



# Guiding Deep Learning with Expert Knowledge for Dense Stereo Matching

Waseem Iqbal<sup>1</sup>  · Jens-André Paffenholz<sup>1</sup>  · Max Mehlretter<sup>2</sup> 

Received: 14 February 2023 / Accepted: 9 July 2023  
© The Author(s) 2023

## Abstract

Dense depth information can be reconstructed from stereo images using conventional hand-crafted as well as deep learning-based approaches. While deep-learning methods often show superior results compared to hand-crafted ones, they commonly learn geometric principles underlying the matching task from scratch and neglect that these principles have already been intensively studied and were considered explicitly in various models with great success in the past. In consequence, a broad range of principles and associated features need to be learned, limiting the possibility to focus on important details to also succeed in challenging image regions, such as close to depth discontinuities, thin objects and in weakly textured areas. To overcome this limitation, in this work, a hybrid technique, i.e., a combination of conventional hand-crafted and deep learning-based methods, is presented, addressing the task of dense stereo matching. More precisely, the input RGB stereo images are supplemented by a fourth image channel containing feature information obtained with a method based on expert knowledge. In addition, the assumption that edges in an image and discontinuities in the corresponding depth map coincide is modeled explicitly, allowing to predict the probability of being located next to a depth discontinuity per pixel. This information is used to guide the matching process and helps to sharpen correct depth discontinuities and to avoid the false prediction of such discontinuities, especially in weakly textured areas. The performance of the proposed method is investigated on three different data sets, including studies on the influence of the two methodological components as well as on the generalization capability. The results demonstrate that the presented hybrid approach can help to mitigate common limitations of deep learning-based methods and improves the quality of the estimated depth maps.

**Keywords** Image matching · Depth estimation · 3D reconstruction · Hybrid technique

## 1 Introduction

The availability of depth information is beneficial or even a crucial prerequisite for many applications, such as robotics (Häne et al. 2011), 3D reconstruction (Krutikova et al. 2017), semantic segmentation (Badrinarayanan et al. 2017) and pedestrian tracking (Nguyen and Heipke 2020). Besides

the acquisition via active sensors, the reconstruction of depth from stereo images is particularly popular, inter alia due to the possibility to obtain dense depth estimates and the comparably inexpensive sensor setup. However, such dense stereo matching approaches are commonly challenged by occlusions, weakly textured surfaces and depth discontinuities in the scene. Latter may lead to the problem that depth estimates corresponding to small and thin (parts of) objects in an image are falsely detected as noise and filtered out, resulting in the disappearance of these parts of the scene in the depth map. While inaccurate depth discontinuities and missing object parts may only have a minor negative influence on the overall quantitative results, these areas are highly relevant, for example, in navigation-related tasks to avoid collisions.

Especially in the pre-deep learning era, many hand-crafted, i.e., non-data driven, approaches have been presented in the literature that address these challenging

---

✉ Waseem Iqbal  
Waseem.Iqbal@tu-clausthal.de

Jens-André Paffenholz  
Jens-Andre.Paffenholz@tu-clausthal.de

Max Mehlretter  
mehlretter@ipi.uni-hannover.de

<sup>1</sup> Institute of Geo-Engineering, Clausthal University of Technology, Clausthal-Zellerfeld, Germany

<sup>2</sup> Institute of Photogrammetry and GeoInformation, Leibniz University Hannover, Hannover, Germany

scenarios explicitly. For this purpose, various feature description methods have been proposed and adapted for image matching that are robust against certain geometric and radiometric conditions (Zabih and Woodfill 1994; Yinan et al. 2012; Kabbai et al. 2015). An assumption commonly made to cope with weakly textured surfaces and depth discontinuities is that a direct relation exists between intensity gradients and depth discontinuities. This assumption can be used to generally encourage smoothness in a depth map and to foster discontinuities only for pixels that have assigned a strong intensity gradient in the associated RGB image (Hirschmüller 2008). However, such hand-crafted features and smoothness assumptions are commonly optimized in an isolated manner, not considering the further components of dense stereo matching methods. As such an approach lacks an overall optimization goal, such hand-crafted features are often not optimal with respect to the entire matching process.

This issue is also demonstrated by the overall superior performance of recent deep learning-based methods, compared to non-data driven approaches. However, deep learning-based methods often learn well-known geometric principles from scratch and consider image regions, that have been characterized as being challenging before, only implicitly by learning solutions from training data. As such challenging regions commonly constitute only a small part of an image, they are highly underrepresented in the training set and without explicitly focusing on these regions, the associated depth estimates are often significantly less accurate and error prone compared to the entirety of an image. Consequently, there are clear trade-offs between traditional computer vision and deep learning-based approaches for dense stereo matching. To combine the advantages of both directions, the discriminative power of data driven methods and the possibility of expert knowledge-based techniques to model geometric principles explicitly, first hybrid methods have been presented in the literature in recent years for vision tasks in general (Tianyu et al. 2018) and for dense stereo matching in particular (Pang et al. 2017; Stucker and Schindler 2022).

Following this concept, a hybrid method addressing dense stereo matching is proposed in the present work. On the one hand, the goal is to avoid false predictions of depth discontinuities in weakly textured regions of an image. On the other hand, correct discontinuities are to be sharpened and small (parts of) objects are to be preserved in the depth map. To achieve these objectives, a twofold strategy is applied, facilitating a Convolutional Neural Network (CNN) with expert knowledge: First, it is investigated whether feature maps generated with a hand-crafted method based on expert knowledge can provide valuable information to a CNN. For this purpose, a basic set of information is provided to the CNN as additional input to

the stereo image pair, allowing the CNN to focus on relevant details. Second, the assumption that image edges, i.e., strong intensity gradients, and depth discontinuities coincide is modeled explicitly, by training a neural network to predict depth discontinuities from RGB images. The depth discontinuities obtained in this way are used to guide the deep learning-based matching process, by providing this rough representation of the scene's geometry to the neural network at different stages and in multiple resolutions. Thus, in this work, expert knowledge is incorporated into a deep learning-based approach in two ways: Feature maps computed based on expert knowledge are used as additional input to a neural network and the assumption that gradients in the RGB images and discontinuities in the depth maps coincide is based on expert knowledge, too. In summary, the present work contains the following main contributions:

- A strategy to consider the information obtained with a hand-crafted feature description method as additional input to a deep learning-based dense stereo matching method, supplementing the RGB stereo image pairs.
- A deep learning-based method to predict a pixel's probability of being located next to a depth discontinuity based on a RGB image, making explicit use of the assumption that image edges and depth discontinuities often coincide.
- An approach to consider the predicted information on depth discontinuities in the matching process at multiple scales.

## 2 Related Work

The task of computing depth from stereo image pairs is already being studied for many years (Barnard and Fischler 1982), leading to a wide variety of methods on this topic presented in the literature. To allow for a sound comparison of these stereo matching algorithms, Scharstein and Szeliski (2002) have presented a taxonomy and categorization scheme, which distinguishes the following four steps: matching cost computation, cost aggregation, disparity computation and disparity refinement. While the taxonomy was originally developed to subdivide hand-crafted methods, many deep learning-based approaches are designed in a similar fashion (Mayer et al. 2016; Kendall et al. 2017). These methods often consist of an end-to-end trainable sequence of individual components, in which each component can be mapped to one of the steps in the taxonomy. Thus, we will also make use of this taxonomy in the following to review and discuss deep learning-based techniques and in the context of the description of our own method.

## 2.1 Deep Learning-Based Dense Stereo Matching

While in earlier years only individual steps of the taxonomy have been realized using deep learning, such as the matching cost computation (Zbontar and LeCun 2016; Luo et al. 2016), a rapid transition took place to model the complete matching process as a single neural network (Mayer et al. 2016; Kendall et al. 2017). In DispNet, presented by Mayer et al. (2016), features are extracted from the left and the right image, respectively, using a Siamese network structure. These features are then combined via correlation to build a cost volume, which is further processed using multiple layers of 2D convolutions. In the final step, disparity estimates per pixel are obtained via regression, which is realized as another 2D convolutional layer. GC-Net proposed by Kendall et al. (2017) presents a similar concept, but concatenates the feature maps from the left and right image instead of correlating them. In the following, multiple 3D convolutions and transposed convolutions, arranged in an encoder–decoder scheme with skip connections, are applied to obtain a cost volume. Disparity estimates per pixel are regressed from this cost volume by applying a differentiable version of the argmin operation. Although both methods show clear improvements in the overall results compared to hand-crafted and partially deep learning-based approaches, depth estimates in weakly textured areas and depth discontinuities often remain inaccurate and error prone.

