GEOMETRY-BASED REGULARISATION FOR DENSE IMAGE MATCHING VIA UNCERTAINTY-DRIVEN DEPTH PROPAGATION

Mark Höllmann [1,*], Max Mehltretter [2], Christian Heipke [2]

[1] Plan-Based Robot Control Group, German Research Center for Artificial Intelligence (DFKI), Osnabrück, Germany
mark.hoellmann@dfki.de
[2] Institute of Photogrammetry and GeoInformation, Leibniz University Hannover, Germany
(mehltretter, heipke)@ipi.uni-hannover.de

Commission II, WG II/2

KEY WORDS: Dense Image Matching, Depth Reconstruction, Regularisation, Triangle Mesh, Confidence

ABSTRACT:

In the present work, an uncertainty-driven geometry-based regularisation for the task of dense stereo matching is presented. The objective of the regularisation is the reduction of ambiguities in the depth reconstruction process, which exist due to the ill-posed nature of this task. Based on cost and uncertainty information computed beforehand, pixels are selected, whose depth information can be determined correctly with a high probability. This depth information assumed to be of high confidence is initially used to construct a triangle mesh, which is interpreted as surface approximation of the imaged scene and allows to propagate the confident depth information of the triangle vertices within local neighbourhoods. The proposed method further computes confidence scores for propagated depth estimates, which are used to fuse this depth information with the previously computed cost information, introducing a regularisation into the data term of global optimisation methods. Furthermore, based on the propagated depth information the local smoothness assumption of global optimisation methods is adjusted. Instead of fronto-parallel planes, the method presumes planes, which are parallel to the propagated depth information. The performance of the proposed regularisation approach is evaluated in combination with a global optimisation method. For a quantitative and qualitative evaluation two commonly employed and well-established stereo datasets are used. The proposed method shows significant improvements in accuracy on both datasets and for two different cost computation methods. Especially in unstructured areas, artefacts in the disparity maps are reduced.

1. INTRODUCTION

The reconstruction of depth information from a stereo image pair is a well known task in the field of photogrammetry. In the special case of dense image matching, the depth information of all or at least a large majority of the pixels in a reference image is determined. Depth reconstruction in general has to deal with the challenge that a projection, which takes place when taking a picture, reduces the number of dimensions from 3D to 2D. This leads to a loss of information and characterises the inverse problem, the depth reconstruction, as ill-posed. To reconstruct 3D information nevertheless, correspondences must be matched in at least two images. This is a challenging task, especially in areas of occlusion or with low repetitive structures, where correspondences can not be identified unambiguously. As can be seen in Figure 1, a relatively large percentage of pixels belongs to such challenging areas, making this issue highly relevant. To reduce the ambiguities in a matching process, regularisation approaches, such as local smoothness constraints, are commonly introduced.

For this purpose, in the present work a novel uncertainty-driven method for geometry-based regularisation is presented. Based on a set of pixels for which depth can be correctly estimated with a high probability, the surface of the depicted scene is approximated using a triangle mesh, which constitutes a geometric model. The challenge for this procedure is to create a geometric model fitting the surface of the depicted scene, which is crucial to correctly propagate depth information within a local neighbourhood. In the second step, the propagated depth information is fused with the matching costs computed by an arbitrary dense image matching algorithm taking into account the uncertainties, which are assigned to both kinds of information. Additionally, a local smoothness assumption is derived from the geometric model for a further regularisation of the subsequent

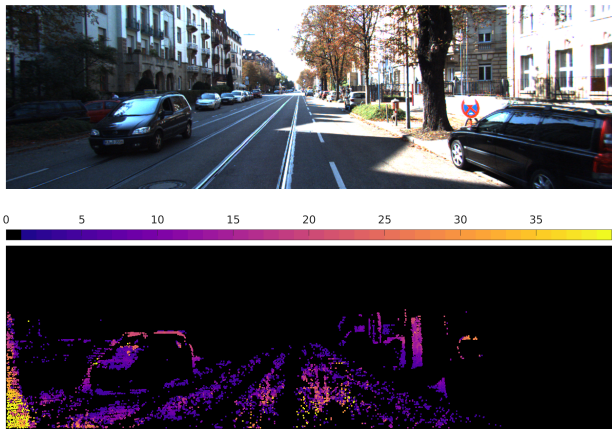


Figure 1. **Example of the KITTI dataset.** RGB reference image and corresponding error map for the disparity estimation (showing the error in pixels), obtained using the Census transformation and semi-global matching. The example shows that a relatively large percentage of pixels belongs to areas which are challenging for dense image matching algorithms, such as the unstructured areas on the street.

*Corresponding author

disparity optimisation process. In consequence, the main assumption of our novel regularisation approach is the piece-wise planarity of the depicted scene, which is true for a large majority of different geometries. Thus, the main contributions of this paper are:

- An uncertainty-driven propagation scheme that allows to disseminate confident depth information over a local neighbourhood with respect to the scene geometry.
- A two-step strategy to integrate the propagated depth information into the data and the smoothness term of a global optimisation process for the purpose of regularisation.
- An extensive evaluation of the behaviour of the proposed method for uncertainty-driven geometry-based regularisation on two common stereo datasets using two different cost computation methods.

The remainder of this work is structured as follows: After the discussion of related work in chapter 2 the methodology of our geometry-based regularisation is presented in Section 3. Afterwards, the proposed regularisation approach is evaluated in Section 4, based on the two well-known and commonly employed stereo datasets Middlebury v3 (Scharstein et al., 2014) and KITTI 2015 (Menze, Geiger, 2015). This work closes with a conclusion and an outlook on further investigations to be carried out in future, in Section 5.

2. RELATED WORK

Motivated by the ill-posed nature of the dense image matching task (Terzopoulos, 1986, Poggio et al., 1987), in recent years, a wide variety of different regularisation procedures were proposed in the literature to minimise the number of ambiguous matches. The objective of all of these methods is to improve the accuracy of an estimated depth map by introducing certain assumptions regarding the depicted scene into the depth reconstruction process, often realised as additional energy term that has to be optimised. The most common of these assumptions is certainly the piece-wise planarity of the scene. For this purpose, different kinds of approaches can be distinguished, which are reviewed in the subsequent paragraphs.

