SDBF-Net: Semantic and Disparity Bidirectional Fusion Network for 3D Semantic Detection on Incidental Satellite Images

Zhibo Rao*, Mingyi He*‡, Zhidong Zhu*, Yuchao Dai*, Renjie He*†
* Northwestern Polytechnical University, Xian 710129, China
† Nanyang Technological University, 639798, Singapore
‡ Email address: myhe@nwpu.edu.cn (Mingyi He)

***Abstract*—In this paper, we propose a conceptually simple, flexible, and general framework for the semantic stereo task on incidental satellite images. Our method efficiently detects the objects in an incidental satellite image for generating a high-quality segmentation map, and more accurately match the left-right incidental satellite images for obtaining a more accurate disparity map at the same time. The method, called semantic and disparity bidirectional fusion network (SDBF-Net), consists of three main modules: the Semantic Segmentation Module (SSM), the Stereo Matching Module (SMM), and the Fusion Module (FM). The semantic segmentation module takes advantage of the capacity of global context information by extending the receptive field to produce the initial segmentation map. The stereo matching module applies the 3D convolutional operation to regularize the feature map of left-right images to generate the initial disparity map. The fusion module fuses the initial segmentation and disparity map to obtain the refined segmentation and disparity map. Extensive quantitative and qualitative evaluations on the US3D dataset demonstrate the superiority of our proposed SDBF-Net approach, which outperforms state-of-the-art semantic stereo approaches significantly.**

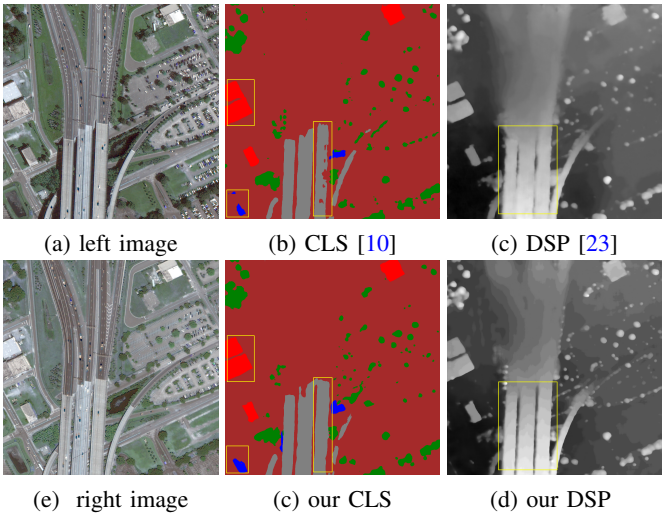
I. Introduction

Semantic segmentation and pairwise stereo is the most promising and latest research directions in the computer vision field, which has a significant impact on the other applications such as autonomous driving for vehicles [1]–[3], object detection and recognition in remote sensing images [4], [5], and 3D model reconstruction and understanding [6]–[8]. However, many researchers viewed this question as two tasks in isolation, namely semantic segmentation, and stereo matching. In this work, we consider the semantic segmentation and the pairwise stereo match as one integrated problem, with focus on the fusion of the two tasks and aim to train the fusion network for state-of-the-art performance.

Convolutional neural networks (CNNs) have exhibited impressive power in computer vision, especially semantic segmentation and stereo matching. Many researchers have designed straightforward and advanced networks for the two challenging tasks [9]–[14].

For semantic segmentation, J. Long *et al.* applied the fully convolutional networks (FCN) without any fully connected layers to solve the problem caused by different resolution, and achieved much higher accuracy which largely better than other competitors [9]. The FCN followers used dilated convolutions [10], [15], spatial pyramid pooling (SPP) [16], [17], or feature pyramid structure [18]–[20] to extend the receptive field of networks for improving the performance. Follow those works, R. Girshick *et al.* proposed the region-based CNN (R-CNN) which adopted a manageable number of candidate object regions to search region of interest (RoI) and evaluated the convolutional networks on each RoI [21]. The followers of R-CNN exploit the attention [22] or mask mechanism [11] to boost the speed or increase the accuracy of the instance. On the other hand, C. Hazirbas *et al.* utilized the depth map to combine with RGB image for promoting the scene understanding [14].

