Synthesising Light Field Volume Visualisations Using Image Warping in Real-Time

Seán K. Martin[1], Seán Bruton[1], David Ganter[1], and Michael Manzke[1]

School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland
{martins7, sbruton, ganterd, manzkem}@tcd.ie

Abstract. We extend our prior research on light field view synthesis for volume data presented in the conference proceedings of VISIGRAPP 2019 [13]. In that prior research, we identified the best Convolutional Neural Network, depth heuristic, and image warping technique to employ in our light field synthesis method. Our research demonstrated that applying backward image warping using a depth map estimated during volume rendering followed by a Convolutional Neural Network produced high quality results. In this body of work, we further address the generalisation of Convolutional Neural Network applied to different volumes and transfer functions from those trained upon. We show that the Convolutional Neural Network (CNN) fails to generalise on a large dataset of head magnetic resonance images. Additionally, we speed up our implementation to enable better timing comparisons while remaining functionally equivalent to our previous method. This produces a real-time application of light field synthesis for volume data and the results are of high quality for low-baseline light fields.

Keywords: Light Fields · View Synthesis · Volume Rendering · Depth Estimation · Image Warping · Angular Resolution Enhancement.

1 INTRODUCTION

In our previous work we demonstrated a method for high quality light field synthesis on a specific volume and transfer function [13]. Lacking a baseline for our particular problem, we evaluated multiple different strategies for each step of our method. To be specific, multiple CNNs, depth heuristics and warping methods were experimented with. Although this provided detailed results for our specific dataset, we had minimal tests for generalisation of the method to other volumes and transfer functions. We demonstrated that training the CNN on a heart Magnetic Resonance Imaging (MRI) would not be sufficient to then use the CNN on a head MRI. But, it is possible that the CNN could be trained on head MRIs and generalise to work on other unseen MRIs, or be trained on a set of transfer functions and generalise to work on similar transfer functions. To investigate this idea of generalisability, we performed additional experiments on the best performing strategy from our previous work. We present results for a dataset of multiple head MRIs with different transfer functions, a far more variable case than our original dataset of a heart MRI. Furthermore, the implementation has been rewritten to take better advantage of shader functions, leading to far more efficient approach and more accurate timings than the previous work.

As before, we aim towards low baseline light field displays with a fast light field synthesis approach for volumetric data. Although light field display technology is still in its infancy, devices such as The Looking Glass [2] and a near-eye light field display [6] show promising results. Generating a light field allows for intuitive exploration of volumetric data. In gross anatomy, studies have shown that students respond better to three dimensional (3D) anatomical atlases for rapid identification of anatomical stuctures [16], but is not necessary for deep anatomical knowledge. In this sense, a light field synthesis approache fits very well for rapid exploration of anatomical structure. For instance, on the microscopic level, histology volume reconstruction [24] could be readily visualised in 3D without the need for glasses. However, it is impossible to render a light field from volumetric data using ray tracing techniques at interactive rates without an extensive and expensive hardware setup. To move towards light field volume visualisation at interactive rates, we use a view synthesis method from a single sample view to avoid rendering every view in the light field.

Unlike conventional images, which record a two dimensional (2D) projection of light rays, a light field describes the distribution of light rays in free space [33] described by a four dimensional (4D) plenoptic function. Capturing an entire light field is infeasible, so a light field is generally sampled by capturing a large uniform grid of images of a scene. [7]. However, capturing this 4D data leads to an inherent resolution trade off between dimensions. As such, the angular resolution, the number of images in the grid, is often low compared to the spatial resolution, the size of each image in the grid. Because of this, angular resolution enhancement, or view synthesis, from a limited number of reference views is of great interest and benefit to a light field application.

Angular resolution enhancement for light fields of natural images (images taken by a real camera) has many strong solutions [14,34,4]. However, for light fields from volumetric data, we face some unique challenges. Natural image approaches generally assume a unique depth in the scene, but volume renderings frequently feature semi-transparent objects, resulting in an ill-defined depth. Additionally, as opposed to natural light field imaging where the limitation is the resolution of the sensor, for volumetric data the limitation is time. Speed of the approach is vital to us, as an exact light field rendering can be produced with enough time. As such, many existing approaches are not relevant without heavy modification. For instance, the recent state of the art approach by [32] takes 2.57 seconds to produce a single 541×376 view, while an entire light field from a volume could be produced in a tenth of this time with a powerful Graphics Processing Unit (GPU) for a small volume.

Our proposed method synthesises a light field from a single volume rendered sample image, represented as a 2D grid of images captured by a camera moving along a plane. As the first step in this process, a depth heuristic is used to estimate a depth map during volume ray casting [36]. This depth map is converted to a disparity map using the known virtual camera parameters. Backward image warping is performed using the disparity map to shift information from the single sample image to all novel view locations. Finally, the warped images are passed into a CNN to improve the visual consistency of the synthesised views. This method is demonstrated for an 8×8 angular resolution light field (a grid of 64 images) but there is no inherent limitation on this other than GPU memory.

We previously demonstrated that a CNN increases the visual consistency of synthesised views, especially for those views at a large distance from the sample reference view [13]. However, by testing on a large dataset of head MRIs, we further back up our previous results on the lack of generalisation of the CNN. The CNN must be retrained to be used on a specific volume and transfer function, its performance does not transfer between sets. This is a significant limitation, but the CNN could be ommitted for particular use cases if a lower quality estimate is acceptable. Alternatively, for use cases such as education, a library of specifically trained models could be produced.

The re-implemented method is very fast, an order of magnitude faster than rendering each view in the light by traditional volume ray casting. For low baseline light fields, the results are of high quality, as the single sample view provides sufficient information to estimate the light field. The method is particularly beneficial for large volumes because the time to synthesise a light field is independent of the size and complexity of the volume and rendering techniques once a sample view has been produced. In high precision fields where exact results are required, the synthesis could be used for exploratory purposes and a final render performed when the camera is held fixed. In this manner, inevitable errors in the visualisation technique will not lead to medical misdiagnosis or other serious problems.

