Domain Adaptation via Image Style Transfer

Amir Atapour-Abarghouei and Toby P. Breckon

Abstract While recent growth in modern machine learning techniques has led to remarkable strides in computer vision applications, one of the most significant challenges facing learning-based vision systems is the scarcity of large, high-fidelity datasets required for training large-scale models. This has necessitated the creation of transfer learning and domain adaptation as a highly-active area of research, wherein the objective is to adapt a model trained on one set of data from a specific domain to perform well on previously-unseen data from a different domain. In this chapter, we use monocular depth estimation as a means of demonstrating a new perspective on domain adaptation. Most monocular depth estimation approaches either rely on large quantities of ground truth depth data, which is extremely expensive and difficult to obtain, or alternatively predict disparity as an intermediary step using a secondary supervisory signal leading to blurring and other artefacts. Training a depth estimation model using pixel-perfect synthetic depth images can resolve most of these issues but introduces the problem of domain shift from synthetic to real-world data. Here, we take advantage of recent advances in image style transfer and its connection with domain adaptation to predict depth from a single colour image based on training over a large corpus of synthetic data obtained from a virtual environment. Experimental results point to the impressive capabilities of style transfer used as a means of adapting the model to unseen data from a different domain.

Amir Atapour-Abarghouei
Department of Computer Science, Durham University, UK. e-mail: amir.atapour-abarghouei@durham.ac.uk

Toby P. Breckon
Departments of Engineering and Computer Science, Durham University, UK. e-mail: toby.breckon@durham.ac.uk

1 Introduction

Recent advances in modern machine learning techniques have resulted in a significant growth in various computer vision applications readily deployed in real-world scenarios. However, the bias occasionally present within the datasets used to train these machine learning models can lead to notable issues. Such learning-based models often approximate functions capable of performing classification and prediction based tasks by capturing the underlying data distribution from which their training data is sampled. However, even small variations between the distributions of the training and the test data can negatively affect the performance of the approach. Such concerns have led to the creation of the field of transfer learning and domain adaptation [40], with a large community of researchers actively addressing the problem of data domain shift.

In this chapter, our primary focus is on the use of image style transfer as a domain adaptation technique. We utilise one of the fastest-growing and most challenging areas of research, namely monocular depth estimation, within computer vision as a means to demonstrate the efficacy of the domain adaptation via image style transfer.

As 3D imagery has become more prevalent within computer vision, accurate and efficient depth estimation is now of paramount importance within many vision-based systems. While plausible depth estimation has been possible for many years using conventional strategies such as stereo correspondence [47], structure from motion [14, 10], depth from shading and light diffusion [52, 58, 1] and alike, such techniques often suffer from a myriad of issues such as intensive computational and calibration requirements, depth inhomogeneity and missing depth information, often resulting in the need for a post-processing stage to create more accurate and complete scene depth [4, 8, 3, 5, 35, 43, 6, 2]. Learning-based monocular depth estimation can offer a way to circumvent such issues as a novel alternative to many of these outdated approaches [6, 31, 34, 16, 21, 64, 19, 59].

Supervised learning-based monocular depth estimation approaches take advantage of off-line training on ground truth depth data to make depth prediction possible [31, 34, 16, 17, 67]. However, since ground truth depth is often scant and expensive to acquire in the real world, the practical use of many such approaches is heavily constrained.

There are, however, other monocular depth estimation approaches that do not require direct ground truth depth, but instead utilise a secondary supervisory signal during training which indirectly results in producing the desired depth [21, 64, 19, 59, 12]. Training data for these approaches is abundant and easily obtainable but they suffer from undesirable artefacts, such as blurring and incoherent content, due to the nature of their secondary supervision. However, an often overlooked fact is that the same technology that facilitates training large-scale deep neural networks can also assist in acquiring synthetic data for these neural networks [39, 49]. Nearly photo-realistic graphically rendered environments primarily used for gaming can be used to capture homogeneous synthetic depth images which are then utilised in training a depth estimating model.

While the use of such synthetic data is not novel and can resolve the issue of data scarcity [2, 32, 18, 49], the variations between synthetic and real-world images can lead to notable issues during deployment since any model trained on synthetic data cannot be expected to perform equally well when tested on naturally-sensed real-world images. Here, we intend to demonstrate the possibility of using style transfer as a domain adaptation technique. In this vein, in Sections 2 and 3, we briefly outline the relevant areas of domain adaptation, image style transfer and their underlying connections and subsequently move on to practically demonstrating the applicability of style transfer in domain adaptation in the context of monocular depth estimation trained on synthetic imagery.

2 Domain Adaptation via Manximum Mean Discrepancy

The main objective of domain adaptation is to transfer a model that has encapsulated the underlying distribution of a set of labelled data from the source domain so that it can perform well on previously-unseen unlabelled data from the target domain [40].

Within the current literature, this is often accomplished by minimising the distance between the source and target distributions. One of the most common metrics used to measure the distance between the two distributions is Maximum Mean Discrepancy (MMD), which is the difference between probability measures based on embedding probabilities in a reproducing kernel Hilbert space [23, 53].

Assume there exist two sample sets $X = \{x_1, ..., x_n\}$ and $Y = \{y_1, ..., y_m\}$ with x_i and y_i independently and identically distributed from p and q respectively. As described in [23], in the two-sample testing problem, MMD can be used as a test statistic by drawing samples from distributions p and q and fitting a smooth function, which is large on points drawn from p and small on point drawn from q [23]. MMD is the difference between the mean function values on the two samples. This means when the samples are from different distributions, the MMD will be large and when the distributions are equal ($p = q$), the population MMD vanishes. More formally, the squared MMD is as follows [33]:

$$\begin{aligned}MMD^2[X,Y] = ||\mathbb{E}_x[\phi(x)] - \mathbb{E}_y[\phi(y)]||^2 = ||\frac{1}{n}\sum_{i=1}^{n}\phi(x_i) - \frac{1}{m}\sum_{j=1}^{m}\phi(y_j)||^2 = \\ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\phi(x_i)^T\phi(x'_i) + \frac{1}{m^2}\sum_{j=1}^{m}\sum_{j'=1}^{m}\phi(y_j)^T\phi(y'_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\phi(x_i)^T\phi(y_j)\end{aligned} \tag{1}$$

