Combining Implicit-Explicit View Correlation for Light Field Semantic Segmentation

Ruixuan Cong[1,2,3], Da Yang[1,2,3], Rongshan Chen[1,2,3], Sizhe Wang[1,2,3], Zhenglong Cui[1,2,3], Hao Sheng[1,2,3*]
[1] State Key Laboratory of Virtual Reality Technology and Systems,
School of Computer Science and Engineering, Beihang University
[2] Beihang Hangzhou Innovation Institute Yuhang
[3] Faculty of Applied Sciences, Macao Polytechnic University
{congrx, da.yang, rongshan, sizhewang, zhenglong.cui, shenghao}@buaa.edu.cn

Abstract

Since light field simultaneously records spatial information and angular information of light rays, it is considered to be beneficial for many potential applications, and semantic segmentation is one of them. The regular variation of image information across views facilitates a comprehensive scene understanding. However, in the case of limited memory, the high-dimensional property of light field makes the problem more intractable than generic semantic segmentation, manifested in the difficulty of fully exploiting the relationships among views while maintaining contextual information in single view. In this paper, we propose a novel network called LF-IENet for light field semantic segmentation. It contains two different manners to mine complementary information from surrounding views to segment central view. One is implicit feature integration that leverages attention mechanism to compute inter-view and intra-view similarity to modulate features of central view. The other is explicit feature propagation that directly warps features of other views to central view under the guidance of disparity. They complement each other and jointly realize complementary information fusion across views in light field. The proposed method achieves outperforming performance on both real-world and synthetic light field datasets, demonstrating the effectiveness of this new architecture.

1. Introduction

Semantic segmentation is a pixel-level task that assigns a class label to each pixel of the given image, serving as a key fundamental of visual understanding. Due to the partial visibility incurred by occlusion as well as high intra-class variation with diverse appearances, viewpoints and scales, accurate segmentation is a fairly challenging problem. A series of image segmentation methods [4, 11, 41, 43] have been proposed to address these challenges. Furthermore, [1, 7, 23, 37] take depth information into consideration to overcome the deficiency of single image. Recently, [16, 24] employ light field to achieve impressive performance, providing a new perspective for semantic segmentation.

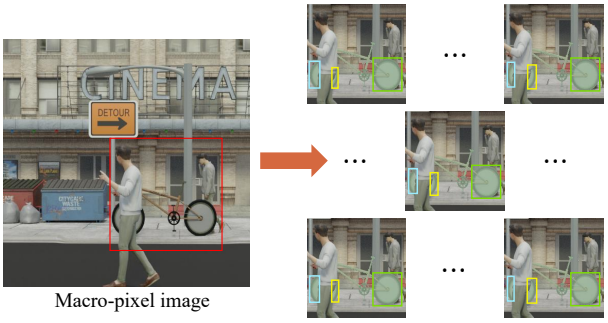
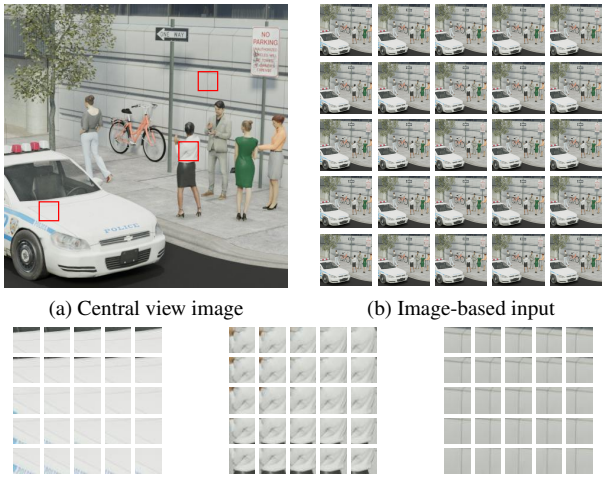


Figure 1. Illustration of light field imaging. Red rectangle shows an occlusion scene, in which the front wheel of bicycle is divided into two areas (enclosed in blue and yellow boxes) and the back wheel is complete (enclosed in green box). Since viewpoints are arranged on a regular grid in angular plane, the location and scale of these areas are regularly changed across views, which is a unique advantage of light field. Influenced by the pedestrian with big disparity, the changes near the front wheel are significant.

Compared to traditional imaging system, 4D light field records intensity for rays in terms of position and direction, yielding a regularly distributed multi-view image array. The information embedded in additional angular dimensions is beneficial for detail analysis to thoroughly parse scenes. As shown in Fig. 1, the front wheel of bicycle is occluded, forming two small areas that are hard to assign labels. With the transformation of viewpoint, the scale of areas changes accordingly. Capturing such regular change with the help

*Corresponding author



(a) Central view image (b) Image-based input

(c) Patch-based input

Figure 2. Illustration of input form for light field with an angular resolution of 5×5. Red rectangles in (a) show three patches with similar color, different disparity and semantic labels. (c) are corresponding patch arrays composed of all SAIs. Super-resolution and disparity estimation can use patch-based input because the former only needs to consider surrounding pixels, and the latter emphasizes pixel matching across views. On the contrary, semantic segmentation cannot assign correct labels based on local information in (c). It requires context information from (b).

of disparity facilitates eliminating ambiguity around object boundaries and occlusion. To this end, it is meaningful to introduce light field into semantic segmentation.

As an emerging research field, light field semantic segmentation can absorb foundations from the great progress of generic semantic segmentation. For instance, a straightforward solution is to organize light field into a 2D macro-pixel image (MacPI) and then apply image semantic segmentation. Video semantic segmentation is also available because light field can be converted into a sub-aperture image (SAI) array, which is similar in form to video sequence. Since depth and disparity are used interchangeably, the disparity contained in light field is workable for RGB-D semantic segmentation. Nevertheless, directly applying these three kinds of methods cannot make full use of advantages of 4D light field. First, treating light field as 2D image inevitably ignores angular information. Second, the regular 2D angular information between SAIs is more compact and intact than 1D temporal information in video. Third, RGB-D-based methods merely extract depth as input, lacking further process about light field. Consequently, it is necessary to design a framework tailored for light field.

