DDRNet: Depth Map Denoising and Refinement for Consumer Depth Cameras Using Cascaded CNNs

Shi Yan[1], Chenglei Wu[2], Lizhen Wang[1], Feng Xu[1], Liang An[1], Kaiwen Guo[3], and Yebin Liu[1]

[1] Tsinghua University, Beijing, China
[2] Facebook Reality Labs, Pittsburgh, USA
[3] Google Inc, USA

Abstract. Consumer depth sensors are more and more popular and come to our daily lives marked by its recent integration in the latest Iphone X. However, they still suffer from heavy noises which dramatically limit their applications. Although plenty of progresses have been made to reduce the noises and boost geometric details, due to the inherent illness and the real-time requirement, the problem is still far from been solved. We propose a cascaded Depth Denoising and Refinement Network (DDRNet) to tackle this problem by leveraging the multi-frame fused geometry and the accompanying high quality color image through a joint training strategy. The classic rendering equation is delicately exploited in our network in an unsupervised manner. Experimental results indicate that our network achieves real-time denoising and refinement on various categories of static and dynamic scenes. Thanks to the well decoupling of the low and high frequency information in the cascaded network, we achieve superior performance over the state-of-the-art techniques.

Keywords: Depth enhancement · Consumer depth camera · Unsupervised learning · Convolutional Neural Networks· DynamicFusion

1 Introduction

Consumer depth cameras have enabled lots of new applications in computer vision and graphics, ranging from live 3D scanning to virtual and augmented reality. However, even with tremendous progresses in improving the quality and resolution, current consumer depth cameras still suffer from heavy sensor noises.

During the past decades, in view of the big quality gap between depth sensors and traditional image sensors, researchers have made great efforts to leverage RGB images or videos to bootstrap the depth quality. While RGB-guided filtering methods show the effectiveness [22, 34], a recent trend is on investigating the light transport in the scene for depth refinement with RGB images, which is able to capture high frequency geometry and reduce the texture-copy artifacts [43, 12, 46, 3]. Progresses have also been made to push these methods to run in real time [44, 30]. In these traditional methods, before refinement, a smooth

filtering is usually carried out on the raw depth to reduce the sensor noise. However, this simple spatial filtering may alter the low-dimensional geometry in a non-preferred way. This degeneration can never be recovered in the follow-up refinement step, as only high-frequency part of the depth is modified.

To attack these challenges, we propose a new cascaded CNN structure to perform depth image denoising and refinement in order to lift the depth quality in low frequency and high frequency simultaneously. Our network consists of two parts, with the first focusing on denosing while the second aiming at refinement. For the denoising net, we train a CNN with a structure similar to U-net [36]. Our first contribution is on how to generate training data. Inspired by the recent progress on depth fusions [19, 26, 11], we generate reference depth maps from the fused 3D model. With fusion, heavy noise present in single depth map can be reduced by integrating the truncated signed distant function (TSDF). From this perspective, our denoising net is learning a deep fusion step, which is able to achieve better depth accuracy than heuristic smoothing.

Our second contribution is the refinement net, structured in our cascade end-to-end framework, which takes the output from the denoising net and refine it to add high-frequency details. Recent progresses in deep learning have demonstrated the power of deep nets to model complex functions between visual components. One challenge to train a similar net to add high-frequency details is that there is no ground truth depth map with desired high-frequency details. To solve this, we propose a new learning-based method for depth refinement using CNNs in an unsupervised way. Different from traditional methods, which define the loss directly on the training data, we design a generative process for RGB images using the rendering equation [20] and define our loss on the intensity difference between the synthesized image and the input RGB image. Scene reflectance is also estimated through a deep net to reduce the texture-copy artifacts. As the rendering procedure is fully differentiable, the image loss can be effectively back propagated throughout the network. Therefore, through these two components in our DDRNet, a noisy depth map is enhanced both in low frequency and high frequency.

We extensively evaluate our proposed cascaded CNNs, demonstrating that our method can produce depth map with higher quality in both low and high frequency, compared with the state-of-the-art methods. Moreover, the CNN-based network structure enables our algorithm to run in real-time. And with the progress of deep-net-specific hardware, our method is promising to be deployed on mobile phones. Applications of our enhanced depth stream in the DynamicFusion systems [26, 11] are demonstrated, which improve the reconstruction performance of the dynamic scenes.

2 Related Work

Depth Image Enhancement As RGB images usually capture a higher resolution than depth sensors, many methods in the past have focused on leveraging the RGB images to enhance the depth data. Some heuristic assumptions are usu-

ally made about the correlation between color and depth. For example, some work assume that the RGB edges are coinciding with depth edges or discontinuities. Diebel and Thrun [9] upsample the depth with a Markov-Random Field. Depth upsampling with color image as input can be formulated as an optimization problem which maximizes the corrleation between RGB edges and depth discontinuities [31]. Another way to implement this heuristics is through filtering [23],e.g. with joint bilateral upsampling filter [22]. Yang et al [45] propose a depth upsampling method by filtering a cost space joint-bilaterally with a stereo image to achieve the resolution upsampling. Similar joint reconstruction ideas with stereo images and depth data are investigate by further constraining the depth refinement with photometric consistency from stereo matching [49]. With the development of modern hardwares and also the improvements in filtering algorithms, variants of joint-bilateral or multilateral filtering for depth upsampling can run in real-time [6, 10, 34]. As all of these methods are based on the heuristic assumption between color and depth, even producing plausible results, refined depth maps are not metrically accurate, and texture-copy artifacts are inevitable as texture variations are frequently mistaken for geometric detail.

