

Unsupervised Monocular Depth Estimation for Night-time Images using Adversarial Domain Feature Adaptation

Madhu Vankadari¹, Sourav Garg², Anima Majumder¹, Swagat Kumar^{3,1}*,
and Ardhendu Behera³

¹ TATA Consultancy Services, India

{`madhu.vankadari,anima.majumder`}@tcs.com

² Queensland University of Technology, Australia

{`s.garg`}@qut.edu.au

³ Edge Hill University, U.K

{`swagat.kumar,beheraa`}@edgehill.ac.uk

Abstract. In this paper, we look into the problem of estimating per-pixel depth maps from unconstrained RGB monocular *night-time* images which is a difficult task that has not been addressed adequately in the literature. The state-of-the-art day-time depth estimation methods fail miserably when tested with night-time images due to a large domain shift between them. The usual photometric losses used for training these networks may not work for night-time images due to the absence of uniform lighting which is commonly present in day-time images, making it a difficult problem to solve. We propose to solve this problem by posing it as a domain adaptation problem where a network trained with day-time images is adapted to work for night-time images. Specifically, an encoder is trained to generate features from night-time images that are indistinguishable from those obtained from day-time images by using a PatchGAN-based adversarial discriminative learning method. Unlike the existing methods that directly adapt depth prediction (network output), we propose to adapt *feature maps* obtained from the encoder network so that a pre-trained day-time depth decoder can be directly used for predicting depth from these *adapted features*. Hence, the resulting method is termed as “Adversarial Domain Feature Adaptation (ADFA)” and its efficacy is demonstrated through experimentation on the challenging Oxford night driving dataset. To the best of our knowledge, this work is a first of its kind to estimate depth from unconstrained night-time monocular RGB images that uses a completely unsupervised learning process. The modular encoder-decoder architecture for the proposed ADFA method allows us to use the encoder module as a feature extractor which can be used in many other applications. One such application is demonstrated where the features obtained from our adapted encoder network are shown to outperform other state-of-the-art methods in a visual place recognition problem, thereby, further establishing the usefulness and effectiveness of the proposed approach.

* SK will like to thank NVIDIA GPU Grant Program for their support.



Fig. 1: The depth predictions of the proposed method on Oxford Night driving images. Top to bottom: (a) Input RGB night-time image. (b) Corresponding ground truth depth map generated from the LIDAR points. (c) The depth predictions using the proposed method

1 Introduction

Estimating depth from RGB images is a challenging problem which finds applications in a wide range of fields such as augmented reality [30], 3D reconstruction [16], self-driving cars [19], place recognition [11], etc. The recent success of deep learning methods has spurred the research in this field leading to the creation of several new benchmarks that now outperform traditional methods which rely on handcrafted features and exploit camera geometry and/or camera motion for depth and pose estimation from monocular or stereo sequence of images (video). These learning methods can be broadly classified into two categories: supervised and unsupervised. The supervised learning methods [10] [7] necessitate explicit availability of ground truth information (Laser or LiDAR range data) which may not always be feasible in many real-world scenarios. This is overcome by the unsupervised methods [42] [39] [5] that harness the spatial and/or temporal consistency present in image sequences to extract the underlying geometry to be used as the implicit supervision signal required for training the models. Many of these methods were shown to provide very impressive results on several popular datasets such as KITTI [15] and Cityscapes [9] containing only day-time images. In contrast, there are a very few works that aim to solve the night-time depth estimation problem, which is comparatively more challenging owing to factors such as low visibility and non-uniform illumination arising from multiple (street lights, traffic lights) and possibly, moving light sources (car headlights). For instance, authors in [22] exploit the inherent motion component in burst shot (several successive shots with varying camera settings, also known as “auto-bracketing”) to estimate depth from images taken under low-light condition. Similarly, Zhu et al. [43] present a deep learning based method for estimating motion flow, depth and pose from images obtained from event cameras that return a time-stamped event tuple whenever a change in pixel intensity is detected. In another work, Kim et al. [24] propose a deep network for estimating depth from thermal images taken

during the night time. To the best of our knowledge, there is no reported work that addresses the problem of estimating depth and pose directly from a single ordinary RGB monocular night-time image. The deep learning models trained on day-time monocular [42] or stereo images [5] fail miserably on night-time images due to the inherent large domain shift between these images. The domain shift refers to the change from day-time conditions (well-lit and uniform illumination) to night-time conditions comprising low illumination/visibility with non-uniform illumination caused by unpredictable appearance and disappearance of multiple point-light sources (e.g., street lamps or car headlights, etc.).

