Zhang F, Zhao J, Zhang Y *et al.* A survey on 360° images and videos in mixed reality: Algorithms and applications. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 38(3): 473-491 May 2023. DOI: 10.1007/s11390-023-3210-1

A Survey on 360° Images and Videos in Mixed Reality: Algorithms and Applications

Fanglue Zhang[1] (张方略), *Member, ACM, IEEE*, Junhong Zhao[1,*] (赵军红)
Yun Zhang[2] (张　赟), *Senior Member, CCF*, and Stefanie Zollmann[3]

[1] *School of Engineering and Computer Science, Victoria University of Wellington, Wellington 6012, New Zealand*

[2] *College of Media Engineering, Communication University of Zhejiang, Hangzhou 310018, China*

[3] *Department of Computer Science, University of Otago, Dunedin 9054, New Zealand*

E-mail: fanglue.zhang@vuw.ac.nz; j.zhao@vuw.ac.nz; zhangyun@cuz.edu.cn; stefanie.zollmann@otago.ac.nz

Received March 6, 2023; accepted May 24, 2023.

Abstract Mixed reality technologies provide real-time and immersive experiences, which bring tremendous opportunities in entertainment, education, and enriched experiences that are not directly accessible owing to safety or cost. The research in this field has been in the spotlight in the last few years as the metaverse went viral. The recently emerging omnidirectional video streams, i.e., 360° videos, provide an affordable way to capture and present dynamic real-world scenes. In the last decade, fueled by the rapid development of artificial intelligence and computational photography technologies, the research interests in mixed reality systems using 360° videos with richer and more realistic experiences are dramatically increased to unlock the true potential of the metaverse. In this survey, we cover recent research aimed at addressing the above issues in the 360° image and video processing technologies and applications for mixed reality. The survey summarizes the contributions of the recent research and describes potential future research directions about 360° media in the field of mixed reality.

Keywords 360° image, mixed reality, 360° image processing, virtual reality scene reconstruction, virtual reality content manipulation

1 Introduction

Advances in computer graphics and mixed reality (MR) have allowed people to virtually teleport to a safari park on the other side of the real world or watch sports games with the feeling of being right in the middle of the action[1, 2]. This real-time, immersive experience provides tremendous opportunities in entertainment, education, and enriched experiences that are not directly accessible owing to safety or cost. The research in this field has been in the spotlight in the last few years as the metaverse went viral. The recently emerging omnidirectional video streams[3], i.e., 360° videos, provide an affordable way to capture and present dynamic real-world scenes. Recent research has successfully established theoretical and algorithmic foundations to capture[4], interpret[5, 6], stabilize[7, 8], and present 360° videos[9–12], which enable delivering astonishingly-good-quality, immersive, and panoramic content. Early MR applications based on 360° images and videos allow only limited interaction with the dynamically-captured real-world content[2, 13, 14]. For example, when enjoying sports games in a VR (virtual reality) headset, users are restricted to fixed virtual positions, and their interaction is limited to turning their head. In the last decade, fueled by the rapid development of artificial intelligence and computational photography technologies, the re-

Survey

Special Section of CVM 2023

The work is supported by the Marsden Fund Council managed by Royal Society of New Zealand under Grant Nos. MFP-20-VUW-180 and UOO1724, Zhejiang Province Public Welfare Technology Application Research under Grant No. LGG22F020009, and the Key Lab of Film and TV Media Technology of Zhejiang Province of China under Grant No. 2020E10015.

*Corresponding Author

search interest in MR systems using 360° videos with richer and more realistic experiences is dramatically increased to unlock the true potential of the metaverse. In this field, the following two research problems have been attracting a lot of attention.

- How to leverage the full field-of-view (FoV) information for better scene understanding and reconstruction?
- How to reconstruct an immersive and interactive environment for MR applications from real-world captures?

In this survey, we cover recent researches aimed at addressing the above issues in the 360° image and video processing technologies and applications for mixed reality. Section 2 describes the methodology used to conduct this survey. As a special kind of images, 360° images capture every single point around the camera in every possible viewing direction, which can be naturally defined on a sphere. However, to be compatible with the conventional 2D imaging pipelines, the raw 360° images need to be transformed into 2D planar representations preserving the omnidirectional information[10] (as shown in Fig.1). However, the methods designed for normal 2D images cannot be trivially extended to work for 360° images. In Section 3, we cover the current research efforts in 360° scene analysis and processing, where we first briefly introduce the different ways of effectively representing the spherical domain data to support 360° image processing (Subsection 3.1), followed by the review of the methods in semantic understanding (Subsection 3.2), depth estimation (Subsection 3.3), and temporal domain analysis (Subsection 3.4). The understanding and analysis methods of real-world 360° images are fundamental building blocks for MR applications introduced in Section 4. We describe the recent advances in building a more realistic virtual space from 360° content by allowing 6-degree-of-freedom navigation (Subsection 4.1), recovering the light condition (Subsection 4.2), and AR content mapping and localization (Subsection 4.3). We also review the recent advances in 360° content manipulation that allows more interactive MR applications. Finally, we conclude the survey and highlight open problems in this field in Section 5.

2 Methodology

We began the survey by conducting a literature search using keywords, followed by a literature selection based on the article types and their relevance to the defined research problems.

We first identified the related papers using Google Scholar for searching the literature such as papers, chapters, technical reports, theses, reviews, and books. Considering that the researchers have used a range of terms to refer to 360° images/videos and VR/MR applications, the initial keywords we used in searching are: “Augmented Reality” OR “Mixed Reality” OR “Virtual Reality” OR “VR” AND “Omnidirectional Video” OR “Panoramic Video” OR “Virtual Reality Video” OR “Immersive Video” OR “Panoramic Image” OR “360” OR “Omnidirectional Image”. We searched for matched keywords appearing in everything in the article. The initial search resulted in 1 010 papers published between 1970 and January 2023.

Second, we defined exclusion criteria to filter out irrelevant documents for our survey. In this paper, we intended to focus on applied research studies and the exclusion criteria were set as follows: theses, monographs, books, theoretical papers, duplicated publications, and review papers. For the remaining articles, we removed the articles with titles that are clearly

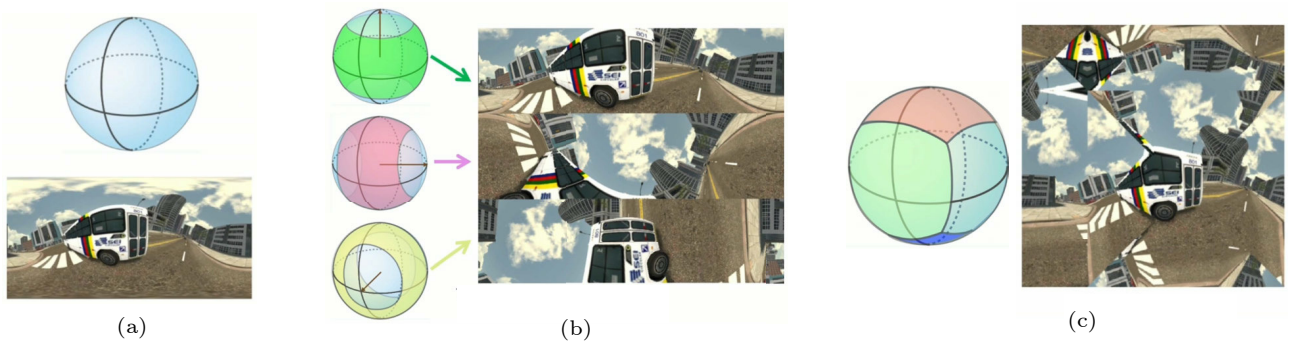


Fig.1. Different projections used in [10]. (a) Equirectangular projection. (b) Tri-cylindrical projection. (c) Cube-Padding projection. It has been demonstrated that utilizing different projection methods can provide complementary information to improve the performance of optical flow estimation using deep neural networks.

not relevant to the aforementioned research problems. For example, we removed the papers about 360° video streaming, 360° video acquisition, and VR sickness alleviation. We then further evaluated whether to include the articles by reading the abstract. The keyword search may not return all of the relevant studies, because we required the 360° related words and VR/MR related words appear simultaneously. Consequently, we discovered more studies by tracing the articles that cite the relevant articles we selected from our search. To limit the range of the research fields in the papers citing the relevant articles, we only traced the citations of the papers that are published in major computer graphics and VR venues, including ACM SIGGRAPH, ACM TOG, IEEE VR, IEEE TVCG, IEEE ISMAR, EuroGraphics, Pacific Graphics and CHI. Finally, 128 papers were collected.

3 360° Content Analysis and Understanding

Research on the next-generation MR applications based on 360° media relies heavily on the development of advanced content understanding and analysis algorithms to enable more functionalities. This section first reviews the fundamental representation used for properly processing 360° media defined in the spherical domain, and covers the methods specially designed for extracting both low-level and semantic-level scene information from 360° images and videos.

3.1 360° Image Representation

360° images, also known as omnidirectional images, were first introduced in 1970[15–18]. 360° images have a field of view that covers the entire sphere. Naturally, it is defined as a signal distributed on a sphere and thus can be considered as a spherical image. Each pixel or point of a spherical image is normally described by the view direction from the center of this sphere to the point itself. Note that this representation can satisfy the need for image-based lighting technology (IBL)[19], where 360° images are used as Environment Maps[19] to provide light rays from all directions. However, to adapt to the existing conventional 2D image processing pipeline, the spherical images have to be converted to a 2D plane while preserving the omnidirectional information. As shown in Fig.1(a), the equirectangular projection is the most common approach to mapping pixels from a sphere to a 2D rectangle. But similar to other projections, such as cube map[20] and cylindrical projection[21], there is no projection that can map the spherical surface to a 2D plane that is both an equal-area and conformal (angle-preserving) map. The reason is that the sphere is not developable[22]. Considering this mathematical fact, several classical researches[23, 24] attempted to find a trade-off between the competing goals of preserving angle and area when mapping textures on spherical surfaces.

Equirectangular Projection. Equirectangular projection maps meridians to vertical straight lines of constant spacing, and circles of latitude to horizontal straight lines of constant spacing. It has become a standard for many datasets[25–27] due to the simple relationship between a pixel in a rectangular map and its corresponding location on the sphere. But this projection is neither equal-area nor conformal because of the distortions introduced by the direct mapping from view angles to 2D positions. Therefore, when using this equirectangular representation to process 360° images, researchers usually use a set of local tangent images of sampled points on the spherical surface when extracting spatial features[28]. Coors *et al.*[28] and Zhao *et al.*[29] deformed their convolutional filters by projecting the pixels of a patch in local tangent images back to the equirectangular image to get the sampled points that have the equivalent spherical distance. In this way, the convolutional networks retain the original pixel connectivity and enable the transfer of perspective models to 360° images.

Tangent Image and Icosahedron. Eder *et al.*[30] proposed using a set of tangent images to represent an omnidirectional image, facilitating transferable 360° image and video processing tasks. They generated a few distortion-mitigated planar images tangent to a subdivided icosahedron to ensure that traditional computer vision methods like sparse feature detection and Simultaneous Localization and Mapping (SLAM)[31] can be applied to spherical images. Such a tangent image solution is effective in other important vision tasks such as optical flow estimation[31]. To further reduce the variance of spatial resolving power in representing 360° images, Lee *et al.*[32] developed a spherical polyhedron-based representation, SpherePHD, for deep omnidirectional image learning. They also designed a convolutional kernel on the polyhedron grid and an associated pooling strategy. In concurrent work, Zhang *et al.*[33] converted the spherical input to an unfolded icosahedron mesh and proposed Hexagonal filters to fit the existing deep neural net-

works for 2D image analysis. More recently, to ensure such a representation can be applied to higher-resolution images, Yoon *et al.*[34] proposed a continuous representation where each pixel is defined as a subdivided icosahedron. Wu *et al.*[35] provided a series of operations defined on such dense spherical triangle image elements.