Depth Fusion With multiple frames as input, different methods have been proposed to fuse them to improve the depth quality or obtain a better quality scan. Cue et al. [8] has proposed a multi-frame superresolution technique to estimate higher resolution depth images from a stack of aligned low resolution images. Taking into account the sensors' noise characteristics, the signed distance function is employed with an efficient data structure to scan scenes with an RGBD camera [16]. KinectFusion [27] is the first method to show real-time hand-held scanning of large scenes with a consumer depth sensor. Better data structures that exploit spatial sparsity in surface scans, e.g. hierarchical grids [7] or voxel hashing schemes [28], have been proposed to scan larger scenes in real time. These fusion methods are able to effectively reduce the noises in the scanning by integrating the TSDF. Recent progresses have extended the fusion to dynamic scenes [26, 11]. The scan from these depth fusion methods can achieve very clean 3D reconstruction, which improves the accuracy of the original depth map. Based on this observation, we employ depth fusion to generate a training data for our denoising net. By feeding lots of the fused depth as our training data to the the network, our denoising net effectively learns the fusion process. In this sense, our work is also related to Riegler et al. [35], where they designed an OctNet to perform the learning on signed distance function. Differently, our denoising net directly works on depth and by special design of our loss function, our net can effectively reduce the noise in the original depth map. Besides, high frequency geometric detail is not dealt with in OctNet, while by our refinement net we can achieve detailed depth maps.

Depth Refinement with Inverse Rendering To model the relation between color and depth in a physically correct way, inverse rendering methods have been proposed to leverage RGB images to improve depth quality by investigating the light transport process. Shape-from-shading (SfS) techniques have been investi-

gated on how to extract the geometric detail from a single image [17, 48]. One challenge to directly apply SfS is that the light and reflectance are usually unknown when capturing the depth map. Recent progresses have shown that SfS can refine coarse image-based geometry models [4], even if they were captured under general uncontrolled lighting with multi-view cameras [43, 42] or an RGB-D camera [12, 46]. In these work, illumination and albedo distributions, as well as refined geometry are estimated via inverse rendering optimization. Optimizing all these unknowns are very challenging by traditional optimization schemes. For instance, if the reflectance is not properly estimated, the texture-copy artifact can still exist. In our work, we employ a specifically structured network to tackle the challenge of reflectance and geometry separation problem. Our network structure can be seen as a regularizer which constrain the inverse rendering loss to back propagate only learnable gradient to train our refinement net. Also with a better reflectance estimation method than previous work, the reflectance influence can be further alleviated, resulting in a CNN network which extracts only geometry-related information to improve the depth quality.

Learning-based and Statistical Methods Data driven methods are another category to solve the depth upsampling/refinement problem. Data-driven priors are also helpful for solving the inverse rendering problem. Barron and Malik [2] jointly solve reflectance, shape and illumination, based on priors derived statistically from images. Similar concepts were also used for offline intrinsic image decomposition of RGB-D data [1]. Khan *et al.* [21] learn weighting parameters for complex SfS models to aid facial reconstruction. Wei and Hirzinger [40] use deep neural networks to learn aspects of the physical model for SfS. Note that even our method is also learning based, our refinement net does not take any training data. Instead, the refinement net relies on a pre-defined generative process and thus an inverse rendering loss for the training process. The closest idea to our paper is the encoder-decoder structure used for image-based face reconstruction [38, 33]. They take the traditional rendering pipeline as a generative process, defined as a fixed decode. Then, a reconstruction loss can be optimized to train the encoder, which directly regress from a input RGB image. However, these methods all require a predefined geometry and reflectance subspace, usually modeled by linear embedding, to help train a meaningful encode, while our method can work with general scenes captured by RGBD sensor.

3 Method

We propose a new framework for jointly training a denoising net and a refinement net from a consumer-level camera to improve depth map both in low frequency and high frequency. The proposed pipeline features our novelties both in training data creation and cascaded CNNs architecture design. To obtain ground-truth high-quality depth data for training is very challenging. We thus have formulated the depth improvement problem into two pixel-wise regression tasks, while each one focuses on lifting the quality in different frequency domains. This also enables us to combine supervised learning unsupervised learning to solve the issue

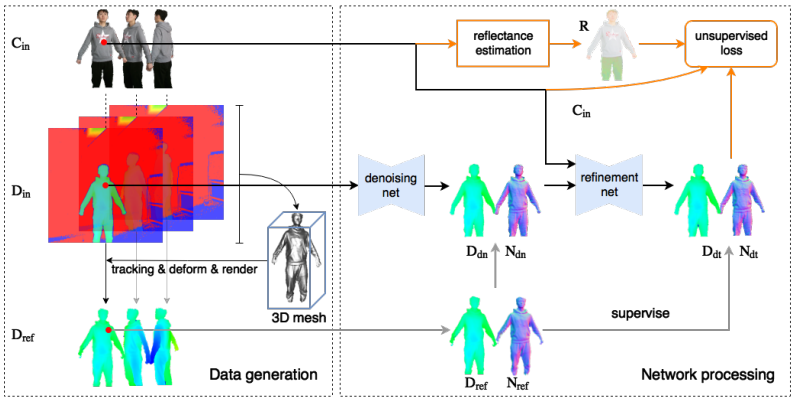

Fig. 1. The pipeline of our method. The black lines are the forward pass during test, the gray lines are supervise signal, and the orange lines are related to the unsupervised loss. Note that every loss function has a input mask W, which is omitted in this figure. N_{ref}, N_{dt} are reference normal map and refined normal map with detail, D_{dn} and D_{dt} are denoised and refined output.

