

# A learned joint depth and intensity prior using Markov Random fields

Daniel Herrera Castro, Juho Kannala, Peter Sturm, Janne Heikkilä

▶ **To cite this version:**

Daniel Herrera Castro, Juho Kannala, Peter Sturm, Janne Heikkilä. A learned joint depth and intensity prior using Markov Random fields. 3DV 2013 - International Conference on 3D Vision, Jun 2013, Seattle, United States. pp.17-24, 10.1109/3DV.2013.11 . hal-00880486

**HAL Id: hal-00880486**
**https://hal.science/hal-00880486**

Submitted on 6 Nov 2013

**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.

# A Learned Joint Depth and Intensity Prior using Markov Random Fields

Daniel Herrera C.[1], Juho Kannala[1], Peter Sturm[2], and Janne Heikkilä[1]

[1]Center for Machine Vision Research
University of Oulu
{dherrera, jkannala, jth}@ee.oulu.fi
[2]INRIA Grenoble, Rhône-Alpes
peter.sturm@inria.fr

## Abstract

*We present a joint prior that takes intensity and depth information into account. The prior is defined using a flexible Field-of-Experts model and is learned from a database of natural images. It is a generative model and has an efficient method for sampling. We use sampling from the model to perform inpainting and upsampling of depth maps when intensity information is available. We show that including the intensity information in the prior improves the results obtained from the model. We also compare to another two-channel inpainting approach and show superior results.*

## 1. Introduction

Prior models are useful in many image processing applications (*e.g.* denoising, inpainting, stereo, optical flow, etc.) They encapsulate our knowledge and assumptions about the structure of images. Image priors can be applied to any modality (*e.g.* color, infrared, depth). Each modality, however, has its own statistics and thus follows a different model. This paper develops a joint prior for both intensity and depth, thus taking advantage of the implicit relations between the two channels.

There are many cases where an image can contain information from more than one modality. One of the most common cases is a color image with a depth map (an RGBD image). This is a natural result of stereo algorithms and depth cameras (Time-of-Flight, Kinect, etc.) Depth maps are especially interesting because they are a convenient and efficient way of representing the 3D structure of a scene. And they have become the leading standard for the representation and transmission of scenes in 3DTV [14].

It is clear that a relation between intensity and depth exists. Yet modelling the exact relation is not trivial. For example, depth discontinuities often lead to intensity discontinuities but not the other way around. Stereo methods typically model this joint prior implicitly in the algorithm.

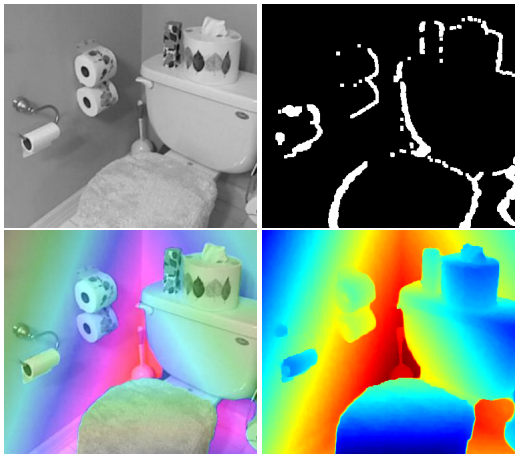


Figure 1: Inpainting a depth map using our prior. Top: intensity and mask of pixels to be inpainted. Bottom: intensity with depth map overlayed (left) and depth (right).

Traditionally, image priors have been hand crafted by researchers based on their assumptions and understanding of the images. For example, robust smoothness priors have been applied based on the assumption that depth maps are mostly smooth with sparse sharp discontinuities. However, this introduces a personal bias. Moreover, there are relations in the data that might not be obvious. Recently, machine learning techniques have been used to learn a flexible model of the data automatically [11]. This allows the algorithm to find complex and subtle relations. The prior models are often learned from a database of real world images, thus capturing the natural image statistics.

We present a joint prior model for depth and intensity images. We learn this prior from a database of images. This prior can be applied to many different image processing applications. Figure 1 shows the result of applying the prior to inpaint the missing areas of a depth map acquired with Kinect.

### 1.1. Motivation

An explicit prior that takes intensity and depth into account is highly desirable. It could be used in a broad range of applications. For example, if intensity and depth are available, the prior can be used to restore (*e.g*. inpainting, denoising, super resolution) both channels simultaneously or only parts of them. In the case of image-based rendering, the prior could be evaluated over the synthesized view to reason on its probability and correct for artefacts.

