Multiscale-VR: Multiscale Gigapixel 3D Panoramic Videography for Virtual Reality

Jianing Zhang⁕, Tianyi Zhu⁕, Anke Zhang⁕,Xiaoyun Yuan⁕, Zihan Wang, Sebastian Beetschen, Lan Xu, Xing Lin, Qionghai Dai, *Senior Member, IEEE* and Lu Fang, *Senior Member, IEEE*

Abstract—Creating virtual reality (VR) content with effective imaging systems has attracted significant attention worldwide following the broad applications of VR in various fields, including entertainment, surveillance, sports, etc. However, due to the inherent trade-off between field-of-view and resolution of the imaging system as well as the prohibitive computational cost, live capturing and generating multiscale 360° 3D video content at an eye-limited resolution to provide immersive VR experiences confront significant challenges. In this work, we propose Multiscale-VR, a **multiscale unstructured** camera array computational imaging system for high-quality gigapixel 3D panoramic videography that creates the six-degree-of-freedom multiscale interactive VR content. The Multiscale-VR imaging system comprises scalable cylindrical-distributed global and local cameras, where global stereo cameras are stitched to cover 360° field-of-view, and unstructured local monocular cameras are adapted to the global camera for flexible high-resolution video streaming arrangement. We demonstrate that a high-quality gigapixel depth video can be faithfully reconstructed by our deep neural network-based algorithm pipeline where the global depth via stereo matching and the local depth via high-resolution RGB-guided refinement are associated. To generate the immersive 3D VR content, we present a three-layer rendering framework that includes an original layer for scene rendering, a diffusion layer for handling occlusion regions, and a dynamic layer for efficient dynamic foreground rendering. Our multiscale reconstruction architecture enables the proposed prototype system for rendering highly effective 3D, 360° gigapixel live VR video at 30 fps from the captured high-throughput multiscale video sequences. The proposed multiscale interactive VR content generation approach by using a heterogeneous camera system design, in contrast to the existing single-scale VR imaging systems with structured homogeneous cameras, will open up new avenues of research in VR and provide an unprecedented immersive experience benefiting various novel applications.

Index Terms—virtual reality, camera array, gigapixel imaging, computational photography, deep learning, depth reconstruction

1 Introduction

Virtual Reality (VR), a technology enabling users figuratively step inside a generated immersive 3D world, has revolutionized a variety of areas, from psychology, artistry, education, sports to entertainment, surveillance, tourism, gaming. Regardless of the fact that VR related products, such as 360° panoramic cameras (Google Jump [1], Facebook Surround 360 [2], Insta360 [3], Gopro Max [4], etc.) and VR headsets (HTC vive [5], Oculus Quest [6], Samsung HMD Odyssey [7], etc.), are emerging in an endless stream, the bottleneck of generating multi-scale high-quality VR video to provide the fully natural and interactive viewing experiences for all users remains unsolved.

For capturing an immersive VR video, the most widely used architecture in existing VR cameras is the omnidirectional stereo solutions. Early works [2], [3], [8] only enable three degrees-of-freedom (3-DoF) VR video capture, where the virtual viewpoint is fixed and the VR contents are rendered under the fixed distance from the eyes. Since the occlusion cue is missing in such 3-DoF VR cameras, the rendered VR scene suffers from unnatural and disorienting visual results even with a slight head roll or tilt. Some recent solutions capture and render full 6-DoF VR content using more sophisticated devices, such as a spinning light field camera array [9] or a professional 16-camera configuration [10] to release the head movement constraint.

Examining these VR cameras, for the hardware architecture, the highly structured hardware design with uniform camera configurations can hardly be scalable while being adaptive to the scene. For the algorithmic solution, the omnidirectional stereo matching requires dense overlapping between adjacent views to estimate the depth information of the whole scene. Moreover, they pay little attention to **the invariant spatial resolution at different scales, i.e., the experience of consistent high resolution regardless of the distance of local contents within the panoramic scene**. Restricted by the structured and homogeneous hardware design, existing VR cameras merely support single-scale capture; thus the VR scene farther away from the human eyes suffers from lower quality due to the limited spatial resolution. We claim that a globally consistent high-resolution across the whole panoramic scene plays a vital role in VR cameras, as the visual parallax detail at different scale across the VR scene is not only important for immersive VR experiences but also brings new effects for VR imaging such as virtual zooming into the specific regions where human eyes spot on. The naive super-resolution post-processing to existing VR contents cannot recover the missing high-resolution details of the local scenes, since the scale-gap among global and local scenes may be significant for real-world outdoor scenarios.

- *J. Zhang, T. Zhu, A. Zhang, Z. Wang, X. Yuan, S. Beetschen, L. Xu and L. Fang are with the Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen, 518071 China. X. Lin and Q. Dai are with the Dept. of Automation, Tsinghua University, Beijing, 100084 China. L. Xu and X. Yuan are also with Dept. of ECE, Hong Kong University of Science and Technology.*

- *This work is supported in part by Natural Science Foundation of China (NSFC) under contract No. 61722209 and 6181001011, in part by Shenzhen Science and Technology Research and Development Funds (JCYJ20180507183706645).*

- *⁕: Contributed Equally. The correspondence author is Lu Fang (E-mail: fanglu@sz.tsinghua.edu.cn)*

978-1-7281-5230-1/20/$31.00 ©2020 IEEE

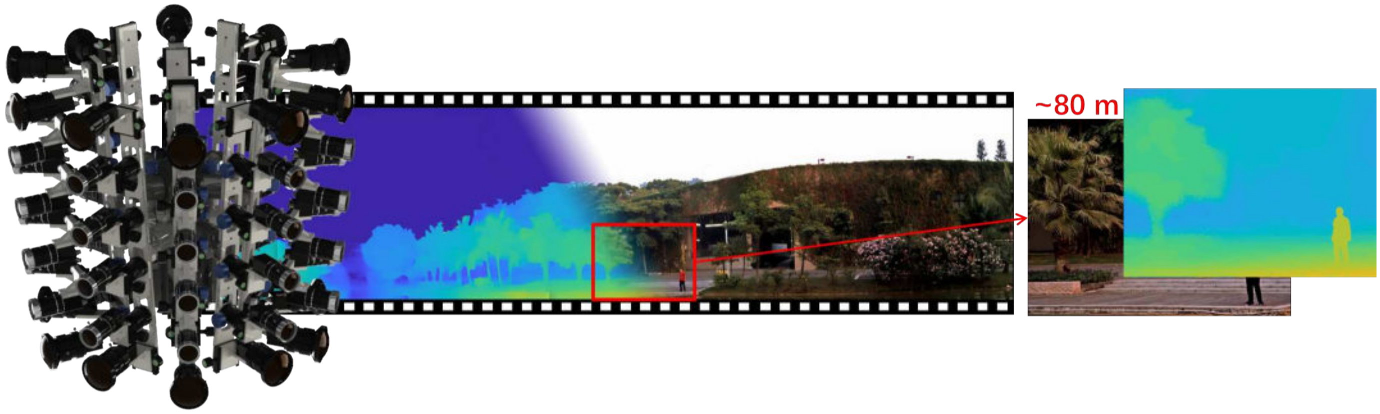


Fig. 1. Architecture of the multiscale 3D virtual reality (VR) video generator. To generate the unprecedented immersive user experience with the multiscale high-resolution 3D VR video, we propose the cylindrical-distributed unstructured camera array system that hybrids of five wide-angle stereo pairs as the global cameras along with local telephoto lens cameras as the source of high-resolution. The designed deep learning algorithms effectively process the captured high-throughput video streams for high-quality depth reconstruction and 3D VR video rendering. Different from the existing VR capture systems, our approach enables the users to zoom into arbitrary local regions of interests while still perceiving high-resolution real-world scenes in 3D.

Also, upgrading existing structured VR capture devices by simply increasing the number of cameras still suffers from the inherent trade-off between field-of-view and spatial resolution, let alone achieving the eye-limited resolution at multiple scale since the human eyes are extremely close to the VR display in headset and sensitive to the "screen door" effect.

In this paper, we attack the above challenges and propose the "Multiscale-VR Camera". For the first time, it delivers the ability of zoom-in to local regions at a great distance away from the camera, allowing multi-scale, gigapixel-level, and 3D panoramic videography for VR content generation. As illustrated in Fig.1, our method can capture fine-detailed visual parallax at different scales across the whole VR scene to enable the new virtual zooming effect for VR imaging. It relies on a novel design of heterogeneous cylindrical-distributed camera array, where up to 10 global stereo micro-cameras with wide-angle lenses (denoted as global camera) are stitched to cover up to 360° field-of-view (FOV), and a set of unstructured local monocular micro-cameras with telephoto lenses (denoted as local camera) are warped to the global camera for high-resolution video streaming. All these global and local cameras work in a relatively compensated yet scalable and flexible way, where different global stereo camera pairs and the local cameras are allocated in an unstructured manner without any pre-calibration. Our unique unstructured setup enables a superior VR immersive experience for the users to zoom into the local regions of interests perceiving high-resolution visual details. To reconstruct a large-scale depth video, we further propose a learning-based depth estimation scheme where the global depth via stereo matching and the local depth via high-resolution RGB-guided refinement are associated. To generate the immersive 3D VR content, we present the three-layer rendering schematic that includes an original layer for scene rendering, a diffusion layer for handling occlusion regions, and a dynamic layer for efficient dynamic foreground rendering. The technical contributions of the proposed Multiscale-VR are summarized as follows:

- **New Concept.** We propose a novel VR camera methodology for generating the 360°, 6-DoF, and multi-scale interactive VR videos that allows the users to zoom into regions-of-interest while still providing high-resolution live dynamic scenes in 3D.
- **New Hardware.** We build a novel hybrid and unstructured camera array as the multi-scale VR videography platform to achieve both scalable field-of-view and high-resolution, which equips the VR system with high-scalability based on the combination of the global stereo cameras and the unstructured local monocular cameras.
- **New Algorithm.** We design a novel algorithmic pipeline, including multi-stage learning-based depth estimation with plane-based optimization and local depth refinement, as well as multi-layer rendering, for gigapixel-level 3D panoramic VR immersive experience under the challenging unstructured and scalable setting.