of lacking ground truth training data. For denoising part, a function $\mathcal{D}$ mapping a noisy depth map D_{in} to a smoothed one D_{dn} with high-quality low frequency is learned by a CNN with the supervision of near-groundtruth depth maps D_{ref}, created from a state of the art of dynamic fusion. For refinement part, an unsupervised shading-based criterion is developed based on inverse rendering to train and a function $\mathcal{R}$ to map D_{dn} and the corresponding RGB image C_{in} to an improved depth map D_{dt} with rich geometric details. The albedo map for each frame is also estimated by a CNN in a way similar to [25]. We concurrently train cascaded CNNs from supervised depth data and unsupervised shading cues to achieve state-of-the-art performance on the task of single image depth enhancement. The detailed pipeline can be visualized in Figure 1.

3.1 Dataset

Previous methods usually take a shortcut to obtain the training data by synthesizing [37, 39]. However, what if noise characteristic varies from sensor to sensor, or even the noise source is untraceable? In this case, how to generate ground-truth (or near-ground-truth) depth map becomes a major problem.

Data generation. In order to learn the real noise distribution of different consumer depth cameras, we need to collect a training dataset of raw depth data with corresponding target depth maps, which act as the supervised signal of our denoising net. To achieve this, we use the non-rigid dynamic fusion pipeline proposed by [11], which is able to reconstruct complete and good quality geometries of dynamic scenes from single RGB-D camera. The captured scene could be static or dynamic and we do not impose any assumptions on the type of motions. Besides, the camera is allowed to move freely during the capture. The reconstructed geometry is well aligned with input color frames. To this end, we

first capture a sequence of synchronized RGB-D frames $\{D_t, C_t\}$. Then we run the non-rigid fusion pipeline [11] to produce a complete and improved mesh, and deform it using the estimated motion to each corresponding frame. Finally the target reference depth map $\{D_{ref,t}\}$ is generated by rasterization at each corresponding view point. Besides, we also produce a foreground mask $\{W_t\}$ using morphological filtering, which indicates the region of interest in the depth.

Content and novelty. Using the above method, we contribute a new dataset of human bodies, including color image, raw depths with real noises and the corresponding reference depths with sufficient quality. Our training dataset contains 36840 views of aligned RGB-D data along with high quality D_{ref} rendered from fused model, among which 11540 views are from structured light depth sensor and 25300 views are from time-of-flight depth sensor. Our validation dataset contains 4010 views. Training set contains human bodies with various clothes poses under different lighting conditions. Moreover, to verify how our method generalized to other scenes, objects such as furniture and toys are also included in the test set. Existing public datasets, *eg.* Face Warehouse, Biwi Kinect face and D3DFACS, lack geometry details, thus do not meet our requirement for surface refinement. ScanNet consists of a huge amount of 3D indoor scenes, but has no human body category. Our dataset fills the blank in human body surface reconstruction. Dataset and training code will be public available.

3.2 Depth Map Denoising

The denoising net $\mathcal{D}$ is trained to remove the sensor noise in depth map D_{in} given the reference depth map D_{ref}. Our denoising net architecture is inspired by DispNet[24] with skip connections and multi-scale predictions, as shown in Fig. 2. The denoising net consists of three parts: encoder, nonlinearity and decoder. The encoder aims to successively extract low-resolution high-dimensional features from D_{in}. To add nonlinearity to the network without performance degradation, several residual blocks with pre-activation are stacked sequentially between encoder and decoder part. The decoder part upsamples encoded feature maps to the original size, together with skip connections from the encoder part. These skip connections is useful to preserve geometry details in D_{in}. The whole denoising net adopts the residual learning strategy to extract the latent clean image from noisy observation. Not only does this direct pass set a good initialization, it turns out that residual learning is able to speed up the training process of deep CNN as well. Instead of the "unpooling + convolution" operation, our upsampling uses transpose convolution with trainable kernels. Note that the combination of bilinear up-sampling and transpose convolution in our upsampling pass help to inhibit checkerboard artifacts[41, 29]. Our denoising net is fully convolutional with receptive field up to 256. As a result, it is able to handle almost all types of consumer sensor inputs with different size.

The first loss for our denoising net is defined on the depth map itself. For example, per-pixel L1 and L2 loss on depth are used for our reconstruction term:

$$\ell_{rec}(D_{dn}, D_{ref}) = \|D_{dn} - D_{ref}\|_1 + \|D_{dn} - D_{ref}\|_2, \tag{1}$$

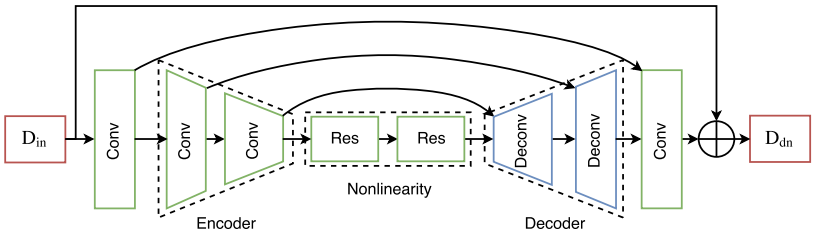

Fig. 2. The structure of our denoising net consists of encoder, nonlinear and decoder. There are three upsampling levels and one direct skip to keep captured value.

where $D_{dn} = \mathcal{D}(D_{in})$ is the output denoised depth map, and D_{ref} is the reference depth map. It is known that L2 and L1 loss may produce blurry results, however they accurately capture the low frequencies[18] which meets our purpose.

