Semi-Supervised Semantic Segmentation for Light Field Images Using Disparity Information

Shansi Zhang, *Graduate Student Member, IEEE*, Yaping Zhao, *Graduate Student Member, IEEE*, and Edmund Y. Lam, *Fellow, IEEE*

***Abstract*— Light field (LF) images enable numerous applications due to their ability to capture information for multiple views. Semantic segmentation is an essential task for LF scene understanding. However, existing supervised methods heavily rely on a large number of pixel-wise annotations. To relieve this problem, we propose a semi-supervised LF semantic segmentation method that requires only a small subset of labeled data and harnesses the LF disparity information. First, we design an unsupervised disparity estimation network, which can determine the disparity map for every view. With the estimated disparity maps, we generate pseudo-labels along with their weight maps for the peripheral views when only the labels of central views are available. We then merge the predictions from multiple views to obtain more reliable pseudo-labels for unlabeled data, and introduce a disparity-semantics consistency loss to enforce structure similarity. Moreover, we develop a comprehensive contrastive learning scheme that includes a pixel-level strategy to enhance feature representations and an object-level strategy to improve segmentation for individual objects. Our method demonstrates state-of-the-art performance on the benchmark LF semantic segmentation dataset under a variety of training settings and achieves comparable performance to supervised methods when trained under 1/2 protocol.**

***Index Terms*— Light field (LF), semi-supervised semantic segmentation, disparity map, contrastive learning.**

I. Introduction

SEMANTIC segmentation, an essential task in computer vision, aims to assign a semantic class to each pixel within an image, facilitating a comprehensive understanding and analysis of a scene. It plays an important role in numerous applications, such as autonomous driving, robot navigation, and medical image analysis. Nowadays, the advancements in deep learning techniques have significantly improved the performance and efficiency of semantic segmentation. As a result, a variety of learning-based methods have been developed to address semantic segmentation for different types of input data, including single RGB images [1], [2], RGB-D images [3], [4], and video sequences [5], [6]. Despite these advancements, the field of semantic segmentation for light field (LF) images still remains relatively unexplored. LF images, which capture both intensities and directions of the light rays [7], offer a wealth of information for multiple views. This unique characteristic provides rich geometric information about the scene and enables many applications [8], such as post-capture refocusing [9], depth estimation [10], [11], [12], 3D scene reconstruction [13], and salient object detection [14], [15].

Existing methods for LF semantic segmentation [16], [17], [18] all employ supervised learning approach, and thus inevitably suffer from a heavy demand on per-view pixel-wise annotations. To relieve this labor-intensive effort, a semi-supervised method, which requires only a small subset of labeled data along with a larger set of unlabeled data, is highly desirable. Different from conventional images, LF images present an opportunity for unsupervised disparity estimation through the matching of correspondences among different views, which is typically trained by the photo-consistency loss. The LF disparity information could be highly valuable in facilitating LF semantic segmentation. However, most existing LF disparity estimation methods [11], [19], [20], [21], [22], [23] can only output the disparity map of the central view, which restricts their potential in assisting semantic segmentation. To overcome this limitation, we develop an unsupervised disparity estimation network that can estimate disparity map for every view. We then leverage the disparity information to enhance the performance of semantic segmentation in scenarios with limited labeled data.

To the best of our knowledge, this is the first attempt at semi-supervised semantic segmentation for LF images, specifically harnessing the inherent disparity information available in LF images to assist in the training of semantic segmentation when labeled data is insufficient. We propose to leverage disparity information in the following approaches: 1) The disparity map of each peripheral (non-central) view is used to generate its pseudo-label and weight map when only the label of the central view is available, thereby enhancing data usage efficiency. 2) The predictions of multiple views from a teacher network are merged in accordance with the estimated disparity map to generate more reliable pseudo-labels for unlabeled data. 3) The disparity map serves as a structure reference for the predicted probability map of unlabeled data. 4) Feature alignment within the segmentation network is facilitated by using the disparity map when integrating features

Manuscript received 28 July 2023; revised 20 May 2024 and 31 July 2024; accepted 31 July 2024. Date of publication 16 August 2024; date of current version 22 August 2024. This work was supported in part by the Research Grants Council of Hong Kong under Grant GRF 17201822; and in part by ACCESS—AI Chip Center for Emerging Smart Systems, Hong Kong, SAR. The associate editor coordinating the review of this article and approving it for publication was Prof. Hichem Sahbi. *(Corresponding author: Edmund Y. Lam.)*

The authors are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, SAR, China (e-mail: shansi@connect.hku.hk; zhaoyp@connect.hku.hk; elam@eee.hku.hk).

Digital Object Identifier 10.1109/TIP.2024.3441930

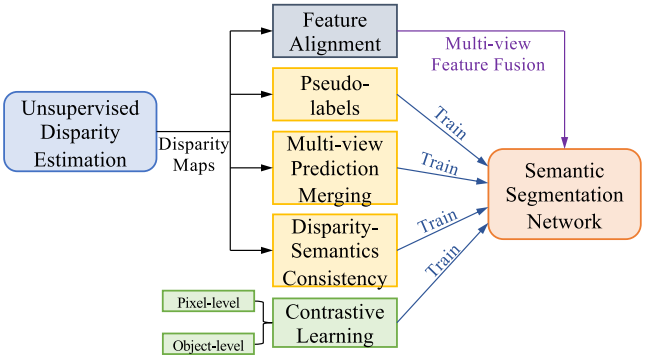


Fig. 1. The overall structure of our semi-supervised semantic segmentation method. The estimated disparity maps from the unsupervised disparity estimation network facilitate feature alignment for multi-view feature fusion within the segmentation network. Moreover, the disparity maps are used to generate pseudo-labels for peripheral views, merge multi-view predictions, and provide a disparity-semantics consistency constraint for unlabeled data to enhance the training of the segmentation network. In addition, contrastive learning at both the pixel level and object level is employed during training to further enhance the segmentation performance.

from multiple views. Additionally, we introduce contrastive learning at both pixel and object levels by collecting high-quality pixel-level and object-level features. In contrast to the existing methods [24], [25], [26], [27] that employ contrastive learning only at the pixel level, our contrastive learning scheme goes beyond and incorporates an object-level strategy, which imposes constraints on each semantic class from a global perspective to further improve the segmentation performance. Fig. 1 illustrates the overall structure of our semi-supervised semantic segmentation method. We evaluate our method on the challenging benchmarks for LF semantic segmentation: UrbanLF-Real and UrbanLF-Syn datasets [17]. Our main contributions are summarized as follows:

- We develop an unsupervised disparity estimation network capable of estimating the disparity map for each view. Leveraging this disparity information, we generate pseudo-labels with their pixel-wise weights for unlabeled views, merge the predictions from multiple views and introduce a disparity-semantics consistency loss, thereby facilitating semantic segmentation learning.
- We propose a comprehensive contrastive learning scheme that incorporates a pixel-level strategy to enhance feature representations and an object-level strategy to improve object segmentation.
- The experimental results demonstrate that our method outperforms other semi-supervised methods across various training settings and achieves comparable performance to supervised methods under 1/2 training protocol.

II. Related Work

A. Semi-Supervised Semantic Segmentation

Most existing semi-supervised semantic segmentation methods were developed for single images. Hung et al. [28] employed the Generative Adversarial Network (GAN), where a discriminator was trained to output confidence maps for the pseudo-labels of unlabeled data. Mittal et al. [29] proposed a two-branch framework, which combines a GAN-based branch to enhance segmentation performance and a semi-supervised classification branch to remove false-positive predictions from the segmentation map. There are some methods focusing on consistency regularization. For example, Chen et al. [30] proposed a cross-pseudo supervision strategy, where two segmentation networks were trained, with each network providing pseudo-labels to guide the training of the other network. Zhong et al. [27] explored consistency in the label space by enforcing the predicted labels invariant to augmentations.

The Mean Teacher method [31] is commonly used in semi-supervised semantic segmentation. Liu et al. [32] introduced a new auxiliary teacher and generated pseudo-labels using an ensemble of two teachers. In addition, contrastive learning is often employed to enhance feature representations. Some methods [25], [26], [27] used the InfoNCE loss [33]. In particular, Wang et al. [26] proposed to utilize unreliable pseudo-labels and construct a memory bank to store negative samples. Alonso et al. [24] proposed a scheme that uses only positive samples. It incorporates a memory bank to store high-quality feature vectors from labeled data and a class-wise attention module to determine the weight for each vector. Its contrastive loss enforces similar pixel-level feature representations within each class. However, the above-mentioned contrastive learning schemes all operate at the pixel level, disregarding the global segmentation performance of individual classes. In contrast, our work extends the contrastive learning to the object level to further improve object segmentation.

B. Light Field Semantic Segmentation

The earlier studies on LF segmentation primarily focused on pixel grouping without explicitly incorporating semantic information. For instance, Wanner et al. [34] developed a variational framework to assign labels consistently for all the views. Hog et al. [35] proposed a ray-based graph structure for interactive LF segmentation, leveraging Markov random field-based optimization for improved efficiency. Lv et al. [36] presented a LF-hypergraph representation using super-pixels and developed a segmentation algorithm based on graph-cut optimization.

