UrbanLF: A Comprehensive Light Field Dataset for Semantic Segmentation of Urban Scenes

Hao Sheng, *Member, IEEE*, Ruixuan Cong, Da Yang, Rongshan Chen, Sizhe Wang, and Zhenglong Cui

***Abstract*— As one of the fundamental technologies for scene understanding, semantic segmentation has been widely explored in the last few years. Light field cameras encode the geometric information by simultaneously recording the spatial information and angular information of light rays, which provides us with a new way to solve this issue. In this paper, we propose a high-quality and challenging urban scene dataset, containing 1074 samples composed of real-world and synthetic light field images as well as pixel-wise annotations for 14 semantic classes. To the best of our knowledge, it is the largest and the most diverse light field dataset for semantic segmentation. We further design two new semantic segmentation baselines tailored for light field and compare them with state-of-the-art RGB, video and RGB-D-based methods using the proposed dataset. The outperforming results of our baselines demonstrate the advantages of the geometric information in light field for this task. We also provide evaluations of super-resolution and depth estimation methods, showing that the proposed dataset presents new challenges and supports detailed comparisons among different methods. We expect this work inspires new research direction and stimulates scientific progress in related fields. The complete dataset is available at https://github.com/HAWKEYE-Group/UrbanLF.**

***Index Terms*— Semantic segmentation, light field, dataset, super-resolution, depth estimation.**

I. Introduction

SEMANTIC segmentation that aims to predict semantic labels for every pixel in the image, has attracted great attention as a basic task of computer vision. It splits an image into some coherent semantically meaningful regions and plays an important role in visual scene understanding. Accurate and reliable semantic segmentation has benefited many popular and challenging applications like autonomous driving [1], medical image analysis [2] and geographic information system [3].

Existing semantic segmentation methods can be divided into four categories based on the type of input data. Early works [4]–[7] focus on exploiting the visual cues from a single RGB image with hand-crafted features or some feature learning techniques. With the rise of the requirement for real-time applications, semantic segmentation has also been applied to video sequences [8]–[11]. The key is to make effective use of temporal context in the video to balance the trade-off between quality and speed. Some other approaches utilize geometric and structural information in 3D data to further improve accuracy, generally falling into two categories. One is RGB-D semantic segmentation [12]–[15], that leverages depth data to recalibrate the RGB feature. The other concentrates on extracting more representative features directly from 3D point sets, dubbed as point-cloud semantic segmentation [16]–[19].

However, there are some defects in these algorithms. For single image and video-based methods, the limited information in RGB images does not allow to fully analyze geometric constraints, making it difficult to show promising results in challenging scenes with low color discrimination and complex occlusion. For RGB-D-based methods, depth maps captured by sensors are partially noisy and hard to accurately align with the RGB pixels, which may result in undesirable results. The limited distance measurement range can make this phenomenon more obvious in outdoor scenes. For point-cloud-based methods, the dataset size is generally small since acquiring and annotating points is much more complicated than images, restricting the development of deep learning methods. In addition, the quality of the data can not be well controlled.

In this paper, we propose a new comprehensive light field (LF) dataset named UrbanLF for semantic segmentation. A 4D LF [20] not only contains intensity but also direction of light rays. The additional directional information implicitly defines the geometry of the scene. Fig. 1 parametrizes the LF as $L(x, y, s, t)$ with two parallel planes, where (x, y) and (s, t) are the spatial coordinates and angular coordinates respectively. Theoretically speaking, LF benefits semantic segmentation in several ways. First, LF can be seen as sub-aperture images where the viewpoints are arranged on a regular grid in angular plane. Some occluded pixels in target view can be obtained from other views. Second, LF contains depth information [21] which has been proven useful for this problem.

Manuscript received 18 January 2022; revised 26 April 2022 and 17 June 2022; accepted 26 June 2022. Date of publication 30 June 2022; date of current version 28 October 2022. This work was supported in part by the National Key R&D Program of China under Grant 2018YFB2100500; in part by the National Natural Science Foundation of China under Grant 61872025; in part by the Science and Technology Development Fund, Macau SAR, under File 0001/2018/AFJ; in part by the Open Fund of the State Key Laboratory of Software Development Environment under Grant SKLSDE-2021ZX-03; and in part by the HAWKEYE Group. This article was recommended by Associate Editor S. T. Kim. *(Corresponding author: Ruixuan Cong.)*

The authors are with the State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, also with the Beihang Hangzhou Innovation Institute Yuhang, Xixi Octagon City, Yuhang District, Hangzhou, 310023, China, and also with the Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR 999078, China (e-mail: shenghao@buaa.edu.cn; congrx@buaa.edu.cn; da.yang@buaa.edu.cn; rongshan@buaa.edu.cn; sizhewang@buaa.edu.cn; zhenglong.cui@buaa.edu.cn).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCSVT.2022.3187664.

Digital Object Identifier 10.1109/TCSVT.2022.3187664

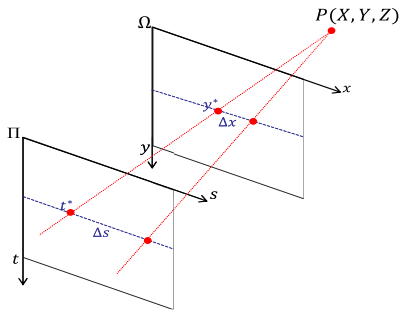


Fig. 1. The 4D light field parametrization with two parallel planes. (x, y) are the coordinates in spatial plane Ω. (s, t) are the coordinates in angular plane Π.

Third, the pixel parallax of LF is available for reflection layer removal [22] and rain streak removal [23] so as to reduce the performance degradation of image understanding.

Standard benchmarks [24], [25] have proven their importance for the development in the respective fields. They can offer guidance on research by giving detailed evaluations and objective comparisons of different methods. Moreover, driven by the big success of deep learning, most top-performing methods are nowadays built upon deep neural networks. A major factor is the availability of large-scale datasets which allow networks to develop full potential. In brief, it is necessary to create a large benchmark dataset to support LF semantic segmentation.

By far, existing LF datasets almost do not contain annotations for semantic segmentation, hindering the development of related field. Commercial plenoptic cameras that can capture LF in a single shot are available and continuous progress on rendering engine guarantees the quality of synthetic data. With the help of these imaging technologies, we collect 1074 LF images with complex urban scenes as well as pixel-wise annotations. The dataset contains not only real-world samples but also synthetic samples. Besides, we specifically design two baselines for LF semantic segmentation and provide a performance evaluation of semantic segmentation methods on the proposed dataset. The results show our baselines achieve better performance than non-LF-based methods, demonstrating the effectiveness of LF for this problem. We also try using the dataset for other fields such as super-resolution and depth estimation. We hope that our dataset can make a contribution to LF community.

In summary, the contributions of our work are as follows:

- We create a large-scale LF dataset called UrbanLF, including 824 real-world samples and 250 synthetic samples with annotations into 14 classes. To the best of our knowledge, it is the first time to propose such a large LF dataset for semantic segmentation.
- We propose two new LF semantic segmentation baselines and provide a systematic experiment on UrbanLF to compare with state-of-the-art RGB, video and RGB-D-based methods. The solid performance of our baselines confirms the superiority of LF to this problem.
- We design a comparative experiment to explore the effectiveness of synthetic data. The experimental results demonstrate that synthetic data can complement real-world data to boost model performance.
- We provide other experiments to explore the potential of UrbanLF for other tasks and give a comprehensive analysis to show the advantages and drawbacks of different methods. The results show that UrbanLF can also be used for super-resolution and depth estimation.

II. Related Work

A. Light Field Datasets

In the past few years, LF has gradually developed into one of the mainstream research fields of the computer vision community. Owing to the potential capabilities from additional directional information of the light, a large variety of fields have tried using LF as input rather than a single image, introducing a series of datasets that can be classified into real-world LF captured by a camera array, a gantry or a plenoptic camera and synthetic LF by Blender [26] or other software. In some applications like depth estimation [24], [27]–[29], super-resolution [30]–[32], saliency detection [33]–[35] and view synthesis [36], [37], LF datasets have been widely used with remarkable results. While in other applications like quality assessment [38], [39], video processing [40] and intrinsic decomposition [41], this attempt still is in preliminary stage with initial success. The summary of these LF datasets is shown in Table I.

Very recently, [42] proposes the first LF dataset for semantic segmentation. It provides 400 real-world macropixel images and corresponding central view images with annotations for 3 foreground objects. The small dataset size and semantic class number constrain its application. As a comparison, our UrbanLF has more samples and richer annotations, which allows the deep learning model well generalizing and presents more challenges to this field.

B. Semantic Segmentation Datasets

Various types of datasets have been proposed for semantic segmentation and gradually improved in terms of size, annotation richness, scene variability and complexity. These datasets play an important role in the overall progress in this field.

