Unsupervised Enhancement of Real-World Depth Images Using Tri-Cycle GAN

Alona Baruhov
Technion – Israel Institute of Technology
alonabar@campus.technion.ac.il

Guy Gilboa
Technion – Israel Institute of Technology
guy.gilboa@ee.technion.ac.il

Abstract

Low quality depth poses a considerable challenge to computer vision algorithms. In this work we aim to enhance highly degraded, real-world depth images acquired by a low-cost sensor, for which an analytical noise model is unavailable. In the absence of clean ground-truth, we approach the task as an unsupervised domain-translation between the low-quality sensor domain and a high-quality sensor domain, represented using two unpaired training sets. We employ the highly-successful Cycle-GAN to this task, but find it to perform poorly in this case. Identifying the sources of the failure, we introduce several modifications to the framework, including a larger generator architecture, depth-specific losses that take into account missing pixels, and a novel Tri-Cycle loss which promotes information-preservation while addressing the asymmetry between the domains. We show that the resulting framework dramatically improves over the original Cycle-GAN both visually and quantitatively, extending its applicability to more challenging and asymmetric translation tasks.

1. Introduction

Depth information acquired by low-cost depth cameras is typically prone to severe errors and degradations. This low image quality limits the performance of depth-based computer vision algorithms, and challenges most image enhancement methods. In this work, we aim to enhance these depth images and bring them closer to the output of high-quality depth cameras. We focus on enhancing real-world depth images, as produced for example by the Intel Realsense R200 (see Figure 1, left). Due to its small size and low operating power, this camera suffers from substantial noise and artifacts, exhibiting complex and non-random patterns. The absence of any analytic model for these degradations prohibits the use of many classical methods, such as probabilistic and model-based reconstruction methods [1]. Furthermore, it makes simulating realistic degraded depth maps impractical, eliminating the possibility of generating pairs of high- and low-quality images for supervised machine learning algorithms.

As an alternative, we propose a novel approach which eases the requirement for aligned ground-truth image pairs, formulating the task as an unsupervised domain-translation problem between a low-quality sensor domain and a high-quality sensor domain. Several works [2,3,4] have recently shown great success in handling such unsupervised domain translation problems. Following their success, we employ a similar approach to the challenging depth enhancement task. We base our approach on the Cycle-GAN framework, and develop a fully unsupervised method for training the enhancement network. To the best of our knowledge, this is the first work to formulate this depth enhancement task as an unsupervised translation task.

We focus on the low-power RealSense R200 stereo camera as our low-quality depth sensor. As the high-quality sensor, we select the time-of-flight Microsoft Kinect 2, which is a significantly higher-powered and more accurate camera with substantially less noise. Our aim therefore becomes to bring the quality of the RealSense images to that of the Kinect 2 images via unsupervised domain translation.

Unfortunately, we find the original Cycle-GAN to perform poorly on this task, as depicted in Figure 1 (center). The main sources of this deficiency are the increased complexity of the task, as well as the asymmetry between the domains, manifested by the lack of information equivalence between them. To address these issues, we introduce several modifications to the framework. First, we replace the relatively small generative architecture with a much larger one, with sufficient representational capacity to handle the translation task. Next, we employ depth-specific losses which...
take into account missing pixels. Finally, we propose the Tri-Cycle loss as an alternative information-retention metric for asymmetric domains. Combining these components, our modified Cycle-GAN framework significantly improves over "vanilla" Cycle-GAN in this task, producing much more detailed and less noisy images, as demonstrated in Figure 1 (right). Our main contributions are therefore:

1. Developing a training method for depth enhancement networks, capable of handling real-world depth with severe degradation, and without requiring labeled data.
2. Presenting architectural design principles for CNNs aimed at processing highly degraded depth data with strong non-Gaussian noise, missing pixels, and structured artifacts.
3. Proposing the Tri-Cycle loss that extends the applicability of Cycle-GANs to asymmetric tasks which may not satisfy the information-preservation assumption.

