Computational Visual Media
https://doi.org/10.1007/s41095-020-0162-z Vol. 6, No. 1, March 2020, 3–28

Review Article

VR content creation and exploration with deep learning: A survey

Miao Wang[1,2] (✉), Xu-Quan Lyu[1], Yi-Jun Li[1], and Fang-Lue Zhang[3]

Abstract Virtual reality (VR) offers an artificial, computer generated simulation of a real life environment. It originated in the 1960s and has evolved to provide increasing immersion, interactivity, imagination, and intelligence. Because deep learning systems are able to represent and compose information at various levels in a deep hierarchical fashion, they can build very powerful models which leverage large quantities of visual media data. Intelligence of VR methods and applications has been significantly boosted by the recent developments in deep learning techniques. VR content creation and exploration relates to image and video analysis, synthesis and editing, so deep learning methods such as fully convolutional networks and general adversarial networks are widely employed, designed specifically to handle panoramic images and video and virtual 3D scenes. This article surveys recent research that uses such deep learning methods for VR content creation and exploration. It considers the problems involved, and discusses possible future directions in this active and emerging research area.

Keywords virtual reality; deep learning; neural networks; 360° image and video virtual content

1 Introduction

Virtual reality (VR) is an artificial, computer generated simulation of a real life environment. It immerses the viewer into a computer-generated 3D environment in which they can explore and interact. Over the near 60-year history of VR, the availability and flexibility of displays and other devices has increased, facilitating the prevalence of VR. Starting from 2014 when consumer-grade head-mounted displays (HMD) such as Oculus Rift [1] and HTC Vive [2] became available for commercial use, VR has entered a new era. This technology has now reached the critical mass of technical maturity and content pervasiveness needed to drive the growth that will embed VR within the multiple sectors of the economy, such as entertainment, education, and tourism.

A VR environment can be created from real life images/videos, or computer-generated 3D models and scenes. The key aspects of providing an immersive VR experience to users are the fidelity of the VR content, and the realism of VR interaction; we need to utilize artificial intelligence approaches to achieve high-quality VR environment construction, to analyze the rich information in VR content, and to properly understand user actions. Recently, AI technologies have leapt forward with the development and application of deep neural networks. The continuously evolving capabilities of deep learning systems has catalyzed their uptake in VR research, especially in VR content creation and exploration tasks. Because deep learning systems are able to represent and compose information in a deep hierarchical fashion with simple non-linear building blocks, they can learn very powerful models from the large quantities of visual media data currently available today. Deep learning methods are likely to become mainstream in the coming years in several sub-fields of VR research.

VR content creation and exploration have attracted research interest in recent decades. VR content can

1 State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China. E-mail: M. Wang, miaow@buaa.edu.cn (✉); X.-Q. Lyu, aincrad@buaa.edu.cn; Y.-J. Li, yaoling@buaa.edu.cn.

2 Peng Cheng Laboratory, Shenzhen 518000, China.

3 School of Engineering and Computer Science, Victoria University of Wellington, New Zealand. E-mail: fanglue.zhang@ecs.vuw.ac.nz.

Manuscript received: 2020-01-30; accepted: 2020-03-02

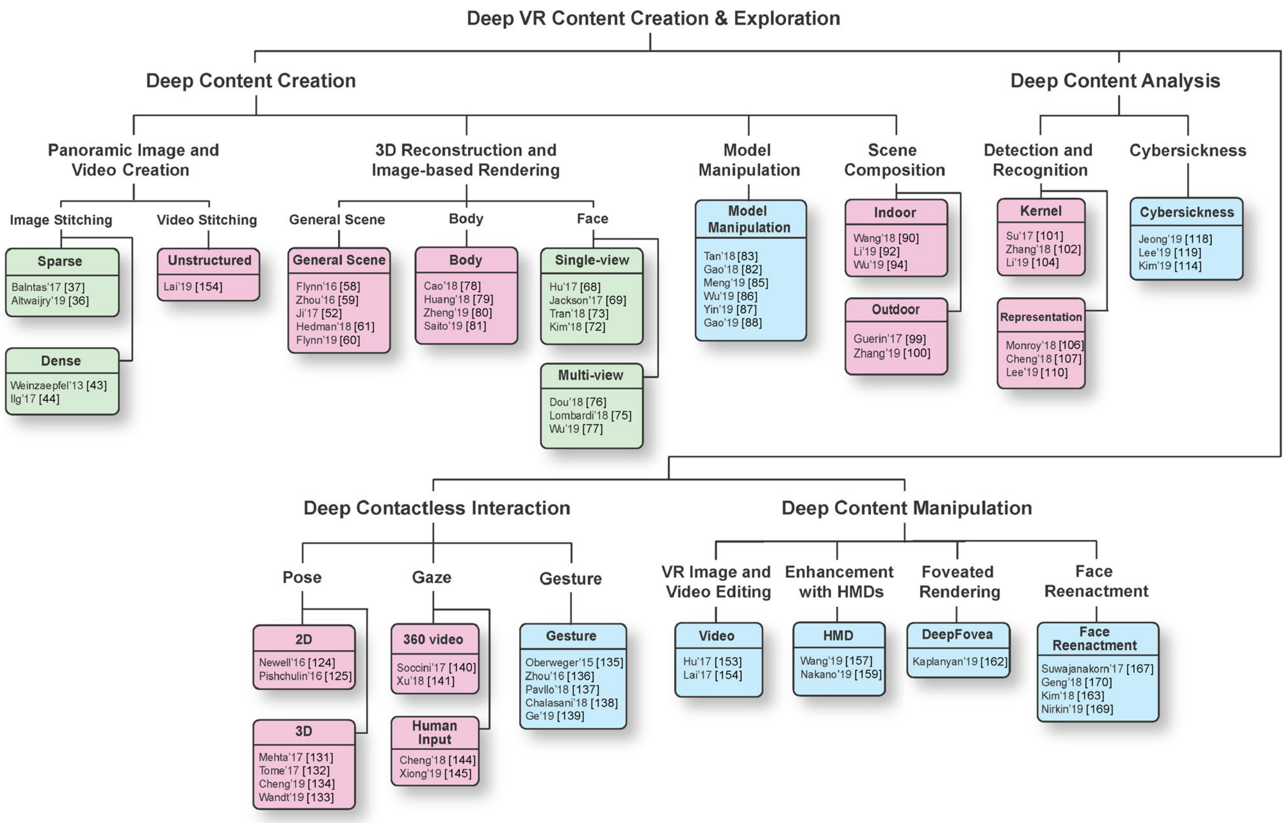


Fig. 1 Taxonomy of deep learning-based VR content creation and exploration techniques in this paper. Representative works are listed in colored boxes.

be computationally generated from either normal field-of-view (NFOV) photos and videos using image stitching [3] and scene reconstruction [4] techniques, or 3D modeling [5] and scene composition [6] methods with a rendering process [7]. After that, the virtual scenes need to be analyzed and understood via computer vision algorithms such as object detection and scene parsing to provide semantic information for further object-aware or scene-aware manipulation and rendering. A virtual scene can be presented to the user using normal 2D displays or head-mounted displays (HMDs) for interactive exploration. Meanwhile, the user's face, pose, gesture, and/or gaze can be recorded, recognized, and tracked by surrounding sensors for accurate and intelligent interaction with the virtual environment. Deep learning methods can be integrated into each stage of the whole pipeline to improve the capability, effectiveness, and efficiency of VR systems, while reducing the amount of labor, expense, and redundancy.

Deep learning-based methods have been investigated and successfully applied to many computer vision tasks for visual media, analyzing images, videos, and geometric models and scenes. For example, single-stage real-time object detection YOLO [8] and two-stage instance-level Mask-RCNN networks [9] can detect and recognize objects in an image; segmentation and parsing [10–12] networks were invented to assign pixel-wise category labels to images to understand scenes; scene graph creation networks [13, 14] further understand possible relationships between objects. To manipulate image content, style transfer [15–17] and image-to-image translation [18–21] networks have been developed. Nevertheless, VR content is much more complex to analyze and manipulate than normal photos and videos for two reasons: firstly, the resolution and field-of-view of 360° images and videos presented in VR are much larger, and so include more scene content. The presence of 360° image and video using latitude–longitude projection can also produce severe distortion. Secondly, virtual content is usually experienced immersively with head-mounted displays, where editing artifacts attract much greater visual attention or lead to uncomfortable viewing experiences. Higher quality is required of the content generated for VR than for normal media. To this end, deep learning methods for VR content creation and exploration are specifically designed with larger receptive fields for spatial-awareness and high-efficiency for handling 360° video data.

In this survey, we cover recent papers that leverage deep learning methods for VR content creation and exploration. The survey is structured as follows: Section 2 considers deep models for VR content creation. Section 3 reviews recent papers in VR content analysis. Section 4 describes recent deep learning work on VR content exploration and interaction. Section 5 introduces VR content manipulation methods using deep neural networks. Finally, Section 6 draws conclusions and discusses possible future directions and trends in this active research field.

2 VR content creation with deep learning

The creation of high-fidelity VR content forms the foundation of immersive VR experience. Generally, two types of sources are employed for computational VR content creation: real-life images and videos, and objects and scenes created automatically or interactively using computers. Deep neural networks are utilized in 360° image and video generation, and scene composition. Here, we review representative techniques and recently proposed algorithms.

2.1 Panoramic image and video creation

Panoramic image and video, or 360° image/video, synthesized from real-life photos, are typically used for consumer-level VR applications [23], and can be viewed even using mobile phones. Raw images and video captured by various devices are seamlessly stitched to generate a panoramic scene for VR presence. Normally, 360° images presented in VR are stereoscopic. Thus solutions for capturing and rendering stereo imagery have been proposed with both fixed camera arrays [24–27] and casual photography [28–30]. The raw images and video are then warped and stitched to make a 360° panorama [31] for VR display. The normal pipeline in conventional panorama stitching techniques [3, 32–34] consists of 2D transformation estimation and seamless stitching with blending [35]. However, they cannot produce acceptable results if correctly matched feature points are lacking. Deep learning-based sparse and dense image matching methods have been proposed to overcome that limitation. For sparse matching, deep features [36, 37] are used for effective correspondence matching. Deep homography estimation methods [38–40] take source and target images as input, and output displacement vectors at image corners. An unsupervised approach for homography estimation was proposed in Ref. [40], which uses a triple loss to ensure content awareness. Correspondence prediction on pixels or mesh cells can be utilized to solve the local content matching problem for transformation estimation. Ye et al. [41] presented the DeepMeshFlow model to predict a sparse motion field from a pair of images with associated motions at mesh vertices. For dense matching [42], various optical flow estimation methods using deep learning have been investigated [43, 44]. To learn more, we refer readers to a survey of optical flow estimation using convolutional neural networks (CNN) [45].

Recently, unstructured video stitching methods have been explored [22, 46, 47]. Lai et al. [22] proposed a neural network for video captured by a linear camera array (see Fig. 2). It casts the stitching problem in terms of spatial interpolation, and presents a pushbroom interpolation layer with the assistance of flow estimation to seamlessly stitch multi-view videos.

Obtained panorama content with a 360° field-of-view sometimes needs to be further processed, for example due to unsuitable camera setup during data acquisition: e.g., if the cameras are tilted when capturing data, the stitched result will be mis-oriented. Jung et al. [48] proposed a CNN-based method to predict the elevation and azimuth angles of the up-vector for a given VR image. To support model training, a large scale dataset of VR images with different orientations was generated, by combining random rotations and resizing of high resolution VR images from the SUN360 dataset [49].

2.2 3D reconstruction and image-based rendering

Geometric information is missing in real-life images and videos. To support interactive navigation and exploration in certain VR applications, 3D geometry needs to be reconstructed from the raw data either implicitly or explicitly. Also, image-based rendering

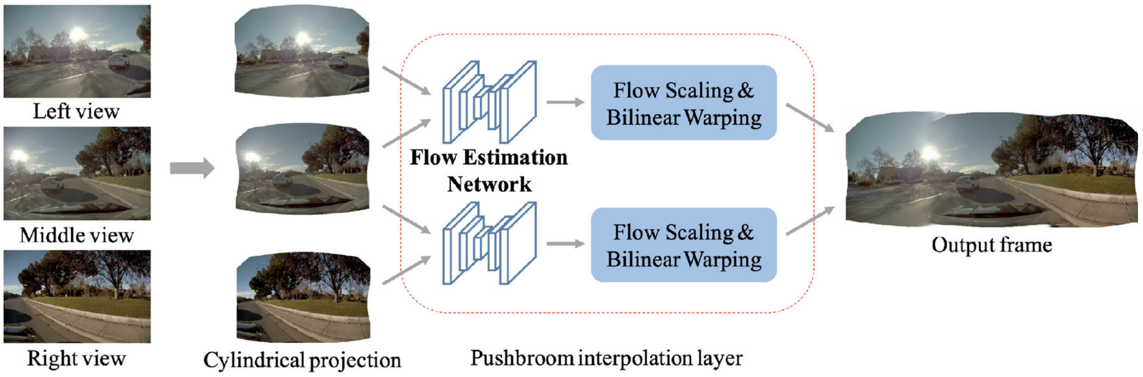

Left view
Middle view
Right view
Cylindrical projection
Flow Estimation Network
Flow Scaling & Bilinear Warping
Flow Scaling & Bilinear Warping
Pushbroom interpolation layer
Output frame

Fig. 2 Video stitching for linear camera arrays using CNNs with a pushbroom interpolation layer. Reproduced with permission from Ref. [22], © The British Machine Vision Association 2019.