In order to realize an overall extraction of 4D information in light field, an effective way is to model spatial relationships in each SAI separately, and then perform interactions across all views along the angular dimension. This modeling mechanism is commonly used in research of light field like super-resolution [33, 38] and disparity estimation [27, 32]. Considering prohibitive inference cost and limited memory usage, each SAI is cropped into multiple small patches for calculation. However, as illustrated in Fig. 2, it is catastrophic and unsuitable for semantic segmentation because independent small patches discard valuable contextual information [35]. Resizing all SAIs to an extremely low scale along the spatial dimension is an alternative manner, but it gives up resolution and granularity which are critical for dense prediction tasks.

In the light of above issues, we present a well-engineered framework, which includes an implicit branch and an explicit branch to fully explore structural information in light field for robust semantic segmentation of central view. The implicit branch only processes a few SAIs and utilizes self-attention and cross-attention mechanisms to calculate similarity for feature integration. The explicit branch processes all SAIs to estimate disparity for subsequent feature propagation. In brief, our framework realizes feature enhancement for central view through implicit feature integration and explicit feature propagation. It is worth noting that two branches transmit supplemental information to each other. Specifically, implicit branch leverages estimated disparity from explicit branch to adjust the weight of cross-attention, enhancing the perception of variation among views. On the other hand, the features to be warped in explicit branch derive from implicit branch. The output features of two branches are fused for final prediction.

Our contributions can be summarized as follows. (1) We present a network called LF-IENet which incorporates implicit and explicit view correlation to exploit light field. The former learns a unified representation within target view and across views. The latter uses disparity to propagate features to target view. (2) The proposed network exists information interaction between two manners, acting as a supplementary item for one another rather than standalone and jointly improving the utilization efficiency of light field. (3) Extensive experiments on the light field semantic segmentation dataset confirm the effectiveness of our method.

2. Related Work

In this section, we briefly review generic semantic segmentation and light field semantic segmentation that are closely related to our work. Since the proposed method uses disparity cue, we additionally introduce light field depth estimation and its application to other tasks.

2.1. Generic Semantic Segmentation

As the pioneering work of image semantic segmentation, FCN [21] firstly adopts fully convolution networks to make pixel-wise predictions combined with the success of deep learning. To make high-accuracy predictions, recent

researches present various schemes following FCN. PSPNet [41] and DeepLab [4] utilize parallel adaptive pooling operations and atrous spatial pyramid pooling to capture multi-scale context, respectively. [11, 15, 42] propose different attention mechanisms to focus on the similarity between pixels for relational context. More recent works like SETR [43] and SegFormer [34] apply transformer to extract global context and achieve outstanding performance.

Video semantic segmentation emphasizes real-time prediction for each frame in a video sequence, which requires a trade-off between quality and speed. Therefore, researchers devote efforts into two directions. Methods in the first category [45,46] concentrate on improving the segmentation efficiency. They reuse the high-level features extracted from key frame, and propagate them along temporal dimension through optical flow. The second category [20,26,30] aim at high-accuracy segmentation result. They incorporate temporal context from several consecutive frames, and interact with target frame to enforce consistency.

RGB-D semantic segmentation is another branch of semantic segmentation. It introduces depth information to distinguish instances which share similar color and texture. [7,23,36] feed RGB and depth data to two parallel streams, where the output features are seamlessly integrated for prediction. In contrast, some methods propose well-engineered layers to replace general convolutional layers by applying depth values, such as S-Conv [6] and ShapeConv [1]. Furthermore, [28, 37] use multitask learning strategy to combine semantic segmentation and depth estimation, in which depth is treated as one of the supervised signals.

2.2. Light Field Semantic Segmentation

Compared with generic semantic segmentation, the research related to light field semantic segmentation is in a preliminary stage. Chen et al. [16] present the first CNN-based method to achieve this goal. They utilize feature extraction backbones to process MacPI, obtaining the result of entire light field. However, only utilizing 2D spatial information is one-sided. Sheng et al. [24] adopt light field to refine the result of central view image. On the basis of existing image semantic segmentation works [35,41], they insert a new branch that excavates structural information from EPIs to refine features of central view. Nevertheless, lines in EPIs are susceptible to noise and occlusion and information involved in more than half of views is abandoned.

Different from the currently only two light field semantic segmentation works mentioned above, the proposed method fully explores spatial and angular information from the light field. It adopts two ways to enhance feature representation of central view. One is to perform feature interaction between side view and center view for geographic relation integration. The other is to utilize the estimated disparity to warp features of side view to central view.

2.3. Light Field Disparity Estimation

The goal of light field disparity estimation is to compute a displacement map that represents the spatial projection coordinate transformation between adjacent views. After long-term research and accumulation, massive methods [12,19,22,25,27,32,40,44] have been proposed to boost the development of this field. They can be divided into conventional methods and learning-based methods, EPI-based methods and non-EPI-based methods, full-supervised methods and unsupervised methods.

Disparity information has been exploited to help other light field tasks. For example, it is used to warp multi-view features to align central view for superior super-resolution result [8, 17]. The intractable problem of light field dimensionality reduction can also be solved with disparity for high-efficiency data compression [3, 10].