This work is organized as follows. We begin in Section 3.1 with a discussion of the specific challenges of real-world depth and its complex noise sources. We formulate the enhancement problem as an unsupervised translation task in Section 3.2 and discuss the limitations of the original Cycle-GAN which prevent it from producing reasonable recovery results in this case. We next describe our modifications to the Cycle-GAN framework, including the network architecture and considerations in designing it (Section 3.2.1), the depth-specific losses (Section 3.2.2), and the Tri-Cycle loss (Section 3.2.3). As discussed later, the Tri-Cycle loss can be interpreted as a nonlinear generalization of the Moore-Penrose inverse for asymmetrical translation problems, and in our view is the main innovation in this work. We continue by providing experimental results on several datasets in Section 4, demonstrating the effectiveness of the improved Cycle-GAN both visually and quantitatively. We discuss and conclude in Section 5.

2. Related Work

Depth map completion and enhancement have received considerable attention over the past years. Depth completion methods can generally be divided into two categories: color-guided and non-guided methods. Color-guided approaches 5 6 7 8 assume the existence of a color image aligned with the corrupted depth image, and rely on the fact that both share much of the structural information — such as object edges — to deduce a dense depth map from the low-quality input. For example, 5 uses a CNN to estimate surface normals and edges from the color image, and subsequently combines them with the low quality depth image in a post-process. Other works, such as 6, directly infer the underlying relation between depth and color, and output an enhanced depth map in a single end-to-end process.

When aligned color is not available, either due to a lack of a color sensor, the absence of an alignment between the depth and color streams, low-light conditions, or (as in the RealSense case) the existence of a projected pattern in the visible image, a non-guided completion method must be used 9 10 11 12. For example, Sparse Depth Sensing 9 reconstructs dense depth maps from very sparse measurements by modeling the scene as a piecewise-planar map, and formulating the recovery task as a compressed sensing problem regularized by sparse second derivatives.

Sparsity-Invariant CNNs 10 take a different approach, and learn an image-to-image enhancement network based on sparse convolutions, which consider only valid depth values when computing convolution outputs. However, in a follow-up work 11, the authors note that sparse convolutions rapidly lose their effectiveness after only a few convolutional layers, and thus elect to fill-in missing pixels using a deep architecture based on traditional convolutions instead. The work introduces a unique sparse training strategy which synthetically varies the density of valid pixels in the input during training (though in relatively simple patterns), and is found to outperform 10 even at the density for which it was trained. In parallel, a GAN-based approach was proposed in 12, introducing an adversarial loss to the supervised depth completion task. The added adversarial loss is shown to notably improve both realism and accuracy of the recovered images compared to previous methods.

Despite the convincing results of all these works in their respective tasks, we note that they all adopt a supervised approach to the training process, relying on the availability of ground-truth images alongside the degraded ones. This is often achieved by limiting the method to simple degradations which can be synthetically reproduced, such as i.i.d depth noise and randomly distributed missing pixels. In the case of real-world depth images, however, such assumptions often do not hold. Thus, in this work we take a different approach, and formulate the enhancement task as an unsupervised problem which does not require ground-truth images. In this way, we address the task of enhancing depth maps captured by real world low-quality depth cameras, and develop a framework for handling this challenging task.

3. Improving Depth Images Using Cycle-GAN

3.1. The Challenge of Real-World Depth

Low power, small form-factor cameras such as the Intel RealSense R200 suffer from significant noise and artifacts in the captured depth maps. As an active stereo camera, the main sources of error include inaccuracies in the pattern matching — due to the algorithm itself or to insufficient information in the scene — as well as from shadowing due to the different viewpoints of the two sensors. These are all amplified by the small camera baseline and the low power of
Following previous work, we adopt the highly successful Cycle-GAN \cite{gans} \cite{cycle-gan} \cite{cycle-gan2} as the basis for our domain translation framework. The Cycle-GAN simultaneously learns two generative networks for translating in both directions, and uses cycle-consistency to encourage information-preservation by the translation in the absence of ground-truth targets. Specifically, given the two domains $L$ and $H$, the loss function of the Cycle-GAN is given by:

$$
\mathcal{L}(G_{L \rightarrow H}, G_{H \rightarrow L}) = \\
\lambda_1 \mathcal{L}_H(G_{L \rightarrow H}) + \lambda_2 \mathcal{L}_L(G_{H \rightarrow L}) + \\
\lambda_3 \mathcal{E}_{I_H \in H} \| G_{L \rightarrow H}(G_{H \rightarrow L}(I_H)) - I_H \| + \\
\lambda_4 \mathcal{E}_{I_L \in L} \| G_{H \rightarrow L}(G_{L \rightarrow H}(I_L)) - I_L \| + \\
\lambda_5 \mathcal{E}_{I_H \in H} \| G_{L \rightarrow H}(I_H) - I_H \| + \\
\lambda_6 \mathcal{E}_{I_L \in L} \| G_{H \rightarrow L}(I_L) - I_L \|. 
$$

