IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 July 2024, accepted 1 August 2024, date of publication 15 August 2024, date of current version 11 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3444452

RESEARCH ARTICLE

Unpaired Depth Super-Resolution in the Wild

ALEKSANDR SAFIN[1], **MAXIM KAN**[1], **NIKITA DROBYSHEV**[1], **OLEG VOYNOV**[1,2], **ALEXEY ARTEMOV**[3], **ALEXANDER FILIPPOV**[4], **DENIS ZORIN**[5], **AND EVGENY BURNAEV**[1,2]

[1]Skolkovo Institute of Science and Technology, 121205 Moscow, Russia
[2]AIRI, 105064 Moscow, Russia
[3]Technical University of Munich, 80333 Munich, Germany
[4]AI Foundation and Algorithm Lab, 121205 Moscow, Russia
[5]New York University, New York City, NY 10003, USA

Corresponding author: Alexey Artemov (artonson@yandex.ru)

Evgeny Burnaev and Oleg Voynov were supported by the Analytical center under the RF Government (subsidy agreement 000000D730321P5Q0002, Grant No. 70-2021-00145 02.11.2021).

ABSTRACT Depth images captured with commodity sensors commonly suffer from low quality and resolution and require enhancing to be used in many applications. State-of-the-art data-driven methods for depth super-resolution rely on registered pairs of low- and high-resolution depth images of the same scenes. Acquisition of such real-world paired data requires specialized setups. On the other hand, generating low-resolution depth images from respective high-resolution versions by subsampling, adding noise and other artificial degradation methods, does not fully capture the characteristics of real-world depth data. As a consequence, supervised learning methods trained on such artificial paired data may not perform well on real-world low-resolution inputs. We propose an approach to depth super-resolution based on learning from *unpaired data*. We show that image-based unpaired techniques that have been proposed for depth super-resolution fail to perform effective hole-filling or reconstruct accurate surface normals in the output depth images. Aiming to improve upon these approaches, we propose an unpaired learning method for depth super-resolution based on a learnable degradation model and including a dedicated enhancement component which integrates surface quality measures to produce more accurate depth images. We propose a benchmark for unpaired depth super-resolution and demonstrate that our method outperforms existing unpaired methods and performs on par with paired ones. In particular, our method shows 28% improvement in terms of a perceptual $\mathrm{MSE}_{\mathrm{v}}$ quality measure, compared to state-of-the-art unpaired depth enhancement techniques adapted to perform super-resolution [e.g., Gu et al. (2020)]. The implementation of our method is publicly available at https://github.com/keqpan/udsr.

INDEX TERMS Depth data, enhancement, generative networks, super-resolution, unsupervised learning.

I. INTRODUCTION

Depth images are commonly used in a variety of applications, from 3D scene reconstruction to robotic navigation, user interfaces and photo effects. Depth sensors are becoming standard for everyday devices such as phones and tablets, immensely expanding availability of this type of data and the range of its applications.

However, when acquired with commodity depth cameras, raw depth images come with multiple limitations, most importantly, limited spatial resolution, severe noise levels, and many gaps. Reliably addressing these flaws has been attracting increasing research interest aimed at enhancing the resolution and quality of depth images.

Following image processing, where the most success has been achieved through deep learning [2], [3], [4], [5], [6], convolutional neural networks (CNNs) were applied to depth super-resolution and enhancement [7], [8], [9], [10], [11], [12], [13]; commonly, they are trained from paired datasets of low- and high-quality target depth maps. However, acquiring a large real dataset of this type is challenging and requires a customized, calibrated hardware setup; as a result, such methods commonly rely on downsampling of high-resolution data for constructing training instances; this approach is

The associate editor coordinating the review of this manuscript and approving it for publication was Yongjie Li.

ineffective for training super-resolution models targeting real depth images heavily contaminated by holes and noise, as we illustrate in Figure 1.

One way to circumvent this issue is through the use of *unpaired* learning methods [14], [15], [16], where the model is trained on two datasets, a dataset of inputs and a dataset of targets, with their elements not necessarily forming pairs. While, in principle, one can directly apply existing image-based unpaired methods to depth data, these methods tend to significantly underperform in practice as they fail to capture the distinctive characteristics of depth data: unlike color photos, depth scans often contain gaps; compared to RGB images where pixels take values from a large but finite palette, the range of depth values is, in principle, continuous and unbounded; additionally, perceptually-relevant differences in depth scans are best captured by depth-specific measures.

We propose, to the best of our knowledge, the first learning-based method for depth super-resolution for real-world sensor data, using *unpaired* data for training, i.e., a set of raw sensor depth images and a set of high-quality, high-resolution depth images generated by depth fusion, without correspondences between images from these sets.

A key ingredient of our method is the introduction of depth enhancement, *i.e.*, a hole-filling and surface denoising method, into the super-resolution pipeline; we demonstrate that coupling depth enhancement and super-resolution tasks yields significant improvements over the baselines. To implement this, similarly to recent literature (*e.g.*, [14]), we design a two-stage approach for training our method, including (1) an unpaired training stage for a depth degradation model, and (2) a supervised training stage for an enhancement model.

Importantly for evaluating depth super-resolution methods, most existing RGB-D scan datasets cannot be used as they offer either high- or low-resolution sensor depth only. To this end, we propose a paired dataset providing both real-world RGB-D scans and high-resolution, high-quality reference depth, that we construct by ray-tracing 3D reconstructions of indoor scenes in ScanNet [17] obtained by depth fusion [18]. Basing on these, we develop a depth super-resolution and enhancement benchmark, extending a standard evaluation methodology with perceptual measures.

Our evaluation shows that our method outperforms several state-of-the-art image-to-image translation approaches applied to depth in a pure enhancement mode. Likewise our approach outperforms straightforward combinations of deep unpaired enhancement (e.g., [1]) and bicubic upsampling, emphasizing the need for a close integration of enhancement and super-resolution parts.

To summarize, our contributions are as follows:

- We introduce *UDSR*, the first dedicated method for learning-based, unpaired depth super-resolution. In comparison to state-of-the-art unpaired depth enhancement techniques adapted for super-resolution (*e.g.*, [1]), our method demonstrates an impressive 28% improvement in *depth super-resolution performance*, as measured by a perceptual MSE_v quality measure.
- A key component of our approach is a novel unpaired depth enhancement algorithm that efficiently performs unpaired learning, incorporates RGB guidance, and optimizes depth-specific performance measures. As a result, we achieve superior denoising and inpainting results for depth images. Compared to the state-of-the-art unpaired depth enhancement method by Gu et al. [1], our algorithm showcases an impressive 63% gain in *depth enhancement performance* when evaluated using the perceptual MSE_v measure.
- We also introduce a new benchmark for real-world depth super-resolution and enhancement, utilizing existing datasets such as [17] and [19]. Additionally, we propose a robust methodology for comparing paired and unpaired approaches.

II. RELATED WORK

A. DEPTH SUPER-RESOLUTION (SR)

Depth super-resolution has been approached from multiple perspectives: filter-based [20], [21], optimization-based [22] and [23], and data-driven [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [24], [25], [26], [27], [28], [29], [30].

In non-learning context, single-frame RGB-guided depth SR has been tackled with joint bilateral filters [21] and filters with adaptive smoothing [20]; optimization-based shape-from-shading approaches [22], [23] relying on photometric constraints and a number of priors. While such techniques can likely generalize across sensors, their ability to exploit RGB and depth image-specific characteristics such as distribution of depth values is limited.

Among data-driven approaches, CNNs have been used in combination with optimization-based methods [7], [28], [29], joint filtering methods [24], [30], as well as with progressive or hierarchical multi-scale fusion of RGB and depth features [8], [10], [27], [31]; we include an explicit RGB guidance mechanism in our enhancement step, but without applying any optimization to network predictions. Recent trends also include applying to depth SR attention-based and image transformer architectures [11], [32]. More recently, perceptually-based depth SR [9] enabled more accurate surface reconstruction; we integrate their loss function in our training framework. Reference [11] introduced non-linear downsampling degradations to improve robustness of their depth SR method; in contrast, our method automatically captures relevant degradation patterns by a learned depth-to-depth translation step. SRFBN [4] is an established supervised image SR method often used as a strong baseline for evaluating depth SR; we compare against this approach in our work. The most recent and concurrent work [33] is the first which considers depth super-resolution on real sensor data. It is trained on their own collected paired dataset of low- and high-resolution depth maps. However, they rely on image colorization [34] to inpaint holes in input low-resolution

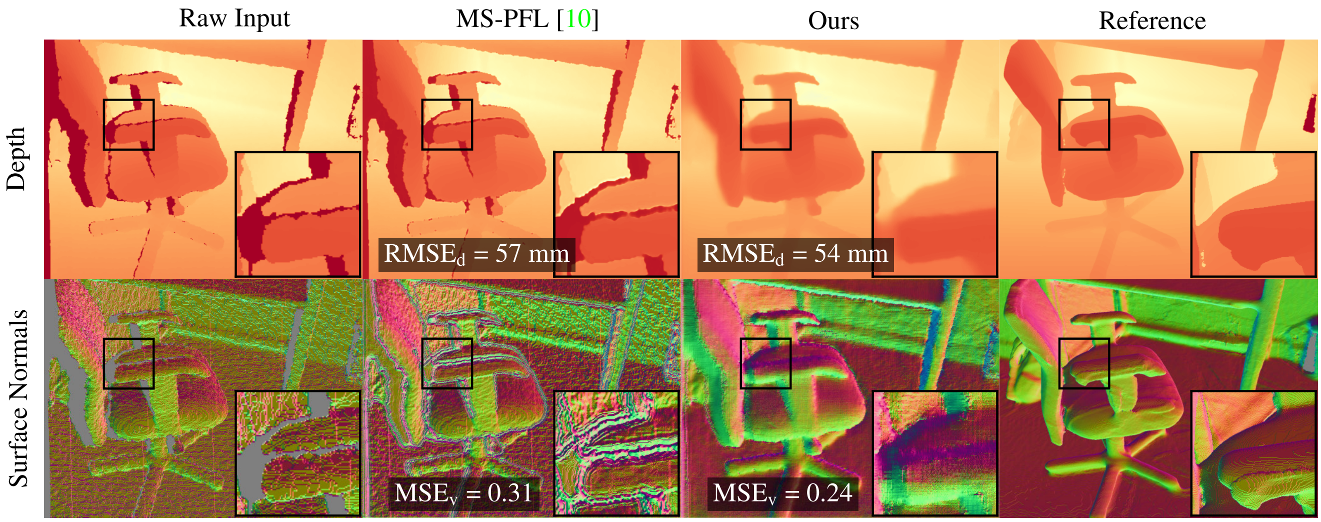


FIGURE 1. State-of-the-art depth super-resolution methods are designed for clean and complete images, but produce noisy, incomplete results in the wild. In contrast, our novel *unpaired* super-resolution method inpaints holes and produces normals closer (MSE$_v$ [9]) to the reference data. RMSE$_d$ is the depth error averaged over the area of valid (non-hole) pixels. We note that MS-PFL [10] is designed to perform depth SR, but not hole-filling.

depth maps as pre-processing stage. This could be inefficient with big holes and areas of drastic change in depth. Moreover, they do not use depth normals neither for evaluation nor in their approach, which can lead to noisy surfaces reconstructed from their output depth.

Importantly, all these methods need registered pairs of input-output images of the same scene; differently, in this work we explore an *unpaired learning scenario* where the sets of source and target depth images may depict distinct environments.