(IBR) techniques can be used to faithfully synthesize or enhance images from novel viewpoints. The reconstructed and rendered scene can be interactively experienced with HMDs and controllers. Here, we review existing deep learning-based work for several major categories of reconstruction methods, namely for general scenes, human faces, and human bodies.

2.2.1 General scenes

The complexity of geometric data and the availability of large datasets make it tempting and very desirable to resort to machine learning techniques. To improve the reconstruction of earlier 3D geometry estimation methods [50, 51], 3D neural networks [52] were proposed to reconstruct dense 3D point clouds from unstructured photo collections. Using point clouds, Delaunay tetrahedralization methods [53, 54] can be used to create mesh-based models. Xi and Chen [55] proposed a multi-view regularization-based method for piecewise planar scene reconstruction. Given multiple images from different viewpoints, this method recovers planar segments and a depth map under a planar shape constraint for each image using CNNs, and then composes the above features to reconstruct the 3D scene with multi-view regularization. Various scene reconstruction methods have been evaluated in a benchmark [56].

Image-based rendering methods take images and videos as input to create realistic scenes from novel viewpoints. In early work, the unstructured lumigraph rendering [57] represents the scene as a simple geometric proxy, and then re-projects and blends the input images using new viewpoints. Several studies investigated CNN-based methods in this field. Flynn et al. [58] proposed the end-to-end DeepStereo framework with two towers of layers for depth and color prediction. DeepStereo directly produces the pixels of the unseen view using pixels from neighboring views. Zhou et al. [59] proposed a network to predict appearance flows of 2D coordinate offset vectors to reconstruct a new target view, and further generalized it to combine multiple single-view predictions. DeepView [60] represents the scene as multi-plane images (MPIs) learned with gradient descent. The method is aware of occlusion and improves results on challenging scenes with thin structures and high depth complexity. Hedman et al. [61] proposed DeepBlending, a CNN-based IBR blending system with per-view geometry refinement and geometry-aware mesh simplification for quality improvement. Given a set of photos from several views, the proposed blending network takes selected warped view mosaics and a global mesh rendering, and outputs weights for blending each pixel from the views (see Fig. 3). Multi-view image fusion [62] was proposed to fuse misaligned photos from different camera sensors. It can transfer details from a high-quality DSLR image to images taken by a VR180 camera [63]. It then learns to predict optical flows at different granularities with a novel cascaded feature extraction stage, and fuses features hierarchically.

2.2.2 Faces

3D reconstruction of humans is an important topic in VR content creation. The face is the most significant visual aspect of a human in visual perception, as it can convey messages, identity, emotion, and intent of humans [65]. While traditional methods can render highly realistic faces, such methods are not widely applied in many VR application scenarios, as they depend on physically accurate estimation of geometry and shading models

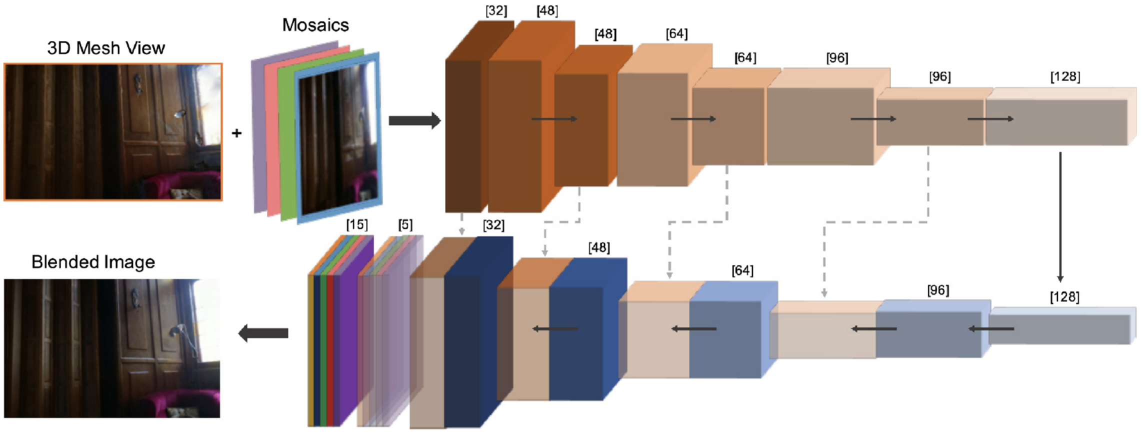


Fig. 3 The network architecture of DeepBlending for free-viewpoint image-based rendering. Reproduced with permission from Ref. [61], © Association for Computing Machinery 2018.

considering colors, facial landmarks, and edges, which involves high computational costs. If estimation of facial parameters is not accurate, the perceptual quality of the face rendering results will be degraded. On the other hand, traditional methods usually require multiple images of the same person to achieve sufficient smoothness and consistency. Deep learning methods have made remarkable progress in solving the above problems in real time.

Face reconstruction and modeling methods can be classified into two categories according to the type of input: one uses lightweight setups that mostly work with a single monocular RGB or RGB-D camera, and achieve real-time performance by making a trade-off between face reconstruction quality and speed [65], while the other uses sophisticated multi-view setups with off-line processing to provide higher quality face modeling.

Typically, CNN-based deep neural networks are widely used for modelling a face from monocular input. Tran et al. [66] proposed a method for 3D face reconstruction from a single image using CNN to regress the shape and texture parameters with a discriminative 3D morphable face model (3DMM) [67]. Hu et al. [68] presented an architecture that can model a complete 3D head with hair from a single image by integrating the latest methods of facial segmentation, shape modeling, and high-fidelity appearance inference. They also use a deep CNN for hairstyle retrieval. Jackson et al. [69] proposed a CNN architecture using a novel volumetric representation that can reconstruct the entire 3D facial geometry for arbitrary poses and expressions of a face. In order to derive the reconstructed face shape from a single image in a coarse-to-fine manner, Richardson et al. [70] proposed an end-to-end CNN framework containing two sub-networks. Their first sub-network, CoarseNet, provides coarse facial geometry recovery. It is then followed by FineNet for facial feature refinement. Dou et al. [71] introduced an approach for end-to-end 3D face reconstruction (UH-E2FAR) from a single image. With a multi-task loss function and a fusion module, neutral 3D facial shape and expressive 3D facial shape are reconstructed. Tewari et al. [64] proposed a novel model-based deep convolutional autoencoder to reconstruct the face from a single image (see Fig. 4). Their proposed network combines an encoder for semantic parameter extraction (e.g., pose, shape, expression, skin reflectance, and illumination) and a differentiable model-based decoder. Kim et al. [72] invented a deep convolutional inverse rendering framework, InverseFaceNet, for joint estimation of facial pose, shape, expression, reflectance, and illumination from a single input image, in real time. By using a novel loss function, InverseFaceNet directly measures model-space similarity in parameter space, which significantly improves reconstruction accuracy. Tran et al. [73] presented a deep convolutional encoder–decoder framework to provide detailed 3D reconstructions of faces viewed under extreme conditions (including occlusion) by using bump maps to represent the coarse 3D face shape with wrinkles. To provide high fidelity texture details in the reconstructed face model, GANFIT was proposed with GANs to train a very powerful generator of facial texture in UV space; it is integrated in 3DMMs fitting approach [74].

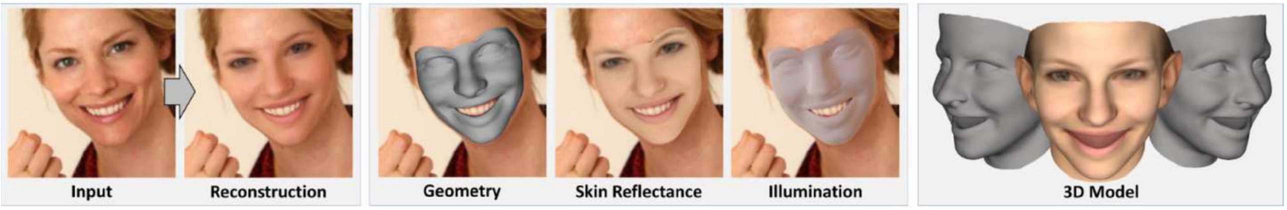


Fig. 4 Model-based neural network for unsupervised monocular reconstruction. The proposed face autoencoder enables unsupervised end-to-end learning of semantic parameters including geometry, illumination, expression, etc. Reproduced with permission from Ref. [64], © Institute of Electrical and Electronics Engineers 2017.

Due to variations in input image quality and insufficient facial information, it could be challenging to accurately reconstruct a 3D face from a single image. Therefore, some authors have suggested reconstructing face models using multi-view face images. Lombardi et al. [75] presented a deep appearance model for rendering a human face using a data-driven rendering pipeline. It utilizes a multi-view capture setup to learn a joint representation of the facial geometry and appearance, and then uses a deep variational autoencoder for predicting vertex positions and view-specific textures. Dou and Kakadiaris [76] proposed a recurrent model for multi-view 3D face reconstruction. They use a subspace for the 3D facial shape representation and a deep recurrent neural network which consists of both a deep convolutional neural network (DCNN) and a recurrent neural network (RNN). The DCNN extracts the identity and expression of the face from each image alone, while the RNN fuses features related to identity from the DCNN and aggregates identity-specific contextual information. Wu et al. [77] proposed an approach to predict 3DMM [67] parameters with an end-to-end CNN from a set of multi-view facial images as input, which can generate the minimized photometric reprojection error between each observed image and the generated image.

2.2.3 Bodies

Human body reconstruction is used for avatar creation and animation. CNN-based methods have been proposed to leverage parametric human body models, starting from single or multiple photos or RGBD images. Cao et al. [78] introduced a cascaded 3D fully convolutional network to reconstruct implicit surface representations from noisy and incomplete depth maps in a two-stage process (see Fig. 5). Huang et al. [79] proposed a multi-view CNN for sparse human performance capture. The method maps 2D images to a 3D volumetric field encoding the probabilistic distribution of surface points of the captured subject. From the 3D volumetric field, a clothed human body can be reconstructed at arbitrary resolutions. Recently, single image body reconstruction methods have been proposed. Generalizing convolutional methods, the image-guided volume-to-volume translation network DeepHuman [80] learns a dense semantic representation from a skinned multi-person linear model to reconstruct a human from a single image. Different scales of image features are fused in 3D space via volumetric feature transformations to recover details of the subject's outer surface geometry. Details on the frontal areas of the outer surface are further refined via a normal map refinement network. Pixel-aligned implicit functions [81] have been explored to represent local alignment of pixels from 2D images to a global context for the target 3D object. With such a representation, highly comprehensive clothed humans with detailed hairstyles can be inferred with both 3D surface and texture from single or multiple images.

2.3 3D model manipulation

The aforementioned methods reconstruct geometry from real-life images and videos. In the cases where the environment is designed to respond to user input, or a customized VR environment needs to be provided, we need further 3D model manipulation methods. These include shape deformation and transformation; they are also widely used in animation. Deep learning-based methods have already surpassed traditional methods in some specific domains, which shows the feasibility of a machine learning to deform and transform shapes.

Research has provided evidence that compressing the shapes to a latent feature space is beneficial for later manipulation and can lead to improved results. Tan et al. [83] proposed a CNN-based auto-encoder which can deal with meshes with irregular connectivity. The method first uses an effective

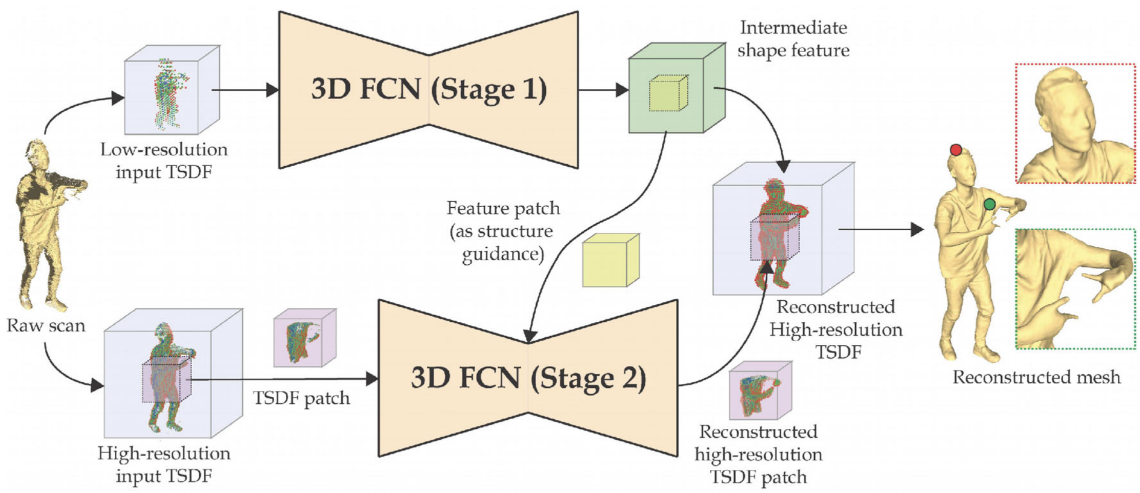

3D FCN (Stage 1)
Intermediate shape feature
Low-resolution input TSDF
Raw scan
Feature patch (as structure guidance)
Reconstructed High-resolution TSDF
Reconstructed mesh
High-resolution input TSDF
TSDF patch
3D FCN (Stage 2)
Reconstructed high-resolution TSDF patch