Cube-Padding. A cube map is another traditional approach to mapping a spherical image to a 2D plane[20, 36]. Although it has the disadvantage of discontinuity along face boundaries, as shown in Fig.2, it can map the sphere to standard 2D images, and thus can reuse the well-trained 2D deep neural networks. To alleviate the issues when processing the objects lying across the cube map face boundaries, Cheng *et al.*[37] introduced Cube-Padding tailor-made for 360° videos, where each face is extended by considering a wider FoV. As shown in Fig.2, the padding operation is also performed on the spatial feature maps. A similar strategy is also utilized in [38].

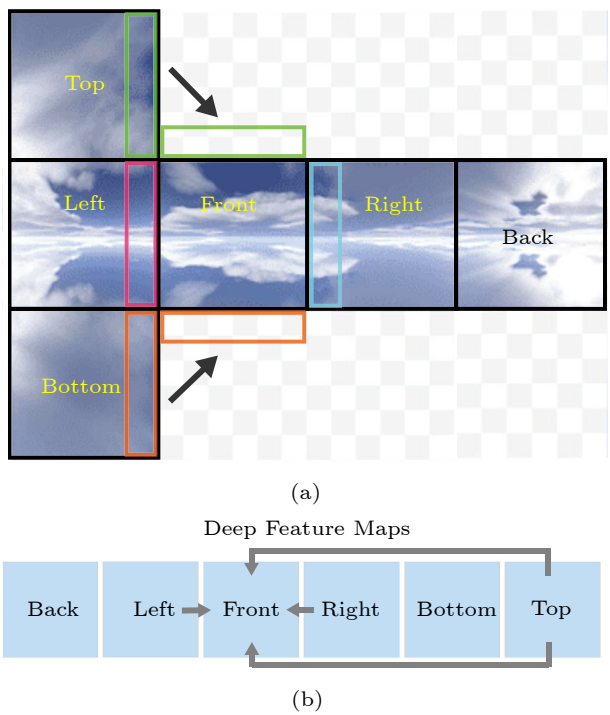


Fig.2. (a) Cube-Padding layout used in [37]. (b) The feature maps are also padded in different layers to ensure the spatial connectivity along the face boundaries.

Fusion of Different Projections. Since no such projection can fully preserve the original spatial relationships among the pixels distributed on a sphere, researchers[10, 39] proposed to learn to fuse the complementary information that can be obtained from different projections into a single final result. Li *et al.*[10] leveraged the fusion of results from different projections to estimate 360° optical flows and demonstrated that the performance consistently outperforms the single-projection optical flows that were fused. As shown in Fig.1, besides the standard equirectangular projection, they also used cylindrical and cube map projection to fuse. A new projection method called spherical padding was proposed in [39, 40], where the padding is fused with equirectangular projection for learning tasks. Geometric embeddings that directly use spherical coordinates as descriptors can also be fused with planar projections. In [41], the feature extracted from the two different domains is successfully combined for the depth estimation task.

3.2 Semantic Understanding

Semantic Segmentation. The usability of existing deep neural networks for semantic 2D image segmentation is limited by the distortion of the spatial relationships in the spherical image representation. We have reviewed different representations of 360° images to facilitate the use of the newest deep learning models in Section 2. Based on the delicately designed icosahedron-based representation for image segmentation networks, Zhang *et al.*[33] introduced how to perform fast interpolation for orientation-aware filter convolutions on the sphere and presented a weight transfer scheme from classical convolutional layers. Holistic scene modeling is a specialized problem in 360° image understanding, since the image captures the complete FoV in one shot to provide a wide range of context. Sun *et al.*[42] developed a deep learning based method, HoHoNet, with a horizon-to-dense module for recovering 2D per-pixel modalities, which effectively encodes the spatial features of spherical pixels and leads to a good performance in the holistic scene segmentation task. Based on the observation that most real-life 360° images have similar spatial layouts due to the common camera position and orientation when capturing 360° images (Fig.3), Yang *et al.*[43] considered the omnidirectional semantic segmentation from the context-aware perspective and proposed to leverage inherent long-range contextual priors when predicting the semantic information for all the pixels. The self-attention learning scheme has been successfully applied to omnidirectional segmentation. Zhang *et al.*[44] modified the original structure of the transformer encoder to adapt to the spatial relationship among the pixels in an equirectangular im-

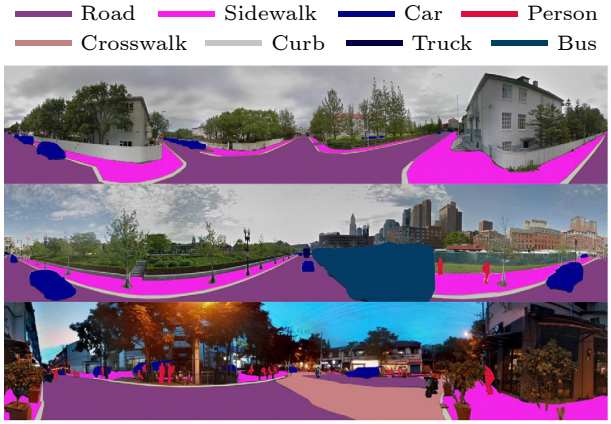


Fig.3. Panoramic segmentations results of [43] for scenes containing similar class labels.

age. Graph CNNs have the freedom of defining connectivities between elements, which were used in the work of Defferrard *et al.*[45] for climate event segmentation. To provide datasets for evaluating the omnidirectional image segmentation methods, Armeni *et al.*[27] and Ros *et al.*[46] rendered equirectangular images and semantic segmentation maps from 3D scenes. Yang *et al.*[47] provided a manually annotated semantic segmentation dataset, including 400 panoramas with annotations.

Object Detection. Object detection is a classical computer vision task. To transfer the 2D methods to the spherical domain, Coors *et al.*[28] and Su and Grauman[48] directly integrated the deformed kernels to existing 2D convolutional network architectures. Their method significantly increases the object detection accuracy. The criteria for measuring 2D object detection accuracy are either inaccurate or inefficient for the 360° object detection. Spherical criteria including both spherical bounding boxes and spherical IoU (SphIoU) were introduced in [49]. To solve the bias due to the sphere-to-plane projection while detecting objects using 2D-based convolutional networks, Cao *et al.*[50] proposed a data augmentation scheme that randomly rotates the spherical images before projection. They also introduced the FoV-IoU criterion that computes the intersection-over-union of two field-of-view bounding boxes in a spherical image for supervising the training of the network. Compared with outdoor data, indoor 360° images and videos are more common because of the prevalent room-scale VR/MR applications. Indoor 360° object detection datasets[51] are built to support the training of deep object detection models for indoor scenes[52].

Saliency. When watching 360° videos, although the content has a full FoV, users are only able to see a limited view range by rotating their head. Saliency detection can provide guidance to help the users concentrate on important events and content, and thus plays a key role in improving MR experiences. Cheng *et al.* built a convolutional network to predict saliency regions in 360° videos based on their special cube-padding representation[37]. In their model, they employed convolutional Long Short-Term Memory (LSTM) modules to process temporal cues. The work of [53] adopts Vision Transformer along with deformable convolution to encode the omnidirectional imagery to predict temporally continuous saliency results. This method alleviates geometric projection errors and outperforms other methods designed for 2D saliency by a large margin. Cubemap representation was directly used in convolutional networks in [38] for saliency value prediction. Ma *et al.* obtained a higher accuracy in saliency prediction using their two-stage deep learning framework[54]. They first coarsely predicted salient candidate regions via semantical saliency and then projected the regions to distortion-free image patches to conduct semantical saliency ranking to accurately locate salient objects. Zhang *et al.*[53] provided a dynamic saliency dataset by annotating the 360° videos from the Sports-360 dataset[55], and used a spherical convolutional kernel defined on a spherical crown in their deep network. Attention-based methods have shown promising results in the 360° saliency detection task. In Dahou *et al.*'s work[56], attention-based learning is performed on equirectangular images and then fused with the learned features from cubemap faces for the final saliency estimation. Considering the special characteristics of 360° video-based applications, Qiao *et al.*[57] produced a viewport-based saliency dataset (Fig.4) and trained a deep model to predict fixations in a given view window. Chao *et al.*[58] utilized a multi-FoV solution and adaptive losses to solve the salient map prediction problem, which is similar to the fixation prediction problem. The audio information is another important factor for users' attention when watching 360° videos. [59], one of the pioneer studies, could be a starting point for advancing saliency detection immersive media.

3.3 Depth Estimation

Monocular. When 2D depth estimation deep neu-

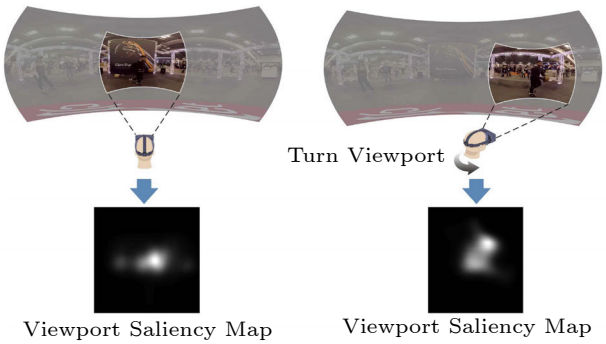


Fig.4. Viewport-based saliency detection[57] © IEEE. The saliency results consider not only the global scene information but also the content in the users' viewing window.

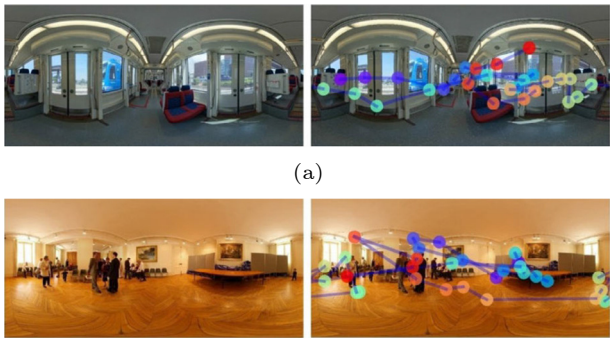
ral networks are applied on 360° images, their performance usually suffers from the distortion of equirectangular projection. Therefore, researchers proposed architectures that can eliminate the inaccuracy of depth estimation caused by spherical-to-2D projections[40]. Bifuse and Bifuse++[39, 40] are end-to-end two-branch networks, which incorporate both equirectangular and cubemap projections. In both of the two methods, to share the information, bi-projection fusion blocks with learnable masks to balance the information are used as the bridge across two projections when learning to predict depth values of 360° pixels. Differing from Bifuse, Unifuse[60] only performs fusion in an unidirectional manner where the distortion-free cubemap features are only fused with equirectangular features in its encoder, since the final result needs to be an equirectangular map. Zhuang *et al.*[61] proposed to use dilated convolution kernels to extend the receptive field and an adaptive channel-wise fusion module to obtain diverse attention areas along different channels. The geometric features can be a good complementary to the convolutional features. In the work of OmniFusion[41], geometric embeddings are learned to help the depth estimation. It also uses the powerful Transformer Encoder to globally aggregate patch-wise and geometric information. A photorealistic synthetic dataset, SynDepth360, was built to evaluate the 360° depth estimation method[62]. To address the challenge in high-resolution 360° depth estimation, Rey-Area *et al.*[63] used state-of-the-art perspective monocular depth estimators on icosahedron faces and then optimally aligned individual depth maps to generate the high-quality 360° depth map. To mitigate the discontinuities along object boundaries on the depth maps, Serrano *et al.*[64] proposed to make use of a layered representation to fix the issues of missing information and jagged silhouettes using the raw depths captured by RGBD camera. It successfully improved the quality of their 6-DoF application.

