Accepted Manuscript

Generative Adversarial Framework for Depth Filling via Wasserstein Metric, Cosine Transform and Domain Transfer

Amir Atapour-Abarghouei, Samet Akcay,
Grégoire Payen de La Garanderie, Toby P. Breckon

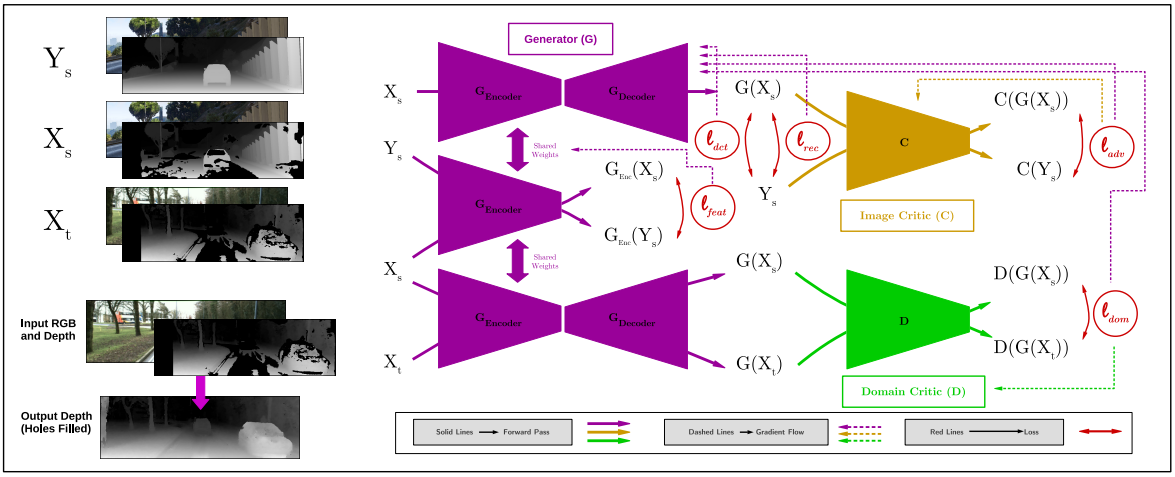
PII: S0031-3203(19)30074-3
DOI: https://doi.org/10.1016/j.patcog.2019.02.010
Reference: PR 6816

To appear in: *Pattern Recognition*

Received date: 1 March 2018
Revised date: 18 December 2018
Accepted date: 17 February 2019

Please cite this article as: Amir Atapour-Abarghouei, Samet Akcay, Grégoire Payen de La Garanderie, Toby P. Breckon, Generative Adversarial Framework for Depth Filling via Wasserstein Metric, Cosine Transform and Domain Transfer, *Pattern Recognition* (2019), doi: https://doi.org/10.1016/j.patcog.2019.02.010

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.


Y_s
X_s
X_t
Input RGB and Depth
Output Depth (Holes Filled)
Generator (G)
G_Encoder
G_Decoder
Shared Weights
G(X_s)
G_Enc(X_s)
G_Enc(Y_s)
G(X_t)
C
C(G(X_s))
C(Y_s)
Image Critic (C)
D
D(G(X_s))
D(G(X_t))
Domain Critic (D)
ℓ_dct
ℓ_rec
ℓ_feat
ℓ_adv
ℓ_dom
Solid Lines → Forward Pass
Dashed Lines → Gradient Flow
Red Lines → Loss

Highlights

- Depth completion can be performed by learning the context and contents of the scene.
- Depth holes can be predicted based on the scene and capture device characteristics.
- Using absolute deviations loss in frequency domain (DCT) improves reconstruction.
- Adversarial training (Wasserstein metric) can improve training and mode selection.
- Using domain transfer, models trained on synthetic data are used in the real world.

