Self-supervised Learning for Single View Depth and Surface Normal Estimation

Huangying Zhan, Chamara Saroj Weerasekera, Ravi Garg, Ian Reid

Abstract— In this work we present a self-supervised learning framework to simultaneously train two Convolutional Neural Networks (CNNs) to predict depth and surface normals from a single image. In contrast to most existing frameworks which represent outdoor scenes as fronto-parallel planes at piece-wise smooth depth, we propose to predict depth with surface orientation while assuming that natural scenes have piece-wise smooth normals. We show that a simple depth-normal consistency as a soft-constraint on the predictions is sufficient and effective for training both these networks simultaneously. The trained normal network provides state-of-the-art predictions while the depth network, relying on much realistic smooth normal assumption, outperforms the traditional self-supervised depth prediction network by a large margin on the KITTI benchmark.

I. INTRODUCTION

Recovering 3D scene structure from image data is long-studied problem in the robotics and computer vision communities. Decades of research has been focused on recovering the 3D structure of the scene from multiple 2D images as a geometric inverse problem, but the recent surge in deep learning has produced promising solutions for recovering scene geometry even from a single RGB image. Supervised deep neural networks trained on large indoor RGBD data-sets like NYUD [1] and SUN-RGBD [2] can currently produce high fidelity depth and surface normal predictions from single RGB image.

While the availability of cheap range sensors like Kinect makes large-scale supervised training of single-view depth and normal predictions possible for indoor scenes, the absence of high quality, dense depth sensing outdoors makes supervised learning difficult. To address this problem, Garg *et al.* [3] introduced a self-supervised learning paradigm where large stereo data-set can be used to train the single view depth prediction networks by using the traditional dense stereo loss function of [4] as the supervisory signal. Various extensions of the self-supervised learning for depth estimation have been proposed in recent times, including the use of robust image alignment costs, Godard *et al.* [5]. [6] extended the framework to enable training depth predictors using monocular sequence whereas [7], [8] advocate using stereo sequences to train a network for unsupervised two-frame visual odometry, because this addresses the issue of metric scale (which is otherwise unobservable). [8] additionally propose to learn good features to match and [9] advocate to evaluate the quality of warp using Generative Adversarial Networks [10].
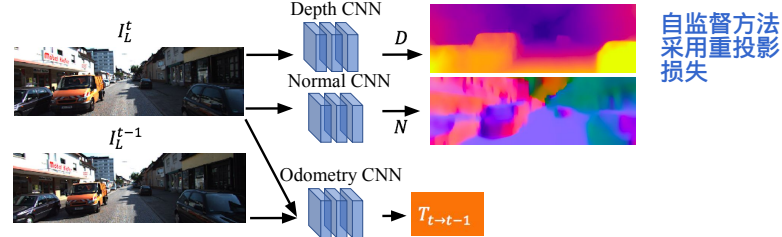
Fig. 1. Our test-time setup where depths and surface normals are predicted from a single image, and ego-motion is predicted from two views. At train-time, all three networks are trained in a self-supervised manner from stereo image sequence data.

Most of these self-supervised frameworks minimise a loss based on image data even though they are predicting a depth map. This means that they must regularise the predicted depth maps during training in regions where there is no strong photometric information; this is usually done by encouraging the predicted depth maps to be piece-wise smooth, or constant with depth discontinuities aligning image edges. These assumptions are rarely realistic and lead to fronto-parallel planar artifacts in the estimated structures in homogeneous regions.

In this work, we propose to address these issues by jointly learning two convolutional neural networks predicting depths and surface normals from a single input image in a self-supervised framework. Estimating surface normals in conjunction with depth-maps allows for a richer geometric reasoning where we can relax the piece-wise smooth/constant depth-map assumption to allow for smooth or planer surfaces in the scene. Similar to [7], [8], in addition to surface prediction networks, we learn a two-frame relative camera pose estimation network to predict visual odometry which allows us to use both stereo and temporal information available in KITTI dataset for accurate depth prediction. Our training schema is described in detail in Section III and the test-time setup is depicted in Fig. 1.

Very closely related to our work are the recently published works of [11] and [12]. In these works, the authors also advocate replacing the piece-wise smooth depth assumption with that of piece-wise smooth normals, however our proposed framework is different in the following aspects:

- Most importantly, unlike our proposed method, [11], [12] do not explicitly learn to predict surface normals from a single image. Instead, these methods estimate the normals from the predicted depths and propose to regularize them to iteratively refine the depth predictions. This amounts to imposing a hard constraint on the

normals to be the function of depth, limiting the normal accuracy, since the normals computed from depth are bound to have severe depth discretization artifacts and are very noisy. We show in our experiments that the combination of a dedicated network for normals and a soft constraint between inverse depths and normals leads to better predictions.