CamVid [51] is the first collection of videos with semantic labels. It includes five sequences captured from the perspective of a driving automobile. Cityscapes [25] contains a larger set of stereo videos recorded in street scenes. The inaccurate depth maps obtained through stereo image pairs are rarely used for RGB-D semantic segmentation. It also has 20000 weakly annotated images as extra training data. Mapillary Vistas [52] collects single images captured at various conditions by different devices. It covers a wide variety of street scenes and is only used for RGB semantic segmentation. SYNTHIA [53] is a synthetic collection of photo-realistic images rendered from a virtual city created by Unity. It has 3 subsets with different types of data that can be selected depending on the needs. NYUDv2 [54] and SUNRGBD V1 [55] are widely used for RGB-D semantic segmentation. The former consists of RGB-D images taken from the Kinect and the latter combines images from [54], [56], [57] with new samples captured from 4 different sensors. As public standard point-cloud datasets, Semantic3D [58] consists of dense and complete points from

TABLE I

THE SUMMARY OF DIFFERENT LIGHT FIELD DATASETS. **R** SIGNIFIES THE NUMBER OF REAL SCENES. **S** SIGNIFIES THE NUMBER OF SYNTHETIC SCENES

Dateset	Year	Device	Num(R/S)	Application	Detail
LFSD [33]	2014	Lytro I	100/-	Saliency Detection	It provides 60 indoor and 40 outdoor samples with ground truth saliency map.
LIFFAD [43]	2014	Lytro I	4826/-	Face Recognition	It provides normal face samples from 80 subjects and corresponding artefact face samples generated by photo print and electronic display.
HCI-New [24]	2016	Blender	-/28	Depth Estimation	It provides 4 stratified, 4 training, 16 additional, 4 test(no depth) samples in high and low resolution. It focuses on five challenges, namely occlusion boundaries, fine structures, low texture, smooth surfaces and camera noise.
Heber et al. [27]	2016	POV-Ray	-/25	Depth Estimation	It provides a random scene generator to generate an arbitrary amount of data with ground truth depth as required.
Johannsen et al [44]	2016	Lytro II Blender	4/24	Scene Reconstruction	It provides ground truth depth and camera orientation. The scenes are challenging with reflection and partially transparent occlusion.
Wang et al. [45]	2016	Lytro II	1200/-	Material Recognition	It provides images from 12 material categories each with 100 samples. The images are classified and labeled per pixel by Photoshop.
EPFL [30]	2016	Lytro II	118/-	Super Resolution	It provides 10 categories of images, illustrating specific aspects of LF imaging.
STF-Lytro [31]	2016	Lytro II	353/-	Super Resolution	It provides natural images that are divided into 9 categories.
Kalantari et al. [36]	2016	Lytro II	130/-	View Synthesis	It provides 100 training and 30 test samples. Some of them are from STF-Lytro.
MPI [38]	2017	Gantry	5/9	Quality Assessment	It applies 4 distortions with different severity levels, resulting in 350 LF images with quality scores. Canon is fixed on a motorized linear stage for real samples. Cameras with off-axis asymmetric frustums are used for synthetic samples.
InterDigital [40]	2017	Array	12/-	Video Processing	It provides a set of synchronized sequences captured by a 4x4 camera rig at 30fps.
Srinivasan et al. [37]	2017	Lytro II	3343/-	View Synthesis	It provides images of flowers and plants with complex occlusions and wide ranges of relative depths which are randomly split into 3243 for training and 100 for testing.
DDFF 12-Scene [28]	2018	Lytro II	720/-	Depth Estimation	It provides 600 samples from six scenes for training and 120 samples from other six scenes for test. A RGB-D sensor is fixed on camera for ground truth depth.
Alperovich et al. [41]	2018	Blender	-/175	Intrinsic Decomposition	It provides a custom random light field generator that in theory can synthesize an arbitrary amount of LF images with diffuse and specular intrinsic components.
INRIA [32]	2018	Lytro I Lytro II	109/-	Super Resolution	It provides 63 samples by Lytro I and 46 samples by Lytro II.
MVLF [46]	2019	Lytro II	850/-	Feature Detection	It contains between 3 and 5 LF with different camera poses in each scene, resulting in 4211 samples in total. Images are challenging with Lambertian and non-Lambertian surfaces, occlusion, specularity, subsurface scattering, fine detail, and transparency.
6-DOF Blur [47]	2019	Lytro II UnrealCV	200/200	Light Field Debluring	It provides 360 training samples and 40 test samples. A camera motion model is applied to training samples to create the blurry dataset.
Xu et al. [48]	2019	Lytro II Array	49/-	Object Recognition	It provides images with 7 transparent objects placed in 7 different background scenes. Each combination is captured by 2 devices.
SLFD [29]	2019	Blender	-/53	Depth Estimation	SLFD is a sparse dataset with a disparity range [-20,20] pixels.
DLFD [29]			-/43		DLFD is a dense dataset with a disparity range [-4,4] pixels.
Sintel [49]	2020	Blender	-/2	Scene Flow	It is a LF video dataset. Each sequence consists of 50 frames with ground truth optical flow and disparity map.
HFUT [34]	2020	Lytro II	640/-	Saliency Detection	It is a larger, higher-quality and more challenging dataset with more variations in illuminance, scale, and position as well as more regions for saliency annotations.
DUTLF [35]	2020	Lytro II	4204/-	Saliency Detection	It is a versatile dataset with large-scale, high diversity and broad coverage. It contains 2957 training samples and 1247 test samples with 10 salient objects categories.
PE-LFID [39]	2020	Gantry	5/9	Quality Assessment	It uses all light field scenes from [38]. 4 perceptual encryption methods with 6 different levels of encryption are employed to generate encrypted light field images.
LFDOF [50]	2021	Lytro II	839/-	All-in-focus Restoration	It first selects 839 LF images from [46], then generates 15 spatially-varying defocused images for each scene, resulting in 11986 defocused and all-in-focus image pairs.
Chen et al. [42]	2021	Raytrix-R8 Lytro II	400/	Semantic Segmentation	It provides central view images from 400 real-world scenes and encodes sub-aperture images into macropixel images at different spatial and angular resolutions with annotations for 3 objects. Some scenes are from [45] and [34].
UrbanLF		Lytro II Blender	824/250	Semantic Segmentation	It provides annotations for 14 semantic classes in urban scenes. For synthetic samples, it also provides ground truth depth. It is split into training, validation and test set at a ratio of 7:1:2.

urban and rural scenes and SensatUrban [59] includes richly annotated points from cities with available RGB color. The summary of these datasets is shown in Table II.

In contrast, our UrbanLF has some distinguishable features. It includes a large number of densely and regularly sampled LF images which contribute to new LF semantic segmentation methods. It has two kinds of data with a small domain gap, in which the synthetic samples can be regarded as a good supplement to real-world samples and provide accurate depth information for RGB-D semantic segmentation.

C. Semantic Segmentation Algorithms

1) RGB Semantic Segmentation: Semantic segmentation has reached a new stage with the introduction of FCN [60] that leverages convolutional layers instead of fully connected layers to get the final predictions. Standard FCN segmentation model utilizes the encoder-decoder structure so as to split the task into two stages. Firstly, the encoder uses ConvNets like ResNet [61] to encode semantic information into feature maps, then the decoder recovers the prediction details gradually through the context information. In order to further improve

TABLE II
THE SUMMARY OF DIFFERENT DATASETS FOR SEMANTIC SEGMENTATION. **E** SIGNIFIES THE NUMBER OF CLASS FOR EVALUATION

Dateset	Year	Type	Resolution	Num (train/val/test)	Class(E)	Data				
						image	video	depth	point cloud	light field
CamVid [51]	2009	real outdoor 2D	960×720	701 (367/101/233)	32(11)	✓	✓	×	×	×
NYUDv2 [54]	2012	real indoor 3D	640×480	1449 (795/-/654)	894(40)	✓	×	✓	×	×
SUNRGBD V1 [55]	2015	real indoor 3D	561×427, 591×441 681×531, 730×530	10335 (5285/-/5050)	63(37)	✓	×	✓	×	×
Cityscapes [25]	2016	real outdoor 3D	2048×1024	5000 (2975/500/1525)	30(19)	✓	✓	✓	×	×
SYNTHIA-Rand [53]	2016	synthetic outdoor 3D	960×720	13407	11(11)	✓	×	✓	×	×
SYNTHIA-Rand-Cityscapes [53]				9000	23(23)	✓	×	✓	×	×
SYNTHIA-Seq [53]				56000	13(13)	✓	✓	✓	×	×
Mapillary Vistas [52]	2017	real outdoor 2D	1920×1080, etc.	25000 (18000/2000/5000)	124(116)	✓	×	×	×	×
Semantic3D [58]	2017	real outdoor 3D	×	30 (15/-/15)	9(8)	×	×	✓	✓	×
SensatUrban [59]	2021	real outdoor 3D	×	43 (30/7/6)	31(13)	×	×	✓	✓	×
UrbanLF-Real		real outdoor 4D	623×432×9×9	824 (580/80/164)	14(14)	✓	×	×	×	✓
UrbanLF-Syn		synthetic outdoor 4D	640×480×9×9	250 (172/28/50)	14(14)	✓	×	✓	×	✓

accuracy, some approaches focus on solving the problem of limited receptive field. Yu *et al.* [6] exploit atrous convolution to enlarge the receptive field while keeping the resolution of the feature maps to preserve the spatial information. HRNet [62] generates reliable high-resolution features through repeatedly fusing the representations from multi-resolution convolution streams. Other approaches achieve this by capturing multi-scale context information. PSPNet [5] proposes the pyramid pooling module (PPM) that adopts average pooling layers with different scales. DeepLabV2 [63] proposes the atrous spatial pyramid pooling (ASPP) that adopts atrous convolutions with different rates. Ca-crfs Net [64] uses spatial pyramid pooling to ensemble multi-scale features and proposes cascaded conditional random fields to learn boundary information. Besides, some recent works combine Transformer [65] with segmentation to achieve state-of-the-art performance. SETR [66] is a new segmentation model that replaces the traditional stacked convolution layers with a pure transformer. In OCR [67], a transformer encoder-decoder framework is used to rephrase the object-contextual representation scheme.