Stereoscopic. Stereoscopic omnidirectional images are difficult to capture and produce. Different from monocular depth estimation, stereoscopic depth estimation heavily relies on the matching information between the left and right views. Won *et al.*[65] built an omnidirectional wide-baseline stereo system that can estimate 360° dense depth maps. Their hardware configuration of a few cameras using ultrawide FOV lenses is flexible and effective in capturing a 360° 3D environment. Wang *et al.*[66] proposed to use a top-down layout of two cameras to capture stereo images and provided a deep neural network based on cost volumes to estimate depth values using the vertical disparity. However, such stereo images cannot be directly used in MR systems due to their different viewpoints.

3.4 Temporal Domain Analysis

Optical Flow. Despite the success of 2D optical flow estimation methods such as RAFT[67] and PWC-Net[68], generalizing these methods beyond narrow FOV videos remains challenging. Due to the lack of ground truth 360° optical flow data, Bhandari *et al.*[69] projected existing 2D optical flow to an equirectangular image to generate pseudo-ground truth data to train the neural networks. Yuan and Christian[31] represented the spherical image by a set of tangent images and their method can adopt any 2D optical flow estimation method for each tangent image. However, their final combined result suffers from the discontinuity along the image boundaries. Most recently, Li *et al.*[10] proposed a multi-projection fusion framework that learns to fuse the complementary motion information under the equirectangular, cube-padding and cylindrical projections. The first large-scale omnidirectional optical flow dataset is also provided in this work for the evaluation of panoramic optical flow estimation methods.

Gaze Prediction and Scanpath. Research on the temporal dynamics of eye gaze in omnidirectional images/videos is crucial to understand how people perceive and interact with this kind of immersive content. Gaze prediction can benefit the data compression for 360° video transmission[70, 71] and improve the watching experience by tailoring the interactions for specific users[72]. Some scanpaths generated by the above methods are shown in Fig.5. The gaze data col-



(a)

(b)

Fig.5. Scanpaths generated for two given 360° images by the method introduced in [70]. (a) and (b) are two examples.© IEEE.

lection in 360° videos needs a large amount of time and effort. Jin *et al.*[25] built a dataset containing synchronized head and gaze behaviors while watching 360° videos in the VR headset, which has a strong diversity covering different types of content and behaviors and outperforms the earlier smaller dataset proposed in [26]. In a pioneering work in this field[73], Xu *et al.* tackled the gaze prediction task in dynamic 360° videos by learning from video frames and their saliency maps at different scales by CNNs. Then the LSTM features are combined together to predict the gaze displacement from one moment to the next moment. The head movements can be an approximation of where the users are watching in VR-headset based applications. Yang *et al.*[74] proposed an approach, Hi-Bayes-LSTM, integrating hierarchical Bayesian inference into the LSTM network to generate a head motion trajectory by learning from their large-scale dataset. Rondon *et al.*[75] considered both past positions and video content and used Structural-RNNs to model the related information as a spatio-temporal graph. Successive eye movements of users when they are watching 360° images are called visual scanpaths, which are a temporal-aware description of saliency. In some early attempts, saliency volumes[76] were proposed to capture the temporal nature of eye-gaze scanpaths in a single image, on which a sampling strategy can be applied to generate a predicted scanpath. de Belen *et al.*[77] leveraged convolutional LSTMs to model the temporal dependencies of gaze positions. By sequentially sampling their output, a reasonable scanpath can be produced. ScanpathNet shows promising performance in several eye-tracking benchmark datasets. To address the challenge of acquiring a large number of scanpaths, a GAN-based model was introduced by Martin *et al.*[70] for mimicking virtual observers to reproduce human watching behaviors. To enable the instant feedback on the content editing for 6-DoF videos, Griffin *et al.*[78] built a pipeline to support immersive editing in a VR headset. Specifically, they used 360° RGBD videos as the data to be edited in a 6-DoF way in their implementation and developed interaction techniques for this paradigm.

The assessment of 360° video quality based on both spatial and temporal features is an important research topic. We recommend readers to refer to [79] for a complete survey on the 360° video perception and assessment.

4 Applications of 360° Content for MR

We have been witnessing the beginning of a new paradigm for highly-realistic mixed reality experiences. As one of the most important resources of MR content, a key to enabling a higher degree of freedom of interaction is the image/video generation algorithms given user inputs. This section introduces the recent progress in reconstructing a more vivid MR environment that allows richer interactions.

4.1 6-DoF Panoramic Video

Recent research on 6-degrees-of-freedom (6-DoF) media focuses on generating images for novel viewpoints and view directions from a given image/video. Using the 360° camera for capturing the source media can greatly simplify the process, due to its complete view of the environment. Although the newest NeRF-based method[80] has been able to reconstruct 360° unbounded scenes, the current neural radiance fields are built upon perspective images and are not able to generate large-scope locomotion for MR applications yet. Grid-based warping[81] is a conventional way to produce images for novel viewpoints from a set of 360° images. With a delicately designed structure-from-motion method working on 360° images[82] proposed by Baker *et al.*, the geometric information can be better reconstructed to support the 6-DoF generation. To satisfy the framerate required in MR applications while synthesizing novel panoramic frames, Chen *et al.*[83] proposed to directly synthesize 360° RGB images using the recovered depth from the input 360° videos (see Fig.6). A ray-marching based depth interpolation and refinement scheme using selected existing views ensures their high visual quality

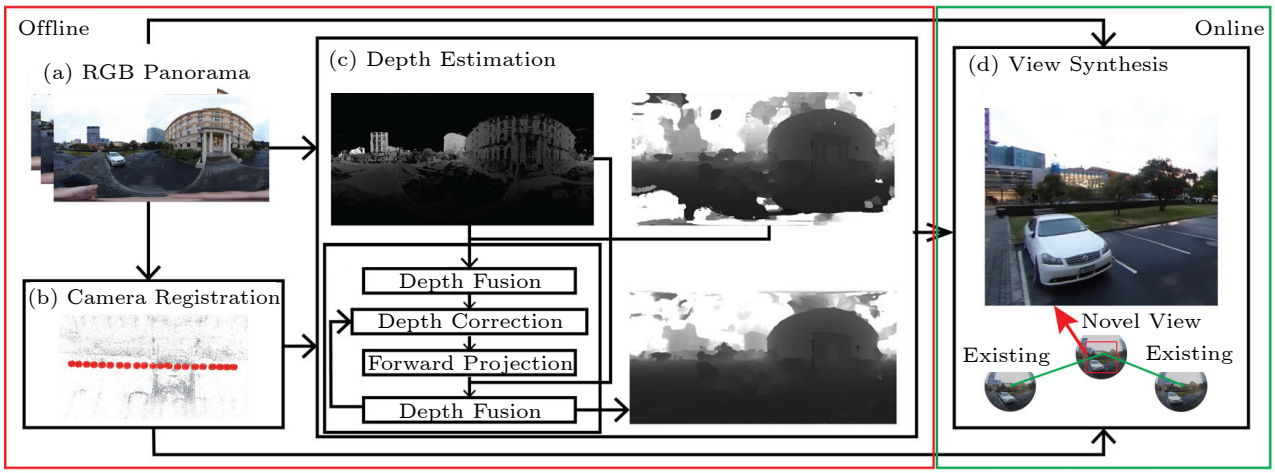


Fig.6. Workflow of the method proposed in [83]. Its offline stage processes the input panoramic images and estimates the depth maps. The online stage uses an efficient depth generation method to guide the view synthesis.

and fast view synthesis. View prediction from single panoramic images is more challenging. Waidhofer *et al.*[84] proposed a lightweight method to generate a 6-DoF experience from only one single panoramic image, which can be captured with consumer-grade 360 cameras. They introduced a multi-cylinder image representation that can be produced using their CNN-based model and synthesized novel views by blending the cylindrical layers for the novel viewpoint. Their method has been extensively evaluated in a user study. They developed and presented a concept for editing 6-DoF immersive videos in VR.

4.2 360° Environment Map Estimation

A 360° environment map is often used to store the high dynamic range (HDR) environment lighting information for illuminating virtual objects into a photograph for MR. The high quality of an environment map is crucial to realize the seamless blending of the virtual object with the real-world background. An earlier attempt was presented in the work of DiVerdi *et al.*[85]. They combined landmarks and frame-to-frame components in the vision-based orientation tracking to map the video frames to a cubemap for environment map estimation. Recent methods use deep learning to estimate environment maps from photographs, which is a lightweight replacement of traditional environment map capturing methods, e.g., using a 360° omnidirectional camera or light probe to obtain high-fidelity lighting information. To address this challenge, physically-inspired deep learning approaches have been explored[86–88]. They often input a foreground object with unknown geometry or material as a reference, and then intrinsically decompose into properties defining its appearance (geometry, material, and lighting). Environment maps[86, 88], or both reflectance and environment maps[87] are reconstructed from sequential deep neural network models that are designed based on these intrinsic decomposition schemes.

As an alternative, in the absence of an exemplar object in the captured photograph, the typical lighting cues such as shadow, highlight, and shading relationship are taken into consideration instead to infer the environment lighting. A significant amount of research[89–94] has been conducted in this direction due to its unconstrained and flexible setting without any specific reference object required, making it a rapidly developing research topic.

Various deep-learning solutions were proposed to address different critical issues arising from the environment map reconstruction from limited field-of-view (FoV) photographs. Some researchers studied on either outdoor[89–94] and/or indoor[95–103] scenes, beginning with the seminal work by Hold-Geoffroy *et al.*[89] and Gardner *et al.*[95] to estimate outdoor and indoor lighting, respectively. Fig.7 shows some results of the work of [96]. Besides, some studies customized the deep learning methods specifically for mobile mixed reality[99, 104].

Outdoor environment map estimation focuses on recovering outdoor daytime lighting, which is mainly influenced by sun position, atmospheric conditions, and weather. The early studies[89, 90] often presupposed an analytical sky model and trained CNN models to regress its parameters from the FoV image. Hold-Geoffroy *et al.*[91], for the first time, proposed a

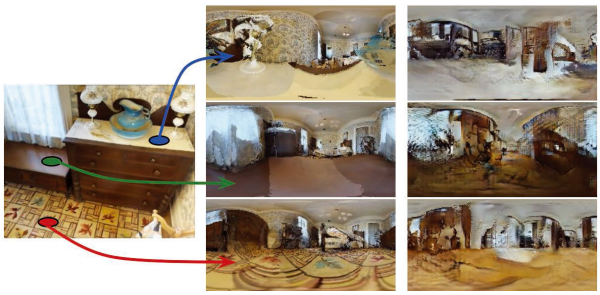


Fig.7. Spatially varying environment map reconstruction in [96]. (a) Input image. (b) Inferred spatially varying environment maps conditioned on a selected position. (c) Another group of results.©IEEE

data-driven deep sky model trained end-to-end across multiple datasets and was successfully used to reconstruct a plausible HDR outdoor environment map. The model was improved lately by Yu *et al.*[92] with HDSky, which conducts a hierarchical disentanglement of sun and sky representation learning, to generate more realistic and diverse outdoor environment map for all-weather conditions and enables light editing (see Fig.8). The spatial-varying (local) effects are also considered in this area and investigated by [93, 94] to realize location-dependent outdoor environment map estimation by either integrating geometric information estimated from intrinsics[93] or dedicatedly disentangling global and local lighting representations[94] for the associated predictions.