(1)

Here, $G_{L \rightarrow H}$ and $G_{H \rightarrow L}$ are the two learned translators, $I_L$ and $I_H$ are images from the two domains, and $\mathcal{L}_L$ and $\mathcal{L}_H$ are adversarial losses \cite{adversarial} for their respective domains, each incorporating a learned discriminator working against the generator (we omit the full definition of this loss for conciseness). The first two losses in this formulation guide the translators to output images in their correct domains (represented via the exemplar images from each domain), the next two losses are the cycle-consistency losses, and the final two losses are the identity losses which regularize the training process, and were introduced in \cite{gans}.

The nature of the depth data, however, poses significant challenges to the Cycle-GAN framework. First, the low-quality depth exhibits significantly stronger and more complex noise patterns than traditionally used with the Cycle-GAN, and large missing regions create severe discontinuities in the data. Furthermore, we observe that the information preservation assumption made by the Cycle-GAN design does not actually hold in our case — specifically, the two depth domains are in fact not equivalent, with the high-quality domain containing distinctly more information than the low-quality one. Thus, the cycle-consistency constraint which forms the basis of the Cycle-GAN becomes problematic in this case. To address these issues, we modify several key aspects of the original Cycle-GAN formulation, enabling it to successfully handle this challenging task. In the next sections we detail these modifications.

### 3.2. Unsupervised Depth Image Improvement

Considering the huge challenge in producing pairs of input-output images for real-world depth enhancement, we believe that the most viable path for training such a process is unsupervised learning. In this approach, the problem is re-stated as a translation problem between two domains — a low-quality domain $L$ and a high-quality domain $H$, represented by two unaligned, freely captured training sets. Such translation tasks have recently received significant attention, and have shown remarkable results in many translation problems \cite{gans} \cite{cycle-gan} \cite{cycle-gan2} \cite{unsupervised} \cite{unsupervised2}.

Following previous work, we adopt the highly successful Cycle-GAN \cite{gans} \cite{cycle-gan} \cite{cycle-gan2} as the basis for our domain translation framework. The Cycle-GAN simultaneously learns two generative networks for translating in both directions, and uses cycle-consistency to encourage information-preservation by the translation in the absence of ground-truth targets. Specifically, given the two domains $L$ and $H$, the loss function of the Cycle-GAN is given by:

$$
\mathcal{L}(G_{L \rightarrow H}, G_{H \rightarrow L}) = \\
\lambda_1 \mathcal{L}_H(G_{L \rightarrow H}) + \lambda_2 \mathcal{L}_L(G_{H \rightarrow L}) + \\
\lambda_3 \mathcal{E}_{I_H \in H} \| G_{L \rightarrow H}(G_{H \rightarrow L}(I_H)) - I_H \| + \\
\lambda_4 \mathcal{E}_{I_L \in L} \| G_{H \rightarrow L}(G_{L \rightarrow H}(I_L)) - I_L \| + \\
\lambda_5 \mathcal{E}_{I_H \in H} \| G_{L \rightarrow H}(I_H) - I_H \| + \\
\lambda_6 \mathcal{E}_{I_L \in L} \| G_{H \rightarrow L}(I_L) - I_L \|. 
$$

(1)

Here, $G_{L \rightarrow H}$ and $G_{H \rightarrow L}$ are the two learned translators, $I_L$ and $I_H$ are images from the two domains, and $\mathcal{L}_L$ and $\mathcal{L}_H$ are adversarial losses \cite{adversarial} for their respective domains, each incorporating a learned discriminator working against the generator (we omit the full definition of this loss for conciseness). The first two losses in this formulation guide the translators to output images in their correct domains (represented via the exemplar images from each domain), the next two losses are the cycle-consistency losses, and the final two losses are the identity losses which regularize the training process, and were introduced in \cite{gans}.

The nature of the depth data, however, poses significant challenges to the Cycle-GAN framework. First, the low-quality depth exhibits significantly stronger and more complex noise patterns than traditionally used with the Cycle-GAN, and large missing regions create severe discontinuities in the data. Furthermore, we observe that the information preservation assumption made by the Cycle-GAN design does not actually hold in our case — specifically, the two depth domains are in fact not equivalent, with the high-quality domain containing distinctly more information than the low-quality one. Thus, the cycle-consistency constraint which forms the basis of the Cycle-GAN becomes problematic in this case. To address these issues, we modify several key aspects of the original Cycle-GAN formulation, enabling it to successfully handle this challenging task. In the next sections we detail these modifications.

