Depth Restoration via Joint Training of a Global Regression Model and CNNs

Gernot Riegler*[1]
riegler@icg.tugraz.at
René Ranftl*[1]
ranftl@icg.tugraz.at
Matthias Rüther[1]
ruether@icg.tugraz.at
Thomas Pock[1,2]
pock@icg.tugraz.at
Horst Bischof[1]
bischof@icg.tugraz.at

[1] Institute for Computer Graphics and Vision
Graz University of Technology
Graz, Austria

[2] Digital Safety & Security Department
Austrian Institute of Technology
Vienna, Austria

Abstract

Denoising and upscaling of depth maps is a fundamental post-processing step for handling the output of depth sensors, since many applications that rely on depth data require accurate estimates to reach optimal accuracy. Adapting methods for denoising and upscaling to specific types of depth sensors is a cumbersome and error-prone task due to their complex noise characteristics. In this work we propose a model for denoising and upscaling of depth maps that adapts to the characteristics of a given sensor in a data-driven manner. We introduce a non-local Global Regression Model which models the inherent smoothness of depth maps. The Global Regression Model is parametrized by a Convolutional Neural Network, which is able to extract a rich set of features from the available input data. The structure of the model enables a complex parametrization, which can be jointly learned end-to-end and eliminates the need to explicitly model the signal formation process and the noise characteristics of a given sensor. Our experiments show that the proposed approach outperforms state-of-the-art methods, is efficient to compute and can be trained in a fully automatic way.

1 Introduction

The increasing availability of cheap consumer depth sensors enables a multitude of novel applications, for example in gaming and gesture control. Depth sensors such as the Microsoft Kinect have found their way into the mass market and Time of Flight (ToF) cameras, a type of sensor that measures depth using the runtime of light, have recently become more and more popular. Due to physical and manufacturing constraints, many consumer depth sensors, and especially ToF cameras, provide noisy and low-resolution output. In order to generate useful depth estimates, the data provided by the sensor is typically subject to post-processing steps such as denoising and upscaling.

*Authors contributed equally

Pages 58.1-58.12
DOI: https://dx.doi.org/10.5244/C.29.58

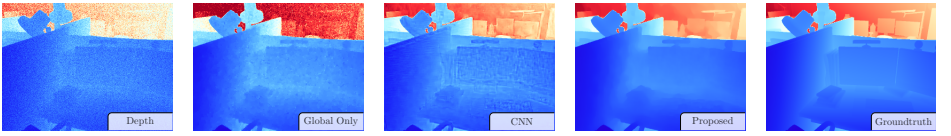


Figure 1: Our proposed method that combines a GRM with a CNN achieves better results than the individual components (best viewed in color and zoomed in).

Methods for denoising and upscaling of signals, be it natural images, or depth data, crucially rely on an accurate model of the observed noise. This is due to the fact that denoising and upscaling are ill-posed problems, which require an in-depth understanding of the signal formation process in order to yield accurate solutions.

The noise characteristics of depth estimates strongly vary with the type of the sensor used for acquisition. Depth estimates from structured light sensors, such as the Kinect v1 will show different characteristic noise [23] compared to a ToF sensor [10], which is for example integrated in the Kinect v2. Moreover, the noise in the estimated depth maps shows a complex dependency on the depth and the form of the imaged surface. As a result it is in practice difficult to accurately model the sensor noise.

We propose in this work a novel approach to circumvent the need to explicitly model the noise characteristics of a depth sensor, if accurate reference data can be acquired. This is in fact often possible: Manufacturers can either build or buy accurate depth sensors in order to calibrate low-cost sensors before shipping, or use several measurements from a depth sensor to build a high-quality model of a scene in an offline step [14]. Our approach utilizes Convolutional Neural Networks (CNNs) [19] to parametrize a Global Regression Model (GRM), which is motivated by non-local variational restoration models [2, 13]. We show how the parameters of the CNN as well as the parameters of the GRM can be learned using a joint end-to-end training procedure, recently proposed in the context of figure-ground segmentation [30], for the task of depth restoration. This training procedure results in a model that combines the discriminative power of CNNs with the strong regularization properties of a global model in a common framework.

Figure 1 shows an exemplar denoising result of the proposed approach and compares it to the result of its individual parts, *i.e.* the global model without CNN parametrization, as well as a plain CNN. We can observe that the proposed joint system performs better than its individual components.