- Both [11], [12] propose to regularize second order depth discontinuity along with the normal discontinuity which is redundant. We show that no additional prior on depth is required for learning state-of-the-art normals.
- Both [11], [12] use monocular setup for training whereas we use stereo information to produce metric visual odometry given a frame pair addressing depth-translation scale ambiguity.
- We introduce a depth and normal consistency term over time and penalize the inconsistent depth and normal predictions for two consecutive frames of the video sequences during training. This leads to improvement in accuracy of the estimations.

In summary, the proposed method to the best of our knowledge, is the first self-supervised framework to jointly train for single view depth, single view normal, and two frame visual odometry prediction without depth/translation scale ambiguity. Our system produces state-of-the-art surface normal estimation and significantly better depth prediction compared to the corresponding depth smoothness based self-supervised framework on KITTI benchmark dataset.

II. RELATED WORK

Depth estimation from a monocular color image is a long-standing problem in computer vision. Early work mainly aimed at estimating surface normals from a single image and in turn integrated them to form depth maps [13]. Normals were estimated using shape from shading [14], shape-from-defocus [15] and other low-level image features [16], [17] or perspective geometry based reasoning using vanishing points and lines [18]. In contrast, learning based approaches such as [19], [20], especially those using neural nets [21], [22], [23], [24] in recent times have achieved state-of-the art performance in depth/surface normal prediction. In the following section we give a brief overview of both supervised and unsupervised methods for learning to predict geometry from single image, the latter of which is more relavent to this work.

Supervised learning of single-view geometry Supervised learning based methods for depth prediction rely upon availability of sensory depth data, such as that acquired by the Kinect sensor and LIDAR sensor (for indoor and outdoor scenes respectively). One of the early learning based methods to achieve reasonable success in single image depth estimation is Make3D [25], [26] which relied upon a set of handcrafted features to map patches from the input image to some feature space, and in turn learn to regress from that feature space to depth values. They also use an additional pairwise depth smoothness prior, modeled as either a Gaussian or Laplacian distribution with a data dependent learned

variance, in order to regularize and globally optimize for a depth map, with the belief that the process of recovering a depth map requires global reasoning on the image.

Recently, deep learning-based methods dominate this area [27], [28], [29]. For example, [27] train a multi-scale Convolutional Neural Network, operating at coarse and fine image resolutions, to regress a depth map from a single image, and in [21] they extend their network to a three-scale architecture and regress for depth maps, normal maps, and semantic labels in real-time from a single image. The semantic label maps were predicted from a single RGB-D image as the additional depth channel improved results. In [30] the latter work was extended to jointly predict depth, surface normals and surface curvature, which improved the results of all three tasks.

Liu *et al.* [31] proposed to formulate depth estimation as a deep continuous Conditional Random Fields (CRF) learning problem. Given the continuous nature of the depth values, they learn the unary depth values and weightings for the pairwise smoothness potential functions via CNNs in an end-to-end framework. [32] used a fully convolutional network architecture based on ResNet [33] with a novel upsampler for decoding the depth map at input resolution. Kendall *et al.* [29] adapted the DenseNet architecture for several regression tasks including depth prediction, and showed that jointly predicting pixelwise depths and confidences, where the output is modeled as a multivariate Gaussian distribution, improves depth estimation results. [34] combined shallow convolutional networks with regression forests to reduce the need for large training sets. Recently in [35], it was proposed that sharper predictions at depth boundaries can be achieved by emphasizing local depth error gradients. This same phenomenon was observed in [21], [36] that also emphasized local depth errors during training. Our inverse-depth normal consistency terms also emphasizes local depth errors based on the predicted normals and achieves a similar effect but in a more implicit unsupervised fashion (i.e. with no dedicated sensory depth data).

Self-supervised learning of single-view geometry Recent work have started to incorporate multi-view geometry based loss functions for depth regression resulting in a self-supervised learning framework for inferring depth from single image. This stream of work aims to replace the more explicit sensory-data based ground truth supervision with a good image alignment loss between different views observing the same scene (by using stereo data or monocular videos for supervision). The ubiquity and relatively low price of RGB cameras and availability of large datasets of recorded videos increase the attractiveness of this approach.