#### 3.2.1 Network Architecture

Since a large part of the difficulty in low-quality depth comes from the high number of missing pixels, one may be tempted to consider architectures based on sparse convolutions \cite{sparse} \cite{sparsely}, which are a type of layer specifically designed for inpainting problems. In a masked convolution, only known pixels contribute to the result of the convolution, with each output feature normalized by the num-
Kinect 2 camera suffers from noise as well. Specifically, mainly on the structured noise of the RealSense camera, the involved, computing similarity over the entire image may and identity losses. However, when missing pixels are in-

The Cycle-GAN uses image similarity as a central compo-
tecture is detailed in Table (1).

3.2.2 Depth-Specific Losses

The Cycle-GAN uses image similarity as a central compo-
ent in the training process, in both the cycle-consistency and identity losses. However, when missing pixels are in-

Table 1. Generator architecture. Each layer consists of a 3 × 3 convolu-
tion, Leaky ReLU, and instance normalization [21], with a stride of 1 or 2 depending on the output size. The + operator represents channel-wise concatenation,  up(·) denotes ×2 nearest-
neighbor upsampling, and  conv(·) denotes a size-maintaining con-
volution with Leaky ReLU and instance normalization.

![Figure 3. U-Net architecture with skip connections (schematic).](image)

![Figure 4. Kinect 2 image with scattered holes (in dark blue).](image)
Formally, given a known depth image \( Z \) with valid pixel mask \( M := (z_{i,j} \neq 0)_{i,j} \), and given a second depth image \( \hat{Z} = f(Z) \), we define the masked similarity loss as
\[
d(Z, \hat{Z}; M) = \| M \circ (Z - \hat{Z}) \|_1 ,
\]
where \( \circ \) denotes element-wise (Hadamard) multiplication.

We note that \( \hat{Z} \) is not symmetric in \( Z \) and \( \hat{Z} \). We use a non-symmetric loss since the valid pixel mask \( M = M(\hat{Z}) \) of the output depth is non-differentiable in the network parameters and is unstable near \( 0 \), and thus optimizing with respect to it would be impractical. Furthermore, the symmetric variant would encourage the formation of holes in the output, and in fact has a trivial global minimum at \( \hat{Z} = 0 \). In contrast, the asymmetric similarity generally prefers filling-in holes in the output image, while still allowing isolated holes to form owing to the robust \( H \)-norm.

Finally, another issue we have observed with the original Cycle-GAN is range preservation. Specifically, for any solution \((G_{H \rightarrow L}, G_{L \rightarrow H})\) of \((1)\), the solution \((G_{H \rightarrow L} + z_0, G_{L \rightarrow H} - z_0)\), where \( z_0 \) represents a depth shift by \( z_0 \) of the non-zero values, is also equally valid. To counter this effect, we add a small masked similarity loss to the \( L \rightarrow H \) translation, requiring that the high-quality image be close to the low-quality one where it is non-zero. Formally, this loss is given by:
\[
\mathbb{E}_{Z \in L} \| M_L \circ (G_{L \rightarrow H}(Z_L) - Z_L) \|_1 ,
\]
where \( M_L \) is the valid mask of the low-quality image. We note that since this image typically has significantly more holes than the expected high-quality output, this loss essentially just maintains the overall distance of the known objects in the scene, without affecting the visual properties of the output image.

### 3.2.3 Tri-Cycle Loss

The Cycle-GAN measures information preservation by passing images through a full cycle of the domain translation process, and requiring the result to be an identity operator. However, particularly in the \( H \rightarrow L \) translation, this transform is in fact a one-to-many mapping, as the low-quality image may degrade in many different ways. An example of this is shown in Figure 5. Formally, for a 3D scene \( S \) and viewpoint \( V \), the depth image is ideally a projection \( Z = P(S; V) \) of the scene on the camera plane. However, due to noise and errors in the capturing process, we obtain a measured depth image which can very roughly be expressed as \( \hat{Z} = M \circ (Z + n) \), with \( n \) the depth noise and \( M \) a mask image. Thus, for a fixed scene and viewpoint, we may measure any one of many possible depth images \( \{ \hat{Z} \} \). In the presence of holes and strong depth errors, this set can become of significant size, in contrast to e.g., a color camera where the noise model can be approximated as a Gaussian or Poisson source with typically low variance, leading to a relatively compact set.