One possible solution will be to apply image-to-image translation methods, such as Cycle-GAN [44] or MUNIT [21], to map night-time images to day-time images and then use a pre-trained day-time depth model to estimate depth from these translated images. Some of these image translation techniques have been used in the context of night-time images. For instance, the authors in [2] use night-to-day image translation for solving the place recognition problem required for localization. Similarly, authors in [4] [40] explore image translation techniques to generate synthetic labeled data to reduce the requirement of real-world images for training depth estimation models. Many of these models trained on simulated images do not generalize well to natural images due to the inherent domain shift and hence, employ several domain adaptation techniques to improve their applicability to real-world situations [4] [40] [31]. These approaches have several limitations. For instance, many of these methods use two different deep networks - one for image translation and another for depth estimation, making it computationally heavy and with possibly, inferior performance due to the cascading error effect of using two models in a cascade. Since the image translation module is trained independent of the depth network module, it may not learn depth-specific attributes required for preserving structural information during image translation. This may, in turn, introduce artifacts which might not be understood by the depth estimation module leading to poor depth prediction for the input night-time image. Secondly, it is difficult to generate synthetic night-time images that can capture all the vagaries of real-world night conditions as one can observe in the Synthia dataset [33]. Many of the simulated night-time images in this dataset appear almost like day-time images and using them for night-time depth prediction may not give desired results. Finally, these methods have been applied so far to day-time images for depth estimation.

In this paper, we propose a PatchGAN-based domain adaptation technique for estimating depth from monocular night images by using a single encoder-decoder type deep network model. Specifically, an encoder network is trained to generate night-time *features* which are indistinguishable from those obtained from day-time images. This is achieved by using an adversarial discriminative learning [36] that uses day-time encoded features as the reference. These adapted night features could then be used directly with a decoder network pre-trained on day-time images for depth estimation. Since the domain *features* are adapted through adversarial learning, this method is termed as “*Adversarial Domain Feature Adaptation (ADFA)*” method to distinguish it from other methods that

attempt to adapt depth predictions directly [4] [40] [31]. PatchGAN networks [23] [37] have been shown to provide superior performance compared to conventional GANs by capturing high frequency local structural information and hence, form a natural choice of GAN architecture for the proposed method.

The resulting outcome of our approach is shown qualitatively in Figure 1. We are able to obtain reliable depth maps shown in Figure 1(c) from monocular night-time images shown in Figure 1(a). This is also evident from the interpolated ground-truth depth maps obtained from the LIDAR point clouds as shown in Figure 1(b). The efficacy of the proposed approach is demonstrated by applying it to the challenging Oxford night-time driving dataset [29]. The modular encoder-decoder architecture provides the flexibility of using the encoder module as a feature extractor to extract or select useful features from input images. Such feature extractors are used in several applications such as pose estimation [18], Visual Place Recognition (VPR)[12][11], object detection [41] and segmentation [6]. We demonstrate one such application where the adapted features obtained from our encoder module are shown to provide superior place recognition accuracy compared to other state-of-the-art feature representations available in the literature.

In short, the main contributions made in this paper may be summarized as follows:

- We propose a novel PatchGAN-based domain *feature* adaptation method for estimating depth from unconstrained monocular night-time RGB images, which is considered to be more difficult compared to day-time images. To the best of our knowledge, this is the first instance where adversarial discriminative domain feature adaptation is being used for estimating depth from unconstrained night-time monocular RGB images and this may act as a stepping-stone for future research in this field.
- We also propose an image translation-based method for night-time depth estimation by using a combination of an image translating network (e.g. CycleGAN [44]) and a standard day-time depth estimation network (such as [18]) in cascade. This serves to highlight the difficulties involved in such methods and hence, provides a strong motivation in favour of the proposed work.
- The usefulness and effectiveness of our method is further established by demonstrating that the features obtained using the proposed ADFA method outperform other state-of-the-art feature representations in a visual place recognition problem.

Rest of this paper is organized as follows. The proposed method is described in the next section. The experimental evaluation of our approach on various datasets is discussed in Section 3. The concluding remarks and future scope of this work is presented in Section 4. Our code will be made available at <https://github.com/madhubabuv/NightDepthADFA>