Gradient-based approaches build on the assumption that depth discontinuities mainly occur close to object borders or edges which are visible within an image as colour or grey-scale gradients. Consequently, the majority of an image is assumed to be planar. In global optimisation methods this is frequently implied via gradient-based penalties (Hirschmuller, 2008). However, this approach only allows to introduce relative assumptions regarding the scene geometry, since no support points with known depth are utilised.

The second group of methods utilise plane proposals to penalise or exclude potential matches which imply a depth far away from this plane. Commonly, one plane is estimated per superpixel (Yamaguchi et al., 2014, Guney, Geiger, 2015, Scharstein et al., 2017) or colour segment (Hong, Chen, 2004, Wei, Quan, 2004) of a reference image, implying a planar surface within this region. More sophisticated methods estimate multiple plane proposals based on different segmentation strategies which are subsequently fused into a consistent representation (Li et al., 2016) or differentiate between planar and non-planar scene structures to only regularise image segments for which the assumption of piece-wise planarity applies (Gallup et al., 2010). If support points with known depth are used, plane-based regularisation methods allow absolute assumptions regarding the scene structure. Otherwise, relative assumptions regarding the relations between adjacent segments may be introduced.

The last group that is reviewed within this section uses a set of support points with known depth to build up a triangle mesh (Geiger et al., 2010, Bulatov et al., 2011). Similar to the plane-based approach, the area within a triangle is assumed to be planar and deviations from the triangle surface are penalised. In general, this kind of procedure allows a more detailed approximation of the real scene geometry than the usage of segment-based plane proposals. However, it is typically more sensitive to support points with erroneous depth values.

For the sake of completeness, it should be noted that there are further regularisation approaches that utilise semantic information to compensate for ambiguities in the matching process. Such information is, for example, used to imply a certain shape for parts of the scene geometry, such as generic vehicle models (Guney, Geiger, 2015), or to favour matching of pixels belonging to the same type of object (Stathopoulou, Remondino, 2019). However, since this kind of assumptions generalise badly to different application domains in general, they are assumed to be outside of the scope of the present work.

For most of the previously discussed methods, support points with known depth are beneficial or even crucial, to allow the implication of absolute assumptions regarding the scene geometry. For this purpose, sparse correspondences from various feature matching approaches are commonly employed (Bulatov et al., 2011, Mehltretter et al., 2018). However, depending on the degree of sparsity and the distribution within the image space, the original scene geometry can only be approximated roughly. Furthermore, no quality measure for these correspondences is given, leading to the assumption that all of them are equally reliable. To overcome this limitation, first approaches are presented which use intermediate information from the dense image matching process to obtain more correspondences together with an uncertainty measure for the purpose of regularisation (Spyropoulos et al., 2014).

3. METHODOLOGY

In the present work, a novel triangle mesh-based approach for the regularisation of dense depth reconstruction from epipolar rectified stereo image pairs is presented. Our method leverages support points extracted from an initial disparity map using uncertainty information. This procedure results in a much larger set of support points compared to sparse feature matching and additionally provides uncertainty information which is propagated together with the depth information.

In more detail, the proposed method is divided into three major steps (c.f. Fig. 2 and Fig. 3). First, pixels with reliable disparity are determined, in the sense that their disparity can be estimated unambiguously (c.f. Sec. 3.1). These pixels are used to approximate the surface of the depicted scene using a triangle mesh (c.f. Sec. 3.2). In this step, inter alia, the results of an image segmentation are used to remove surface inconsistent triangles. Based on the resulting surface approximation, the energy terms used in global disparity optimisation procedures are defined (c.f. Sec. 3.3). A flowchart of the standard dense image matching pipeline (Scharstein, Szeliski, 2002), extended by

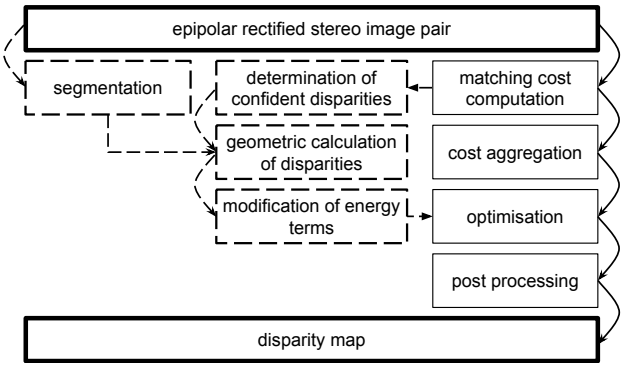


Figure 2. **Flowchart of the proposed method.** The standard elements of the dense image matching pipeline are shown in solid lines, the elements for geometry-based regularisation, as introduced in the present work, in dashed lines.

our geometry-based regularisation approach, is shown in Figure 2. The flowchart illustrates that our geometry-based regularisation seamlessly integrates into standard dense image matching algorithms.

3.1 Identification of pixels with reliable disparity

In the first step, those pixels are determined, whose disparities can be estimated reliably. On the one hand, the disparities from as many pixels as possible should be used to approximate the surface of the scene, since the more pixels are used the more details can be approximated. On the other hand, the disparities have to be determinable unambiguously, because errors in these disparities are propagated through the local neighbourhood and may lead to errors in the final disparity map. In general, a compromise between the density and the error rate has to be found.

The proposed method uses an initial disparity map $\mathbf{D}_{init}$ and an initial confidence map $\mathbf{K}_{init}$ to determine those pixels whose disparities can be estimated with high reliability. For this purpose, pixels with reliable disparity can be identified using a threshold k_{min}, which is defined as minimal confidence. In this context, confidence reflects the likelihood that the estimated disparity is correct. Consequently, the error rate and the density of the resulting confident disparity map $\mathbf{D}_{conf}$ can be balanced via the value of k_{min}. Both, the initial disparity and the initial confidence map, are determined based on the cost volume, which results from the matching cost computation step (c.f. Fig. 2).

The main advantage of this procedure compared to the usage of sparse feature matching, is the significantly higher density of the resulting disparity map $\mathbf{D}_{conf}$ and, therefore, the availability of a higher number of support points for the subsequent processing steps. Additionally, the usage of the cost volume creates no overhead, because it is created in any case during dense image matching. The initial disparity map $\mathbf{D}_{init}$ is directly extracted from the cost volume using a winner takes all strategy, which selects for every pixel the disparity with minimal cost along the disparity axis.