To address these limitations, various further methods have been presented in the literature, most of them adapting the basic concepts of Mayer et al. (2016) and Kendall et al. (2017). For example, in Ilg et al. (2018) a custom layer and auxiliary loss functions are added, in Shaked and Wolf (2017) a reflective confidence algorithm is proposed and in Pang et al. (2017) a residual learning method is introduced to improve the performance of DispNet and MC-CNN (Zbontar and LeCun 2016) to handle occlusions and areas close to depth discontinuities. On the other hand, Kang et al. (2019) state that DispNet shows a lower accuracy for large disparities and introduce dilated convolutions to broaden the receptive field and to consider multi-scale contextual information without introducing additional parameters that need to be trained. Based on the enlarged receptive field, this method is also beneficial in weakly textured areas. Moreover, they incorporate disparity gradient information as a gradient regularizer in the loss function to preserve geometric details. However, this strategy influences the training process only and does not provide additional information or guidance at test time, which clearly limits its impact. Cheng et al. (2019) adapt PSM-Net by adding spatial pooling to compute an improved cost volume. That leads to a 30% reduction in depth error in NYU v2 (Silberman et al. 2012) and KITTI Odometry (Geiger et al. 2012) data sets. Even so, due to the spatial pooling, the smoothing effect is evident

from the disparity map, especially at the object boundaries and smaller structures. Shamsafar et al. (2021) introduce a new cost volume to learn the similarity of unary features at a particular disparity, which significantly reduces the End-Point-Error (EPE). However, object boundaries and small structures are not well-handled. Tosi et al. (2021) propose a different approach of stereo matching that improves the depth accuracy near object boundaries, by employing a 2D or 3D PSM Network (Chang and Chen 2018) as a backbone, which takes a stereo image pair as input and outputs a low-resolution disparity map. This disparity map is used in SMD-Net to initialize a mixture density per pixel, which is represented by a set of Gaussian components. The idea is to model the distribution of possible disparities for each pixel and the output is the final high-resolution disparity map. Post-processing is carried out to address over-smoothing. Finally, Lipson et al. (2021) introduce a convolutional recurrent model that allows to iteratively propagate depth information across the image. By carrying out depth propagation at multiple resolutions simultaneously, a globally consistent high-resolution depth map is obtained. Due to the fact that the results are state of the art and the associated program code is publicly available, this method is used for comparison in Sect. 5.

## 2.2 Hybrid Approaches

Hybrid approaches combine aspects of traditional non-data driven and deep learning-based approaches with the goal to make use of the benefits of both. Such approaches have successfully been developed for various photogrammetry and computer vision-related tasks already (O'Mahony et al. 2020), including image classification (Tianyu et al. 2018), panoramic vision (Verma et al. 2020), video stabilization (Liu et al. 2021) and image-based brain tumor detection (Saba et al. 2020), which mainly fuse hand-crafted features with such that have been extracted with a CNN before an actual classification is done. However, also further hybrid strategies are presented in the literature, for example, using geometric information in the loss function (Kang et al. 2019) or as a traditional prior, such as brightness constancy, gradient constancy and image-driven smoothness assumptions (Xiang et al. 2018).

Also in the context of reconstructing the 3D geometry of a depicted scene, first hybrid approaches have been proposed. Stucker and Schindler (2022) and Stucker et al. (2022) present hybrid methods in the context of binocular dense stereo matching and the estimation of a digital surface model using an occupancy field, respectively. Both follow a similar approach, estimating initial depth information with a non-data driven technique first, which is refined via deep learning afterwards. The idea is that the basic concept of 3D reconstruction from images does not need to be learned

from scratch, rather letting the neural network focus on the improvement and correction of particularly challenging parts of an image by providing a reasonably good initialization of the desired result. While both methods show convincing results, their performance is potentially limited by the fact that the fusion of the non-data driven and the deep learning-based method is realized as a two-step approach of initialization and refinement instead of aiming for a full integration. Such a two-step approach requires that errors made in the first step can be detected and corrected in the second step and prevents optimizing the entire method end-to-end. On the other hand, Zhang et al. (2019) address the cost aggregation step of the dense stereo matching taxonomy by incorporating the semi-global aggregation scheme of Hirschmüller (2008) into a CNN. In this way, the general validity of the well-designed aggregation scheme of semi-global matching is combined with the adaptation capability of a weighting scheme that is optimized for a certain domain.

Closest to the method presented in this work, are the ones of Qi et al. (2020) and Ilg et al. (2018). In Qi et al. (2020), a method for monocular depth estimation is presented in which a neural network is trained to predict depth and surface normals jointly. Both predictions are refined using edge maps extracted from the RGB image using the Canny detector (Canny 1986) and thus establish an explicit connection between image edges and depth discontinuities as well as to abrupt changes in the direction of the surface normals. However, the proposed direct incorporation of extracted image edges does not improve the results significantly close to depth discontinuities, probably because the assumption that image edges and depth discontinuities coincide is not always valid. Ilg et al. (2018) follow a similar approach for the tasks of disparity, optical flow and scene flow estimation, but jointly predict occlusion, motion and depth boundaries

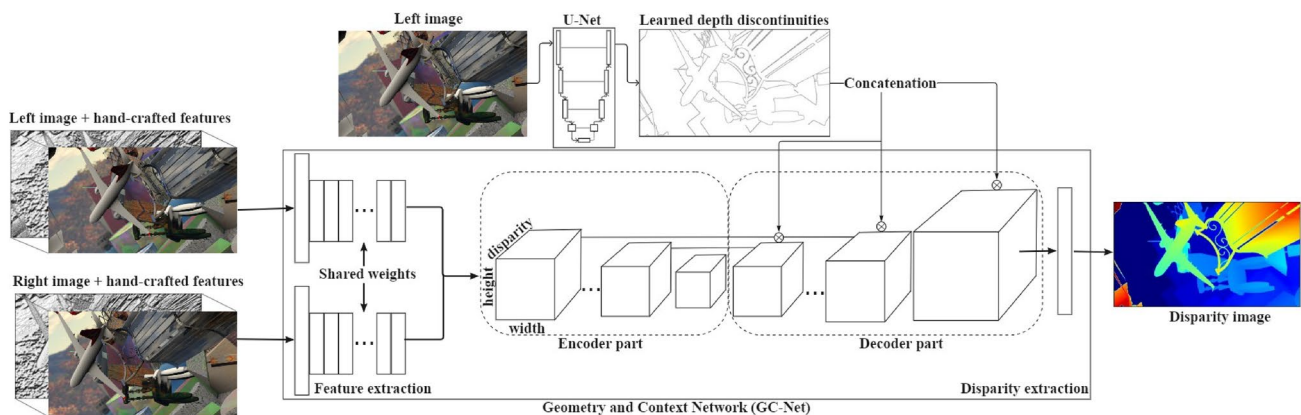
from stereo image pairs using a CNN instead of including image edges directly. In contrast, no temporal information and only the left image of a stereo pair is used as basis to predict depth discontinuities in the present work. In addition, we employ an auxiliary loss term that allows us to learn this task from labeled reference depth discontinuities directly.

### 3 Methodology

In this section, the developed hybrid methodology for the task of dense stereo matching and its components are presented. This method is specifically designed to focus on some of the most challenging areas in the context of disparity estimation, namely, depth discontinuities, thin structures and weakly textured areas. For this purpose, two novel components are presented: A strategy to use hand-crafted features as an additional input of a CNN is described in Sect. 3.2, a technique to predict depth discontinuities and to integrate this information into the matching process is described in Sect. 3.3. Both components are described with respect to GC-Net (Kendall et al. 2017), which is reviewed in Sect. 3.1. Note that the proposed components are generally applicable to all methods that follow the same basic realization of the matching taxonomy (cf. Sect. 2). An overview of the complete method is given in Fig. 1.

#### 3.1 Baseline

As basis for our hybrid method, we use GC-Net proposed by Kendall et al. (2017). GC-Net implements the complete dense stereo matching taxonomy as a single CNN that can be trained in an end-to-end manner to estimate disparity maps from epipolar rectified stereo image pairs. As GC-Net



**Fig. 1** Overview of the proposed hybrid method. GC-Net is used as a basis to estimate a disparity map from an epipolar rectified stereo image pair. The input images are supplemented by an additional feature channel which is obtained by applying the LBP transformation

on the respective RGB image. Depth discontinuities are predicted from the left stereo image using a U-Net architecture and are fed into the decoder of GC-Net at multiple stages to guide the up-sampling process



follows a purely data driven approach and builds the basis for many methods presented afterwards (cf. Sect. 2.1), it appears to be a well-suited baseline for investigating the effects of the hybrid concepts developed in the context of this work, while at the same time offering the possibility to apply these concepts to other dense stereo matching methods in a straightforward manner. The architecture of GC-Net consists of four basic modules, which are also shown in Fig. 1: feature extraction, encoder, decoder and disparity extraction. First, features are extracted from the left and right image, respectively, using several 2D convolutional layers that are arranged in a Siamese network structure. The extracted features are assembled to a 4D volume (a 3D volume with two spatial dimensions and one disparity dimension, for which every entry is a feature vector) by concatenating feature vectors from the left and right image for all potential point correspondences. This volume is then further processed by an encoder–decoder structure that consists of multiple 3D convolutional and transposed convolutional layers. Skip connections that link layers of the encoder and the decoder are used to guide the up-sampling process. The result is a cost volume that has the same spatial resolution as the RGB input images and a disparity map is extracted by applying a soft argmin operation over the disparity dimension of the cost volume.