To address this, we propose using an asymmetric loss function for promoting information preservation, which does not require the two translations to be inverses. Instead, this loss essentially requires that when performing a full cycle of the form \( L \rightarrow H \rightarrow L \), we simply produce a low-quality image which could feasibly reproduce the high-quality one, but not necessarily the same one we began with.

To this end, we regard the Kinect camera as a high-quality camera with relatively low noise, and hence consider the volume of the set \( \{ \hat{Z} \} \) to be negligible. However, this does not hold for the low quality RealSense camera, where the capturing process is substantially less stable, and multiple frames of the same scene may display large variation.

Given a low-quality depth image \( Z \) captured from the underlying scene \( S = S(Z) \) and viewpoint \( V = V(Z) \), we denote by \( \psi(Z) \) the set of all low quality images which could have been captured under the same conditions:
\[
\psi(Z) = \left\{ \hat{Z} = M \circ (P(S; V) + n) \mid (M,n) \sim \mathcal{P}_L(S,V) \right\} .
\]
Here, \( \mathcal{P}_L(S,V) \) denotes the joint distribution of plausible low-quality depth noise and hole patterns corresponding to the underlying scene. The set \( \psi(Z) \) forms the equivalency set of \( Z \), and as previously noted, has a non-negligible volume due to the properties of the low-quality depth camera.

Returning to the cycle-consistency assumption broken in the \( L \rightarrow H \rightarrow L \) case, as the second translation is a one-to-many mapping. Clearly, requiring this cycle to be the identity mapping is an unnecessarily difficult constraint. We thus propose to relax the cycle-consistency constraint in this case, such that the output belongs to the equivalency set of the input, rather than equal it. This translates to the constraint:
\[
G_{H \rightarrow L}(G_{L \rightarrow H}(Z_L)) \in \psi(Z_L) .
\]
Unfortunately, enforcing this constraint directly is impractical, as the set \( \psi(Z_L) \) is a complex, non-convex set with no analytical form. However, if we apply \( G_{H \rightarrow L} \) to both sides of this expression, and by using the fact that \( G_{L \rightarrow H}(\hat{Z}_L) = G_{L \rightarrow H}(Z_L) \) iff \( \hat{Z}_L \in \psi(Z_L) \), we can re-write the above as:
\[
G_{L \rightarrow H}(G_{H \rightarrow L}(G_{L \rightarrow H}(Z_L))) = G_{L \rightarrow H}(Z_L) .
\]
\[ \mathcal{L}(G_{L \rightarrow H}, G_{H \rightarrow L}) = \lambda_1 \mathcal{L}_H(G_{L \rightarrow H}) + \lambda_2 \mathcal{L}_L(G_{H \rightarrow L}) + \lambda_3 \mathbb{E}_{Z_H \in H} \|M_H \circ (G_{L \rightarrow H}(G_{H \rightarrow L}(Z_H)) - Z_H)\|_1 + \lambda_4 \mathbb{E}_{Z_L \in L} \|G_{H \rightarrow L}(G_{H \rightarrow L}(Z_L)) - Z_L\|_1 + \lambda_5 \mathbb{E}_{Z_H \in H} \|M_H \circ (G_{L \rightarrow H}(Z_H)) - Z_H\|_1 + \lambda_6 \mathbb{E}_{Z_L \in L} \|G_{H \rightarrow L}(Z_L) - Z_L\|_1 + \lambda_7 \mathbb{E}_{Z_L \in L} \|M_L \circ (G_{L \rightarrow H}(Z_L)) - Z_L\|_1 + \lambda_8 \mathbb{E}_{Z_L \in L} \|G_{L \rightarrow H}(G_{H \rightarrow L}(G_{L \rightarrow H}(Z_L))) - G_{L \rightarrow H}(Z_L)\|_1 \]  

(Adversarial)  

(H cycle)  

(L cycle)  

(H identity)  

(L identity)  

(Depth preserve)  

(Tri-cycle)  

Table 2. Full loss function for the depth enhancement Tri-Cycle GAN.