Fig. 5 Two-stage 3D-CFCN architecture for 3D body reconstruction from low quality raw depth scans. An intermediate representation is produced with a fused low-resolution TSDF volume for stage 1 of the 3D-FCN. Then the network regresses a complete low-resolution TSDF and predicts TSDF patches for further refinement. Next, for each patch to be refined, the corresponding block is taken from a fused high-resolution input TSDF to infer a detailed high-resolution local TSDF volume that can substitute for the corresponding block in the regressed TSDF for quality improvement. Reproduced with permission from Ref. [78], © Springer Nature Switzerland AG 2018.

representation [84] for the deformation of shapes, and then utilizes a CNN-based auto-encoder to encode the deformation representation into the latent space. After that, sparsity regularization is introduced to help identify sparse localized deformation by applying it to weights in the fully-connected layers. These enable the method to extract intuitive localized deformation components while being insensitive to noise. Meng et al. [85] introduced a voxel variational autoencoder (VAE) network for robust point segmentation which considers both spatial distribution of points and group symmetry. The network first transforms an unstructured point cloud to a voxel grid, and employs radial basis functions which are symmetric around point samples to handle sparse distributions of points. A kernel-based interpolated VAE architecture is then used to effectively encode the local geometry within each voxel. Robustness is further enhanced by extending the group equivalent CNN to 3D; this improves the expressive capacity without increasing the number of parameters. Gao et al. [82] proposed an approach based on generative adversarial networks (GAN) to automatically transfer deformations between unpaired shape datasets. Source and target shapes are first encoded to their own latent spaces by two convolutional VAEs respectively to get more compact and effective representations. Afterwards, GAN is employed to find the mapping between source shapes and deformed target shapes in the latent space. To make the mapping more reliable, reverse mapping from target shapes to source shapes is also utilized (see Fig. 6). Wu et al. [86] presented SAGNet which is a structure-aware generative model that enables 3D shape generation with separate control over geometry and structure.

Motivated by recent advances in analyzing data using neural networks, GANs have been used to

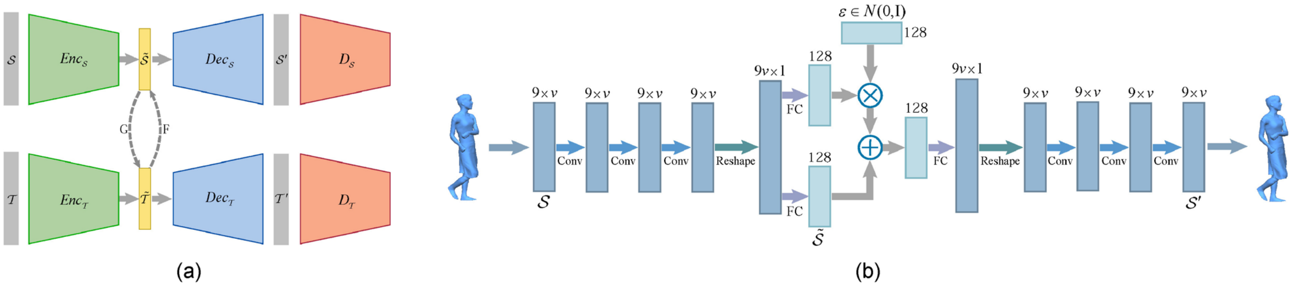


Fig. 6 Automatic unpaired shape deformation transfer. (a) Architecture of the proposed VC-GAN network for deformation transfer. (b) Convolutional variational autoencoder. Reproduced with permission from Ref. [82], © Association for Computing Machinery 2018.

analyze and encode latent relationships between geometry and structure instead of performing unsupervised training with geometric models. During analysis, two branches for structure and geometry exchange information between them, which allows the system to learn the dependency between structure and geometry. Yin et al. [87] introduced LOGAN, a general-purpose shape transformation learning method based on unpaired domains. It includes an auto-encoder and a translator. The CNN-based auto-encoder is designed to encode shapes from the two unpaired domains in a common latent space; latent codes are created to represent multi-scale shape features. The GAN-based translator is devised to perform transformations by operating in the latent space. It contains both an adversarial loss to enforce cross-domain translation and a feature preservation loss to help preserve features during translation. Gao et al. [88] presented SDM-NET, a deep generative neural network that can produce structured deformable meshes. It includes a two-level variational auto-encoder, where one level learns a deformable model of part geometries and the other level learns the part structure of a shape collection and part geometries to ensure coherence between global shape structure and surface details. Its inspiration comes from the observation that a 3D shape can be decomposed into a set of parts even though the overall structure of the 3D shape could be complex, and furthermore, each part is homeomorphic to a box, and so can be recovered by deforming the box.

2.4 3D scene composition

3D virtual scenes that allow daily human activities are in high demand for various VR applications, e.g., interior design and 3D game production [89]. Their popularity has been constantly increasing. Although specialized knowledge is required to design a scene, deep neural networks can learn certain patterns to assist designers. The literature takes two different approaches for indoor and outdoor scene composition.

For indoor scenes, handling the various types of indoor objects is the main challenge. Fu et al. [89] presented an indoor scene synthesis system that adaptively creates 3D scenes using only a small number of object categories specified by users. Since it requires some professional knowledge to design the layout, they exploit a database of 2D floor plans to learn layout examples for scene synthesis and extract object relationships to ensure the functional plausibility of synthetic scenes. They used an activity-associated object relation graph which captures relations between objects to enable adaptive object suggestion. Wang et al. [90] presented a system based on CNNs for synthesizing indoor scenes from scratch. Since a large dataset of 3D scenes has become available [91], they are able to train their CNN model to select and place objects to generate a room scene by iteratively adding objects.

Although convolutional networks cannot directly work in 3D space, they utilize the observation that most objects in a room are arranged on a 2D ground plane, so can use a semantically-enriched representation of scenes based on orthographic top-down views. GRAINS [92] is a generative recursive neural network with a variational auto-encoder inspired by the work of Ref. [93]. In the latter work, they developed a generative recursive auto-encoder for learning hierarchical structures of 3D indoor scenes. The work of Ref. [92] further enables generation of a plausible 3D scene from a random vector. Its auto-encoder performs scene object grouping while encoding information about objects' spatial properties, and scene generation during decoding which turns a randomly sampled code from the learned distribution into a plausible indoor scene hierarchy. Wu et al. [94] proposed a data-driven method to automatically and efficiently generate floor plans for residential buildings given only the boundary. To do so, they created a large-scale dataset (RPLAN) consisting of real floor plans from residential buildings. A living-room-first strategy based on Refs. [95] and [90] was used to determine room connections and positions; it improves the plausibility of resulted floor plans. An encoder–decoder network was then applied to predict wall positions.

Functionality analysis for indoor models and scenes using neural networks is an emerging research topic in virtual reality and robotics. Hu et al. [97] proposed a deep model to predict the functionality of an isolated 3D object, and to generate possible interaction contexts related to the object. The method uses voxels to represent models and 3D-CNN to build the networks. Yan et al. [98] presented a recurrent neural network to predict parts and motion from point clouds. In their work, point-wise displacements for input shape are predicted by an interleaved LSTM-based encoder–decoder. To complete the scene with existing 3D objects, while considering possible interactions among them, a localization and completion network is used [96] (see Fig. 7). In this work, an omni-directional depth image encoding a 360° field of view is used to regress the positions for new objects. Using them, the system predicts corresponding objects which fit the geometry of related objects in the scene.

For outdoor scenes, deep learning based methods tend to focus on terrain manipulation. Guérin et al. [99] proposed an automatic terrain synthesis pipeline driven by real world examples. Each terrain synthesizer is a conditional generative adversarial network. It takes a terrain sketch with visually important features such as summits and valleys as input, and generates the entire elevation of a terrain. Since large training sets are required, they presented a method to automatically extract sketches from real world terrains, to avoid heavy labor needed for artificial sketches. An erosion meta-simulation is also trained to efficiently apply erosion to terrains in their work. Zhang et al. [100] proposed an example-based method to rapidly generate vegetation in outdoor natural environments. The method utilizes a VGG-network to learn the relationships between terrain and vegetation distribution. They treat a 3D scene as a combination of height maps and vegetation density maps for CNN training. Once local information has been extracted from the terrain and vegetation distribution, an initial feature map of the target vegetation distribution is produced, based on patch matching and the vegetation density map. Finally a forward neural network is applied to predict the result.

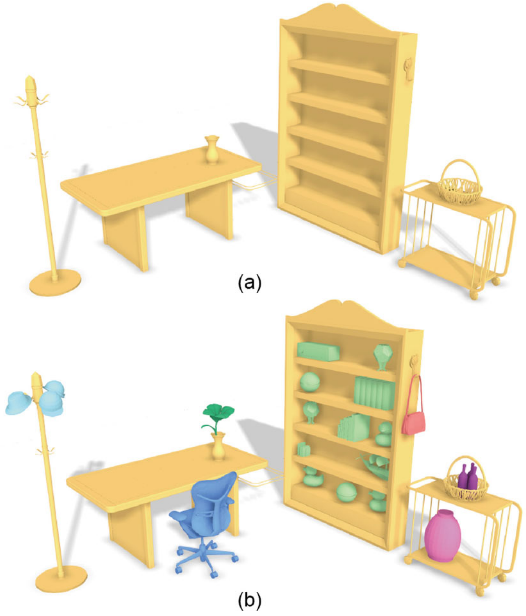


Fig. 7 Given a scene with furniture (a), each piece of furniture is adorned with objects (b) by localizing and completing possible interactions with it. Reproduced with permission from Ref. [96], © Institute of Electrical and Electronics Engineers 2019.

3 Deep VR content analysis

In order to allow sophisticated interactions with VR applications, semantics need to be extracted by deep learning based approaches. On the other hand, to improve the comfort of VR users while exploring or making actions, the factors affecting VR adoption need to be analyzed.

3.1 Detection and recognition

For VR content provided by real life images and videos, object detection and recognition are essential, to extract semantics that are very valuable. Both are typical computer vision tasks, so can be handled by deep learning methods. Nowadays, CNNs are publicly considered to be one of the most effective and powerful tools in computer vision. It is thus a natural idea to use CNNs to solve problems of object detection and recognition in panoramic images and video. However, the distortion caused by sphere-to-plane projection can reduce the performance of CNNs in 360° image and video. To alleviate the distortion when directly applying object detection methods, two types of strategies are used. One is to change the form of the kernel used in the CNN to adapt to the equirectangular representation. The other is to change the representation of the panoramic image or to augment the input with semantics.

To adapt the CNN structure to equirectangular form, Su and Grauman [101] proposed a method called SphConv, which leverages the CNN to extract features on 360° image or video. It improves both efficiency and accuracy by dynamically adapting the kernels' sizes when performing convolution operations on equirectangular images. However, it cannot achieve kernel parameter sharing and may suffer from model bloat. Zhang et al. [102] introduced a new type of spherical CNN. It defines the kernel on a spherical crown, which allows kernel parameters to be shared. Also, taking into account the common format used for 360° videos, they proposed a method to re-sample kernels. They then proposed a spherical U-Net for saliency detection in 360° videos. Coors et al. [103] presented SphereNet, a novel framework aimed at encoding rotation invariance into CNN architectures, improving performance in detection and classification tasks in omnidirectional images. To achieve this, sampling grid locations for convolutional kernels are adjusted based on the geometry of the spherical image representation. The connectivity of spherical images is retained by SphereNet. Li et al. [104] proposed a baseline model, DDS, which includes a distortion-adaptive module and a multi-scale context module based on ResNet-50 to deal with detection problems in images suffering from projection distortion, large-scale complex scenes, and small salient objects. Su and Grauman [105] presented the kernel transformer network (KTN) which adapts a source CNN model trained on perspective images to 360° images by learning a relationship between two different kinds of kernels: an input kernel of the source CNN is transformed to suit 360° images through a learned function (see Fig. 8). Once the function has been learned, the KTN can be applied to multiple source CNNs with the same architecture without the need for retraining.

Transforming the spherical signals to the 2D domain can be better than previous strategies in some situations, as it can utilize well trained CNN models defined over the normal 2D domain. Monroy et al. [106] presented an architectural extension for CNNs to work on omni-directional images by subdividing the omni-directional image into equally-sized undistorted patches by rendering six viewing frustums. By recording the spherical coordinates for each pixel in these patches, they are able to project all pixels to the equirectangular representation, using a CNN to combine the results from the six patches. Cheng et al. [107] proposed a cube padding method which renders panoramic image on cubemaps to avoid distortion. It mitigates disconnectivity between faces by padding (see Fig. 9). Cube padding is feasible in almost all CNN structures and performs better in object detection and recognition tasks. Yang et al. [108] proposed a multi-projection YOLO method which is a variant of the original YOLO detector [109]. It adopts stereographic projection to preserve linear structures, so can reduce the severe geometric distortions caused by equirectangular projection. To further alleviate the impact of distortion near image boundaries, they adopted four sub-windows with an overlap of 90° and used soft selection to select detection results from multiple predicted windows. Lee et al. [110] presented a spherical polyhedron-based representation of omni-directional images (SpherePHD) which aims to overcome shape distortion in the equirectangular representation, and discontinuity at image boundaries of the cubemap representation.