However, with only the depth reconstruction constraint, the high-frequency noise in small local patch could still remain after passing denoising net. To prevent this, we design a *normaldot* term to remove the high-frequency noise further. Specifically, this term is designed to constrain the normal direction of the denoised depth map to be consistent with the reference normal direction. We define the dot production of reference normal N^i_{ref} and tangential direction as the second loss term for our denoising net. Since each neighbouring depth point j ($j \in \mathcal{N}(i)$) could potentially define a 3D tangential direction,we sum over all possible directions, and the final normaldot term is formulated as:

$$\ell_{dot}(D_{dn}, N_{ref}) = \sum_i \sum_{j \in \mathcal{N}(i)} \left[< P^i - P^j, N^i_{ref} > \right]^2, \tag{2}$$

where P^i is the 3D coordinate of D^i_{dn}. This term explicitly drives the network to consider the dependence between neighboring pixels $\mathcal{N}(i)$, and to learn locally the joint distributions of the neighboring pixels. Therefore, the final loss function for training the denoising net is defined as:

$$\mathcal{L}_{dn}(D_{dn}, D_{ref}) = \lambda_{rec}\ell_{rec} + \lambda_{dot}\ell_{dot}, \tag{3}$$

where $\lambda_{rec}, \lambda_{dot}$ defines the strength of each loss term.

In order to get N_{ref} from the depth map D_{ref}, a *depth to normal* (d2n) layer is proposed. d2n layer is fully differentiable and has been employed several times in our end-to-end framework as shown in Figure 1.

3.3 Depth Map Refinement

Although denoising net is able to effectively remove the noises, the denoised depth map, even getting improved in low frequency, still lacks details compared with RGB images. To add high-frequency details to the denoised depth map, we adopt a relatively small fully convolutional network based on hypercolumn architecture[14, 33].

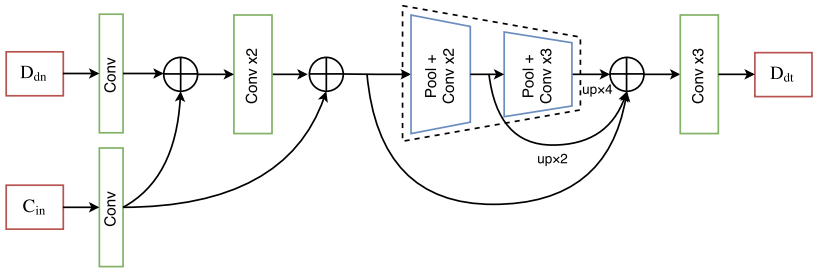

Fig. 3. Refinement net structure. The convolved feature maps from D_{dn} are complemented with the corresponding feature maps from C_{in} possessing the same resolution.

Denote the single channel intensity map of color image C_{in} as I. [4] The hypercolumn descriptor for a pixel is extracted by concatenating the features at its spatial location in several layers, from both D_{dn} and I of the corresponding color image with high-frequency details. We first combine the spectral features from D_{dn} and I, then fuse these features in the spatial domain by max-pooling and convolutional down-sampling, which end with multi-scale fused feature maps. The pooling and convolution operation after hypercolumn extraction constructs a new set of sub-bands by fusing the local features of other hypercolumns in the vicinity. This transfers fine structure from color map domain to depth map domain. Three post-fusion convolutional layers is introduced to learn a better channel coupling. *tanh* function is used as the last activation to limit the output to the same range of the input. In brief, high frequency features in the color image are extracted, and used as guidance, to extrude local detailed geometry from the denoised surfaces by the proposed refinement net shown in Fig. 3. As high frequency details are mainly inferred from small local patches, a shallow network with relative small reception field has enough capacity. Without post-processing as in other two-stage pipelines[37], our refinement net generates high-frequency details on depth map in a single forward pass.

Many SfS-based refinement approaches [44, 13] demonstrate that color images can be used to estimate the incident illumination, which is parameterized by the rendering process of an image. For Lambertian surface and low-frequency illumination, we can express the reflected irradiance B as the function of the surface normal N, the lighting condition $\boldsymbol{l}$ and the albedo R as follows:

$$B(\boldsymbol{l}, N, R) = R \sum_{b=1}^{9} l_b H_b(N), \tag{4}$$

where $H_b : \mathbb{R}^3 \mapsto \mathbb{R}$ is the basis function of spherical harmonics(SH) that takes unit surface normal N as input. $\boldsymbol{l} = [l_1, \cdots, l_9]^T$ are the nine 2nd order SH coefficients which represent the low-frequency scene illumination.

Based on Eq. 4, a per-pixel shading loss is designed. It penalizes both intensity and gradient of the difference value between the rendered image and the

[4] Intensity image I plays the same role as C_{in}. We study I for simplicity.