### 3.2 Hand-Crafted Features as Basis

The central idea of the first component of our hybrid method is to incorporate feature maps obtained with a hand-crafted method into GC-Net at an early stage. Thus, a basic set of coarse information is provided, allowing the CNN to focus on learning supplementary features that capture details relevant to the matching task. This concept is realized by applying a hand-crafted feature description method on both RGB images of a stereo pair, appending the resulting feature maps to the respective RGB image as a fourth image channel before feeding the so-extended images to the neural network. On the one hand, this approach provides additional guidance during training. On the other hand, the expert domain knowledge used to develop a hand-crafted feature description method is also available for the network at test time. Consequently, this information does not need to be encoded in the neural network itself, as it would, for example, be necessary if the expert knowledge would only be used to adapt the loss function. Finally, the introduction of this information as an additional image channel only marginally increases the number of parameters of the neural network to be trained.

Following Anwer et al. (2018), who propose a hybrid method similar to the one described above but for image classification, we use Local Binary Patterns (LBP) as presented in Ojala et al. (2002) as hand-crafted feature

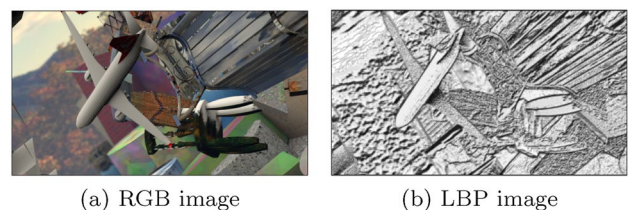
description method. This choice is further supported by the results of preliminary experiments that have been carried out in the context of the present work, showing that LBP outperforms other types of features, such as Histogram of Oriented Gradients (HOG) features or edge maps extracted with the Canny detector, when being combined with a CNN for the task of dense stereo matching. An ablation study on this aspect is presented in Sect. 5.4.

LBP is an effective descriptor for local texture patterns. It involves dividing an image into small regions from which robust features are extracted. It encodes the local structure of an image by comparing the gray-level intensity of each pixel with its eight neighboring pixels. The resulting binary number is then converted into a decimal value, which represents the texture pattern encoded by that pixel.

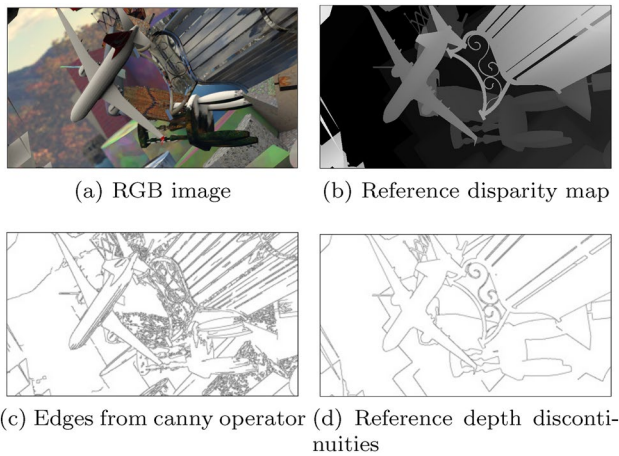
A visual example of a feature map obtained via LBP is shown in Fig. 2. LBP is chosen to be part of our method due to its discriminative power, its computational simplicity, and its robustness to monotonic gray level changes caused, for example, by variations in illumination and contrast.

### 3.3 Predicted Depth Discontinuities as Guidance for the Decoder

The second component of our hybrid method is based on the assumption that image edges, i.e., strong intensity gradients of an image, coincide with discontinuities in the corresponding depth map. However, as this assumption does not always hold true (see Fig. 3), we propose to train a CNN to predict depth discontinuities from a single RGB image, instead of making direct use of image edges to guide the matching process. More precisely, the probability per pixel of the input image is predicted that the respective pixel is located next to a depth discontinuity, resulting in a single-channel feature map  $Y$  that has the same spatial dimensions as the input image. Thus, only the potential position of a depth discontinuity is considered, not its magnitude (which would be much more error prone as shown by the general issues of monocular depth estimation methods with predicting absolute depth values). As functional model for this component, we use a symmetric U-Net (Ronneberger et al. 2015) with



**Fig. 2** Example of the LBP transformation. The figure shows an RGB image and its transformation using LBP



**Fig. 3** Example for the relation between image edges and depth discontinuities. It can be seen that the majority of image gradients coincide with depth discontinuities. However, there also exist image edges that are caused by texture and do not correspond to a geometric edge, and, vice versa, depth discontinuities that do coincide with an image edge. Latter is commonly caused by geometric structures that are not reflected in the appearance, e.g., due to objects with similar textures that are partially overlapping in the image

VGG16 (Simonyan and Zisserman 2014) as encoder and a maximum down-sampling factor of  $2^5$ .

In our experiments, this network is either trained with the dice or the weighted binary cross-entropy loss function as optimization objective. Using the dice loss commonly leads to sharp separations between distinct classes and thus allows for precise localization of depth discontinuities. On the other hand, the weighted binary cross-entropy loss allows us to account for class imbalances, which have to be expected with the present problem definition, as pixels close to depth discontinuities are commonly highly underrepresented compared to all other pixels. As these regions close to depth discontinuities are of special interest in the present work, a weighting factor  $\alpha$  is applied to pixels in these regions in the weighted binary cross-entropy loss function. This weight is determined based on the inverse class frequency of pixels close to depth discontinuities in the training set, guiding the model to focus on these pixels. Both variants, using the dice and the weighted binary cross-entropy loss function, are compared experimentally in Sect. 5.5 and are defined as follows:

$$\mathcal{L}_{dice} = 1 - \frac{2 \sum_{i=1}^N y_i \cdot \hat{y}_i}{\sum_{i=1}^N y_i + \hat{y}_i}, \quad (1)$$

$$\mathcal{L}_{wbce} = -\frac{1}{N} \sum_{i=1}^N \alpha \hat{y}_i \log y_i + (1 - \alpha)(1 - \hat{y}_i) \log(1 - y_i) \quad (2)$$

where  $N$  is the number of pixels with a reference disparity available,  $y$  is the predicted probability in  $[0, 1]$  and  $\hat{y}$  is the true class membership, whereas  $\hat{y} \in \{0, 1\}$  with 1 meaning that a pixel is located next to a depth discontinuity and 0 that it is not. Thus, in the dice loss, the numerator is the sum of true positive predictions and the denominator is the total number of positives in the prediction and the reference, meaning that the dice loss decreases as the intersection of the prediction and the reference increases. In this context, the true class memberships  $\hat{y}$  are derived from the reference disparity map by computing the first derivative of the disparity map using a Canny operator, followed by a binarization of the obtained gradients with thresholds of 10 (lower threshold) and 30 (higher threshold). These thresholds are determined empirically.

To use the estimated information on depth discontinuities as guidance for the matching process, the predicted feature map  $Y$  is incorporated into the decoder part of GC-Net after each up-sampling of the 4D feature volume (see Fig. 1). For this purpose,  $Y$  is actually predicted in multiple resolutions by adapting the resolution of the RGB input image to the spatial dimensions of the respective feature volume. In this context, the parameters of the described CNN for predicting depth discontinuities are shared over all resolutions. The incorporation itself is realized by concatenating the predicted probability of a pixel being located close to a depth discontinuity to the feature vectors of all entries in the feature volume corresponding to this pixel (note that there are multiple entries associated with one pixel of the left image along the disparity dimension). The intuition behind this guidance strategy is that relevant high-frequency information might get lost during the down-sampling process in the encoder. While higher resolution feature maps from the encoder are available via skip connections, this high-frequency information cannot always be fully recovered when up-sampling in the decoder. In this context, the predicted depth discontinuities provide a rough initialization for the 3D geometry in the desired higher resolution of an up-sampling operation, facilitating the accurate localization of a correct depth discontinuity and the avoidance of an incorrect one. This strategy promises to be particularly helpful for estimating the full resolution cost volume, which has the same spatial dimensions as the input stereo images, as no skip connection is linked to this volume in GC-Net, due to the fact that the first feature volume in the encoder is already down-sampled by a factor of two.

The described combination of U-Net used to predict depth discontinuities and GC-Net is trained via a two-part loss function:

$$\mathcal{L} = \mathcal{L}_{L1} + \mathcal{L}_{depth\_discon} \quad (3)$$

where  $\mathcal{L}_{L1}$  is the mean L1 distance between the estimated disparity and the true disparity of a pixel, considering all pixels for which the true disparity is available.  $\mathcal{L}_{depth\_discon}$  is the loss function used to optimize the depth discontinuity prediction, using either the dice or the weighted binary cross-entropy loss function, as defined in Eqs. 1 and 2, respectively.

## 4 Experimental Setup

In this section, the experimental setup is described which is used to evaluate the methodology proposed in Sect. 3, including the data sets used (Sect. 4.1), the training (Sect. 4.2) as well as the test settings (Sect. 4.3).

### 4.1 Data Sets

For the experiments conducted in the context of the present work, three different data sets are used: Sceneflow FlyingThings3D (Mayer et al. 2016), InStereo2k (Bao et al. 2020) and Middlebury v3 (Scharstein et al. 2014). The Sceneflow data set is a collection of about 27 thousand synthetic stereo image pairs with a resolution of  $540 \times 960$  pixels that show a variety of scenes with randomly located objects and for which a dense ground truth is available. InStereo2k and Middlebury v3 contain 2050 and 15 stereo image pairs, respectively, which have a varying resolution between  $659 \times 497$  and  $900 \times 750$  pixels, showing various real indoor scenes. Both data sets provide a reference for the disparity for about 90% of the pixels which is captured via structured light. As the method of learning to predict depth discontinuities presented in this work relies on the availability of a dense ground truth for the disparity during training, the experiments are limited to indoor data sets. Adapting this part of our method to also be applicable to outdoor data, which commonly provides sparse reference data, will be subject of future work.

