frontiers | Frontiers in Environmental Science

TYPE Original Research
PUBLISHED 07 October 2022
DOI 10.3389/fenvs.2022.996513

OPEN ACCESS

EDITED BY
Hao Sheng,
Beihang University, China

REVIEWED BY
Wenhui Zhou,
Hangzhou Dianzi University, China
Yuan Xu,
Beijing University of Chemical Technology, China

*CORRESPONDENCE
Tongyu Zhu,
zhutongyu@nlsde.buaa.edu.cn

SPECIALTY SECTION
This article was submitted to Environmental Informatics and Remote Sensing,
a section of the journal
Frontiers in Environmental Science

RECEIVED 17 July 2022
ACCEPTED 29 July 2022
PUBLISHED 07 October 2022

CITATION
Yang D, Zhu T, Wang S, Wang S and Xiong Z (2022), LFRSNet: A robust light field semantic segmentation network combining contextual and geometric features.
Front. Environ. Sci. 10:996513.
doi: 10.3389/fenvs.2022.996513

LFRSNet: A robust light field semantic segmentation network combining contextual and geometric features

Da Yang[1,2], Tongyu Zhu[1,2]*, Shuai Wang[1,2], Sizhe Wang[1,2] and Zhang Xiong[1,2,3]

[1]State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing, China, [2]Beihang Hangzhou Innovation Institute Yuhang, Hangzhou, China, [3]Faculty of Applied Sciences, Macao Polytechnic University, Macao, Macao SAR, China

Light field (LF) semantic segmentation is a newly arisen technology and is widely used in many smart city applications such as remote sensing, virtual reality and 3D photogrammetry. Compared with RGB images, LF images contain multi-layer contextual information and rich geometric information of real-world scenes, which are challenging to be fully exploited because of the complex and highly inter-twined structure of LF. In this paper, LF Contextual Feature (LFCF) and LF Geometric Feature (LFGF) are proposed respectively for occluded area perception and segmentation edge refinement. With exploitation of all the views in LF, LFCF provides glimpse of some occluded areas from other angular positions besides the superficial color information of the target view. The multi-layer information of the occluded area enhances the classification of partly occluded objects. Whereas LFGF is extracted from Ray Epipolar-Plane Images (RayEPIs) in eight directions for geometric information embedding. The solid geometric information refines object edges, especially for occlusion boundaries with similar colors. At last, Light Field Robust Segmentation Network (LFRSNet) is designed to integrate LFCF and LFGF. Multi-layer contextual information and geometric information are effectively incorporated through LFRSNet, which brings significant improvement for segmentation of the occluded objects and the object edges. Experimental results on both realworld and synthetic datasets proves the state-of-the-art performance of our method. Compared with other methods, LFRSNet produces more accurate segmentation under occlusion, especially in the edge regions.

KEYWORDS

light field, semantic segmentation, contextual feature, geometric feature, smart city

1 Introduction

Semantic segmentation, which assigns semantic labels for each pixel in an image, has drawn great attention in recent years. With high-level understanding of images, it facilitates many smart city applications like remote sensing (Li F et al., 2022; Li Y et al., 2022), object tracking (Zhang et al., 2020; Wang H et al., 2021), virtual reality (Gao et al., 2022; Gu et al., 2022) and 3D photogrammetry (Frangez et al., 2022; Wang H et al., 2022). Based on fully convolutional networks, semantic segmentation is successfully conducted with single images, videos and RGB-D data considering different application scenarios. Compared with single images, videos provide information from other perspectives and RGB-D data provide direct geometric information (depth maps, etc), both of which greatly promote performance in semantic segmentaion. Light field (LF) captures both intensity and directions of light rays in the scene. Sub-aperture images (SAIs) in LF are regularly sampled on angular domain. Compared with videos, geometric information can be deduced more easily from LF. And different from RGB-D data, LF images embed geometric information without need of additional depth sensors and provide multi-perspective observation. Hence the introduction of LF can boost the development of semantic segmentation.

Although LF can greatly benefit semantic segmentation, this field develops rather slowly because of lack in relevant datasets. Recently, the first large-scale LF semantic segmentation dataset named UrbanLF (Sheng et al., 2022) was proposed. Two state-of-the-art methods, PSPNet (Zhao et al., 2017) and OCR (Yuan et al., 2020), were modified by the authors to adapt for LF data and work as baseline methods. In both methods (PSPNet-LF and OCR-LF), geometric features were extracted from epipolar-plane images (EPIs) in four directions. The geometric features were then integrated into the original networks through attention mechanism for final segmentation of the center view. However, in these two modifications, more than half of the SAIs are ignored due to the star-like input structure. The contextual and geometric information of LF are not fully explored.

In this paper, Light Field Robust Segmentation Network (LFRSNet) is designed to fully expoit LF in semantic segmentation. In LFRSNet, Light Field Contextual Feature (LFCF) and Light Field Geometric Feature (LFGF) are proposed. LFCF is extracted through perception of multi-layer information from all the SAIs, which benefits the classification of the occluded objects. Whereas LFGF is extracted based on Ray Epipolar-Plane Images (RayEPIs) in eight directions, which is more robust to occlusion and is beneficial to the segmentation along occlusion boundaries. LFRSNet integrates LFCF, LFGF and the intial features from the center view adaptively with attention mechanism. Our method achieves state-of-the-art performance on UrbanLF (Sheng et al., 2022) dataset. On subset UrbanLF-Syn with ground-truth disparity, based purely on multiple perpectives of LF, LFRSNet also outperforms state-of-the-art RGB-D methods.

In summary, the main contributions of this paper are concluded as follows:

- LFCF is introduced based on an angular-distance-aware context-perception mechanism to provide perception of multi-layer information, which promotes classification accuracy of occluded objects.
- RayEPI is proposed for robustness in occlusion areas and LFGF is extracted from RayEPIs in eight directions, which benefits semantic segmentation around occlusion boudaries.
- LFRSNet is designed by adaptively combining LFCF and LFGF with attention mechnism, which outperforms state-of-the-art methods on the public dataset.

The rest of this paper is organized as follows. In Section 2, the related works are briefly reviewed. In Section 3, two specific semantic segmentation features from LF, LFCF and LFGF, are first introduced. Then the architecture of LFRSNet is proposed. The experimental results are presented in Section 4 and the conclusion is given in Section 5.

2 Related work

Semantic segmentation has long been studied by researchers. Because of various practical conditions, it is investigated with all kinds of datatypes, like single images, videos, RGB-D data, etc. Due to the fact that LF can be organized as image sequences with apparent regularity in SAIs, methods developed for videos can inspire the research in LF semantic segmentation. And different from RGB-D data, LF contains geometric information without need of additional depth sensors. It is necessary to include RGB-D based method into the discussion to learn the exploitation of geometric information. Hence in this section, previous works in semantic segementation based on single images, videos, RGB-D data are first reviewed for latter experiments and analysis. Then the lage-scale LF semantic segmentation dataset UrbanLF (Sheng et al., 2022) and its propsoed baseline methods are introduced.

2.1 Single image semantic segmentation

FCN (Shelhamer et al., 2017) first introduces fully convolutional networks to semantic segmentation. Exploiting both global and local clues, PSPNet (Zhao et al., 2017) produces pyramid pooling module (PPM) which is widely used in research. Different from other methods that encode the input image as a low-resolution representation, HRNet (Wang J et al., 2021) keeps high-resolution representations through the whole process, which causes significant sensitivity

to small objects. Deeplabv2 (Chen et al., 2018) proposes atrous spatial pyramid pooling (ASPP) for robust segmentation of objects at multiple scales. OCR (Yuan et al., 2020) presents object-contextual representations which characterize a pixel by exploiting the representation of the corresponding object class. Cacrfs Net (Ji et al., 1938) designs a cascaded CRFs and integrates it into the decoder of semantic segmentation model to learn boundary information from multi-layers. SETR (Zheng et al., 2021) replaces traditional convolution layers with a pure transformer to encode an image as a sequence of patches and a simple decoder is enough to reach state-of-the-art performance. Because of the limited scene information reserved by single images, the state-of-the-art single image semantic segmentation methods still suffers from inferior segmentation boundaries.

