

Investigating 3D holoscopic visual content upsampling using super-resolution for cultural heritage digitization

Abdelhak Belhi^{1,2,✉}, Abdelaziz Bouras¹, Taha Alfaqheri³,
Akuha Solomon Aondoakaa³, Abdul H. Sadka³

¹ CSE, Qatar University, Doha, Qatar

{abdelhak.belhi, abdelaziz.bouras}@qu.edu.qa

² DISP Laboratory, University Lumière Lyon 2, Lyon, France

³ Brunel University, London, United Kingdom

{taha.alfaqheri, akuha.aondoakaa, abdul.sadka}@brunel.ac.uk

Abstract.

Through this paper, we aim at investigating the impact of using deep learning-based technologies such as super-resolution on Holoscopic 3D (H3D) images. Holoscopic 3D imaging is a technology that aims at providing cost-effective alternatives for 3D content viewing and consumption without requiring a special headgear or posture. The technique is using a special lens array fitted to standard DSLR or mirrorless cameras to generate or capture 3D content. The output is a Holoscopic 3D image that can be displayed in lightfield displays or Multiview displays following a post-processing procedure. The main advantage of this technique is its cost-effectiveness in viewing and interacting with 3D content. However, one of its drawbacks is the low spatial density of the commercial cameras CMOS sensors and the lens induced imperfections. The latter can be fixed in software using some distortion correction techniques. However, the former is still challenging in terms of techniques that result in naturally looking output. Mitigating such issues with hardware will lead to higher costs and the technique loses its main advantage. Our approach consists of designing a framework that leverages software tools in order to upscale the output of H3D cameras whilst solving the low spatial density problem of H3D images. We also investigate the impact of deep learning-based video motion interpolation on the output quality of the cultural H3D imaging framework.

Keywords: Cultural heritage, Deep learning, Super-resolution, 3D holoscopic imaging.

1 Introduction

With the growing need for new and more attractive mediums for content consumption and interaction, 3D technologies introduced an attractive solution [1]. For a long time, there were only two types of 3D content, either designed or scanned. However, there were no effective ways to preserve the spatial information when displaying or reproducing the captured asset. Mostly, the 3D copy was rendered, and viewpoints were

displayed on 2D screens. Several approaches were used such as stereoscopy, Multiview, Virtual Reality, etc. [2].

The impact of 3D cinema pushed major players in content creation and multimedia hardware manufacturing to promote 3D vision through 3D cameras and 3D screens (mainly stereoscopic). The commercialized 3D technology on consumer level TVs is based on stereoscopic vision which relies on feeding a left image to the left eye and a right image to the right eye through either a spatial or temporal multiplexing using a special type of polarized glasses [3]. The next big improvement to 3D vision was Multiview displays (autostereoscopic) which consists of displaying many pairs of videos (for left and right eyes) so that the viewer can perceive a pair of views from each position within the specified view angle without wearing any headgear or glasses. Unfortunately, these two solutions have some drawbacks related to the comfort of the viewer (eye or head fatigue) or low-quality output [1] as these solutions rely on, or fool the human brain in thinking that the image viewed is in 3D either by wearing a special type of glasses or by looking to a screen from a certain angle. As a solution for preserving the spatial information in 3D when displaying the asset, some chose 3D printing in order to replicate the shape of the asset, but this is unfortunately impractical in the real world [4].

The main drawback of Multiview and stereoscopic displays is the fact that they do not provide a true 3D representation of the content. They rather rely on the human brain to fuse the pairs of images which can lead to headaches and eye fatigue etc. Some research work addressed these issues, but some intrinsic eye fatigue will always exist with stereoscopic 3D technologies [1].

To solve these limitations, many researchers are looking for alternatives to capture and display true 3D content. The main developed techniques rely on either holography or holoscopic imaging [3, 5]. Holography, however, is still at development levels as there are multiple limitations on how to control the light fields [1]. Holoscopic imaging (or also integral imaging) in contrast, provides a simple, cost-effective alternative. The principle requires a special macro lens array (MLA) fitted to a camera sensor. Each lens in the array captures the scene from a slightly shifted angle. At the display stage, the process is reversed, and the viewer will perceive a true 3D representation of the content without wearing any kind of headgear or having to look at the screen from a specific posture [3].