3.2 Joint disparity and uncertainty propagation

In the next step, the previously determined pixels with reliable disparities are used as support points to propagate their disparities over the local neighbourhoods. The objective of this procedure is, to compute correct disparities for as many pixels as possible. One crucial prerequisite for a correct propagation is the identification of correct local neighbourhoods, which, on the other hand, requires a detailed surface approximation of the scene. For this purpose, our method constructs a triangle mesh $\mathbf{T}$ via Delaunay triangulation based on the support points, because in general, such a mesh preserves a majority of detail contained in the scene. However, this triangle mesh has to be filtered, because some triangles may stretch over depth discontinuities, e.g. at object boundaries, and the disparities of some support points may be incorrect. These particular triangles are not surface consistent, leading to the propagation of wrong disparities to the corresponding local neighbourhoods.

3.2.1 Filtering via image segmentation In order to remove triangles which extend beyond depth discontinuities at object boundaries, a segmentation of the reference image is used. It is assumed that each segment corresponds to at maximum one object in the image. Each triangle which extends beyond the boundary of a segment is declared as non-surface consistent and is removed from the triangle mesh.

3.2.2 Filtering via disparity analysis In order to eliminate triangles, which are build on support points with incorrect disparities, the disparity values of all triangle vertices are analysed. The basic idea of this procedure is to identify and remove triangles built over areas which violate the local smoothness assumption. First, the absolute disparity values of the vertices belonging to one triangle are compared. If the difference of two vertices of a triangle is higher than a defined threshold Δd, the triangle is assumed to be surface inconsistent. Afterwards, the relative disparity difference of the vertices belonging to one triangle are analysed. The relative disparity difference is defined as the absolute disparity difference normalised by the distance of the corresponding vertices in image coordinates. If the relative disparity difference is higher than a predefined threshold Δd_{rel}, the triangle is also assumed to be surface inconsistent. All surface inconsistent triangles are removed from the mesh and excluded from further processing.

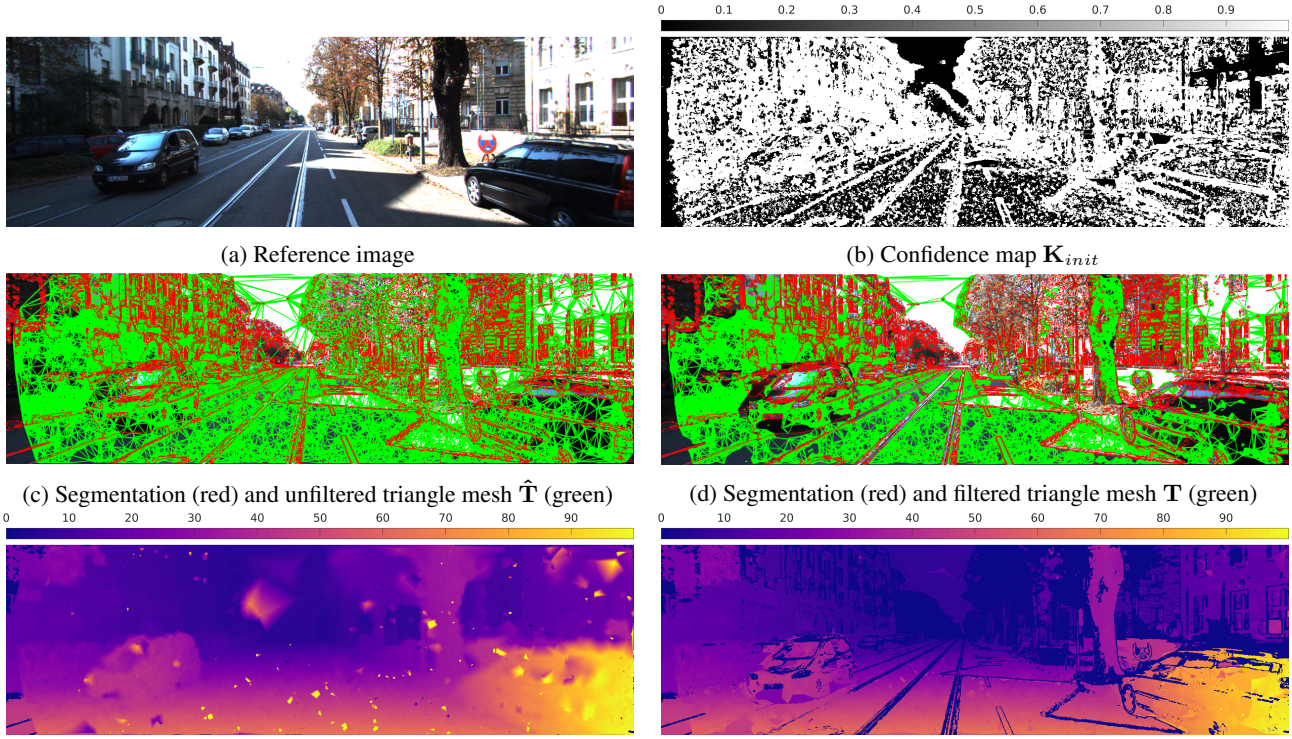
3.2.3 Interpolation Based on the known disparity values $\mathbf{D}_{conf}$ of the triangle vertices, the disparity values of all pixels are linearly interpolated, which are located within the triangles remaining after the filtering step. For this purpose, one plane is utilised per triangle, which is defined based on the corresponding triangle vertices. Each pixel within a triangle is then assigned the disparity value that the plane possesses at the corresponding position, resulting in the disparity values $\mathbf{D}_{\mathbf{T},in}$.

