Capture, Reconstruction, and Representation of the Visual Real World for Virtual Reality

Christian Richardt[1] [0000−0001−6716−9845], James Tompkin[2] [0000−0003−2218−2899], and Gordon Wetzstein[3] [0000−0002−9243−6885]

[1] University of Bath
christian@richardt.name
https://richardt.name/
[2] Brown University
james_tompkin@brown.edu
http://jamestompkin.com/
[3] Stanford University
gordon.wetzstein@stanford.edu
http://www.computationalimaging.org/

Abstract. We provide an overview of the concerns, current practice, and limitations for capturing, reconstructing, and representing the real world visually within virtual reality. Given that our goals are to capture, transmit, and depict complex real-world phenomena to humans, these challenges cover the opto-electromechancial, computational, informational, and perceptual fields. Practically producing a system for real-world VR capture requires navigating a complex design space and pushing the state of the art in each of these areas. As such, we outline several promising directions for future work to improve the quality and flexibility of real-world VR capture systems.

Keywords: Cameras · Reconstruction · Representation · Image-Based Rendering · Novel-view synthesis · Virtual reality

1 Introduction

One of the high-level goals of virtual reality is to reproduce how the real world *looks* in a way which is indistinguishable from reality. Achieving this arguably-quixotic goal requires us to solve significant problems across capture, reconstruction, and representation, and raises many questions: “Which camera system should we use to sample enough of the environment for our application?”; “How should we model the world and which algorithm should we use to recover these models?”; “How should we store the data for easy compression and transmission?”, and “How can we achieve simple and high-quality rendering for human viewing?”. Solving any one of these problems is a challenge, and this challenge is exacerbated by the interplay between the questions. This provides us with a complex design space to navigate if we wish to build practical and high-quality systems for real-world VR reproduction.

Yet, significant progress has been made over the past 5 years in our ability to capture and display the visual properties of the real world, driven most recently by the need to provide content for low-cost VR headsets. Many compelling applications are now within reach: the broad area of telepresence, e.g., for VR communications; for remote operation, e.g., medical robotics; for cultural heritage and virtual tourism; and for storytelling via documentaries, movies, and games. We wish the tools for these applications to be simple in all stages of authorship, and for the applications to be comfortable and easy to use, requiring only novice intuition (the "can my grandparents use this?" test).

So, broadly, what issues concern us when we wish to capture the world in a visually indistinguishably way?

Objects, Scenes, and Subjects. Typically *what* we wish to capture helps determine *how* we should capture it for later analysis or virtual reality presentation. Capturing a single *object* in a studio has long been accomplished with so-called *outside-in* multi-camera systems: cameras are placed to encircle or ensphere an object and allow its multi-view capture. Generally, the more cameras we have, the higher the quality of reproduction. Further, such controlled environment conditions allow higher-quality capture than unconstrained settings, e.g., outdoors.

To capture *scenes*, we use an *inside-out* camera system in which multiple cameras face the world. These may be arranged into a circle or sphere to capture 360° for immersive VR, though planar configurations to densely capture narrower fields of view are also common (often called *light field cameras*). The distance between the cameras—or baseline—determines for how near and how far away we can reconstruct the geometry of the scene. Further, VR capture typically predicates that the camera and its paraphernalia are not visible in the scene, which informs the capture scenario design.

One special class of object exists for which much specific work has been directed: people. As social creatures, we wish to represent ourselves realistically, especially as many potential applications are driven by social interaction. Often, these methods use databases of human shape and appearance to create efficient and high-quality representations.

Photons, Rays, and Waves. Given an intended capture scenario, we next wish to maximize our capture fidelity. Our lens systems play a significant part in overall quality, and typically the camera configuration and lens systems are co-designed. At the camera sensor, output fidelity spans three major axes: spatial resolution, temporal resolution, and spectral resolution (i.e., color). Each of these must be tempered by our ability to store, process, and eventually transmit these data. Given that our output is to a human observer, then there is an eventual limit beyond which no additional captured information is perceived (which, arguably, we are fast approaching [78]).

Geometry and Appearance. The raw output from a multi-camera system must be reconstructed into a representation which is comfortable and easy to view. For video, existing real-world capture for VR is typically monoscopic and 360°, which requires stitching multiple camera images via a spherical proxy geometry. Current state-of-the-art systems produce stereoscopic 360° images, which have complex disparity challenges. Both of

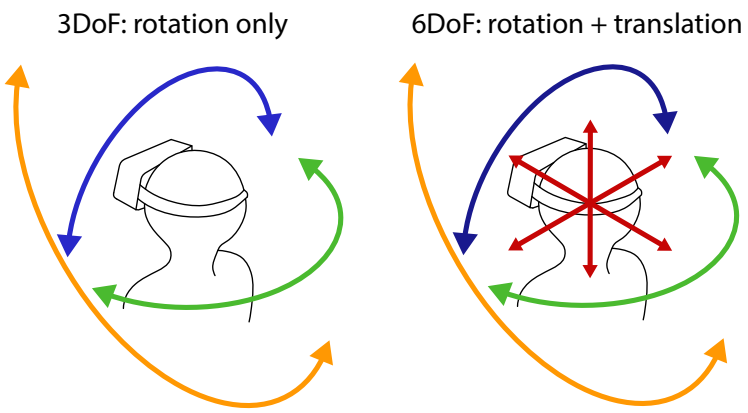

Fig. 1: VR systems provide either three or six degrees of freedom (DoF) for head motion. **Left:** '3DoF' lets a user look around the virtual world from a fixed point. **Right:** '6DoF' lets the user move in the virtual world like in the real world.

these are so-called '3DoF' or three-degrees-of-freedom representations describing the three rotation angles available within a VR headset. This is insufficient to represent human motion within the headset as humans can also translate along three axes in free space (Figure 1).

As such, one near-term goal of VR production is '6DoF', which allows realistic response to human head and body motions and solves the disparity challenges of the stereo 360° format by allowing stereo rendering in any viewing direction. The range of 6DoF movement available is a key concern: the larger the 'headbox' required, the larger the camera baseline must be to accommodate the eventual range of user motion. Sparser sampling of the scene requires us to reconstruct more sophisticated scene models to 'fill in the gaps' during rendering. Thus, depth and geometry reconstruction become critical pieces of the processing, transmission, and render systems. Likewise, complex scene object which display view-dependent appearance effects, such as shiny or translucent objects, make this reconstruction and depiction problem harder.

Many representations exist, including simple proxy geometry, depth, layered images, voxels, point clouds, signed distance fields, and textured 3D geometry. Each has a complementary rendering system, e.g., image-based rendering or ray casting. Each also has different compression and storage methods for efficient transmission. Further, state-of-the-art 'neural rendering' learned representations now also exist, with a 'neural' version of each classic representation.

Humans as Creators. Producing VR content requires the ability to *edit* the captured material. These operations could be color matching between cameras, editing scene content to remove unwanted artifacts, or editing the perceptual result such as adjusting disparity for comfort. As such, any reconstructed representation must be malleable to the editing tasks. More sophisticated productions may wish to edit content by integrating captured and purely virtual content, which requires a more sophisticated recon-

struction such as to represent real-world occlusions and integrate object lighting with illumination capture.

Humans as Consumers, Visual Expectation, and Avoiding the Visual Gap. Human perception has limitations which may be exploited. Many compelling applications—and even many *users*—may not need flawless reproduction, and the positive impact of, say, stronger personal connection through immersive telecommunications is likely to outweigh any failures in subtle appearance reproduction. Further, 'indistinguishable from reality' is different from 'photorealistic': we typically know when we are looking at camera-captured media, yet, this is sufficient for much of our storytelling. Likewise, 'perceptually realistic' is different from 'indistinguishable from reality' as it lightens the burden on scrutiny. Some 'non-photorealistic' depictions may help us *avoid entirely* complex challenges of fidelity and representation.

That said, VR is still a new and hopeful technology, and current limitations which are easy to overlook at this nascent stage—especially to technologists developing these techniques—may be more significant barriers to adoption. VR sickness is one such issue; here, 6DoF reconstruction and rendering holds promise to significantly reduce its effects and make VR more accessible.

2 Current Practice

Capturing the real world for rendering in virtual reality [59,85] is fundamentally about creating novel views of a scene given only a sparsely sampled set of images [7,13,21,30,55,92]. These techniques are closely related to 3D reconstruction [16] and image-based rendering (IBR) [20,93], and have many important applications in VR, including telepresence and digitizing avatars; capturing faces, hands, or whole body performances; and capturing cinematic experiences with dedicated VR camera rigs. Novel-view synthesis and image-based rendering are active and long-running fields of research that have produced a large variety of techniques and systems working towards the goal of capturing the real world in all its visual fidelity. Many of the proposed systems share a similar high-level structure, which is embodied by the **VR Capture Pipeline**:

Capture → Reconstruction → Representation → Compression → Rendering

In this section, we will look at each stage of this pipeline and provide an overview of the range of VR capture techniques and their trade-offs. For any particular approach or system, the most important design choice is the data representation to be used, as this constrains many of the other pipeline stages. In particular, the choice constrains reconstruction, compression, and rendering.

2.1 Capture

Most virtual reality capture approaches rely on one or more color cameras to capture the visual appearance and dynamics of a scene (see examples in Figure 2). Sometimes, special cameras are used, such as RGBD cameras which capture depth maps in addition to color footage, or special attachments like mirrors.