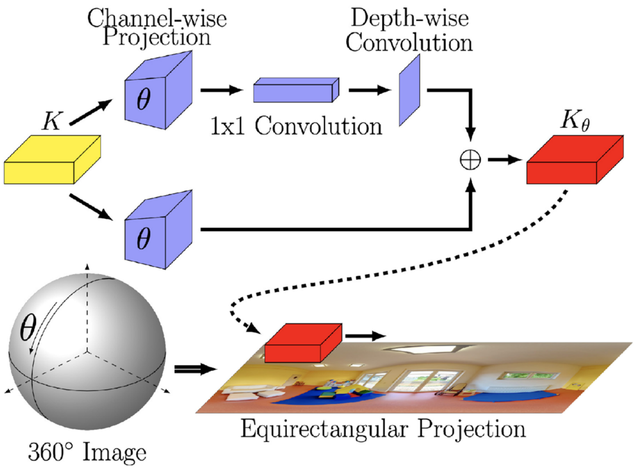


Fig. 8 Kernel transformer networks (KTN) for compact spherical convolution in 360° images. KTNs consist of row dependent channel-wise projections where the kernel is resized to the target size, and depth separable convolution blocks. A source kernel K and an angle θ are fed into the KTN to generate the output kernel K_θ for convolution with the image in equirectangular projection. Reproduced with permission from Ref. [105], © Institute of Electrical and Electronics Engineers 2019.

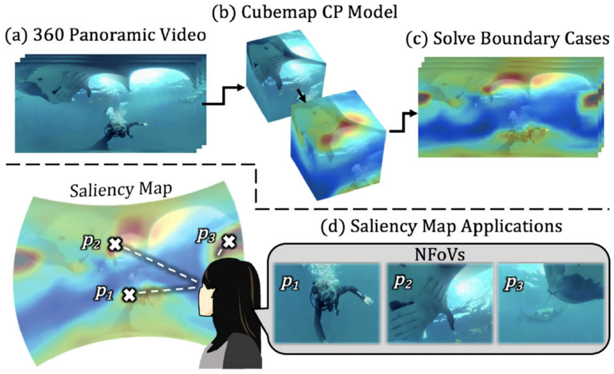


Fig. 9 Saliency detection in 360° video using cube padding. (a) A frame in equirectangular projection. (b) Cubemap projection with cube padding (CP) to mitigate distortion and cuts at image boundaries. High-quality saliency maps are predicted on the cubemap. (c) Predicted equirectangular saliency map. (d) Desirable NFOVs obtained from the high-quality saliency map. Reproduced with permission from Ref. [107], © Institute of Electrical and Electronics Engineers 2018.

Beyond the object level, neural networks have been proposed to determine scene-level geometry and layout from panoramas. LayoutNet learns to predict room layout from panorama images [111]. Assisted with Manhattan lines, the model decomposes the task into boundary map prediction and corner map prediction. The final predicted layout is reconstructed with Manhattan constraints. Later, HorizonNet was proposed to reconstruct 3D scene layout with a 1D representation that encodes the whole-room layout for a panoramic scene [112]. In this framework, recurrent neural networks are used to capture global information, and the positions of the floor–wall and ceiling–wall boundaries; the existence of wall–wall boundaries are encoded in the proposed 1D representation. At the same time, DuLa-Net was proposed with a two-branch network to analyze an input panorama in panorama-view and ceiling-view projections, and to fuse the learned features for the final 3D layout projection [113]. With the proposed dual-projection architecture, complex layout shapes beyond cuboids and L-shapes can be correctly reconstructed from a single panorama.

3.2 Cybersickness analysis

When a user experiences a virtual environment, cybersickness may occur, which causes symptoms similar to motion sickness. The most common symptoms include discomfort, headache, queasiness, nausea, vomiting, pallor, fatigue, drowsiness, disorientation, and apathy [115, 116]. Disequilibrium between organs of the human body and the visual information acquired by the eyes may cause cybersickness. Additionally, screen resolution, size of field of view, and latency play a role in cybersickness, too.

Crosstalk between the sensory and cognitive systems is the main factor in cybersickness. It is difficult to quantify cybersickness by measuring sensory and cognitive systems objectively. Traditional cybersickness evaluation methods usually collect answers using questionnaires but lack objectivity. Analyzing electroencephalogram (EEG) data is another possible approach. Traditional machine learning algorithms (e.g., support vector machines) have proved effective in obtaining highly accurate measurements (up to 95% accuracy [117]) for complex EEG data. Deep learning is also a powerful tool in helping to solve the problem, due to the maturity of emotion recognition and pattern analysis high performance analysis of EEG data. Jeong et al. [118] compared DNN and CNN in measuring cybersickness from EGG data. They also proposed a data preprocessing method for recommending an optimal weight set for EGG data to give the highest accuracy when learning from it with deep models. In addition, Lee et al. [119] proposed a method using 3D CNN to predict the degree of motion sickness when watching a 360° stereoscopic video. It takes the user's eye movements, motion velocity, and video depth as features. Wang et al. [120] proposed an LSTM model using dynamic information from normal-state posture signals to measure and quantize the amount of VR sickness in real time while allowing adaptive interactions in virtual environments. Kim et al. [114] developed an EEG-driven VR cybersickness level predict prediction model. It uses CNNs to encode a cognitive representation of the EEG spectrogram and an RNN-based model to predict cybersickness by learning from VR video sequences (see Fig. 10). Hu et al. [121] presented a computational model which can predict the discomfort level in terms of a given scene and camera trajectory. It is combined with a path planning method to optimize the camera trajectory to mitigate perceptual sickness.

4 Contactless interaction with deep learning

Aiming to improve the flexibility of content exploration, modern VR systems attempt to allow natural interactions between humans and the VR environment. In this survey, we review recent works on human pose estimation, hand gesture recognition, and gaze prediction, to which deep learning approaches have been applied.

4.1 Human pose estimation

Human pose estimation is a fundamental building block for methods translating natural body movements into functional actions in a VR environment. It focuses on estimation the locations of body parts and their connections [122]. Human pose estimation may involve 2D pose estimation or 3D pose estimation where human anatomical keypoints or parts are represented in 2D or 3D respectively.

There are two types of application scenario in 2D pose estimation: *single-person* pose estimation

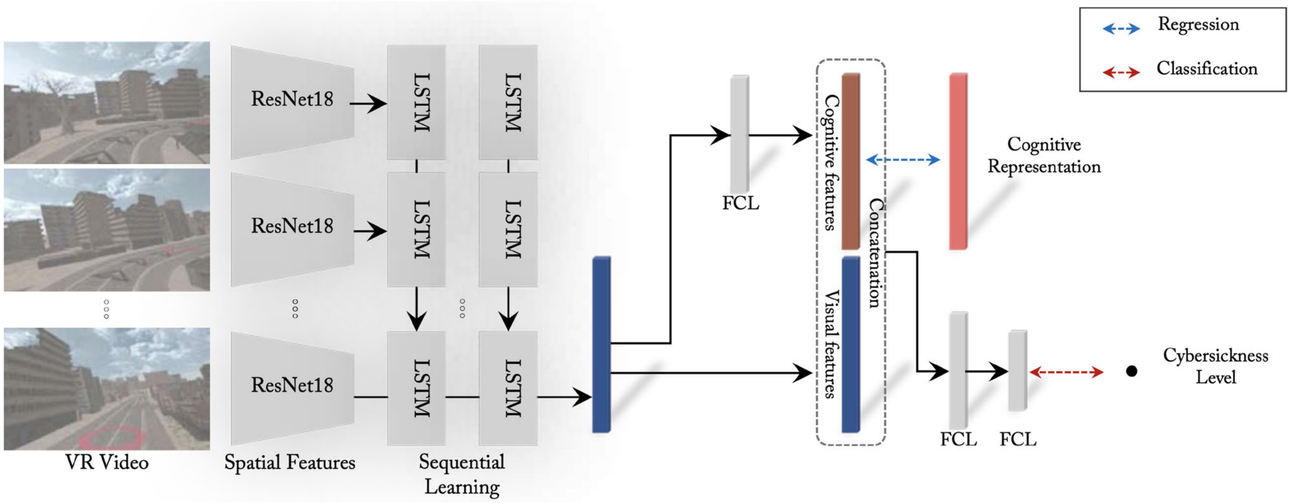


Fig. 10 Cybersickness learning predictor. Input VR video is fed into a CNN–RNN network to extract features which the fully connected layer represents as cognitive features. These are concatenated with visual features to predict cybersickness level. Reproduced with permission from Ref. [114], © Institute of Electrical and Electronics Engineers 2019.

and *multi-person* pose estimation. For 2D single-person pose estimation from a single image, Toshev and Szegedy [123] proposed DeepPose, the first deep neural network for human pose estimation. Newell et al. [124] proposed a "stacked hourglass" convolutional network architecture based on successive steps of pooling and upsampling for 2D human pose estimation. For multi-person pose estimation, Pishchulin et al. [125] proposed the DeepCut algorithm that jointly solves single-person and multi-person 2D articulated human pose estimation from a single image by inferring the number of people in the image, and then subsequently estimates their body poses using CNN-based part detectors. Cao et al. [126] presented a method to detect 2D multi-person poses in a single image. Their method uses non-parametric part affinity fields (PAFs) to learn to associate body parts with individuals. Fang et al. [127] proposed a regional multi-person pose estimation framework consisting of a symmetric spatial transformer network, a parametric pose non maximum-suppression module, and a pose-guided proposal generator, to handle cases with inaccurate and redundant detection results. Jin et al. [128] proposed a framework consisting of SpatialNet and TemporalNet for multi-person pose estimation and tracking. SpatialNet detects body parts and associates part-level data to each frame. TemporalNet tracks trajectories of human instances across consecutive frames.

For 3D pose estimation, Bogo et al. [129] proposed the first deep learning-based method, "Keep it SMPL". This method uses a single unconstrained image as input to automatically estimate the 3D pose of a human body as well as its 3D shape. Initially it uses DeepCut [125] to predict 2D body joint locations, and then fits a skinned multi-person linear model (SMPL) [130] from the 2D joints by minimizing errors between the projected 3D model joints and detected 2D joints. Mehta et al. [131] presented VNect, the first real-time model to capture the full 3D skeletal pose of a person from a single RGB image. Their method combines a CNN-based pose regressor with kinematic skeleton fitting; it offers stable, temporally consistent results. Tome et al. [132] proposed a CNN architecture to jointly solve 2D landmark detection and full 3D pose estimation from a single image. More recently, RepNet was proposed with an adversarial training process based on 2D re-projection [133] to tackle the overfitting problem. It was trained in a weakly supervised manner without 2D to 3D correspondences and camera parameters. Cheng et al. [134] proposed a method to handle occlusion by filtering out unreliable estimates of occluded keypoints when training their 2D and 3D temporal convolutional networks.

4.2 Hand gesture recognition

Hand gestures are postures or movements of the user's hands, and provide a common and natural way to interact with VR environments. Recognizing hand gestures efficiently and accurately is a critical component in contactless VR interaction.

Traditionally, it is challenging to accurately estimate hand poses, as the hand has many degrees of freedom, and fingers may occlude each other. It can also be cost-consuming to obtain highly accurate hand pose signals. With recent developments in deep learning, researchers have focused on solving the pose estimation problem by learning regression mapping functions between image appearance and hand pose representations. Oberweger et al. [135] compared different CNN architectures for 3D hand joint localization from a depth map, then introduced a constrained prior hand model and applied a joint-specific refinement stage to improve joint localization accuracy while reducing computational time. Zhou et al. [136] proposed a deep model that confirms the geometric validity of pose prediction by using a forward kinematics-based layer, which fully exploits prior knowledge in a generative model for hand geometry estimation. Pavllo et al. [137] proposed a real-time neural network that uses a motion capture system containing cameras and active markers to track hands and fingers. It copes well with occlusion and completely reconstructs hand posture. Chalasani et al. [138] proposed a method to solve the gesture recognition problem from a sequence of egocentric images. It consists of an ego-hand encoder network and an RNN; the encoder network finds ego-hand features and the RNN distinguishes temporally discriminative features. Ge et al. [139] proposed a graph CNN-based method to reconstruct the 3D shape and pose of a hand. A large-scale synthetic 3D hand shape and pose dataset, and a small-scale real-life hand dataset, were both introduced to train the network in a weakly-supervised manner.

4.3 Gaze prediction

Understanding where a user is looking, in a virtual environment, greatly benefits VR content creation from both commercial and technological perspectives. There are two main types of gaze prediction applications: one predicts gaze fixations in 360° image and video content, and the other predicts 3D gaze information from user input, e.g., iris contours or user facial images.

Much of the gaze prediction literature for 360° video focuses on VR content and history scanpath analysis to predict future gaze points. Deep learning methods can effectively extract features from video and predict gaze. Soccini [140] proposed a deep CNN model using image features and head movements as input to infer 2D coordinates of gaze points in the imaging plane. Xu et al. [141] fed images and corresponding saliency maps into a CNN and used LSTMs to encode the history scanpath as features. They then used these features to predict gaze displacements between successive gaze points.