### 4.2 Training Procedure

The presented combination of GC-Net and U-Net is trained on twelve thousand synthetic stereo image pairs of the Sceneflow data set first, as commonly done in the literature (Mayer et al. 2016; Kendall et al. 2017), before fine-tuning it on two thousand real stereo image pairs of the InStereo2k data set. In each iteration of the training process, a sample of  $256 \times 512$  pixels is randomly cropped from a stereo image pair and the contained RGB values are normalized to the interval  $[-1, 1]$ , before feeding it to the neural network. The batch size is equal to one, i.e., the gradients are computed and the neural network parameters are updated after each sample. In this context, one sample per stereo image pair is

seen in every epoch. This training strategy is applied on both data sets. To compute the loss, only pixels with a known reference are considered. Moreover, the disparity range considered during training is limited to  $[0, 191]$  pixels, meaning that pixels with a ground truth disparity outside of this range are discarded during the training process, whereas this limit does not apply at test time. For optimizing the ability to predict depth discontinuities, only the feature map  $Y$  in full resolution is evaluated in the loss term  $\mathcal{L}_{depth\_discon}$ . Using an early stopping strategy, i.e., the training is stopped when the validation loss does not decrease within three consecutive epochs, the parameter values resulting in the lowest validation loss are used for testing. Using the weighted binary cross-entropy loss function for  $\mathcal{L}_{depth\_discon}$  in Eq. 3, pixels that are located close to depth discontinuities are weighted by 0.9, while all other pixels are weighted by 0.1. These weights are determined based on the inverse frequency of the two classes in the training set. Lastly, all convolutional layers are initialized with the Xavier normal initializer (Glorot and Bengio 2010) and RMSprop (Tieleman and Hinton 2012) with a learning rate of 0.001 is used as optimizer. To compare the proposed method against the current state-of-the-art, RAFT-Stereo (Lipson et al. 2021) has been trained following the same procedure and using the same data described above, while using the model-related hyperparameter settings described in the original publication.

### 4.3 Evaluation Procedure

To evaluate the effectiveness of the presented method and its individual components, the following variants are compared: As *Baseline*, the original version of GC-Net is used without applying any changes. *4th channel (LBP)* refers to a variant that only considers hand-crafted features as additional input as described in Sect. 3.2, but neglects the prediction of depth discontinuities. *U-Net with dice* refers to a variant in which the baseline is extended by the prediction of depth discontinuities using the dice loss function as described in Sect. 3.3, but without adding additional input. Lastly, two variants that incorporate both components into the baseline are referred to as *LBP + U-Net with dice* and *LBP + U-Net with wbce*, using the dice and the weighted binary cross-entropy loss function for learning to predict depth discontinuities, respectively.

To compute quantitative results, twenty random stereo image pairs from each of the Sceneflow FlyingThings3D and InStereo2k data sets, and all fifteen stereo image pairs from the Middlebury v3 data set are evaluated using the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE) and the Pixel Error Rate (PER) as metrics. The same image pairs are being used for all experiments. In the MAE, all deviations are weighted equally, making it easy to interpret the results of this metric:

$$MAE = \frac{1}{N} \sum_{i=1}^N |d_i - \hat{d}_i|, \quad (4)$$

where  $N$  is the number of pixels for which a ground truth for the disparity is available,  $d$  is the predicted disparity and  $\hat{d}$  is the ground truth disparity.

In the RMSE, on the other hand, large errors have a higher weight than small ones, allowing to identify the presence of such large errors if the RMSE is evaluated together with the MAE:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (d_i - \hat{d}_i)^2}. \quad (5)$$

Finally, the PER allows for a more detailed analysis of the error distribution by providing the percentage of pixels that achieve a specific level of accuracy, defined based on the specified threshold  $\tau$ :

$$PER = \frac{1}{N} \sum_{i=1}^N \alpha_i, \quad (6)$$

$$\alpha_i = \begin{cases} 1 & \text{if } |d_i - \hat{d}_i| > \tau \\ 0 & \text{else} \end{cases}, \quad (7)$$

whereas in this work, the PER is evaluated using 1, 3, and 5 pixels for  $\tau$ .

Besides the general evaluation considering all pixels of the test images, also a more detailed analysis of the results is carried out, assessing predictions for pixels close to depth discontinuities only. For this purpose, a pixel is considered to be situated near a depth discontinuity if the difference in the disparity between this and the surrounding pixels exceeds a threshold of two pixels, averaging the disparity over a  $9 \times 9$  neighborhood (Scharstein and Szeliski 2002).

## 5 Results and Discussion

In this section, the results of the dense stereo matching method described in Sect. 3 are presented and discussed. To give an overview of the general performance, the method is first evaluated on test data from the two data sets that are also used for training in Sect. 5.1. The generalization capability is analyzed in Sect. 5.2 by testing on samples from a data set which has not been seen during training. As the general issue of dense stereo matching methods with depth discontinuities is one of the main motivations of this work, these parts of the test images are analyzed in detail in Sect. 5.3. Ablation studies on the influence of the hand-crafted feature description method used as a fourth image channel and of the loss function used for learning to predict depth discontinuities are presented in Sects. 5.4 and 5.5, respectively. Lastly, in Sect. 5.6, our approach is compared against the current state-of-the-art, namely, RAFT-Stereo presented by Lipson et al. (2021).

### 5.1 General Performance

The first set of experiments evaluates the presented method on the Sceneflow and on the InStereo2k data sets, using the pre-trained parameter values to produce the results of the former and the fine-tuned parameter values for the latter. As shown in Table 1, both components of the presented method have a positive impact on the results of the Sceneflow data set, leading to a clear improvement of all variants compared to the *Baseline*, whereas the consideration of both components leads to the best results. A similar trend can be observed for the results of the InStereo2k data set. However, the consideration of hand-crafted features as additional input leads to a slight decrease in performance. This might be caused by the fact that the indoor scenes of this data set contain many weakly textured and texture-less surfaces

**Table 1** Overall quantitative results

Variants			Sceneflow					InStereo2k				
4th channel (LBP)	Discontinuities via dice	Discontinuities via wbce	Pixel Error Rate [%]			MAE [px]	RMSE [px]	Pixel Error Rate [%]			MAE [px]	RMSE [px]
			$\tau = 1$	$\tau = 3$	$\tau = 5$			$\tau = 1$	$\tau = 3$	$\tau = 5$		
			13.55	8.05	6.55	2.24	9.86	14.34	5.42	3.43	1.00	2.91
✓			6.90	3.01	2.13	0.63	<b>3.44</b>	15.74	6.34	3.97	1.11	3.07
	✓		<b>6.64</b>	2.93	2.04	0.63	3.59	12.38	4.19	2.50	0.95	2.56
✓	✓		6.67	2.98	2.04	0.62	3.50	<b>12.14</b>	<b>3.81</b>	<b>2.06</b>	<b>0.70</b>	<b>1.86</b>
✓		✓	7.15	<b>2.74</b>	<b>1.92</b>	<b>0.61</b>	3.45	12.37	4.36	2.72	1.05	3.75
RAFT-Stereo (Lipson et al. 2021)			6.12	2.66	1.84	0.55	3.31	7.83	1.90	0.75	0.42	0.85

The results of the baseline and the four variants of the method presented in this work are compared to RAFT-Stereo on the Sceneflow and the InStereo2k data sets. The shown results have been produced using the parameter values obtained by training on the Sceneflow data set and by additionally fine-tuning on the InStereo2k data set, respectively. The best values of our method per column are shown in bold