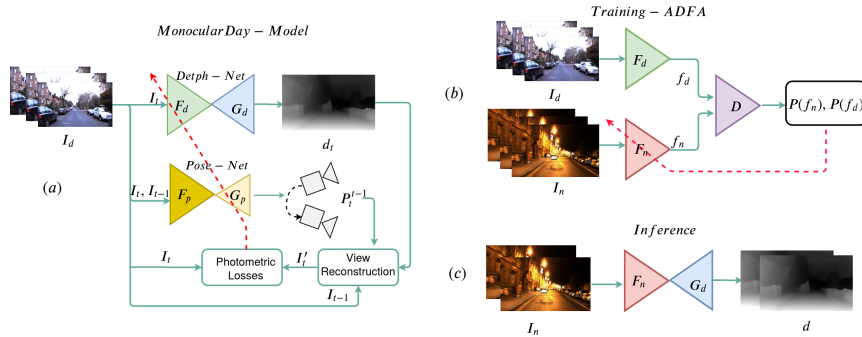


Fig. 2: Architectural overview of the proposed method. (a) The monocular day-model consists of a Depth-Net (F_d, G_d) and a Pose-Net (F_p, G_p) to predict per-pixel depth-map d_t and 6-DoF pose P_t^{t-1} , respectively. The day-model is trained using photometric losses calculated from images reconstructed from view-reconstruction module. (b) A new encoder F_n takes a night-time image I_n and predicts f_n . The f_n features are adapted to look-like day features f_d using adversarial learning with a Patch-GAN based discriminator D . The red dotted line is drawn to indicate modules trained using back propagation. (c) Finally, the new-encoder F_n and the day depth-decoder G_d are used together to predict depth for night-time images

2 Proposed method

We propose to solve the depth estimation problem for night-time images by posing it as a domain adaption problem in which a model pre-trained on day-time images is adapted to work for night-time images as well. The overall approach is shown in Figure 2. It consists of three steps. First, an encoder-decoder type deep network model (F_d, G_d) is trained on day-time images to estimate depth directly from RGB images by using one of the existing methods as in [18], [37], [39], [17], [42]. This is shown in Figure 2(a). The second step involves training a new image encoder F_n with night-time images using adversarial discriminative learning that uses F_d as the generator. This is shown in Figure 2(b). The third and the final step involves using the new encoder F_n in conjunction with the day-time decoder G_d for estimating depth directly from night-time images as shown in Figure 2(c).

The above three components of the proposed ADFA method are described in detail in the following subsections.

2.1 Learning F_d and G_d from day-time images

Estimating depth from monocular day-time images is an active field of research where deep learning methods have been applied successfully and several new benchmarks have been reported in the literature [10], [7], [28], [42], [39], [37], [5], [27]. These deep networks have an encoder-decoder type architecture as shown

in Figure 2(a). Such an architecture allows us to decompose the entire pipeline into two sub-networks, one for encoding (or extracting) features from input images and another for mapping these features to depth information. In unsupervised methods, the image reconstruction error is used as the loss function for training the entire model thereby avoiding the necessity of having the explicit ground truth depth information. The images are reconstructed by using spatial and/or temporal cues obtained from stereo or monocular sequence of images. The methods that use only temporal cues (such as optical flow) incorporate an additional network to estimate pose or ego motion required for image reconstruction [42],[39]. The Depth-Net as shown in Figure 2(a) is composed of a series of convolutional and deconvolutional layers with different filter sizes. Given a monocular day-time image I_d , the image encoder F_d generates, say, L number of convolutional feature maps with different shapes and sizes, one from each layer. This feature map is represented as $F_d(I_d) = f_d = \{f_d^i\}$, $i = 1, 2, \dots, L$, where L is the total number of convolutional layers used in the image encoder. These feature maps are then passed to a depth-decoder G_d to predict per-pixel depth map \mathcal{D} of the input image I_d . One can use any of the existing methods (supervised or unsupervised) to learn the functions F_d and G_d . In this work, we have used the state-of-the art depth-net model [18] as our F_d and G_d which are trained on the day-time monocular images. Since only monocular sequence of images are used for training, an additional pose network is required to estimate ego motion of the camera required for reconstructing images in the temporal domain. The encoder network F_d is used to train a new encoder F_n for night-images using an adversarial learning as explained in the next section.

2.2 Learning F_n using night-time images

Once the day-time image encoder F_d and depth decoder G_d are learned, our objective is to learn an image encoder F_n that can generate the features maps f_n from a night-time image I_n which are indistinguishable from the day-time feature maps f_d obtained from the day-time encoder F_d . There is no direct supervision signal available for computing the loss function from f_d and f_n as the input day and night images are *unpaired*. Here, the term *unpaired* means that these two images are not taken at the same time or at the same place. The encoder F_n is trained to reduce the distance between the distributions of day and night feature spaces by using an adversarial training approach proposed in [36]. In this approach, the image encoder F_n acts as a *generator* trying to generate feature maps from a night image I_n , which look similar to the day-time feature maps f_d obtained from a day-time image I_d using a day-time encoder F_d . These generated features maps are then evaluated by a *discriminator* network D that tries not to get fooled by the generator by assigning correct labels to them. In this way, the generator learns to generate day-like feature maps from the night-time images by playing a zero-sum min-max game with the discriminator.

