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, [2]Facebook Reality Labs, [3]Google Inc

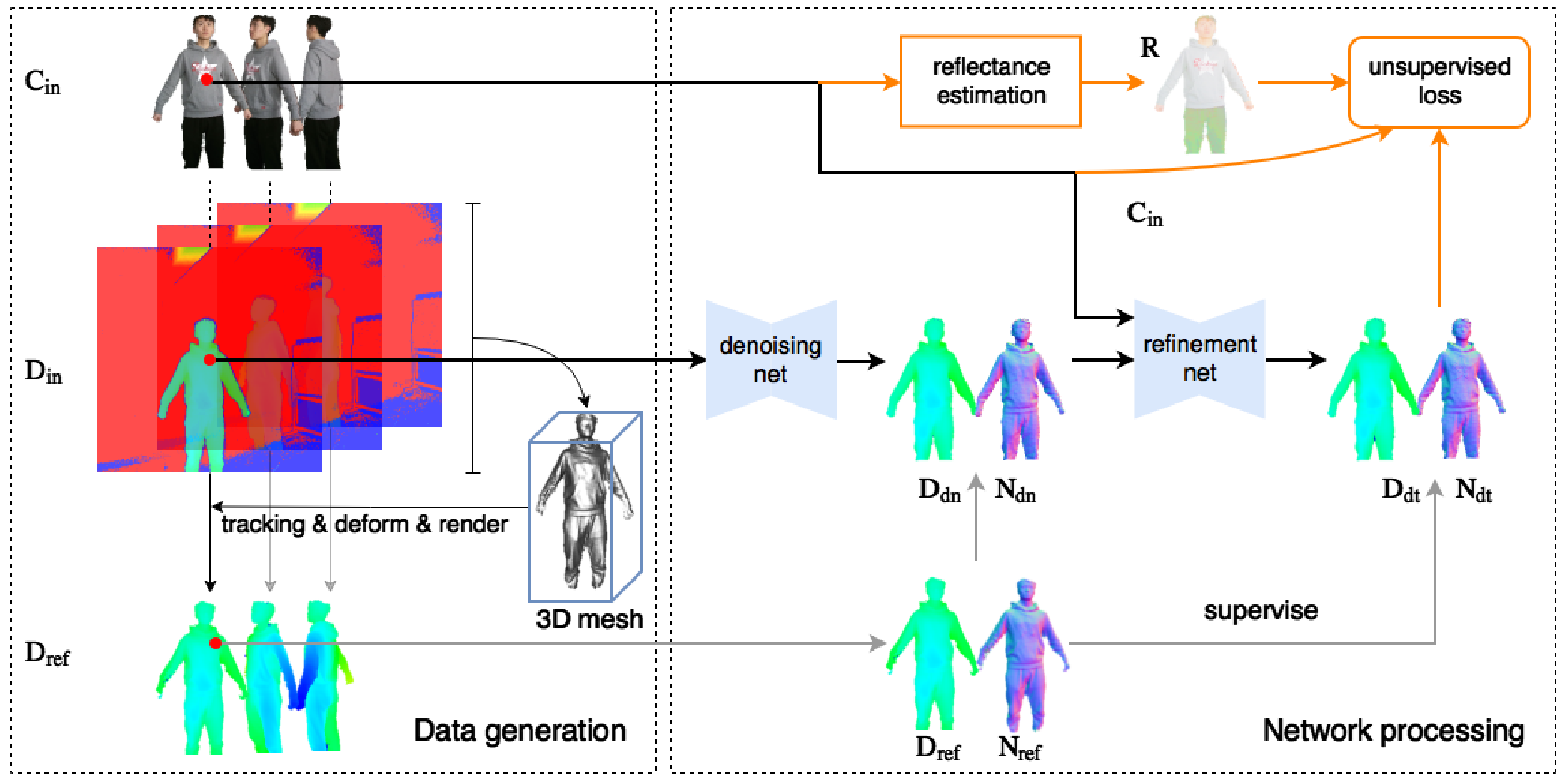
Motivation

Consumer depth sensors suffer from heavy noises which limit their applications. Neural networks are capable of modeling complex functions and meet the real-time requirement. By leveraging single frame depth map and the accompanying high quality color image through a joint training strategy, we can achieve enhanced depth map.

Challenge

The challenge is to generate details and keep depth merically accurate. To lift the depth quality in low frequency and high frequency simultaneously, we propose a cascaded CNN structure to perform depth image denoising and refinement. To transfer surface details, we design a generative process together with an unsupervised loss based on rendering equation.

Pipeline *and* Dataset



The proposed pipeline features our novelties in training data creation and cascaded archietecture design. **Unlike synthesized data**, to capture noise pattern from real scenes, we employ multi-frame depth fusion technique to generate data. **We formulate the problem** into two regression tasks specific in different frequency domains. The cascaded CNNs learn from supervised depth data and unsupervised shading cues in an end-to-end way.

