DeFeat-Net: General Monocular Depth via Simultaneous Unsupervised Representation Learning

Jaime Spencer, Richard Bowden, Simon Hadfield
Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey
{jaime.spencer, r.bowden, s.hadfield}@surrey.ac.uk

Abstract

In the current monocular depth research, the dominant approach is to employ unsupervised training on large datasets, driven by warped photometric consistency. Such approaches lack robustness and are unable to generalize to challenging domains such as nighttime scenes or adverse weather conditions where assumptions about photometric consistency break down.

We propose DeFeat-Net (Depth & Feature network), an approach to simultaneously learn a cross-domain dense feature representation, alongside a robust depth-estimation framework based on warped feature consistency. The resulting feature representation is learned in an unsupervised manner with no explicit ground-truth correspondences required.

We show that within a single domain, our technique is comparable to both the current state of the art in monocular depth estimation and supervised feature representation learning. However, by simultaneously learning features, depth and motion, our technique is able to generalize to challenging domains, allowing DeFeat-Net to outperform the current state-of-the-art with around 10% reduction in all error measures on more challenging sequences such as nighttime driving.

1. Introduction

Recently there have been many advances in computer vision tasks related to autonomous vehicles, including monocular depth estimation [22, 83, 73] and feature learning [13, 61, 65]. However, as shown in Figure 1, these approaches tend to fail in the most complex scenarios, namely adverse weather and nighttime conditions.

In the case of depth estimation, this is usually due to the assumption of photometric consistency, which starts to break down in dimly-lit environments. Feature learning can overcome such strong photometric assumptions, but these approaches tend to require ground truth pixel-wise correspondences and obtaining this ground truth in cross-seasonal situations is non-trivial. Inconsistencies between GPS measurements and drift from Visual Odometry (VO) makes automatic pointcloud alignment highly inaccurate and manual annotation is costly and time-consuming.

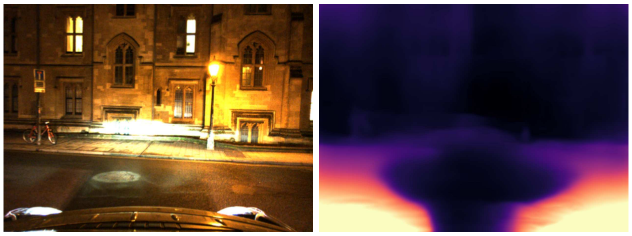

Figure 1. Left: Challenging lighting conditions during nighttime driving. Right: A catastrophic failure during depth map estimation for a current state-of-the-art monocular depth estimation framework, after being trained specifically for this scenario.

We make the observation that depth estimation and feature representation are inherently complementary. The process of estimating the depth for a scene also allows for the computation of ground-truth feature matches between any views of the scene. Meanwhile robust feature spaces are necessary in order to create reliable depth-estimation systems with invariance to lighting and appearance change.

Despite this relationship, all existing approaches tackle these challenges independently. Instead, we propose DeFeat-Net, a system that is capable of jointly learning depth from a single image in addition to a dense feature representation of the world and ego-motion between consecutive frames. What's more, this is achieved in an entirely self-supervised fashion, requiring no ground truth other than a monocular stream of images.

We show how the proposed framework can use the existing relationships between these tasks to complement each other and boost performance in complex environments. As has become commonplace [23], the predicted depth and ego-motion can be used to generate a correspondence map between consecutive images, allowing for the use of photometric error based losses. However, these correspondences can also be used as positive examples in relative metric learning losses [65]. In turn, the learnt features can provide a more robust loss in cases where photometric errors fail, *i.e*. nighttime conditions.

The remainder of the paper provides a more detailed description of the proposed DeFeat-Net framework in the context of previous work. We extensively show the benefits of our joint optimization approach, evaluating on a wide variety of datasets. Finally, we discuss the current state-of-the-art and opportunities for future work. The contributions of this paper can be summarized as:

1. We introduce a framework capable of jointly and simultaneously learning monocular depth, dense feature representations and vehicle ego-motion.
2. This is achieved entirely self-supervised, eliminating the need for costly and unreliable ground truth data collection.
3. We show how the system provides robust depth and invariant features in all weather and lighting conditions, establishing new state-of-the-art performance.

2. Related Work

Here we review some of the most relevant previous work, namely in depth estimation and feature learning.

2.1. Depth Estimation.

Traditionally, depth estimation relied on finding correspondences between every pixel in pairs of images. However, if the images have been stereo rectified, the problem can be reduced to a search for the best match along a single row in the target image, known as disparity estimation. Initial methods for disparity estimation relied on hand-crafted matching techniques based on the Sum of Squared Differences (SSD), smoothness and energy minimization.

Supervised. Ladickỳ [33] and Žbontar [79] showed how learning the matching function can drastically improve the performance of these systems. Mayer *et al*. [46] instead proposed DispNet, a Fully Convolutional Network (FCN) [40] capable of directly predicting the disparity map between two images, which was further extended by [50]. Kendall *et al*. [30] introduced GC-Net, where the disparity is processed as a matching cost-volume in a 3D convolutional network. PSMNet [9] and GA-Net [81] extended these cost-volume networks by introducing Spatial Pooling Pyramid (SPP) features and Local/Semi-Global aggregation layers, respectively.

Estimating depth from a single image seemed like an impossible task without these disparity and perspective cues. However, Saxena [58] showed how it is possible to approximate the geometry of the world based on superpixel segmentation. Each superpixel's 3D position and orientation is estimated using a trained linear model and an MRF. Liu *et al*. [38, 39] improve on this method by instead learning these models using a CNN, while Ladickỳ *et al*. [34] incorporate semantic information as an alternative cue.

Eigen *et al*. [14, 15] introduced the first methods for monocular depth regression using end-to-end deep learning by using a scale-invariant loss. Laina [35] and Cao [7] instead treated the task of monocular estimation as a classification problem and introduced a more robust loss function. Meanwhile, Ummenhofer *et al*. [66] introduced DeMoN, jointly training monocular depth and egomotion in order to perform Structure-from-Motion (SfM). In this paper we go one step further, jointly learning depth, egomotion and the feature space used to support them.