Gaze prediction methods based on user input can be categorized as *model-based* and *appearance-based* approaches [142]. Model-based methods fit geometric eye models, detecting eye features using dedicated devices. However, the working distance between the user and the camera is limited. Appearance-based methods learn non-linear mappings between user input and corresponding gaze points. Deep learning-based methods can more effectively model the non-linear gaze prediction mapping than traditional methods. Lu et al. [143] proposed a model-driven method called synthetic iris appearance fitting (SIAF). It analyzes iris shape to predict 3D gaze direction. Cheng et al. [144] proposed an asymmetric regression-evaluation network architecture (ARE-Net) to estimate eye gaze. A sub-module of the asymmetric regression network (AR-Net) uses a new asymmetric strategy to estimate both eyes' 3D gaze directions, and a sub-module of the evaluation network (E-Net) evaluates the two eyes' performance to adjust the strategy adaptively during the optimization process. Furthermore, Cheng et al. [142] constructed a coarse-to-fine adaptive network named CA-Net. This architecture uses face images to estimate gaze direction, and then predicts corresponding residuals from eye images to refine gaze direction (see Fig. 11). Xiong et al. [145] provided mixed effects neural networks (MeNets) which adapts the mixed effects strategy from statistics to a DNN architecture for gaze estimation from eye images. It improves prediction accuracy by 10%–20% on many publicly available benchmarks.

5 Deep VR content manipulation

In recent visual media content manipulation research, we have experienced a paradigm shift from axiomatic modeling to data-driven modeling based on deep neural networks. The unprecedented performance of deep architectures has caused them to be widely used in content manipulation for VR applications.

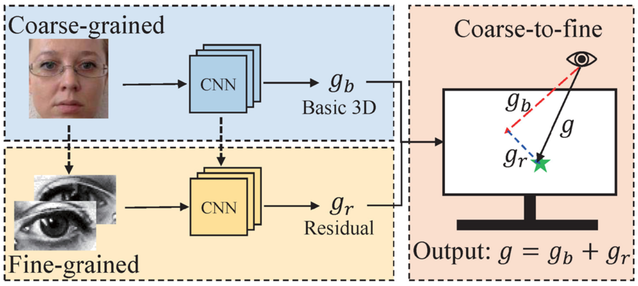


Fig. 11 Coarse-to-fine gaze prediction. Coarse-grained features are extracted from face images to estimate basic gaze direction g_b while fine-grained features are extracted from eye images to estimate gaze residual g_r. g_r is used to refine g_b to give the output gaze direction g. Reproduced with permission from Ref. [142], © Association for the Advancement of Artificial Intelligence 2020.

5.1 VR image and video editing

Generative adversarial networks (GANs) have been successfully applied to many image and video editing applications, e.g., image-to-image translation and inpainting. The image-to-image translation task converts the input from one domain (e.g., an edge map or segmentation map) to another (e.g., a photo-realistic image). Deep models such as pix2pix [146], CycleGAN [147], and StarGAN [148] were invented for such tasks. The image inpainting task, also known as image completion, aims to complete content inside missing regions of an image. Yu et al. [149] proposed a coarse-to-fine framework for inpainting large missing regions in an image using GANs with contextual attention. Li et al. [150] and Wu et al. [151] proposed deep generative models to restore occluded parts of an input portrait based on GANs. A survey of GANs in image synthesis and editing is provided in Ref. [152].

The outstanding improvements in these generative tasks has led to widespread interest in deep convolutional networks for VR image/video editing. In order to relieve viewers from frequently selecting view pilots while watching a 360° sports video, Hu et al. [153] proposed a deep learning-based agent, Deep 360 Pilot, for piloting through 360° sports videos automatically by using an RNN to choose the main subject to view and to regress a viewing angle shift for next move according to the chosen subject and previous viewing angle path: see Fig. 12. Lai et al. [154] presented a system to generate a normal field-of-view (NFOV) hyperlapse from a panoramic video. It uses fully convolutional networks to obtain initial semantic labels for each frame independently. To select valuable normal field-of-view segments from a 360° video and summarize it as a concise and informative subset of subshots spatially and temporally, Yu et al. [155] proposed a deep ranking model, composition view score. It produces a spherical score map for 360° video segments and uses a sliding window kernel to decide which view is suitable for highlighting. Lee et al. [156] proposed a past–future memory network with two external memories. The memories are used for storing previously chosen subshots and future candidate subshots' embeddings for temporal summarization of 360° videos.

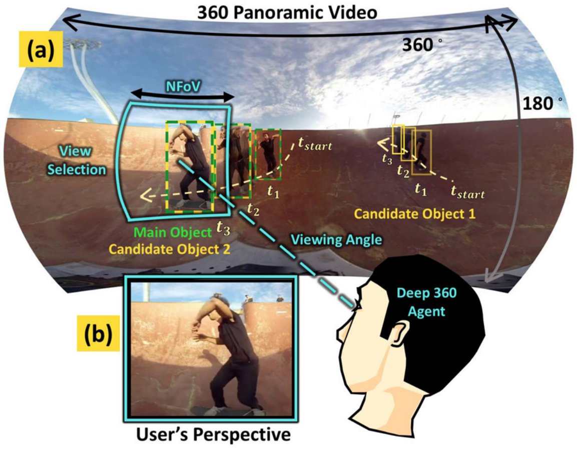


Fig. 12 Deep 360 Pilot for NFOV selection in 360° sports video. (a) Three overlapping panoramic frames sampled from a 360° skateboarding video with two skateboarders. The proposed "Deep 360 Pilot" selects a viewing angle, and NFOV center. (b) NFOV from a viewer's perspective. Reproduced with permission from Ref. [153], © Institute of Electrical and Electronics Engineers 2017.

5.2 Image and video enhancement with HMDs

Head-mounted displays (HMDs) block the real world from the viewer to provide an immersive experience of the virtual environment. However, when wearing such a device, the eye region of user's face is partially occluded. This partial face reduces immersion in teleconferencing, or other VR education and entertainment applications. Thies et al. [158] proposed the FaceVR system, a novel image-based method that performs real-time facial motion capture of an actor mounted with an HMD to enable VR teleconferencing based on self-reenactment. A new data-driven approach based on random ferns for real-time eye-gaze tracking is also presented in this work. Wang et al. [157] proposed an automatic face image completion solution using GANs. It learns to complete the HMD occluded region by referring to an occlusion-free image of the same person (see Fig. 13). Recently, real-time image-to-image

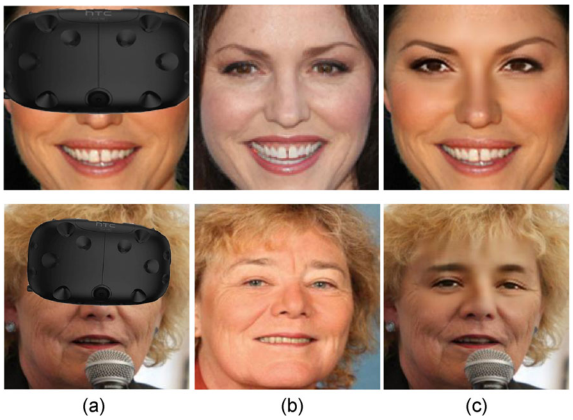


Fig. 13 Facial image completion for HMD occlusion using with a reference image. (a) Input photo with occluded region. (b) Reference image. (c) Image completion result. Reproduced with permission from Ref. [157], © Institute of Electrical and Electronics Engineers 2019.

translation of virtual content for HMD experience has been addressed. Nakano et al. [159] proposed a system using GAN-based image-to-image translation to change the appearance of one type of food into another dynamically and interactively. This problem successfully manipulates gustatory sensations of the user (see Fig. 14).

5.3 Foveated rendering

The spatial acuity of the human visual system varies across the retina, and is highest in the fovea, a small region with enhanced spatial acuity near the center of the retina [160]. Therefore, it is feasible to render realistic scenes with an acceptable computational cost in a spatially adaptive manner. Deep learning has been adopted to solve this problem very recently.

The earlier traditional method was proposed by Guenter et al. [161] whose system exploits foveation by separating the scene to be rendered into three nested eccentricity layers centered around the current gaze point. The inner layer is rendered at the highest resolution while peripheral layers are rendered at progressively lower resolution. Its rendering quality has been surpassed by the latest deep learning-based system. Recently, Kaplanyan et al. [162] presented a novel GAN-based method, DeepFovea, for foveated rendering and video compression. DeepFovea is able to reconstruct plausible peripheral video without noticeable quality degradation. It only requires a small fraction of color information provided by each frame. Since realistic rendering needs a huge amount of computation, DeepFovea can significantly reduce its workload.

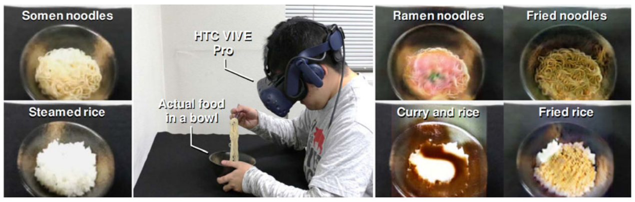


Fig. 14 GAN-based real-time food-to-food translation. Left: input food images. Center: user with an HMD experiencing gustatory manipulation. Right: translated food image examples. Reproduced with permission from Ref. [159], © Institute of Electrical and Electronics Engineers 2019.

5.4 Face reenactment

Face reenactment has been widely used in the film industry to animate virtual CG avatars in recent years. Since real-time markerless facial performance capture based on commodity sensors was invented [164], research using this field for VR has been increasing, and deep neural networks for automatic high-fidelity facial appearance generation are being investigated.

Face2Face [164] is a representative work for face reenactment. It is a real-time system which just takes monocular video as input and can manipulate the facial expression of a target video driven by a source actor. A novel image-based mouth synthesis approach is used to generate a realistic mouth interior and a sub-space deformation transfer technique inspired by Sumner and Popović [165] was also proposed in this paper. Although the method does not utilize deep learning, it produces plausible results and facilitates the development of face reenactment. Later, Olszewski et al. [166] introduced a system using deep convolutional networks to extract high-fidelity expressions in real time. It is able to produce realistic facial expressions and visual speech animation. Their model learns a direct mapping function that can transfer a high-dimensional image to lower-dimensional animation controls for a rigged 3D character. The highly accurate lip and eye motions enable applications like natural face-to-face conversations. Suwajanakorn et al. [167] presented a system for video synthesis from audio for the region around the mouth. It first converts input audio to a time-varying sparse mouth shape based on RNN and learns the mapping from raw audio features to mouth shape. It then synthesizes high-quality mouth texture at each time instant and finally composites generated photo-realistic mouth texture into the mouth region of the target video. When synthesizing, their method borrows the rest of the head and torso from other footage to make the head motions appear natural and consistent with the input speech. To keep structural consistency of faces in face reenactment, Wu et al. [168] proposed a method that first maps the source face onto a boundary latent space, then transforms the source boundary to adapt to the target boundary, and finally decodes the transformed boundary to generate the reenacted face. The Face Swapping GAN (FSGAN) was proposed to swap and reenact faces in a subject agnostic manner [169]. An RNN-based approach which adjusts for pose and expression variations was proposed, assisted by a face completion network and a face blending network to generate realistic face swapping results.

In portrait video synthesis, GAN-based models have shown their capability to encode essential features of the input data and to restore photo-realistic images. Geng et al. [170] introduced a method based on a warp-guided generative model (wg-GAN) for real-time photo-realistic facial animation that closely matches the expressions in source frames. The model first performs global 2D warps on the target portrait photo with a set of control points transferred from the motion parameters of the source portrait. Since the global structural movements of the facial expression can be well captured by 2D facial landmarks and preserved in 2D warps, it then generates a per-pixel displacement map by extracting the facial region and interpolating the 2D facial landmarks. Kim et al. [163] presented an approach based on a conditional GAN to achieve photo-realistic re-animation of portrait videos. The proposed network takes synthetic renderings of a parametric face model obtained by using a monocular face reconstruction with both source video and target video as input. It then automatically translates it into a full-frame photo-realistic output video with control of the target's head pose, facial expression, and eye motion (see Fig. 15). A space–time network architecture that takes short sequences of conditioned input frames of head and eye gaze is designed to keep temporal stability. Later, Kim et al. [171] presented a style-preserving visual dubbing approach based on recurrent GANs; it modifies facial expressions of a target actor to match the speech in a foreign language while maintaining the style of the target actor. Taking into consideration the idiosyncrasies and demeanor of different people, the network captures the spatio-temporal co-activation of facial expressions of unsynchronized source and target videos. They train their networks using cycle-consistency and mouth expression losses in an unsupervised manner. To generate the final results, they synthesize photo-realistic video frames using a layered neural face renderer.