This requirement readily translates to a loss function, which we term the Tri-Cycle loss due to the application of three consecutive translations in its definition. Incorporating an \(L_1\) distance norm, and accumulating over the entire low-quality domain, this loss becomes:

\[ \mathcal{L}_{tri}(G_{L \rightarrow H}, G_{H \rightarrow L}) = \mathbb{E}_{Z_L \in L} \|G_{H \rightarrow L}(G_{H \rightarrow L}(G_{L \rightarrow H}(Z_L))) - G_{L \rightarrow H}(Z_L)\|_1. \]  

It is interesting to note the similarity between the above Tri-Cycle loss and the linear Generalized Inverse. In the linear case, we may consider the inversion of a dimension-reducing matrix (i.e., a many-to-one mapping) \(A : x \in \mathbb{R}^m \rightarrow y \in \mathbb{R}^n\) with \(m < n\), with the set \(\psi(x)\) in this case being the Affine set of all vectors \(x\) mapped to the same \(y\). The Penrose conditions for inverting such a matrix \([22]\) essentially require that the inverse transform map every \(y = Ax\) to one of the vectors which would have been mapped to it by \(A\), i.e., \(A^+ y = A^+ A x \in \psi(x)\), which is formalized by the condition \(AA^+ A x = A x\). Indeed, our Tri-Cycle constraint follows very similar reasoning. In this sense, we may view the mapping \(G_{H \rightarrow L}\) as a generalized inverse of \(G_{L \rightarrow H}\), and the Tri-Cycle formulation as seeking one of these mappings as part of the optimization process.

3.3. Full Loss Function and Optimization Method

Our full depth enhancement architecture optimizes a combined penalty consisting of all the losses discussed in the previous sections. The full loss function is given in Table 2. We optimize using ADAM \([23]\) with batch size 1, with each batch consisting of a pair of random low-quality and high-quality images sampled from \(L \times H\). We use a constant learning rate of 0.001, and augment the examples with random crops, 90-degree rotations, horizontal and vertical flips, and random shifts in depth.

4. Experimental Evaluation and Results

4.1. Synthetic Experiments

To quantify the performance of our Cycle-GAN framework, we use a high-quality dataset of rendered depth images, to which we apply noise in a process simulating a depth camera. Our dataset is based on the Physically Based Rendering Dataset \([24]\) consisting of 568,793 depth images randomly sampled from the SUNCG set of 3D scenes \([25]\). We further filter the data by removing images with very low standard deviation (\(\sigma < 400\)mm) or with more than 15% distant pixels (\(d > 5000\)mm), as we are emulating a depth camera with limited range. The resulting synthetic dataset contains around 120,000 images, see Figure 6 (top).

For the depth noise, we apply several degradations typical of depth cameras. Unfortunately, it is extremely difficult to emulate the highly structured noise of the RealSense camera. However, our process includes several noise sources which are common to depth cameras such as the RealSense and Kinect. These include structural noise, generated by adding Gaussian noise to a down-sampled version of the image, followed by nearest-neighbor upsampling; object boundary noise, produced by removing pixels near object edges with a probability of \(p = 80\%\); depth-
adaptive noise, generated as random Gaussian noise with a distance-dependent standard deviation $\sigma \sim z^3$; and depth-adaptive holes, generated by randomly eliminating pixels from the image with a probability $p \sim z^2$. Figure 6 (bottom) shows a few noisy images produced by this process.

We train a translation network to convert between the noisy and noiseless depth domains. Our experiments compare the performance of our full Cycle-GAN framework to the original formulation, as well as to the original Cycle-GAN but with the larger generator architecture. In addition, we compare our results to those of the recent Sparse Depth Sensing depth enhancement algorithm [9].

For the quantitative comparison, it is well-known that traditional measures such as PSNR are unreliable as image quality estimators, particularly when adversarial and perceptual losses are involved [26, 27, 28, 29]. Indeed, in most cases methods which directly optimize MSE will outperform perceptual methods in terms of PSNR, while in reality they produce over-smoothed images which lack detail. Thus, to more accurately quantify recovery of detail, we instead propose a patch-based normalized cross-correlation (PNCC) measure. This measure computes the similarity between two images by computing the normalized cross-correlations between their local patches (with overlap), and averaging the results. Formally, given two images $X$ and $Y$, we define $PNCC(X, Y)$ in terms of a block size $b$ and a step size $s$. Denoting by $I_{i,j}^b$ the patch of image $I$ beginning at $(i, j)$ and extending to $(i+b-1, j+b-1)$ (inclusive), the similarity between $X$ and $Y$ is computed as:

$$PNCC(X, Y) = \frac{1}{N} \sum_{i \in [0, s, 2s, \ldots]} \sum_{j \in [0, s, 2s, \ldots]} \rho(X_{i,j}^b, Y_{i,j}^b) \quad (8)$$

Here, $\rho(\cdot, \cdot)$ is the normalized cross correlation function and $N$ is the total number of patches in the sum.