Unsupervised - Stereo Training. In order to circumvent the need for costly ground truth training data, an increasing number of approaches have been proposed using photometric warp errors as a substitute. For instance, DeepStereo [17] synthesizes novel views using raw pixels from arbitrary nearby views. Deep3D [74] also performs novel view synthesis, but restricts this to stereo pairs and introduces a novel image reconstruction loss. Garg [18] and Godard [23] greatly improved the performance of these methods by introducing an additional autoencoder and left-right consistency losses, respectively. UnDeepVO [37] additionally learns monocular VO between consecutive frames by aligning the predicted depth pointclouds and enforcing consistency between both stereo streams. More recently, there have been several approaches making use of GANs [1, 53]. Most notably, [62] uses GANs to perform day-night translation and provide an additional consistency to improve performance in nighttime conditions. However, the lack of any explicit feature learning makes it challenging to generalize across domains.

Unsupervised - Monocular Training. In order to learn unsupervised monocular depth without stereo information, it is necessary to learn a surrogate task that allows for the use of photometric warp losses. Zhou *et al*. [82, 83] introduced some of the first methods to make use of VO estimation to warp the previous and next frames to reconstruct the target view. Zhan [80] later extended this by additionally incorporating a feature based warp loss. Babu *et al*. [3, 44] proposed an unsupervised version of DeMoN [66]. Other published methods are based upon video processing with RNNs [69] and LSTMs [51] or additionally predicting scene motion [67] or optical flow [29, 70, 78].

The current state-of-the-art has been pushed by methods that incorporate additional constraints [68] such as temporal [45], semantic [10], edge & normal [75, 76], cross-task [84] and cycle [52, 73] consistencies. Godard *et al*. [22]

expanded on these methods by incorporating information from the previous frame and using the minimum reprojection error in order to deal with occlusions. They also introduce an automasking process which removes stationary pixels in the target frame. However, they still compute photometric losses in the original RGB colourspace, making it challenging to learn across domains.

2.2. Feature Learning

Hand-Crafted. Initial approaches to feature description typically relied on heuristics based on intensity gradients in the image. Since these were computationally expensive, it became necessary to introduce methods capable of finding interesting points in the image, *i.e*. keypoints. Some of the most well-know methods include SIFT [41] and its variant RootSIFT [2], based on a Difference of Gaussians and Non-Maxima Suppression (NMS) for keypoint detection and HOG descriptors.

Research then focused on improving the speed of these systems. Such is the case with SURF [5], BRIEF [6] and BRISK [36]. ORB features [56] improved the accuracy, robustness and speed of BRIEF [6] and are still widely used.

Sparse Learning. Initial feature learning methods made use of decision trees [55], convex optimization [63] and evolutionary algorithms [31, 32] in order to improve detection reliability and discriminative power. Intelligent Cost functions [24] took this a step further, by using Gaussian Processes to learn appropriate cost functions for optical/scene flow.

Since the widespread use of deep learning, several methods have been proposed to learn feature detection and/or description. Balntas *et al*. [4] introduced a method for learning feature descriptors using in-triplet hard negative mining. LIFT [77] proposes a sequential pipeline consisting of keypoint detection, orientation estimation and feature description, each performed by a separate network. LF-Net [49] builds on this idea, jointly generating dense score and orientation maps without requiring human supervision.

On the other hand, several approaches make use of networks with shared encoder parameters in order to simultaneously learn feature detection and description. Georgakis *et al*. [20] learn 3D interest points using a shared Fast R-CNN [21] encoder. Meanwhile, DeTone introduced SuperPoint [12] where neither decoder has trainable parameters, improving the overall speed and computational cost. More recently, D2-Net [13] proposed a *describe-then-detect* approach where the network produces a dense feature map, from which keypoints are detected using NMS.

Dense Learning. Even though SuperPoint [12] and D2-Net [13] produce dense feature maps, they still focus on the detection of interest points and don't use their features in a dense manner. Weerasekera *et al*. [72] learn dense features in the context of SLAM by minimizing multi-view matching cost-volumes, whereas [60] use generative feature learning with scene completion as an auxiliary task to perform visual localisation.

The Universal Correspondence Network [11] uses optical correspondences to create a pixel-wise version of the contrastive loss. Schmidt [59] instead propose semi-supervised training with correspondences obtained from KinectFusion [47] and DynamicFusion [48] models. Fathy [16] and Spencer [65] extended the pixel-wise contrastive loss to multiple scale features through a coarse-to-fine network and spatial negative mining, respectively. On the other hand, SDC-Net [61] focuses on the design of the network architecture, increasing the receptive field through stacked dilated convolution, and apply the learnt features to optical flow estimation.

In this work we attempt to unify state-of-the-art feature learning with monocular depth and odometry estimation. This is done in such a way that the pixel-wise correspondences from monocular depth estimation can support dense feature learning in the absence of ground-truth labels. Meanwhile, computing match-costs in the learned feature space greatly improves the robustness of the depth estimation in challenging cross-domain scenarios.

3. Methodology

The main objective of DeFeat-Net is to jointly learn monocular depth and dense features in order to provide more robust estimates in adverse weather conditions. By leveraging the synergy between both tasks we are able to do this in a fully self-supervised manner, requiring only a monocular stream of images. Furthermore, as a byproduct of the training losses, the system additionally learns to predict VO between consecutive frames.

Figure 2 shows an overview of DeFeat-Net. Each training sample is composed of a target frame I_t and a set of support frames I_{t+k}, where $k \in \{-1, 1\}$. Using the predicted depth for I_t and the predicted transforms to I_{t+k} we can obtain a series of correspondences between these images, which in turn can be used in the photometric warp and pixel-wise contrastive losses. The code and pre-trained models for this technique will be available at https://github.com/jspenmar/DeFeat-Net.

3.1. Networks

DispNet. Given a single input image, I_t, its corresponding depth map is obtained through

$$D_t = \frac{1}{a\ \Phi_D(I_t) + b}, \tag{1}$$

where a and b scale the final depth to the range $[0.1, 100]$. Φ_D represents the disparity estimation network, formed by a ResNet [25] encoder and decoder with skip connections.