2 Related Work

2.1 VR Cameras and Displays

High-quality VR content capturing and displaying are critical for the immersive 3D experience. Early VR cameras [11] typically stitch two fish-eye cameras to capture 360° panoramic videos, which is convenient to employ but the immersion experience is limited due to the lack of depth information. Recent research on VR cameras includes the use of a rapidly spinning two camera rig [12] or a ring camera array and omnidirectional stereo technique to capture stereo panoramic videos, such as Google Jump [1] [8], 1st generation Facebook Surround 360 camera [2], Kandao Obsidian R [13] and Insta360 pro 2 [3], etc. However, distortion is introduced when viewing omnidirectional stereo VR content in a headset [14], because only 3-DoF VR content is captured, leading to unnatural motion parallax when the head is moved and introducing vertical stretching because the ring diameter is usually much larger than the pupillary distance. To release such head movement constraint, recent solutions [9], [10], [15], [16] are proposed for 6-DoF VR content capturing. Hedman *et al.* [15], [16] propose causal 3D photography and use only a smartphone to capture 6-DoF VR content but only static scenes are supported. More sophisticated and structured devices are employed for 6-DoF videography, such as spinning light field camera array [9], 2nd generation Facebook Surround 360 camera [10], etc. However, these structured VR cameras above rely on structured setup and pay less attention to the spatial resolution at different scale so that the VR scene farther away from the human eyes suffers from worse immersive effect due to the depth ambiguity. Besides, image-based rendering (IBR) is also widely used to enhance the immersive VR experience. Huang *et al.* [17] used structure from

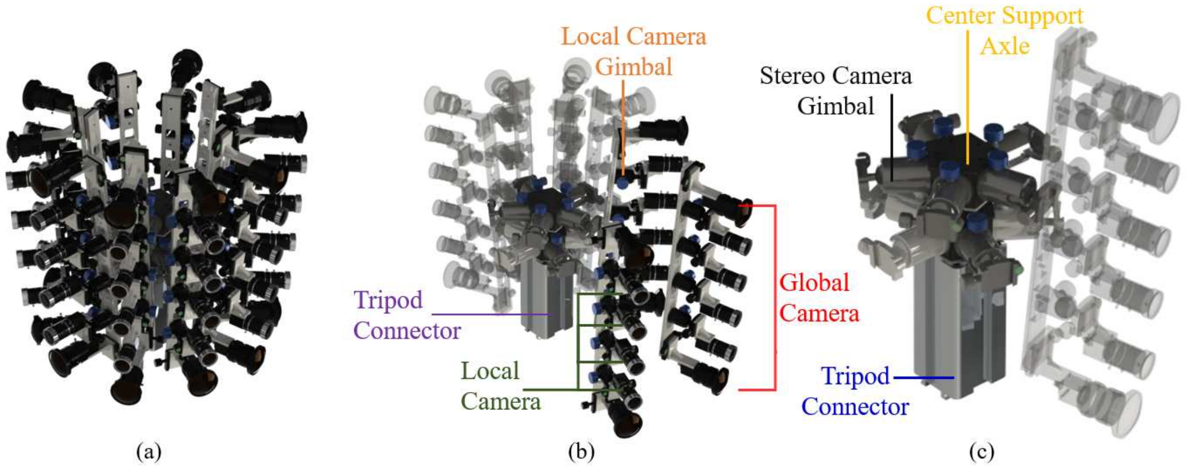


Fig. 2. Illustration of our hardware design. (a) The scalable multiscale-360 capturing system composes of up to ten groups of the global stereo camera, and multiple monocular local cameras are embedded in the middle of the global stereo pairs for detailed views. (b) The global and local spherical gimbals empower the adjustment ability of the multi-scale capturing system for locating sets of cameras over particular scenes. (c) The pluggable middle supporting-structure enables scalable setup over a maximum of 360-degree FOV.

motion and frame warping to synthesize new views for 6-DoF VR experiences, while Luo *et al.* [18] designed a robotic arm camera and a novel scene representation model to convert the real scene into VR content with head-motion parallax. Hedman *et al.* [19] proposed a two-scale surface reconstruction pipeline to enable tiled image-based free-viewpoint rendering. However, such IBR-based methods still struggle to recover the fine details visual parallax at different scales across the scene.

Compared with VR cameras, the history of VR displays dates back to the 60s when Morton Heilig invented the Telesphere Mask, the first head mount VR display. In recent years, with the development of VR technology, various light-weight VR display prototypes [20], [21], [22] are proposed and commercial VR headsets are emerging in an endless stream, such as Vive [5], Samsung HMD Odyssey [7] and Oculus Quest [6], etc. However, such commercial headsets are still far from providing immersive experiences because of missing real the depth-of-field. Thus, researchers improve the immersive experience by combining the eye tracking [23], [24], salience and motion detection [25], [26], [27] or multi-focal displays [28], [29]. Besides, efficient rendering algorithms for more realistic and comfortable experience are also proposed [30], [31].

Distinctively, our Multiscale-VR system firstly enables multi-scale, 6-DoF and 3D panoramic VR videography under a challenging unstructured and scalable setting. It is worth noting that our VR camera can recover the visual parallax at different scales across the whole VR scene and enables a new virtual zooming effect for VR imaging.

2.2 Gigapixel Imaging

Gigapixel imaging [32], [33], [34], [35], [36] aims to capture large scale scenes with extremely high resolution, breaking through the spatial-temporal bandwidth product of the single-lens optical system and providing panoramic immersive experiences. Kopf *et al.* [32] are the first to capture and stitch gigapixel images using a motor-controlled camera mount with a DSLR camera. This method is robust but is only capable of capturing static gigapixel images. Wilburn *et al.* [37] utilize a larger camera array and an image stitching algorithm [38] to generate image streams at high resolution. More sophisticated video stitching [39] and video filtering [40] techniques can be further applied to the camera array to generate more immersive results. Brady *et al.* [34] build the first gigapixel camera AWARE2, which adopts a two-stage multi-scale optical imaging system: a spherical objective lens for capturing a large FOV image of the whole scene and 98 micro-optic cameras each relaying a part of the image onto its sensor. However, the required precision of this camera is very high, especially the front spherical lens, and the whole system remains quite bulky. Electrical power and volume are substantial as well, at more than 10 W operating power per megapixel and the mass exceeding $100kg$ [41], [42]. Stamenov *et al.* [43] utilize fiber-coupled focal planes to replace conventional focal planes in this spherical lens camera array, which reduces the camera volume significantly (around $1/10$ of the conventional camera), yet the image is impacted by Moiré effects due to the imperfect fiber bundle microstructure. Yuan *et al.* [36] propose the multi-scale gigapixel videography, which is much more flexible compared with the previous gigapixel capturing methods.

Comparably, the proposed Multiscale-VR system is the first to combine the gigapixel imaging techniques above to VR videography, so as to provide gigapixel-level 3D panoramic VR immersive experience.

2.3 Depth Estimation

Depth estimation technique is vital to recover the visual parallax for immersive VR experiences. Specifically, recovering the depth from stereo images is widely used in modern VR cameras. Traditional methods usually estimate dense correspondence explicitly between the stereo images. Hirschmuller *et al.* [44] propose SGBM (semi-global block matching) [44], which introduces an efficient approximation of a global cost calculation. Bleyer *et al.* [45] present slanted support windows and estimate a 3D plane at each pixel onto its support region, which is able to reconstruct highly slanted surfaces with sub-pixel precision. Li *et al.* [46] further apply the patch match propagation and constrained random search in a coarse-to-fine way, which improves both efficiency and accuracy performance. Besides, in recent years, deep neural networks are utilized to estimate the depth information from a single image or stereo image pair [47], [48], [49], [50]. These learning-based methods achieve impressive depth estimation results but most of them can only estimate depth images at relatively low resolution and need extra fine-tuning process when used in different scenes.

In our Multiscale-VR system, we propose a novel smooth and multi-scale depth estimation algorithm for our unique VR camera, consisting of a learning-based strategy for the large-scale global scenes and a refinement scheme for the local region-of-interest.

3 System Overview

Our new Multiscale-VR system attempts to bring aspects inherent in 3D panoramic VR videography to the unstructured and scalable setting, so as to recover the visual parallax at different scale across the whole VR scene at gigapixel level and enable a new virtual zooming effect for VR imaging. In doing so, we introduce our novel design methodology and VR camera architecture (Sec. 3.1), as well as the components in our software pipeline (Sec. 3.2).

3.1 Hardware Design

As illustrated in Fig.2, our Multiscale-VR relies on a novel hybrid cylindrical distributed camera array for multi-scale, gigapixel-level and 3D panoramic VR videography, which encodes the following unique features.