2.2 Video semantic segmentation

Videos provide multiple perspectives of the scene, which facilitates semantic segmentation. To lower the cost of deep networks in per-frame evaluation, (Zhu et al., 2017) performs the expensive convolutional sub-network only on sparse key frames and propagates their deep feature maps to other frames based on a flow field. Information across frames are shared by reusing stable features extracted from deep layers in (Carreira et al., 2018). Jain et al. (2019) designs a reference branch to extract high-detail features on a reference keyframe and an update branch to perform a temporal update at each video frame, which achieves high accuracy at low inference cost. TDNet (Hu et al., 2020) achieves fast and accurate video semantic segmentation by approximating features of high-level layers with the composition of features extracted from several shallower sub-networks. (Zhuang et al., 2021). proposes a distortion-aware feature correction method, which improves video segmentation performance at a low price. TMANet (Wang S et al., 2021) adaptively integrates the long-range temporal relations over the video sequence based on the self-attention mechanism. Although containing multi-perspective information, the changes between frames in videos are not regular, which makes it hard to fully exploit multiple frames.

2.3 RGB-D semantic segmentation

Different from videos, RGB-D data provide direct geometric information like depth maps, rather than multi-perspective observation of the scene. With additional depth information, depth-aware convolution and depth-aware average pooling are propsed by DCNN (Wang and Neumann, 2018) to seamlessly incorporate geometry into CNN. Based on attention mechanism, ACNet (Hu et al., 2019) selectively gathers features from RGB and depth branches. A novel Pattern-Affinitive Propagation framework is proposed to jointly predict depth, surface normal and semantic segmentation in (Zhang et al., 2019). SA-Gate (Chen et al., 2020) introduces a novel Separation-and-Aggregation Gating operation to filter and recalibrate RGB and depth representations before cross-modality aggregation. MTI-Net (Vandenhende et al., 2020) utilizes depth data as a supervised signal and a multi-task learning framework is adopted to jointly train multi-modal tasks to improve single-task performance. To avoid separate process of RGB and 3D spatial information, spatial information guided convolution is proposed in SGNet (Chen et al., 2021), which allows efficient RGB feature and 3D spatial information integration. RGB-D data provide accurate geometric information of the scene. However, on the one hand, the capture of RGB-D data requires additional depth sensors and the calibration between depth information and RGB data is also hard to be conducted correctly. On the other hand, the lack of multi-perspective information limits the comprehensive understanding of the scene.

2.4 LF semantic segmentation

Different from videos and RGB-D data, SAIs in LF are uniformally sampled in angular domain. Reliable geometric information can be extracted from LF and meanwhile it provides multi-perspective observation of the scene. This property of LF facilitates many applications, like light field super-resolution (Zhang et al., 2021; Wang Y et al., 2022), disparity estimation (Shin et al., 2018; Huang et al., 2021), etc. The performance in semantic segmentation can also be greatly promoted with the introduction of LF. Recently, the first large-scale LF semantic segmentation dataset (namely UrbanLF) is constructed by (Sheng et al., 2022). Two state-of-the-art methods PSPNet (Zhao et al., 2017) and OCR (Yuan et al., 2020) are modified by the authors to deal with LF. Through simple modification, the resulting models, PSPNet-LF and OCR-LF, easily surpasses other state-of-the-art methods. It is obvious that the potential of LF semantic segmentation is not fully excavated. In this paper, we also dig into this problem with a semantic segmentation method specially designed for LF.

3 Light field robust segmentation network

A 4D LF is denoted as $L \in \mathbb{R}^{U\times V\times X\times Y}$ where $U \times V$ is the angular resolution and $X \times Y$ is the spatial resolution. An SAI $L_{(u,v)} \in \mathbb{R}^{X\times Y}$ is extracted by fixing the angular coordinate at (u, v). The task of LF semantic segmentation is to assign semantic labels to each pixel of the center SAI $L_{(u_c,v_c)}$.

SAIs in an LF image are uniformly sampled in the angular domain, which leads to linear displacement of the pixels across

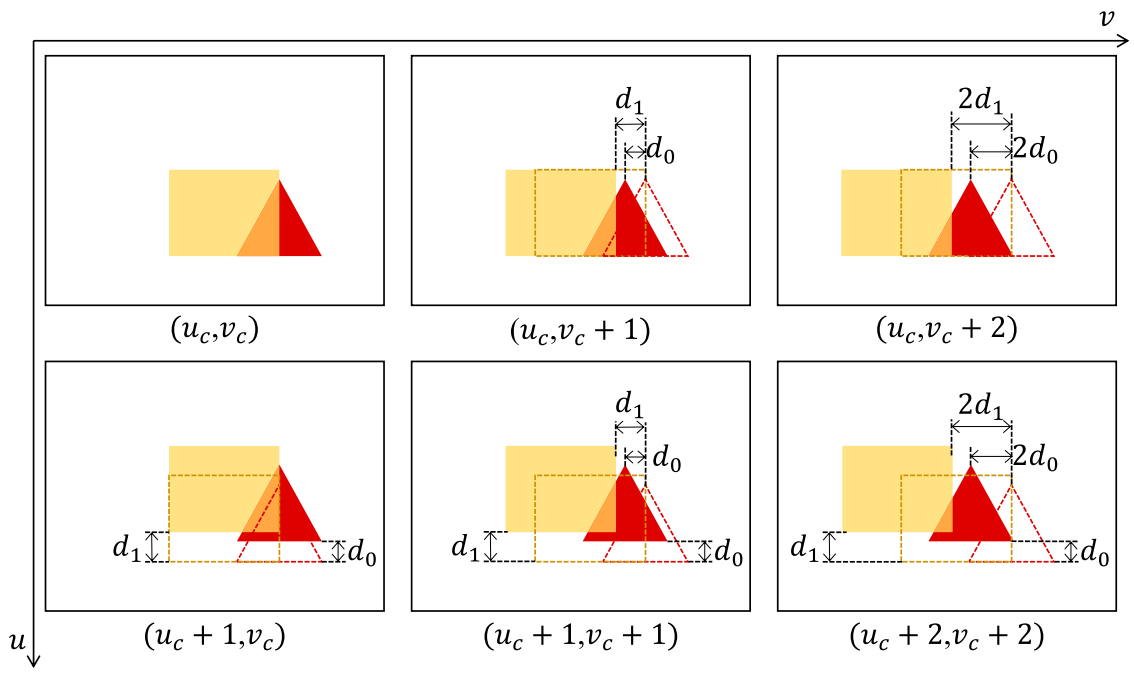

FIGURE 1
The yellow rectangle lies in front of the red triangle in the scene and it is set semitransparent to make the triangle visible during the analysis. Their disparities are d_1 and d_0 respectively, with $d_0 < d_1$. As the view point moves rightwards from (u_c, v_c), the occluded part of the triangle gradually becomes visible, which facilitates classification. The occluded object may not be entirely visible with view shift in only one direction. By taking other directions into consideration, as in view ($u_c + 1$, v_c), the left corner of the triangle hidden in (u_c, $v_c + 2$) is exposed. Combining ($u_c + 1$, v_c) and (u_c, $v_c + 2$), it is easier to correctly classify the triangle to be an equilateral triangle, rather than a right triangle. With relative displacements of the objects in all the SAIs, the performance in classification can be greatly improved.

views. It results in two special properties. One is that the relative position of the objects in the SAIs changes when viewpoint shifts. The other is that the geometric information of the scene can be easily deduced from LF. Considering these properties, two specific features, LFCF and LFGF, are proposed respectively to produce multi-layer contextual information and robust geometric information of the scene to promote performance in LF semantic segmentation. By adaptively combining LFCF and LFGF, LFRSNet is proposed based on attention mechanism, which shows significant improvement for classification of the occluded objects and segmentation of the object edges.

3.1 Light field contextual feature (LFCF)

In LF images, objects of different distances to the camera plane have different disparities. Hence objects shift in different speeds across SAIs and the occluded part of an object in the center view may be observed in other SAIs. Due to the linear structure of LF, objects shift in diverse directions and various degrees in different surrounding views. Therefore many different occluded areas of the objects can be perceived from other angular positions, which greatly benefits the classification of occluded objects. With perception to both the occluders and the occluded areas, LFCF is proposed accordingly to capture this multi-level contextual information of the scene.Due to the uniform sampling in the angular domain, a scene point has consistent disparity between each pair of adjacent SAIs in an LF image. Objects in an SAI $L_{(u,v)}$ moves linearly across SAIs:

$$L(u, v, x, y) = L(u + \Delta u, v + \Delta v, x - \Delta u \times d, y - \Delta v \times d). \quad (1)$$

d denotes the disparity of pixel L (u, v, x, y). According to Eq. 1, when the angular distances (Δu and Δv) between the SAIs grows, the displacements of the objects in the scene inscrease. The relative displacements among the objects also rise with growing (Δu, Δv):

$$Displacement_{relative} = (\Delta u \times \Delta d, \Delta v \times \Delta d), \quad (2)$$

with Δd signifing the difference in disparities between two objects.