Unlike a regular GAN discriminator which assigns a single scalar value for a given input, a patch-based discriminator [23] assigns a grid of $m \times n$ scalar values for a given feature map. Each value of this grid is a probability ranging from 0

(night) to 1 (day) and it corresponds to a patch of the input feature map. This allows the discriminator to evaluate the input feature maps locally thereby, providing superior distinguishing ability compared to normal GAN discriminators. In addition, the patch-based discriminators are fully convolutional and hence, are computationally much faster compared to the other discriminator models that use fully-connected layers along with the convolutional layers [37].

Instead of training a single discriminator network on the feature maps obtained from the final convolutional layer of the image encoder as is done in [31][36], we train multiple discriminators, one for each layer of the encoder network to constrain the solution space further. Hence, the proposed multi-stage patch-based discriminator is composed of L number of discriminators where each discriminator D_i , takes feature maps (f_n^i, f_d^i) obtained from the i -th convolutional layer of the encoder networks (F_n, F_d) as input. This multi-stage discriminator is shown to provide superior domain adaptation performance which will be discussed later in the experiments section.

2.3 Training Losses

The proposed method is an unsupervised learning approach which neither uses any explicit ground truth nor paired day-night image examples to calculate losses for training. Instead, we entirely rely on adversarial losses calculated using the discriminator module. The loss functions to learn F_n and D can be expressed as follows:

$$\mathcal{L}_{GAN}(F_n, D) = \min_{F_n} \max_D V(F_n, D) = \mathbb{E}_{f_d \sim F_d(I_d)} [\log(D(f_d))] + \mathbb{E}_{f_n \sim F_n(I_n)} [\log(1 - D(f_n))] \quad (1)$$

$$\min_{F_n} L_{F_n}(F_n, D, I_n) = \frac{1}{L} \sum_{i=1}^L - \mathbb{E}_{f_n \sim F_n(I_n)} \left[\sum_{m,n} \log [D_i(f_n^i)]_{m,n} \right] \quad (2)$$

$$\begin{aligned} \min_D L_D(F_d, F_n, D, I_d, I_n) &= \frac{1}{L} \sum_{i=1}^L - \mathbb{E}_{f_d \sim F_d(I_d)} \left[\sum_{m,n} \log [D_i(f_d^i)]_{m,n} \right] \\ &\quad - \mathbb{E}_{f_n \sim F_n(I_n)} \left[\sum_{m,n} \log \left(1 - [D_i(f_n^i)]_{m,n} \right) \right] \quad (3) \end{aligned}$$

The details about our experimental setup and various experiments conducted are explained in the following section.

3 Experiments and Results

In this section, we provide various experimental results to establish the efficacy of the proposed method for estimating depth from night-time monocular RGB images. We use the publicly available Oxford Robotcar dataset [29] for evaluating the performance of our method. This dataset is used to perform two sets of

experiments. The first experiment is carried out to analyze the depth estimation performance of the proposed method while the second experiment is performed to demonstrate the flexibility of using the encoder for solving a Visual Place Recognition (VPR) problem. The overview of dataset used and the details of experiments performed are described next in this section.

3.1 Oxford Robotcar Dataset : Training and Testing Data Setup

Oxford RobotCar dataset [29] is a popular outdoor-driving dataset comprising of images collected during different seasons, weather conditions and at different timings of day and night. The data collection is carried out over a period of one year by setting cameras in all the four directions. The images captured from the front-view stereo cameras are of resolution 1280×960 . We have used the left images of the front stereo-camera (Bumblebee XB3) data from the sequences captured on 2014-12-16-18-44-24 for night-time and 2014-12-09-13-21-02 for day-time images for depth estimation. The training is performed on the images from the first 5 splits of the day and night-time sequences after cropping the car-hood from the images and downscaling them to 256×512 . The static images where the car has stopped at signals are not considered for the experiments and thus, the total number of images left for training is close to 20,000. We have randomly sampled a total of 498 images for testing from the 6th split of night-driving sequence.

For VPR, we have used day and night sequences as 2014-12-09-13-21-02 and 2014-12-10-18-10-50 respectively from the Oxford Robotcar dataset, where the query (night) sequence is different from that used in the network training. We only used the first 6000 *stereo-left* image frames from each of these traverses which were uniformly sub-sampled using the GPS data to maintain consecutive frame distance of approximately 2 meters. The day traverse is used as the reference traverse against which each of the query (night) image representations is compared with Euclidean distance to retrieve the best match. The night images do not overlap geographically with the night data used for training the model employed for feature extraction for VPR experiments. The evaluation is done using the GPS data by calculating the recall rate for the localization radius varying between 0 to 100 meters. Here, recall rate is defined as the ratio of correctly matched images within the given radius of localization to the total number of query images.