Using stereo pairs for training, [3][5] deploy an auto-encoder like framework where the authors propose to predict the disparity (inverse depth) of the left image, using which the right image of the stereo pair can be warped to synthesize the left image. The photometric difference between the input (left) image and the warped image is minimized to train the single view depth predictor. An inverse depth smoothness prior on the predicted depths is used to regularize the

solution, encouraging piece-wise smooth depth maps. [6] extended the above framework to jointly estimate depth and ego-motion using monocular videos - upto a scale. Methods like [7][8] proposed to combine the advantages of using both spatial and temporal information available in KITTI sequences for improving depth predictions while solving the scaling ambiguity issue. A large body of work since have been targeted to use better loss functions, in particular the image alignment loss [5], [9] propose to use SSIM and GANs respectively for image matching. Enforcing temporal consistency in the predicted depths by aligning the back-projected depth maps via differentiable approximation of ICP has been studied in [37].

While most of the self supervised approaches have mainly focused on getting accurate depth-maps, little attention has been devoted to use other scene representations. We are aware of two recent works [11], [12] which incorporate the surface orientation (normal) estimation for single view geometric understanding. Similar to [6], [11][12] learn depth from monocular sequences using a self-supervised photometric loss but additionally they compute surface normals from the predicted depths using a weighted mean cross product [38]. They propose to regularize the inverse depths and the normals computed from the depth predictions simultaneously. We believe that this is redundant and a separate normal prediction is beneficial then relying on the normals to be computed from predicted depth. The differences between our work and [11], [12] is detailed in section I and a extensive comparison of the proposed work with these methods is described in section IV.

III. FRAMEWORK FOR JOINT LEARNING OF DEPTHS AND SURFACE NORMALS

In this section, we present our system which consists of three CNNs. One each for per-pixel single-view depth prediction, for single view surface normal prediction and a pose-net which takes two images – consecutive images of a monocular video – as input to predict the camera motion (vehicle's egomotion in KITTI) between these two frames in metric units. Our system is trained in a self-supervised manner, which means no ground truth data (depths or surface normals) is required for training. Instead we use stereo sequences for training for depths and surface normals from a single image where two consecutive stereo-pairs $\left(\{I_L^t, I_R^t\}, \{I_L^{t-1}, I_R^{t-1}\}\right)$ form a single training instance.

The goal is to predict the depth map D and surface normal map $\hat{N}$ of I_L^t (the left image at time t) which we define as the reference image I for a particular training instance. At the same time we also want to predict $T_{t\rightarrow t-1}$ which is the relative pose (ego-motion) between the left/right image at time t and the left/right image at time $t-1$.

Our proposed loss function to train these three networks jointly consists of six terms:

$$\lambda_1 L_P + \lambda_2 L_{DN} + \lambda_3 L_N + \lambda_4 L_{NS} + \lambda_5 L_{DC} + \lambda_6 L_{NC}, \quad (1)$$

where the λ's are the relative weights for losses used for training. L_P denotes the photometric alignment cost involv-

ing the scene's depth observed by the left camera at time t and $t-1$ with the estimated egomotion, L_{DN} enforces the estimated depths and normals to be consistent, L_N enforces the predicted normals to face the camera and L_{NS} is a smoothness prior which favors the predicted normals to be piece-wise smooth. Additionally, assuming the scene is rigid, two temporal geometric consistency terms L_{DC} and L_{NC} enforce the estimated depths and normals at the two time instances to be consistent given the egomotion. Each of these terms are elaborated in the following sections.

A. Enforcing Multi-View Photometric Consistency

We enforce multi-view photometric consistency (both spatially and temporally) by minimizing the photometric cost that aligns image I_L^t and I_R^t (the left and right stereo images at time t) as well as the photometric cost that aligns image I_L^t and I_L^{t-1} (the left images at time t and $t-1$).

It follows that the photometric loss based on dense image alignment can be written as follows,

$$L_P = \sum_p |I_L^t(p) - W(I_R^t, p')| + |I_L^t(p) - W(I_L^{t-1}, p'')| \quad (2)$$

where, $W(I_R^t, p')$ and $W(I_L^{t-1}, p'')$ are the synthesized left images reconstructed from I_R^t and I_L^{t-1} respectively. $W(.,.)$ is a differentiable bilinear interpolation function [39] for indexing into a particular non-integer location in a given input image. The corresponding pixel p' in I_R^t for a pixel p in I_L^t is defined by the camera intrinsics, K; the known stereo baseline, $T_{L\rightarrow R}$; and the depth map, D as follows,

$$p' = K T_{L\rightarrow R} D(p) K^{-1} p \quad (3)$$

where p represents an integer pixel location in I_L^t (the reference image) in homogeneous coordinates. Similarly, the corresponding pixel p'' in I_L^{t-1} for a pixel in p in I_L^t is given by,

$$p'' = K T_{t\rightarrow t-1} D(p) K^{-1} p \quad (4)$$

The relative camera transformation, $T_{t\rightarrow t-1}$, is represented by 6 parameters in $se3$, and consists of rotation and translation components.