3. The Proposed LF-IENet

In this section, we describe the proposed LF-IENet for light field semantic segmentation. Given a 4D light field $L \in \mathbb{R}^{U\times V\times H\times W\times 3}$ composed of $U\times V$ SAIs with $H\times W$ spatial resolution, only central view image $L_{\boldsymbol{a}_c} \in \mathbb{R}^{H\times W\times 3}$ has annotated semantic labels $Y_{\boldsymbol{a}_c} \in \mathbb{R}^{H\times W\times C_{cls}}$, in which C_{cls} is the total number of semantic categories. The remaining SAIs except $L_{\boldsymbol{a}_c}$ are denoted as reference view images $L_{\boldsymbol{a}_r^i} \in \mathbb{R}^{H\times W\times 3}$ $(i = 1, \cdots, UV-1)$. Our proposed network only predicts semantic labels $\overline{Y}_{\boldsymbol{a}_c} \in \mathbb{R}^{H\times W\times C_{cls}}$ for central view with the help of reference views. Note that here $\boldsymbol{a}$ represents angular coordinates (u, v) and we achieve semantic segmentation using SAIs distributed on a square array of angular dimensions, *i.e.*, $U = V = A$.

The arrangement of this section is organized as follows. In Section 3.1, we introduce the core idea and framework of the proposed LF-IENet from a global perspective. The implicit and explicit branches that utilize structural information of light field are elaborated in Section 3.2 and Section 3.3, respectively. In Section 3.4, we present the loss functions used to train the proposed network.

3.1. Motivation and Framework

Light field records the same scene from multiple regularly varying views, containing a wealth of visual information and encoding geometric information. The surrounding view image in different angular directions possesses subpixel offsets in specific spatial direction. As shown in Fig. 1, the blue and yellow regions of central view can not only correlate the similar features from green region of same view, but also from blue, yellow and green regions of other views. Besides, the spatial location of these regions varies with the change of view. Therefore, a rigorous exploration of the complementary information within a single view and across views is crucial for high-quality semantic segmentation.

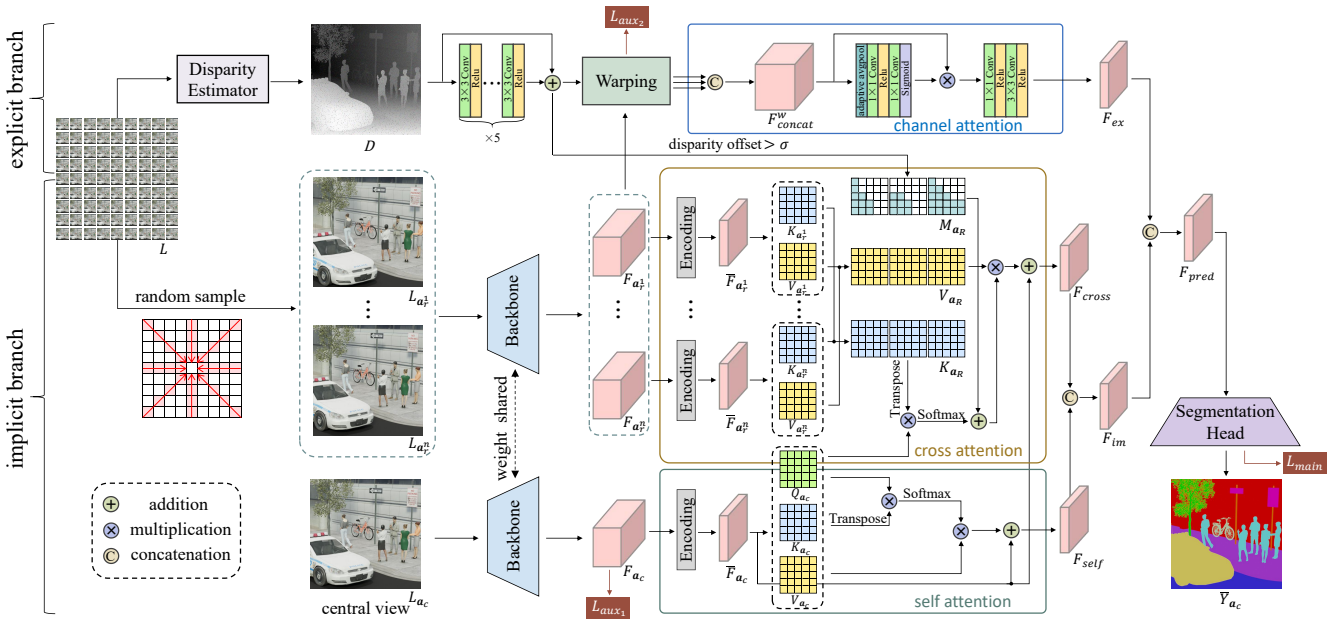


Figure 3. Architecture of the proposed LF-IENet. Central SAI and n ($n \leq 4$) reference SAIs are fed into a shared backbone to extract features. At the same time, a disparity estimator predicts disparity value for light field image. To fully exploit complementary information from reference views to reinforce central view, we adopt two different approaches, called implicit feature integration and explicit feature propagation. The former uses self-attention and cross-attention to integrate similar context information, in which the cross-attention focuses more on inconsistent regions across views. The latter propagates feature based on the estimated disparity and relative angular position. Loss functions for training are viewed with red rectangles. Note that n reference views are randomly sampled from one of the eight image stacks, whose viewpoints change along horizontal, vertical, left and right diagonal directions in order.

One feasible scheme is to apply the powerful attention mechanism [29] to light field. It can implicitly retrieve useful information from reference views and central view by computing the feature correlation to enhance the representation of central view. However, this solution is incapable of fully utilizing 2D angular information in light field because of ignoring the relative position association among viewpoints. Feature propagation based on disparity is another valid solution. It firstly computes a disparity prior from light field, and then explicitly produces the features of central view by propagating the features of reference views under the guidance of estimated disparity. Obviously, the accuracy of light field disparity estimation determines the quality of propagated features. In this paper, we touch on the joint learning of complementary information among views in the aforementioned implicit and explicit manners, which are both important and effective for semantic segmentation of central view image.