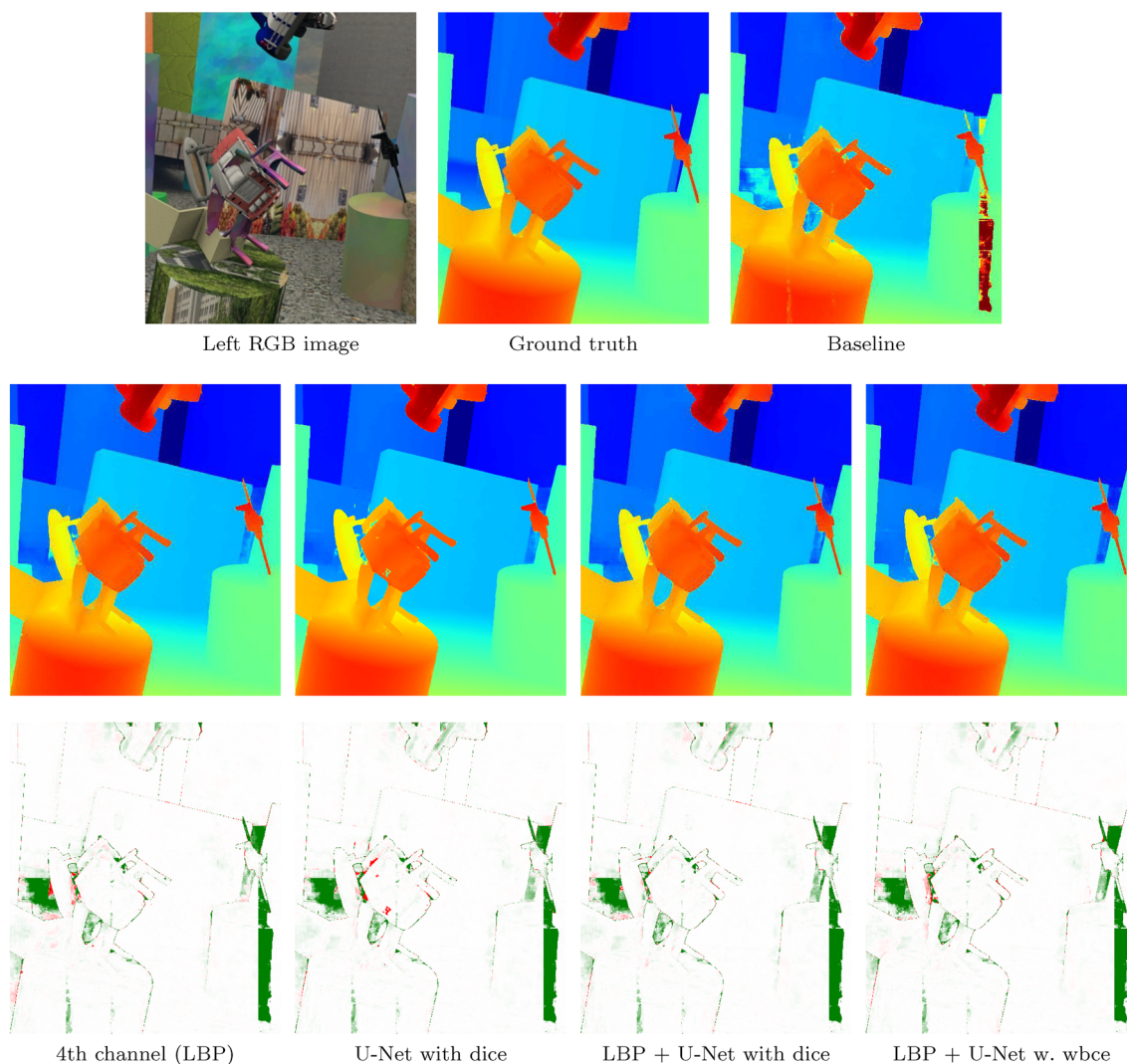
for which the LBP descriptor does not provide any helpful information, whereas the objects in the Sceneflow data set are commonly highly textured.

The qualitative results on the Sceneflow data set shown in Fig. 4 demonstrate that clear improvements are achieved close to depth discontinuities as well as on smooth surfaces. While the former is addressed explicitly by the presented method, the latter is supported implicitly, as a low probability for a depth discontinuity in the prediction of the U-Net suppresses the prediction of such a discontinuity in the final disparity map. Both are also clearly visible in the qualitative results on the InStereo2k data set in Fig. 5, whereas the best results are achieved with the complete model trained with

the dice loss function, which is also indicated by the numeric results in Table 1. Moreover, also the reconstruction of fine structures can be improved, particularly visible, for example, at the handle of the box in the upper part of the image and at the plate of the table on the right hand side.

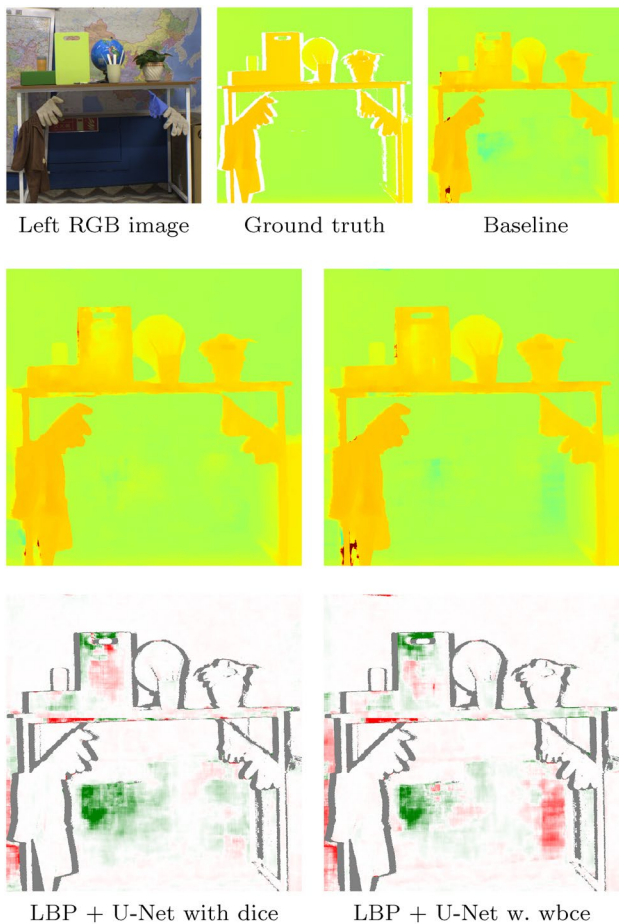
## 5.2 Generalization Capability

To investigate the generalization capability of the presented method, it is tested on samples of the Middlebury data set, from which no samples are used for training. For this experiment, the parameters fine-tuned on the InStereo2k data set are used for all variants. As shown in Table 2, the quality of



**Fig. 4** Qualitative comparison on the Sceneflow data set. These results are produced with parameter values obtained by training on the Sceneflow data set. The second row shows disparity maps, where dark blue to dark red represent small to large disparities, respectively. The difference maps (third row) show the differences in the disparity maps between the respective variant and the baseline. Ranging

from dark green ( $\geq 10$  pixels) over white to dark red ( $\leq -10$  pixels), improvements and deteriorations compared to the baseline are shown. Pixels for which no reference depth is available are displayed in the disparity and difference maps in white and gray, respectively. Significant differences can especially be seen in weakly textured areas, for thin objects and at depth discontinuities



**Fig. 5** Qualitative comparison on the InStereo2k data set. These results are produced with parameter values obtained by training on the Sceneflow data set and fine-tuning on the InStereo2k data set. The second and third row show the disparity and difference maps of the two variants of our full method, respectively, varying only the loss function used for training. For an explanation of the color coding, refer to Fig. 4

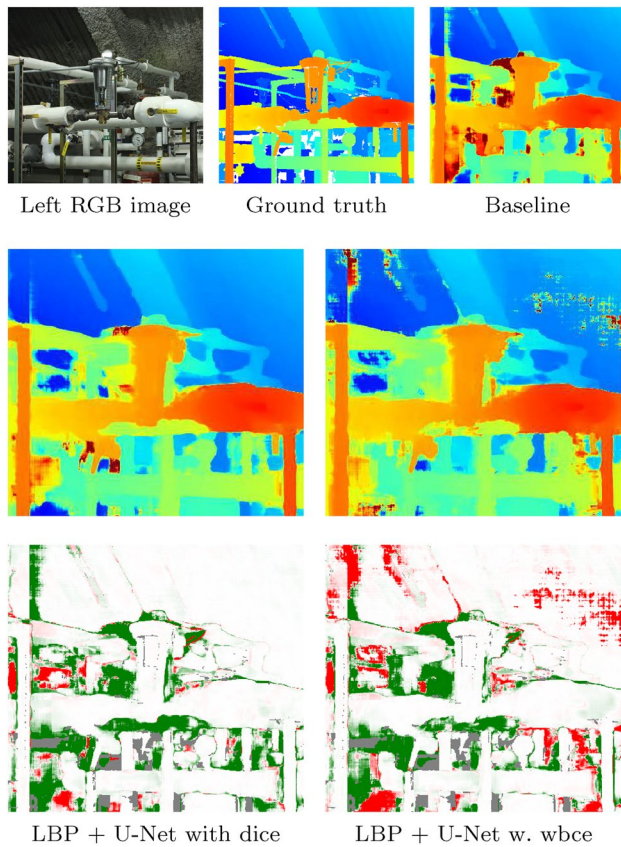
the estimated disparity maps decreases significantly for all evaluated variants compared to the results on the Sceneflow and the InStereo2k data sets. This is to be expected, as a relevant domain gap exists between this test set and the data used for training, which is characterized by differences in the acquisition setup (e.g., different cameras and base lengths) as well as in the depicted scenes (e.g., different objects and depth ranges). However, our complete method trained with the dice loss demonstrates the best performance, although the improvement to the *Baseline* is less prominent. On the Middlebury data set, the influence of the hand-crafted features as the fourth image channel is ambiguous: on the one hand, the predictions get less accurate, visible by the increase of the MAE and the PER for all three thresholds. On the other hand, the number of large errors is reduced, as indicated by the smaller value for the RMSE.