As shown in Figure 1, the red triangle lies behind the yellow rectangle and has smaller disparity. Assume the disparities of the triangle and rectangle are d_0 an d_1 respectively. The relative displacement between them across views is ($\Delta u \times (d_1 - d_0)$, $\Delta v \times (d_1 - d_0)$). When the view point moves rightwards from (u_c, v_c) to

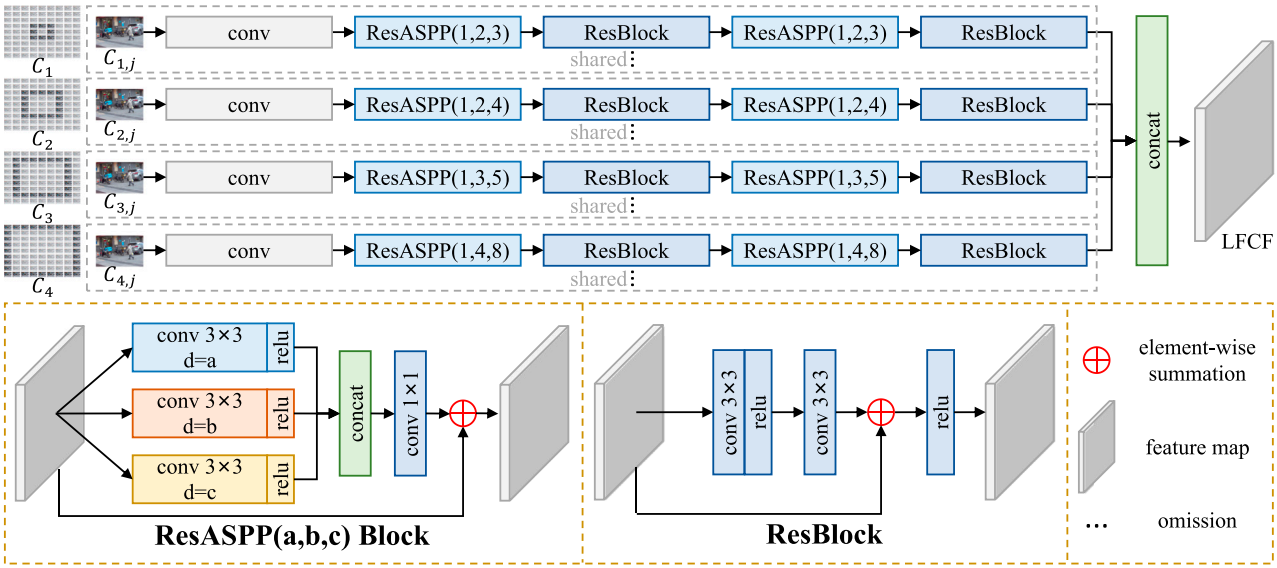

FIGURE 2
Given an LF image, the surrounding views are first grouped into circles around the center view according to their angular distances to the center view. For a 9 × 9 LF, four circles can be constructed. The initial contextual features of SAIs are extracted separately with a sequence of "conv-ResASPP-ResBlock-ResASPP-ResBlock." For SAIs in different circles, the dilation rates of the convolutional layers in ResASPP vary to handle the linearly growing displacement range of the occluded objects. Then the initial contextual features of 80 SAIs are concatenated to form LFCF. Parameters are shared among branches for SAIs in the same circle.

$(u_c, v_c + 1)$, the two objects relatively shift for $(0, (d_1 - d_0))$ and part of the occluded area becomes visible. When the view point continues to shift to $(u_c, v_c + 2)$, the relative displacement grows to $(0, 2 \times (d_1 - d_0))$ and more of the occluded area moves out from behind the rectangle. With more and more occluded area being in sight, the classification of the rectangle gets easier.

Due to the relatively small disparities of LF images, the occluded objects cannot always be entirely visible when the viewpoint shifts to the border views in a certain direction. However, Δu and Δv can be either positive and negative integers. Specifically for the center view of a 9 × 9 LF where $(u_c, v_c) = (5, 5)$, $\Delta u, \Delta v \in \{\pm 1, \pm 2, \pm 3, \pm 4\}$. Hence, the search for the occluded part of an object can be continued in other directions. As shown in Figure 1, the relative displacement of the objects in the lower view $(u_c + 1, v_c)$ is $((d_1 - d_0), 0)$. And the left corner of the triangle is exposed. The relative displacements in various directions reveal complementary information of the occluded part of an object.

Combining the additionally exposed parts from the surrounding views, the triangle can be classified to be an equilateral triangle with high confidence, rather than a right triangle. The perception of the occluded part from all the SAIs can greatly boost the performance of semantic segmentation.

LFCF is introduced to represent this multi-layer contextual information, covering both the occluders and the occluded parts of objects. Different from the contextual features for single images, which mainly covers the relationships among the surrounding areas inside the view, LFCF additionally learns the comprehensive representation of the objects themselves from LF. Specifically, LFCF complements the occluded objects in the center view with the information from the areas that become visible in other SAIs, like the occluded half of the triangle in Figure 1.

According to Eq. 2, the relative displacement of the objects increases linearly with the angular distance between views. To cover as much information of the target objects from other views as possible, an angular-distance-aware context-perception mechanism is designed to extract LFCF. As shown in Figure 2, all the SAIs other than the center view participate in LFCF extraction. The surrounding views are seperated into four groups based on their maximum angular distances to the center view, which form four circles around the center view in a 9 × 9 LF. Let C_i denote a circle around the center view, with $i = 1$ indicating the most inside circle. Objects in SAIs of the same circle C_i share similar degrees of relative displacements according to Eq. 2.

As shown in Figure 2, ResASPP Block from (Wang et al., 2019) and ResBlock from (He et al., 2016) are adopted for initial contextual feature extraction:

$$\begin{aligned} ResASPP_{a,b,c}(x) &= conv_{1\times1}\left(cat\left(\{relu\left(conv_{3\times3,d}(x)\right)|d=a,b,c\}\right)\right)+x,\\ ResBlock(x) &= relu\left(conv_{3\times3}\left(relu\left(conv_{3\times3}(x)\right)\right)+x\right). \end{aligned} \tag{3}$$

d denotes dilation rate of the convolution. When not specifically marked, the dilation rate of the convolution is 1. The SAIs in the

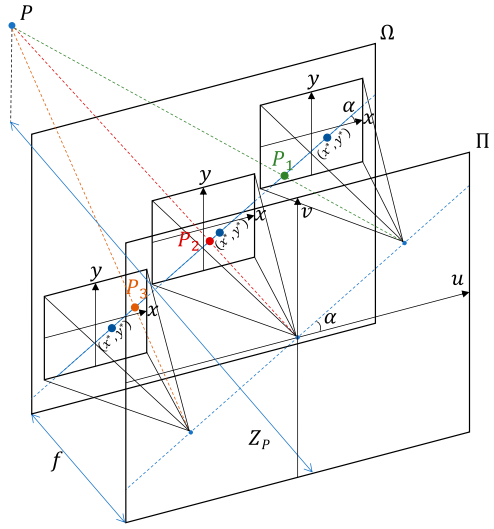

FIGURE 3
(*x*, *y*) and (*u*, *v*) denote coordinates in spatial plane Ω and angular plane Π, respectively. To construct EPIs with direction α, the views lie on the line with direction α through the center view in angular plane Π are selected. By varying (*x**, *y**), different groups of line segments, which are parallel to the line above, in the selected view images in spatial plane Ω can be extracted. Each group of line segments form an EPI with direction α. To construct RayEPIs, only the views lie on the ray with direction α out from the center view in angular plane Π are selected. The rest of steps are the same as the construction of EPIs.

same circle are first fed into an convolutional layer for initial feature extraction. Then two groups of alternate ResASPP Block and ResBlock are sequentially applied to extract the initial contextual feature of the *j*th SAI $C_{i,j}$ in circle C_i:

$$f_{initcontext}\left(C_{i,j}\right) = ResRes_{a,b,c}\left(ResRes_{a,b,c}\left(conv\left(C_{i,j}\right)\right)\right). \quad (4)$$

$ResRes_{a,b,c}$ denotes a ResBlock following a ResASPP Block. *a*, *b* and *c* are the dilation rates of the convolutions in ResASPP Blocks. Varying dilation rates are used in ResASPP to cover different displacement ranges caused by change in angular distance. Specifically, for SAIs in C_1 to C_4, (*a*, *b*, *c*) are set to (1, 2, 3), (1, 2, 4), (1, 3, 5) and (1, 4, 8) respectively. Paremeters are shared among branches for SAIs in the same circle. In the end, the initial contextual features from all the SAIs in all the circles are concatenated to form LFCF:

$$LFCF = concat\left(\left\{f_{initcontext}\left(C_{i,j}\right)\right\}\right), i \in \{1,2,3,4\}, C_{i,j} \in C_i. \quad (5)$$

LFCF provides comprehensive perception of objects in the center view with both the superficial information of the occluders and glimpse of the occluded parts from surrounding views. With multi-layer information of the scene from LFCF, the occluded objects can be more accurately classified.