The overall pipeline of our proposed method is illustrated in Fig. 3. It takes entire light field as input and outputs the result of central view. There are two branches in the network. In the implicit branch, it firstly uses a feature extraction backbone to process several view images including central view image. Then self-attention and cross-attention are performed to reinforce feature representation capability, in which self-attention is designed to gather context information within central view, while cross-attention aims at exploiting similarity from other views to compensate central view. In the explicit branch, it firstly computes an initial disparity map through mature light field disparity estimation technologies. According to photo-consistency assumption, features of other reference views are propagated to central view. Finally, the output features of two branches F_{im} and F_{ex} are aggregated to produce the final feature of central view F_{pred}, which is then fed to a segmentation head to get the predicted segmentation map.

3.2. Implicit Feature Integration

Limited by computation and memory, We select n SAIs instead of all reference views to calculate cross-attention. Following [39], we explore the complementary information from surrounding reference view images that have horizontal, vertical or diagonal sub-pixel shifts. A total of $n+1$ images including central view $\{L_{\boldsymbol{a}_r^1}, \cdots, L_{\boldsymbol{a}_r^n}, L_{\boldsymbol{a}_c}\}$ are fed to the same backbone to extract feature $\{F_{\boldsymbol{a}_r^1}, \cdots, F_{\boldsymbol{a}_r^n}, F_{\boldsymbol{a}_c}\}$, each of which has a size of $\mathbb{R}^{c \times h \times w}$. After different parallel encoding layers, these features are further condensed into $\{\overline{F}_{\boldsymbol{a}_r^1}, \cdots, \overline{F}_{\boldsymbol{a}_r^n}, \overline{F}_{\boldsymbol{a}_c}\}$ with a size of $\mathbb{R}^{c_v \times h \times w}$. Then we generate $Q_{\boldsymbol{a}_c} \in \mathbb{R}^{N \times c_q}$, $K_{\boldsymbol{a}_c} \in \mathbb{R}^{N \times c_k}$, $V_{\boldsymbol{a}_c} \in \mathbb{R}^{N \times c_v}$ of central view image and $K_{\boldsymbol{a}_r^i} \in \mathbb{R}^{N \times c_k}$, $V_{\boldsymbol{a}_r^i} \in \mathbb{R}^{N \times c_v}$ of

each reference image, in which $c_q = c_k$, $N = hw$. Finally we utilize self-attention to enhance feature of central view, which is defined as:

$$F_{self} = R(\text{softmax}(Q_{\boldsymbol{a}_c} \cdot K^T_{\boldsymbol{a}_c}) \cdot V_{\boldsymbol{a}_c}) + \overline{F}_{\boldsymbol{a}_c} \quad (1)$$

where $F_{self} \in \mathbb{R}^{c_v \times h \times w}$, R denotes the reshape conversion from $N \times c_v$ to $c_v \times h \times w$.

In order to make reference view images better supplement central view image, cross-attention should pay more attention to inconsistent regions among views, which are not available via self-attention. As the yellow box and blue box in Fig. 1 show, inconsistent regions can be expressed as regions with large disparity. To this end, we generate an attention mask $M_{\boldsymbol{a}^i_r} \in \mathbb{R}^{h \times w}$ for each reference view by computing the disparity offset of each pixel, which is formulated as follows:

$$M_{\boldsymbol{a}^i_r} = \begin{cases} 1, & d * ||\boldsymbol{a}^i_r - \boldsymbol{a}_c||_2 > \sigma \ / \ \text{stride} \\ 0, & otherwise \end{cases} \quad (2)$$

where $d \in \mathbb{R}^{h \times w}$ is the disparity value of central view image, $||\cdot||_2$ is the euclidean distance, σ is the threshold value, $\text{stride} = H \ / \ h$ denotes the output stride of the backbone. Then all key maps, value maps and attention masks of reference views are sequentially concatenated, permuted, and reshaped to obtain $K_{\boldsymbol{a}_R} \in \mathbb{R}^{N_R \times c_k}$, $V_{\boldsymbol{a}_R} \in \mathbb{R}^{N_R \times c_v}$ and $M_{\boldsymbol{a}_R} \in \mathbb{R}^{1 \times N_R}$, respectively, in which $N_R = nhw$. The cross-attention weight is defined as:

$$W_{cross} = \text{softmax}(Q_{\boldsymbol{a}_c} \cdot K^T_{\boldsymbol{a}_R}) + \omega \ / \ N_R * M_{\boldsymbol{a}_R} \quad (3)$$

where $W_{cross} \in \mathbb{R}^{N \times N_R}$, ω denotes the additional weight for pixels with large disparity. We multiply W_{cross} and $V_{\boldsymbol{a}_R}$ to integrate the angular relation, thus enhancing the feature of central view. The process is formulated as:

$$F_{cross} = R(W_{cross} \cdot V_{\boldsymbol{a}_R}) + \overline{F}_{\boldsymbol{a}_c} \quad (4)$$

where $F_{cross} \in \mathbb{R}^{c_v \times h \times w}$. In this way, cross-attention tends to get more information from inconsistent areas across views with the aid of estimated disparity originating from explicit branch. Finally, we obtain the output feature of implicit branch F_{im} by concatenating F_{self} and F_{cross} along the channel dimension.

3.3. Explicit Feature Propagation

To cast accurate geometric correspondences in feature space, the paramount task is to obtain a high-quality disparity prior. Due to the lack of ground truth disparity for supervision, we leverage OAVC [12], a conventional method with superior performance as disparity estimator to predict the disparity value $D \in \mathbb{R}^{H \times W}$ of central view, in which the label range is set to 256. Note that OAVC runs on the CPU, which decreases the workload on the GPU and leads to a massive reduction in computation time. To further improve the precision of estimated disparity, we apply 5 cascaded 3 × 3 convolution layers with a local skip residual connection to continuously calibrate D, then obtain $d \in \mathbb{R}^{h \times w}$ with the same resolution as $F_{\boldsymbol{a}^i_r}$ through downsampling operation for subsequent feature propagation.