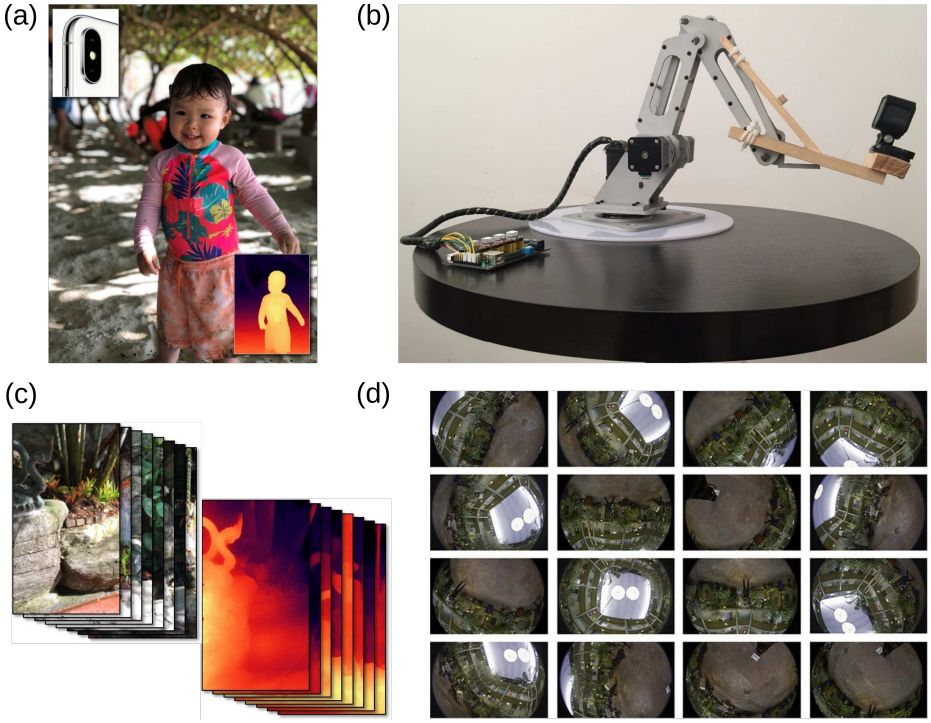

Fig. 2: Visual overview of capture approaches: (a) one static (RGBD) camera, (b) one moving camera, (c) one moving RGBD camera, and (d) a multi-camera capture with 16 cameras. Figures reproduced from: (a) Kopf *et al*. 2019 [49], (b) Luo *et al*. 2018 [58], (c) Hedman *et al*. 2018 [33], and (d) Parra Pozo *et al*. 2019 [78].

One Static Camera can capture a partial view of a larger scene, typically with perspective lenses. The content captured in this fashion can still be compelling, as demonstrated by Facebook's 3D photos [49], which are captured by dual-lens cameras on commodity mobile phones to provide depth in addition to color. However, wider views require wider camera optics, such as fisheye lenses ($> 90°$) or catadioptic systems [1] for omnidirectional video.

One Moving Camera can capture more of a static scene by sweeping over it across time. Traditional panorama stitching approaches [6,101] assume a camera that rotates around its optical center, so that it captures all light rays converging at a single point in space—the center of the panorama. By translating the camera in space, even more light rays can be captured, for instance for omnidirectional stereo [5,79,86], layered depth panoramas [123], or 3D photography [32]. More elaborate setups have a camera moving along the surface of a plane or sphere to capture different portions of a light field [58,63,75].

One Moving RGBD Camera makes it easier to reconstruct the geometry of the scene from the captured depth maps. A pioneer in this category is the KinectFusion approach [69], which reconstructs a global truncated signed distance field (TSDF) representation of a scene from registered input depth maps alone. There are many more recent variants that improve on the scale and robustness of this kind of scene reconstruction [17,71,114]. Alternatively, Instant 3D Photography [33] aligns multiple RGBD images captured with a dual-lens camera into a consistent textured 3D panoramic surface.

Multi-Camera Rigs are required for video capture and to capture multiple viewing directions simultaneously. Consumer 360° cameras are now commercially available as commodity devices that stitch two or more video streams into a single 360° video [53,81,111]. Stereo cameras capture two viewpoints side by side, and their baseline can be magnified in post production [124]. Multiple viewpoints can also be interpolated and manipulated in a post-process after video capture [56]. A ring of video cameras captures sufficient information for compelling omnidirectional stereo video [3,9,87], while a rotating camera rig can even capture live omnidirectional stereo video [48]. The Facebook Manifold camera [78] has 16 cinema cameras in a large sphere configuration to evenly capture views in all directions. Light fields [25,30,55] are based on a dense sampling of viewpoints, which requires many co-located cameras. A different camera setup distributes cameras on a dome or around a capture volume, for example to capture objects and people in a light stage [19] or as volumetric video [15].

2.2 Reconstruction

Reconstruction interprets and combines the information contained in the captured imagery to create a unified model. The first step is often camera calibration and structure from motion, i.e., characterizing the imaging devices used, including their lens distortion, and determining which views of a scene they captured. Multiple structure-from-motion implementations are publicly available, including Bundler [96], VisualSfM [117], AliceVision [41,65], MVE [27], Theia [99] and COLMAP [88], with the latter currently enjoying the widest use. However, general-purpose structure-from-motion tools do not perform well for the kind of inside-out capture commonly used for environment capture [5,32]. This has led to the development of specifically tailored structure-from-motion solutions that assume camera motion on a spherical surface [100,109], which is a good match for handheld [5,32,86] or spherical [58,75] capture approaches. One of the outputs of structure from motion is also a sparse 3D point cloud of feature points in the scene, which can be useful for image alignment [53] or view warping [36].

Once the viewpoints are reconstructed, the next step is generally to combine all the captured information into a single model of the scene. In classical panorama stitching, this is achieved by aligning and blending the individual input views on a spherical or cylindrical image surface [6,101]. While still panoramas can hide alignment artifact to some degree using clever stitching or blending approaches [32,121,122], this becomes much harder for panoramic videos, as the visual content, and hence any artifacts, keep changing over time. To address this, the stitching needs to vary over time in accordance with the scene [53,81,111]. To achieve more complex projections, such as the multi-perspective omnidirectional stereo (ODS) projection [39,79], requires dense correspon-

dence between input views so that intermediate views can be synthesized [3,9,86,87]. Most approaches use optical flow for this purpose, as it provides useful flexibility in case of calibration errors or scene motion.

The reconstruction of 3D geometry goes beyond the purely image-based approaches discussed before by recovering the 3D structure of a scene or object. Most approaches start by estimating per-view depth maps using multi-view stereo (MVS) techniques [28,37,89,90] or deep learning [33,72,91], unless depth maps are directly available from RGBD cameras. In theory, these per-view depth maps can be integrated into a global geometry model of the scene [12,32,91] if the camera poses and depth maps are estimated sufficiently accurately. Approaches such as KinectFusion [69] and BundleFusion [17] integrate noisy depth maps over time to improve the accuracy of the surface reconstruction. Having a large number of views also leads to a cleaner geometry reconstruction [15]. Hedman *et al.* [33] introduce a locally varying depth map alignment step to integrate differently normalized depth maps from mobile phones or neural networks into a globally consistent depth map. However, because of calibration and depth estimation errors, better view synthesis results can often be obtained with per-view geometry [11,35,75] that is smoothly blended across the synthesized novel view.

2.3 Representation

Over the years, various approaches have been proposed for representing captured scenes or objects. View synthesis techniques can be classified by how heavily they rely on the input image data vs. proxy geometry in their representation. For example, light field rendering [55] represents one extreme that does not use any geometry at all, but that requires densely sampled input views, whereas conventional 3D rendering with polygon meshes and textures is the other extreme in requiring detailed geometry but few input images (i.e., the textures) for the rendering. Geometric representations of a scene can be either modeled or estimated from the input images, for example using classical 3D computer vision pipelines [31,102]. To provide an overview, Shum *et al.* [92,93] organized representations along a continuum according to how much geometry they use:

- *No geometry* refers to purely image-based approaches, such as panoramas or 360° video.
- *Implicit geometry* comprises approaches using posed images and/or relying on 2D image correspondences, such as optical flow.
- And *explicit geometry* includes textured meshes or point clouds with actual 3D geometry.

Figure 3 contains an updated version of Shum *et al.*'s continuum of representations and Figure 4 illustrates examples from across this continuum.

There is no universally best representation—all have their advantages and disadvantages and provide different trade-offs. There is also often no hard boundary between representations, so there is some overlap; hybrids that combine multiple representations are also possible. In the limit, i.e., with infinite resolution, the representations are theoretically interchangeable. However, any conversion always requires resampling, which

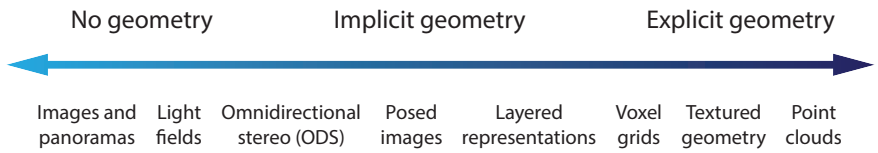

Fig. 3: Updated continuum of image-based rendering representations, inspired by Shum *et al*. [92,93]. Please see the discussion in Section 2.3 for details.

is usually a lossy process that reduces overall fidelity. There are usually also practical limits: for example, the physical size of cameras which limits the maximum camera density that is achievable in practice.

Images and Panoramas provide the most basic snapshot of what a scene or object looked like. They represent a photographic likeness that captures visual appearance of a scene or object from a single point of view with a fixed field of view. Panoramas [6,101] and 360° videos [53,81] capture a wide or even complete field of view. Images and panoramas enjoy great popularity as they are easy to capture with modern mobile phones and consumer cameras, and are straightforward to share. However, their main limitation is that they only provide information for a single point of view (i.e., only 3DoF) and no depth perception, and thus do not support any translational change of viewpoint.

Light Fields represent a dense spatio-angular sampling of a scene [55], generally using a regular 2D grid of camera viewpoints. More general camera configurations are supported by the Lumigraph [30], a closely related variant of light fields. As the comprehensive coverage of an object in a scene is challenging to obtain in practice, Davis *et al*. [18] proposed a guidance approach that helps users in capturing missing viewpoints. Videos captured with a moving camera can also be considered to be a densely sampled light field along the camera path, which can be exploited for particularly accurate scene reconstruction [45,120].

Omnidirectional Stereo (ODS) is a multi-perspective, circular projection [39,79] that has become a popular medium for stereoscopic and 360° VR photos and videos [3,9,86,87]. ODS encodes two panoramic views—one for the left eye and one for the right eye. This has the advantage that there is binocular disparity—and hence the perception of depth—in all viewing directions along the equator, though distortion exists away from the equator (Figure 5). The format is an excellent fit for existing video processing, compression, and transmission pipelines, as both views are encoded in a single top-bottom configuration.