Additionally, we propose to compute a confidence map $\mathbf{K}_{in}$ containing one value for every pixel with an interpolated disparity. The confidence scores describe the probability that an interpolated disparity value is correct. In the first step, the overall confidence of a triangle $t \in \mathbf{T}$ is calculated based on the initial confidence map $\mathbf{K}_{init}$ and the vertices of this triangle $\mathbf{p}_{t,v}$ with $v \in \{A, B, C\}$:

$$k(t, \mathbf{K}_{init}) = \mathbf{K}_{init}(\mathbf{p}_{t,A}) \cdot \mathbf{K}_{init}(\mathbf{p}_{t,B}) \cdot \mathbf{K}_{init}(\mathbf{p}_{t,C}) \,. \quad (1)$$

In more detail, we propose to multiply the individual confidence scores of the corresponding vertices, to obtain the overall confidence k of a triangle. Following the definition in (Bulatov et al., 2011), the confidence of a pixel with interpolated disparity $\mathbf{D}_{\mathbf{T},in}$ further depends on the pixel's position within the triangle:

$$A(\mathbf{p}, \mathbf{T}) = A_{in} \exp\left(-\frac{g(\mathbf{p}, \mathbf{T})}{\sigma_{in}}\right) , \quad (2)$$



(a) Reference image

(b) Confidence map $\mathbf{K}_{init}$

(c) Segmentation (red) and unfiltered triangle mesh $\hat{\mathbf{T}}$ (green)

(d) Segmentation (red) and filtered triangle mesh $\mathbf{T}$ (green)

(e) Disparities $\mathbf{D}_{\hat{\mathbf{T}}}$ computed based on the unfiltered triangle mesh

(f) Disparities $\mathbf{D}_{\mathbf{T}}$ computed based on the filtered triangle mesh

Figure 3. **Overview of the proposed method on an example of the KITTI 2015 dataset.** Based on sparse support points, which are extracted from an initial disparity map using the information from an initial confidence map $\mathbf{K}_{init}$ (c.f. (b), showing low confidence scores in black and high ones in white), a triangle mesh $\mathbf{T}$ is constructed. The triangles are filtered using an image segmentation, which reduces over-smoothing at depth discontinuities and eliminates outliers (c.f. (e) and (f), showing small disparities in blue to large ones in yellow).

where the value of the function depends on the amplitude A_{in} and on the factor σ_{in}, which defines the width of the function, as well as on the distance $g(\mathbf{p}, \mathbf{T})$ to the closest vertex of the triangle. This function has its maxima at the positions of the vertices and its minimum in the centre of the triangle. This models the decreasing significance of the vertices towards the centre. Finally, the combination of Equations 1 and 2 define the confidence K_{in} of the interpolated disparity values:

$$K_{in}(\mathbf{p}, \mathbf{T}, \mathbf{K}_{init}) = k(t, \mathbf{K}_{init}) \cdot A(\mathbf{p}, \mathbf{T}) \,. \tag{3}$$

3.2.4 Extrapolation To increase the coverage of the reference image with triangle mesh-based disparities, not only interpolation but also extrapolation is used. To be more precise, for every pixel which has no interpolated disparity value assigned, but is located within a segment which contains a minimum number of triangles, a disparity value is computed via extrapolation. In consequence, the set of disparities from interpolated areas $\mathbf{D}_{\mathbf{T},in}$ is supplemented by the set of extrapolated disparities $\mathbf{D}_{\mathbf{T},ex}$ to form the set of all triangle mesh-based disparities $\mathbf{D}_{\mathbf{T}}$.

In principle, it is possible to extrapolate disparity values based on a single triangle, following the basic procedure of the previously described interpolation approach. However, the significance of a triangle decreases rapidly with increasing extrapolation distance. To circumvent this problem, not only one triangle is used to estimate a pixel's disparity but all triangles of the image segment the particular pixel is located in. For this purpose, one disparity value is computed per triangle in the corresponding segment and the pixel is assigned the median of these $n_{\mathbf{p},ex}$ disparity values. Additionally, the standard deviation $\sigma_{\mathbf{p}_{ex}}$ of these disparity values is computed and used to estimate the confidence K_{ex} of the extrapolated disparity value:

$$K_{ex}(\sigma_{\mathbf{p}_{ex}}) = \begin{cases} A_{ex} \exp\left(-\frac{\sigma^2_{\mathbf{p}_{ex}}}{\sigma^2_{ex}}\right) & \text{if } n_{\mathbf{p},ex} \geq n_{ex,min} \\ 0 & \text{else} \end{cases}, \tag{4}$$

where, in analogy to Equation 3, A_{ex} encodes the amplitude and σ_{ex} controls the width of the function. The computed confidence is only assigned to a pixel if a minimum number $n_{\mathbf{p},ex}$ of triangles is located within the particular image segment and therefore if a minimum number of extrapolated disparities exist for the pixel $\mathbf{p}$. This minimum number is necessary, since the standard deviation of extrapolated disparities is not significant if too few values are used.

3.3 Definition of the energy term

Commonly, the energy term of global optimisation methods, used to extract a disparity map from a cost volume which is optimal for all pixels $\mathbf{p}$ of a reference image, consists of two sub-energy terms:

$$E_{total}(\mathbf{D}, \mathbf{T}, \mathbf{K}_{init}) = \sum_{\mathbf{p}} E_{data,GR}(\mathbf{p}, \mathbf{D}, \mathbf{T}, \mathbf{K}_{init}) + E_{smooth}(\mathbf{p}, \mathbf{D}, \mathbf{T}) \,, \tag{5}$$

where the data term E_{data} utilises the information from a pixel-wise correspondence identification encoded in a cost volume and the smoothness term E_{smooth} realises a local smoothness

assumption. $\mathbf{D}$ is the disparity map to be optimised, $\mathbf{T}$ is our filtered triangle mesh and $\mathbf{K_{init}}$ is an initial confidence map. Details on both terms are given in the subsequent paragraphs.

3.3.1 Data term As data term E_{data}, we propose to use a weighted average of the cost volume $\mathbf{C}$ and the costs C_T computed from the triangle mesh-based disparities $\mathbf{D_T}$:

$$E_{data}(\mathbf{p}, \mathbf{D}, \mathbf{T}, \mathbf{K}_{init}) = \frac{\mathbf{K}_{init}(\mathbf{p})\mathbf{C}(\mathbf{p}, d_\mathbf{p}) + K_T(\cdot)C_T(\cdot)}{\mathbf{K}_{init}(\mathbf{p}) + K_T(\cdot)}, \tag{6}$$

where $d_\mathbf{p} \in \mathbf{D}$, $d_{\mathbf{T},\mathbf{p}} \in \mathbf{D_T}$ and K_T contains the confidence information estimated based on the triangle mesh:

$$K_T(\mathbf{p}, \mathbf{T}, \mathbf{K}_{init}) = \begin{cases} K_{in}(\mathbf{p}, \mathbf{T}, \mathbf{K}_{init}) & \text{if } \mathbf{p} \in \mathbf{P_{in}} \\ K_{ex}(\sigma_{\mathbf{T},\mathbf{p}_{ex}}) & \text{if } \mathbf{p} \in \mathbf{P_{ex}} \\ 0 & \text{else} \end{cases}, \tag{7}$$

