Learned Collaborative Stereo Refinement

Patrick Knöbelreiter and Thomas Pock
1Institute for Computer Graphics and Vision, Graz University of Technology

Abstract. In this work, we propose a learning-based method to denoise and refine disparity maps of a given stereo method. The proposed variational network arises naturally from unrolling the iterates of a proximal gradient method applied to a variational energy defined in a joint disparity, color, and confidence image space. Our method allows to learn a robust collaborative regularizer leveraging the joint statistics of the color image, the confidence map and the disparity map. Due to the variational structure of our method, the individual steps can be easily visualized, thus enabling interpretability of the method. We can therefore provide interesting insights into how our method refines and denoises disparity maps. The efficiency of our method is demonstrated by the publicly available stereo benchmarks Middlebury 2014 and Kitti 2015.

1 Introduction

Computing 3D information from a stereo pair is one of the most important problems in computer vision. One reason for this is that depth information is a very strong cue to understanding visual scenes, and depth information is therefore an integral part of many vision based systems. For example, in autonomous driving, it is not sufficient to know the objects visible in the scene, but it is also important to estimate the distance to these objects. A lidar scanner is often too expensive and provides only sparse depth estimates. Therefore, it is an interesting alternative to compute depth information exclusively from stereo images. However, the calculation of depth information from images is still a very challenging task. Reflections, occlusions, challenging illuminations etc. make the task even harder. To tackle these difficulties the computation of dense depth maps is usually split up into the four steps (i) matching cost computation, (ii) cost aggregation, (iii) disparity computation and (iv) disparity refinement [27]. In deep learning based approaches (i) and (ii) are usually implemented in a matching convolutional neural network (CNN), (iii) is done using graphical models or 3D regularization CNNs and (iv) is done with a refinement module [29].

There are many approaches to tackle (i)-(iii). However, there are only a few learning-based works for disparity refinement (iv) (see Section [2]). Existing work to refine the disparity map is often based on another CNN using residual connections. In this work we want to overcome these black-box refinement networks with a simple, effective and most important easily interpretable refinement approach for disparity maps. We tackle the refinement problem with a learnable hierarchical variational network. This allows us to exploit both the power of deep learning and the interpretability of variational methods. In order to show the effectiveness of the proposed refinement module, we conduct
Fig. 1: Model Overview. Our model takes three inputs, an initial disparity map, confidence map and the color image. The collaborative hierarchical regularizer iteratively computes a refined disparity map and yields refined confidences and an abstracted color image as a byproduct. The subscripts indicate the level.

experiments on directly refining/denoising winner-takes-all (WTA) solutions of feature matching and as a pure post-processing module on top of an existing stereo method. Fig. 1 shows an overview of our method. Starting from an initial disparity map the final result is iteratively reconstructed.

**Contributions** We propose a learnable variational refinement network which takes advantage of the joint information of the color image, the disparity map and a confidence map to compute a regularized disparity map. We therefore show how our proposed method can be derived from the iterates of a proximal gradient method and how it can be specifically designed for stereo refinement. Additionally, we evaluate a broad range of possible architectural choices in an ablation study. Furthermore, we give insights into how our model constructs the final disparity map by visualizing and interpreting the intermediate iterates. We show the effectiveness of our method by participating on the two complementary public available benchmarks Middlebury 2014 and Kitti 2015.

2 Related Work

We propose a learnable variational model, where we use the modeling power of variational calculus to explicitly guide the refinement process for stereo. Thus, we focus on disparity refinement in the following sections.

**Variational Methods** Variational methods formulate the correspondence problem as minimization of an energy functional comprising a data fidelity term and a smoothness term. We also briefly review variational optical flow methods, because stereo is
Learned Collaborative Stereo Refinement

a special case of optical flow, i.e. it can be considered as optical flow in horizontal direction only. The data-term measures usually the raw intensity difference between the reference view and the warped other view. The regularizer imposes prior knowledge on the resulting disparity map. This is, the disparity map is assumed to be piecewise smooth. Prominent regularizers are the robust Total Variation (TV) and the higher order generalization of TV as e.g. used by Ranftl et al. or by Kuschk and Cremers. Variational approaches have two important advantages in the context of stereo. They naturally produce sub-pixel accurate disparities and they are easily interpretable. However, in order to also capture large displacements a coarse-to-fine warping scheme is necessary. To overcome the warping scheme without losing fine details, variational methods can also be used to refine an initial disparity map. This has e.g. been done by Shekhovtsov et al. who refined the initial disparity estimates coming from a Conditional Random Field (CRF). Similarly, and used a variational method for refining optical flow.