where $\phi(.)$ denotes the feature mapping function from X to $\mathbb{R}$. To reformulate Eqn. 1 in the form of kernel, the function $k(x,y) = \langle\phi(x),\phi(y)\rangle_{\mathscr{H}}$ in a reproducing kernel Hilbert space $\mathscr{H}$ can be applied to the equation [33]:

$$
\begin{aligned}
MMD^2[X,Y] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n} k(x_i, x'_i) + \frac{1}{m^2}\sum_{j=1}^{m}\sum_{j'=1}^{m} k(y_j, y'_j) \\
- \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i, y_j)
\end{aligned}
\tag{2}
$$

where $k(.,.)$ is the kernel function defining a mapping to a higher dimensional feature space. In Section 3, we provide a brief overview of the advances made in modern neural-based style transfer and its connections with domain adaptation via Maximum Mean Discrepancy.

3 Image Style Transfer

Image style transfer via convolutional neural networks first emerged as an effective stylization technique via the work in [20] and various improved and novel approaches capable of transferring the style of one image onto another [29, 54, 11] have been proposed ever since.

Conventionally, the style of an image is represented as a set of Gram matrices [48] that describe the correlations between low-level convolutional features extracted from the image, while the raw values of high-level semantic features often constitute the content of an image. These style and content representations are often extracted from a pre-trained loss network and are subsequently utilised to quantify style and content losses with respect to the target style and content images. More formally, the content loss for a specific layer l of the loss network can be defined as:

$$
\mathcal{L}_{content} = \sum_{i=1}^{N_l}\sum_{j=1}^{M_l} ||f^l_{ij}(x) - f^l_{ij}(c)||^2 \tag{3}
$$

where c and x respectively denote the content and the output stylized images, f represents the loss network [51], $f^l(x)$ is the set of feature maps extracted from layer l after x is passed through f, N_l is the number of feature maps in layer l and M_l denotes the size (height × width) of the feature map. Similarly the style loss for a specific layer l of the loss network can be expressed as:

$$
\mathcal{L}_{style} = \frac{1}{4N_l^2 M_l^2}\sum_{i=1}^{N_l}\sum_{j=1}^{N_l} (\mathcal{G}[f^l_{ij}(x)] - \mathcal{G}[f^l_{ij}(s)])^2 \tag{4}
$$

where s and x respectively represent the style and the output stylized images, and $\mathcal{G}[f^l_{ij}(x)]$ denotes the Gram matrix of the feature maps extracted from layer l after x is passed through f. The overall loss function can subsequently be defined as:

$$
\mathcal{L} = \lambda_c \mathcal{L}_{content}(x,c) + \lambda_s \mathcal{L}_{style}(x,s) \tag{5}
$$

where λ_c and λ_s are coefficients determining the relative weights of the style and content loss components in the overall objective. In the original work in [20], this objective was minimised directly by gradient descent within the image space, and although the results of [20] are impressive, its process is very computationally intensive, leading to the emergence of alternative approaches that use neural networks to approximate the global minimum of the objective in a single forward pass. Such approaches [29, 54, 11] utilise neural networks trained to restyle an input image while preserving its content.

Style transfer can be considered as a distribution alignment process from the content image to the style image [33, 28]. In other words, transferring the style of one image (from the source domain) to another image (from the target domain) is essentially the same as minimising the distance between the source and target distributions. To demonstrate this connection between style transfer and domain adaptation (through MMD), the style loss in Eqn. 4 can be reformulated by expanding the Gram matrices and applying the second order degree polynomial $k(x, y) = (x^T y)$ as follows [33]:

$$\begin{aligned}\mathcal{L}_{style} = \frac{1}{4N_l^2 M_l^2} \sum_{k_1=1}^{M_l} \sum_{k_2=1}^{M_l} (k(f^l_{k_1}(x), f^l_{k_2}(x)) + k(f^l_{k_1}(s), f^l_{k_2}(s)) \\ - 2k(f^l_{k_1}(x), f^l_{k_2}(s))) = \frac{1}{4N_l^2} MMD^2[\mathcal{F}^l(x), \mathcal{F}^l(s)]\end{aligned} \tag{6}$$

where s and x respectively denote the style and the output stylized images, $f^l_k(x)$ denotes the kth column of $f^l(x)$ and $\mathcal{F}^l(x)$ is the set of features extracted from x in which each sample is a column of $f^l(x)$. Consequently, by minimising the style loss (reducing the distance between the style of the target image and desired stylized image), we are in effect reducing the distance between their distributions.

Here, we take advantage of this direct connection between domain adaptation and style transfer to perform monocular depth estimation by adapting our data distribution (i.e. real-world images) to our depth estimation model trained on data from a different distribution (i.e. synthetic images). However, while style transfer by matching Gram matrices is theoretically equivalent to minimising the MMD with the second order polynomial kernel and leads to domain adaptation, we forego the use of conventional style transfer and opt for an adversarially trained style transfer approach [66]. Other than the fact that the adversarially trained style transfer approach originally proposed in [66] is capable of superior performance and more pronounced changes in the style of the output image, the main reason for the choice of this approach is that [66] can transfer the style between two *sets* of unaligned images from different domains, while more conventional neural style transfer techniques such as [29] can only accept *one* specific image to be used as the style image. Within domain adaptation, this is not very desirable, especially since not one but tens of thousands of images representing the same style exist within the target domain. Experiments empirically justifying this choice are included in Section 5.2. In the next section, the approach to monocular depth estimation via style transfer [7] is outlined in greater depth.

4 Monocular Depth Estimation via Style Transfer

Our style transfer based monocular depth estimation approach consists of two stages, relying on two completely separate models trained at the same time to carry out the operations of each stage. The first stage includes directly training a depth estimation model using synthetic data captured from a graphically rendered environment primarily designed for gaming applications [39] (Section 4.1). However, as the eventual objective of the overall model involves estimating depth from real-world images, we attempt to reduce the domain discrepancy between the synthetic data distribution and the real-world data distribution using a model trained to transfer the style of synthetic images to real-world images in the second stage of the overall approach (Section 4.2).

