# A Deep Primal-Dual Network for Guided Depth Super-Resolution

Gernot Riegler, David Ferstl, Matthias Rüther, Horst Bischof
{riegler,ferstl,ruether,bischof}@icg.tugraz.at

Institute for Computer Graphics and Vision
Graz University of Technology
Austria

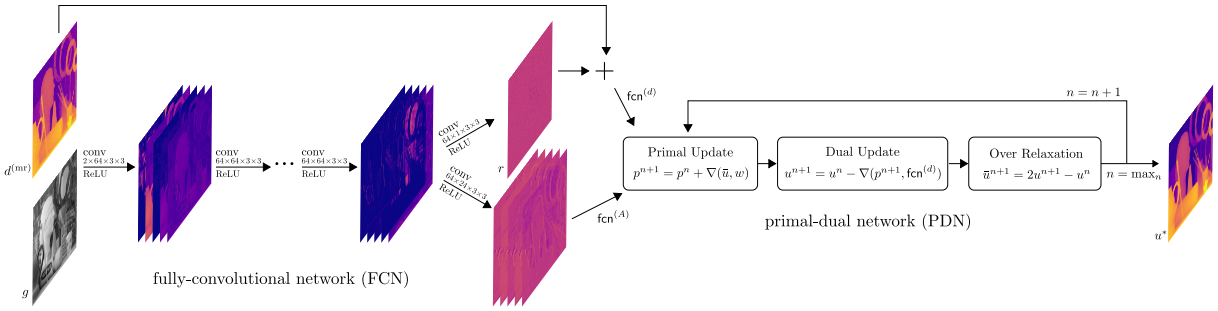


Figure 1: Our *deep primal-dual network* consists of two main parts. A fully-convolutional network (*FCN*) that computes a first HR estimate and weighting coefficients. These outputs are then feed to our primal-dual network (*PDN*), where we unroll the optimization steps of a non-local variational method that incorporates prior knowledge about the data modalities. The complete network is trained end-to-end.

Sensors that measure pixel-wise depth information have become increasingly popular since they are available in the consumer market and enabled a broad range of novel computer vision applications, for example in robot navigation, human pose estimation, and hand pose estimation. Despite their success, these sensors suffer from a low spatial resolution and depth noise due to physical limitations of the measurement principles.

In this paper we present a novel method to increase the spatial and lateral resolution of noisy depth images. For this purpose, we combine a deep fully convolutional network (*FCN*) with a non-local variational method in a *deep primal-dual network* (see Fig. 1) and extend our work presented in [6]. The input to our method is a low resolution, noisy depth map $d^{(\text{lr})}$ and a high-resolution intensity image $g$ that is used as guidance in the upsampling process. This guidance image is essential for higher upsampling factors, as we show in our experiments. The input of the fully convolutional network is a upscaled low resolution depth map using bilinear interpolation. The network is trained to compute only the residual (high frequency parts) to the upscaled low resolution input. Further, the second output of the *FCN* are non-local weighting terms, which are utilized in the subsequent primal-dual network (*PDN*) as weighting coefficients and correspond to discontinuities in the high resolution depth.

In the primal-dual network we compute the optimizer $u_k^*$ of a variational energy functional given in (1) by unrolling the computation steps of the first-order primal-dual scheme by [1].

$$u_k^* = \arg\min_u \lambda D(u, \text{fcn}_{s_k}^{(d)}) + R(u, \text{fcn}_{s_k}^{(A)}). \tag{1}$$

$D$ is the data term that penalizes the deviation from the initial solution, $R$ is the regularization term that encodes smoothness assumptions, and $\lambda$ is a trade-off parameter. We evaluate in this work several popular choices of regularization terms, which are especially suited for depth data and found that a non-local Huber regularization, as given by

$$R(u) = \int_\Omega \int_{\mathcal{N}(x)} w(x,y) |u(x) - u(y)|_\varepsilon \, \mathrm{d}x \, \mathrm{d}y, \tag{2}$$

in combination with a $\ell_2$ data term yields the best trade-off between accuracy and computational requirements. $\Omega$ is the image domain, $\mathcal{N}(x)$ defines the neighborhood of $x$ in which the regularization term should be evaluated, $|\cdot|_\varepsilon$ is the Huber norm, and $w$ are weighting coefficients derived from the fully convolutional network $\text{fcn}^{(A)}$. The benefit of unrolling the optimization algorithm in the network are that we can learn all parameters of the variational method, as well as, all hyper-parameter of the optimization scheme itself. Further, the fully-convolutional network adapts in the joint training to the subsequent primal-dual network.