The model can also be applied to a broad range of input data. The prior can serve as a constraint when modelling a 3D scene from a single color image [18]. On the other hand, if a 3D model is reconstructed from multiple images, it can be rendered to obtain a depth map for each image. Then a prior probability for the 3D model can be obtained via the priors evaluated over these depth maps.

### 1.2. Previous work

*Depth and intensity priors:* Most stereo methods also include a prior on the generated depth map [12]. If a dense depth map is desired, a prior is necessary to regularize areas with low texture. Traditionally, a robust (first order) smoothness prior has been used. However, this favours fronto-parallel planes and leads to a staircase effect, thus higher order priors are recommended [16].

Herrera *et al*. [4] implement a second order smoothness constraint directly in their depth map inpainting algorithm. However, color and depth are used independently without exploiting the joint information. Levin *et al*.'s inpainting algorithm [6] has been succesfully used to inpaint depth maps [13]. Although it was originally meant for colorization, color and depth share a similar relation to intensity and the results are visually pleasing.

Yang *et al*. [17] exploit the relation between intensity and depth to perform super resolution of depth maps. They use bilateral filtering to align the up-sampled depth discontinuities with the high-resolution intensity edges. They obtain visually pleasing results but their formulation cannot be easily extended to other applications. Ghandi *et al*. [2] also start with a low resolution depth map, but they use it to construct a prior for a stereo camera system. Their prior improves the stereo reconstruction, but depends on an active Time-of-Flight camera.

In some situations, as in image-based rendering, the depth map is only an intermediate step. Fitzgibbon *et al*. [1] apply the prior directly on the color of the synthesized view, thus ignoring depth ambiguities arising from similar colors. In the multi-view case, the priors do not necessarily have to be applied on the image level. Gargallo and Sturm [3] implement a multi-view depth map prior in 3D space. It enforces depth map overlap and smoothness with discontinuities. The prior is expensive but is crafted so that it can be efficiently applied to small sets of neighbouring points.

*Natural image statistics:* The literature on this topic is extensive. In the following we only mention approaches directly related to ours. Hyvärinen *et al*. [5] thoroughly examine the statistics of natural intensity images using independent component analysis of small patches. They draw parallels between the obtained filters and human vision receptive fields. They suggest concatenating different channels to extending their approach to color and stereo.

Roth and Black [9] introduced a framework for learning image priors using a high order Markov Random Field (MRF) with large cliques. Instead of modelling independent patches, they model a series of overlapping patches. A filter bank is applied to a patch and each scalar result is fed to an expert function. The model is thus called Fields of Experts (FoE). This model allows them to formulate a joint probability density model for the pixels of an arbitrarily sized image using few parameters. The partition function of their density model is intractable, but they use contrastive divergence for training to avoid it. They show high quality color image inpainting and denoising results using this framework.

Roth and Black obtained results in color images by applying the Fields of Experts to each color channel individually. Although this worked for color images, it ignores the correlations between channels. This is specially important when the channels have different statistics (*e.g*. intensity and depth). However, it is not trivial to extend this to multi-channel images. McAuley *et al*. [7] extended Roth and Black's approach to 3-channel (RGB) images, but they were not able to include the filter coefficients in the learning process. Our method is able to learn the filter coefficients and expert parameters while still handling multi-channel images.

Schmidt *et al*. [10, 11] extended the FoE by using Gaussian scale mixtures (GSM) as expert functions. The model they learned was shown to generate statistics more similar to those of natural images than previously proposed formulations. They also improve the inference stage by using Bayesian minimum mean squared error estimation (MMSE) instead of maximum a posteriori (MAP).

In our approach we extend the model of Schmidt *et al*. to deal with a two-channel image (intensity and depth). The result is a joint prior model that captures the natural relation between intensity and depth. We apply the model to the problem of inpainting semi-dense depth maps. We also

use sampling and MMSE but we alter the inference stage to handle the separate channels.

## 2. A joint MRF prior model

The model we use for our two-channel images is an extension of the single-channel model of [11, 10]. In the following we use the same notation as [10].