Disparity Refinement Here we want to focus on the refinement of an initial disparity map. The initial disparity map can be e.g. the WTA solution of a matching volume or any other output of a stereo algorithm. One important approach of refinement algorithms is the fast bilateral solver (FBS). This algorithm refines the initial disparity estimate by solving an optimization problem containing an $\ell_2$ smoothness- and an $\ell_2$ data-fidelity term. The fast bilateral solver is the most related work to ours. However, in this work we replace the $\ell_2$ norm with the robust $\ell_1$ norm. More importantly, we additionally replace the hand-crafted smoothness term by a learnable multi-scale regularizer. Another refinement method was proposed by Gidaris and Komodakis. They also start with an initial disparity map, detect erroneous regions and then replace and refine these regions to get a high-quality output. Pang et al. proposed to apply one and the same network twice. They compute the initial disparity map in a first pass, warp the second view with the initial disparity map and then compute only the residual to obtain a high quality disparity map. Liang et al. also improved the results by adding a refinement sub-network on top of the regularization network. We want to stress that the CNN based refinement networks do not have a specialized architecture for refinement as opposed to the proposed model.

Learnable optimization schemes Learnable optimization schemes are based on unrolling the iterates of optimization algorithms. We divide the approaches into two categories. In the first category the optimization iterates are mainly used to utilize the structure during learning. For example in 10 iterations of a TGV regularized variational method are unrolled and used for depth super-resolution. However, they kept the algorithm fixed, i.e. the only learnable parameters in the inference part are the step-sizes. Similarly, in unrolling 10k iterations of the FISTA algorithm is proposed. The second category includes methods where the optimization scheme is not only used to provide the structure, but it is also generalized by adding additional learnable parameters directly to the optimization iterates. For example proposed a primal-dual-network for low-level vision problems, where the authors learned the inference part of a Markov Random Field (MRF) model generalizing a primal-dual algorithm. Chen et al. generalized a reaction-diffusion model and successfully learned a model for im-
4 P. Knöbelreiter and T. Pock

age denoising. Based on [6] a generalized incremental proximal gradient method was proposed in [12], where the authors showed connections to residual units [8]. We built on the work of Chen et al., but specially designed the energy terms for the stereo task. Additionally, we allow to regularize on multiple spatial resolutions jointly and make use of the robust \( \ell_1 \) function in our data-terms.

3 Method

We consider images to be functions \( f : \Omega \rightarrow \mathbb{R}^C \), with \( \Omega \subset \mathbb{N}_2^+ \) and \( C \) is the number of channels which is 3 for RGB color images. Given two images \( f_0 \) and \( f_1 \) from a rectified stereo pair, we want to compute dense disparities \( d \) such that \( f_0(x) = f_1(x - \tilde{d}) \), i.e. we want to compute the horizontal shift \( \tilde{d} = (d, 0) \) for each pixel \( x = (x_1, x_2) \) between the reference image \( f_0 \) and the second image \( f_1 \). Here, we propose a novel variational refinement network for stereo which operates solely in 2D image space and is thus very efficient. The input to our method is an initial disparity map \( \tilde{u} : \Omega \rightarrow [0, D] \), where \( D \) is the maximal disparity, a reference image \( f_0 \) and a pixel-wise confidence map \( c : \Omega \rightarrow [0, 1] \). The proposed variational network is a method to regularize, denoise and refine a noisy disparity map with learnable filters and learnable potential functions. Hence, the task we want to solve is the following: Given a noisy disparity map \( \tilde{u} \), we want to recover the clean disparity with \( T \) learnable variational network steps. We do not make any assumptions on the quality of the initial disparity map, i.e. the initial disparity map may contain many strong outliers.

3.1 Collaborative Disparity Denoising

As the main contribution of this paper, we propose a method that performs a collaborative denoising in the joint color image, disparity and confidence space (see Fig. 2). Our model is based on the following three observations: (i) Depth discontinuities co-inside with object boundaries, (ii) discontinuities in the confidence image are expected to be close to left-sided object boundaries and (iii) the confidence image can be used as a pixel-wise weighting factor in the data fidelity term. Based on these three observations, we propose the following collaborative variational denoising model

\[
\min_{\tilde{u}} \mathcal{R}(\tilde{u}) + \mathcal{D}(\tilde{u}),
\]

where \( \tilde{u} = (u^{rgb}, u^d, u^c) : \Omega \rightarrow \mathbb{R}^5 \), i.e. \( \tilde{u} \) contains for every pixel an RGB color information, a disparity value and a confidence value. \( \mathcal{R}(\tilde{u}) \) denotes the collaborative regularizer and it is given by a multi-scale and multi-channel version of the Fields of Experts (FoE) model [25] with \( L \) scales and \( K \) channels.