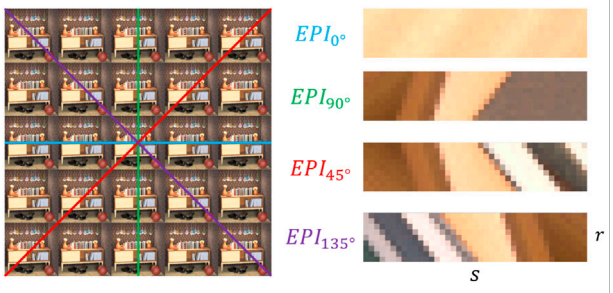

FIGURE 4
A set of multi-direction EPIs in four directions. In EPI_α, abscissa *s* changes in spatial domain and ordinate *r* changes in angular domain. Traditionally, multi-direction EPIs are constructed in four directions: 0°, 45°, 90° and 135°.

3.2 Light field geometric feature (LFGF)

The success of semantic segmentation from RGB-D data has proved the effectiveness of geometric information. LF images contain abundant geometric information of the scene which has been considered by Sheng et al. (2022). However, in the baseline methods proposed in Sheng et al. (2022), geometric information was only extracted from EPIs in four directions, which is not robust in complex scenes. In this subsection, RayEPI is proposed and LFGF is extracted from RayEPIs in eight directions. Compared with normal EPI, RayEPI is more robust to occlusion. Hence geometric information of higher accuracy can be provided by LFGF.

EPIs are constructed by stacking slices of SAIs in certain directions, which directly reflect disparities of the scene. To make full use of angular information in geometric information extraction, researchers propose multi-direction EPIs. As shown in Figure 3, EPI with direction α is defined as

$$EPI_{\alpha,x^*,y^*}(s,r) = L(x^* + s\cos\alpha, y^* + s\sin\alpha,\ r\cos\alpha, r\sin\alpha), \quad s \in \mathbb{R}, r \in \mathbb{R}, \alpha \in [0,\pi), \quad (6)$$

where (*x**, *y**) and α determine the position and direction of the sampling lines in the sampling views used to construct EPIs. Traditionally, multi-direction EPIs are sampled with α of 0, $1/4\pi$, $1/2\pi$ and $3/4\pi$. As illustrated in Figure 4, *s* is the abscissa of EPI_α, which changes in spatial domain. *r* is the ordinate of EPI_α, which changes in angular domain. We assume the coordinate of the center view in LF to be (0, 0), which is also set as the coordinate of the center pixel in each view image.

Normal EPIs have been widely used in previous works (Shin et al., 2018) for its intuitive reflection of geometric information. However, they still have two main disadvantages:

1) In the case of general occlusion, where occlusion edge only exists in one side of the scene point, most of the EPIs are considered to be unreliable since occlusion affects most of

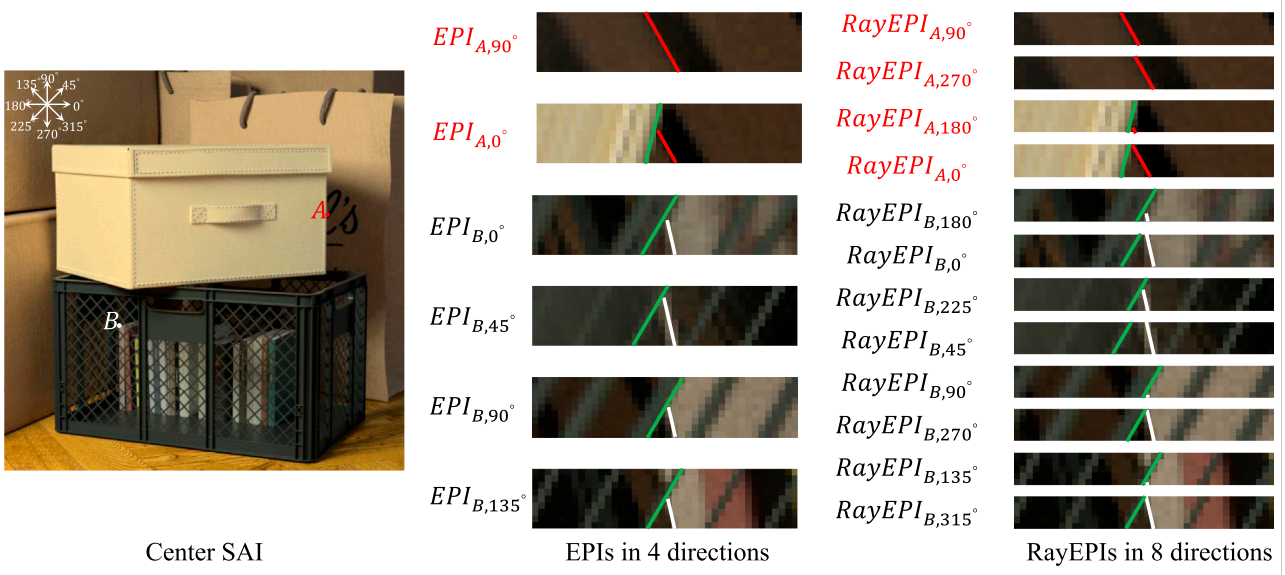

FIGURE 5
In the case of general occlusion, traditional EPIs are affected in most directions. Usually only the EPI which consists of lines parallel to the occlusion edge is occlusion-free. For example, *A* is a scene point in the character on the paper bag, near the vertical occlusion edge. In $EPI_{A,90^\circ}$, the images of *A* form a complete line. While in all the other directions, point *A* is occluded to some extent. Like in $EPI_{A,0^\circ}$, the red line of point *A* is cut off by the green line representing the occlusion edge. Accurate depth estimation can not be made under the influence of occlusion. However, the general occlusion usually only affects half of the EPI. By separating the $EPI_{A,0^\circ}$ into $RayEPI_{A,180^\circ}$ and $RayEPI_{A,0^\circ}$, a reliable depth estimation can still be deduced from $RayEPI_{A,0^\circ}$. In the case of dense occlusion, different from traditional EPIs, one non-occluded RayEPI can be found at least. For example, *B* is a point on the book in the basket. The white lines of point *B* is cut off by green lines of occlusion points in all the trational EPIs. While in $RayEPI_{B,0^\circ}$, $RayEPI_{B,45^\circ}$, $RayEPI_{B,270^\circ}$ and $RayEPI_{B,315^\circ}$, the images of *B* all form complete lines and geometric information of higher accuracy can be extracted from them.

them. For example, *A* is a point near the vertical occlusion edge in Figure 5. Except $EPI_{A,90^\circ}$, wihch consists of lines parallel to the occlusion edge, EPIs in all the other directions are affected by occlusion. Like in $EPI_{A,0^\circ}$, which is composed of lines perpendicular to the occlusion edge, the red line of point *A* is cut off by the green line of the occlusion points. It is hard to estimate the slope of the red line due to the exsitence of occlusion. In traditional EPI-based methods, $EPI_{A,0^\circ}$ and other EPIs affected by occlusion are usually ignored or assigned lower weights. The loss of scene information makes reliable geometric information extraction much harder.

2) In the areas with dense occlusion, where the scene point is surrounded or half surrounded by occlusion edges, EPIs are affected in all the directions. For example, point *B* locates behind the dense wire mesh in Figure 5. In traditional multi-direction EPIs, the white lines of point *B* are affected by green lines of occlusion points in all the directions. So the effect of occlusion cannot be excluded and reliable geometric information cannot be extracted.