2) Video Semantic Segmentation: Video semantic segmentation aims to generate real-time predictions for each frame. The most straightforward approach is to apply an RGB semantic segmentation model to each frame. However, this strategy brings an excessive computational burden. Existing approaches concentrate on exploring the temporal relation between video frames to avoid unnecessary computation. One way is to reuse the features from the key frame to current frame. The challenge is how to propagate information robustly. Carreira *et al.* [8] directly reuse stable features extracted from deep layers to share information across frames. [9], [10], [68] apply an optical flow network to guide the propagation process. The other way is to use the same model for each frame and aggregate them through temporal context for better features. TDNet [11] applies several sub-networks to extract sub-feature groups and gets full features via grouped knowledge distillation loss and attention propagation module. TMANet [69] treats past frames as memory and builds long-range temporal context information to enhance the representation power of features from current frame.

3) RGB-D Semantic Segmentation: RGB-D semantic segmentation takes depth data into consideration to achieve better performance. The majority of approaches treat the depth as an additional input of the network. A two-stream network is used to process RGB images for color and texture information as well as depth images for geometry information, then fuses them for final prediction. ACNet [13] proposes a third branch to process and propagate the fusion features from RGB and depth branches. SA-Gate [14] performs feature aggregation and transfers the fusion features back into RGB branch and depth branch to recalibrate information at each stage. ESANet [70] adopts shallow encoder branches and a 3×1 along with a 1×3 convolution for faster inference with competitive results. Some approaches directly incorporate depth data into explicit operations. DCNN [12] proposes depth-aware convolution and depth-aware average pooling to seamlessly incorporate geometry into CNN. SGNet [71] proposes a S-Conv operator that can adaptively adjust the convolution weights and distributions based on the spatial information. Other approaches [15], [72] treat depth data as a supervised signal and use a multi-task learning framework to jointly train segmentation and depth estimation to improve single-task performance.

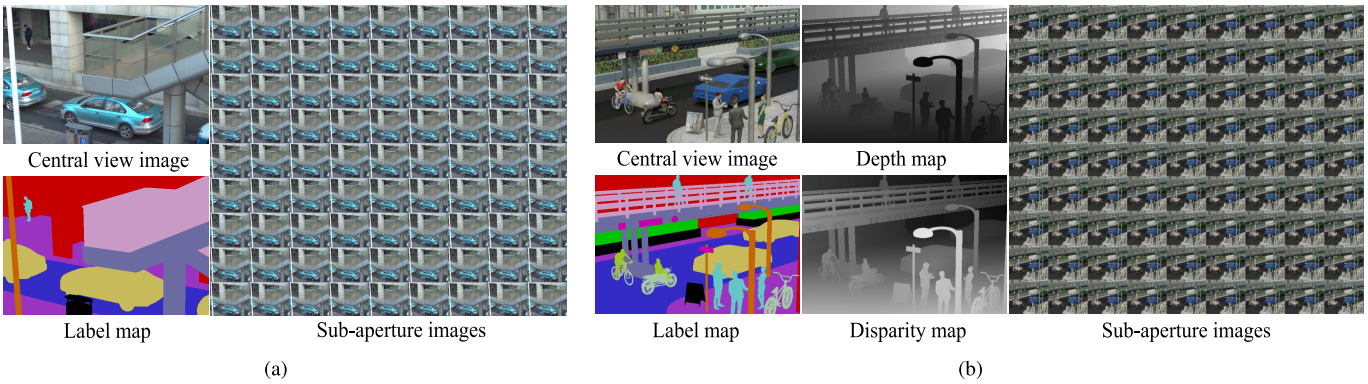


Fig. 2. Data composition of UrbanLF. (a) Real-world samples. The label map is only for central view image and accurate depth information is not available. The angular resolution of sub-aperture images is 9×9. (b) Synthetic samples. The label map, depth map, and disparity map are available for all views. The angular resolution of sub-aperture images is also 9×9.

4) Point-Cloud Semantic Segmentation: Point-cloud semantic segmentation adopts points in 3D space instead of pixels in 2D images and assigns each point with a label. Existing approaches are sorted into three categories according to the data format. 2D-based approaches [16], [73] first convert data into multi-view 2D images, apply 2D CNN architectures to generate downsampled 2D labels and transfer them back to 3D form. Voxel-based approaches [17], [74] first voxelize raw points, apply 3D CNN frameworks for subsequent processing and restore the result to the original 3D point labels. Unlike the aforementioned work, point-based approaches directly process point-cloud without data pre-transformation operation. PointNet [18] applies multi-layer perceptrons to extract point features that aggregate both global and local knowledge and finally outputs per point scores. PointNet++ [75] further explores the local relationship among points to augment features for improving performance.

5) Light Field Semantic Segmentation: Previous works mainly focus on LF segmentation which aims at grouping pixels of different objects without considering semantic information. Wanner *et al.* [76] propose globally consistent multi-label assignment for the first time. Hog *et al.* [77] decrease the running time of Markov random field graph-cut by using a ray-based graph structure. Inspired by superpixel segmentation of 2D images, Zhu *et al.* [78] propose light field superpixel (LFSP) and develop a refocus-invariant LFSP segmentation method. Khan *et al.* [79] segment horizontal and vertical epipolar plane images (EPIs) and combine the angular segmentations in them through view-consistent clustering. Lv *et al.* [80] build a hypergraph representation with LFSPs and present a method via graph-cut optimization. HAMAD *et al.* [81] propose an automatic, adaptive, and view-consistent method based on normalized LF cues and K-means clustering.

[42] is currently the only work that uses LF to explore semantic segmentation. It investigates the advantage of LF angular-spatial information combined with a designed convolutional neural network. The network has an angular model to learn the angle features from macropixel images and applies ASPP to extract multi-scale context features.

III. The UrbanLF Dataset

For providing sufficient data for LF semantic segmentation, we create a new large-scale LF dataset called UrbanLF which includes 824 real-world and 250 synthetic samples. As shown in Fig. 2, each sample is composed of 81 sub-aperture images with an angular resolution of 9×9 and high-quality pixel-wise annotation of central view. The synthetic sample further contains annotation, depth map and disparity map of all views.

We choose urban scenes as the subject of UrbanLF. With the development of urbanization, urban scene understanding has become a research hotspot and has been widely used in advanced applications like crowd detection and traffic analysis. Consequently, it is meaningful to further understand complex urban scenes through the rich information in LF to improve the practical system performance and reliability, offering a good alternative choice for depth data.

A. Data Capture Process

The real-world data are captured by Lytro Illum which is widely used because of its simplicity on carrying and operation. To ensure the quality of the data, we collect LF in the time period with sufficient light so that all objects in the scene can be clearly captured. The density of foreground objects is large to prevent background classes from occupying most of pixels in a single image. We avoid unfavorable weather conditions such as heavy rain or snow because of equipment limitations. We also avoid overly complicated scenes to reduce the adverse impact of unclear structure on annotation due to the limited image resolution. Lytro Illum stores the original data in LF Raw format that is processed with MATLAB Light Field Toolbox [82] in this work. Note that the depth maps obtained from the toolbox are discarded because they are prediction results rather than ground truth data.

The synthetic data are created by Blender using the Cycles and Eevee renderer. We design a virtual urban environment and add various elements in it to acquire images. In order to increase the diversity and complexity of data, each element has multiple models with different textures, colors and shapes and we place many instances in a scene to avoid leaving large

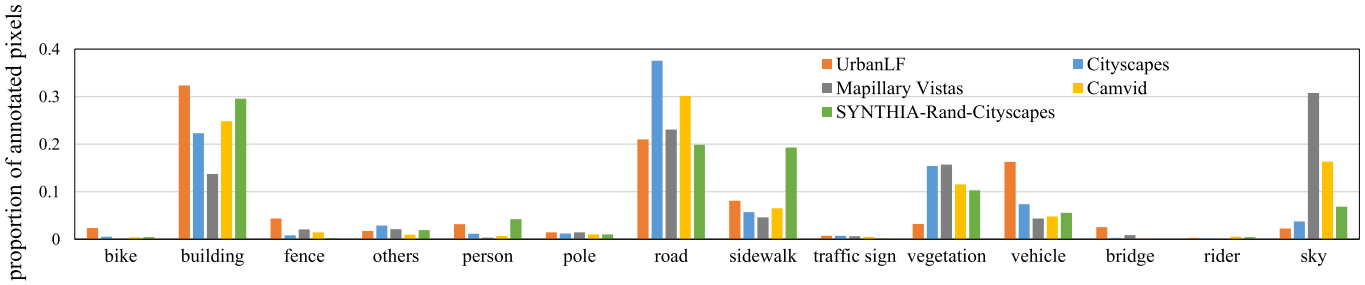


Fig. 3. The proportion of annotated pixels (y-axis) per class (x-axis) in UrbanLF, Cityscapes (fine-grained), Mapillary Vistas, Camvid and SYNTHIA-Rand-Cityscapes. The original labels of latter four datasets are remapped to 14 classes in UrbanLF for a unified comparison. For Cityscapes and Mapillary Vistas, only training and validation sets are counted since test set annotations are not publicly available.

empty space. We use Sun as the type of light to simulate the lighting conditions at different times of the day by changing the values of strength, specular and angle. A camera array composed of 81 virtual cameras with the same configuration is used to collect LF and the disparity can be controlled by changing the distance between adjacent cameras.