Note that the depth CNN predicts inverse depth D_{inv} which is better constrained, e.g. sky in infinite depth is zero is inverse depth. The depth is converted from our predicted inverse depth as follows: $D = 1/(D_{inv} + 10^{-4})$.

B. Normal Smoothness Prior

Relying on photometric loss described in the previous section for learning depth suffers from well known aperture problem. A single pixel from a uniformly colored area in a image can be matched to many pixels with similar color intensity in the next view making the depth estimation ambiguous. To counter the problem, previous works [3] [5] [8] adopt an inverse depth/disparity smoothness as a prior. However, as explained in the introduction such disparity smoothness assumption is not very realistic and strong regularization of disparity discontinuity leads to fronto-parallel artifacts in the predictions.
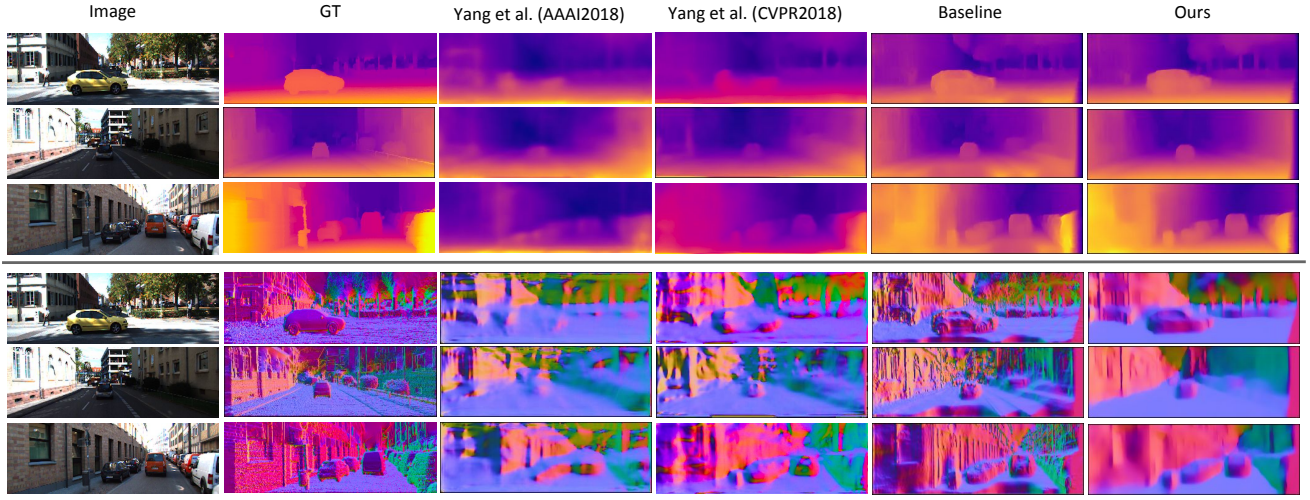
Fig. 2. Qualitative comparison of depths and surface normals between different methods. The ground truth (GT) depths are inpainted from sparse LIDAR ground truth depths. The ground truth surface normals are computed from the inpainted ground truth depths, and are not reliable for all the points (especially the upper part of the images where the LIDAR depths are missing).

In this work we rely on smooth normal assumption whereby we apply edge-aware regularization to the discontinuities in the predicted surface normal by minimizing:

$$L_{NS} = \sum_p |\partial_x \hat{\mathbf{N}}(p)| e^{-|\partial_x I(p)|} + |\partial_y \hat{\mathbf{N}}(p)| e^{-|\partial_y I(p)|} \quad (5)$$

where $\partial_x(.)$ and $\partial_y(.)$ are gradients in horizontal and vertical direction respectively. Similar to previous works [5], [8], we assume that the image edges are very good indicator of scene discontinuity, however we use them to guide the normal smoothness. Eqn. (5) allows for normal discontinuities at the areas where strong image gradient is present while penalizing the normal discontinuities in the homogeneous regions of the image.