Multi-scale. For capturing the visual parallax at different scales of the VR scene, we design a basic hybrid component for the camera array, denoted as a camera column, which consists of two global stereo cameras and some extra local cameras with telephoto lenses attached using local gimbals.

Unstructured. For the adaptability to capture various VR scenarios, both different global stereo camera pairs and the local cameras are allocated in an unstructured and adjustable manner so as to capture different regions of interest for different users.

Scalable. A novel center mechanical structure with flexible global gimbals to connect all the camera columns. Thus, the numbers of both the camera columns and local cameras on each column are scalable for various VR scenarios. Note that our current system supports 10 camera columns with 4 local cameras in each column at maximum to capture 360° FOV immersive VR content.

To enable the above multi-scale, unstructured and scalable VR content capturing, the parameters of all the cameras in our Multiscale-VR system need to be designed accordingly. To capture large-scale VR scenarios, for the global camera, 12mm lens with 2/3" CMOS sensor is adopted to provide adequate FOV (Table.1) and the corresponding baseline of each global stereo camera pair is set to 450mm, for estimating high-quality depth map over the range from 5 to 150 meters. Meanwhile, 12-36mm lenses with $1/1.8$" CMOS sensors are employed for the local cameras to capture high resolution local details. It is worth noting that the focal length of the local camera lens is adjustable to adapt to various VR scenarios.

As for the mechanical layout, a lightweight aluminum alloy stereo camera rack with two thermal-stable polylactic acids (PLA) made structural components is employed to connect the global cameras in each camera column. Beside, 4 extra installation anchors are provided for the local cameras in each rack. To assemble all the camera columns into a sector-shaped camera array, a pluggable center mechanical structure using robust carbon fiber is employed. Furthermore, the whole system can be fitted into a bounding cylinder with a size 0.6 meters in diameter 0.7 meters in vertical height with all the 10 camera columns plugged.

3.2 Reconstruction Pipeline

Our unique multi-scale, unstructured and scalable setup presents considerable challenges on the algorithm side for 3D panoramic VR videography generation. Fig.3 illustrates the high-level components of our reconstruction pipeline, which takes the hybrid video streams captured by all the global camera stereos and the local cameras as input, and generate a multi-scale immersive VR content as output. Specifically, auto white balance and stereo rectification is adopted as pre-processing to keep color consistency and reduce distortion. Furthermore, a brief introduction of each main component of our software pipeline is provided as follows, while more details are discussed in Sec. 4.

Gigapixel Videography. For gigapixel 3D VR videography, we stitch all the global views to generate a large-scale panorama scene and embed all the high-resolution local views seamlessly to enhance the local panorama details.

Depth Estimation. We propose a learning-based global level depth estimation scheme to generate the depth map that is suitable for artifact-free rendering by incorporating semantic information, and refine the local depth map using the high-resolution local cameras as guidance.

Layer-based Rendering. Once the depth is generated, an efficient three layer based-rendering algorithm is proposed, in which the dynamic foreground and almost static background are processed separately for realistic and fluent VR content rendering.

4 Method

4.1 Gigapixel Videography

Recall that under the multiscale and unstructured setup of our system, all the cameras including global and local ones are calibration-free except for the two global cameras in the same column. To provide a seamless and immersive 3D panoramic VR videography and enable the new virtual zooming in experience, we introduce a novel two-stage scheme for unstructured video generation, including a global stage for stereoscopic global camera stitching and a local stage for embedding local cameras.

Global Stitching. Aiming to present a high-resolution panoramic VR scene, a feature-based stitching algorithm [38] is adopted to estimate intrinsic and extrinsic camera parameters for each global camera pair. Furthermore, to reduce the noticeable artifacts caused by the camera localization error and the color inconsistency in those regions near the stitching boundaries, we apply graph-cut [51] to estimate a seam-free mask and reject the non-mask regions when calculating camera poses, followed by the linear Monge-Kantorovitch solution [52] for inter-camera color consistency.

Local Embedding. To enhance the resolution and the details of the stitched panorama, the fusion of the global and local cameras is critical. Therefore, an unstructured embedding scheme [36] is applied to warp all the local cameras to their corresponding global cameras. A cross resolution matching algorithm is first used to find matching points between global-local pairs, then a mesh-based multiple homography model is estimated to represent the warping field. Similarly, the linear Monge-Kantorovitch solution is utilized to map the color style of local cameras to the global panorama for local-global color consistency.

4.2 Depth Estimation

Depth estimation technology plays a pivotal role in our 6-DoF VR system. We first generate a global depth map for the whole

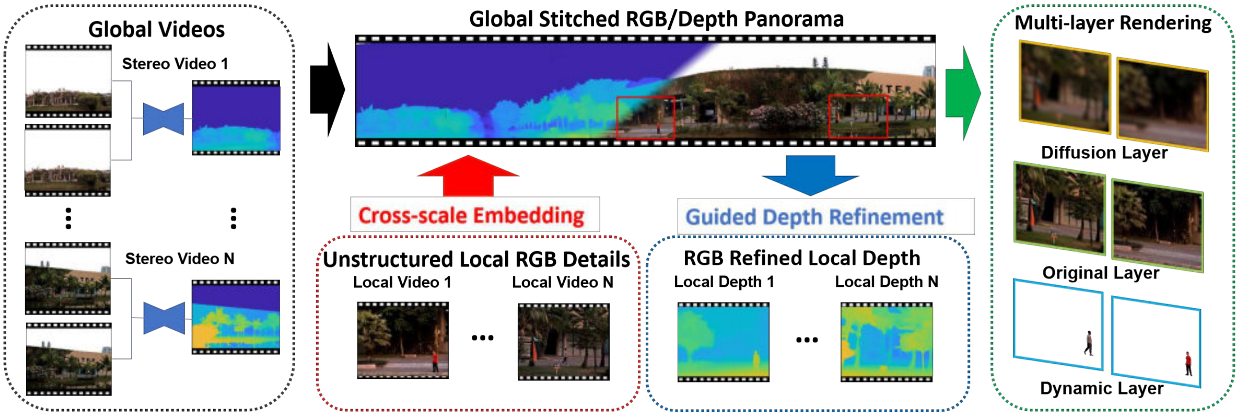


Fig. 3. Multiscale 3D VR video reconstruction pipeline. The proposed imaging system captures the synchronized global stereo videos and high-resolution local videos with the auto white balance for keeping color consistency and stereo rectification for reducing the distortion. The hybrid multi-view videos are assembled using a global stitching scheme for generating stereo videos and a local embedding scheme for matching the high-resolution local videos. The high-quality global depth video is obtained by learning-enabled estimation, and local depth video is refined with the guidance of the high-resolution RGB images. Finally, the reconstructed multiscale RGB-D video is used for rendering the multiscale interactive 3D VR content at eye-limited resolution with the proposed multi-layer rendering approach.

3D scene rendering, semantic information is utilized to help depth estimation. Then, for the regions covered by local cameras, a color image guided refinement step is applied. As shown in Fig.4, our depth estimation pipeline consists of two levels, global level for robust depth map estimation and local level for local depth map refinement.

4.2.1 Global level Depth Estimation

Since the human vision system is sensitive to visual artifacts but insensitive to accurate absolute depth, a good depth map for immersive VR experience must keep the hierarchy of the scene and follow the semantic information. For example, if the depth relationship of the foreground and background is wrong or the building plane is distorted, the overall rendering result will be severely affected. Therefore, for large scale global area depth estimation, we focus more on the hierarchy of the front and back scenes and the correct semantic information. Besides the search range of the disparity in the large scale scene is wide, and widespread repeated texture areas are inevitable, which also increases the estimation difficulty.

Based on such insight, we propose our learning-based global level depth estimation algorithm, which can take the semantic information into consideration and tend to generate a smooth and large scale result rapidly. To generate the depth maps suitable for rendering, we present the spatial propagation layers, plane based correction module and novel hierarchical supervision loss. Our whole global depth estimation scheme is composed of the following parts.

Feature Extraction The shared weight feature pyramid is used in our network to extract the feature maps from two stereo images. To reduce the complexity of solving large scale feature maps, we use the coarse-to-fine strategy to extract four decreasing spatial resolution feature maps. Afterward, we adopt the encoder-decoder structure with skip connection to fuse the different level feature maps, and the SPP structure [53] is adopted to extend the receptive field for a large search range.

4D Cost Volume With the extracted feature maps, we use the distance metric between the left and the right feature maps across each disparity level to construct a 4D (channel number C, height H, width W and disparity D) disparity cost volume. And there are four different scale cost volumes corresponding to the four coarse-to-fine feature pyramids.

Cost Aggregation Decoder 3D convolution layers can extract the semantic information and aggregate the matching cost to improve the disparity quality. Inspired by the PSMnet [53], we adopt the stacked hourglass structure to learn more semantic information.

Disparity Regression We use a differentiable soft-argmin operation to obtain the disparity map from the cost volume. The probability of each candidate disparity is calculated from the predicted cost c_d using $softmax$ operation $\sigma()$, and the predicted disparity is the sum of each disparity d weighted by its probability:

$$\hat{d} = \sum_{d=0}^{D_{max}} d \times \sigma(-c_d). \tag{1}$$

Spatial Propagation The spatial propagation can improve the disparity maps by propagating the accurate estimated disparity across the entire disparity map. We form the spatial propagation layers based on the SPNetwork [54]. The spatial propagation layers can extract the affinity matrix from RGB images and propagate the disparity in four directions with the affinity matrix. With our spatial propagation layers, the small abrupt error areas can be corrected and the output disparity maps are smoother.