For stereo matching, A. Kendall *et al.* introduced cost volume to list all possible disparity, then used a 3D CNN to regularize the cost volume; it obtained the top performance in the benchmarks [12]. The followers of GC-Net took advantages of semantic [24] or edge information [25], multi-



(a) left image (b) CLS [10] (c) DSP [23]
(e) right image (c) our CLS (d) our DSP

Fig. 1: **Results on the US3D dataset.** Different from the traditional matching task, seasonal appearance differences pose challenges for stereo matching on the incidental satellite images. We highlight the advantage of our bidirectional fusion strategy. Note that the completeness of segmentation map is better than the non-fusion segmentation method [10] for the semantic segmentation task, and the clarity of object structure outperforms the non-fusion stereo method [23] for the stereo matching task.

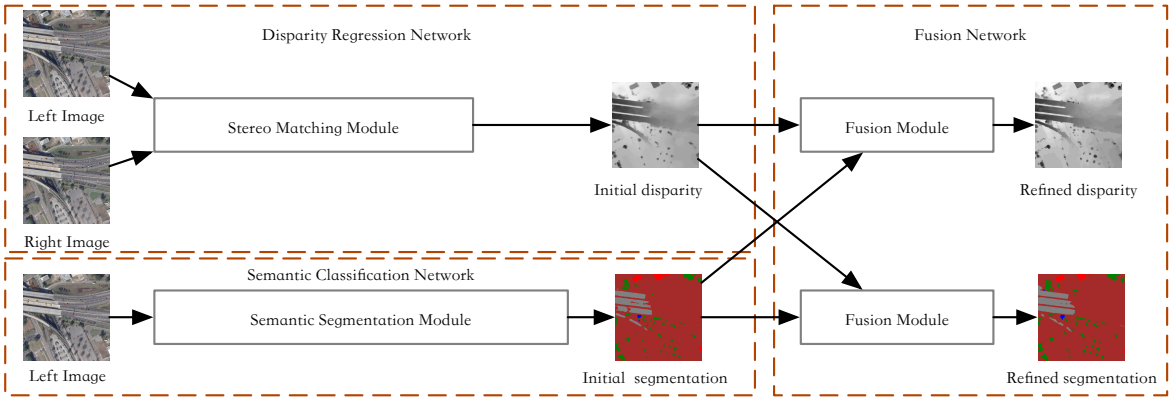


Fig. 2: **Our Semantic and Disparity Bidirectional Fusion Network, SDBF-Net.** The network consists of three modules: SSM (semantic segmentation module), SMM (stereo matching module), and FM (fusion module).

scales structure [26], and refinement process [27] to improve the performance in the occlusion area. On this base, the traditional stereo methods such as the warping or structural similarity (SSIM) function were used as loss function to drive the training process for achieving self-supervision [28], [29].

Albeit the above success in the field of semantic segmentation and stereo matching, the methods based on deep learning still exist some limitations. First, these methods more concentrate on the two tasks alone rather than fusion information; the two results could promote each other via fusing the different dimension information. Second, the seasonal appearance and differences pose are great challenges for the stereo matching task on the incidental satellite images; the semantic information could promote the scene understanding for the matching task. The precise disparity and fine segmentation result should be produced in one network, and they could promote the performance of each other.

In this work, we tackle the above challenges and propose a more elegant network architecture, called Semantic and Disparity Bidirectional Fusion Network (SDBF-Net).

II. Our Method

In this section, we present the semantic and disparity bidirectional fusion network (SDBF-Net). The network architecture is illustrated in Fig. 2. Our SDBF-Net consists of three modules: semantic segmentation module, stereo matching module, and fusion module (noting that the fusion module contains two fusion sub-modules). First, the semantic segmentation module is applied to extract semantic information from the left incidental satellite image for getting the initial segmentation map as shown in Sec. II-A. Next, the stereo matching module is adopted to match the left-right incidental satellite image to obtain the initial disparity map as shown in Sec. II-B. Then, the fusion module fuses the initial disparity and segmentation map to further improve the accuracy as shown in Sec. II-C. Finally, we introduce the loss function as shown in Sec. II-D. The implementation detail is described in the following sub-sections respectively.

A. Semantic Segmentation Module

In this module, we apply a series of 2D convolutional operations to produce the initial segmentation map, and each convolutional operations is followed by a BN layer and a ReLU layer except for the last layer. The semantic segmentation module (SSM) consists of two parts: ResNet-101 part and multi-scale feature learning part, as shown in Fig. 3.