3.2 Experimental setup

The proposed method is implemented using TensorFlow [1]. The network is trained for 40 epochs using a GTX 1080 Alienware-R4 laptop. The learning rate is initially set to 0.0001, then it is reduced by half after $3/5^{th}$ of the total iterations and finally, it is further reduced by half after $4/5^{th}$ of the total iterations. Leaky Relu [38] is used as an activation function in all the layers, except in disparity prediction layers. The predicted disparity is normalized to have the maximum disparity as 30 percent of the input image width by using sigmoid

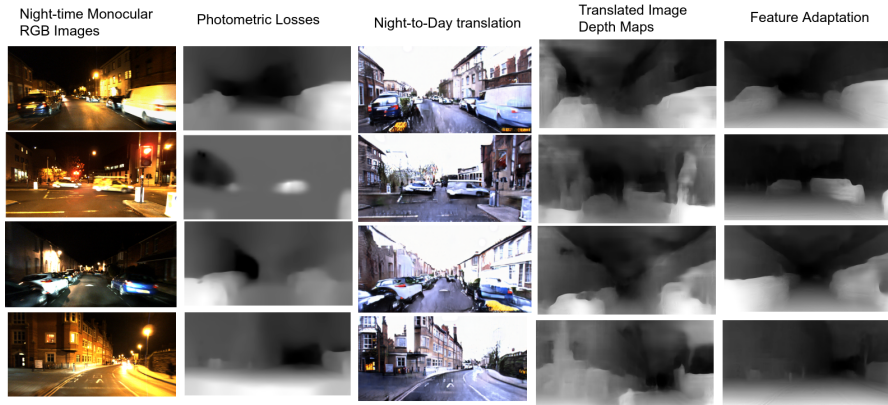


Fig. 3: A qualitative comparison of predicted depth-maps with different experiments. The first column shows the night-time images which are provided as input to different networks. The second column shows the output depth images obtained using photometric losses. As one can observe, these methods fail to maintain the structural layout of the scene. The third column shows the output of an image-translation network (CycleGAN) which are then applied to a day-depth estimation network to obtain depth-maps as shown in the fourth column. These are slightly better compared to the previous case but it introduces several artifacts which degrade the depth estimation in several cases. The last column shows the predictions using the proposed ADFA approach. As one can see, the proposed method provides better predictions compared to these methods and is capable of preserving structural attributes of the scene to a greater extent

as activation function while learning the day-time depth estimation model. The network is trained using the Adam [25] optimizer. Two major experimental studies, one for depth estimation and another for visual place recognition, are carried out under extreme photometric variations using the Oxford dataset [29]. Both the experimental studies along with the qualitative and quantitative analyses are presented below.

3.3 Study 1: Depth Evaluation

In this study, we perform several experiments to establish the efficacy of our proposed method for estimating depth from monocular night-time images. The summary of these experiments is provided in Table 1. The first row of this table shows the outcome of our first experiment where we train a monocular version of Monodepth2 [18] network on Oxford day-time images and then, test it on Oxford night-time images. As expected, the day-time trained model performs poorly on night-time images because of the inherent domain shift present between day-time and night-time images. The second row shows the outcome of another experiment where the same network is trained on the Oxford night-time images and then, tested on a different set of night-time images (test-split). The performance in this case is better than the first experiment but still not good

Table 1: A quantitative performance comparison analysis for depth estimation from night-time images. The top split of the table is evaluated with 60 meters and the lower is evaluate with 40 meters as the maximum depth-range. Higher value is better for the blue color labeled cells and lower value is better for the rest

Method	Error Metric ↓				Accuracy Metric ↑		
	Abs Rel	Sq Rel	RMSE	logRMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [18] (Day)	0.7221	11.5155	14.253	0.663	0.252	0.467	0.644
Monodepth2 [18] (Night)	0.3990	38.8965	23.596	0.408	0.482	0.760	0.894
Cycle-GAN [44]	0.7587	12.7944	13.681	0.663	0.277	0.503	0.688
ADFA (with KITTI)	0.3589	5.1174	11.611	0.384	0.424	0.730	0.914
ADFA (with Oxford)	0.2327	3.783	10.089	0.319	0.668	0.844	0.924
Monodepth2 [18] (Day)	0.6108	6.9513	9.945	0.592	0.267	0.502	0.695
Monodepth2 [18] (Night)	0.2921	7.5395	10.686	0.332	0.588	0.829	0.932
Cycle-GAN [44]	0.6497	7.9346	9.521	0.596	0.298	0.546	0.740
ADFA (with KITTI)	0.2984	3.2349	7.801	0.328	0.495	0.833	0.942
ADFA (with Oxford)	0.2005	2.5750	7.172	0.278	0.735	0.883	0.942