Within the space of indoor environment map estimation, the quality of the reconstruction map, including accuracy of lighting directions and intensity, ambient tones, and realistic high-resolution texture, has been improved progressively with the advancement of neural network design and training schemes. Chalmers *et al.*[98] proposed a novel stacked CNN that is trained with a progressive training scheme from high to low roughness, allowing the generation of reflectance maps with varying materials roughness and improving high-frequency texture recovery. The same framework of [98] is also applicable to outdoor scenarios. Zhao *et al.*[99] proposed to use the dynamic filtering technology in the deep network architecture, to adaptively learn sample-specific lighting features and improve the generalization to wide variations caused by either indoor environments or sensor/lens characteristics of the (mobile) capturing device (see Fig.9). Zhan *et al.*[100] and Xu *et al.*[102] proposed to combine lighting parameter (e.g., spherical gaussian+/harmonics) regression and environment map generation into a uniform framework, leveraging lighting parameters into guiding environment map generation to produce a high-realistic reconstruction. Somanath and Kurz[104] customized the deep learning methods specifically for mobile MR using a light-weight end-to-end EnvMapNet network with an efficient clustering-based adversarial training loss and mask-based projection loss to produce the HDR environment map in real time. To

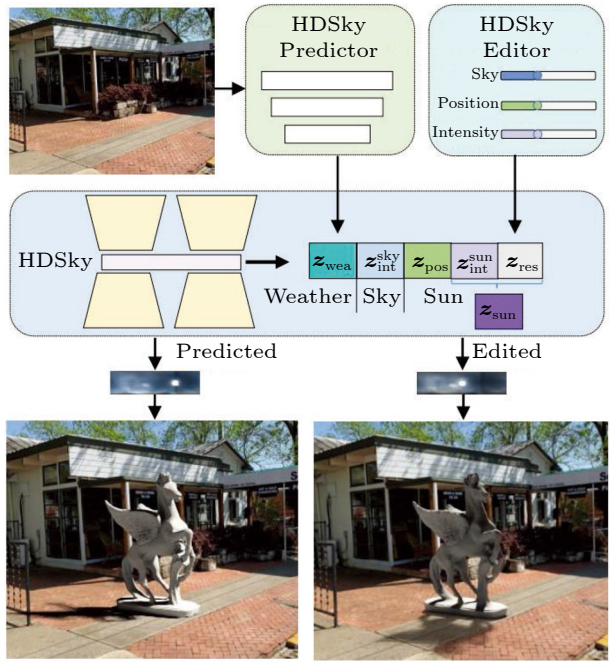


Fig.8. Flowchart of HDSky proposed by Yu *et al.*[92]. The learned latent space is disentangled into several independent factors to help improve environment map reconstruction for all-weather conditions and allow light editing. $z_{wea}, z_{int}^{sky}, z_{int}^{sun}, z_{res}$ are learned latent code vectors.

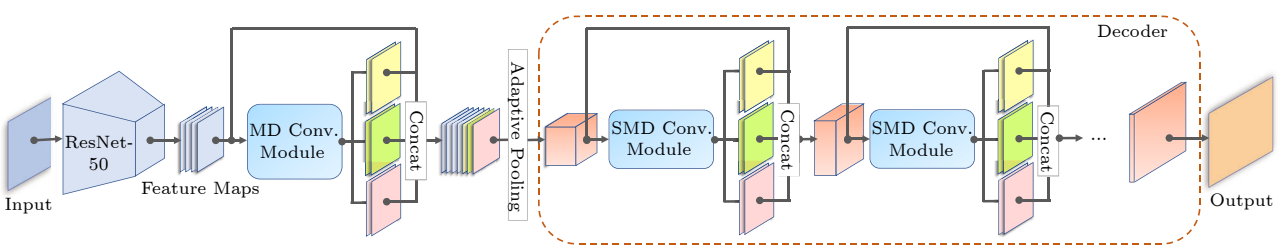


Fig.9. Overall structure of DLNet proposed by Zhao *et al.*[99]. It combines (spherical) multi-scale dynamic convolution modules into the backbone encoder-decoder structure to facilitate sample-specific lighting feature extraction and environment map reconstruction.

obtain high-resolution and diverse environment maps for both indoor and outdoor scenes, Akimoto *et al.*[103] proposed a multi-module ensemble deep learning framework with the CompletionNets module to perform diverse completions and the AdjustmentNet module to match the input image's color and realize outpainting at arbitrary image sizes.

Moreover, the spatial variation of indoor environments with fixed positions has been given particular attention (as opposed to outdoor scenes, in which the sun is assumed to be infinitely far away). The problem was initially addressed by spatially warping the photograph[95], but extended by considering depth estimates in the training process[96, 101]. A stereo pair of photographs with parallax information was also employed to address this issue[97].

4.3 AR Content Mapping

Since the panoramic images carry omnidirectional scene information, they can be leveraged to help mobile AR applications by mapping the content captured by a local view to the representations of the entire real-world scene. Researchers proposed panoramic tracking and mapping systems to model the environment while determining the pose of the camera. Envisor[85] is an early real-time system that aims to map 2D video content to a cube map for environment map generation. Wagner *et al.*[105] described a system that is able to robustly run in real time on a mobile device. They first registered cameras based on image features and then extended the map with new features from new viewing directions. Compared with SLAM, their approach is able to create a dense panoramic map of features, which are mapped during their first observation and do not need to refine again. However, it only works for 3-dimensional rotations. Gauglitz *et al.*[106, 107] presented real-time tracking and mapping approaches that support any type of camera motions in 3D environments, including parallax-inducing and degenerate rotation-only motions, and effectively generalize both a panorama mapping and tracking system and a keyframe-based SLAM system, seamlessly switching between the rotation-based panorama mapping system and the keyframe-based SLAM system depending on the camera movement. Pirchheim *et al.*[108] proposed a hybrid keyframe-based system that can track and localize in a combination of fully triangulated content as well as 360° keyframes with only rotational constraints. More recently, Baker *et al.*[109] proposed a localization and tracking solution that combines spherical Structure-from-Motion (SfM) and 2D tracking, which is suitable for real-time AR applications. Particularly, they introduced a method for computing the absolute pose with a spherical constraint for 360° scene representation.

By mapping the real-world content captured by mobile devices with a 360° image of the same scene, better AR services are enabled on consumer-grade devices. Pan *et al.* proposed a cheap online space carving approach based on Delaunay triangulation to obtain a polygonal textured representation from a set of panoramic images[110]. In their system, a robust feature correspondence estimation component for aligning individual 360° images based on bundle adjustment provides essential support for the subsequent triangulation and reconstruction. Arth *et al.*[111] proposed a system for self-localization on mobile phones using a GPS prior by matching captured content to an online-generated 360° image. It delivers high-quality self-tracking across a wide area (such as a whole city) with six degrees of freedom for an outdoor user. To provide panoramic images more efficiently, Reinisch *et al.*[112] solved the issues of real-time panorama generation for mobile devices. The pixel-mapping process is transferred from CPU to GPU by their shader-based mapping approach. Their application is implemented for Android phones.

The content placement was also applied in telepresence applications, such as teleconference and remote education. PanoInserts[113] is a system that uses smartphone cameras to create a surround representation of all the users' meeting places. Pece *et al.*[113] took a static 360° image of a location into which they combined all the users' live videos from smartphones. In their implementation, a combination of marker- and image-based tracking methods is employed for inserting videos at proper places and they transmitted this representation to a remote viewer. For online AR education, 360Anywhere[114] was developed by Speicher *et al.*, which is a framework for 360 video based multi-user collaboration. It not only allows collaborators to view and annotate a 360 live stream but also supports the projection of annotations in the 360 stream back into the real-world environment in real time. Piumsomboon *et al.*[115] presented a system where VR users can collaborate with local AR users by being immersed in a 360-video through a tangible interface, a combined 360 camera with a 6-DOF tracker. Similar systems are also proposed in [116].

For more remote collaborations using 360 videos, we suggest that readers refer to the survey of Wang *et al.*[117]. Nebeling and Madier[118] also explored how to conduct AR/VR application prototyping by mapping the hand-drawn paper-based content to the virtual space. They proposed a system to rapidly create AR/VR prototypes from the content drawn on equirectangular paper maps and bring them to life on AR/VR devices.

4.4 360° Content Manipulation

With the rise of MR applications, it has been increasingly important to manipulate 360° media in an intuitive and efficient way. Common 360° image/video manipulations refer to the editing of colors, content, motion, direction, etc., which aim to improve the visual quality and meet individual requirements. Recent years have witnessed great success in content manipulation on 2D images and videos, but these techniques cannot be directly applied to 360° media, due to their differences in geometry structure, distance metrics, etc. As a pioneer work in content editing, an inpainting method was proposed by Zhu *et al.*[119] to complete holes in 360° images for applications such as Google Street View. To deal with the distortions, they performed structure-rectifying warping and completed the holes using 2D completion methods[120]. Although effective in many examples, it is specially designed to complete holes in the bottom regions. To remove occlusions in 360° videos, Xu *et al.*[121] proposed a coarse-to-fine optimization to iteratively complete missing pixels and motion information while considering the geometric properties of spherical images. Zhao *et al.*[122] addressed the 360° panorama cloning problem using a coordinate-based method in the spherical domain. Huang *et al.*[123] further explored the composition of 360° stereo images, where they ensured the fundamental geometry of the inserted objects by using the estimated depth information to guide the content manipulation in 3D space (see Fig.10).

Video correction and stabilization have been exploited to improve the sense of immersion and visual comfort. Jung *et al.*[124] proposed an automatic method for upright adjustment of 360° panoramas (see Fig.11). With the Atlanta world assumption and line constraints, the updated north pole direction is estimated by iterative optimization, and the adjusted image can be obtained by resampling the image (Fig.11). Kopf[7] proposed a hybrid 3D-2D method and a new deformed-rotation motion model to remove the shakeness in 360° videos. To remove undesired camera motion while obtaining a smooth and redirected camera path, Tang *et al.*[8] further proposed a joint optimization for stabilization and redirection. Compared with [7], [8] can handle rotation, translation, and strong parallax well, and can produce smoother feature trajectories due to its 3D spherical warping model.