Posed Images have known camera geometry (camera position and orientation) in addition to the image data. This enables scene reconstruction in the form of point clouds using multi-view stereo. Even sparse point clouds are sufficient for providing a compelling overview of community photo collections as demonstrated by Snavely *et al*.'s

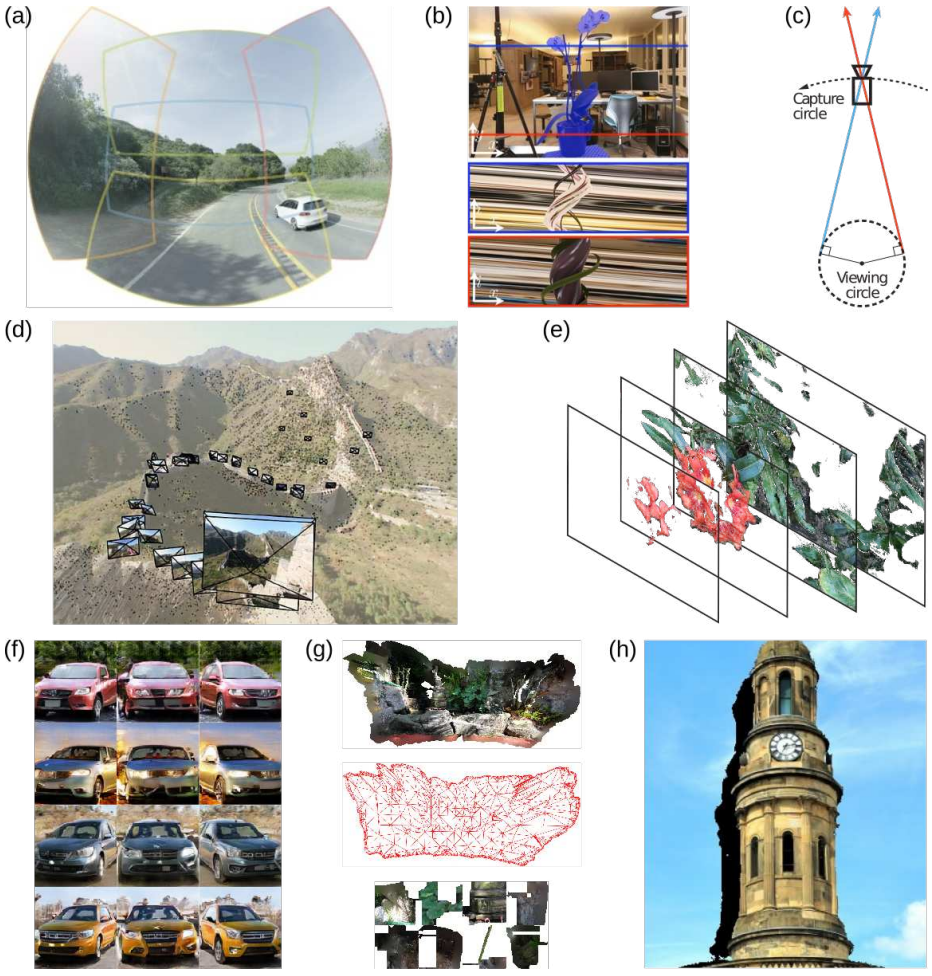

Fig. 4: Visual examples that illustrate the range of image-based rendering representations: (a) panoramas [81], (b) light fields [120], (c) omnidirectional stereo, (d) posed images [96], (e) layered representations such as multiplane images [63], (f) voxel grids with deep features [70], (g) textured geometry [33], and (h) point clouds [72].

PhotoTourism work [96]. Novel views can be interpolated from existing ones by establishing correspondences between adjacent viewpoints. In practice, optical flow is often used for flow-based blending [3,5,58,86,87], which significantly reduces blurry ghosting artifacts and produces results with high visual fidelity.

Layered Representations consist of multiple semi-transparent layers that encapsulate the appearance of a scene or object without any explicit geometry. The underlying core idea goes back to Disney's multiplane camera (1937), in which multiple transparent cel sheets are positioned at different depths from the camera. This allows each cel sheet

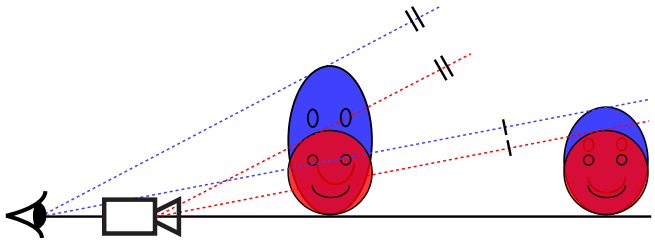

Fig. 5: Omnidirectional stereo introduces vertical distortion as cameras lie on a larger circle than the viewing circle. The red faces, as seen by the camera, appear vertically stretched (blue faces) when rendered using parallel rays for a viewpoint behind the camera. Figure adapted from Anderson *et al*. [3].

to be moved independently and creates the effect of motion parallax over time. Early approaches by Wetzstein *et al*. computed layered representations using custom-tailored optimization frameworks [112,113]. Recently, advances in deep learning have revived and accelerated progress in the reconstruction of so-called multiplane images (MPIs) [25,63,98,124].

Voxel Grids can represent regularly sampled occupancy ('filled or empty'), color, opacity, or distance (e.g., truncated signed distance fields [69]) to enable novel-view synthesis. Managing memory as a resource with voxel grids is critical given their n^3 nature, and octree-based voxel grids are possible. New voxel grids storing deep features [57,70,94,107] aim to enable novel-view synthesis at a higher quality but with a lower memory use. We discuss this emerging work on neural scene representations in more detail in Section 3.

Textured Geometry makes it easy to render novel views in real time with existing 3D graphics pipelines, even on mobile devices. Mesh geometry is particularly good at modeling hard occlusion boundaries, but it needs to be reconstructed accurately from usually noisy depth maps. For the highest quality depth maps, many observations from different viewpoints need to be combined, for example for volumetric video [15] or Google's light fields [75]. One consumer-facing example are Facebook's 3D Photos [49], which are based on an image and lower-resolution depth map from an off-the-shelf mobile phone. The final 3D photo can be looked at from different directions by tilting the phone. Several approaches separate foreground and background objects in a scene into multiple textured layers [32,33,91,123], to preserve clean occlusion boundaries. This generally requires some kind of inpainting to fill the areas behind foreground objects. In the real world, the appearance of objects also often depends on the viewing direction, e.g., when objects are shiny. This effect can be modeled using surface light fields [115] or view-dependent blending [34]. In general, modeling and editing favors geometric approaches, as there are better software tools available for textured meshes than other representations.

Point Clouds represent a scene as an unordered collection of points, which may or may not have colors and/or surface normals. They are readily obtained from structure-from-motion and multi-view stereo tools, RGBD images [72], or Lidar scans of a scenes. However, they are inherently sparse, tend to be noisy and non-uniformly distributed in reconstruction space (rather than camera space), and contain gaps that make them impractical for rendering high-quality novel views (although this is slowly changing thanks to neural re-rendering [62]). Nevertheless, they are often a useful intermediate representation or debugging tool.

2.4 Compression

Raw scene representations can become very large (hundreds of gigabytes). This can make them difficult to store given limited space on disk or in memory, to transmit over networks in a reasonable time, or even to render them in real time. Thus, compression and decompression are indispensable for practical scene capture and rendering systems.

The light fields introduced by Levoy and Hanrahan [55] in 1996 were up to 1.6 GB in size. This would easily fill a large hard drive at the time, and would never fit in memory. However, light fields are highly redundant within images and between images, so they are highly compressible. Levoy and Hanrahan designed a custom light-field compression scheme that combines vector quantization of 2D or 4D tiles (24:1 compression) with gzip entropy encoding (another 5:1 compression) for a total compression of 120:1. This scheme allowed fast random-access decompression entirely in software, so that real-time rendering became feasible.

Recently, image compression techniques such as JPEG have become computationally affordable, even in real-time applications. Existing video codecs, such as h.264 and h.265, can also often be used directly for compressing video-based representations, such as 360° video [53,81] or omnidirectional stereo videos [3,87].

Collet *et al.* [15] encode their volumetric free-viewpoint videos in a standard MPEG-DASH file. Thanks to mesh tracking, their geometry has a temporally consistent parameterization. Therefore, the resulting texture atlases are unwrapped consistently and can be compressed effectively using the standard h.264 video codec. The mesh geometry is encoded as a custom unit inside the video stream and compressed using linear motion prediction, 16-bit quantization of vertex positions and UV coordinates, and Golomb coding.

Google's panoramic light fields require 4–6 GB of image data each [75] and so also need compression. As for the original light fields paper [55], fast random access is required for rendering novel views of the light field. Overbeck *et al.* [75] build on the open-source VP9 codec and encode most light field images relative to a sparse set of reference views, which are like key frames in standard videos. In practice, they decode all reference images when loading the light field from disk and keep them in memory. They also contribute an extension to VP9 that enables random access to individual image tiles. This allows their system to decode any tile from any other image immediately. Most light fields can hence be compressed at high quality by 40×–200×.

2.5 Rendering

The final step of the VR capture pipeline is to render the novel views corresponding to the user's location, so that they see the correct views of the captured scene as they move. Most rendering approaches adopt the standard graphics pipeline, which has the benefit of efficient hardware implementations across a large range of devices, from mobile to desktop setups. This efficient rendering hardware enables rendering in real time, and even hitting the high frame rates of 80–144 Hz required to feed state-of-the-art VR head-mounted displays [50].

Panoramas and omnidirectional stereo content only require a change to perspective projection to be viewed by users. This does not require any explicit geometry and can be implemented in 2D or, equivalently, by using textured spheres viewed from virtual perspective cameras. Many other approaches also use textured geometry directly [17,32,33,69,91]. Even multiplane images [25,63,98,124] can be rendered using textured geometry, by texturing the semi-transparent layers on parallel planes that are appropriately spaced, and using alpha compositing in the z-buffer during rendering.