Loss Function First, we adopt the smooth $L1$ loss for better convergence:

$$\begin{aligned} L_{s1}(d,\hat{d}) &= \frac{1}{N}\sum_{i=1}^{N} l(d_i - \hat{d}_i), \\ l(x) &= \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & |x| \geqslant 1, \end{cases} \end{aligned} \tag{2}$$

where N is the number of the ground truth disparity, d is the ground truth disparity, and $\hat{d}$ is the predicted disparity.

To further improve the disparity map, we propose a hierarchical supervision loss function. For each candidate disparity level, the post-softmax probability of each pixel composes the probability map of the corresponding disparity level, and we refer the produced probability map and its candidate disparity value as candidate disparity map. The final disparity map is the sum of the candidate disparity maps. In other words, candidate disparity maps can be regarded as the result of decomposing the final disparity map and describe the hierarchy of the front and back scenes. Directly supervising the candidate disparity maps

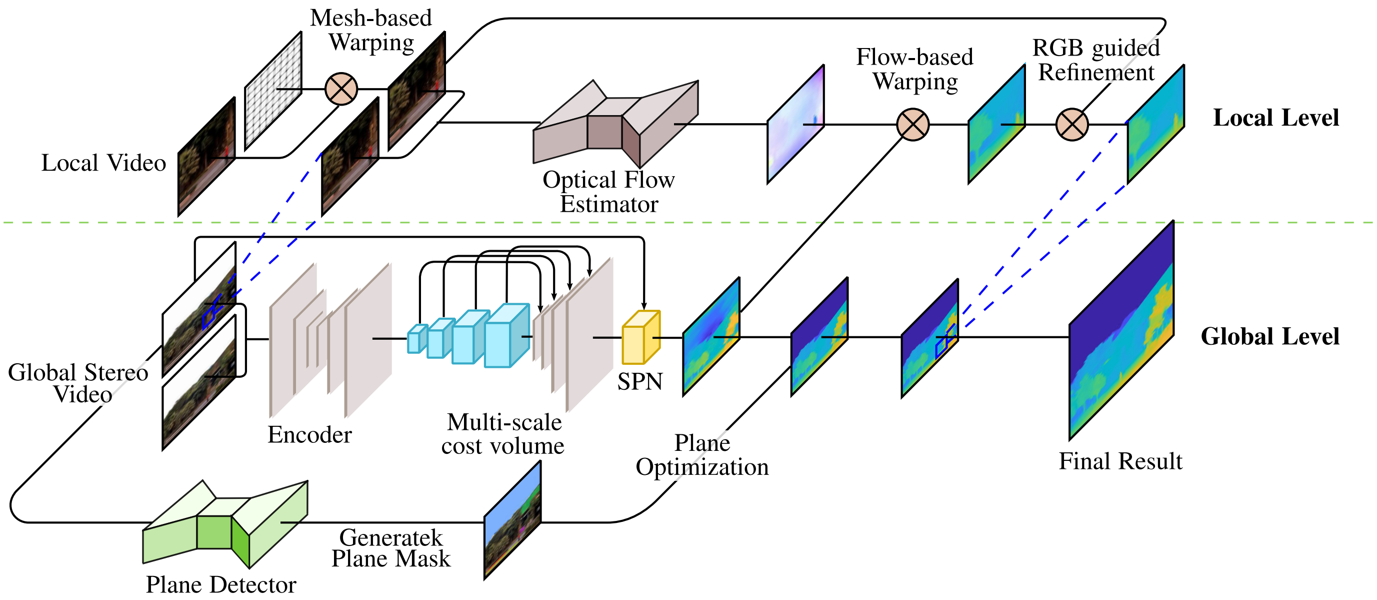


Fig. 4. Depth estimation pipeline of the proposed Multiscale-VR. With the captured video streams from the global stereo and local monocular cameras, the proposed multi-stage deep learning-based algorithm is used to estimate the global level depth from global stereo video streams. By incorporating the spatial propagation network (SPN) and plane detection optimization, the result can follow the semantic information of the scene and is suitable for large scale rendering. To provide the multi-scale VR interaction, the local level depths are reconstructed from the local unconstructed video streams with the flow-based estimator. The RGB guided refinement is further applied to the local depth to generate the visual immersive 3D perception, after which the local area depth is stitched back into the global depth map for cross-scale rendering.

can make the probability distribution of the weighted disparity more concentrated and emphasizes the hierarchy of the scene. Therefore, error-prone details can be corrected and the continuity to objects with the same disparity value can be improved. The ground truth disparity map can be decomposed into D sub-disparity maps, where D is the number of the candidate disparity maps.

$$d_{gt} = \frac{d_{gt} - d_l}{d_u - d_l} \times d_u + \frac{d_u - d_{gt}}{d_u - d_l} \times d_l, \tag{3}$$

d_{gt} is the ground truth disparity, $d_l = max(\{d_c | d_c < d_{gt}\})$ and $d_u = min(\{d_c | d_c \geqslant d_{gt}\})$, where d_c represent each candidate disparity values.

In this way, we can slice the ground truth disparity map and supervise each candidate disparity maps directly:

$$L_{submap}(D, \hat{D}) = \frac{1}{D_{max}} \sum_{i=0}^{D_{max}} l(D - \hat{D}), \tag{4}$$

where D is the sliced ground truth disparity maps and $\hat{D}$ is the weighted candidate disparity maps. And the final loss function is $L_i = L^i_{s1} + w \times L^i_{submap}$, where w is the weight and i represents the loss on the different coarse-to-fine output levels.

Finally, in the training phase, we supervise the four different scale output disparity maps:

$$L = L_1 + 0.5L_2 + 0.25L_3 + 0.125L_4. \tag{5}$$

Plane Based Correction In order to further enhance the smoothness of the plane and remove abnormal disparity value, we use the plane segmentation algorithm [55] to segment the plane areas and correct error values. Using coordinates and initial depth values, we can fit the plane equation by minimizing the following function:

$$\begin{aligned} &\min \sum_{i=0}^{n-1} (ax_i + by_i + cz_i + d)^2 \\ &\text{s.t.} \quad a^2 + b^2 + c^2 = 1, \end{aligned} \tag{6}$$

where (x_i, y_i) and z_i are coordinate and depth value of the pixel i respectively, and n is the total number of the pixels. With plane equation, we can correct the disparity values of the plane.

For better visual effects, we refine the disparity of objects containing semantic information. For example, we set the disparity value of the segmented sky regions to 0 and treat the distant person as a plane since the body depth can be ignored in large scale scenes.

4.2.2 Local Depth refinement

Through our global depth estimation pipeline, we can generate the high-quality depth maps suitable for rendering but high-resolution local areas call for more detailed local depth maps. Therefore, we propose an RGB guided refinement method to refine the depth map in local regions. Our embedding scheme can embed the local image into the global image seamlessly, but it cannot warp the local image pixel by pixel. Also, the local image and its corresponding global region are not perfectly aligned. Therefore, we cannot directly use local images to guide global images aligned disparity maps. Therefore, before the RGB guided refinement, it is necessary to warp the disparity maps in the local region to align it with the local RGB image. Considering that after mesh-based homography warping operation, the local image and its corresponding global region is very similar, we estimate the optical flow between the two images by PWCnet [56]. Based on the optical flow, we warp the disparity maps to achieve the alignment between the local RGB images and its disparity maps:

$$P_w(x, y) = P(x + f_x, y + f_y), \tag{7}$$

where P_w is the disparity map after warping, (x, y) is the coordinate position, and f_x, f_y is the x,y components of optical flow.

With aligned local RGB images, we adopt the bilateral solver [57] to refine the local disparity maps efficiently based on the structure of high resolution local RGB images. Assume that the target disparity map is t, the per-pixel confidence map of t is c and