To efficiently edit the visual appearance of 360° panoramas, Zhang *et al.*[125] proposed the first stroke-based edit propagation on 360° panoramas. To produce seam-free and visually pleasing results, they constructed the manifold structure with a spherical distance metric for each pixel. Zhang *et al.*[126] further accelerated the edit propagation using the function interpolation with an adaptive sampling strategy in the spherical domain (see Fig.12). To enhance the details of 360° panoramas, Wong[127] proposed a view-adaptive asymmetric detail enhancement solution, which improves the panoramic viewing experience with a less computational cost. For better 360° video viewing experiences, friendly user interactions are very im-

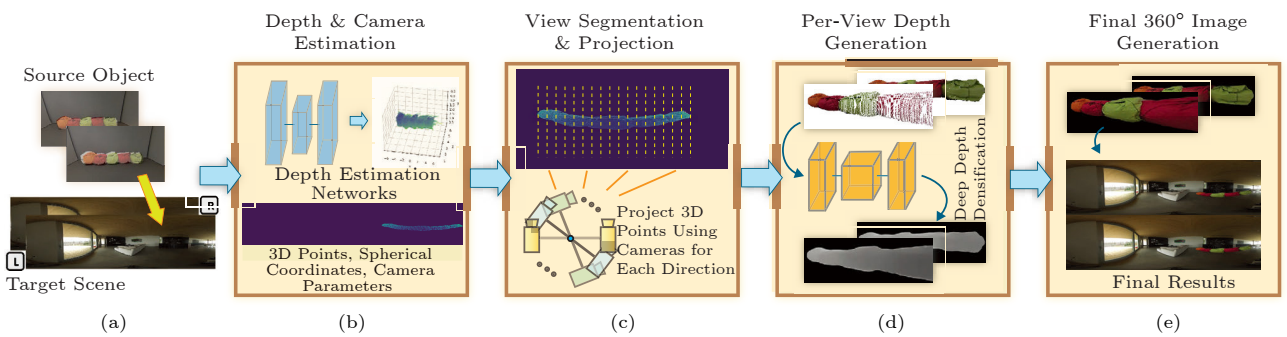


Fig.10. Pipeline of the method in [123]. (a) A stereo object. (b)–(d) Our approach manipulates the image content with guidance from 3D space to avoid distance metric issues when compositing the object into the target omnidirectional stereo background image. (e) Result.

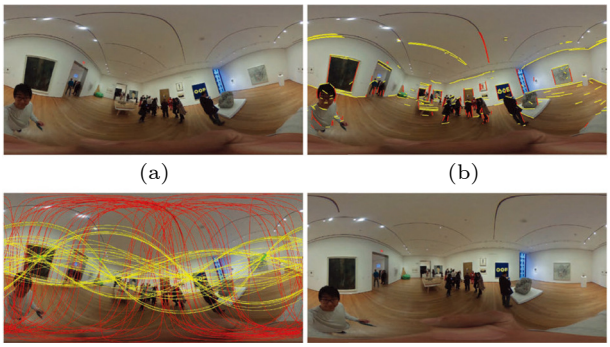


Fig.11. Overall process[124] © Springer. (a) Input image. (b) Detected vertical and horizontal lines. (c) Great circles from the classified lines. (d) Adjustment result.

portant. Kang and Cho[14] proposed interactive and automatic navigation for 360° video playback, which enables users to watch interesting events without intensive adjustment of viewing directions. However, their method only calculates a single camera path, which may ignore other regions of interest. To further improve the navigation and playback experience, Wang *et al.*[128] designed a novel diverse constraint, which helps to compute diverse content-aware normal field-of-view (NFoV) virtual camera paths using a coarse-to-fine dynamic programming optimization (See Fig.13). In their system, the user watches an NFoV video extracted from a 360° video and changes their view during playback. All NFoV camera paths are precomputed by the proposed content-aware and diverse virtual camera path optimization. The user can continue to watch using the new camera path. Li *et al.*[129] introduced the recently popular bullet comment function into 360° videos. They designed several bullet comment display methods and controller-based methods for bullet comment insertion, which help to add more interactivity and sociability to the 360° viewing experience.

5 Conclusions and Discussions

There have been great advances in 360° media-based mixed reality technologies in the last decade. This paper reviews the recent methods in 360° image representation, understanding, and reconstruction, and their applications in immersive visual technology. This paper is expected to contribute to the development of the algorithms in the field of 360° image processing and applications and deepen the understanding of mixed reality techniques based on real-world content.

360° media based mixed reality technology is an emerging research field. The challenges of realizing highly realistic immersive experiences using 360° images have not been fully addressed. The remaining open problems for future research in this field include the followings:

- *Holistic Scene Understanding and 3D Reconstruction.* The current image understanding and 3D reconstruction methods designed for 2D videos cannot be trivially applied to 360° videos due to the omnidirectional representation. How to reconstruct the full-FoV 3D information accurately and extract globally-consistent visual features needs to be addressed.
- *Friendly User-Content Interactions.* Although the existing content manipulation method is able to provide basic content operations on 2D images/videos

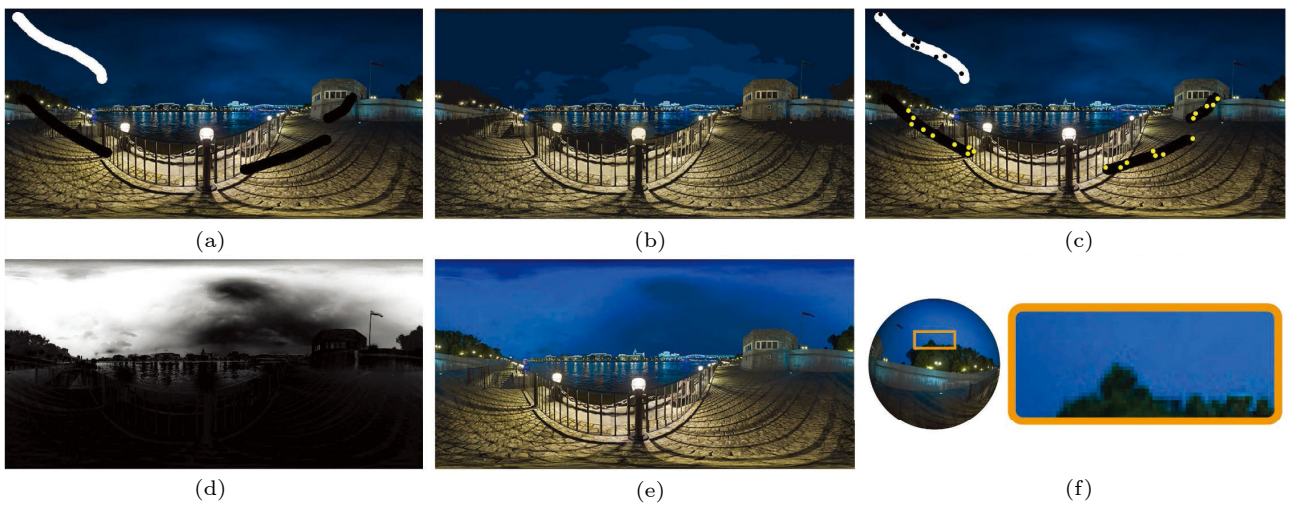


Fig.12. Flow chart of the method in [126]. (a) Input image. (b) (c) The input image is first quantized and the samples are adaptively collected. (d) RBF interpolation. (e) Edit for each pixel. (f) Final result.

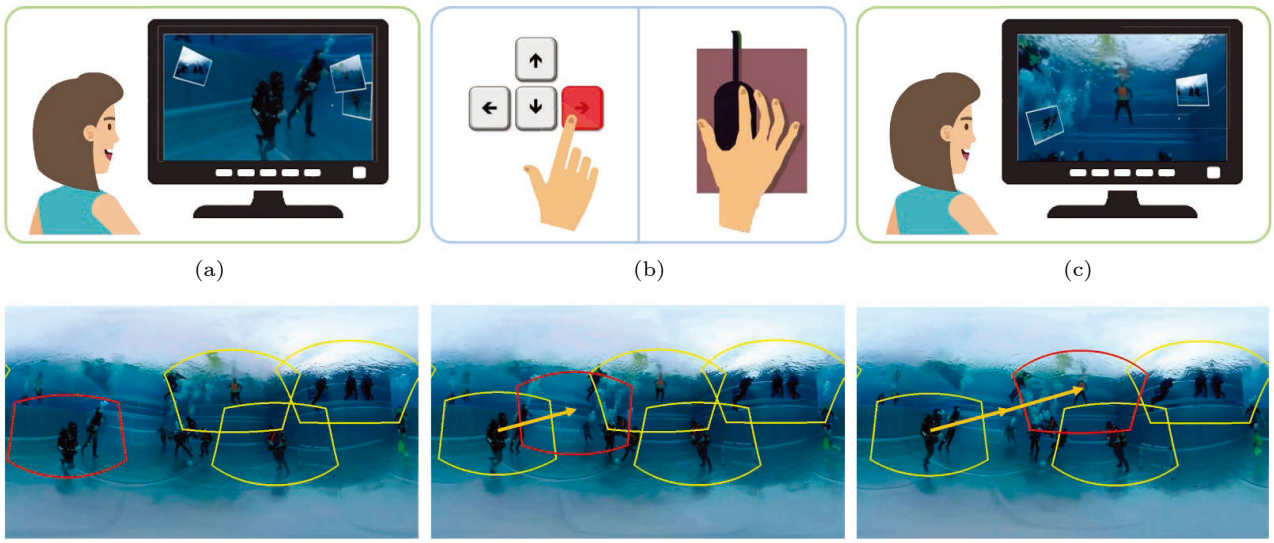


Fig.13. Overview of the system in [128]. (a)(b)(c) The user watches an NFoV video and changes their view. The corresponding view boundary is marked in red and other candidate views are marked in yellow in (d). (e)(f) The orange arrows indicate the trajectory of the transitioning view.

such as inpainting and reshuffling, it is still challenging to generate photo-realistic results efficiently especially for dynamics objects.

- 360° *Video Capture and Rendering*. The captured omnidirectional content is the basis of a high-quality immersive experience. There have been a variety of 360° camera prototypes and products. However, the available solutions still come with artifacts such as distortions and visible seams along stitching boundaries between the frames captured by neighboring cameras. A better capture and rendering model is needed in this field.

References

[1] Friston S, Ritschel T, Steed A. Perceptual rasterization for head-mounted display image synthesis. *ACM Trans. Graphics*, 2019, 38(4): Article No. 97. DOI: 10.1145/3306346.3323033.

[2] Tursun O T, Arabadzhiyska-Koleva E, Wernikowski M, Mantiuk R, Seidel H P, Myszkowski K, Didyk P. Luminance-contrast-aware foveated rendering. *ACM Trans. Graphics*, 2019, 38(4): Article No. 98. DOI: 10.1145/3306346.3322985.

[3] Schroers C, Bazin J C, Sorkine-Hornung A. An omnistereoscopic video pipeline for capture and display of real-world VR. *ACM Trans. Graphics*, 2018, 37(3): Article No. 37. DOI: 10.1145/3225150.

[4] Matzen K, Cohen M F, Evans B, Kopf J, Szeliski R. Low-cost 360 stereo photography and video capture. *ACM Trans. Graphics*, 2017, 36(4): Article No. 148. DOI: 10.1145/3072959.3073645.

[5] Habermann M, Xu W P, Zollhöfer M, Pons-Moll G, Theobalt C. LiveCap: Real-time human performance capture from monocular video. *ACM Trans. Graphics*, 2019, 38(2): Article No. 14. DOI: 10.1145/3311970.

[6] Xu W P, Chatterjee A, Zollhöfer M, Rhodin H, Mehta D, Seidel H P, Theobalt C. MonoPerfCap: Human performance capture from monocular video. *ACM Trans. Graphics*, 2018, 37(2): Article No. 27. DOI: 10.1145/3181973.