In the case of general occlusion, once a scene point is captured by the center view, it is usually occluded in half of the views at most. So the line of occlusion points only affects half of an EPI, like in $EPI_{A,0^\circ}$. Therefore, we divide the EPI into two parts along the rays in opposite directions, named RayEPIs. A pair of RayEPIs are defined as follows:

$$\begin{aligned} RayEPI_{\alpha,x^*,y^*}(s,r_\alpha) &= L\left(x^* + s\cos\alpha,\, y^* + s\sin\alpha,\, r_\alpha\cos\alpha,\, r_\alpha\sin\alpha\right),\\ RayEPI_{\alpha+\pi,x^*,y^*}(s,r_{\alpha+\pi}) &= L\left(x^* + s\cos\alpha,\, y^* + s\sin\alpha,\, r_{\alpha+\pi}\cos\alpha,\, r_{\alpha+\pi}\sin\alpha\right),\\ & s\in\mathbb{R}, r_\alpha\in\mathbb{R}_0^+, r_{\alpha+\pi}\in\mathbb{R}_0^-, \alpha\in[0,\pi), \end{aligned} \tag{7}$$

where (x^*, y^*), α, s and r have the same meaning as their counterparts in Eq. 6. Following the construction of traditional multi-direction EPIs, RayEPIs are also sampled with α of 0, $1/4\pi$, $1/2\pi$ and $3/4\pi$, which results in RayEPIs in eight directions.

For general occlusion, where occlusion exists on one side of the center view, RayEPI in the opposite direction is usually free from occlusion and reliable geometric information can be generated. For example, $EPI_{A,0^\circ}$ is divided into $RayEPI_{A,0^\circ}$ and $RayEPI_{A,180^\circ}$ in Figure 5. $RayEPI_{A,180^\circ}$ is severely occluded, while $RayEPI_{A,0^\circ}$ is not affected by occlusion. Thus, different from the traditional EPI whose information is wasted entirely, the RayEPIs provide as much useful information as possible.

For dense occlusion, at least one occlusion-free RayEPI can be found for geometric information extraction. For example in Figure 5, point *B* is occluded in all the traditional EPIs. However, $RayEPI_{B,0^\circ}$,

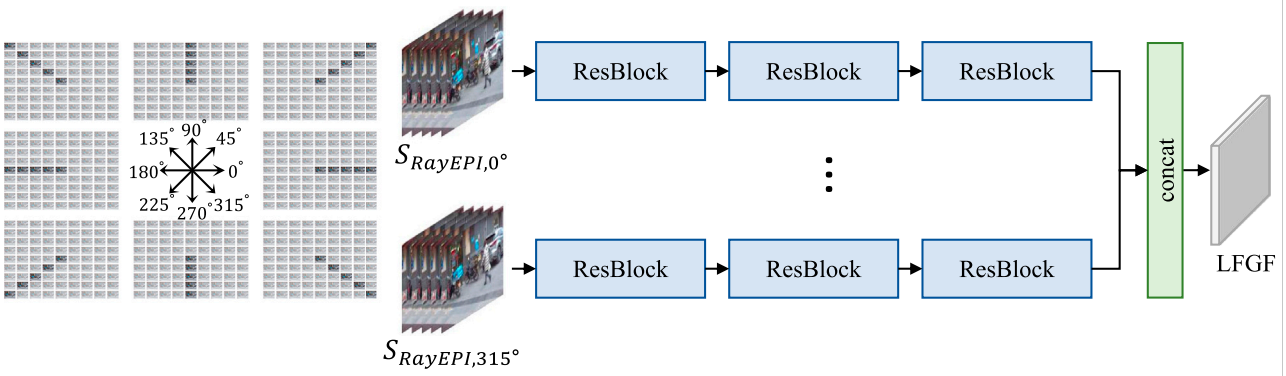

FIGURE 6
LFGF is extracted from RayEPIs in eight directions. In this paper, RayEPIs are constructed by stacking SAIs in certain directions. RayEPIs are separatly fed into three sequential ResBlocks for initial geometric feature extraction. The initial geometric features from eight RayEPIs are then directly concatenated to retain all the useful information.

$RayEPI_{B,45°}$, $RayEPI_{B,270°}$ and $RayEPI_{B,315°}$ are not affected by occlusion. Thus different from traditional EPIs which fail in this case, reliable geometric information can be extracted from RayEPIs.

Algorithm 1. Construction of RayEPIs.

Initialization: $L \in \mathbb{R}^{U\times V\times X\times Y}$
Create empty lists $\{S_{RayEPI,\alpha}\}$ with $\alpha \in \{1/4k\pi | k \in \mathbb{N}, 0 \le k \le 7\}$ to save SAIs in direction α
$u_c = (U+1)/2$
$v_c = (V+1)/2$
$n_{views} = (V+1)/2$
for $k \leftarrow 1$ to n_{views} **do**
  $S_{RayEPI,0}$ append $L_{(u_c,v_c+(k-1))}$
  $S_{RayEPI,1/4\pi}$ append $L_{(u_c-(k-1),v_c+(k-1))}$
  $S_{RayEPI,1/2\pi}$ append $L_{(u_c-(k-1),v_c)}$
  $S_{RayEPI,3/4\pi}$ append $L_{(u_c-(k-1),v_c-(k-1))}$
  $S_{RayEPI,\pi}$ append $L_{(u_c,v_c-(k-1))}$
  $S_{RayEPI,5/4\pi}$ append $L_{(u_c+(k-1),v_c-(k-1))}$
  $S_{RayEPI,3/2\pi}$ append $L_{(u_c+(k-1),v_c)}$
  $S_{RayEPI,7/4\pi}$ append $L_{(u_c+(k-1),v_c+(k-1))}$
end for
Reverse $S_{RayEPI,1/2\pi}$
Reverse $S_{RayEPI,3/4\pi}$
Reverse $S_{RayEPI,\pi}$
Reverse $S_{RayEPI,5/4\pi}$
Output: RayEPI Image stacks $\{S_{RayEPI,\alpha}\}$ for L, with $\alpha \in \{1/4k\pi | k \in \mathbb{N}, 0 \le k \le 7\}$

Following previous works, as shwon in Figure 6, RayEPIs are constructed by stacking view images in certain directions in this paper to reduce consumption in constructing RayEPIs directly with Eq. 7. As depicted in Algorithm 1, the construction of RayEPIs starts from the center view. SAI in each direction α is saved by the corresponding image stack $S_{RayEPI,\alpha}$ view-by-view. Then the orders of images in $S_{RayEPI,1/2\pi}$, $S_{RayEPI,3/4\pi}$, $S_{RayEPI,\pi}$ and $S_{RayEPI,5/4\pi}$ are reversed to ensure the disparities of the same pixels in all the RayEPIs to be identical.

With RayEPIs in eight directions, initial geometric features from different directions are separately extracted through three sequential ResBlocks, the receptive field of which is sufficient to cover the displacements of the pixels in LF. The intial geometric features are then concatenated together to form the final LFGF:

$$LFGF = concat\left(\{ResBlocks\left(RayEPI_{\alpha}\right)\}\right), \alpha \in \{1/4k\pi, \quad k = 0, 1, 2, 3, 4, 5, 6, 7\}, \tag{8}$$

where *ResBlocks* (·) denotes three sequential ResBlocks. The detailed structure of ResBlock is depicted in Figure 2 and Eq. 3. In this way, all the useful geometric information is retained in LFGF.

With the robust geometric information from LFGF, the shapes of objects can be more accurately estimated, which not only benefits classification of the object, but also promotes performance of the segmention along boundaries.

3.3 Architecture of LFRSNet

LFCF contains multi-layer information of the scene, which is beneficial in the classification of occluded objects. And LFGF provides robust geometric information, which boosts performance along occlusion boundaries. In this subsection, Light Field Robust Segmentation Network is proposed accordingly through adaptive combination of the features with attention mechanism. Based on LFCF and LFGF, LFRSNet produces robust semantic segmentation, which both classifies occluded objects better, but also produces clearer and more accurate occlusion boundaries.

As shown in Figure 7, following previous methods (Zhao et al., 2017; Chen et al., 2018), ResNet101 (He et al., 2016) is

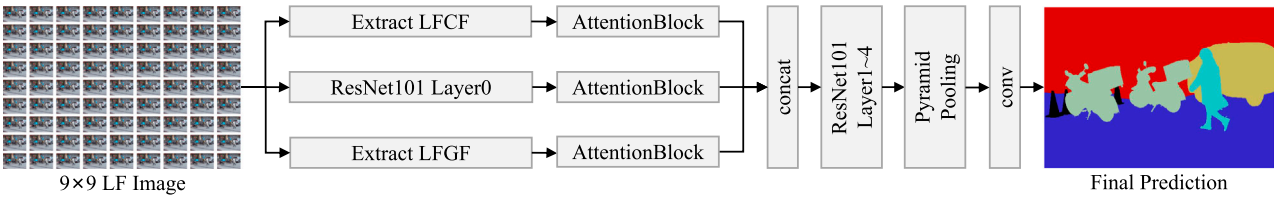

FIGURE 7
The overall architecture of LFRSNet. ResNet101 is used as the backbone of LFRSNet. LFCF and LFGF are respectively extracted according to Eqs 5, 8. The two features are then adaptively concatenated with the initial feature of the center view from Layer0 of ResNet101. The integrated features are fed into the rest Bottleneck Blocks of ResNet101 (Layer1 to Layer4) and the output of Layer4 is further fed into a PPM and another convolutional layer for the final pixel-level prediction.