C. Depth - Normal Consistency

For enforcing consistency of predicted depth with predicted surface normals, we use the inverse-depth-normal consistency term proposed in [40] as a soft constraint in the form of a loss for training our networks. This loss is based on the geometric relationship between the predicted normal $\hat{\mathbf{N}}(p)$ of the scene corresponding to point p in the reference image, its predicted depth $D(p)$ and the predicted depth $D(q)$ of p's neighbour q. The depth-normal consistency term can be written as:

$$\langle \hat{\mathbf{N}}(p), D(q)\tilde{\mathbf{X}}(q) - D(p)\tilde{\mathbf{X}}(p) \rangle = 0 \quad (6)$$

where $\langle , \rangle$ is the dot product operator and $\tilde{\mathbf{X}}(p) = K^{-1}p$, $\tilde{\mathbf{X}}(q) = K^{-1}q$. Note that $\hat{\mathbf{N}}(p)$ is normalized to have unit magnitude and is expressed in Cartesian coordinates. Eqn. (6) can be simplified as:

$$D(q)\langle \hat{\mathbf{N}}(p), \tilde{\mathbf{X}}(q) \rangle - D(p)\langle \hat{\mathbf{N}}(p), \tilde{\mathbf{X}}(p) \rangle = 0 \quad (7)$$

By dividing the above equation by $D(p)D(q)$ we get:

$$D_{inv}(p)\langle \hat{\mathbf{N}}(p), \tilde{\mathbf{X}}(q) \rangle - D_{inv}(q)\langle \hat{\mathbf{N}}(p), \tilde{\mathbf{X}}(p) \rangle = 0 \quad (8)$$

Finally, we minimize the following energy L_{DN}, penalizing inconsistency between predicted inverse depths and normals:

$$L_{DN} = G(p) \sum_{p,q \in \mathcal{N}(p)} |D_{inv}(p)c_{pq} - D_{inv}(q)c_{pp}| \quad (9)$$

$$c_{pq} = \langle \hat{\mathbf{N}}(p), \tilde{\mathbf{X}}(q) \rangle, \quad c_{pp} = \langle \hat{\mathbf{N}}(p), \tilde{\mathbf{X}}(p) \rangle \quad (10)$$

In our experiments, the neighbourhood $\mathcal{N}(p)$ comprises just the pixel itself and its two neighbours immediately below and to the right. The image-edge based weight $G(p) = e^{-\alpha |\nabla I(p)|_2^\beta}$ reduces regularization at image edges, under the assumption that these regions align with depth discontinuities. α and β are tunable parameters, which we use $[\alpha, \beta] = [1, 1]$.

It is easy to note that, in the special case when $c_{pq} = c_{pp} = -1$, i.e. the normal $\hat{\mathbf{N}} = (0, 0, -1)^T$ is pointed directly at the camera, Eqn. (9) reduces to the traditionally used inverse depth smoothness prior in unsupervised learning methods such as [3][5][8].

D. Fixing Surface Normal Direction Ambiguity

It is to be noted that the surface normal prediction network in our framework only rely upon the depth-normal consistency loss L_{DN}. However, normals estimated from the depth maps using Eqn. (6) have directional ambiguity. i.e. given the depth map the computed surface normal can face the camera or be in the opposite direction. To fix this ambiguity, we first compute an approximated surface normal $\hat{\mathbf{N}}_{\mathbf{c}}(p)$ from the predicted depth using mean cross product[1] and then penalize the deviation of the predicted normals $\hat{\mathbf{N}}(p)$ from these approximated normals my minimizing:

$$L_N = \frac{1}{2N} \sum_p ||\hat{\mathbf{N}}(p) - \hat{\mathbf{N}}_{\mathbf{c}}(p)||_2^2. \quad (11)$$

---

[1]It should be noted that this approximation was used in [11] to compute normals from depth maps.

Fig. 3. Qualitative comparison between surface normals computed from CNN depths (Stereo+Normal+Temporal) and surface normals predicted from the Normal CNN, showing the importance of having a dedicated Normal CNN. Left: Groundtruth (GT); Middle: Computed normals from predicted depths; Right: Predicted normals.

E. Enforcing Temporal Consistency of Predicted Geometry

Finally, while the accumulated loss terms defined in sections III-A-III-D makes the simultaneous learning of depth, normal and egomotion well posed, we enforce temporal consistency in our predictions to improve the accuracy of our predictions. Using the rigid scene assumption as the cameras move in space over time we want the predicted depths and normals at time t to be consistent with the respective predictions at time $t-1$. This is done by correctly transforming the scene geometry (inverse depth and normal maps) from frame t to frame $t-1$ much like the image warping over time as described in section III-A. In particular, we define two consistency-terms L_{DC} and L_{NC}:

$$L_{DC} = \sum_p |D_{inv}^t(p) - W(D_{inv}^{t-1'}, p'')| \qquad (12)$$

$$L_{NC} = \sum_p |\hat{\mathbf{N}}^t(p) - W(\hat{\mathbf{N}}^{t-1'}, p'')| \qquad (13)$$

where,

$$[X^{t-1'}(p), Y^{t-1'}(p), D^{t-1'}(p)] = T_{t\to t-1}^{-1} D^{t-1}(p) K^{-1} p \qquad (14)$$

$$\hat{\mathbf{N}}^{t-1'}(p) = R_{t\to t-1}^{-1} \hat{\mathbf{N}}^{t-1}(p). \qquad (15)$$