B. DEPTH ENHANCEMENT

Depth Enhancement [1], [12] is an umbrella task that encompasses denoising [35], [36], [37], [38], completion [39], [40], and inpainting [41], [42], [43], [44], [45]. Among the methods in this group, our approach implements a data generation technique similar to LapDEN [12] who render 3D reconstructions of scenes in ScanNet [17] to obtain training data for their depth enhancement method; we also compare our method to this approach.

C. UNPAIRED IMAGE SR AND ENHANCEMENT

More recently, image-based approaches have resorted to using unpaired learning methods to more accurately model image acquisition and processing artefacts; these formulations bypass the need for paired data, greatly simplifying construction of training datasets. Some approaches use independent sets of low-resolution and high-resolution images to learn the SR mapping without correspondences between the images, commonly employing cycle consistency [46]. [47] trains a super-resolution model in a cycle-consistent manner on a set of LR images acquired with one device and pairs of LR-HR images acquired with a different device. [48] proposes to embed the two domains of the low- and high-resolution images into a shared latent space and find the translation between these domains using the shortest path assumption regularization. Other approaches rely on self-supervised training from low-resolution images only [49], [50], [51], [52], [53]. Reference [51] trains a denoising model for real-world images in a self-supervised manner by breaking spatial correlations of the sensor noise. Reference [52] employs bursts of noisy images as training data for the same task. Reference [53] trains a denoising model that decomposes the image into a noise free and a noisy components with a technique similar to the cycle consistency, by combining the predicted components using a pre-defined noisy image formation model and feeding this image back to the denoising model. Most unpaired systems are trained in multiple stages: Cycle-in-Cycle [16] learns image cleaning during the first cycle and SR in the unpaired setting using the second cycle, Bulat et al. [14] learns degradation using an unpaired setup, further performing supervised training for SR. Maeda [15] decomposes SR mapping into a cleaning step trained in an unpaired way and a pseudo-supervised SR network; similarly, our method uses two-stage training.

D. UNPAIRED DEPTH ENHANCEMENT

Unpaired Depth Enhancement is similar in spirit to unpaired image-to-image translation but requires considerable adaptation of existing image-based methods. Gu et al. pioneered a GAN-based unpaired depth enhancement [1] with a four-stage learnable approach, involving hole prediction, image adaptation, degradation, and final enhancement. Compared to Gu et al. which focuses on depth enhancement (specifically hole filling) but does not address super-resolution (beyond trivial bicubic scaling) or denoising, our algorithm integrates super-resolution and enhancement. From a self-supervised perspective, [13] leverages photometric constraints to recover high-quality depth but requires a non-standard acquisition setup. Learning from unsynchronized low- and high-quality depth frames, [54] proposes a self-supervised approach employing temporal and spatial alignment. Reference [55] proposes an approach for super-resolution of dToF depth videos with the guidance of high-resolution RGB frames. Among these methods, we extensively compare to single-image unpaired enhancement [1], adapting this method for depth SR via complementing it with various upsampling methods.

E. RGB-D DATASETS FOR DEPTH SR

Among RGB-D datasets, Middlebury [56], NYU-Depth V2 [57], SUN RGB-D [58], and the synthetic ICL-NUIM [59] provide RGB-D frames but cannot serve as evaluation data for depth super-resolution since they either do not provide real-world sensor depth or lack corresponding ground truth. Matterport3D [60] is a large-scale dataset with high-quality depth but lacks corresponding depth from less accurate sensors. ToF-Mark [61] contains depth maps from a low-accuracy time-of-flight sensor and a high-accuracy structured light scanner but only provides three pairs of high- and low-resolution images, making it suitable for qualitative evaluation only. Similarly, Redwood [62] consists of RGB-D sequences obtained using a consumer Asus Xtion Live depth camera and point clouds from industrial-grade laser scanner but captures only five scenes. The largest to date collection [63] is at the time of writing the manuscript not available.

In the context of depth enhancement, [12] synthesized a paired dataset from ScanNet [17], a large-scale collection of RGB-D scans, using its complete 3D reconstructed models obtained using BundleFusion [18]. These models were ray-casted to obtain image pairs of the same resolution for training neural networks targeting denoising and hole-filling. We extend this approach to our task, additionally creating high-resolution depth images from renderings of 3D reconstructions (Section IV). We additionally conduct experiments using synthetic indoor data in InteriorNet [19].

III. UNPAIRED DEPTH SUPER-RESOLUTION FRAMEWORK

A. UNPAIRED DEPTH SUPER-RESOLUTION

1) TASK FORMULATION

RGB-D sensors capture the surface of the scene using pairs (I, D^L) of sensor data, where I is a color image and D^L a depth image, acquired jointly. Following the existing practice in depth SR literature, we assume color images to come at high resolution and satisfactory quality.

A common approach to depth SR is to fit a mapping from sensed instances (I, D^L) to desired depth D^H using a neural network [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [24]. We refer to this learning-based formulation as *supervised (paired)* depth SR. As an alternative formulation, in this work we consider *unpaired* depth SR, a previously unexplored task that does not rely on having paired image data during training; instead, our formulation assumes that source and target image datasets are entirely independent. In practice, this means that sensor data pairs (I^L, D^L) and target instances (I^H, D^H) can capture non-overlapping segments of the scene or even separate scenes; they may be collected using differing, non-registered sensors, or at different periods of time (*e.g.*, allowing the scenes to undergo changes). Such data is significantly easier to collect, as no alignment or additional hardware are necessary, and facilitates re-use of existing datasets; as a consequence, our formulation enables a broader scope of applications.

2) FRAMEWORK OVERVIEW

The input to our algorithm is a single RGB-D image with a low-resolution, noisy, incomplete depth; as an output, we produce a single high-resolution, denoised, complete depth image. To achieve our objective, we designed a learning-based framework to learn depth SR without paired, registered training instances; we made a series of algorithmic decisions to address the challenges of our task. The four main components of our method are:

1) *Framework Architecture* (Section III-B). We follow a two-step approach to depth super-resolution: first, we upsample the input depth image to the desired output resolution; second, we process the upsampled depth image using our enhancement algorithm to produce the final result. In our system, only the enhancement part is trained by solving a two-stage unpaired learning task.
2) *Enhancement Algorithm* (Section III-C). We pre-train four neural sub-networks using complementary sub-tasks to provide rich photometric and geometric features, and integrate these models within our learning-based enhancement algorithm u_{enh}. Our best performing learning configuration predicts an enhanced depth image from the upsampled input RGB-D image, its feature maps, and an intermediate depth estimate.
3) *Unpaired Translation Algorithm* (Section III-D). We construct supervised data for training our enhancement algorithm by synthesizing a realistic *pseudo-source* RGB-D instance for each *target* (high-resolution) RGB-D image. To this end, we pre-train a deep translation network g_{H2L} using an unpaired, cycle-consistent adversarial learning approach, minimizing discrepancy between sets of source and target instances.
4) *Construction of Datasets* (Section IV). We develop a methodology to directly evaluate paired and unpaired depth SR algorithms and construct three benchmarks of up to 38,000 instances for our quantitative comparisons.

In the next sections, we describe each component in detail and provide a rationale for our algorithmic choices.

B. TASK DECOMPOSITION AND FRAMEWORK ARCHITECTURE

Our depth SR algorithm is designed to increase the spatial resolution (*i.e.*, do upsampling) and to suppress noise, fill in

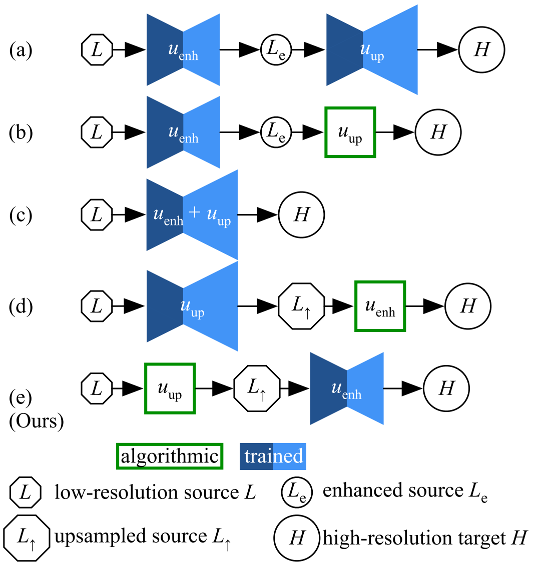


FIGURE 2. Architectural options we considered (see also Table 1 for notation). Our architectural choice (e) enjoys both a simple design and high performance in experiments. ***Enhanced source*** L_e refers to a set of RGB-D images with low-resolution but enhanced depth (used only in this scheme).

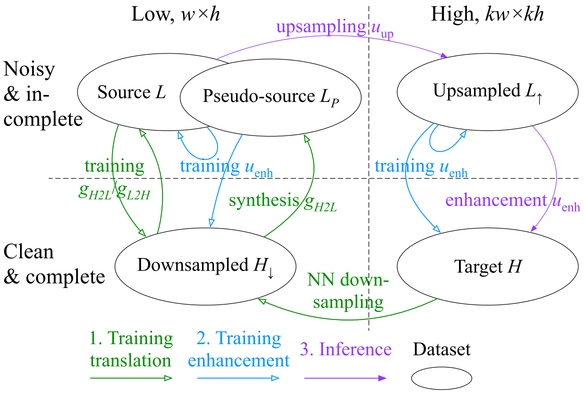


FIGURE 3. Overview of the conceptual architecture and principal stages of our learning framework. (1) We start with the training of the unpaired translation algorithm and obtain bidirectional mappings g_{H2L}, g_{L2H} between the source L and the downsampled $H_\downarrow$ sets. In the same stage, we synthesize the pseudo-source L_P set by translating instances in $H_\downarrow$. (2) Next, we train the enhancement algorithm u_{enh} at low and high resolution using a multi-task objective. (3) During inference for a source instance from L, we perform upsampling by u_{up} and enhancement by u_{enh} to compute the final prediction.

TABLE 1. Overview of the qualitatively different sets of RGB-D images (I, D) used within the present work. "Low-quality" depth images are prone to noise and incomplete depth values; "high-quality" depth images are clean and complete.

Symbol	Depth Properties		Set Meaning
	Resolution	Quality	
L	$w \times h$	low	Source RGB-D images
$L_\uparrow$	$kw \times kh$	low	L with upsampled depth
H	$kw \times kh$	high	Target RGB-D images
$H_\downarrow$	$w \times h$	high	H with downsampled depth
L_P	$w \times h$	≈low	Pseudo-source RGB-D images

missing areas, and resolve detail (*i.e.*, do enhancement) in an input depth image. We have considered a number of alternatives for implementing these diverse modifications in the unpaired learning context. We summarize our resulting configuration and compare it to alternatives below, provide more detail in Sections III-C and III-D, and discuss the effect of various choices in Section V.

1) TASK DECOMPOSITION

We decompose the depth SR mapping into a sequence $u_{\text{sr}} = u_{\text{enh}} \circ u_{\text{up}}$ where a *trainable* enhancement algorithm u_{enh} succeeds (an optional) *non-trainable* upsampling operation u_{up} (see Figure 2 (e)); separating these stages is in line with recent successful approaches to unpaired image SR [14]. Compared to complementing a *trainable* enhancement algorithm with a separate *trainable* upsampling network (*c.f.* [14], Figure 2 (a)), our approach requires training only a single model; adding a *non-trainable* upsampling operation (Figure 2 (b)) yields results inferior to our algorithm, according to our experiments. In comparison to integrating the enhancement and upsampling stages inside a *single trained* network (Figure 2 (c)), or complementing a *trained* upsampling network with a *non-trained* enhancement operation (Figure 2 (d)), our approach enjoys greater simplicity by avoiding the need for in-network upsampling [15]. Our upsampling operation u_{up} is bicubic interpolation: we upscale the input $w \times h$ depth image to the output $kw \times kh$ resolution (k being the SR factor). The enhancement algorithm u_{enh} takes in an RGB-D image with the upsampled depth and produces a refined depth image at the same spatial resolution. Making upsampling optional in this way enables using our algorithm at different SR factors, including keeping the input resolution of the image ($k = 1$).