Similar observations can also be made on the qualitative results shown in Fig. 6. However, a clear foreground fattening effect can be observed, which leads to a blurry reconstruction of the scene's geometry and negatively affects the sharpness of the estimated depth discontinuities. The explicit prediction of depth discontinuities using the presented U-Net-based approach mitigates this effect, independent of the used loss function. The issue that the depth discontinuities are still not as sharp as for the Sceneflow data set is probably caused by the domain gap between training and test data, as well as by the fact that the InStereo2k data set is frequently missing a reference for the disparity for pixels close to depth discontinuities (visible, for example, in Fig. 5). As the InStereo2k data set is used for fine-tuning the parameters of our neural network, learning sharp depth discontinuities is a challenging task under these circumstances. Lastly, using the weighted binary cross-entropy loss, artifacts and noise can be observed in the weakly textured areas in the background of the image, which does not apply using the dice loss.

**Table 2** Quantitative results on the generalization capability

Variants			Middlebury			MAE [px]	RMSE [px]
			Pixel Error Rate [%]				
4th channel (LBP)	Discontinuities via dice	Discontinuities via wbce	$\tau = 1$	$\tau = 3$	$\tau = 5$		
			✓				
			37.35	22.75	18.04	5.10	11.21
	✓		36.30	24.24	20.29	8.04	21.34
✓	✓		<b>30.06</b>	<b>18.02</b>	<b>14.03</b>	<b>4.22</b>	<b>10.50</b>
✓		✓	33.51	21.14	17.34	5.97	14.71
RAFT-Stereo (Lipson et al. 2021)			20.31	9.08	5.92	1.35	3.52

The results of the baseline and the four variants of the method presented in this work are compared to RAFT-Stereo on the Middlebury data set, for which no samples have been seen during training. The shown results have been produced using the parameter values obtained by training on the Sceneflow data set and fine-tuning on the InStereo2k data set. The best values of our method per column are shown in bold



**Fig. 6** Qualitative comparison on the Middlebury data set. These results are produced with parameter values obtained by training on the Sceneflow data set and fine-tuning on the InStereo2k data set. The second and third row show the disparity and difference maps of the two variants of our full method, respectively, varying only the loss function used for training. Significant differences can especially be seen in weakly textured areas and at depth discontinuities. For an explanation of the color coding, refer to Fig. 4

### 5.3 Behavior Close to Depth Discontinuities

So far, the discussed quantitative results are average values over all pixels of an image, not distinguishing between different parts of the depicted scenes. To allow for a more detailed analysis of the error close to depth discontinuities—the parts of the scene that are given special emphasis in the methodology—the MAE and RMSE for the corresponding parts in the images are determined isolated from the rest of these images and are shown in Table 3. These results demonstrate that both components of our methodology have a positive effect on the disparity estimation close to depth discontinuities and lead to clear improvements compared to the baseline. In this context, especially the clear benefit of using LBP as additional input on the Middlebury data set is to be highlighted, which is contrary to what can be observed from the results for the entire images (cf. Table 2). This indicates that this component has a positive influence on the disparity estimation close to depth discontinuities, but a negative one on other regions of an image, an issue that we will further investigate in future research. Furthermore, it can be stated that the usage of the dice loss is to be preferred over the weighted binary cross-entropy loss on the Middlebury data set, as the latter shows significantly worse results considering both, the entire image as well as depth discontinuities only, probably caused by a worse generalization capability. Lastly, it is to be noted that although the presented method leads to superior results compared to the baseline, a clear gap can still be seen between the overall results and the results that only consider pixels close to depth discontinuities. Consequently, depth discontinuities remain one of the major challenges in dense stereo matching and require further research in the future.

**Table 3** Quantitative results for pixels close to depth discontinuities

Variants			Sceneflow		Middlebury	
4th channel (LBP)	Discontinuities via dice	Discontinuities via wbce	MAE [px]	RMSE [px]	MAE [px]	RMSE [px]
			6.65	14.32	8.55	16.21
✓			3.93	<b>10.61</b>	<b>7.43</b>	14.05
	✓		3.99	11.03	10.73	22.32
✓	✓		3.91	10.76	7.93	<b>13.56</b>
✓		✓	<b>3.81</b>	10.71	8.85	16.80
RAFT-Stereo (Lipson et al. 2021)			4.91	10.51	5.55	9.66

The results of the baseline and the four variants of the method presented in this work are compared to RAFT-Stereo on the Sceneflow and the Middlebury data sets. The best values of our method per column are shown in bold



**Table 4** Quantitative comparison of different feature description methods used as fourth image channel

Variants	Sceneflow				
	Pixel Error Rate [%]			MAE [px]	RMSE [px]
	$\tau = 1$	$\tau = 3$	$\tau = 5$		
Baseline	13.55	8.05	6.55	2.24	9.86
HOG	7.48	3.18	2.20	0.66	3.59
Canny	6.92	3.10	2.18	0.65	3.57
LBP	<b>6.90</b>	<b>3.01</b>	<b>2.13</b>	<b>0.63</b>	<b>3.44</b>

The numeric values shown in the table correspond to the defined Sceneflow test set (see Sect. 4.1). The best values per row are shown in bold

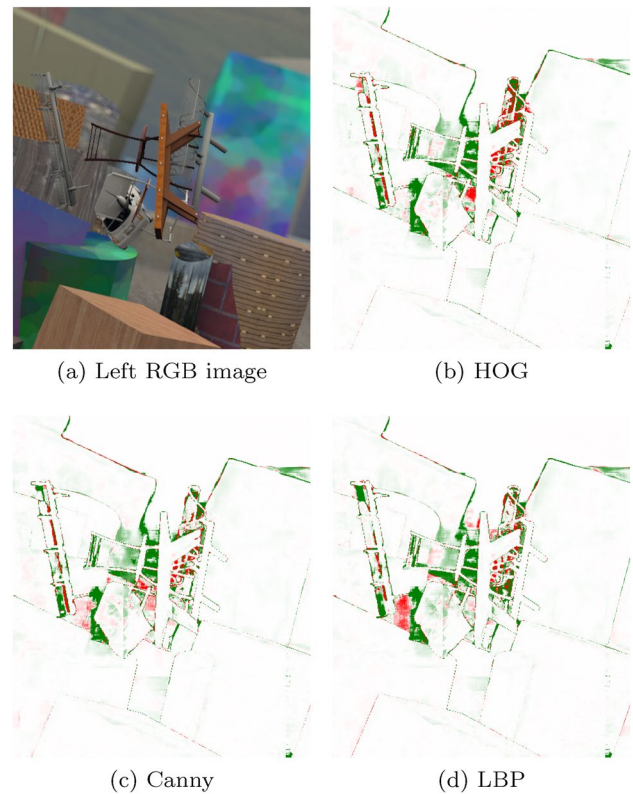
#### 5.4 Influence of the Feature Description Method Used as Fourth Image Channel

In this section, the concept of using a hand-crafted feature description method as a fourth image channel (cf. Sect. 3.2) is investigated in more detail. For this purpose, the proposed usage of Local Binary Patterns (LBP) is compared against the employment of Histograms of Oriented Gradients (HOG) and feature maps resulting from the Canny edge detector. All three feature description methods focus on image gradients and thus follow the general concept of this work to make use of the assumption that image gradients and depth discontinuities are related. To highlight the influence of this part of our method, the prediction of depth discontinuities (cf. Sect. 3.3) is neglected for producing the results discussed in this section.

As demonstrated by the quantitative results in Table 4, the general idea of providing additional guidance to the purely data driven baseline using a hand-crafted feature description method leads to improvements independent of the actually used kind of features. The same is also shown by the difference maps between the baseline and the variants of our method that use one of the listed feature description methods as the fourth image channel in Fig. 7. Especially for pixels close to actual depth discontinuities and for pixels that are falsely identified as such by the baseline, significant improvements can be observed. Having a closer look at the quantitative results, the usage of LBP leads to the best results, although only minor differences can be seen in comparison with the usage of HOG and Canny edge detector.

#### 5.5 Influence of the Loss Function Used for Predicting Depth Discontinuities

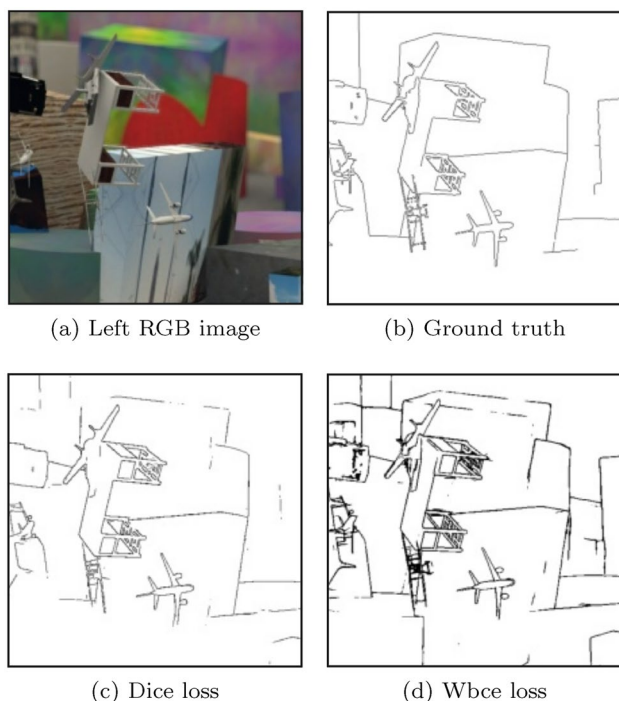
In this section, the influence of the loss function is investigated that is used for training the described U-Net to predict depth discontinuities (cf. Sect. 3.3). As shown and discussed in the previous results, the usage of the