Modern graphics pipelines are also programmable using shaders, which provides an opportunity to influence the rendering more locally depending on the viewing direction, for example. Flow-based blending has been used to interpolate novel views on the fly [58] and per pixel or light ray [86], also in a view-dependent fashion [5]. When many input views are combined to synthesize novel views, they also require spatial blending to ensure smooth transitions [75]. Ultimately, the decision of how to blend multiple observations of a single surface point can even be optimized using a deep neural network [34]. However, evaluating the neural network per frame at run time noticeably impacts the overall frame rate that is achievable with this approach, which does not yet reach real-time rates.

3 Neural Scene Representations and Rendering

Over the last few years, a new class of algorithms has emerged that has great potential for capturing, representing, and rendering real scenes in virtual environments—neural scene representations and rendering. The idea behind these algorithms is similar to classical approaches: given a set of input views, distill these into an intermediate representation, and then render the scene from novel viewpoints using the intermediate representation. However, a neural representation differs from a classical scene representation, such as a polygon mesh, a 3D point cloud, an implicit function, or a voxel grid, in being differentiable with respect to its parameters. In combination with a differentiable renderer that takes the neural scene representation as well as a camera position and orientation, i.e., a pose, as input and computes a 2D image from the camera's perspective, neural scene representations allow for end-to-end optimization of the representation supervised only on the images.

For example, Sitzmann *et al*. [94] recently proposed a voxel representation where each voxel is located in a Cartesian grid and stores a feature vector. A differentiable renderer with occlusion reasoning then projects these 3D features into 2D images and a 2D image-to-image translation rendering network then converts the projected features

into the RGB values of the final images. During training, the weights of the voxel features are optimized given only a set of posed RGB images of the scene. Once optimized, the neural voxels can be rendered into 2D RGB images given an arbitrary camera pose. Due to the fact that the intermediate voxel representation is inherently defined in a three-dimensional space, all projected views will be approximately consistent across different camera perspectives. This can be interpreted as choosing a neural network architecture that is aware of the 3D structure of the scene, simply by choosing an adequate scene representation. Several different classes of neural scene representations have been proposed over the last few years, which we briefly outline in the following.

Image-based Rendering with Deep Flow Prediction and Learned Image Blending. Recently, deep learning has been used to aid image-based rendering via learning subtasks, such as the prediction of occlusion-aware optical flow between views [42,43,76,125] and/or the computation of the blending weights [26,34]. While this approach can achieve photorealism, it depends on a dense set of high-resolution photographs to be available at rendering time and requires an error-prone reconstruction step to obtain the geometric proxy.

Unstructured or Weakly Structured Latent Representations. Other approaches aim at distilling an intermediate representation, or *embedding*, from the images. The benefit of such an approach is that the input views may not be necessary during inference anymore, after the embedding is learned. This is beneficial for multiple reasons: the used computational resources (such as memory) can be optimized; embeddings have the potential to disentangle different effects, such as lighting, shading, geometry etc., which can make them more interpretable or potentially even editable [51,61,118]; embeddings can sometimes also be interpolated or new examples within this latent space could even be generated. Therefore, learning structured embeddings is a topic of great interest.

Several approaches have been proposed that rely on embedding views into a latent space, but without enforcing any geometrical constraints [22,24,104]. Weakly structured embeddings [14,84,116], such as learning rotation-equivariant features by explicitly rotating the latent space feature vectors, have also been proposed. However, all of these approaches have in common that there are little to no guarantees that the synthesized views create consistent perspective projections because the underlying network structures do not enforce or capture the 3D structure of the scene explicitly. In other words, the choice of embedding captures the structure of the data weakly or not at all.

Using Proxy Geometries and Neural Textures. In many applications, such as reconstructing faces [126], hands, or whole body performances, we have detailed prior knowledge of the types of objects in the scene. For example, the image or 3D model of a face can be well described by a blendshape—a low-dimensional geometric basis function representation that only requires a few coefficients to model the face. Nonlinear optimization can be used to fit a blendshape representation to an image or video or a face. Similarly, parametric representations of hands or bodies exist and can be used to fit a 3D proxy geometry to 2D images or videos. Although such proxy geometries represent a good first-order approximation of the underlying shape, many subtle details

of the appearance, like the interior of the mouth, facial hair, or other perceptually important details, are typically not modeled in a convincing manner. However, such proxy geometries have great potential for neural scene representations because they can be rendered using existing computer graphics pipelines from arbitrary perspectives. To adequately model the appearance using a neural scene representation, a clever idea is to use the proxy geometry and texture it with a 2D texture containing learnable features. These features can then be optimized in a training stage for a given example image or video and later re-rendered. This approach uses little memory, because we do not have to learn a 3D model but only a texture and the neural renderer is a simple 2D (convolutional) neural network that computes an image-to-image translation from 2D features to RGB pixel values. This idea has so far been applied to deep video portraits [46] and it has been explored as a more general concept of neural textures [105].

One of the limitations of parametric representations is that they exist only for specific types of objects, such as faces, hands, and bodies. However, the idea of proxy geometry can also be applied to more general 3D computer vision pipelines. In this case, the reconstructed geometry is often coarse, it can have holes or it may be missing other parts. Yet, such an incomplete or noisy point cloud or mesh can still be easily rendered into arbitrary camera poses and an image-to-image translation network could then learn how to map from the incomplete projection of the point cloud to a photorealistic image. This idea was recently explored by Martin-Brualla *et al*. [60] and represents another example of combining proxy geometry with a differentiable (part of a) renderer.

Multiplane Image Representations. Many recent proposals on neural scene representations are based on the idea of decomposing a set of input views, or a light field, into a layered representation that can be re-projected into the input views but also into novel views. This is another example of using proxy geometry along with learnable parts, but the proxy geometry is a simple set of planes that can easily be projected into different cameras using homographies. Wetzstein *et al*. [52,112,113] optimized such representations for display application from densely sampled input light fields via computed tomography or non-negative matrix and tensor factorizations. More recently, deep learning based approaches have been proposed to optimize such representations using the input of small baseline stereo cameras [98,124], from single-input image [106], or from four input images [25] with learned gradient descent. The primary challenge in these deep learning based approaches is to work with a set of sparsely sampled input views and ensure that the views synthesized in between these given images look perceptually realistic. Another related approach recently proposed guided camera placement for such irregularly sampled light fields [63]. Most of these layered view synthesis approaches optimize RGBA color values at each position of the layers. The additional alpha channel allows for transparency-aware "soft" reconstructions by blending the layers for perspectives from different camera perspectives [80]. Figure 6 shows a visual comparison of classic and learned multiplane view synthesis approaches.

One benefit for multiplane image representations is that they are simple and, once optimized, enable real-time rendering of the layered representation. A downside is that novel views can only be synthesized over a limited baseline, i.e., we cannot synthesize novel views that look at the layers from the side.

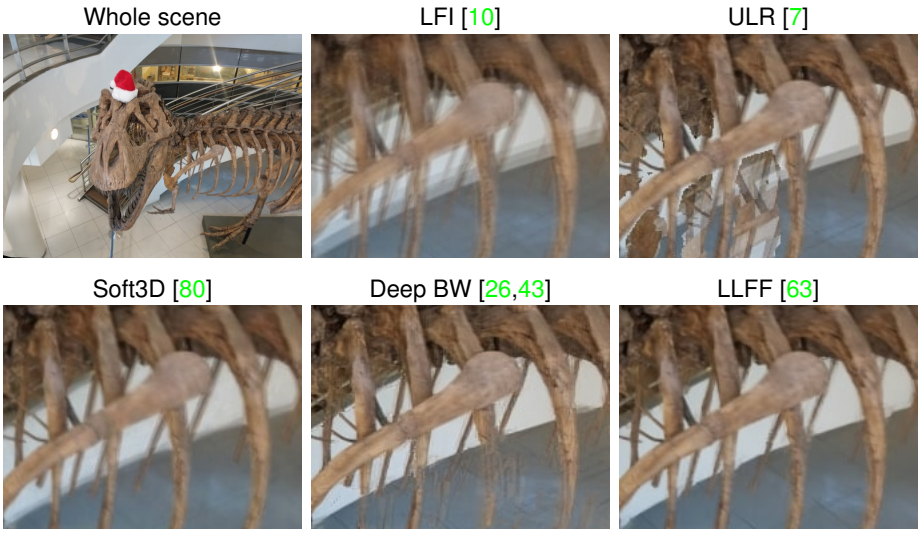

Fig. 6: Visual comparison of view synthesis results on the challenging T-Rex scene between traditional and neural rendering approaches. Light field interpolation [10] fails to align objects at different depths. Unstructured lumigraph rendering [7] suffers from poorly reconstructed geometry due to the thin ribs. Soft 3D reconstruction [80] shows blurry views caused by depth uncertainty. Deep backwards warping [26,43] exhibits visual artifacts near occlusion boundaries like the thin ribs. Local light-field fusion [63] smoothly blends neighboring local light fields to render novel views to minimize visual artifacts. Figure adapted from Mildenhall *et al.* [63].

Deep Voxel Representations. Another specific type of proxy geometry is a voxel grid, which overcomes the limited-baseline issue of sparse layered representations. For example, Sitzmann at al. [94] proposed an occlusion-aware volume renderer in combination with a grid of features that is trained only on posed 2D RGB images of a scene. Nguyen-Phuoc *et al.*'s HoloGAN [70] shows that deep voxel representations can also be learned from natural images in an unsupervised manner. The implicit deep 3D features enable disentangling of 3D pose and object identity, which can further be decomposed into shape and appearance. A different variant of this idea was recently shown to be able to generate real-time 3D reconstructions of human faces and actors [57,110]. One of the downsides of voxel representations is the relatively high memory footprint required to store all the voxels. Representing RGBA values or feature vectors on a Cartesian grid has the benefits of allowing convolutions and other intuitive operations to be performed on the grid [74], for example using a convolutional neural network, but another downside is that values have to be stored at all locations of the grid, even if there is no object there.