Our image prior is based on the flexible Fields of Experts model [9]. It consists of a high-order MRF whose clique potentials model the responses to a bank of linear filters $\mathbf{f}_i$. The probability density of an image $\mathbf{x}$ under the FoE is,

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} e^{-\epsilon||\mathbf{x}||^2/2} \prod_{k=1}^{K}\prod_{i=1}^{N} \phi(\mathbf{f}_i^\top \mathbf{x}_{(k)};\boldsymbol{\alpha}_i), \quad (1)$$

where $\mathbf{x}_{(k)}$ denotes the $k^{\text{th}}$ maximal clique of $\mathbf{x}$ (as explained below), $\phi$ is an expert function, $\boldsymbol{\alpha}_i$ are the parameters for expert $i$, $\mathbf{f}_i$ are the coefficients of a linear filter, and $Z(\boldsymbol{\theta})$ is the partition function that depends on all model parameters $\boldsymbol{\theta} = \{\mathbf{f}_i, \boldsymbol{\alpha}_i | i = 1,\dots,N\}$. The factor $e^{-\epsilon||\mathbf{x}||^2/2}$ is a broad Gaussian ($\epsilon = 10^{-8}$) that regularizes the model even if the experts do not fully constrain the image space [15].

In our case, vector $\mathbf{x}^\top = (\mathbf{u}^\top \mathbf{v}^\top)$ denotes the elements of the two-channel image so that vector $\mathbf{u}$ contains the intensity values of all pixels and vector $\mathbf{v}$ contains the depth values. Further, the intensity and depth values of the $k^{\text{th}}$ maximal clique are denoted by $\mathbf{x}_{(k)}^\top = (\mathbf{u}_{(k)}^\top \mathbf{v}_{(k)}^\top)$. In general, the maximal cliques are square patches of the two-channel image and one may assume that the number of cliques, $K$, is equal to the number of pixels[1].

An important difference from [10] is that each filter $\mathbf{f}_i$ has twice the number of coefficients. In other words, since each patch $\mathbf{x}_{(k)}$ has two channels, each filter has coefficients for both channels, *i.e.* $\mathbf{f}_i^\top = (\mathbf{h}_i^\top \ \mathbf{g}_i^\top)$. The input to the expert function thus has two parts, *i.e.* $\mathbf{f}_i^\top \mathbf{x}_{(k)} = \mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)}$.

Following [15], we use Gaussian scale mixtures (GSMs) [8] as our expert functions because they are flexible and allow very fast sampling. The GSM expert function is defined as

$$\phi(\mathbf{f}_i^\top \mathbf{x}_{(k)};\boldsymbol{\alpha}_i) = \sum_{j=1}^{J} \alpha_{ij} \cdot \mathcal{N}(\mathbf{f}_i^\top \mathbf{x}_{(k)}); 0, \sigma_i^2/s_j), \quad (2)$$

where $\alpha_{ij}$ are the weights of the Gaussian components with scale $s_j$ and base variance $\sigma_i^2$. Following the results obtained by [11] we use a fixed base variance and a wide range of 15 scales $s = \exp(0, \pm1, ..., \pm5, \pm7, \pm9)$ to support a broad range of shapes.

[1]The details of boundary handling are as in [10] but for simplicity one may assume here that the boundary cliques are non-square

## 3. Learning

Since the partition function $Z(\boldsymbol{\theta})$ is unknown in (1), a contrastive divergence method is used for estimation of the model parameters $\boldsymbol{\theta}$, as in [10]. A prerequisite for using contrastive divergence is a fast and rapidly mixing sampling procedure which allows to sample from (1).

To allow fast sampling, a set of hidden variables $\mathbf{z}$ is introduced in [10]. Each discrete variable $z_{ik}$ represents the Gaussian that is active for GSM $i$ and clique $k$. Given $\mathbf{z}$ eq. (1) simplifies to

$$\begin{aligned} p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) &\propto \mathcal{N}\left(\mathbf{x}; \mathbf{0}, \left(\epsilon\mathbf{I} + \sum_{i=1}^{N} \mathbf{W}_i \mathbf{Z}_i \mathbf{W}_i^\top\right)^{-1}\right), \\ &\propto \mathcal{N}\left(\mathbf{x}; \mathbf{0}, \left(\mathbf{W}\mathbf{Z}\mathbf{W}^\top\right)^{-1}\right). \end{aligned} \quad (3)$$

where

$$W = [\mathbf{W}_1, \dots, \mathbf{W}_N, \mathbf{I}] \quad (4)$$

$$Z = \begin{bmatrix} \mathbf{Z}_1 & & \dots & 0 \\ & \ddots & & \vdots \\ \vdots & & \mathbf{Z}_N & \\ 0 & \dots & & \epsilon\mathbf{I} \end{bmatrix} \quad (5)$$

In the original formulation [10, 11], $\mathbf{W}_i$ are filter matrices that correspond to a convolution of the image with filter $i$, and $\mathbf{Z}_i = diag\{s_{z_{ik}}/\sigma_i^2\}$ are diagonal matrices with entries for each expert and clique. Each entry corresponds to the scale of the selected Gaussian for expert $i$ and clique $k$, *i.e.* $z_{ik}$. This clever formulation allows to express the product of expert functions on individual cliques as an efficient series of convolution matrices and use samples from a normal distribution.