Regarding LF semantic segmentation, Jia et al. [16] created a dataset and designed a convolutional neural network (CNN) with spatial-angular feature extraction to achieve semantic segmentation. However, a drawback of their dataset is that each LF image contain only a small number of foreground objects, which restricts its generalization and applicability in highly complex real-world scenes. Sheng et al. [17] developed a more challenging dataset comprising 14 semantic classes, addressing the limitation of the previous dataset. Additionally, they proposed two baseline networks, namely PSPNet-LF and OCR-LF, which include an additional branch dedicated to extracting features from the stacked views in different directions. Yang et al. [37] proposed the LFRSNet, which incorporates contextual feature extraction for occluded region perception and geometric feature extraction for segmentation edge refinement. Cong et al. [18] developed the LF-IENet, which combines implicit feature integration with self-attention and cross-attention and explicit feature propagation using disparity maps. The above-mentioned methods all adopt supervised training and primarily focus on the design of complex

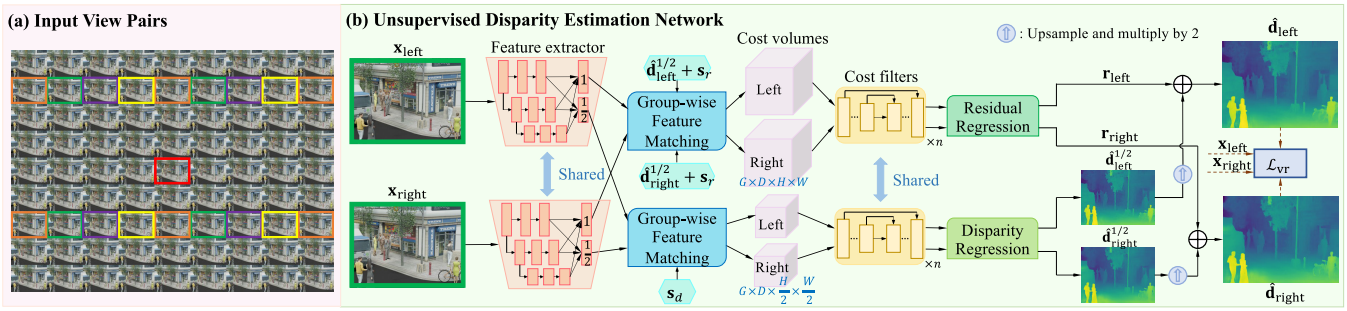


Fig. 2. (a) Input view pairs. The two views for stereo matching are from the same row and framed by the same color, ensuring that the disparity maps for every view can be obtained by our framework. (b) Unsupervised disparity estimation network. The two views $\mathbf{x}_{\text{left}}$ and $\mathbf{x}_{\text{right}}$ are fed to a shared feature extractor. The extracted 1/2-scale feature maps and full-scale feature maps of the two views are matched using the initial disparity samples from $\mathbf{s}_d$ and the residual samples from $\mathbf{s}_r$ to obtain the 1/2-scale initial disparity maps $\hat{\mathbf{d}}^{1/2}_{\text{left}}$ and $\hat{\mathbf{d}}^{1/2}_{\text{right}}$, and the residual maps $\mathbf{r}_{\text{left}}$ and $\mathbf{r}_{\text{right}}$. $\hat{\mathbf{d}}^{1/2}_{\text{left}}$ and $\hat{\mathbf{d}}^{1/2}_{\text{right}}$ are upsampled and added to $\mathbf{r}_{\text{left}}$ and $\mathbf{r}_{\text{right}}$ to obtain the final disparity maps $\hat{\mathbf{d}}_{\text{left}}$ and $\hat{\mathbf{d}}_{\text{right}}$. The view reconstruction loss $\mathcal{L}_{\text{vr}}$ is employed for unsupervised training.

network architecture to achieve semantic segmentation for the central view of each LF image. Unlike them, our objective is to develop effective training strategies for semi-supervised LF semantic segmentation. The key focus is on leveraging LF disparity information to enhance data usage efficiency and generate more reliable pseudo-labels, thus reliving the dependence on a large number of pixel-wise annotations during training.

III. Proposed Method

In this section, we first describe our unsupervised disparity estimation method in Sec. III-A. Following that, we present our semi-supervised LF semantic segmentation method that utilizes disparity information in Sec. III-B. Finally, we introduce our contrastive learning scheme in Sec. III-C.

A. Unsupervised Disparity Estimation

A LF image is represented as $\mathbf{X} \in \mathbb{R}^{U\times V\times 3\times H\times W}$, where $U \times V$ corresponds to the angular resolution, and $H \times W$ represents the spatial resolution. The angular position is indexed by (u, v), while the spatial position is indexed by (h, w). For convenience, a single view at an arbitrary angular position is denoted as $\mathbf{x} = \mathbf{X}(u, v) \in \mathbb{R}^{3\times H\times W}$. According to the LF geometry, the relationship between the view at (u_1, v_1) with disparity map $\mathbf{d}_1$ and the view at (u_2, v_2) is given by

$$\mathbf{X}(u_1, v_1, h, w) = \mathbf{X}\big(u_2, v_2, h + \mathbf{d}_1(h, w)\cdot(u_1 - u_2), w + \mathbf{d}_1(h, w)\cdot(v_1 - v_2)\big). \quad (1)$$

Note that the warping operation is performed according to this relationship.

Fig. 2 illustrates the architecture of our unsupervised disparity estimation network. In the case of a 9×9 LF image, two views that are from the same row and framed by the same color are paired and input to the network. The left view $\mathbf{x}_{\text{left}}$ and the right view $\mathbf{x}_{\text{right}}$ are first fed to a shared feature extractor. The feature extractor consists of three branches to extract full-scale, 1/2-scale and 1/4-scale features using residual blocks. These features at different scales are fused by rescaling and addition to integrate multi-scale information. The resulting fused features at the 1/2 scale and full scale are then used for subsequent feature matching. The 1/2-scale feature maps extracted from the two views are matched according to a set of initial disparity samples (represented as a vector $\mathbf{s}_d$) to derive the 1/2-scale cost volumes. The left cost volume is constructed by warping the right feature map using each sample from $\mathbf{s}_d$ to align with the left feature map. Group-wise correlation [38] is computed to measure the feature similarities within each group. The right cost volume is constructed in a similar way. However, in this case, the left feature map is warped to align with the right feature map. The two cost volumes, each with a shape of $G \times D \times \frac{H}{2} \times \frac{W}{2}$ (where G is the number of groups and D is the number of disparity samples), are fed to several cascaded cost filters separately. Each cost filter comprises 3D convolutional layers and follows a U-shape structure, where the cost volumes are first downsampled and subsequently upsampled to recover the original spatial resolution. Skip connections are incorporated at different scales to facilitate information flow. Following that, disparity regression [39] is employed to obtain the initial 1/2-scale disparity maps $\hat{\mathbf{d}}^{1/2}_{\text{left}}$ and $\hat{\mathbf{d}}^{1/2}_{\text{right}}$ for the two input views.

The initial disparity maps are not sufficiently accurate due to the coarse disparity sampling at a small scale. To address this issue, a refinement branch with residual samples (represented as a vector $\mathbf{s}_r$) at a finer scale is introduced to improve the accuracy. The full-scale feature maps extracted from the two views are matched using the addition of each residual sample from $\mathbf{s}_r$ and the initial disparity maps $\hat{\mathbf{d}}^{1/2}_{\text{left}}$ and $\hat{\mathbf{d}}^{1/2}_{\text{right}}$ to construct the full-scale cost volumes, each with a shape of $G \times D \times H \times W$. Then, the two cost volumes are fed to the cost filters shared with the coarse branch, followed by a residual regression to obtain the residual maps $\mathbf{r}_{\text{left}}$ and $\mathbf{r}_{\text{right}}$. Finally, the refined disparity maps for the two views are obtained by

$$\hat{\mathbf{d}}_{\text{left}} = 2 \times \mathcal{U}(\hat{\mathbf{d}}^{1/2}_{\text{left}}) + \mathbf{r}_{\text{left}}, \quad (2)$$

$$\hat{\mathbf{d}}_{\text{right}} = 2 \times \mathcal{U}(\hat{\mathbf{d}}^{1/2}_{\text{right}}) + \mathbf{r}_{\text{right}}, \quad (3)$$

where $\mathcal{U}(\cdot)$ denotes upsampling operation. Note that $\hat{\mathbf{d}}_{\text{left}}$ and $\hat{\mathbf{d}}_{\text{right}}$ should be divided by the baseline distance between the two input views, which is 4 in this case, to obtain the disparity maps for the adjacent views. With this input strategy, the disparity maps for each view can be obtained.

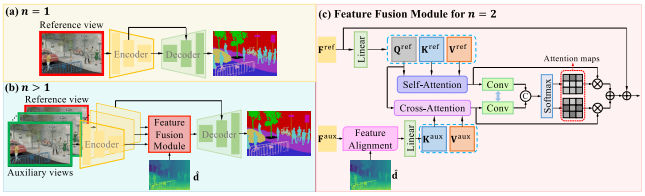


Fig. 3. (a) Semantic segmentation network for $n = 1$. (b) Semantic segmentation network for $n > 1$. (c) Feature fusion module for $n = 2$.

The unsupervised disparity estimation is trained by the view reconstruction loss, expressed as

$$\mathcal{L}_{\text{vr}} = \|\mathbf{x}_{\text{left}} - \mathcal{W}(\mathbf{x}_{\text{right}}; \hat{\mathbf{d}}_{\text{left}})\|_1 + \|\mathbf{x}_{\text{right}} - \mathcal{W}(\mathbf{x}_{\text{left}}; \hat{\mathbf{d}}_{\text{right}})\|_1 \tag{4}$$

where $\mathcal{W}(\mathbf{x}_{\text{right}}; \hat{\mathbf{d}}_{\text{left}})$ denotes warping $\mathbf{x}_{\text{right}}$ to align with $\mathbf{x}_{\text{left}}$ using $\hat{\mathbf{d}}_{\text{left}}$, and $\|\cdot\|_1$ is the ℓ_1 norm. This loss term is also applied to $\hat{\mathbf{d}}^{1/2}_{\text{left}}$ and $\hat{\mathbf{d}}^{1/2}_{\text{right}}$.

To improve the smoothness of estimated disparity maps, we apply a structure-aware smoothness loss [40], [41], [42] to $\hat{\mathbf{d}}_{\text{left}}$ and $\hat{\mathbf{d}}_{\text{right}}$, with

$$\mathcal{L}_{\text{sm}} = \left\| |\nabla \hat{\mathbf{d}}_{\text{left}}| \times \exp(-\eta |\nabla \mathbf{x}_{\text{left}}|) \right\|_1 + \left\| |\nabla \hat{\mathbf{d}}_{\text{right}}| \times \exp(-\eta |\nabla \mathbf{x}_{\text{right}}|) \right\|_1, \tag{5}$$

where ∇ is a gradient operator and η is a hyperparameter to adjust the degree of structure preservation. Thus, the full loss for training the disparity estimation network is

$$\mathcal{L}_{\text{disp}} = \mathcal{L}_{\text{vr}} + \lambda_{\text{sm}} \mathcal{L}_{\text{sm}}, \tag{6}$$

where λ_{sm} is the coefficient of the smoothness loss.

B. Semi-Supervised LF Semantic Segmentation Using Disparity Information

Given a labeled dataset $\mathcal{X}^l = \{(\mathbf{X}^l, \mathbf{y}_c)\}$, where $\mathbf{y}_c$ denotes the segmentation label of the central view, and an unlabeled dataset $\mathcal{X}^u = \{\mathbf{X}^u\}$, our objective is to train a semantic segmentation model for LF images using both labeled and unlabeled data.