Fig. 4. Estimated albedo map and relighted uniform albedo using estimated lighting coefficients. The light source is above the head and the lighting estimation is accurate.

corresponding intensity image:

$$\ell_{sh}(N_{dt}, N_{ref}, I) = \|B(\boldsymbol{l}^*, N_{dt}, R) - I\|_2 + \lambda_g \|\nabla B(\boldsymbol{l}^*, N_{dt}, R) - \nabla I\|_2, \quad (5)$$

where N_{dt} represents the normal map of the regressed depth from the refinement net, λ_g is the weight to balance shading loss term, R is the albedo map estimated using Nestmeyer's "CNN + filter" method[25]. Then, the light coefficients $\boldsymbol{l}^*$ can be computed by solving the least squares problem:

$$\boldsymbol{l}^* = \arg\min_{\boldsymbol{l}} \|B(\boldsymbol{l}, N_{ref}, R) - I\|_2^2. \quad (6)$$

Here N_{ref} is calculated by the aforementioned d2n layer in section 3.2. To show the effectiveness of our estimated illumination, a per-pixel albedo image is calculated by $R_I = I / \sum_{b=1}^{9} l_b H_b(N_{ref})$, as shown in Figure 4. Note that pixels at grazing angles are excluded in the lighting estimation, as both shading and depth are unreliable in these regions. Additionally, to constrain the refined depth to be close to the reference depth map, a fidelity term is added:

$$\ell_{fid}(D_{dt}, D_{ref}) = \|D_{dt} - D_{ref}\|_2. \quad (7)$$

Furthermore, a smoothness term is added to regularize the refined depth. More specifically, we minimize the anisotropic total variation of the depth:

$$\ell_{smo}(D_{dt}) = \sum_{i,j} |D_{dt}^{i+1,j} - D_{dt}^{i,j}| + |D_{dt}^{i,j+1} - D_{dt}^{i,j}|. \quad (8)$$

With all the above terms, the final loss for our refinement net is expressed as:

$$\mathcal{L}_{dt}(D_{dt}, D_{ref}, I) = \lambda_{sh}\ell_{sh} + \lambda_{fid}\ell_{fid} + \lambda_{smo}\ell_{smo}, \quad (9)$$

where $\lambda_{sh}, \lambda_{fid}, \lambda_{smo}$ defines the strength of each loss term. The last two additional terms are necessary, because they constrain the output depth map to be smooth and also close to our reference depth, as the shading loss would not be able to constrain the low frequency component.

3.4 End-to-End Training

We train our denoising net and the refinement net jointly. To do so, we define total loss $\mathcal{L}_{total}$ as the sum of $\mathcal{L}_{dn}$ and $\mathcal{L}_{dt}$ of the cascaded networks. The denoising net is supervised by temporal fused reference depth map, and the refinement

CNN is trained in a novel unsupervised manner. By incorporating supervision signals in both the middle and the output of the network, we achieve a steady convergence during the training phase. In the forward pass, each batch of input depth maps is propagated through the denoising net, generating smoothed depth maps without noise patterns. First, reconstruction L1/L2 term and *normaldot* term are added to $\mathcal{L}_{total}$. Then, the denoised depth maps, together with the corresponding color images, are fed to our refinement net, which generates refined depth maps with high-frequency details. Shading, fidelity and smooth terms are added to $\mathcal{L}_{total}$. In the backward pass, the gradient of the loss $\mathcal{L}_{total}$ are back-propagated through both network. All the weights λ are fixed during training.

There are two types of consumer depth camera data in our training and validation set: structured light (K1) and time-of-flight (K2). We train the variants of our model for 15 epochs with batch-size 32 on a TitanX GPU, on K1/K2 dataset respectively. To augment our training set, each RGB-D map are randomly cropped, flipped and re-scaled to the resolution of 256×256. Considering that depth map is 2.5D in nature, the intrinsic matrix should be changed accordingly during data augmentation. This enables the network to learn more object-independent statistics. For efficiency, we implement our d2n layer as a single CUDA layer. We choose*Adam* optimizer to compute gradients, with 0.9 and 0.999 exponential decay rate for the 1st and 2nd moment estimates. Base learning-rate is set to 0.001. All convolution weights are initialized by Xavier algorithm, and weight decay is used for regularization.

4 Experiments

4.1 Evaluation

In this section, we evaluate the effectiveness of our cascade depth denoising and refinement framework, and analyze the contribution from each loss term. To the best of our knowledge, there is no public dataset for human body that contains raw and ground-truth depth maps with rich details from consumer depth cameras. We thus will evaluate the performance of our method on our own validation set, and also evaluate on other objects other than human to test the generalization ability of our network, which can be seen in Figure 5. One can see that although refined in an unsupervised manner, our results are comparable to the high quality fused depth map [11] obtained using consumer depth camera only, and preserve thin structures such as fingers and folds in clothes better.

4.2 Ablation Study

The Role of Cascade CNN. To verify the necessity of our cascade CNNs, we replace our denoising net by a traditional preprocessing procedure, *eg.* bilateral filter, and still keep the refinement net to refine the filtered depth. We call this two-stage method as "Base+Ours refine", and it is trained from scratch with shading and fidelity loss. As we can see in the middle of Figure 6, "Base+Ours

refine" is not able to preserve distinctive structures of clothes in the presence of widespread structured noise. Unwanted high frequency noise leads to inaccurate estimation of illuminance, therefore shading loss term will keep fluctuating during training. This training process will end up with non-optimal model parameters. However, in our cascade design, denoising net sets a good initialization for refinement net and achieves better result.

Supervision of Refinement Net. For our refinement net, there are two choices for regularization depth map in fidelity loss formulation, using reference depth map D_{ref} or the denoised depth map D_{dn}. When using only output of denoising net D_{dn} in an unsupervised manner, scene illumination is also estimated using D_{dn}. We denote this unsupervised framework as "Ours unsupervised". Output of these two choices are shown in Fig. 7. In the unsupervised case, refinement net could produce reasonable result, but D_{dt} may stray from input.

4.3 Comparison With Other Methods

Compared with other non-data-driven methods, deep neural networks allow us to optimize non-linear loss and to add data-driven regularization, while keeping the inference time constant. Besides, our method would also benefit from the progress of deep-net-specific hardware, making it more promising both in terms of run-time and quality. Fig. 8 shows examples of the qualitative comparison of different methods for depth map enhancement. Our method outperforms other methods by capturing better structure of the geometry with cleaner and high-fidelity geometric details.

