

Received October 25, 2018, accepted November 27, 2018, date of publication December 10, 2018, date of current version January 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2885773

Weakly Supervised Deep Depth Prediction Leveraging Ground Control Points for Guidance

LIANG DU^{1,2}, JIAMAOL^{1,2}, (Member, IEEE), XIAOQING YE^{1,2},
AND XIAOLIN ZHANG^{1,2,3}, (Member, IEEE)

¹Bio-Vision System Laboratory, State Key Laboratory of Transducer Technology, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China

²School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

³School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China

Corresponding author: Liang Du (duliang@mail.ustc.edu.cn)

This work was supported by the Shanghai Committee of Science and Technology, China, under Grant 18ZR1447500.

ABSTRACT Despite the tremendous progress made in learning-based depth prediction, most methods rely heavily on large amounts of dense ground-truth depth data for training. To solve the tradeoff between the labeling cost and precision, we propose a novel weakly supervised approach, namely, the Guided-Net, by incorporating robust ground control points for guidance. By exploiting the guidance from ground control points, disparity edge gradients, and image appearance constraints, our improved network with deformable convolutional layers is empowered to learn in a more efficient way. The experiments on the KITTI, Cityscapes, and Make3D datasets demonstrate that the proposed method yields a performance superior to that of the existing weakly supervised approaches and achieves results comparable to those of the semisupervised and supervised frameworks.

INDEX TERMS Computer vision, stereo image processing, stereo vision, weakly supervised learning.

I. INTRODUCTION

Depth estimation has been extensively studied for decades due to its indispensable role in autonomous driving [26], 3D reconstruction [34], scene understanding and human pose estimation [37]. Traditional depth prediction algorithms take advantage of multiple views or observations with shading and motion because estimating depth from a single image is an ill-posed problem. With the development of deep learning, recent interest has focused on learning-based methods to perform single/stereo depth prediction, and promising performances have been reported with supervised learning-based approaches [6], [23]. However, a substantial drawback of supervised learning methods is their dependence on a vast amount of ground-truth depth data for training. Active RGB-D cameras are usually adopted in indoor scenes, and LiDAR laser systems are popularly applied for outdoor environments. However, noise and sparsity are unavoidable in outdoor measurements, resulting in a much sparser ground truth with potential errors and missing details. An alternative method is to use synthesized dense ground-truth datasets [25] that lack realistic image characteristics.

In contrast to the abovementioned labeled data-driven methods, weakly supervised frameworks [9], [11] have recently been exploited to directly predict disparity maps from images without the need for ground truth. However, the resulting depth is far from accurate due to a lack of ground truth. To compensate for this loss of accuracy, semisupervised methods [20] have been introduced to incorporate semi-dense laser depth measurements with rectified stereo image pairs. The significant improvement over single weakly supervised depth prediction demonstrates the vital role of supervision. However, these methods require a large quantity of ground-truth depth data, and the network can be affected by the noisy erroneous labels acquired by LiDAR.

In this paper, we propose a weakly supervised learning framework that takes advantage of ground control points (GCPs) for guidance in the training process. GCPs, which are points with high reliability of true disparity, are first incorporated into the graph cut [15] or Markov random field (MRF) constraint [42]. In this work, we perform an oriented FAST and rotated BRIEF (ORB) feature point matching across rectified binocular images with our novel

bad-match-filtering method (DDS) to further improve the confidence. In our autoencoder framework, the feature-point-based GCPs are complemented by the combined training loss and guide the network to learn a better image reconstruction. Note that our approach does not require explicit ground-truth depth data, such as in semisupervised methods, to achieve remarkable results. Consequently, we use the term “guidance” rather than supervision to demonstrate our GCP-fused approach. The right image is required in only the training stage; it is not needed for prediction. An example of our approach can be seen in Fig. 1.

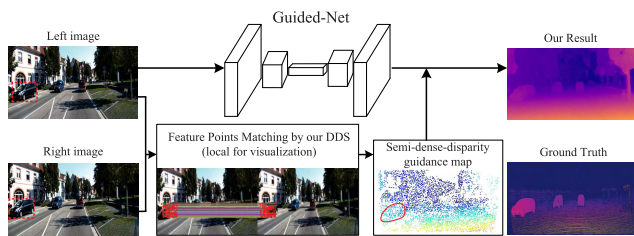


FIGURE 1. Our Guided-Net for weakly supervised depth prediction. From left to right: rectified stereo image pairs, our improved skip-layer network with deformable convolutional layers, part of the feature points matching map based on our novel DDS method for visualization, our semi-dense guidance map and the predicted result.

In summary, we propose the following contributions:

- A novel weakly supervised learning framework for single-image depth estimation that incorporates feature-point-based GCPs to guide the learning task.
- A novel algorithm based on disparity distribution statistics (DDS) for stereo feature point matching, which can generate both robust and accurate matches.
- A combination of training losses that includes the guidance loss, edge loss and reconstructed appearance loss.
- State-of-the-art performance in challenging outdoor scenes, such as KITTI [10], that achieves a performance superior even to those of several semisupervised and supervised approaches.

II. RELATED WORK

Recent years have witnessed increasing developments in learning-based depth estimation approaches motivated by the availability of large outdoor datasets such as KITTI [10] and Cityscapes [4]. Various methods can be applied, for example, stereo matching, multi-view overlapped images and single-image prediction.

A. LEARNING-BASED STEREO MATCHING

A convolutional neural network (CNN) was first leveraged to learn local feature representations for stereo matching [3], [49], [50]. The dot product was adopted on top of the convolutional layers to measure the similarity in [50]. Chen *et al.* [3] adopted multiscale patches as input to fuse hierarchical features for the final decision. However, these methods merely utilize CNNs in the matching cost computation stage; they still rely on traditional postprocessing

methods for accurate disparity computation. End-to-end stereo and flow networks are trained by a synthetic dataset to directly obtain the correspondence between two given binocular images [25]. Kendall *et al.* [18] incorporated contextual information using 3D convolutions over the cost volume in an end-to-end regression framework. Despite the remarkable performances achieved by the abovementioned approaches, they all rely on a pair of images and dense ground-truth disparity maps, which are almost impossible to obtain in real world outdoor scenes; thus, their training is constrained to relatively small datasets.

B. MONOCULAR DEPTH ESTIMATION

Recently, learning-based monocular depth estimation methods have been proposed to solve the ill-posed problem by learning the potential cues from appearance and distance. Saxena *et al.* [31] adopted a discriminatively trained MRF with multiscale local and global features under the supervision of ground-truth depth. Simonyan and Zisserman [38] addressed monocular depth prediction by applying two sub-networks, one for coarse global estimation based on the entire image and the other for further refinement. Rather than using hand-crafted features, Liu *et al.* [23] learned the unary and pairwise potentials of continuous CRF in a unified deep framework. Laina *et al.* [21] proposed a fully convolutional architecture to model the ambiguous mapping between monocular images and depth maps. Semantic segmentation was further integrated to improve the performance of monocular depth estimation [22], [43]. Kendall and Gal [17] presented a novel Bayesian deep learning framework to learn a mapping for aleatoric uncertainty from the input data, and it is composed on top of the epistemic uncertainty models. These authors derived their framework for monocular depth estimations. Xu *et al.* [46] addressed monocular depth estimations by a novel framework based on continuous CRFs for fusing multiscale representations derived from CNN side outputs.

Semisupervised methods attempt to learn depth estimation with limited supervision and to improve the performance of weakly supervised strategies. Kuznetsov *et al.* [20] used semi-dense LiDAR depth as the ground truth for supervised learning in combination with an image appearance loss. However, mounting labeled data are still required. Deep3D [45] was trained directly on stereo pairs to predict a probabilistic disparity-like map as an intermediate output and minimize the pixelwise reconstruction error.