1) Semantic Segmentation Network: The current semantic segmentation networks for LF images [16], [17], [18] are limited to segmenting only the central view. In order to achieve segmentation for every view, we can adapt a semantic segmentation network designed for conventional images by treating each individual view as a separate input, as depicted in Fig. 3(a), where the encoder can be any pretrained backbone. Suppose that n is the total number of views input to the network. In this case, $n = 1$.

Considering that a LF image comprises multiple views, it is possible to fuse the multi-view features to obtain richer semantic information, as illustrated in Fig. 3(b). When $n > 1$, one reference view (the view to be segmented) and $n-1$ auxiliary views are fed to a shared encoder. The extracted features from the reference view and the auxiliary views are combined using a feature fusion module. The fused features are then passed through a decoder to obtain the segmentation map for the reference view.

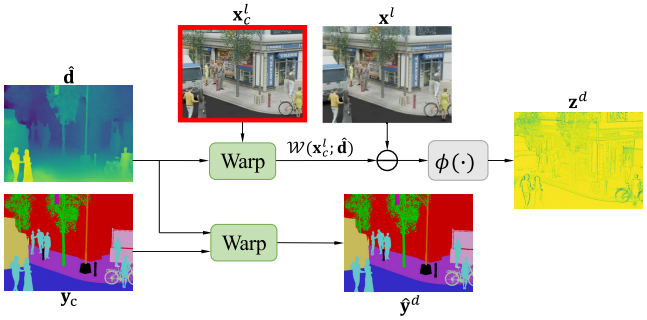


Fig. 4. Pseudo-label generated by disparity. The central label $\mathbf{y}_c$ is warped according to the estimated disparity map $\hat{\mathbf{d}}$ of the view $\mathbf{x}^l$ to obtain its pseudo-label $\hat{\mathbf{y}}^d$. The central view $\mathbf{x}_c^l$ is also warped using $\hat{\mathbf{d}}$, and the image warping error map is passed through a function $\phi(\cdot)$ to obtain the weight map $\mathbf{z}^d$ for the pseudo-label $\hat{\mathbf{y}}^d$.

The feature fusion module is shown in Fig. 3(c), where we take $n = 2$ as an example to illustrate its structure. To align the feature map of the auxiliary view $\mathbf{F}^{\text{aux}}$ with the feature map of the reference view $\mathbf{F}^{\text{ref}}$, a warping operation on $\mathbf{F}^{\text{aux}}$ is performed using the estimated disparity map of the reference view. Then, the query $\mathbf{Q}^{\text{ref}}$, key $\mathbf{K}^{\text{ref}}$, and value $\mathbf{V}^{\text{ref}}$ of the reference view, as well as the key $\mathbf{K}^{\text{aux}}$ and value $\mathbf{V}^{\text{aux}}$ of the auxiliary view are derived through linear layers. Self-attention within $\mathbf{F}^{\text{ref}}$ is computed with $\text{softmax}(\frac{\mathbf{Q}^{\text{ref}}(\mathbf{K}^{\text{ref}})^{\text{T}}}{\tau})\mathbf{V}^{\text{ref}}$ (τ is a learnable scale factor), and cross-attention between $\mathbf{F}^{\text{ref}}$ and $\mathbf{F}^{\text{aux}}$ is computed with $\text{softmax}(\frac{\mathbf{Q}^{\text{ref}}(\mathbf{K}^{\text{aux}})^{\text{T}}}{\tau})\mathbf{V}^{\text{aux}}$. The feature maps after self-attention and cross-attention are further processed by convolution layers and a softmax function to obtain their attention maps for pixel-wise fusion.

2) Supervised Loss: The segmentation network, parameterized by θ, predicts the pixel-wise probability distribution $P_\theta(\mathbf{x}) \in \mathbb{R}^{K \times H \times W}$ for each view $\mathbf{x}$, where K is the number of semantic classes, with each class indexed by k ranging from 1 to K. When training the network using the central view $\mathbf{x}_c^l$ of the labeled data along with its one-hot encoded label $\tilde{\mathbf{y}}_c$, we employ the cross-entropy (CE) loss, expressed as

$$\mathcal{L}_{\text{ce}} = -\frac{1}{HW} \sum_{k,h,w} \tilde{\mathbf{y}}_{c_{(k,h,w)}} \log P_\theta(\mathbf{x}_c^l)_{(k,h,w)}, \tag{7}$$

and the multi-class dice loss [43], with

$$\mathcal{L}_{\text{dice}} = 1 - \frac{1}{K} \sum_{k=1}^{K} \frac{2 \sum_{h,w} P_\theta(\mathbf{x}_c^l)[k]_{(h,w)} \times \tilde{\mathbf{y}}_c[k]_{(h,w)}}{\sum_{h,w} P_\theta(\mathbf{x}_c^l)[k]_{(h,w)} + \sum_{h,w} \tilde{\mathbf{y}}_c[k]_{(h,w)}}, \tag{8}$$

where $P_\theta(\mathbf{x}_c^l)[k]$ and $\tilde{\mathbf{y}}_c[k]$ denote the prediction and the label of class k, respectively. Thus, the supervised loss is written as

$$\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{dice}}. \tag{9}$$

3) Pseudo-Labels Generated by Disparity: As only the central view of each LF image has segmentation label, there might not be enough annotations available in the semi-supervised setting. With the disparity estimation network introduced in Sec. III-A, the pseudo-labels for any other view can be generated by using its estimated disparity map and the segmentation label of the central view.

Fig. 4 provides an example of generating the pseudo-label and its pixel-wise weight for a peripheral view $\mathbf{x}^l$. The central

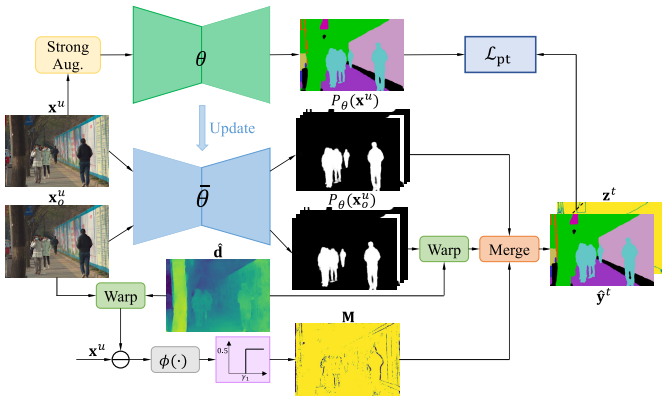


Fig. 5. Mean Teacher structure. The pseudo-label for view $\mathbf{x}^u$ is obtained from the teacher network $\bar{\theta}$ by merging the probability maps of $\mathbf{x}^u$ and an additional view $\mathbf{x}_o^u$. The probability map of $\mathbf{x}_o^u$ is warped to align with the probability map of $\mathbf{x}^u$ using the estimated disparity map $\hat{\mathbf{d}}$ of $\mathbf{x}^u$. The merging mask $\mathbf{M}$ is derived from the corresponding image warping error.

label $\mathbf{y}_c$ is warped using the estimated disparity map $\hat{\mathbf{d}}$ of $\mathbf{x}^l$ and the view position distance to obtain the pseudo-label $\hat{\mathbf{y}}^d = \mathcal{W}(\mathbf{y}_c; \hat{\mathbf{d}})$ for $\mathbf{x}^l$ (superscript d denotes that the pseudo-label is generated by disparity). Label warping is similar to image warping, but it employs nearest neighbor interpolation. For the regions with occlusion or incorrect disparity estimation that results in significant warping errors, the corresponding pseudo-label pixels are not accurate, and therefore smaller weights should be assigned to them. To determine the weight for each pixel, the central view $\mathbf{x}_c^l$ is warped to align with $\mathbf{x}^l$ using $\hat{\mathbf{d}}$. The weight map $\mathbf{z}^d$ is derived according to the image warping error using a function $\phi(\cdot)$, with

$$\mathbf{z}^d = \phi(|\mathbf{x}^l - \mathcal{W}(\mathbf{x}_c^l; \hat{\mathbf{d}})|) = \exp\Big(-\frac{|\mathbf{x}^l - \mathcal{W}(\mathbf{x}_c^l; \hat{\mathbf{d}})|}{\sigma}\Big), \quad (10)$$

where σ is to adjust the decay rate. The weight assigned to each pixel falls within the range of 0 to 1 and decreases as the image warping error increases. Typically, the regions with smaller weights are commonly observed near the object boundaries where the occlusion occurs. This can be visualized in the weight map $\mathbf{z}^d$, as shown in Fig. 4.

Once the central label is provided, the pseudo-labels and their corresponding weight maps for other views can be obtained. These pseudo-labels provide additional guidance for the segmentation network in the absence of ground truth annotations. In this way, multiple training pairs are available within each LF image, which greatly improves data usage efficiency. By combining various data augmentations, more abundant training data can be utilized by the segmentation network. The weighted CE loss $\ell_{\text{wce}}(\cdot)$ is applied when using these pseudo-labels, with

$$\begin{aligned}\mathcal{L}_{\text{pd}} &= \ell_{\text{wce}}\big(P_\theta(\mathbf{x}^l), \tilde{\mathbf{y}}^d, \mathbf{z}^d\big) \\ &= -\frac{1}{HW}\sum_{h,w}\mathbf{z}^d_{(h,w)}\sum_k \tilde{\mathbf{y}}^d_{(k,h,w)} \log P_\theta(\mathbf{x}^l)_{(k,h,w)} \end{aligned} \quad (11)$$

where $\tilde{\mathbf{y}}^d$ is the one-hot representation of $\hat{\mathbf{y}}^d$.

4) Pseudo-Labels Generated by Mean Teacher: We employ the Mean Teacher method [31] to generate pseudo-labels for unlabeled data in $\mathcal{X}^u$, as depicted in Fig. 5. The parameters of

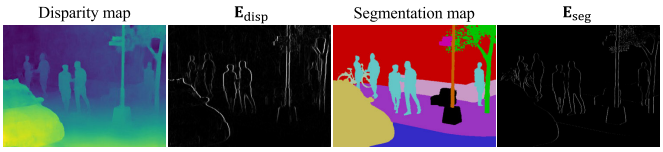


Fig. 6. The estimated disparity map along with its edge map $\mathbf{E}_{\text{disp}}$, and the predicted segmentation map along with its edge map $\mathbf{E}_{\text{seg}}$.

the teacher network, denoted as $\bar{\theta}$, are updated using an exponential moving average of the student network's parameters, with

$$\bar{\theta} = \mu\bar{\theta} + (1-\mu)\theta, \quad (12)$$

where $\mu \in (0, 1)$ is a smoothing coefficient.