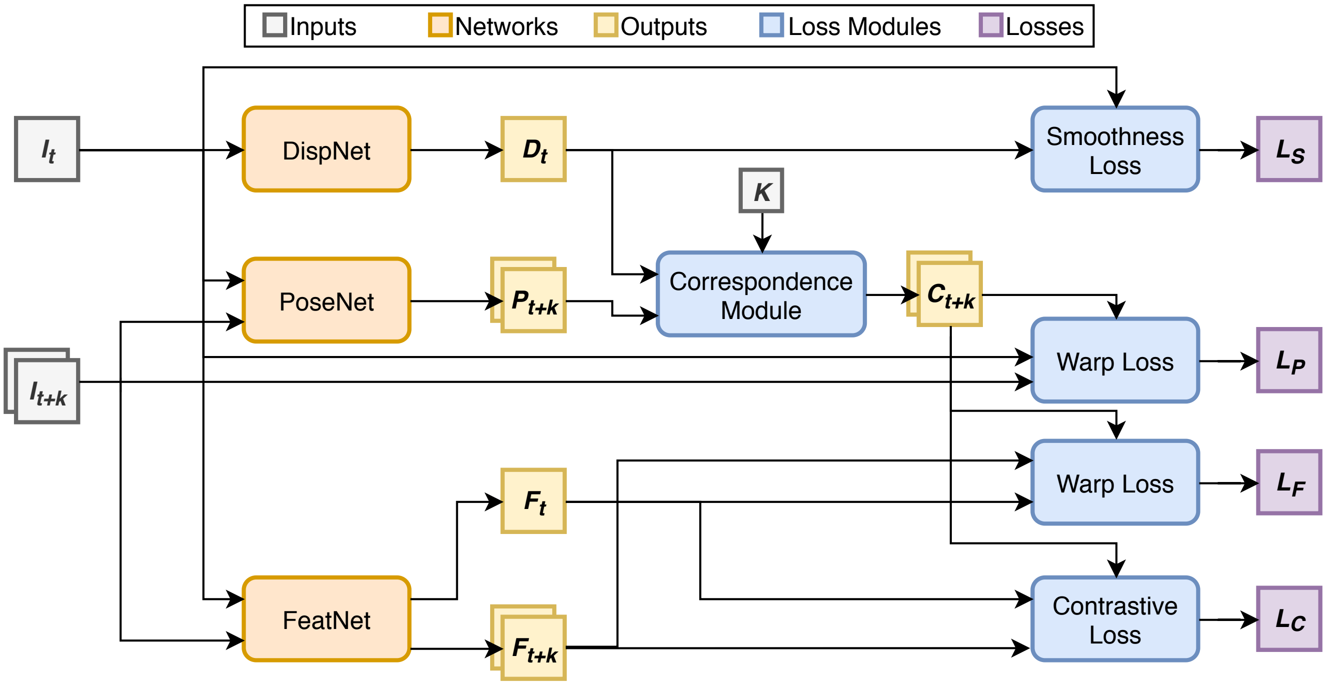

Figure 2. Overview of *DeFeat-Net* which combines complementary networks to simultaneously solve for feature representation, depth and ego-motion. The introduction of feature warping improves the robustness in complex scenarios.

This decoder also produces intermediate disparity maps at each stage, resulting in four different scales.

PoseNet. Similarly, the pose prediction network Φ_P consists of a multi-image ResNet encoder, followed by a 4-layer convolutional decoder. Formally,

$$P_{t\to t+k} = \Phi_P(I_t, I_{t+k}), \tag{2}$$

where $P_{t\to t+k}$ is the predicted transform between the cameras at times t and $t+k$. As in [22, 68] the predicted pose is composed of a rotation in axis-angle representation and a translation vector, scaled by a factor of 0.001.

FeatNet. The final network produces a dense n-dimensional feature map of the given input image, $\Phi_F : \mathbb{N}^{H\times W\times 3} \mapsto \mathbb{R}^{H\times W\times n}$. As such, we define the corresponding L2-normalized feature map as

$$F = ||\Phi_F(I)|| \, . \tag{3}$$

In this case, Φ_F is composed of a residual block encoder-decoder with skip connections, where the final encoder stage is made up of an SPP [9] with four scales.

3.2. Correspondence Module

Using the predicted D_t and $P_{t\to t+k}$ we can obtain a set of pixel-wise correspondences between the target frame and each of the support frames. Given a 2D point in the image p and its homogeneous coordinates $\dot{p}$ we can obtain its corresponding location q in the 3D world through

$$q = \pi^{-1}(\dot{p}) = K_t^{-1}\dot{p}\; D_t(p), \tag{4}$$

where π^{-1} is the backprojection function, K_t is the camera's intrinsics and $D_t(p)$ the depth value at the 2D pixel location estimated using (1).

We can then compute the corresponding point $c_{t\to t+k}$ by projecting the resulting 3D point onto a new image with

$$c_{t\to t+k}(p) = \pi(\dot{q}) = K_t P_{t\to t+k}\dot{q}, \tag{5}$$

where $P_{t\to t+k}$ is the transform to the new coordinate frame, *i.e.* the next or previous camera position from (2). Therefore, the final correspondences map is defined as

$$C_{t\to t+k} = \{c_{t\to t+k}(p) : \forall p\} \, . \tag{6}$$

These correspondences can now be used in order to determine the sampling locations for the photometric warp loss and the positive matches in a pixel-wise contrastive loss to learn an appropriate feature space.

3.3. Losses

Once again, it is worth noting that DeFeat-Net is entirely self-supervised. As such, the only ground truth inputs required are the orginal images and the camera's intrinsics.

Pixel-wise Contrastive. In order to train Φ_F, we make use of the well established pixel-wise contrastive loss [11, 59, 65]. Given two feature vectors from the dense feature maps, $f_1 = F_1(p_1)$ and $f_2 = F_2(p_2)$, the contrastive loss is defined as

$$l(y, f_1, f_2) = \begin{cases} \frac{1}{2}(d)^2 & \text{if } y = 1 \\ \frac{1}{2}\{\max(0, m-d)\}^2 & \text{if } y = 0 \\ 0 & otherwise \end{cases} \tag{7}$$

with y as the label indicating if the pair is a correspondence, $d = ||f_1 - f_2||$ and m the target margin between negative pairs. In this case, the set of positive correspondences is given by $C_{t\to t+k}$. Meanwhile, the negative examples are generated using one of the spatial negative mining techniques from [65].

From both sets, a label mask Y is created indicating if each possible pair of pixels is a positive, negative or should be ignored. As such, the final loss is defined as

$$L_C = \sum_{p_1} \sum_{p_2} l(Y(p_1, p_2), F_t(p_1), F_{t+k}(p_2)). \qquad (8)$$

This loss serves to drive the learning of a dense feature space which enables matching regardless of weather and seasonal appearance variations.

