Deformable Kernel Networks for Joint Image Filtering
Beomjun Kim, Jean Ponce, Bumsub Ham

To cite this version:
Beomjun Kim, Jean Ponce, Bumsub Ham. Deformable Kernel Networks for Joint Image Filtering. 2019. <hal-01857016v3>

HAL Id: hal-01857016
https://hal.archives-ouvertes.fr/hal-01857016v3
Submitted on 4 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Deformable Kernel Networks for Joint Image Filtering

Beomjun Kim, Jean Ponce, Fellow, IEEE, Bumsub Ham, Member, IEEE

Abstract—Joint image filters are used to transfer structural details from a guidance picture used as a prior to a target image, in tasks such as enhancing spatial resolution and suppressing noise. Previous methods based on convolutional neural networks (CNNs) combine nonlinear activations of spatially-invariant kernels to estimate structural details and regress the filtering result. In this paper, we instead learn explicitly sparse and spatially-variant kernels. We propose a CNN architecture and its efficient implementation, called the deformable kernel network (DKN), that outputs sets of neighbors and the corresponding weights adaptively for each pixel. The filtering result is then computed as a weighted average. We also propose a fast version of DKN that runs about fifteen times faster for an image of size $640 \times 480$. We demonstrate the effectiveness and flexibility of our models on the tasks of depth map upsampling, saliency map upsampling, cross-modality image restoration, texture removal, and semantic segmentation. In particular, we show that the weighted averaging process with sparsely sampled $3 \times 3$ kernels outperforms the state of the art by a significant margin.

Index Terms—Joint filtering, convolutional neural networks, depth map upsampling, cross-modality image restoration, texture removal, semantic segmentation

1 INTRODUCTION

Image filtering with a guidance signal, a process called guided or joint filtering, has been used in a variety of computer vision and graphics tasks, including depth map upsampling [1], [2], [3], [4], [5], [6], cross-modality image restoration [7], [8], [9], texture removal [6], [10], [11], [12], scale-space filtering [6], dense correspondence [13], [14] and semantic segmentation [15]. For example, high-resolution color images can be used as guidance to enhance the spatial resolution of depth maps [4]. The basic idea behind joint image filtering is to transfer structural details from the guidance image to the target one, typically by estimating spatially-variant kernels from the guidance. Concretely, given the target image $f$ and the guidance image $g$, the filtering output $\hat{f}$ at position $p = (x, y)$ is expressed as a weighted average [4], [16]:

$$\hat{f}_p = \sum_{q \in N(p)} W_{pq}(f, g) f_q,$$

where we denote by $N(p)$ a set of neighbors (defined on a discrete regular grid) near the position $p$. The filter kernel $W$ is a function of the guidance image $g$ [2], [3], [4], [7], the target image $f$ itself [11], [16], or both [5], [6], normalized so that

$$\sum_{q \in N(p)} W_{pq}(f, g) = 1.$$

Classical approaches to joint image filtering mainly focus on designing the filter kernels $W$ and the set of neighbors $N$ (i.e., sampling locations $q$). They use hand-crafted kernels and sets of neighbors without learning. For example, the bilateral filter [16] uses spatially-variant Gaussian kernels to encode local structures from the guidance image. The guided filter [7] also leverages the local structure of the guidance image, but uses matting Laplacian kernels [19], enabling a constant processing time. These filters use regularly sampled neighbors for aggregating pixels, and do not handle inconsistent structures in the guidance and target images [6]. This causes texture-copying artifacts [3], especially in the case of data from different sensors. To address this problem, the SD filter [6] constructs spatially-variant kernels from both guidance and target images to exploit common structures, and formulates joint image filtering as an optimization problem. This type of approaches (e.g., [3], [10], [20]) computes a filtering output by optimizing an objective function that involves solving a large linear system. This is equivalent to filtering an image by an inverse matrix [7], whose rows correspond to a filter kernel, leveraging global structures in the guidance image. Optimization-based methods can be considered as implicit weighted-average filters. Learning-based approaches using convolutional neural networks (CNNs) [5], [21], [22] are also becoming increasingly popular. The networks are trained using large quantities of data, reflecting natural image priors and often outperforming traditional methods by large margins. These methods do not use a weighted averaging process with spatially-variant kernels as in (1). They combine instead nonlinear activations of spatially-invariant kernels learned from the networks. That is, they approximate spatially-variant kernels by mixing the activations of spatially-invariant ones nonlinearly (e.g., via the ReLU function [23]).

In this paper, we propose to exploit spatially-variant kernels explicitly to encode the structural details from both guidance and target images as in classical approaches, but learn the kernel weights in a completely data-driven way. We also learn the set of neighbors, building an adaptive and sparse neighborhood system for each pixel, which may be difficult to design by hand. To implement this idea, we propose a CNN architecture and its efficient implementation, called a deformable kernel network (DKN), for learning sampling locations of the neighboring pixels and their corresponding kernel weights at every pixel. We also propose a fast version of DKN (FDKN), achieving a $15 \times$ speed-up compared to the plain DKN for an image of size $640 \times 480$, while
retaining its superior performance. We show that the weighted averaging process using sparsely sampled $3 \times 3$ kernels is sufficient to obtain a new state of the art in a variety of applications, including depth map upsampling (Figs. 1 and 7), saliency map upsampling (Fig. 8), cross-modality image restoration (Fig. 9), texture removal (Fig. 10), and semantic segmentation (Fig. 11).

**Contributions.** The main contributions of this paper can be summarized as follows:

- We introduce a generic model for joint image filtering and its implementation, the DKN, that computes the set of neighbors and their corresponding weights adaptively for individual pixels (Section 3).
- We propose a fast version of DKN (FDKN) that runs about \( \frac{1}{10} \) times faster than the DKN while retaining its superior performance (Section 3).
- We achieve a new state of the art on several tasks, clearly demonstrating the advantage of our approach to learning both kernel weights and sampling locations (Section 4). We also provide an extensive experimental analysis to investigate the influence of all the components and parameters of our model (Section 5).