Generative Adversarial Framework for Depth Filling via Wasserstein Metric, Cosine Transform and Domain Transfer

Amir Atapour-Abarghouei , Samet Akcay, Grégoire Payen de La Garanderie, Toby P. Breckon

Department of Computer Science, Durham University, UK

Abstract

In this work, the issue of depth filling is addressed using a self-supervised feature learning model that predicts missing depth pixel values based on the context and structure of the scene. A fully-convolutional generative model is conditioned on the available depth information and full RGB colour information from the scene and trained in an adversarial fashion to complete scene depth. Since ground truth depth is not readily available, synthetic data is instead used with a separate model developed to predict where holes would appear in a sensed (non-synthetic) depth image based on the contents of the RGB image. The resulting synthetic data with realistic holes is utilized in training the depth filling model which makes joint use of a reconstruction loss which employs the Discrete Cosine Transform for more realistic outputs, an adversarial loss which measures the distribution distances via the Wasserstein metric and a bottleneck feature loss that aids in better contextual feature execration. Additionally, the model is adversarially adapted to perform

*Corresponding Author

Email addresses: amir.atapour-abarghouei@durham.ac.uk (Amir Atapour-Abarghouei), toby.breckon@durham.ac.uk (Toby P. Breckon)

well on naturally-obtained data with no available ground truth. Qualitative and quantitative evaluations demonstrate the efficacy of the approach compared to contemporary depth filling techniques. The strength of the feature learning capabilities of the resulting deep network model is also demonstrated by performing the task of monocular depth estimation using our pre-trained depth hole filling model as the initialization for subsequent transfer learning.
Keywords: Depth image, Hole filling, Self-supervised learning, Generative model, Adversarial training, Feature distance, Domain adaptation.

---

1. Introduction

The world is visually diverse, irregular and contrastingly structured at the same time. Three-dimensional scenes containing depth information are highly applicable within visual systems such as autonomous driving, augmented reality, environment modelling and alike. Moreover, recent achievements in depth capture technologies, including time-of-flight cameras, stereo correspondence and structured light devices, have made depth accessible in any scene understanding process. However, complete (hole-free) scene depth cannot be acquired facilely using commercial devices and even high-performance depth sensing solutions suffer from a range of environmental noise issues that preclude the recovery of hole-free scene depth under all conditions. This work is an exploration into whether a state-of-the-art learning based approach is capable of understanding the structures and intricacies of a scene, just as humans are, to predict the missing parts of scene depth as a standalone real-time portion of any visual system.

Image completion is considered challenging as it is inherently ill-posed. RGB completion approaches can achieve plausible results, using either local or non-

local information [1, 2, 3, 4, 5]. However, due to the differences between depth and colour images (e.g., absence of granular texture, object separation, and in-scene transferability of varying depth sub-regions), conventional colour image inpainting is considerably less effective within the depth modality [6].

Some depth filling techniques leverage classic image inpainting approaches to complete depth [7]. There have also been attempts to fill a target region in one of a set of multi-view photographs [8], to fill depth using exemplar-based image completion [9], and a myriad of approaches utilizing filters [10], temporal-based methods [11], reconstruction-based methods [12], and others [13, 14, 15].

Deep neural networks have recently been successfully utilized for image stylization [16, 17], super-resolution [18, 19, 20], and colorization [21]. In the realm of image completion, Pathak et al. [22] propose a context encoder which can predict missing regions in a colour image using an adversarial [23] and a reconstruction loss ($_2$). Although the model produces promising results, the absence of fine texture and the existence of visible artefacts near the boundaries of the target region point to flaws in the learning mechanism within the framework.

In a related work, Yeh et al. [24] utilize an analogous framework with similar loss functions to map the input image with missing or corrupted regions to a latent vector, which in turn is passed through their generator that recovers the target content. Despite the large amount of corruption applied to the input, the model generates perceptually plausible outputs. Nevertheless, blurring effects and unwanted artefacts persist in spite of the low resolution of the images.