Photometric and Feature Warp. We also use the correspondences in a differentiable bilinear sampler [28] in order to generate the warped support frames and feature maps

$$I_{t+k\to t} = I_{t+k}\langle C_{t\to t+k}\rangle \qquad (9)$$

$$F_{t+k\to t} = F_{t+k}\langle C_{t\to t+k}\rangle \qquad (10)$$

where $\langle\rangle$ is the sampling operator. The final warp losses are a weighted combination of SSIM [71] and L1, defined by

$$\Psi(I_1, I_2) = \alpha\,\frac{1 - SSIM(I_1, I_2)}{2} + (1-\alpha)\ ||I_1 - I_2|| \qquad (11)$$

$$L_P = \Psi(I_t, I_{t+k\to t}), \qquad (12)$$

$$L_F = \Psi(F_t, F_{t+k\to t}), \qquad (13)$$

The photometric loss L_P serves primarily to support the early stages of training when the feature space is still being learned.

Smoothness. As an additional regularizing constraint, we incorporate a smoothness loss [27]. This enforces local smoothness in the predicted depths proportional to the strength of the edge in the original image, ∂I_t. This is defined as

$$L_S = \frac{\lambda}{N} \sum_p |\partial D_t(p)|\, e^{-||\partial I_t(p)||}, \qquad (14)$$

where λ is a scaling factor typically set to 0.001. This loss is designed to avoid smoothing over edges by reducing the weighting in areas of strong intensity gradients.

3.4. Masking & Filtering

Some of the more recent improvements in monocular depth estimation have arisen from explicit edge-case handling [22]. This includes occlusion filtering and the masking of stationary pixels. We apply these automatic procedures to the correspondences used to train both the depth and dense features.

Minimum Reprojection. As the camera capturing the monocular stream moves throughout the scene, various elements will become occluded and disoccluded. In terms of a photometric error based loss, this means that some of the correspondences generated by the system will be invalid. However, when multiple consecutive frames are being used, *i.e.* $k \in \{-1, 1\}$, different occlusions occur in each image.

By making the assumption that the photometric error will be greater in the case where an occlusion is present, we can filter these out by simply propagating the correspondence with the minimum error. This is defined as

$$C_{t\to t+k} = \begin{cases} c_{t\to t-1} & \text{where } \Psi(I_t, I_{t\to t-1}) < \Psi(I_t, I_{t\to t+1}) \\ c_{t\to t+1} & \text{otherwise} \end{cases} \qquad (15)$$

Automasking. Due to the nature of the training method and implicit depth priors (*i.e.* regions further away change less) stationary frames or moving objects can cause holes of infinite depth in the predicted depth maps. An automasking procedure is used to remove these stationary pixels from contributing to the loss,

$$\mu = \left[\min_k \Psi(I_t, I_{t+k}) < \min_k \Psi(I_t, I_{t+k\to t})\right], \qquad (16)$$

where μ is the resulting mask indicating if a correspondence is valid or not and [] is the Iverson bracket. In other words, pixels that exhibit lower photometric error to the unwarped frame than to the warped frame are masked from the cost function.

4. Results

Each subsystem in DeFeat-Net follows a U-Net structure with a ResNet18 encoder pretrained on ImageNet, followed by a 7 layer convolutional decoder similar to [23]. The code and pre-trained models will be available at https://github.com/jspenmar/DeFeat-Net. In all our experiments, the warp loss parameter is set to $\alpha = 0.85$ as per [28].

On the KITTI dataset [19] we follow the Eigen-Zhou evaluation protocol of [23, 83]. This dataset split provides 39,810 training images and 4,424 validation images. These images are all from a single domain (sunny daytime driving).

We also make use of the RobotCar Seasons dataset [57]. This is a curated subset of the larger RobotCar dataset [43], containing 49 sequences. The dataset was specifically chosen to cover a wide variety of seasons and weather conditions, leading to greater diversity in appearance than KITTI.

Unlike the KITTI dataset, which provides sparse ground-truth depth from LiDAR, RobotCar Seasons does not include any depth ground truth. Our proposed technique is unsupervised, and can still be trained on this varied dataset, but the lack of ground truth makes quantitative evaluation on RobotCar Seasons impossible. To resolve this, we returned to the original RobotCar dataset and manualy created a validation dataset comprising of 12,000 images with

Method	Abs-Rel	Sq-Rel	RMSE	RMSE-log	A1	A2	A3
LEGO [75]	0.162	1.352	6.276	0.252	-	-	-
Ranjan [54]	0.148	1.149	5.464	0.226	0.815	0.935	0.973
EPC++ [42]	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2depth (M) [8]	0.141	1.026	5.291	0.215	0.816	0.945	0.979
Monodepth V2 [22]	**0.123**	0.944	5.061	**0.197**	**0.866**	**0.957**	**0.980**
DeFeat	0.126	**0.925**	**5.035**	0.200	0.862	0.954	**0.980**

Table 1. Monocular depth evaluation on the KITTI dataset

Method	μ_+	Global μ_-	Global AUC	Local μ_-	Local AUC
ORB [56]	N/A	N/A	85.83	N/A	84.06
ResNet [26]	8.5117	25.9872	94.77	11.1335	68.26
ResNet-L2	0.341	1.0391	99.25	0.4371	71.80
VGG [64]	4.0077	12.6543	92.94	5.9088	70.03
VGG-L2	0.3905	1.2235	99.57	0.565	77.06
SAND-G [65]	**0.093**	0.746	**99.73**	0.266	87.06
SAND-L	0.156	0.592	98.88	0.505	**94.34**
SAND-GL	0.183	0.996	99.28	0.642	93.34
DeFeat	0.105	1.113	99.10	0.294	83.64

Table 2. Learned feature evaluation on the KITTI dataset

their corresponding ground-truth LiDAR depth maps, split evenly across day and night driving scenarios.

4.1. Single Domain Evaluation

We first evaluate our approach on the KITTI dataset, which covers only a single domain. For evaluation of depth accuracy, we use the standard KITTI evaluation metrics, namely the absolute relative depth error (ABS_REL), the relative square error (SQ_REL) and the root mean square error (RMSE). For these measures, a lower number is better. We also include the inlier ratio measures (A1, A2 and A3) of [23] which measure the fraction of relative depth errors within 25%, 25^2% and 25^3% of the ground truth. For these measures, a larger fraction is better.