(1) ResNet-101 Part

The first step of SSM is to extract the local context from the left incidental satellite image I. The deep feature maps are more robust to photometric differences (e.g., lighting effects and perspective effects). We adopt the standard ResNet-101 as the backbone. Because ResNet architecture is not the critical point in this paper, we suggest interested readers read ResNet paper [30] to get detail information. As shown in [30], we could get the deep feature maps with the size $(H/8) \cdot (W/8) \cdot \cdot 1024$.

(2) Multi-Scale Feature Learning Part

The next step of SSM is to extract the multi-scale features from the deep feature maps. To further extract hierarchical contextual information using an atrous spatial pyramid pooling (ASPP) block with 3×3 filters and dilated rate of 3, 6, 12 and 18 respectively. The exploitation of dilated convolution can enlarge the receptive field with less computational memory cost and less spatial resolution decrease. We concatenate the deep feature maps and hierarchical contextual information, and then fuse them via passing one 1024-channel and one 512-channel convolutional layers. Then, we utilize one 256-channel, one 128-channel, and one C-channel 2D deconvolutional layer (both the stride is 2) to recover the size of feature maps for producing the high-quality score map, where C denotes the number of categories. After this part, we could obtain the initial segmentation map with the size $H \cdot W \cdot C$.

B. Stereo Matching Module

In this module, we apply a series of 2D or 3D convolutional operations to obtain the initial disparity map, and each convolutional operations is followed by a BN layer and a ReLU layer except for the last layer. The stereo matching module

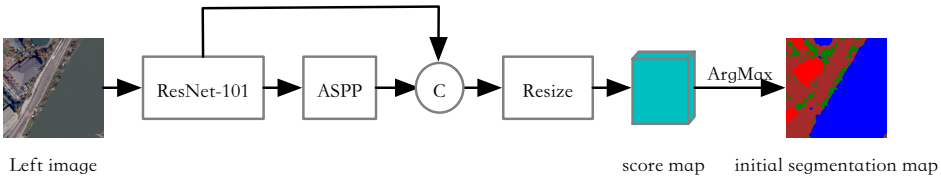


Fig. 3: **The semantic segmentation module.** The module contains two main parts: ResNet-101 and multi-scale feature learning.

(SMM) contains four parts: deep feature extraction part, cost volume construction part, feature matching part, and disparity regression part, as shown in Fig. 4.

(1) Deep Feature Extraction Part

The first step of SMM is to extract the deep unary feature maps $\{\mathbf{F}_i\}_{i=1}^{N=2}$ of the left-right images $\{\mathbf{I}_i\}_{i=1}^{N=2}$ for building the cost volume. We first pass the left-right incidental satellite images through four 32-channel 2D convolutional layers (the stride is 1, except for the first layer which is 2), three 32-channel residual blocks, one 32-channel 2D convolutional layers (the stride is 2), and fifteen 64-channel residual blocks (both 3×3 filters) to encode them. Moreover, we extract hierarchical contextual information from these feature maps using the spatial pyramid pooling (SPP) block. Finally, we concatenate hierarchical contextual information and previous feature maps, and fuse them via one 128-channel 2D convolutional layers and 32-channel 2D convolutional layers (the last layer does not contain the BN layer and ReLU layer). We pass the left-right incidental satellite images through this part with the same weights. Therefore, we could obtain the left-right unary feature maps with the size $(H/4) \cdot (W/4) \cdot 32$.

(2) Cost Volume Construction Part

The second step of SMM is to build a 3D cost volume by listing all possible positions. Follow our previous work [31], we adopt the unary feature maps of the left-right incidental satellite to form the cost volume by concatenating the unary feature maps with the traversed right unary feature maps. Denote f_L, f_R of the $\{\mathbf{F}_i\}_{i=1}^{N=2}$ extracted from $\{\mathbf{I}_i\}_{i=1}^{N=2}$ by the deep feature extraction part, the left-to-right feature volume at the pixel position (u, v) with the potential disparity d in the disparity range D could be expressed as:

$$V(d,u,v,f) = \text{stack}\{f_L(u,v) \,||\, f_R(u-d,v)\} \tag{1}$$

where $||$ denotes the concatenation operation, f denotes the feature dimension of the unary feature maps, V denotes the cost volume, and $\text{stack}\{\cdot\}$ denotes the stack operation. In this way, we construct the cost volume V with the size $(D/4) \cdot (H/4) \cdot (W/4) \cdot 64$.