2) ARCHITECTURE OF THE LEARNING FRAMEWORK

Conceptually, we reframe our unpaired learning task as two arguably easier tasks: a data synthesis task, concerned with *generating appropriate training (pseudo-source) images,* and a supervised regression learning task, which addresses *learning an enhancement mapping from the pseudo-sources* and bears similarity to recent successful approaches to image SR [15]. We elaborate on the design of our framework below; we visually accompany this design by Figure 3 and present an overview for the symbols denoting the different sets of RGB-D data used throughout this work in Table 1.

The input domain for our depth SR method is the set of source RGB-D images with low-resolution depth *(source L)*, that we upsample using u_{up} to obtain RGB-D images with upsampled depth *(upsampled $L_\uparrow$)*. Our central learning goal is to construct an enhancement algorithm u_{enh} that transforms the upsampled $L_\uparrow$ to the set of target RGB-D images with high-resolution depth *(target H)*. In this context, our first objective is to construct a collection of training RGB-D instances with source-like and target-like depth, suitable for supervised training of the enhancement algorithm. To serve as such source-like and target-like images, we define two sets with $w \times h$ depth: pseudo-source RGB-D images *(pseudo-source L_P)* and target RGB-D images with downsampled depth *(downsampled $H_\downarrow$)*. The downsampled $H_\downarrow$ is obtained by nearest-neighbour downsampling of the depth images in the target set H. Pseudo-sources L_P are constructed by training an unpaired image-to-image translation method minimizing discrepancy between the distributions of the pseudo-sources L_P and the sources L. Our second objective is to use the paired data in the pseudo-source L_P and downsampled $H_\downarrow$ sets to learn an effective enhancement method. During training, we strive to maintain equally high enhancement performance for the pseudo-source L_P, source L, and upsampled $L_\uparrow$ sets despite differences in their distributions and resolution. Note that while the source L and the upsampled $L_\uparrow$ sets are related by the upsampling operation u_{up}, including instances from both sets for training the pseudo-supervised model remains crucial for upscaling factors $k > 1$ to explicitly account for distribution shift stemming from increase in resolution.

Following this design, we decompose our learning framework into two interrelated trainable parts: the unpaired translation algorithm, implementing the bidirectional translation between the downsampled $H_\downarrow$ and the source L sets, and the enhancement algorithm, implementing the enhancement transformation u_{enh}. We train them consecutively using two stages. In the first stage, we train the translation algorithm using the source L and the downsampled $H_\downarrow$ in an unpaired manner, obtaining a deep translation network $g_{H2L} : H_\downarrow \to L$ (Section III-D), and freeze the weights of g_{H2L}. In the second stage, we use g_{H2L} to process samples in the downscaled $H_\downarrow$ and obtain the set of pseudo-sources L_P. We train the enhancement algorithm using the synthesized pseudo-sources L_P, sources L, and downsampled targets $H_\downarrow$ (Section III-C), and fine-tune it using the upsampled sources $L_\uparrow$ and the targets H. Note that while both trainable parts can technically be trained jointly in a shared training loop, in practice this would impose upon the system a heavy memory footprint, reducing batch sizes and learning rates; we thus opted on successive training. We give details of these components below.

C. ENHANCEMENT ALGORITHM

Our enhancement algorithm u_{enh} is designed to leverage multiple complementary sensed and predicted images as presented in a data-flow Figure 4. More specifically, we include raw input color and depth images, estimate an accurate intermediate monocular depth image from the input color image, and construct two additional photometric and geometric feature maps extracted from the input color and depth images, respectively. During training, we (1) pre-train a separate RGB guidance network (feature extractor) using RGB-D data with low-resolution depth: source L and downsampled $H_\downarrow$; (2) train an enhancement network using a mix of source L, pseudo-source L_P and target H sets in a shared training loop. For inference with an input RGB-D image, we (1) upsample its input depth to the target resolution ($kw \times kh$), (2) compute feature maps of the input RGB and depth images using convolutional feature extractors, (3) estimate an intermediate depth image from the RGB image with the RGB guidance network, and (4) estimate the final depth image from the input RGB-D image concatenated with their feature maps and the intermediate depth image, using the enhancement network.

We describe the supervised training procedure for our enhancement algorithm below; we provide details on the construction of its training data in Section III-D. We assume that we have access to RGB-D images from the source L, pseudo-source L_P, target H, and downsampled $H_\downarrow$ sets. We obtain depth images in the pseudo-source L_P by applying the translation network (Section III-D) to those in the target $H_\downarrow$; depth images in the downsampled $H_\downarrow$ are obtained by nearest-neighbour downsampling of the depth from the target H. The distributions of the source L and pseudo-source L_P sets are assumed to be close to identical.

1) RGB GUIDANCE NETWORK

Color guidance has been established as an important visual cue heavily utilised for RGBD-based depth SR [8] due to the strong correlation between color and depth images. Following this intuition, we exploit a simple color guidance mechanism, that we demonstrate to serve as an effective signal, particularly for recovering accurate depth in regions where sensor depth measurements are missing or unreliable. To achieve this, we define an RGB guidance CNN f_{rgb} used to estimate a depth image from the respective monocular color image. We implement f_{rgb} using a U-Net-like [64] encoder-decoder architecture with two encoders, one for the source L and one for the downsampled $H_\downarrow$ sets, and a shared decoder. We pre-train f_{rgb} by minimizing mean absolute error between the monocular depth estimate $f_{\text{rgb}}(I)$ and the reference depth image D for a given RGB-D image $(I, D) \in L$ or $(I, D) \in H_\downarrow$, and freeze its weights after pre-training. We describe the architecture used for f_{rgb} in the experimental Section V-A.

2) FEATURE EXTRACTORS

We separately mention photometric and geometric feature extractors, represented by two smaller CNNs trained jointly with the RGB guidance CNN and the enhancement CNN, respectively. Their outputs, photometric x_I and geometric x_D convolutional features, are extracted from the raw input RGB and depth images, respectively, and used as additional inputs to our enhancement algorithm (see Figure 4).

3) ENHANCEMENT NETWORK F_{enh}

We define an enhancement CNN f_{enh} as our estimator of the final high-resolution depth image $\widehat{D}^H$ using the diverse available visual data. For an input RGB-D image (I, D), we compute a depth estimate $f_{\text{rgb}}(I)$ using the pre-trained

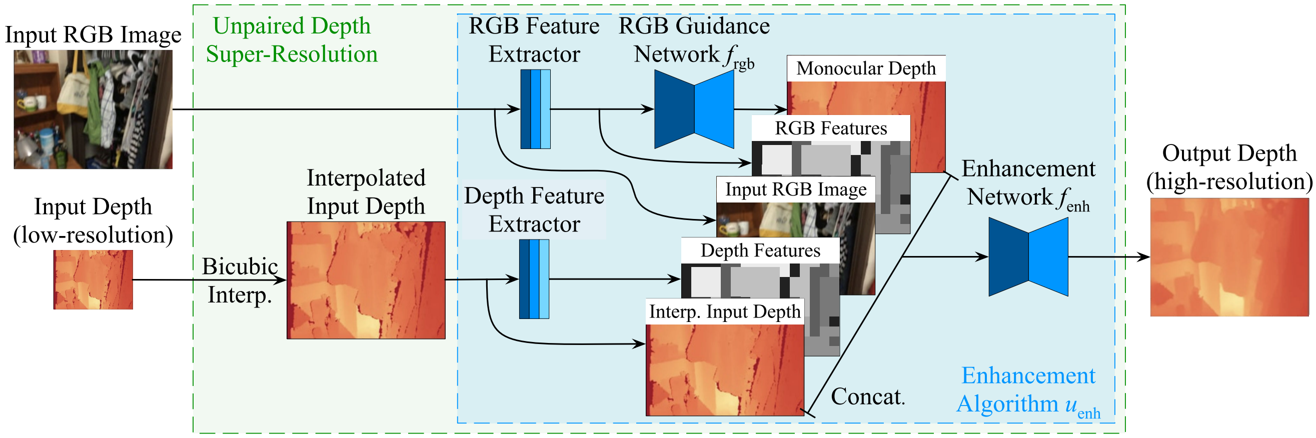

Input RGB Image
Unpaired Depth Super-Resolution
RGB Feature Extractor
RGB Guidance Network f_{rgb}
Monocular Depth
RGB Features
Input RGB Image
Input Depth (low-resolution)
Bicubic Interp.
Interpolated Input Depth
Depth Feature Extractor
Depth Features
Interp. Input Depth
Enhancement Network f_{enh}
Output Depth (high-resolution)
Concat.
Enhancement Algorithm u_{enh}

FIGURE 4. Scheme of depth super-resolution with our enhancement algorithm u_{enh}. We upsample depth via Bicubic interpolation to the target resolution of the color image I, obtaining $u_{\text{up}}(D)$. We use RGB guidance network f_{rgb} to extract a monocular depth estimate $f_{\text{rgb}}(I)$ and run feature extractors to obtain photometric and geometric convolutional features (x_I, x_D). Finally, the enhancement network f_{enh} predicts a refined depth image by processing the concatenated data $(u_{\text{up}}(D), x_D, I, x_I, f_{\text{rgb}}(I))$. We provide results of ablative studies assessing contributions of individual components in Table 7 and give information on parameters of their training configuration in Table 1 in Supplementary.

image guidance network f_{rgb} and extract photometric and geometric convolutional features x_I, x_D, respectively. The 5-tuple $(u_{\text{up}}(D), x_D, I, x_I, f_{\text{rgb}}(I))$ is fed in as an input data for predicting the output $\widehat{D}^H$ using f_{enh} (see Figure 4). The network architecture used for f_{enh} is specified in Section V-A.

4) LOSS TERMS FOR TRAINING F_{enh}

In our training scenario, the network is unlikely to automatically pick up relevant patterns in absence of "true" large-scale paired data; we thus seek to give our regression model the desired properties by optimizing it using a combination of loss terms, that we find to directly impact performance for our algorithm. We describe these properties and the respective loss terms here.

We define pixel-weighted mean absolute error (MAE) and mean squared error (MSE) by computing integral of the pixelwise deviation in the predicted and reference depth images

$$L_p(D_1, D_2) = \left\| w_p \odot (D_1 - D_2) \right\|_p, \qquad \text{MAE} = L_1, \text{MSE} = L_2, \tag{1}$$

where $\odot$ denotes pixel-wise multiplication, and use a combination of these with our *depth-based loss term*

$$\begin{aligned}\mathcal{L}_{\text{depth}} &= \lambda_{\text{depth},1}\text{MAE} + \lambda_{\text{depth},2}\text{MSE} \\ &= \left[\lambda_{\text{depth},1}\ \lambda_{\text{depth},2}\right]\left[\text{MAE}\ \text{MSE}\right]^{\mathsf{T}} = \boldsymbol{\lambda}_{\text{depth}}\boldsymbol{L}_{\text{depth}}^{\mathsf{T}}\end{aligned} \tag{2}$$

where $\boldsymbol{\lambda}_{\text{depth}}$ weights p-norms of pixelwise deviations.