\[
\mathcal{R}(\tilde{u}; \theta) = \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{x \in \Omega} \phi_k^l \left( \left( K_k^l A^l \tilde{u} \right)(x) \right),
\]

where \( A^l \) are combined blur and downsampling operators, \( K_k^l \) are linear convolution operators and \( \phi_k^l : \mathbb{R} \rightarrow \mathbb{R} \) are non-linear activation functions. The vector \( \theta \) holds
Learned Collaborative Stereo Refinement

Fig. 2: Collaborative Disparity Denoising. Our method produces three outputs: (a) the refined disparity map, (b) the refined confidence map and (d) the refined color image. (c) shows the ground-truth image for comparison (black pixels = invalid). Note how our method is able to preserve fine details such as the spokes of the motorcycle.

the parameters of the regularizer which will be detailed later. Note that multiple levels allow the model to operate on different spatial resolutions and therefore enables the denoising of large corrupted areas. Intuitively, the collaborative regularizer captures the statistics of the joint color, confidence and disparity space. Hence, it will be necessary to learn the linear operators and the non-linear potential functions from data. It will turn out that the combination of filtering in the joint color-disparity-confidence space at multiple hierarchical pyramid levels and specifically learned channel-wise potential functions make our model very powerful.

\( D(u) \) denotes the collaborative data fidelity term and it is defined by

\[
D(u; \theta) = \frac{\lambda}{2} \| u^{rgb} - f^0 \|^2 + \mu \| u^c - c \|_1 + \nu \| u^d - \tilde{d} \|_{u^c, 1},
\]

where \( \theta \) is again a placeholder for the learnable parameters. The first term ensures that the smoothed color image \( u^{rgb} \) does not deviate too much from the original color image \( f^0 \), the second term ensures that the smoothed confidence map stays close to the original confidence map. Here we use an \( \ell_1 \) norm in order to deal with outliers in the initial confidence map. The last term is the data fidelity term of the disparity map. It is given by an \( \ell_1 \) norm which is pixel-wise weighted by the confidence measure \( u^c \). Hence, data fidelity is enforced in high-confidence regions and suppressed in low-confidence
regions. Note that the weighted $\ell_1$ norm additionally ties the disparity map with the confidence map during the steps of the variational network.

**Proximal Gradient Method (PGM)** We consider a PGM \cite{19} whose iterates are given by

$$u_{t+1} = \text{prox}_{\alpha_t D}(u_t - \alpha_t \nabla R(u_t)),$$

where $\alpha_t$ is the step-size, $\nabla R(u_t)$ is the gradient of the regularizer which is given by

$$\nabla R(u) = \sum_{l=1}^{L} \sum_{k=1}^{K} (K_l^k A_l^k)^T \rho^l_k (K_l^k A_l^k u),$$

where $\rho^l_k = \text{diag}((\phi^l_k)'^T). \text{prox}_{\alpha_t D}$ denotes the proximal operator with respect to the data fidelity term, which is defined by

$$\text{prox}_{\alpha_t D}(\tilde{u}) = \arg \min_u D(u) + \frac{1}{2\alpha_t} \|u - \tilde{u}\|^2_2.$$

Note that the proximal map allows to handle the non-smooth data fidelity terms such as the $\ell_1$ norm. Additionally, there is a strong link between proximal gradient methods and residual units which allows to incrementally reconstruct a solution (see Fig. 1). We provide the computation of the prox-terms in the supplementary material.

**Variational Network** Our collaborative denoising algorithm consists of performing a fixed number of $T$ iterations of the proximal gradient method Eq. (4). In order to increase the flexibility we allow the model parameters to change in each iteration.

$$u_{t+1} = \text{prox}_{\alpha_t \mathcal{D}(\cdot,\theta_t)}(u_t - \alpha_t \nabla R(u_t, \theta_t)), \quad 0 \leq t \leq T - 1$$

Following \cite{6,12} we parametrize the derivatives of the potential functions using Gaussian radial basis functions (RBF)

$$\rho^l_k(s) = \beta^l_k \sum_{b=1}^{B} w^l_{k,b} \exp \left( -\frac{(s - \gamma_b)^2}{2\sigma^2} \right),$$

where $\gamma_b$ are the means regularly sampled on the interval $[-3, 3]$, $\sigma$ is the standard deviation of the Gaussian kernel and $\beta^l_k$ is a scaling factor. The linear operators $K_l^k$ are implemented as multi-channel 2D convolutions with convolution kernels $\kappa_{l,t}^{k}$. In summary, the parameters in each step are given by $\theta_t = \{\kappa_{l,t}^{k}, \beta^l_k, w^l_{k,b}, \mu^l, \nu^l, \lambda^l, \alpha_t\}$.