Quantitative Comparison. To evaluate our method and compare with other methods quantitatively, we need a dataset with ground truth depth map. Currently, multi-view stereo methods and laser scanner are able to capture static scene with high resolution and quality. We thus obtain ground truth depth value by multi-view stereo [32](for K1) and Mantis Visions F6 laser scanner(for K2). Meanwhile, we simultaneously capture the RGB-D sequence of the same static scene by a consumer depth camera to evaluate our enhancement method. The size of validation set is limited due to the high scan cost. Therefore, we also contribute a larger validation set labeled with the near-ground-truth depth obtained using mentioned method in 3.1. We evaluate the error of the enhanced depth map using our method and other methods on both validation set. Ground truth densely scanned 3D model is rescaled and aligned with our 3D model, which is recovered from the network output, using iterative closest point(ICP)[5]. Then the root mean squared error(RMSE) and the mean absolute error (MAE) between these two point clouds are calculated in Euclidean space. We have trained two sets of models on K1 data and K2 data respectively. Quantitative comparison with other methods are summarized in Table 2 and Table 1 for two types of data and model. Results shows that our method performs the best result in terms of both metrics on the validation set.

Runtime Performance. At test time, our whole processing procedure includes data pre-processing and cascade CNN predicting. The listed preprocessing steps include: depth-to-color alignment, dilating and eroding of raw depth map

to fill some holes, and resampling color image if needed. Regressing all depth map pixels using a TitanX Pascal graphics card is fast and takes only 10.8ms (256×256 input) or 20.4ms (640×480 input). Same test is run for 182.56ms (256×256 input) or for 265.8ms (640×480 input) per frame on an Intel Core i7-6900K CPU with 3.20GHz(64G Ram). It is worth mentioning that without denoising CNN, a variant of our method, *ie.* "Base+Ours refine" reaches a speed of 9.6ms per frame for 640×480 sized inputs.

Fig. 5. Qualitative results on validation set. From left to right in each row: aligned RGB image, raw depth map, output of our denoising net D_{dn} and output of our refinement net D_{dt} with added high-frequency details. D_{dn} captures the low-dimensional geometry without noise, D_{dt} shows fine-grained details. Although trained on human body dataset, our model also produce high-quality depth map on general objects in arbitrary scenes, *eg.* the backpack sequence. The last row shows typical failure case of our network, which may suffer from non-Lambertian reflectance.

Table 1. Sequence average score in terms of RMSE on our K2 validation set obtained by laser scanner. Our method achieves the best result on the validation set

Method	seq.1	seq.2	seq.3	seq.4	seq.5
Wu *et al.* [44]	27.60	22.19	21.34	22.41	25.67
Ours D_{dn}	19.03	19.25	18.49	18.37	18.76
Ours D_{dt}	18.97	19.41	18.38	18.50	18.61

4.4 Limitation

Similar to other real-time methods, our real-time DDRNet considers simplified light transport model. This simplification is effective but will impose intensity image's texture on depth map. With the learning-based framework, the texture-copy artifacts can be alleviated due to the fact that our neural network can balance the fidelity and shading loss term during training. Another limitation of the simplified reflectance model comes with non-diffuse surface assumption. As we only consider second order spherical harmonics representation, non-diffuse surfaces are still challenging for our method.

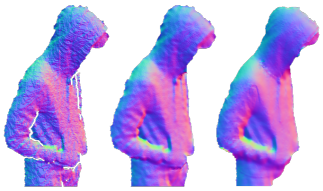

Fig. 6. Left: normal map of D_{in}. **Middle**: Base+Ours refine, bilateral filter can't remove wavelet noise, refinement result suffers from high-frequency. **Right**: Ours.

Fig. 7. left: C_{in} and D_{in}. **middle**: Ours unsupervised, output depth does not match input value in stripes area in the cloth. **right**: Ours with more reliable result.

Table 2. Quantitative comparison results on K1 validation set. The error metrics are in *mm* and are only computed on pixels where output depth map has value.

Method	Near-GT set		GT set	
	MAE	RMSE	MAE	RMSE
He *et al.* [15]	46.5	14.7	41.1	15.2
Wu *et al.* [44]	14.5	4.3	15.7	4.4
Ours	10.9	4.1	11.0	3.6
Base+Ours refine	15.7	4.1	15.8	4.4
Ours unsupervised	16.1	5.2	14.9	5.5

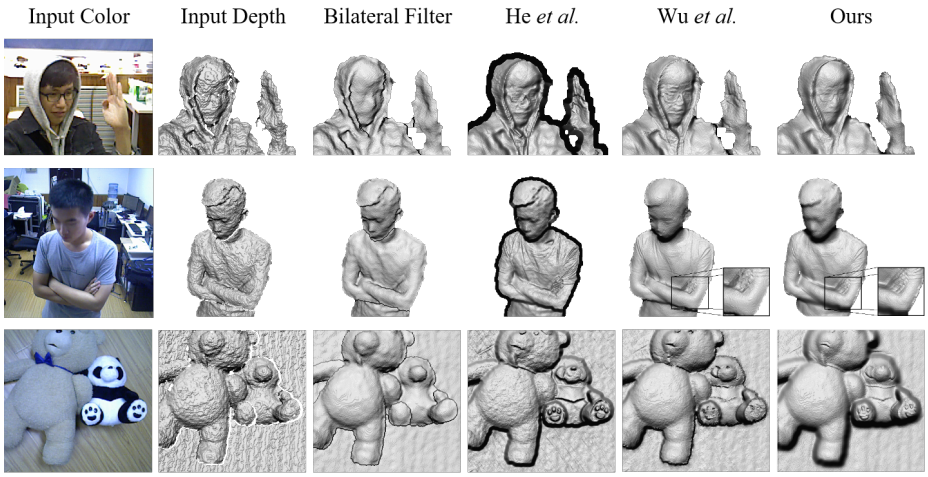

Fig. 8. Comparison of color-assisted depth map enhancement between bilateral filter, He *et al.* [15], Wu *et al.* [44] and our method. The closeup of the finger region demonstrates the effectiveness of unsupervised shading term in our refinement net loss.