In our case, the argument of the expert functions consists of two separate parts, intensity and depth, each with separate filters. The derivation presented in [10] must be extended to account for this. We found that our formulation can be expressed in the same form as eq. (3) by defining $\mathbf{W}_i^\top = \left[\mathbf{H}_i^\top \mathbf{G}_i^\top\right]$, where $\mathbf{H}_i$ and $\mathbf{G}_i$ are filter matrices that correspond to a convolution of the intensity and depth channels with filters $\mathbf{h}_i$ and $\mathbf{g}_i$, respectively. Other aspects of the extension, like the derivatives used for gradient descent, can be derived in a more straightforward manner for the two-channel case. We include the full derivation in the supplemental material for brevity. This allows us to use the same auxiliary-variable Gibbs sampler as described in [10].

## 4. Inpainting two-channel images

We apply our learned joint prior to the problem of inpainting images containing intensity and depth. Inpainting

was also explored as an application in [11]. Because our model can be expressed in a form similar to that in [11], we are able to use a similar sampling strategy. In the inpainting case we seek to conditionally sample values for the missing regions given the known regions of the image. The missing regions can be in the intensity channel, the depth, or both. The most common situation is having a complete intensity image with an incomplete depth map.

To perform this conditional sampling we note that eq. (3) treats the image as a one dimensional vector $\mathbf{x}$ where the intensity and depth channels are simply concatenated. The clique neighbourhoods and the filter coefficients are implicitly coded in the covariance matrix $\Sigma = \left(\mathbf{W}\mathbf{Z}\mathbf{W}^\top\right)^{-1}$. It is possible, however, to reorder $\mathbf{x}$ as long as the covariance matrix is reordered in the same way. Thus we can separate the variables that we want to sample $\mathbf{x}_A$, from those that are known $\mathbf{x}_B$. Leading to

$$\mathbf{x}' = \begin{bmatrix}\mathbf{x}_A \\ \mathbf{x}_B\end{bmatrix}, \quad \Sigma' = \begin{bmatrix}\mathbf{A} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{B}\end{bmatrix}^{-1}, \tag{6}$$

where the submatrix $\mathbf{A}$ has as many rows and columns as the vector $\mathbf{x}_A$ has elements, and so for the others. Once the matrices $\mathbf{W}_i$ have been constructed as described in section 3, we can follow [11]. The final conditional probability is then

$$\begin{aligned} p(\mathbf{x}_A|\mathbf{x}_B, \mathbf{z}; \boldsymbol{\theta}) &\propto \exp\left(-\frac{1}{2}\begin{bmatrix}\mathbf{x}_A \\ \mathbf{x}_B\end{bmatrix}^\top \begin{bmatrix}\mathbf{A} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{B}\end{bmatrix}\begin{bmatrix}\mathbf{x}_A \\ \mathbf{x}_B\end{bmatrix}\right) \\ &\propto \exp\left(-\frac{1}{2}(\mathbf{x}_A + \mathbf{A}^{-1}\mathbf{C}\mathbf{x}_B)^\top \mathbf{A}(\mathbf{x}_A + \mathbf{A}^{-1}\mathbf{C}\mathbf{x}_B)\right) \\ &\propto \mathcal{N}(\mathbf{x}_A; -\mathbf{A}^{-1}\mathbf{C}\mathbf{x}_B, \mathbf{A}^{-1}) \end{aligned} \tag{7}$$

Finally, using a Gibbs sampler we are able to alternatively sample $\mathbf{z}$ and $\mathbf{x}_A$. In line with the findings of [11] we use MMSE estimation to get a value of $\mathbf{x}$ for each chain.

The final inpainting procedure consists of first initializing the missing regions with some value (we explore different options in the experiments). This initial values are used to sample $\mathbf{z}$. Given $\mathbf{z}$, a new estimate for $\mathbf{x}$ is sampled from Eq. (7). These last two steps are alternated to obtain different samples for $\mathbf{x}$. The first samples are discarded as a burn-in to initialize the Gibbs chain. The remaining samples are averaged to obtain the inpainted result from the chain. We use five different initializations, because the inpainted result from the chains don't always agree in ambiguous regions we use the median of the chains' averaged result as the final inpainting.