Given the high cost of acquiring ground-truth depth data, weakly supervised learning approaches are attracting increasing attention. Garg *et al.* [9] exploited a coarse-to-fine architecture for depth estimation based on reconstruction loss; nevertheless, the framework is not fully differentiable and requires linear approximation. Godard [11] *et al.* employed an image reconstruction loss and left-right consistency to perform multiscale encoder-decoder training. The carefully designed loss function ensured detailed depth prediction, but large erroneous regions in disparity discontinuities and

thin objects remained. Tonioni *et al.* [40] use the off-the-shelf stereo and confidence algorithms to train the network. Yin and Shi [47] proposed a jointly weakly supervised learning framework for monocular depth, egomotion estimation and optical flow from video sequences. Geometric relationships are extracted over the predictions of individual modules and then combined as an image reconstruction loss, reasoning about static and dynamic scene parts separately. An adaptive geometric consistency loss is proposed to increase robustness towards outliers and non-Lambertian regions, which resolves occlusions and texture ambiguities. Zhan *et al.* [51] constructed an end-to-end visual odometry and depth estimation network by using stereo sequences, enabling the use of both spatial and temporal photometric warp error, and constrains the scene depth and camera motion to be in a common, real-world scale. They also improved the depth estimation network by using the deep feature-based warping loss instead of the photometric warp loss. Kundu *et al.* [19] proposed a novel unsupervised domain adaptation method, AdaDepth, for adapting depth predictions from synthetic RGB-D pairs to natural scenes, and they demonstrated AdaDepth's efficiency in adapting learned representations from synthetic to real scenes via empirical evaluations of challenging datasets. To this end, we propose a guidance strategy to improve the performance of weakly supervised approaches. We first extract semi-dense but reliable GCPs by means of real-time ORB feature point matching based on our DDS. Then, the guidance maps are used to enforce an additional loss by minimizing the error between the control point disparity and the estimation. This strategy does not require supervision from ground-truth depth or video sequences; thus, we call it guidance rather than supervision.

C. GROUND CONTROL POINTS

GCPs were first used in global stereo matching [42] and Structure from Motion [28] to add constraints to global optimization. Traditional handcrafted selection of GCPs can be implemented by a variety of confidence measures [14], [36] as well as semi-dense key point matching. Recently, an ultra-robust outlier filtering algorithm for feature point matching named GMS [1] enabled the translation of high match numbers into high match quality. Alternatively, many methods leverage deep learning to compute the confidence and select GCPs. CNN was first utilized for confidence measurement by discriminating disparity patches whose difference with the median value is lower than a threshold [35]. Poggi and Mattoccia [30] enforced the local consistency assumption by exploiting a deep patch-based network for confidence measurement. GCPs learned by a supervised learning approach were further integrated with an MRF-stereo framework to improve the matching accuracy [39]. In our work, to make training easier, we adopt real-time modified ORB feature point matching for selecting GCPs instead of learning-based strategies. Specifically, we constrain the searching range according to the left-right consistency and maximum disparity of the calibrated image pairs and further eliminate

outliers based on the probability distributions of the semi-dense guidance points for disparity.

III. THE PROPOSED APPROACH

This section introduces our multiple semi-dense disparity guidance map and illustrates how it is incorporated into our improved depth prediction framework. A novel depth estimation training loss is presented by adopting robust feature point matching to guide the learning process, which enables us to train on stereo image pairs without requiring supervision from the ground-truth depth. During prediction, only the left image is required to generate the final disparity map. An overview of our approach is presented in Fig. 2.

During the training process, given a pair of rectified stereo images, our goal is to predict the disparity value for each pixel in the left image, which represents the difference in the horizontal coordinate of the same point projecting on the left and right images. Accordingly, if given a left image and left disparity map, the right image can be reconstructed, and vice versa. The warping functions of the left and right images are as follows:

$$\begin{aligned} I_l(x_l) &= I_r(x_r + d_r) \\ I_r(x_r) &= I_l(x_l - d_l) \end{aligned} \quad (1)$$

A. SEMI-DENSE DISPARITY GUIDANCE MAP BASED ON DDS

Depth measurement by LiDAR has some shortcomings, such as measurement noise, requirement for calibration precision between sensors, sparsity of the distance measurements, limited range of the measurement and unaffordable high cost. In contrast to the LiDAR-dependent semisupervised monocular depth prediction approach [20], we obtain a semi-dense guidance map by robust feature point matching. Compared with sparse key feature point matching in SLAM [29], which extracts only the high-confidence points for camera position computation, we require dense feature points to form the guidance map. However, the traditional feature point matching methods based on SIFT [2] and RANSAC [7] are not sufficiently dense and accurate to produce both robust and evenly distributed matching points, as introduced in [1].

First, our method achieves basic feature matches by leveraging the characteristics of rectified stereo image pairs. Stereo-rectified images have provided useful a priori geometric constraints for feature point matching, and they are used here for filtering incorrect matches to adopt more reliable matching points computed by ORB. The details are explained in the following section. We limit the searching range of the stereo images according to the consistency of rectified pairs. Specifically, the matching points should be in the same row, and the column coordinate of the matching point in the left image should always be greater than the corresponding point in the right image due to the nonnegative disparity values. Furthermore, we consider the influence of the maximum disparity, which helps to avoid some wrong matches. A novel method called DDS (disparity distribution statistics), which

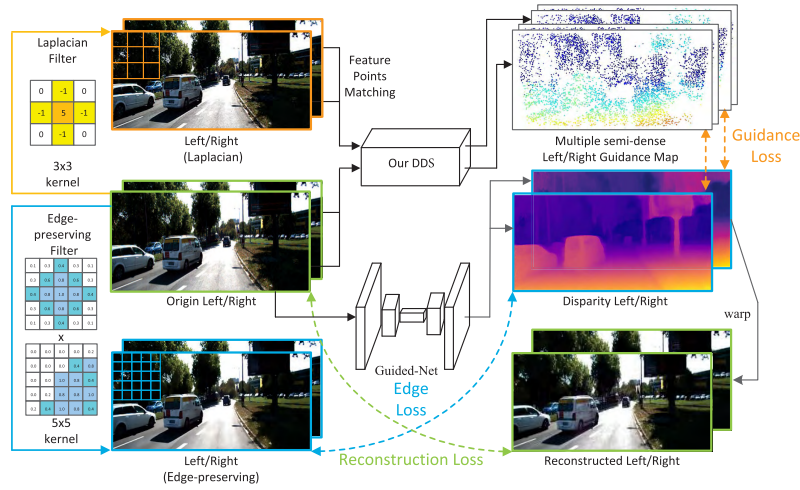


FIGURE 2. The proposed framework for depth prediction guided by semi-dense ground control points. Our DDS filters the incorrect matches and provides a multiple semi-dense guidance map based on multi-actance images. Training losses consist of three items: the guidance loss computed from multiple semi-dense guidance maps and disparity maps, the edge loss computed from edge-preserving images and disparity maps, and the reconstruction loss computed from the original and reconstructed images. Note that during the prediction stage, only the original left-hand image is required.

is a statistics-based bad-match-filtering method according to the disparity distribution probability, is proposed to obtain sufficiently dense feature point matches and to filter the outliers for accuracy. It is a reliable method of encapsulating disparity smoothness as the statistical likelihood of a certain number of matches in a region.

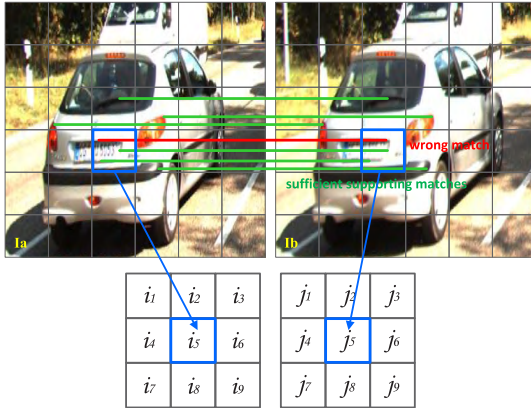


FIGURE 3. Nine regions around cells $\{i, j\}$ used in score evaluation in GMS [1]. The red match is the wrong match ignored by GMS, which also has sufficient supporting matches.