6 Conclusions and future directions

The creation and exploration of virtual content in VR is a fundamental research topic, which serves and supports various applications utilizing an immersive virtual environment. This paper has reviewed representative deep learning works in VR content creation and exploration mostly from the last five years. We can see that deep neural networks are

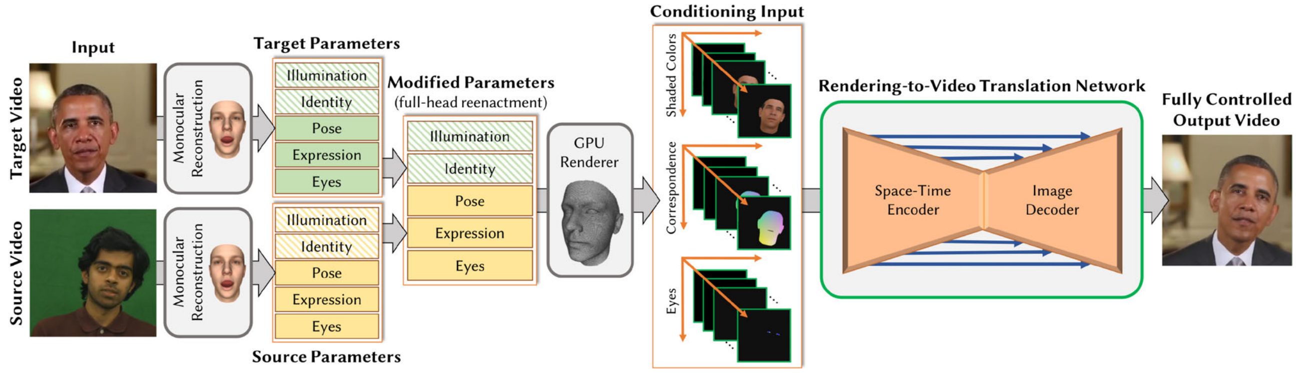


Fig. 15 Deep video portrait architecture for face reenactment. The method enables a source actor to control a target video portrait in terms of head pose, expression, etc. Left: a low-dimensional parametric representation of both videos is obtained using monocular face reconstruction. Center: head pose, expression, and eye gaze are transferred in parameter space. Right: rendering the photo-realistic video portrait of the target actor from input images. Obama video courtesy of the White House (public domain). Reproduced with permission from Ref. [163], © Association for Computing Machinery 2018.

most utilized in VR content creation using real world images and videos. They have been used in almost all stages of the reconstruction process, including raw image stitching, and 3D scene and human reconstruction from monocular, stereo, and RGB-D data. We have also reviewed deep neural networks for VR content analysis, which contribute to more natural interactions between users and their VR environment; they include both understanding of the VR environment and analysis of the user's body and actions. We wish to clarify that some research fields in VR content exploration heavily dependent on hardware have not been included in this survey, e.g., haptic technology for VR.

6.1 Future research directions

Leveraging deep learning techniques in VR constitutes a new research area which holds great potential for further visual media research. Here, we list some remaining challenges and open problems for future research.

6.1.1 *Motion parallax for dynamic 360° image and video*

Although several CNN-based view interpolation methods have been proposed [29, 51, 58–60], computing motion parallax for dynamic 360° video is still challenging [28, 172, 173]. Scene geometry estimation from casual photography is not accurate enough for high-quality view synthesis and rendering. Furthermore, dynamic objects in the scene can lead to complex occlusions. Combining scene semantics, object motion prediction, and view synthesis could be a promising solution.

6.1.2 *360° image and video synthesis with GANs*

Image-to-image translation methods using GANs are powerful tools for interactive face, body, object, and natural scene synthesis. Compared to normal field-of-view images, high-quality 360° images and videos are rarer and more difficult to capture and create. Fully automatic or interactive panoramic image and video synthesis with semantics and guided examples is required [21, 174]. Convolution kernels and neural network architectures considering wide fields-of-view are expected.

6.1.3 *Human-centric scene functionality analysis*

As discussed in Section 2, 3D scene functionality analysis is a new research topic [96, 97]. Functionality-aware shape modeling and scene analysis with comprehensive understanding of scene semantics is essential for VR applications. More importantly, human behavior is a key factor for indoor scene functionality processing. Combining human behavior and object-to-object interaction is of great importance. Due to the requirement of analyzing local and global contexts of the scene, graph neural networks [175] could be utilized to process large-scale complex data.

6.1.4 *Intelligent contactless interaction*

Human interaction is deeply involved in VR environments, and contactless user interaction has gained much attention. Current contactless interaction is based on gaze, gesture, or body pose signals. However, most current research considers such input signals separately. Gaze prediction in 3D environments considering the semantics of the scene has the potential to improve prediction accuracy. In addition, combining gaze and gesture recognition helps to better understand the user's intent. More complex interaction modes are expected in future VR content exploration.

Acknowledgements

The authors would like to thank the reviewers. This work was supported by the National Natural Science Foundation of China (Grant Nos. 61902012, 61932003). Fang-Lue Zhang was supported by a Victoria Early-Career Research Excellence Award.

References

[1] Oculus Rift. Available at https://www.oculus.com/.

[2] HTC Vive. Available at https://www.vive.com/cn/.

[3] Szeliski, R. Image alignment and stitching: A tutorial. *Foundations and Trends®in Computer Graphics and Vision* Vol. 2, No. 1, 1–104, 2006.

[4] Snavely, N.; Seitz, S. M.; Szeliski, R. Photo tourism: Exploring photo collections in 3D. *ACM Transactions on Graphics* Vol. 25, No. 3, 835–846, 2006.

[5] Huang, J.; Shi, X.; Liu, X.; Zhou, K.; Wei, L.-Y.; Teng, S.-H.; Bao, H.; Guo, B.; Shum, H.-Y. Subspace gradient domain mesh deformation. *ACM Transactions on Graphics* Vol. 25, No. 3, 1126–1134, 2006.

[6] Xu, K.; Chen, K.; Fu, H.; Sun, W.-L.; Hu, S.-M. Sketch2Scene: Sketch-based co-retrieval and co-placement of 3D models. *ACM Transactions on Graphics* Vol. 32, No. 4, Article No. 123, 2013.

[7] Nah, J. H.; Lim, Y.; Ki, S.; Shin, C. Z2 traversal order: An interleaving approach for VR stereo rendering on tile-based GPUs. *Computational Visual Media* Vol. 3, No. 4, 349–357, 2017.

[8] Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 779–788, 2016.

[9] He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN In: Proceedings of the IEEE International Conference on Computer Vision, 2961–2969, 2017.

[10] Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431–3440, 2015.

[11] Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; Torralba, A. Scene parsing through ADE20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 633–641, 2017.

[12] Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2881–2890, 2017.

[13] Xu, D.; Zhu, Y.; Choy, C. B.; Fei-Fei, L. Scene graph generation by iterative message passing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5410–5419, 2017.

[14] Dai, B.; Zhang, Y.; Lin, D. Detecting visual relationships with deep relational networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3076–3086, 2017.

[15] Gatys, L. A.; Ecker, A. S.; Bethge, M. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2414–2423, 2016.

[16] Johnson, J.; Alahi, A.; Li, F. F. Perceptual losses for real-time style transfer and super-resolution. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9906.* Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 694–711, 2016.

[17] Luan, F.; Paris, S.; Shechtman, E.; Bala, K. Deep photo style transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4990–4998, 2017.

[18] Isola, P.; Zhu, J.; Zhou, T.; Efros, A. A. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1125–1134, 2017.

[19] Zhu, J. Y.; Park, T.; Isola, P.; Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, 2242–2251, 2017.

[20] Choi, Y.; Choi, M.; Kim, M.; Ha, J. W.; Kim, S.; Choo, J. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8789–8797, 2018.

[21] Wang, M.; Yang, G.-Y.; Li, R.; Liang, R.-Z.; Zhang, S.-H.; Hall, P. M.; Hu, S.-M. Example-guided style-consistent image synthesis from semantic labeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1495–1504, 2019.

[22] Lai, W.-S.; Gallo, O.; Gu, J.; Sun, D.; Yang, M.-H.; Kantz, J. Video stitching for linear camera arrays. In: Proceedings of the British Machine Vision Conference, 2019.

[23] Rhee, T.; Petikam, L.; Allen, B.; Chalmers, A. MR360: Mixed reality rendering for 360° panoramic videos. *IEEE Transactions on Visualization and Computer Graphics* Vol. 23, No. 4, 1379–1388, 2017.

[24] Anderson, R.; Gallup, D.; Barron, J. T.; Kontkanen, J.; Snavely, N.; Hernández, C.; Agarwal, S.; Seitz, S. M. Jump: Virtual reality video. *ACM Transactions on Graphics* Vol. 35, No. 6, Article No. 198, 2016.

[25] Overbeck, R. S.; Erickson, D.; Evangelakos, D.; Pharr, M.; Debevec, P. A system for acquiring, processing, and rendering panoramic light field stills for virtual reality. *ACM Transactions on Graphics* Vol. 37, No. 6, Article No. 197, 2019.

[26] Schroers, C.; Bazin, J. C.; Sorkine-Hornung, A. An omnistereoscopic video pipeline for capture and display of real-world VR. *ACM Transactions on Graphics* Vol. 37, No. 3, Article No. 37, 2018.

[27] Matzen, K.; Cohen, M. F.; Evans, B.; Kopf, J.; Szeliski, R. Low-cost 360 stereo photography and video capture. *ACM Transactions on Graphics* Vol. 36, No. 4, Article No. 148, 2017.

[28] Bertel, T.; Campbell, N. D. F.; Richardt, C. MegaParallax: Casual 360° panoramas with motion parallax. *IEEE Transactions on Visualization and Computer Graphics* Vol. 25, No. 5, 1828–1835, 2019.

[29] Hedman, P.; Alsisan, S.; Szeliski, R.; Kopf, J. Casual 3D photography. *ACM Transactions on Graphics* Vol. 36, No. 6, Article No. 234, 2017.

[30] Hedman, P.; Kopf, J. Instant 3D photography. *ACM Transactions on Graphics* Vol. 37, No. 4, Article No. 101, 2018.

[31] Wei, L.; Zhong, Z.; Lang, C.; Yi, Z. A survey on image and video stitching. *Virtual Reality & Intelligent Hardware* Vol. 1, No. 1, 55–83, 2019.

[32] Brown, M.; Lowe, D. G. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision* Vol. 74, No. 1, 59–73, 2007.

[33] Zhang, Y.; Lai, Y. K.; Zhang, F. L. Content-preserving image stitching with piecewise rectangular boundary constraints. *IEEE Transactions on Visualization and Computer Graphics* DOI: 10.1109/TVCG.2020.2965097, 2020.

[34] Zhang, Y.; Lai, Y. K.; Zhang, F. L. Stereoscopic image stitching with rectangular boundaries. *The Visual Computer* Vol. 35, Nos. 6–8, 823–835, 2019.

[35] Zhu, Z.; Lu, J. M.; Wang, M. X.; Zhang, S. H.; Martin, R. R.; Liu, H. T.; et al. A comparative study of algorithms for realtime panoramic video blending. *IEEE Transactions on Image Processing* Vol. 27, No. 6, 2952–2965, 2018.

[36] Altwaijry, H.; Veit, A.; Belongie, S. Learning to detect and match keypoints with deep architectures. In: Proceedings of the British Machine Vision Conference, 2016.

[37] Balntas, V.; Lenc, K.; Vedaldi, A.; Mikolajczyk, K. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3852–3861, 2017.

[38] DeTone, D.; Malisiewicz, T.; Rabinovich, A. Deep image homography estimation. *arXiv preprint* arXiv:1606.03798, 2016.

[39] Nguyen, T.; Chen, S. W.; Shivakumar, S. S.; Taylor, C. J.; Kumar, V. Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robotics and Automation Letters* Vol. 3, No. 3, 2346–2353, 2018.

[40] Zhang, J.; Wang, C.; Liu, S.; Jia, L.; Wang, J.; Zhou, J. Content-aware unsupervised deep homography estimation. *arXiv preprint* arXiv:1909.05983, 2019.

[41] Ye, N.; Wang, C.; Liu, S.; Jia, L.; Wang, J.; Cui, Y. DeepMeshFlow: Content adaptive mesh deformation for robust image registration. *arXiv preprint* arXiv:1912.05131, 2019.

[42] Revaud, J.; Weinzaepfel, P.; Harchaoui, Z.; Schmid, C. DeepMatching: Hierarchical deformable dense matching. *International Journal of Computer Vision* Vol. 120, No. 3, 300–323, 2016.

[43] Weinzaepfel, P.; Revaud, J.; Harchaoui, Z.; Schmid, C. DeepFlow: Large displacement optical flow with deep matching. In: Proceedings of the IEEE International Conference on Computer Vision, 1385–1392, 2013.

[44] Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; Brox, T. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1647–1655, 2017.

[45] Tu, Z. G.; Xie, W.; Zhang, D. J.; Poppe, R.; Veltkamp, R. C.; Li, B. X.; Yuan, J. A survey of variational and CNN-based optical flow techniques. *Signal Processing: Image Communication* Vol. 72, 9–24, 2019.

[46] Lin, K. M.; Liu, S. C.; Cheong, L. F.; Zeng, B. Seamless video stitching from hand-held camera inputs. *Computer Graphics Forum* Vol. 35, No. 2, 479–487, 2016.

[47] Wang, M.; Shamir, A.; Yang, G. Y.; Lin, J. K.; Yang, G. W.; Lu, S. P.; Hu, S.-M. BiggerSelfie: Selfie video expansion with hand-held camera. *IEEE Transactions on Image Processing* Vol. 27, No. 12, 5854–5865, 2018.

[48] Jung, R.; Lee, A. S. J.; Ashtari, A.; Bazin, J. C. Deep360Up: A deep learning-based approach for automatic VR image upright adjustment. In: Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces, 1–8, 2019.