### 4 Computing Inputs

Our proposed refinement method can be applied to an arbitrary stereo method, provided it comes along with a cost-volume, which is the case for the majority of existing stereo methods.
Learned Collaborative Stereo Refinement

**Probability volume** Assume we have given a cost-volume $v : \Omega \times \{0, \ldots, D-1\} \rightarrow \mathbb{R}$, where smaller costs mean a higher likelihood of the respective disparity values. In order to map the values onto probabilities $p : \Omega \times \{0, \ldots, D-1\}$, we make use of the “softmax” function, that is

$$p(x, d) = \frac{\exp\left(\frac{-v(x,d)}{\eta}\right)}{\sum_{d' = 0}^{D-1} \exp\left(\frac{-v(x,d')}{\eta}\right)},$$

where $\eta$ influences the smoothness of the probability distribution.

**Initial disparity map** From Eq. 9 we can compute the WTA solution by a pixel-wise arg max over the disparity dimension, i.e.,

$$\tilde{d}(x) \in \arg \max_{d} p(x, d).$$

Moreover, we compute a sub-pixel accurate disparity map $\tilde{d}(x)$ by fitting a quadratic function to the probability volume. This is equivalent to perform one step of Newton’s algorithm:

$$\tilde{d}(x) = \tilde{d}(x) - \frac{\delta^+(p(x, \cdot))(\tilde{d}(x))}{\delta^-(\delta^+(p(x, \cdot)))(\tilde{d}(x))},$$

where $\delta^{(+,-)}$ denote standard forward and backward differences in the disparity dimension. Furthermore, we compute the refined value of the probabilities, denoted as $\tilde{p}(x)$, via linear interpolation in the probability volume.

**Initial Confidence Measure** The computation of a confidence measure of the stereo results is important for many applications and a research topic on its own [9]. Here we take advantage of the probabilistic nature of our matching costs $\tilde{p}(x)$. Moreover, we make use of geometric constraints by using a left-right (LR) consistency check, where the left and right images are interchanged. This allows us to identify occluded regions. We compute the probability of a pixel being not occluded as

$$p_o(x) = \frac{\max(\varepsilon - \text{dist}_{lr}(x), 0)}{\varepsilon} \in [0, 1],$$

where

$$\text{dist}_{lr}(x) = |\tilde{d}_l(x) + \tilde{d}_r(x + \tilde{d}_l(x))|$$

is the disparity difference between the left prediction $\tilde{d}_l$ and the right prediction $\tilde{d}_r$ and the parameter $\varepsilon$ acts as a threshold and is set to $\varepsilon = 3$ in all experiments. The final confidence measure is given by

$$c(x) = \tilde{p}(x)p_o(x) \in [0, 1].$$

Thus, we define our total confidence as the product of the matching confidence and the LR confidence. Most of the pixels not surviving the LR check are pixels in occluded regions. To get a good initialization for these pixels as well, we inpaint the disparities of these pixels from the left side. The experiments show that this significantly increases the performance of the model (see Table 1).
5 Learning

In this section we describe our learning procedure for the collaborative denoising model. To remove scaling ambiguities we require the filter kernels $\kappa_{k}^{l,t}$ to be zero-mean and to have an $\ell_2$ norm $\leq 1$. Moreover, we constrain the weights of the RBF kernels to have an $\ell_2$ norm $\leq 1$, too. This is defined with the following convex set:

$$\Theta = \{ \theta_t : \|\kappa_{k}^{l,t}\| \leq 1, \sum_{k} \kappa_{k}^{l,t} = 0, \|w_{k}^{l,t}\| \leq 1 \}$$  \hspace{1cm} (15)

For learning, we define a loss function that measures the error between the last iterate of the disparity map $u^{d}_{s,T}$ and the ground-truth disparity $d^{*}_{s}$. Note that we do not have a loss function for the confidence and the color image. Their aim is rather to support the disparity map to achieve the lowest loss. We use a truncated Huber function of the form

$$\min_{\theta_t \in \Theta} \sum_{s=1}^{S} \sum_{i=1}^{MN} \min \left( |u^{d}_{s,T}(x, \theta) - d^{*}_{s}(x)|_{\delta}, \tau \right)$$  \hspace{1cm} (16)

where $\tau$ is a truncation value, $s$ denotes the index of the training sample and

$$|r|_{\delta} = \begin{cases} \frac{r^2}{2\delta} & \text{if } |r| \leq \delta \\ |r| - \frac{\delta}{2} & \text{else} \end{cases}$$  \hspace{1cm} (17)

is the Huber function.

Implementation details We implemented our model in the PyTorch machine learning framework. We train the refinement module for 3000 epochs with a learning rate of $10^{-3}$ with a modified projected Adam optimizer [10]. We dynamically adjusted the stepsize computation to be constant within each parameter block in our constraint set $\Theta$ to ensure that we perform an orthogonal projection onto the constraint set. After 1500 epochs we reduce the truncation value $\tau$ from $\infty$ to 3.