For each view $\mathbf{x}^u$, its probability map output by the teacher network is represented as $P_{\bar{\theta}}(\mathbf{x}^u)$. During training, the predictions of different views for the same objects may have differences, especially if the network is not robust enough. To generate more reliable pseudo-labels and weights by considering multi-view predictions, the probability map of an additional view $\mathbf{x}_o^u$ is also predicted, written as $P_{\bar{\theta}}(\mathbf{x}_o^u)$. $P_{\bar{\theta}}(\mathbf{x}_o^u)$ is warped using the estimated disparity map $\hat{\mathbf{d}}$ of $\mathbf{x}^u$ to align with $P_{\bar{\theta}}(\mathbf{x}^u)$, expressed as $\mathcal{W}\big(P_{\bar{\theta}}(\mathbf{x}_o^u); \hat{\mathbf{d}}\big)$. The merging mask $\mathbf{M}$ is obtained using the function $\phi(\cdot)$ in Eq. 10, with

$$\mathbf{M}_{(h,w)} = \begin{cases} 0.5, & \phi(|\mathbf{x}^u - \mathcal{W}(\mathbf{x}_o^u; \hat{\mathbf{d}})|) > \gamma_1 \\ 0, & \text{otherwise} \end{cases}, \quad (13)$$

where γ_1 is a threshold to eliminate the influence of occlusion and inaccurate disparity estimation. Then, the rectified probability map for $\mathbf{x}^u$ is derived by merging $P_{\bar{\theta}}(\mathbf{x}^u)$ and $\mathcal{W}\big(P_{\bar{\theta}}(\mathbf{x}_o^u); \hat{\mathbf{d}}\big)$, with

$$P'_{\bar{\theta}}(\mathbf{x}^u) = (\mathbf{1} - \mathbf{M}) \times P_{\bar{\theta}}(\mathbf{x}^u) + \mathbf{M} \times \mathcal{W}\big(P_{\bar{\theta}}(\mathbf{x}_o^u); \hat{\mathbf{d}}\big). \quad (14)$$

The pseudo-label $\hat{\mathbf{y}}^t$ (superscript t denotes that the pseudo-label is generated by the teacher network) for view $\mathbf{x}^u$ is obtained by selecting the class with the maximum probability from $P'_{\bar{\theta}}(\mathbf{x}^u)$ at each pixel, and its weight map $\mathbf{z}^t$ is determined by

$$\mathbf{z}^t = \max_k \big(P'_{\bar{\theta}}(\mathbf{x}^u)_{(k,h,w)}\big) \times \mathbb{1}\Big(\max_k \big(P'_{\bar{\theta}}(\mathbf{x}^u)_{(k,h,w)}\big) > \gamma_2\Big), \quad (15)$$

where $\mathbb{1}(\cdot)$ is an indicator function, and γ_2 is a threshold to remove the pseudo-label pixels with low confidence.

For the pixels with low confidence ($\mathbb{1}(\mathbf{z}^t = 0)$), we apply an entropy loss $\mathcal{L}_{\text{en}}$ to minimize their entropies,with

$$\mathcal{L}_{\text{en}} = \frac{1}{N_{\text{en}}}\sum_{k,h,w} P_\theta\big(\mathcal{A}(\mathbf{x}^u)\big) \log P_\theta\big(\mathcal{A}(\mathbf{x}^u)\big)_{(k,h,w)} \mathbb{1}(\mathbf{z}^t = 0), \quad (16)$$

where $\mathcal{A}(\cdot)$ represents strong data augmentations [44] and N_{en} is the number of pixels included to calculate the entropy loss. Thus, the total loss using pseudo-labels generated by the teacher network is the combination of the weighted CE loss and the entropy loss, expressed as

$$\mathcal{L}_{\text{pt}} = \ell_{\text{wce}}\Big(P_\theta\big(\mathcal{A}(\mathbf{x}^u)\big), \tilde{\mathbf{y}}^t, \mathbf{z}^t\Big) + \lambda_{\text{en}}\mathcal{L}_{\text{en}} \quad (17)$$

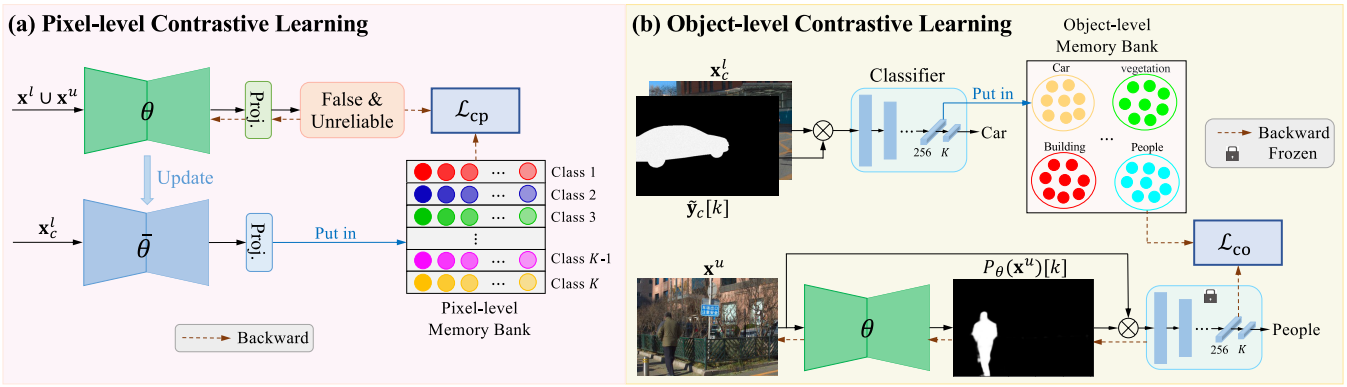


Fig. 7. (a) Pixel-level contrastive learning. A pixel-level memory bank is constructed using the high-quality feature vectors of each class from the teacher network. The loss $\mathcal{L}_{\text{cp}}$ enforces the feature vectors from the student network that lead to false and unreliable predictions to be similar to those in the memory bank. (b) Object-level contrastive learning. A classifier is trained using the labeled data by taking $\mathbf{x}_c^l \times \tilde{\mathbf{y}}_c[k]$ as input. The object-level feature vectors of each class corresponding to correct classifications with the largest confidences are selected to construct the object-level memory bank. The loss $\mathcal{L}_{\text{co}}$ enforces the feature vectors of the segmented objects $\mathbf{x}^u \times P_\theta(\mathbf{x}^u)[k]$ to be similar to the feature vectors of class k in the memory bank.

where $\tilde{\mathbf{y}}^t$ is the one-hot representation of $\hat{\mathbf{y}}^t$, and λ_{en} is the coefficient of the entropy loss.

5) Disparity-Semantics Consistency: The disparity map and the semantic segmentation map exhibit similar boundary structures. Given this, the estimated disparity map can be used as a structure reference for the segmentation of unlabeled data. Thus, we propose a disparity-semantics consistency loss to constrain the edges extracted from the disparity map and the segmentation map to be similar. The normalized disparity edge map $\mathbf{E}_{\text{disp}}$ is derived by

$$\mathbf{E}_{\text{disp}} = \min\left(\frac{\nabla\hat{\mathbf{d}} - \min(\nabla\hat{\mathbf{d}})}{\max(\nabla\hat{\mathbf{d}}) - \min(\nabla\hat{\mathbf{d}})} \times 2, \mathbf{1}\right), \tag{18}$$

where $\nabla\hat{\mathbf{d}} = |\hat{\mathbf{d}} - \mathcal{T}_{\mapsto}(\hat{\mathbf{d}})|$, with $\mathcal{T}_{\mapsto}(\cdot)$ denoting shifting the input by one pixel along the horizontal direction, and multiplying by 2 aims to enhance the edges of objects in the distance. Moreover, we utilize the edge mask extracted from the input view $\mathbf{x}^u$ as a reference for identifying edges and removing non-edge pixels that arise due to inaccurate disparity estimation.

The segmentation edge map $\mathbf{E}_{\text{seg}}$ is obtained by calculating cosine similarity between the predicted probability distributions of adjacent pixels, with

$$\mathbf{E}_{\text{seg}} = \mathbf{1} - \text{sim}\Big(P_\theta(\mathbf{x}^u), \mathcal{T}_{\mapsto}\big(P_\theta(\mathbf{x}^u)\big)\Big), \tag{19}$$

where $\text{sim}(\cdot)$ represents calculating cosine similarity between two vectors, with $\text{sim}(\mathbf{p}_1, \mathbf{p}_2) = \frac{\langle \mathbf{p}_1, \mathbf{p}_2 \rangle}{\|\mathbf{p}_1\|_2 \|\mathbf{p}_2\|_2}$.

Fig. 6 provides an example of extracted edge maps from the estimated disparity map and the predicted segmentation map. Note that the edges in the vertical direction are also derived, but we omit the procedures for simplicity. The disparity-semantics consistency loss is written as

$$\mathcal{L}_{\text{cs}} = \|\mathbf{E}_{\text{disp}} - \mathbf{E}_{\text{seg}}\|_1, \tag{20}$$

which aims to enforce similar probability distributions within each object and distinct probability distributions near the object boundaries. Hence, it can also be interpreted as a semantics smoothness loss.

C. Contrastive Learning

Instead of simply employing contrastive learning at the pixel-level, we propose a comprehensive contrastive learning scheme that incorporates an improved pixel-level strategy and a novel object-level strategy.

1) Pixel-Level Contrastive Learning: We follow the positive-only scheme [24] to construct a memory bank using high-quality pixel-level features of each class obtained from the teacher network, as shown in Fig. 7(a). A projection layer is introduced to both the student and teacher networks to output the feature vectors. During each training step, the memory bank is updated using the top-K feature vectors from the labeled data that lead to correct predictions with the highest confidence. In order to focus more on challenging pixels and reduce memory consumption during training, we pick out the vectors that correspond to false predictions or correct predictions with low confidence from the student network to calculate loss.