According to 4D light field structure, given a disparity d of central view $\boldsymbol{a}_c$, the spatial coordinates (h^i_r, w^i_r) of pixel $L(u_c, v_c, h_c, w_c)$ in reference view $\boldsymbol{a}^i_r$ can be calculated via the following transformation:

$$(h^i_r, w^i_r) = (h_c, w_c) + d(h_c, w_c) * (\boldsymbol{a}_c - \boldsymbol{a}^i_r) \quad (5)$$

Thus, we successively warp each reference feature to central view to get an aligned feature group $\{F^w_{\boldsymbol{a}^1_r}, \cdots, F^w_{\boldsymbol{a}^n_r}\}$, each of which has a size of $\mathbb{R}^{c \times h \times w}$. Taking into account that each reference view contributes differently to final representation, we adopt channel attention to generate weights for aggregating aligned features. Specifically, all parallel aligned features are firstly concatenated along the channel dimension, then an adaptive pooling is used to squeeze spatial information. After two convolution layers and a sigmoid function, we obtain weights and perform multiplication. Finally, two convolution layers are performed to compact the channel of feature. The overall process is formulated as:

$$F^w_{concat} = [F^w_{\boldsymbol{a}^1_r}, \cdots, F^w_{\boldsymbol{a}^n_r}] \quad (6)$$

$$W_{channel} = \phi(H_{1\times1}(H_{1\times1}(\text{AdaptPool}(F^w_{concat})))) \quad (7)$$

$$F_{ex} = H_{3\times3}(H_{1\times1}(W_{channel} * F^w_{concat})) \quad (8)$$

where $[\cdot]$ represents the concatenation operation, $H_{s\times s}$ represents convolution layer with kernel size $s \times s$, ϕ represents the sigmoid function. F_{ex} represents the output feature of explicit branch, containing effective information from all reference views to augment central view.

3.4. Loss Functions

The total loss L_{total} for training our proposed LF-IENet in an end-to-end fashion consists of three parts. The first term L_{main} is the main cross-entropy loss at the end of the network. The second term L_{aux_1} is an auxiliary cross-entropy loss at $F_{\boldsymbol{a}_c}$ to help optimize the feature extraction process in the backbone. In order to constrain the propagated features in explicit branch close to the feature of central view, we also add an auxiliary cross-entropy loss at each propagated feature $F^w_{\boldsymbol{a}^i_r}$ as the third term L_{aux_2}. It also trains cascaded convolution layers which are used to rectify the estimated disparity. Therefore, the total loss can be formulated as:

$$\begin{aligned} L_{total} &= L_{main} + \lambda_1 * L_{aux_1} + \lambda_2 * L_{aux_2} \\ &= CE(S_{pred}, Y_{\boldsymbol{a}_c}) + \lambda_1 * CE(S_{\boldsymbol{a}_c}, Y_{\boldsymbol{a}_c}) \\ &\quad + \lambda_2 * \sum_{i=1}^{n} CE(S^w_{\boldsymbol{a}^i_r}, Y_{\boldsymbol{a}_c}) \end{aligned} \quad (9)$$

where CE denotes cross-entropy loss function, S_{pred} , $S_{\boldsymbol{a}_c}$ and $S^w_{\boldsymbol{a}^i_r}$, each of which has a size of $\mathbb{R}^{H\times W\times C_{cls}}$, denote segmentation logits of F_{pred}, $F_{\boldsymbol{a}_c}$ and $F^w_{\boldsymbol{a}^i_r}$, respectively. λ_1 and λ_2 denote the weighting factor of two auxiliary losses that require adjusting to deliver the best performance. Note that we try to use mean average error (MAE) loss in image space between warped reference images and central view image to further optimize the estimated disparity, but it has little effect on the accuracy of final segmentation result.

4. Experiments

4.1. Experimental Setup

Dataset. All the experimental results are produced on UrbanLF dataset [24], the only available light field semantic segmentation benchmark up to now. There are two subsets in it, consisting of 824 real-world samples (UrbanLF-Real) and 250 synthetic samples (UrbanLF-Syn). Each sample is composed of 81 SAIs with an angular resolution of 9 × 9 and pixel-wise annotation of central view for 14 classes. We evaluate all methods on these two subsets using the standard train-val-test split.

Evaluation Metrics. Following [24], we use pixel accuracy (Acc), mean pixel accuracy (mAcc) and mean intersection-over-union (mIoU) to evaluate performance. Results produced by both single-scale and multi-scale testing strategies are recorded. For multi-scale testing, random horizontal flipping and multi-scale scaling with four factors (0.75, 1.0, 1.25, 1.5) are exploited. In addition, we calculate the number of parameters to evaluate the efficiency of methods.

Models & Baselines. The proposed LF-IENet can flexibly combine with different backbones. We choose two widely used backbones: ResNet-50 [13] and HRNetV2-W48 [31]. The latter has better feature extraction ability but is less efficient in terms of model size. We apply them into our light field semantic segmentation framework, resulting in two models: LF-IENet^4-Res50 and LF-IENet^3-HR48, in which the number of reference views is 4 and 3, respectively.

Implementation details. We implement our methods with the open source mmsegmentation toolbox [9]. By default, we set $\sigma = 0.8, \omega = 0.2, \lambda_1 = 0.4, \lambda_2 = 0.1$. Backbones are pretrained on ImageNet [18] and the rest modules are weighted through normal initialization. We adopt random scaling, cropping, flipping, and photometric distortion for data augmentation, in which the cropping size is set as 432 × 432 for UrbanLF-Real and 480 × 480 for UrbanLF-Syn. Networks are trained using SGD optimizer with momentum 0.9 and weight decay 5e-4 for 80k iterations. The learning rate is initialized as 0.01 and decreases according to "poly" schedule. Two NVIDIA RTX 3090 are used for distributed training. During inference stage, we sequentially take reference views from each image stack as input and calculate average segmentation score as the final prediction.