$D^{t-1'}(p) = 1/D_{inv}^{t-1'}(p)$ and $\hat{\mathbf{N}}^{t-1'}(p)$ in the above two equations are the transformed depths and normals from frame $t-1$ to frame t (based on the predicted pose $T_{t\to t-1}$ where $R_{t\to t-1}$ is the rotation component) for use in L_{DC} and L_{NC}.

To further clarify, it is important to bring the depth/normals which are estimated in the camera reference frame at any given time step into the current reference frame before applying consistency of depth/normal over time. For depth consistency we transform the estimated 3D points (back-projection of the depth map) at time $t-1$ to the camera reference frame of time t using the estimated pose $T_{t\to t-1}$ before warping. Similarly, the normals need to transformed from one reference frame to the other using the estimated rotation $R_{t\to t-1}$.

F. Network Architecture

For both the depth CNN and normal CNN, we use the same architecture except for the last prediction layer, which has a 1-channel output for the depth CNN while a 3-channel output for the normal CNN. Following previous works [3][5][8], we use a fully convolutional neural network with skip-connections. The network consists of an encoder and a decoder. For the encoder, we use the ResNet50 [41] variant (ResNet50-1by2), which has much less learnable parameters and also faster computation speed. For the decoder network,

we have studied two architectures, including the simple bilinear upsampler used in [3][8] and bilinear upsampling with convolutions, i.e. upsample coarser feature maps, concatenate with features in ResNet50-1by2 (skip connection), and apply convolution (including batch normalization, and ReLU activation). Our experiments show that although both upsamplers have similar quantitative performance for depth estimation as observed in [8], the latter architecture is able to produce sharper predictions, especially useful for predicting surface normals. We use sigmoid activation at the end of the depth and normal network. We additionally apply a L2-normalization on the surface normals to get unit normals. For the pose network, we use the one proposed in [8].

IV. EXPERIMENTS AND EVALUATIONS

In this section we describe the details of our experimental evaluation of our proposed framework. We evaluate our framework on KITTI dataset [42][43] and compare with prior art on both depth esimtation and surface normal estimation. Moreover, we present an ablation study to show the contribution of each component in our framework.

A. Implementation Details

We train our CNNs with the Caffe [44] framework. Adam optimizer [45] is used with the following settings, $[\beta_1, \beta_2, \epsilon] = [0.9, 0.999, 10^{-8}]$. The initial learning rate is 0.001 and is decreased to one tenth manually when the training loss converges. We found the loss weightings for Eqn. (1) via grid search and by referring to previous loss weightings adopted by [40], [8], and are as follows: $[\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6] = [1, 13, 1, 0.7, 1, 0.01]$.

B. Depth Evaluation

We evaluate our depth estimation result on the Eigen split provided by [27] for fair comparison with prior works. We follow the same depth evaluation protocol as in [3], [5]. Our results for depth estimation using our full framework (Stereo+Normal+Temporal) are presented in Table I and is compared against relevant prior works and our own 'Baseline' network that are all trained using the photometric loss with inverse depth smoothness prior. Our full method performs best in all measures compared to all the baselines. This reaffirms the importance of the more realistic normal smoothness prior over the traditional inverse-depth smoothness prior.

We also compare against our method without temporal photometric and geometric (depth and normal) consistency (Stereo+Normal) which leads to less accurate results as seen in Table I, signifying the importance of temporal geometric consistency especially for reconstructing far points in the scene.

In Fig. 2 it is apparent that the depth maps predicted using our proposed framework are superior, particularly to reconstruct the road (Rows 1-3) and building in Row 3 when compared with the Baseline approach, and other prior works. While the results of [12], [11] are blurry our proposed methods is able to retain sharp edges with correct reconstruction of non-fronto parallel planes.