[49] Xiao, J. X.; Ehinger, K. A.; Oliva, A.; Torralba, A. Recognizing scene viewpoint using panoramic place representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2695–2702, 2012.

[50] Furukawa, Y.; Ponce, J. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 32, No. 8, 1362–1376, 2010.

[51] Goesele, M.; Snavely, N.; Curless, B.; Hoppe, H.; Seitz, S. M. Multi-view stereo for community photo collections. In: Proceedings of the IEEE 11th International Conference on Computer Vision, 1–8, 2007.

[52] Ji, M. Q.; Gall, J.; Zheng, H. T.; Liu, Y. B.; Fang, L. SurfaceNet: An end-to-end 3D neural network for multiview stereopsis. In: Proceedings of the IEEE International Conference on Computer Vision, 2326–2334, 2017.

[53] Ummenhofer B.; Brox, T. Global, dense multiscale reconstruction for a billion points. In: Proceedings of the IEEE International Conference on Computer Vision, 1341–1349, 2015.

[54] Jancosek, M.; Pajdla, T. Multi-view reconstruction preserving weakly-supported surfaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3121–3128, 2011.

[55] Xi, W. J.; Chen, X. J. Reconstructing piecewise planar scenes with multi-view regularization. *Computational Visual Media* Vol. 5, No. 4, 337–345, 2019.

[56] Knapitsch, A.; Park, J.; Zhou, Q.-Y.; Koltun, V. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics* Vol. 36, No. 4, Article No. 78, 2017.

[57] Buehler, C.; Bosse, M.; McMillan, L.; Gortler, S.; Cohen, M. Unstructured lumigraph rendering. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, 425–432, 2001.

[58] Flynn, J.; Neulander, I.; Philbin, J.; Snavely, N. Deep stereo: Learning to predict new views from the world's imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5515–5524, 2016.

[59] Zhou, T. H.; Tulsiani, S.; Sun, W. L.; Malik, J.; Efros, A. A. View synthesis by appearance flow. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9908*. Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 286–301, 2016.

[60] Flynn, J.; Broxton, M.; Debevec, P.; DuVall, M.; Fyffe, G.; Overbeck, R.; Snavely, N.; Tucker, R. DeepView: View synthesis with learned gradient descent. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2367–2376, 2019.

[61] Hedman, P.; Philip, J.; Price, T.; Frahm, J. M.; Drettakis, G.; Brostow, G. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics* Vol. 37, No. 6, Article No. 257, 2018.

[62] Trinidad, M. C.; Brualla, R. M.; Kainz, F.; Kontkanen, J. Multi-view image fusion. In: Proceedings of the IEEE International Conference on Computer Vision, 4101–4110, 2019.

[63] Introducing vr180 cameras. Available at https://vr.google.com/vr180/.

[64] Tewari, A.; Zollhofer, M.; Kim, H.; Garrido, P.; Bernard, F.; Perez, P.; Theobalt, C. MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision, 1274–1283, 2017.

[65] Zollhöfer, M.; Thies, J.; Garrido, P.; Bradley, D.; Beeler, T.; Pérez, P.; Stamminger, M.; Nießner, M.; Theobalt, C.. State of the art on monocular 3D face reconstruction, tracking, and applications. *Computer Graphics Forum* Vol. 37, No. 2, 523–550, 2018.

[66] Tran, A. T.; Hassner, T.; Masi, I.; Medioni, G. Regressing robust and discriminative 3D morphable models with a very deep neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5163–5172, 2017.

[67] Blanz, V.; Vetter, T. A morphable model for the synthesis of 3D faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, 187–194, 1999.

[68] Hu, L.; Saito, S.; Wei, L.; Nagano, K.; Seo, J.; Fursund, J.; Sadeghi, I.; Sun, C.; Chen, Y.-C.; Li, H. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics* Vol. 36, No. 6, Article No. 195, 2017.

[69] Jackson, A. S.; Bulat, A.; Argyriou, V.; Tzimiropoulos, G. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In: Proceedings of the IEEE International Conference on Computer Vision, 1031–1039, 2017.

[70] Richardson, E.; Sela, M.; Or-El, R.; Kimmel, R. Learning detailed face reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1259–1268, 2017.

[71] Dou, P.; Shah, S. K.; Kakadiaris, I. A. End-to-end 3D face reconstruction with deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5908–5917, 2017.

[72] Kim, H.; Zollhofer, M.; Tewari, A.; Thies, J.; Richardt, C.; Theobalt, C. InverseFaceNet: Deep monocular inverse face rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4625–4634, 2018.

[73] Tran, A. T.; Hassner, T.; Masi, I.; Paz, E.; Nirkin, Y.; Medioni, G. G. Extreme 3D face reconstruction: Seeing through occlusions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3935–3944, 2018.

[74] Gecer, B.; Ploumpis, S.; Kotsia, I.; Zafeiriou, S. GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1155–1164, 2019.

[75] Lombardi, S.; Saragih, J.; Simon, T.; Sheikh, Y. Deep appearance models for face rendering. *ACM Transactions on Graphics* Vol. 37, No. 4, Article No. 68, 2018.

[76] Dou, P. F.; Kakadiaris, I. A. Multi-view 3D face reconstruction with deep recurrent neural networks. *Image and Vision Computing* Vol. 80, 80–91, 2018.

[77] Wu, F.; Bao, L.; Chen, Y.; Ling, Y.; Song, Y.; Li, S.; Ngan, K. N.; Liu, W. MVF-Net: Multi-view 3D face morphable model regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 959–968, 2019.

[78] Cao, Y. P.; Liu, Z. N.; Kuang, Z. F.; Kobbelt, L.; Hu, S.M. Learning to reconstruct high-quality 3D shapes with cascaded fully convolutional networks. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11213.* Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 626–643, 2018.

[79] Huang, Z.; Li, T. Y.; Chen, W. K.; Zhao, Y. J.; Xing, J.; LeGendre, C.; Luo, L.; Ma, C.; Li, H. Deep volumetric video from very sparse multi-view performance capture. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11220.* Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 351–369, 2018.

[80] Zheng, Z.; Yu, T.; Wei, Y.; Dai, Q.; Liu, Y. DeepHuman: 3D human reconstruction from a single image. In: Proceedings of the IEEE International Conference on Computer Vision, 7739–7749, 2019.

[81] Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Li, H.; Kanazawa, A. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2304–2314, 2019.

[82] Gao, L.; Yang, J.; Qiao, Y. L.; Lai, Y. K.; Rosin, P. L.; Xu, W. W.; Xia, S. Automatic unpaired shape deformation transfer. *ACM Transactions on Graphics* Vol. 37, No. 6, Article No. 237, 2018.

[83] Tan, Q.; Gao, L.; Lai, Y.-K.; Yang, J.; Xia, S. Mesh-based autoencoders for localized deformation component analysis. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018.

[84] Gao, L.; Lai, Y. K.; Yang, J.; Zhang, L. X.; Xia, S. H.; Kobbelt, L. Sparse data driven mesh deformation. *IEEE Transactions on Visualization and Computer Graphics* DOI: 10.1109/TVCG.2019.2941200, 2019.

[85] Meng, H.-Y.; Gao, L.; Lai, Y.-K.; Manocha, D. VV-Net: Voxel VAE net with group convolutions for point cloud segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, 8500–8508, 2019.

[86] Wu, Z.; Wang, X.; Lin, D.; Lischinski, D.; Cohen-Or, D.; Huang, H. SAGNet: Structure-aware generative network for 3D-shape modeling. *ACM Transactions on Graphics* Vol. 38, No. 4, Article No. 91, 2019.

[87] Yin, K.; Chen, Z.; Huang, H.; Cohen-Or, D.; Zhang, H. LOGAN: Unpaired shape transform in latent overcomplete space. *ACM Transactions on Graphics* Vol. 38, No. 6, Article No. 198, 2019.

[88] Gao, L.; Yang, J.; Wu, T.; Yuan, Y.-J.; Fu, H.; Lai, Y.-K.; Zhang, H. SDM-NET: Deep generative network for structured deformable mesh. *ACM Transactions on Graphics* Vol. 38, No. 6, Article No. 243, 2019.

[89] Fu, Q.; Chen, X. W.; Wang, X. T.; Wen, S. J.; Zhou, B.; Fu, H. B. Adaptive synthesis of indoor scenes via activity-associated object relation graphs. *ACM Transactions on Graphics* Vol. 36, No. 6, Article No. 201, 2017.

[90] Wang, K.; Savva, M.; Chang, A. X.; Ritchie, D. Deep convolutional priors for indoor scene synthesis. *ACM Transactions on Graphics* Vol. 37, No. 4, Article No. 70, 2018.

[91] Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; Funkhouser, T. Semantic scene completion from a single depth image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1746–1754, 2017.

[92] Li, M.; Patil, A. G.; Xu, K.; Chaudhuri, S.; Khan, O.; Shamir, A.; Tu, C.; Chen, B.; Cohen-Or, D.; Zhang, H. Grains: Generative recursive autoencoders for indoor scenes. *ACM Transactions on Graphics* Vol. 38, No. 2, Article No. 12, 2019.

[93] Li, J.; Xu, K.; Chaudhuri, S.; Yumer, E.; Zhang, H.; Guibas, L. GRASS: Generative recursive autoencoders for shape structures. *ACM Transactions on Graphics* Vol. 36, No. 4, Article No. 52, 2017.

[94] Wu, W. M.; Fu, X. M.; Tang, R.; Wang, Y. H.; Qi, Y. H.; Liu, L. G. Data-driven interior plan generation for residential buildings. *ACM Transactions on Graphics* Vol. 38, No. 6, Article No. 234, 2019.

[95] Ritchie, D.; Wang, K.; Lin, Y.-A. Fast and flexible indoor scene synthesis via deep convolutional generative models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6182–6190, 2019.

[96] Zhao, X.; Hu, R. Z.; Liu, H. S.; Komura, T.; Yang, X. Y. Localization and completion for 3D object interactions. *IEEE Transactions on Visualization and Computer Graphics* DOI: 10.1109/TVCG.2019.2892454, 2019.

[97] Hu, R. Z.; Yan, Z. H.; Zhang, J. W.; van Kaick, O.; Shamir, A.; Zhang, H.; Huang, H. Predictive and generative neural networks for object functionality. *ACM Transactions on Graphics* Vol. 37, No. 4, Article No. 151, 2018.

[98] Yan, Z.; Hu, R.; Yan, X.; Chen, L.; Van Kaick, O.; Zhang, H.; Huang, H. RPM-Net: Recurrent prediction of motion and parts from point cloud. *ACM Transactions on Graphics* Vol. 38, No. 6, Article No. 240, 2019.

[99] Guérin, É.; Digne, J.; Galin, É.; Peytavie, A.; Wolf, C.; Benes, B.; Martinez, B. Interactive example-based terrain authoring with conditional generative adversarial networks. *ACM Transactions on Graphics* Vol. 36, No. 6, Article No. 228, 2017.

[100] Zhang, J.; Wang, C. B.; Li, C.; Qin, H. Example-based rapid generation of vegetation on terrain via CNN-based distribution learning. *The Visual Computer* Vol. 35, Nos. 6–8, 1181–1191, 2019.

[101] Su, Y.-C.; Grauman, K. Learning spherical convolution for fast features from 360 imagery. In: Proceedings of the Advances in Neural Information Processing Systems 30, 529–539, 2017.

[102] Zhang, Z. H.; Xu, Y. Y.; Yu, J. Y.; Gao, S. H. Saliency detection in 360° videos. In: Proceedings of the European Conference on Computer Vision, 488–503, 2018.

[103] Coors, B.; Condurache, A. P.; Geiger, A. SphereNet: Learning spherical representations for detection and classification in omnidirectional images. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11213.* Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 518–533, 2018.

[104] Li, J.; Su, J. M.; Xia, C. Q.; Tian, Y. H. Distortion-adaptive salient object detection in 360° omnidirectional images. *IEEE Journal of Selected Topics in Signal Processing* Vol. 14, No. 1, 38–48, 2020.

[105] Su Y.-C.; Grauman, K. Kernel transformer networks for compact spherical convolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 9442–9451, 2019.

[106] Monroy, R.; Lutz, S.; Chalasani, T.; Smolic, A. SalNet360: Saliency maps for omni-directional images with CNN. *Signal Processing: Image Communication* Vol. 69, 26–34, 2018.

[107] 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1420–1429, 2018.

[108] Yang, W.; Qian, Y.; Kämäräinen, J.-K.; Cricri, F.; Fan, L. Object detection in equirectangular panorama. In: Proceedings of the 24th International Conference on Pattern Recognition, 2190–2195, 2018.

[109] Redmon J.; Farhadi, A. YOLO9000: Better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7263–7271, 2017.

[110] Lee, Y.; Jeong, J.; Yun, J.; Cho, W.; Yoon, K.-J. SpherePHD: Applying CNNs on a spherical polyhedron representation of 360deg images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 9181–9189, 2019.

[111] Zou, C.; Colburn, A.; Shan, Q.; Hoiem, D. LayoutNet: Reconstructing the 3D room layout from a single RGB image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2051–2059, 2018.

[112] Sun, C.; Hsiao, C. W.; Sun, M.; Chen, H. T. HorizonNet: Learning room layout with 1D representation and pano stretch data augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1047–1056, 2019.

