
38. D. Scharstein, H. Hirschmller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In X. Jiang, J. Hornegger, and R. Koch, editors, *GCPR*, volume 8753 of *Lecture Notes in Computer Science*, pages 31–42. Springer, 2014.
39. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.
40. J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017.
41. 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, 2017.
42. C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
43. X. Wang, D. Fouhey, and A. Gupta. Designing deep networks for surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 539–547, 2015.
44. 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.
45. R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):191139, 1980.
46. 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.
47. Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
48. N. Young, R. Wang, J. Stuckler, and D. Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *15th European Conference on Computer Vision (ECCV)*, 2018.
49. J. Zbontar and Y. LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1592–1599, 2015.
50. J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32):2, 2016.
51. 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
52. H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE Int. Conf. Comput. Vision (ICCV)*, pages 5907–5915, 2017.

53. 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.
54. J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.
55. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.