4.1 Stage 1: Depth Estimation Model

Here, we consider monocular depth estimation as an image-to-image mapping problem, with the RGB image used as the input to our mapping function, and scene depth produced as its output. Using modern convolutional neural networks, image-to-image translation and prediction problems have become significantly more tractable and can yield remarkably high-quality results. An overly simplistic solution to a translation problem such as depth estimation would be employing a network that attempts to minimise a reconstruction loss (Euclidean distance) between the pixel values of the output and the ground truth. However, since monocular depth estimation is an inherently multi-modal problem (a problem that has several global solutions instead of a unique global optimum since several plausible depth values can correspond with a single RGB view), any model trained to predict depth based on a sole reconstruction loss tends to generate values that are the average of all the possible modes in the predictions. This averaging can lead to blurring effects in the outputs.

As a result, many such prediction-based approaches [2, 66, 7, 42, 61, 60, 27, 63] and other generative models [15, 55] make use of adversarial training [22] to alleviate the blurry output problem since the use of an adversarial loss generally forces the model to select a single mode from the distribution instead of averaging all possible modes and generate more realistic results without blurring.

A Generative Adversarial Network (GAN) [22] is capable of producing semantically sound samples by creating a competition between a generator, which attempts to capture the underlying data distribution, and a discriminator, which judges the output of the generator and penalises unrealistic images and artefacts. Both networks are trained simultaneously to achieve an equilibrium [22]. Whilst most generative models generate images from a latent noise vector as the input to the generator, the model presented here is solely conditioned on an input image (RGB).

More formally, the generative model learns a mapping from the input image, x (RGB view), to the output image, y (scene depth), $G : x \to y$. The generator, G, attempts to produce fake samples, $G(x) = \tilde{y}$, which cannot be distinguished from real ground truth samples, y, by the discriminator, D, which is adversarially trained to detect the fake samples produced by the generator.

Many other approaches following a similar framework incorporate a random noise vector z or drop-outs into the generator training to prevent deterministic mapping and induce stochasticity [27, 42, 37, 57]. However, since deterministic mapping is not of concern in a problem such as depth estimation, no random noise or drop-out is required. Empirical experiments demonstrate no significant difference in the output distribution could be achieved even if stochasticity is encouraged within the model using these strategies.

4.1.1 Loss Function

The objective of the monocular depth estimation model is achieved via minimising a loss function consisting of two components. The first is a simple reconstruction loss, which forces the generator to capture the structural and contextual content of the scene and output depth images which are as close as possible to the ground truth depth information. To accomplish this, we use the L_1 loss:

$$\mathcal{L}_{rec} = ||G(x) - y||_1 \tag{7}$$

While the use of a reconstruction loss can help the network to internally model the structure and content of the scene, it can also lead to the generator optimising towards averaging all possible output depth values rather than selecting one, which can lead to blurring effects within the output depth image. Consequently, the second component of the overall loss function, an adversarial loss, is introduced to incentivise the generator to create shaper and higher quality depth images:

$$\mathcal{L}_{adv} = \min_G \max_D \mathop{\mathbb{E}}_{x,y \sim \mathbb{P}_d(x,y)} [\log D(x,y)] + \mathop{\mathbb{E}}_{x \sim \mathbb{P}_d(x)} [\log(1 - D(x, G(x)))] \tag{8}$$

where $\mathbb{P}_d$ denotes the data distribution defined by $\tilde{y} = G(x)$ and x is the input to the generator and y the ground truth. Subsequently, the overall loss function is as follows:

$$\mathcal{L} = \lambda \mathcal{L}_{rec} + (1 - \lambda)\mathcal{L}_{adv} \tag{9}$$

with λ being the weighting coefficient selected empirically.

4.1.2 Implementation Details

In order to obtain the synthetic data required to train the depth estimation model, corresponding colour and disparity images are captured using a camera view placed in front of a virtual car as it automatically drives around a graphically-rendered virtual environment, with images captured every 60 frames with randomly varying height, field of view, weather and lighting conditions at different times of day. From the overall dataset of 80,000 corresponding pairs of colour and disparity images captured in this manner, 70,000 are used for training and 10,000 are set aside for testing. The depth estimation model trained using the synthetic dataset generates disparity images which can be converted to depth using the known camera parameters and scaled to the depth range of the KITTI image frame [38].

An important aspect of any depth estimation problem is that the overall structure and the high frequency information present within the RGB view of the scene (input) and the depth image (output) are aligned as they ultimately represent the exact same scene. As a result, much information (e.g. structure, geometry, object boundaries and alike) is shared between the input and output. Consequently, we utilise the capabilities of skip connections within the architecture of the generator [42, 57, 45, 25, 9] to accurately preserve high-frequency scene content. The generator, therefore, can take advantage of the opportunity to directly pass geometric information between corresponding layers in the encoder and the decoder without having to go through every single layer in between and possibly losing precious details in the down-sampling and up-sampling processes.

The generator consists of an architecture similar to that of [45] with the exception that skip connections exist between every pair of corresponding layers in the encoder and decoder. As for the discriminator, the basic architecture used in [44] is deployed. Both the generator and discriminator utilise convolution-BatchNorm-ReLu modules [26] with the discriminator using leaky ReLUs ($slope = 0.2$).

All technical implementation is performed using *PyTorch* [41], with Adam [30] providing the optimisation ($\beta_1 = 0.5$, $\beta_2 = 0.999$, $\alpha = 0.0002$). The weighting coefficient in the overall loss function in Eqn. 9 was empirically chosen to be $\lambda = 0.99$.

4.2 Stage 2: Style Transfer as Domain Adaptation

The monocular depth estimation model presented in Section 4.1 can perform very well on unseen images from the test set of synthetic data captured from the virtual environment. However, since the model is only trained on synthetic images and the synthetic and real-world images are from different domains, directly estimating depth from RGB images captured in the real-world remains challenging, which is why domain adaptation via style transfer is an important component of the overall approach.