(3) Feature Matching Part

The third step of SMM is to regularize the cost volume by a series of 3D convolutional operations. In this step, we use the multi-scale 3D CNN to aggregate the feature information alone the disparity dimension as well as spatial dimensions. The multi-scale 3D CNN is very similar to a 3D version U-Net, which consists of four level 3D CNN for aggregating the neighboring information. We adopt the down-sampling process (32-channel, 64-channel, 96-channel, 128-channel 3D convolutional layers with the stride 2) to encode the cost volume V and the up-sampling process (96-channel, 64-channel, 32-channel, 32-channel 3D deconvolutional layers with the stride 2) to decode the encoded feature maps. Moreover, we link the same level to build the residual structure promising the critical information is not lost. Then, we utilize one 16-channel 3D deconvolutional layer and one 1-channel 3D deconvolutional layer (the stride both is 2) to recover the size of cost volume V. After the feature matching part, we could get the regularized cost volume with the size $D \cdot H \cdot W$.

(4) Disparity Regression Part

The final step of SMM is to regress the initial disparity $\hat{d}_{\text{init}}$ by the ArgMin operation. First, we convert the cost volume V to the probability volume P via softmax operation. Then, we calculate the sum of each disparity d weighted with the probability volume P. The regression process could be expressed as:

$$\hat{d}_{\text{init}} = \sum_{d=d_{\min}}^{d_{\max}} d \times P(d) = \sum_{d=d_{\min}}^{d_{\max}} d \times \text{softmax}(d) \tag{2}$$

where $P(d)$ is the probability estimation for all pixels at disparity d. As shown in [12], the above disparity regression is more robust the classification-based methods, and the regression-based method could produce the sub-pixel estimation. After the disparity regression part, we could obtain the initial disparity map $\hat{d}_{\text{init}}$ with the size $H \cdot W$.

C. Fusion Module

In this module, we will apply a series of 2D convolutional operations to fuse the segmentation and disparity maps for further improving the accuracy, and each convolutional operations is followed by a BN layer and a ReLU layer except for the last layer. The fusion module (FM) consists of semantic fusion part and disparity fusion part, as shown in Fig. 5.

(1) Semantic Fusion Part

In the semantic fusion part, we apply a residual learning network at the end of SDBF-Net. The initial segmentation map, the left input image, and the initial disparity map are concatenated as a 10-channel input, which is then passed through one 32-channel 2D convolutional layers, three 32-channel residual blocks and one 6-channel convolutional layer (both 3×3 filters) to learn the segmentation residual. The initial segmentation map is added back to obtain the refined segmentation map. Note that, the last layer does not contain the BN layer and ReLU layer for learning the negative residual.

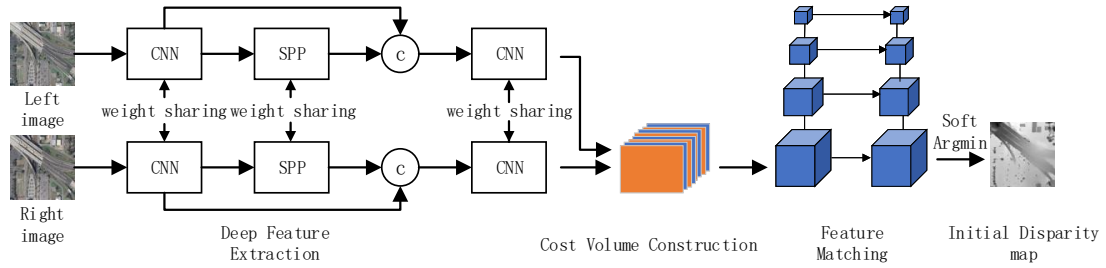


Fig. 4: **The stereo matching module.** The module contains four main parts: deep feature extraction, cost volume construction, feature matching, and disparity regression.

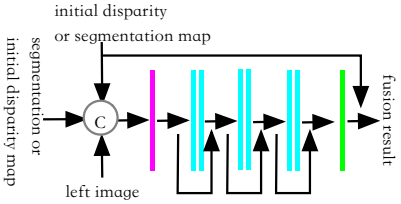


Fig. 5: **The fusion module.** The module contains two parts: semantic fusion and disparity fusion. Note that semantic fusion and disparity fusion are only different in the last layer.

(2) Disparity Fusion Part

In the disparity fusion part, the network structure is very similar to the semantic fusion part except for the last layer. The last layer in this part is one 1-channel convolutional layer.

After the fusion module, we could obtain the refined segmentation and the refined disparity map with the same size as the input incidental satellite images.

D. Loss Function

In the above sections, the stereo matching is a regression problem, but the semantic segmentation is a classification problem. We design the different loss function for the two problems.