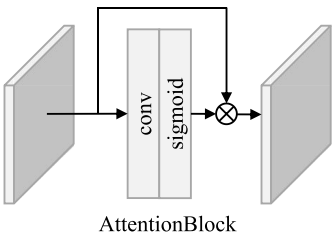

FIGURE 8
The architecture of AttentionBlock used in LFRSNet. To accelerate the processing speed, a simple structure with one convlutional layer followed by sigmoid is adopted for weight map generation.

adoted as the backbone of LFRSNet. The initial feature F_{init} of the center view $L_{(5,5)}$ is extracted with Layer0 of ResNet101:

$$F_{init} = ResNet101_{Layer0}(L_{(5,5)}). \tag{9}$$

LFCF and LFGF are extracted seperately based on Eqs 5, 8.

To adaptively integrate LFCF, LFGF and F_{init}, as shown in Figure 8, a simple Attention Block is adopted:

$$\begin{aligned} LFCF_{atten} &= sigmoid(conv(LFCF)) \circ LFCF, \\ F_{initatten} &= sigmoid(conv(F_{init})) \circ F_{init}, \\ LFGF_{atten} &= sigmoid(conv(LFGF)) \circ LFGF, \end{aligned} \tag{10}$$

where $\circ$ denotes the Hadamard product of two matrices. Then the attention feature maps are concatenated

$$F_{concat} = concat(LFCF_{atten}, F_{initatten}, LFGF_{atten}) \tag{11}$$

and fed into the rest Bottleneck Blocks of ResNet101 (Layer1 to Layer4) for feature fusion and refinement:

$$F_{fused} = ResNet101_{Layer1\sim4}(F_{concat}). \tag{12}$$

In the end, Pyramid Pooling Module (PPM) from (Zhao et al., 2017) is adopted to integrate the global and local clues for prediction with higher reliability. After a final convolutional layer, reliable pixel-level semantic prediction for the center view of LF is produced:

$$P = conv(PPM(F_{fused})), P \in \mathbb{R}^{C \times X \times Y}. \tag{13}$$

C indicates the number of semantic classes to predict. In this paper, C is set to 14.

4 Experimental results

In this section, we first introduce the implementation details of the experiments. Then comparison with the state-of-the-art methods, including methods for single images, videos, RGB-D data and LF, is conducted. Our LFRSNet outperforms other methods in all the metrics. Ablation study is performed at last to verify the contribution of our designs.

4.1 Implementation details

The large-scale LF semantic segmentation dataset UrbanLF (Sheng et al., 2022) is used in the experiments. UrbanLF includes two subsets: UrbanLF-Real and UrbanLF-Syn. The former is captured with LF camera Lytro Illum and the latter is rendered with Blender. UrbanLF-Syn contains ground-truth disparity and depth labels. Experiments are conducted separately on the two subsets to incorporate RGB-D based methods on UrbanLF-Syn in the comparison.

The training data is augmented with random flipping (left-right, up-down), scaling and cropping. Our network is implemented in Pytorch (1.7.0) and trained with one NVIDIA RTX 3090 GPU. Following (Sheng et al., 2022), SGD optimizer is adopted with an initial learning rate of 0.01. Momentum and weight decay are set to 0.9 and 0.0005 respectively. The "poly" learning rate policy is adopted, where the learning rate is multiplied by $(1 - \frac{iter}{maxiter})^{0.9}$ in each iteration. Comparisons are conducted on the center view of LF with pixel Accuracy

TABLE 1 Comparison with the state-of-the-art methods on UrbanLF-Real. Bold texts indicate the best results and italics indicate the second best results. LFRSNet outperforms other methods in all metrics.

Method	DataType	Acc(%) ↑	mAcc(%) ↑	mIoU (%) ↑
PSPNet Zhao et al. (2017)	Single	91.21	83.87	76.34
OCR Yuan et al. (2020)	Single	92.02	85.17	78.60
TDNet Hu et al. (2020)	Video	91.05	83.38	76.48
TMANet Wang H. et al. (2021)	Video	91.04	83.54	75.91
PSPNet-LF Sheng et al. (2022)	LF	92.14	84.86	78.10
OCR-LF Sheng et al. (2022)	LF	*92.51*	*86.31*	*79.32*
LFRSNet	LF	**92.83**	**87.10**	**79.98**

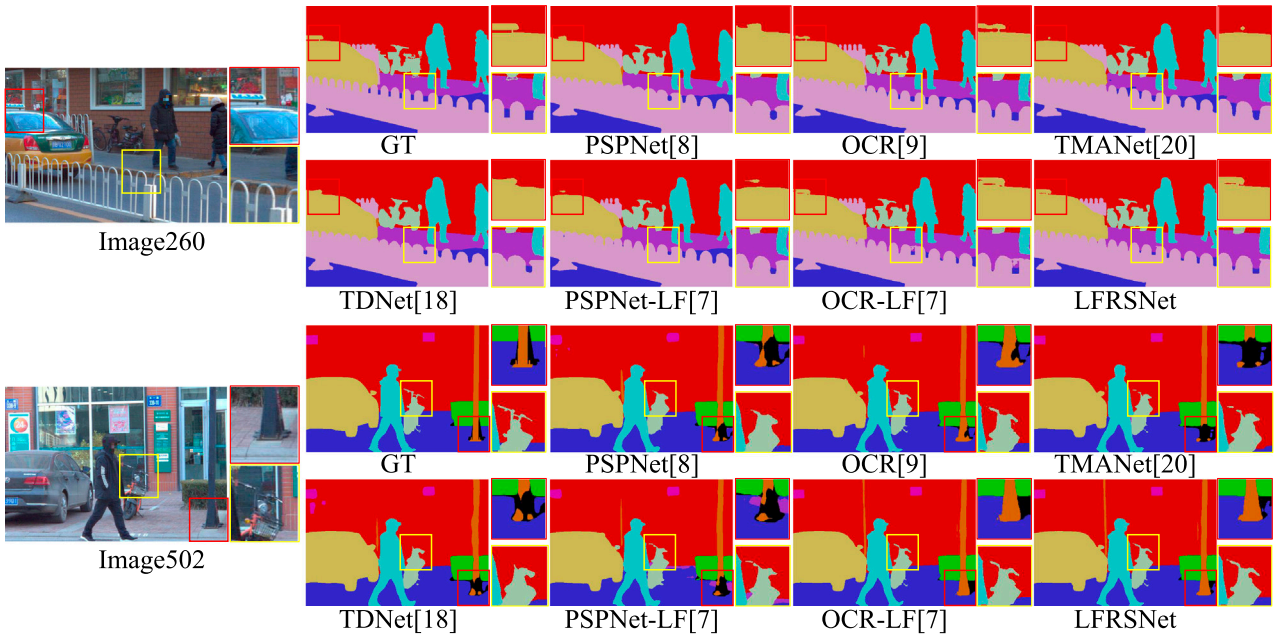

FIGURE 9
Visual results of Image260 and Image502 in UrbanLF-Real. With full exploitation of all the SAIs in LF, LFRSNet improves significantly in the misclassification between neighboring objects with similar colors. The more robust geometric information from RayEPIs also enhances the identification of fine structures.

(Acc ↑), mean pixel Accuracy (mAcc ↑) and mean Intersection-over-Union (mIoU ↑).

4.2 Comparison with the state-of-the-art methods

4.2.1 Comparison on realworld data

Six state-of-the-art semantic segmentation methods are used for comparison on realworld data UrbanLF-Real, including methods for single images (PSPNet (Zhao et al., 2017) and OCR (Yuan et al., 2020)), videos (TDNet (Hu et al., 2020) and TMANet (Wang S et al., 2021)) and LFs (PSPNet-LF (Sheng et al., 2022) and OCR-LF (Sheng et al., 2022)). Note that for video-based methods, SAIs are organized in S-shape, starting from the top-left view and scanning horizontally, to form pseudo videos.

As shown in Table 1, LFRSNet achieves the highest scores on all the metrics. Video-based methods (TDNet and TMANet) perform inferior to PSPNet and OCR. The reason lies in the relatively small baseline of LF cameras. Difference between adjacent SAIs is much smaller than that between frames of a video. Hence the ability to construct long-range connections cannot be fully exerted. The methods modified for LF data

TABLE 2 Comparison with the state-of-the-art methods on UrbanLF-Syn. Bold texts indicate the best results and italics indicate the second best results. LFRSNet outperforms other methods in all metrics.