Yang et al. [25] propose a joint optimization framework composed of two separate networks, a content encoder, based on [22], which is tasked to preserve contextual structures within the image, and a texture network, which enforces

similarity of the fine texture within and without the target region using neural patches [26]. The model is capable of completing higher resolution images than its two earlier counterparts [22, 24] but at the cost of greater inference time since the final output is not achievable via a single forward pass.

Regarding depth images, advances have been made in monocular depth estimation [27, 28, 29, 30, 31] and depth super-resolution [32]. Here, we utilize a generative model trained on synthetic data [33, 34] to complete depth. Since the model is expected to synthesize large portions of depth, it has to adapt to learning image structures and semantics. In existing works on learning-based colour image completion [22, 24, 25], training requires large datasets. The complete image is often considered as the ground truth, and the input is created by adding noise or sparse corruptions [24], removing rectangular blocks [22, 24, 25], or cutting random regions from the image [22]. In the realm of depth filling, however, no such large datasets exist that contain large quantities of ground truth (hole-free) depth. Consequently, synthetic data needs to be acquired from a graphically rendered virtual environment primarily designed for a gaming application [35].

Since depth holes are neither random nor manually created, they are *predictable*, in that they occur due to specific scene features or the capture device. For instance, featureless surfaces such as blank walls and roads, reflective objects, and depth discontinuities, among others can cause depth holes. As a result of this *predictability*, the location of a hole occurrence can be learned via a separate model trained to predict where holes would be in a depth image based on the features present in the scene and the assumption of a specific capture approach.

When high-quality ground truth exists, a model can be naively trained based on a simplistic reconstruction loss ($_1$ or $_2$). However, due to the multi-modality

of image completion, a model trained in this way tends to generate the average of the multiple possible modes in the predictions, which results in an output containing blurring effects. This is why the techniques in [22, 24, 25] and other generative models [36, 37] leverage adversarial training [23] as this assists with mode selection to generate realistic results. However, approaches using Generative Adversarial Networks (GAN) [22, 23, 24, 25] suffer from certain flaws such as unstable training, difficulties in reaching an equilibrium, and vanishing gradients due to premature discriminator optimality and other issues [38, 39, 40]. Here, we utilize an improved adversarial framework [39] that avoids such issues.

Even though an adversarial loss can help diffuse blurring effects, the goal of the adversary should be generating a more realistic image across the board and blurring artefacts still occasionally make their way to the output. This is because the generator feels safer averaging than selecting values. To ease the burden of de-blurring on the adversary, we propose the addition of a loss term based on the Discrete Cosine Transform (DCT) in addition to the conventional $_1$ loss. The DCT preserves an accurate representation of the image structure in its spatial frequency content, which is why it has long been used in de-blurring [41], compression [42] and alike. We utilize the absolute deviations loss ($_1$) in the frequency domain, as this error is far more obvious when the DCT is applied to a blurry averaged image. As seen in Figure 1, the $_1$ distance between the original image and the blurry image, both in the spatial domain, is not very large, but when the same images are transformed into the frequency domain using the DCT, the $_1$ error is much larger and therefore a better indicator of blurring effects.

The task of our generator consists of two stages: reducing the input into a compact representation of itself in the feature space (encoding) and reconstructing the

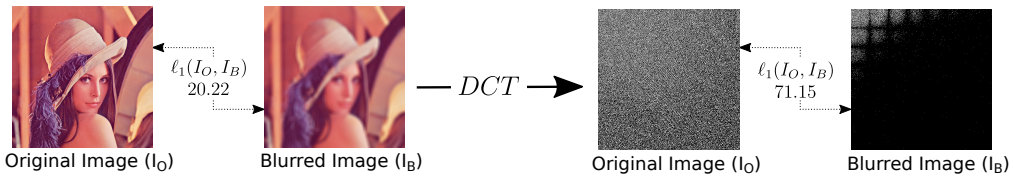


Figure 1: A demonstration of how the DCT makes the absolute deviations loss more susceptible to blurring. Note that the $_1$ distance between the images in the frequency domain is higher and therefore a great tool to identify blurring.