Previous work [1] has proved that true matched neighborhoods view the same 3D region and thus share many similar features across both images, which results in a number of supporting matches in the neighborhood. In contrast, false matching neighborhoods view different 3D regions and have far fewer similar features, which reduces the matching support. In the GMS algorithm [1], after each feature in image I_b has found its nearest neighbor in image I_a as shown in Fig. 3, all

the matches are filtered according to the following formula:

$$cellpair\{i, j\} \in \begin{cases} \mathcal{T}, & \text{if } S_{ij} > \tau_i = \alpha \sqrt{n_i} \\ \mathcal{F}, & \text{otherwise} \end{cases} \quad (2)$$

where thresholding S_{ij} is used to divide $cellpairs\{i, j\}$ into true and false sets \mathcal{T}, \mathcal{F} . $\alpha = 6$ is experimentally determined and n_i is the average number of features (of the 9 grid-cells in Fig. 3) present in a single grid-cell. The score S_{ij} for $cellpairs\{i, j\}$ is:

$$S_{ij} = \sum_{k=1}^9 |X_{ijk}| \quad (3)$$

where $|X_{ijk}|$ is the number of matches between cells i^k, j^k shown in Fig. 3.

Different from [1], we converted the disparity smoothness constraints into statistical measures to eliminate the error matching. Each reliable match should have both sufficient supporting matches and satisfy the disparity distribution according to the constant-local-disparity [24]. The disparity calculation formula for the left matched feature point is as follows:

$$\begin{cases} \mathcal{D}^l = X^l - X^r \\ \mathcal{D}_{ps_1}^l = X_{ps_1}^l - X_{ps_1}^r \\ \mathcal{D}_{ps_2}^l = X_{ps_2}^l - X_{ps_2}^r \\ \dots \\ \mathcal{D}_{ps_n}^l = X_{ps_n}^l - X_{ps_n}^r \end{cases} \quad (4)$$

$$\begin{aligned} \Rightarrow \Delta \mathcal{D}^l &= \mathcal{D}^l - \mathcal{D}_{ps_n}^l \\ &= (X^l - X_{ps_n}^l) - (X^r - X_{ps_n}^r) \\ |X^l - X_{ps_n}^l| &\in [0, \delta_{local}) \end{aligned} \quad (5)$$

where X^l and X^r are the horizontal ordinates of the matching features in the left and right images, \mathcal{D}^l is the disparity of the left feature points, and δ_{local} is a relatively smaller positive value. Here, ps_n is the label of the n potential supporting feature matches in the δ_{local} area in the left image, and $\Delta\mathcal{D}^l$ is the difference between the disparity value of the matching feature point and its potential support ps_n .

$$\text{if : } |\Delta\mathcal{D}^l| \in [0, \delta_D) \quad (6)$$

$$\text{then : } |X^r - X_{ps_n}^r| \in [0, \delta_D - \delta_{local}) \quad (7)$$

where δ_D is a relatively smaller disparity value. According to formulas (5) and (6), if the disparity value of the left matching point is closer to the value ($|\Delta\mathcal{D}^l| \in [0, \delta_D)$) around the local area ($|X^l - X_{ps_n}^l| \in [0, \delta_{local})$), we can obtain formula (7). According to formula (7), the smaller the value of δ_D is, the closer the potential supporting feature matching point is distributed around the matching point. If the disparity difference $\Delta\mathcal{D}^l$ between the matched point and the potential supporting point satisfies $\Delta\mathcal{D}^l - \delta_{local} \leq \delta_{local}$, this potential supporter is true. If the disparity difference between the matched point and the potential supporting point is too large, ($\Delta\mathcal{D}^l - \delta_{local} > \delta_{local}$), this potential support point is filtered. Similar to the rationale used in formula (2), if there are sufficient such supporting points with similar disparity, the matching point is a reliable point. For GMS:

$$\text{if : } \begin{cases} |X^l - X_{ps_n}^l| \in [0, \delta_{grid}) \\ |X^r - X_{ps_n}^r| \in [0, \delta_{grid}) \end{cases} \quad (8)$$

$$\text{then : } |\Delta\mathcal{D}^l| \in [0, 2\delta_{grid}) \quad (9)$$

where δ_{grid} is the size of the grids. From the formula, we can see that $\Delta\mathcal{D}^l$ is dependent only on the δ_{grid} . In other words, even when the supporting points around the match are sufficient, the disparity smoothness in a local area cannot be guaranteed. As shown in Fig. 3, for those areas with repeated textures during stereo matching, the method of [1] may ignore incorrect matches that also have sufficient neighborhood supporting matches, because the local smoothness of the disparity map is not considered. In conclusion, for stereo feature point matching, the smoother and denser the disparity feature points around the match are, the more reliable this match is.

According to our deduction above, we use the disparity-statistical method to determine whether sufficient dense points with similar disparity exist around a matching point. We first compute the disparity map according to all the matching points. Then, for each matching point, we compute the histogram of disparity values in its local neighbor region. If a disparity value similar to this point occurs very frequently, then many reliable supporting disparity points exist around this point.

More specifically, we use the grid framework for fast statistics. We divide the disparity map by nonoverlapping $k \times k$ grid regions. All the disparity values are divided into n classes (uniformly distributed n intervals from minimum to maximum disparity). By counting the number of points in each class, we obtain the disparity histogram. Then, we

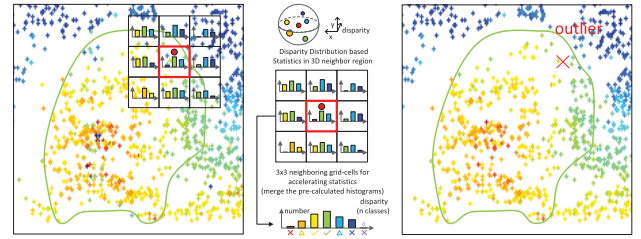


FIGURE 4. Our DDS stereo matching method to obtain both robust and accurate feature point matches based on the statistics of the disparity distribution. The wrong red point is filtered according to our DDS, which considers the local smoothness of the disparity map. Note that the left disparity map is the one without using our DDS.

calculate the disparity histogram in each grid. As shown in Fig. 4, we define the 3×3 neighbor grid regions as the local neighboring regions of the red point and then merge their histograms. If the red point belongs to the top S high frequency classes, we regard it as a reliable point. The higher the frequency of the red point's class is, the more supporting points that red point has. For those background regions of the image, the disparity distribution has a large span; consequently, the disparity histogram is uniform, and the potential right point may not belong to the top S high frequency classes. In such situations, we further consider whether the number of the points in its disparity class exceeds β of the number of top 1 high frequency class. When the red point does not satisfy the disparity distribution mentioned above, it will be rejected as an outlier. Based on experiments, we set the grid size to 30×30 , $n = 10$, $S = 3$, and $\beta = 1/3$ are experimentally determined to limit the $|\Delta\mathcal{D}^l|$ between the red point and its supporting point to a small value, which guarantees the smoothness of the local disparity. Our DDS algorithm is shown in Algorithm 1.

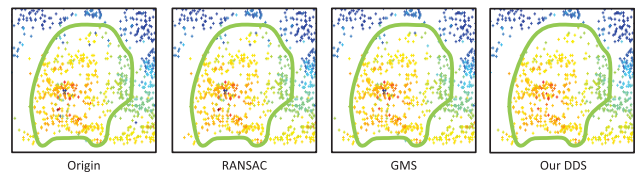


FIGURE 5. Our method handles outliers successfully. Left to right: Original semi-dense disparity map produced by ORB with the limitation of the rectified stereo image, the disparity maps with additional RANSAC [7], GMS [1] and our DDS method. Note that the semi-dense guidance map produced by our DDS has fewer incorrect disparities, which makes the loss function more reliable.

Fig. 5 provides an example illustrating the comparison of the normal stereo feature point matching method without any filtering algorithm with RANSAC [7], with GMS [1] and our DDS method for stereo matching. The result in Fig. 5 shows that the guidance points extracted by our approach are more accurate and reliable than those obtained without using our method. We evaluated our DDS on the KITTI 2015 stereo 200 training set. According to the available disparity points in the images, our DDS increases the accuracy by nearly 10% compared to that of GMS.