Method	DataType	Acc(%) ↑	mAcc(%) ↑	mIoU (%) ↑
PSPNet Zhao et al. (2017)	Single	89.39	84.48	75.78
OCR Yuan et al. (2020)	Single	91.50	86.96	79.36
TDNet Hu et al. (2020)	Video	89.06	83.43	74.71
TMANet Wang H. et al. (2021)	Video	89.47	82.94	74.27
MTINet Vandenhende et al. (2020)	RGB-D	91.24	86.94	79.10
SA-Gate Chen et al. (2020)	RGB-D	*92.10*	87.04	79.53
PSPNet-LF Sheng et al. (2022)	LF	90.55	85.91	77.88
OCR-LF Sheng et al. (2022)	LF	92.01	*87.71*	*80.43*
LFRSNet	LF	**92.32**	**87.94**	**80.87**

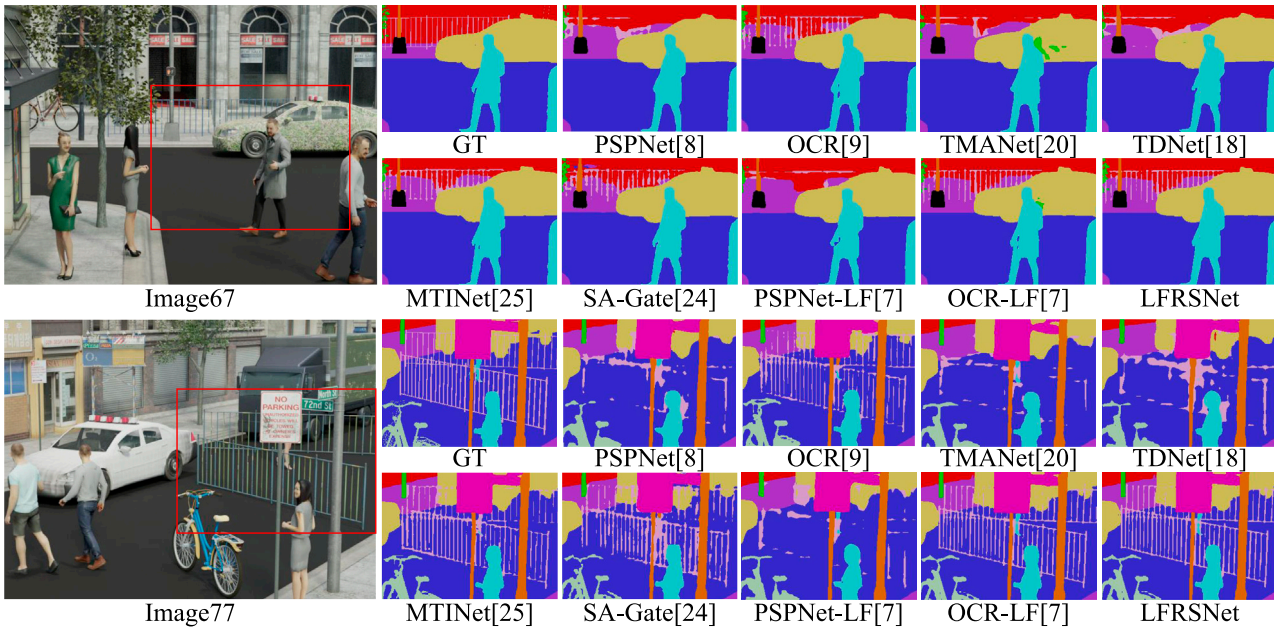

FIGURE 10
Visual results of Image67 and Image77 in UrbanLF-Syn. With the correct depth maps provided by the synthetic data, RGB-D based methods outperform most of the state-of-the-art methods. However, with full exploitation of all the SAIs in the LF, LFCF provides a more comprehensive understanding of the multi-layer scene for LFRSNet. LFGF from RayEPIs also provides robust geometric information of complex occlusion. Thus the classic fine structure of the thin fence is better handled, even compared with RGB-D based methods.

outperform their single-image version with an additional geometric information extraction branch. However, on the one hand, they only exploit less than half of the SAIs in an LF. On the other hand, the use of EPI restricts their performance along occlusion boundaries. With RayEPIs in eight directions and full exploitation of all the SAIs, LFRSNet produces the best results.

As shown in Figure 9, Image260 and Iamge502 of UrbanLF-Real are used as visual examples. In the red boxes of Image260, a taxi light is on the roof of the car. It shares the same color with the window frame in the background. Hence most of the methods fails to fully find the contour of it. Benefitting from the high-resolution feature space, OCR manages to partly recover the taxi light. And with introduction of geometric information from LF, OCR-LF produces better reconstruction. LFGF, compared with the feature from EPI in OCR-LF, further provides geometric information more robust to occlusion. Hence LFRSNet draws the most complete contour of the taxi light. As for the areas in the yellow boxes of Image 260, LFCF extracts multi-layer contextual information. Especially for this scenerio, the area occluded by the

TABLE 3 Investigation of LFRSNet with different designs.

Method	Acc(%) ↑	mAcc(%) ↑	mIoU (%) ↑
LFRSNet	92.83	87.10	79.98
Without LFCF	92.40 (−0.43)	86.12 (−0.98)	79.13 (−0.85)
Without LFGF	92.62 (−0.21)	86.56 (−0.54)	79.63 (−0.35)
Without LFCF&LFGF	91.30 (−1.53)	83.93 (−3.17)	76.54 (−3.44)

thin fence can almost all be observed from other SAIs. Therefore LFRSNet segments the fence clearly and doesnot diffuse it to the backgroud like OCR-LF does.

In Image502, a traffic cone is placed behind the pole of the street lamp in the red boxes. Methods for videos (TMANet and TDNet) fail to separate the two objects. Leveraging geometric information from multi-direction EPIs, OCR-LF clearly reconstructs the bottom of the street lamp. And benefitting from the multi-layer information from LFCF, LFRSNet further reduces the area of the ground that is misclassified to the same class as the traffic cone (class *Others*). Based on the comprehensive geometric information from multi-direction RayEPIs, the complex occlusion boundaries of the handlebar in the yellow boxes are also identified more completely.

4.2.2 Comparison on synthetic data

Because UrbanLF-Syn provides ground-truth depth and disparity maps, other than previous six methods adopted on realworld data, two state-of-the-art RGB-D based methods, MTINet (Vandenhende et al., 2020) and SA-Gate (Chen et al., 2020), are introduced in the comparison.

As shown in Table 2, LFRSNet outperforms all the other methods. With reliable depth information, RGB-D based method SA-Gate achieves second best performance in Acc. However, depth information can only describe the superficial layer of the scene. Without use of other views, the improvement is not uniform across semantic classes. Hence OCR-LF surpasses SA-Gate in both mAcc and mIoU. It also proves that the extracted geometric features from multiple perspectives (like EPIs and RayEPIs) contains geometric information more-comprehensive than depth maps.

Iamge67 and Image77 of UrbanLF-Syn is used as visual examples in Figure 10. The red box circles a classic hard area of the fences with complex occlusion and fine structures. Methods for videos (TMANet and TDNet) and PSPNet completely fail in the reconstruction of the thin fence. With geometric information from EPIs, PSPNet-LF corrects some obvious errors in classification of large objects but is still unable to reconstruct the fence. Taking advantage of correct depth information, sensitivity to fine structures is greatly improved in RGB-D based methods, MTINet and SA-Gate, which indeed outperform most methods. However, the comprehensive perception of the geometric and contextual information of the scene introduced by LFCF and LFGF make LFRSNet better handle this complex structure. The thin fences are identified more completely by LFRSNet. Especially in Image67, although provided with depth ground-truth, the similar color of the fence and the background still poses difficulties to RGB-D based methods. Whereas with observation from different views helps LFRSNet to better seperate the slender rods of the fence from the sidewalk.

4.3 Ablation study

In this subsection, LFRSNet is compared with its variations different in architecture to investigate the potential benefits of the proposed LF features. The experiments are performed on UrbanLF-Real.

In LFRSNet, LFCF is constructed to represent the multi-layer information of the scene and LFGF contains comprehensive geometric information. As shown in Table 3, three variants are investigated: *without LFCF*, *without LFGF* and *without LFCF&LFGF*. In *without LFCF* and *without LFGF*, LFCF and LFGF are removed from LFRSNet respectively. And in *without LFCF&LFGF*, semantic segmentation is performed with neither LFCF nor LFGF.

