Multi-view Semantic Information Guidance for Light Field Image Segmentation

Yiming Li[1], Ruixuan Cong[1,2], Sizhe Wang[1,2], Mingyuan Zhao[1],
Yang Zhang[4], Fangping Li[1], Hao Sheng[1,2,3*]
[1] State Key Laboratory of Virtual Reality Technology and Systems,
School of Computer Science and Engineering, Beihang University, Beijing 100191, P.R.China
[2] Beihang Hangzhou Innovation Institute Yuhang, Xixi Octagon City,
Yuhang District, Hangzhou 310023, P.R.China
[3] Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR 999078, P.R.China
[4] College of Information Science and Technology,
Beijing University of Chemical Technology, Beijing 100029, P.R.China
{Liyiming914, congrx, sizhewang, mingyuanzhao, shenghao}@buaa.edu.cn

yang_zh@mail.buct.edu.cn dclifp@126.com

Abstract

One of the great important fields of computer vision is semantic segmentation. As for single image semantic segmentation, due to limited available information, it appears poor performance when the occlusion and similar color interference occur, and has difficulty exploiting the rich scene information. In comparison, the special micro-len array structure of light field camera can record multi-view information of the scene, which provides us with a new solution to solve this issue. In this paper, we propose a multi-view semantic information guidance network (MSIGNet) for light field semantic segmentation. It can effectively utilize semantic information from multi-view images to guide pixel feature of center view image. First, we extract feature of each view image and further obtain semantic probability. Then all probabilities are aggregated through a self-adaptive multi-view probability fusion module. Last, the resulting coarse fusion representation interacts with center view feature to obtain the refined segmentation result. The proposed method shows excellent performance on both real-world and synthetic light field datasets.

1. Introduction

Semantic segmentation is an important task in the field of computer vision, especially showing great significance for scene understanding. The main aim of it is to assign a class label to each pixel of the specified image [24]. Early scholars carry out conventional methods to solve this issue.

*represents the corresponding author.

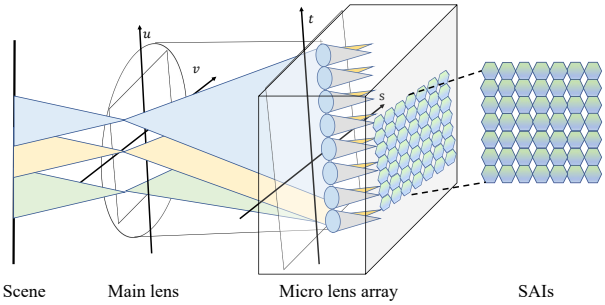

Figure 1. Illustration of the light field camera imaging. The light emitted by the object first passes through the main lens of the camera and then passes through the microlen array to form a light field sub-aperture image array on the surface of the sensor.

The proposal of Fully Convolutional Network (FCN) [16] has greatly promoted the development of semantic segmentation. Based on this, some approaches explore the multi-scale context relationship [3, 30] or introduce the attention mechanism [7,13,18,31] to improve accuracy. Furthermore, other methods attempt utilizing additional data to change the situation. With the help of depth information, RGB-D-based methods [25] achieve superior results. Video-based methods [8] use multi-frame information to increase effectiveness while ensuring accuracy. Recently, [21] opens up a new research direction for semantic segmentation called light field semantic segmentation. It shows outstanding segmentation performance through the 4D scene information. Inspired by this, we focus on this novel issue of light field semantic segmentation in this paper.

As illustrated in Fig. 1, a light field camera records the

scene as a 4D light field with two parallel planes. It can capture the intensity and direction of light rays at the same time. The direction information is recorded by the microlens at different position and represented as sub-aperture images (SAIs), which clearly show that light field provides scene geometry information. By aggregating pixels at the same spatial position from different viewpoints, we can form the macro-pixel image (MacPI). Another visualization for light field is Epipolar Plane Image (EPI) representation. It contains multiple 2D image slices and reflects scene depth information through slope of line. Regardless of any aforementioned approaches used to describe light field, regular transformation relation always lies in different views, which makes it possible to use other view images to supplement the details of center view image.