where $\mathbf{P_{in}}$ and $\mathbf{P_{ex}}$ are the sets of pixels with interpolated and extrapolated disparity values, respectively. According to (Bulatov et al., 2011), the costs C_T which are computed based on the triangle mesh, depend on the distance between the current disparity value $d_\mathbf{p}$ of a pixel and the disparity $d_{\mathbf{T},\mathbf{p}}$ which was obtained via our propagation scheme:

$$C_T(d_\mathbf{p}, d_{\mathbf{T},\mathbf{p}}) = \min\left(\frac{|d_\mathbf{p} - d_{\mathbf{T},\mathbf{p}}|}{d_0}, 1\right), \tag{8}$$

where d_0 defines the maximum distance to the prediction before the cost is set to maximum value. Finally, the cost volume is weighted by the initial confidences K_{init} and the costs computed from the triangle mesh-based disparities is weighted by the propagated confidences K_{in} and K_{ex}, respectively. Due to the confidence weighted average, the influence of the triangle mesh-based disparities is high in areas with a high propagated confidence and low in areas with a low propagated confidence. Consequently, low-confident and therefore probably wrong information originating from the proposed regularisation approach has a low impact on the final disparity estimation.

3.3.2 Smoothness term To realise the smoothness assumption of piece-wise planarity of a depicted scene, we force adjacent pixels to have disparity values located on one common plane. For this purpose, the smoothness term is designed according to (Scharstein et al., 2017) and prefers planes which are parallel to the gradient implied by the triangle mesh-based disparities, instead of using fronto-parallel planes:

$$E_{smooth}(\mathbf{p}, \mathbf{D}, \mathbf{T}) = \sum_{\mathbf{p}_j \in \mathcal{N}} V_{GR}(d_\mathbf{p}, d_{\mathbf{p}_j}, \mathbf{T}), \tag{9}$$

where pixels $\mathbf{p}_j$ are located within a local neighbourhood of $\mathbf{p}$. If the confidence scores of the triangle mesh-based disparities at the positions of the pixels $\mathbf{p}$ and $\mathbf{p}_j$ are higher than a minimum confidence $k_{min,in}$ for interpolated areas or $k_{min,ex}$ for extrapolated areas, their gradient is used to infer a disparity plane. Otherwise a fronto-parallel plane is used:

$$V_{GR}(\cdot) = \begin{cases} V(d_\mathbf{p} + j_\mathbf{p}, d_{\mathbf{p}_j}) & \text{if } \mathbf{p} \in \mathbf{P_a} \\ & \wedge K_a(\mathbf{p}, \cdot) > k_{min,a} \\ & \wedge \mathbf{p}_j \in \mathbf{P_b} \\ & \wedge K_b(\mathbf{p}_j, \cdot) > k_{min,b} \\ V(d_\mathbf{p}, d_{\mathbf{p}_j}) & \text{else} \end{cases}, \tag{10}$$

where $a, b \in \{in, ex\}$, $j_\mathbf{p} = d_{\mathbf{p}_j} - d_\mathbf{p}$ and the first-order smoothness term V with penalties P_1 and P_2 is defined as:

$$V(d, \hat{d}) = \begin{cases} 0 & \text{if } d = \hat{d} \\ P_1 & \text{if } |d - \hat{d}| = 1 \\ P_2 & \text{else} \end{cases}. \tag{11}$$

4. EXPERIMENTAL RESULTS

In this section our approach for uncertainty-driven geometry-based regularisation is evaluated. In the first step, optimal hyper-parameters are determined using a parameter study. Afterwards, the proposed method and the determined parameters are validated using two different well-known stereo datasets: The Middlebury v3 dataset shows indoor scenes and provides a dense ground truth, which was generated using structured light. The KITTI 2015 dataset, on the other hand, shows different scenes of urban and rural traffic. Thus, the scenes are fundamentally different from those contained in the Middlebury dataset. For approximately 30 % of the pixels, reference disparities are available, which were captured using a laser-scanner.

Additionally to the two datasets, we also test our approach on two different cost computation methods. While the Census transformation (Zabih, Woodfill, 1994) is a hand-crafted similarity measure, MC-CNN (Zbontar, LeCun, 2016) predicts the matching cost of image patches based on a Convolutional Neural Network. In both cases, the final disparity map is obtained using semi-global matching (Hirschmuller, 2008). The initial confidence map (c.f. Sec. 3.1) is obtained using CVA-Net (Mehltretter, Heipke, 2019), since this method shows superior results and can operate directly on the cost volume. For the task of image segmentation (c.f. Sec. 3.2), the Quick Shift Algorithm (Vedaldi, Soatto, 2008) is used, because it can handle an unknown and varying number of segments, is robust against noise and is independent from any seed values.

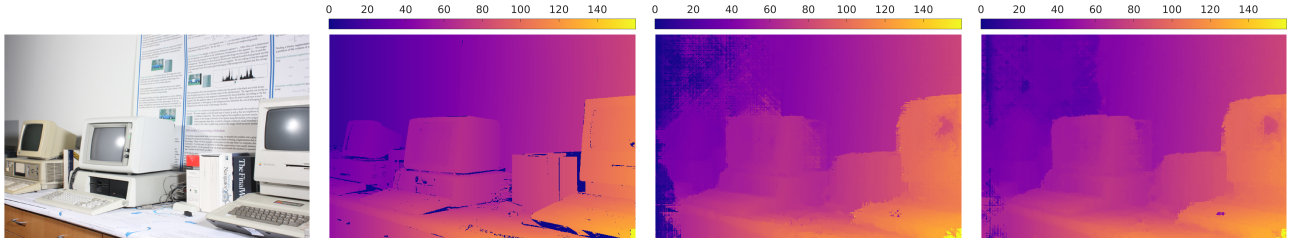
4.1 Parameter study

In a first step, the optimal parameters of the proposed geometry-based regularisation approach are determined using Monte Carlo simulations. The quality of the triangle mesh, the data term and the smoothness term have a linear dependency, in the sense that they only depend on the previously mentioned steps. In consequence, three different Monte Carlo simulations are performed to obtain optimal hyper-parameter values for the individual processing steps. The hyper-parameter values of previous simulations are kept constant afterwards. This significantly decreases the overall amount of necessary runs, compared to the determination of all hyper-parameter values in just one such simulation. Each Monte Carlo simulation is performed with 200 iterations using the first seven images of the Middlebury dataset and the Census transformation as cost computation method.