2 Related Work

There exists a large body of work discussing denoising of natural images. Among the most successful and widely used are methods based on collaborative filtering [6] or sparse coding [9]. These methods implicitly assume a specific noise type, which is typically an additive Gaussian distribution. This assumption is reasonable for natural images, but not necessarily for other types of images, such as depth maps. However, adapting these models to different noise distributions such as the noise encountered in SAR imagery [27] is challenging. Moreover, these methods are modeled to the intrinsic statistics of natural images, which are different from the statistics of depth data.

In contrast, variational models based on Total Variation like the famous ROF [31], TV-L1 [5, 24], and in particular TGV-L2 [3] are well suited for depth data since their prior

assumption fits to the smooth and textureless characteristics of depth maps. However, those methods have the drawback that they model a specific noise distribution and the parametrization has to be chosen manually.

Another category of methods tackle this ill-posed problem by learning a mapping from the degenerated signal to the original signal, *e.g.* from noisy input to clean depth maps, in a data-driven manner. Similar to our method are CNN-based approaches for natural image denoising [15], inpainting [37] and super-resolution [8], but also the conceptually related Filter Forests [11] have shown good results for denoising natural images and depth maps.

While latter methods are learning based, they apply the inference on individual patches. On the other hand, holistic approaches that learn a global and a local model in a common system are more similar to our method and have shown excellent results for a variety of tasks. Jancsary *et al.* [16] propose to learn the potentials of a Gaussian Random Field using a decision tree. The approach is conceptually similar to our method, but showed inferior results in image restoration when compared to state-of-the-art methods. There exist also a few works with the similar idea of combining a CNN with a global model. Ning *et al.* [25] use an Energy-Based model for sequentially training a CNN and a discrete graphical model for image segmentation. Tompson *et al.* [35] utilize a CNN together with a single iteration of belief propagation on a graphical model for human body poses estimation. Baltrusaitis *et al.* [1] use a perceptron to parametrize the potential of a Gaussian Random Field which are used as patch-experts in subsequent applications. In contrast to those methods, our approach optimizes a full global model jointly with a CNN. We achieve this by back-propagating the gradients through the GRM and the CNN in a bi-level formulation.

For the depth upscaling task, state-of-the-art methods rely on a guidance image, based on the assumption that depth discontinuities often occur at edges in the image domain. Diebel and Thrun [7] applied a Markov Random Field to exploit this assumption. They weighted the smoothness of their model with the gradient magnitude of the registered high-resolution intensity image. Similar ideas have been also proposed in approaches that use joint bilateral filters [9, 39]. Park *et al.* [26] formulated the depth map upscaling as a constrained least-squares optimization problem by combining non-local means and an edge-weighting scheme derived from the intensity image. A higher-order variational method with a strong regularization term was introduced by Ferstl *et al.* [12]. Recently, Yang *et al.* [38] demonstrated good results by applying a pixel-wise autoregressive model. They designed the coefficients of the model with respect to non-local correlations in the intensity and depth data. Our approach can also easily incorporate high-resolution guidance images for depth map restoration, but can further learn to distinguish which image gradients correlate with depth discontinuities in a data-driven manner.

3 Global Regression Model

Due to its ill-posed nature, denoising and upscaling is typically modeled as a regularized energy minimization problem:

$$u^* = \arg\min_u E(u;w,I) = \arg\min_u R(u;w_r) + D(u,I;w_d)\ , \tag{1}$$

where I is the input data, *e.g.* the low-resolution noisy depth map acquired by the sensor, and u^* is the restored estimate. The data term $D(u;.)$ encapsulates knowledge about the signal formation process, whereas the regularization term $R(u;.)$ imposes prior knowledge about

desirable solutions. Note that the energy depends on a set of parameters $w = [w_r, w_d]^T$. This is typically a small number of parameters that influence the amount of smoothing, or the probability of seeing a depth edge at a certain position in the image.

The explicit form of $R(u;.)$, $D(u;.)$ and the parametrization of the energy requires expert knowledge of the image formation process in order to yield good estimates u^*. To address this problem, we propose to parametrize the energy by a complex and highly non-linear function, and learn the parameters w from ground-truth data. As a result, the model adapts the parameters in a data-driven way. Moreover, the error-prone manual selection of suitable parameters w is completely eliminated.