In recent years, with the appearance of the commercial and industrial light field cameras Lytro [9] and Raytrix [17], complete 4D light field structure information represented as $L(u, v, s, t)$ [15] can be obtained by their microlens array. Specifically, the (u, v) records angular resolution, (s, t) is the image resolution. The abundant geometric information embedded in light field can improve the neural network performance. As a result, a series of experiments about light field in conjunction with computer vision have been continuously increased in recent years, such as depth estimation [27], image super-resolution [29], intrinsic decomposition [2]. This attempt even extends to semantic segmentation. For instance, [14] tries to organize the light field into a 2D MacPI and then leverages a 2D semantic segmentation technique to process it. [21] proposes a large-scale light filed dataset tailored for semantic segmentation, and designs a method through using disparity extracted from EPI as additional information to solve the segmentation problem. However, none of above consider the supplement of 4D information to center image feature.

In light of above analysis, we propose a multi-view semantic information guidance network called MSIGNet for light field semantic segmentation, which uses other view information in SAIs to guide the segmentation of the center image. First, it adopts backbone to extract features of the SAIs and generates semantic prediction probability from an FCN module. Followed by a self-adaptive attention module, the probability of all view images are aggregated to form a coarse segmentation result on the basis of attention weight matrix. For the sake of establishing the relationship between image pixel and semantic class, the initial center image feature is interacted with semantic class representation derived from the coarse result. In the end, the proposed network fuses pixel feature and the relationship to acquire refined semantic segmentation result of center view.

In summary, our main contribution can be summarized as follows:

1) We propose a network that can use semantic information of light field multi-view images to guide segmentation of the center image.

2) We design a self-adaptive multi-view probability fusion module to fuse semantic prediction probability of different views via channel attention mechanism.

3) We prove the effectiveness of the proposed MSIGNet through extensive experiments on existing light field semantic segmentation datasets.

2. Related Work

In this section, we first review the light field image, which is the foundation of our work. And then we briefly demonstrate the development of semantic segmentation and light field semantic segmentation.

2.1. Light Field Image

Light field is a complete representation of the ray of light in a 3D world. E.Adelson [1] proposes to utilize a 7D function $L(x, y, z, \theta, \varphi, \lambda, t)$ to describe the light field. The (x, y, z) show the position in 3D space, (θ, φ) demonstrate the horizontal angle and vertical angle of the light, λ is the wavelength, and t represents the time of observation. Considering that the 7D function brings too many parameters, leading to a heavy computation burden, Marc Levoy [15] reduces the dimension of the 7D light field function and use a 4D light field model (u, v, s, t) composed of two parallel planes to describe the light field. In this model, (u, v) and (s, t) are the coordinates of two points on two parallel planes respectively. From these two coordinates, we can ensure a ray of light. Based on this 4D light field model, the commercial Lytro [4] and industrial Raytrix [17] light field cameras come out one after another. They greatly simplify the difficulty of capturing light field, conveniently forming SAI array which usually consists of 9×9 view images. Different images in a SAI array are the result of observing the same scene from different perspectives, and there is a regular transformation relationship between them.

2.2. Semantic Segmentation

Since the FCN [16] is proposed, semantic segmentation has reached to a new stage. Based on FCN, in order to improve the accuracy of prediction, some approaches attempt to explore multi-scale relational context. PSPNet [30] acquires the multi-scale feature by applying different scale pooling operations in pyramid pooling module. DeepLab [3] uses dilated convolutions with different sizes in atrous spatial pyramid pooling module to obtain various feature representation relationships. There are also some methods exploring different attention mechanisms to enhance feature representation ability. CCNet [13] adopts the recurrent criss cross attention module to get the weight of the pixel feature and reduce the computation burden. DANet [7] uses

self-attention to acquire the feature relation between spatial dimension and channel dimension. In recent years, some transformer-based approaches have been proposed. SETR [31] and DPT [18] make great efforts to move the vision transformer [6] to the semantic segmentation problem and achieve excellent results.

2.3. Light Field Semantic Segmentation

The composition of light field and semantic segmentation is in the preliminary stage. Chen et al. [14] synthesize light field image as MacPI, and use the method tailored for 2D image segmentation to process the MacPI. However, this approach loses mass of angular information. UrbanLF [21] proposes the first mature light field semantic segmentation dataset and introduces a method that extracts the disparity feature from four EPI stacks to guide the semantic segmentation of the center image. This method exploits the spatial geometric information contained in light field and proves its effectiveness on light field dataset. However, it only leverages disparity information in light field image and wastes half of view images. Different from these two light field semantic segmentation methods, our approach makes full use of spatial and angular information of light field to guide the semantic segmentation of the center view image by learning the multi-view semantic feature.