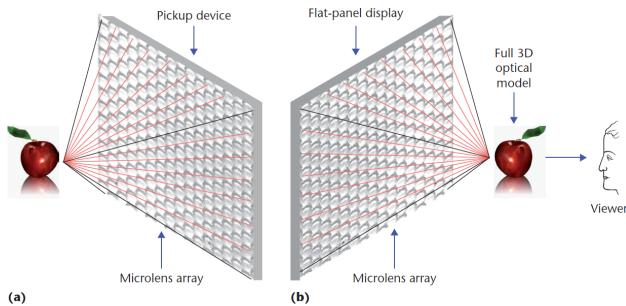


Fig. 1. 3D Holoscopic capture and display principle

Thanks to many research works and with the advanced lens manufacturing technologies, the H3D technique is nowadays an efficient alternative for 3D content capturing and consumption. The technique is currently being adapted for cultural heritage digital preservation in multiple initiatives such as the CEPROQHA Project, funded by the Qatar National Research Fund [6, 7]. The project is introducing multiple advances in 3D holoscopic imaging mainly in the capture, processing, and display of 3D holoscopic content. The technique is currently being adapted to fit the requirements of cultural data preservation such as improving the output quality [6].

Unfortunately, the H3D technique still presents numerous challenges related to the output quality which is primarily related to limitations in the capturing hardware used. Although high-density CMOS sensors can address this challenge, the main advantage of the H3D technique being its cost-effectiveness will be lost. Mitigating the low-density issue through software is thus a must. Recent successes in deep learning proven that such limitations could be addressed through software-based solutions. In fact, recent super-resolution techniques that are designed to upscale and increase the spatial density of visual content can be adapted to mitigate the limitations found in the H3D imaging technique.

The contributions discussed and presented in the present paper could be summarized as follows:

- Construction of an H3D dataset of image and video of cultural content collected in a professional studio environment (to get the best possible conditions).
- A review and performance comparison of most notable Single Image Super-Resolution (SISR) techniques on H3D visual content with highlights of the advantages of each technique over the others.
- Design and implementation of an H3D post-processing framework for content up-sampling through super-resolution and first experiments on H3D video frame interpolation.

In this paper, we aim at investigating the impact of applying super-resolution techniques for 3D holoscopic images. The context of our study is cultural heritage digitization where there is an increasing need for digital preservation and multimedia tools. The main driver of this study is related to limitations found in the H3D acquisition framework and that we strongly believe could be addressed cost-effectively by software and deep learning-based tools. These limitations are mainly due to the low pixel density induced by the use of commercial cameras and which need to be solved through software to preserve the main advantages of the H3D technique.

The rest of the present paper is organized as follows. In section two, we present a literature review of the different state-of-the-art techniques used within this study. Section three describes the tools and methods we used to design and implement our H3D post-processing framework which is based on single image super-resolution and video motion interpolation. In section four, we define our experimental setup in addition to a presentation and a discussion of our results. In section five, we draw our conclusions and give some perspectives on future work to improve our framework.

2 Related work

In this section, we present the background and the research advances related to the context of our study. This includes the impact of digital technologies on cultural heritage assets, technical details about the 3D holoscopic technology, an introduction to deep learning and some details about the single image super-resolution.

2.1 Cultural heritage digitization

Cultural heritage is the most effective medium for history and knowledge transfer between generations and civilizations [7]. These assets are often exhibited in museums, archeological sites, and art galleries. Unfortunately, these assets face lots of risks due to their degradation or due to other external factors. Nowadays, digital preservation attracts a lot of attention especially due to the proven performance of IT infrastructures and the new high-quality ways of content consumption [8, 9]. This means that cultural collections will be accessible to larger audiences from anywhere in the world. This also means that efficient alternatives to physical preservation can be developed which will decrease costs and efficiently improve the impact of cultural heritage through innovative exhibition ways such as virtual museums or galleries, VR, AR, virtual interaction etc. However, in museums particularly, interaction with cultural assets is nowadays almost not existent as these assets are often protected with glass shields, fixed, or away from visitors. Having these assets in a digital form will effectively enable such interaction, but the need to preserve the fine details of the assets is very challenging. The reason is that 3D scanning technologies are very costly (1000 USD per scan using the CultLab3D scanner) [10], and photogrammetry is often inefficient as it requires a lot of trial and error which often results in lost details, etc. [11]. 3D holoscopic imaging, in contrast, provides a new paradigm of capture and display that can be applied to cultural heritage assets [12]. The acquisition gear is relatively cheap in comparison with other technologies such as 3D scanning, and the output is theoretically a true 3D representation of the asset that can be displayed in either lightfield or Multiview displays [12]. However, there are still challenges regarding the quality of the output and the interaction with the assets.