[113] Yang, S.-T.; Wang, F.-E.; Peng, C.-H.; Wonka, P.; Sun, M.; Chu, H.-K. DuLa-Net: A dual-projection network for estimating room layouts from a single RGB panorama. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3363–3372, 2019.

[114] Kim, J.; Kim, W.; Oh, H.; Lee, S.; Lee, S. A deep cybersickness predictor based on brain signal analysis for virtual reality contents. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 10580–10589, 2019.

[115] Kolasinski, E. M. Simulator sickness in virtual environments. Technical Report. Army Research Inst for the Behavioral and Social Sciences Alexandria VA, 1995.

[116] Wang, M.; Zhang, X. J.; Liang, J. B.; Zhang, S. H.; Martin, R. R. Comfort-driven disparity adjustment for stereoscopic video. *Computational Visual Media* Vol. 2, No. 1, 3–17, 2016.

[117] Yu, Y. H.; Lai, P. C.; Ko, L. W.; Chuang, C. H.; Kuo, B. C.; Lin, C. T. An EEG-based classification system of Passenger's motion sickness level by using feature extraction/selection technologies. In: Proceedings of the International Joint Conference on Neural Networks, 1–6, 2010.

[118] Jeong, D.; Yoo, S.; Yun, J. Cybersickness analysis with EEG using deep learning algorithms. In: Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces, 827–835, 2019.

[119] Lee, T. M.; Yoon, J. C.; Lee, I. K. Motion sickness prediction in stereoscopic videos using 3D convolutional neural networks. *IEEE Transactions on Visualization and Computer Graphics* Vol. 25, No. 5, 1919–1927, 2019.

[120] Wang, Y. Y.; Chardonnet, J. R.; Merienne, F. VR sickness prediction for navigation in immersive virtual environments using a deep long short term memory model. In: Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces, 1874–1881, 2019.

[121] Hu, P.; Sun, Q.; Didyk, P.; Wei, L. Y.; Kaufman, A. E. Reducing simulator sickness with perceptual camera control. *ACM Transactions on Graphics* Vol. 38, No. 6, Article No. 210, 2019.

[122] Gong, W. J.; Zhang, X. N.; Gonzàlez, J.; Sobral, A.; Bouwmans, T.; Tu, C. H.; Zahzah, E.-h. Human pose estimation from monocular images: A comprehensive survey. *Sensors* Vol. 16, No. 12, 1966, 2016.

[123] Toshev, A.; Szegedy, C. DeepPose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1653–1660, 2014.

[124] Newell, A.; Yang, K. Y.; Deng, J. Stacked hourglass networks for human pose estimation. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9912.* Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 483–499, 2016.

[125] Pishchulin, L.; Insafutdinov, E.; Tang, S. Y.; Andres, B.; Andriluka, M.; Gehler, P.; Schiele, B. DeepCut: Joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4929–4937, 2016.

[126] Cao, Z.; Simon, T.; Wei, S.-E.; Sheikh, Y. Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7291–7299, 2017.

[127] Fang, H.-S.; Xie, S.; Tai, Y.-W.; Lu, C. RMPE: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, 2334–2343, 2017.

[128] Jin, S.; Liu, W.; Ouyang, W.; Qian, C. Multi-person articulated tracking with spatial and temporal embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5664–5673, 2019.

[129] Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; Black, M. J. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9909.* Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 561–578, 2016.

[130] Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M. J. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics* Vol. 34, No. 6, Article No. 248, 2015.

[131] Mehta, D.; Sridhar, S.; Sotnychenko, O.; Rhodin, H.; Shafiei, M.; Seidel, H.-P.; Xu, W.; Casas, D.; Theobalt, C. VNect: Real-time 3D human pose estimation with a single RGB camera. *ACM Transactions on Graphics* Vol. 36, No. 4, Article No. 44, 2017.

[132] Tome, D.; Russell, C.; Agapito, L. Lifting from the deep: Convolutional 3D pose estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2500–2509, 2017.

[133] Wandt, B.; Rosenhahn, B. RepNet: Weakly supervised training of an adversarial reprojection network for 3D human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7782–7791, 2019.

[134] Cheng, Y.; Yang, B.; Wang, B.; Yan, W.; Tan, R. T. Occlusion-aware networks for 3D human pose estimation in video. In: Proceedings of the IEEE International Conference on Computer Vision, 723–732, 2019.

[135] Oberweger, M.; Wohlhart, P.; Lepetit, V. Hands deep in deep learning for hand pose estimation. *arXiv preprint* arXiv:1502.06807, 2015.

[136] Zhou, X.; Wan, Q.; Zhang, W.; Xue, X.; Wei, Y. Model-based deep hand pose estimation. *arXiv preprint* arXiv:1606.06854, 2016.

[137] Pavllo, D.; Porssut, T.; Herbelin, B.; Boulic, R. Real-time marker-based finger tracking with neural

networks. In: Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces, 651–652, 2018.

[138] Chalasani, T.; Ondrej, J.; Smolic, A. Egocentric gesture recognition for head-mounted AR devices. In: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality Adjunct, 109–114, 2018.

[139] Ge, L.; Ren, Z.; Li, Y.; Xue, Z.; Wang, Y.; Cai, J.; Yuan, J. 3D hand shape and pose estimation from a single RGB image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 10833–10842, 2019.

[140] Soccini, A. M. Gaze estimation based on head movements in virtual reality applications using deep learning. In: Proceedings of the IEEE Virtual Reality, 413–414, 2017.

[141] Xu, Y.; Dong, Y.; Wu, J.; Sun, Z.; Shi, Z.; Yu, J.; Gao, S. Gaze prediction in dynamic 360° immersive videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5333–5342, 2018.

[142] Cheng, Y.; Huang, S.; Wang, F.; Qian, C.; Lu, F. A coarse-to-fine adaptive network for appearance-based gaze estimation. *arXiv preprint* arXiv:2001.00187, 2020.

[143] Lu, F.; Gao, Y.; Chen, X. W. Estimating 3D gaze directions using unlabeled eye images via synthetic iris appearance fitting. *IEEE Transactions on Multimedia* Vol. 18, No. 9, 1772–1782, 2016.

[144] Cheng, Y. H.; Lu, F.; Zhang, X. C. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11218*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 105–121, 2018.

[145] Xiong, Y.; Kim, H. J.; Singh, V. Mixed effects neural networks (MeNets) with applications to gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7743–7752, 2019.

[146] Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A. A. Image-toimage translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1125–1134, 2017.

[147] Zhu, J. Y.; Park, T.; Isola, P.; Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, 2223–2232, 2017.

[148] Choi, Y.; Choi, M.; Kim, M.; Ha, J. W.; Kim, S.; Choo, J. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8789–8797, 2018.

[149] Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T. S. Generative image inpainting with contextual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5505–5514, 2018.

[150] Li, Y.; Liu, S.; Yang, J.; Yang, M.-H. Generative face completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3911–3919, 2017.

[151] Wu, X.; Li, R. L.; Zhang, F. L.; Liu, J. C.; Wang, J.; Shamir, A.; Hu, S.-M. Deep portrait image completion and extrapolation. *IEEE Transactions on Image Processing* Vol. 29, 2344–2355, 2020.

[152] Wu, X.; Xu, K.; Hall, P. A survey of image synthesis and editing with generative adversarial networks. *Tsinghua Science and Technology* Vol. 22, No. 6, 660–674, 2017.

[153] 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1396–1405, 2017.

[154] Lai, W. S.; Huang, Y. J.; Joshi, N.; Buehler, C.; Yang, M. H.; Kang, S. B. Semantic-driven generation of hyperlapse from 360 degree video. *IEEE Transactions on Visualization and Computer Graphics* Vol. 24, No. 9, 2610–2621, 2018.

[155] Yu, Y.; Lee, S.; Na, J.; Kang, J.; Kim, G. A deep ranking model for spatio-temporal highlight detection from a 360 video. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018.

[156] Lee, S.; Sung, J.; Yu, Y.; Kim, G. A memory network approach for story-based temporal summarization of 360° videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1410–1419, 2018.

[157] Wang, M.; Wen, X.; Hu, S.-M. Faithful face image completion for HMD occlusion removal. In: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality Adjunct, 251–256, 2019.

[158] Thies, J.; Zollhöfer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. FaceVR: Real-time gaze-aware facial

reenactment in virtual reality. *ACM Transactions on Graphics* Vol. 37, No. 2, Article No. 25, 2018.

[159] Nakano, K.; Horita, D.; Sakata, N.; Kiyokawa, K.; Yanai, K.; Narumi, T. DeepTaste: Augmented reality gustatory manipulation with GAN-based real-time food-to-food translation. In: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality, 212–223, 2019.

[160] Levoy, M.; Whitaker, R. Gaze-directed volume rendering. *ACM SIGGRAPH Computer Graphics* Vol. 24, No. 2, 217–223, 1990.

[161] Guenter, B.; Finch, M.; Drucker, S.; Tan, D.; Snyder, J. Foveated 3D graphics. *ACM Transactions on Graphics* Vol. 31, No. 6, Article No. 164, 2012.

[162] Kaplanyan, A. S.; Sochenov, A.; Leimkühler, T.; Okunev, M.; Goodall, T.; Rufo, G. DeepFovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Transactions on Graphics* Vol. 38, No. 6, Article No. 212, 2019.

[163] Kim, H.; Carrido, P.; Tewari, A.; Xu, W.; Thies, J.; Niessner, M.; Pérez, P.; Richardt, C.; Zollhöfer, M.; Theobalt, C. Deep video portraits. *ACM Transactions on Graphics* Vol. 37, No. 4, Article No. 163, 2018.

[164] Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; NieBner, M. Face2Face: Real-time face capture and reenactment of RGB videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2387–2395, 2016.

[165] Sumner, R. W.; Popović, J. Deformation transfer for triangle meshes. *ACM Transactions on Graphics* Vol. 23, No. 3, 399–405, 2004.

[166] Olszewski, K.; Lim, J. J.; Saito, S.; Li, H. High-fidelity facial and speech animation for VR HMDs. *ACM Transactions on Graphics* Vol. 35, No. 6, Article No. 221, 2016.

[167] Suwajanakorn, S.; Seitz, S. M.; Kemelmacher-Shlizerman I. Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics* Vol. 36, No. 4, Article No. 95, 2017.

[168] Wu, W.; Zhang, Y. X.; Li, C.; Qian, C.; Loy, C. C. ReenactGAN: Learning to reenact faces via boundary transfer. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11205.* Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 622–638, 2018.

[169] Nirkin, Y.; Keller, Y.; Hassner T. FSGAN: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE International Conference on Computer Vision, 7184–7193, 2019.

[170] Geng, J. H.; Shao, T. J.; Zheng, Y. Y.; Weng, Y. L.; Zhou, K. Warp-guided GANs for single-photo facial animation. *ACM Transactions on Graphics* Vol. 37, No. 6, Article No. 231, 2019.

[171] Kim, H.; Elgharib, M.; Zollhöfer, M.; Seidel, H. P.; Beeler, T.; Richardt, C.; Theobalt, C. Neural style-preserving visual dubbing. *ACM Transactions on Graphics* Vol. 38, No. 6, Article No. 178, 2019.

[172] Huang, J. W.; Chen, Z. L.; Ceylan, D.; Jin, H. L. 6-DOF VR videos with a single 360-camera. In: Proceedings of the IEEE Virtual Reality, 37–44, 2017.

[173] Serrano, A.; Kim, I.; Chen, Z. L.; DiVerdi, S.; Gutierrez, D.; Hertzmann, A.; Masia, B. Motion parallax for 360° RGBD video. *IEEE Transactions on Visualization and Computer Graphics* Vol. 25, No. 5, 1817–1827, 2019.

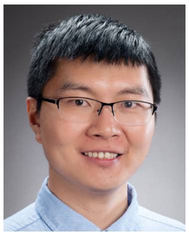
[174] Park, T.; Liu, M.-Y.; Wang, T.-C.; Zhu, J.-Y. Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2337–2346, 2019.

[175] Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S. A comprehensive survey on graph neural networks. *arXiv preprint* arXiv:1901.00596, 2019.

Miao Wang is an assistant professor at the State Key Laboratory of Virtual Reality Technology and Systems, Research Institute for Frontier Science, Beihang University, and Peng Cheng Laboratory, China. He received his Ph.D. degree from Tsinghua University in 2016. In 2013–2014, he visited the Visual Computing Group in Cardiff University as a joint Ph.D. student. In 2016–2018, he worked as a postdoc researcher at Tsinghua University. His research interests lie in virtual reality and computer graphics, with particular focus on content creation for virtual reality.

Xu-Quan Lyu is a master student with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, China. His research interests include virtual reality and augmented reality.

Yi-Jun Li is a Ph.D. student with the State Key Laboratory of Virtual Reality Technology and System, School of Computer Science and Engineering, Beihang University, China. His research interests are virtual reality, with particular focus on virtual scene navigation and 360° image and video processing.



Fang-Lue Zhang is currently a lecturer at Victoria University of Wellington, New Zealand. He received his bachelor degree from Zhejiang University in 2009, and his doctoral degree from Tsinghua University in 2015. His research interests include image and video editing, computer vision, and computer graphics. He is a member of IEEE and ACM. He received a Victoria Early-Career Research Excellence Award in 2019.