4.2. Comparison to State-of-the-art

To prove the effectiveness of our proposed networks, we compare them with the state-of-the-art methods: PSPNet-LF [24], OCR-LF [24]. Considering that the available light field semantic segmentation approaches are rare, the comparison is also extended to other generic methods following [24], including two image-based methods: DeepLabv3$^+$ [4], SETR [43], three video-based methods: TDNet [14], DAVSS [46], TMANet [30], three RGB-D-based methods: MTINet [28], SA-Gate [7], ESANet [23].

Results on UrbanLF-Syn. Tab. 1 presents the quantitative results on the UrbanLF-Syn subset which has obvious changing across views as well as little image noise. Our LF-IENet^3-HR48 achieves a state-of-the-art mIoU of 81.78% with single-scale testing. When multi-scale testing is performed, it also ranks 1st place with highest scores on every metric. Even if compared with RGB-D-based methods using ground truth depth, it gains better accuracy. Moreover, our light-weight LF-IENet^4-Res50 outperforms PSPNet-LF with a heavy ResNet-101 backbone on metric of mAcc and mIoU. The outstanding performance indicates the proposed method effectively leverages complementary information from reference views to reinforce central view image for robust light field semantic segmentation. Fig. 4 shows the qualitative results. It is observed that our methods gain good segmentation results at the edges and occlusions.

Results on UrbanLF-Real. As shown in Tab. 2, our LF-IENet achieves higher accuracy compared with generic semantic segmentation methods. Compared with light field-based methods, our LF-IENet^3-HR48 is slightly worse than OCR-LF with 0.13% mIoU margin applying single-scale testing. The performance gap is highly associated with data quality and focus of method. Specifically, light field captured by plenoptic camera contains plenty of noise, which has an extremely negative effect on disparity estimation. Meanwhile, the small baseline causes nearly identical SAIs, making it dispensable to exploit complementary information across views. Therefore, real-world samples are not conducive to the advantages of the proposed method. On the contrary, OCR-LF puts more effort into mining contextual information inside central view image without considering reference views during decoder stage, thus it gets the highest scores at the expense of large parameters. It is worth noting that the situation changes during multi-scale testing. Differences between view images are magnified through big scaling factors in favor of our approach, so that LF-IENet^3-HR48 obtains the best performance in this case. Our LF-IENet^4-Res50 is worse than PSPNet-LF on almost all metrics. In addition to aforementioned reasons, it is relevant to the capability of feature extraction backbone. Fig. 5 shows the qualitative results of our models and other two light field-based methods.

Method	Backbone	Type	Params.	Acc	mAcc	mIoU	Acc*	mAcc*	mIoU*
DeepLabv3$^+$ [5]	ResNet-101	Image	59.3M	89.60	83.55	75.39	90.99	85.35	78.05
SETR [43]	Vit-Large	Image	97.0M	90.97	85.26	77.69	91.74	86.60	79.32
TDNet [14]	ResNet-50	Video	65.3M	89.06	83.43	74.71	89.79	84.32	76.39
DAVSS [46]	Xception-65	Video	56.0M	89.47	82.94	74.27	90.94	85.15	77.33
TMANet [30]	ResNet-50	Video	33.4M	89.76	84.44	76.41	90.99	86.30	78.87
MTINet [28]	HRNetV2-W48	RGB-D	98.7M	91.24	86.94	79.10	91.86	87.34	80.01
ESANet [23]	ResNet-34	RGB-D	46.9M	91.81	86.26	79.43	92.63	86.97	80.97
SA-Gate [7]	ResNet-101	RGB-D	110.9M	92.10	87.04	79.53	93.18	88.51	81.72
PSPNet-LF [24]	ResNet-101	LF	127.8M	90.55	85.91	77.88	91.55	87.54	80.09
OCR-LF [24]	HRNetV2-W48	LF	137.4M	92.01	87.71	80.43	93.06	89.20	82.77
LF-IENet^4 (Ours)	ResNet-50	LF	94.6M	90.42	86.17	78.27	91.38	87.66	80.33
LF-IENet^3 (Ours)	HRNetV2-W48	LF	117.4M	**92.41**	**88.31**	**81.78**	**93.26**	**89.25**	**83.32**

Table 1. Comparison with state-of-the-art image, video, RGB-D and LF-based methods on UrbanLF-Syn. Our methods achieve outstanding performance with proper model size. The best results are in bold and the second best results are underlined. * signifies multi-scale testing.

Method	Backbone	Type	Params.	Acc	mAcc	mIoU	Acc*	mAcc*	mIoU*
DeepLabv3$^+$ [5]	ResNet-101	Image	59.3M	91.02	83.53	76.27	91.50	84.30	77.35
SETR [43]	Vit-Large	Image	96.9M	92.16	84.27	77.74	92.71	84.93	79.05
TDNet [14]	ResNet-50	Video	65.3M	91.05	83.38	76.48	91.79	84.85	78.36
DAVSS [46]	Xception-65	Video	56.0M	91.04	83.54	75.91	91.74	84.54	77.68
TMANet [30]	ResNet-50	Video	33.4M	91.67	84.13	77.14	91.87	84.55	77.91
PSPNet-LF [24]	ResNet-101	LF	127.8M	92.14	84.86	78.10	92.77	85.73	79.55
OCR-LF [24]	HRNetV2-W48	LF	137.4M	**92.51**	**86.31**	**79.32**	92.68	86.58	80.06
LF-IENet^4 (Ours)	ResNet-50	LF	94.6M	92.01	85.10	78.09	92.38	85.52	79.08
LF-IENet^3 (Ours)	HRNetV2-W48	LF	117.4M	92.09	86.03	79.19	**92.82**	**86.87**	**80.49**

Table 2. Comparison with state-of-the-art image, video and LF-based methods on UrbanLF-Real. Our methods achieve outstanding performance with proper model size. The best results are in bold and the second best results are underlined. * signifies multi-scale testing.

4.3. Ablation Studies

This section introduces ablation studies to validate the effectiveness of prominent components in our method. All experiments are performed on UrbanLF-Syn with ResNet-50 backbone. We only report mIoU of single-scale testing.