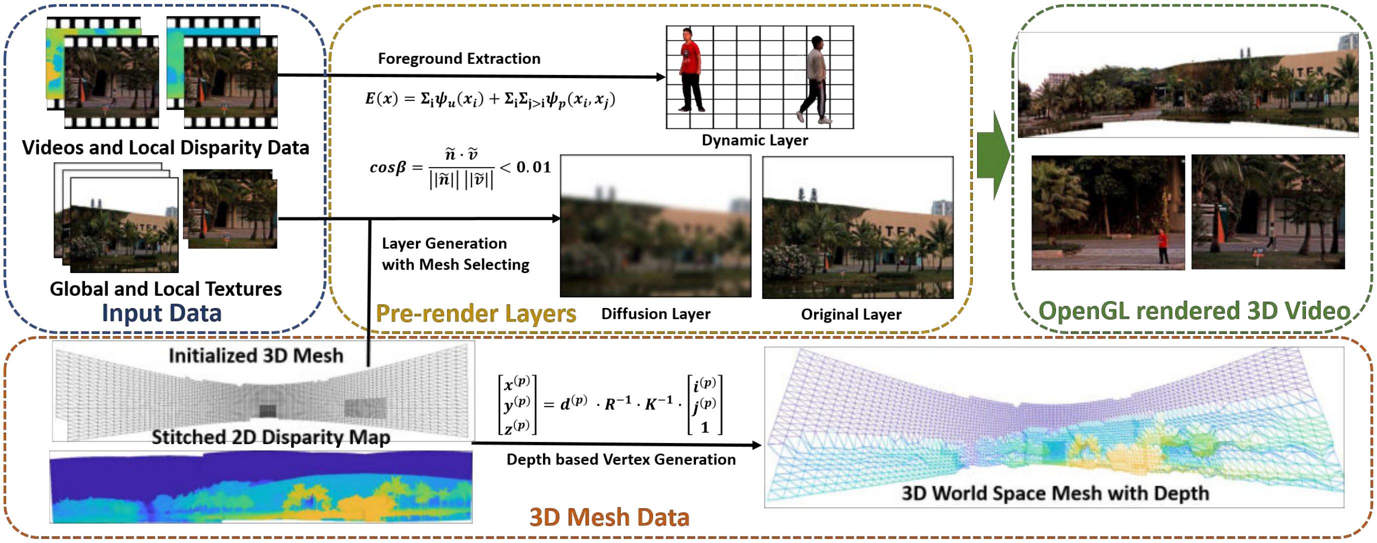

Foreground Extraction
$E(x) = \Sigma_i \psi_u(x_i) + \Sigma_i \Sigma_{j>i} \psi_p(x_i, x_j)$
Dynamic Layer
Videos and Local Disparity Data
$cos\beta = \frac{\tilde{n} \cdot \tilde{v}}{||\tilde{n}||\,||\tilde{v}||} < 0.01$
Layer Generation with Mesh Selecting
Global and Local Textures
Input Data
Pre-render Layers
Diffusion Layer
Original Layer
OpenGL rendered 3D Video
Initialized 3D Mesh
Stitched 2D Disparity Map
$\begin{bmatrix} x^{(p)} \\ y^{(p)} \\ z^{(p)} \end{bmatrix} = d^{(p)} \cdot R^{-1} \cdot K^{-1} \cdot \begin{bmatrix} i^{(p)} \\ j^{(p)} \\ 1 \end{bmatrix}$
Depth based Vertex Generation
3D World Space Mesh with Depth
3D Mesh Data

Fig. 5. Our multi-layer based rendering pipeline. With the reconstructed high-quality depth map, the three-layer rendering schematic is proposed to generate the 3D VR content, which comprises an original layer for scene rendering, a diffusion layer for handling occlusion regions, and a dynamic layer for moving object updating.

the refined disparity map x is obtained by solving the following function:

$$\min \sum_{i,j} \hat{W}_{i,j}(x_i - x_j)^2 + \sum_i c_i(x_i - t_i)^2, \tag{8}$$

where $\hat{W}_{i,j}$ is an affinity matrix which can be obtained from the reference image in the YUV color space. Since the resolution of the local image is higher, the RGB guided refinement can add more details of semantic structure into disparity maps.

4.3 Layer-based Rendering

To render our gigapixel 3d videography in realtime , we propose an efficient 3-layer rendering scheme, as illustrated in Fig.5, which comprises of (1) Original layer for high-resolution 3d video rendering; (2) Diffusion layer for occlusion handling; and (3) Dynamic layer for efficient dynamic foreground rendering.

Original Layer. The original layer aims to render the high-resolution videos. We backproject the stitched disparity map to 3D world coordinates to generate the background mesh, and draw the stitched high quality panorama on it:

$$\begin{bmatrix} x^{(p)} \\ y^{(p)} \\ z^{(p)} \end{bmatrix} = d^{(p)} \cdot R^{-1} \cdot K^{-1} \cdot \begin{bmatrix} i^{(p)} \\ j^{(p)} \\ 1 \end{bmatrix}, \tag{9}$$

where $\{K, R\}$ denotes the intrinsic and extrinsic camera parameters, $j^{(p)}$ and $i^{(p)}$ is the pixel location of the point p in image plane, $d^{(p)}$ is the calculated depth value of the pixel, $x^{(p)}$, $y^{(p)}$ and $z^{(p)}$ denote the rendering location of pixel p. For regions covered by local cameras, we increase the mesh vertex density for better depth quality while zooming in.

Diffusion Layer. Rendering using single layer mesh easily yields stretched triangle artifacts at depth edges when moving the viewpoints, as shown in Fig.10(a). To refine these artifact, we first tear the mesh by removing the meshes whose normal direction has a large angle to the view direction:

$$\cos\beta = \frac{\tilde{\mathbf{n}} \cdot \tilde{\mathbf{v}}}{||\tilde{\mathbf{n}}||\,||\tilde{\mathbf{v}}||} < 0.01, \tag{10}$$

where $\tilde{\mathbf{n}}$ is the normal vector of the face, $\tilde{\mathbf{v}}$ denotes the view direction from the face center to the optical center, β represent the angle between them.

TABLE 1
Hardware Specification

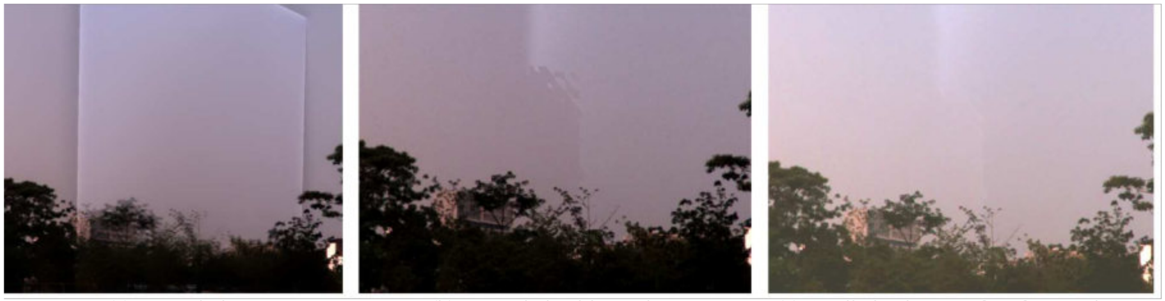
Global Stereo System	
Global Camera	XIMEA MC050CG-SY
Resolution	2464 × 2056
FOV	39.8° × 30.4° (12mm)
Baseline	45cm
Local System	
Local Camera	XIMEA MC031CG-SY
Resolution	2064 × 1544
FOV	11.2° × 8.4° (36mm)
Capturing System	
Number of servers	2
CPU	Intel Core i7-8700K
Graphics Card	NVIDIA 1080 Ti

After tearing, the stretched triangles disappear, but holes will occur while moving the viewpoint. To solve this problem, we add a diffusion layer to inpaint these holes. Inspired by the two-layer rendering scheme proposed by Hedman *et al.* [15], we fill the hole by adding a diffusion layer on the back of the original layer with a blurred panoramic image as texture.

Dynamic Layer. In our efficient implementation, we only update the mesh for the dynamic foreground, which is separated from the background first. The foreground can be initially extracted by the Gaussian mixture model (GMM) background subtraction [58], [59]. Since the dynamic mask generated by GMM is coarse in the object boundary, we adopt efficient inference dense CRF model [60] to obtain a sharp boundary mask.

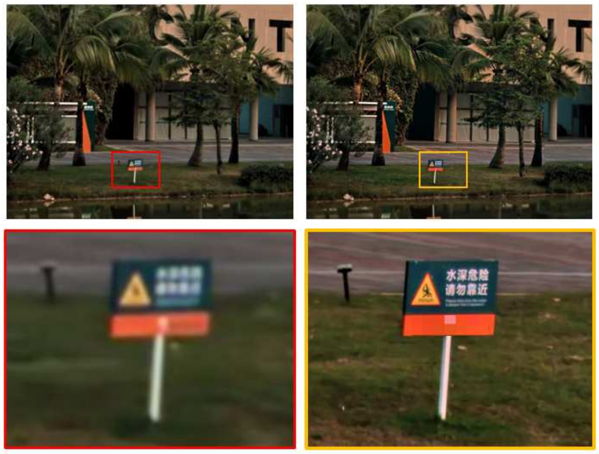
With the high quality dynamic masks, for each new frame, we re-calculate the 3D vertices belonging to the dynamic masks to render the dynamic objects.

The whole layer based rendering pipeline can generate high quality panoramic rendering result, especially in the local areas, which improves the visual effects and supports our zooming in function. Besides, with the diffusion layer, the artifact caused by occlusion can be eliminated and the dynamic layer can efficiently update dynamic areas.



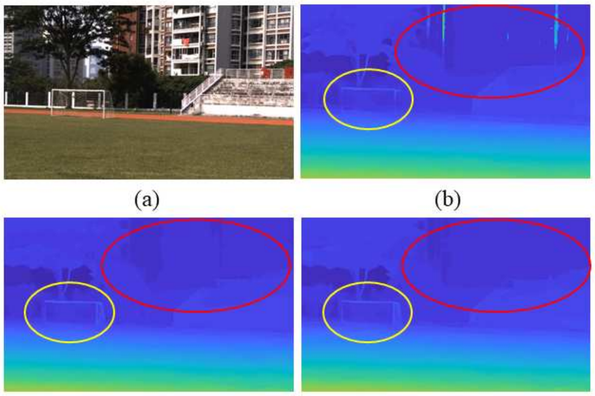
(a) Raw stitch (b) Raw stitch with graph-cut (c) Applied color transfer after graph-cut

Fig. 6. Evaluating the performance of the proposed global stereo stitching algorithm. (a) The original raw stitch suffers from severe camera location error and unnatural artifact. (b) Raw stitch with graph-cut algorithm optimization still suffers from color inconsistency. (c) Our full global stitching achieves seamless results for immersive VR experience.