While the two depth images may be close in the Euclidean sense, appearance of their respective surfaces (as captured, *e.g.*, by a rendering) can vary significantly due to geometric noise in local surface orientation. Thus, we assess surface quality using perceptual losses [9] defined as p-norms of pixelwise-weighted difference in surface renderings of depth images averaged over three orthogonal light directions e_i:

$$R_p(D_1, D_2) = \frac{1}{3}\sum_{i=1}^{3}\left\| w_p \odot (N_1 - N_2)\cdot e_i \right\|_p, \qquad \text{MAE}_{\text{v}} = R_1, \quad \text{MSE}_{\text{v}} = R_2, \tag{3}$$

where N_1 and N_2 are finite-difference estimates of per-pixel normals from the depth images D_1 and D_2, respectively (computed as in [9]). We define our *surface-based loss term* via

$$\begin{aligned}\mathcal{L}_{\text{surf}} &= \lambda_1\text{MAE}_{\text{v}} + \lambda_2\text{MSE}_{\text{v}} \\ &= \left[\lambda_{\text{surf},1}\ \lambda_{\text{surf},2}\right]\left[\text{MAE}_{\text{v}}\ \text{MSE}_{\text{v}}\right]^{\mathsf{T}} = \boldsymbol{\lambda}_{\text{surf}}\boldsymbol{L}_{\text{surf}}^{\mathsf{T}}\end{aligned} \tag{4}$$

We specify our pixelwise weighting functions $w(u)$, $u = (i, j)$ for use within the depth- and surface-based loss terms Equations (1) and (4) according to the general expression

$$\begin{aligned}w(u) &= w_1\chi_{A_1}(u) + w_2\chi_{A_2}(u) \\ &= \left[w_1\ w_2\right]\left[\chi_{A_1}(u)\ \chi_{A_2}(u)\right]^{\mathsf{T}} \\ &= \boldsymbol{w}\boldsymbol{\chi}(u),\end{aligned} \tag{5}$$

where $w_1, w_2 \in \mathbb{R}$, the function $\chi_{A_1}(\cdot)$ is an indicator function of the set A_1 of *valid* pixels, and $\chi_{A_2}(\cdot)$ is an indicator function of the set A_2 of pixels corresponding to *gaps* (indicated as *NaN*s in input depth readings). As we would like to plausibly inpaint gaps, we typically set w_2 to a larger value compared to w_1. We specify particular values for per-pixel weights w_p in Section V.

To encourage sharp depth discontinuities at object boundaries, we use the *edge-based regularizer term* [65]

$$\mathcal{R}_{\text{edge}}(I, D) = \left\|\nabla_h D\right\|_1 e^{-\|\nabla_h I\|_1} + \left\|\nabla_v D\right\|_1 e^{-\|\nabla_v I\|_1}, \tag{6}$$

where ∇_h and ∇_v denote finite differences computed in horizontal and vertical direction.

To aid noise suppression and impose spatial smoothness on the normals N for an RGB-D image (I, D), we additionally optimize the *total variation regularizer term* [16], [66] expressed by

$$\mathcal{R}_{\text{smooth}}(N) = \|\nabla_h N\|_2 + \|\nabla_v N\|_2 . \tag{7}$$

5) TRAINING PROCEDURE FOR F_{enh}

Our core idea is to implement a shared enhancement mapping for both source real-world L and pseudo-source synthesized L_P sets using a single model, enabling it to treat these sets in a unified way; this leads us to include two critical ingredients in designing our procedure for training the enhancement network f_{enh}. First, to effectively perform enhancement we require the network to minimize a *pseudo-supervised objective,* aiming to reconstruct output depth in $H_\downarrow$ from respective synthesized images in L_P. For pseudo-source samples in L_P, as accurate reference depth and normals are available, we minimize

$$\begin{aligned}\mathcal{L}^{L_P}_{\text{enh}}(f_{\text{enh}}) = \mathcal{L}_{\text{depth}}(D^{L_P}, \widehat{D}^{H_\downarrow}) + \mathcal{L}_{\text{surf}}(D^{L_P}, \widehat{D}^{H_\downarrow})\\ + \lambda^{L_P}_{\text{smooth}}\mathcal{R}_{\text{smooth}}(\widehat{D}^{H_\downarrow}),\end{aligned} \tag{8}$$

where $\widehat{D}^{H_\downarrow} = f_{\text{enh}}(I, D^{L_P})$ is the refined depth image produced by the enhancement CNN. However, this alone does not guarantee achieving similar performance for real-world images in L due to a discrepancy between L and L_P. We thus additionally minimize a *self-supervised objective* over instances from L which serves as a fixed point constraint. For these instances, normals tend to be very noisy, and we thus exclude the surface-based term $\mathcal{L}_{\text{surf}}$ and instead focus on detecting object contours using $\mathcal{R}_{\text{edge}}$ by minimizing

$$\begin{aligned}\mathcal{L}^{L}_{\text{enh}}(f_{\text{enh}}) = \mathcal{L}_{\text{depth}}(D^{L}, \widehat{D}^{L}) + \lambda^{L}_{\text{edge}}\mathcal{R}_{\text{edge}}(\widehat{D}^{L})\\ + \lambda^{L}_{\text{smooth}}\mathcal{R}_{\text{smooth}}(\widehat{D}^{L}),\end{aligned} \tag{9}$$

where we expect the output depth image $\widehat{D}^L = f_{\text{enh}}(I, D^L)$ to replicate the input depth D^L.

Note that each of the terms $\mathcal{L}_{\text{depth}}(D^{L_P}, \widehat{D}^{H_\downarrow})$, $\mathcal{L}_{\text{surf}}(D^{L_P}, \widehat{D}^{H_\downarrow})$, and $\mathcal{L}_{\text{depth}}(D^L, \widehat{D}^L)$ incorporate 1-norm and 2-norms weighted by respective weights $\boldsymbol{\lambda}^{L_P}_{\text{depth}}$, $\boldsymbol{\lambda}^{L_P}_{\text{surf}}$, and $\boldsymbol{\lambda}^{L}_{\text{depth}}$.

Our full enhancement objective is

$$\mathcal{L}_{\text{enh}} = \mathcal{L}^{L_P}_{\text{enh}} + \mathcal{L}^{L}_{\text{enh}} \tag{10}$$

where $\mathcal{L}^{L_P}_{\text{enh}}$ leverages the available pseudo-supervision and $\mathcal{L}^{L}_{\text{enh}}$ helps to achieve good performance for real instances. Technically, we include an equal number of samples from both L_P and L in each mini-batch while performing gradient descent.

6) TRAINING AT HIGH RESOLUTION

We pre-train the enhancement network at $w \times h$ resolution using instances in L_P, $H_\downarrow$, and L; after that, we fine-tune it using $kw \times kh$ high-resolution images. We keep the same objective in Equations (8) to (10) but set new hyperparameters giving heavier weight to depth and surface terms.

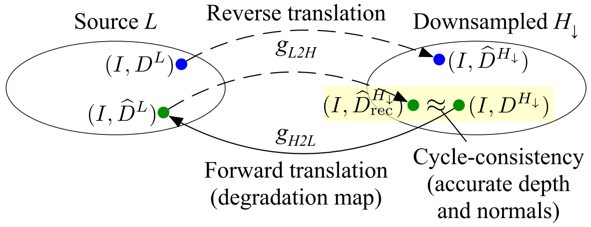


FIGURE 5. We use the source *L* and the downsampled $H_\downarrow$ as the two sets of RBG-D images involved in unpaired training of the translation network.

Section V describes specific values for weighting, details of our parameter choices, and results of an ablation study involving alternatives.

D. UNPAIRED TRANSLATION ALGORITHM

We design our translation step to perform data-driven construction of pseudo-sources L_P required for training the enhancement algorithm u_{enh} from the downsampled targets $H_\downarrow$. To this end, we take inspiration from the cycle-consistent learning paradigm [46] and its recent applications to photo and depth SR [1], [14], [15], [16], and construct two deep models for performing forward and reverse translation between the downsampled $H_\downarrow$ and the source L. Note that performing this training under the cycle-consistent adversarial framework does not require having paired training instances as such models minimize objectives formulated in terms of distributions. Once the forward translation network has been trained, we may proceed with generation of pseudo-source instances L_P via passing through it images from the downsampled $H_\downarrow$.

1) TRANSLATION NETWORKS

We define a forward translation network g_{H2L} to learn a desired *degradation mapping* from the high-quality downsampled $H_\downarrow$ to the low-quality source L, and a concurrent reverse translation network g_{L2H} which learns to translate the source L to the downsampled $H_\downarrow$. Depth images in both sets have the same $w \times h$ spatial resolution. Operation of these networks can be described in terms of relations

$$\widehat{D}^L = g_{H2L}(I, D^{H_\downarrow}) \quad \text{and} \quad \widehat{D}^{H_\downarrow} = g_{L2H}(I, D^L), \tag{11}$$

where $\widehat{D}^L$ is the translated pseudo-source depth for a downsampled $(I, D^{H_\downarrow}) \in H_\downarrow$, and $\widehat{D}^{H_\downarrow}$ is the translated downsampled depth of a source RGB-D image $(I, D^L) \in L$. We additionally define a reconstructed depth image $\widehat{D}^{H_\downarrow}_{\text{rec}}$ via

$$\widehat{D}^{H_\downarrow}_{\text{rec}} = g_{L2H}(I, g_{H2L}(I, D^{H_\downarrow})). \tag{12}$$

2) OBJECTIVE FOR UNPAIRED TRAINING

We train the forward and reverse translation networks using several distinct loss terms. We formulate an *adversarial loss term* by adapting an existing image-to-image translation objective developed in the context of image SR [15]. While a straightforward adaptation already allows to optimize translation networks (generators) g_{L2H} and g_{H2L} along with discriminators d^H_{depth} and d^L_{depth}, to further raise performance for RGB-D data we develop an improved formulation by integrating three modifications (we investigate their effect in Section V). First, in a similar vein to [15], we substitute the original $D^{H\downarrow}$ with a reconstructed $\widehat{D}^{H\downarrow}_{\text{rec}}$ for the discriminator d^H_{depth} operating in the downsampled $H_\downarrow$: we aim to relax the requirements for the generator and to stabilize training. Second, to facilitate reconstruction of faithful normals, we extend the adversarial part with surface normals discriminator networks d^H_{norm} and d^L_{norm} operating on images in the downsampled $H_\downarrow$ and the source L, respectively (we compute normals using finite differences). Finally, we formulate all adversarial losses in terms of the Least Squares GAN [67] that is known to improve GAN performance by preventing gradient saturation via a least squares penalty. We provide the full formulation of our adversarial loss term $\mathcal{L}_{\text{adv}}$ in the Supplementary material. Figure 6 illustrates the data flow used in training of our translation network.

By the cycle-consistency property, we expect the reconstructed depth image $\widehat{D}^{H\downarrow}_{\text{rec}}$ from Equation (12) to approximate the original image well: $\widehat{D}^{H\downarrow}_{\text{rec}} \approx D^{H\downarrow}$. To implement cycle-consistent training, we enforce the accurate reconstruction of the original depth image when undergoing a composition of forward and reverse mapping by minimizing the *cycle-consistency term* (see Figure 5)

$$\mathcal{L}_{\text{cycle}} = \text{MAE} + \text{MSE}_{\text{v}}, \tag{13}$$

emphasizing the need to fit both depth and normals components accurately (MAE and MSE$_{\text{v}}$ are defined in Equation (1) and Equation (2), respectively). As the downsampled $H_\downarrow$ is the only set that provides accurate reference depth and normals, we compute $\mathcal{L}_{\text{cycle}}(\widehat{D}^{H\downarrow}_{\text{rec}}, D^{H\downarrow})$ for instances in $H_\downarrow$ only.