(1) Loss Function in Semantic Segmentation

For semantic segmentation is a classical classification problem, we train the semantic segmentation module and its fusion module with cross-entropy loss. First, we convert the output to the probability volume $\mathcal{P}$ via the softmax operation. Then, we use the cross-entropy function to calculate the loss. The process could be represented as:

$$\begin{aligned} Loss_i &= \sum_{\mathcal{P}_i} \left(\sum_{i=1}^{C} -\mathcal{P}_i\left(i,p\right) \cdot \log Q\left(i,p\right) \right) \\ Loss_r &= \sum_{\mathcal{P}_r} \left(\sum_{i=1}^{C} -\mathcal{P}_r\left(i,p\right) \cdot \log Q\left(i,p\right) \right) \end{aligned} \tag{3}$$

where p denotes the spatial image coordinate, $\mathcal{P}_i$ or $\mathcal{P}_r$ denotes the probability volume before or after fusion module, $\mathcal{P}_i\left(i,p\right)$ or $\mathcal{P}_r\left(i,p\right)$ denotes a voxel in the probability volume $\mathcal{P}_i$ or $\mathcal{P}_r$, Q denotes the ground truth binary occupancy volume, which is generated by the one-hot encoding of the ground truth with the number of categories C, and $Q\left(i,p\right)$ denotes corresponding voxel to $\mathcal{P}_i\left(i,p\right)$ or $\mathcal{P}_r\left(i,p\right)$.

(2) Loss Function in Stereo Matching

For stereo matching is a regression problem in this paper, we train the stereo matching module and its fusion module with L_1 loss. Because the labels of the many datasets is sparse, we average our loss over the labeled pixels. The process could be represented as:

$$\begin{aligned} Loss_i &= \frac{1}{N} \sum_{n=1}^{N} ||d\left(p\right) - \hat{d}_i\left(p\right)||_1 \\ Loss_r &= \frac{1}{N} \sum_{n=1}^{N} ||d\left(p\right) - \hat{d}_r\left(p\right)||_1 \end{aligned} \tag{4}$$

where N denotes the number of the valid labels, p denotes the spatial image coordinate, and $d\left(p\right)$ denotes the value of ground truth at the pixel p.

III. Experiments

In this section, we will evaluate the performance of our method on an emerging semantic stereo dataset: the urban semantic 3D (US3D) [32]. First, we present our implementation details about the training process and testing process as shown in Sec. III-A. Then, we compare the contribution of the different component in SDBF-Net as shown in Sec. III-B. Finally, we quantize the performance of our method and compare with the state-of-art methods on the data fusion contest pairwise semantic stereo challenge benchmark as shown in Sec. III-C.

A. Implementation Details

(1) Dataset

The US3D dataset [32] includes 100 square kilometer coverage for the United States cities of Jacksonville, Florida and Omaha, Nebraska. The training set contains 4292 pairs of epipolar rectified images and the corresponding labels of the semantic segmentation and disparities with the size 1024×1024. The testing set includes 50 testing pairs without ground truth of semantic labels and disparity maps.

For semantic segmentation task, the classification labels are based on the LSA specification, and the pixels are classified each pixel into the following six categories: ground, high vegetation/trees, building roof, elevated road/bridge, water and unlabeled.