To encourage comparison and future work, our code and models are available at our project webpage: https://github.com/jun0kim/deformablekernelnetwork.

## 2 Related work

Here we briefly describe the context of our approach, and review representative works related to ours.

### 2.1 Joint image filtering

We categorize joint image filtering into explicit/implicit weighted-average methods and learning-based ones. First, explicit joint filters compute the output at each pixel by a weighted average of neighboring pixels in the target image, where the weights are estimated from the guidance and/or target image \([4], [7], [11]\). The bilateral \([16]\) and guided \([7]\) filters are representative methods that have been successfully adapted to joint image filtering. They use hand-crafted kernels to transfer fine-grained structures from the guidance image. It is, however, difficult to manually adapt the kernels to new tasks, and these methods may transfer erroneous structures to the target image \([5]\). Second, implicit weighted-average methods formulate joint filtering as an optimization problem, and minimize an objective function that usually involves fidelity and regularization terms \([2], [3], [6], [10], [20], [24]\). The fidelity term encourages the filtering output to be close to the target image, and the regularization term, typically modeled using a weighted L2 norm \([20]\), gives the output having a structure similar to that of the guidance image. Although, unlike explicit ones, implicit joint filters exploit global structures in the guidance image, hand-crafted regularizers may not reflect structural priors in the guidance image. Moreover, optimizing the objective function involves solving a large linear system, which is time consuming, even with preconditioning \([25]\) or multigrid methods \([20]\). Finally, learning-based methods can further be categorized into dictionary- and CNN-based approaches. Dictionary-based methods exploit the relationship between paired target patches (e.g., low- and high-resolution patches for upsampling), additionally coupled with the guidance image \([26], [27]\). In CNN-based methods \([5], [21], [22]\), an encoder-decoder architecture is used to learn features from the target and guidance images, and the filtering output is then regressed directly from the network. Learning-based techniques require a large number of ground-truth images for training.

Our method borrows from both explicit weighted-average methods and CNN-based ones. Unlike existing explicit joint filters \([4], [7], [11]\), that use hand-crafted kernels and neighbors defined on a fixed regular grid, we leverage CNNs to learn the set of neighbors and their corresponding weights adaptively. Our method differs from previous CNN-based ones \([5], [21], [22]\) in that we learn spatially-variant kernels for each pixel to obtain filtering results as a weighted average.

### 2.2 Variants of the spatial transformer \([28]\)

Recent works introduce more flexible and effective CNN architectures. Jaderberg et al. propose a novel learnable module, the spatial transformer \([28]\), that outputs the parameters of the desired spatial transformation (e.g., affine or thin plate spline) given a feature map or an input image. The spatial transformer makes a standard CNN network for classification invariant to a set of geometric transformation, but it has a limited capability of handling local transformations. Choy et al. introduce a convolutional version of the spatial transformer \([29]\). They learn local transformation parameters for normalizing orientation and scale in feature matching. Most similar to ours are the dynamic filter network \([30]\) and its variants (the adaptive convolution network \([31]\) and the kernel prediction networks \([32], [33], [34]\)), where a set of local transformation parameters is generated adaptively conditioned on the input image. The main differences between our model and these works are three-fold. First, our network is more general in that it is not limited to learning spatially-variant kernels, but it also learns the sampling locations of neighbors. This allows us to aggregate
sparse but highly related samples only, enabling an efficient implementation in terms of speed and memory and achieving state-of-the-art results even with kernels of size $3 \times 3$ on several tasks. Note that the adaptive convolution and kernel prediction networks require much larger neighbors (e.g., $21 \times 21$ in [32], [34], $41 \times 41$ in [31], and $8 \times 5 \times 5$ in [33]). We show that learning sampling locations of neighbors clearly boosts the performance significantly compared to learning kernel weights only. Second, our model is generic and generalizes well to other tasks. It can handle multi-modal images (e.g., RGB/D images in depth map upsampling and RGB/cost images in semantic segmentation), and thus is applicable to various tasks including depth and saliency map upsampling, cross-modality image restoration, texture removal, and semantic segmentation. In contrast, the adaptive convolution network is specialized to video frame interpolation, and kernel prediction networks are applicable to denoising Monte Carlo renderings [32], [34] or burst denoising [33] only. Finally, our model learns spatially-variant kernels to compute residual images, not a final result, while the deformable convolutional network uses it to compute intermediate feature maps.

## 3 Proposed Approach

In this section, we briefly describe our approach to learning both kernel weights and sampling locations for joint image filtering (Section 3.1), and present a concrete network architecture (Section 3.2) and its efficient implementation using a shift-and-stitch approach (Section 3.3). We then describe a fast version of DKN (Section 3.4).

### 3.1 Overview

Our network mainly consists of two parts (Fig. 2): We first learn spatially-variant kernel weights and spatial sampling offsets w.r.t. the regular grid. To this end, a two-stream CNN [36], where each sub-network has the same structure (but different parameters), takes the guidance and target images to extract feature maps that are used to estimate the kernel weights and the offsets. We then compute a weighted average using the learned kernel weights and sampling locations computed from the offsets to obtain a residual image. Finally, the filtering result is obtained by combining the residuals with the target image. Our network is fully convolutional, does not require fixed-size input images, and it is trained end-to-end.

**Weight and offset learning.** It is hard to get supervision directly for kernel weights and offsets for individual pixels. We instead train our network with ground-truth target images and learn the kernel weights and offsets guided by an explicit weighted average. This is a kind of meta-supervision that specifies how the kernel weights and offsets behave. In particular, constraints on weight and offset regression (sigmoid and mean subtraction layers in Fig. 2) guide the learning process. For weight regression, we apply a sigmoid layer that makes all elements larger than 0 and less than 1. We then subtract the mean value from the output of the sigmoid layer so that the regressed weights should be similar to high-pass filter with kernel weights adding to 0. For offset regression, we do not apply the sigmoid layer since relative offsets (for $x$, $y$ positions) from locations on a regular grid can have negative values.