Qualitative Result

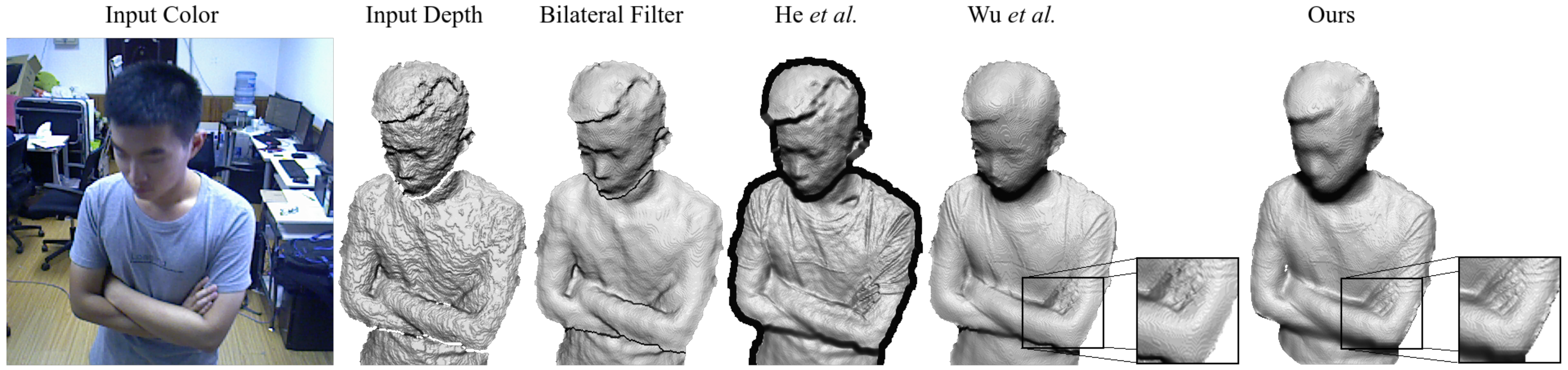


Fig. 1. Comparison of color-assisted depth map enhancement between bilateral filter, He *et al.* , Wu *et al.* and our method. The closeup of the finger demonstrates the effectiveness of unsupervised shading term.

Acknowledgement

Supported by the National key foundation for exploring scientific instrument of China No.2013YQ140517, and the National NSF of China grant No.61522111, No.61531014, No.61671268 and No.61727808.

Denoising Method

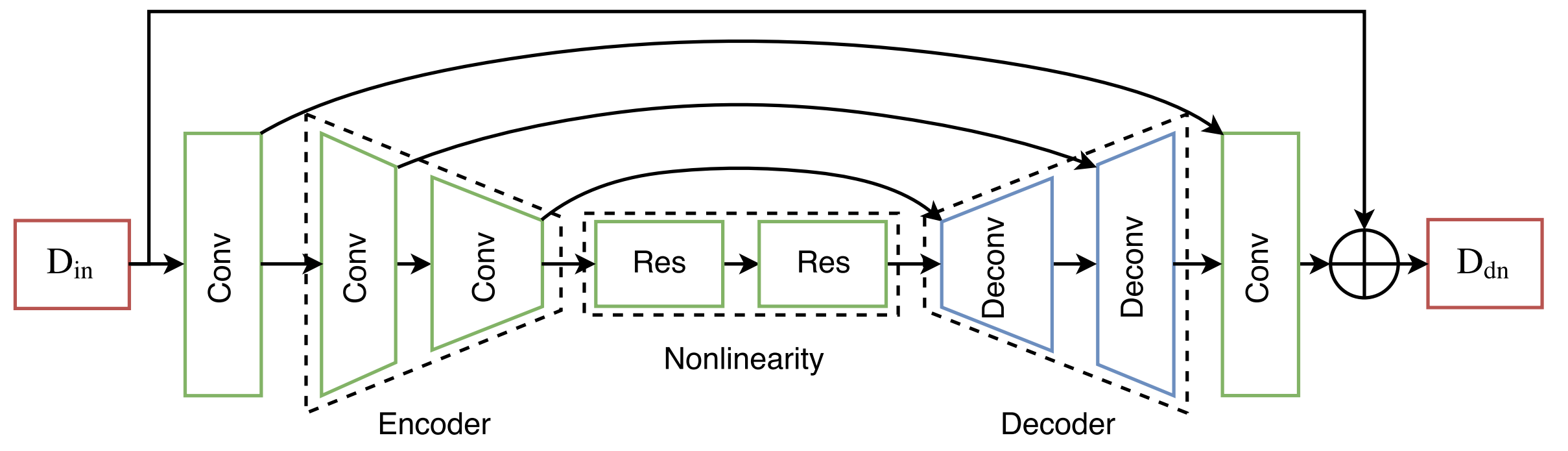


Fig. 2. The structure of our denoising net. We adopt residual learning strategy to initialize denoising net and resolve vanishing gradients.