**Fig. 7** Comparison of different feature description methods used as fourth image channel. For an example from the Sceneflow data set, the differences between the baseline and our method using only a fourth image channel (cf. Sect. 3.2) are shown. For an explanation of the color coding, refer to Fig. 4

weighted binary cross-entropy and the dice loss function often leads to clearly different results, which is particularly visible in Figs. 5 and 6. While the wbce loss leads to artifacts and noise in the weakly textured background of the images, these issues do not occur with the dice loss. Comparing the predicted information with the two loss functions shown in Fig. 8, it can be seen that the variant with the dice loss leads to sharper but partially incomplete depth discontinuities, while the wbce-based variant tends to predict more blurry discontinuities and even some discontinuities that do not exist in the true geometry of the scene. Especially latter leads to such artifacts and noise in the disparity estimations that are described above, as the predicted information is used as guidance, encouraging GC-Net to predict incorrect depth discontinuities. These characteristics of the two variants associated with the two different losses are also indicated by the numeric results shown in Table 5. Evaluating the ability of the two variants to distinguish between pixels that are located next to depth discontinuities and such that are not, the dice loss variant shows a higher precision, while the wbce loss variant shows a higher recall. To take into account the





**Fig. 8** Example for the prediction of depth discontinuities. The images show the ground truth and the depth discontinuities predicted using parameter values trained with the dice and the weighted binary cross-entropy loss, respectively

**Table 5** Quantitative comparison of weighted binary cross-entropy and dice loss

	Discontinuities via dice	Discontinuities via wbce
Balanced Accuracy [%]	<b>83.6</b>	70.1
Recall [%]	99.2	<b>99.7</b>
Precision [%]	<b>99.2</b>	96.8
F1 score [%]	<b>99.2</b>	98.2

The numeric values shown in the table correspond to the defined Sceneflow test set (see Sect. 4.1). The best values per row are shown in bold

highly unbalanced nature of the data—only about 10% of all pixels are located close to depth discontinuities—also the balanced accuracy, being the mean of the true positive rate and the true negative rate, is provided. With a balanced accuracy of 70.1% and 83.6% for the variants based on wbce and dice loss, respectively, this metric demonstrates that the ability to predict depth discontinuities is still limited. Addressing similar issues with respect to imbalanced data, Jadon (2020) and Taghanaki et al. (2019) propose a more sophisticated loss function and a combination of different loss functions, respectively, both being

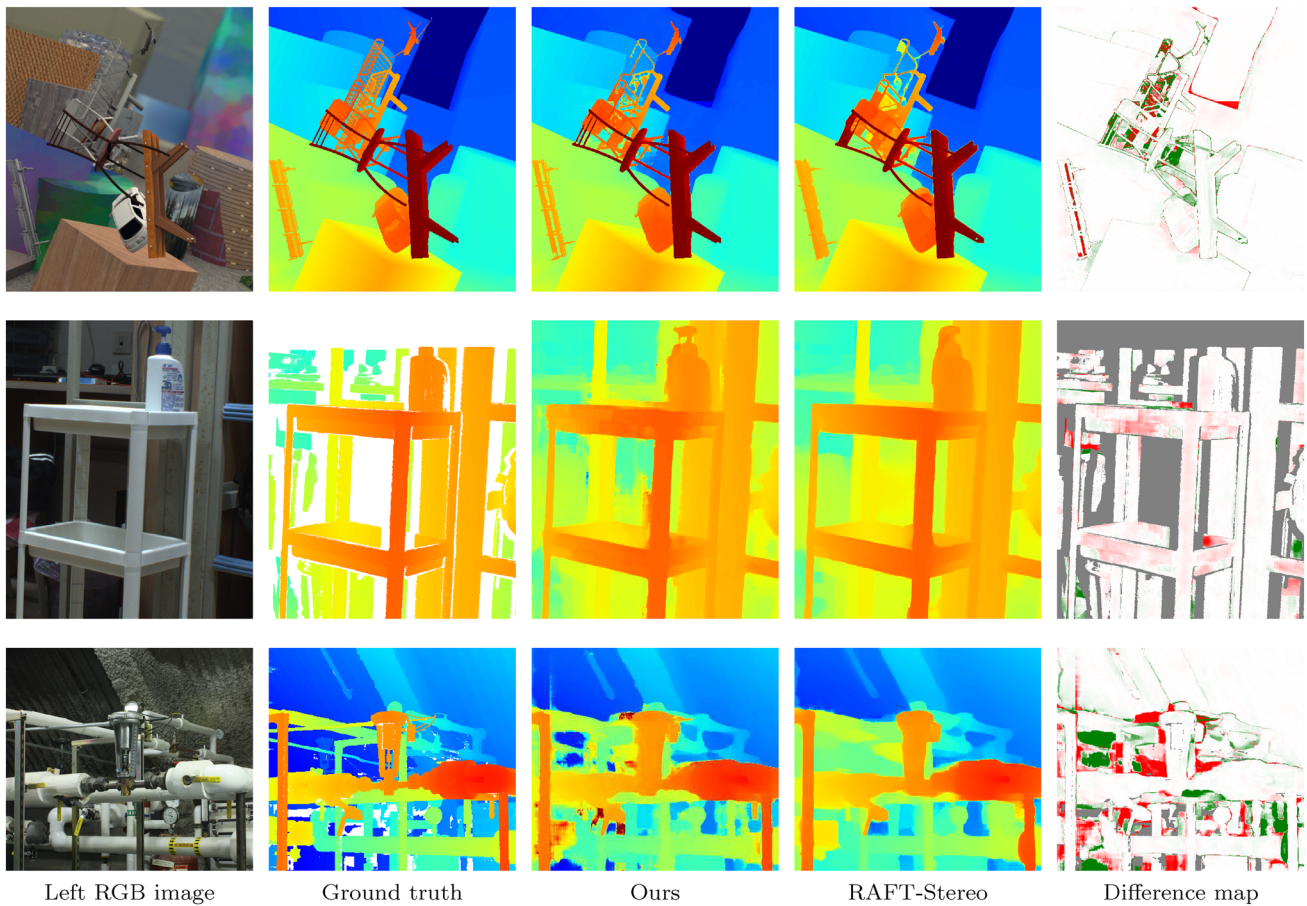
promising directions to further improve the performance of our method in future work.

## 5.6 Comparison Against State-of-the-Art

While the results shown and discussed so far clearly reveal the improvements of the method presented in this work compared to the baseline it is based on, this section focuses on the comparison against the current state-of-the-art. For this purpose, RAFT-stereo (Lipson et al. 2021) is trained and tested under the same conditions and on the same data (cf. Sect. 4.2). Although we use a relatively simple method as baseline, the results achieved on the Sceneflow data set are comparable to those of RAFT-stereo, as shown in Table 1. Even slightly better results can be achieved for pixels close to depth discontinuities with respect to the mean absolute error (cf. Table 3) and as shown in Fig. 9. However, RAFT-Stereo outperforms our approach on the InStereo2k and the Middlebury data sets. It is worth noting that in the InStereo2k data set, which is used for fine-tuning, almost no ground truth is available for the disparity of pixels close to depth discontinuities. While this information is crucial during training to make full use of the concept our method is based on, i.e. learning to estimate the position of depth discontinuities from image gradients, RAFT-Stereo does not focus on these pixels, potentially explaining the differences in the results. However, our method is still able to outperform RAFT-Stereo in certain areas, as visible in the difference maps shown in Fig. 9: On the InStereo2k data set, the disparity estimates of pixels in the background can be improved, by counteracting over-smooth predictions. On the Middlebury data set, the results of RAFT-Stereo also suffer from the foreground fattening effect described in Sect. 5.2, again leading to better results of our method in the background, but more artifacts in the foreground. All in all, these results show that the method presented in this work is also relevant in the context of the current state-of-the-art in dense stereo matching.

## 6 Conclusions

In this work, a novel method for dense stereo matching is presented that supplements deep learning with feature information obtained based on expert knowledge. For this purpose, a twofold strategy is applied: First, the RGB input images are described via Local Binary Patterns and the resulting feature descriptors are attached to the respective images as an additional image channel. In this way, a basic set of feature information is provided to the neural network, allowing it to focus on details that supplement this set. Second, the assumption that edges in an image and discontinuities in the corresponding depth map coincide is modeled



**Fig. 9** Qualitative comparison against the state-of-the-art. From top to bottom, the rows show results from the Sceneflow, InStereo2k and Middlebury data sets, respectively. The results on the Sceneflow data set were computed with parameter values obtained by training on the Sceneflow data set only, the other results were computed with

parameter values obtained by additional fine-tuning on the InStereo2k data set. While *Ours* refers to the variant *LBP + U-Net with dice*, the difference maps show the differences between the results of RAFT-Stereo and ours. For an explanation of the color coding, refer to Fig. 4

explicitly by training a neural network that predicts latter from RGB images. The information obtained in this way is used to guide the deep learning-based matching process.