For the sake of keeping the consistency between the real-world and synthetic data, the resolution of these two parts should be as near as possible. The limited number of sensors in Lytro Illum makes the spatial resolution of the real-world images only 623×432, so we finally select 640×480 as the spatial resolution of the synthetic images. In addition, the synthetic data contain densely sampled LF with disparity in a range from -0.47 to 1.55 between adjacent views that is basically as similar as the real-world data.

B. Class Selection and Image Annotation

Taking into account practical applications, the frequency of objects and the compatibility with existing urban scene datasets, we define 14 classes for evaluation, i.e., *bike*, *building*, *fence*, *others*, *person*, *pole*, *road*, *sidewalk*, *traffic sign*, *vegetation*, *vehicle*, *bridge*, *rider* and *sky*. Please refer to the supplementary material for detailed definition. We provide fine annotations that accurately reflect details in the scene, including the contour of the object, the scale of the object and the occlusion relation between different objects.

For real-world data, the annotations of central sub-aperture images are realized by human labour via LabelMe [83]. To guarantee the quality, the annotation time is more than one hour on average for an image. Furthermore, three participants are responsible for checking all annotations so as to avoid inconsistencies caused by the different understanding of the label scheme and definition of classes among annotators.

For synthetic data, Blender generates completely accurate label maps, depth maps and disparity maps of the scene, greatly reducing the demand for manual effort. The extra semantic annotations and depth information for all 81 views broaden the scope of application of synthetic data.

C. Dataset Splitting

UrbanLF is split into training, validation and test set approximately at a ratio of 7:1:2. Following this scheme, the real-world data consist of 580 training, 80 validation and 164 test samples, while the corresponding number in synthetic data are 172, 28 and 50 respectively. The training and validation set are publicly available and the test set is withheld for benchmarking. We divide the data by stratified sampling instead of random sampling. Specifically, each set is composed of samples with the same distribution ratio in the following properties: 1) the light condition, 2) the number of instances, 3) the shooting angle. This balanced way helps to comprehensively train and test the model.

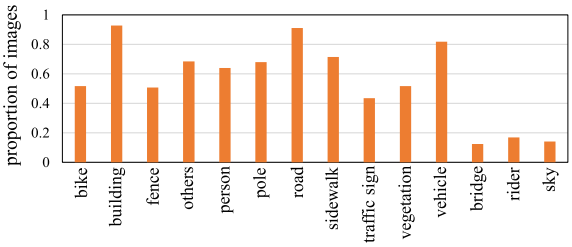


Fig. 4. The proportion of images (y-axis) that have specific class (x-axis) in UrbanLF.

D. Statistical Analysis

We conduct statistical analysis from three aspects to give a comprehensive introduction to UrbanLF.

1) Distribution of Classes: We compare our UrbanLF with widely-used datasets that focus on urban scenes with semantic pixel-wise annotations, i.e. Cityscpes [25], Mapillary Vistas [52], CamVid [51] and SYNTHIA-Rand-Cityscpes [53]. The original labels of each dataset are remapped to the aforementioned 14 classes for a unified comparison. As shown in Fig. 3, the statistical results of these datasets are relatively consistent, in which background entities like *building* and *road* occupy more pixels than foreground objects like *bike*, *traffic sign* and *rider*. This imbalanced class distribution is in line with urban scenes. The difference comes from the characteristics of scenes in the datasets. UrbanLF mainly covers traffic and street scenes, resulting in more *vehicle*, more *building* and fewer *vegetation*. Cityscpaes with inner-city traffic of roads and intersections contains the most *road*. Mapillary Vistas with a wide vertical field of view contains the most *sky*. SYNTHIA-Rand-Cityscpes with lots of street blocks contains the most *sidewalk*. Fig. 4 shows the proportion of images that have specific class in UrbanLF. It can be observed that the majority

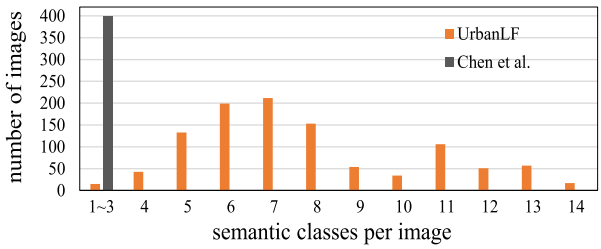


Fig. 5. The number of images (y-axis) that have specific number of classes (x-axis) in UrbanLF and Chen *et al.* [42].

of classes appear in at least half of the images, and only the value for *bridge*, *rider* and *sky* is less than 20%.

2) Scene Complexity: We report the scene complexity from the number of semantic classes per image. As shown in Fig. 5, it is obviously that UrbanLF has a high diversity of scene complexity, where the number of semantic class per image is in a wide range of [1, 14] rather than 3 classes at most in the only published LF dataset for semantic segmentation [42]. Moreover, nearly half of images in UrbanLF contain at least 8 classes, meaning that hard samples occupy a considerable share of the dataset.

3) Shooting Angle: The shooting angle is one of the key factors that have a great influence on the outcome of the images. A low-angle shot mainly takes the sky as background and creates a sense of depth. An eye-level shot is a standard shooting angle and accords with custom of human visual. A high-angle shot captures the object from above and makes it look flat. The shooting angle transformation in datasets may present new challenges. With this in mind, our UrbanLF covers various shooting angles to achieve a comprehensive visual effect, consisting of 89, 767 and 218 images at low, eye-level and high shooting angle respectively. Fig. 6 shows some representative samples.

IV. Benchmarks

In this section, we first introduce the algorithm benchmarking for semantic segmentation, including representative baselines, experimental setup and result analysis. With the rich resources provided by UrbanLF, our benchmark extends to super-resolution and depth estimation. The benchmark website will be available online after publication. We also apply the proposed dataset to other tasks like LF segmentation and LFSP, please refer to Appendix C of supplemental material.

A. Semantic Segmentation

1) Representative Baselines: We evaluate 12 state-of-the-art methods on UrbanLF, including 4 RGB-based methods: PSPNet [5], DeepLabv3+ [7], OCR [67], SETR [66], 4 video-based methods: Accel [10], TDNet [11], DAVSS [68], TMANet [69], 4 RGB-D-based methods: ACNet [13], MTI-Net [72], SA-Gate [14], ESANet [70]. They cover most of the representative methods and offer open source code. Note that we do not evaluate point-cloud-based methods due to the limitation of data content.

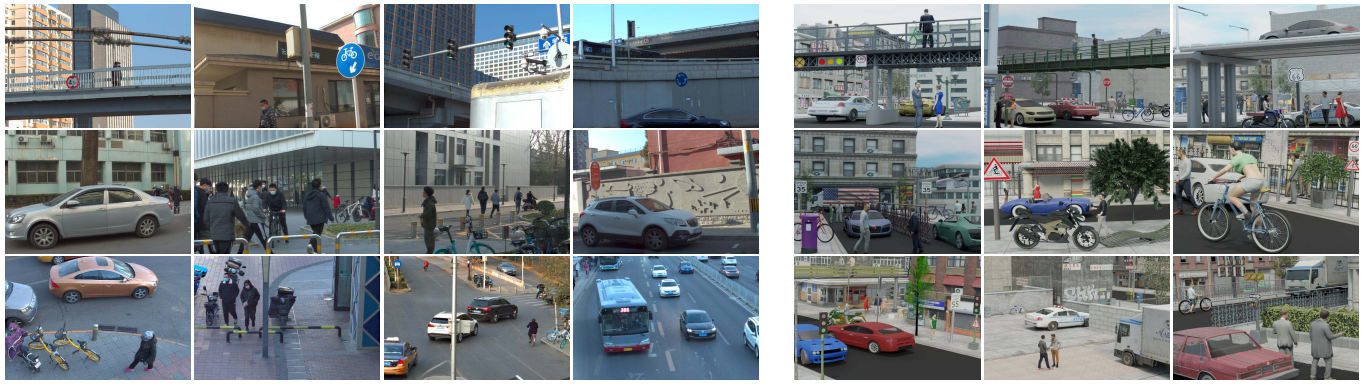
TABLE III
Data and Method Involved in the Experiment

Exp	Train	Val	Test	Method
I	Real(train)	Real(val)	Real(test)	RGB,Video,LF
II	Real(train)+ Syn(train+val)	Real(val)	Real(test)	RGB,Video,LF
III	Syn(train)	Syn(val)	Syn(test)	RGB,Video, RGB-D,LF

We also design two new LF-based methods to prove the benefits of LF for semantic segmentation. They rely on PSPNet and OCR respectively and increase a spatial branch along with a feature fusion module on the original basis. Different from [42], we explore the possibility of using partial sub-aperture images as input to reduce memory consumption. As illustrated in Fig. 7, our baselines apply the encoder-decoder structure. There are two independent branches in the encoder. One is RGB branch that extracts color features from central view image. The other is spatial branch that extracts spatial features from image stacks in four directions of horizontal, vertical, $\frac{1}{4}\pi$, and $\frac{3}{4}\pi$. PSPNet-LF adopts ResNet as the backbone and OCR-LF adopts HRNet as the backbone. We apply a channel attention operation to further refine the two features and use element-wise add as input of decoder to convert the fusion feature into the final segmentation result.