enough as the presence of temporal intensity gradient makes it difficult to use the existing photometric losses for training the network. The third row of this table shows the outcome of yet another experiment where we use image translation for depth estimation. In this approach, we use Cycle-GAN [44] for translating night-time Oxford images into day-time images and then use a day-time trained Monodepth2 model for estimating depth from these translated images. The performance of this approach is similar to the above methods (worse in terms of ‘Abs Rel’ metric and better in terms of ‘RMSE’ metric) indicating that image translation is not adequate for solving the night-time depth estimation problem. Moreover, it is a computationally expensive method that uses two independent networks in cascade unlike the above methods that use only one network for this task. We now apply our proposed ADFA method to adapt the depth model used in the first experiment above and the outcome is shown in the fifth row of this table. As one can see, it provides significant improvement over the previous three approaches, thereby establishing the superiority of our approach. In this case, day-time encoder-decoder pair (F_d , G_d) and night-time encoder (F_n) are trained using images from Oxford dataset and then tested using night-time images from the same dataset. We also perform another experiment where the day-time encoder-decoder network (F_d , G_d) is trained on the KITTI dataset, but the night-time encoder (F_n) is trained and then tested on night-time Oxford images. The corresponding result is shown in the fourth row and is labeled as ‘ADFA (with KITTI)’. While its performance is worse than ADFA (Oxford), it is better than all other methods mentioned above. It is worth to mention that this is an extreme case of domain adaption where not only there is a domain variation from day to night, but also a place variation from KITTI to Oxford. It only demonstrates the resilience of our approach whose performance degrades gracefully in the face of this extreme domain variation. Even though Monodepth2 [18]

Table 2: Ablation study to determine the number of day-encoder convolutional layers to be used during the adversarial learning. The best performance is achieved by skipping the first two layers (without cnv-1,2) features

Method	Abs Rel	Sq Rel	RMSE	logRMSE
Full conv layers	0.2071	2.8971	7.619	0.282
without cnv-1	0.2038	2.7908	7.461	0.280
without cnv-1,2	0.2005	2.575	7.172	0.278
without cnv-1,2,3	0.2260	2.574	7.283	0.300

model has been used as our base network architecture for providing the above performance analysis, ADFA is a generic approach which could be applied to any other deep network model with similar effect.

A qualitative performance comparison of these methods is shown in Figure 3. The first column shows the input night-time images selected randomly from the test set. The second column shows the depth estimation results obtained by using methods such as Monodepth2 [18] that use photometric losses for training. The third column shows the images obtained after image translation by using methods such as Cycle-GAN [44]. The fourth column shows the depth map obtained from these translated images by a pre-trained day-time depth network. We can clearly see that image translation introduces several artifacts leading to poor depth estimation results. The last column shows the depth prediction results obtained by using our proposed ADFA method. One can clearly notice the improvements achieved through our proposed domain feature adaptation method.

The front LMS laser sensor data with INS data is used to prepare the ground-truth needed for testing images using the official code-base released with the dataset. The maximum depth range is set to 60m in the first half of the Table 1 and changed to 40m in the second half. The scale is calculated using the ground-truth depth data, as it is done in [42,39]. In addition, an ablation study is carried out to determine the optimal number of night-time encoder F_n layers to be constrained for the best performance and the results are shown in Table 2. We observed that a model trained by skipping the first two layers of the day-encoder gives the best-performance and the same model is used to report the final results.

To the best of our knowledge, the proposed work is the first attempt at solving the depth estimation problem for unconstrained night-time monocular images for which no priors are available in the literature. However, there are some cases, shown in the Figure 4, where the model is observed to provide poor or failed prediction results. Some of the failure cases include night-time images with very low-illumination conditions, blurred image regions and saturated regions (bright light spots). It is also difficult for our method to deal with small and narrow structures such as traffic poles. The failure case with low-illuminated night-time images could be due to the absence of such extreme conditions in day-time images on which the day encoder-decoder model is trained. The problems

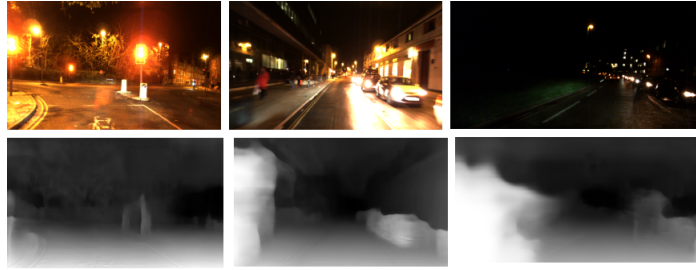


Fig. 4: Failure cases of the proposed depth prediction approach. The model is not able to predict accurate depth for blurred image regions, traffic signal poles and very low illuminated regions of the image

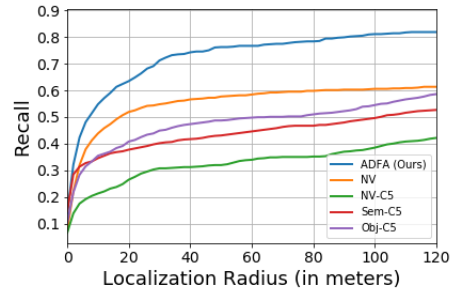


Fig. 5: Visual Place Recognition Performance Benchmark: It can be observed that the feature representations derived from our depth encoder perform the best as compared to other approaches

associated with small structures could be dealt by incorporating some semantic information (if available) into the training data. These limitations will provide a fertile ground for further research in this field.