Method	Supervision	Error metric				Accuracy metric		
		Abs Rel	SqRel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Train set mean	Depth	0.361	4.826	8.102	0.377	0.638	0.804	0.894
Zhou *et al.* [6]	Mono.	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Yang *et al.*[11]	Mono.	0.182	1.481	6.501	0.267	-	-	-
Yang *et al.*[12]	Mono.	0.162	1.352	6.276	0.252	-	-	-
Garg *et al.* [3]	Stereo	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Godard *et al.* [5]	Stereo	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Stereo+Inv.Depth.Smoothness (Baseline)	Stereo	0.150	1.409	5.800	0.249	0.800	0.923	0.964
Stereo+Normal	Stereo	0.135	1.194	5.718	0.237	0.816	0.929	0.968
Stereo+Normal+Temporal	Stereo	0.133	1.083	5.580	0.229	0.816	0.932	0.971

TABLE I

SELF-SUPERVISED DEPTH ESTIMATION PERFORMANCE COMPARISON ON KITTI DATASET (TRAIN&TEST ON EIGEN SPLIT, 80M CAP). THE RESULTS ARE EVALUATED ON CROPPED REGION USED IN [5]. OUR ABLATION STUDY IS PRESENTED ON THE BOTTOM OF THE TABLE.

C. Surface Normal Evaluation

There is no surface normal ground truth available in KITTI dataset. In particular, the depth ground truth in Eigen split provided by KITTI raw dataset is very sparse and unsuitable to generate surface normal. To have better evaluation on surface normal prediction, [11][12] use KITTI's official stereo split which contains 200 high quality disparity images from which reasonably high quality surface normals can be generated for evaluation. Following [11][12], we use the KITTI split for surface normal evaluation and inpaint the depth ground truth following the approach used in [1]. We use the mean cross product to generate surface normal ground truth from these inpainted depths. As, 92 out of 200 images in KITTI split are used in the training set of Eigen split, we only use the remaining 108 images for evaluation. Moreover, we follow the depth evaluation protocol [5] that only use the centre part of the prediction and evaluate normals only at the pixels where the ground truth depths exist to reduce the normal errors introduces due to inpainting.

The quantitative evaluation of unsupervised normal prediction frameworks is presented in Table II, where we compare against surface normals estimated via different methods. The bottom-most row (Stereo+Normal+Temporal (CNN)) shows our best result. This is the normal predicted by the Normal CNN using our full loss function. We show that using inverse depth-normal consistency (Stereo+Normal) gives better surface normals than the inverse depth smoothness (Baseline) and the result is further improved by using temporal geometric consistency. On the bottom part of the table, '- (CNN)' are the surface normals predicted from CNNs. We can see that the surface normals predicted from the CNNs are better than the corresponding results which are computed from predicted

depths in all cases, signifying the importance of a dedicated Normal CNN.

Fig. 2 (bottom) compares the normals predicted by our framework with that of [12], [46] and our 'Baseline' approach (where the normals are computed using the mean cross product rule). It can be easily seen that while the inverse depth smoothness regularization produces detailed but very noisy normal maps, our predicted normals are of a significantly high quality preserving the normal edges while being largely smooth/constant on the smooth objects/road and buildings.

It is important to note that unlike the proposed method, our 'Baseline', [11] and [12] do not have an explicit normal prediction network which allows deviation from the predicted depths. Our claim is that a dedicated network and a soft constraint allowing deviation from depth normal consistency is critical in good normal estimation performance. A clear visual indication is shown in Fig. 3 where we compare the predicted normals $\hat{N}$ by our Normal CNN with the ones $\hat{N}_c$ which are computed via the mean cross product of the corresponding depth predictions coming from our joint training framework. It can be seen that even after the joint training of depths and normals, computing the normals from the predicted depth leaves us with very noisy undesirable output. This effect is additionally quantified in Table II.

V. CONCLUSION

In this work we have proposed to simultaneously learn a single view depth and normal prediction network, along with a 2-frame visual odometry network in a self-supervised manner from stereo sequence training data. We show that these three networks can be learned together using a soft depth-normal consistency constraint while assuming the surface normals to be piece-wise smooth, to give state of the art surface normal predictions and significantly improved depth predictions when compared to prediction reliant on inverse-depth smoothness prior currently prevalent for self-supervised learning.

Method	Error metric		Accuracy metric		
	Mean	Median	11.25°	22.5°	30°
Yang *et al.*[11]	37.44	24.32	0.275	0.477	0.560
Yang *et al.* [12]	35.69	22.33	0.293	0.502	0.585
Baseline (Computed)	36.03	24.00	0.283	0.481	0.565
Stereo+Normal (Computed)	33.43	21.15	0.305	0.519	0.607
Stereo+Normal+Temporal (Computed)	32.01	20.17	0.319	0.534	0.622
Stereo+Normal (CNN)	30.37	19.13	0.335	0.551	0.640
Stereo+Normal+Temporal (CNN)	30.23	19.11	0.336	0.551	0.638

TABLE II

SURFACE NORMAL EVALUATED ON KITTI SPLIT (108/200 SAMPLES, EXCLUDING 92 SAMPLES IN EIGEN SPLIT). WE EVALUATED ON CENTRE CROPPED REGION AS DEPTH EVALUATION IN [5].