# 5. Experiments

We trained our model using a publicly available image database. We then applied this learned model to the problems of depth map inpainting and upsampling. The following sections describe the results obtained.

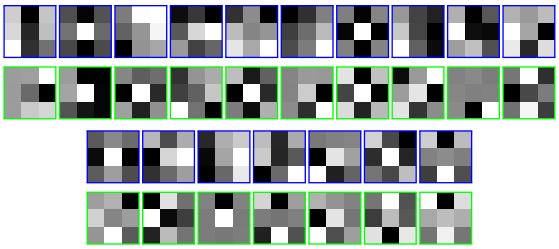


Figure 2: Obtained filters. Blue (1st and 3rd rows) are the intensity filters and green (2nd and 4th rows) are the corresponding disparity filters.

## 5.1. Learning the model

We used the NYU depth dataset V2 [13] to train our model. This dataset contains registered color and depth images taken with a Kinect. It contains only indoor scenes. Therefore the learned model will capture the statistics of indoor images.

There are two practical issues to consider with a Kinect dataset. First, Kinect does not reconstruct all surface types which leads to holes in the resulting depth map. Silberman *et al.* [13] use a colorization approach [6] to inpaint the missing areas in their depth maps. Because we cannot train with missing data we use the inpainted versions for training, even if this slightly modifies the statistics of the images. In the next section we compare our inpainting results with this colorization approach. Second, the alignment between color and depth edges is not always pixel perfect. This is a known artefact of the Kinect. To reduce the effect of this misalignment, we downsample the images by a factor of 4. Thus reducing the misalignment to less than a pixel. This downsampling during training improves the results considerably.

We follow a similar training procedure as used by Schmidt *et al.* We use a bank of 17 3x3 filters. We extracted 2000 patches randomly from the downsampled images. Each patch was 50x50 pixels. For both training and inpainting the second channel contained values in the disparity space (*i.e.* $1/z$) because it better represents the accuracy of the Kinect. The disparity values were also scaled to match the $[0...255]$ range of the intensity channel. Note that this doesn't affect accuracy because the values were not quantized.

Figure 2 shows the learned intensity and disparity filters. Some filters seem to correspond to image derivatives but others are not so easy to interpret. Moreover, sometimes the shape of the intensity and disparity filters is very similar, but this is not always the case. This suggests that some filters model the case when intensity and disparity edges align, while others model the cases when they do not.

### 5.2. Depth inpainting

We compare our approach with two other inpainting strategies. First, we train the single channel FoE model of [11] on the disparity images of our database and use the model for inpainting. We use their own implementation for training and inpainting. The comparison to the single-channel FoE model shows how the addition of the intensity channel improves the inpainting with this method. Second, we compare with the colorization approach of [6]. This shows how our approach compares to another one that uses information from both channels.

We use images from [13] as well as our own images for testing. Kinect labels pixels with no depth, thus it is easy to obtain a mask of pixels to be inpainted. However, because of the inaccuracy of Kinect at depth discontinuities we dilate this mask with a 10x10 rectangle. The resulting inpainting problem is more challenging because the holes are considerably bigger, but this ensures that the corresponding intensity discontinuity will be found inside the hole.

The approach of [11] and the one presented here use Gibbs sampling to obtain samples from the learned distribution. This means that the initialization of the empty regions can influence the result considerably. We used 5 chains with different initializations to explore this effect: zero disparity, max disparity, gaussian noise, median filtering, and linear scanline interpolation. These produce very different starting conditions, as the median filter produces very sharp edges and the linear interpolation is very smooth. The results of the one and two channel algorithms are shown in Figure 3.

The single channel version is clearly much more dependant on the initialization. Because it has no intensity information, it aligns the depth edges randomly. Our approach consistently aligns the depth edges to the intensity edges, as is desired. Yet, there are areas that remain ambiguous for inpainting due to the presence of multiple intensity edges, *e.g.* the top-right corner of Figure 3. To obtain the final result we take the pixel-wise median value over the results obtained with the 5 initializations.

Figure 4 shows some examples of inpainted areas using the three different approaches. Our two channel approach produces clearly superior results to the one channel version of [11]. The depth and intensity edges are aligned and the resulting depth map is more realistic. Levin *et al.* results show bleeding at the edges (*e.g.* see the dinner chair backrests). However, our approach presents smoothing at the depth discontinuities. The edges are not pixel sharp but extend smoothly over 3 pixels. The nature of the artefacts presented by these approaches is different and the quality of the results is similar.