We choose the parametrization to be given by a CNN, since this choice has several advantages: (1) CNNs are able to learn highly discriminative features directly from the input data, without the need for hand-crafted features. This makes them widely applicable and easily adaptable to novel input modalities. (2) They are trained using gradient-based optimization. Specifically, the back-propagation rule allows for efficient computation of the gradient of a large number of parameters. Moreover, the back-propagation rule can be adapted to accommodate for a global model [30]. This enables a joint end-to-end training of the complete model. (3) CNNs can be efficiently evaluated on GPUs, which allows for an overall fast execution, provided that the global model can be solved efficiently.

GRM Definition We assume that all information, which is provided by the depth sensor, is encapsulated in I_k, *i.e.* I_k is a multi-channel image that either includes depth and possibly intensity images. In the case of a ToF sensor I_k could for example be the depth map and the infrared image. We propose to use a Global Regression Model (GRM) to estimate a depth map from the input data I_k:

$$E(u; f(w, I_k)) = R(u, h(w_h, I_k)) + \frac{\exp(w_\lambda)}{2} \|u - g(w_g, I_k))\|^2. \tag{2}$$

Here, $R(u, h(w_h, I_k))$ is a regularization term, which is parametrized by the function $h(w_h, I_k)$. This term introduces prior knowledge, which will be learned from available training data via the parametrization $h(w_h, I_k)$. Similarly, the function $g(w_g, I_k)$ can be used to transform the input data to a form such that the global model can make reliable estimates. This function, again, will be learned from training data in our joint training procedure, which we detail later. The scalar smoothness parameter w_λ, which is also learned, allows us to find a trade-off between prior knowledge and the likelihood of the given estimate $g(w_g, I_k)$. We apply the exponential function to this parameter to ensure that the weighting is non-negative. It is important to note that the functions $h(w_h, I_k)$ and $g(w_g, I_k)$ may share parameters, *i.e.* subsets of w_h and w_g can be equal, *i.e.* the weights of a CNN.

As a regularizer, we propose a non-local pairwise model:

$$R(u) = \sum_{i=1}^{N} \sum_{j>i}^{N} n_\varepsilon(h_{ij}(w_h, I_k) \cdot (u_i - u_j)), \tag{3}$$

where n_ε denotes a convex, twice differentiable potential function. Further, we impose the constraint that the factors $h_{ij}(w_h, I_k)$ are non-negative to ensure an overall convex regularizer (3). We use a twice continuously differentiable approximation of the Huber function [18]:

$$n_\varepsilon(t) = [|t| \leq \varepsilon] \cdot \left(-\tfrac{1}{8\varepsilon^3} t^4 + \tfrac{3}{4\varepsilon} t^2 + \tfrac{3\varepsilon}{8} \right) + [|t| > \varepsilon] \cdot |t| \,, \tag{4}$$

where the small positive constant ε defines the threshold between the linear portion of the potential and the smooth polynomial part, and $[.]$ denotes to the Iversion bracket.

The definition for the regularizer is similar to the discretized non-local Total Variation [2, 13], where the non-smooth ℓ_1 penalty was replaced by a smooth penalty. The Huber approximation has two important merits when compared to the ℓ_1 penalty: Like the ℓ_1 penalty it allows for sharp discontinuities in the depth maps, but does not favor piecewise constant solutions, which alleviates the problem of stair-casing on slanted surfaces [36]. Second, in contrast to the ℓ_1 penalty, the proposed potential function is twice continuously differentiable, which makes a gradient-based training scheme feasible [30]. We present a visual comparison of the different penalties in the supplemental material.

The regularizer (3) is parametrized by functions $h_{ij}(w_h, I_k)$, *i.e.* the strength of the interaction between pixels i and j is determined by a complex non-linear relationship. This is different from previous approaches for modeling non-local interactions [13, 36], where the interactions are given by simple bilateral or even constant weights [33]. To ensure that the model remains tractable, we adopt a translation-invariant parametrization for the regularization term:

$$h_{ij}(w_h, I_k) = [\mathrm{d}(i,j) \leq T] \exp(-(h_{\mathrm{d(i,j)}}(w_h, I_k))^2) \tag{5}$$

where $\mathrm{d}(i,j)$ denotes the distance between pixels i and j in the image plane. The resulting parametrization of edges between pixels is thus determined by their distance alone, not by their absolute position. This definition fits very well with the translation invariant nature of CNNs. Moreover, setting the parametrization to 0 outside of a certain range T allows for efficient optimization of the model. We use Nesterov's Accelerated Gradient method [22] to optimize the GRM.