Algorithm 1 DDS Feature Matcher**Input:** A pair of rectified stereo images;**Output:** All reasonable matches \mathcal{P} ;

- 1: Detect feature points and perform feature points matching;
- 2: Filter outliers according to the a priori geometric constraints of the rectified stereo image;
- 3: Calculate the disparity values of the I matching points and generate the disparity map;
- 4: Divide I points into n classes (uniformly distributed n intervals from minimum to maximum disparity);
- 5: Divide the disparity map by nonoverlapping M grid regions ($k \times k$) and compute the disparity histograms of the n classes in each region;
- 6: Define the point i in grid m as \mathcal{P}_m^i , and its class is \mathcal{C}_m^i . The number of points in this class is \mathcal{N}_m^i ;
- 7: For each grid, merge the histograms of the 3×3 grids. Sort the merged histogram by the number of points in the class, from large to small. Here, ϑ_m are the sets of the top S high frequency classes. Define the maximum number of classes in each region as N_m^{\max} , and β as its coefficient;
- 8: $\mathcal{P} = \emptyset$
- 9: **for** $m = 0$ to M **do**
- 10: **for** $i = 0$ to I_m **do**
- 11: **if** $\mathcal{C}_m^i \in \vartheta_m$ **then**
- 12: $\mathcal{P} = \mathcal{P} \cup \mathcal{P}_m^i$;
- 13: **else**
- 14: **if** $\mathcal{N}_m^i \geq \beta \times N_m^{\max}$ **then**
- 15: $\mathcal{P} = \mathcal{P} \cup \mathcal{P}_m^i$;
- 16: **end if**
- 17: **end if**
- 18: **end for**
- 19: **end for**
- 20: **return** \mathcal{P} ;

To compare the computational efficiency of our method with that of the GMS, we use the OpenCV ORB features and set the feature number to 10000 following by [1]. The GMS takes 1 ms in single thread CPU-time, and our DDS takes approximately 5ms. The computational efficiency of our grid-based algorithm is close to that of the GMS, which also performs better than other state-of-the-art and traditional methods mentioned in [1].

Fig. 9 and Fig. 11 show the qualitative stereo feature point matching results based on our DDS.

B. MONOCULAR DEPTH ESTIMATION NETWORK WITH DEFORMABLE CONVOLUTION

We adopt the encoder-decoder scheme for our monocular depth prediction network. The architecture is shown in Fig. 6. Note that the semi-dense guidance map and the right image are used only in the training process to compute the loss and are not fed into the network. During prediction, there is no need for right images, which is why we categorize our

approach as a monocular branch. The input of our network is a left RGB image. The right RGB image as well as the left and right semi-dense guidance map, are used directly at the output of our network. The encoder resembles a ResNet-50 [12] architecture and subsamples the input image in 5 stages. The first stage convolves the image to half of the input resolution and each successive stage stacks multiple residual blocks. After reaching the largest receptive field, the decoder is applied to obtain the output with the same resolution of the input image using the corresponding residual blocks.

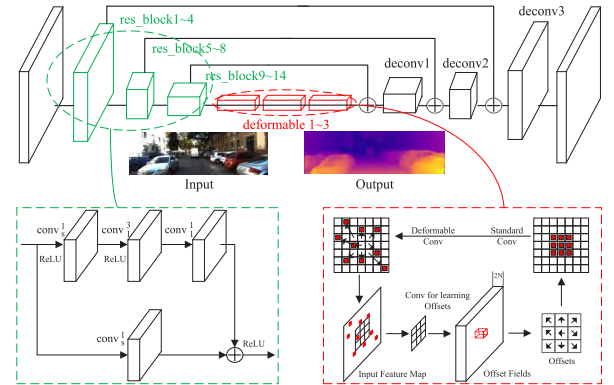


FIGURE 6. Our Guided-Net deep residual network with deformable convolution. The blocks colored in red denote the deformable convolutional layers. The blocks colored in green denote the *res_block*s, with stride s . The residual is obtained from 3 successive convolutions, while the first convolution applies stride s . An additional convolution applies the same stride s and projects the input to the number of channels of the residual.

Inspired by [11], we incorporate the skip-layer connection into our network to merge both edge and context information. We also replace the common convolution layers with deformable convolution [5] layers because deformable convolution can enhance the geometric transformation modeling capability of CNNs. The process is illustrated in Fig. 6. The deformable convolution adds 2D offsets to the regular grid sampling locations in the standard convolution, which enables free-form deformation of the sampling grid. The offsets are learned from the preceding feature maps via additional convolutional layers. The offsets are obtained by applying a convolutional layer over the same input feature map. The convolution kernel is of the same spatial resolution and dilation as those of the current convolutional layers. The output offset fields have the same spatial resolution as the input feature map. The channel dimension of $2N$ corresponds to N 2D offsets. The convolutional kernels for generating the output features and the offsets are learned simultaneously during training. Our model succeeds in preserving clear and detailed contours after adopting the deformable layers. The details of our network architecture are shown in Table 1.

C. TRAINING LOSS FUNCTION FOR THE NETWORK

We present a novel loss function that incorporates three weakly supervised losses, including our novel guidance and

TABLE 1. Our network architecture. We denote a convolution with a filter size of $k \times k$ and a stride s by conv_s^k . The same notation applies to pooling, deconvolution and deformable convolution layers, e.g., max_pool_s^k , deconv_s^k and deformable_s^k , respectively. Each convolution layer is followed by batch normalization except for the last layer in the network. We use ReLU activation functions on the output of the convolutions except at the inputs to the sum operation of the residual blocks, where the ReLU functions come after the sum operation. res_block_s denotes the residual block with stride s at its first convolution layer (see Figs. 6 for details). The input in the table corresponds to the input of each layer where + is a concatenation.

Layer	Channels I/O	Scaling	Input
conv_2^{17}	3/64	2	RGB
max_pool_2^{13}	64/64	4	conv1
res_block_{11}	64/256	4	max_pool1
res_block_{21}	256/256	4	res_block1
res_block_{31}	256/256	4	res_block2
res_block_{42}	256/512	8	res_block3
res_block_{51}	512/512	8	res_block4
res_block_{61}	512/512	8	res_block5
res_block_{71}	512/512	8	res_block6
res_block_{82}	512/1024	16	res_block7
res_block_{91}	1024/1024	16	res_block8
res_block_{101}	1024/1024	16	res_block9
res_block_{111}	1024/1024	16	res_block10
res_block_{121}	1024/1024	16	res_block11
res_block_{131}	1024/1024	16	res_block12
res_block_{142}	1024/2048	32	res_block13
deformable_{13}^3	2048/2048	32	resblock14
deformable_{23}^3	2048/2048	32	deformable1
deformable_{33}^3	2048/1024	32	deformable2
deconv_{13}^3	1024/512	16	deformable3
deconv_{23}^3	512/256	8	deconv1+resblock13
deconv_{33}^3	256/128	4	deconv2+resblock7
deconv_{43}^3	128/64	2	deconv3+resblock3
conv_{31}^3	64/2	2	deconv4

edge loss. As shown in Fig. 2, edge loss is based on the edge gradients of the image processed by a bilateral filter and the disparity image, which includes less weak texture. Guidance loss is based on multiple semi-dense guidance maps and disparity maps. Guidance left and edge left are the enhanced images obtained using the Laplacian filter and edge-preserving filter. The combined loss function is defined in the following formula:

$$\mathcal{E}(I_l, I_r) = w_g \mathcal{E}_{\text{guidance}}(I_l, I_r) + w_e \mathcal{E}_{\text{edge}}(I_l, I_r) + w_r \mathcal{E}_{\text{reconstruction}}(I_l, I_r) \quad (10)$$

where w_g , w_e and w_r are the corresponding weights of the three weakly supervised loss items. The loss of guidance based on multiple semi-dense guidance maps for disparity is $\mathcal{E}_{\text{guidance}}$, the loss of the edge based on the disparity gradient with less weak texture is $\mathcal{E}_{\text{edge}}$, and the loss of the reconstructed image appearance error is $\mathcal{E}_{\text{reconstruction}}$. Note that we discard the loss of the left-right consistency used in [11], because we symmetrically calculate all the loss functions for the left and right images, which implicitly enforces left-right consistency. I_l and I_r are the rectified left and right images. The detailed meanings are explained in the following.