Effect of implicit feature integration. Our method leverages cross-attention to implicitly integrate information from reference views to assist central view. To validate the benefit of this design, as shown in Tab. 3, we first introduce a baseline (*model-1*) with only self-attention in the implicit branch, then insert cross-attention into the implicit branch of baseline (*model-3*), achieving a 1.05% mIoU improvement. Moreover, when cross-attention is removed from LF-IENet (*model-5*), the performance of the resulting *model-2* degrades by 0.66% mIoU, proving the effect of implicit feature integration from another perspective.

Effect of explicit feature propagation. Similar to the previous ablation study, we investigate the impact of explicit feature propagation by comparing *model-2* to *model-1* and *model-5* to *model-3*. It can be observed in Tab. 3 that *model-2* and *model-5* obtain a 0.81% and 0.42% higher mIoU, respectively. The performance gap demonstrates that feature propagation also contributes to strengthening the feature of central SAI. More importantly, it provides additional complementary information for central view image in a different way from implicit feature integration.

Effect of attention mask of cross-attention. The attention mask in the proposed network aims to highlight the focus on inconsistent areas across views, which cannot be achieved by self-attention on central view image. To prove the effectiveness introduced by attention mask, we simply remove it without other modification to the original cross-attention weights. Tab. 3 shows that the mIoU score of *model-4* is degraded by 0.22% compared with *model-5*, indicating that it makes sense to pay extra attention to large disparity regions.

Effect of the quality of estimated disparity. In the proposed LF-IENet, feature propagation in the explicit branch is implemented based on disparity, and the implicit branch also applies disparity to enhance attention weight. To study the impact of disparity, we conduct experiments with three kinds of disparity value, including ground truth, result from OAVC [12] and SubFocal [2]. As shown in Tab. 4, the performance is positively related to the accuracy of disparity prior. Once the ground truth disparity is available, the network gains the highest mIoU score, outperforming 0.23%

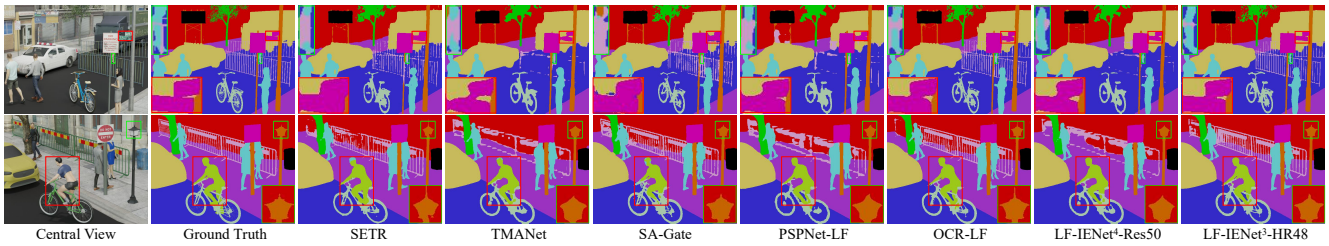


Figure 4. Comparison in terms of qualitative results on UrbanLF-Syn. Best viewed with zooming in red and green.

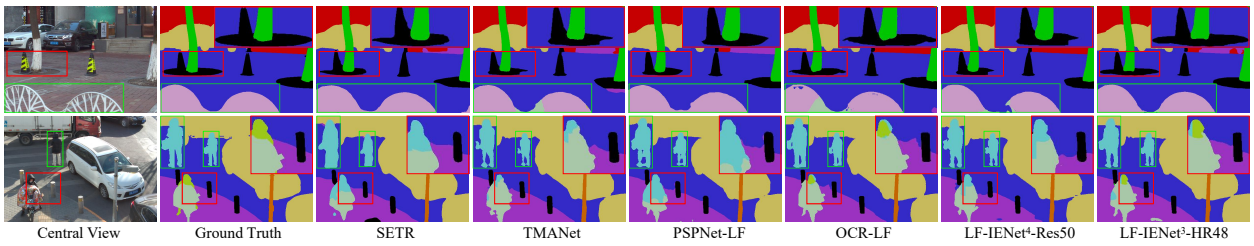


Figure 5. Comparison in terms of qualitative results on UrbanLF-Real. Best viewed with zooming in red and green.

Model	Implicit	Explicit	Mask	Params.	mIoU
1				96.3M	76.80
2		✓		95.3M	77.61
3	✓		✓	97.3M	77.85
4	✓	✓		94.6M	78.05
5	✓	✓	✓	94.6M	**78.27**

Table 3. Ablation study with implicit feature integration, explicit feature propagation, attention mask for LF-IENet^4-Res50. Note that all models keep self-attention in the implicit branch and the channel number of variants is increased to make up the model size. In addition, *model-2* keeps all reference features from the implicit branch. *model-3* keeps the disparity map from the explicit branch.

Methods	Backbone	Disparity	MSE	mIoU
LF-IENet^4	ResNet-50	Ground Truth	-	**78.50**
LF-IENet^4	ResNet-50	OAVC	0.11	78.27
LF-IENet^4	ResNet-50	SubFocal	0.18	78.19

Table 4. Ablation study on UrbanLF-Syn with different estimated disparity. MSE denotes the mean square error of disparity values.

than our default setting (*i.e.,* OAVC). This points out a new direction for us to further optimize the proposed method.

Effect of the number of reference views. Theoretically speaking, with the increase in the number of reference view images, the size and performance of the proposed network rises. Here we investigate the relationship among these factors. Corresponding results are listed in Tab. 5. Restricted by computation and memory, the maximum number of reference views is 4 and our network obtains the highest mIoU score in this case. Moreover, the performance tends to be saturated with more reference views.

Methods	Backbone	View(Ref)	Params.	mIoU
LF-IENet^1	ResNet-50	1	59.3M	77.40
LF-IENet^2	ResNet-50	2	71.0M	77.82
LF-IENet^3	ResNet-50	3	82.8M	78.09
LF-IENet^4	ResNet-50	4	94.6M	**78.27**

Table 5. Ablation study with the number of reference view images.