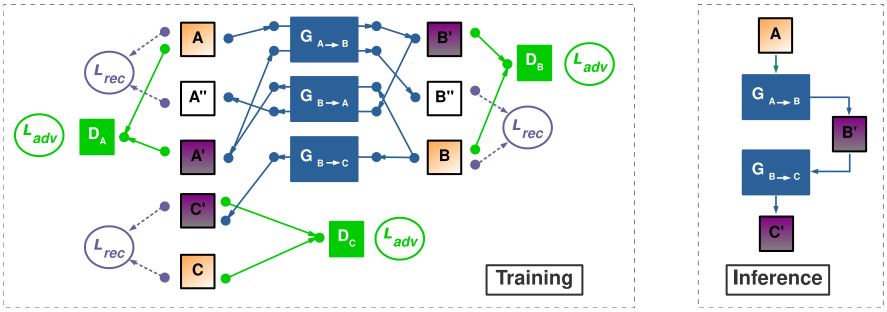

Fig. 1: Outline of the approach to monocular depth estimation via domain adaptation using [66]. Domain A (real-world RGB) is transformed into B (synthetic RGB) and then to C (pixel-perfect depth). A, B and C represent ground truth images, A', B' and C' are the generated images and A" and B" denote images cyclically regenerated via [66].

The objective of the style transfer component of the approach, therefore, is to learn a mapping function $\mathcal{D} : X \rightarrow Y$ from the source domain X (real-world images) to the target domain Y (synthetic images) in a way that the distributions $\mathcal{D}(X)$ and Y are identical. When images from X are mapped into Y, their corresponding depth information can be inferred using the monocular depth estimation model presented in Section 4.1 that is specifically trained on images from Y.

Within the existing literature, there have been various successful attempts at transforming images from one domain to another [66, 36, 46, 50]. Here, the proposed approach relies on the idea of image style transfer using generative adversarial networks, as proposed in [66], to reduce the discrepancy between the source domain (real-world data) and the target domain (synthetic data on which the depth estimation model in Section 4.2 trained). This approach uses adversarial training [22] and cycle-consistency [21, 56, 65, 62] to translate between two sets of unaligned images from different domains.

More formally, the objective is to map images between the two domains X and Y with the respective distributions of $x \sim \mathbb{P}_d(x)$ and $y \sim \mathbb{P}_d(y)$. The mapping functions are approximated using two separate generators, G_{XtoY} and G_{YtoX} and two discriminators D_X (discriminating between $x \in X$ and $G_{YtoX}(y)$) and D_Y (discriminating between $y \in Y$ and $G_{XtoY}(x)$). The loss contains two components: an adversarial loss [22] and a cycle consistency loss [66]. The general pipeline of the approach (along with the depth estimation model 4.1) is seen in Figure 1, with three generators $G_{A_{to}B}$, $G_{B_{to}A}$ and $G_{B_{to}C}$, and three discriminators D_A, D_B and D_C.

4.2.1 Loss Function

Since there are two generators to constrain the content of the images, there are two mapping functions. The use of an adversarial loss guarantees the style of one domain is transferred to the other. The loss for G_{XtoY} with D_Y is represented as follows:

Method	Training Data	Error Metrics (lower, better)				Accuracy Metrics (higher, better)		
		Abs. Rel.	Sq. Rel.	RMSE	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Eigen et al. Coarse	K	0.214	1.605	6.563	0.292	0.673	0.884	0.957
Eigen et al. Fine	K	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu et al.	K	0.202	1.614	6.523	0.275	0.678	0.895	0.965
Zhou et al.	K	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Zhou et al.	K+CS	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Godard et al.	K	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Godard et al.	K+CS	0.124	1.076	5.311	0.219	0.847	0.942	0.973
DST Approach	K+S*	**0.110**	**0.929**	**4.726**	**0.194**	**0.923**	**0.967**	**0.984**

Table 1: Comparing the results of depth estimation via style transfer (DST) against other approaches over the KITTI dataset using the data split in [17]. For the training data, K represents KITTI [38], CS Cityscapes [13] and S* the synthetic data captured from a virtual environment.

$$\mathcal{L}_{adv-XtoY} = \min_{G_{XtoY}} \max_{D_Y} \mathop{\mathbb{E}}_{y \sim \mathbb{P}_d(y)} [\log D_Y(y)] + \mathop{\mathbb{E}}_{x \sim \mathbb{P}_d(x)} [\log(1 - D_Y(G_{XtoY}(x)))] \quad (10)$$

where $\mathbb{P}_d$ is the data distribution, X the source domain with samples x and Y the target domain with samples y. Similarly, for G_{YtoX} and D_X, the adversarial loss is as follows:

$$\mathcal{L}_{adv-YtoX} = \min_{G_{YtoX}} \max_{D_X} \mathop{\mathbb{E}}_{x \sim \mathbb{P}_d(x)} [\log D_X(x)] + \mathop{\mathbb{E}}_{y \sim \mathbb{P}_d(y)} [\log(1 - D_X(G_{YtoX}(y)))] \quad (11)$$

To constrain the adversarial loss of the generators to force the model to produce contextually coherent images rather than random semantically meaningless content from the target domain, a cycle-consistency loss is added that encourages the model to become capable of bringing an image x that is translated into the target domain Y using G_{XtoY} back into the source domain X using G_{YtoX}. In essence, after a full cycle: $G_{YtoX}(G_{XtoY}(x)) = x$ and vice versa. Consequently, the cycle-consistency loss is as follows:

$$\mathcal{L}_{cyc} = ||G_{YtoX}(G_{XtoY}(x)) - x||_1 + ||G_{XtoY}(G_{YtoX}(y)) - y||_1 \quad (12)$$

Subsequently, the joint loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{adv-XtoY} + \mathcal{L}_{adv-YtoX} + \lambda \mathcal{L}_{cyc} \quad (13)$$

with λ being the weighting coefficient selected empirically.