The determined parameters are presented in Table 1. For the Census transformation, a square with a side length of $h_C = 5$ px is used as transformation region. The penalties P_1 and P_2 of semi-global matching are chosen according to (Mehltretter et al., 2018).

4.2 Validation on Middlebury v3

Building on the determined parameters, the proposed approach is validated on the eight remaining image pairs of the Middlebury v3 dataset. For the purpose of consistency, the definition



(a) Reference image (b) Ground truth (c) Without regularisation (d) With our regularisation

Figure 4. **Qualitative results on the Middlebury v3 dataset.** The disparity maps (showing disparities in pixel, from small ones in blue to large ones in yellow) are obtained using the Census transformation with semi-global matching. It can be seen that our regularisation approach smooths the disparity map and reduces artefacts, while preserving sharp object boundaries and depth discontinuities.

creation of the triangle mesh			
h_s	8 px	h_r	6
k_{min}	90 %	Δd	5 px
Δd_{rel}	2		
modification of the data term			
d_0	50 px	$n_{\mathbf{p},ex}$	15
σ_{in}	10 px	σ_{ex}	0.5 px
modification of the smoothness term			
$k_{min,in}$	0.0	$k_{min,ex}$	0.5
external parameters			
h_C	5 px	P_1	5/7
P_2	30/7		

Table 1. **Utilised parameter values.** These parameter values of the proposed geometry-based regularisation method are estimated via Monte Carlo simulation.

of an error is taken from (Menze, Geiger, 2015) for both validation sections: A pixel's disparity is assumed to be correct if the difference between the estimated and the reference disparity is lower than 3 px or lower than 5 % of the true disparity.

For the purpose of evaluation, different configurations of the proposed method are tested. Configuration 1 represents the baseline and does not employ the regularisation approach proposed in the present work. Configuration 2 extends configuration 1 by the modification of the data term (c.f. Eq. 6), using interpolated disparities. In addition, configuration 3 also considers extrapolated disparities. Configurations 4 and 5 correspond to configurations 2 and 3, respectively, but further employ our proposed smoothness term (c.f. Eq. 9). As already mentioned in the introduction of this section, all five configurations are performed twice, once with the Census transformation and once with MC-CNN as cost computation method. For all configurations, the overall error rate ϵ_M as well as the error rate within unstructured areas ϵ_{unstr} are computed and listed in Table 2. For the identification of unstructured areas, the definition and the method described in (Scharstein, Szeliski, 2002) are used.

The quantitative results for this dataset (c.f. Tab. 2) show that the proposed data and smoothness terms both have a positive impact and reduce the overall error rate, if only interpolated disparities are used. This effect of an reduced error rate can be observed for both kinds of evaluated cost computation methods. The disparity map corresponding to the image 'vintage' resulting from configuration 4 (c.f. Fig. 4d) supports the quantitative results. Compared to the disparity map created without the proposed method (c.f. Fig. 4c), improvements can especially be seen in planar, unicoloured regions, like the wall or the desk.

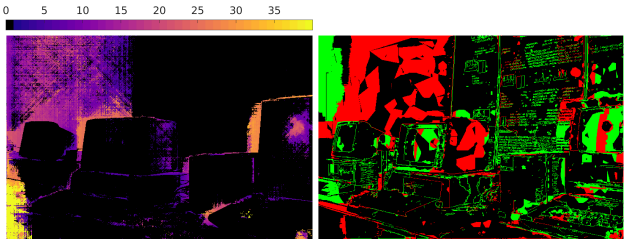


Figure 5. **Analysis of the extrapolated disparities.** The left image shows the error in pixel corresponding to the disparity map shown in Figure 4c, with a low error in black to a large one in yellow. The right image differentiates between pixels with no extrapolated disparity (black), correctly extrapolated disparity (green) and incorrectly extrapolated disparity (red), according to the error definition introduced at the beginning of Section 4.2. It can be seen that pixels with incorrect disparity estimates are located in the same areas for the extrapolation and the original dense stereo matching method.

Even though the error rate of the extrapolated disparities is lower than 40 % (c.f. Tab. 3), these disparity values have only a minor impact on the results. The reason for this is that the confidence scores of extrapolated disparities are high in areas where the dense stereo matching method can estimate the correct disparity already without the proposed regularisation. In areas where the baseline estimates incorrect disparities, the confidence of extrapolated disparities is also low (c.f. Fig. 5).

The error within unstructured areas $\epsilon_{M,CT,unstr}$ decreases, when the Census transformation and the proposed smoothness term are used. The additional usage of the data term, on the other hand, increases the error in this areas. Consequently, the proposed smoothness term seems to be well suited for unstructured areas if employed together with the estimated hyper-parameters and the Census transformation. However, using the MC-CNN as cost computation method, both energy terms increase the error. This indicates, that the estimated hyper-parameters do not generalise well over different cost computation methods for unstructured areas. However, if not only the impact on unstructured areas but the overall error rate and the qualitative changes on the disparity map are considered, it can be stated that the chosen parameters are well-suited for cost volumes originating from different cost computation methods.

4.3 Cross-validation on KITTI 2015

For the purpose of cross-validation, the KITTI 2015 dataset is used. As before, the five configurations are evaluated for the two cost computation methods. The corresponding quantitative results can also be found in Table 2. For the cost volumes

#	KITTI				Middlebury			
	$\epsilon_{K,CT}$	$\epsilon_{K,CT,unstr}$	$\epsilon_{K,MC}$	$\epsilon_{K,MC,unstr}$	$\epsilon_{M,CT}$	$\epsilon_{M,CT,unstr}$	$\epsilon_{M,MC}$	$\epsilon_{M,MC,unstr}$
1	14.6 %	20.0 %	9.2 %	15.7 %	21.0 %	29.8 %	16.6 %	39.4 %
2	10.4 %	11.2 %	8.7 %	16.7 %	19.4 %	32.4 %	16.3 %	39.9 %
3	10.4 %	11.2 %	8.7 %	16.7 %	19.4 %	32.4 %	16.3 %	39.9 %
4	10.3 %	11.6 %	9.0 %	17.4 %	18.7 %	27.8 %	15.9 %	41.7 %
5	10.3 %	11.8 %	9.0 %	17.1 %	18.7 %	27.9 %	15.8 %	41.7 %