3. Approach

In this section, we will describe well-designed MSIGNet elaborately. In section 3.1, we introduce the overall architecture of the proposed network framework. In section 3.2, we introduce the self-adaptive multi-view probability fusion module based on channel attention mechanism. In section 3.3, we introduce the detailed implementation that uses fusion semantic class feature representation to guide the segmentation of center view image.

3.1. Network Architecture

Considering that the light field records structured multi-view information embedded in SAIs, and contains a strong as well as regular complementarity among different viewpoints. We extract multi-view semantic information of light field to guide the image segmentation of the center view, and further design an innovative network according to this idea. The overall network framework is shown in Fig. 2.

The proposed MSIGNet takes K light field SAIs as input. After feeding to the feature extraction backbone, we can get the feature map of K SAIs, with a resolution reduced to $1/8$ of the original image size. Here we denote the i-th feature map as $F_{ref_i} \in \mathbb{R}^{C\times H\times W}$ $(i = 1, \ldots, K)$ for simplicity. In particular, the feature map of the center image from the K SAIs is represented as $F_{cen} \in \mathbb{R}^{C\times H\times W}$. For the SAI feature map array, we apply an FCN module to acquire the semantic prediction probability $P_{r_i} \in \mathbb{R}^{cls\times H\times W}$ $(i = 1, \ldots, K)$ of each view, in which the cls is the number of semantic classes. Considering the relevancy level distinction between different SAIs and center view image, we design a channel attention module to generate a self-adaptive weight matrix, which aims at fusing all probabilities to form a coarse semantic segmentation prediction probability $P_{\overline{r}} \in \mathbb{R}^{cls\times H\times W}$. Then, F_{cen} interacts with $P_{\overline{r}}$ to learn the semantic class representation for the center view image, resulting in $SCR_{cen} \in \mathbb{R}^{C\times cls}$. Through associating SCR_{cen} with F_{cen} to acquire the context relation between pixels and semantic classes, F_{cen} is further optimized to obtain $\overline{F}_{cen} \in \mathbb{R}^{C\times H\times W}$. Processed by a segmentation head, we generate the final refined semantic segmentation probability $P_{cen} \in \mathbb{R}^{cls\times H\times W}$ of center view image.

3.2. Self-adaptive Multi-view Probability Fusing

The multi-view images in light field SAIs can provide abundant information for semantic segmentation of center view image. However, due to the divergence between SAIs, different view images offer supplements to the center view image in varying degree. Therefore, we propose a self-adaptive multi-view probability fusion module to aggregate multi-view information. Specifically, we fuse the semantic probability P_{r_i} of each view from the FCN module to generate the coarse semantic segmentation result $P_{\overline{r}}$. During the fusion process, we give greater weight to more meaningful view images. Because they are more complementary to center view image, which can benefit more to the semantic segmentation of the center view image.

The input of the self-adaptive multi-view probability fusion module is P_{r_i}. First, we concatenate them along the channel dimension. Then, a global average pooling operation is adopted to generate the initial channel attention. Followed by two fully connected layers and a sigmoid layer, we obtain the final channel attention weight that models the importance of different view semantic probabilities. The process is formulated as:

$$W = softmax(FC_2(AvgPool([P_{r_1}, \ldots, P_{r_K}]))) \quad (1)$$

where $W \in \mathbb{R}^{K*cls\times 1\times 1}$ represents the channel attention weight, $[\cdot]$ denotes the concatenation operation, $AvgPool$ denotes global average pooling and FC_2 represents two cascaded fully connected layers. Finally, we fuse all prediction probabilities to build the coarse segmentation result $P_{\overline{r}} \in \mathbb{R}^{cls\times H\times W}$, which is defined as:

$$P_{\overline{r}} = C_{1\times 1}(W \odot [P_{r_1}, \ldots, P_{r_K}]) \quad (2)$$

where $C_{1\times 1}$ represents 1×1 convolution layer that shrinks the channel dimension from $K * cls$ to cls.