2 BACKGROUND

2.1 Light Field Representation

The first description of the light field was the 7D plenoptic function recording the intensity of light rays travelling in every direction (θ, ϕ) through every location in space (x, y, z) for every wavelength λ, at every time t [1]. As a simplification, the plenoptic function is assumed to be mono-chromatic and time invariant, reducing it to a 5D function $L(x, y, z, \theta, \phi)$. In a region free of any occluders, the radiance along a ray remains constant along a straight line, and the 5D plenoptic function reduces to the 4D function $L(x, y, \theta, \phi)$ [7,3]. This 4D plenoptic function is termed the 4D light field, or simply the light field, the radiance along rays in empty space.

The most common light field representation is to parameterise a ray by its intersection with two planes, a *uv* plane and an *st* plane [7]. As such the 4D light field maps rays passing through a point (u, v) on one plane and (s, t) on another plane to a radiance value:

$$L : \mathbb{R}^4 \to \mathbb{R}, \quad (u, v, s, t) \mapsto L(u, v, s, t)$$

With this parameterisation, a light field can be effectively sampled using arrays of rendered or digital images by considering the camera to sit on the *uv* plane, and capture images on the *st* plane. It was later shown that if the disparity between neighbouring views in a light field sampling is less than one pixel, novel views can be generated without ghosting effects by the use of linear interpolation [9]. Since sampling rates are rarely this high, to achieve such a densely sampled light field, view synthesis is a necessity.

2.2 Light Field View Synthesis

Light field view synthesis is the process of producing unseen views in the light field from a set of sample views. This is a well studied problem [35,18,30], see [33] for a review of light field view synthesis, as well as multiple other light field image processing problems. However, because speed is essential to our problem, much of the existing literature is unusable without significant modifications. Nonetheless, we recount prominent literature here and elucidate why the majority of it can not be used directly.

A wide range of view synthesis approaches take advantage of the inherent structure of light fields. A set of state of the art structure based approaches revolve around using the properties of Epipolar-Plane Images (EPIs) [34,32]. In this context an EPI is a 2D slice of the 4D light field in which one image plane co-ordinate is fixed and one camera plane co-ordinate is fixed. For instance, fixing a horizontal co-ordinate t^* on the image plane, and v^* on the camera plane produces an EPI:

$$E : \mathbb{R}^2 \to \mathbb{R}, \quad (u,s) \mapsto L(u,v^*,s,t^*)$$

A line of constant slope in an EPI corresponds to a point in 3D space where the slope of the line is related to the depth of the point. For Lambertian objects, this means that the instensity of the light field should not change along a line of constant slope [30]. This strong sense of structure is the basis of accurate view synthesis using EPIs. Other structure based approaches transform the light field to other domains with similar strong contraints [23,28]. Despite the good results from these approaches, the necessary domain transfer or slicing transforms are too slow for our purpose. As an example, [34] takes roughly 16 minutes to synthesise a $64 \times 512 \times 512$ light field which could be rendered in under a second.

For large-baseline light fields, such as that required for the Looking Glass [2], methods based on multiplane images [22] represent the state of the art [14,38]. These revolve around the principle of splitting reference views into multiple images at different depth planes to gain more information and simplify the aggregation process with great success. In volume rendering, a similar idea has been applied to view synthesis [10], by subdividing the original ray-casting into a layered representation. The layered information can be effectively transformed to a novel view, and combined via compositing an efficient manner. Although this produces fast single view synthesis for a low number of layers, and would be very effective for stereo magnification, the number of views required for a light field adds significant difficulty to the fast application of this paradigm.

As such, fast depth-based warping approaches are most relevant to our problem of synthesising low-baseline light fields, as we can cheaply estimate some form of depth during a volume rendering. With this in mind, the most relevant body of work to the problem at hand is by Srinivasan et al. [26]. This is fast, synthesising a $187 \times 270 \times 8 \times 8$ light field in under one second on a NVIDIA Titan X GPU. For comparison, a state of the art depth based approach [4] takes roughly 12.3 seconds to generate a single novel view from four input images of 541×376 resolution. Srinivasan et al.'s [26] method is fast, uses deep learning to account for specular highlights, and only requires a single input view to synthesise a full light field. This is very relevant for volume rendering, as speed is essential, surfaces are often anisotropically shaded, and rendering sample views is especially slow.

A single image does not carry enough information to truthfully reconstruct the light field, so at best a good estimate will result from this process. By taking advantage of redundant information in the light field representation, the Srinivasan et al. [26] achieve a high quality estimate, and a method summary follows here. First, a 3D CNN is applied to a single sample view to estimate the depth in the scene for a full light field. Backward warping is then applied to the sample view using the depth maps to estimate a Lambertian light field. Finally, the Lambertian light field is passed through a 3D CNN to help account for specular highlights. This produces high quality results for objects from specific categories, such as flowers. We base our approach on the method of [26], but provide modifications to increase suitability for volume rendering.

Firstly, we avoid the expensive CNN based depth estimation step by estimating depth during volume rendering for a single view. Warping information from sample volume rendered views to synthesise new views [12,15] is feasible when rendered images do not change dramatically between viewpoints. Zellmann et al. [36] proposed to warp images based on depth heuristics. Due to alpha compositing resulting in transparent surfaces without single depth values, the authors present multiple depth heuristics for image warping. Returning depth value at the voxel where the accumulated opacity along the ray reaches 80% during ray tracing achieved the best balance between speed and quality.

Secondly, we propose to apply a 2D CNN to improve the quality of the novel views as the 3D CNN from [26] is too slow for this problem. Although we have volumetric information in a light field, remapping the 3D volume to a 2D structure is faster and current deep learning architectures are often unable to fully exploit the power of 3D representations [20]. Due to limitations of 3D CNNs, Wang et al. [29] demonstrate how to map a 4D light field into a 2D VGG network [25] instead of using a 3D CNN. This is beneficial as the weights of a pre-trained 2D model can be updated. Additionally, although the 4D filters in 3D CNNs are intuitive to use on a 4D light field, the number of parameters quickly explode.

3 LIGHT FIELD SYNTHESIS

Our goal is to quickly synthesise a low-baseline light field for visualisation purposes from volumetric data. We demonstrate this for an 8×8 angular resolution light field, but there is no inherent limitation on this size other than GPU memory. The pipeline of the method has not changed from our original description in [13]. The following steps are involved in our light field synthesis (Figure 1).