The architecture of *without LFCF* is much simpler compared with PSPNet-LF. However, by replacing EPI with RayEPI, *without LFCF* outperforms PSPNet-LF in all three metrics (0.26% in Acc, 1.26% in mAcc, 1.03% in mIoU). Different from previous methods, all SAIs in LF are exploited in LFRSNet. The improvement of *without LFGF* over *without LFCF* proves that the multi-layer contextual information from all views in LF contributes more compared with geometric information from RayEPIs. Removing both features, LFRSNet degenerates to PSPNet with attention mechanism, which performs slightly superior to the original PSPNet. *Without LFCF* significantly surpasses *without LFCF&LFGF* (1.10% in Acc, 2.19% in mAcc, 2.59% in mIoU), verifying the effectiveness of geometric information in LFGF.

5 Conclusion

In this paper, LFRSNet is designed to fully exploit LF in semantic segmentation. Sepcifically, LFCF is introduced based on an angular-distance-aware context-perception mechanism for multi-layer information extraction. LFGF is proposed based on RayEPI for robust and comprehensive geometric information representation. Extensive experimental results show that our method outperforms other state-of-the art methods, especially in the edge regions. As demonstrated by the experiments, LFGF and LFCF contains abundant geometric and contextual information, which greatly facilitates understanding of the realworld scenes, especially for smart city applications that contains complex scenarios. In the future, we will further try to apply the informative features of LF to other smart city tasks, like 3D reconstruction, understanding of remote sensing images, etc.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

DY finished most of the writing and part of the experiments. The idea and design of the method were proposed together by DY and TZ. ShW and SiW reproduced state-of-the-art methods in Section 4. ZX conducted part of the ablation experiments. All authors contributed to the article and approved the submitted version.

Funding

This study is partially supported by the National Key R&D Program of China (No.2019YFB2102200), the National Natural Science Foundation of China (No.61872025), and the Science and Technology Development Fund, Macau SAR(File no.0001/2018/AFJ) and the Open Fund of the State Key Laboratory of Software Development Environment (No. SKLSDE2021ZX-03). Thank you for the support from HAWKEYE Group.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Carreira, J., Pătrăucean, V., Mazare, L., Zisserman, A., and Osindero, S. (2018). "Massively parallel video networks," in European Conference on Computer Vision (ECCV), 680–697.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40 (4), 834–848. doi:10.1109/tpami.2017.2699184

Chen, L.-Z., Lin, Z., Wang, Z., Yang, Y.-L., and Cheng, M.-M. (2021). Spatial information guided convolution for real-time rgbd semantic segmentation. *IEEE Trans. Image Process.* 30, 2313–2324. doi:10.1109/tip.2021.3049332

Chen, X., Lin, K.-Y., Wang, J., Wu, W., Qian, C., Li, H., et al. (2020). "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in European Conference on Computer Vision (ECCV), 561–577.

Frangez, V., Salido-Monzú, D., and Wieser, A. (2022). Assessment and improvement of distance measurement accuracy for time-of-flight cameras. *IEEE Trans. Instrum. Meas.* 71, 1–11. doi:10.1109/tim.2022.3167792

Gao, B., Mai, Z., Tu, H., and Duh, H. (2022). Effects of transfer functions and body parts on body-centric locomotion in virtual reality. *IEEE Trans. Vis. Comput. Graph.*, 1. doi:10.1109/tvcg.2022.3169222

Gu, X., Li, S., Yi, K., Yang, X., Liu, H., and Wang, G. (2022). Role-exchange playing: An exploration of role-playing effects for anti-bullying in immersive virtual environments. *IEEE Trans. Vis. Comput. Graph.*, 1–15. doi:10.1109/tvcg.2022.3184986

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.

Hu, P., Caba, F., Wang, O., Lin, Z., Sclaroff, S., and Perazzi, F. (2020). "Temporally distributed networks for fast video semantic segmentation," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8815–8824.

Hu, X., Yang, K., Fei, L., and Wang, K. (2019). "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in IEEE International Conference on Image Processing (ICIP), 1440–1444.

Huang, Z., Hu, X., Xue, Z., Xu, W., and Yue, T. (2021). "Fast light-field disparity estimation with multi-disparity-scale cost aggregation," in IEEE/CVF International Conference on Computer Vision (ICCV), 6300–6309.

Jain, S., Wang, X., and Gonzalez, J. E. (2019). "Accel: A corrective fusion network for efficient semantic segmentation on video," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8858–8867.

Ji, J., Shi, R., Li, S., Chen, P., and Miao, Q. (1938). Encoder-decoder with cascaded crfs for semantic segmentation. *IEEE Trans. Circuits Syst. Video Technol.* 31 (5), 1926. doi:10.1109/tcsvt.2020.3015866

Li, F., Song, M., Xue, B., and Yu, C. (2022). Abundance estimation based on band fusion and prioritization mechanism. *IEEE Trans. Geosci. Remote Sens.* 60, 1–21. doi:10.1109/tgrs.2022.3187867

Li, Y., Lai, X., Wang, M., and Zhang, X. (2022). C-SASO: A clustering-based size-adaptive safer oversampling technique for imbalanced sar ship classification. *IEEE Trans. Geoscience Remote Sens. (TGRS)* 60, 1–12. doi:10.1109/TGRS.2022.3187751

Shelhamer, E., Long, J., and Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (4), 640–651. doi:10.1109/tpami.2016.2572683

Sheng, H., Cong, R., Yang, D., Chen, R., Wang, S., and Cui, Z. (2022). Urbanlf: A comprehensive light field dataset for semantic segmentation of urban scenes. *IEEE Trans. Circuits Syst. Video Technol.*, 1. doi:10.1109/tcsvt.2022.3187664

Shin, C., Jeon, H.-G., Yoon, Y., Kweon, I. S., and Kim, S. J. (2018). "Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4748–4757.

Vandenhende, S., Georgoulis, S., and Van Gool, L. (2020). "Mti-net: Multi-scale task interaction networks for multi-task learning," in European Conference on Computer Vision (ECCV), 527–543.

Wang, H., Wang, W., and Liu, J. (2021). "Temporal memory attention for video semantic segmentation," in IEEE International Conference on Image Processing (ICIP), 2254–2258.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., et al. (2021). Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 43 (10), 3349–3364. doi:10.1109/tpami.2020.2983686

Wang, L., Wang, Y., Liang, Z., Lin, Z., Yang, J., An, W., et al. (2019). "Learning parallax attention for stereo image super-resolution," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12 242–312 251.

Wang, S., Sheng, H., Zhang, Y., Wu, Y., and Xiong, Z. (2021). "A general recurrent tracking framework without real data," in IEEE/CVF International Conference on Computer Vision (ICCV), 13 199–213 208.

Wang, H., Zhang, S., Zhang, X., Zhang, X., and Liu, J. (2022). Near-optimal 3-d visual coverage for quadrotor unmanned aerial vehicles under photogrammetric constraints. *IEEE Trans. Ind. Electron.* 69 (2), 1694–1704. doi:10.1109/tie.2021.3060643

Wang, Y., Wang, L., Wu, G., Yang, J., An, W., Yu, J., et al. (2022). Disentangling light fields for super-resolution and disparity estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1. doi:10.1109/tpami.2022.3152488

Wang, W., and Neumann, U. (2018). "Depth-aware cnn for rgb-d segmentation," in European Conference on Computer Vision (ECCV), 144–161.

Yuan, Y., Chen, X., and Wang, J. (2020). "Object-contextual representations for semantic segmentation," in European Conference on Computer Vision (ECCV), 173–190.

Zhang, S., Chang, S., and Lin, Y. (2021). End-to-end light field spatial super-resolution network using multiple epipolar geometry. *IEEE Trans. Image Process.* 30, 5956–5968. doi:10.1109/tip.2021.3079805

Zhang, Y., Sheng, H., Wu, Y., Wang, S., Lyu, W., Ke, W., et al. (2020). Long-term tracking with deep tracklet association. *IEEE Trans. Image Process.* 29, 6694–6706. doi:10.1109/tip.2020.2993073

Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., and Yang, J. (2019). "Pattern-affinitive propagation across depth, surface normal and semantic segmentation," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4101–4110.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). "Pyramid scene parsing network," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6230–6239.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., et al. (2021). "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6877–6886.

Zhu, X., Xiong, Y., Dai, J., Yuan, L., and Wei, Y. (2017). "Deep feature flow for video recognition," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4141–4150.

Zhuang, J., Wang, Z., and Wang, B. (2021). Video semantic segmentation with distortion-aware feature correction. *IEEE Trans. Circuits Syst. Video Technol.* 31 (8), 3128–3139. doi:10.1109/tcsvt.2020.3037234