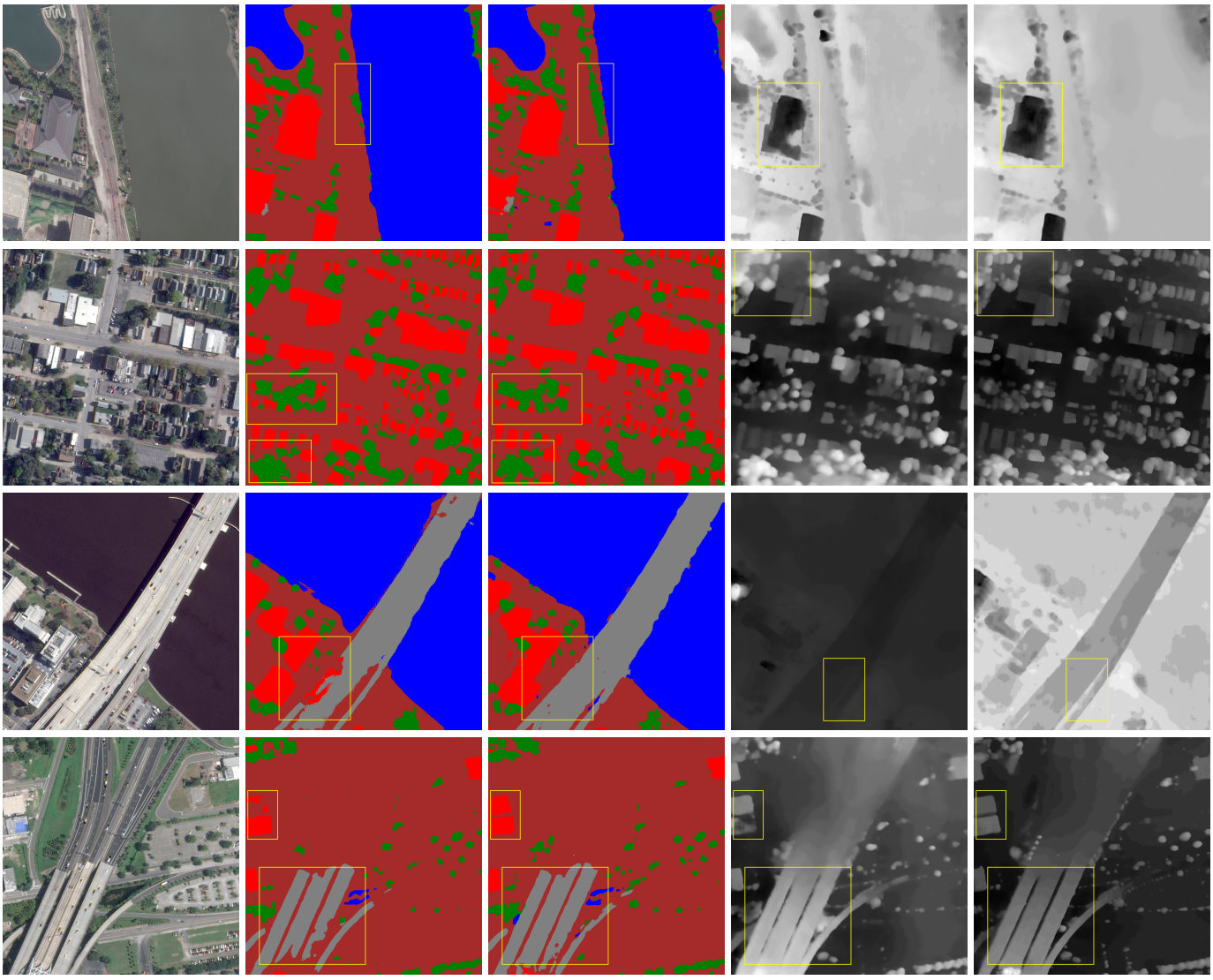


Fig. 6: **The Results of our SDBF-Net on the US3D dataset.** We highlight the result of before and after fusion. The completeness of the fusion segmentation map is obviously better than the initial segmentation map. The clarity of the fusion disparity map exceeded outperforms the initial disparity map.

For disparity prediction task, the disparity label is a 32-bit floating point image, where each pixel value represents disparity in pixel.

(2) Training and Testing

At training time, we implement SDBF-Net in Tensorflow, and train the model with the ADAM optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) on the US3D dataset [32]. For all iterations, we set a batch size of 2, the learning rate 1×10^{-3}, the disparity $D = 128$, and the number of category $C = 6$. Moreover, we normalize input image with pixel intensities level ranging from 0 to 1, and randomly crop them into 448×448. The training procedure is performed on four NVIDIA 1080Ti GPUs, and which contains three stages: semantic segmentation module training stage, stereo matching module training stage, and fusion module training stage for 1000, 200, and 200 epochs respectively.

At testing time, we link the three modules and push the full size left-right incidental satellite images into the network. Then, we apply the fusion results as the final results and upload the final results to the CodaLab benchmark website for evaluating our model.

(3) Evaluation Metric

The evaluation metrics are following previous works [32]. For semantic segmentation task, the mean intersection over union (mIoU) is applied for assessing the performance. For stereo matching task, the average end-point error (EPE) and the fraction of erroneous pixels (D_1) are adopted for evaluating the consistency. For a better evaluation, the US3D dataset provides a new evaluation criterion called mIoU-3, the concrete

calculating methods are as follows:

$$\mathrm{mIoU}_t = \frac{1}{C}\sum_c \frac{TP_t}{TP + FP + FN} \quad (5)$$

where t denotes threshold of correct disparity, TP denotes matched pairs of segments, FP denotes unmatched predicted segments, FN denotes unmatched ground truth segments, and C denotes the number of categories. mIoU-3 means FP must have both the correct semantic label and disparity error less than a given threshold 3 pixels.