Deep Point Cloud Representations. Differentiable point clouds have the potential to overcome some of the memory limitations of layered or voxel-based representations

Fig. 7: Interpolating latent code vectors of cars and chairs in the ShapeNet dataset while rotating the camera around the model. Features smoothly transition from one model to another. Figure reproduced from Sitzmann *et al*. [95].

by adaptively changing the positions of the points [119]. This relates to the approach of Martin-Brualla *et al*. [60] who used a fixed (i.e., non-optimizable) proxy geometry. Neural re-rendering of point clouds is a promising postprocessing step that not only fills in gaps in rasterized point cloud renderings [2], but also also provides control over scene appearance [62]. A challenge of working with differentiable point clouds is to update the locations of the 3D points by backpropagating through a differentiable renderer, such as a splatting algorithm.

Continuous Neural Representations. Finally, differentiable continuous scene representations have also been explored. For example, Park *et al*. [77] recently proposed to model a signed distance function as a neural network and train it to learn an object's shape supervised on a 3D model of that object. Sitzmann *et al*. [95] introduced a differentiable renderer for such continuous scene representations to be able to train it in an end-to-end manner supervised only by posed 2D RGB images. Moreover, their approach allows for the scene representation to be generalized across object classes, enabling interpolation of the representations (see Figure 7), generating entirely new objects of a specific class, or fitting a 3D representation to a single 2D RGB image.

In summary, different variants of neural scene representation are emerging and show great potential to applications in capturing, representing, and rendering real environments in VR and beyond.

4 Limitations of Current Practice and Future Research Directions

Thus far, we have described current practice; next, we will discuss limitations of current practice and potential future research directions to overcome them. One useful framing device to help conceptualize these limitations is by what type of artifacts they introduce and by how much it affects the overall experience (Figure 8). Model-based approaches tend to introduce world inconsistencies which make them look fake, e.g., incorrect geometry, missing translucency or specular reflections, or suffering from Uncanny Valley effects in the case of humans. Image-based approaches introduce a different axis of artifacts relating to resolution and sampling, interpolation and warping, and tell-tale image compression errors. Navigating the design space between pure model-based (e.g., classic computer graphics) and pure image-based techniques (e.g., dense light fields) exposes our world reproduction to many artifacts; throughout these operations, our goal is to remain within some human-perception tolerable region which minimizes both world fakeness and image artifacts.

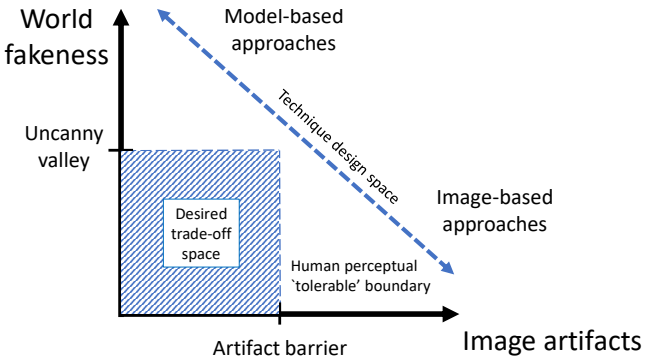

Fig. 8: One way to conceptualize our design-space trade-off is as one between model- and image-based methods. Model-based techniques tend to suffer 'uncanny valley' effects where the world appears only *almost* real, which is off-putting for human reproduction. On the other hand, image-based techniques exhibit characteristic artifacts which reduce quality and are easy to identify as 'unreal', such as sampling and compression artifacts. Our goal is to find a representation which minimizes both effects. Reproduced from whiteboard discussions with Brian Cabral and colleagues at the Dagstuhl RealVR seminar (July 2019).

Saturating the Senses (Vision): Current technologies for capture fall short of convincing real-world depiction purely from a raw pixel perspective, but are perhaps closer than you might anticipate. While perception varies from person to person, for spatial resolution, Facebook's Manifold camera [78] with 8K RED sensors over 180° (≈0.0225° resolution) is approaching the needed spatial resolution to match 20/20 visual acuity (0.0167° resolution). Temporal sampling of 60 Hz (16.6 ms per frame) is also approaching experimental rates for the task of individual image recognition (at least 14 ms [44]), though flickering artifacts can be seen at higher framerates. Static human eye dynamic range is relatively low (100:1, or 6.5 stops) vs. the approximately 15 stops available on the sensor [23], though eye dynamic contrast is extremely large (10^{14}, or 46 stops) and sensitive to very low luminance levels (10^{-6} cd/m2). While signal-to-noise ratios are improving at high ISO levels, the sensor would still struggle to produce a non-noisy image at this light level. This high sensitivity and low noise improves color reproduction. Binocular stereo provides depth cues from eye vergence, which is reproduced through multi-view renderings of the multi-camera reconstructions of the world geometry.

On the display side, current technologies also fall sort but may be closer than anticipated. The current best-in-class headset (Valve Index, September 2019 [108]) has a display panel resolution of 1440×1600 pixels per eye, and an LCD panel with dense subpixels to reduce the 'screen door' effect. Over a field of view of ≈130°, this provides 0.09° resolution. Temporal resolution is up to 144 Hz. One limitation with headsets is eye accommodation to allow focusing. This can be accomplished with very dense sensors for near-eye light field displays [52], or to use eye tracking and variable-depth displays [50].

These limits highlight the general quality issue with reconstructing imagery into VR-rendered representations: that the fidelity of the representation must match the fidelity of the display. Capture pixel resolution currently outstrips display resolution, but our true problem is one of reconstructing a representation which, when rendered, still saturates the display. This is easier for image-based rather than model-based reconstruction methods as they are 'closer to the camera', but image-based methods can limit another of our key senses: motion.

Saturating the Senses (Motion): Vision combines with the vestibular and proprioception senses to provide human beings with an awareness of motion. This must be keenly attended to in VR, with reproduction of motion parallax and occlusion required to achieve 6DoF video. However, the 'headbox' of allowable motion is limited by the baseline of the capturing camera system. Most 6DoF camera systems have baselines smaller than one meter; practically building and using a larger camera system is challenging. Beyond this, content must be hallucinated (or inpainted) in a plausible way [124], for which we can only expect 'good at best' quality and which becomes harder and harder as we move farther outside the baseline.

Motion also requires fast display pose estimation; head-motion rotation velocity in daily life can achieve 9 radians per second [8]. Current outside-in tracking systems (Valve Index) and inside-out (Oculus Quest [73]) provide millimeter-level tracking at sufficient framerates to meet their display framerates, though precise details are unspecified. Future work in this area will aim to track more of the human body beyond the head (hands, full-body pose) to allow greater interaction with our reconstructed VR worlds [50].

Capturing Everything Easily: Casual capture is another area of persistent need. Professional cameras and workflows are expensive and require expertise, and few systems exist to allow novice users to capture a scene with cheap hardware. For static scenes, casual capture can exploit the space/time swap: that space can be traded for time by moving the camera [5]. This lets the user 'sweep over' the scene from different poses, say with a smartphone, to complete the capture [69].

However, much interesting content is of dynamic scenes, which requires algorithms and representations to be temporally consistency. This is a much stricter requirement on accuracy as human sensitivity to perceptual effects over time is strong, e.g., any flickering at the edges of objects within a geometry reconstruction is particularly noticeable. Representations which explicitly accommodate time can also help; one example is spatiotemporal atlases for time-varying texture and geometry [82], but future work is needed for other representations along the image-to-model-based spectrum.

Complex material acquisition is another limitation and area of significant future work. Beyond simple Lambertian diffuse texture, we need to represent materials with shiny or glossy reflectances via 4D BRDFs, and to be able to capture transparent materials like glass. Without these effects, the world can look fake as objects do not visually respond realistically to human motion. Some methods exist for spatially-varying BRDF capture of objects from smartphones [68], but scaling these to whole scenes requires sampling many directions and is typically not possible for dynamic scenes. This sug-

gests further work in data-driven methods to fit known material models to sparse scene samples.

Similarly, we also wish to capture complex illumination, typically for editing applications like object insertion or relighting [61]. 360° imagery can act as environment reflection maps for lighting objects [103], but illumination estimation from perspective views is an ill-posed problem and requires learning-based methods [29,54]. Future work into plausible illumination reconstruction using inverse or differentiable rendering techniques holds promise.

Finally, to go beyond visual reproduction and into complex editing and interaction, capture must extend beyond representations of geometry and appearance and into object and scene semantics [47], context via hierarchical scene and relationship graphs [83], and even the capture of other physical properties such as aural properties, functional properties like mechanical actuation and articulation, and material properties like mass and elasticity.

Big Data Problems: Cameras which saturate our senses will produce terabytes of data for video sequences of just a few minutes in length. This question of compression is one that may initially seem tantalizingly simple because of the high level of redundancy in the data, e.g., for light field or 6DoF imagery, where each view is 'just a little bit different'. However, the minor differences in these samples are often what makes the viewing effect convincing.

Some representations focus on compression, storage, and transmission factors, such as formats that fit within a classic 2D video pipeline, like side-by-side omnidirectional stereo or RGB+D 360° representations [91]. Here, changes in time are well-handled. However, compressing in screen space in 2D can significantly limit flexibility and ultimately quality, making correct occlusion and motion parallax difficult. Scene-space reconstruction and parameterization allows these effects; however, for geometry-based reconstruction representations, these changes require more work to maintain temporal handling for easy compression. Work in 4D geometry from video addresses some of these challenges for human avatars [66,67], but representing larger dynamic scenes is still complex.

Our discussion in this chapter has also implicitly considered pre-recorded and post-processed data, but one significant area of future work is in producing high-quality representations and systems for live or streamed experiences. Future work is needed to produce real-time reconstruction approaches which exploit any and every natural redundancy in the data to reduce computation time and network transmission.