5 Applications

It is known that Real-time single frame depth enhancement is applicable for low-latency system without temporal accumulation. We compare the impact of using depth stream refined by our method with that using raw depth stream captured by Kinect camera on Dynamic Fusion [11] and DoubleFusion[47]. The temporal window in fusion systems would smooth out noise, but it will also wipe out high-frequency details. The time in TSDF fusion blocks the whole system from tracking detailed motions. However, our method can shorten the smooth window and provide timely update of fast changing surface details (*eg.* deformation of clothes and body gestures), as shown in red circles in Fig. 9 and the supplementary video. Moreover, a per-frame improved depth could help general tracking and recognition tasks over dynamic scenes(*eg.* in interactive scenarios).

Fig. 9. Application on DynamicFusion(left) and DoubleFusion(right) using our enhanced depth stream. **Left**: color image, **Middle**: fused geometry using raw depth stream, **Right**: "instant" geometry using our refined depth stream.

6 Conclusion

We presented the first end-to-end trainable network for depth map denoising and refinement of RGB-D data captured with consumer depth cameras. We proposed a near-groundtruth training data generation pipeline, based on the depth fusion techniques. Enabled by the separation of low/high frequency parts in network design, as well as the collected fusion data, our cascaded CNNs achieves state-of-the-art result in real-time. Compared with available methods, our method achieved higher quality reconstruction in terms of both low dimensional geometry and high frequency detailed components, which leads to superior performance quantitatively and qualitatively. Finally, with the popularity of integrating depth sensors into cellphones, we believe that our deep-net-specific algorithm is able to run on these portable devices for various quantitative measurement and qualitative visualization applications.

References

1. Barron, J.T., Malik, J.: Intrinsic scene properties from a single rgb-d image. In: Proc. CVPR. pp. 17–24. IEEE (2013)
2. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. Tech. rep., EECS, UC Berkeley (May 2013)
3. Beeler, T., Bickel, B., Beardsley, P., Sumner, B., Gross, M.: High-quality single-shot capture of facial geometry. Proc. SIGGRAPH **29**(3) (2010)
4. Beeler, T., Bradley, D., Zimmer, H., Gross, M.H.: Improved reconstruction of deforming surfaces by cancelling ambient occlusion pp. 30–43 (2012)
5. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Robotics-DL tentative. pp. 586–606. International Society for Optics and Photonics (1992)
6. Chan, D., Buisman, H., Theobalt, C., Thrun, S.: A noise-aware filter for real-time depth upsampling. In: ECCV Workshop on multi-camera & multi-modal sensor fusion (2008)
7. Chen, J., Bautembach, D., Izadi, S.: Scalable real-time volumetric surface reconstruction. ACM Trans. Graph. **32**(4), 113:1–113:16 (Jul 2013). https://doi.org/10.1145/2461912.2461940, http://doi.acm.org/10.1145/2461912.2461940
8. Cui, Y., Schuon, S., Thrun, S., Stricker, D., Theobalt, C.: Algorithms for 3d shape scanning with a depth camera. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(5), 1039–1050 (May 2013). https://doi.org/10.1109/TPAMI.2012.190
9. Diebel, J., Thrun, S.: An application of markov random fields to range sensing. In: Proceedings of the 18th International Conference on Neural Information Processing Systems. pp. 291–298. NIPS'05, MIT Press, Cambridge, MA, USA (2005), http://dl.acm.org/citation.cfm?id=2976248.2976285
10. Dolson, J., Baek, J., Plagemann, C., Thrun, S.: Upsampling range data in dynamic environments. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1141–1148 (June 2010). https://doi.org/10.1109/CVPR.2010.5540086
11. Guo, K., Xu, F., Yu, T., Liu, X., Dai, Q., Liu, Y.: Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. ACM