(a) (b)

Fig. 7. The effectiveness of the hybrid architecture. Without the unconstructed local views, the global panoramic result cannot provide the details of local regions (a). The adapting and embedding of the local views to the global views offers ten times zoom in to local regions while preserves an eye-limited spatial resolution (b).



(a) (b) (c) (d)

Fig. 8. Evaluation of the global depth estimation. (a) Color images of the captured scene. (b) With the SPN layer, without hierarchical supervision. (c) Without the SPN layer, with hierarchical supervision. (d) With both the SPN layer and hierarchical supervision.

5 Experiment

5.1 Hardware

We built the Multiscale-VR system with the devices listed in Table. 1. Each stereo camera pair, as well as unstructured local cameras, are fixed on a 450mm-baseline aluminium shelf by structural components. Then, we assembled the camera array on a carbon-fibre made center support axle, which is connected to the tripod. A pre-calibration process is required within each stereo camera pair and no other part has to be calibrated. During video capturing, all the video streams are sent to two computing servers through USB 3.1 ports and compressed to H265 format for further processing.

5.2 RGB Texture

As shown in Fig.6, the stitching without graph-cut and color correction leads to severe artifacts, while a further enhancement using graph-cut still suffers from unnatural color jumps. Comparably, our global stitching scheme achieves seamless and globally consistent results to generate more immersive VR content.

The local views are warped to the global background by cross-scale matching and warping. After wrapping, we correct the color using linear Monge-Kantorovitch color transfer [52]. Fig.7 demonstrates the local to global embedding result, (a) shows the global views part, which is blurry and lacks details, (b) shows the embedded local view. From the magnified area, we can observe the significant improvement regarding the spatial resolution by our local embedding scheme.

5.3 Depth

While most stereo matching networks use sceneflow [61] to pretrain the model and use KITTI [62] to fine-tune, it does not apply for our system, as the disparity search range in our scenario (disparity > 384) is much larger than the datasets. Thus, we keep those pixels with large disparity by release the disparity threshold. For generalization, we add the Middlebury [63], eth3d [64] and HR-VS [63] datasets for training.

In Fig.12(b), we show the smooth and continuous stitched disparity map of our global scene, which is suitable for large-scale rendering. Our method is also able to guarantee the integrity of the plane with repeated color textures (e.g., meadow, tarmac). With the hierarchical supervision strategy and the spatial propagation layers, our method also can refine the detail and correct the abrupt disparity areas. In Fig.8, we evaluate our hierarchical supervision strategy and the spatial propagation layers for global depth estimation. With the hierarchical supervision strategy, the buildings in the disparity map are formed as a whole without abrupt disparity error. The spatial propagation strategy further propagates the accurate disparity to make the plane areas intact.

We further evaluate our RGB-guided local depth refinement scheme, which can not only enrich the semantic details of the disparity map, but also align the local disparity map to its corresponding color image. As shown in Fig.9, the results without local depth refinement suffer from severe artifacts, i.e., the excessively slender arms and the misaligned head region of the

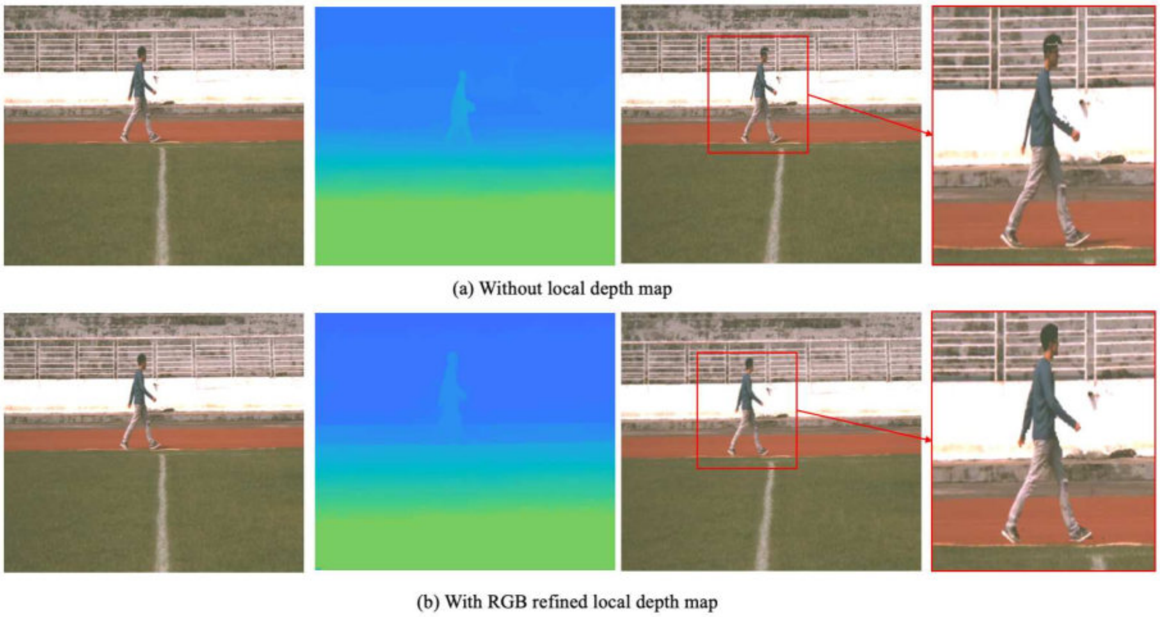

(a) Without local depth map
(b) With RGB refined local depth map

Fig. 9. Evaluating the performance of the local depth refinement scheme. (a,b) The results without and with local refinement, respectively. From left to right: the original RGB images captured by the local camera; the corresponding depth map; the rendering results. Note that with our local refinement, the depth detail of the local region is significantly improved, especially in the head and arm regions of the captured person.

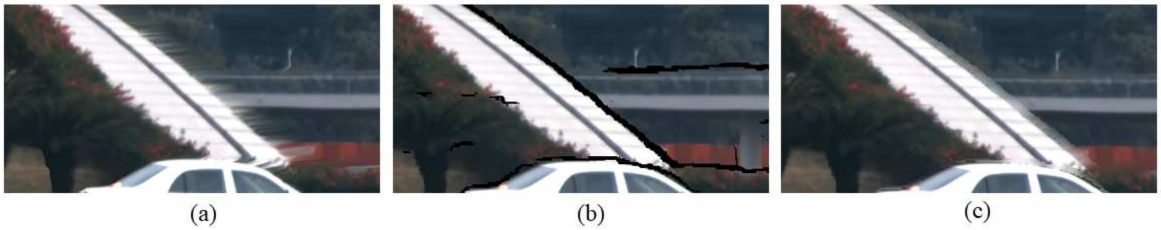


Fig. 10. Rendering on depth discontinuous regions. (a) Without using the discard strategy, the edges are pulled out with a stretching artifact. (b) With discard strategy but the lack of under-course diffusion layer brings hiatus to the image. (c) Our rendering result by applying the diffusion layer after enabled discard strategy. Note that the disturbing stretching artifact or fissure are successfully suppressed.

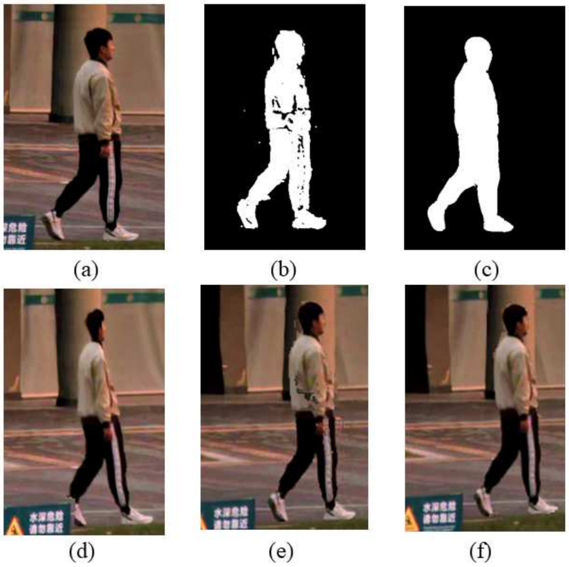


Fig. 11. Evaluation of our dynamic layer updating. (a) The input RGB image. (b) Coarse-grained mask generated by GMM. (c) Dense CRF optimized mask. (d) Rendering result without a dynamic mask. (e) Rendering result with a dynamic mask generated by GMM. (f) Rendering result with a dynamic mask optimized by dense CRF. Note that the dense-CRF optimized mask has the sharper outline of the person, and the rendering results combined with this mask is significantly more vivid and accurate in the local view.

captured person. Comparably, our refined local scheme solves the aforementioned problems, leading to a better visual effect with more accurate semantic details.

5.4 Rendering

The proposed rendering pipeline can reduce the stretching artifact caused by depth discontinuity. As shown in Fig 10, there are noticeable stretching artifacts around the depth of discontinuous areas. With our discarding strategy, we replace the meshes of these areas with diffusion layer to make the rendering results of these areas smoother, and subsequently improve visual effect.

For dynamic layer updating, the dense CRF algorithm is used to refine foreground masks generated by GMM. As shown in Fig.11(b), the GMM can provide the initial rough mask. After dense CRF optimization, the outline of mask in Fig.11(c) is more faithful, which improves the rendering results of dynamic layer. As shown in Fig.11 (d)(e)(f), our refined mask rejects the artifacts, leading to more immersive rendering result in the local view.