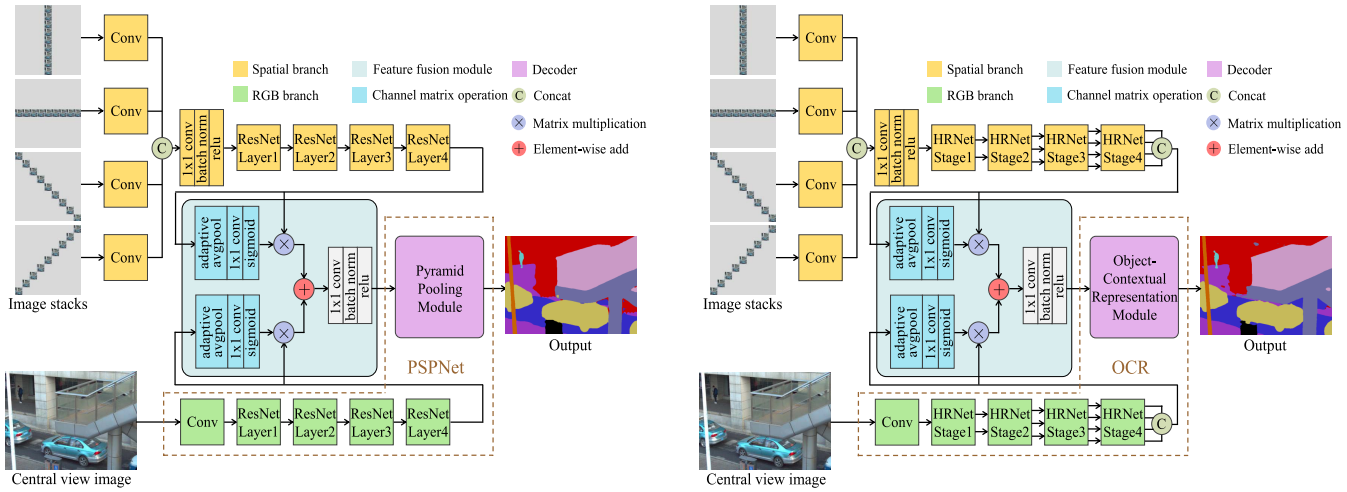
2) Experimental Setup: There are three experiments on UrbanLF in total. We make the first experiment on the real-world part of the proposed dataset. In the second experiment the model is trained along with some synthetic samples. The aim of this work is to show that synthetic data helps to improve segmentation results on real-world data. To achieve this, we crop the synthetic image to the same resolution as the real-world image, then build batches with images from two domains. Since the real-world data do not provide depth information, we conduct the third experiment on the synthetic part own to extend the evaluation scope to RGB-D-based methods. The details of experiments are shown in Table III.

We follow the original experimental settings of each method. For RGB-based methods, we only use the sub-aperture image in central view and corresponding annotation. For RGB-D-based methods, we additionally use the depth map. For video-abased methods, we choose the order in [84] that horizontally scans the sub-aperture images from left to right starting from the view on the left superior corner to create the pseudo video. For our two baselines, a SGD optimizer with initial learning rate 0.01, momentum 0.9 and weight decay 0.0005 is used to train the network. We employ a poly learning rate policy where the learning rate is multiplied by $(1-\frac{iter}{max\_iter})^{0.9}$. As for data augmentation, we apply random scaling, flipping and cropping to both central view image and image stacks. The comparison is done with pixel accuracy (Acc), mean pixel accuracy (mAcc) and mean intersection-over-union (mIoU) in full resolution of central view. We adopt single-scale testing and multi-scale testing at the same time. The testing strategies for the latter are horizontal flipping and multi-scale scaling with a factor (0.75, 1.0, 1.25, 1.5).



(a) (b)

Fig. 6. Example central view images with different shooting angles from UrbanLF. Top: low-angle shot. Medium: eye-level shot. Bottom: high-angle shot. (a) Real-world samples. (b) Synthetic samples.



(a) (b)

Fig. 7. The overview of our proposed baselines for LF semantic segmentation. We implement them by making a few modifications to PSPNet and OCR. The network contains two parallel branches for RGB and image stacks, a feature fusion module and a decoder. (a) PSPNet-LF. (b) OCR-LF.

We also report average inference time in the third experiment to evaluate speed. For a fair and consistent comparison, all experiments are conducted with single-scale testing strategy and a batch size of one on a NVIDIA RTX 2080Ti. For video-based methods, we compute inference time per frame in the pseudo video as statistical results.

3) Result Analysis: The quantitative results of the former two experiments are presented in Table IV. Our modification to OCR and PSPNet by additionally using image stacks as input is particularly effective. OCR-LF achieves the highest scores on almost every metric. PSPNet-LF shows remarkable performance with multi-scale testing. The following are OCR and SETR. All methods achieve improvement on Acc, mAcc and mIoU while exploiting the extra synthetic images for training. The specific increments are highlighted in bold. Table V presents the results of the third experiment. In terms of accuracy, OCR-LF obtains the highest scores on mAcc and mIoU. PSPNet-LF also has improvements compared with PSPNet. When depth data are available, ACNet achieves superior performance on Acc and SA-Gate obtains the second highest scores on every metric. As for speed, DAVSS achieves the shortest inference time by reusing and warping keyframe features. ESANet with an efficient ResNet-34-based encoder obtains the second fastest speed. PSPNet-LF and OCR-LF have the longest inference time due to utilizing many sub-aperture images, about twice as long as PSPNet and OCR respectively. On the whole, LF-based methods can obtain comparable results through leveraging the implicit geometry information in the sub-aperture images. However, they achieve start-of-the-art performance at the cost of low inference speed. It is worthy of exploring how to further reduce memory usage while retaining high accuracy. Due to effectively using the extra depth information, the performance of RGB-D-based methods is generally superior to other methods, which is consistent among different datasets. Furthermore, adding synthetic data and multi-scale testing help to boost the performance.

The qualitative results of experiment I and experiment III are shown in Fig. 8 and Fig. 9. From Fig. 8, it is observed that our proposed baselines improve the case of inaccurate

TABLE IV

QUANTITATIVE RESULTS ON URBANLF-REAL FOR SEMANTIC SEGMENTATION. THE TOP TEN ROWS SHOW THE COMPARISON OF TRAINING THE MODEL ON REAL-WORLD SAMPLES ONLY, WHILE THE BOTTOM TEN ROWS SHOW THE EFFECT OF EXTENDING TRAINING SETS WITH SYNTHETIC SAMPLES. ACC (%), MACC (%) AND MIOU (%) ARE REPORTED. THE BEST RESULTS ARE IN RED AND THE SECOND BEST RESULTS ARE IN BLUE. THE IMPROVEMENTS ARE IN BOLD WITH NUMERICAL VALUE IN BRACKET. * SIGNIFIES MULTI-SCALE TESTING

Method	Backbone	Type	Acc	mAcc	mIoU	Acc*	mAcc*	mIoU*
PSPNet [5]	ResNet-101	RGB	91.21	83.87	76.34	91.74	84.68	77.75
DeepLabv3$^+$ [7]	ResNet-101	RGB	91.02	83.53	76.27	91.50	84.30	77.35
SETR [66]	ViT-Large	RGB	**92.16**	84.27	77.74	**92.71**	84.93	79.05
OCR [67]	HRNetV2-W48	RGB	92.02	**85.17**	**78.60**	92.43	**85.77**	**79.65**
Accel [10]	ResNet-101	Video	89.15	80.69	71.64	90.07	81.47	73.56
TDNet [11]	ResNet-50	Video	91.05	83.38	76.48	91.79	84.85	78.36
DAVSS [68]	Xception-65	Video	91.04	83.54	75.91	91.74	84.54	77.68
TMANet [69]	ResNet-50	Video	91.67	84.13	77.14	91.87	84.55	77.91
PSPNet-LF	ResNet-101	LF	92.14	84.86	78.10	**92.77**	85.73	79.55
OCR-LF	HRNetV2-W48	LF	**92.51**	**86.31**	**79.32**	92.68	**86.58**	**80.06**
PSPNet [5]	ResNet-101	RGB	91.73(**0.52**)	84.47(**0.60**)	77.71(**1.37**)	91.94(**0.20**)	84.97(**0.29**)	78.52(**0.77**)
DeepLabv3$^+$ [7]	ResNet-101	RGB	91.33(**0.31**)	83.86(**0.33**)	76.70(**0.43**)	91.95(**0.45**)	84.65(**0.35**)	77.96(**0.61**)
SETR [66]	ViT-Large	RGB	**92.72**(**0.56**)	85.55(**1.28**)	79.06(**1.32**)	93.23(**0.52**)	86.37(**1.44**)	80.51(**1.46**)
OCR [67]	HRNetV2-W48	RGB	92.56(**0.54**)	**86.56**(**1.39**)	**79.90**(**1.30**)	93.04(**0.61**)	**86.96**(**1.19**)	80.83(**1.18**)
Accel [10]	ResNet-101	Video	89.40(**0.25**)	82.30(**1.61**)	72.85(**1.21**)	90.40(**0.33**)	82.81(**1.34**)	74.76(**1.20**)
TDNet [11]	ResNet-50	Video	91.48(**0.43**)	84.25(**0.87**)	77.52(**1.04**)	92.06(**0.27**)	85.46(**0.61**)	79.20(**0.84**)
DAVSS [68]	Xception-65	Video	91.96(**0.92**)	85.21(**1.67**)	77.31(**1.40**)	92.47(**0.73**)	86.22(**1.68**)	79.02(**1.34**)
TMANet [69]	ResNet-50	Video	91.84(**0.17**)	84.81(**0.68**)	78.13(**0.99**)	92.44(**0.57**)	85.79(**1.24**)	79.54(**1.63**)
PSPNet-LF	ResNet-101	LF	92.69(**0.55**)	86.01(**1.15**)	79.45(**1.35**)	**93.29**(**0.52**)	**86.96**(**1.23**)	**80.92**(**1.37**)
OCR-LF	HRNetV2-W48	LF	**92.95**(**0.44**)	**86.94**(**0.63**)	**80.40**(**1.08**)	**93.27**(**0.59**)	**87.26**(**0.68**)	**81.21**(**1.15**)