To obtain a quantitative performance measure of the different algorithms we created several artificial holes in the original images. We selected 20 areas of varying structure (flat, single edge, and corners). Then we inpainted a rectangular area around this region with the different algorithms. The Kinect disparities for these areas were considered as ground truth. Some examples of these inpainting results are presented in Figure 6. Table 1 shows the obtained PSNR for each case. The single channel case performs marginally better in the case of small holes (where the edge position cannot vary much), but for medium and larger holes our algorithm is better. Our algorithm also produces better results than Levin *et al.* [6], especially for small and medium holes.



(a) 10x10 (b) 20x20 (c) 40x40

Figure 5: Artificial hole sizes. Green area is the size of the hole to be inpainted. Size in pixels.

| Method | PSNR (dB) | | |
|---|---|---|---|
| | 10x10 | 20x20 | 40x40 |
| 1 channel [11] | **43.98** | 41.54 | 42.19 |
| Levin [6] | 40.91 | 41.60 | 42.42 |
| Ours | 43.81 | **44.15** | **42.77** |

Table 1: Performance with different hole sizes. Kinect data was used as ground truth. Hole size in pixels.

### 5.3. Depth map upsampling

There are cases when the intensity image is available in high resolution, but we only have a low resolution depth map. Upsampling the depth map to match the intensity channel's resolution can be seen as an inpainting task where half (or more) of the pixels are missing. To explore this application we discarded every other value from the original depth map, effectively halving the resolution. The missing values were then inpainted with the three approaches. We considered the original disparity values as ground truth.

Figure 7 shows the absolute difference between the inpainted areas and the original depth map. Table 2 shows the PSNR for each method. Our method clearly produces less errors at the depth discontinuities but also less errors overall, resulting in a higher PSNR.

## 6. Conclusions

We have developed a prior model for images with intensity and depth. The model is able to learn the joint statistics of both channels from a database of images. We applied the prior to the problem of depth map inpainting and upsampling. The results show that including the intensity channel in the prior model improves the results considerably.

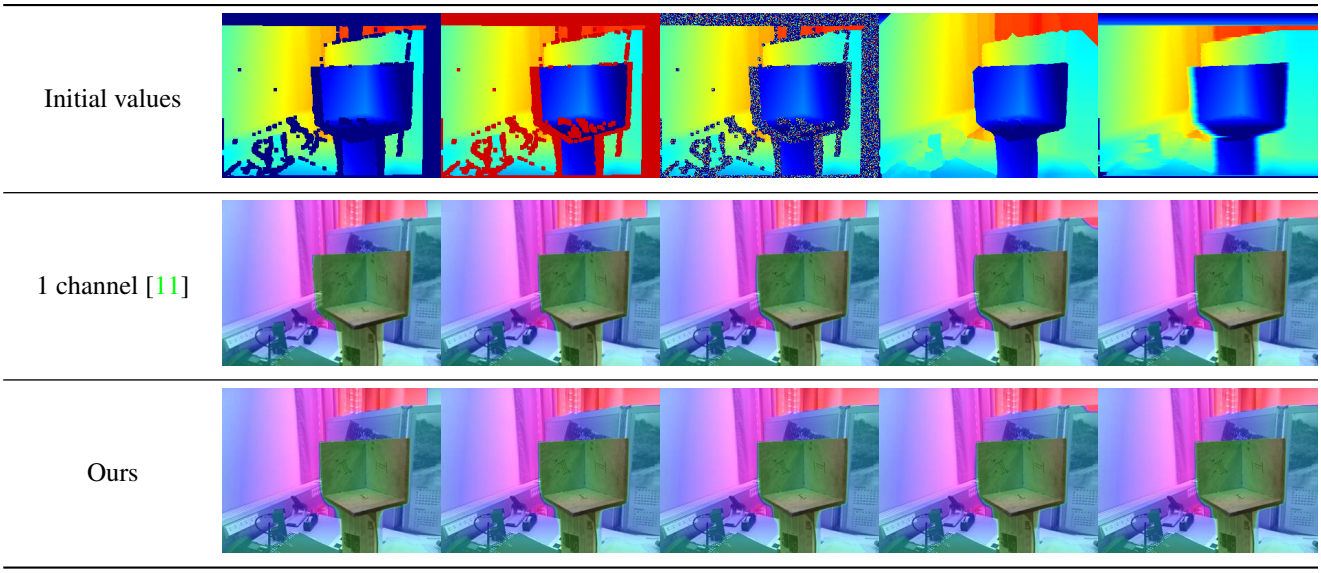



Figure 3: Results of different initializations. The last two rows show an overlay of the inpainted disparity map on top of the original intensity channel. Notice how the depth discontinuities are inpainted in different places when using only 1 channel, whereas our approach aligns them with the intensity edges. 1st row: initial disparity values. 2nd row: results using only the disparity channel [13]. 3rd row: results using our approach.