1 https://pytorch.org
Fig. 4: Qualitative results on the Kitti 2015 test set. Top-to-bottom: Reference image, disparity map which is color coded with blue = far away to yellow = near, error map, where blue = correct disparity, orange = incorrect disparity.

6 Experiments

We split the experiments into two parts. In the first part we evaluate architectural choices based on the WTA result of a matching network and compare with the FBS [1]. In the second part, we use the best architecture and train a variational network for refining the disparity maps computed by the CNN-CRF method [11]. We use this method to participate in the public available stereo benchmarks Middlebury 2014 and Kitti 2015. To ensure a fair comparison we choose methods with similar numbers of parameters and runtimes. Fig. 3 shows how our method constructs the final result. The method recovers step-by-step fine details with the guidance of the confidences and the color image.

Kitti 2015 The Kitti 2015 dataset [17] is an outdoor dataset specifically designed for autonomous driving. It contains 200 images with available ground-truth to train a model and 200 images with withheld ground-truth which is used for testing the models on previously unseen data. The ground-truth is captured using a laser scanner and is therefore sparse in general. The cars are densified by fitting CAD models into the laser point-cloud. We report the badX error metric for occluded (occ) and non-occluded (noc) pixels with $X = 3$. In the badX measure the predicted disparity $\hat{d}$ is treated incorrect, if the distance to the ground-truth disparity $d^*$ is larger than $X$.

Middlebury 2014 The Middlebury 2014 stereo dataset [26] is orthogonal to the Kitti 2015 dataset. It consists of 153 high resolution indoor images with highly precise dense ground-truth. The challenges in the Middlebury dataset are large, almost untextured regions, huge occluded regions, reflections and difficult lighting conditions. The generalization capability of the method is evaluated on a 15 images test-set with withheld ground-truth data. We report all available metrics, i.e., bad{$0.5, 1, 2, 4$} errors, the average error (avg) and the root-mean-squared error (rms).
6.1 Ablation Study

To find the most appropriate hyperparameters for the proposed method, we generate our initial disparity map with a simple feature network. The learned features are then compared using a fixed matching function for a predefined number of discrete disparities.

**Feature Network** Our feature network is a modified version of the U-Net [24, 15] which we use to extract features suitable for stereo matching. We kept the number of parameters low by only using 64 channels at every layer. The output of our feature network is thus a 64-dimensional feature vector for every pixel. The exact architecture is shown in the supplementary material.

**Feature Matching** Next, we use the extracted features $\psi^0$ from the left image and $\psi^1$ from the right image to compute a matching score volume $	ilde{p} : \Omega \times \{0, \ldots, D-1\} \rightarrow \mathbb{R}$ with

$$\tilde{p}(x,d) = \langle \psi^0(x), \psi^1(x - \tilde{d}) \rangle. \quad (18)$$

We follow Section 4 to compute the inputs for the variational network.

**Ablation Study** We systematically remove parts of our method in order to show how the final performance is influenced by the individual parts. Table 1 shows an overview of all experiments. First we investigate the influence of our data-terms, the disparity data-term, the confidence data-term and the RGB image data-term. The study shows that each of the data-terms positively influence the final performance. Especially, adding the original input image significantly increases the performance. This can be e.g. seen in Fig. 3, where the information of how the basket needs to be reconstructed, is derived from the input image. In the second part of the study, we evaluate different variational network architectures. To make the comparison as fair as possible, we chose the variants such that the total number of parameters is approximately the same for all architectures. The experiments show, that a compromise between number of steps, pyramid levels and filter-size yields the best results. The best performing model is the model VN$_{7,5}^4$, where

| Model | WTA FBF | VN$_7^5$ | VN$_7^5$ | VN$_7^5$ | VN$_7^5$ | VN$_7^5$ | VN$_7^5$ | VN$_7^5$ | VN$_7^5$ |
|-------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| Conf  | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       |
| Img   | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       |
| OccIp | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       |
| Joint | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       | ✓       |

Table 1: Ablation study on the Kitti 2015 dataset. Conf = Confidences, Img = Image, OccIp = Occlusion inpainting, Joint = joint training. The super-script indicates the number of steps and the filter-size while the sub-script indicates the number of levels in the variational network. VN$_{7,5}^4$ is therefore a variational network with 7 steps and 4 levels.
6.2 Benchmark Performance

We use our method on top of the CNN-CRF [11] stereo method for the official test set evaluation (see Table 2). We set the temperature parameter $\eta = 0.075$ in all experiments. The average rank is computed with all published methods in the respective benchmarks with a runtime $\leq 20s$. This yields in total 71 methods on the Kitti benchmark and 50 methods on the Middlebury benchmark.