3.3. Multi-view Semantic Representation Guidance

The single image semantic segmentation only utilizes the information itself, which seriously limits the segmentation

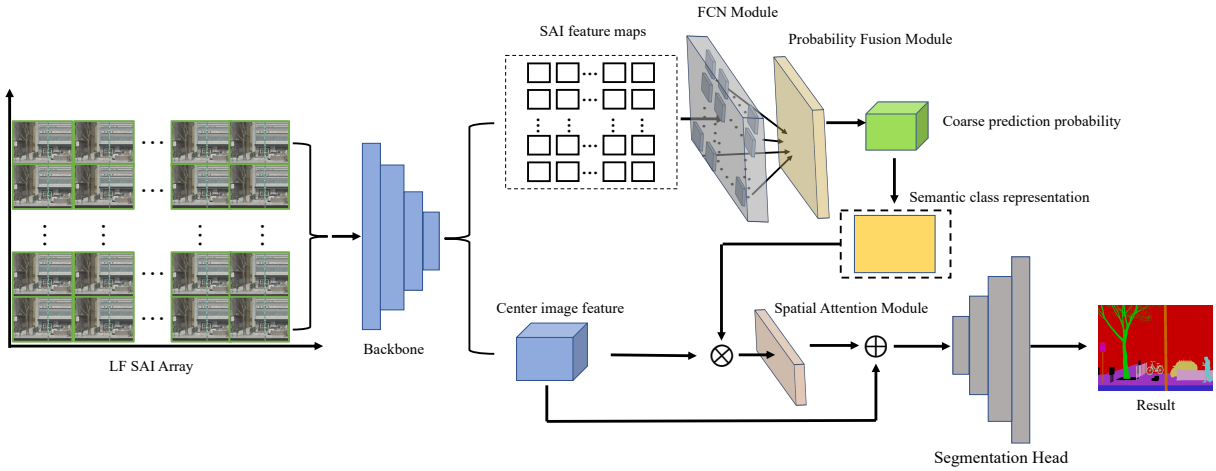

Figure 2. Architecture of the proposed network. Our MSIGNet takes K SAIs as input. They are processed by a feature extraction backbone to generate K SAI feature maps, including the center view image feature. After passing to an FCN module to obtain semantic probabilities, a self-adaptive multi-view probability fusion module is applied to generate a coarse result on the basis of channel attention mechanism. In order to introduce semantic class representation into center view, the coarse fusion probability interacts with center image feature for optimization. Finally, the resulting feature is fed to a segmentation head to form refined semantic segmentation result of center view.

inference performance. As a comparison, the special structure of light field stores 4D scene geometric information, making it possible to take advantage of the multi-view cues to supplement center view image and form a more distinct pixel feature. Therefore, we use aggregated coarse semantic probability $P_{\overline{r}}$ from the multi-view images to guide the image segmentation of the center view.

Specifically, we reshape $P_{\overline{r}}$ to form $P_{cls} \in \mathbb{R}^{cls \times HW}$, in which each line of P_{cls} stands for the semantic class probability of all pixels. At the same time, we reshape the F_{cen} to $F'_{cen} \in \mathbb{R}^{C \times HW}$, which represents the multi-dimension feature of all pixels in center image. Then we get the semantic class representation for center view image $SCR_{cen} \in \mathbb{R}^{C \times cls}$ by the following formulation:

$$SCR_{cen} = F'_{cen} \otimes {P_{cls}}^{T} \tag{3}$$

where each column in SCR_{cen} represents multi-dimension feature of each semantic class. We further interact SCR_{cen} with F_{cen} to calculate the similarity between the pixel of center image and the semantic class through cross-attention mechanism. The Q, K, V matrix are generated by:

$$f(x) = \begin{cases} Q = f_q(F_{cen}) \\ K = f_k(SCR_{cen}) \\ V = f_v(SCR_{cen}) \end{cases} \tag{4}$$

Then we get the class similarity $W_{cls} \in \mathbb{R}^{HW \times cls}$ by:

$$W_{cls} = softmax(\frac{QK^T}{\sqrt{C}}) \tag{5}$$

After weighting value matrix V via W_{cls}, the obtained result is combined with F_{cen} through a local skip residual connection to acquire optimized feature $\overline{F}_{cen} \in \mathbb{R}^{C \times H \times W}$. The overall process is formulated as:

$$\overline{F}_{cen} = W_{cls} \otimes V + F_{cen} \tag{6}$$

Finally, the optimized feature is fed to a segmentation head to achieve the final refined semantic segmentation result of center view $P_{cen} \in \mathbb{R}^{cls \times H \times W}$.