[7] Kopf J. 360° video stabilization. *ACM Trans. Graphics*, 2016, 35(6): Article No. 195. DOI: 10.1145/2980179.2982405.

[8] Tang C Z, Wang O, Liu F, Tan P. Joint stabilization and direction of 360° videos. *ACM Trans. Graphics*, 2019, 38(2): Article No. 18. DOI: 10.1145/3211889.

[9] Silva R M A, Feijó B, Gomes P B, Frensh T, Monteiro D. Real time 360° video stitching and streaming. In *Proc. the ACM SIGGRAPH 2016 Posters*, Jul. 2016, Article No. 70. DOI: 10.1145/2945078.2945148.

[10] Li Y H, Barnes C, Huang K, Zhang F L. Deep 360° optical flow estimation based on multi-projection fusion. In *Proc. the 17th European Conference on Computer Vision*, Oct. 2022, pp.336–352. DOI: 10.1007/978-3-031-19833-5_20.

[11] Jung R, Lee A S J, Ashtari A, Bazin J C. Deep360Up: A deep learning-based approach for automatic VR image upright adjustment. In *Proc. the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, Mar. 2019. DOI: 10.1109/VR.2019.8798326.

[12] Li D Z Y, Langlois T R, Zheng C X. Scene-aware audio for 360° videos. *ACM Trans. Graphics*, 2018, 37(4): Article No. 111. DOI: 10.1145/3197517.3201391.

[13] Rhee T, Petikam L, Allen B, Chalmers A. MR360: Mixed reality rendering for 360° panoramic videos. *IEEE Trans. Visualization and Computer Graphics*, 2017, 23(4): 1379–1388. DOI: 10.1109/TVCG.2017.2657178.

[14] Kang K, Cho S. Interactive and automatic navigation for 360° video playback. *ACM Trans. Graphics*, 2019, 38(4): Article No. 108. DOI: 10.1145/3306346.3323046.

[15] Rees D W. Panoramic television viewing system. United States Patent, No. 3505465, 1970.

[16] Yagi Y, Kawato S, Tsuji S. Real-time omnidirectional image sensor (COPIS) for vision-guided navigation. *IEEE Trans. Robotics and Automation*, 1994, 10(1): 11–22. DOI: 10.1109/70.285581.

[17] Gledhill D, Tian G Y, Taylor D, Clarke D. Panoramic imaging—A review. *Computers & Graphics*, 2003, 27(3): 435–445. DOI: 10.1016/S0097-8493(03)00038-4.

[18] Yagi Y, Yachida M. Real-time omnidirectional image sensors. *International Journal of Computer Vision*, 2004, 58(3): 173–207. DOI: 10.1023/B:VISI.0000019684.35147.fc.

[19] Debevec P. Image-based lighting. In *Proc. the ACM SIGGRAPH 2005 Courses*, Jul. 2005, Article No. 3-es. DOI: 10.1145/1198555.1198709.

[20] Tarini M, Hormann K, Cignoni P, Montani C. PolyCube-maps. *ACM Trans. Graphics*, 2004, 23(3): 853–860. DOI: 10.1145/1015706.1015810.

[21] McMillan L, Bishop G. Plenoptic modeling: An image-based rendering system. In *Proc. the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, Aug. 1995, pp.39–46. DOI: 10.1145/218380.218398.

[22] Hilbert D, Cohn-Vossen S. Geometry and the Imagination, Volume 87. Providence: American Mathematical Soc. , 2021.

[23] Nadeem S, Su Z Y, Zeng W, Kaufman A, Gu X F. Spherical parameterization balancing angle and area distortions. *IEEE Trans. Visualization and Computer Graphics*, 2017, 23(6): 1663–1676. DOI: 10.1109/TVCG.2016.2542073.

[24] Poranne R, Tarini M, Huber S, Panozzo D, Sorkine-Hornung O. Autocuts: Simultaneous distortion and cut optimization for UV mapping. *ACM Trans. Graphics*, 2017, 36(6): Article No. 215. DOI: 10.1145/3130800.3130845.

[25] Jin Y L, Liu J H, Wang F X, Cui S G. Where are you looking? A large-scale dataset of head and gaze behavior for 360-degree videos and a pilot study. In *Proc. the 30th ACM International Conference on Multimedia*, Oct. 2022, pp.1025–1034. DOI: 10.1145/3503161.3548200.

[26] David E J, Gutiérrez J, Coutrot A, Da Silva M P, Le Callet P. A dataset of head and eye movements for 360° videos. In *Proc. the 9th ACM Multimedia Systems Conference*, Jun. 2018, pp.432–437. DOI: 10.1145/3204949.3208139.

[27] Armeni I, Sax S, Zamir A R, Savarese S. Joint 2D-3D-semantic data for indoor scene understanding. arXiv: 1702.01105, 2017. https://arxiv.org/abs/1702.01105, Jul. 2023.

[28] Coors B, Condurache A P, Geiger A. SphereNet: Learning spherical representations for detection and classification in omnidirectional images. In *Proc. the 15th European Conference on Computer Vision*, Sept. 2018, pp.525–541. DOI: 10.1007/978-3-030-01240-3_32.

[29] Zhao Q, Zhu C, Dai F, Ma Y K, Jin G Q, Zhang Y D. Distortion-aware CNNs for spherical images. In *Proc. the 27th International Joint Conference on Artificial Intelligence*, Jul. 2018, pp.1198–1204.

[30] Eder M, Shvets M, Lim J, Frahm J M. Tangent images for mitigating spherical distortion. In *Proc. the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp.12423–12431. DOI: 10.1109/CVPR42600.2020.01244.

[31] Yuan M Z, Christian R. 360° Optical flow using tangent images. In *Proc. the 32nd British Machine Vision Conference*, Nov. 2021.

[32] Lee Y, Jeong J, Yun J, Cho W, Yoon K J. SpherePHD: Applying CNNs on a spherical PolyHeDron representation of 360° images. In *Proc. the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2019, pp.9173–9181. DOI: 10.1109/CVPR.2019.00940.

[33] Zhang C, Liwicki S, Smith W, Cipolla R. Orientation-aware semantic segmentation on icosahedron spheres. In *Proc. the 2019 IEEE/CVF International Conference on Computer Vision*, Oct. 27–Nov. 2, 2019, pp.3532–3540. DOI: 10.1109/ICCV.2019.00363.

[34] Yoon Y, Chung I, Wang L, Yoon K J. SphereSR: 360° image super-resolution with arbitrary projection via continuous spherical image representation. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp.5667–5676. DOI: 10.1109/CVPR52688.2022.00559.

[35] Wu G, Shi Y H, Sun X Y, Wang J, Yin B C. SMSIR: Spherical measure based spherical image representation. *IEEE Trans. Image Processing*, 2021, 30: 6377–6391. DOI: 10.1109/TIP.2021.3079797.

[36] Li J S, Wen Z Y, Li S H, Zhao Y K, Guo B C, Wen J T. Novel tile segmentation scheme for omnidirectional video. In *Proc. the 2016 IEEE International Conference on Image Processing (ICIP)*, Sept. 2016, pp.370–374. DOI: 10.1109/ICIP.2016.7532381.

[37] Cheng H T, Chao C H, Dong J D, Wen H K, Liu T L, Sun M. Cube padding for weakly-supervised saliency prediction in 360° videos. In *Proc. the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp.1420–1429. DOI: 10.1109/CVPR.2018.00154.

[38] Monroy R, Lutz S, Chalasani T, Smolic A. SalNet360: Saliency maps for omni-directional images with CNN. *Signal Processing: Image Communication*, 2018, 69: 26–34. DOI: 10.1016/j.image.2018.05.005.

[39] Wang F E, Yeh Y H, Tsai Y H, Chiu W C, Sun M. BiFuse++: Self-supervised and efficient bi-projection fusion for 360° depth estimation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2023, 45(5): 5448–5460. DOI: 10.1109/TPAMI.2022.3203516.

[40] Wang F E, Yeh Y H, Sun M, Chiu W C, Tsai Y H. BiFuse: Monocular 360 depth estimation via bi-projection fusion. In *Proc. the 2020 IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, Jun. 2020, pp.459–468. DOI: 10.1109/CVPR42600.2020.00054.

[41] Li Y Y, Guo Y L, Yan Z X, Huang X Y, Duan Y, Ren L. OmniFusion: 360 monocular depth estimation via geometry-aware fusion. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp.2791–2800. DOI: 10.1109/CVPR52688.2022.00282.

[42] Sun C, Sun M, Chen H T. HoHoNet: 360 indoor holistic understanding with latent horizontal features. In *Proc. the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2021, pp.2573–2582. DOI: 10.1109/CVPR46437.2021.00260.

[43] Yang K L, Hu X X, Fang Y C, Wang K W, Stiefelhagen R. Omnisupervised omnidirectional semantic segmentation. *IEEE Trans. Intelligent Transportation Systems*, 2022, 23(2): 1184–1199. DOI: 10.1109/TITS.2020.3023331.

[44] Zhang J M, Yang K L, Ma C X, Reiß S, Peng K Y, Stiefelhagen R. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp.16896–16906. DOI: 10.1109/CVPR52688.2022.01641.

[45] Defferrard M, Milani M, Gusset F, Perraudin N. DeepSphere: A graph-based spherical CNN. In *Proc. the 8th International Conference on Learning Representations*, Apr. 2019.

[46] Ros G, Sellart L, Materzynska J, Vazquez D, Lopez A M. The Synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2016, pp.3234–3243. DOI: 10.1109/CVPR.2016.352.

[47] Yang K L, Hu X X, Bergasa L M, Romera E, Wang K W. PASS: Panoramic annular semantic segmentation. *IEEE Trans. Intelligent Transportation Systems*, 2020, 21(10): 4171–4185. DOI: 10.1109/TITS.2019.2938965.

[48] Su Y C, Grauman K. Learning spherical convolution for fast features from 360° imagery. In *Proc. the 31st International Conference on Neural Information Processing Systems*, Dec. 2017, pp.529–539.

[49] Zhao P Y, You A S, Zhang Y X, Liu J Y, Bian K G, Tong Y H. Spherical criteria for fast and accurate 360° object detection. In *Proc. the 34th AAAI Conference on Artificial Intelligence*, Feb. 2020, pp.12959–12966. DOI: 10.1609/aaai.v34i07.6995.

[50] Cao M, Ikehata S, Aizawa K. Field-of-view IoU for object detection in 360° images. arXiv: 2202.03176, 2022. https://arxiv.org/abs/2202.03176, Jul. 2023.

[51] Chou S H, Sun C, Chang W Y, Hsu W T, Sun M, Fu J D. 360-indoor: Towards learning real-world objects in 360° indoor equirectangular images. In *Proc. the 2020 IEEE Winter Conference on Applications of Computer Vision*, Mar. 2020, pp.834–842. DOI: 10.1109/WACV45572.2020.9093262.

[52] Guerrero-Viu J, Fernandez-Labrador C, Demonceaux C, Guerrero J J. What's in my room? Object recognition on indoor panoramic images. In *Proc. the 2020 IEEE International Conference on Robotics and Automation (ICRA)*, May 2020, pp.567–573. DOI: 10.1109/ICRA40945.2020.9197335.

[53] Zhang Z H, Xu Y Y, Yu J Y, Gao S H. Saliency detection in 360° videos. In *Proc. the 15th European Conference on Computer Vision*, Sept. 2018, pp.504–520. DOI: 10.1007/978-3-030-01234-2_30.