1. Render a reference view by direct volume rendering and use a depth heuristic to estimate a depth map during ray casting.
2. Convert the depth map to a disparity map using the intrinsic camera parameters.
3. Apply backward image warping to the reference view using the disparity map to approximate a light field.
4. Apply a CNN to the warped images to improve visual consistency. This is modelled as a residual function which is added to the approximate light field from the previous step.

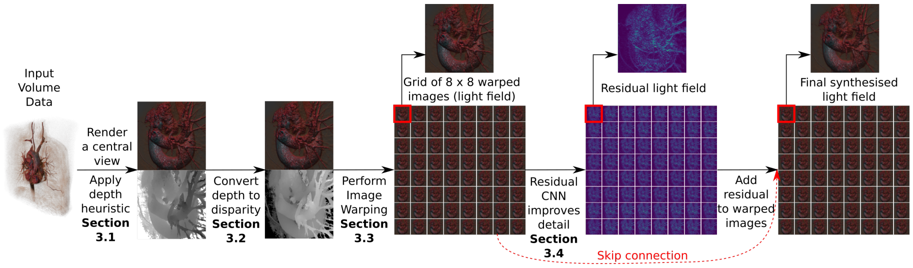


Fig. 1: Our proposed light field synthesis method can be broken down into distinct stages, including an initial depth heuristic calculation stage and a final CNN stage acting as a residual function to improve fine-grained detail. Extracted from our previous work [13].

3.1 Volume Depth Heuristics

To apply image warping based on a depth map, we quickly estimate depth values during the ray casting process. Due to the semi-transparent structures that frequently occur in volume rendering, there is no well defined depth and a heuristic is a necessity. An obvious approach is to use the depth of the first non-transparent voxel along the ray, but this is often ineffective due to near transparent volume information close to the camera. Using isosurfaces gives a good view of depth, but these must be recalculated during runtime if the volume changes.

In our previous work [13], we proposed to modify the best performing single pass depth heuristic from [36]. The heuristic from [36] is to save the depth of the voxel at which the opacity accumulated along a ray exceeds a pre-defined opacity threshold. A depth map formed in this way frequently lacks information at highly transparent regions as the opacity threshold needs to be set to a large value. To counteract this limitation, we applied a two-layered approach. A depth value is saved when a ray accumulates a low threshold opacity and overwritten if the ray accumulates the high threshold opacity. The layered approach can be repeated multiple times, but two layers was enough to capture the volume in most cases. This improved the quality of the depth map over isosurfaces and the best single pass method from Zelmann et al. [36]. A more detailed comparison of different depth heuristics is provided in our previous work [13].

3.2 Converting Depth to Disparity

This mathematical conversion is the same process as our previous paper [13]. During rendering, a depth value from the Z-buffer $Z_b \in [0,1]$ is converted to a pixel disparity value using the intrinsic camera parameters as follows. The depth buffer value Z_b is converted into normalised device co-ordinates, in the range $[-1,1]$, as $Z_c = 2 \cdot Z_b - 1$. Then, perspective projection is inverted to give depth in eye space as

$$Z_e = \frac{2 \cdot Z_n \cdot Z_f}{Z_n + Z_f - Z_c \cdot (Z_f - Z_n)} \tag{1}$$

Where Z_n and Z_f are the depths of the camera's near and far clipping planes in eye space, respectively. Note that Z_n should be set as close to the visualised object as possible to improve depth buffer accuracy, while Z_f has negligible effect on the accuracy. Given eye depth Z_e, it is converted to a disparity value d_r in real units using similar triangles [31] as

$$d_r = \frac{B \cdot f}{Z_e} - \Delta x \tag{2}$$

Where B is the camera baseline, or distance between two neighbouring cameras in the grid, f is the focal length of the camera, and Δx is the distance between two neighbouring cameras' principle points. Again, using similar triangles, the disparity in real units is converted to a disparity in pixels as

$$d_p = \frac{d_r W_p}{W_r} \tag{3}$$

Where d_p and d_r denote the disparity in pixels and real world units respectively, W_p is the image width in pixels, and W_r is the image sensor width in real units. If the image sensor width in real units is unknown, W_r can be computed from the camera field of view θ and focal length f as $W_r = 2 \cdot f \cdot \tan(\frac{\theta}{2})$.

3.3 Disparity Based Image Warping

The mathematical description of this operation is identical to our previous work [13]. To synthesise a novel view, a disparity map $D : \mathbb{R}^2 \mapsto \mathbb{R}$ is used to relate pixel locations in a novel view to those in the reference view. Let $I : \mathbb{R}^2 \mapsto \mathbb{R}^3$ denote a reference Red Green Blue (RGB) colour image at grid position $(u_r,\ v_r)$ with an associated pixel valued disparity map D. Then a synthesised novel view I' at grid position $(u_n,\ v_n)$ can be formulated as:

$$\begin{aligned} I'(x + d \cdot (u_r - u_n),\ y + d \cdot (v_r - v_n)) &= I(x,\ y) \\ \text{where} \quad d &= D(x,\ y) \end{aligned} \tag{4}$$

As opposed to the non-surjective hole producing forward warping operation, we apply a surjective backward warping. For each pixel in the novel view, information is inferred from the reference view. This results a hole free novel view but it is generally an oversampled from the reference view. Pixels in the novel view that require information from outside the border of the reference view instead used the closest border pixel. This strategy effectively stretches the border of the reference view in the absence of information. Since are dealing with low baseline light fields, this strategy is not overly restrictive.

3.4 Convolutional Neural Network

To improve the fine-grained details of the synthesised light field, we apply a CNN to the resulting images from image warping. The network is framed as a residual function that predicts corrections to be made to the warped images to reduce the synthesis loss in terms of mean squared error. The residual light field has full range over the colour information to allow for removal of predicted erroneous information and addition of predicted improvements. Srinivasan et al. [26] applied 3D convolutions to achieve a similar goal. Because the light field is 4D, 3D CNNs which use 4D filters are intuitive to apply to this problem. However, using 2D convolutions is advantageous as less parameters are required for a 2D CNN. Additionally, more pretrained models and better optimisation tools exist for 2D CNNs than 3D CNNs. Wang et al. [29] previously demonstrated strong evidence that 3D CNNs can be effectively mapped into 2D architectures.