Table 3 lists the quantitative results of the synthetic experiment in terms of PNCC. Left to right: base Cycle-GAN, Cycle-GAN with improved generator, our result with Tri-Cycle loss, result of [9].

| Base       | Improved Net   | Tri-Cycle     | Sparse Sensing |
|------------|----------------|--------------|----------------|
| 0.668      | 0.869          | 0.879        | 0.736          |

Table 3. Quantitative results of the synthetic experiment in terms of PNCC. Left to right: base Cycle-GAN, Cycle-GAN with improved generator, our result with Tri-Cycle loss, result of [9].

4.2. Experiments with Real Depth Data

To demonstrate our method in more real conditions, we created a dataset of real-world depth images captured in an office setting. The images were captured independently using the RealSense and Kinect 2 cameras, with no synchronization between them. After some basic filtering (e.g., removing similar images or images with very little content) we arrived at a dataset consisting of just over 1,000 images from each camera. We note that due to the unconstrained manner in which this dataset was captured, we have no ground-truth for these images, and thus can only perform a qualitative evaluation of the results. On the other hand, the construction of this dataset makes it truly unsupervised, and thus well-representative of a real-world scenario. Fig-

![Figure 7. Example enhancement results on the synthetic dataset.](image-url)
Figure 8 shows an example from this dataset, with additional examples provided in the results figures.

Figure 8 demonstrates the effects of the generator architecture on the results. To isolate the parameters for this experiment, we do not employ any of the new losses in this case, and only vary the network architecture. We consider the following architectures: (1) original network; (2) original network with increased number of channels; (3) original network with increased number of layers; and (4) the final network. As can be seen, the original architecture is essentially unusable for this task, producing strong artifacts and providing no visible enhancement. Increasing the number of channels or the number of layers each have a notable effect in terms of reducing artifacts and filling-in holes, though with limited success. Combining both modifications produces the best results, with the fewest visible artifacts and the most accurate hole filling. It is thus clear that both width and depth are crucial for handling the challenges of low-quality depth. We note that in particular, the increased number of channels in the earlier network layers deviates from the standard practice for CNNs [30], though proves advantageous in this case.

Finally, Figure 9 shows some recovery results of our full Tri-Cycle GAN framework. As before, we compare our results to those of Sparse Depth Sensing [9]. We also show results with and without the Tri-Cycle loss, to demonstrate its effects on recovery performance. As can be seen, the method [9] struggles with these images, exhibiting over-smoothness, jagged object edges, and intensification of outlier pixels leading to unnatural holes in objects. Clearly, the degradation model assumed by this method is too simplistic for this task. Continuing with the Cycle-GAN, increasing the network size has a significant effect on the results, though the output still suffers from visible artifacts and missing regions. Adding the Tri-Cycle loss leads to a notable improvement in the results, producing more realistic and detailed images with fewer artifacts and missing pixels. Indeed, as many of these artifacts are in regions which were strongly corrupted in the input image, we attribute these improvements to the Tri-Cycle loss, which relaxes the requirement to recover the exact degraded input by the inverse translation.

4.3. Experiments with the DROT Dataset

The Depth Restoration Occlusionless Temporal dataset, or DROT [13], is a carefully captured and post-processed set of RealSense and Kinect 2 images, which are nearly pixel-level aligned. The dataset consists of 112 image sets which have been recorded in a studio setting, employing

\[1\]The dataset also includes color, Kinect 1, and 3D DAVID images, though we do not use these in this work.
Figure 10. Example aligned image set from the DROT dataset.

Table 4. Quantitative results for the DROT dataset in terms of PNCC. Left to right: base Cycle-GAN, our result, result of [9].

| Base  | Tri-Cycle | Sparse Sensing |
|-------|-----------|----------------|
| 0.213 | 0.633     | 0.611          |

A highly accurate calibration process between the cameras. Figure 10 shows an example from this dataset. We use this dataset to quantify the performance of our method on actual RealSense depth maps. It should be noted, though, that due to the controlled environment and specifically-chosen scene and materials, this dataset exhibits much lighter degradations than those our method was intended to handle.