Parametrization We propose to use a CNN to parameterize energy (2). In general, a CNN consists of several layers that perform a linear transformation of the data, *e.g.* convolution, followed by a non-linearity, *e.g.* rectified linear unit (ReLU) [21]. In the remainder of this work we will use the following CNN architecture: First, we have a convolutional layer with 32 filters, each of size 9×9 pixels. On these feature maps, we apply a ReLU as non-linearity. In the second layer we again have a convolutional layer with 32 filters, each of size 5×5 pixels and use a ReLU as non-linearity. In the last layer, we have two different outputs. One for the data-term $g(w_g, I_k)$ and one for the regularization $h(w_h, Ik)$. Both are implemented as convolutions of size 3×3 and only differ in the number of output channels.

The network is directly applied to the raw input data, *i.e.* depth map and optionally an intensity image. For the denoising task no further pre-processing steps are involved. For upscaling, we resize the input data to the desired output resolution using bicubic interpolation as in [8].

Joint Training We assume that K pairs of sensor images I_k together with their ground-truth v_k are given for training the model. We formulate the training task as a bi-level problem [30]:

$$\underbrace{\min_{w \in W} \frac{1}{K} \sum_{k=1}^{K} L(u^*(f(w, I_k)), v_k)}_{\text{HL}} \quad \text{s.t. } u^*(f(w, I_k)) = \underbrace{\arg\min_{u \in \mathbb{R}^N} E(u; f(w, I_k))}_{\text{LL}}. \tag{6}$$

We will call HL the higher-level problem and LL the lower-level problem. This formulation of the training problem has an intuitive interpretation: The task of the training procedure is

to find parameters w for the energy $E(u; f(w, I_k))$, such that the minimizer $u^*(f(w, I_k))$ of the energy yields low training loss $L(u^*(f(w, I_k)), v_k))$. Note that if the LL problem is assumed to be of the form (1), the parametrization $f(w, I_k)$ may influence the data term, the regularizer or both. Specifically, in view of (2), we have $f(w, I_k) = [h(w_h, I_k),\ g(w_g, I_k),\ w_\lambda]^T$.

In practice bi-level problems are challenging to optimize even for a small number of variables w and u, due to their highly non-convex nature. The following proposition, however, provides conditions which will allow us to compute gradients of the HL problem, and simultaneously satisfy the constraint given by the LL problem, even for large-scale problems*:

Proposition 1. *Let $E(u; f(w, I_k))$ be strongly convex and twice differentiable with respect to u. Further, let $E(u; f(w, I_k))$ be differentiable with respect to f and let $f(w, I_k)$ be differentiable with respect to w. Then the gradient of a differentiable loss L with respect to the parameters w is well-defined and is given by*

$$\frac{\partial L}{\partial w} = -\sum_{k=1}^{K} \left(\left[\left(\nabla_u^2 E \right)^{-1} \frac{\partial L}{\partial u_k} \right]^T \frac{\partial^2 E}{\partial u \partial w} \right) \Bigg|_{u_k = u_k^*}. \tag{7}$$

Remark. Note that

$$\frac{\partial^2 E(u_k)}{\partial u_k \partial w} = \frac{\partial^2 E(u_k; f(w, I_k))}{\partial u_k \partial f} \frac{\partial f(w, I_k)}{\partial w}, \tag{8}$$

which shows that a necessary condition for the computation of the gradient (7) is differentiability with respect to the parametrization $f(w, I_k)$.

An interesting property of this scheme is that it can be interpreted as back-propagation: If the parametrization is a CNN, the gradient for a single training example can be computed, by back-propagating the quantity

$$\Delta E = -\left(\left[\left(\nabla_u^2 E \right)^{-1} \frac{\partial L}{\partial u_k} \right]^T \frac{\partial^2 E}{\partial u_k \partial f} \right) \Bigg|_{u_k = u_k^*} \tag{9}$$

into the network. Hence, integration of the global model within existing CNN frameworks is straight-forward.