We used the model VN$_7^{[1]}$, on the Kitti dataset. As shown in Table 2 we reduce the bad3 error in both, occluded and in non-occluded regions. Fig. 4 shows qualitative results with the corresponding error maps on the Kitti test set.

On the Middlebury benchmark we use the model VN$_7^{[11]}$ for all evaluations. We compare the errors on the training set with the errors on the test set (Table 2) and observe that: (i) Our method achieves the best scores in all metrics among the compared methods on the training set. (ii) This positive trend is transferred to the test set for the average error and the RMS error. (iii) The bad$_{\{0.5, 1\}}$ errors on the test set are reduced and (iv) the bad$_{\{2, 4\}}$ errors slightly increase on the test set compared to [11]. One reason for this is the limited amount of training data for these very high-resolution images. Fig. 5...
Fig. 5: Qualitative results on the Middlebury test set. Left: Color-coded disparity maps ranging from blue = far away to red = near. Right: Error maps, where white = correct and black = incorrect. The top row shows the initial disparity map (=input to the VN) and the bottom row shows our refined result.

shows a qualitative example of the Middlebury test set. Note that the tabletop is nice and smooth while the sharp edges of the objects are very well preserved.

7 Conclusion & Future Work

We have proposed a learnable variational network for efficient refinement of disparity maps. The learned collaborative and hierarchical refinement method allows the use of information from the joint color, confidence and disparity space from multiple spatial resolutions. In an ablation study, we evaluated a broad range of architectural choices and demonstrated the impact of our design decisions. Our method can be applied on top of any other stereo method, with full use of additional information of a full cost volume. We demonstrated this by adding the variational refinement network on top of the CNN-CRF method and have shown improved results. Furthermore, we have proven the effectiveness of our method by participating in the publicly available stereo benchmarks of Middlebury and Kitti. In future work, we would like to include a matching score during the refinement process and perform data augmentation to increase the training set for learning.

Acknowledgements. This work was partly supported from the ERC starting grant HOMEMOVIS (No.640156).
References

1. Barron, J.T., Poole, B.: The fast bilateral solver. In: European Conference on Computer Vision (ECCV). pp. 617–632 (2016)
2. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sciences pp. 183–202 (2009)
3. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: European Conference on Computer Vision (ECCV). pp. 25–36 (2004)
4. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision pp. 120–145 (2011)
5. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5410–5418 (2018)
6. Chen, Y., Yu, W., Pock, T.: On learning optimized reaction diffusion processes for effective image restoration. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5261–5269 (2015)
7. Gidaris, S., Komodakis, N.: Detect, replace, refine: Deep structured prediction for pixel wise labeling. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5248–5257 (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (June 2016)
9. Hu, X., Mordohai, P.: A quantitative evaluation of confidence measures for stereo vision. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 2121–2133 (2012)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Knöbelreiter, P., Reinbacher, C., Shekhovtsov, A., Pock, T.: End-to-end training of hybrid cnn-crf models for stereo. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2339–2348 (2017)
12. Kohler, E., Klatzer, T., Hammernik, K., Pock, T.: Variational networks: Connecting variational methods and deep learning. In: German Conference on Pattern Recognition (GCPR). pp. 281–293 (2017)
13. Kuschk, G., Cremers, D.: Fast and accurate large-scale stereo reconstruction using variational methods. In: IEEE International Conference on Computer Vision Workshop. pp. 700–707 (2013)
14. Liang, Z., Feng, Y., Guo, Y., Liu, H., Chen, W., Qiao, L., Zhou, L., Zhang, J.: Learning for disparity estimation through feature constancy. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2811–2820 (2018)
15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3431–3440 (2015)
16. Maurer, D., Stoll, M., Bruhn, A.: Order-adaptive and illumination-aware variational optical flow refinement. In: British Machine Vision Conference (2017)
17. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3061–3070 (2015)
18. Pang, J., Sun, W., Ren, J.S., Yang, C., Yan, Q.: Cascade residual learning: A two-stage convolutional neural network for stereo matching. In: IEEE International Conference on Computer Vision Workshop. pp. 887–895 (2017)
19. Parikh, N., Boyd, S., et al.: Proximal algorithms. Foundations and Trends® in Optimization pp. 127–239 (2014)
20. Ranftl, R., Gehrig, S., Pock, T., Bischof, H.: Pushing the limits of stereo using variational stereo estimation. In: IEEE Intelligent Vehicles Symposium. pp. 401–407 (2012)
21. Ranftl, R., Bredies, K., Pock, T.: Non-local total generalized variation for optical flow estimation. In: European Conference on Computer Vision (ECCV). pp. 439–454 (2014)
22. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Epicflow: Edge-preserving interpolation of correspondences for optical flow. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1164–1172 (2015)
23. Riegler, G., Rüther, M., Bischof, H.: Atgv-net: Accurate depth super-resolution. In: European Conference on Computer Vision (ECCV). pp. 268–284 (2016)
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 234–241 (2015)
25. Roth, S., Black, M.J.: Fields of experts. International Journal of Computer Vision (2009)
26. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nesci, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: German Conference on Pattern Recognition (GCPR). pp. 31–42 (2014)
27. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision (2002)
28. Shekhovtsov, A., Reinbacher, C., Graber, G., Pock, T.: Solving dense image matching in real-time using discrete-continuous optimization. Computer Vision Winter Workshop (2016)
29. Tulyakov, S., Ivanov, A., Fleuret, F.: Practical deep stereo (pds): Toward applications-friendly deep stereo matching. In: Proceedings of Advances in Neural Information Processing Systems. pp. 5871–5881 (2018)
30. Vogel, C., Knöbelreiter, P., Pock, T.: Learning energy based inpainting for optical flow. In: Asian Conference on Computer Vision (ACCV) (2018)
31. Vogel, C., Pock, T.: A primal dual network for low-level vision problems. In: German Conference on Pattern Recognition (GCPR) (2017)
32. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l1 optical flow. In: German Conference on Pattern Recognition (GCPR) (2007)
33. Zbontar, J., LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. Journal of Machine Learning Research (2016)
1 Proximal Operators for the data terms