**Residual connection.** The main reason behind using a residual connection is that the filtering result is largely correlated with the target image, and both share low-frequency content [17], [37], [38], [39]. Focusing on learning the residuals also accelerates training speed while achieving better performance. Note that contrary to [17], [37], [38], [39], we obtain the residuals by a weighted averaging process with the learned kernels, instead of estimating them directly from the network output. Empirically, the kernels learned with the residual connection have the same characteristics as the high-pass filters widely used to extract important structures (e.g., object boundaries) from images. Note also that we can train DKN without the residual connection. In
this case, we compute the filtering result as the weighted average in (1).

3.2 DKN architecture

We design a fully convolutional network to learn the kernel weights and the sampling offsets for individual pixel. A detailed description of the network structure is shown in Table 1.

Feature extraction. We adapt a similar architecture in [31] for feature extraction that consists of 7 convolutional layers. We input the guidance and target images to each of the sub-networks that gives a feature map of size $128 \times 1 \times 1$ for a receptive field of size $51 \times 51$. We use the ReLU [12] as an activation function. Batch normalization [40] is used for speeding up training and regularization. In case of joint upsampling tasks (e.g., depth map upsampling), we input a high-resolution guidance image (e.g., a color image) and a low-resolution target image (e.g., a depth map) upsampled using bicubic interpolation.

Weight regression. For each sub-network, we add a $1 \times 1$ convolutional layer on top of the feature extraction layer. It gives a feature map of size $k^2 \times 1 \times 1$, where $k$ is the size of the filter kernel, which is used to regress the kernel weights. To estimate the weights, we apply a sigmoid layer to each feature map of size $k^2 \times 1 \times 1$, and then combine the outputs by element-wise multiplication (see Fig. 2). We could use a softmax layer as in [31], but empirically find that it does not perform as well as the sigmoid layer. The softmax function encourages the estimated kernel to have only a few non-zero elements, which is not appropriate for image filtering. The estimated kernels should be similar to high-pass filters, with kernel weights adding to 0. To this end, we subtract the mean value from the combined output of size $k^2 \times 1 \times 1$. For our model without a residual connection, we apply instead L1 normalization to the output of size $k^2 \times 1 \times 1$. Since the sigmoid layer makes all elements in the combined output larger than 0, applying L1 normalization forces the kernel weights to add to 1 as in (2).

Offset regression. Similar to the weight regression case, we add a $1 \times 1$ convolutional layer on top of the feature extraction layer. The resulting two feature maps of size $2k^2 \times 1 \times 1$ are combined by element-wise multiplication. The final output contains relative offsets (for $x$, $y$ positions) from locations on a regular grid. In our implementation, we use $3 \times 3$ kernels, but the filtering result is computed by aggregating 9 samples sparsely chosen from a much larger neighborhood. The two main reasons behind the use of small-size kernels are that (1) the size of the receptive field and the reliability of samples are much more important than the total number of samples aggregated, and (2) this enables an efficient implementation in terms of speed and memory. A similar finding is noted in [41], which shows that only high-confidence samples should be chosen when estimating foreground and background images in image matting. Note that offset regression is closely related to nonlocal means [42] in that both select which pixels to aggregate instead of immediate neighbors. Note also that learning offsets can be thought of as a generalization of “self-supervised” correspondence models in stereo matching [43] and optical flow estimation [44]. For example, in the case of stereo matching, a model is trained to produce a flow field such that a right image is reconstructed by a left one according to that flow field. Our model with filter kernels of size $k \times k$ computes $k^2$ correspondences for each pixel within input images, and also learns the corresponding weights to produce a filtering output.

![Fig. 3. Illustration of irregular sampling of neighboring pixels using offsets: (a) regular sampling $q$ on discrete grid; (b) learned offsets $\Delta q$; (c) deformable sampling locations $s(q)$ with the offsets $\Delta q$. The learned offsets are fractional and the corresponding pixel values are obtained by bilinear interpolation.](image_url)

**Weighted average.** Given the learned kernel $K$ and sampling offsets $\Delta q$, we compute the residuals $\hat{f}_p - f_p$ as a weighted average:

$$\hat{f}_p = f_p + \sum_{q \in \mathcal{N}(p)} K_{ps}(f,g) f_q(s(q)),$$

where $\mathcal{N}(p)$ is a local $3 \times 3$ window centered at the location $p$ on a regular grid (Fig. 3(a)). We denote by $s(q)$ the sampling position computed from the offset $\Delta q$ (Fig. 3(b)) of the location $q$ as follows.

$$s(q) = q + \Delta q,$$

The sampling position $s(q)$ predicted by the network is irregular and typically fractional (Fig. 3(c)). We use bilinear interpolation [28] to sample corresponding (sub) pixels $f_q(s(q))$ as

$$f_q(s(q)) = \sum_{t \in \mathcal{R}(s(q))} G(s,t) f_t,$$

where $\mathcal{R}(s(q))$ enumerates all integer locations in a local 4-neighborhood system to the fractional position $s(q)$, and $G$ is a two-dimensional bilinear kernel. We split the kernel $G$ into two one-dimensional ones [28], [35] as

$$G(s,t) = g(s_x,t_x)g(s_y,t_y),$$

where $g(a,b) = \max(0,1 - |a - b|)$. Note that the residual term in (3) is exactly the same as the explicit joint filters in (1), but we aggregate pixels from the sparsely chosen locations $s(q)$ with the learned kernels $K$. Note that aggregating pixels from a fractional grid is not feasible in current joint filters including DJF [17].

When we do not use a residual connection, we compute the filtering result $\hat{f}_p$ directly as a weighted average using the learned kernels and offsets:

$$\hat{f}_p = \sum_{q \in \mathcal{N}(p)} K_{ps}(f,g) f_q(s(q)).$$

**Loss.** We train our model by minimizing the $L_1$ norm of the difference between the network output $\hat{f}$ and ground truth $f^g$ as follows.

$$L(f^g, \hat{f}) = \sum_{p} |f^g_p - \hat{f}_p|_1.$$
Network architecture details. “BN” and “Res.” denote the batch normalization [40] and residual connection, respectively. We denote by “DownConv” convolution with stride 2. The inputs of our network are 3-channel guidance and 1-channel target images (denoted by D). For the model without the residual connection, we use an L1 normalization layer (denoted by “L1 norm.”) instead of subtracting mean values for weight regression.