4. Experiments

4.1. Training Settings

Dataset All experiments involved in this article are based on UrbanLF dataset [21]. As a light field semantic segmentation dataset, UrbanLF has currently the largest scale and most semantic categories in the world. UrbanLF can be divided into two subsets termed as UrbanLF-Real and UrbanLF-Syn, respectively. In UrbanLF-Real, it contains 824 real world data samples, and each light field sample includes 9×9 SAIs, and ground-truth segmentation value of center view image. UrbanLF-Syn contains 250 synthetic samples, in which each light field sample includes 9×9 SAIs, ground-truth segmentation value and depth information as well as disparity information of all 9×9 views. In this paper, we evaluate our approach in both UrbanLF-Real and UrbanLF-Syn.

Experiment Details We implement our MSIGNet with the open-source mmsegmentation framework [5]. We set the learning rate as 0.01, the momentum as 0.9, and the weight decay as 0.0005. We select SGD function [10] as

Method	Backbone	Type	Params	Acc	mAcc	**mIoU**↓	Acc*	mAcc*	mIoU*
DAVSS [32]	Xception-65	Video	56.0M	91.04	83.54	75.91	91.74	84.54	77.68
DeepLabv3$^+$ [3]	ResNet-101	RGB	59.3M	91.02	83.53	76.27	91.50	84.30	77.35
PSPNet [30]	ResNet-101	RGB	65.6M	91.21	83.87	76.34	91.74	84.68	77.75
TDNet [12]	ResNet-50	Video	65.3M	91.05	83.38	76.48	91.79	84.85	78.36
MSIGNet(Ours)	ResNet-101	**LF**	**65.6M**	91.12	83.91	76.65	91.66	85.46	77.81
TMANet [26]	ResNet-50	Video	33.4M	91.67	84.13	77.14	91.87	84.55	77.91
SETR [31]	Vit-Large	RGB	96.9M	92.16	84.27	77.74	92.71	84.93	79.05
MSIGNet(Ours)	HrNetV2-W48	**LF**	76.2M	92.08	85.31	77.95	92.47	85.95	79.12
PSPNet-LF [21]	ResNet-101	**LF**	127.8M	92.14	84.86	**78.10**	92.77	85.73	**79.55**

Table 1. Comparison with different types of state-of-the-art methods for semantic segmentation on UrbanLF-Real. Our methods achieve excellent performance with proper model size. * signifies multi-scale testing. LF-based methods and notable values are in bold. All methods are ranked based on the ascending order of mIoU value from top to bottom.

Method	Backbone	Type	Params	Acc	mAcc	**mIoU**↓	Acc*	mAcc*	mIoU*
DAVSS [32]	Xception-65	Video	56.0M	89.47	82.94	74.27	90.94	85.15	77.33
TDNet [12]	ResNet-50	Video	65.3M	89.06	83.43	74.71	89.79	84.32	76.39
DeepLabv3$^+$ [3]	ResNet-101	RGB	59.3M	89.60	83.55	75.39	90.99	85.35	78.05
PSPNet [30]	ResNet-101	RGB	65.6M	89.39	84.48	75.78	90.76	85.64	78.16
TMANet [26]	ResNet-50	Video	33.4M	89.76	84.44	76.41	90.99	86.30	78.87
MSIGNet(Ours)	ResNet-101	**LF**	**65.6M**	89.94	85.13	76.79	91.19	87.02	79.46
SETR [31]	Vit-Large	RGB	97.0M	90.97	85.26	77.69	91.74	86.60	79.32
PSPNet-LF [21]	ResNet-101	**LF**	127.8M	90.55	85.91	77.88	91.55	87.54	80.09
OCR [28]	HRNetV2-W48	RGB	70.4M	91.50	86.96	79.36	92.44	88.18	81.22
ESANet [20]	ResNet-34	RGB-D	46.9M	91.81	86.26	79.43	92.63	86.97	80.97
MSIGNet(Ours)	HrNetV2-W48	**LF**	76.2M	92.00	87.71	**80.33**	93.12	89.37	**82.70**

Table 2. Comparison with different types of state-of-the-art methods for semantic segmentation on UrbanLF-Syn. Our methods achieve excellent performance with proper model size. * signifies multi-scale testing. LF-based methods and notable values are in bold. All methods are ranked based on the ascending order of mIoU value from top to bottom.

the optimizer. Two NVIDIA RTX 3090 are used for distributed training. For the dataset, we adopt random scaling, cropping, flipping, and photometric distortion for data augmentation [22]. The image in UrbanLF-Real is cropped to 432×432 and the image in UrbanLF-Syn is cropped to 480×480. We perform 80k training iterations and take one validation for every 2000 iterations.