To evaluate the quality of the learned feature representations, we follow the protocol of [65]. We compute the average distance in the feature space for the positive pairs from the ground-truth (μ_+), and the negative pairs (μ_-). Naturally a smaller distance between positive pairs, and a larger distance between negative pairs, is best. We also compute the Area Under the Curve (AUC) which can be interpreted as the probability that a randomly chosen negative sample will have a larger distance than the corresponding positive ground truth match. Therefore, higher numbers are better. Following [65] all three errors are split into both local (within 25 pixels) and global measurements.

The results of the depth evaluation are shown in Table 1 and the feature evaluation is shown in Table 2. We can see that in this single-domain scenario, the performance of our technique is competitive with MonodepthV2 and clearly outperforms most other state-of-the-art techniques for monocular depth estimation. The results for [22] were obtained by training a network using the code provided by the authors.

Regarding the features, L2 denotes the L2-normalized versions, whereas G, L & GL represent the different negative mining variants from [65]. We can also see that despite being unsupervised, our learned feature space is competitive with contemporary supervised feature learning techniques and greatly outperforms pretrained features when evaluating locally. It is interesting, however, to note that the simple act of L2-normalizing can improve the global performance of the pretrained features.

Our feature space tends to perform better in the Global evaluation metrics than the local ones. This is unsurprising as the negative samples for the contrastive loss in (7) are obtained globally across the entire image.

4.2. Multi-Domain Evaluation

However, performance in the more challenging RobotCar Seasons dataset demonstrates the real strength of jointly learning both depth and feature representations. RobotCar Seasons covers multiple domains, where traditional photometric based monocular depth algorithms struggle and where a lack of cross-domain ground-truth has historically made feature learning a challenge. For this evaluation, we select the best competing approach from Table 1 (MonodepthV2) and retrain both it and DeFeat-Net on the RobotCar Seasons dataset. All techniques are trained from scratch.

The results are shown in Table 3 and example depth map comparisons are shown in Figure 3. We can see that in this more challenging task, the proposed approach outperforms the previous state of the art technique across all error measures. While for the daytime scenario, the improvements are modest, on the nighttime data there is a significant improvement with around 10% reduction in all error measures.

We believe that the main reason behind this difference is

Test domain	Method	Abs-Rel	Sq-Rel	RMSE	RMSE-log	A1	A2	A3
Day	Monodepth V2 [22]	0.271	3.438	9.268	0.329	**0.600**	0.840	0.932
Day	**DeFeat**	**0.265**	**3.129**	**8.954**	**0.323**	0.597	**0.843**	**0.935**
Night	Monodepth V2 [22]	0.367	4.512	9.270	0.412	0.561	0.790	0.888
Night	**DeFeat**	**0.335**	**4.339**	**9.111**	**0.389**	**0.603**	**0.828**	**0.914**

Table 3. Monocular depth evaluation on the RobotCar dataset

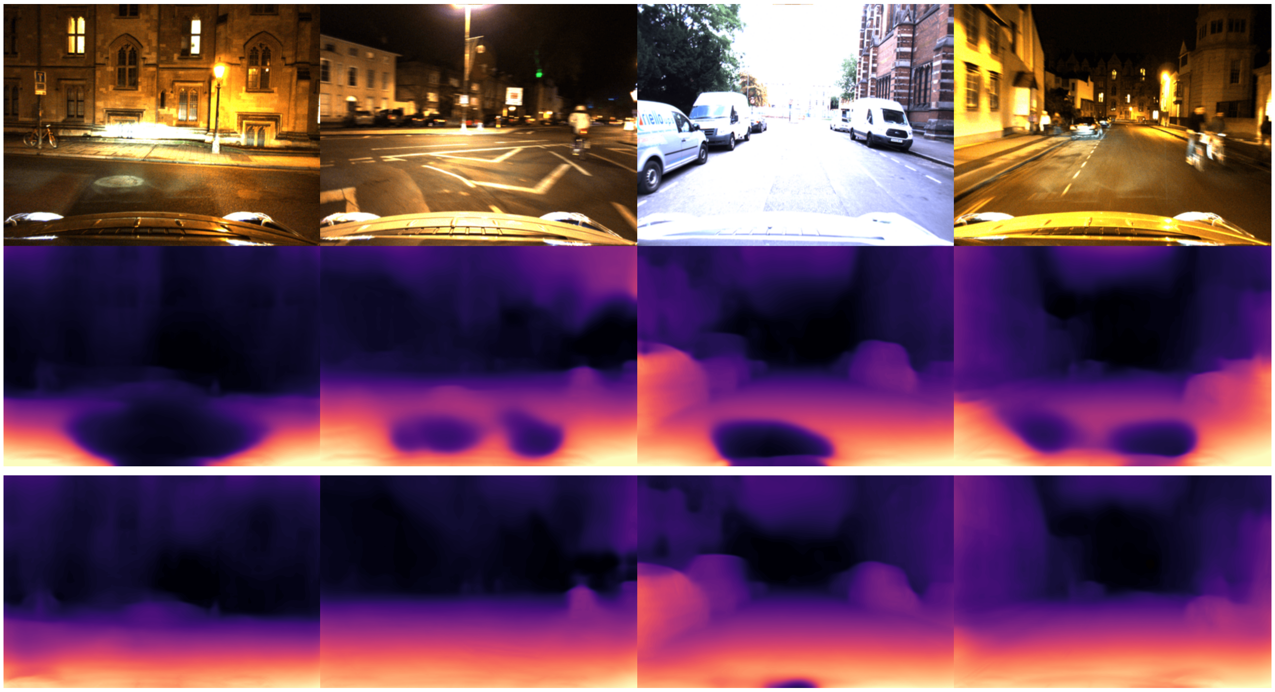

Figure 3. Top: input images from the RobotCar dataset. Middle: estimated depth maps from Monodepth V2 [22]. Bottom: estimated depth maps from *DeFeat-Net*.

that in well-lit conditions, the photometric loss is already a good supervision signal. In this case, incorporating the feature learning adds to the complexity of the task. However, nighttime scenarios make photometric matching less discriminative, leading to weaker supervision. Feature learning provides the much needed invariance and robustness to the loss, leading to the significant increase in performance.