<table>
<thead>
<tr>
<th>Feature extraction</th>
<th>Output</th>
<th>Weight regression</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>D x 51 x 51</td>
<td>Conv(1 x 1)</td>
<td>k^2 x 1 x 1</td>
</tr>
<tr>
<td>Conv(7 x 7)-BN-ReLU</td>
<td>32 x 45 x 45</td>
<td>Sigmoid</td>
<td>k^2 x 1 x 1</td>
</tr>
<tr>
<td>DownConv(2 x 2)-ReLU</td>
<td>32 x 22 x 22</td>
<td>Mean subtraction or L1 norm. (w/o Res.)</td>
<td>k^2 x 1 x 1</td>
</tr>
<tr>
<td>Conv(5 x 5)-BN-ReLU</td>
<td>64 x 18 x 18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DownConv(2 x 2)-ReLU</td>
<td>64 x 9 x 9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conv(5 x 5)-BN-ReLU</td>
<td>128 x 5 x 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conv(3 x 3)-ReLU</td>
<td>128 x 3 x 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conv(3 x 3)-ReLU</td>
<td>128 x 1 x 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE 1

3.3 FDKN architecture

We now present a fast version of DKN (FDKN) that achieves a 15x speed-up over the DKN for an image of size 640 x 480 while retaining the same level of quality (Fig. 5). The basic idea is to remove DownConv layers while maintaining the same receptive field size as the DKN, making it possible to obtain the filtering output in a single forward pass. Simply removing these layers leads to an increase in the total number of convolutions (see Section 6). We instead leverage resampling [46] (Fig. 6) that enables covering large receptive fields using small-size convolutions compared to the DKN. This also reduces the total number of network parameters to estimate. Particularly, we resample guidance and interpolated target images separately with stride 4 in each channel, and obtain resampled ones of size H/4 x W/4 x 16C and H/4 x W/4 x 16, respectively, where C is the number of channels. This has an effect of enlarging receptive fields by factor of 4 in each dimension. We may use dilated convolutions [47]

Fig. 4. Efficient implementation using a shift-and-stitch approach. We denote d by the maximum range of the sampling location s(q). We shift the input images and compute the filtering result for each shifted input. We then stitch them up to get a filtering result that has the same resolution as the inputs. Our approach makes it possible to reuse the storage for kernel weights, offsets, and resampled pixels. See text for details. (Best viewed in color.)

Total of N forward passes in case of an image of size \( N = HW \) pixels, where H and W are height and width, respectively.

We use instead the shift-and-stitch approach [31], [45] that stitches the network outputs from shifted versions of the input (Fig. 4). We can obtain the same result as the pixel-wise implementation in 16 forward passes. We first shift input images \( x \) pixels to the left and \( y \) pixels up, once for every \( (x, y) \) where \( \{(x, y) | 0 \leq x, y \leq 3 \} \), and obtain a total of 16 shifted inputs. Each shifted input goes through the network that gives the kernel weights K and the offsets \( \Delta q \) of size \( k^2 x N/16 \) and \( 2k^2 x N/16 \), respectively. The next step is to obtain image values \( f_{s(q)} \) using the sampling function \( s(q) \) from the target image. To this end, starting from every location \( (x, y) \) in the target image, we sample patches of size \( d x d \) with stride 4 in each dimension, each of which gives the output of size \( d^2 x N/16 \). The patch size corresponds to the maximum range of the sampling position \( s(q) \). For an efficient implementation, we restrict the range (e.g., to \( 15 \times 15 \) in our experiment). We then sample \( k^2 \) pixels using the sampling position \( s(q) \) from patches of size \( d x d \), obtaining \( f_{s(q)} \) of size \( k^2 x N/16 \) for each shifted input. To compute a weighted average, we apply element-wise multiplication between the kernel weights \( K \) and the corresponding sampled pixels \( f_{s(q)} \) of size \( k^2 x N/16 \) followed by column-wise summation, resulting in an output of size \( 1 x N/16 \).

Finally, we stitch 16 outputs of size \( 1 x N/16 \) into a single one to get the final output. Note that one can stitch kernel weights and offsets first and then compute a weighted average, but this requires a large amount of memory. We stitch instead the outputs after the weighted average, and reuse the storage for kernel weights, offsets, and sampled pixels.
that support large receptive fields without loss of resolution, but empirically find that runtime gain is marginal (see Section 6). The FDKN takes as inputs resampled guidance and target images, and extracts features for weight and offset regression by 6 convolutional layers of size $3 \times 3$, which maintains a receptive field size comparable with the DKN. Different from the DKN, FDKN regresses kernel weights and offsets for all pixels simultaneously, allowing to obtain the filtering output in a single forward pass. After estimating weights and offset locations, we combine the residuals computed by a weighted average with the interpolated target image. We show in Table 2 the detailed description of the network structure.

4 EXPERIMENTS

In this section we present a detailed analysis and evaluation of our approach. We describe implementation details including our experimental protocol and training/testing setups (Section 4.1). We then apply our models to the tasks of joint depth or saliency image upsampling (Section 4.2), cross-modality image restoration and texture removal\(^1\) (Section 4.3) and semantic segmentation (Section 4.4), and compare them to the state of the art in each case. The results for all comparisons have been obtained from the source code provided by the authors.

4.1 Experimental details

4.1.1 Joint image upsampling and noise removal

Following the experimental protocol of [5], [17], we train our models on the tasks of depth map upsampling and depth noise removal using a large number of RGB/D image pairs, and test them on depth map upsampling and other joint filtering tasks (e.g., cross-modality image restoration) that require selectively transferring the structural details from the guidance image to the target one.