We provide the solution of the proximal operators used in the data terms. The proximal operator is defined as the optimization problem

$$\text{prox}_{\alpha f}(\tilde{u}) = \arg\min_u f(u) + \frac{1}{2\alpha} \|u - \tilde{u}\|^2.$$  

(19)

In the main paper we used the proximal operator with $\ell_1$ and $\ell_2$ functions. The $\ell_2$ prox is used for the color image and the $\ell_1$ prox for the confidences and the disparities. We will present the closed form solutions for these two cases in the next paragraphs.

Proximal operator for $\ell_2$ functions

First, we present the result of the proximal operator for the following $\ell_2$ function

$$f(u) = \frac{\lambda}{2} \|u - u_0\|^2.$$  

(20)

Inserting Eq. (20) into Eq. (19) and setting the derivative w.r.t. $u$ to zero, we can compute the optimal solution $u^*$ with

$$u^* = \frac{\tilde{u} + \tau \alpha u_0}{1 + \lambda \alpha},$$  

(21)

where for the color image data term, $u_0 = I_0$ and $\tilde{u} = u^{rgb}$.

Proximal operator for $\ell_1$ functions

Similarly, we compute the proximal operator of the following weighted $\ell_1$ function

$$f(u) = \gamma \|u - u_0\|_{w,1} = \gamma \sum_{x \in \Omega} w(x) \|u(x) - u_0(x)\|.$$  

(22)

The absolute function is not differentiable at 0 and therefore the optimality condition requires the sub-differential to contain 0. The closed form solution of the proximal operator Eq. (19) with $f$ being the $\ell_1$ function as defined in Eq. (22) is given by

$$u^* = u_0 + \max(0, |\tilde{u} - u_0| - \tau \gamma w) \text{sign}(\tilde{u} - u_0).$$  

(23)

Thus, for the confidence data term $w = 1$ and $u_0 = c$ and for the disparity data term $w = c$ and $u_0 = \tilde{d}$. 
2 Derivative of the quadratic fitting

In the joint training of our feature network and the regularization network we need to backpropagate the gradient through the refined disparities. Therefore, we must compute the gradient of our sub-pixel accurate disparity map w.r.t. the probability volume. The gradient is non-zero only for the supporting points of the quadratic function (shown in blue in Fig. 1). Precisely, it is given by

\[
\frac{\partial \hat{d}(x)}{\partial p(x, d)} = \begin{cases} 
\frac{\delta^+(p(x, \cdot))(\hat{d}(x))}{(\delta^+(p(x, \cdot)))(\hat{d}(x))^{\frac{3}{2}}} & \text{if } d = \hat{d}(x) \\
\frac{\delta^-(p(x, \cdot))(\hat{d}(x))}{(\delta^-(p(x, \cdot)))(\hat{d}(x))^{\frac{3}{2}}} & \text{if } d = \hat{d}(x) - 1 \\
\frac{\delta^-(-p(x, \cdot))(\hat{d}(x))}{(\delta^-(-p(x, \cdot)))(\hat{d}(x))^{\frac{3}{2}}} & \text{if } d = \hat{d}(x) + 1 \\
0 & \text{else,}
\end{cases}
\] (24)

where \(\delta^{(+, -, c)}\) are standard forward-, backward- and central-differences in the disparity dimension. Note, that we overcome the problem of the non-differentiable \(\arg\min\) function with the fitting of the quadratic function. Fig. 1 shows a visualization of the quadratic fitting procedure.