In summary, our efficient rendering pipeline can take advantage of our high quality depth maps, and generate the vivid VR content. More results are shown in our Supplementary Video.

6 LIMITATION

Although our Multiscale VR system is able to produce unique scalable VR experiences, computational complexity is still a

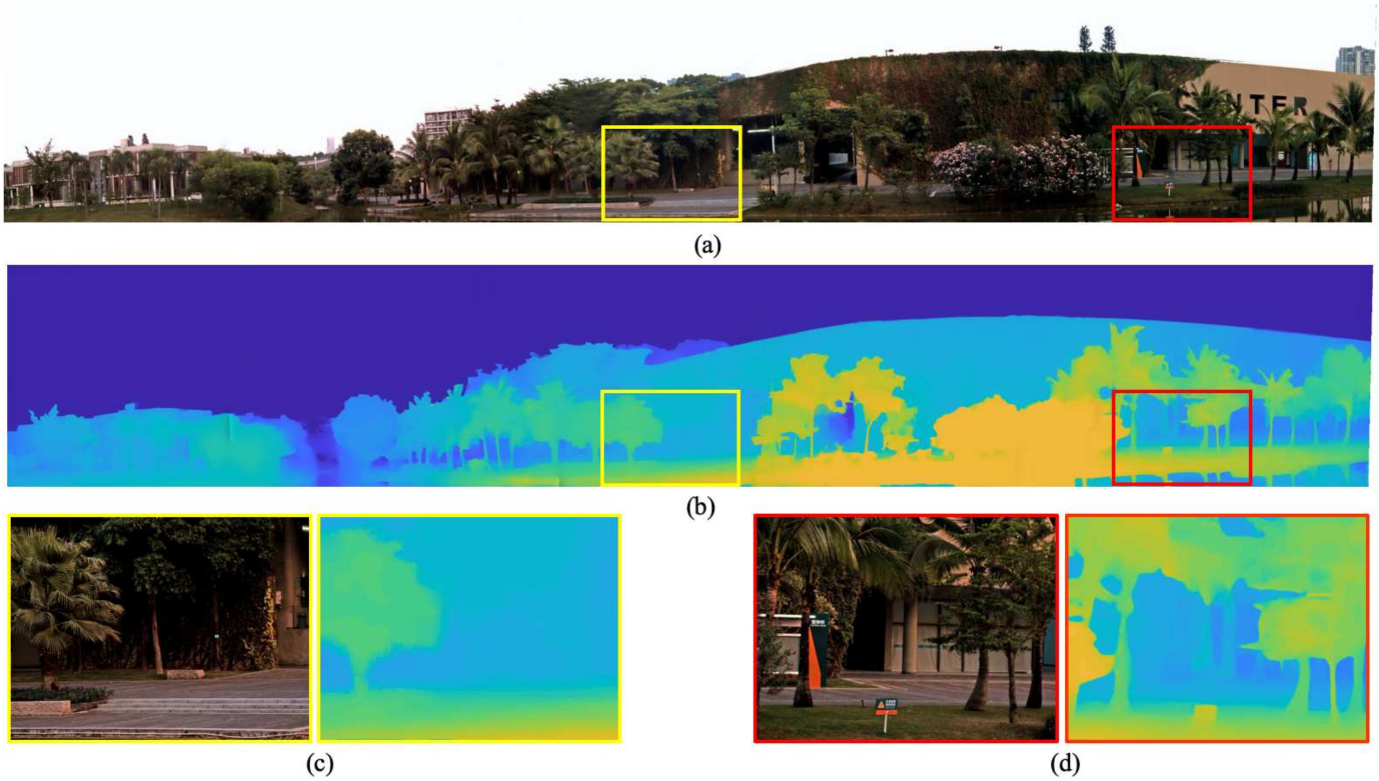


Fig. 12. Experimental result on the “Campus” scene with the proposed Multiscale-VR content generation architecture. (a) The “Campus” scene is dynamically captured with the multiscale unstructured camera array system at video framerate, where the captured unconstructed global and local video sequences are seamless stitched and fused to generate the multiscale gigapixel panoramic video. (b) In order to create the 3D perception, the corresponding depth video is efficiently reconstructed with the proposed depth estimation pipeline based on pre-trained deep neural network models. (c, d) The proposed Multiscale-VR content generation architecture provides high-resolution, i.e., 0.0037-degree of angular resolution, allowing immersive retinal resolution 3D scenes.

stumbling block for live broadcasting our system’s high-resolution immersive VR content. Further optimization can be made by improving algorithm efficiency and develop a video streaming module suitable for home users to enjoy interactive VR content. For example, in VR sport applications, audiences can zoom in their regions of interests of the playfield to observe details of their favourite player’s performance. In terms of the mechanical structure, the current system volume is mainly limited by the use of standard mechanical structure and components. The proposed hybrid camera array for VR imaging could be made much more compact and portable by adopting higher integration mechanical design and replace standard mechanical components with customized modules.

7 Conclusion

We demonstrate the multi-scale unstructured cylindrical distributed camera array for high-resolution 3D virtual reality content generation. The captured gigapixel video sequences are efficiently processed with the gigapixel videography schemes followed by deep learning-enabled depth estimation. We captured various scenes with our system and demonstrated the effectiveness of the proposed approach for creating interactive gigapixel virtual reality content, with which the users can zoom in to regions of interest and enjoy high-resolution 3D scenes with an unprecedented immersive experience. We believe the proposed Multiscale-VR content generation architecture will empower the new generation of VR and near-eye display system to arbitrary move the virtual viewpoint in the 3D scene and will promote its popularity in the areas of entertainment, surveillance, sports, cultural heritage and more.

References

[1] R. Anderson, D. Gallup, J. T. Barron, J. Kontkanen, N. Snavely, C. Hernández, S. Agarwal, and S. M. Seitz, “Jump: virtual reality video,” *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, p. 198, 2016.

[2] Facebook, “Facebook Surround 360,” https://github.com/facebook/Surround360, 2019.

[3] Insta360, “Insta360 Pro 2,” https://www.insta360.com/product/insta360-pro2/, 2019.

[4] I. GoPro, “Gopro Max,” https://gopro.com/en/us/shop/cameras/max/, 2019.

[5] H. Corporation, “Htc Vive,” https://www.vive.com/us/, 2019.

[6] L. Facebook Technologies, “Oculus Quest,” https://www.oculus.com/quest/, 2019.

[7] Samsung, “Samsung HMD Odyssey,” https://www.samsung.com/us/computing/hmd/windows-mixed-reality/hmd-odyssey-windows-mixed-reality-headset-xe800zba-hc1us/, 2019.

[8] YI, “Yi halo,” https://www.yitechnology.com/yi-halo-vr-camera, 2019.

[9] R. S. Overbeck, D. Erickson, D. Evangelakos, and P. Debevec, “The making of welcome to light fields VR,” in *ACM SIGGRAPH 2018 Talks*. ACM, 2018, p. 63.

[10] A. P. Pozo, M. Toksvig, T. F. Schrager, J. Hsu, U. Mathur, A. Sorkine-Hornung, R. Szeliski, and B. Cabral, “An integrated 6DoF video camera and system design,” *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, p. 216, 2019.

[11] Samsung, “Samsung Gear 360,” https://www.samsung.com/global/galaxy/gear-360/, 2017.

[12] R. Konrad, D. G. Dansereau, A. Masood, and G. Wetzstein, “SpinVR: towards live-streaming 3D virtual reality video,” *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, p. 209, 2017.

[13] Kandao, “Kandao obsidian R,” https://www.kandaovr.com/en/Obsidian/, 2019.

[14] V. Couture, M. S. Langer, and S. Roy, "Analysis of disparity distortions in omnistereoscopic displays," *ACM Transactions on Applied Perception (TAP)*, vol. 7, no. 4, p. 25, 2010.

[15] P. Hedman, S. Alsisan, R. Szeliski, and J. Kopf, "Casual 3D photography," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, p. 234, 2017.

[16] P. Hedman and J. Kopf, "Instant 3D photography," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 101, 2018.

[17] J. Huang, Z. Chen, D. Ceylan, and H. Jin, "6-DOF VR videos with a single 360-camera," in *IEEE Virtual Reality*. IEEE, 2017, pp. 37–44.

[18] B. Luo, F. Xu, C. Richardt, and J.-H. Yong, "Parallax360: stereoscopic 360 scene representation for head-motion parallax," *IEEE transactions on Visualization and Computer Graphics*, vol. 24, no. 4, pp. 1545–1553, 2018.

[19] P. Hedman, T. Ritschel, G. Drettakis, and G. Brostow, "Scalable inside-out image-based rendering," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, p. 231, 2016.

[20] M. Ogata, H. Wada, J. van Baar, and R. Raskar, "A unified calibration method with a parametric approach for wide-field-of-view multiprojector displays," in *Proceedings of the IEEE Virtual Reality Conference*. IEEE, 2009, pp. 235–236.

[21] D. Lanman and D. Luebke, "Near-eye light field displays," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, p. 220, 2013.

[22] N. Matsuda, A. Fix, and D. Lanman, "Focal surface displays," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 86, 2017.

[23] N. Padmanaban, R. Konrad, T. Stramer, E. Cooper, and G. Wetzstein, "Optimizing virtual reality for all users through gaze-contingent and adaptive focus displays," *Proceedings of the National Academy of Sciences*, 2017. [Online]. Available: http://www.pnas.org/content/early/2017/02/07/1617251114.abstract

[24] N. Padmanaban, R. Konrad, and G. Wetzstein, "Autofocals: Evaluating gaze-contingent eyeglasses for presbyopes," *Science Advances*, vol. 5, no. 6, p. eaav6187, 2019.