[54] Ma G X, Li S, Chen C L Z, Hao A M, Qin H. Stage-wise salient object detection in 360° omnidirectional image via object-level semantical saliency ranking. *IEEE Trans. Visualization and Computer Graphics*, 2020, 26(12): 3535–3545. DOI: 10.1109/TVCG.2020.3023636.

[55] Hu H N, Lin Y C, Liu M Y, Cheng H T, Chang Y J, Sun M. Deep 360 pilot: Learning a deep agent for piloting through 360° sports videos. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp.1396–1405. DOI: 10.1109/CVPR.2017.153.

[56] Dahou Y, Tliba M, McGuinness K, O'Connor N. ATSal: An attention based architecture for saliency prediction in 360° videos. In *Proc. the 2021 Pattern Recognition. ICPR International Conference on Pattern Recognition*, Jan. 2021, pp.305–320. DOI: 10.1007/978-3-030-68796-0_22.

[57] Qiao M L, Xu M, Wang Z L, Borji A. Viewport-dependent saliency prediction in 360° video. *IEEE Trans. Multimedia*, 2021, 23: 748–760. DOI: 10.1109/TMM.2020.2987682.

[58] Chao F Y, Zhang L, Hamidouche W, Déforges O. A multi-FoV viewport-based visual saliency model using adaptive weighting losses for 360° images. *IEEE Trans. Multimedia*, 2021, 23: 1811–1826. DOI: 10.1109/TMM.2020.3003642.

[59] Zhang Y, Chao F Y, Hamidouche W, Deforges O. PAV-SOD: A new task towards panoramic audiovisual saliency detection. *ACM Trans. Multimedia Computing, Communications, and Applications*, 2023, 19(3): Article No. 101. DOI: 10.1145/3565267.

[60] Jiang H L, Sheng Z, Zhu S Y, Dong Z L, Huang R. UniFuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics and Automation Letters*, 2021, 6(2): 1519–1526. DOI: 10.1109/LRA.2021.3058957.

[61] Zhuang C Q, Lu Z D, Wang Y Q, Xiao J, Wang Y. ACDNet: Adaptively combined dilated convolution for monocular panorama depth estimation. In *Proc. the 36th AAAI Conference on Artificial Intelligence*, Feb. 22–Mar. 1, 2022, pp.3653–3661. DOI: 10.1609/aaai.v36i3.20278.

[62] Feng Q, Shum H P H, Morishima S. 360 depth estimation in the wild—the depth360 dataset and the SegFuse network. In *Proc. the 2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, Mar. 2022, pp.664–673. DOI: 10.1109/VR51125.2022.00087.

[63] Rey-Area M, Yuan M Z, Richardt C. 360MonoDepth: High-resolution 360° monocular depth estimation. In *Proc. the 2022 IEEE/CVF Conference on Computer Vi-*

sion and Pattern Recognition, Jun. 2022, pp.3752–3762. DOI: 10.1109/CVPR52688.2022.00374.

[64] Serrano A, Kim I, Chen Z L, DiVerdi S, Gutierrez D, Hertzmann A, Masia B. Motion parallax for 360° RGBD video. *IEEE Trans. Visualization and Computer Graphics*, 2019, 25(5): 1817–1827. DOI: 10.1109/TVCG.2019.2898757.

[65] Won C, Ryu J, Lim J. SweepNet: Wide-baseline omnidirectional depth estimation. In *Proc. the 2019 International Conference on Robotics and Automation (ICRA)*, May 2019, pp.6073–6079. DOI: 10.1109/ICRA.2019.8793823.

[66] Wang N H, Solarte B, Tsai Y H, Chiu W C, Sun M. 360SD-net: 360° stereo depth estimation with learnable cost volume. In *Proc. the 2020 IEEE International Conference on Robotics and Automation (ICRA)*, May 2020, pp.582–588. DOI: 10.1109/ICRA40945.2020.9196975.

[67] Teed Z, Deng J. RAFT: Recurrent all-pairs field transforms for optical flow. arXiv: 2003.12039, 2020. https://arxiv.org/abs/2003.12039, Jul. 2023.

[68] Sun D Q, Yang X D, Liu M Y, Kautz J. PWC-net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proc. the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp.8934–8943. DOI: 10.1109/CVPR.2018.00931.

[69] Bhandari K, Zong Z L, Yan Y. Revisiting optical flow estimation in 360 videos. In *Proc. the 25th International Conference on Pattern Recognition (ICPR)*, Jan. 2021, pp.8196–8203. DOI: 10.1109/ICPR48806.2021.9412035.

[70] Martin D, Serrano A, Bergman A W, Wetzstein G, Masia B. ScanGAN360: A generative model of realistic scanpaths for 360° images. *IEEE Trans. Visualization and Computer Graphics*, 2022, 28(5): 2003–2013. DOI: 10.1109/TVCG.2022.3150502.

[71] Yu M, Lakshman H, Girod B. A framework to evaluate omnidirectional video coding schemes. In *Proc. the 2015 IEEE International Symposium on Mixed and Augmented Reality*, Sept. 29–Oct. 3, 2015, pp.31–36. DOI: 10.1109/ISMAR.2015.12.

[72] Wang S B, Yang S S, Li H L, Zhang X D, Zhou C, Xu C R, Qian F, Wang N B, Xu Z B. SalientVR: Saliency-driven mobile 360-degree video streaming with gaze information. In *Proc. the 28th Annual International Conference on Mobile Computing and Networking*, Oct. 2022, pp.542–555. DOI: 10.1145/3495243.3517018.

[73] Xu Y Y, Dong Y B, Wu J R, Sun Z Z, Shi Z R, Yu J Y, Gao S H. Gaze prediction in dynamic 360° immersive videos. In *Proc. the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp.5333–5342. DOI: 10.1109/CVPR.2018.00559.

[74] Yang L, Xu M, Guo Y C, Deng X, Gao F Y, Guan Z Y. Hierarchical Bayesian LSTM for head trajectory prediction on omnidirectional images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2022, 44(11): 7563–7580. DOI: 10.1109/TPAMI.2021.3117019.

[75] Rondón M, Sassatelli L, Aparicio-Pardo R, Precioso F. TRACK: A new method from a re-examination of deep architectures for head motion prediction in 360° videos. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2022, 44(9): 5681–5699. DOI: 10.1109/TPAMI.2021.3070520.

[76] Assens M, Giro-i-Nieto X, McGuinness K, O'Connor N E. SaltiNet: Scan-path prediction on 360 degree images using saliency volumes. In *Proc. the 2017 IEEE International Conference on Computer Vision Workshops*, Oct. 2017, pp.2331–2338. DOI: 10.1109/ICCVW.2017.275.

[77] de Belen R A J, Bednarz T, Sowmya A. ScanpathNet: A recurrent mixture density network for scanpath prediction. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp.5006–5016. DOI: 10.1109/CVPRW56347.2022.00549.

[78] Griffin R, Langlotz T, Zollmann S. 6DIVE: 6 degrees-of-freedom immersive video editor. *Frontiers in Virtual Reality*, 2021, 2: 676895. DOI: 10.3389/frvir.2021.676895.

[79] Xu M, Li C, Zhang S Y, Le Callet P. State-of-the-art in 360° video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing*, 2020, 14(1): 5–26. DOI: 10.1109/JSTSP.2020.2966864.

[80] Barron J T, Mildenhall B, Verbin D, Srinivasan P P, Hedman P. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp.5460–5469. DOI: 10.1109/CVPR52688.2022.00539.

[81] Huang J W, Chen Z L, Ceylan D, Jin H L. 6-DOF VR videos with a single 360-camera. In *Proc. the 2017 IEEE Virtual Reality (VR)*, Mar. 2017, pp.37–44. DOI: 10.1109/VR.2017.7892229.

[82] Baker L, Mills S, Zollmann S, Ventura J. CasualStereo: Casual capture of stereo panoramas with spherical structure-from-motion. In *Proc. the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, Mar. 2020, pp.782–790. DOI: 10.1109/VR46266.2020.00102.

[83] Chen R S, Zhang F L, Finnie S, Chalmers A, Rhee T. Casual 6-DoF: Free-viewpoint panorama using a handheld 360° camera. *IEEE Trans. Visualization and Computer Graphics*, 2022: 1. DOI: 10.1109/TVCG.2022.3176832.

[84] Waidhofer J, Gadgil R, Dickson A, Zollmann S, Ventura J. PanoSynthVR: Toward light-weight 360-degree view synthesis from a single panoramic input. In *Proc. the 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2022, pp.584–592. DOI: 10.1109/ISMAR55827.2022.00075.

[85] DiVerdi S, Wither J, Hollerer T. Envisor: Online environment map construction for mixed reality. In *Proc. the 2008 IEEE Virtual Reality Conference*, Mar. 2008, pp.19–26. DOI: 10.1109/VR.2008.4480745.

[86] Park J, Park H, Yoon S E, Woo W. Physically-inspired deep light estimation from a homogeneous-material object for mixed reality lighting. *IEEE Trans. Visualization and Computer Graphics*, 2020, 26(5): 2002–2011. DOI: 10.1109/TVCG.2020.2973050.

[87] Georgoulis S, Rematas K, Ritschel T, Gavves E, Fritz M, Van Gool L, Tuytelaars T. Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2018, 40(8): 1932–1947. DOI: 10.1109/TPAMI.2017.2742999.

[88] Wei X, Chen G J, Dong Y, Lin S, Tong X. Object-based illumination estimation with rendering-aware neural networks. In *Proc. the 16th European Conference on Computer Vision*, Aug. 2020, pp.380–396. DOI: 10.1007/978-3-030-58555-6_23.

[89] Hold-Geoffroy Y, Sunkavalli K, Hadap S, Gambaretto E, Lalonde J F. Deep outdoor illumination estimation. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Jul. 2017, pp.2373–2382. DOI: 10.1109/CVPR.2017.255.

[90] Zhang J S, Sunkavalli K, Hold-Geoffroy Y, Hadap S, Eisenman J, Lalonde J F. All-weather deep outdoor lighting estimation. In *Proc. the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2019, pp.10150–10158. DOI: 10.1109/CVPR.2019.01040.

[91] Hold-Geoffroy Y, Athawale A, Lalonde J F. Deep sky modeling for single image outdoor lighting estimation. In *Proc. the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2019, pp.6920–6928. DOI: 10.1109/CVPR.2019.00709.

[92] Yu P P, Guo J, Huang F, Zhou C, Che H W, Ling X, Guo Y W. Hierarchical disentangled representation learning for outdoor illumination estimation and editing. In *Proc. the 2021 IEEE/CVF International Conference on Computer Vision*, Oct. 2021, pp.15293–15302. DOI: 10.1109/ICCV48922.2021.01503.

[93] Zhu Y J, Zhang Y D, Li S, Shi B X. Spatially-varying outdoor lighting estimation from intrinsics. In *Proc. the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2021, pp.12829–12837. DOI: 10.1109/CVPR46437.2021.01264.

[94] Tang J J, Zhu Y J, Wang H Y, Chan J H, Li S, Shi B X. Estimating spatially-varying lighting in urban scenes with disentangled representation. In *Proc. the 17th European Conference on Computer Vision*, Oct. 2022, pp.454–469. DOI: 10.1007/978-3-031-20068-7_26.

[95] Gardner M A, Sunkavalli K, Yumer E, Shen X H, Gambaretto E, Gagné C, Lalonde J F. Learning to predict indoor illumination from a single image. *ACM Trans. Graphics*, 2017, 36(6): Article No. 176. DOI: 10.1145/3130800.3130891.