4.2.2 Implementation Details

The architecture of the generators is similar to that of the network proposed in [29] with two convolutional layers followed by nine residual blocks [24] and two up-convolutions that bring the image back to its original input size. As for the

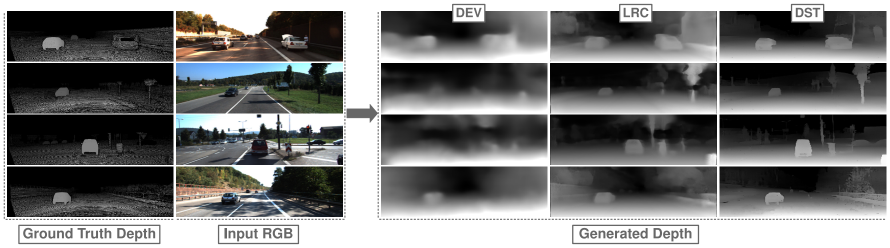

Fig. 2: Qualitative comparison of the results of contemporary state-of-the-art approaches of depth and ego-motion from video (DEV) [64], estimation based on left/right consistency (LRC) [21] and depth via style transfer (DST) over the KITTI split.

discriminators, the same architecture is used as was in Section 4.1. Moreover, the discriminators are updated based on the last 50 generator outputs and not just the last generated image [66, 50].

All technical implementation is performed using *PyTorch* [41], with Adam [30] providing the optimisation ($\beta_1 = 0.5$, $\beta_2 = 0.999$, $\alpha = 0.0001$). The weighting coefficient in the overall loss function in Eqn. 13 was empirically chosen to be $\lambda = 10$.

5 Experimental Results

In order to demonstrate the efficacy of style transfer used as domain adaptation technique for the task of monocular depth estimation, in this section, the depth estimation approach is evaluated using ablation studies and both qualitative and quantitative comparisons with state-of-the-art monocular depth estimation methods. The KITTI dataset [38] and locally-captured data are used for evaluations.

5.1 Comparisons against Contemporary Approaches

To evaluate the performance of the monocular depth estimation approach in Section 4 and demonstrate the capability of style transfer used as domain adaptation, 697 images from the data split suggested in [17] are used as the test set. As demonstrated in Table 1, the monocular depth estimation model trained on synthetic data and adapted using style transfer (DST) performs better than contemporary monocular depth estimation approaches directly trained on real-world images [34, 21, 64, 17] with lower error and higher accuracy. Some of the comparators [21, 64] use a combination of different datasets for training and fine-tuning to boost performance, while the approach presented here only relies on synthetic data for training.

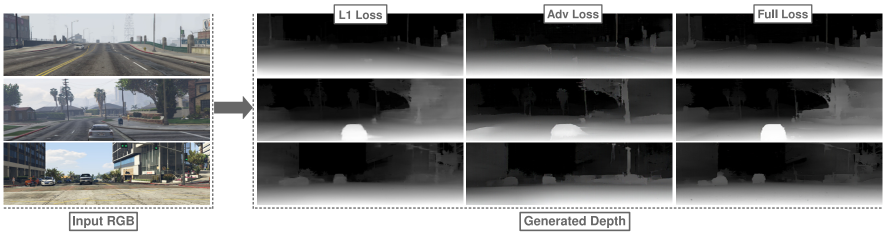

Fig. 3: Demonstrating the importance of the different components of the loss function in the depth estimation model (Section 4.1).

The data split of 200 images in KITTI [38] is also used to provide better qualitative evaluation, since the ground truth disparity images within this split are of considerably higher quality and provide CAD models as replacements for moving cars. As seen in Figure 2, compared to other approaches [21, 64] trained on similar data domains, monocular depth estimation via style transfer leads to sharper and more crisp outputs in which object boundaries and thin structures are better preserved.

5.2 Ablation Studies

Ablation studies are integral in demonstrating the necessity of the components of the approach. The monocular depth estimation model presented in Section 4.1 utilises a combination of reconstruction and adversarial losses (Eqn. 9). In order to test the importance of each loss component, the model is separately trained using the reconstruction loss only and the adversarial loss only. Figure 3 demonstrates the effects of removing parts of the training objective. The model based only on the reconstruction loss produces contextually sound but blurry results, while the adversarial loss generates sharp outputs that contain artefacts. When the approach is trained using the full overall loss function, it creates more accurate results without unwanted effects. Further numerical and qualitative evidence of the efficacy of a combination of a reconstruction and adversarial loss can be found in [27].

Another important aspect of the ablation studies involves demonstrating the necessity of domain adaptation (Section 4.2) within the overall pipeline. As indicated in Table 2, due to the differences in the domains of the synthetic and natural data, the depth estimation model directly applied to real-world data does not produce numerically desirable results, which points to the importance of domain adaptation to the approach. Similarly, Figure 4 qualitatively demonstrates that when no style transfer is used in the approach, the generated depth outputs contain significant inaccuracies and undesirable artefacts.

While the connection between domain adaptation by minimising the Maximum Mean Discrepancy with the second order polynomial kernel and neural style transfer by matching Gram matrices is briefly outlined in Sections 2 and 3, the approach

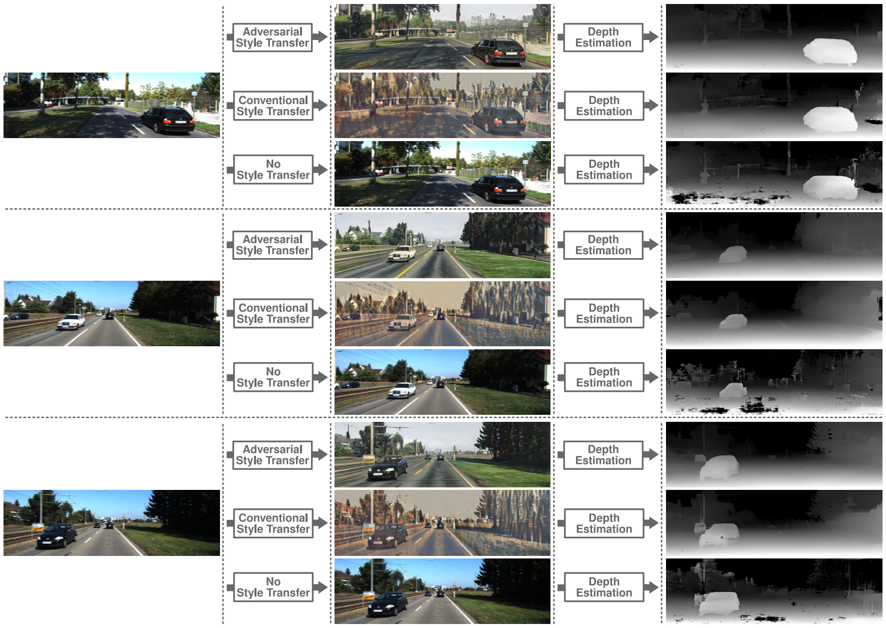

Fig. 4: Exemplar results demonstrating the importance of style transfer. Examples include results of depth estimation with style transfer via cycle-consistent adversarial training [66], conventional style transfer approach of [29] and without any style transfer.

presented here does not use conventional neural style transfer and instead requires an adversarial discriminator [66] to carry out style transfer for domain adaptation.