The proposed scheme for gradient computation allows to handle large-scale data, *i.e.* a large amount of model parameters, as well as typical image sizes of u. For a large body of training data, repeated evaluation of the gradient quickly becomes infeasible, since the LL problem has to be solved for each training image in every gradient evaluation. This problem can be alleviated by using batch stochastic gradient descent, where in each step only a random subset of the training images has to be considered. A basic stochastic gradient descent scheme, which shows the necessary computations for the gradient evaluation is summarized in Algorithm 1.

while not converged
1. Sample B instances from training set
2. $\Delta w \leftarrow 0$
3. For each $b \in B$ solve
   - $u_b^* = \arg\min_{u \in \mathbb{R}^N} E(u; f(w^k, I_b))$
   - Compute $\Delta L = \frac{\partial L}{\partial u_b}(u_b^*, v_b)$
   - Compute $H = \nabla_u^2 E(u_b^*, f(w^k, I_b))$
   - Solve the linear system $H\gamma = \Delta L$
   - Compute $\Delta E = \gamma^T \frac{\partial^2 E}{\partial u_b \partial f}(u_b^*, f(w^k, I_b))$
   - Compute Δw_b by backpropagating ΔE
   - Set $\Delta w \leftarrow \Delta w - \Delta w_b$
4. Update $w^{k+1} \leftarrow w^k - \alpha \Delta w$

Algorithm 1: GRM Gradient Evaluation

*Proof in the supplemental material.

Note that energy (2) is strongly convex with modulus $exp(w_\lambda)$ and fulfills the necessary conditions on differentiability. Thus, Algorithm 1 can be used to learn the GRM. The exact expressions for the necessary gradients and the Hessian are given in the supplemental material.

Practical Considerations The GRM was integrated as a custom layer in the Caffe framework [17]. The back-propagation step to compute the weights of the parametrization is entirely handled by the Caffe framework and allows to incorporate a diverse set of layer types into the parametrization. We use stochastic gradient descent with a Nesterov-style momentum term [34] to train the CNN. To allow for an overall faster training procedure, we pre-train the CNN on a training set of patches having a size of 32×32 pixel. For the data term, the training target is given by the ground-truth depth maps, while for the pairwise potentials ground-truth data was generated by extracting edges from the depth maps. In the pre-training phase we optimize the joint quadratic loss over the depth estimates and gradient estimates:

$$\begin{aligned} L([g(w_g,I_k),h(w_h,I_k)]^T,v_k) =& \tfrac{1}{2}\|g(w_g,I_k)-v_k\|^2 \\ &+ \tfrac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{P}\|h_{ij}(w_h,I_k)-((v_k)_i-(v_k)_j)\|^2. \end{aligned} \tag{10}$$

The pre-training was carried out for $250,000$ iteration with an initial learning rate of 10^{-5} and the momentum set to 0.95. After $200,000$ iterations the learning rate was decreased to 10^{-6}. The pre-trained weights are used to initialize the full model (CNN+GRM), where we optimize the quadratic loss

$$L(u_k^*(f(w,I_k)),v_k) = \tfrac{1}{2}\|u_k^*(f(w,I_k))-v_k\|^2. \tag{11}$$

The full model was trained for $60,000$ iterations with an initial learning rate of 10^{-7}. We again decreased the learning rate after $55,000$ iterations to 10^{-8}.

4 Evaluation

This section presents quantitative and qualitative results of our method. We perform an extensive evaluation for two tasks: depth map denoising and upscaling. We compare our method to current state-of-the-art algorithms for different noise characteristics and validate the flexibility and effectiveness of our proposed method.

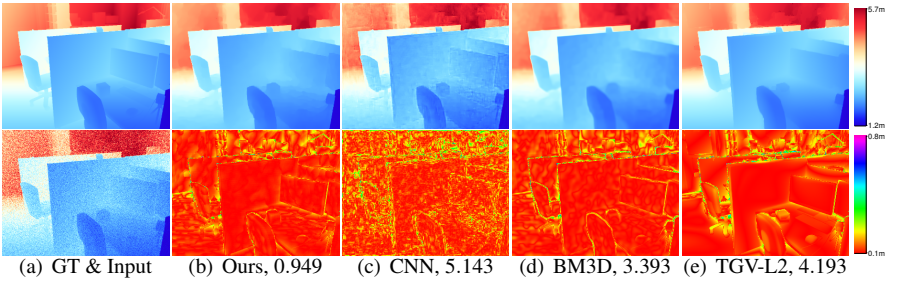
Denoising For the depth map denoising experiments, we use the New Tsukuba dataset [20, 28]*. The dataset consists of $1,800$ realistically rendered depth maps paired with aligned color images.We split the dataset into a training set that consists of the first $1,500$ images and a test set that includes the remaining 300 images. To simulate the acquisition process of a depth sensor, we add multiplicative Gaussian noise with $\sigma \in \{0.1, 0.5, 0.7\}$ to the depth maps and additive Gaussian noise with $\sigma = 0.2$ to the intensity images.