\(^1\) We only show qualitative results in this task since ground truth is not available.
Quantitative comparison with the state of the art on depth map upsampling in terms of average RMSE. Numbers in bold indicate the best performance and underscored ones are the second best.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>4x, 8x, 16x</td>
<td>4x, 8x, 16x</td>
<td>4x, 8x, 16x</td>
<td>4x, 8x, 16x</td>
</tr>
<tr>
<td>Bicubic Int.</td>
<td>4.44, 7.39, 11.87</td>
<td>5.07, 9.22, 14.27</td>
<td>8.16, 14.22, 22.32</td>
<td>6.54, 8.80, 12.17</td>
</tr>
<tr>
<td>MRF [51]</td>
<td>4.26, 7.43, 11.80</td>
<td>4.90, 9.03, 14.19</td>
<td>7.84, 13.98, 22.20</td>
<td>8.81, 11.77, 15.75</td>
</tr>
<tr>
<td>JBU [4]</td>
<td>2.44, 3.81, 6.13</td>
<td>2.99, 5.06, 7.51</td>
<td>4.07, 8.29, 13.35</td>
<td>5.88, 7.63, 10.97</td>
</tr>
<tr>
<td>TGV [3]</td>
<td>3.39, 5.41, 12.03</td>
<td>4.48, 7.58, 17.46</td>
<td>6.98, 11.23, 28.13</td>
<td>12.01, 36.78, 43.89</td>
</tr>
<tr>
<td>Park [2]</td>
<td>2.82, 4.08, 7.26</td>
<td>4.09, 6.19, 10.14</td>
<td>5.21, 9.56, 18.10</td>
<td>9.28, 12.22, 16.51</td>
</tr>
<tr>
<td>FBS [15]</td>
<td>2.58, 4.19, 7.30</td>
<td>3.03, 5.77, 8.48</td>
<td>4.29, 8.94, 14.59</td>
<td>11.96, 12.29, 13.08</td>
</tr>
<tr>
<td>DMSG [21]</td>
<td>1.88, 3.45, 6.28</td>
<td>2.30, 4.17, 7.22</td>
<td>3.02, 5.38, 9.17</td>
<td>5.32, 7.24, 10.11</td>
</tr>
<tr>
<td>DJF [5]</td>
<td>2.14, 3.77, 6.12</td>
<td>2.54, 4.71, 7.66</td>
<td>3.54, 6.20, 10.21</td>
<td>5.51, 7.52, 10.63</td>
</tr>
<tr>
<td>DJFR [17]</td>
<td>1.98, 3.61, 6.07</td>
<td>2.22, 4.54, 7.48</td>
<td>3.38, 5.86, 10.11</td>
<td>5.50, 7.43, 10.48</td>
</tr>
<tr>
<td>FDKN w/o Res.</td>
<td>1.12, 2.23, 4.52</td>
<td>0.85, 2.19, 5.15</td>
<td>1.88, 3.67, 7.13</td>
<td>3.38, 5.02, 7.74</td>
</tr>
<tr>
<td>FDKN</td>
<td>1.08, 2.17, 4.50</td>
<td>0.82, 2.10, 5.05</td>
<td>1.86, 3.58, 6.96</td>
<td>3.36, 4.96, 7.74</td>
</tr>
<tr>
<td>DKN w/o Res.</td>
<td>1.26, 2.16, 4.32</td>
<td>0.99, 2.21, 5.12</td>
<td>1.66, 3.36, 6.78</td>
<td>3.36, 4.82, 7.48</td>
</tr>
<tr>
<td>DKN</td>
<td>1.23, 2.12, 4.24</td>
<td>0.96, 2.16, 5.11</td>
<td>1.62, 3.26, 6.51</td>
<td>3.30, 4.77, 7.59</td>
</tr>
</tbody>
</table>

**Training.** We sample 1,000 RGB/D image pairs of size $640 \times 480$ from the NYU v2 dataset [50]. We use the same image pairs as in [5], [17] to train the networks. We divide each image in the 1,000 training pairs into two halves of size $320 \times 480$ due to the lack of GPU memory, and use the corresponding 2,000 RGB/D image pairs as training samples. We train different models for joint image upsampling and other tasks. For joint image upsampling, the models are trained with a batch size of 1 for 40k iterations, giving roughly 20 epochs over the training data. We synthesize low-resolution depth images ($4 \times 8 \times 16 \times$) from ground truth by bicubic downsampling. The models for depth noise removal are similarly trained but with 4k iterations under the guidance of high-resolution RGB images. Noisy depth images are synthesized by adding Gaussian noise with zero mean and variance of 0.005. We use the Adam optimizer [52] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. As learning rate we use 0.001 and divide it by 5 every 10k iterations. Data augmentation and regularization techniques such as weight decay and dropout [23] are not used, since 1,000 RGB/D image pairs from the NYU dataset have proven to be sufficient to train our models. All networks are trained end-to-end using PyTorch [53].

**Testing.** We can obtain filtering results of DKN for inputs of any size by stitching coarse outputs, which is far more efficient than a brute-force implementation. Filtering results for FDKN are obtained in a single pass. The inputs of our network are 3-channel guidance and 1-channel target images. In case of a 1-channel guidance image (e.g., RGB/NIR image restoration), we create a 3-channel image by duplicating the single channel three times. For multi-channel target images (e.g., in texture removal), we apply our model separately in each channel and combine the outputs.

### 4.1.2 Semantic segmentation

Following the experimental protocol of [54], we plug FDKN in DeepLab-v2 [55], which uses ResNet-101 [38] pretrained for ImageNet classification, as a part of CNNs for semantic segmentation\(^2\), instead of applying a fully connected conditional random field (CRF) [56] to refine segmentation results. That is, we integrate DeepLab-v2 and our model and train the whole network end-to-end, avoiding an offline post-processing using CRFs.

---

\(^2\) We use a PyTorch version available online: https://github.com/isht7/pytorch-deeplab-resnet

**Training.** We use the Pascal VOC 2012 dataset [57] that contains 1,464, 1,449, and 1,456 images for training, validation and test, respectively. Following [54], [55], we augment the training dataset by the annotations provided by [58], resulting in 10,582 images, and use 1,449 images in the validation set for evaluation. We train the network using a softmax log loss with a batch size of 1 for 20k iterations. The SGD optimizer with momentum of 0.9 is used. As learning rate, we use the scheduling method of [55] with learning rate of $2.5 \times 10^{-4}$ and $2.5 \times 10^{-5}$ for DeepLab-v2 and FDKN, respectively.