To demonstrate that a discriminator can reasonably perform domain adaptation via style transfer, experiments are carried out with the style transfer approach proposed in [29], which improves on the pioneering style transfer work of [20] by training a generator that can transfer a specific style (that of our synthetic domain in this work) onto a set of images of a specific domain (real-world images) by minimising content and style losses (Eqns. 3 and 4). An overview of the entire pipeline using [29] (along with the monocular depth estimation model in Section 4.1) is seen in Figure 5.

Whilst [66] is capable of transferring the style between two large *sets* of unaligned images from different domains, the neural style transfer approach in [29] requires *one* specific image to be used as the target style image. In this work, the target domain consists of tens of thousands of images representing the same style. Consequently, a number of synthetic images that contain a variety of objects, textures and colours that represent their domain are collected and a single image that holds the desired style is created by pooling features from the images.

To evaluate the performance of the approach regarding the effects of domain adaptation via style transfer, the data split of 200 images in the KITTI dataset [38] is used. Experiments are carried out with both style transfer techniques in [66] and [29], in addition to using real-world images as direct inputs to the depth estimation

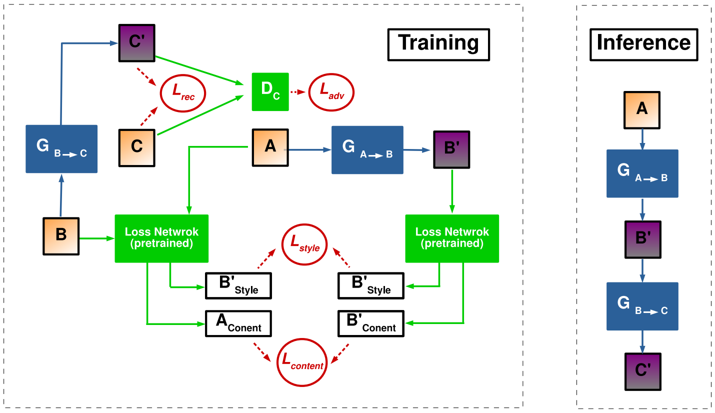

Fig. 5: Outline of the approach to monocular depth estimation via domain adaptation using [29]. Images from domain A (real-world) are transformed into B (synthetic) and then to C (pixel-perfect depth). A, B and C denote ground truth images and A', B' and C' represent generated images.

model without any domain adaptation. As seen in the results presented in Table 2, using direct real-world inputs without any domain adaptation via style transfer results in significant anomalies in the output while translating images into synthetic space using [66] before depth estimation leads to notably improved results. The qualitative results provided in Figure 4 also point to the same conclusion.

5.3 Generalisation

The use of domain adaptation via style transfer can make the model more robust and less susceptible to domain shift in the presence of unseen data. Considering that the images used in the training procedure of the monocular depth estimation model (Section 4.1) are captured from a synthetic environment [39] and the data used to train the style transfer component of the approach (Section 4.2) are from the KITTI dataset [38], we evaluate the generalisation capabilities of the approach using additional data captured locally in an urban environment. As clearly seen in Figure

Method	Training Data	Error Metrics (lower, better)				Accuracy Metrics (higher, better)		
		Abs. Rel.	Sq. Rel.	RMSE	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
w/o domain adaptation	K+S*	0.498	6.533	9.382	0.609	0.712	0.823	0.883
w/ the approach in [29]	K+S*	0.154	1.338	6.470	0.296	0.874	0.962	0.981
w/ the approach in [66]	K+S*	**0.101**	**1.048**	**5.308**	**0.184**	**0.903**	**0.988**	**0.992**

Table 2: Ablation study over the KITTI dataset using the KITTI split. The approach is trained using, KITTI (K) and synthetic data (S*). The approach provides the best results when it includes domain adaptation via style transfer using the technique in [66].

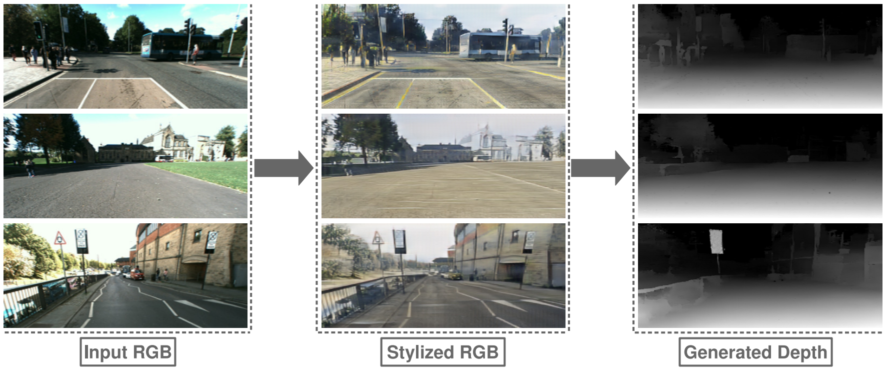

Fig. 6: Qualitative results of the proposed approach on urban driving scenes captured locally without further training.

6, the approach is easily capable of generating sharp, coherent and visually plausible depth without any training on the unseen images from the new data domain.