TABLE V

QUANTITATIVE RESULTS ON URBANLF-SYN FOR SEMANTIC SEGMENTATION. ACC (%), MACC (%), MIOU (%) AND TIME (MS) ARE REPORTED. THE BEST RESULTS ARE IN RED AND THE SECOND BEST RESULTS ARE IN BLUE. * SIGNIFIES MULTI-SCALE TESTING

Method	Backbone	Type	Acc	mAcc	mIoU	Acc*	mAcc*	mIoU*	Time
PSPNet [5]	ResNet-101	RGB	89.39	84.48	75.78	90.76	85.64	78.16	74.5
DeepLabv3$^+$ [7]	ResNet-101	RGB	89.60	83.55	75.39	90.99	85.35	78.05	80.5
OCR [67]	HRNetV2-W48	RGB	91.50	86.96	79.36	92.44	88.18	81.22	43.7
Accel [10]	ResNet-101	Video	87.56	80.52	70.48	89.20	82.67	74.07	25.7
TDNet [11]	ResNet-50	Video	89.06	83.43	74.71	89.79	84.32	76.39	35.2
DAVSS [68]	Xception-65	Video	89.47	82.94	74.27	90.94	85.15	77.33	**12.7**
TMANet [69]	ResNet-50	Video	89.76	84.44	76.41	90.99	86.30	78.87	54.4
ACNet [13]	ResNet-50	RGB-D	**92.53**	86.62	78.56	**93.56**	87.95	80.90	40.1
MTINet [72]	HRNetV2-W48	RGB-D	91.24	86.94	79.10	91.86	87.34	80.01	53.6
ESANet [70]	ResNet-34	RGB-D	91.81	86.26	79.43	92.63	86.97	80.97	**19.9**
SA-Gate [14]	ResNet-101	RGB-D	**92.10**	**87.04**	**79.53**	**93.18**	**88.51**	**81.72**	56.4
PSPNet-LF	ResNet-101	LF	90.55	85.91	77.88	91.55	87.54	80.09	152.6
OCR-LF	HRNetV2-W48	LF	92.01	**87.71**	**80.43**	93.06	**89.20**	**82.77**	82.6

label assignment in the complex occlusion areas. With the help of the complementary information provided by the multi-view images, PSPNet-LF successfully segments the occluded trash can and OCR-LF obtains better reconstruction quality near occlusion boundary compared with OCR. Fig. 9 provides the results on the synthetic samples. Benefiting from the depth cue, the RGB-D-based methods distinguish wheels and clothes from road with similar colors at high accuracy, achieving better performance than other methods. Our baselines get similar visual results through utilizing the implicit depth information in LF. It's worth noting that all methods fail to recover the spokes of the bike wheel and the exact boundary of arms and fingers, leaving a margin for future research.

B. Super-Resolution

1) Representative Baselines: We select 3 representative LF spatial super-resolution (LFSSR) methods, including LF-ATO [85], LF-InterNet [86] and LF-DFnet [87]. LF-ATO applies an all-to-one architecture and appends structural consistency regularization to preserve parallax relationship. LF-InterNet combines the separately extracted spatial and angular features through repetitive interactions. LF-DFnet incorporates and encodes the angular information through

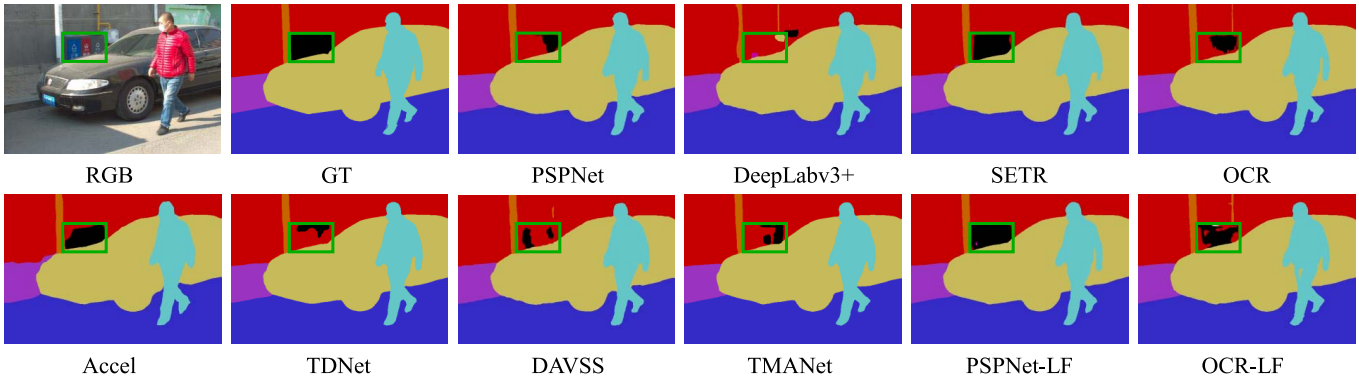


Fig. 8. Qualitative results on UrbanLF-Real for semantic segmentation. Green rectangles highlight the occlusion areas.

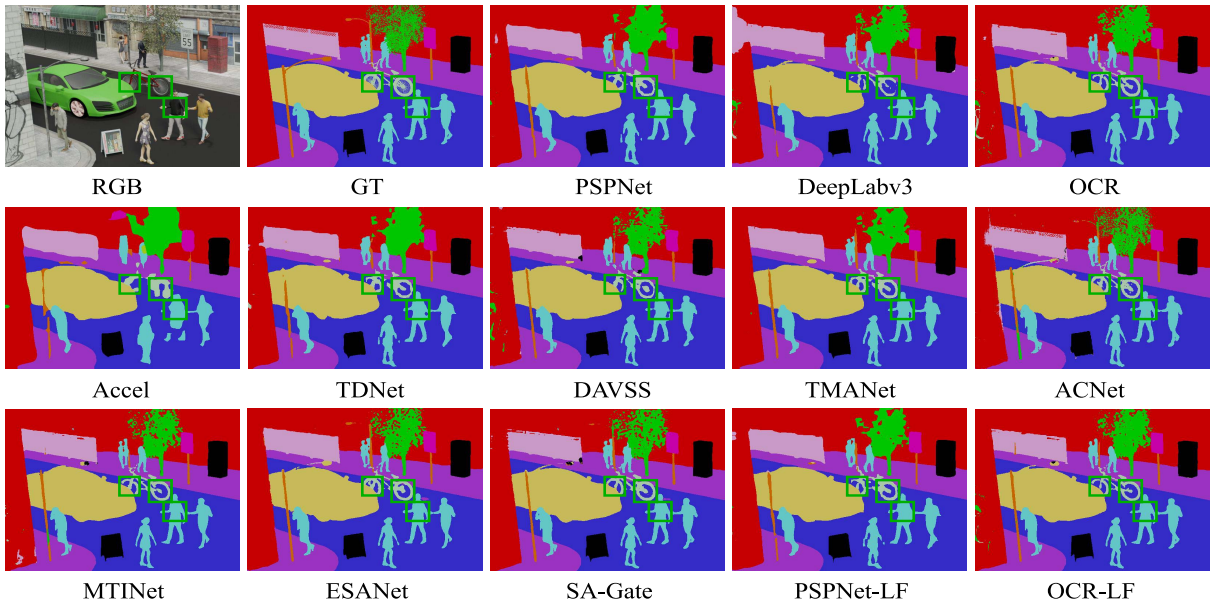


Fig. 9. Qualitative results on UrbanLF-Syn for semantic segmentation. Green rectangles highlight the areas of different objects with similar colors.

deformable convolution. They are all deep learning methods and have proven effectiveness on many LF datasets.

2) Experimental Setup: Following the general setting, we train the models with both real-world samples and synthetic samples from UrbanLF, validate and test them on two parts independently. Considering that sharing the same test set with other tasks will expose the ground truth, we extra collect 80 real-world and 30 synthetic samples as new test data. The bicubic interpolation with a factor of 2 and 4 is applied to generate low resolution images of different scales. All these methods have open source code and we follow the original settings. The comparison is done with peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) averaged over all sub-aperture images.

3) Result Analysis: Table VI shows the quantitative results. LF-DFnet gets the highest PSNR and SSIM scores on both parts for $\times 2$ and $\times 4$ LFSSR. Fig. 10 and Fig. 11 show the qualitative results. For $\times 2$ LFSSR, LF-DFnet preserves most texture details and obtains the best reconstruction performance. LF-InterNet generates similar results. LF-ATO produces artifacts near the fence and misses plenty of text textures. For $\times 4$ LFSSR, LF-DFnet and LF-InterNet still achieve better visual quality when the input LF images are more seriously degraded and the problem becomes more ill-posed.