The source L and downsampled $H_\downarrow$ sets have different range and distribution of depth values; to prevent both networks g_{L2H} and g_{H2L} from learning systematic shifts, we regularize translation via the two *range regularization terms*

$$\begin{aligned}\mathcal{R}^L_{\text{range}}(D^L) &= \text{MAE}\big(g_{L2H}(I, D^L), D^L\big),\\ \mathcal{R}^{H\downarrow}_{\text{range}}(D^{H\downarrow}) &= \text{MAE}\big(g_{H2L}(I, D^{H\downarrow}), D^{H\downarrow}\big).\end{aligned} \tag{14}$$

We additionally aim to prevent the reverse translation network g_{L2H} from distorting depth images in the clean downsampled $H_\downarrow$ by introducing the *idempotency regularization term*

$$\mathcal{R}_{\text{idt}}(D^{H\downarrow}) = \text{MAE}(g_{L2H}(I, D^{H\downarrow}), D^{H\downarrow}). \tag{15}$$

The system is trained by optimizing the following final translation objective:

$$\begin{aligned}\mathcal{L}_{\text{trans}} &= \mathcal{L}_{\text{adv}} + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}}\\ &\quad + \lambda^L_{\text{range}}\mathcal{R}^L_{\text{range}} + \lambda^{H\downarrow}_{\text{range}}\mathcal{R}^{H\downarrow}_{\text{range}} + \lambda_{\text{idt}}\mathcal{R}_{\text{idt}}.\end{aligned} \tag{16}$$

For brevity, we specify formulation for our adversarial loss $\mathcal{L}_{\text{adv}}$ in the Supplemental. We describe the architecture for our models and hyperparameter choices in the experimental Section V-A.

IV. METHODOLOGY AND DATA CONSTRUCTION FOR EVALUATION OF PAIRED AND UNPAIRED APPROACHES

The goals of our evaluation are (1) to compare our unpaired depth SR method with paired and unpaired methods on as equal terms as possible, and (2) to study the performance of our method in a realistic unpaired training scenario. For this, we develop a framework that enables directly performing such comparisons; we use this framework for creating a benchmark based on two datasets containing RGB-D scans of indoor scenes: ScanNet [17], which contains scans captured with Structure depth sensor, and InteriorNet [19], which contains high-quality simulated RGB-D scans.

A. DATASET CONSTRUCTION METHODOLOGY

We start from a paired dataset containing aligned low- (L) and high-quality RGB-D images (such as $H_\downarrow$ or H) of the same 3D environments (scenes) and adapt it as follows (see Figure 7):

1) We split the dataset into training, validation and test sets, each using a different subset of scenes.
2) We further split the training set into two disjoint parts, also using distinct scenes, Train A and Train B.
3) The unpaired training set U includes low-quality images in Train A and high-quality images in Train B.
4) The paired training set P combines low-quality images in Train B and high-quality images in Train B.
5) Test set T, obtained at step 1, is the same for both paired and unpaired methods.

This approach enables directly comparing unsupervised methods trained on U with supervised methods trained on P. We further develop a depth SR and enhancement benchmark, that we describe below.

B. BENCHMARK VARIETIES

Our benchmark consists of three parts that serve distinct purposes, as summarized by Table 2. *ScanNet-RenderScanNet* is built from raw sensor RGB-D images from ScanNet and high-quality depth rendered from 3D reconstructions of ScanNet scenes obtained automatically using depth fusion [18]. This benchmark aims to provide uniform training data for both paired and unpaired methods, and is split into three sub-parts *Train A, Train B,* and *Val,* following the data construction framework. Additionally, training on *ScanNet-RenderScanNet* allows to study unpaired methods in a controlled training scenario when the low- and high-quality

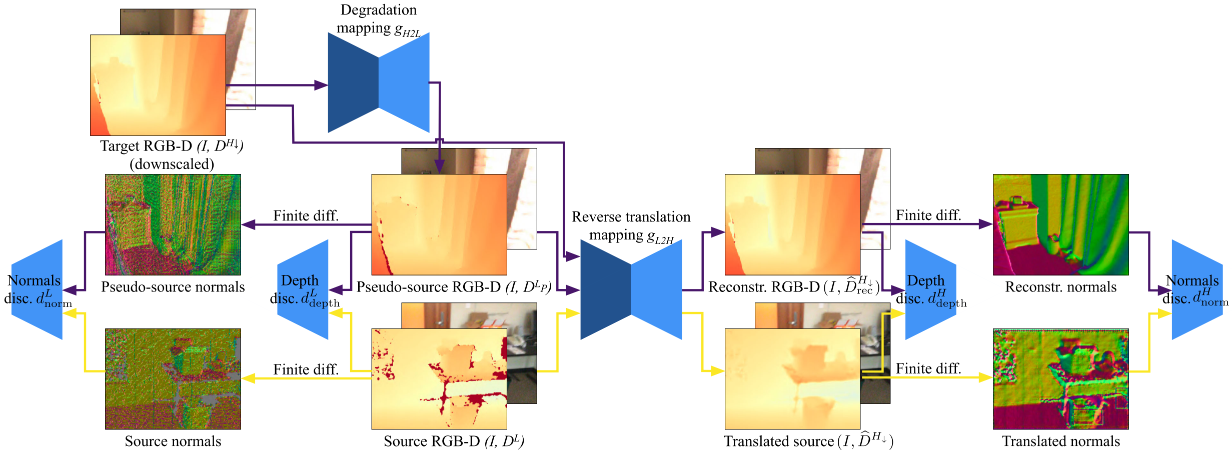

Degradation mapping g_{H2L}
Target RGB-D (I, $D^{H\downarrow}$) (downscaled)
Finite diff.
Normals disc. d^L_{norm}
Pseudo-source normals
Depth disc. d^L_{depth}
Pseudo-source RGB-D (I, D^{L_P})
Reverse translation mapping g_{L2H}
Reconstr. RGB-D $(I, \widehat{D}^{H_\downarrow}_{\mathrm{rec}})$
Finite diff.
Depth disc. d^H_{depth}
Reconstr. normals
Normals disc. d^H_{norm}
Finite diff.
Source normals
Source RGB-D (I, D^L)
Finite diff.
Translated source $(I, \widehat{D}^{H_\downarrow})$
Translated normals

FIGURE 6. During training of the the translation network, we use normals-based discriminator networks in addition to depth-based discriminators; this leads to improved performance in our evaluation.

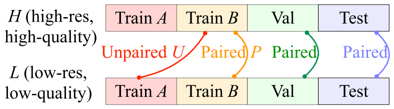


FIGURE 7. Unpaired methods are trained on disjoint parts A and B, while paired ones have training pairs.

TABLE 2. Summary of the used training and validation datasets. "Rendered [18]" refers to rendering depth images from 3D reconstructions of ScanNet scenes obtained using BundleFusion [18].

Split	Set	Source	RGB	Depth	Volume
ScanNet-RenderScanNet (paired and unpaired)					
Train A	L	ScanNet [17]	1280 × 960	640 × 480	6 221
Train B	L	ScanNet [17]	1280 × 960	640 × 480	6 221
	H	Rendered [18]	1280 × 960	1280 × 960	6 221
Val	L	ScanNet [17]	1280 × 960	640 × 480	2 945
	H	Rendered [18]	1280 × 960	1280 × 960	2 945
ScanNet-InteriorNet (unpaired only)					
Train A	L	*Train A* (L) in *ScanNet-RenderScanNet*			
Train B	H	InteriorNet [19]	1280 × 960	1280 × 960	13 744
Testing dataset					
Test	L	ScanNet [17]	1280 × 960	640 × 480	501
	H	Rendered [18]	1280 × 960	1280 × 960	501

data differ only in quality but not in semantics or other properties which could impair training stability. *ScanNet-InteriorNet* contains real-world sensor data from ScanNet and photorealistic images and high-quality virtual scans of procedurally generated scenes from InteriorNet [19]. With this benchmark, we aim to evaluate unpaired methods in a realistic training scenario when the low- and high-quality data differ not only in quality, but also in semantic content, distribution of depth values, etc. *Testing dataset* is a hold-out set for evaluation of both paired and unpaired depth map super-resolution methods; it consists of full-sized sensor RGB-D images from ScanNet and their respective high-quality renders.

V. EXPERIMENTAL EVALUATION

A. EXPERIMENTAL SETUP

We evaluate our method in depth map super-resolution with a scaling factor $k = 2$, training it on either *ScanNet-RenderScanNet* or *ScanNet-InteriorNet,* and testing on *Testing dataset.* As our method is capable of performing depth enhancement at the original spatial resolution, we present an additional evaluation on depth enhancement task with $k = 1$. Finally, we study the individual contributions of different components of our method in an ablation study. We specify the experimental setup here and give details specific to each experiment in respective Sections V-B and V-F.

1) NETWORK ARCHITECTURES

Translation networks g_{H2L} and g_{L2H} accept an RGB and a depth image as input and produce a depth image as output. Both networks share the same architecture that consist of convolutional RGB and depth feature encoders computing features that are concatenated and processed by 9 ResNet [68] blocks, and a convolutional decoder that produces a final depth image. For feature normalization in these models, we apply GroupNorm [69]. All discriminators follow the CycleGAN [46] discriminator architecture. We use Spectral Normalization [70] for weight regularization but perform no feature normalization in discriminators.

The RGB Guidance Network f_{rgb} architecturally includes a feature extractor (2 downsampling blocks, 6 ResNet blocks, and 2 upsampling blocks) and the depth prediction (a vanilla U-Net [64]) sub-networks. We use Instance Normalization [71] in both these parts.

The enhancement network f_{enh} architecturally is a vanilla U-Net [64].

We include information on the specific network parameters and training configurations in Supplemental, Table S1. All networks in our algorithm are trained from scratch.

2) DATA PREPROCESSING AND AUGMENTATION

All super-resolution methods in our experiments are based on fully-convolutional neural networks and can be naturally trained on crops of arbitrary size, and we build *Train B* and *Val* sub-parts from crops instead of full-sized images in order to increase the size of these subsets, as we explain further.

For the three datasets (ScanNet, RenderScanNet, and InteriorNet), we select RGB-D frames with the maximum depth value not exceeding 5.1 m, corresponding to the 85th depth range percentile in ScanNet [17]. Such filtering is consistent with the depth range where commercial depth sensors have sufficient precision while simultaneously allowing to reduce discrepancy better between images in synthetic and real datasets and providing common threshold for data normalization, which is a common practice for deep learning.

The depth fusion process [18] is not perfect, and rendering of reconstructed meshes has been shown to produce local misalignments [12] and create unnatural regions in output images. To reduce this effect, we applied a filtering procedure similar to the one used in [12]: we extracted 320 × 320 and 640 × 640 patches from input RGB-D frames and measured structured similarity [72] (SSIM) between the raw and rendered depth patch and its downsampled version, selecting images with SSIM greater than 0.8, and discarding pairs if the rendering produces misaligned result relatively to the sensor depth. We also store full frames with at least one patch selected.

To improve hole inpainting performance, we randomly add N rectangular holes to each training RGB-D frame, sampling N uniformly in [10, 75] range. The holes have random sizes (h_n, w_n) where h_n is uniformly sampled in the range $H/128 \dots H/8$, and w_n from $W/128 \dots W/8$ where (H, W) represent the dimensions of the input depth image. We perform this during each training iteration with probability 0.9.