| Method | PSNR (dB) |
|---|---|
| 1 channel [11] | 34.51 |
| Levin [6] | 32.39 |
| Ours | **35.61** |

Table 2: Performance of inpainting methods in the upsampling task of Fig. 7. Kinect data was used as ground truth.

Thus showing that the model captures the joint distribution between intensity and depth. We also show better results than another inpainting method that utilizes both channels. Moreover, whereas their algorithm is specifically aimed at inpainting, ours is a generative prior model that can be applied to many different problems.

## References


[1] A. Fitzgibbon, Y. Wexler, A. Zisserman, et al. Image-based rendering using image-based priors. In *Proc. ICCV*, volume 2, pages 1176–1183, 2003. 2

[2] V. Gandhi, J. Cech, and R. Horaud. High-resolution depth maps based on ToF-stereo fusion. In *ICRA*, pages 4742–4749, 2012. 2

[3] P. Gargallo and P. Sturm. Bayesian 3d modeling from images using multiple depth maps. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 885–891. IEEE, 2005. 2

[4] D. Herrera C., J. Kannala, and J. Heikkila. Generating dense depth maps using a patch cloud and local planar surface models. In *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2011*. IEEE, 2011. 2

[5] A. Hyvärinen, J. Hurri, and P. Hoyer. *Natural Image Statistics*. Springer, 2009. 2

[6] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. In *SIGGRAPH*, 2004. 2, 4, 5, 6, 7, 8

[7] J. McAuley, T. Caetano, A. Smola, and M. Franz. Learning high-order mrf priors of color images. In *International Conference on Machine Learning*. ACM, 2006. 2

[8] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. In *IEEE TIP*, volume 12, pages 1338–1351, 2003. 3

[9] S. Roth and M. Black. Fields of experts. *International Journal of Computer Vision*, 82, 2009. 2, 3

[10] U. Schmidt. Learning and Evaluating Markov Random Fields for Natural Images. Master's thesis, 2010. 2, 3

[11] U. Schmidt, Q. Gao, and S. Roth. A generative perspective on MRFs in low-level vision. In *CVPR*, 2010. 1, 2, 3, 4, 5, 6, 7, 8

[12] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 519–528. IEEE, 2006. 2

[13] N. Silberman, P. Kohli, D. Hoiem, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. 2, 4, 5, 6




Figure 4: Inpainting results for the different algorithms.

[14] A. Smolic, K. Müller, N. Stefanoski, J. Ostermann, A. Gotchev, G. Akar, G. Triantafyllidis, and A. Koz. Coding algorithms for 3dtv - a survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(11), 2007. 1

[15] Y. Weiss and W. Freeman. What makes a good model of natural images? In *CVPR*, 2007. 3

[16] O. Woodford, P. Torr, I. Reid, and A. Fitzgibbon. Global stereo reconstruction under second-order smoothness priors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(12):2115–2128, 2009. 2

[17] Q. Yang, R. Yang, J. Davis, and D. Nistér. Spatial-depth super resolution for range images. In *CVPR*, 2007. 2

[18] L. Zhang, G. Dugas-Phocion, J. Samson, and S. Seitz. Single-view modelling of free-form scenes. *The Journal of Visualization and Computer Animation*, 13(4):225–235, 2002. 2