1) GUIDANCE LOSS BASED ON MULTIPLE SEMI-DENSE GUIDANCE MAPS

The guidance loss measures the deviation of the predicted disparity map obtained by CNN from the semi-dense guidance disparity map generated by robust matching feature points at the given pixels. Because of the different number of GCPs obtained from images with different textures, the loss is unstable. Therefore, we propose multiple semi-dense guidance maps to allow the network to learn more from those images with less texture, which leads to a reduction in the influence on the loss by the quantity variance of the GCPs. As shown in Fig. 2, a Laplacian filter is applied to obtain a sharper image and denser guidance map. Our guidance loss is defined as:

$$\mathcal{E}_{\text{guidance}} = \zeta \mathcal{E}_{\text{guidance}} + (1 - \zeta) \mathcal{E}_{\text{guidance}}^* \quad (11)$$

where ζ and $1 - \zeta$ are the weights, which are adaptively changed during the training process. The weight is equal to the ratio of n_{actual} to n_{total} , where n_{total} is the total number of all feature points and n_{actual} is the actual number of the matched feature points.

$$\zeta = \frac{n_{\text{actual}} + n_{\text{actual}}^*}{2n_{\text{total}}} \quad (12)$$

Note that * indicates the result based on the original image or the enhanced image. $\mathcal{E}_{\text{guidance}}$ and $\mathcal{E}_{\text{guidance}}^*$ are the losses based on multiple semi-dense guidance maps with multiacutance images.

$$\mathcal{E}_{\text{guidance}} = \frac{1}{n_{\text{actual}}} \left(\sum_{x \in d_{gl}} \|d_l(x) - d_{gl}(x)\| + \sum_{x \in d_{gr}} \|d_r(x) - d_{gr}(x)\| \right) \quad (13)$$

$$\mathcal{E}_{\text{guidance}}^* = \frac{1}{n_{\text{actual}}^*} \left(\sum_{x \in d_{gl}^*} \|d_l(x) - d_{gl}^*(x)\| + \sum_{x \in d_{gr}^*} \|d_r(x) - d_{gr}^*(x)\| \right) \quad (14)$$

d and d^* are semi-dense guidance maps generated by the original stereo pairs and the pairs after sharpening by the Laplacian filter. d_l , d_r are the predicted left and right disparity maps, d_{gl} , d_{gr} , d_{gl}^* and d_{gr}^* are the multiple semi-dense guidance maps. Only points with available guidance disparities are involved in the computation of this loss.

2) EDGE LOSS BASED ON EDGE GRADIENTS

Although the disparities should be locally smooth with an L1 penalty on the disparity gradients and depth discontinuities often occur at image gradients, texture gradients exist inside the objects in RGB images and should not be taken into the final loss function. Inspired by [6], we use an edge-preserving filter to constrain the weak texture gradients, to make the smoothness loss used in [11] more reliable. We select the bilateral filter, which is a nonlinear, edge-preserving, and noise-reducing smoothing filter for images. The bilateral filter replaces the intensity of each pixel with

a weighted average of intensity values from nearby pixels. Our edge loss is defined as:

$$\begin{aligned} \mathcal{E}_{edge} &= \frac{1}{N} \sum |\Theta_l^{xx}| \text{Exp}(-\|\Psi_l^{xx}\|) \\ &\quad + |\Theta_l^{yy}| \text{Exp}(-\|\Psi_l^{yy}\|) + |\Theta_l^{xy}| \text{Exp}(-\|\Psi_l^{xy}\|) \\ &\quad + \frac{1}{N} \sum |\Theta_r^{xx}| \text{Exp}(-\|\Psi_r^{xx}\|) \\ &\quad + |\Theta_r^{yy}| \text{Exp}(-\|\Psi_r^{yy}\|) + |\Theta_r^{xy}| \text{Exp}(-\|\Psi_r^{xy}\|) \quad (15) \end{aligned}$$

where Ψ_l^{xx} , Ψ_l^{yy} and Ψ_l^{xy} are the gradients of $\mathcal{B} * I_l$ (the left image after bilateral filter \mathcal{B} is applied) along the x , y and xy diagonal directions, and Θ_l^{xx} , Θ_l^{yy} and Θ_l^{xy} are the gradients of the predicted disparity map along the x , y and xy diagonal directions, respectively. The 5×5 kernel of \mathcal{B} is shown in Fig. 2. N is the number of pixels, and r indicates the right pixels. Note that the summation is for all pixels.

3) RECONSTRUCTED APPEARANCE MATCHING LOSS

We evaluate the direct image appearance loss at the sets of image pixels. Linear interpolation is used for subpixel-level warping. The reconstructed appearance matching loss is denoted as follows:

$$\begin{aligned} \mathcal{E}_{reconstruction} &= \frac{1}{N} \sum_{x \in \delta} |I_l(x) - I_r(x - d_l(x))| \\ &\quad + \frac{1}{N} \sum_{x \in \delta} |I_r(x) - I_l(x + d_r(x))| \quad (16) \end{aligned}$$

where δ denotes the set of available x in the whole image, and N is the number of pixels.

All three functions are symmetrically formulated based on both the left and right images, which ensures consistency in the predicted depth maps.

IV. EXPERIMENTAL RESULTS

To evaluate the performance of our method, we quantitatively and qualitatively compare the state-of-the-art weakly supervised and semisupervised monocular depth estimation methods and non-deep learning stereo matching methods. The evaluation focuses on five main components. (1) Comparison with the baseline weakly supervised monocular depth prediction approach [11] and its variant models. (2) Comparison with state-of-the-art monocular depth estimation approaches, including supervised, weakly supervised and semisupervised methods. (3) Comparison with state-of-the-art GCPs approaches. (4) An ablation study conducted with variants of our method. (5) The generalization ability to other datasets. Note that the rectified stereo image pairs are employed only in the training stage without any ground-truth depth. Our weakly supervised method utilizes only the ground truth to evaluate the performance.

A. IMPLEMENTATION DETAILS

Before training our network, we apply our DDS for rectified stereo image pairs to acquire the semi-dense guidance maps.

The input stereo images are resized to 512×256 before being fed into our DDS framework for 10000 feature points. Based on our DDS, we can obtain approximately 6000- 8000 accurate matches for each image. After sharpening for guidance loss, the number of matches is increased by approximately 10%. The model is trained for 60 epochs on an NVIDIA GTX 1070 with 8 GB memory, which allows for a batch size of 5. To train the network on the KITTI raw dataset, we use stochastic gradient descent with a learning rate of 0.01 and a momentum of 0.9. We set the weights of the guidance loss, edge loss, and reconstruction loss as $w_g = 1.0$, $w_e = 0.1$ and $w_r = 1.0$, respectively. Note that in the experiment, our guidance loss based on multiple semi-dense guidance maps results in large gradients if added at the beginning of training, which could cause divergence of the model. Therefore, we set $w_g = 0$ during the first five epochs and recover it to 1.0 gradually during the next five epochs to obtain a convergent optimization.

B. DATASET AND EVALUATION CRITERIA

Evaluations are conducted mainly on stereo image pairs with ground truth on the 2015 KITTI raw data and Cityscapes [4] benchmark. The sequences contain stereo images taken from a car driving in an urban scenario. We evaluate our approach on the KITTI raw dataset, which is split into 28 testing scenes as proposed by Liu *et al.* [23]. The remaining sequences are adopted for training and validation. The detailed depth evaluation metrics used in our comparison are derived from [6]. Note that the first three experiments are conducted using the same evaluation code and depth range as Godard *et al.* [11].

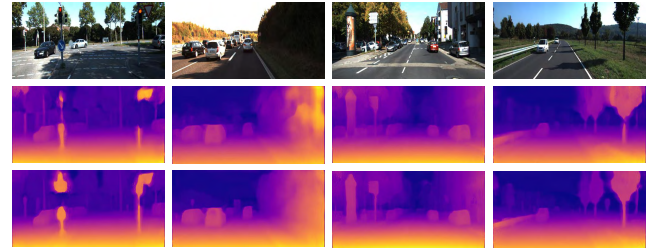


FIGURE 7. Qualitative results and comparison with the state-of-the-art weakly supervised method [11]. The first rows are the left images from KITTI, the second row is the weakly supervised method [11], and the last row corresponds to our method. Note that our method provides more precise details because of our novel semi-dense guidance map based on DDS, as well as the loss function.