It is interesting to note that the proposed approach is especially robust with regards to the number of estimated outliers. The A1, A2 and A3 error measures are fairly consistent between the day and night scenarios for the proposed technique. This indicates that even in areas of uncertain depth (due to under-exposure and over-saturation), the proposed technique fails gracefully rather than producing catastrophically incorrect estimates.

Since previous state-of-the-art representations cannot be trained unsupervised, and RobotCar Seasons does not provide any ground-truth depth, it is not possible to repeat the feature comparison from Table 2 in the multi-domain scenario. Instead Figure 4 compares qualitative examples of the learned feature spaces. For these visualizations, we find the linear projection that best shows the correlation between the feature map and the images and map it to the RGB color cube. This dimensionality reduction removes a significant amount of discriminative power from the descriptors, but allows for some form of visualization.

In all cases, the feature descriptors can clearly distinguish scene structures such as the road. It is interesting to note that a significant degree of context has been encoded in the features, and they are capable of easily distinguishing a patch in the middle of the road, from one on the left or right, and from a patch of similarly colored pavement. The feature maps trained on the single domain KITTI dataset can sometimes display more contrast than those trained on RobotCar Seasons. Although this implies a greater degree of discrimination between different image regions, this is likely because the latter representation can cover a much broader range of appearances from other domains. Regarding, the nighttime features, it is interesting that those trained on a single domain seem to exhibit strange behaviour around external light sources such as the lampposts, traffic lights and headlights. This is likely due to the bias in the training data, with overall brighter image content.

4.3. Ablation

Finally, for each dataset we explore the benefits of concurrent feature learning, by re-training with the FeatNet subsystem disabled. As shown in Table 4, the removal of

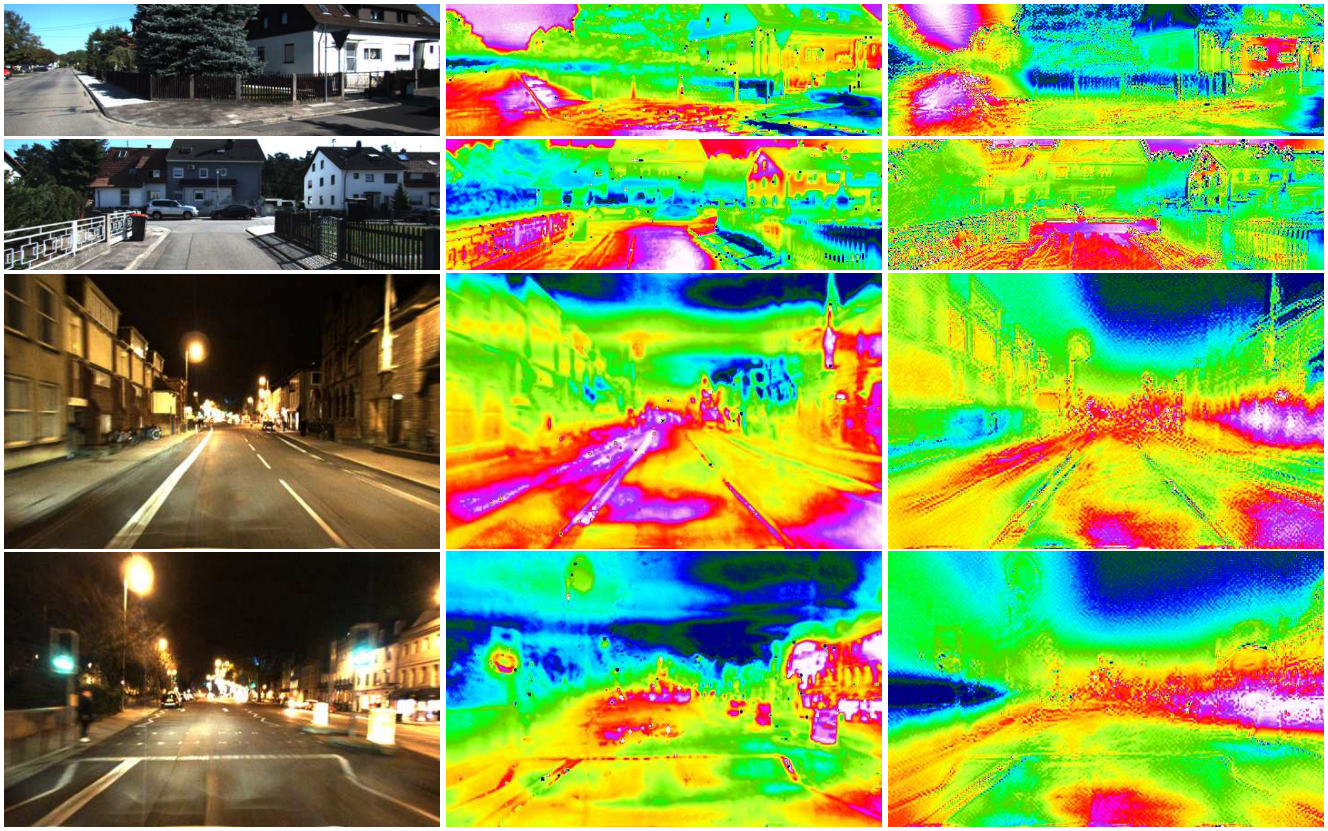

Figure 4. Feature space visualizations for *DeFeat-Net* trained on the single-domain KITTI dataset (centre) and multi-domain RobotCar Seasons dataset (right).

Dataset	Method	Abs-Rel	Sq-Rel	RMSE	RMSE-log	A1	A2	A3
KITTI	DeFeat (no feat)	**0.123**	0.948	5.130	**0.197**	**0.863**	**0.956**	**0.980**
KITTI	**DeFeat**	0.126	**0.925**	**5.035**	0.200	0.862	0.954	**0.980**
RobotCar Day	DeFeat (no feat)	0.274	3.885	**8.953**	0.335	**0.640**	**0.853**	0.934
RobotCar Day	**DeFeat**	**0.265**	**3.129**	8.954	**0.323**	0.597	0.843	**0.935**
RobotCar Night	DeFeat (no feat)	0.748	13.502	**8.956**	0.657	0.393	0.624	0.759
RobotCar Night	**DeFeat**	**0.335**	**4.339**	9.111	**0.389**	**0.603**	**0.828**	**0.914**

Table 4. Performance with and without concurrent feature learning, on each dataset

the concurrent feature learning from our technique causes a small and inconsistent change on the KITTI and RobotCar Day data. However, on the RobotCar Night data, our full approach drastically outperforms the version which does not learn a specialist matching representation. For many error measures, the performance doubles in these challenging scenarios, and the reduction in outliers causes a three-fold reduction in the Sq-Rel error.