For our proposed method we evaluated a non-local regularizer (NL), where the distance threshold T was set to 2. This corresponds to a non-local neighborhood of 5×5 pixels. Further, we compare this to a simpler local model (L), where each pixel is only connected to its direct neighbors in a 4-connected neighborhood.

*The supplemental material includes additional evaluations on the NYU2 dataset [32].

	Ours (NL)	Ours (L)	CNN+GRM	CNN	K-SVD	SAR-BM3D	BM3D	TGV-L2	TV-L1
$\sigma = 0.2$	**2.052**	2.175	2.802	4.880	4.968	2.715	2.260	2.664	2.781
$\sigma = 0.5$	**3.335**	3.538	3.835	6.916	6.696	7.252	4.133	4.782	4.949
$\sigma = 0.7$	**3.965**	4.084	4.384	10.901	7.900	9.326	5.200	5.621	6.191

Table 1: Quantitative evaluation of our method on the New Tsukuba dataset. The error is measured as *RMSE* in *cm*.



(a) GT & Input (b) Ours, 0.949 (c) CNN, 5.143 (d) BM3D, 3.393 (e) TGV-L2, 4.193

Figure 2: Qualitative results on New Tsukuba, $\sigma = 0.2$. The first column presents the ground-truth and the noisy input depth map. The remaining columns show the denoising results in the first row and the absolute error in the second row, respectively. The numbers in the sub-captions denote to the RMSE in *cm*.

The results of our evaluation are summarized in Table 1. In addition to our proposed methods with non-local (NL) and local (L) regularization, we evaluate the performance of the pre-trained CNN alone and of the CNN with the GRM on top of it, but without joint training. Finally, we report the results of several state-of-the-art denoising methods, *i.e.* K-SVD [9], SAR-BM3D [27], BM3D [6], TGV-L2 [3] and TV-L1 [5, 24]. All numerical results are in terms of the root mean squared error (RMSE) in *cm*.

It can be seen that our approach outperforms all other methods. The proposed non-local regularization improves the results over the simple local model. We can further observe that the joint training of the full model leads to a significant improvement, as the parameters of the CNN adapt to the GRM. The performance of the other methods significantly depends on the noise distribution they model. K-SVD has almost a twice as high RMSE when compared to TV-L1, or TGV-L2, which has a very strong and suitable regularization term.

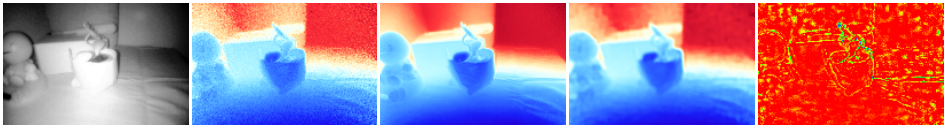
In Figure 2 we present additional qualitative results of a subset of the evaluated methods. We can observe that BM3D has problems with over-smoothing at edges and multiplicative noise, especially in the background. While TGV-L2 generates visually very appealing results, it produces larger errors near depth discontinuities and in the background. In contrast, our method is accurate near depth boundaries and handles the increased noise in the background better.

Different Noise Characteristics One of the main advantages of our proposed method is that the CNN parametrization is able to adapt to different noise characteristics. We evaluate this behavior with two more data dependent noise distributions and Salt & Pepper noise on the New Tsukuba dataset. The first noise type is given as in [26] by adding Gaussian noise with $\mu = 0$ and a depth dependent sigma $\sigma_d = 0.5d$. For the second noise type, we use a depth-dependent Poisson noise with $\lambda = 10^{-3}$ as in [11]. Finally, we evaluate Salt & Pepper noise, where we set 35% of the pixels to either the maximum or the minimum depth value.

The numerical results of this experiment are shown in Table 2. Our method is able to

		Ours (NL)	CNN	K-SVD	SAR-BM3D	BM3D	TGV-L2	TV-L1
Local Variance	$\sigma_d = 0.5d$	**2.730**	3.913	5.741	4.616	3.178	3.042	4.006
Poisson	$\lambda = 10^{-3}$	**3.4016**	8.660	6.695	7.279	4.129	10.595	4.958
Salt & Pepper	$p = 0.35$	**10.484**	18.880	81.685	89.929	77.010	80.764	97.420

Table 2: Quantitative results on different noise distributions. The error is measured as RMSE in *cm*.



(a) IR (b) Depth (c) GT (d) Ours (e) Error

Figure 3: Qualitative results of our method on ToF data. RMSE to ground-truth is 0.49*cm*.

learn a good parametrization for all different noise characteristics. In contrast, the compared methods inherently have problems with noise types, which differ from their model assumptions. This is especially pronounced in the case of Salt & Pepper noise, that can not be handled by any method, except the CNN and our proposed method.