C. COMPARISON WITH THE STATE-OF-THE-ART

First, we compare our results with the state-of-the-art weakly supervised depth prediction method [11] on the 200 training images of the KITTI benchmark (the training dataset is the same as that in [6], [11], and [20], namely, the KITTI raw dataset). The quantitative results are given in Table 2. As demonstrated in Table 2, for all metrics and setups, our approach performs better than the baseline work. We outperform the best setup of Godard *et al.* [11] by 40% in terms of RMSE and by 17 pixels with regard to D1-all. We also

TABLE 2. Comparison of ours and the baseline weakly supervised depth prediction models. Results on the KITTI 2015 stereo 200 training set disparity images [10]. For training, K is the KITTI dataset [10] and CS is the Cityscapes dataset [4]. Res denotes that the residual network is adopted instead of the VGG model [38]. Compared to the state-of-the-art weakly supervised depth prediction methods, our model performs the best. For Abs Rel, Sq Rel, RMSE, RMSE (log), and D1-all, lower values are better; for the remaining measures, higher values are better.

Method	Dataset	input	Abs Rel	Sq Rel	RMSE	RMSE(log)	D1-all	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Godard [11]	K	mono	0.124	1.388	6.125	0.217	30.272	0.841	0.936	0.975
Godard [11]	CS	mono	0.699	10.060	14.445	0.542	94.757	0.053	0.326	0.862
Godard [11]	CS+K	mono	0.104	1.070	5.417	0.188	25.523	0.875	0.956	0.983
Godard Res [11]	CS+K	mono	0.097	0.896	5.093	0.176	23.811	0.879	0.962	0.986
Ours	K	mono	0.063	0.842	4.300	0.135	17.113	0.953	0.982	0.991
Ours	CS	mono	0.261	4.010	9.208	0.316	67.638	0.671	0.873	0.948
Ours	CS+K	mono	0.062	0.839	4.231	0.134	16.878	0.954	0.982	0.991
Ours Res	CS+K	mono	0.059	0.738	4.092	0.130	16.776	0.956	0.983	0.991

TABLE 3. Quantitative results of our method and approaches reported in the literature for the test set of the KITTI Raw dataset used by Eigen et al. [6]. For the RMSE, RMSE (log), and ARD, lower values are better, and for the remaining parameters, higher values are better. For supervision, “Depth” means that the ground-truth depth is used in the method; “Mono.” means that monocular sequences are used in the training; and “Stereo” means that stereo sequences with known stereo camera poses are used in the training. “Synthetic” means that synthetic RGB-D datasets are used in the training.

Method (cap : 0-80 m)	Supervision	RMSE	RMSE(log)	Abs Rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Kundu et al. [19] (Unsupervised)	Synthetic	7.157	0.295	0.214	0.665	0.882	0.950
Eigen et al. [6] fine	Depth	6.307	0.282	0.203	0.702	0.890	0.958
Liu et al. [23]	Depth	6.471	0.273	0.201	0.680	0.898	0.967
Kundu et al. [19](Semisupervised)	Depth+Synthetic	5.578	0.237	0.167	0.771	0.922	0.971
Fu et al. [8]	Depth	2.727	0.120	0.072	0.932	0.984	0.994
Zhou et al. [52]	Mono.	6.856	0.283	0.208	0.678	0.885	0.957
Yin et al. [47]	Mono.	5.857	0.233	0.155	0.793	0.931	0.973
Garg et al. [9]	Stereo	5.849	0.246	0.152	0.784	0.921	0.967
Godard et al. [11]	Stereo	5.927	0.247	0.148	0.803	0.922	0.964
Zhan et al. [51] (Full-NYUv2)	Stereo	5.585	0.229	0.135	0.820	0.933	0.971
Kuznetsova et al. [20]	Stereo+Depth	4.621	0.189	0.113	0.862	0.960	0.986
Ours	Stereo	4.675	0.186	0.091	0.905	0.959	0.978

TABLE 4. Comparison with state-of-the-art methods for the KITTI dataset. These methods have been trained and tested with two input images instead of one, and the best results are shown in bold.

Method	Coverage	Abs Rel	Sq Rel	RMSE	RMS log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Garg [9]	100%	0.177	1.169	5.285	0.282	0.727	0.896	0.958
Godard [11]	100%	0.068	0.835	4.392	0.146	0.942	0.978	0.989
SGM [13]	87%	0.064	0.506	3.030	0.150	0.955	0.979	0.989
Ours	100%	0.057	0.612	3.852	0.127	0.958	0.985	0.992

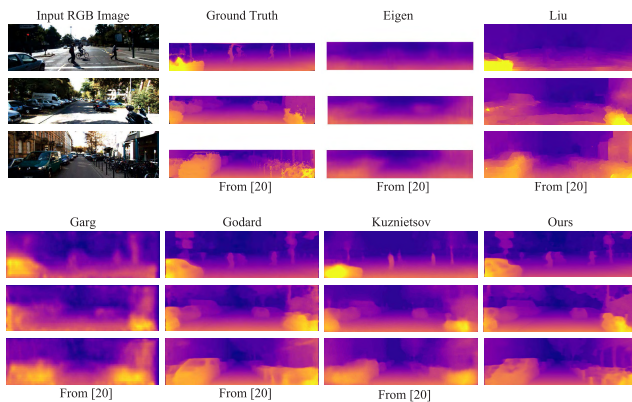


FIGURE 8. Qualitative results and comparison with state-of-the-art methods. The ground truth has been interpolated for visualization. Our results show more details and clearer contours.

qualitatively compare the output of our method with the baseline work in Fig. 7. Next, similar comparison experiments are conducted on more state-of-the-art approaches and our method. The results are given in Table 3 and Fig. 8.

TABLE 5. Analysis of training labels inferred for 8 sequences of KITTI 12. For [41], [27], and our proposal we report the accuracy A for the predicted labels (computed for points with available ground-truth), the average density D on the 8 sequences.

KITTI 12	CENSUS [48]		MC-CNN [50]		SGM [13]	
Method	A	D	A	D	A	D
Tosi et al. [41]	88.9%	33.8%	85.4%	29.4%	81.3%	21.5%
Mostegel et al. [27]	98.5%	8.4%	97.0%	12.4%	88.6%	12.5%
KITTI 12	GMS [1]		DDS		DDSm	
Method	A	D	A	D	A	D
Ours	99.3%	6.3%	99.5%	6.1%	99.4%	6.7%

Our method outperforms stereo supervised methods such as [6], [11], and [51], the supervised approaches such as [6] and [23], and all mono supervised approaches. Without the aid of LiDAR depth data, our model is able to achieve results comparable with those of the semisupervised method in [20]. Qualitative results are shown in Fig. 11.

Following the methods of [9] and [11], we also implemented a stereo version of our model (see Table 4), where the network’s inputs are both left and right views.

TABLE 6. Comparison of different settings. Results on the KITTI 2015 stereo 200 training set disparity images. Baseline uses reconstructed appearance loss, disparity smoothness loss used in [11] without the bilateral filter, and the skip-layer network without deformable convolutional layers. For Abs Rel, Sq Rel, RMSE, RMSE (log), and D1-all, lower is better, and for the remaining, higher is better.

Variants	Abs Rel	Sq Rel	RMSE	RMS log	D1-all	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Baseline	0.098	0.911	5.124	0.180	23.852	0.877	0.958	0.984
Base + GT	0.071	0.794	4.280	0.141	19.332	0.940	0.981	0.990
Base + Mostegel [27]	0.069	0.792	4.278	0.141	16.559	0.941	0.981	0.990
Base + GMS [1]	0.074	0.812	4.452	0.145	21.438	0.931	0.976	0.988
Base + DDS	0.072	0.801	4.367	0.142	19.902	0.939	0.979	0.989
Base + DDSm	0.068	0.790	4.283	0.140	18.832	0.942	0.980	0.989

TABLE 7. Comparison of different settings. The results are shown for the KITTI 2015 stereo 200 training set disparity images. The baseline uses the reconstructed appearance loss, disparity smoothness loss used in [11] without the bilateral filter, and the skip-layer network without deformable convolutional layers. Our full approach is shown in bold. For Abs Rel, Sq Rel, RMSE, RMSE (log), and D1-all, lower values are better, whereas for the remaining parameters, higher values are better.