Loss consists of 2 parts, reconstruction term constrains depth and normaldot term constrains normal direction, which remove noise in local patch.

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

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

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

Refinement Method

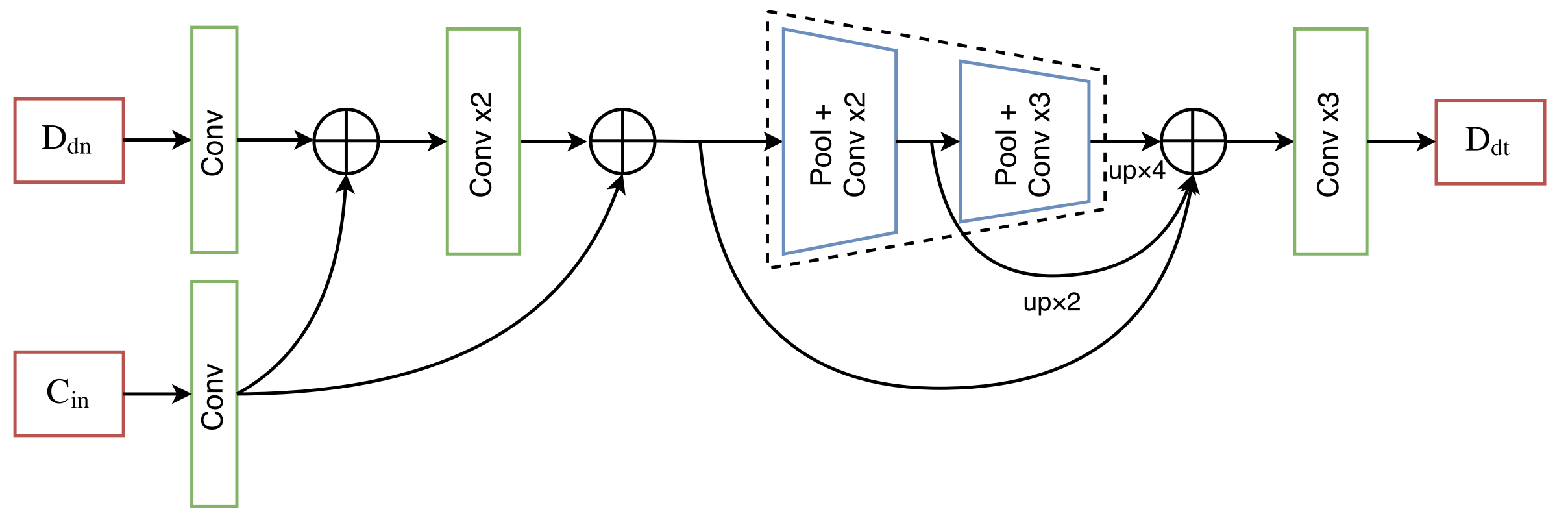


Fig. 3. Refinement net structure. The convolved feature maps from depth map are complemented with the corresponding feature maps from RGB.

Reflected Irradiance is a function of normal, lighting and albedo. $B(\boldsymbol{l}, N, R) = R\sum_{b=1}^{9} l_b H_b(N)$

Light coefficients are computed by solving least square problem. Albedos are estimated from RGB.

Loss consists of 2 parts, shading term extrudes surface details and fidelity term keeps depth faithful.

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

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

Quantitative Result

Table 1. Average score of depth and normal error and on our ToF validation set.

Method	Depth difference					Normal difference				
	seq. 1	seq. 2	seq. 3	seq. 4	seq. 5	Mean ↓	Median ↓	3.0↑	5.0↑	10.0↑
Wu *et al.*	27.60	22.19	21.34	22.41	25.67	11.20	5.02	29.81	50.24	76.62
Or-El *et al.*	27.14	25.42	22.89	21.31	26.08	10.03	4.12	35.43	56.57	79.99
Ours D_{dn}	19.03	**19.25**	18.49	**18.37**	18.76	**9.36**	**3.40**	**45.33**	**66.79**	**84.69**
Ours D_{dt}	**18.97**	19.41	**18.38**	18.50	**18.61**	9.55	3.54	43.77	64.98	83.69

The ground truth 3D model is obtained from laser scanner. After reprojection, rescaling and ICP, RMSE and MAE are calculated for depth map. We also report the angular difference of normals in degree.

Implementation Details

Our training set contains 11540 views of structured light data and 25300 views of time-of-flight data.To adapt to different intrinsic parameters, we augment the intrinsic matrix and its 2.5D depth map accordingly. The forward pass takes only 20.4ms for 640 × 480 input on TitanX. Get the model and code at: `github.com/neycyanshi/DDRNet`

Conclusion

Thanks to the well decoupling of low and high frequency information, as well as the dataset generated from real scenes, our work produces clean results with sufficient geometry details. With the popularity of integrating depth sensors into cellphones, our deep-net-specific algorithm is able to run on these portable devices for various applications.