To test remapping the 3D network from [26] into a 2D network, we compared four CNNs architectures to improve the synthesised light field [13]. The CNNs were tested on a single heart MRI with a fixed transfer function from different camera positions. Each network was evaluated based on the difference to the baseline result of pure geometrical image warping in terms of Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM). In the experiments, all networks improved the resulting light field quality measured by SSIM. As such, the 3D CNN Srinivasan et al. [26] could be effectively remapped into a 2D architecture. In our tests, the best performing network in terms of quality was a slightly modified Enhanced Deep Super-Resolution (EDSR) network [8] taking an angular remapped input from [29], achieving an average of 0.923 SSIM on the validation set [13].

We briefly describe the 3D occlusion prediction network from Srinivasan et al. [26], and the modified EDSR network [8] we use in its place. The 3D network is structured as a residual network with $3 \times 3 \times 3$ filters that have access to every view. The input to the 3D network is all warped images and colour mapped disparity maps. Our EDSR network is the same as the original EDSR network from [8], bar removal of spatial upscaling at the last layer and application of tanh activation at the final layer. Spatial upscaling is removed as we only require angular resolution enhancement, while tanh is applied to allow residual output to add and remove information. To map the light field input into the three colour channel RGB input required for EDSR, angular remapping from [29] is applied. Angluar remapping transforms an $n \times m$ angular resolution light field with $x \times y$ spatial resolution into an $(n \cdot x) \times (m \cdot y)$ image. In this remapped image, the uppermost $n \times m$ pixels contain the upper-left pixel from each of the original $n \times m$ light field views. There are two significant differences in this EDSR network from the network of Srinivasan et al. [26], besides being a 2D network. Firstly, the disparity map is not input to the network, reducing the number of input channels. Secondly, the 3×3 filters used in this network only consider the nearest neighbours to a view and this local connectivity is very efficient.

4 IMPLEMENTATION

Note that these test sytem specifications differ from our previous work [13]. All new experiments were performed on a computer with 8GB memory, an Intel(R) Xeon(R) CPU E5-1620 v3 @ 3.50GHz Central Processing Unit (CPU), and a NVIDIA GeForce RTX 2080 GPU running on Ubuntu 16.04. For deep learning, the PyTorch library [17] was used with Cuda 10.0, cuDNN 7.1.2, and NVIDIA driver version 410.104.

4.1 Speed Improvements

Compared to our previous work [13], we have improved the implementation and speed of multiple aspects of this research, but the result is functionally identical. Firstly, the light field capturing approach was improved over rendering each view in the light field separately. This enables faster data capture times, and more importantly allows for a fairer baseline comparison of our method's speed. This was achieved by creating two large textures containing the entry and exit positions into the volumetric data for each ray in the light field, and casting all of these rays in one step. Traversing the rays in a single pass made better advantage of the highly parallel nature of GPUs. The resulting method is roughly twice as fast as rendering views individually, leading to a fairer speed comparison.

Additionally, the image warping procedure was improved considerably. Instead of performing the image warping on the CPU using NumPy and PyTorch, fixed function GPU shader code was implemented. First, the depth map from the reference view is converted to a disparity in a fragment shader, and saved to a depth texture. Then, the reference view colour texture and depth texture are passed to another fragment shader which performs backward warping for each novel view, saving the result into multiple viewports of a large colour texture. Unsurprisingly, this is far faster than the CPU warping, close to two orders of magnitude faster.

4.2 Datasets

Using a 2D array of outward looking (non-sheared) perspective views with fixed field of view [7], synthetic light field datasets were captured in the volume visualisation framework Inviwo [27]. The cameras were shifted along an equidistant grid to keep their optical axes parallel, removing the need to rectify the images to a common plane. Sampling was performed uniformly, with the camera fixed to lie within a certain distance of the central object. Additionally, a plane with its normal vector aligned with the camera view direction is used to clip the volume, revealing detailed structures inside the volume and demonstrating the accuracy of the depth heuristic. See Figure 2 for the central sub-aperture image of five captured light fields for each dataset.

In our prior research [13], the validity of our method was demonstrated for one specific dataset, an MRI of a heart with visible aorta and arteries with a resolution of $512 \times 512 \times 96$ [21]. Each of the 2000 sample training light fields and 100 validation light fields were captured at 512×512 spatial resolution. This was a difficult dataset because the heart has a rough surface, and the aorta and arteries create intricate structures which are difficult to reconstruct. However, the colours in the transfer function were quite dark with low contrast. Additionally, the CNN could learn a specific single case and generalisation was not fully tested. To address the limitations in these experiments, we tested a new dataset with far more variety.

Using multiple head MRIs from a large scale neurological study [19] we captured a highly varying dataset. Many hand designed functions are applied to these images, as well as randomly generated transfer functions. The density values in each MRI volume were normalised to a common data range, so the transfer functions acted similarly

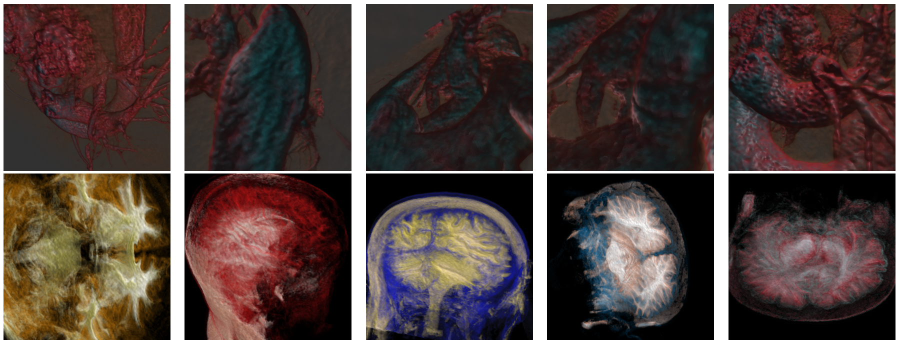


Fig. 2: Sample training light field central sub-aperture views. Heart MRI images are from our previous work [13]. It should be immediately apparent that the dataset of head MRIs has far more variation.

across the volumes. The generalisation of the deep learning could be evaluated by running the network against unseen MRIs volumes and unseen transfer functions during training. In this manner, we could test if the volume or transfer function had more of an effect on the generalisation of learning. The head MRI volumes are smaller than the heart dataset, so these were captured at 256×256 spatial resolution. 2000 training light fields are captured, with 400 validation light fields for non-training volumes and 400 light fields for non-training transfer functions.