Variants	Abs Rel	Sq Rel	RMSE	RMS log	D1-all	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Baseline	0.098	0.911	5.124	0.180	23.852	0.877	0.958	0.984
Base + Deformable	0.092	0.827	5.044	0.177	22.787	0.881	0.960	0.985
Base + Bilateral	0.096	0.886	5.109	0.178	22.993	0.880	0.959	0.985
Base + De + Bi	0.089	0.803	4.978	0.175	20.662	0.898	0.961	0.986
Base + DDSm	0.068	0.790	4.283	0.140	18.832	0.942	0.980	0.989
Full Approach	0.059	0.738	4.092	0.130	16.776	0.956	0.983	0.991

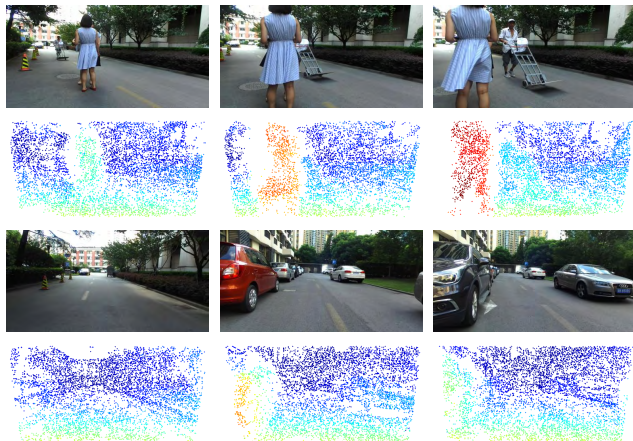


FIGURE 9. The qualitative results of our model predicted on the Cityscapes dataset [4] (top) and Make3D [33] (bottom) without pretraining.

To further prove the advantage of our DDS compared with other state-of-the-art unsupervised GCP methods, such as those in [27] and [41], we perform an additional experiment followed by [41].

As shown in Table 5, the pseudo-ground-truth data generated by [27] performs slightly better than our DDS in term of density. However, the distribution of high-confidence GCPs on the disparity map is heavily dependent on the basic stereo algorithm. The GCPs generated by dense stereo matching methods can only provide ground-truth guidance (as the one provided by laser) because limited monocular feature information is contained in such GCPs. In contrast to the methods of [27] and [41], DDS is a feature-point-based matching method. The GCPs based on DDS provide accurate and robust disparity values as well as distribution information on the



FIGURE 10. Actual experimental results of the disparity guidance maps based on our DDS.

key feature points for every monocular image, which can help the network extract and learn more important and robust features from monocular images. This factor ensures that the DDS is a more beneficial alternative than other methods used for monocular depth estimation. We have also evaluated training with different GCPs methods in Table 6. As shown in Table 6, our superior experiment generates indicators that

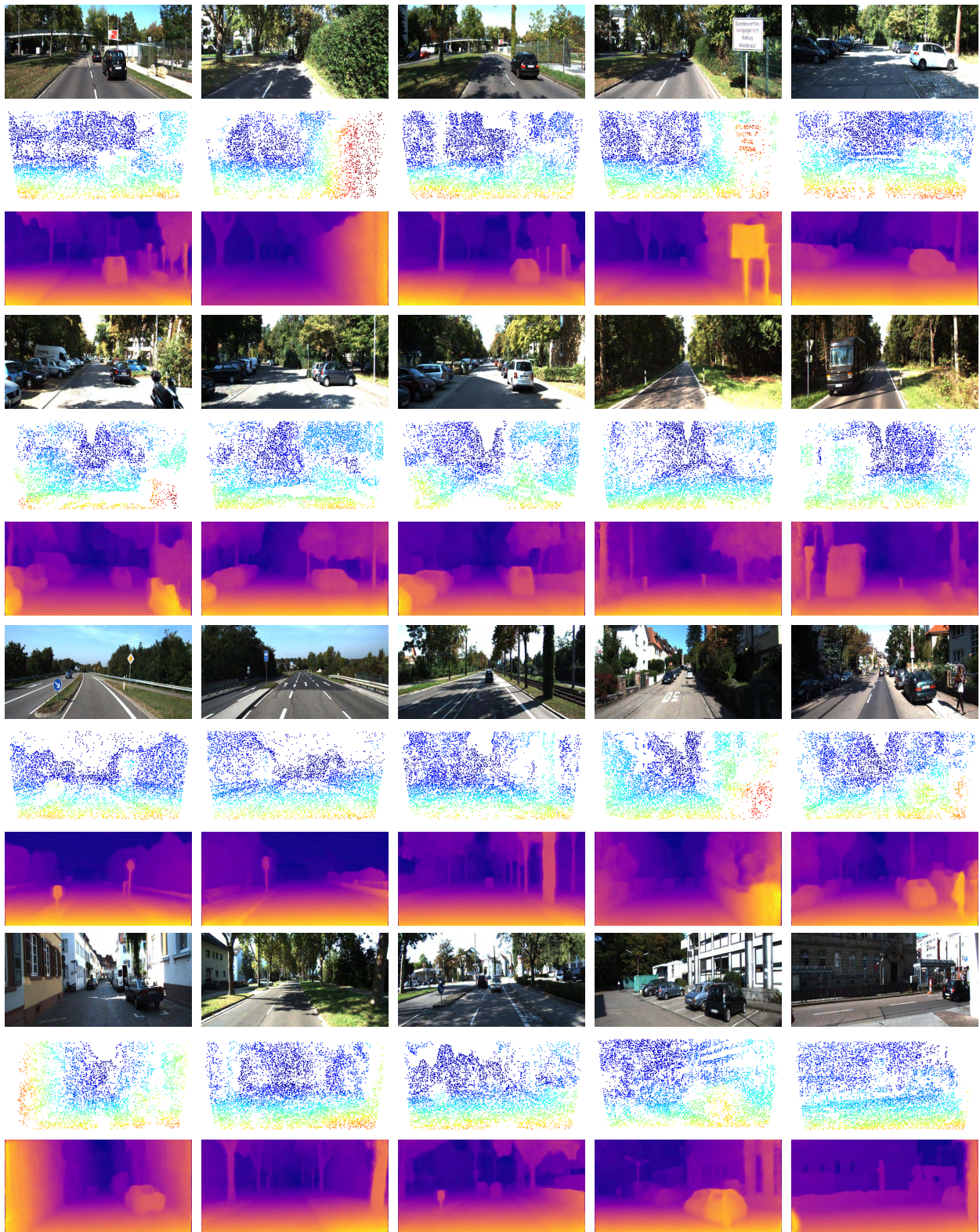


FIGURE 11. Qualitative results of our approach on the test set of the KITTI Raw dataset used by Eigen *et al.* [6] and the guidance maps based on our DDS (only for visualization).

demonstrate the advantage of using the feature-point-based GCPs provided by our method. Qualitative results are shown in Fig. 9.

D. ABLATION STUDY

For the ablation study, we analyze the contributions of the various design choices of our approach. As shown in Table 7,

TABLE 8. Results for the Make3D dataset [33]. Following Zhou et al. [52], we apply the model trained on KITTI to the test set without any of the Make3D data for training. Following the evaluation protocol of [52], we match the scale of our depth predictions to the ground truth and compute the errors only for pixels in a central image crop with a ground-truth depth of less than 70 meters. For supervision, “Depth” means that the ground-truth depth is used in the method; “Mono.” means that monocular sequences are used in the training; and “Stereo” means that stereo sequences with known stereo camera poses are used in the training. “Synthetic” means that synthetic RGB-D datasets are used in the training.