TABLE VI
AVERAGE PSNR/SSIM VALUES ON URBANLF-REAL AND URBANLF-SYN FOR $\times 2$ AND $\times 4$ LFSSR

Method	×2		×4	
	Real	Syn	Real	Syn
Bicubic	31.35/0.950	31.95/0.945	24.60/0.825	25.57/0.826
LF-ATO [85]	38.45/0.987	35.01/0.970	30.22/0.926	28.82/0.888
LF-InterNet [86]	39.99/0.990	39.46/0.987	31.71/0.945	31.87/0.937
LF-DFnet [87]	40.05/0.991	39.69/0.988	32.01/0.947	32.21/0.941

C. Depth Estimation

1) Representative Baselines: We compare 3 representative methods, including Spinning Parallelogram Operator (SPO) [88], EPINet [89] and LFattNet [90]. They play a

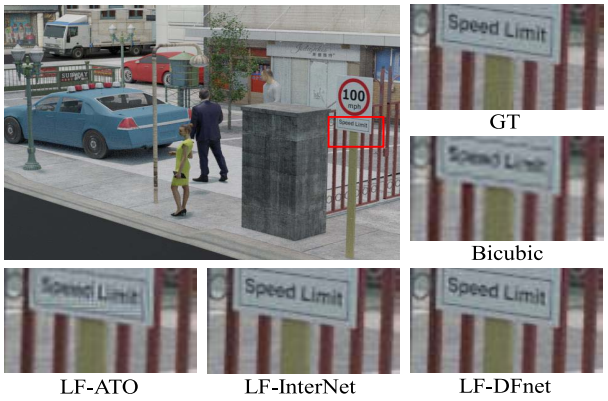


Fig. 10. Qualitative results on UrbanLF-Syn for ×2 LFSSR. The red rectangle is zoomed for better viewing.

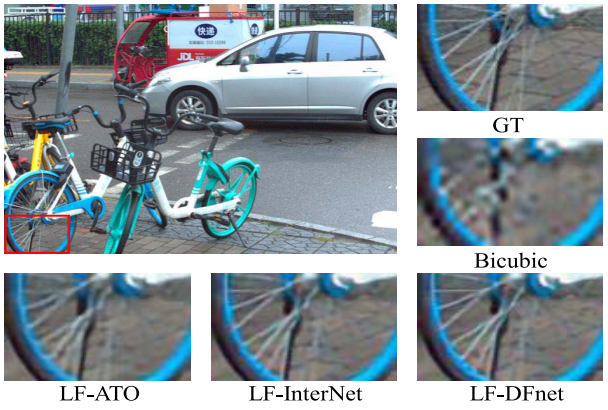


Fig. 11. Qualitative results on UrbanLF-Real for ×4 LFSSR. The red rectangle is zoomed for better viewing.

leading part in the development of this field. SPO finds the lines indicating depth information from EPIs. EPINet uses the FCN framework to exploit the characteristics of epipolar geometry. LFattNet introduces a view selection module to infer the contribution of each view by generating an attention map.

2) Experimental Setup: The experiment is only performed on the synthetic part of UrbanLF. We exclude all samples that contain the *sky* because the true depth of this class can not be accurately measured. We also create a new test set to avoid depth data leakage owing to data sharing among tasks. After redistributing the data, there are 170 samples for training, 30 samples for validation and 30 samples for test with corresponding disparity map. We adopt the settings in the original publication and the disparity label range is set to 64 for SPO. As for evaluation, we only estimate the disparity of central view and use the mean square disparity error (MSE) and the bad pixel ratio (BadPix) with three thresholds (0.01, 0.03 and 0.07 pixels). For these metrics, small value signifies good performance. Since EPINet applies the convolutional layer without zero-padding, we crop 15 bordering pixels for a fair comparison.

3) Result Analysis: Table VII shows the quantitative results. LFattNet achieves the best MSE performance and SPO achieves the best BadPix performance. However, there is still much room for improvement. The BadPix scores are generally too high, indicating that the prediction of most pixels is not accurate enough. Fig. 12 shows the qualitative results. We can observe high errors caused by the shadows in all methods. SPO has difficulty in recovering weak texture areas thus it fails to make predictions for road, bridge and other areas with similar colors, resulting in high MSE scores. EPINet and LFattNet struggle with reconstructing fine structures such as the outline of tree and thin gaps between leaves. Their performance also deteriorates at different levels on car surface with specular highlights. Judging from the results, we conclude that our meticulously designed urban scenes include various combinations of open challenges to further stimulate advanced research in depth estimation.

TABLE VII
QUANTITATIVE RESULTS ON URBANLF-SYN FOR DEPTH ESTIMATION IN TERMS OF MSE*100 AND BADPIX 0.01, 0.03, 0.07 (%)

Method	MSE	BadPix(0.01)	BadPix(0.03)	BadPix(0.07)
SPO [88]	8.62	68.95	42.28	30.12
EPINet [89]	1.95	90.81	73.35	34.00
LFattNet [90]	1.72	86.59	63.85	39.32

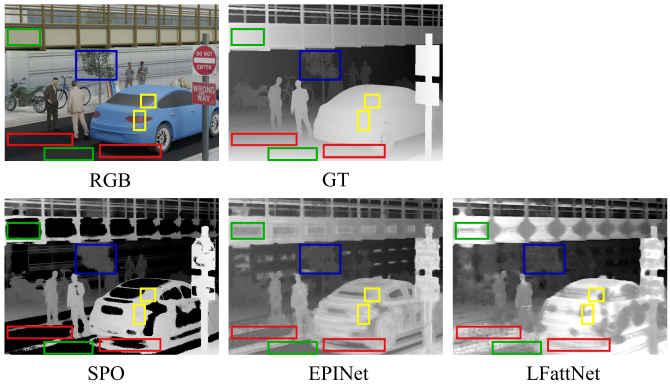


Fig. 12. Qualitative results on UrbanLF-Syn for depth estimation. Shadow areas are framed in red. Fine structures are framed in blue. Text areas are framed in green. Specular areas are framed in yellow.

V. DISCUSSION

LF semantic segmentation is a challenging and meaningful topic. However, due to the lack of large-scale datasets, it has not been well explored up to now. The key to constructing such a dataset is to ensure both quantity and quality of the data and UrbanLF fills this blank.

Considering the characteristics of the LF geometry, our baselines encode sub-aperture image stacks to learn the angular and spatial information for semantic segmentation. Since sub-aperture images share information, acting as a supplementary item for one another, our baselines solve the problem of inaccurate prediction in occluded regions of central view. The implicit depth information is also useful for distinguishing different objects with similar colors in RGB space. The results on UrbanLF are better than those of RGB and video-based methods and are comparable to those of RGB-D-based methods, proving that LF does benefit this topic.

Although applying LF to semantic segmentation is the main contribution of UrbanLF, it is applicable to other fields of research as well. The complex urban scenes present challenges for super-resolution and depth estimation. In future work we plan to introduce multiple data content like intrinsic layers to make UrbanLF suitable for more tasks.

VI. Conclusion

In this paper, we introduce a brand new LF dataset called UrbanLF, including 824 real-world and 250 synthetic urban scene samples with ground truth pixel-wise annotations. Through evaluating several state-of-the-art methods on three tasks, we highlight that the proposed dataset supports detailed comparisons among different methods. Furthermore, we specially design two baselines for LF semantic segmentation and get outstanding performance. We also find that synthetic samples can supplement real-world samples to solve the problem of limited available data caused by cumbersome and error-prone manual annotation. As the largest and the most diverse LF dataset for semantic segmentation, we hope that UrbanLF attracts more researchers into related fields.

References

[1] J. Levinson *et al.*, "Towards fully autonomous driving: Systems and algorithms," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2011, pp. 163–168.

[2] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.

[3] K.-T. Chang, "Geographic information system," in *International Encyclopedia of Geography: People, the Earth, Environment and Technology: People, the Earth, Environment and Technology*. Atlanta, GA, USA: Wiley, 2017, pp. 1–9, doi: 10.1002/9781118786352.wbieg0152.

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017.

[5] 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. 2881–2890.

[6] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[7] 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.

[8] J. Carreira, V. Patraucean, L. Mazare, A. Zisserman, and S. Osindero, "Massively parallel video networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 649–666.

[9] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2349–2358.

[10] S. Jain, X. Wang, and J. E. Gonzalez, "Accel: A corrective fusion network for efficient semantic segmentation on video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8866–8875.

[11] 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. 8818–8827.

[12] W. Wang and U. Neumann, "Depth-aware CNN for RGB-D segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 135–150.

[13] X. Hu, K. Yang, L. Fei, and K. Wang, "ACNET: Attention based network to exploit complementary features for RGBD semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1440–1444.

[14] X. Chen *et al.*, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 561–577.

[15] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, and J. Yang, "Pattern-affinitive propagation across depth, surface normal and semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4106–4115.

[16] A. Boulch, J. Guerry, B. Le Saux, and N. Audebert, "SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks," *Comput. Graph.*, vol. 71, pp. 189–198, Apr. 2018.

[17] L. Tchapmi, C. Choy, I. Armeni, J. Gwak, and S. Savarese, "SEGCloud: Semantic segmentation of 3D point clouds," in *Proc. Int. Conf. 3D Vis. (DV)*, Oct. 2017, pp. 537–547.

[18] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[19] Z. Song, L. Zhao, and J. Zhou, "Learning hybrid semantic affinity for point cloud segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Dec. 1, 2021, doi: 10.1109/TCSVT.2021.3132047.