[25] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in VR: How do people explore virtual environments?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 4, pp. 1633–1642, 2018.

[26] A. Serrano, V. Sitzmann, J. Ruiz-Borau, G. Wetzstein, D. Gutierrez, and B. Masia, "Movie editing and cognitive event segmentation in virtual reality video," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 47, 2017.

[27] A. Serrano, I. Kim, Z. Chen, S. DiVerdi, D. Gutierrez, A. Hertzmann, and B. Masia, "Motion parallax for 360 RGBD video," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 5, pp. 1817–1827, 2019.

[28] W. Cui and L. Gao, "All-passive transformable optical mapping near-eye display," *Scientific Reports*, vol. 9, 2019.

[29] S. Lee, Y. Jo, D. Yoo, J. Cho, D. Lee, and B. Lee, "Tomographic near-eye displays," *Nature Communications*, vol. 10, no. 1, p. 2497, 2019.

[30] N. Padmanaban, T. Ruban, V. Sitzmann, A. M. Norcia, and G. Wetzstein, "Towards a machine-learning approach for sickness prediction in 360 stereoscopic videos," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 4, pp. 1594–1603, 2018.

[31] R. Konrad, A. Angelopoulos, and G. Wetzstein, "Gaze-contingent ocular parallax rendering for virtual reality," in *arXiv*, 2019.

[32] J. Kopf, M. Uyttendaele, O. Deussen, and M. F. Cohen, "Capturing and viewing gigapixel images," in *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3. ACM, 2007, p. 93.

[33] O. S. Cossairt, D. Miau, and S. K. Nayar, "Gigapixel computational imaging," in *Proceedings of the IEEE International Conference on Computational Photography*. IEEE, 2011, pp. 1–8.

[34] D. J. Brady, M. E. Gehm, R. A. Stack, D. L. Marks, D. S. Kittle, D. R. Golish, E. Vera, and S. D. Feller, "Multiscale gigapixel photography," *Nature*, vol. 486, no. 7403, p. 386, 2012.

[35] J. Holloway, M. S. Asif, M. K. Sharma, N. Matsuda, R. Horstmeyer, O. Cossairt, and A. Veeraraghavan, "Toward long-distance subdiffraction imaging using coherent camera arrays," *IEEE Transactions on Computational Imaging*, vol. 2, no. 3, pp. 251–265, 2016.

[36] X. Yuan, L. Fang, Q. Dai, D. J. Brady, and Y. Liu, "Multiscale gigapixel video: a cross resolution image matching and warping approach," in *Proceedings of the IEEE International Conference on Computational Photography*. IEEE, 2017, pp. 1–9.

[37] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," in *ACM Transactions on Graphics (TOG)*, vol. 24, no. 3. ACM, 2005, pp. 765–776.

[38] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 59–73, 2007.

[39] W.-S. Lai, O. Gallo, J. Gu, D. Sun, M.-H. Yang, and J. Kautz, "Video stitching for linear camera arrays," *arXiv preprint arXiv:1907.13622*, 2019.

[40] N. Bonneel, J. Tompkin, D. Sun, O. Wang, K. Sunkavalli, S. Paris, and H. Pfister, "Consistent video filtering for camera arrays," in *Computer Graphics Forum*, vol. 36, no. 2. Wiley Online Library, 2017, pp. 397–407.

[41] D. S. Kittle, D. L. Marks, and D. J. Brady, "Automated calibration and optical testing of the AWARE-2 gigapixel multiscale camera," in *IS&T/SPIE Electronic Imaging*, 2013.

[42] D. S. Kittle, D. L. Marks, H. S. Son, J. Kim, and D. J. Brady, "A testbed for wide-field, high-resolution, gigapixel-class cameras," *Review of Scientific Instruments*, vol. 84, no. 5, p. 053107, 2013.

[43] I. Stamenov, A. Arianpour, S. J. Olivas, I. P. Agurok, A. R. Johnson, R. A. Stack, R. L. Morrison, and J. E. Ford, "Panoramic monocentric imaging using fiber-coupled focal planes," *Optics Express*, vol. 22, no. 26, pp. 31 708–31 721, 2014.

[44] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2. IEEE, 2005, pp. 807–814.

[45] M. Bleyer, C. Rhemann, and C. Rother, "Patchmatch stereo-stereo matching with slanted support windows," in *Proceedings of the British Machine Vision Conference*, vol. 11, 2011, pp. 1–11.

[46] Y. Li, Y. Hu, R. Song, P. Rao, and Y. Wang, "Coarse-to-fine PatchMatch for dense correspondence," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2233–2245, 2018.

[47] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[48] F. Guney and A. Geiger, "Displets: resolving stereo ambiguities using object knowledge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4165–4175.

[49] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1592–1599.

[50] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5695–5703.

[51] V. Kwatra, A. Schödl, I. Essa, G. Turk, and A. Bobick, "Graphcut textures: image and video synthesis using graph cuts," *ACM Transactions on Graphics (ToG)*, vol. 22, no. 3, pp. 277–286, 2003.

[52] F. Pitie, "The linear Monge-Kantorovitch colour mapping for example-based colour transfer," 12 2007, pp. 1 – 9.

[53] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410–5418.

[54] S. Liu, S. De Mello, J. Gu, G. Zhong, M.-H. Yang, and J. Kautz, "Learning affinity via spatial propagation networks," in *Proceedings of the Advances in Neural Information Processing Systems*, 2017, pp. 1520–1530.

[55] C. Liu, K. Kim, J. Gu, Y. Furukawa, and J. Kautz, "Planercnn: 3d plane detection and reconstruction from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4450–4459.

[56] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *CVPR*, 2018.

[57] J. T. Barron and B. Poole, "The fast bilateral solver," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 617–632.

[58] Z. Zivkovic *et al.*, "Improved adaptive Gaussian mixture model for background subtraction." in *Proceedings of the International Conference on Pattern Recognition*. Citeseer, 2004, pp. 28–31.

[59] Z. Zivkovic and F. Van Der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, 2006.

[60] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Proceedings of the Advances in Neural Information Processing Systems*, 2011, pp. 109–117.

[61] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, 2016, pp. 4040–4048.

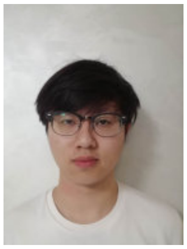
[62] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[63] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proceedings of the German Conference on Pattern Recognition*. Springer, 2014, pp. 31–42.

[64] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3260–3269.

Jianing Zhang is currently a MS student in Tsinghua University. He received his BS degree from Northeastern University of China. His research interests include Computational Imaging and Computer Graphics.

Tianyi Zhu received the BS degree in Department of Automation from Tsinghua University in 2017. He is currently studing towards PhD degree in Tsinghua-Berkeley Shenzhen Institute, Tsinghua University. His research interests include Computational Imaging, Light Field Imaging and Computer Graphics.



Anke Zhang received the BS and ME degree in Mechanical Engineering from University of California Berkeley in 2017 and 2018. He studied towards PhD degree in Tsinghua-Berkeley Shenzhen Institute from 2018. His research interested includes computational photography, imaging related deep learning and ultra-fast imaging.

Zihan Wang received the B.E. degrees in Telecommunications Engineering from Xidian University, Xi'an, China, and Heriot-Watt University, Edinburgh, UK, in 2019. He is currently pursuing the M.S. degree in Tsinghua-Berkeley Shenzhen Institute from 2019.

Xiaoyun Yuan received the B.S. degree in Department of Electronics Science and Technology, School for Information Science and Technology from University of Science and Technology of China, Hefei, China, in 2014. He is currently pursuing the PhD degree in the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong SAR.

Sebastian Beetschen received the B.E. degree from the Swiss Institute of Technology Lausanne (EPFL) and did an exchange program at the Technical University of Delft in the Netherlands. He is currently pursuing the M.S. degree in Tsinghua-Berkeley Shenzhen Institute from 2019.

Lan Xu received the B.S. degree in Department of Information and Communication, College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China, in 2015. He is currently pursuing the Ph.D. degree in the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong SAR. His current research interests include computer vision and computer graphics.

Xing Lin is currently a research scientist in the Beijing-Tsinghua Innovation Center for Future Chips, Tsinghua University. He received a B.E. degree in Electronic Engineering from Xidian University in 2010, and the Ph.D. degree in Automation from Tsinghua University in 2015. From 2012 to 2013, he was a visiting student at MIT Media Lab. He was a research associate with the Howard Hughes Medical Institute at Stanford University and a postdoctoral scholar at the University of California, Los Angeles (UCLA). His research interests are in the areas of optoelectronic computing, computational imaging, computer vision, and visual computing.

Dr. Qionghai Dai is a professor in the Department of Automation, and an adjunct professor in the School of Life Science, Tsinghua University. Dr. Dai is also an academician of Chinese Academy of Engineering. Dr. Dai is expertised in computational photography and stereoscopic vision.

Dr. Lu Fang is currently an Associate Professor in Tsinghua University. She received Ph.D from Hong Kong University of Science and Technology in 2011, and B.E. from University of Science and Technology of China in 2007, respectively. Her research interests include Computational Imaging and 3D Vision. Dr. Fang is currently IEEE Senior Member.