![Fig. 1: Visualization of the quadratic fitting. We select the points next to the maximum value and fit a quadratic function. Computing the extremum of the quadratic functions yields the refined disparity and the refined probability.](image)

3 Architecture of the feature network

The detailed architecture of our feature network is shown in Table 1. It is a variant of a U-Net with two spatial resolutions and 64 channels in every layer. In the contraction part
Learned Collaborative Stereo Refinement

Table 1: Detailed architecture of our multi-level feature network. KS denotes the kernel size, Resolution contains the spatial resolution of the input and output, respectively and Channels contain the number of input and output feature channels, respectively. We use curly brackets to indicate a concatenation of feature maps.

| Layer   | KS             | Resolution        | Channels | Input       |
|---------|----------------|-------------------|----------|-------------|
| conv00  | 3 $W \times H$ | $W / 2 \times H$ | 3 / 64   | Image       |
| conv01  | 3 $W \times H$ | $W / 2 \times H$ | 64 / 64  | conv00      |
| pool0   | 2 $W \times H$ | $W / 4 \times H$ | 64 / 64  | conv01      |
| conv10  | 3 $W \times H$ | $W / 4 \times H$ | 64 / 64  | pool0       |
| conv11  | 3 $W \times H$ | $W / 4 \times H$ | 64 / 64  | conv10      |
| pool1   | 2 $W \times H$ | $W / 8 \times H$ | 64 / 64  | conv10      |
| conv20  | 3 $W \times H$ | $W / 8 \times H$ | 64 / 64  | pool1       |
| conv21  | 3 $W \times H$ | $W / 8 \times H$ | 64 / 64  | conv20      |
| deconv1 | 3 $H \times W$ | $H / 2 \times H$ | 64 / 64  | conv21      |
| conv12  | 3 $W \times H$ | $W / 4 \times H$ | 128 / 64 | {deconv1, conv11} |
| conv13  | 3 $W \times H$ | $W / 8 \times H$ | 64 / 64  | conv12      |
| deconv0 | 3 $H \times W$ | $H / 2 \times H$ | 64 / 64  | conv12      |
| conv02  | 3 $W \times H$ | $W \times H$     | 128 / 64 | {deconv0, conv01} |
| conv03  | 3 $W \times H$ | $W \times H$     | 64 / 64  | conv02      |

the resolution is reduced every 2 layers. In the expansion part features are upsampled with transposed convolutions and concatenated with the extracted features of the same resolution to keep fine details.

4 Learned filters and activation functions

We give further insights into what and how our model learns by visualizing the filters and activation functions learned by our model. Fig. 2a shows selected filter kernels. The first row contains filters for the disparity map, the second row contains filters for the RGB color image and the third row contains filters for the confidence map. Note that the learned filters are interpretable because they contain structure. This suggests that our model captures statistics of how to appropriately denoise disparity maps, confidence maps and color images jointly.

Fig. 2b shows the learned activation functions. We can integrate the learned activation functions (blue) to get the potential functions (green) used in our energy. For example the third potential function has the shape of a truncated Huber function. This means the function has learned to be robust against outliers.

5 Additional Qualitative Results

We use the supplementary material to show additional qualitative results of our method. The results shown in Figs. 3 and 4 impressively show how our simple denoising/refinement scheme is able to reconstruct high quality disparity maps.
Fig. 2: Visualization of filters (left) and activation functions (right) of our model. Filters: Top to bottom: Filters of the disparity map, filters of the RGB color image and filters of the confidence map. Activation Functions: Visualization learned activation functions. The first row (blue) shows the activation functions in the derivative space. The second row (green) shows the integrate of the same functions which correspond to the potential functions in the energy domain. The structure is even better visible if zoom is decreased in a pdf viewer.

The results on the Middlebury benchmark show, how information is exploited from both, the confidences and the input image to reconstruct the best possible disparity map. Note how the blue channel in Fig. 3 seems to be an indicator for regions being occluded, because they appear always next to left sided object boundaries.

As shown in Fig. 4, the results on Kitti show similar properties. Note here how fine details in the background are accurately captured. The confidence maps and the images guide the disparity map where discontinuities are likely to occur and where smoothing is the better option. The former is always close to object boundaries, and the latter is on the street, on cars or in the background.
Fig. 3: Results of VN7\textsuperscript{7,11} on half size (H) Middlebury images. Left to right: Disparity map, confidences and color image. Our model learns to use object edges to guide the denoising of the disparity map. Best viewed with zoom on the PC.
Fig. 4: Results of $\text{VN}_{7.5}$ on the Kitti 2015 dataset. Starting at top-left in clock-wise order: VN disparity maps, confidence map, color images, ground-truth. Note how strong object boundaries are detected in the confidence image and in the RGB image to guide the disparity refinement. Best viewed on the PC.