Perceptual Realism: We can capture to saturate human senses, and we can capture precise geometry, illumination, and materials for physically-based and model-based representations, but this still leaves many other challenges relating to perceptual effects. One classic effect discussed above is the uncanny valley [64], which hypothesizes perceptual effects to robot appearance and now has a casual understanding relating to digital human avatars. This effect is poorly understood and hard to quantify, particularly for geometry. Approaches to hide the problem with stylization can be successful. However, principled progress requires future work in better computational models of human perception, especially models which are differentiable and so can be used to optimize

a (neural) reconstruction function. These would allow faster exploration of a complex design space for capture, reconstruction, representation, and ultimately display by narrowing to only those effects we will see (and see comfortably).

Neural Rendering: Neural rendering techniques are plentiful and promising, with the space of possible techniques currently being explored. Neural scene representations exist to aid capture, stylization, hallucination, and view synthesis. That said, there are a few caveats with scene generality, representation size, real-time rendering, and editability.

The scene generality limitation is that many current techniques learn a neural representation which is specific to just one object or scene. This approach allows high-quality rendering, but new scenes require retraining the networks from new training data [34,94]. Future work to increase generalizability should look at how to resolve the trade-off between network capacity and quality. This problem is also related to the size of the neural representation and the ability to render quickly. Each network has millions of parameters and requires a large GPU to process, which limits their applicability via memory, rendering, and distribution costs. Future work should investigate efficient and compressible representations for neural scene rendering.

Finally, one benefit of physically-based scene models is that they can be edited easily. This is not the case for most neural representations, which have obtuse 'black box' representations which are difficult to inspect let alone edit. Current works attempt to build interfaces to help understand the generation process for 2D image synthesis [4] and steer it towards exposing more useful controls [40]; these must be extended for neural scene rendering to be useful for VR. Going forward, effort should be placed in constraining the learned representations to be implicitly editable in predictable ways.

Tools and Workflow: The question of editable representations—both classic and neural—highlights the critical role of tools and workflow in the capture and reconstruction pipeline. This area is often overlooked by academia but is vitally important; it requires partnerships between industry and academia to make reliable progress in tools that are actually adopted.

One question is the ease of use: hardware and software tools are currently catered towards experts, and casual capture is important, but both sets of users would benefit from representations which allow easier processing and manipulation of data. The problem is that different representations are better or worse for different tasks, and so the choice of production capture and representation often depends upon the scene content. For example, distant content can be better as an image-based representation to maximize image quality (suggesting 360° video), but up-close content can be better as depth- or model-based reconstruction to maximize realistic motion effects (suggesting depth or multi-view camera capture). The 'chicken and egg' problem of not knowing which systems to use until the content is captured lacks flexibility, and this has knock-on effects for the post-production workflow that also depends on the representation.

If the best representation for the job is task specific, then the underlying question could be one of how to allow easy *conversion* between representations. Most conversion operations require resampling, which is often lossy. This conversion must also happen

quickly and be memory efficient to allow engineers and artists to flexibly pick the best representation for the task at hand.

Editing tools themselves for multi-view content are also nascent, with limitations both in low-level reconstruction for matting and depth/layer decomposition, for interactive user operations like selection, geometry and appearance editing, and for imagery integration operations like compositing and CG insertion. Each of these tasks must be completed in spatio-temporally consistent ways, which requires high-accuracy reconstruction and typically some form of additional explicit consistency constraint.

While the craft has made significant progress over the past five years, we also lack higher-level tools and understanding for storytelling in VR, especially to allow novice users to express their experiences and ideas. In principle, VR can be a powerful method of 'experiential storytelling'—to put the viewer in the shoes of another. Efforts to reduce the cost and increase the ease of use and accessibility of our creative tools will democratize the capture, reconstruction, and editing pipeline and help more people tell effective stories.

Trust and Privacy: One new area of research is in building reconstructions and representations which respect privacy. This topic is often seen as more pressing for augmented reality systems (above VR systems), for which real-world capture, reconstruction, and representation techniques are equally as important. Scene capture can often include other humans who have not given permission for their representation, and these people may need to be anonymized in a realistic way [38]. Recent work has also looked at ways to prevent information leakage via *unwanted* scene reconstructions of people's homes derived from AR/VR headset tracking structure-from-motion systems [97]. Likewise, future work is needed on the security of information contained within neural representations.

5 Conclusion

If VR is to become more than just a technology for synthetic scenes—to use its powerful telepresence capability to impact domains across industry, commerce, healthcare, and the arts—then we must be able to capture and reproduce the dynamic visual world with high fidelity. We have discussed a range of existing and state-of-the-art solutions to capturing, reconstructing, representing, transmitting, and rendering the world for VR applications. Challenges remain: finding the sweet spots in the complex design space is difficult, and fundamental trade-offs about capture sampling and reproduction quality still remain. However, new neural representations which combine geometry proxies and learned appearance representation functions offer one potential approach to overcoming these trade-offs with domain-specific data-driven representations. Even with these challenges, the field of VR has made significant progress in the past five years, and capturing and distributing the real world is now easier than ever. We await the coming progress over the next five years with bated breath, particularly for VR's potential to improve the quality of telecommunications and ultimately reduce our carbon footprint by reducing the need for travel.

References

1. Aggarwal, R., Vohra, A., Namboodiri, A.M.: Panoramic stereo videos with a single camera. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3755–3763 (June 2016). doi:10.1109/CVPR.2016.408
2. Aliev, K.A., Ulyanov, D., Lempitsky, V.: Neural point-based graphics (2019), arXiv:1906.08240
3. Anderson, R., Gallup, D., Barron, J.T., Kontkanen, J., Snavely, N., Hernandez, C., Agarwal, S., Seitz, S.M.: Jump: Virtual reality video. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) **35**(6), 198:1–13 (November 2016). doi:10.1145/2980179.2980257
4. Bau, D., Zhu, J.Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., Torralba, A.: Seeing what a GAN cannot generate. In: Proceedings of the International Conference on Computer Vision (ICCV) (2019)
5. Bertel, T., Campbell, N.D.F., Richardt, C.: MegaParallax: Casual 360° panoramas with motion parallax. IEEE Transactions on Visualization and Computer Graphics **25**(5), 1828–1835 (May 2019). doi:10.1109/TVCG.2019.2898799
6. Brown, M., Lowe, D.G.: Automatic panoramic image stitching using invariant features. International Journal of Computer Vision **74**(1), 59–73 (2007). doi:10.1007/s11263-006-0002-3
7. Buehler, C., Bosse, M., McMillan, L., Gortler, S., Cohen, M.: Unstructured lumigraph rendering. In: Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). pp. 425–432 (2001). doi:10.1145/383259.383309
8. Bussone, W.: Linear and angular head accelerations in daily life. Ph.D. thesis, Virginia Tech (2005)
9. Cabral, B.: VR capture: Designing and building an open source 3D-360 video camera. SIGGRAPH Asia Keynote (December 2016)
10. Chai, J.X., Tong, X., Chan, S.C., Shum, H.Y.: Plenoptic sampling. In: Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). pp. 307–318 (2000). doi:10.1145/344779.344932
11. Chaurasia, G., Duchêne, S., Sorkine-Hornung, O., Drettakis, G.: Depth synthesis and local warps for plausible image-based navigation. ACM Transactions on Graphics **32**(3), 30:1–12 (July 2013). doi:10.1145/2487228.2487238
12. Chaurasia, G., Sorkine-Hornung, O., Drettakis, G.: Silhouette-aware warping for image-based rendering. Computer Graphics Forum (Proceedings of Eurographics Symposium on Rendering) **30**(4), 1223–1232 (June 2011). doi:10.1111/j.1467-8659.2011.01981.x
13. Chen, S.E., Williams, L.: View interpolation for image synthesis. In: Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). pp. 279–288 (1993). doi:10.1145/166117.166153
14. Cohen, T.S., Welling, M.: Transformation properties of learned visual representations. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015)
15. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **34**(4), 69:1–13 (July 2015). doi:10.1145/2766945
16. Curless, B., Seitz, S., Bouguet, J.Y., Debevec, P., Levoy, M., Nayar, S.K.: 3D photography. In: SIGGRAPH Courses (2000), http://www.cs.cmu.edu/~seitz/course/3DPhoto.html
17. Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., Theobalt, C.: BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. ACM Transactions on Graphics **36**(3), 24:1–18 (May 2017). doi:10.1145/3054739
18. Davis, A., Levoy, M., Durand, F.: Unstructured light fields. Computer Graphics Forum (Proceedings of Eurographics) **31**(2), 305–314 (May 2012). doi:10.1111/j.1467-8659.2012.03009.x

19. Debevec, P.: The light stages and their applications to photoreal digital actors. In: SIGGRAPH Asia Technical Briefs (2012)
20. Debevec, P., Bregler, C., Cohen, M.F., McMillan, L., Sillion, F., Szeliski, R.: Image-based modeling, rendering, and lighting. In: SIGGRAPH Courses (2000), https://www.pauldebevec.com/IBMR99/
21. Debevec, P.E., Taylor, C.J., Malik, J.: Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. In: Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). pp. 11–20 (August 1996). doi:10.1145/237170.237191
22. Dosovitskiy, A., Springenberg, J.T., Tatarchenko, M., Brox, T.: Learning to generate chairs, tables and cars with convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(4), 692–705 (April 2017). doi:10.1109/TPAMI.2016.2567384
23. DXOMARK: RED Helium 8K DxOMark sensor score: 108 — a new all-time-high score! https://www.dxomark.com/red-helium-8k-dxomark-sensor-score-108-a-new-all-time-high-score2/, (Accessed on 30 October 2019)
24. Eslami, S.M.A., Jimenez Rezende, D., Besse, F., Viola, F., Morcos, A.S., Garnelo, M., Ruderman, A., Rusu, A.A., Danihelka, I., Gregor, K., Reichert, D.P., Buesing, L., Weber, T., Vinyals, O., Rosenbaum, D., Rabinowitz, N., King, H., Hillier, C., Botvinick, M., Wierstra, D., Kavukcuoglu, K., Hassabis, D.: Neural scene representation and rendering. Science **360**(6394), 1204–1210 (June 2018). doi:10.1126/science.aar6170
25. 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 International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2367–2376 (June 2019)
26. Flynn, J., Neulander, I., Philbin, J., Snavely, N.: DeepStereo: Learning to predict new views from the world's imagery. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5515–5524 (June 2016). doi:10.1109/CVPR.2016.595
27. Fuhrmann, S., Langguth, F., Goesele, M.: MVE: A multi-view reconstruction environment. In: Proceedings of the Eurographics Workshop on Graphics and Cultural Heritage. pp. 11–18 (2014). doi:10.2312/gch.20141299
28. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: Proceedings of the International Conference on Computer Vision (ICCV). pp. 873–881 (December 2015). doi:10.1109/ICCV.2015.106
29. Garon, M., Sunkavalli, K., Hadap, S., Carr, N., Lalonde, J.F.: Fast spatially-varying indoor lighting estimation. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
30. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F.: The lumigraph. In: Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). pp. 43–54 (August 1996). doi:10.1145/237170.237200
31. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2004). doi:10.1017/CBO9780511811685
32. Hedman, P., Alsisan, S., Szeliski, R., Kopf, J.: Casual 3D photography. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) **36**(6), 234:1–15 (November 2017). doi:10.1145/3130800.3130828
33. Hedman, P., Kopf, J.: Instant 3D photography. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **37**(4), 101:1–12 (July 2018). doi:10.1145/3197517.3201384
34. Hedman, P., Philip, J., Price, T., Frahm, J.M., Drettakis, G.: Deep blending for free-viewpoint image-based rendering. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) **37**(6), 257:1–15 (November 2018). doi:10.1145/3272127.3275084