6 Limitations

The monocular depth estimation approach discussed in this chapter is capable of generating high quality and accurate depth with minimal anomalies by taking advantage of domain adaptation via image style transfer. However, the very component of the approach that enables it to generate highly accurate pixel-perfect depth, namely style transfer, can also bring forth certain shortcomings within the overall pipeline. The most significant issue is that of adapting to sudden lighting changes and saturation during style transfer. The two domains of images used here (synthetic and real-world images) significantly vary in intensity differences between lit areas and shadows, as is very common in images captured using different cameras in different environments. As a result, image regions containing shadows can be wrongly construed as elevated surfaces or foreground objects post style transfer, leading to inaccurate depth estimation of said regions. Examples in Figure 7 demonstrate how such issues can arise.

Moreover, despite the fact that holes (missing regions) are generally considered undesirable in depth images [4, 8, 3, 35, 43], certain areas within the scene depth should remain without depth values (e.g. very distant objects and sky). However, a supervised monocular depth estimation approach such as the one discussed in this chapter is incapable of distinguishing the sky from other extremely saturated objects within the scene even with style transfer, which can lead to creation of small holes where they do not belong.

Additionally, while the approach in [66] has been demonstrated to be very powerful in mapping between two sets of unaligned images with similar content, it can be

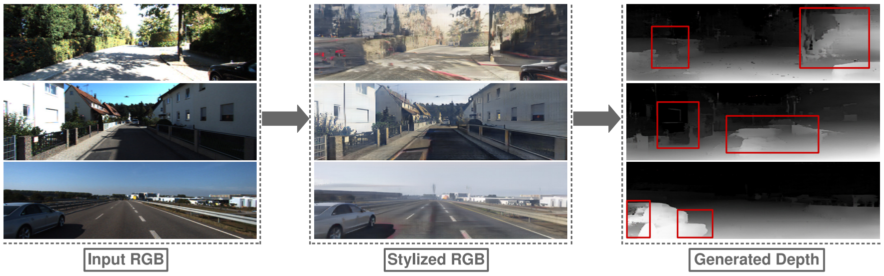

Fig. 7: Examples of failures, mainly due to light saturation and shadows.

very susceptible to wrongly synthesising meaningless content. This can especially happen if certain content (scene objects, geometry, structure and alike) is commonly found in images from one domain but not the other. Under these circumstances, the adversarial discriminator in [66] tends to encourage the generator to synthesise content in the latter domain to compensate for the discrepancies induced by the differences in overall scene content. Since in domain adaptation via style transfer, the objective is to transform the style of the images and not their content, this issue can lead to significant issues in terms of unwanted artefacts and anomalies within the output.

7 Conclusion

In this chapter, we have primarily focused on demonstrating the viability of image style transfer as a domain adaptation technique in computer vision applications. The aim of a domain adaptation approach is to adapt a model trained on one set of data from a specific domain to perform well on previously-unseen data from a different domain. In this vein, we have selected the problem of monocular depth estimation for our experiments since large quantities of ground truth depth data required for training a directly supervised monocular depth estimation approach is extremely expensive and difficult to obtain, leading to a greater need for domain adaptation.

Taking advantage of pixel-perfect synthetic depth data captured from a graphically rendered urban environment designed for gaming applications, an effective depth estimation model can be trained in a directly supervised manner. However, such a model cannot be expected to perform well on previously-unseen real-world images as the data distributions to which images from these two domains (synthetic images and real-world images) belong are vastly different. Since modern advances in neural style transfer can theoretically be linked to minimising the Maximum Mean Discrepancy between two distributions with the second order polynomial kernel, we make use of a adversarially trained cycle-consistent approach capable of transferring styles between two unaligned sets of images to adapt our real-world data to fit into the distribution approximated by the generator in our depth estimation model. Despite certain isolated issues, experimental evaluations and comparisons against contemporary monocular depth estimation approaches demonstrate that style transfer is indeed a highly effective method of domain adaptation.

References

1. Abrams, A., Hawley, C., Pless, R.: Heliometric stereo: Shape from sun position. Proc. Euro. Conf. Computer Vision pp. 357–370 (2012)
2. Atapour-Abarghouei, A., Akcay, S., Payen de La Garanderie, G., Breckon, T.: Generative adversarial framework for depth filling via wasserstein metric, cosine transform and domain transfer. Pattern Recognition **91**, 232–244 (2019)
3. Atapour-Abarghouei, A., Breckon, T.: DepthComp: Real-time depth image completion based on prior semantic scene segmentation. In: Proc. British Machine Vision Conference (2017)
4. Atapour-Abarghouei, A., Breckon, T.: A comparative review of plausible hole filling strategies in the context of scene depth image completion. Computers and Graphics **72**, 39–58 (2018)
5. Atapour-Abarghouei, A., Breckon, T.: Extended patch prioritization for depth filling within constrained exemplar-based RGB-D image completion. In: Proc. Int. Conf. Image Analysis and Recognition, pp. 306–314 (2018)
6. Atapour-Abarghouei, A., Breckon, T.: Veritatem Dies Aperit-temporally consistent depth prediction enabled by a multi-task geometric and semantic scene understanding approach. arXiv preprint arXiv:1903.10764 (2019)
7. Atapour-Abarghouei, A., Breckon, T.P.: Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2800–2810 (2018)
8. Atapour-Abarghouei, A., Payen de La Garanderie, G., Breckon, T.: Back to butterworth - a Fourier basis for 3D surface relief hole filling within RGB-D imagery. In: Proc. Int. Conf. Pattern Recognition, pp. 2813–2818. IEEE (2016)
9. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Analysis and Machine Intelligence **39**(12), 2481–2495 (2017)
10. Cavestany, P., Rodriguez, A., Martinez-Barbera, H., Breckon, T.: Improved 3d sparse maps for high-performance structure from motion with low-cost omnidirectional robots. In: Proc. Int. Conf. Image Processing, pp. 4927–4931 (2015)
11. Chen, T.Q., Schmidt, M.: Fast patch-based style transfer of arbitrary style. In: Workshop in Constructive Machine Learning, pp. 1–5 (2016)
12. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: Advances in Neural Information Processing Systems, pp. 730–738 (2016)
13. 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: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
14. Ding, L., Sharma, G.: Fusing structure from motion and LiDAR for dense accurate depth map estimation. In: Proc. Int. Conf. Acoustics, Speech and Signal Processing, pp. 1283–1287. IEEE (2017)
15. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: Advances in Neural Information Processing Systems, pp. 658–666 (2016)
16. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proc. Int. Conf. Computer Vision, pp. 2650–2658 (2015)
17. 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)

18. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 4340–4349 (2016)
19. Garg, R., Carneiro, G., Reid, I.: Unsupervised CNN for single view depth estimation: Geometry to the rescue. In: Proc. Euro. Conf. Computer Vision, pp. 740–756. Springer (2016)
20. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2414–2423 (2016)
21. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 6602 – 6611 (2017)
22. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
23. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. Machine Learning Research **13**, 723–773 (2012)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. Conf. Computer Vision and Pattern Recognition, pp. 770–778 (2016)
25. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
26. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proc. Int. Conf. Machine Learning, pp. 448–456 (2015)
27. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 5967–5976 (2017)
28. Jing, Y., Yang, Y., Feng, Z., Ye, J., Song, M.: Neural style transfer: A review. arXiv preprint arXiv:1705.04058 (2017)
29. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proc. Euro. Conf. Computer Vision, pp. 694–711 (2016)
30. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: Proc. Int. Conf. Learning Representations (2014)
31. Ladicky, L., Shi, J., Pollefeys, M.: Pulling things out of perspective. In: Proc. Conf. Computer Vision and Pattern Recognition, pp. 89–96 (2014)
32. Le, T.A., Baydin, A.G., Zinkov, R., Wood, F.: Using synthetic data to train neural networks is model-based reasoning. In: Proc. Int. Joint Conf. Neural Networks, pp. 3514–3521. IEEE (2017)
33. Li, Y., Wang, N., Liu, J., Hou, X.: Demystifying neural style transfer. In: Proc. Int. Joint Conf. Artificial Intelligence, pp. 2230–2236. AAAI Press (2017)
34. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. IEEE Trans. Pattern Analysis and Machine Intelligence **38**(10), 2024–2039 (2016)
35. Liu, J., Gong, X., Liu, J.: Guided inpainting and filtering for kinect depth maps. In: Proc. Int. Conf. Pattern Recognition, pp. 2055–2058. IEEE (2012)
36. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: Advances in Neural Information Processing Systems, pp. 469–477 (2016)
37. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. In: Proc. Int. Conf. Learning Representations (2016)
38. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 3061–3070 (2015)
39. Miralles, R.: An open-source development environment for self-driving vehicles. In: Universitat Oberta de Catalunya, pp. 1–31 (2017)
40. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowledge and Data Engineering **22**(10), 1345–1359 (2010)
41. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: Advances in Neural Information Processing Systems, pp. 1–4 (2017)

42. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2536–2544 (2016)
43. Qi, F., Han, J., Wang, P., Shi, G., Li, F.: Structure guided fusion for depth map inpainting. Pattern Recognition Letters **34**(1), 70–76 (2013)
44. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 pp. 1–16 (2015)
45. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015)
46. Rosales, R., Achan, K., Frey, B.J.: Unsupervised image translation. In: Proc. Int. Conf. Computer Vision, pp. 472–478 (2003)
47. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Computer Vision **47**(1-3), 7–42 (2002)
48. Schwerdtfeger, H.: Introduction to Linear Algebra and the Theory of Matrices. P. Noordhoff Groningen (1950)
49. Shah, S., Dey, D., Lovett, C., Kapoor, A.: AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In: Field and Service Robotics, pp. 621–635 (2017)
50. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2242–2251 (2017)
51. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Proc. Int. Conf. Learning Representations pp. 1–14 (2015)
52. Tao, M.W., Srinivasan, P.P., Malik, J., Rusinkiewicz, S., Ramamoorthi, R.: Depth from shading, defocus, and correspondence using light-field angular coherence. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1940–1948 (2015)
53. Tolstikhin, I.O., Sriperumbudur, B.K., Schölkopf, B.: Minimax estimation of Maximum Mean Discrepancy with radial kernels. In: Advances in Neural Information Processing Systems, pp. 1930–1938 (2016)
54. Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.S.: Texture networks: Feed-forward synthesis of textures and stylized images. In: Proc. Int. Conf. Machine Learning, pp. 1349–1357 (2016)
55. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. In: Proc. Int. Conf. Computer Vision, pp. 3352–3361. IEEE (2017)
56. Wang, F., Huang, Q., Guibas, L.J.: Image co-segmentation via consistent functional maps. In: Proc. Int. Conf. Computer Vision, pp. 849–856 (2013)
57. Wang, X., Gupta, A.: Generative image modeling using style and structure adversarial networks. In: Proc. Euro. Conf. Computer Vision, pp. 318–335. Springer (2016)
58. Woodham, R.J.: Photometric method for determining surface orientation from multiple images. Optical Engineering **19**(1), 191139 (1980)
59. Xie, J., Girshick, R., Farhadi, A.: Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In: Proc. Euro. Conf. Computer Vision, pp. 842–857. Springer (2016)
60. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 4076 – 4084 (2017)
61. Yeh*, R.A., Chen*, C., Lim, T.Y., G., S.A., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with deep generative models. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 6882–6890 (2017). * equal contribution
62. Yi, Z., Zhang, H., Gong, P.T.: DualGAN: Unsupervised dual learning for image-to-image translation. In: Proc. Int. Conf. Computer Vision, pp. 2868–2876 (2017)
63. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–15 (2018)

64. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proc. Conf. Computer Vision and Pattern Recognition, pp. 6612–6619 (2017)
65. Zhou, T., Krahenbuhl, P., Aubry, M., Huang, Q., Efros, A.A.: Learning dense correspondence via 3D-guided cycle consistency. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 117–126 (2016)
66. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. Proc. Int. Conf. Computer Vision pp. 2242 – 2251 (2017)
67. Zhuo, W., Salzmann, M., He, X., Liu, M.: Indoor scene structure analysis for single image depth estimation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 614–622 (2015)