5. Conclusion

In this work, we explore light field semantic segmentation from a brand-new perspective. In order to take full advantage of additional angular information embedded in light field, we proposed two various manners, *i.e.,* implicit feature integration based on similarity and explicit feature propagation relying on disparity to enhance features of central view image. The implicit and explicit combination framework can exert their own advantages and gain superior performance. We also find that leveraging disparity cue to calibrate weight of cross-attention to rise attention on inconsistent regions across views, further boosts performance. In the future, we will extend the proposed method in two aspects: (1) accurate estimation of disparity prior. (2) efficient model with fewer parameters and higher inference speed.

Acknowledge

This study is partially supported by the National Key R&D Program of China (No.2019YFB2101600), 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] Jinming Cao, Hanchao Leng, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7088–7097, 2021. 1, 3

[2] Wentao Chao, Xuechun Wang, Yingqian Wang, Liang Chang, and Fuqing Duan. Learning sub-pixel disparity distribution for light field depth estimation. *arXiv preprint arXiv:2208.09688*, 2022. 7

[3] Jie Chen, Junhui Hou, and Lap-Pui Chau. Light field compression with disparity-guided sparse coding based on structural key views. *IEEE Transactions on Image Processing*, 27(1):314–324, 2017. 3

[4] 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, 3, 6

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 7

[6] Lin-Zhuo Chen, Zheng Lin, Ziqin Wang, Yong-Liang Yang, and Ming-Ming Cheng. Spatial information guided convolution for real-time rgbd semantic segmentation. *IEEE Transactions on Image Processing*, 30:2313–2324, 2021. 3

[7] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *European Conference on Computer Vision*, pages 561–577. Springer, 2020. 1, 3, 6, 7

[8] Zhen Cheng, Zhiwei Xiong, Chang Chen, Dong Liu, and Zheng-Jun Zha. Light field super-resolution with zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10010–10019, 2021. 3

[9] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 6

[10] Elian Dib, Mikael Le Pendu, Xiaoran Jiang, and Christine Guillemot. Local low rank approximation with a parametric disparity model for light field compression. *IEEE Transactions on Image Processing*, 29:9641–9653, 2020. 3

[11] 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, 3

[12] Kang Han, Wei Xiang, Eric Wang, and Tao Huang. A novel occlusion-aware vote cost for light field depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3, 5, 7

[13] 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. 6

[14] 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. 6, 7

[15] 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. 3

[16] 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. 1, 3

[17] Keunsoo Ko, Yeong Jun Koh, Soonkeun Chang, and Chang-Su Kim. Light field super-resolution via adaptive feature remixing. *IEEE Transactions on Image Processing*, 30:4114–4128, 2021. 3

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 6

[19] Titus Leistner, Hendrik Schilling, Radek Mackowiak, Stefan Gumhold, and Carsten Rother. Learning to think outside the box: Wide-baseline light field depth estimation with epi-shift. In *2019 International Conference on 3D Vision (3DV)*, pages 249–257. IEEE, 2019. 3

[20] Jiangtong Li, Wentao Wang, Junjie Chen, Li Niu, Jianlou Si, Chen Qian, and Liqing Zhang. Video semantic segmentation via sparse temporal transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 59–68, 2021. 3

[21] 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. 2

[22] Jiayong Peng, Zhiwei Xiong, Yicheng Wang, Yueyi Zhang, and Dong Liu. Zero-shot depth estimation from light field using a convolutional neural network. *IEEE Transactions on Computational Imaging*, 6:682–696, 2020. 3

[23] 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. 1, 3, 6, 7

[24] 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*, 2022. 1, 3, 6, 7

[25] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo Kim. Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field im-

ages. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4748–4757, 2018. 3

[26] Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, and Luc Van Gool. Coarse-to-fine feature mining for video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3126–3137, 2022. 3

[27] Yu-Ju Tsai, Yu-Lun Liu, Ming Ouhyoung, and Yung-Yu Chuang. Attention-based view selection networks for light-field disparity estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12095–12103, 2020. 2, 3

[28] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *European Conference on Computer Vision*, pages 527–543. Springer, 2020. 3, 6, 7

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[30] 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. 3, 6, 7

[31] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 6

[32] 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, 3

[33] Yingqian Wang, Longguang Wang, Gaochang Wu, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Disentangling light fields for super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[34] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3

[35] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *European conference on computer vision*, pages 173–190. Springer, 2020. 2, 3

[36] Guodong Zhang, Jing-Hao Xue, Pengwei Xie, Sifan Yang, and Guijin Wang. Non-local aggregation for rgb-d semantic segmentation. *IEEE Signal Processing Letters*, 28:658–662, 2021. 3

[37] Junning Zhang, Qunxing Su, Bo Tang, Cheng Wang, and Yining Li. Dpsnet: Multitask learning using geometry reasoning for scene depth and semantics. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 1, 3

[38] Shuo Zhang, Song Chang, and Youfang Lin. End-to-end light field spatial super-resolution network using multiple epipolar geometry. *IEEE Transactions on Image Processing*, 30:5956–5968, 2021. 2

[39] 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. 4

[40] Shuo Zhang, Hao Sheng, Chao Li, Jun Zhang, and Zhang Xiong. Robust depth estimation for light field via spinning parallelogram operator. *Computer Vision and Image Understanding*, 145:148–159, 2016. 3

[41] 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, 3

[42] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 267–283, 2018. 3

[43] 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, 6, 7

[44] Wenhui Zhou, Enci Zhou, Gaomin Liu, Lili Lin, and Andrew Lumsdaine. Unsupervised monocular depth estimation from light field image. *IEEE Transactions on Image Processing*, 29:1606–1617, 2019. 3

[45] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2349–2358, 2017. 3

[46] 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. 3, 6, 7