B. Ablations

To verify the effectiveness of our design, we conducted experiments with different settings to evaluate SDBF-Net, including the use of fusion module, the backbone network of semantic segmentation module, etc. The results are shown in Tab. I and Tab. II.

TABLE I: **Evaluation of the semantic segmentation module with different settings.** Computed mIoU on the US3D test set.

Model	Setting		mIoU	Time (s)
	backbone	ASPP		
SSM	ResNet-50	-	0.646	**0.151**
SSM	ResNet-50	√	0.723	0.156
SSM	ResNet-101	-	0.739	0.234
SSM	ResNet-101	√	0.759	0.239
SSM + Fusion	ResNet-101	√	**0.767**	0.245

TABLE II: **Evaluation of the stereo matching module with different settings.** Computed the end-point-error and percentage of three-pixel-error on the US3D test set.

Model	Setting		D_1 (%)	EPE	Time (s)
	SPP	3D CNN			
SMM	-	-	40.32	8.77	**0.143**
SMM	√	-	40.40	8.63	0.147
SMM	-	√	10.77	1.55	0.696
SMM	√	√	10.55	1.50	0.701
SMM + Fusion	√	√	**8.02**	**1.31**	0.713

As shown in Tab. I and Tab. II, it qualitatively demonstrates the advanced nature of using the SDBF-Net we proposed. First, the fusion module has a significant performance improvement for semantic segmentation task and stereo matching task in mIoU and EPE respectively. Second, SPP, ASPP, and 3D CNN enhance the scene understanding ability of the model effectively. The results prove the rationality of our designed model.

C. US3D

To evaluate the performance of our model, we use our best model trained in the ablation experiments to calculate the segmentation and disparity map and submit the results to the CodaLab evaluation server. Because the US3D is an emerging dataset, many researchers submit their results without the related papers. Thus, we use the baseline result provided in [32] as a reference to assess our model as shown in Tab. III and present our high-quality results in Fig. 6.

TABLE III: **Performance comparison with other state-of-the-art methods.** We computed the end-point-error, mIoU, and percentage of three-pixel-error on the US3D test set.

Model	mIoU	D_1 (%)	EPE	Time (s)
CU-PSM [33]	0.772	-	-	-
ICNet [34] + SGM [35]	0.700	43.00	10.34	-
DeepLab v3 [10] + SGM [35]	0.750	43.00	10.34	-
ICNet [34] + iResNet-i2 [23]	0.700	33.00	3.05	-
DeepLab v3 [10] + iResNet-i2 [23]	0.750	33.00	3.05	-
ICNet [34] + DenseMapNet [36]	0.700	35.00	3.51	-
DeepLab v3 [10] + DenseMapNet [36]	0.750	35.00	3.51	-
MRFCNet [37]	**0.790**	9.06	1.39	-
SDBF-Net	0.767	**8.02**	**1.31**	**1.08**

As shown in Tab. III and Fig. 6, our method achieve superiority performance compared to the baseline, which proves the effectiveness of our method. Moreover, The D_1 and EPE of our method rank first in the CodaLab evaluation server [1] (id: raoxi36, the screenshot of our comprehensive score mIoU-3 as shown in Fig. 7) before August 27, 2019. The estimated results clearly demonstrate that SDBF-Net significantly improves for the semantic stereo task. Our approach takes full advantage of fusion technology to largely promote the result of $D1$ compared with other state-of-the-art methods.

#	User	Entries	Date of Last Entry	Prediction ▲
1	raoxi36	11	08/27/19	0.7561 (1)
2	JiKun63	8	07/29/19	0.7373 (2)
3	sghuffar23	1	08/07/19	0.7365 (3)

Fig. 7: **The screenshot of our comprehensive score in the codalab server.** The comprehensive score mIoU-3 contains two parts: mIoU and D_1.

IV. Conclusions

In this paper, we have presented a novel network framework SDBF-Net for the semantic stereo task on incidental satellite images. The method utilizes the semantic segmentation and stereo matching modules to predict the dense-pixel initial segmentation and disparity map. Moreover, the fusion module fuses the initial segmentation and disparity map for generating the refined segmentation and disparity map. Extensive experiments demonstrate the simple fusion strategy could obtain a great improvement for disparity prediction task and a small promotion for semantic segmentation.