image from these compact features (decoding). Up-convolutions, of any kind, are fraught with intrinsic unpredictability and can lead to bad salient edges and absence of fine texture. As a result, ensuring that the reconstruction starts from a correct and viable feature representation is paramount. We use the feature representations produced in the generator bottleneck in our loss to make sure the scene representation is correctly captured before reconstruction. While the sole use of this as a loss function is inadvisable and can lead to high-frequency artefacts, it is a helpful complement to the reconstruction and adversarial losses.

Our approach is meant to fill holes in depth images acquired via commercially and computationally inexpensive tools (a stereo camera and established stereo correspondence approaches such as [43, 44]) and not in pixel-perfect synthetic depth images only. Therefore, as part of our training procedure, it is vital to guide the model toward capturing the distribution of the natural data. With this in mind, a domain transfer network is trained within the framework to rectify the model such that real-world images can be viable inputs during inference. In short, the contributions of our work are as follows:

- *Novelty* - no comparable approach utilizes a generative model using the Earth Mover's distance to complete depth via the Discrete Cosine Transform based on a synthetic training corpus with predicted holes.

- *Accuracy* - the approach is far more efficient and accurate than comparators (conventional image completion techniques) within a side-by-side comparison framework (Tables 2 and 3; Figures 10 and 11).
- *Representation Learning* - our model is capable of learning better semantics and context as illustrated by superior sharp and artefact-free qualitative outputs when performing monocular depth estimation (Figures 12 and 13).

In the following section, we present an overview of the literature relevant to this work. Section 3 provides a discussion on the data preparation process and Section 4 contains a detailed outline of the proposed hole filling approach. Results are evaluated in Section 5 and the work is finally concluded in Section 6.

2. Related Work

There have recently been remarkable strides made in complex learning-based computer vision problems such as image classification [45, 46, 47, 48, 49, 50], semantic segmentation [51, 52, 53], and image generation [23, 36, 38, 39, 54, 55, 56]. Inspired by the capabilities of recent generative models [22, 23, 36, 38, 39], we attempt to complete depth images by learning the details of a scene.

Generative Adversarial Networks (GAN) have revolutionized the field and are capable of producing semantically sound samples by creating a competition between a generator (G), which endeavours to capture the data distribution, and a discriminator (Dis), which judges the generator output and penalizes unrealistic images. Both networks are trained simultaneously to achieve an equilibrium. More formally put, this competition follows the minimax objective [23]:

$$\min_{G} \max_{Dis} \; \mathop{\mathbb{E}}_{x \; \mathbb{P}_r} [log(Dis(x))] + \mathop{\mathbb{E}}_{\tilde{x} \; \mathbb{P}_g} [log(1 - Dis(\tilde{x}))], \tag{1}$$

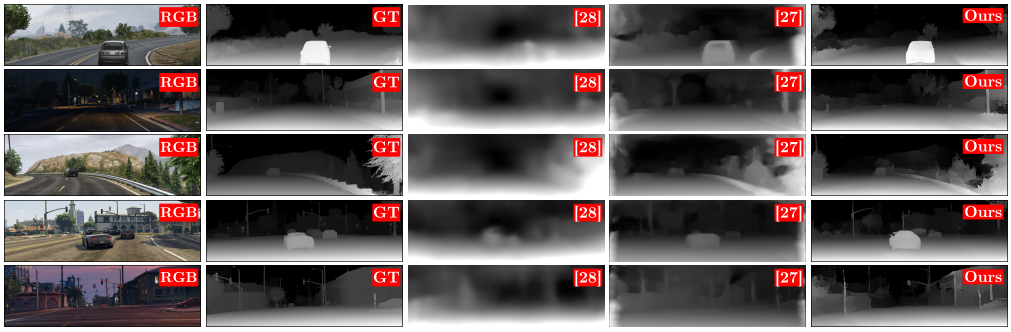


Figure 12: The results of our approach re-purposed to estimate depth from an RGB image compared against [27] and [28] with synthetic data.

over a single epoch of the dataset without any layer freezing. The results are compared with state-of-the-art approaches [27, 28, 29]. Qualitative results based on synthetic images used as inputs are seen in Figure 12.