4.3 Training Procedure

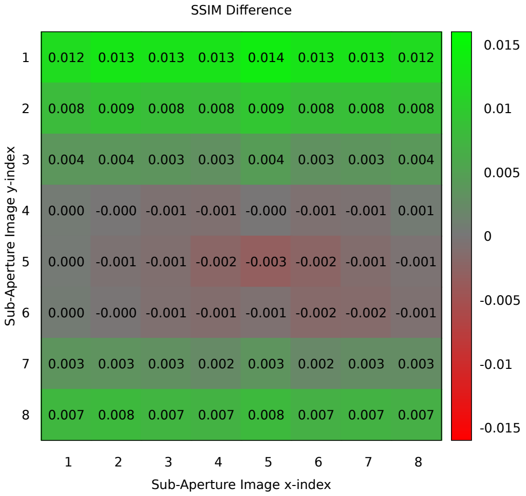
The training procedure for the convolutional network is identical to that described in [13]. The CNNs are trained by minimising the per-pixel mean squared error between the ground truth views and the synthesised views. For the heart dataset, to increase training speeds and the amount of available data, four random spatial patches of size 128×128 were extracted from each light field at every training epoch. Training colour images for both datasets have a random gamma applied as data augmentation. Network optimisation was performed with Stochastic gradient descent and Nesterov momentum. An initial learning rate of 0.1 was updated during learning by cosine annealing the learning rate with warm restarts [11]. Gradients were clipped based on the norm at a value of 0.4 and an L2 regularisation factor of 0.0001 was applied. Training takes about 14 hours using the 2D CNN architecture with eight CPU cores used for data loading and image warping.

5 EXPERIMENTS

5.1 Generalisation of Deep Learning

We previously tested multiple CNNs on a single volume and transfer function, with different camera positions [13]. As discussed in Section 3, the best performing network in terms of quality was the EDSR network [8] taking an angular remapped input.

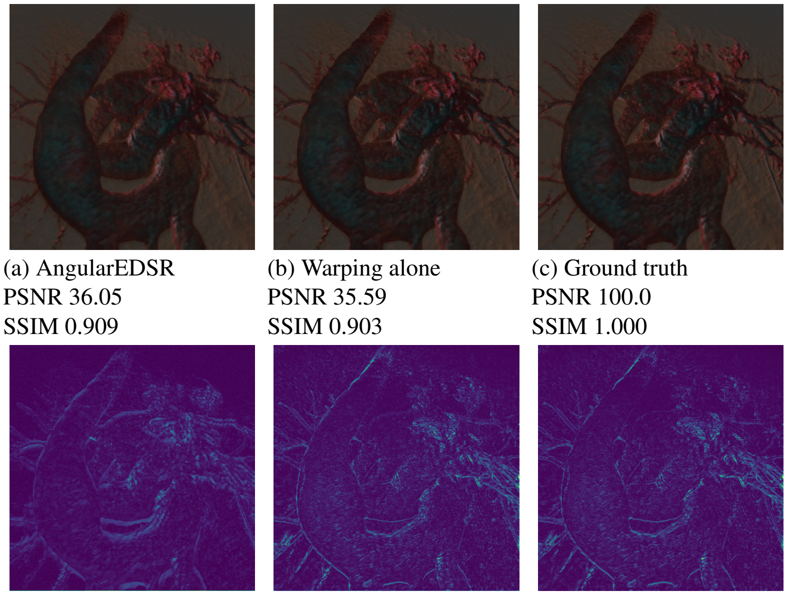
This resulted in a residual function which improved the visual quality of novel views, especially for those far from the reference view, see Figure 3. Additional evaluation performed with PSNR andthe Learned Perceptual Image Patch Similarity (LPIPS) metric [37] using the deep features of AlexNet [5] to form a perceptual loss function agrees with the per image values for SSIM. See Figure 4 for the bottom right sub-aperture view of this light field from a validation set along with difference images to visualise the effect of the CNN.



	1	2	3	4	5	6	7	8
1	0.012	0.013	0.013	0.013	0.014	0.013	0.013	0.012
2	0.008	0.009	0.008	0.008	0.009	0.008	0.008	0.008
3	0.004	0.004	0.003	0.003	0.004	0.003	0.003	0.004
4	0.000	-0.000	-0.001	-0.001	-0.000	-0.001	-0.001	0.001
5	0.000	-0.001	-0.001	-0.002	-0.003	-0.002	-0.001	-0.001
6	0.000	-0.000	-0.001	-0.001	-0.001	-0.002	-0.002	-0.001
7	0.003	0.003	0.003	0.002	0.003	0.002	0.003	0.003
8	0.007	0.008	0.007	0.007	0.008	0.007	0.007	0.007

Fig. 3: The difference in SSIM per image location after applying the EDSR network to the warped images. Images far away from the reference view exhibited lower loss, but the CNN caused a degradation in quality of the reference image. Position $(5,5)$ is the location of the reference view. Extracted from our previous work [13].

Although the learnt residual function worked well for a single volume, our initial tests indicated that this learnt function would not generalise for multiple transfer functions and volumes [13]. In contrast, the image warping procedure and depth heuristic did generalise well across volumes and transfer functions. Using our dataset consisting of multiple different head MRIs and transfer functions from [19], we found that the learning certainly did not generalise. Due to the lack of a clear pattern in this widely varying data, the CNN quickly learnt to regress the residual function to a blank output. This demonstrates that the CNN could not predict effective changes for multiple volumes and transfer functions. Although this a significant limitation of this method, it is an extremely important result. In particular, it is necessary to train the CNN to handle specific volumes with fixed transfer functions. For some cases, such as an educational demonstration, this would be feasible as a specific set of volumes could be selected that elucidate a topic.