Let's define $\mathcal{F}_\theta^k = \{\mathbf{f}_\theta^k\}$ as a set of feature vectors of class k from the student network associated with false and unreliable predictions, and define $\mathcal{F}^k = \{\mathbf{f}^k\}$ as a set of feature vectors of class k from the memory bank. The objective is to maximize the similarity between the feature vectors from $\mathcal{F}_\theta^k$ and $\mathcal{F}^k$. Cosine similarity is used as the measurement. Thus, the pixel-level contrastive loss is expressed as

$$\mathcal{L}_{\text{cp}} = \frac{1}{K|\mathcal{F}_\theta^k||\mathcal{F}^k|} \sum_{k=1}^{K} \sum_{\mathbf{f}_\theta^k \in \mathcal{F}_\theta^k} \sum_{\mathbf{f}^k \in \mathcal{F}^k} 1 - \text{sim}(\mathbf{f}_\theta^k, \mathbf{f}^k). \tag{21}$$

2) Object-Level Contrastive Learning: The dice loss in Eq. 8 aims to guide the segmentation of each semantic class from a global perspective. However, when dealing with the unlabeled data, using only the weighted CE loss and the pixel-level contrastive loss ignores the global segmentation performance of each class. Therefore, we introduce an object-level contrastive learning strategy to impose global constraints on each semantic class.

We first train an auxiliary classifier using the labeled data to classify the K object classes according to their shapes and features. The input of the classifier is $\mathbf{x}_c^l \times \tilde{\mathbf{y}}_c[k]$, which

contains only the objects of class k. Label smoothing is applied to $\tilde{\mathbf{y}}_c$ in order to approximate the probability predictions of the segmentation network. The trained classifier is utilized to obtain high-quality object-level features from the labeled data. For each class, we select the feature vectors that lead to correct classifications with a confidence greater than a specified threshold to construct an object-level memory bank, as shown in Fig. 7(b).

When training the student network with the unlabeled data, $\mathbf{x}^u \times P_\theta(\mathbf{x}^u)[k]$ is input to the frozen classifier. The feature vector extracted by the classifier is expected to be similar to the feature vectors of class k stored in the memory bank, which enforces that the segmented objects can be correctly classified with high confidence.

Let's define $\mathcal{Q}_\theta^k = \{\mathbf{q}_\theta^k\}$ as a set of feature vectors extracted from $\mathbf{x}^u \times P_\theta(\mathbf{x}^u)[k]$ by the classifier, and define $\mathcal{Q}^k = \{\mathbf{q}^k\}$ as a set of object-level feature vectors of class k from the memory bank. The object-level contrastive loss encourages the feature vectors from $\mathcal{Q}_\theta^k$ to be similar to those in $\mathcal{Q}^k$, with

$$\mathcal{L}_{\text{co}} = \frac{1}{K|\mathcal{Q}_\theta^k||\mathcal{Q}^k|}\sum_{k=1}^{K}\sum_{\mathbf{q}_\theta^k\in\mathcal{Q}_\theta^k}\sum_{\mathbf{q}^k\in\mathcal{Q}^k} 1-\text{sim}(\mathbf{q}_\theta^k,\mathbf{q}^k). \quad (22)$$

With the aforementioned loss terms, the segmentation network is trained using the combined loss, with

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{pd}} + \mathcal{L}_{\text{pt}} + \lambda_{\text{cs}}\mathcal{L}_{\text{cs}} + \lambda_{\text{cp}}\mathcal{L}_{\text{cp}} + \lambda_{\text{co}}\mathcal{L}_{\text{co}}, \quad (23)$$

where λ_{cs}, λ_{cp} and λ_{co} are the coefficients of $\mathcal{L}_{\text{cs}}$, $\mathcal{L}_{\text{cp}}$ and $\mathcal{L}_{\text{co}}$, respectively.

IV. Experiments

A. Dataset

We use the UrbanLF dataset [17], [48], which is currently the only publicly available semantic segmentation dataset for LF images. This dataset focuses on urban scenes with 14 semantic classes and includes two subsets: UrbanLF-Real and UrbanLF-Syn. The UrbanLF-Real subset contains 580 real-world LF images for training and the UrbanLF-Syn subset contains 280 synthetic LF images for training. Each LF image consists of 9×9 sub-aperture images with pixel-wise annotation for the central view. Besides the training set, there are 80 real-world LF images and 40 synthetic LF images with ground-truth labels available for evaluation. They are randomly split into an approximate ratio of $1:2$ for validation and testing, respectively.

B. Implementation Details

Regarding the disparity estimation network, the initial 1/2-scale disparity range was set to $[-1, 3]$ with a sampling interval of 0.2, and the residual range was set to $[-0.2, 0.6]$ with a sampling interval of 0.04. These ranges were determined according to the disparity range observed across the entire dataset. The input views were randomly cropped into 256×256 patches during training. The hyperparameter η in Eq. 5 was set to 50 and the coefficient λ_{sm} in Eq. 6 was set to 0.1. The learning rate was initially set to 5×10^{-4} and decayed by multiplying 0.9 after every 50 epochs. Our disparity estimation network was trained by the Adam optimizer [49] for about 200 epochs.

Regarding the semantic segmentation network, we employed the DeepLabv3+ [1] architecture with ResNet-101 [50] and HRNetV2-W48 [51] backbones as the baseline model. The backbones were pre-trained on the ImageNet dataset [52]. During training, for each labeled LF image, we utilized the central view with its ground-truth label and any other view with its pseudo-label generated by its disparity map at each iteration. The two views were augmented differently to increase diversity. σ in Eq. 10 was set to 0.5. For each unlabeled LF image, we utilized a randomly selected view with its pseudo-label obtained by combining the prediction of any other view at each iteration. The thresholds in Eq. 13 and Eq. 15 were set to $\gamma_1 = 0.9$ and $\gamma_2 = 0.3$, respectively. When using the pseudo-labels for training, strong data augmentations were applied to the views fed to the student network. These augmentations include color jitter that adjusts the brightness, contrast and saturation within a specific range, Gaussian blur and gray-scale transformation. A form of weak data augmentation, namely image flipping, was also applied to the input views during training. We set $\lambda_{\text{en}} = 0.01$ in Eq. 17, and $\lambda_{\text{cs}} = 0.1$, $\lambda_{\text{cp}} = 0.1$, $\lambda_{\text{co}} = 0.1$ in Eq. 23. The effects of the hyperparameters on the overall performance are studied in Sec. IV-D.3. The initial learning rates for the pre-trained backbone and the other modules were set to 1×10^{-4} and 1×10^{-3}, respectively. They were both multiplied by 0.9 every 50 epochs. The segmentation network was trained using the Adam optimizer for about 400 epochs.

The classifier used in the object-level contrastive learning employed the ResNet-50 architecture. It was trained using labeled data under 1/2, 1/4 and 1/8 partition protocols, respectively. Strong and weak data augmentations were applied during training.

C. Comparisons

1) Evaluation on LF Disparity Estimation: We compared our unsupervised disparity estimation method with several other unsupervised methods, including two unsupervised stereo matching methods: PANet [45] and ES3Net [46], as well as three unsupervised LF disparity estimation methods: UnMono [10], UnOcc [11] and UnPlug [47]. PANet and ES3Net employed the same input strategy as our method. However, they output disparity maps only for the left view in a single inference. UnOcc and UnPlug are limited to estimating disparity maps only for the central views. As a result, they are not applicable to our semi-supervised semantic segmentation framework. All these methods were trained on the synthetic LF images without using any ground-truth disparity maps.

Table I lists the quantitative results of various methods on the UrbanLF-Syn test set, along with their number of parameters and FLOPs. The results are reported in terms of mean square error (MSE) and bad pixel ratios (BPR) with thresholds of 0.07, 0.03 and 0.01. We can observe that our method obtains better quantitative results than the other methods. Compared to the LF disparity estimation methods, our method can be

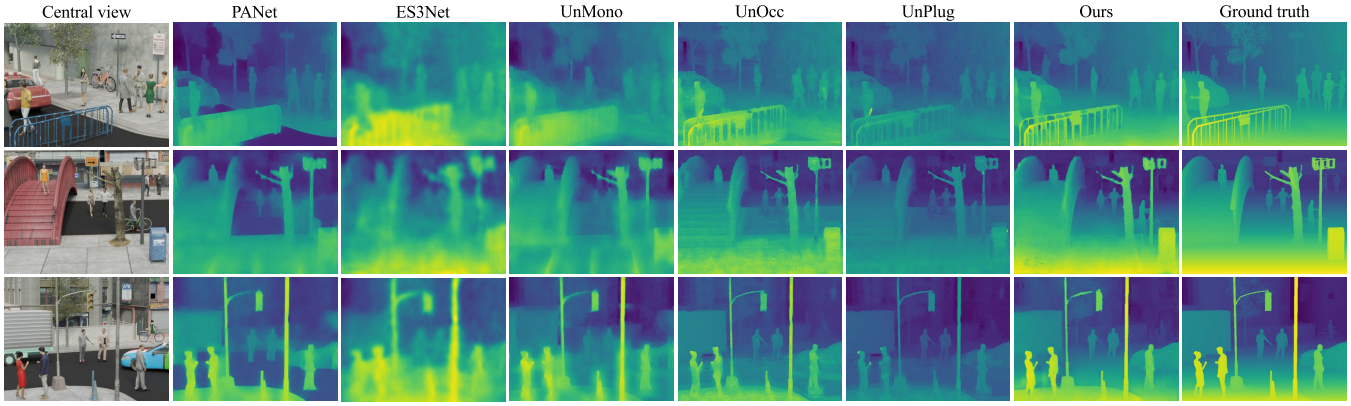


Fig. 8. The estimated disparity maps of different unsupervised methods for the central view of each LF image. Zoom in for best view.