These findings reinforce the observation that the frequently used photometric warping loss is insufficient for estimating depth in challenging real-world domains.

5. Conclusions & Future Work

This paper proposed DeFeat-Net, a unified framework for learning robust monocular depth estimation and dense feature representations. Unlike previous techniques, the system is able to function over a wide range of appearance domains, and can perform feature representation learning with no explicit ground truth. This idea of co-training an unsupervised feature representations has potential applications in many areas of computer vision beyond monocular depth estimation.

The main limitation of the current approach is that there is no way to enforce feature consistency across seasons. Although depth estimation and feature matching work robustly within any given season, it is currently unclear weather feature matching between different seasons is possible. It would be interesting in the future to explore cross-domain consistency as an additional training constraint. However, this will necessitate the collection of new datasets with cross seasonal alignments.

Acknowledgements

This work was partially funded by the EPSRC under grant agreements (EP/R512217/1, EP/S016317/1 and EP/S035761/1). We would also like to thank NVIDIA Corporation for their Titan Xp GPU grant.

References

[1] Filippo Aleotti, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Generative adversarial networks for unsupervised monocular depth prediction. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11129 LNCS, pages 337–354, 2019.

[2] Relja Arandjelovic. Three Things Everyone Should Know to Improve Object Retrieval. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2911–2918, 2012.

[3] Madhu Babu, Swagat Kumar, Anima Majumder, and Kaushik Das. UnDEMoN 2.0: Improved Depth and Ego Motion Estimation through Deep Image Sampling. *arXiv preprint*, nov 2018.

[4] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *British Machine Vision Conference 2016, BMVC 2016*, volume 2016-Septe, pages 119.1–119.11, 2016.

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3951 LNCS, pages 404–417, 2006.

[6] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. BRIEF: Binary robust independent elementary features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6314 LNCS, pages 778–792, 2010.

[7] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating Depth from Monocular Images as Classification Using Deep Fully Convolutional Residual Networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11):3174–3182, nov 2016.

[8] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth Prediction without the Sensors: Leveraging Structure for Unsupervised Learning from Monocular Videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:8001–8008, jul 2019.

[9] Jia-Ren Chang and Yong-Sheng Chen. Pyramid Stereo Matching Network. *CVPR*, 2018.

[10] Po-Yi Chen, Alexander H. Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards Scene Understanding: Unsupervised Monocular Depth Estimation With Semantic-Aware Representation. In *CVPR*, pages 2619–2627, 2019.

[11] Christopher B. Choy, Jun Young Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. In *Advances in Neural Information Processing Systems*, pages 2414–2422. Neural information processing systems foundation, 2016.

[12] Daniel Detone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 2018-June, pages 337–349, dec 2018.

[13] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. *CVPR*, may 2019.

[14] David Eigen and Rob 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*, volume 2015 Inter, pages 2650–2658, nov 2015.

[15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, volume 3, pages 2366–2374, 2014.

[16] Mohammed E Fathy, Quoc-Huy Tran, M. Zeeshan Zia, Paul Vernaza, and Manmohan Chandraker. Hierarchical Metric Learning and Matching for 2D and 3D Geometric Correspondences. *ECCV*, 2018.

[17] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deep stereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 5515–5524, jun 2016.

[18] Ravi Garg, B. G. Vijay Kumar, Gustavo Carneiro, and Ian Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9912 LNCS, pages 740–756, mar 2016.

[19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[20] Georgios Georgakis, Srikrishna Karanam, Ziyan Wu, Jan Ernst, and Jana Kosecka. End-to-End Learning of Keypoint Detector and Descriptor for Pose Invariant 3D Matching. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1965–1973, 2018.

[21] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2015 Inter, pages 1440–1448, 2015.

[22] Clément Godard, Oisin Mac Aodha, and Gabriel Brostow. Digging Into Self-Supervised Monocular Depth Estimation. *ICCV*, jun 2019.

[23] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 6602–6611, sep 2017.

[24] Simon Hadfield and Richard Bowden. Scene flow estimation using intelligent cost functions. In *BMVC 2014 - Proceedings of the British Machine Vision Conference 2014*, 2014.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 770–778. IEEE Computer Society, dec 2016.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 770–778, 2016.

[27] Philipp Heise, Sebastian Klose, Brian Jensen, and Alois Knoll. PM-Huber: PatchMatch with Huber Regularization for Stereo Matching. *ICCV*, 2013.

[28] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial Transformer Networks. *NIPS*, jun 2015.

[29] Joel Janai, Fatma Güney, Anurag Ranjan, Michael Black, and Andreas Geiger. Unsupervised Learning of Multi-Frame Optical Flow with Occlusions. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11220 LNCS, pages 713–731, 2018.

[30] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-End Learning of Geometry and Context for Deep Stereo Regression. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-Octob, pages 66–75, mar 2017.

[31] Tomáš Krajník, Pablo Cristóforis, Keerthy Kusumam, Peer Neubert, and Tom Duckett. Image features for visual teach-and-repeat navigation in changing environments. *Robotics and Autonomous Systems*, 88:127–141, feb 2017.

[32] Tomáš Krajník, Pablo Cristóforis, Matías Nitsche, Keerthy Kusumam, and Tom Duckett. Image features and seasons revisited. In *2015 European Conference on Mobile Robots, ECMR 2015 - Proceedings*, pages 1–7. IEEE, sep 2015.

[33] Lubor Ladický, Christian Häne, and Marc Pollefeys. Learning the Matching Function. *arXiv preprint*, feb 2015.

[34] Lubor Ladický, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 89–96, 2014.

[35] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, pages 239–248, jun 2016.

[36] Stefan Leutenegger, Margarita Chli, and Roland Y. Siegwart. BRISK: Binary Robust invariant scalable keypoints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2548–2555, 2011.