**Testing.** We upsample 21-channel outputs (20 object classes and background) of DeepLab-v2 before a softmax layer using a high-resolution color image. We apply the FDKN separately in each channel.

### 4.2 Joint image upsampling experiments

We train different models to upsample depth images for scale factors $4 \times 8 \times 16 \times$ with RGB/D image pairs from the NYU v2 dataset [50], and apply them to the tasks of depth map and saliency map upsampling. The inputs to our models are a high-resolution color image and a low-resolution depth image upsampled using bicubic interpolation.

**Depth map upsampling.** We test our models on depth map upsampling with the following four benchmark datasets. These datasets feature aligned color and depth images.

- **Middlebury dataset** [48], [59]: We use the 30 RGB/D image pairs from the 2001-2006 datasets provided by Lu [49].
- **Lu dataset** [49]: This provides 6 RGB/D image pairs acquired by the ASUS Xtion Pro camera [60].
- **NYU v2 dataset** [50]: It consists of 1,449 RGB/D image pairs captured with the Microsoft Kinect [61]. We exclude the 1,000 pairs used for training, and use the rest (449 pairs) for evaluation.
- **Sintel dataset** [18]: This dataset provides 1,064 RGB/D image pairs from the 2001-2006 datasets provided by Lu [49]. We exclude the 1,000 final-pass dataset for testing.

We compare our method with the state of the art in Table 3. It shows the average root mean squared errors (RMSE) between upsampling results and ground truth. The results of DJF [5] and its residual version (DJFR [17]) are obtained by the provided models.
trained with the NYU v2 dataset. DMSG [21] uses the Middlebury and Sintel datasets for training the network. For fair comparison of DMSG with other CNN-based methods including ours, we retrain the DMSG model using the same image pairs from the NYU v2 dataset as in [5], [17]. All other numbers are taken from [5], [17]. From this table, we can see that (1) our models outperform the state of the art including CNN-based methods [5], [17], [21] by significant margins in terms of RMSE, even without the residual connection (DKN w/o Res. and FDKN w/o Res.); (2) they perform well on both synthetic and real datasets (e.g., the Sintel and NYU v2 datasets), and generalize well to other images (e.g., on the Middlebury dataset) outside the training dataset; and (3) FDKN retains the superior performance of DKN, and even outperforms DKN for the Lu dataset. Figure 7 shows a visual comparison of the upsampled depth images $(8 \times)$. The better ability to extract common structures from the target and guidance images by our models here is clearly visible. Specifically, our results show a sharp depth transition without the texture-copying artifacts. In contrast, artifacts are clearly visible even in the results of DJFR, which tends to over-smooth the results and does not recover fine details. This confirms once more the advantage of using spatially-variant kernels and an adaptive neighborhood system in joint image filtering.

Saliency map upsampling. To evaluate the generalization ability of our models on other tasks, we apply them trained with the NYU v2 dataset to saliency map upsampling without fine-
tuning. We downsample saliency maps ($\times 8$) in the DUT-OMRON dataset [63], and then upsample them under the guidance of high-resolution color images. We show in Table 4 a comparison of weighted F-measure [62] between upsampled images and the ground truth. Figure 8 shows examples of the upsampling results by the state of the art and our models. The results show that our models outperform others including a CNN-based method [17].

### 4.3 Other joint filtering experiments

Following [5], [17], we use depth denoising as a proxy task for cross-modality image restoration and texture removal, since ground-truth datasets for these tasks are not available. We train the networks for denoising depth images with RGB/D image pairs from the NYU v2 dataset [50]. The models are then applied to the tasks of cross-modality image restoration and texture removal without fine-tuning. We do not use the residual connection for noise removal tasks, since we empirically find that it does not help in this case. For cross-modality image restoration and texture removal tasks, all previous works (e.g., [5], [7], [9], [17]) we are aware of for these tasks offer qualitative results only. For comparison, average RMSE for GF [7], Yan [9], SDF [6], DJF [5], and DKN on depth noise removal are 5.34, 12.53, 7.56, 2.63, and 2.46, respectively, in the test split of the NYU v2 dataset, showing that our model again outperforms the others.

**Cross-modality image restoration.** For flash/non-flash denoising, we set the flash and non-flash images as guidance and target ones, respectively. Similarly, we restore the color image guided by the flash NIR image in RGB/NIR denoising. Examples for flash/non-flash and RGB/NIR restoration are shown in Fig. 9. Qualitatively, our models outperform other state-of-the-art methods [5], [6], [7], and give comparable results to those of Yan [9] that is specially designed for this task. In particular, they preserve edges while smoothing noise without artifacts. This demonstrates that our models trained with RGB/D images can generalize well for others with different modalities.

**Texture removal.** We set the textured image itself for guidance and target, and apply our models repeatedly to remove small-scale textures. We show examples in Fig. 10. Compared to the state of the art, our models remove textures without artifacts while maintaining other high-frequency structures such as image boundaries and corners. A plausible explanation of why networks trained for denoising depth images work for texture removal is that textures can be considered as patterned noise. Repeatedly applying our networks thus removes them. A similar finding can be found in [5]. We empirically find that 4 iterations are enough to get satisfactory results, and use the same number of iterations for all experiments.

### 4.4 Semantic segmentation experiments

CNNs commonly use max pooling and downsampling to achieve invariance, but this degrades localization accuracy especially at object boundaries [45]. DeepLab [55] overcomes this problem using probabilistic graphical models. It applies a fully connected CRF [56] to the response of the final layer of a CNN. Zheng et al. [64] interpret CRFs as recurrent neural networks which are then plugged in as a part of a CNN, making it possible to train the whole network end-to-end. Recently, Wu et al. [54] have proposed a layer to integrate guided filtering [7] into CNNs.