Table 2. **Quantitative results.** Overall error ϵ and error in unstructured areas ϵ_{unstr} on images of the KITTI 2015 K and the Middlebury v3 dataset M, using two different cost computation methods: Census transformation CT and MC-CNN MC. The following configurations are evaluated: #1: SGM, #2: SGM and modification of the data term with interpolated disparities, #3: SGM and modification of the data term with interpolated and extrapolated disparities, #4: SGM and modification of the data and smoothness term with interpolated disparities, #5: SGM and modification of the data and smoothness term with interpolated and extrapolated disparities

Dataset	Error type	Error value
KITTI	$\epsilon_{in,K,CT}$	3.5 %
	$\epsilon_{ex,K,CT}$	32.9 %
	$\epsilon_{in,K,MC}$	5.1 %
	$\epsilon_{ex,K,MC}$	32.9 %
Middlebury	$\epsilon_{in,M,CT}$	9.1 %
	$\epsilon_{ex,M,CT}$	38.3 %
	$\epsilon_{in,M,MC}$	6.7 %
	$\epsilon_{ex,M,MC}$	32.3 %

Table 3. **Error of interpolated ϵ_{in} and the extrapolated ϵ_{ex} disparities.** Datasets: KITTI (K), Middlebury (M). Cost computation methods: Census transformation (CT), MC-CNN (MC).

created by the Census transformation, the results match with the results of the Middlebury dataset: The proposed data and smoothness terms both improve the results. Again, the extrapolated disparities only have a minor influence on the results, even though the majority of the extrapolated disparities are estimated correctly (c.f. Table 3).

Using cost volumes obtained via MC-CNN, the error decreases when the data term is modified. In consequence, it can be stated that the estimated hyper-parameters of the data term are suitable for different datasets and cost computation methods. However, the error increases when the proposed smoothness term is also considered. This indicates that the estimation of hyper-parameters for this processing step requires further investigations in order to obtain values which show general validity.

In contrast to the results on the Middlebury dataset, the error within unstructured areas $\epsilon_{K,CT,unstr}$ can be reduced significantly on the KITTI dataset, using the Census transformation and the proposed data term. This effect is probably caused by the more complex scenes of the KITTI dataset, which pose a greater challenge to dense stereo matching algorithms. In such cases, the need of a suitable regularisation approach becomes clear. The improvements can also be seen when inspecting the results qualitatively, especially when comparing the disparity map created with the proposed method (c.f. Fig. 6d) with the disparity map resulting from the baseline approach (c.f. Fig.6c). Especially the disparity estimations in planar and uni-coloured areas, such as the street, are improved. This is also confirmed by the two corresponding error maps, shown in the Figures 6e and 6f: Especially on the street, the number of erroneous disparity estimations is significantly reduced. Additionally, the absolute difference between estimated and reference disparity is minimised in the area close to the left image border, which is only visible in the left reference image.

To conclude, the disparity maps obtained with our proposed method for geometry-based regularisation are smoother and less noisy, compared to the results of the baseline approach. While reducing the number of outliers and artefacts within disparity maps, our method preserves sharp object boundaries and does not over-smooth depth discontinuities.

5. CONCLUSION

In the present work, an uncertainty-driven geometry-based regularisation approach for the task of dense stereo matching is presented. Based on an initial disparity and an initial confidence map, reliable support points are identified which are used to construct a triangle mesh. After filtering for surface inconsistent triangles, this mesh is used to propagate disparity and uncertainty information within local neighbourhoods. The propagated information is introduced to the energy term of global disparity optimisation methods via novel data and smoothness terms.

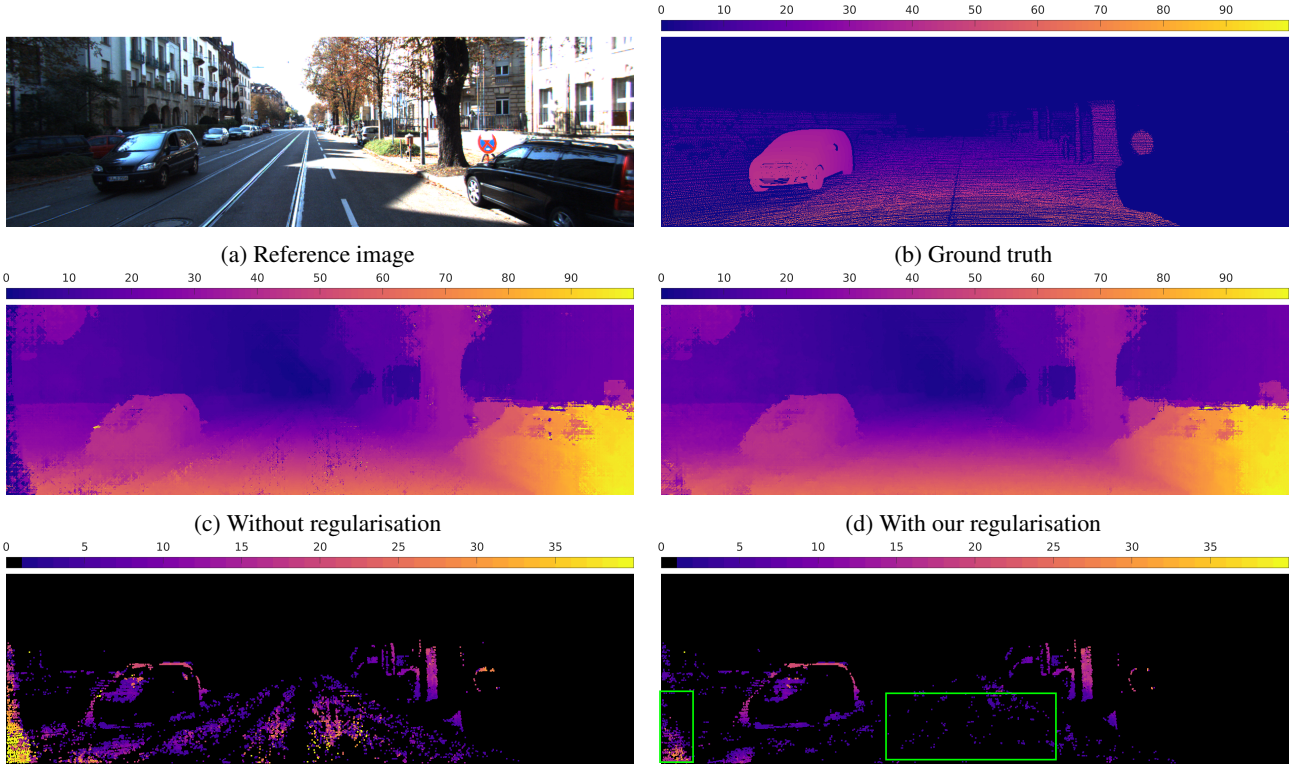
The experimental results show that our propagation scheme together with the proposed data term is able to improve the results for different datasets and cost computation methods, if only interpolated disparities are used. While this improvement is greater when using the Census transformation, due to the poorer initial disparity estimates, it is still clearly evident for the cost computation via MC-CNN. Also the proposed smoothness term improves the results for most of the evaluated examples, but not for all. The corresponding hyper-parameters are sensitive to different cost computation methods and further investigations are necessary to obtain values with general validity.