[37] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. UnDeepVO: Monocular Visual Odometry Through Unsupervised Deep Learning. In *Proceedings - IEEE International Conference on Robotics and Automation*, pages 7286–7291, sep 2018.

[38] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 5162–5170, nov 2015.

[39] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):2024–2039, feb 2016.

[40] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 3431–3440, 2015.

[41] David G Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[42] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ramkant Nevatia, and Alan Yuille. Every Pixel Counts ++: Joint Learning of Geometry and Motion with 3D Holistic Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, jul 2019.

[43] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 Year , 1000km : The Oxford RobotCar Dataset. *IJRR*, 3(December 2015), 2015.

[44] V. Madhu Babu, Kaushik Das, Anima Majumdar, and Swagat Kumar. UnDEMoN: Unsupervised Deep Network for Depth and Ego-Motion Estimation. In *IEEE International Conference on Intelligent Robots and Systems*, pages 1082–1088, aug 2018.

[45] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, feb 2018.

[46] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 4040–4048, 2016.

[47] Richard A Newcombe. KinectFusion : Real-Time Dense Surface Mapping and Tracking. Technical report, 2013.

[48] Richard A Newcombe, Dieter Fox, and Steven M Seitz. DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time. *Computer Vision and Pattern Recognition (CVPR)*, pages 343–352.

[49] Yuki Ono, Pascal Fua, Eduard Trulls, and Kwang Moo Yi. LF-Net: Learning local features from images. In *Advances in Neural Information Processing Systems*, volume 2018-Decem, pages 6234–6244, 2018.

[50] Jiahao Pang, Wenxiu Sun, Jimmy SJ. Ren, Chengxi Yang, and Qiong Yan. Cascade Residual Learning: A Two-Stage Convolutional Neural Network for Stereo Matching. In *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, volume 2018-Janua, pages 878–886, aug 2018.

[51] Viorica Patraucean, Ankur Handa, and Roberto Cipolla. Spatio-temporal video autoencoder with differentiable memory. *ICLR*, nov 2016.

[52] Andrea Pilzer, Stéphane Lathuilière, Nicu Sebe, and Elisa Ricci. Refine and Distill: Exploiting Cycle-Inconsistency

and Knowledge Distillation for Unsupervised Monocular Depth Estimation. *CVPR*, pages 9768–9777, 2019.

[53] Andrea Pilzer, Dan Xu, Mihai Puscas, Elisa Ricci, and Nicu Sebe. Unsupervised adversarial depth estimation using cycled generative networks. In *Proceedings - 2018 International Conference on 3D Vision, 3DV 2018*, pages 587–595, jul 2018.

[54] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J. Black. Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation. *CVPR*, pages 12240–12249, 2019.

[55] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3951 LNCS, pages 430–443, 2006.

[56] Ethan Rublee and Gary Bradski. ORB: an efcient alternative to SIFT or SURF. *ICCV*, 2011.

[57] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. *CVPR*, 2018.

[58] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3D: Learning 3D scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840, may 2009.

[59] Tanner Schmidt, Richard Newcombe, and Dieter Fox. Self-Supervised Visual Descriptor Learning for Dense Correspondence. *IEEE Robotics and Automation Letters*, 2(2):420–427, 2017.

[60] Johannes L. Schonberger, Marc Pollefeys, Andreas Geiger, and Torsten Sattler. Semantic Visual Localization. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6896–6906, 2018.

[61] René Schuster, Oliver Wasenmüller, Christian Unger, and Didier Stricker. SDC - Stacked Dilated Convolution: A Unified Descriptor Network for Dense Matching Tasks. *CVPR*, 2019.

[62] Aashish Sharma, Robby T. Tan, and Loong-Fah Cheong. Depth Estimation in Nighttime using Stereo-Consistent Cyclic Translations. *arXiv preprint*, sep 2019.

[63] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Descriptor learning using convex optimisation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7572 LNCS, pages 243–256, 2012.

[64] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[65] Jaime Spencer, Richard Bowden, and Simon Hadfield. Scale-Adaptive Neural Dense Features : Learning via Hierarchical Context Aggregation. *CVPR*, 2019.

[66] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. DeMoN: Depth and motion network for learning monocular stereo. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 5622–5631, 2017.

[67] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. SfM-Net: Learning of Structure and Motion from Video. *arXiv preprint*, apr 2017.

[68] Chaoyang Wang, Jose Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning Depth from Monocular Videos Using Direct Methods. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, nov 2018.

[69] Rui Wang, Stephen M. Pizer, and Jan-Michael Frahm. Recurrent Neural Network for (Un-)supervised Learning of Monocular VideoVisual Odometry and Depth. *CVPR*, pages 5555–5564, 2019.

[70] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. UnOS: Unified Unsupervised Optical-flow and Stereo-depth Estimation by Watching Videos. Technical report, 2019.

[71] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[72] Chamara Saroj Weerasekera, Ravi Garg, Yasir Latif, and Ian Reid. Learning Deeply Supervised Good Features to Match for Dense Monocular Reconstruction. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11365 LNCS, pages 609–624. Springer Verlag, 2019.

[73] Alex Wong, Byung-Woo Hong, and Stefano Soatto. Bilateral Cyclic Constraint and Adaptive Regularization for Unsupervised Monocular Depth Prediction. *CVPR*, pages 5644–5653, 2019.

[74] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9908 LNCS, pages 842–857, apr 2016.

[75] Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. LEGO: Learning Edge with Geometry all at Once by Watching Videos. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 225–234, mar 2018.

[76] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 7493–7500, nov 2018.

[77] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: Learned invariant feature transform. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9910 LNCS, pages 467–483, 2016.

[78] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1983–1992. IEEE Computer Society, dec 2018.

[79] Jure Žbontar and Yann Lecun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17, oct 2016.

[80] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction. *CVPR*, 2018.

[81] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip H. S. Torr. GA-Net: Guided Aggregation Net for End-to-end Stereo Matching. *CVPR*, apr 2019.

[82] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 6612–6621, 2017.

[83] Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qixing Huang, and Alexei A. Efros. Learning dense correspondence via 3D-guided cycle consistency. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 117–126, apr 2016.

[84] Yuliang Zou, Zelun Luo, and Jia Bin Huang. DF-Net: Unsupervised Joint Learning of Depth and Flow Using Cross-Task Consistency. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11209 LNCS, pages 38–55, sep 2018.