The training of such a deep network requires of course a large training set for supervision. Therefore, we generate high-quality depth maps and corresponding color images with a physically based renderer in large quantities. The training procedure is then two-fold: First, we pre-train the fully-convolutional network and subsequently train the complete model end-to-end using an Euclidean loss, as it relates to our evaluation metric.

In our experimental evaluation we show the influence of the energy functional and the non-local neighborhood size on the performance of our method. Further, we compare our method on two standard benchmarks for depth super-resolution to other recent approaches: On the noisy Middlebury images as proposed by [5] and on the realistic ToFMark dataset [2]. For an excerpt of our quantitative and qualitative results see Tab. 1 and Fig. 2, respectively. With this novel combination we are able create visually appealing results and outperform state-of-the-art on both datasets.

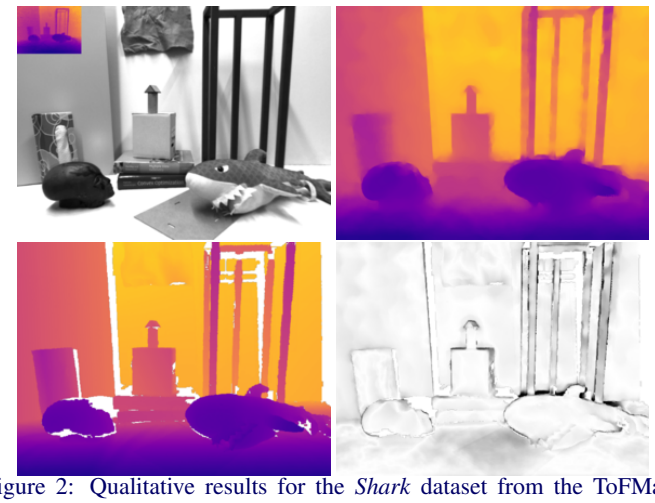

Figure 2: Qualitative results for the *Shark* dataset from the ToFMark benchmark [2]. The first row shows the input and the output of our method, respectively. In the second row we present the ground-truth depth and the error of our method.

| | Books | Devil | Shark |
|---|---|---|---|
| NN | 30.46 | 27.53 | 38.21 |
| Bilinear | 29.11 | 25.34 | 36.34 |
| Kopf *et al.* [4] | 27.82 | 24.30 | 34.79 |
| He *et al.* [3] | 27.11 | 23.45 | 33.26 |
| Ferstl *et al.* [2] | 24.00 | 23.19 | 29.89 |
| FCN-PDN ($d^{(\text{mr})}$ & $g$) | 23.74 | 20.47 | 28.81 |

Table 1: Quantitative results on the ToFMark benchmark [2].

[1] Antonin Chambolle and Thomas Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40 (1):120–145, 2011.

[2] David Ferstl, Christian Reinbacher, René Ranftl, Matthias Rüther, and Horst Bischof. Image Guided Depth Upsampling using Anisotropic Total Generalized Variation. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[3] Kaiming He, Jian Sun, and Xiaoou Tang. Guided Image Filtering. In *European Conference on Computer Vision (ECCV)*, 2010.

[4] Johannes Kopf, Michael F. Cohen, Dani Lischinski, and Matthew Uyttendaele. Joint Bilateral Upsampling. *ACM Transactions on Graphics (TOG)*, 26(3):96, 2007.

[5] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S. Brown, and In-So Kweon. High Quality Depth Map Upsampling for 3D-TOF Cameras. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[6] Gernot Riegler, Matthias Rüther, and Horst Bischof. ATGV-Net: Accurate Depth Super-Resolution. In *European Conference on Computer Vision (ECCV)*, 2016.

**Acknowledgment:** This work was supported by the Austrian Research Promotion Agency project TOFUSION (FIT-IT Bridge program)