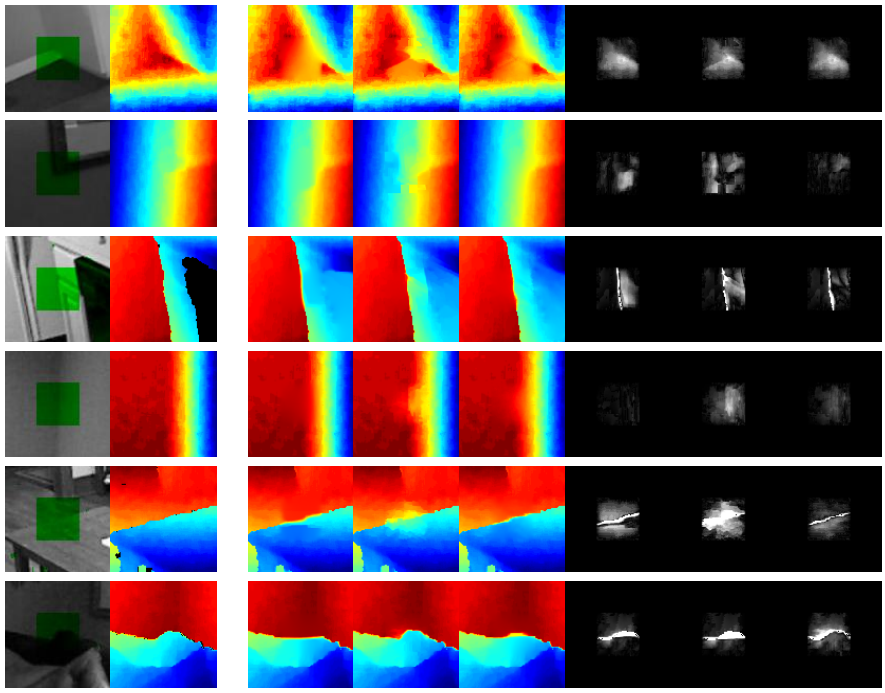



Figure 6: Comparison of inpainted holes with ground truth. (a) Intensity channel (green represents area to be inpainted). (b) Ground truth. (c-e) Inpainting using 1 channel [11], Levin [6] and our approaches respectively. (f-h) Difference between the inpaintings and the ground truth.

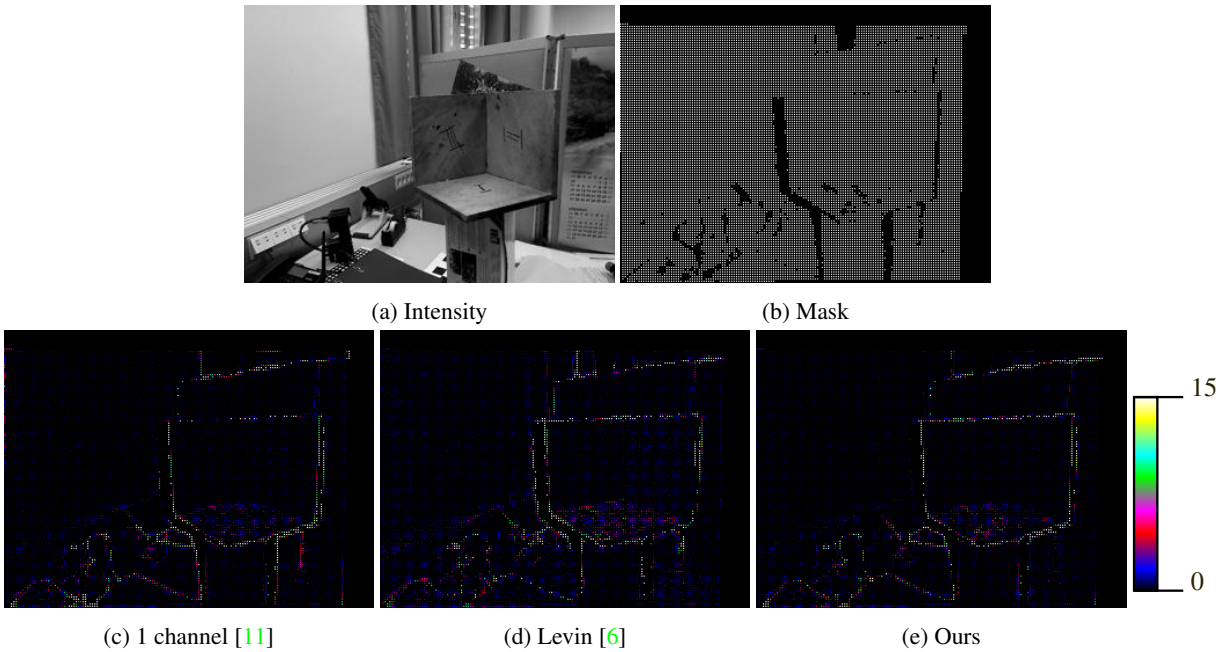



Figure 7: Upsampling of a depth map. Half of the depth values were discarded and inpainted. The bottom row shows the absolute error of the inpainted results. Our approach shows less errors at the depth discontinuities but also lower errors overall.