---

**Fig. 8.** Visual comparison of saliency map upsampling ($8 \times$) on the DUT-OMRON dataset [63].

**Fig. 9.** Examples of cross-modality noise reduction for (top) flash/non-flash denoising and (bottom) RGB/NIR denoising. Our models preserve textures while smoothing noise. GF and DJF tend to over-smooth the textures. Artifacts are clearly visible in the results of SDF. The method of [9], specially designed for this task, gives the best results.
Instead of using CRFs [55], [56], [64] or guided image filtering [7], we apply the FDKN to the response of the final layer of DeepLab v2 [55], before CRFs. Note that FDKN is fully differentiable and the whole network is trained end-to-end. We show in Table 5 mean intersection-over-union (IoU) scores for the validation set in the Pascal VOC 2012 benchmark [57]. To the baseline method (DeepLab v2 w/o CRF [55]), we add CRF [56], guided filtering [54], or FDKN layers. This table shows that our model quantitatively yields better accuracy in terms of mean IoU than other state-of-the-art methods. Examples of semantic segmentation are shown in Fig. 11. Our model outperforms other methods qualitatively as well: It improves the localization of object boundaries (first row) and refines incorrect labels (second row).

5 Discussion

In this section, we conduct an ablation analysis on different components in our models, and show the effects of different parameters for depth map upsampling ($8 \times$) on the NYU v2 dataset. We also discuss other issues including kernel visualization, runtime, and upscaling factors for training and testing.

Network architecture. We show the average RMSE for five variants of our models in Table 6. The baseline model learns kernel weights from the guidance image only. The first row shows that this baseline already outperforms the state of the art (see Table 3). From the second row, we can see that our models trained using the target image only gives better results than the baseline, indicating that using the guidance image only is not enough to fully exploit common structures. The third row demonstrates that constructing kernels from both guidance and target images boosts performance. For example, the average RMSE of DKN decreases from 5.92 to 5.03 for the $3 \times 3$ kernel. We can also see that the
effect of learning kernel weights and offsets from both inputs is significant, and combining all components including the residual connection gives the best results. For example, learning the offsets significantly boosts the performance of DKN from 5.03 to 3.36 for the $3 \times 3$ kernel, and the residual connection further decreases the average RMSE to 3.26. Figure 12 shows a visual comparison of using different networks for depth upsampling. Note that learning to predict the spatial offsets is important because (1) learning spatially-variant kernels for individual pixels would be very hard otherwise, unless using much larger kernels to achieve the same neighborhood size, which would lead to an inefficient implementation, and (2) contrary to current architectures including DJF [5] and DMSG [21], this allows sub-pixel information aggregation.

**Kernel size.** Table 6 also compares the performances of networks with different size of kernels. We enlarge the kernel size gradually from $3 \times 3$ to $25 \times 25$ and compute the average RMSE. From the third row, we observe that the performance improves until size of $15 \times 15$. Increasing size further does not give additional performance gain. This indicates that aggregating pixels from a $15 \times 15$ window is enough for the task. For offset learning, we restrict the maximum range of the sampling position to $15 \times 15$ for all experiments. That is, the filtering results from the third to last rows are computed by aggregating 9, 25 or 49 samples sparsely chosen from a $15 \times 15$ window. The last row of Table 6 suggests that our final models also benefit from using more samples. The RMSE for DKN decreases from 3.26 to 3.19 at the cost of additional runtime. For comparison, DKN with kernels of size $3 \times 3, 5 \times 5$ and $7 \times 7$ take 0.17, 0.18 and 0.19 seconds, respectively, with a Nvidia Titan XP. A $3 \times 3$ size offers a good compromise in terms of RMSE and runtime and this is when we have used in all experiments.

**Feature channels.** In Table 7, we compare the effects of the number of feature channels in terms of RMSE, runtime, the number of network parameters, and model size. We use our DKN and FDKN models including the residual connection and a fixed size of $3 \times 3$ kernels. We vary the number of channels $n_i$ ($i = 1, 2$) in the final two layers for feature extraction. For each network, numbers in bold indicate the best performance and underscored ones are the second best.

![Image](image.png)

**Fig. 12.** Visual comparison of different networks for depth upsampling ($8 \times$) using the kernel of size $3 \times 3$. (a) RGB image. Results for weight regression using (b) RGB, (c) depth and (d) RGB and depth images. Results for weight and offset regression using (e) RGB, (f) RGB and depth images, and (g) RGB and depth images with the residual connection. (h) Ground truth.

<table>
<thead>
<tr>
<th>Weight learning</th>
<th>Offset learning</th>
<th>Res.</th>
<th>$3 \times 3$</th>
<th>$5 \times 5$</th>
<th>$7 \times 7$</th>
<th>$15 \times 15$</th>
<th>$19 \times 19$</th>
<th>$25 \times 25$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGB</td>
<td>Depth</td>
<td>RGB</td>
<td>Depth</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

**TABLE 6**
Average RMSE comparison (DKN/FDKN) of different components and size of kernels (from $3 \times 3$ to $25 \times 25$). From the third row, we can see that aggregating pixels from a $15 \times 15$ window is enough. We thus restrict the maximum range of offset locations to $15 \times 15$. For example, results for $7 \times 7$ in the forth row are computed using 49 pixels sparsely sampled from a $15 \times 15$ window. We omit the results for $15 \times 15, 19 \times 19$ and $25 \times 25$ kernels, since they are equal to or beyond the maximum range of offset locations.

**TABLE 7**
Quantitative comparison of using different number of channels for feature extraction. The results of DKN and FDKN are separated by “^”. We denote by $n_i$ ($i = 1, 2$) the number of channels in the final two layers for feature extraction. For each network, numbers in bold indicate the best performance and underscored ones are the second best.
and second examples can extract diagonal and vertical edges, respectively.