Upscaling Dong *et al.* [8] showed in their work that a CNN can be effectively utilized for natural image upscaling. They resize the low-resolution image in a pre-processing step to the output resolution using bicubic interpolation and learn the mapping from the pre-processed images to the ground truth high-resolution images. In our experiment we follow the same principle, as we first upscale the low-resolution depth map by bicubic interpolation and then apply a CNN, but with the GRM on top of it.

To evaluate the upscaling capabilities of our method, we use the New Tsukuba dataset with the same train and test split as before. We resize each depth map by a factor of $s \in \{2,4\}$ and add multiplicative Gaussian noise with zero mean and $\sigma = 0.2$ to the low-resolution data. We compare our non-local model (NL) and the local model (L) with the results of a plain CNN, which is similar to the method of [8], and a state-of-the-art image-guided depth map upscaling method by Ferstl *et al.* [12].

	Ours (NL)	Ours (L)	CNN	Ferstl *et al.*
×2	**2.940**	3.042	6.427	3.834
×4	**4.530**	4.813	8.411	5.506

Table 3: Upscaling results of our method with non-local (NL) and local (L) regularization compared to the CNN [8] and the method by Ferstl *et al.* [12]. The error is measured as RMSE in *cm*.

We present the quantitative results in Table 3. We can see that our proposed method yields the lowest overall RMSE, and that the non-local regularization gives better results than the local variant. In contrast to natural image super-resolution, the CNN alone is not able to reach state-of-the-art performance.

ToF Denoising In a last experiment we qualitatively evaluate our approach for denoising on a consumer ToF depth camera [29]. To generate train and test data, we constructed five different static scenes and recorded each scene from several different view-points. The ground-truth is obtained by taking the per-pixel median over 500 images of each scene for a fixed view-point. We present qualitative results of our approach on a test image in Figure 3.

5 Conclusion

In this paper we introduced a Global Regression Model (GRM) for denoising and upscaling of depth maps. The model uses a complex non-linear parametrization of the GRM which is given by a Convolutional Neural Network (CNN). The parameters of the model can be trained jointly in an end-to-end fashion, which enables the model to adapt to the characteristics of a given sensor in a data-driven manner. Our experiments show that the model is indeed able to learn different noise characteristics and consistently outperforms specialized state-of-the-art methods for different noise types. We further showed the applicability to depth map upscaling and qualitative results on real Time-of-Flight data. In future research we want to incorporate even more powerful regularization terms into the GRM, such as a smooth approximation of TGV. Finally, we see potential of our method for computer vision tasks that can benefit from a joint learning of a global model and its parametrization.

Acknowledgments This work was supported by *Infineon Technologies Austria AG* and the Austrian Research Promotion Agency (FFG) under the *FIT-IT Bridge* program, project #838513 (TOFUSION).

References

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Continuous Conditional Neural Fields for Structured Regression. In *ECCV*, 2014.

[2] Sébastien Bougleux, Abderrahim Elmoataz, and Mahmoud Melkemi. Local and Nonlocal Discrete Regularization on Weighted Graphs for Image and Mesh Processing. *IJCV*, 84(2):220–236, 2009.

[3] Kristian Bredies, Karl Kunisch, and Thomas Pock. Total Generalized Variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526, 2010.

[4] Derek Chan, Hylke Buisman, Christian Theobalt, Sebastian Thrun, et al. A Noise-aware Filter for Real-time Depth Upsampling. In *ECCV*, 2008.

[5] Tony F. Chan and Selim Esedoglu. Aspects of Total Variation Regularized L1 Function Approximation. *SIAM Journal on Applied Mathematics*, 65(5):1817–1837, 2005.

[6] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen O. Egiazarian. Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. *TIP*, 16(8):2080–2095, 2007.

[7] James Diebel and Sebastian Thrun. An Application of Markov Random Fields to Range Sensing. In *NIPS*, 2005.

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a Deep Convolutional Network for Image Super-Resolution. In *ECCV*, 2014.

[9] Michael Elad and Michal Aharon. Image Denoising Via Sparse and Redundant Representations Over Learned Dictionaries. *TIP*, 15(12):3736–3745, 2006.

[10] Dragos Falie and Vasile Buzuloiu. Noise Characteristics of 3D Time-of-Flight Cameras. In *International Symposium on Signals, Circuits and Systems*, 2007.