Method (cap : 0-70 m)	Supervision	RMSE	RMSE(log)	Abs Rel	Sq Rel
Kundu et al. [19] (Unsupervised)	Synthetic	11.567	/	0.647	12.341
Train set mean	Depth	12.27	0.307	0.876	13.98
Liu et al. [23]	Depth	10.05	0.165	0.475	6.562
Kundu et al. [19] (Semisupervised)	Depth+Synthetic	9.559	/	0.452	5.710
Karsch et al. [16]	Depth	8.389	0.149	0.428	5.079
Laina et al. [21]	Depth	5.683	0.084	0.204	1.840
Xu et al. (10K) [46]	Depth	4.38	0.065	0.184	/
Kendall et al. [17] (Aleatoric & Epistemic)	Depth	4.08	0.063	0.149	/
Zhou et al. [52]	Mono.	10.47	0.478	0.383	5.321
Godard et al. [11]	Stereo	11.76	0.193	0.544	10.94
Ours	Stereo	8.027	0.188	0.381	4.715

in the first row, only the reconstructed appearance loss, disparity smoothness loss without the bilateral filter, and skip-layer network are used for training the baseline. Next, we add deformable convolution (Base + Deformable), a bilateral filter (Base + Bilateral) and both (Base + De + Bi) together. Then, we use our multi-acutance DDS (DDSm). The combination of all the variants achieves outperforming results (denoted by the last row in Table 7). Our ablation results clearly demonstrate the benefits of employing our novel methods for depth prediction.

E. GENERALIZATION TO OTHER DATASETS

As demonstrated in Table 2, we train our model on Cityscapes [4], test it on the KITTI 2015 stereo 200 training set, and achieve state-of-the-art results. Our model trained on the KITTI raw dataset is also quantitatively evaluated on Make3D, as shown in Table 8. In addition, we train our model on the KITTI raw dataset and test it on the Cityscapes dataset to further validate the generalization ability. The qualitative results on images in the Cityscapes [4] and Make3D [33] datasets are shown in Fig. 10.

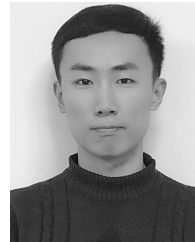
V. CONCLUSION

In this paper, we propose a novel weakly supervised approach that incorporates feature-point-based GCPs for guidance. Supervised and semisupervised learning requires costly depth measuring sensors and strongly relies on the ground truth of the dataset without leveraging the characteristics of the stereo image pairs. Our weakly supervised method can predict more accurate depth maps than those of the state-of-the-art weakly supervised methods while maintaining real-time processing.

REFERENCES

- [1] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, and M.-M. Cheng, “GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2828–2837.
- [2] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli, “On the use of sift features for face authentication,” in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, 2006, p. 35.
- [3] Z. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang, “A deep visual correspondence embedding model for stereo matching costs,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 972–980.
- [4] M. Cordts et al., “The cityscapes dataset for semantic urban scene understanding,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3213–3223.
- [5] J. Dai et al., “Deformable convolutional networks,” *CoRR*, vol. 1, no. 2, p. 3, 2017.
- [6] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2366–2374.
- [7] M. A. Fischler and R. C. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [8] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep ordinal regression network for monocular depth estimation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2002–2011.
- [9] R. Garg, V. K. BG, G. Carneiro, and I. Reid, “Unsupervised CNN for single view depth estimation: Geometry to the rescue,” in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2016, pp. 740–756.
- [10] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3354–3361.
- [11] C. Godard, O. M. Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in *Proc. CVPR*, vol. 2, 2017, p. 7.
- [12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
- [13] H. Hirschmüller, “Accurate and efficient stereo processing by semi-global matching and mutual information,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2005, pp. 807–814.
- [14] X. Hu and P. Mordohai, “A quantitative evaluation of confidence measures for stereo vision,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2121–2133, Nov. 2012.
- [15] X. Huang, C. Yuan, and J. Zhang, “Graph cuts stereo matching based on patch-match and ground control points constraint,” in *Proc. Pacific Rim Conf. Multimedia. Cham, Switzerland: Springer*, 2015, pp. 14–23.
- [16] K. Karsch, C. Liu, and S. B. Kang, “Depth transfer: Depth extraction from video using non-parametric sampling,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2144–2158, Nov. 2014.
- [17] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5574–5584.
- [18] A. Kendall et al., “End-to-end learning of geometry and context for deep stereo regression,” *CoRR*, pp. 66–75, Oct. 2017.
- [19] J. N. Kundu, P. K. Uppala, A. Pahuja, and R. V. Babu. (2018). “AdaDepth: Unsupervised content congruent adaptation for depth estimation.” [Online]. Available: <https://arxiv.org/abs/1803.01599>

- [20] Y. Kuznetsov, J. Stückler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Feb. 2017, pp. 6647–6655.
- [21] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 3D Vis. 4th Int. Conf. (3DV)*, 2016, pp. 239–248.
- [22] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 1253–1260.
- [23] F. Liu, C. Shen, G. Lin, and I. D. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.
- [24] D. Marr and T. Poggio, "Cooperative computation of stereo disparity," *Science*, vol. 194, no. 4262, pp. 283–287, 1976.
- [25] N. Mayer et al., "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4040–4048.
- [26] F. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3061–3070.
- [27] C. Mostegel, M. Rumpler, F. Fraundorfer, and H. Bischof, "Using self-contradiction to learn confidence measures in stereo vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4067–4076.
- [28] P. Moulon, P. Monasse, and R. Marlet, "Global fusion of relative motions for robust, accurate and scalable structure from motion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3248–3255.
- [29] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.
- [30] M. Poggi and S. Mattoccia, "Learning to predict stereo reliability enforcing local consistency of confidence maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jul. 2017, pp. 4541–4550.
- [31] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1161–1168.
- [32] A. Saxena, M. Sun, and A. Y. Ng, "Learning 3-D scene structure from a single still image," in *Proc. IEEE 11th Int. Conf. Comput. Vis. (ICCV)*, Oct. 2007, pp. 1–8.
- [33] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2009.
- [34] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, 2002.
- [35] A. Seki and M. Pollefeys, "Patch based confidence prediction for dense disparity map," in *Proc. BMVC*, vol. 2, 2016, p. 4.
- [36] C. Shi, G. Wang, X. Yin, X. Pei, B. He, and X. Lin, "High-accuracy stereo matching based on adaptive ground control points," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1412–1423, Apr. 2015.
- [37] J. Shotton et al., "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1297–1304.
- [38] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: <https://arxiv.org/abs/1409.1556>
- [39] A. Spyropoulos, N. Komodakis, and P. Mordohai, "Learning to detect ground control points for improving the accuracy of stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1621–1628.
- [40] A. Tonioni, M. Poggi, S. Mattoccia, and L. Di Stefano, "Unsupervised adaptation for deep stereo," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Oct. 2017, pp. 1614–1622.
- [41] F. Tosi, M. Poggi, A. Tonioni, L. D. Stefano, and S. Mattoccia, "Learning confidence measures in the wild," in *Proc. 28th Brit. Mach. Vis. Conf. (BMVC)*, London, U.K., 2017, pp. 4–7.
- [42] L. Wang and R. Yang, "Global stereo matching leveraged by sparse ground control points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3033–3040.
- [43] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Towards unified depth and semantic prediction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2800–2809.
- [44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [45] J. Xie, R. Girshick, and A. Farhadi, "Deep3d: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 842–857.
- [46] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation," in *Proc. CVPR*, vol. 1, 2017, pp. 161–169.
- [47] Z. Yin and J. Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2018, pp. 1983–1992.
- [48] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 1994, pp. 151–158.
- [49] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4353–4361.
- [50] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, nos. 1–32, p. 2, 2016.
- [51] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Mar. 2018, pp. 340–349.
- [52] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. CVPR*, vol. 2, 2017, p. 7.



LIANG DU received the B.S. degree from Harbin Engineering University, China, in 2016. He is currently pursuing the master's degree in information and communication engineering with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China. His current research interests include monocular depth estimation and object detection.



JIAMAO LI received the Ph.D. degree from the Tokyo Institute of Technology, Japan, in 2012. He is currently an Associate Professor with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, China. His current research interests include computer vision, machine vision, 3D micro-imaging, and artificial intelligence.



XIAOQING YE received the B.S. degree from Wuhan University, China, in 2014. She is currently pursuing the Ph.D. degree in information and communication engineering with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China. Her current research interests include stereo vision, 3D reconstruction, and autonomous driving.



XIAOLIN ZHANG received the Ph.D. degree from Yokohama National University, in 1995. He was a Professor with the Tokyo Institute of Technology, Japan, from 2012 to 2013. He is currently a Professor with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, China. His research interests include bionics, brain science, computer vision, and artificial intelligence.

...