The proposed extrapolation scheme, on the other hand, has only a minor impact on the result. Since correctly extrapolated disparities are mainly located in areas, where dense image matching algorithms are able to estimate the correct disparity without any regularisation, only minor improvements can be achieved. However, in order to further increase the density of propagated disparity and uncertainty information, extrapolation may be beneficial if the mentioned limitations can be overcome. For this purpose, further investigations on other extrapolation strategies have to be carried out in future work.

Finally, our regularisation approach helps to obtain smoother and less noisy disparity maps, while preserving sharp object boundaries. Especially in complex scenes, the proposed method shows convincing results and supports dense stereo matching algorithms to obtain correct disparity estimations even in challenging cases, such as unstructured areas.

ACKNOWLEDGEMENTS

This work was supported by the MOBILISE initiative of the Leibniz University Hannover and TU Braunschweig and the German Research Foundation (DFG) as a part of the Research Training Group i.c.sens [GRK2159].



(a) Reference image (b) Ground truth

(c) Without regularisation (d) With our regularisation

(e) Error map without regularisation (f) Error map with our regularisation

Figure 6. **Qualitative results on the KITTI 2015 dataset.** The disparity maps are obtained using the Census transformation with semi-global matching, showing the disparity in pixels from small disparities in blue to large ones in yellow. The error maps show the difference between the disparity estimates and the ground truth in pixel from a small error in black to a large one in yellow. It can be seen that our regularisation approach mainly improves the disparity map in the low-textured area on the street and minimises the error close the left image border which can only be seen in the left image (marked with green rectangles).

Gefördert im Niedersächsischen Vorab durch:
Sponsored by the Ministry of Science and Culture of Lower Saxony:
Sponsored by VolkswagenStiftung:

Niedersächsisches Ministerium für Wissenschaft und Kultur

VolkswagenStiftung

REFERENCES

Bulatov, D., Wernerus, P., Heipke, C., 2011. Multi-view Dense Matching supported by Triangular Meshes. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(6), 907-918.

Gallup, D., Frahm, J.-M., Pollefeys, M., 2010. Piecewise Planar and Non-Planar Stereo for Urban Scene Reconstruction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 1418–1425.

Geiger, A., Roser, M., Urtasun, R., 2010. Efficient large-scale Stereo Matching. *Asian Conference on Computer Vision*, Springer, 25–38.

Guney, F., Geiger, A., 2015. Displets: Resolving Stereo Ambiguities using Object Knowledge. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4165–4175.

Hirschmuller, H., 2008. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2), 328-341.

Hong, L., Chen, G., 2004. Segment-based Stereo Matching using Graph Cuts. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 74–81.

Li, A., Chen, D., Liu, Y., Yuan, Z., 2016. Coordinating Multiple Disparity Proposals for Stereo Computation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4022–4030.

Mehltretter, M., Heipke, C., 2019. CNN-based Cost Volume Analysis as Confidence Measure for Dense Matching. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2070–2079.

Mehltretter, M., Kleinschmidt, S. P., Wagner, B., Heipke, C., 2018. Multimodal Dense Stereo Matching. *German Conference on Pattern Recognition*, Springer, 407–421.

Menze, M., Geiger, A., 2015. Object Scene Flow for Autonomous Vehicles. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3061–3070.

Poggio, T., Torre, V., Koch, C., 1987. Computational Vision and Regularization Theory. *Readings in Computer Vision*, Elsevier, 638–643.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P., 2014. High-resolution Stereo

Datasets with Subpixel-accurate Ground Truth. *German Conference on Pattern Recognition*, Springer, 31–42.

Scharstein, D., Szeliski, R., 2002. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision*, 47(1-3), 7-42.

Scharstein, D., Taniai, T., Sinha, S. N., 2017. Semi-global Stereo Matching with Surface Orientation Priors. *Proceedings of the International Conference on 3D Vision*, IEEE, 215–224.

Spyropoulos, A., Komodakis, N., Mordohai, P., 2014. Learning to detect Ground Control Points for Improving the Accuracy of Stereo Matching. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1621–1628.

Stathopoulou, E.-K., Remondino, F., 2019. Multi-view Stereo with Semantic Priors. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W15, 1135–1140.

Terzopoulos, D., 1986. Regularization of Inverse Visual Problems Involving Discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 413–424.

Vedaldi, A., Soatto, S., 2008. Quick Shift and Kernel Methods for Mode Seeking. *Proceedings of the European Conference on Computer Vision*, Springer, 705–718.

Wei, Y., Quan, L., 2004. Region-Based Progressive Stereo Matching. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE.

Yamaguchi, K., McAllester, D., Urtasun, R., 2014. Efficient Joint Segmentation, Occlusion Labeling, Stereo and Flow Estimation. *Proceedings of the European Conference on Computer Vision*, Springer, 756–771.

Zabih, R., Woodfill, J., 1994. Non-Parametric Local Transforms for Computing Visual Correspondence. *Proceedings of the European Conference on Computer Vision*, Springer, 151–158.

Zbontar, J., LeCun, Y., 2016. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *Journal of Machine Learning Research*, 17(1), 2287–2318.