(a) AngularEDSR
PSNR 36.05
SSIM 0.909

(b) Warping alone
PSNR 35.59
SSIM 0.903

(c) Ground truth
PSNR 100.0
SSIM 1.000

(d) Difference
of (a) and (b)

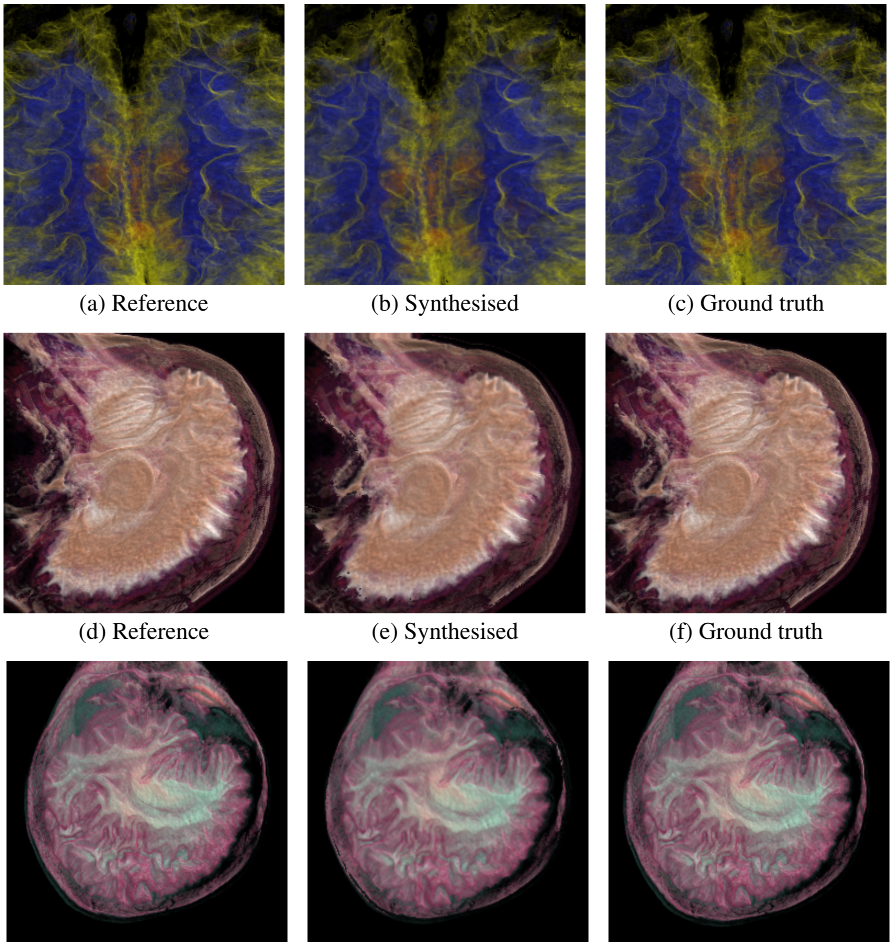
(e) Difference
of (a) and (c)

(f) Difference
of (b) and (c)

Fig. 4: The bottom right view in the light field which Figure 3 presents results for. Figure (d) visualises the residual applied by the CNN to the warped images to improve visual quality. The CNN detects broad edges to improve, such as the central arch of the aorta, but fails to improve finer details such as the arteries in the top right of the image. Extracted from our previous work [13].

5.2 Example Synthesised Light Fields

To investigate the method performance, examples of a low, middling, and high quality synthesised light fields from the validation sets are presented in Figures 5 and 6. For the head MRI dataset, Figure 5(b) is a poor reconstruction due to a large number of cracks appearing in the wavy semi-transparent structure of the cerebral cortex. Figure 5(e) is a reasonably well synthesised view, but skull and brain's borders are inaccurately estimated. Figure 5(h) is an accurate synthesis as the information is moved very well to the novel view, though the image is blurry. For the heart MRI dataset, Figure 6(b) is a poor reconstruction due to the opaque structure that should be present in the centre of the view causing a large crack in the image. Figure 6(e) is a reasonably well synthesised view, though some arteries lose their desired thickness and the image is not very sharp. Figure 6(h) is an accurate synthesis, although some errors are seen around object borders, such as on the arch of the aorta.



(a) Reference (b) Synthesised (c) Ground truth

(d) Reference (e) Synthesised (f) Ground truth

(g) Reference (h) Synthesised (i) Ground truth

Fig. 5: Example synthesised upper-left images for the head MRI. The first row has low performance (25.8 PSNR, 0.60 SSIM) due to the abundant translucent structures. The second row has middling performance (25.3 PSNR, 0.75 SSIM) since the background has erroneously been picked up in the novel view information. The third row has high performance (28.0 PSNR, 0.86 SSIM) with small inaccuracies, including a general blurriness and splitting on the lower edge of the head.

5.3 Time Performance

All timing performances are reported on 300 $512 \times 512 \times 8 \times 8$ light fields captured in Inviwo of the heart MRI discussed in Section 4.2. In our previous work [13], we assumed that the light field was rendered view by view. With this assumption, rendering such a light field takes 1.25s on average with 93ms std dev. With the improved method taking advantage of large textures, rendering a light field takes 612ms on average, with