[11] Sean Fanello, Cem Keskin, Pushmeet Kohli, Shahram Izadi, Jamie Shotton, Antonio Criminisi, Ugo Pattaccini, and Tim Paek. Filter Forests for Learning Data-Dependent Convolutional Kernels. In *CVPR*, 2014.

[12] David Ferstl, Christian Reinbacher, René Ranftl, Matthias Rüther, and Horst Bischof. Image Guided Depth Upsampling using Anisotropic Total Generalized Variation. In *ICCV*, 2013.

[13] G. Gilboa and S. Osher. Nonlocal Operators with Applications to Image Processing. *Multiscale Modeling and Simulation*, 7(3):1005–1028, 2009.

[14] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, and Andrew Fitzgibbon. KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera. In *ACM Symposium on User Interface Software and Technology*, 2011.

[15] Viren Jain and Sebastian Seung. Natural Image Denoising with Convolutional Networks. In *NIPS*, 2009.

[16] Jeremy Jancsary, Sebastian Nowozin, Toby Sharp, and Carsten Rother. Regression Tree Fields - An Efficient, Non-parametric Approach to Image Labeling Problems. In *CVPR*, 2012.

[17] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[18] Karl Kunisch and Thomas Pock. A Bilevel Optimization Approach for Parameter Learning in Variational Models. *SIAM Journal on Imaging Sciences*, 6(2):938–983, 2013.

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[20] Sarah Martull, Martin Peris, and Kazuhiro Fukui. Realistic CG Stereo Image Dataset with Ground Truth Disparity Maps. In *ICPR Workshops*, 2012.

[21] Vinod Nair and Geoffrey E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *ICML*, 2010.

[22] Yurii Nesterov. A Method of Solving a Convex Programming Problem with Convergence Rate O(1/sqr(k)). *Soviet Mathematics Doklady*, 27:372–376, 1983.

[23] Chuong V. Nguyen, Shahram Izadi, and David Lovell. Modeling Kinect Sensor Noise for Improved 3D Reconstruction and Tracking. In *3D Imaging, Modeling, Processing, Visualization and Transmission*, 2012.

[24] Mila Nikolova. A Variational Approach to Remove Outliers and Impulse Noise. *Journal of Mathematical Imaging and Vision*, 20(1-2):99–120, 2004.

[25] Feng Ning, Damien Delhomme, Yann LeCun, Fabio Piano, Léon Bottou, and Paolo E. Barbano. Toward automatic phenotyping of developing embryos from videos. *TIP*, 14 (9):1360–1371, 2005.

[26] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S. Brown, and In-So Kweon. High Quality Depth Map Upsampling for 3D-TOF Cameras. In *ICCV*, 2011.

[27] Sara Parrilli, Mariana Poderico, Cesario Vincenzo Angelino, and Luisa Verdoliva. A Nonlocal SAR Image Denoising Algorithm Based on LLMMSE Wavelet Shrinkage. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2):606–616, 2012.

[28] Martin Peris, Sara Martull, Atsuto Maki, Yasuhiro Ohkawa, and Kazuhiro Fukui. Towards a Simulation Driven Stereo Vision System. In *ICPR*, 2012.

[29] *Camboard Pico*. PMD Technologies. Germany.

[30] René Ranftl and Thomas Pock. A Deep Variational Model for Image Segmentation. 2014.

[31] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear Total Variation Based Noise Removal Algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[32] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*, 2012.

[33] Deqing Sun, Stefan Roth, and Michael J. Black. Secrets of Optical Flow Estimation and Their Principles. In *CVPR*, 2010.

[34] Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, 2013.

[35] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In *NIPS*, 2014.

[36] Manuel Werlberger, Thomas Pock, and Horst Bischof. Motion Estimation with Non-Local Total Variation Regularization. In *CVPR*, 2010.

[37] Junyuan Xie, Linli Xu, and Enhong Chen. Image Denoising and Inpainting with Deep Neural Networks. In *NIPS*, 2012.

[38] Jingyu Yang, Xinchen Ye, Kun Li, Chunping Hou, and Yao Wang. Color-Guided Depth Recovery From RGB-D Data Using an Adaptive Autoregressive Model. *TIP*, 23(8): 3443–3458, 2014.

[39] Qingxiong Yang, Ruigang Yang, James Davis, and David Nistér. Spatial-Depth Super Resolution for Range Images. In *CVPR*, 2007.