**Runtime.** Table 8 shows runtime comparisons on the same machine. We report the runtime for DMSG [21], DJFR [17] and our models with a Nvidia Titan XP and for other methods with an Intel i5 3.3 GHz CPU. Our current implementation for DKN takes on average 0.17 seconds for images of size 640 × 480. For comparison, the brute-force implementation takes 720 seconds, requiring about 4000× more time. DKN is slower than DMSG [21] and DJFR [17], but yields a significantly better RMSE (Fig. 14 and Table 3). FDKN runs about 15× faster than the DKN and as fast as DJFR with significantly higher accuracy.

**DownConv for DKN.** The DKN without “DownConv” layers can be implemented in a single forward pass, but requires more parameters (∼ 0.47M) to maintain the same receptive field size, with no runtime gain in experiments and a total number of convolutions increasing from 0.6M to 1M at each pixel. We have also tried dilated convolutions [65] that support large receptive fields without loss of resolution. When using the same receptive field size and learnable parameter numbers (∼ 0.5M), even though the runtime decreases from 0.17 to 0.15 seconds, the average RMSE for dilated convolutions increases from 3.26 to 4.30 for depth upsampling (∗ 8) on the NYU v2 dataset, and the storage requirements are multiplied by 16. FDKN thus appears to be the preferable alternative.

**Upscaling factors for training and testing.** Table 9 compares the average RMSE between DKN and DJFR [17] on the NYU dataset [50], when the scale factors for training and test are different. It shows that the performance is degraded for both methods in this case. This may be handled by a scale augmentation technique during training [37]. Another observation is that our model gives a similar result when the scale factor for testing is smaller than that used for training. In contrast, weighted averaging smooths the output and acts as a regularizer, suggesting that our model is not seriously affected by the number of training samples. To demonstrate this, we evaluate the average RMSE in Table 10 when varying the size of the training dataset. For a single training sample, the direct regression may overfit the particular characteristics of training data, especially when the number of training samples is small. In contrast, weighted averaging smooths the output and acts as a regularizer, suggesting that our model is not seriously affected by the number of training samples. To demonstrate this, we evaluate the average RMSE in Table 10 when varying the size of the training data. We train the DKN, FDKN and DMSG [21], where the filtering output is directly regressed from input images, for depth map upsampling (∗ 8) while gradually increasing the number of training samples from 10 to 1,000 in the NYU v2 dataset [50]. We
test them in the same configuration as in Table 3. Table 10 shows that our models are more robust to the size of training data and generalize better to other images (e.g., on the Sintel dataset [18]) outside the training dataset than the direct regression approach, even with more learnable parameters (1.1M for DKN and 0.6M for FDKN vs. 0.43M for DMSG [21]). In particular, the DKN trained with only 10 images outperforms the state of the art by a significant margin for all test datasets (see Table 3). The kernels learned by direct regression are defined implicitly and hard to visualize. In contrast, our method learns sparse kernels (i.e., where to aggregate) explicitly. We can interpret and visualize why kernels learned by our model give smooth results while preserving edges (Fig. 13), and this also gives a clue for tuning hyper-parameters. For example, we can reduce the maximum range of offset locations (i.e., the size of the filter kernel) and the number of weights (i.e., the total number of samples to aggregate), when the weights are concentrated on central parts of the kernels, and a few of them are highly confident, respectively. Note that sparsely aggregating sub-pixel information is not feasible for direct regression approaches (e.g., DMSG [21] and DJF [5]).

Finally, our model can be applied to any tasks requiring an explicit weighted averaging processing beyond (joint) image filtering, as confirmed for the task of semantic segmentation in Section 4.4.

6 Conclusion

We have presented a CNN architecture for joint image filtering that is generic and applicable to a great variety of applications. Instead of regressing the filtering results directly from the network, we use spatially-variant weighted averages where the set of neighbors and the corresponding kernel weights are learned end-to-end in a dense and local manner. We have also presented an efficient implementation that gives much faster runtime than the brute-force one. A fast version further achieves an additional 15 × speed-up without much (if any) loss in performance. Our models generalize well to images that have different modalities from the training data without much (if any) loss in performance. Our models generalize better to other images (e.g., that our models are more robust to the size of training data and test them in the same configuration as in Table 3. Table 10 shows that our models are more robust to the size of training data and generalize better to other images (e.g., on the Sintel dataset [18]) outside the training dataset than the direct regression approach, even with more learnable parameters (1.1M for DKN and 0.6M for FDKN vs. 0.43M for DMSG [21]). In particular, the DKN trained with only 10 images outperforms the state of the art by a significant margin for all test datasets (see Table 3). The kernels learned by direct regression are defined implicitly and hard to visualize. In contrast, our method learns sparse kernels (i.e., where to aggregate) explicitly. We can interpret and visualize why kernels learned by our model give smooth results while preserving edges (Fig. 13), and this also gives a clue for tuning hyper-parameters. For example, we can reduce the maximum range of offset locations (i.e., the size of the filter kernel) and the number of weights (i.e., the total number of samples to aggregate), when the weights are concentrated on central parts of the kernels, and a few of them are highly confident, respectively. Note that sparsely aggregating sub-pixel information is not feasible for direct regression approaches (e.g., DMSG [21] and DJF [5]).

Finally, our model can be applied to any tasks requiring an explicit weighted averaging processing beyond (joint) image filtering, as confirmed for the task of semantic segmentation in Section 4.4.

6 Conclusion

We have presented a CNN architecture for joint image filtering that is generic and applicable to a great variety of applications. Instead of regressing the filtering results directly from the network, we use spatially-variant weighted averages where the set of neighbors and the corresponding kernel weights are learned end-to-end in a dense and local manner. We have also presented an efficient implementation that gives much faster runtime than the brute-force one. A fast version further achieves an additional 15 × speed-up without much (if any) loss in performance. Our models generalize well to images that have different modalities from the training dataset, as demonstrated by our experiments. Finally, we have shown that the weighted averaging process with sparsely sampled 3 × 3 kernels is sufficient to set new state-of-the-art results on several tasks.

Acknowledgments

The authors would like to thank Yijun Li for helpful discussion. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (No. 2017R1C1B2005584), the Louis Vuitton/ENS chair on artificial intelligence and the NYU/Inria collaboration agreement.

References