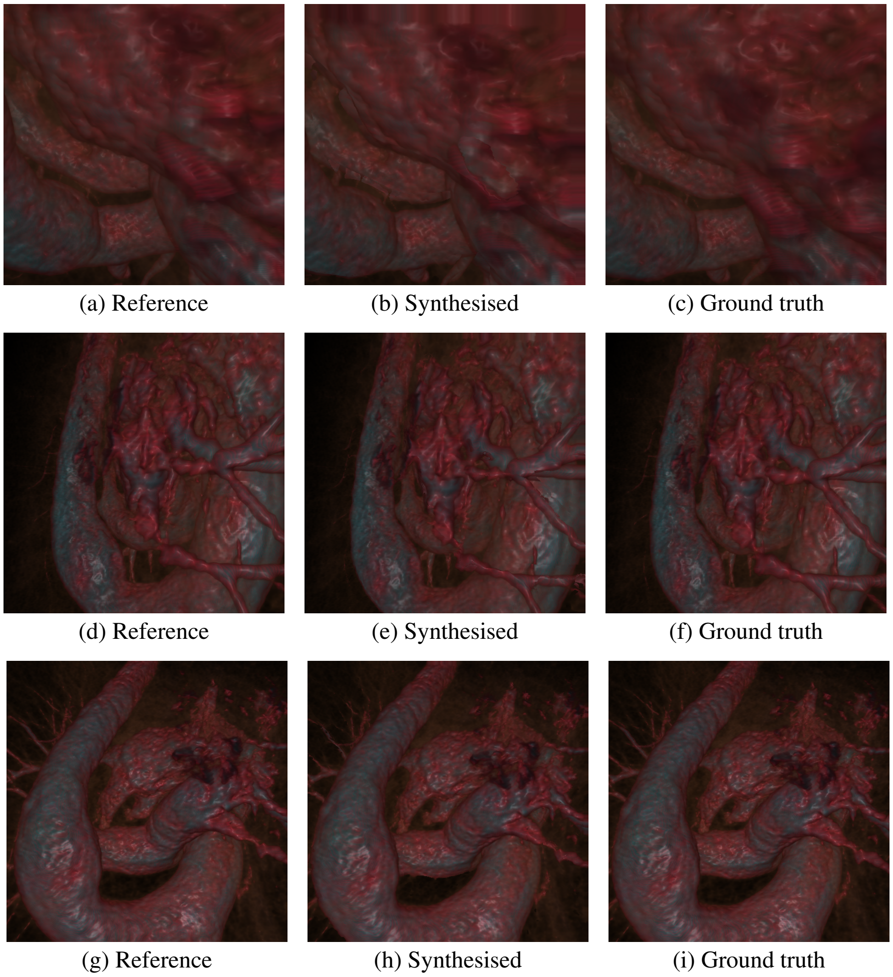

(a) Reference
(b) Synthesised
(c) Ground truth
(d) Reference
(e) Synthesised
(f) Ground truth
(g) Reference
(h) Synthesised
(i) Ground truth

Fig. 6: Example synthesised upper-left images for the heart MRI. The first row has low performance (29.0 PSNR, 0.90 SSIM) due to the translucent structures at the front of the view. The second row has middling performance (31.7 PSNR, 0.86 SSIM) since the arteries are not perfectly distinguished from the aorta. The third row has high performance (35.3 PSNR, 0.91 SSIM) with small inaccuracies, such as on the lower right edge of the aorta.

standard devation of 20ms. Using our method is an nearly an order of magnitude faster than this, taking 72ms on average, with negligible standard deviation of less than 0.1ms. This is 72ms is broken down in Table 1, with average values given over all 300 light fields. The timing for the CNN includes the time to transform the values in the GPU texture to a PyTorch CUDA tensor, which could potentially be alleviated by reusing the GPU memory. The GLSL shader code implementation of our image warping operation is two orders of magnitude faster than our previous pytorch implementation [13] with a time reduction from 2.77 seconds to 10ms. Note that the whole pipeline takes only 32ms to synthesise a light field without applying a CNN. As such, a 512 × 512 volume could be visualised in a 512 × 512 × 8 × 8 light field at interactive rates of 30 frames per second.

Table 1: Average timing values in ms.

Operation	Average Time (ms)
Rendering the reference view	19ms
Computing the depth heuristic	1ms
Converting depth to disparity	2ms
Image warping	10ms
CNN	40ms

As aforementioned, the time for light field synthesis has far less deviation than directly volume rendering a light field, because the latter depends heavily on the complexity of the scene. A CNN performs the same operations regardless of input volume size and complexity, which results in steady performance. The fixed function backward warping and disparity conversion shaders also perform the same operation regardless of the volume size and complexity. For a fixed spatial resolution, the only time variable is rendering the single sample view. Accordingly, this method could be applied to very large complex volumes with expensive rendering techniques.

6 CONCLUSION

Synthesising light fields by image warping produces a useful visualisation method that runs in real-time producing a 512 × 512 × 8 × 8 light field in an eight of the time to render the light field by ray casting. Additionally, the rendering time for our method is fixed after rendering the central view regardless of the size of the volume or the transfer function applied. A signification limitation of this approach is that the CNN must be retrained for every combination of volume and transfer function. However, a purely image warping based visualisation would still be useful for exploratory purposes as the exact result can be rendered when the camera is held fixed. There is still significant limitations in the quality of the result beyond low-baseline light fields, as a single sample view is not sufficient to extrapolate for a wide-baseline. For wide-baselines, we believe that a fruitful avenue could be to combine multiplane image approaches [10,14,38] for a new volume targeted approach.

ACKNOWLEDGEMENTS

This research has been conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number 13/IA/1895.