[96] Song S R, Funkhouser T. Neural Illumination: Lighting prediction for indoor environments. In *Proc. the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2019, pp.6911–6919. DOI: 10.1109/CVPR.2019.00708.

[97] Srinivasan P P, Mildenhall B, Tancik M, Barron J T, Tucker R, Snavely N. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *Proc. the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp.8077–8086. DOI: 10.1109/CVPR42600.2020.00810.

[98] Chalmers A, Zhao J H, Medeiros D, Rhee T. Reconstructing reflection maps using a stacked-CNN for mixed reality rendering. *IEEE Trans. Visualization and Computer Graphics*, 2021, 27(10): 4073–4084. DOI: 10.1109/TVCG.2020.3001917.

[99] Zhao J H, Chalmers A, Rhee T. Adaptive light estimation using dynamic filtering for diverse lighting conditions. *IEEE Trans. Visualization and Computer Graphics*, 2021, 27(11): 4097–4106. DOI: 10.1109/TVCG.2021.3106497.

[100] Zhan F N, Zhang C G, Yu Y C, Chang Y, Lu S J, Ma F Y, Xie X S. EMLight: Lighting estimation via spherical distribution approximation. In *Proc. the 35th AAAI Conference on Artificial Intelligence*, Feb. 2021, pp.3287–3295. DOI: 10.1609/aaai.v35i4.16440.

[101] Zhan F N, Yu Y C, Wu R L, Zhang C G, Lu S J, Shao L, Ma F Y, Xie X S. GMLight: Lighting estimation via geometric distribution approximation. arXiv: 2102.10244, 2021. https://arxiv.org/abs/2102.10244v1, Jul. 2023.

[102] Xu J P, Zuo C Y, Zhang F L, Wang M. Rendering-aware HDR environment map prediction from a single image. In *Proc. the 36th AAAI Conference on Artificial Intelligence*, Feb. 22,–Mar. 1, 2022, pp.2857–2865. DOI: 10.1609/aaai.v36i3.20190.

[103] Akimoto N, Matsuo Y, Aoki Y. Diverse plausible 360-degree image outpainting for efficient 3DCG background creation. In *Proc. the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2022, pp.11431–11440. DOI: 10.1109/CVPR52688.2022.01115.

[104] Somanath G, Kurz D. HDR environment map estimation for real-time augmented reality. In *Proc. the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2021, pp.11293–11301. DOI: 10.1109/CVPR46437.2021.01114.

[105] Wagner D, Mulloni A, Langlotz T, Schmalstieg D. Real-time panoramic mapping and tracking on mobile phones. In *Proc. the 2010 IEEE Virtual Reality Conference (VR)*, Mar. 2010, pp.211–218. DOI: 10.1109/VR.2010.5444786.

[106] Gauglitz S, Sweeney C, Ventura J, Turk M, Hlerer T. Live tracking and mapping from both general and rotation-only camera motion. In *Proc. the 2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Nov. 2012, pp.13–22. DOI: 10.1109/ISMAR.2012.6402532.

[107] Gauglitz S, Sweeney C, Ventura J, Turk M, Hollerer T. Model estimation and selection towards unconstrained real-time tracking and mapping. *IEEE Trans. Visualization and Computer Graphics*, 2014, 20(6): 825–838. DOI: 10.1109/TVCG.2013.243.

[108] Pirchheim C, Schmalstieg D, Reitmayr G. Handling pure camera rotation in keyframe-based SLAM. In *Proc. the 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2013, pp.229–238. DOI: 10.1109/ISMAR.2013.6671783.

[109] Baker L, Ventura J, Zollmann S, Mills S, Langlotz T. SPLAT: Spherical localization and tracking in large spaces. In *Proc. the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, Mar. 2020, pp.809–817. DOI: 10.1109/VR46266.2020.00105.

[110] Pan Q, Arth C, Reitmayr G, Rosten E, Drummond T. Rapid scene reconstruction on mobile phones from panoramic images. In *Proc. the 10th IEEE International Symposium on Mixed and Augmented Reality*, Oct. 2011, pp.55–64. DOI: 10.1109/ISMAR.2011.6092370.

[111] Arth C, Klopschitz M, Reitmayr G, Schmalstieg D. Real-time self-localization from panoramic images on mobile devices. In *Proc. the 10th IEEE International Symposium on Mixed and Augmented Reality*, Oct. 2011, pp.37–46. DOI: 10.1109/ISMAR.2011.6092368.

[112] Reinisch G, Arth C, Schmalstieg D. Panoramic mapping on a mobile phone GPU. In *Proc. the 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2013, pp.291–292. DOI: 10.1109/ISMAR.2013.6671810.

[113] Pece F, Steptoe W, Wanner F, Julier S, Weyrich T, Kautz J, Steed A. PanoInserts: Mobile spatial teleconferencing. In *Proc. the 2013 SIGCHI Conference on Human Factors in Computing Systems*, Apr. 2013, pp.1319–1328. DOI: 10.1145/2470654.2466173.

[114] Speicher M, Cao J C, Yu A, Zhang H H, Nebeling M. 360Anywhere: Mobile ad-hoc collaboration in any environment using 360 video and augmented reality. *Proceedings of the ACM on Human-Computer Interaction*, 2018, 2(EICS): 9. DOI: 10.1145/3229091.

[115] Piumsomboon T, Lee G A, Irlitti A, Ens B, Thomas B H, Billinghurst M. On the shoulder of the giant: A multi-scale mixed reality collaboration with 360 video sharing and tangible interaction. In *Proc. the 2019 CHI Conference on Human Factors in Computing Systems*, May 2019, Article No. 228. DOI: 10.1145/3290605.3300458.

[116] Teo T, Lawrence L, Lee G A, Billinghurst M, Adcock M. Mixed reality remote collaboration combining 360 video and 3D reconstruction. In *Proc. the 2019 CHI Conference on Human Factors in Computing Systems*, May 2019, Article No. 201. DOI: 10.1145/3290605.3300431.

[117] Wang P, Bai X L, Billinghurst M, Zhang S S, Zhang X Y, Wang S X, He W P, Yan Y X, Ji H Y. AR/MR remote collaboration on physical tasks: A review. *Robotics and Computer-Integrated Manufacturing*, 2021, 72: 102071. DOI: 10.1016/j.rcim.2020.102071.

[118] Nebeling M, Madier K. 360proto: Making interactive virtual reality & augmented reality prototypes from paper. In *Proc. the 2019 CHI Conference on Human Factors in Computing Systems*, May 2019, Article No. 596. DOI: 10.1145/3290605.3300826.

[119] Zhu Z, Martin R R, Hu S M. Panorama completion for street views. *Computational Visual Media*, 2015, 1(1): 49–57. DOI: 10.1007/s41095-015-0008-2.

[120] He K M, Sun J. Image completion approaches using the statistics of similar patches. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2014, 36(12): 2423–2435. DOI: 10.1109/TPAMI.2014.2330611.

[121] Xu B B, Pathak S, Fujii H, Yamashita A, Asama H. Spatio-temporal video completion in spherical image sequences. *IEEE Robotics and Automation Letters*, 2017, 2(4): 2032–2039. DOI: 10.1109/LRA.2017.2718106.

[122] Zhao Q, Wan L, Feng W, Zhang J W, Wong T T. 360 panorama cloning on sphere. arXiv: 1709.01638, 2017. https://arxiv.org/abs/1709.01638, Jul. 2023.

[123] Huang K, Zhang F L, Zhao J H, Li Y H, Dodgson N. 360° stereo image composition with depth adaption. arXiv: 2212.10062, 2022. https://arxiv.org/abs/2212.10062, Jul. 2023.

[124] Jung J, Kim B, Lee J Y, Kim B, Lee S. Robust upright adjustment of 360 spherical panoramas. *The Visual Computer*, 2017, 33(6): 737–747. DOI: 10.1007/s00371-017-1368-7.

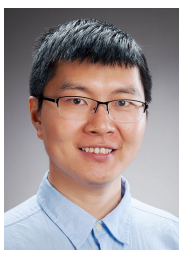
[125] Zhang Y, Zhang F L, Lai Y K, Zhu Z. Efficient propagation of sparse edits on 360° panoramas. *Computers & Graphics*, 2021, 96: 61–70. DOI: 10.1016/j.cag.2021.03.005.

[126] Zhang Y, Zhang F L, Zhu Z, Wang L D, Jin Y. Fast edit propagation for 360 degree panoramas using function interpolation. *IEEE Access*, 2022, 10: 43882–43894. DOI: 10.1109/ACCESS.2022.3168665.

[127] Wong K M. View-adaptive asymmetric image detail enhancement for 360-degree stereoscopic VR content. In *Proc. the 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, Mar. 2022, pp.23–26. DOI: 10.1109/VRW55335.2022.00012.

[128] Wang M, Li Y J, Zhang W X, Richardt C, Hu S M. Transitioning360: Content-aware NFOV virtual camera paths for 360° video playback. In *Proc. the 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Nov. 2020, pp.185–194. DOI: 10.1109/ISMAR50242.2020.00040.

[129] Li Y J, Shi J C, Zhang F L, Wang M. Bullet comments for 360°video. In *Proc. the 2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, Mar. 2022. DOI: 10.1109/VR51125.2022.00017.



Fanglue Zhang is currently a senior lecturer with Victoria University of Wellington, Wellington. He received his Bachelor's degree from Zhejiang University, Hangzhou, in 2009, and his Ph.D. degree from Tsinghua University, Beijing, in 2015. His research interests include image and video editing, mixed reality, and image-based graphics. He is a member of ACM and IEEE. He received Victoria Early Career Research Excellence Award in 2019. He is on the editorial board of Computer & Graphics. He is a committee member of IEEE Central New Zealand sector.

Junhong Zhao is currently a research fellow with School of Engineering and Computer Science of Victoria University of Wellington, Wellington. She completed her Ph.D. degree in 2015 at the Institute of Electronics of the Chinese Academy of Sciences, Beijing. She worked at the Institute of Information Engineering of the Chinese Academy of Sciences, Beijing, as an assistant researcher from 2015 to 2017. From 2018 to 2022, she was working with the Computational Media Innovation Centre (CMIC) at Victoria University of Wellington, Wellington, as a postdoctoral research fellow. Her research interests include machine learning, image processing and computer vision.

Yun Zhang is currently a professor at the Communication University of Zhejiang, Hangzhou. He received his Ph.D. degree from Zhejiang University, Hangzhou, in 2013. Before that, he received Bachelor's and Master's degrees from Hangzhou Dianzi University, Hangzhou, in 2006 and 2009, respectively. He visited the Visual Computing Group of Cardiff University in 2018 and 2023, and the Computational Media Innovation Centre (CMIC) of Victoria University of Wellington in 2019. His research interests include computer graphics, image and video editing, and computer vision. He is a senior member of CCF.

Stefanie Zollmann is an associate professor at the University of Otago, Dunedin. Before, she worked at Animation Research Limited on cross reality (XR) visualization and tracking technology for sports broadcasting. She worked as postdoctoral researcher at the Graz University of Technology, Styria, where she also obtained her Ph.D. degree in 2013. She received a University of Otago Early Career Award for Distinction in Research in 2020 and she is on the editorial board of renown journals such as TVCG and Computer & Graphics. Her main research interests are XR for sports and media, and visualization techniques for augmented reality, and also include capturing for XR.