TABLE I

QUANTITATIVE RESULTS OF UNSUPERVISED DISPARITY ESTIMATION METHODS. MSE (×100), BPRS (%), PARAMETERS AND FLOPS ARE REPORTED. **BOLD**: BEST

Method	Parameters	FLOPs	LFUrban-Syn				HCI			
			MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)
PANet [45]	7.82M	33.34G	3.676	43.03	67.89	88.59	2.995	47.19	74.51	91.13
ES3Net [46]	0.023M	0.23G	2.442	49.98	79.17	93.40	3.177	53.35	78.43	92.48
UnMono [10]	1.87M	28.55G	2.552	46.50	70.99	89.62	3.003	34.69	63.95	85.91
UnOcc [11]	1.90M	130.52G	1.039	21.46	42.48	74.59	**1.623**	**18.17**	46.65	80.20
UnPlug [47]	3.17M	623.08G	1.501	24.75	54.52	85.35	2.236	32.99	62.64	86.55
Ours	1.61M	451.85G	**0.536**	**13.63**	**36.29**	**73.04**	1.928	20.19	**43.83**	**78.43**

TABLE II

QUANTITATIVE RESULTS OF SEMI-SUPERVISED SEMANTIC SEGMENTATION METHODS ON THE URBANLF-REAL AND URBANLF-SYN TEST SETS. MIOU (%), MACC (%) AND ACC (%) ARE REPORTED. **BOLD**: BEST

Dataset	Method	1/2 (430)			1/4 (215)			1/8 (108)		
		mIoU	mAcc	Acc	mIoU	mAcc	Acc	mIoU	mAcc	Acc
UrbanLF-Real	CPS [30]	70.27	78.01	88.66	68.50	76.91	86.42	64.12	73.86	83.99
	PL-Contra [24]	71.77	79.05	90.38	70.99	78.54	89.91	66.37	74.97	86.12
	PS-MT [32]	72.76	79.99	89.88	71.38	78.65	89.10	65.47	74.19	83.14
	UPL [26]	72.51	79.27	90.12	71.18	78.47	88.72	66.58	75.54	85.05
	Ours	**75.27**	**81.95**	**91.38**	**74.02**	**80.70**	**90.05**	**70.81**	**78.38**	**88.62**
UrbanLF-Syn	CPS [30]	65.19	75.85	83.94	63.19	75.27	83.81	60.74	72.20	81.72
	PL-Contra [24]	68.01	77.57	85.33	64.80	76.22	84.06	62.57	73.86	82.35
	PS-MT [32]	68.29	78.08	86.85	66.62	77.15	85.73	63.10	75.17	83.72
	UPL [26]	68.54	78.60	86.73	66.10	76.81	84.99	63.94	74.38	82.16
	Ours	**71.40**	**82.30**	**88.62**	**70.12**	**80.62**	**87.65**	**66.52**	**77.65**	**85.89**

trained using any view pair within each LF image, as depicted in Fig. 2(a). This flexibility allows us to overcome limitations imposed by the availability of specific view positions, leading to enhanced data usage efficiency. Moreover, our coarse-to-fine structure facilitates superior performance compared to other stereo matching methods. Fig. 8 provides the visual comparisons, which showcase that our method realizes more accurate disparity estimation while preserving clearer object boundaries.

In addition, we trained and evaluated these methods on the HCI dataset [53]. The average quantitative results of several test scenes are listed in Table I, showing that our method exhibits relatively good performance compared to other unsupervised methods.

2) Comparisons With Semi-Supervised Semantic Segmentation Methods: As we are the first to study semi-supervised semantic segmentation for LF images, we primarily made comparisons with the state-of-the-art semi-supervised semantic segmentation methods for conventional images, including CPS [30], PL-Contra [24], PS-MT [32] and UPL [26]. These methods were trained using labeled central views and unlabeled views with pseudo-labels generated by their respective proposed approaches. They all employed DeepLabv3+ [1] with the ResNet-101 backbone as the segmentation network. Regarding our segmentation network, we set $n = 1$, meaning that only the reference view was taken as input. Therefore, fare comparisons were conducted by ensuring that the views used for training and the network architecture are consistent

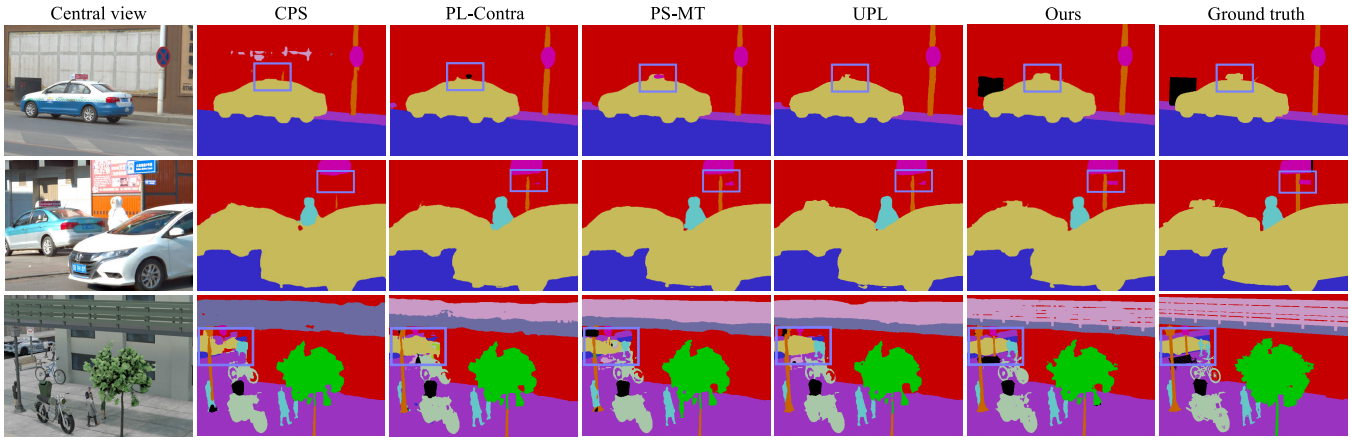


Fig. 9. Segmentation maps obtained by different semi-supervised methods trained under 1/2 protocol. The first two rows display real-world LF images and the third row displays a synthetic LF image. The purple rectangles highlight the advantages of our method over the other methods. Zoom in for best view.

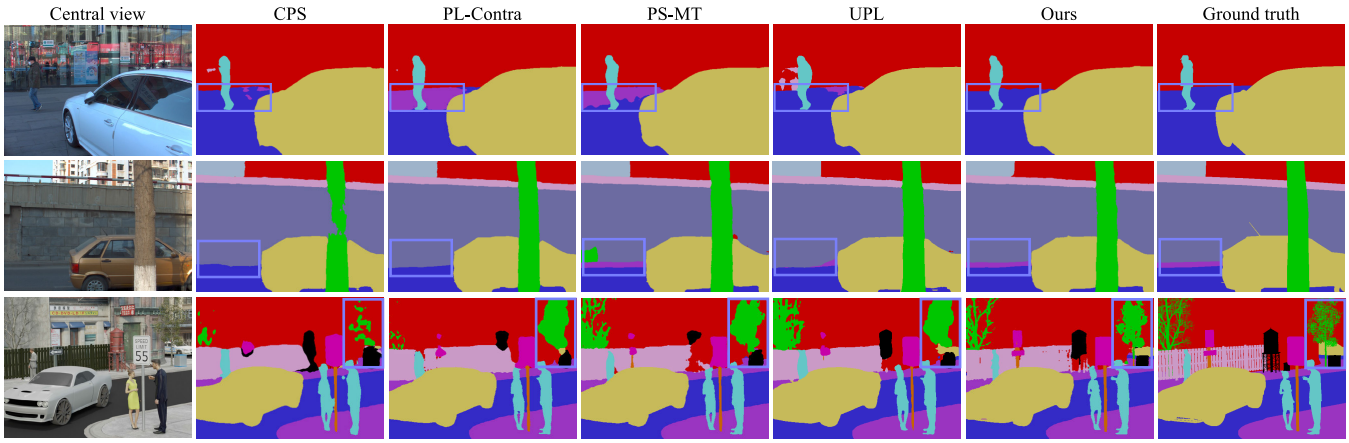


Fig. 10. Segmentation maps obtained by different semi-supervised methods trained under 1/4 protocol. The first two rows display real-world LF images and the third row displays a synthetic LF image. The purple rectangles highlight the advantages of our method over the other methods. Zoom in for best view.

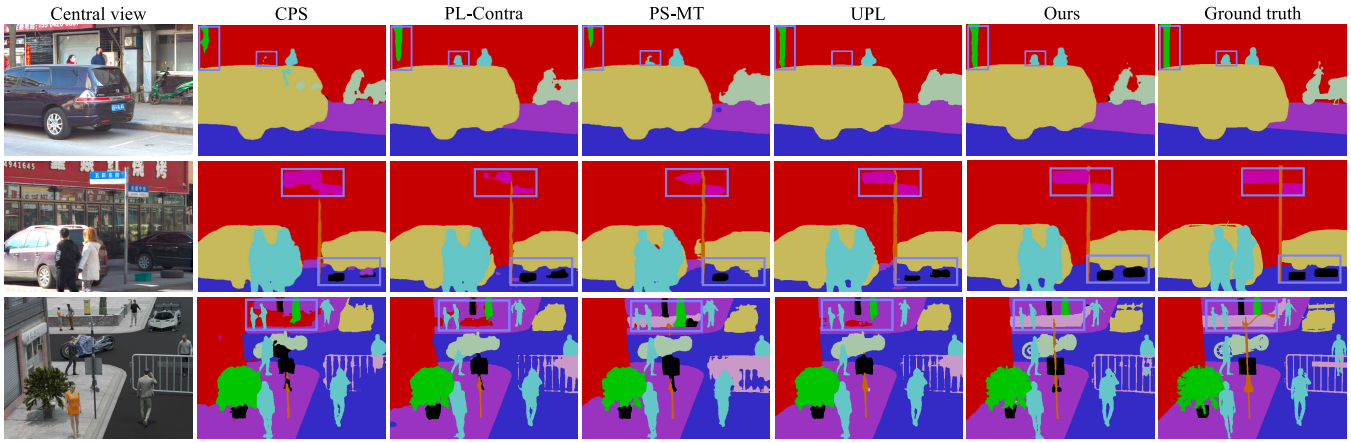


Fig. 11. Segmentation maps obtained by different semi-supervised methods trained under 1/8 protocol. The first two rows display real-world LF images and the third row displays a synthetic LF image. The purple rectangles highlight the advantages of our method over the other methods. Zoom in for best view.

across all the methods. We trained all these methods under 1/2, 1/4 and 1/8 partition protocols, with 430, 215 and 108 labeled LF images, respectively. The mean of Intersection over Union (mIoU), mean pixel accuracy (mAcc) and pixel accuracy (Acc) are used as the evaluation metrics.