V. Acknowledgment

This work was supported in part by Natural Science Foundation of China (61420106007, 61671387 and 61871325).

[1] https://competitions.codalab.org/competitions/20212#results

REFERENCES

[1] C. Chen, S. Ari, K. Alain, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2722–2730. 1

[2] G. Andreas, L. Philip, and U. Raquel, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361. 1

[3] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3061–3070. 1

[4] J. Zhang, Y. Dai, F. Porikli, and M. He, "Multi-scale salient object detection with pyramid spatial pooling," in *Asia-pacific Signal and Information Processing Association Summit and Conference*, 2018. 1

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241. 1

[6] B. Li, Y. Dai, and M. He, "Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference," in *Pattern Recognition*, 2018. 1

[7] B. Li, C. Shen, Y. Dai, A. V. D. Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1119–1127. 1

[8] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *The European Conference on Computer Vision (ECCV)*, 2018. 1

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2015. 1

[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *The European Conference on Computer Vision (ECCV)*, 2018, pp. 833–851. 1, 6

[11] K. He, G. Gkioxari, P. Dollr, and R. Girshick, "Mask r-cnn," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018. 1

[12] A. Kendall, H. Martirosyan, S. Dasgupta, and P. Henry, "End-to-End Learning of Geometry and Context for Deep Stereo Regression," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 66–75. 1, 3

[13] A. Valada, G. L. Oliveira, T. Brox, and W. Burgard, "Robust semantic segmentation using deep fusion," in *Robotics: Science and Systems (RSS 2016) Workshop, Are the Sceptics Right? Limits and Potentials of Deep Learning in Robotics*, 2016. 1

[14] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian Conference on Computer Vision (ACCV)*, 2016, pp. 213–228. 1

[15] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," in *arXiv preprint*, 2017. 1

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–16, 2014. 1

[17] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6230–6239. 1

[18] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic Feature Pyramid Networks," in *arXiv preprint*, 2019. 1

[19] Y. Li, X. Chen, Z. Zhu, L. Xie, G. Huang, D. Du, and X. Wang, "Attention-guided Unified Network for Panoptic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[20] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944. 1

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *International Conference on Neural Information Processing Systems (NIPS)*, 2015, pp. 91–99. 1

[23] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, and J. Zhang, "Learning for disparity estimation through feature constancy," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2811–2820. 1, 6

[24] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "Segstereo: Exploiting semantic information for disparity estimation," in *The European Conference on Computer Vision (ECCV)*, 2018, pp. 636–651. 1

[25] X. Song, X. Zhao, H. Hu, and L. Fang, "EdgeStereo: A Context Integrated Residual Pyramid Network for Stereo Matching," in *Asian Conference on Computer Vision (ACCV)*, 2018. 1

[26] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5410–5418. 1

[27] X. Cheng, P. Wang, and R. Yang, "Learning depth with convolutional spatial propagation network," in *arXiv preprint*, 2018. 1

[28] Y. Zhong, Y. Dai, and H. Li, "Self-Supervised Learning for Stereo Matching with Self-Improving Ability," in *arXiv preprint*, 2017. 2

[29] C. Zhou, H. Zhang, X. Shen, and J. Jia, "Unsupervised Learning of Stereo Matching," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1576–1584. 2

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. 2

[31] Z. Zhu, M. He, Y. Dai, Z. Rao, and B. Li, "Multi-scale Cross-form Pyramid Network for Stereo Matching," in *IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2019. 3

[32] M. Bosch, K. Foster, G. Christie, S. Wang, G. D. Hager, and M. Brown, "Semantic stereo for incidental satellite images," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 1524–1532. 4, 5, 6

[33] R. Qin, X. Huang, W. Liu, and C. Xiao, "Pairwise stereo image disparity and semantics estimation with the combination of u-net and pyramid stereo matching network," in *IEEE Geoscience and Remote Sensing Society*, 2019. 6

[34] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *The European Conference on Computer Vision (ECCV)*, 2018. 6

[35] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2008. 6

[36] A. Rowel, "Fast disparity estimation using dense networks," in *International Conference on Robotics and Automation (ICRA)*, 2018. 6

[37] H. Chen, Lin M., H. Zhang, G. Yang, G.-S Xia, X. Zheng, and L. Zhang, "Multi-level fusion of the multi-receptive fields contextual networks and disparity network for pairwise semantic stereo," in *IEEE Geoscience and Remote Sensing Society*, 2019. 6