3.4 Study 2: Visual Place Recognition: Day versus Night

The depth estimation network trained using our proposed approach is able to learn appearance-robust features within the encoder. This is particularly useful for visual place recognition under significant appearance variations, for example, day versus night. The state-of-the-art VPR methods use deep-learned representations either based on end-to-end training [8][32][3] or indirectly derived from the internal layer representations [14] [35][2]. For the performance benchmark presented in this section, we directly compare the convolutional features based image representations, extracted from different networks. In this way, the repeatability of activation patterns across day and night appearance conditions can be directly evaluated.

Figure 5 shows the performance comparison among different place representation methods. This includes flattened *conv5* representations from four differ-



Fig. 6: Qualitative Results: For night time query images (top row), Ground Truth (GT) match (second row) and matches obtained from different methods are displayed (subsequent rows) including successful matches using our proposed representation (third row)

ent networks trained on different tasks: *Ours-C5* uses the encoder output from our proposed network, trained to predict depth for night-time images; *NV-C5* uses VGG [34] based NetVLAD [3], trained for place recognition; *Obj-C5* uses ResNet50 [20], trained for object recognition; and *Sem-C5* uses the encoder output of RefineNet [26] which is based on ResNet101 [20] and trained for dense semantic segmentation. The latter has also been effectively used for state-of-the-art place recognition descriptor LoST [13]. The flattened *conv5* representations expect a similar viewpoint between the compared pairs of images; for sake of completion, we also include a viewpoint-invariant representation in our comparisons: NetVLAD as *NV* which uses 4096-dimensional descriptors. It can be observed that the feature representations based on our depth encoder perform the best. While there is a significant margin in performance for the flattened *conv5* comparisons, the proposed representation also outperforms the end-to-end learnt viewpoint-invariant NetVLAD representation.

Figure 6 shows qualitative results for visual place recognition under significant appearance variations. The first row shows four query images captured under night time conditions; their corresponding Ground Truth (GT) day-time image matches are shown in the second row. In subsequent columns, image matches obtained through different representation methods are displayed with the third row comprising successful matches based on our proposed representation. The incorrect matches using other methods in the first column seem to indicate a

bias in their selection based on the presence of a vehicle in the query image. In the second row, it can be observed that most of the retrieved matches comprise buildings viewed from far with an oblique viewpoint, however, only the proposed representation is able to obtain the correct match. We believe that learning to predict depth per pixel for night time imagery enables the latent representations to be more robust to perceptual aliasing caused by appearance variations. Moreover, our proposed depth-estimation network is trained in a completely unsupervised manner, where other vision-based tasks like object recognition and semantic segmentation would require labeled night-time data if they were to be used for extracting appearance-invariant image representations for place recognition.

4 Conclusions and Future Scope

This paper discusses the problem of estimating depth from night-time images, which suffers from poor visibility, non-uniform and unpredictable variation in illumination arising from multiple and possibly, moving light sources. The problem is tackled by applying a patchGAN-based domain adaptation technique that allows an encoder to adapt the *features* obtained from the night-time images to acquire the attributes of day-time features so that a decoder trained on day-time images could be directly used for estimating depth from these adapted features. The proposed novel approach is completely unsupervised as it does not necessitate the availability of either explicit ground truth signals (obtained from range sensors) or implicit supervision signals obtained from multi-view (spatial / temporal) images. Unlike many of the existing methods, the proposed method also does not require generating simulated data which is considerably difficult for night-time images. The efficacy of the proposed approach is demonstrated through extensive analyses on the challenging Oxford RobotCar dataset. Its usefulness is also demonstrated through its application to a visual place recognition problem where the feature representation obtained from our depth encoder is shown to outperform those obtained from the existing state-of-the-art methods. The proposed approach has some limitations which will form the scope for future investigations. As shown in Figure 4, our method can not deal with saturated regions (bright lights), very low illuminated regions and thin structures, such as traffic signal poles. This could be solved to some extent by incorporating semantic information in the learning process. Secondly, instead of learning two separate encoders - one for day-time images and the other for night-time images, it would be good to have one encoder model which could be trained to learn context-specific features which are unique to different styles rather than image-specific features. These context-specific features could then provide the necessary semantics to deal with the above failure cases.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine

- learning. In: OSDI. vol. 16, pp. 265–283 (2016)
2. Anoosheh, A., Sattler, T., Timofte, R., Pollefeys, M., Van Gool, L.: Night-to-day image translation for retrieval-based localization. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 5958–5964. IEEE (2019)
 3. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5297–5307 (2016)
 4. Atapour-Abarghouei, A., Breckon, T.P.: Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2800–2810 (2018)
 5. Babu, V.M., Das, K., Majumdar, A., Kumar, S.: Undemon: Unsupervised deep network for depth and ego-motion estimation. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1082–1088. IEEE (2018)
 6. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* **39**(12), 2481–2495 (2017)
 7. Cao, Y., Wu, Z., Shen, C.: 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 (2018)
 8. Chen, Z., Jacobson, A., Sünderhauf, N., Upcroft, B., Liu, L., Shen, C., Reid, I., Milford, M.: Deep learning features at scale for visual place recognition. In: Robotics and Automation (ICRA), 2017 IEEE International Conference on. pp. 3223–3230. IEEE (2017)
 9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
 10. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems. pp. 2366–2374 (2014)
 11. Garg, S., Babu, V., Dharmasiri, T., Hausler, S., Suenderhauf, N., Kumar, S., Drummond, T., Milford, M., et al.: Look no deeper: Recognizing places from opposing viewpoints under varying scene appearance using single-view depth estimation. arXiv preprint arXiv:1902.07381 (2019)
 12. Garg, S., Harwood, B., Anand, G., Milford, M.: Delta descriptors: Change-based place representation for robust visual localization. *IEEE Robotics and Automation Letters* (2020)
 13. Garg, S., Suenderhauf, N., Milford, M.: Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics. In: Proceedings of Robotics: Science and Systems XIV (2018)
 14. Garg, S., Sünderhauf, N., Milford, M.: Semantic-geometric visual place recognition: A new perspective for reconciling opposing views. *The International Journal of Robotics Research* (2019)
 15. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research* **32**(11), 1231–1237 (2013)
 16. Geiger, A., Ziegler, J., Stiller, C.: Stereoscan: Dense 3d reconstruction in real-time. In: IEEE Intelligent Vehicles Symposium. Baden-Baden, Germany (June 2011)

17. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6602–6611. IEEE (2017)
18. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.: Digging into self-supervised monocular depth estimation. arXiv preprint arXiv:1806.01260 (2018)
19. Handa, A., Whelan, T., McDonald, J., Davison, A.J.: A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In: Robotics and automation (ICRA), 2014 IEEE international conference on. pp. 1524–1531. IEEE (2014)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
21. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 172–189 (2018)
22. Im, S., Jeon, H.G., So Kweon, I.: Robust depth estimation from auto bracketed images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2946–2954 (2018)
23. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5967–5976. IEEE (2017)
24. Kim, N., Choi, Y., Hwang, S., Kweon, I.S.: Multispectral transfer network: Unsupervised depth estimation for all-day vision. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
26. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: IEEE conference on computer vision and pattern recognition (CVPR). vol. 1, p. 3 (2017)
27. Luo, C., Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R., Yuille, A.: Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. arXiv preprint arXiv:1810.06125 (2018)
28. Luo, Y., Ren, J., Lin, M., Pang, J., Sun, W., Li, H., Lin, L.: Single view stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 155–163 (2018)
29. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research* **36**(1), 3–15 (2017)
30. Marchand, E., Uchiyama, H., Spindler, F.: Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics* **22**(12), 2633–2651 (2016)
31. Nath Kundu, J., Krishna Uppala, P., Pahuja, A., Venkatesh Babu, R.: Adadepth: Unsupervised content congruent adaptation for depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2656–2665 (2018)
32. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018)
33. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3234–3243 (2016)

34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
35. Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., Milford, M.: On the performance of convnet features for place recognition. In: Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on. pp. 4297–4304. IEEE (2015)
36. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7167–7176 (2017)
37. Vankadari, M., Kumar, S., Majumder, A., Das, K.: Unsupervised learning of monocular depth and ego-motion using conditional patchgans. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. pp. 5677–5684. International Joint Conferences on Artificial Intelligence Organization (7 2019). <https://doi.org/10.24963/ijcai.2019/787>, <https://doi.org/10.24963/ijcai.2019/787>
38. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)
39. Yin, Z., Shi, J.: GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2 (2018)
40. Zhao, S., Fu, H., Gong, M., Tao, D.: Geometry-aware symmetric domain adaptation for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9788–9798 (2019)
41. Zhao, Z.Q., Zheng, P., Xu, S.t., Wu, X.: Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems **30**(11), 3212–3232 (2019)
42. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017)
43. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 989–997 (2019)
44. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017)