The quantitative results on the UrbanLF-Real and UrbanLF-Syn test sets are listed in Table II. It is evident that our method consistently outperforms the other methods under a variety of training settings on both sets, and the advantage of our method becomes more obvious as the amount of labeled data decreases. Fig. 9, Fig. 10 and Fig. 11 present the segmentation maps of the central views obtained by different methods under 1/2, 1/4 and 1/8 partition protocols, respectively. In each figure, the first two rows display real-world LF images and the

TABLE III

QUANTITATIVE RESULTS OF SUPERVISED SEMANTIC SEGMENTATION METHODS AND OUR SEMI-SUPERVISED METHOD TRAINED UNDER 1/2 PROTOCOL ON THE URBANLF-REAL AND URBANLF-SYN TEST SETS. MIOU (%), MACC (%), ACC (%), PARAMETERS, FLOPS AND RUNTIME (S) ARE REPORTED

Method	Backbone	Parameters	FLOPs	Runtime	UrbanLF-Real			UrbanLF-Syn		
					mIoU	mAcc	Acc	mIoU	mAcc	Acc
Supervised										
DeepLabv3+ [1]	ResNet-101	58.75M	59.49G	0.017	74.12	80.76	90.72	72.02	81.97	88.67
PSPNet-LF [17]	ResNet-101	127.8M	429.4G	0.153	74.04	80.82	90.07	71.32	81.18	88.39
OCR-LF [17]	HRNetV2-W48	137.4M	228.1G	0.120	76.23	83.34	91.31	73.82	83.25	90.31
LF-IENet [18]	HRNetV2-W48	117.4M	719.7G	0.354	76.98	83.83	91.66	74.04	83.83	90.64
Semi-supervised (1/2)										
Ours ($n = 1$)	ResNet-101	58.75M	59.49G	0.017	75.27	81.95	91.38	71.40	82.30	88.62
	HRNetV2-W48	79.94M	78.29G	0.032	76.44	82.78	91.42	73.17	83.52	89.60
Ours ($n = 2$)	ResNet-101	78.78M	79.43G	0.043	75.40	82.04	91.27	71.60	82.68	88.69
	HRNetV2-W48	80.10M	146.7G	0.064	76.63	82.76	90.90	73.22	83.13	89.31
Ours ($n = 3$)	ResNet-101	78.78M	112.9G	0.053	75.31	82.06	91.24	71.54	82.26	88.79
	HRNetV2-W48	80.10M	215.0G	0.094	76.50	82.47	91.41	73.36	83.04	89.42

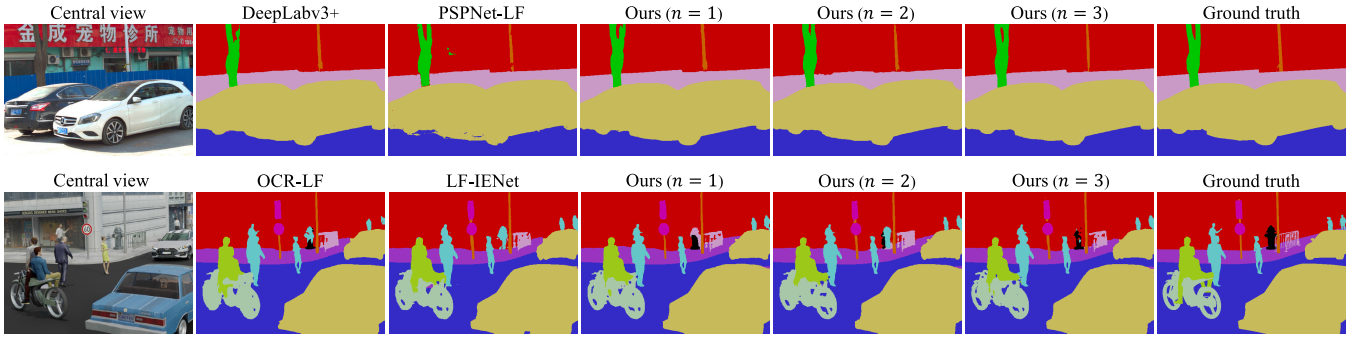


Fig. 12. Segmentation maps obtained by supervised methods and our semi-supervised method with different network configurations ($n = 1, 2, 3$) trained under 1/2 protocol for a real-world LF image and a synthetic LF image. The methods in the first row employ the ResNet-101 backbone, while the methods in the second row employ the HRNetV2-W48 backbone. Zoom in for best view.

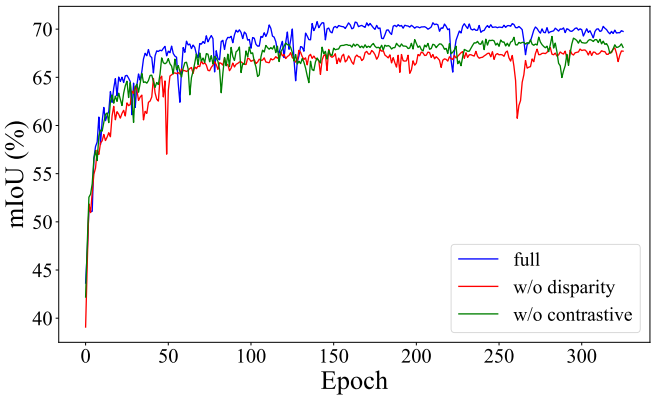


Fig. 13. The changes in mIoU on the validation set during training for convergence comparison.

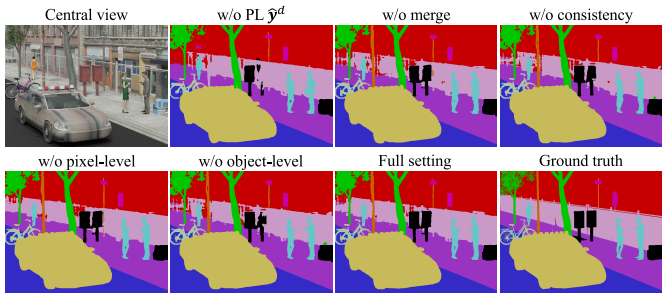
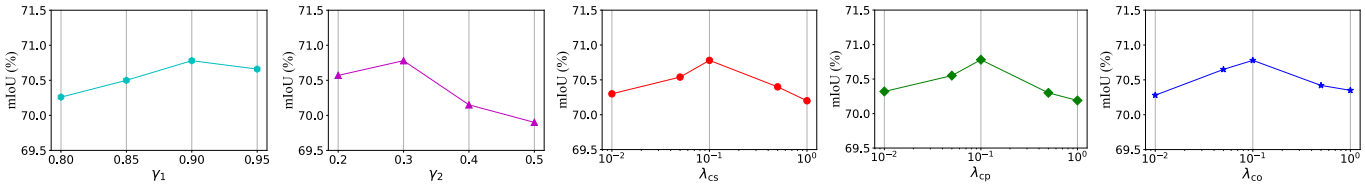


Fig. 14. Visual comparisons of various configurations in ablation studies.

third row displays a synthetic LF images. It can be seen that our method demonstrates superior segmentation performance for various objects, including trees, traffic signs and cars, as highlighted by the purple rectangles. Notably, our method distinguishes between the road and sidewalk regions more accurately, as shown in the first two rows of Fig. 10. Moreover, our method excels in segmenting objects with smaller sizes, such as people, trees and traffic signs in the synthetic images. Consequently, our training strategies that leverage LF disparity information have proven to be effective in enhancing the performance of LF semantic segmentation under semi-supervised settings.

3) Comparisons With Supervised Semantic Segmentation Methods: We also trained several models under a supervised setting. These models include DeepLabv3+ [1], which takes a single view as input, as well as PSPNet-LF [17], OCR-LF [17] and LF-IENet [18], which leverage multiple views to perform segmentation for the central view. DeepLabv3+ and PSPNet-LF employ the ResNet-101 backbone, while OCRNet-LF and LF-IENet employ the HRNetV2-W48 backbone. Table III lists their quantitative results on both the UrbanLF-Real and UrbanLF-Syn test sets, along with the



(a) mIoU vs. λ_{cs} (b) mIoU vs. λ_{cp} (c) mIoU vs. λ_{co} (d) mIoU vs. γ_1 (e) mIoU vs. γ_2

Fig. 15. The effects of hyperparameters γ_1, γ_2, λ_{cs}, λ_{cp}, and λ_{co} on mIoU. We selected the value of each hyperparameter that leads to the highest mIoU as our default setting.

number of parameters, FLOPs and runtime. In addition to the results of these supervised methods, we also provide the results of our semi-supervised method with different network configurations ($n = 1, 2, 3$) trained under $1/2$ protocol. We can observe that our semi-supervised method obtains comparable quantitative results to the supervised methods that employ the same backbone as ours. Furthermore, incorporating multiple views as input with feature alignment and fusion typically leads to a modest improvement in mIoU, but it also results in increased computation cost and memory consumption. Therefore, we did not conduct further experiments with larger values of n. Fig. 12 presents the visual comparisons of various methods that use the same backbone, which further illustrates that our semi-supervised method trained under $1/2$ protocol achieves comparable segmentation performance to the supervised methods on both real and synthetic LF images.

It is worth noting that our method is capable of performing segmentation for every view within a LF image, regardless of the number of views utilized, whereas other LF semantic segmentation models [16], [17], [18], including PSPNet-LF, OCR-LF and LF-IENet, are limited to segmenting only the central view.

D. Ablation Studies

1) Effectiveness of Using LF Disparity Information: We conducted ablation studies to investigate the impact of using LF disparity information, as described in Sec. III-B. We trained additional models with different configurations using the ResNet-101 backbone under $1/8$ partition protocol. The quantitative results on the UrbanLF-Real test set are recorded in Table IV. These results clearly indicate that when the pseudo-labels generated by disparity $\hat{\mathbf{y}}^d$ (PL $\hat{\mathbf{y}}^d$) are not utilized, there is a significant decline in performance. Additionally, even when the pseudo-labels $\hat{\mathbf{y}}^d$ are used, not incorporating the weight map $\mathbf{z}^d$ also leads to a noticeable decrease in performance. These findings validate the effectiveness of employing pseudo-labels $\hat{\mathbf{y}}^d$ for the peripheral views and performing image warping to determine pixel-wise weights for the pseudo-labels.

When the multi-view prediction merging is excluded, the performance experiences a slight decline. When the merging is included but without the merging mask $\mathbf{M}$, the performance has a more noticeable decline. This is primarily due to the fact that regions with occlusion and inaccurate disparity estimation lead to incorrect merging of the warped probability map. Moreover, the performance also suffers a degradation when the disparity-semantics consistency loss $\mathcal{L}_{cs}$ is not included.

TABLE IV
ABLATION STUDIES ON THE IMPACT OF LF DISPARITY INFORMATION. **BOLD**: BEST

PL $\hat{\mathbf{y}}^d$	Weight $\mathbf{z}^d$	Merging	Mask $\mathbf{M}$	$\mathcal{L}_{cs}$	mIoU (1/8)
✗	✗	✓	✓	✓	69.18
✓	✗	✓	✓	✓	70.33
✓	✓	✗	✗	✓	70.42
✓	✓	✓	✗	✓	70.17
✓	✓	✓	✓	✗	70.38
✓	✓	✓	✓	✓	**70.81**

TABLE V
ABLATION STUDIES ON THE CONTRASTIVE LEARNING SCHEME. **BOLD**: BEST

Pixel-level	Object-level	mIoU (1/8)
✗	✗	69.35
✗	✓	70.39
✓	✗	70.28
✓	✓	**70.81**

Fig. 13 plots the changes in mIoU on the validation set during training to show the convergence comparison, which indicates that leveraging LF disparity information with our proposed approaches accelerates the learning of semantic segmentation and leads to better performance. Fig. 14 presents the visual comparisons of various configurations in the ablation studies, demonstrating that the inclusion of pseudo-labels $\hat{\mathbf{y}}^d$, multi-view prediction merging, and disparity-semantics consistence loss contributes to improved segmentation performance.