3) QUALITY MEASURES

We calculate seven performance measures by comparing the super-resolved or enhanced depth images against their high-quality counterparts. RMSE, the root mean squared error, emphasizes large deviations. MAE, the mean absolute error, quantifies an average error without taking outliers into account. RMSE_h and MAE_h, the errors averaged over pixels with missing input depth value, assess the inpainting performance of the method. RMSE_d and MAE_d, the errors averaged over pixels with defined input depth value, measure the quality of the method in regions with valid input, and are well-suited for the evaluation of approaches that do not perform inpainting. Finally, MSE_v defined in Equation (3) measures the perceptual similarity between the 3D surface represented using an output depth map and the 3D surface represented by a reference depth map. In all calculations, we exclude pixels with unknown reference depth values. We report all metrics except MSE_v in millimeters.

TABLE 3. For a *paired* scenario *ScanNet-RenderScanNet* (where supervised training is possible), our UDSR outperforms all unpaired methods and quantitatively approaches the supervised SRFBN [4] and MS-PFL [10] methods. Lower = better; we show $\text{MSE}_\text{v} \times 10^{-2}$.

Method	RMSE↓	RMSE_h	RMSE_d	MAE	MAE_h	MAE_d	MSE_v
SRFBN [4]*	363.5	1435.5	129.0	109.3	1392.1	**27.1**	25.8
MS-PFL [10]*	299.5	1163.8	117.7	93.7	1110.6	29.3	24.9
Gu [1] + SRBFN [4]*	107.7	194.1	94.7	56.8	121.2	51.7	28.1
SRFBN [4]§	76.9	176.4	58.5	22.2	92.5	16.4	16.4
MS-PFL [10]§	75.5	168.4	60.4	30.2	113.3	23.8	16.1
Bicubic + Gu [1]	114.7	323.0	78.5	56.3	246.5	43.1	52.0
Gu [1] + Bicubic	108.0	193.8	95.0	57.1	121.4	52.1	28.5
UDSR (Ours)	**86.2**	**172.6**	**74.7**	**45.5**	**113.8**	40.6	**20.2**

§ supervised algorithms trained on (L, H)
* supervised algorithms trained on $(H_\downarrow, H)$
non-marked methods Gu *et al.* [1] and UDSR are unpaired

4) TRAINING COMPETITOR METHODS

For a fair comparison between competitors and our approach we re-train all competitor methods on the datasets available in our benchmark; we do not use datasets available from respective authors as they are unable to provide equal conditions for our comparisons.

B. SUPER-RESOLUTION ON SCANNET-RENDERSCANNET

1) METHODS AND TRAINING

We compare our method with three other learning-based methods. MS-PFL [10] is a state-of-the-art supervised depth SR method based on progressive multi-scale fusion of features extracted from input depth and RGB images. SRFBN [4] is an established supervised method for RGB image super-resolution based on recurrent connections which allow to incorporate high-level information flow. SRFBN often serves as a strong baseline for evaluating depth super-resolution methods (for instance, in [10], [11]). Following [10], we modified SRFBN to take a single-channel depth tensor as input and produce a single-channel output instead of a three-channel RGB one. Lastly, Gu et al. [1] is the only existing unpaired method for depth enhancement with a state-of-the-art performance. We adapt this method for super-resolution as explained below.

To more fully characterize performance of supervised methods, we include results obtained by training these methods to predict clean high-resolution targets in H using two distinct variants of input data (see Table 3). In the first variant, we use synthesized low-resolution inputs from the downsampled $H_\downarrow$ as done commonly by depth SR literature; the second variant consists in using registered real-world inputs from the source L available in our benchmarks.

To adapt the enhancement method of Gu et al. for super-resolution, we combined it with upsampling in three ways. For the first combination (we denote it "Bicubic + Gu"), we trained the method of Gu et al. on bicubically upsampled sensor depth from the source L as inputs and high-resolution high-quality depth images from H as targets, and for testing applied the method to bicubically upsampled sensor depth. For the second ("Gu + Bicubic"), we trained the method on sensor depth maps from L as inputs and downsampled high-quality depth maps from the downsampled $H_{\downarrow}$ as targets, and for testing applied bicubic interpolation to the output of the method. For the last combination ("Gu + SRFBN"), we again trained the method of Gu et al. on depth maps from L and $H_{\downarrow}$, but for upsampling during testing we used SRFBN trained on pairs from $H_{\downarrow}$ and H.

2) RESULTS

We display statistical evaluation results in Table 3 and visualize example predictions in Figure 8b. Our method outperforms all variants of the unpaired method of Gu et al. adapted for depth SR both quantitatively and qualitatively. All variants, particularly "Bicubic + Gu" where enhancement follows upsampling, produce surfaces with step-like artefacts, which is illustrated by the greyish color of normal maps, and which leads to a low perceptual quality (MSE_v) of the result. Variants where upsampling follows enhancement, *i.e.*, "Gu + Bicubic" and "Gu + SRFBN", produce very similar results, so we only show the qualitative results for the latter. We note that since our translation network g_{H2L} plays a similar role to Gu et al., similar results can be expected from using combinations of the form "Bicubic + g_{H2L}" or "g_{H2L} + Bicubic".

Compared to paired methods trained in the commonly used scenario on downsampled inputs from $H_{\downarrow}$, our method achieves lower RMSE_d and MSE_v but a higher MAE_d. As illustrated by visuals of the normal maps, compared to our algorithm, these methods produce significantly noisier surface, almost as noisy as the input sensor data; this is likely due to the domain shift between the synthetic input during training and the real-world sensor input during testing. This aligns with our initial motivation of exploring unpaired methods against the lack of representative paired data reflecting real-world input.

Compared to paired methods trained in the second scenario on data from L and H, our unpaired method performs only slightly worse quantitatively. At the same time, qualitative results of these methods contain artefacts that are not present in the predictions of our method; for instance, SRFBN produces ringing artefacts around object boundaries, and MS-PFL tends to output an over-smoothed depth.

C. SUPER-RESOLUTION ON SCANNET-INTERIORNET

1) METHODS AND TRAINING

For evaluation using the *ScanNet-InteriorNet* benchmark, we compared our method with the three adaptations of the method of Gu et al. described above.

2) RESULTS

Evaluation results are shown quantitatively in Table 4 and visually in Figure Figure 9b. Our method outperforms all three variants of Gu et al. which suffer from the same issues as during training on *ScanNet-RenderScanNet*. Notably, using ideal target data from InteriorNet leads to a similar or improved quantitative performance for our method, compared to using renders of ScanNet reconstructions.

TABLE 4. For an *unpaired* scenario *ScanNet-InteriorNet* with geometric and semantic differences in source and target datasets, our method outperforms state-of-the-art depth enhancement [1] coupled with trained [4] and untrained depth upsampling steps, across all quality measures we computed. Note: we only show combinations of methods that enable unpaired training. Lower = better; we show $\text{MSE}_\text{v} \times 10^{-2}$.

Method	RMSE↓	RMSE_h	RMSE_d	MAE	MAE_h	MAE_d	MSE_v
Bicubic + Gu [1]	241.2	893.1	109.0	93.5	771.5	50.5	47.9
Gu [1] + SRFBN [4]	107.2	298.2	67.8	45.7	210.0	33.1	42.5
Gu [1] + Bicubic	107.1	297.7	67.9	46.1	209.9	33.6	36.7
UDSR (Ours)	**81.1**	**197.9**	**61.0**	**28.0**	**123.5**	**20.6**	**26.1**

D. ENHANCEMENT ON SCANNET-RENDERSCANNET

1) METHODS AND TRAINING

In the task of depth enhancement with no change in spatial resolution we compared our method with five other learning-based methods: the unpaired method of Gu et al. (which differs from our method both architecturally and in its training procedure), LapDEN [12], a state-of-the-art supervised method for depth enhancement based on deep Laplacian pyramid network, and three unpaired methods for RGB image-to-image translation, CycleGAN [46], U-GAT-IT [73], and NiceGAN [74]. We modified them to take a single-channel depth tensor as input and produce a single-channel output, and for CycleGAN we additionally trained a version with four-channel RGB-D input.

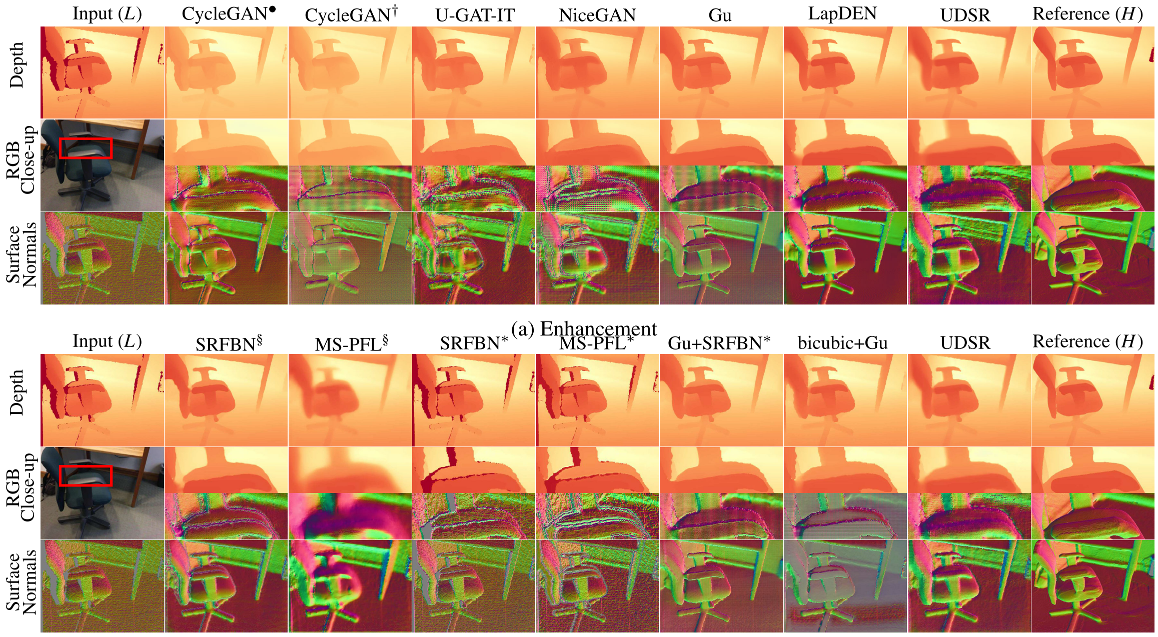
2) RESULTS

The evaluation results are shown in Table 5 and in Figure 8. Quantitatively, our method outperforms the other unpaired methods in all measures. Notably, even for a harder task of 2× depth SR, our method achieves higher scores in all measures except MSE_v compared to other unpaired methods on the task of depth enhancement (no change in spatial resolution). Qualitatively, our method successfully performs denoising, hole inpainting, and yields surface normals close to the target while the other unpaired methods suffer from various artefacts. Gu et al. produces the result with step-like artefacts, which is indicated by the greyish color of the normal map and a high value of MSE_v; NiceGAN and U-GAT-IT suffer from ringing artefacts around the object boundaries; U-GAT-IT and CycleGAN fail to preserve the correct absolute depth value, which is indicated by the shifted colors in depth image visualization and high values of RMSE- and MAE-based measures. Compared to the paired LapDEN, our method performs slightly worse quantitatively and produces the result with slightly noisier surface.