Table 4 details our quantitative results on this dataset, and Figure 11 shows some example results. As can be seen, the original Cycle-GAN remains unusable in this case. However, both our method and Sparse Depth Sensing produce very competitive results, with each exhibiting different visual strengths and artifacts. Specifically, our method produces sharper edges and more accurate geometries and boundaries, whereas [9] produces images with no missing pixels and with negligible depth shift. Perhaps ironically, the main limitation of our method may be its own success — specifically, as our network was trained to produce convincing Kinect 2 images, it has also learned to reproduce its typical artifacts and noise patterns, such as missing pixels in this case. Nonetheless, the visual results as well as the higher PNCC scores of our method support its improved reconstruction of geometry and detail in this case.

5. Conclusions

Enhancing depth images with real-world noise is an immensely challenging task, with few practical solutions at this point. Formulating the problem as an unsupervised translation task dramatically simplifies dataset construction, however, the existing Cycle-GAN framework is found to be insufficient for this complex task. To overcome this, we proposed several modifications to the framework: a much larger generator architecture designed to handle low-quality depth, use of depth-specific masked similarity losses, and importantly, the asymmetric Tri-Cycle loss which promotes information-preservation between non-equivalent domains. We have tested these modifications on three datasets, and found them to dramatically improve over the base Cycle-GAN in all cases, producing sharp, detailed, and realistic-looking images. We conclude that the proposed approach enables effective enhancement of real-world depth images with severe noise and degradations, expanding the applicability of the Cycle-GAN to asymmetric tasks which do not necessarily satisfy the cycle-consistency assumption.

References

[1] C. R. Vogel, *Computational methods for inverse problems*. SIAM, 2002, vol. 23.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1701–1708.

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8789–8797.

[5] Y. Zhang and T. Funkhouser, “Deep depth completion of a single rgb-d image.”

[6] F. Mal and S. Karaman, “Sparse-to-dense: Depth prediction from sparse depth samples and a single image,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 1–8.
[7] L. Pan, Y. Dai, M. Liu, and F. Porikli, “Depth map completion by jointly exploiting blurry color images and sparse depth maps,” in IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 1377–1386.

[8] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, and M. Pollefeys, “Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3313–3322.

[9] F. Ma, L. Carlone, U. Ayaz, and S. Karaman, “Sparse depth sensing for resource-constrained robots,” International Journal of Robotics Research, vol. 38, no. 8, pp. 935–980, 2019.

[10] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, “Sparcity invariant cnns,” in International Conference on 3D Vision (3DV), 2017, pp. 11–20.

[11] M. Jaritz, R. De Charette, E. Wirbel, X. Perrotton, and F. Nashashibi, “Sparse and dense data with cnns: Depth completion and semantic segmentation,” in International Conference on 3D Vision (3DV), 2018, pp. 52–60.

[12] Y. Tsuji, H. Chishiro, and S. Kato, “Non-guided depth completion with adversarial networks,” in International Conference on Intelligent Transportation Systems (ITSC), 2018, pp. 1109–1114.

[13] D. Rotman and G. Gilboa, “A depth restoration occlusionless temporal dataset,” in International Conference on 3D Vision (3DV), 2016, pp. 176–184.

[14] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, “Learning to discover cross-domain relations with generative adversarial networks,” in International Conference on Machine Learning (ICML), 2017, pp. 1857–1865.

[15] Z. Yi, H. R. Zhang, P. Tan, and M. Gong, “Dualgan: Unsupervised dual learning for image-to-image translation,” in IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2868–2876.

[16] Y. Miron and Y. Coscas, “S-flow gan,” arXiv preprint arXiv:1905.08474, 2019.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in Neural Information Processing Systems (NIPS), 2014, pp. 2672–2680.

[18] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, “Image inpainting for irregular holes using partial convolutions,” in European Conference on Computer Vision (ECCV), 2018, pp. 85–100.

[19] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015, pp. 234–241.

[20] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in European Conference on Computer Vision (ECCV), 2016, pp. 483–499.

[21] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, “Instance normalization: The missing ingredient for fast stylization,” arXiv preprint arXiv:1607.08022, 2016.