Model Selection The proposed MSIGNet can be flexibly combined with different feature extraction backbones. After considering the balance between the accuracy and the speed of inference calculation, we choose ResNet-101 [11] and HRNetV2-W48 [23] as the backbones for our experiments. For these two backbones, ResNet-101 can make all network more lightweight, while HRNetV2-W48 shows better ability in the aspect of feature representation. Based on the ResNet-101 and HRNet-48, we acquire two models: MSIGNet-Res101 and MSIGNet-HR48. Note that we use 5 reference views (*i.e.* $K = 5$) in our experiment owing to the memory limit.

Evaluation Criteria We use pixel accuracy (Acc), mean pixel accuracy (mAcc), and mean intersection-over-union (mIoU) [19] to evaluate different methods.

$$Acc = \sum_{i}^{n_c}(\frac{n_{ii}}{s}) \tag{7}$$

$$mAcc = \frac{1}{n_c} \cdot \sum_{i}^{n_c} \frac{n_{ii}}{s} \tag{8}$$

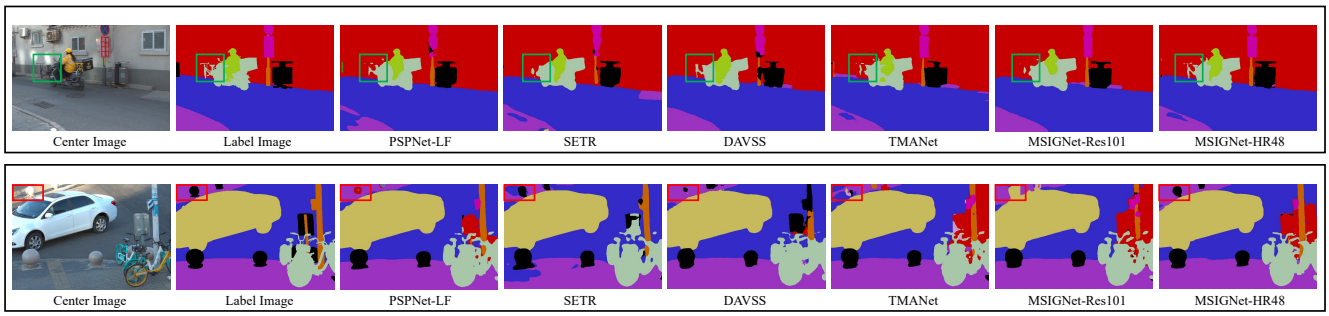

Figure 3. Qualitative result for the proposed MSIGNet on UrbanLF-Real. The colorful boxes show the detailed discrepancy among different semantic segmentation methods.

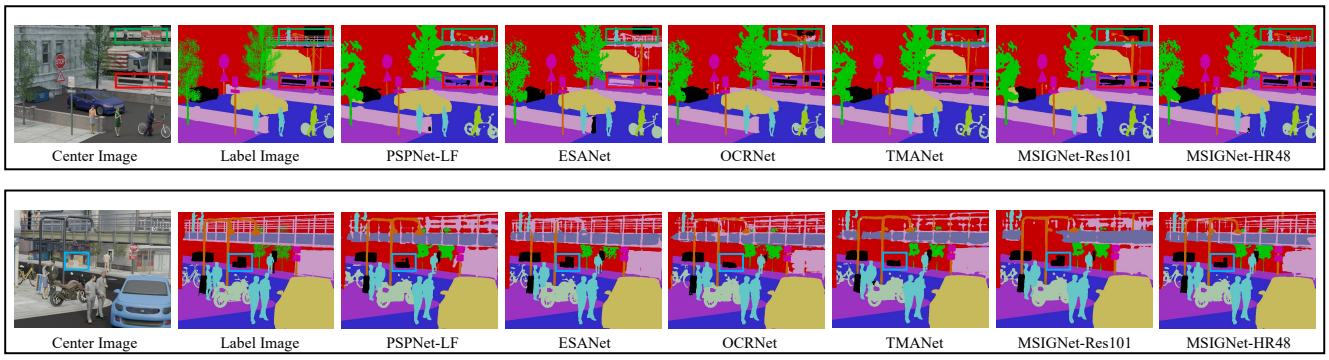

Figure 4. Qualitative result for the proposed MSIGNet on UrbanLF-Syn. The colorful boxes show the detailed discrepancy among different semantic segmentation methods.

$$mIoU = \frac{1}{n_c} \cdot \sum_{i}^{n_c} \frac{n_{ii}}{s_i + \sum_{i}^{n_c} n_{ji} - n_{ii}} \tag{9}$$

where n_{ij} is the total number of pixels whose ground truth semantic label is the i-th class and predict label is the j-th semantic class. s is the total number of pixels. n_c is the total number of semantic classes.