35. Hedman, P., Ritschel, T., Drettakis, G., Brostow, G.: Scalable inside-out image-based rendering. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) **35**(6), 231:1–11 (November 2016). doi:10.1145/2980179.2982420
36. Huang, J., Chen, Z., Ceylan, D., Jin, H.: 6-DOF VR videos with a single 360-camera. In: Proceedings of IEEE Virtual Reality (VR). pp. 37–44 (March 2017). doi:10.1109/VR.2017.7892229
37. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: DeepMVS: Learning multi-view stereopsis. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
38. Hukkelås, H., Mester, R., Lindseth, F.: DeepPrivacy: A generative adversarial network for face anonymization. In: Advances in Visual Computing. pp. 565–578 (2019). doi:10.1007/978-3-030-33720-9_44
39. Ishiguro, H., Yamamoto, M., Tsuji, S.: Omni-directional stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence **14**(2), 257–262 (February 1992). doi:10.1109/34.121792
40. Jahanian, A., Chai, L., Isola, P.: On the "steerability" of generative adversarial networks (2019), arXiv:1907.07171
41. Jancosek, M., Pajdla, T.: Multi-view reconstruction preserving weakly-supported surfaces. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3121–3128 (June 2011). doi:10.1109/CVPR.2011.5995693
42. Ji, D., Kwon, J., McFarland, M., Savarese, S.: Deep view morphing. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7092–7100 (July 2017). doi:10.1109/CVPR.2017.750
43. Kalantari, N.K., Wang, T.C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) **35**(6), 193:1–10 (November 2016). doi:10.1145/2980179.2980251
44. Keysers, C., Xiao, D.K., Földiák, P., Perrett, D.I.: The speed of sight. Journal of Cognitive Neuroscience **13**(1), 90–101 (2001). doi:10.1162/089892901564199
45. Kim, C., Zimmer, H., Pritch, Y., Sorkine-Hornung, A., Gross, M.: Scene reconstruction from high spatio-angular resolution light fields. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **32**(4), 73:1–12 (July 2013). doi:10.1145/2461912.2461926
46. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **37**(4), 163:1–14 (August 2018). doi:10.1145/3197517.3201283
47. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
48. Konrad, R., Dansereau, D.G., Masood, A., Wetzstein, G.: SpinVR: Towards live-streaming 3D virtual reality video. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) **36**(6), 209:1–12 (November 2017). doi:10.1145/3130800.3130836
49. Kopf, J., Alsisan, S., Ge, F., Chong, Y., Matzen, K., Quigley, O., Patterson, J., Tirado, J., Wu, S., Cohen, M.F.: Practical 3D photography. In: Proceedings of CVPR Workshops (2019)
50. Koulieris, G.A., Akşit, K., Stengel, M., Mantiuk, R.K., Mania, K., Richardt, C.: Near-eye display and tracking technologies for virtual and augmented reality. Computer Graphics Forum **38**(2), 493–519 (May 2019). doi:10.1111/cgf.13654
51. Kulkarni, T.D., Whitney, W., Kohli, P., Tenenbaum, J.B.: Deep convolutional inverse graphics network. In: Advances in Neural Information Processing Systems (NIPS). pp. 2539–2547 (2015)
52. Lanman, D., Wetzstein, G., Hirsch, M., Heidrich, W., Raskar, R.: Polarization fields: Dynamic light field display using multi-layer LCDs. ACM Transactions on

Graphics (Proceedings of SIGGRAPH Asia) **30**(6), 186:1–10 (December 2011). doi:10.1145/2070781.2024220

53. Lee, J., Kim, B., Kim, K., Kim, Y., Noh, J.: Rich360: Optimized spherical representation from structured panoramic camera arrays. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **35**(4), 63:1–11 (July 2016). doi:10.1145/2897824.2925983
54. LeGendre, C., Ma, W.C., Fyffe, G., Flynn, J., Charbonnel, L., Busch, J., Debevec, P.: DeepLight: Learning illumination for unconstrained mobile mixed reality. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
55. Levoy, M., Hanrahan, P.: Light field rendering. In: Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). pp. 31–42 (August 1996). doi:10.1145/237170.237199
56. Lipski, C., Linz, C., Berger, K., Sellent, A., Magnor, M.: Virtual video camera: Image-based viewpoint navigation through space and time. Computer Graphics Forum **29**(8), 2555–2568 (December 2010). doi:10.1111/j.1467-8659.2010.01824.x
57. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **38**(4), 65:1–14 (July 2019). doi:10.1145/3306346.3323020
58. Luo, B., Xu, F., Richardt, C., Yong, J.H.: Parallax360: Stereoscopic 360° scene representation for head-motion parallax. IEEE Transactions on Visualization and Computer Graphics **24**(4), 1545–1553 (April 2018). doi:10.1109/TVCG.2018.2794071
59. Magnor, M., Grau, O., Sorkine-Hornung, O., Theobalt, C. (eds.): Digital Representations of the Real World: How to Capture, Model, and Render Visual Reality. A K Peters/CRC Press (May 2015)
60. Martin-Brualla, R., Pandey, R., Yang, S., Pidlypenskyi, P., Taylor, J., Valentin, J., Khamis, S., Davidson, P., Tkach, A., Lincoln, P., Kowdle, A., Rhemann, C., Goldman, D.B., Keskin, C., Seitz, S., Izadi, S., Fanello, S.: LookinGood: Enhancing performance capture with real-time neural re-rendering. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) **37**(6), 255:1–14 (November 2018). doi:10.1145/3272127.3275099
61. Meka, A., Häne, C., Pandey, R., Zollhöfer, M., Fanello, S., Fyffe, G., Kowdle, A., Yu, X., Busch, J., Dourgarian, J., Denny, P., Bouaziz, S., Lincoln, P., Whalen, M., Harvey, G., Taylor, J., Izadi, S., Tagliasacchi, A., Debevec, P., Theobalt, C., Valentin, J., Rhemann, C.: Deep reflectance fields: High-quality facial reflectance field inference from color gradient illumination. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **38**(4), 77:1–12 (July 2019). doi:10.1145/3306346.3323027
62. Meshry, M., Goldman, D.B., Khamis, S., Hoppe, H., Pandey, R., Snavely, N., Martin-Brualla, R.: Neural rerendering in the wild. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
63. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **38**(4), 29:1–14 (July 2019). doi:10.1145/3306346.3322980
64. Mori, M.: The Uncanny Valley. Energy (in Japanese) **7**(4), 33–35 (1970)
65. Moulon, P., Monasse, P., Marlet, R.: Adaptive structure from motion with a Contrario model estimation. In: Proceedings of the Asian Conference on Computer Vision (ACCV). pp. 257–270 (2012). doi:10.1007/978-3-642-37447-0_20
66. Mustafa, A., Volino, M., Guillemaut, J.Y., Hilton, A.: 4D temporally coherent light-field video. In: Proceedings of International Conference on 3D Vision (3DV) (2017)
67. Mustafa, A., Volino, M., Kim, H., Guillemaut, J.Y., Hilton, A.: Temporally coherent general dynamic scene reconstruction (2019), arXiv:1907.08195