References

1. Adelson, E.H., et al.: The plenoptic function and the elements of early vision. In: Computational Models of Visual Processing. pp. 3–20. MIT (1991)
2. Frayne, S.: The looking glass. https://lookingglassfactory.com/ (2018), accessed: 22/11/2018
3. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F.: The lumigraph. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. pp. 43–54. SIGGRAPH '96, ACM (1996). https://doi.org/10.1145/237170.237200, http://doi.acm.org/10.1145/237170.237200
4. Kalantari, N.K., Wang, T.C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. ACM Transactions on Graphics **35**(6), 193:1–193:10 (Nov 2016). https://doi.org/10.1145/2980179.2980251, http://doi.acm.org/10.1145/2980179.2980251
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
6. Lanman, D., Luebke, D.: Near-eye light field displays. ACM Transactions on Graphics (TOG) **32**(6), 220 (2013)
7. Levoy, M., Hanrahan, P.: Light Field Rendering. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 31–42. SIGGRAPH '96, ACM (1996)
8. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: The IEEE conference on computer vision and pattern recognition (CVPR) workshops. vol. 1, p. 4 (2017)
9. Lin, Z., Shum, H.Y.: A geometric analysis of light field rendering. International Journal of Computer Vision **58**(2), 121–138 (2004). https://doi.org/10.1023/B:VISI.0000015916.91741.27, https://doi.org/10.1023/B:VISI.0000015916.91741.27
10. Lochmann, G., Reinert, B., Buchacher, A., Ritschel, T.: Real-time Novel-view Synthesis for Volume Rendering Using a Piecewise-analytic Representation. In: Vision, Modeling and Visualization. The Eurographics Association (2016)
11. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (2017)
12. Mark, W.R., McMillan, L., Bishop, G.: Post-rendering 3d warping. In: Proceedings of the 1997 symposium on Interactive 3D graphics. pp. 7–16. ACM (1997)
13. Martin, S., Bruton, S., Ganter, D., Manzke, M.: Using a Depth Heuristic for Light Field Volume Rendering. pp. 134–144 (May 2019), https://www.scitepress.org/PublicationsDetail.aspx?ID=ZRRCGeI7xV8=&t=1
14. 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. arXiv preprint arXiv:1905.00889 (2019)
15. Mueller, K., Shareef, N., Huang, J., Crawfis, R.: Ibr-assisted volume rendering. In: Proceedings of IEEE Visualization. vol. 99, pp. 5–8. Citeseer (1999)
16. Park, S., Kim, Y., Park, S., Shin, J.A.: The impacts of three-dimensional anatomical atlas on learning anatomy. Anatomy & Cell Biology **52**(1), 76–81 (Mar 2019). https://doi.org/10.5115/acb.2019.52.1.76, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6449593/
17. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)

18. Penner, E., Zhang, L.: Soft 3d reconstruction for view synthesis. ACM Transactions on Graphics **36**(6), 235:1–235:11 (Nov 2017). https://doi.org/10.1145/3130800.3130855, http://doi.acm.org/10.1145/3130800.3130855
19. Poldrack, R.A., Congdon, E., Triplett, W., Gorgolewski, K., Karlsgodt, K., Mumford, J., Sabb, F., Freimer, N., London, E., Cannon, T., et al.: A phenome-wide examination of neural and cognitive function. Scientific data **3**, 160110 (2016)
20. Qi, C.R., Su, H., Nießner, M., Dai, A., Yan, M., Guibas, L.J.: Volumetric and multi-view cnns for object classification on 3d data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5648–5656 (2016)
21. Roettger, S.: Heart volume dataset. http://schorsch.efi.fh-nuernberg.de/data/volume/Subclavia.pvm.sav (2018), accessed: 15/08/2018
22. Shade, J., Gortler, S., He, L.w., Szeliski, R.: Layered depth images. In: Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques. pp. 231–242. SIGGRAPH '98, ACM, New York, NY, USA (1998). https://doi.org/10.1145/280814.280882, http://doi.acm.org/10.1145/280814.280882
23. Shi, L., Hassanieh, H., Davis, A., Katabi, D., Durand, F.: Light field reconstruction using sparsity in the continuous fourier domain. ACM Transactions on Graphics **34**(1), 1–13 (dec 2014). https://doi.org/10.1145/2682631
24. Shojaii, R., Bacopulos, S., Yang, W., Karavardanyan, T., Spyropoulos, D., Raouf, A., Martel, A., Seth, A.: Reconstruction of 3-dimensional histology volume and its application to study mouse mammary glands. Journal of Visualized Experiments: JoVE (89), e51325 (Jul 2014). https://doi.org/10.3791/51325
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
26. Srinivasan, P.P., Wang, T., Sreelal, A., Ramamoorthi, R., Ng, R.: Learning to synthesize a 4d rgbd light field from a single image. In: IEEE International Conference on Computer Vision (ICCV). pp. 2262–2270 (Oct 2017). https://doi.org/10.1109/ICCV.2017.246
27. Sundén, E., Steneteg, P., Kottravel, S., Jonsson, D., Englund, R., Falk, M., Ropinski, T.: Inviwo - an extensible, multi-purpose visualization framework. In: IEEE Scientific Visualization Conference (SciVis). pp. 163–164 (Oct 2015). https://doi.org/10.1109/SciVis.2015.7429514
28. Vagharshakyan, S., Bregovic, R., Gotchev, A.: Light field reconstruction using shearlet transform. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(1), 133–147 (jan 2018). https://doi.org/10.1109/tpami.2017.2653101
29. Wang, T.C., Zhu, J.Y., Hiroaki, E., Chandraker, M., Efros, A.A., Ramamoorthi, R.: A 4d light-field dataset and cnn architectures for material recognition. In: European Conference on Computer Vision. pp. 121–138. Springer (2016)
30. Wanner, S., Goldluecke, B.: Variational light field analysis for disparity estimation and super-resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(3), 606–619 (March 2014). https://doi.org/10.1109/TPAMI.2013.147
31. Wanner, S., Meister, S., Goldluecke, B.: Datasets and benchmarks for densely sampled 4d light fields. In: Vision, Modeling, and Visualization (2013)
32. Wu, G., Liu, Y., Dai, Q., Chai, T.: Learning sheared epi structure for light field reconstruction. IEEE Transactions on Image Processing **28**(7), 3261–3273 (July 2019). https://doi.org/10.1109/TIP.2019.2895463
33. Wu, G., Masia, B., Jarabo, A., Zhang, Y., Wang, L., Dai, Q., Chai, T., Liu, Y.: Light field image processing: An overview. IEEE Journal of Selected Topics in Signal Processing **11**(7), 926–954 (oct 2017). https://doi.org/10.1109/jstsp.2017.2747126
34. Wu, G., Zhao, M., Wang, L., Dai, Q., Chai, T., Liu, Y.: Light field reconstruction using deep convolutional network on epi. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1638–1646 (July 2017). https://doi.org/10.1109/CVPR.2017.178

35. Yoon, Y., Jeon, H.G., Yoo, D., Lee, J.Y., So Kweon, I.: Learning a deep convolutional network for light-field image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 24–32 (Dec 2015). https://doi.org/10.1109/ICCVW.2015.17
36. Zellmann, S., Aumüller, M., Lang, U.: Image-based remote real-time volume rendering: Decoupling rendering from view point updates. In: ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. pp. 1385–1394. ASME (2012)
37. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
38. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)