E. ENHANCEMENT ON SCANNET-INTERIORNET

1) METHODS AND TRAINING

We additionally perform evaluation of depth enhancement using the data available in the *ScanNet-InteriorNet* benchmark, using the same methods as described previously.



(a) Enhancement

(b) Depth SR (with enhancement)

FIGURE 8. Enhancement and depth SR on *ScanNet-RenderScanNet* benchmark. Each column corresponds to a separately trained network. The networks Gu et al. and SRFBN are trained separately from the case *ScanNet-InteriorNet*. § and * mark the models trained on (L, H) and $(H_{\downarrow}, H)$ pairs respectively. $^{\cdot}$ and † mark the models with the depth only input and depth image concatenated with RGB, respectively.

TABLE 5. Enhancement performance statistics using the *ScanNet-RenderScanNet* benchmark indicates that our UDSR significantly outperforms all existing unpaired methods and quantitatively approaches the supervised LapDEN [12]. Lower = better; we show $\mathrm{MSE_v} \times 10^{-2}$.

Method	RMSE↓	$\mathrm{RMSE_h}$	$\mathrm{RMSE_d}$	MAE	$\mathrm{MAE_h}$	$\mathrm{MAE_d}$	$\mathrm{MSE_v}$
LapDEN [12]	76.0	165.6	58.4	22.6	92.7	15.9	7.7
CycleGAN [46]	416.6	450.6	407.5	391.2	389.4	392.5	13.3
U-GAT-IT [73]	292.8	508.0	261.3	249.5	434.5	235.6	14.2
NiceGAN [74]	156.5	410.2	111.5	97.6	324.8	79.5	19.5
Gu *et al.* [1]	108.3	193.9	94.3	57.1	120.2	51.7	33.1
UDSR (Ours)	**77.1**	**169.6**	**63.3**	**33.0**	**110.1**	**27.3**	**11.9**

TABLE 6. Enhancement performance statistics for *ScanNet-InteriorNet* scenario. Our method outperforms all competitors across all measures we computed. Lower = better; we show $\mathrm{MSE_v} \times 10^{-2}$.

Method	RMSE↓	$\mathrm{RMSE_h}$	$\mathrm{RMSE_d}$	MAE	$\mathrm{MAE_h}$	$\mathrm{MAE_d}$	$\mathrm{MSE_v}$
NiceGAN [74]•	2063.7	2824.6	1984.3	1860.0	2759.8	1787.0	59.5
U-GAT-IT [73]•	1300.2	1064.4	1311.4	1254.7	999.3	1276.4	24.4
CycleGAN [46]•	405.9	640.0	359.5	356.8	566.6	336.6	48.2
CycleGAN [46]†	471.9	523.0	457.8	425.1	457.7	421.9	43.9
Gu *et al.* [1]	107.3	290.1	66.8	46.1	201.0	33.1	44.4
UDSR (Ours)	**80.8**	**199.9**	**59.1**	**29.9**	**124.8**	**22.2**	**14.7**

• input: depth image
† input: depth image concatenated with RGB

We consider two input possibilities where we supply depth image or depth image concatenated with RGB as input to CycleGAN [46]; U-GAT-IT [73], and NiceGAN [74] are trained using depth images only.

2) RESULTS

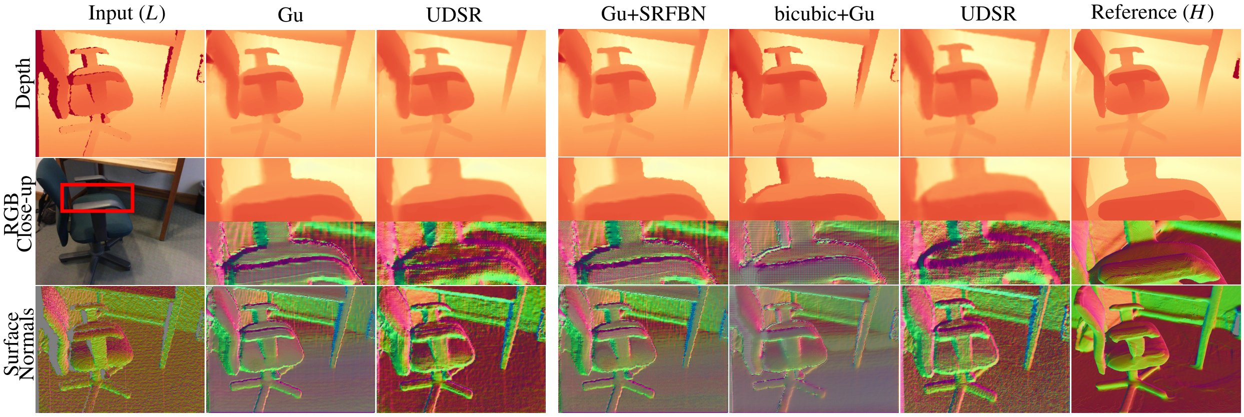
The evaluation results presented in Table 6 demonstrate significant quantitative performance gains across quality measures we compute, particularly for $\mathrm{MSE_v}$. Visual results Figure 9 demonstrate that our method recovers normals better compared to Gu et al., resulting in more visually appealing surface geometry.

F. ABLATION STUDIES

We study the effects of different components of our method by training its several versions on *ScanNet-InteriorNet* benchmark. Results of this ablative study are presented quantitatively in Table 7 and visually in Figure 10.

Removing the RGB guidance network f_{rgb} (UDSR†) and training without augmenting input data using random gaps (UDSR$^{\diamond}$) impairs inpainting performance both qualitatively and quantitatively as assessed by $\mathrm{RMSE_h}$ and $\mathrm{MAE_h}$. These findings are expected, as RGB guidance enables predicting depth in areas where direct range measurements are unavailable, while adding random holes is meant to provide robustness of our approach.

Compared to generating pseudo-examples using an unmodified CycleGAN (UDSR$^{\cdot}$), our surface normals-aware translation algorithm significantly improves performance for the full method w.r.t. all measures. Training without pseudo-examples at all (UDSR$^{\circ}$) produces low-quality surfaces, both qualitatively and as measured by $\mathrm{MSE_v}$.



(a) Enhancement

(b) Depth SR (with enhancement)

FIGURE 9. Enhancement and SR using the *ScanNet-InteriorNet* benchmark. Each column corresponds to a separately trained network.

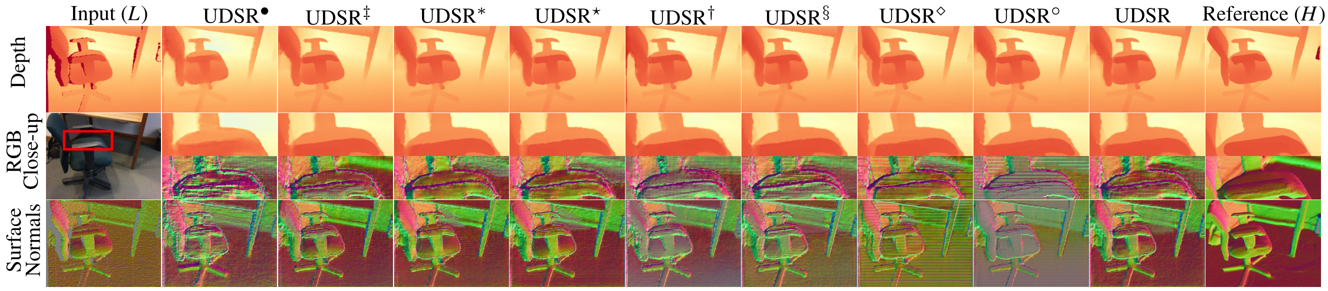


FIGURE 10. Ablation study of our depth SR method using data in *ScanNet-InteriorNet* benchmark (see Table 7 for notation).

TABLE 7. Ablation study of our depth SR algorithm using the data in *ScanNet-InteriorNet* benchmark.

Method	RMSE$_\mathrm{h}$	RMSE$_\mathrm{d}$	MAE$_\mathrm{h}$	MAE$_\mathrm{d}$	MSE$_\mathrm{v}$
UDSR$^\bullet$	238.4	203.8	183.7	188.2	39.3
UDSR‡	213.3	66.1	135.2	31.0	24.4
UDSR*	208.6	69.3	135.0	36.3	23.6
UDSR*	192.0	63.1	119.4	25.4	**23.0**
UDSR†	376.3	66.5	271.5	29.7	41.2
UDSR§	228.4	71.6	151.1	36.6	35.1
UDSR$^\diamond$	342.1	66.2	279.9	28.8	28.8
UDSR$^\circ$	**177.7**	62.8	**104.2**	24.6	43.2
UDSR (Ours)	197.9	**61.0**	123.5	**20.6**	26.1

Modifications of the unpaired translation algorithm:
$^\bullet$ replaced with unmodified CycleGAN,
‡ trained without normals-based discriminators d^H_{norm}, d^L_{norm},
* trained without MSE$_\mathrm{v}$ in cycle-consistency loss $\mathcal{L}_{\mathrm{cycle}}$,
* trained without depth range regularizer $\mathcal{R}^*_{\mathrm{range}}$.
Modifications of the enhancement algorithm:
† without RGB guidance network f_{rgb},
§ trained without normals-based loss term $\mathcal{L}_{\mathrm{surf}}$,
$^\diamond$ trained without additional holes in the input,
$^\circ$ trained without pseudo-examples.

VI. CONCLUSION

We have described a new approach to data-driven depth super-resolution and enhancement, eliminating the need for paired datasets and simplifying the training process using real-world depth data. Our learning-based pipeline efficiently trains a depth enhancement model using diverse data from different 3D environments, featuring varying resolutions and qualities. Throughout it, we have introduced several enhancements to improve robustness, accuracy, and completeness of depth reconstructions.

Additionally, we have introduced a novel benchmark for depth super-resolution, leveraging real-world RGB-D scans from existing collections [17] and higher quality data obtained by rendering reconstructions from multiple scans. This benchmark provides a valuable evaluation framework for comparing depth enhancement methods.

To summarize, our contributions comprise the development of UDSR, a new method for unpaired depth super-resolution, along with an efficient color-guided unpaired depth enhancement algorithm. Furthermore, we have established a comprehensive benchmark for real-world depth SR and enhancement, advancing the state-of-the-art in these domains.

A. FUTURE WORK

Although our method outperformed existing unpaired methods, we believe that performance can be further improved with fine-tuning the models. The improvement of the model robustness and scalability is another possible future direction. Our method uses a simplistic augmentation models where rectangular holes are added into input data to simulate gaps in real acquisitions; real holes, however, are likely to be correlated with edges between foreground and background objects. Richer, more effective augmentations including hard cases could be created by special lighting and scene arrangements. Another direction for future work is the development of paired training and evaluation datasets with high-fidelity reference depth measurements, for more accurate validation and training [33], [63].

REFERENCES

[1] X. Gu, Y. Guo, F. Deligianni, and G.-Z. Yang, "Coupled real-synthetic domain adaptation for real-world deep depth enhancement," *IEEE Trans. Image Process.*, vol. 29, pp. 6343–6356, 2020.

[2] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.

[3] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3106–3121, Dec. 2019.

[4] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3862–3871.

[5] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5790–5799.

[6] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou, "Structure-preserving super resolution with gradient guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7766–7775.

[7] G. Riegler, M. Rüther, and H. Bischof, "ATGV-Net: Accurate depth super-resolution," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham, Switzerland: Springer, 2016, pp. 268–284.

[8] T.-W. Hui, C. C. Loy, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *Computer Vision—ECCV 2018*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham, Switzerland: Springer, 2016, pp. 353–369.