Trans. Graph. **36**(3), 32:1–32:13 (Jun 2017). https://doi.org/10.1145/3083722, http://doi.acm.org/10.1145/3083722
12. Han, Y., Lee, J.Y., Kweon, I.S.: High quality shape from a single rgb-d image under uncalibrated natural illumination. In: Proc. ICCV (2013)
13. Han, Y., Lee, J.Y., Kweon, I.S.: High quality shape from a single rgb-d image under uncalibrated natural illumination. In: IEEE International Conference on Computer Vision. pp. 1617–1624 (2013)
14. Hariharan, B., Arbelaez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization pp. 447–456 (2014)
15. He, K., Sun, J., Tang, X.: Guided image filtering. IEEE Transactions on Pattern Analysis & Machine Intelligence **35**(6), 1397–1409 (2013)
16. Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. The International Journal of Robotics Research **31**(5), 647–663 (2012). https://doi.org/10.1177/0278364911434148, https://doi.org/10.1177/0278364911434148
17. Horn, B.K.: Obtaining shape from shading information. The psychology of computer vision pp. 115–155 (1975)
18. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks (2016)
19. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A.: Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In: Proc. UIST. pp. 559–568. ACM (2011)
20. Kajiya, J.T.: The rendering equation. In: Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques. pp. 143–150. SIGGRAPH '86, ACM, New York, NY, USA (1986). https://doi.org/10.1145/15922.15902, http://doi.acm.org/10.1145/15922.15902
21. Khan, N., Tran, L., Tappen, M.: Training many-parameter shape-from-shading models using a surface database. In: Proc. ICCV Workshop (2009)
22. Kopf, J., Cohen, M.F., Lischinski, D., Uyttendaele, M.: Joint bilateral upsampling. ACM Trans. Graph. **26**(3) (Jul 2007). https://doi.org/10.1145/1276377.1276497, http://doi.acm.org/10.1145/1276377.1276497
23. Lindner, M., Kolb, A., Hartmann, K.: Data-fusion of pmd-based distance-information and high-resolution rgb-images. In: 2007 International Symposium on Signals, Circuits and Systems. vol. 1, pp. 1–4 (July 2007). https://doi.org/10.1109/ISSCS.2007.4292666
24. Mayer, N., Ilg, E., Husser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Computer Vision and Pattern Recognition. pp. 4040–4048 (2016)
25. Nestmeyer, T., Gehler, P.V.: Reflectance adaptive filtering improves intrinsic image estimation. In: CVPR. pp. 1771–1780 (2017)
26. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 343–352 (June 2015). https://doi.org/10.1109/CVPR.2015.7298631
27. Newcombe, R.A., Izadi, S., et al.: Kinectfusion: Real-time dense surface mapping and tracking. In: Mixed and augmented reality (ISMAR), IEEE international symposium on. pp. 127–136 (2011)

28. Nießner, M., Zollhöfer, M., Izadi, S., Stamminger, M.: Real-time 3d reconstruction at scale using voxel hashing. ACM Transactions on Graphics (TOG) **32**(6), 169 (2013)
29. Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. Distill (2016). https://doi.org/10.23915/distill.00003, http://distill.pub/2016/deconv-checkerboard
30. Or El, R., Rosman, G., Wetzler, A., Kimmel, R., Bruckstein, A.M.: Rgbd-fusion: Real-time high precision depth recovery. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
31. Park, J., Kim, H., Tai, Y.W., Brown, M.S., Kweon, I.: High quality depth map upsampling for 3d-tof cameras. In: 2011 International Conference on Computer Vision. pp. 1623–1630 (Nov 2011). https://doi.org/10.1109/ICCV.2011.6126423
32. RealityCapture: https://www.capturingreality.com/ (2017)
33. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image. In: CVPR (2017)
34. Richardt, C., Stoll, C., Dodgson, N.A., Seidel, H.P., Theobalt, C.: Coherent spatiotemporal filtering, upsampling and rendering of rgbz videos. Computer Graphics Forum **31**(2pt1), 247–256 (2012). https://doi.org/10.1111/j.1467-8659.2012.03003.x, http://dx.doi.org/10.1111/j.1467-8659.2012.03003.x
35. Riegler, G., Ulusoy, A.O., Bischof, H., Geiger, A.: Octnetfusion: Learning depth fusion from data. CoRR **abs/1704.01047** (2017), http://arxiv.org/abs/1704.01047
36. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. CoRR **abs/1505.04597** (2015), http://arxiv.org/abs/1505.04597
37. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation (2017)
38. Tewari, A., Zollhöfer, M., Kim, H., Garrido, P., Bernard, F., Pérez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. CoRR **abs/1703.10580** (2017), http://arxiv.org/abs/1703.10580
39. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from Synthetic Humans. In: CVPR (2017)
40. Wei, G., Hirzinger, G.: Learning shape from shading by a multilayer network. IEEE Transactions on Neural Networks **7**(4), 985–995 (1996)
41. Wojna, Z., Ferrari, V., Guadarrama, S., Silberman, N., Chen, L.C., Fathi, A., Uijlings, J.: The devil is in the decoder (2017)
42. Wu, C., Stoll, C., Valgaerts, L., Theobalt, C.: On-set performance capture of multiple actors with a stereo camera. ACM Transactions on Graphics (TOG) **32**(6), 161 (2013)
43. Wu, C., Varanasi, K., Liu, Y., Seidel, H., Theobalt, C.: Shading-based dynamic shape refinement from multi-view video under general illumination pp. 1108–1115 (2011)
44. Wu, C., Zollhöfer, M., Nießner, M., Stamminger, M., Izadi, S., Theobalt, C.: Real-time shading-based refinement for consumer depth cameras. ACM Transactions on Graphics (TOG) **33**(6), 200 (2014)
45. Yang, Q., Yang, R., Davis, J., Nister, D.: Spatial-depth super resolution for range images. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (June 2007). https://doi.org/10.1109/CVPR.2007.383211
46. Yu, L., Yeung, S., Tai, Y., Lin, S.: Shading-based shape refinement of rgb-d images pp. 1415–1422 (2013)

47. Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y.: Doublefusion: Real-time capture of human performance with inner body shape from a depth sensor. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
48. Zhang, Z., Tsa, P.S., Cryer, J.E., Shah, M.: Shape from shading: A survey. IEEE PAMI **21**(8), 690–706 (1999)
49. Zhu, J., Wang, L., Yang, R., Davis, J.: Fusion of time-of-flight depth and stereo for high accuracy depth maps. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (June 2008). https://doi.org/10.1109/CVPR.2008.4587761