4.2. Experimental Results

We compare our method with state-of-the-art light field semantic segmentation methods. Besides, considering there are only a few numbers of light field-based methods, we carry out the comparison with some single image, video, RGB-D-based semantic segmentation approaches. The results are shown in Tab. 1 and Tab. 2.

Results on UrbanLF-Real As listed in Tab. 1, our well-designed MSIGNet makes a higher accuracy value compared with the generic semantic segmentation methods. As for the light field-based approaches, the small baseline in UrbanLF-Real makes different view information similar, leading to limited complementary information between multi-view images and center view image. Therefore, our method is not conducive to superior performance on real-world data samples. Compared with PSPNet-LF, our MSIGNet is worse than it with 0.15% mIoU with single-scale testing. Fig. 3 shows the qualitative results of our proposed method on UrbanLF-Real dataset.

Results on UrbanLF-Syn Tab. 2 shows the reliable results for different types of semantic segmentation methods. Compared to UrbanLF-Real, UrbanLF-Syn has marked distinction between different view images. Thus we can acquire more complementary information from other views, allowing the proposed model realising its full potential. For instance, our MSIGNet-HR48 exceeds PSPNet-LF with a value of 2.45% mIoU with single-scale testing and a value of 2.61% mIoU with multi-scale testing. The outstanding performance demonstrates the effectiveness of our method. Fig. 4 shows the qualitative results of different segmentation methods on UrbanLF-Syn.

4.3. Ablation Studies

This section introduces the ablation studies to validate the effectiveness of different modules in our method. Considering the inference speed, all experiments utilize ResNet-101 backbone with the same training strategy. In addition, we use single scale testing as the test mode.

Attn	Dataset	Acc	mAcc	mIoU
✓	UrbanLF-Syn	89.94	85.13	76.79
×	UrbanLF-Syn	89.70	85.11	76.61
✓	UrbanLF-Real	91.12	83.91	76.65
×	UrbanLF-Real	90.98	81.99	75.22

Table 3. Ablation study with the probability fusion module. ✓ and × indicate using and not using this module, respectively.

Performance of probability fusion module Our method applies a self-adaptive multi-view probability fusion module to fuse different view prediction probabilities to generate the coarse prediction result. As shown in Tab. 3, we abandon the original channel attention mechanism, introducing a baseline that just concatenates all multi-view prediction probabilities to generate the coarse segmentation result. In this situation, we get the 76.61% mIoU in UrbanLF-Syn and 75.22% mIoU in UrbanLF-Real. The mIoU value has been decreased by 0.18% and 1.43% in UrbanLF-Syn and UrbanLF-Real respectively compared to MSIGNet, which demonstrates the effectiveness of our proposed probability fusion module.

Influence of multi-view information Our method uses multi-view information to guide the semantic segmentation of the center view image. We select UrbanLF-Syn as the experiment dataset. As shown in Tab. 4, We take the number of view images as 1, 3, 5 to perform experiments respectively. When the view number is 1, our MSIGNet only gets the mIoU 76.57%. And then we increase the view number to 3, finding that the mIoU value turns to 76.64%. We further set the view number as 5 to acquire 76.79% mIoU. The continuously increased performance shows the significance of introducing multi-view information.

Impact of semantic information guidance Our method applies the semantic information to guide the image segmentation of center view. To prove the effectiveness, we construct a baseline that utilizes an FCN module to generate the segmentation result after acquiring the fused multi-view semantic probability, rather than use the probability from multi-view to further guide the center view image feature. As shown in Tab. 5, without the semantic information guidance, the mIoU decreases by 5.21%. This result proves the effectiveness of the semantic information guidance of the proposed method.