[9] O. Voynov, A. Artemov, V. Egiazarian, A. Notchenko, G. Bobrovskikh, E. Burnaev, and D. Zorin, "Perceptual deep depth super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5652–5662.

[10] C. Xian, K. Qian, Z. Zhang, and C. C. L. Wang, "Multi-scale progressive fusion learning for depth map super-resolution," 2020, *arXiv:2011.11865*.

[11] X. Song, Y. Dai, D. Zhou, L. Liu, W. Li, H. Li, and R. Yang, "Channel attention based iterative residual learning for depth map super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5630–5639.

[12] J. Jeon and S. Lee, "Reconstruction-based pairwise depth dataset for depth image enhancement using CNN," in *Computer Vision—ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham, Switzerland: Springer, 2018, pp. 438–454.

[13] V. Sterzentsenko, L. Saroglou, A. Chatzitofis, S. Thermos, N. Zioulis, A. Doumanoglou, D. Zarpalas, and P. Daras, "Self-supervised deep depth denoising," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Los Alamitos, CA, USA, Oct. 2019, pp. 1242–1251.

[14] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a GAN to learn how to do image degradation first," in *Computer Vision—ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham, Switzerland: Springer, 2018, pp. 187–202.

[15] S. Maeda, "Unpaired image super-resolution using pseudo-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 288–297.

[16] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 701–710.

[17] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2432–2443, doi: 10.1109/CVPR.2017.261. [Online]. Available: http://www.scan-net.org/

[18] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration," *ACM Trans. Graph.*, vol. 36, no. 3, pp. 1–18, May 2017.

[19] W. Li, S. Saeedi, J. McCormac, R. Clark, D. Tzoumanikas, Q. Ye, Y. Huang, R. Tang, and S. Leutenegger, "InteriorNet: Mega-scale multi-sensor photo-realistic indoor scenes dataset," 2018, *arXiv:1809.00716*.

[20] J. Kim, G. Jeon, and J. Jeong, "Joint-adaptive bilateral depth map upsampling," *Signal Process., Image Commun.*, vol. 29, no. 4, pp. 506–513, Apr. 2014.

[21] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 99, p. 96, Jul. 2007.

[22] B. Haefner, S. Peng, A. Verma, Y. Quéau, and D. Cremers, "Photometric depth super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2453–2464, Oct. 2020.

[23] B. Haefner, Y. Quéau, T. Möllenhoff, and D. Cremers, "Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 164–174.

[24] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep joint image filtering," in *Computer Vision—ECCV 2018*. Springer, 2016, pp. 154–169.

[25] Z. Yan, K. Wang, X. Li, Z. Zhang, G. Li, J. Li, and J. Yang, "Learning complementary correlations for depth super-resolution with incomplete data in real world," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 4, pp. 5616–5626, Apr. 2022.

[26] Z. Jiang, H. Yue, Y.-K. Lai, J. Yang, Y. Hou, and C. Hou, "Deep edge map guided depth super resolution," *Signal Process., Image Commun.*, vol. 90, Jan. 2021, Art. no. 116040.

[27] P. Liu, Z. Zhang, Z. Meng, N. Gao, and C. Wang, "PDR-Net: Progressive depth reconstruction network for color guided depth map super-resolution," *Neurocomputing*, vol. 479, pp. 75–88, Mar. 2022.

[28] R. De Lutio, A. Becker, S. D'Aronco, S. Russo, J. D. Wegner, and K. Schindler, "Learning graph regularisation for guided super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1969–1978.

[29] N. Metzger, R. C. Daudt, and K. Schindler, "Guided depth super-resolution by deep anisotropic diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18237–18246.

[30] Z. Zhao, J. Zhang, S. Xu, Z. Lin, and H. Pfister, "Discrete cosine transform network for guided depth map super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5687–5697.

[31] Y. Qiao, L. Jiao, W. Li, C. Richardt, and D. Cosker, "Fast, high-quality hierarchical depth-map super-resolution," in *Proc. 29th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 4444–4453.

[32] X. Ye, B. Sun, Z. Wang, J. Yang, R. Xu, H. Li, and B. Li, "PMBANet: Progressive multi-branch aggregation network for scene depth super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 7427–7442, 2020.

[33] L. He, H. Zhu, F. Li, H. Bai, R. Cong, C. Zhang, C. Lin, M. Liu, and Y. Zhao, "Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9225–9234.

[34] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 689–694, Aug. 2004.

[35] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[36] S. Lefkimmiatis, "Non-local color image denoising with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5882–5891.

[37] S. Lefkimmiatis, "Universal denoising networks: A novel CNN architecture for image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3204–3213.

[38] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, "Toward convolutional blind denoising of real photographs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1712–1722.

[39] H. Wang, M. Wang, Z. Che, Z. Xu, X. Qiao, M. Qi, F. Feng, and J. Tang, "RGB-depth fusion GAN for indoor depth completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6199–6208.

[40] Y. Zhang, X. Guo, M. Poggi, Z. Zhu, G. Huang, and S. Mattoccia, "CompletionFormer: Depth completion with convolutions and vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18527–18536.

[41] Y. Zhang and T. Funkhouser, "Deep depth completion of a single RGB-D image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 175–185.

[42] 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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 85–100.

[43] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[44] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "StructureFlow: Image inpainting via structure-aware appearance flow," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 181–190.

[45] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, "Contextual residual aggregation for ultra high-resolution image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7505–7514.

[46] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[47] X. Xu, P. Wei, W. Chen, Y. Liu, M. Mao, L. Lin, and G. Li, "Dual adversarial adaptation for cross-device real-world image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5657–5666.

[48] S. Xie, Y. Xu, M. Gong, and K. Zhang, "Unpaired image-to-image translation with shortest path regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10177–10187.

[49] M. I. A. Shocher and N. Cohen, "'Zero-shot' super-resolution using deep internal learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3118–3126.

[50] J. Li, Z. Zhang, X. Liu, C. Feng, X. Wang, L. Lei, and W. Zuo, "Spatially adaptive self-supervised learning for real-world image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9914–9924.

[51] W. Lee, S. Son, and K. M. Lee, "AP-BSN: Self-supervised denoising for real-world images via asymmetric PD and blind-spot network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17704–17713.

[52] G. Vaksman and M. Elad, "Patchcraft self-supervised training for correlated image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5795–5804.

[53] R. Neshatavar, M. Yavartanoo, S. Son, and K. M. Lee, "CVF-SID: Cyclic multi-variate function for self-supervised image denoising by disentangling noise from image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17562–17570.

[54] A. Shabanov, I. Krotov, N. Chinaev, V. Poletaev, S. Kozlukov, I. Pasechnik, B. Yakupov, A. Sanakoyeu, V. Lebedev, and D. Ulyanov, "Self-supervised depth denoising using lower-and higher-quality RGB-D sensors," in *Proc. Int. Conf. 3D Vis. (3DV)*, Nov. 2020, pp. 743–752.

[55] Z. Sun, W. Ye, J. Xiong, G. Choe, J. Wang, S. Su, and R. Ranjan, "Consistent direct time-of-flight video depth super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5075–5085.

[56] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Pattern Recognition*. Springer, 2014, pp. 31–42.

[57] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 7576, 2012, pp. 746–760.

[58] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 567–576.

[59] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Hong Kong, May 2014, pp. 1524–1531.

[60] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3D: Learning from RGB-D data in indoor environments," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 667–676.

[61] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 993–1000.

[62] J. Park, Q.-Y. Zhou, and V. Koltun, "Colored point cloud registration revisited," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 143–152.

[63] O. Voynov, G. Bobrovskikh, P. Karpyshev, S. Galochkin, A.-T. Ardelean, A. Bozhcnko, E. Karmanova, P. Kopanev, Y. Labutin-Rymsho, R. Rakhimov, A. Safin, V. Serpiva, A. Artemov, E. Burnaev, D. Tsetserukou, and D. Zorin, "Multi-sensor large-scale dataset for multi-view 3D reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21392–21403.

[64] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015* (Lecture Notes in Computer Science), N. Navab, J. Hornegger, W. M. W. Iii, and A. F. Frangi, Eds., Springer, 2015, pp. 234–241.

[65] C. Zheng, T.-J. Cham, and J. Cai, "T2Net: Synthetic-to-realistic translation for solving single-image depth estimation tasks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 767–783.

[66] G. Riegler, D. Ferstl, M. Rüther, and H. Bischof, "A deep primal-dual network for guided depth super-resolution," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–14.

[67] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2813–2821.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[69] Y. Wu and K. He, "Group normalization," *Int. J. Comput. Vis.*, vol. 128, no. 3, pp. 742–755, Mar. 2020.

[70] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. 6th Int. Conf. Learn. Represent.*, Vancouver, BC, Canada, 2018, pp. 1–29.

[71] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.

[72] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[73] J. Kim, M. Kim, H. Kang, and K. H. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[74] R. Chen, W. Huang, B. Huang, F. Sun, and B. Fang, "Reusing discriminators for encoding: Towards unsupervised image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8165–8174.

ALEKSANDR SAFIN received the B.S. degree from the Higher School of Economics, in 2017, and the M.S. degree from Skoltech, in 2019, where he is currently pursuing the Ph.D. degree, focusing on unpaired learning and 3D reconstruction. Previously, he collaborated with the Huawei Research Center on Computational Photography.

MAXIM KAN received the bachelor's degree in applied electronics from TUSUR, in 2018, and the Master of Science degree from Skoltech, in 2021, building upon his prior academic achievement. He is currently a Deep Learning Engineer based in Moscow, also with Phygitalism. His research interests include generative models and 3D data processing.

NIKITA DROBYSHEV received the B.S. degree from National Research Nuclear University (MEPhI), in 2018, and the M.S. degree from Skoltech, in 2021. He is currently a Deep Learning Engineer and also with Samsung AI Center, Moscow. Prior to joining Samsung, he was a Machine Learning Engineer with Sber.

OLEG VOYNOV received the B.S. and M.S. degrees in applied mathematics and in physics from MIPT, in 2017. He is currently pursuing the Ph.D. degree with Skoltech. He is currently a Researcher with the Skoltech Applied AI Center and AIRI. He was an Intern in computational geophysics with MIPT. He developed deep learning algorithms for 3D surface reconstruction from RGB-D data with Skoltech.

ALEXEY ARTEMOV received the Ph.D. degree from the Institute for Systems Analysis, Russian Academy of Sciences, in 2017. He is currently a Postdoctoral Researcher with the Technical University of Munich focusing on 3D scanning and reconstruction. In 2021, he received the Ilya Segalovich Award for young researchers.

ALEXANDER FILIPPOV received the M.Sc. degree in mathematics from Moscow State University, in 2004, and the Ph.D. degree in computer science from the Russian Academy of Sciences, in 2009. He is currently working as the Director of the AI Foundation and Algorithm Laboratory. His main research interests include deep learning for image and video processing, computational photography, and effective AI for terminal devices.

DENIS ZORIN is currently a Silver Professor in computer science and in mathematics with the Courant Institute of Mathematical Sciences, New York University. His distinctions include ACM Gordon Bell Prize, SIGGRAPH Achievement Award, and a number of best paper awards. His research interests include geometric modeling, geometry processing, and scientific computing.

EVGENY BURNAEV received the M.S. degree from MIPT, in 2006, the Ph.D. degree in theoretical foundations of informatics, in 2008, and the Habilitation (Doctor of Science) degree in mathematical modeling, in 2022. He is currently a Full Professor with Skoltech. His research interests include deep learning for 3D computer vision, manifold learning, and generative modeling.