2) Effectiveness of Contrastive Learning: To validate the effectiveness of our contrastive learning scheme, we trained additional models excluding the pixel-level or object-level strategies. The corresponding quantitative results are listed in Table V, revealing that both the pixel-level and object-level strategies contribute to higher mIoU values, and the combination of them leads to a more significant improvement in performance. The convergence comparison in Fig. 13 further supports this finding. Additionally, the visual comparison in Fig. 14 illustrates that our contrastive learning scheme leads to more accurate segmentation with improved object shapes.

3) Selection of Hyperparameters: We also conducted ablation studies to analyze the effects of hyperparameters γ_1 in Eq. 13, γ_2 in Eq. 15, as well as λ_{cs}, λ_{cp} and λ_{co} in Eq. 23 on the performance. For γ_1 and γ_2, we experimented with $[0.8, 0.85, 0.9, 0.95]$ and $[0.2, 0.3, 0.4, 0.5]$, respectively. For λ_{cs}, λ_{cp} and λ_{co}, we experimented with

[0.01, 0.05, 0.1, 0.5, 1]. The performances of models trained using different values of hyperparameters were evaluated on the validation set. Fig. 15 depicts the relationship between the values of each hyperparameter and the corresponding mIoU. For each hyperparameter, we selected the value that yields the highest mIoU as our default setting.

V. Conclusion

This paper presents a semi-supervised semantic segmentation method for LF images by leveraging disparity information. We first develop an unsupervised disparity estimation network with coarse-to-fine feature matching to estimate disparity maps for each individual views. The estimated disparity maps are utilized to generate pseudo-labels for the peripheral views when only central labels are available, merge the multi-view predictions from the teacher network to obtain more reliable pseudo-labels, provide a structure reference for the segmentation of unlabeled data, and align features for multi-view feature fusion. Moreover, we propose a contrastive learning scheme that operates at both the pixel level and the object level to improve feature representations and object segmentation. The experimental results demonstrate the effectiveness of our method for semi-supervised LF semantic segmentation.

References

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[2] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[3] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, "ShapeConv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7068–7077.

[4] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Gross, "Efficient RGB-D semantic segmentation for indoor scene analysis," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13525–13531.

[5] P. Hu, F. Caba, O. Wang, Z. Lin, S. Sclaroff, and F. Perazzi, "Temporally distributed networks for fast video semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8815–8824.

[6] Z. Liang, X. Dai, Y. Wu, X. Jin, and J. Shen, "Multi-granularity context network for efficient video semantic segmentation," *IEEE Trans. Image Process.*, vol. 32, pp. 3163–3175, 2023.

[7] S. Zhang and E. Y. Lam, "Learning to restore light fields under low-light imaging," *Neurocomputing*, vol. 456, pp. 76–87, Oct. 2021.

[8] S. Zhang and E. Y. Lam, "Light field image restoration via latent diffusion and multi-view attention," *IEEE Signal Process. Lett.*, vol. 31, pp. 1094–1098, 2024.

[9] Y. Wang, J. Yang, Y. Guo, C. Xiao, and W. An, "Selective light field refocusing for camera arrays using bokeh rendering and superresolution," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 204–208, Jan. 2019.

[10] W. Zhou, E. Zhou, G. Liu, L. Lin, and A. Lumsdaine, "Unsupervised monocular depth estimation from light field image," *IEEE Trans. Image Process.*, vol. 29, pp. 1606–1617, 2020.

[11] J. Jin and J. Hou, "Occlusion-aware unsupervised learning of depth from 4-D light fields," *IEEE Trans. Image Process.*, vol. 31, pp. 2216–2228, 2022.

[12] S. Zhang and E. Y. Lam, "Unsupervised disparity estimation for light field videos," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 2620–2624.

[13] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–12, Jul. 2013.

[14] Z. Liang, P. Wang, K. Xu, P. Zhang, and R. W. H. Lau, "Weakly-supervised salient object detection on light fields," *IEEE Trans. Image Process.*, vol. 31, pp. 6295–6305, 2022.

[15] M. Zhang, S. Xu, Y. Piao, and H. Lu, "Exploring spatial correlation for light field saliency detection: Expansion from a single view," *IEEE Trans. Image Process.*, vol. 31, pp. 6152–6163, 2022.

[16] C. Jia et al., "Semantic segmentation with light field imaging and convolutional neural networks," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[17] H. Sheng, R. Cong, D. Yang, R. Chen, S. Wang, and Z. Cui, "UrbanLF: A comprehensive light field dataset for semantic segmentation of urban scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7880–7893, Nov. 2022.

[18] R. Cong, D. Yang, R. Chen, S. Wang, Z. Cui, and H. Sheng, "Combining implicit-explicit view correlation for light field semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 9172–9181.

[19] J. Chen, S. Zhang, and Y. Lin, "Attention-based multi-level fusion network for light field depth estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1009–1017.

[20] J. Shi, X. Jiang, and C. Guillemot, "A framework for learning depth from a flexible subset of dense and sparse light field views," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5867–5880, Dec. 2019.

[21] C. Shin, H.-G. Jeon, Y. Yoon, I. S. Kweon, and S. J. Kim, "EPINET: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4748–4757.

[22] Y.-J. Tsai, Y.-L. Liu, M. Ouhyoung, and Y.-Y. Chuang, "Attention-based view selection networks for light-field disparity estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12095–12103.

[23] S. Zhang, N. Meng, and E. Y. Lam, "Unsupervised light field depth estimation via multi-view feature matching with occlusion prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2261–2273, Apr. 2024.

[24] I. Alonso, A. Sabater, D. Ferstl, L. Montesano, and A. C. Murillo, "Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8219–8228.

[25] D. Kwon and S. Kwak, "Semi-supervised semantic segmentation with error localization network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9957–9967.

[26] Y. Wang et al., "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 4248–4257.

[27] Y. Zhong, B. Yuan, H. Wu, Z. Yuan, J. Peng, and Y.-X. Wang, "Pixel contrastive-consistent semi-supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7273–7282.

[28] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," 2018, *arXiv:1802.07934*.

[29] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high- and low-level consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1369–1379, Apr. 2021.

[30] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2613–2622.

[31] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1195–1204.

[32] Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro, "Perturbed and strict mean teachers for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4258–4267.

[33] R. D. Hjelm et al., "Learning deep representations by mutual information estimation and maximization," in *Proc. Int. Conf. for Learn. Represent. (ICLR)*, 2019.

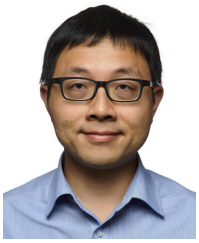
[34] S. Wanner, C. Straehle, and B. Goldluecke, "Globally consistent multi-label assignment on the ray space of 4D light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1011–1018.

[35] M. Hog, N. Sabater, and C. Guillemot, "Light field segmentation using a ray-based graph structure," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 35–50.

[36] X. Lv, X. Wang, Q. Wang, and J. Yu, "4D light field segmentation from light field super-pixel hypergraph representation," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 9, pp. 3597–3610, Sep. 2021.
[37] D. Yang, T. Zhu, S. Wang, S. Wang, and Z. Xiong, "LFRSNet: A robust light field semantic segmentation network combining contextual and geometric features," *Frontiers Environ. Sci.*, vol. 10, p. 1443, Oct. 2022.
[38] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3268–3277.
[39] A. Kendall et al., "End-to-end learning of geometry and context for deep stereo regression," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 66–75.
[40] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6602–6611.
[41] S. Zhang and E. Y. Lam, "An effective decomposition-enhancement method to restore light field images captured in the dark," *Signal Process.*, vol. 189, Dec. 2021, Art. no. 108279.
[42] S. Zhang, N. Meng, and E. Y. Lam, "LRT: An efficient low-light restoration transformer for dark light field images," *IEEE Trans. Image Process.*, vol. 32, pp. 4314–4326, 2023.
[43] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.
[44] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 1597–1607.
[45] L. Wang et al., "Parallax attention for unsupervised stereo correspondence learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2108–2125, Apr. 2022.
[46] I.-S. Fang, H.-C. Wen, C.-L. Hsu, P.-C. Jen, P.-Y. Chen, and Y.-S. Chen, "ES3Net: Accurate and efficient edge-based self-supervised stereo matching network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 4472–4481.
[47] T. Iwatsuki, K. Takahashi, and T. Fujii, "Unsupervised disparity estimation from light field using plug-and-play weighted warping loss," *Signal Process., Image Commun.*, vol. 107, Sep. 2022, Art. no. 116764.
[48] R. Cong et al., "End-to-end semantic segmentation utilizing multi-scale baseline light field," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 5790–5804, Jul. 2024.
[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[51] K. Sun et al., "High-resolution representations for labeling pixels and regions," 2019, *arXiv:1904.04514*.
[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[53] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 19–34.

Shansi Zhang (Graduate Student Member, IEEE) received the B.S. degree in mechanical engineering from Beijing Institute of Technology in 2016 and the M.S. degree in electrical and electronic engineering from Nanyang Technological University in 2019. She is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering, The University of Hong Kong. Her research interests include image processing, computer vision, and computational imaging.

Yaping Zhao (Graduate Student Member, IEEE) received the B.S. degree in automation from Beihang University in 2018 and the M.S. degree from Tsinghua University in 2021. She is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering, The University of Hong Kong. She visited Westlake University in 2021. Her research interests include computational imaging and computer vision.



Edmund Y. Lam (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA. He was a Visiting Associate Professor with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. He is currently a Professor in electrical and electronic engineering with The University of Hong Kong, Hong Kong. He is also the Computer Engineering Program Director and a Research Program Coordinator with the AI Chip Center for Emerging Smart Systems. His research interests focus on computational imaging algorithms, systems, and applications. He is a fellow of Optica, SPIE, IS&T, and HKIE; and a Founding Member of Hong Kong Young Academy of Sciences.