[20] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 1996, pp. 31–42.

[21] Y. Zhang, W. Dai, M. Xu, J. Zou, X. Zhang, and H. Xiong, "Depth estimation from light field using graph-based structure-aware analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4269–4283, Nov. 2020.

[22] Y. Ni, J. Chen, and L.-P. Chau, "Reflection removal on single light field capture using focus manipulation," *IEEE Trans. Comput. Imag.*, vol. 4, no. 4, pp. 562–572, Dec. 2018.

[23] Y. Ding, M. Li, T. Yan, F. Zhang, Y. Liu, and R. W. H. Lau, "Rain streak removal from light field images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 467–482, Feb. 2022.

[24] 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.* Cham, Switzerland: Springer, 2016, pp. 19–34.

[25] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[26] T. Mullen, *Mastering Blender*. Hoboken, NJ, USA: Wiley, 2011.

[27] S. Heber and T. Pock, "Convolutional networks for shape from light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3746–3754.

[28] C. Hazirbas, S. G. Soyer, M. C. Staab, L. Leal-Taixé, and D. Cremers, "Deep depth from focus," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 525–541.

[29] 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.

[30] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *Proc. 8th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, 2016. [Online]. Available: https://infoscience.epfl.ch/record/218363

[31] A. S. Raj, M. Lowney, R. Shah, and G. Wetzstein. (2016). *Stanford LYTRO Light Field Archive*. [Online]. Available: http://lightfields.stanford.edu/LF2016.html

[32] M. Le Pendu, X. Jiang, and C. Guillemot, "Light field inpainting propagation via low rank matrix completion," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1981–1993, Apr. 2018.

[33] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2806–2813.

[34] J. Zhang, Y. Liu, S. Zhang, R. Poppe, and M. Wang, "Light field saliency detection with deep convolutional networks," *IEEE Trans. Image Process.*, vol. 29, pp. 4421–4434, 2020.

[35] Y. Piao, Z. Rong, S. Xu, M. Zhang, and H. Lu, "DUT-LFSaliency: Versatile dataset and light field-to-RGB saliency detection," 2020, *arXiv:2012.15124*.

[36] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–10, Nov. 2016.

[37] P. P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng, "Learning to synthesize a 4D RGBD light field from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2243–2251.

[38] V. K. Adhikarla *et al.*, "Towards a quality metric for dense light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 58–67.

[39] W. Wen, K. Wei, Y. Fang, and Y. Zhang, "Visual quality assessment for perceptually encrypted light field images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2522–2534, Jul. 2021.

[40] N. Sabater *et al.*, "Dataset and pipeline for multi-view light-field video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 30–40.

[41] A. Alperovich, O. Johannsen, M. Strecke, and B. Goldluecke, "Light field intrinsics with a deep encoder-decoder network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9145–9154.

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

[43] R. Raghavendra, K. B. Raja, and C. Busch, "Presentation attack detection for face recognition using light field camera," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 1060–1075, Mar. 2015.

[44] O. Johannsen, A. Sulc, N. Marniok, and B. Goldluecke, "Layered scene reconstruction from multiple light field camera views," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 3–18.

[45] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "A 4D light-field dataset and CNN architectures for material recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 121–138.

[46] D. G. Dansereau, B. Girod, and G. Wetzstein, "LiFF: Light field features in scale and depth," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8042–8051.

[47] J. S. Lumentut, T. H. Kim, R. Ramamoorthi, and I. K. Park, "Fast and full-resolution light field deblurring using a deep neural network," 2019, *arXiv:1904.00352*.

[48] Y. Xu, H. Nagahara, A. Shimada, and R.-I. Taniguchi, "TransCut2: Transparent object segmentation from a light-field image," *IEEE Trans. Comput. Imag.*, vol. 5, no. 3, pp. 465–477, Sep. 2019.

[49] P. David, M. Le Pendu, and C. Guillemot, "Scene flow estimation from sparse light fields using a local 4D affine model," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 791–805, 2020.

[50] L. Ruan, B. Chen, J. Li, and M.-L. Lam, "AIFNet: All-in-focus image restoration network using a light field-based dataset," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 675–688, 2021.

[51] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 88–97, 2009.

[52] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4990–4999.

[53] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3234–3243.

[54] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2012, pp. 746–760.

[55] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 567–576.

[56] A. Janoch *et al.*, "A category-level 3-D object dataset: Putting the Kinect to work," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 141–165.

[57] J. Xiao, A. Owens, and A. Torralba, "SUN3D: A database of big spaces reconstructed using SfM and object labels," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1625–1632.

[58] T. Hackel, N. Savinov, L. Ladicky, J. D. Wegner, K. Schindler, and M. Pollefeys, "Semantic3D.Net: A new large-scale point cloud classification benchmark," 2017, *arXiv:1704.03847*.

[59] Q. Hu, B. Yang, S. Khalid, W. Xiao, N. Trigoni, and A. Markham, "Towards semantic segmentation of urban-scale 3D point clouds: A dataset, benchmarks and challenges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4977–4987.

[60] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[61] 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.

[62] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[63] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[64] J. Ji, R. Shi, S. Li, P. Chen, and Q. Miao, "Encoder-decoder with cascaded CRFs for semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1926–1938, May 2021.

[65] N. Parmar *et al.*, "Image transformer," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4055–4064.

[66] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[67] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 173–190.

[68] J. Zhuang, Z. Wang, and B. Wang, "Video semantic segmentation with distortion-aware feature correction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3128–3139, Aug. 2021.

[69] H. Wang, W. Wang, and J. Liu, "Temporal memory attention for video semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2254–2258.

[70] 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.

[71] L.-Z. Chen, Z. Lin, Z. Wang, Y.-L. Yang, and M.-M. Cheng, "Spatial information guided convolution for real-time RGBD semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 2313–2324, 2021.

[72] S. Vandenhende, S. Georgoulis, and L. Van Gool, "MTI-Net: Multi-scale task interaction networks for multi-task learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 527–543.

[73] Y. Lyu, X. Huang, and Z. Zhang, "Learning to segment 3D point clouds in 2D image space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12255–12264.

[74] G. Riegler, A. O. Ulusoy, and A. Geiger, "OctNet: Learning deep 3D representations at high resolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3577–3586.

[75] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5099–5108.

[76] 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.

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

[78] H. Zhu, Q. Zhang, and Q. Wang, "4D light field superpixel and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6384–6392.

[79] N. Khan, Q. Zhang, L. Kasser, H. Stone, M. H. Kim, and J. Tompkin, "View-consistent 4D light field superpixel segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7811–7819.

[80] 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.

[81] M. Hamad, C. Conti, P. Nunes, and L. D. Soares, "ALFO: Adaptive light field over-segmentation," *IEEE Access*, vol. 9, pp. 131147–131165, 2021.

[82] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1027–1034.

[83] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, 2008.

[84] J. Xiang, M. Yu, G. Jiang, H. Xu, Y. Song, and Y.-S. Ho, "Pseudo video and refocused images-based blind light field image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2575–2590, Jul. 2021.

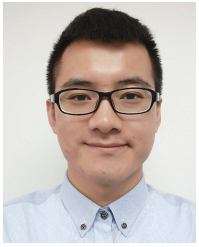
[85] J. Jin, J. Hou, J. Chen, and S. Kwong, "Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2260–2269.

[86] Y. Wang, L. Wang, J. Yang, W. An, J. Yu, and Y. Guo, "Spatial-angular interaction for light field image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 290–308.

[87] Y. Wang *et al.*, "Light field image super-resolution using deformable convolution," *IEEE Trans. Image Process.*, vol. 30, pp. 1057–1071, 2020.
[88] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *Comput. Vis. Image Understand.*, vol. 145, pp. 148–159, Apr. 2016.
[89] 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.
[90] 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.*, vol. 34, 2020, pp. 12095–12103.

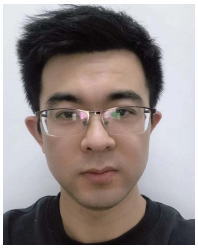
Hao Sheng (Member, IEEE) received the B.S. and Ph.D. degrees from the School of Computer Science and Engineering, Beihang University, China, in 2003 and 2009, respectively. Currently, he is a Professor and a Ph.D. Supervisor with the School of Computer Science and Engineering, Beihang University. He is working on computer vision, pattern recognition, and machine learning.

Ruixuan Cong received the B.S. degree from the School of Computer Science and Engineering, Beihang University, China, in 2021, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering. His research interest is computer vision, and he is particularly interested in light field semantic segmentation.



Da Yang received the B.S. degree from the School of Computer Science and Engineering, Beihang University, Beijing, China, in 2012, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering. His research interest is computer vision, and he is particularly interested in light field super-resolution.

Rongshan Chen received the B.S. degree from the School of Computer Science and Engineering, Beihang University, China, in 2019, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering. His research interest is computer vision, and he is particularly interested in light field depth estimation.



Sizhe Wang received the B.S. degree from the Xi'an University of Technology in 2012 and the M.S. degree from Xi'an Jiaotong University in 2014. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Beihang University, China. His research interest includes computer vision.

Zhenglong Cui received the B.S. degree from the School of Computer Science and Engineering, Beihang University, China, in 2018, where he is currently pursuing the Ph.D. degree. His research interest is computer vision, and he is particularly interested in light field depth estimation.