References

[1] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[2] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," June 2015.

[3] R. Garg, V. K. B G, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*. Springer, 2016, pp. 740–756.

[4] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[5] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017.

[6] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, 2017.

[7] R. Li, S. Wang, Z. Long, and D. Gu, "Undeepvo: Monocular visual odometry through unsupervised deep learning," *arXiv preprint arXiv:1709.06841*, 2017.

[8] 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 *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018, pp. 340–349.

[9] F. Aleotti, F. Tosi, M. Poggi, and S. Mattoccia, "Generative adversarial networks for unsupervised monocular depth prediction."

[10] 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*, 2014, pp. 2672–2680.

[11] Z. Yang, P. Wang, W. Xu, L. Zhao, and R. Nevatia, "Unsupervised learning of geometry with edge-aware depth-normal consistency," in *AAAI*, 2017.

[12] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "Lego: Learning edge with geometry all at once by watching videos," in *CVPR*, 2018.

[13] J.-D. Durou, Y. Quéau, and J.-F. Aujol, "Normal Integration – Part I: A Survey," jun 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01334349

[14] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape from Shading: A Survey," *IEEE Transactions on Pattern Analaysis and Machine Intelligence*, vol. 21, no. 8, pp. 690–706, 1999.

[15] S. Suwajanakorn, C. Hernandez, and S. M. Seitz, "Depth from focus with your mobile phone," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3497–3506.

[16] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic photo pop-up," in *ACM transactions on graphics (TOG)*, vol. 24, no. 3. ACM, 2005, pp. 577–584.

[17] ——, "Recovering surface layout from an image," *International Journal of Computer Vision*, vol. 75, no. 1, pp. 151–172, 2007.

[18] A. Criminisi, I. Reid, and A. Zisserman, "Single view metrology," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 1. IEEE, 1999, pp. 434–441.

[19] L. Ladický, B. Zeisl, and M. Pollefeys, "Discriminatively Trained Dense Surface Normal Estimation," in *ECCV*, 2014, vol. 8693, pp. 468–484.

[20] D. F. Fouhey, A. Gupta, and M. Hebert, "Data-driven 3D primitives for single image understanding," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3392–3399.

[21] 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*, 2015, pp. 2650–2658.

[22] X. Wang, D. F. Fouhey, and A. Gupta, "Designing Deep Networks for Surface Normal Estimation," in *CVPR*, 2015.

[23] A. Bansal, B. Russell, and A. Gupta, "Marr revisited: 2d-3d alignment via surface normal prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5965–5974.

[24] A. Bansal, X. Chen, B. Russell, A. Gupta, and D. Ramanan, "Pixelnet: Towards a general pixel-level architecture," *arXiv preprint arXiv:1609.06694*, 2016.

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

[26] 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*, vol. 31, no. 5, pp. 824–840, 2009.

[27] 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*, 2014, pp. 2366–2374.

[28] 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*. Springer, 2016, pp. 842–857.

[29] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems*, 2017, pp. 5580–5590.

[30] T. Dharmasiri, A. Spek, and T. Drummond, "Joint prediction of depths, normals and surface curvature from rgb images using cnns," in *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*. IEEE, 2017, pp. 1505–1512.

[31] 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*, vol. 38, no. 10, pp. 2024–2039, 2016.

[32] 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*. IEEE, 2016, pp. 239–248.

[33] 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*, 2016, pp. 770–778.

[34] A. Roy and S. Todorovic, "Monocular depth estimation using neural regression forest," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5506–5514.

[35] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," *CoRR*, vol. abs/1803.08673, 2018. [Online]. Available: http://arxiv.org/abs/1803.08673

[36] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5038–5047.

[37] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," *arXiv preprint arXiv:1802.05522*, 2018.

[38] Z. Jia, "Using cross-product matrices to compute the svd," *Numerical Algorithms*, vol. 42, no. 1, pp. 31–61, 2006.

[39] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2017–2025.

[40] C. S. Weerasekera, Y. Latif, R. Garg, and I. Reid, "Dense monocular reconstruction using surface normals," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 2524–2531.

[41] 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*, 2016, pp. 770–778.

[42] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[43] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[44] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

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

[46] Z. Yang, P. Wang, W. Xu, L. Zhao, and R. Nevatia, "Unsupervised learning of geometry with edge-aware depth-normal consistency," *arXiv preprint arXiv:1711.03665*, 2017.