68. Nam, G., Lee, J.H., Gutierrez, D., Kim, M.H.: Practical SVBRDF acquisition of 3D objects with unstructured flash photography. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) **37**(6), 267:1–12 (December 2018). doi:10.1145/3272127.3275017
69. Newcombe, R.A., Davison, A.J., Izadi, S., Kohli, P., Hilliges, O., Shotton, J., Molyneaux, D., Hodges, S., Kim, D., Fitzgibbon, A.: KinectFusion: Real-time dense surface mapping and tracking. In: Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR). pp. 127–136 (October 2011). doi:10.1109/ISMAR.2011.6092378
70. Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., Yang, Y.L.: HoloGAN: Unsupervised learning of 3D representations from natural images. In: Proceedings of the International Conference on Computer Vision (ICCV) (2019)
71. Nießner, M., Zollhöfer, M., Izadi, S., Stamminger, M.: Real-time 3D reconstruction at scale using voxel hashing. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) **32**(6), 169:1–11 (November 2013). doi:10.1145/2508363.2508374
72. Niklaus, S., Mai, L., Yang, J., Liu, F.: 3D Ken Burns effect from a single image. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) **38**(6), 184:1–15 (November 2019). doi:10.1145/3355089.3356528
73. Oculus: From the lab to the living room: The story behind Facebook's Oculus Insight technology and a new era of consumer VR. https://tech.fb.com/the-story-behind-oculus-insight-technology/, (Accessed on 30 October 2019)
74. Olszewski, K., Tulyakov, S., Woodford, O., Li, H., Luo, L.: Transformable bottleneck networks. In: Proceedings of the International Conference on Computer Vision (ICCV) (2019)
75. Overbeck, R.S., Erickson, D., Evangelakos, D., Pharr, M., Debevec, P.: A system for acquiring, compressing, and rendering panoramic light field stills for virtual reality. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) **37**(6), 197:1–15 (2018). doi:10.1145/3272127.3275031
76. Park, E., Yang, J., Yumer, E., Ceylan, D., Berg, A.C.: Transformation-grounded image generation network for novel 3D view synthesis. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 702–711 (July 2017). doi:10.1109/CVPR.2017.82
77. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepSDF: Learning continuous signed distance functions for shape representation. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
78. Parra Pozo, A., Toksvig, M., Filiba Schrager, T., Hsu, J., Mathur, U., Sorkine-Hornung, A., Szeliski, R., Cabral, B.: An integrated 6DoF video camera and system design. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) **38**(6), 216:1–16 (November 2019). doi:10.1145/3355089.3356555, https://github.com/facebook/facebook360_dep
79. Peleg, S., Ben-Ezra, M., Pritch, Y.: Omnistereo: Panoramic stereo imaging. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(3), 279–290 (2001). doi:10.1109/34.910880
80. Penner, E., Zhang, L.: Soft 3D reconstruction for view synthesis. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) **36**(6), 235:1–11 (November 2017). doi:10.1145/3130800.3130855
81. Perazzi, F., Sorkine-Hornung, A., Zimmer, H., Kaufmann, P., Wang, O., Watson, S., Gross, M.: Panoramic video from unstructured camera arrays. Computer Graphics Forum (Proceedings of Eurographics) **34**(2), 57–68 (May 2015). doi:10.1111/cgf.12541
82. Prada, F., Kazhdan, M., Chuang, M., Collet, A., Hoppe, H.: Spatiotemporal atlas parameterization for evolving meshes. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **36**(4), 58:1–12 (July 2017). doi:10.1145/3072959.3073679
83. Qi, M., Li, W., Yang, Z., Wang, Y., Luo, J.: Attentive relational networks for mapping images to scene graphs. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

84. Rhodin, H., Salzmann, M., Fua, P.: Unsupervised geometry-aware representation for 3D human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 765–782 (2018). doi:10.1007/978-3-030-01249-6_46
85. Richardt, C., Hedman, P., Overbeck, R.S., Cabral, B., Konrad, R., Sullivan, S.: Capture4VR: From VR photography to VR video. In: SIGGRAPH Courses (2019). doi:10.1145/3305366.3328028
86. Richardt, C., Pritch, Y., Zimmer, H., Sorkine-Hornung, A.: Megastereo: Constructing high-resolution stereo panoramas. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1256–1263 (June 2013). doi:10.1109/CVPR.2013.166
87. Schroers, C., Bazin, J.C., Sorkine-Hornung, A.: An omnistereoscopic video pipeline for capture and display of real-world VR. ACM Transactions on Graphics **37**(3), 37:1–13 (August 2018). doi:10.1145/3225150
88. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4104–4113 (2016). doi:10.1109/CVPR.2016.445
89. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 501–518 (2016). doi:10.1007/978-3-319-46487-9_31
90. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. 519–528 (June 2006). doi:10.1109/CVPR.2006.19
91. Serrano, A., Kim, I., Chen, Z., DiVerdi, S., Gutierrez, D., Hertzmann, A., Masia, B.: Motion parallax for 360° RGBD video. IEEE Transactions on Visualization and Computer Graphics **25**(5), 1817–1827 (May 2019). doi:10.1109/TVCG.2019.2898757
92. Shum, H., Kang, S.B.: Review of image-based rendering techniques. In: Visual Communications and Image Processing. Proceedings of SPIE, vol. 4067 (2000). doi:10.1117/12.386541
93. Shum, H.Y., Chan, S.C., Kang, S.B.: Image-Based Rendering. Springer (2007). doi:10.1007/978-0-387-32668-9
94. Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhöfer, M.: DeepVoxels: Learning persistent 3D feature embeddings. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2437–2446 (2019)
95. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: Continuous 3D-structure-aware neural scene representations. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
96. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3D. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **25**(3), 835–846 (July 2006). doi:10.1145/1141911.1141964
97. Speciale, P., Schönberger, J.L., Kang, S.B., Sinha, S.N., Pollefeys, M.: Privacy preserving image-based localization. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
98. Srinivasan, P.P., Tucker, R., Barron, J.T., Ramamoorthi, R., Ng, R., Snavely, N.: Pushing the boundaries of view extrapolation with multiplane images. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 175–184 (June 2019)
99. Sweeney, C.: Theia multiview geometry library: Tutorial & reference (2016), http://theia-sfm.org
100. Sweeney, C., Holynski, A., Curless, B., Seitz, S.M.: Structure from motion for panorama-style videos (2019), arXiv:1906.03539

101. Szeliski, R.: Image alignment and stitching: a tutorial. Foundations and Trends in Computer Graphics and Vision **2**(1), 1–104 (January 2006). doi:10.1561/0600000009
102. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer (2010). doi:10.1007/978-1-84882-935-0, http://szeliski.org/Book/
103. Tarko, J., Tompkin, J., Richardt, C.: Real-time virtual object insertion for moving 360° videos. In: Proceedings of the International Conference on Virtual-Reality Continuum and its Applications in Industry (VRCAI) (2019)
104. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Multi-view 3D models from single images with a convolutional network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 322–337 (2016). doi:10.1007/978-3-319-46478-7_20
105. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **38**(4), 66:1–12 (July 2019). doi:10.1145/3306346.3323035
106. Tulsiani, S., Tucker, R., Snavely, N.: Layer-structured 3D scene inference via view synthesis. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018). doi:10.1007/978-3-030-01234-2_19
107. Tung, H.Y.F., Cheng, R., Fragkiadaki, K.: Learning spatial common sense with geometry-aware recurrent networks. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2595–2603 (2019)
108. Valve Corporation: Index headset. www.valvesoftware.com/en/index/headset, (Accessed on 30 October 2019)
109. Ventura, J.: Structure from motion on a sphere. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 53–68 (2016). doi:10.1007/978-3-319-46487-9_4
110. Wei, S.E., Saragih, J., Simon, T., Harley, A.W., Lombardi, S., Perdoch, M., Hypes, A., Wang, D., Badino, H., Sheikh, Y.: VR facial animation via multiview image translation. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **38**(4), 67:1–16 (July 2019). doi:10.1145/3306346.3323030
111. Weissig, C., Schreer, O., Eisert, P., Kauff, P.: The ultimate immersive experience: Panoramic 3D video acquisition. In: Advances in Multimedia Modeling (MMM). Lecture Notes in Computer Science, vol. 7131, pp. 671–681 (2012). doi:10.1007/978-3-642-27355-1_72
112. Wetzstein, G., Lanman, D., Heidrich, W., Raskar, R.: Layered 3D: Tomographic image synthesis for attenuation-based light field and high dynamic range displays. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **30**(4), 95:1–12 (July 2011). doi:10.1145/2010324.1964990
113. Wetzstein, G., Lanman, D., Hirsch, M., Raskar, R.: Tensor displays: Compressive light field synthesis using multilayer displays with directional backlighting. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **31**(4), 80:1–11 (July 2012). doi:10.1145/2185520.2185576
114. Whelan, T., Salas-Moreno, R.F., Glocker, B., Davison, A.J., Leutenegger, S.: ElasticFusion: Real-time dense SLAM and light source estimation. The International Journal of Robotics Research **35**(14), 1697–1716 (2016). doi:10.1177/0278364916669237
115. Wood, D.N., Azuma, D.I., Aldinger, K., Curless, B., Duchamp, T., Salesin, D.H., Stuetzle, W.: Surface light fields for 3D photography. In: Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). pp. 287–296 (2000). doi:10.1145/344779.344925
116. Worrall, D.E., Garbin, S.J., Turmukhambetov, D., Brostow, G.J.: Interpretable transformations with encoder-decoder networks. In: Proceedings of the International Conference on Computer Vision (ICCV). pp. 5737–5746 (2017). doi:10.1109/ICCV.2017.611
117. Wu, C.: VisualSFM: A visual structure from motion system (2011), http://ccwu.me/vsfm/

118. Yang, J., Reed, S.E., Yang, M.H., Lee, H.: Weakly-supervised disentangling with recurrent transformations for 3D view synthesis. In: Advances in Neural Information Processing Systems (NIPS). pp. 1099–1107 (2015)
119. Yifan, W., Serena, F., Wu, S., Öztireli, C., Sorkine-Hornung, O.: Differentiable surface splatting for point-based geometry processing. ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) **38**(6) (November 2019). doi:10.1145/3355089.3356513
120. Yücer, K., Sorkine-Hornung, A., Wang, O., Sorkine-Hornung, O.: Efficient 3D object segmentation from densely sampled light fields with applications to 3D reconstruction. ACM Transactions on Graphics **35**(3), 22:1–15 (March 2016). doi:10.1145/2876504
121. Zaragoza, J., Chin, T.J., Tran, Q.H., Brown, M.S., Suter, D.: As-projective-as-possible image stitching with moving DLT. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1285–1298 (July 2014). doi:10.1109/TPAMI.2013.247
122. Zhang, F., Liu, F.: Parallax-tolerant image stitching. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3262–3269 (June 2014). doi:10.1109/CVPR.2014.423
123. Zheng, K.C., Kang, S.B., Cohen, M.F., Szeliski, R.: Layered depth panoramas. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR) (2007). doi:10.1109/CVPR.2007.383295
124. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **37**(4), 65:1–12 (August 2018). doi:10.1145/3197517.3201323
125. Zhou, T., Tulsiani, S., Sun, W., Malik, J., Efros, A.A.: View synthesis by appearance flow. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 286–301 (2016). doi:10.1007/978-3-319-46493-0_1
126. 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 **37**(2), 523–550 (May 2018). doi:10.1111/cgf.13382