The evaluation carried out in the context of this work reveals two main findings: The proposed method improves the quality of the depth estimation compared to a purely data driven baseline and it achieves results comparable to the current state-of-the-art in dense stereo matching, if sufficient training data with dense ground truth for pixels close to depth discontinuities is available. Furthermore, it is shown that both components of the method contribute to this improvement. These findings also apply to the subset of pixels that are located close to depth discontinuities and in the presence of a domain gap between training and test data. However, a significant impact of such a gap is still visible, leading to a clear deterioration of the results compared to the scenario in which similar data has been seen during training. To mitigate this problem of over-fitting and to improve the generalization capability, we will investigate whether the

complexity of the model used to predict depth discontinuities can be reduced and we will introduce further geometric constraints.

Comparing different optimization objectives for the prediction of depth discontinuities, it is found that dice loss is superior with respect to the majority of evaluated scenarios and quality metrics, although weighted binary cross-entropy performs slightly better on artificial data in some aspects. As the ratio between pixels that are close to depth discontinuities and pixels that are not is highly imbalanced, special care has to be taken about this issue. While weighting the binary cross-entropy term according to this ratio in the training data worked reasonably well, a more sophisticated strategy such as the log-cosh dice loss function proposed by Jadon (2020) or a combination of different loss functions (Taghanaki et al. 2019) might further improve the results. Lastly, we will investigate the applicability of the presented method on terrestrial outdoor data and airborne images in our future work. As reference data for depth in such outdoor

settings is commonly captured by a laser scanner and is thus sparse, this investigation will include the development of a suitable training strategy to learn the prediction of depth discontinuities from sparse reference data.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Data availability** There are no new data obtained for this paper. However, the data sets used in this paper are sourced from open repositories and can be accessed through the following links: SceneFlow (FlyingThings3D): <https://lmb.informatik.uni-freiburg.de/resources/datasets/SceneFlowDatasets.en.html>, InStereo2k: <https://github.com/YuhuaXu/StereoDataset>, and Middlebury: <https://vision.middlebury.edu/stereo/data/>.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Anwer RM, Khan FS, van de Weijer J, Molinier M, Laaksonen J (2018) Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. *ISPRS J Photogramm Remote Sens* 138:74–85
- Badrinarayanan V, Kendall A, Cipolla R (2017) SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell* 39(12):2481–2495
- Bao W, Wang W, Xu Y, Guo Y, Hong S, Zhang X (2020) InStereo2K: a large real dataset for stereo matching in indoor scenes. *Sci China Inf Sci* 63:212101
- Barnard ST, Fischler MA (1982) Computational stereo. *ACM Comput Surv* 14(4):553–572
- Canny J (1986) A computational approach to edge detection. *IEEE Trans Pattern Anal Mach Intell* 8(6):679–698
- Chang JR, Chen YS (2018) Pyramid stereo matching network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5410–5418
- Cheng X, Wang P, Yang R (2019) Learning depth with convolutional spatial propagation network. *IEEE Trans Pattern Anal Mach Intell* 42(10):2361–2379
- Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? the KITTI vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE. 10.1109/cvpr.2012.6248074
- Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. *Proc Int Conf Artif Intell Stat* 9:249–256
- Häne C, Zach C, Lim J, Ranganathan A, Pollefeys M (2011) Stereo depth map fusion for robot navigation. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems, pp 1618–1625
- Hirschmüller H (2008) Stereo processing by semiglobal matching and mutual information. *IEEE Trans Pattern Anal Mach Intell* 30(2):328–341
- Ilg E, Saikia T, Keuper M, Brox T (2018) Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In: Proceedings of the European conference on computer vision, pp 614–630
- Jadon S (2020) A survey of loss functions for semantic segmentation. In: Proceedings of the IEEE conference on computational intelligence in bioinformatics and computational biology
- Kabbai L, Azaza A, Abdellaoui M, Douik A (2015) Image matching based on LBP and SIFT descriptor. In: Proceedings of the IEEE international multi-conference on systems, signals & devices
- Kang J, Chen L, Deng F, Heipke C (2019) Context pyramid network for stereo matching regularized by disparity gradients. *ISPRS J Photogramm Remote Sens* 157:201–215
- Kendall A, Martirosyan H, Dasgupta S, Henry P, Kennedy R, Bachrach A, Bry A (2017) End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE international conference on computer vision, pp 66–75
- Krutikova O, Sisojevs A, Kovalovs M (2017) Creation of a depth map from stereo images of faces for 3D model reconstruction. *Proc Comput Sci* 104:452–459
- Lipson L, Teed Z, Deng J (2021) RAFT-stereo: multilevel recurrent field transforms for stereo matching. In: Proceedings of the international conference on 3D vision, pp 218–227
- Liu YL, Lai WS, Yang MH, Chuang YY, Huang JB (2021) Hybrid neural fusion for full-frame video stabilization. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2299–2308
- Luo W, Schwing AG, Urtasun R (2016) Efficient deep learning for stereo matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5695–5703
- Mayer N, Ilg E, Hausser P, Fischer P, Cremers D, Dosovitskiy A, Brox T (2016) A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4040–4048
- Nguyen U, Heipke C (2020) 3D pedestrian tracking using local structure constraints. *ISPRS J Photogramm Remote Sens* 166:347–358
- Ojala T, Pietikainen M, Maenpää T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell* 24(7):971–987
- O'Mahony N, Campbell S, Carvalho A, Harapanahalli S, Hernandez GV, Krpalkova L, Riordan D, Walsh J (2020) Deep learning vs. traditional computer vision. In: Advances in computer vision, pp 128–144
- Pang J, Sun W, Ren JSJ, Yang C, Yan Q (2017) Cascade residual learning: a two-stage convolutional neural network for stereo matching. In: Proceedings of the IEEE international conference on computer vision workshops, pp 887–895
- Qi X, Liu Z, Liao R, Torr PHS, Urtasun R, Jia J (2020) GeoNet++: iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation. *IEEE Trans Pattern Anal Mach Intell* 44(2):969–984
- Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of the international conference on medical image computing and computer-assisted intervention, pp 234–241
- Saba T, Mohamed AS, El-Affendi M, Amin J, Sharif M (2020) Brain tumor detection using fusion of hand crafted and deep learning features. *Cogn Syst Res* 59:221–230
- Scharstein D, Szeliski R (2002) A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int J Comput Vis* 47(1–3):7–42

- Scharstein D, Hirschmüller H, Kitajima Y, Krathwohl G, Nešić N, Wang X, Westling P (2014) High-resolution stereo datasets with subpixel-accurate ground truth. In: Proceedings of the German conference on pattern recognition, pp 31–42
- Shaked A, Wolf L (2017) Improved stereo matching with constant highway networks and reflective confidence learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4641–4650
- Shamsafar F, Woerz S, Rahim R, Zell A (2021) MobileStereoNet: towards lightweight deep networks for stereo matching. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2417–2426
- Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from RGBD images. In: Proceedings of the European computer vision conference, pp 746–760
- Simonyan K, Zisserman A (2014) A very deep convolutional networks for large-scale image recognition. In: Proceedings of the international conference on learning representations
- Stucker C, Schindler K (2022) ResDepth: a deep residual prior for 3D reconstruction from high-resolution satellite images. *ISPRS J Photogramm Remote Sens* 183:560–580
- Stucker C, Ke B, Yue Y, Huang S, Armeni I, Schindler K (2022) Implicit: city modeling from satellite images with deep implicit occupancy fields. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences V-2-2022*:193–201
- Taghanaki SA, Zheng Y, Zhou SK, Georgescu B, Sharma P, Xu D, Comaniciu D, Hamarneh GCL (2019) Combo loss: handling input and output imbalance in multi-organ segmentation. *Comput Med Imaging Graphics* 75:24–33
- Tianyu Z, Zhenjiang M, Jianhu Z (2018) Combining CNN with hand-crafted features for image classification. In: Proceedings of the IEEE international conference on signal processing, pp 554–557
- Tieleman T, Hinton G (2012) Lecture 6.5 - RMSprop: divide the gradient by a running average of its recent magnitude. In: COURSERA: neural networks for machine learning
- Tosi F, Liao Y, Schmitt C, Geiger A (2021) SMD-Nets: stereo mixture density networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8942–8952
- Verma D, Puri S, Prabhu S, Smriti K (2020) Anomaly detection in panoramic dental X-rays using a hybrid deep learning and machine learning approach. In: Proceedings of the IEEE region 10 conference, pp 263–268
- Xiang X, Zhai M, Zhang R, Qiao Y, El Saddik A (2018) Deep optical flow supervised learning with prior assumptions. *IEEE Access* 6:43222–43232
- Yinan W, Nuo Z, Toshinori W, Hisashi K (2012) An effective method for image matching based on modified LBP and SIFT. In: Proceedings of the international conference on computer vision theory and applications, pp 99–110
- Zabih R, Woodfill J (1994) Non-parametric local transforms for computing visual correspondence. In: Proceedings of the European conference on computer vision, Springer, Berlin, Heidelberg, pp 151–158
- Zbontar J, LeCun Y (2016) Stereo matching by training a convolutional neural network to compare image patches. *J Mach Learn Res* 17(1):2287–2318
- Zhang F, Prisacariu V, Yang R, Torr PHS (2019) GA-Net: guided aggregation net for end-to-end stereo matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 185–194