5. Conclusion

Light field includes the 4D scene geometric information, which provides us with a new solution to alleviate semantic segmentation problem. Previous semantic segmentation methods only focus on the image itself but ignore the other view information. Therefore, we propose a novel network that can adopt the multi-view semantic information from light field to guide the semantic segmentation of the center view image. The proposed MSIGNet primarily consists of three parts. First, it applies a backbone to extract features from SAIs and acquires the prediction probability via FCN module. Then, by using a self-adaptive multi-view probability fusion module, it generates the coarse segmentation result. Finally, the center image feature is interacted with the coarse result to get the final refined semantic segmentation result. We evaluate our approach on the light field semantic segmentation dataset UrbanLF and achieve excellent results.

View-num	Dataset	Acc	mAcc	mIoU
1	UrbanLF-Syn	90.01	85.05	76.57
3	UrbanLF-Syn	89.85	84.97	76.64
5	UrbanLF-Syn	89.94	85.13	76.79

Table 4. Ablation study with the multi-view information. We select UrbanLF-Syn as the experiment dataset. The view-num represents the number of view image used as input in our network.

Guidance	Dataset	Acc	mAcc	mIoU
×	UrbanLF-Syn	87.56	82.23	71.58
✓	UrbanLF-Syn	89.94	85.13	76.79

Table 5. Ablation study with the semantic information guidance. We select UrbanLF-Syn as the experiment dataset. ✓ and × indicate using and not using the semantic information to guide the center view image feature, respectively.

Acknowledgement

This study is partially supported by the National Key R&D Program of China (No.2022YFC3803600), the National Natural Science Foundation of China (No.61872025), and the Open Fund of the State Key Laboratory of Software Development Environment (No.SKLSDE-2021ZX-03). Thank you for the support from HAWKEYE Group.

References

[1] Edward H Adelson, James R Bergen, et al. The plenoptic function and the elements of early vision. *Computational models of visual processing*, 1(2):3–20, 1991. 2

[2] Anna Alperovich, Ole Johannsen, Michael Strecke, and Bastian Goldluecke. Light field intrinsics with a deep encoder-decoder network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9145–9154, 2018. 2

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image

segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1, 2, 5

[4] Donghyeon Cho, Minhaeng Lee, Sunyeong Kim, and Yu-Wing Tai. Modeling the calibration pipeline of the lytro camera for high quality light-field image reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3280–3287, 2013. 2

[5] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 4

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[7] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 1, 2

[8] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41–65, 2018. 1

[9] Todor Georgiev, Zhan Yu, Andrew Lumsdaine, and Sergio Goma. Lytro camera technology: theory, algorithms, performance analysis. In *Multimedia content and mobile devices*, volume 8667, pages 458–467. SPIE, 2013. 2

[10] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International conference on machine learning*, pages 5200–5209. PMLR, 2019. 4

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[12] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks for fast video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8818–8827, 2020. 5

[13] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 1, 2

[14] Chen Jia, Fan Shi, Meng Zhao, Yao Zhang, Xu Cheng, Mianzhao Wang, and Shengyong Chen. Semantic segmentation with light field imaging and convolutional neural networks. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2021. 2, 3

[15] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996. 2

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 2

[17] Christian Perwa and Lennart Wietzke. Raytrix: Light filed technology, 2018. 2

[18] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 1, 3

[19] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 5

[20] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13525–13531. IEEE, 2021. 5

[21] Hao Sheng, Ruixuan Cong, Da Yang, Rongshan Chen, Sizhe Wang, and Zhenglong Cui. Urbanlf: a comprehensive light field dataset for semantic segmentation of urban scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7880–7893, 2022. 1, 2, 3, 4, 5

[22] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 5

[23] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 5

[24] Martin Thoma. A survey of semantic segmentation. *arXiv preprint arXiv:1602.06541*, 2016. 1

[25] Changshuo Wang, Chen Wang, Weijun Li, and Haining Wang. A brief survey on rgb-d semantic segmentation using deep learning. *Displays*, 70:102080, 2021. 1

[26] Hao Wang, Weining Wang, and Jing Liu. Temporal memory attention for video semantic segmentation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2254–2258. IEEE, 2021. 5

[27] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Wei An, and Yulan Guo. Occlusion-aware cost constructor for light field depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19809–19818, 2022. 2

[28] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020. 5

[29] Shuo Zhang, Youfang Lin, and Hao Sheng. Residual networks for light field image super-resolution. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 11046–11055, 2019. 2

[30] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1, 2, 5

[31] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 1, 3, 5

[32] Jiafan Zhuang, Zilei Wang, and Bingke Wang. Video semantic segmentation with distortion-aware feature correction. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):3128–3139, 2020. 5