No domain adaptation to our real-world set was performed during this experiment but the models are evaluated using our real-world test images, nonetheless. As seen in Figure 13, even though our network has never seen a real-world image and data domain has not been transferred, we can see that our approach produces sharper and more crisp depth information despite the anomalies that persist due to domain bias issues. Quantitative analysis using synthetic ground truth images is presented in Table 4. While our approach cannot outperform directly-supervised models trained on similar synthetic data [29], we can see it has succeeded in a task

Method	Mean $_1$ Error	Mean $_2$ Error	PSNR (dB)	SSIM $(-1, 1)$
Result of [28]	28.61	13.68	10.15	0.374
Result of [27]	14.46	3.93	14.22	0.565
Result of [29]	4.22	0.79	24.18	0.793
Our Result	4.97	0.88	22.35	0.778

Table 4: Our pre-trained model tasked with monocular depth estimation compared to [27, 28, 29]. More realistic images have lower error values, but with PNSR and SSIM, higher values are better.

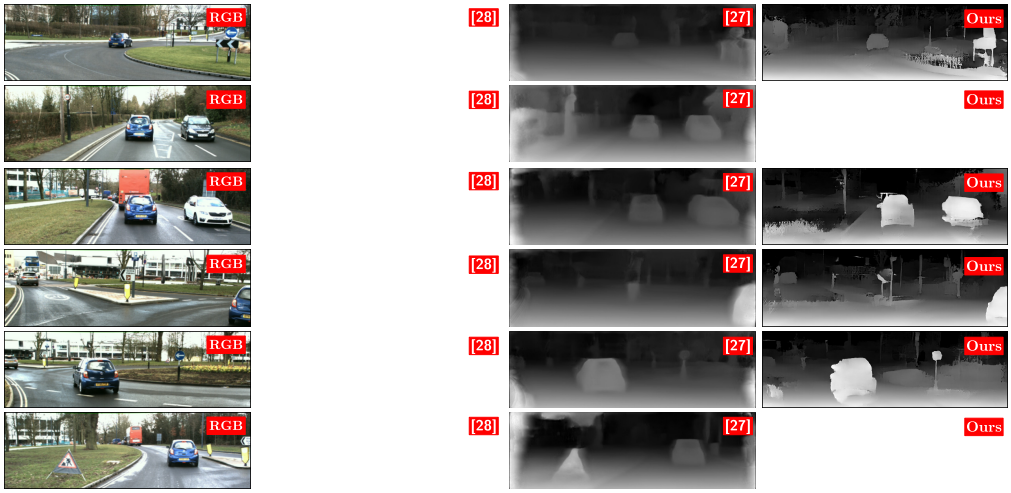


Figure 13: The results of our approach re-purposed to estimate depth from an RGB image compared against [27] and [28] with natural real-world data.

it is not primarily designed for due to its strength in scene feature learning.

6. Conclusion

We have approached the problem of hole filling in depth images from a learning perspective by employ an adversarially trained self-supervised encoder/decoder architecture. It is expected that if enough is learned about the contents and semantics of a scene, missing regions of a depth image can be inferred given the known regions and the full RGB view. Our training is fully self-supervised, i.e. at no point is annotation or human intervention necessary. The ground truth depth used for training is acquired from a graphical environment developed for gaming and a separate model is trained to infer where holes would be if the data were obtained via stereo correspondence. The model objective is to minimize a loss consisting of four loss components: reconstruction, adversarial, bottleneck feature and domain transfer loss, which results in filling depth holes, not only in synthetic depth images but also in real-world data with no ground truth.