2.2 3D Holoscopic Imaging

The 3D holoscopic technology is not recent. Its principle was proposed in 1908 by Lippmann [13]. The technology is often referred to as lightfield imaging. The principle is inspired by Fly's eyes using an evenly spaced macrolens array fitted to a normal camera (either DSLR or mirrorless) [3]. Each of these lenses captures the scene from a slightly shifted angle in comparison with neighboring lenses in the array. The fundamental principle of H3D is described by Fig. 1. The lightfield data is recorded by the CMOS sensor and will be stored as a 2D capture. At the display stage, the same process used for capture is reversed. A MLA is placed in front of the screen, and the object can be reconstructed in space [3, 5].

Many researchers worked on designing and manufacturing a handheld holoscopic camera with similar capabilities as conventional 2D cameras, i.e., camera focal length,

exposure and the ability to capture viewpoints [14]. The raw holoscopic data was thus defined as a projection of the high-dimensional light signal onto the camera sensor plane [15] whilst preserving spatial and angular information. The H3D data structure can be described as a set of light rays captured in 3D space with different angles and directions. In contrast, 2D conventional cameras record pixel values while discarding the information related to the direction of light [15-17].

The acquired angular information can be leveraged to generate multiple image formats, and this represents the main additional feature in comparison with 2D image data. The acquired angular information is used to extract multiple images from several viewpoints with just a single camera capture. Many research studies worked on different holoscopic processing stages; H3D camera capturing stage [3, 14, 18, 19], post-processing/image quality enhancement, reformatting and adaptation [1, 3], and light field visualization.

One of the limitations of the H3D capturing framework is the low spatial density of elemental images. This limitation can be mitigated through hardware-based solutions (larger CMOS sensors) but the technique may lose one of its main advantages which is its cost-effectiveness. To mitigate low spatial resolution effects, many research studies focused on enhancing the low spatial resolution using deep learning techniques; Wang and his colleagues implemented a bidirectional recurrent convolutional neural network, their technique aims to find a *spatial relation* between horizontally or vertically adjacent sub-aperture images of light-field data [20]. They developed a framework to improve light field image resolution by combining SISR deep CNN and elemental Epipolar Plane Image (EPI) enhancement deep CNN. Their primary goal is to generate light field images with more geometric consistency. The researchers in [21] presented a method to synthesize new views from a sparse set of input views. A robust and straightforward super-resolution method for light field images is presented in [22]. Wang and his colleagues used in their method a projection-based Light Field Super-resolution (LFSR) solution without prior information based on a redefinition of the mapping function between disparity and shearing shift. This can provide a more consistent representation of the spatial resolution of 4D light fields, which does not require any additional camera parameters or settings compared with former projection-based LFSR methods. However, this method requires generating a map from 2D lenslet images to 4D light field representation data using geometric optic rules.

The authors of [15], worked on improving both the spatial and the temporal resolutions of light field data using Convolutional Neural Networks (CNNs). They used a Lytro camera to generate the raw data and test the performance of the proposed algorithm. The outcome of this work is similar to the one presented in [23], but the main difference is in the implementation approach. The work in [23] focuses on perspective images while in [15] the authors focus on upsampling the raw light field data. The authors of [21] used two sequential convolutional neural networks to model disparity and color information. The main aim of this work is to increase the spatial resolution at the expense of decreasing the angular information.

In [24], the authors presented a light field image resolution enhancing technique based on deep learning and epipolar plane images (EPI). Their combined framework

provides good spatial resolution by enhancing each sub-aperture image separately using super-resolution SR deep CNN. To ensure consistent light field image geometry, they proposed an Epipolar Plane Image enhancement deep CNN in their implementation.

In this paper, we propose a post-processing framework intended to mitigate the low spatial density limitation induced by the commercial cameras CMOS sensors. We investigate and compare the performance of state-of-the-art 2D super-resolution framework on raw H3D images.

2.2.1. Holoscopic 3D imaging technology

The Holoscopic 3D camera records the angular and spatial information of any given scene; this is made possible due to the omnidirectional micro macrolens array (MLA) placed right before the camera imaging sensor as shown in Fig. 2.

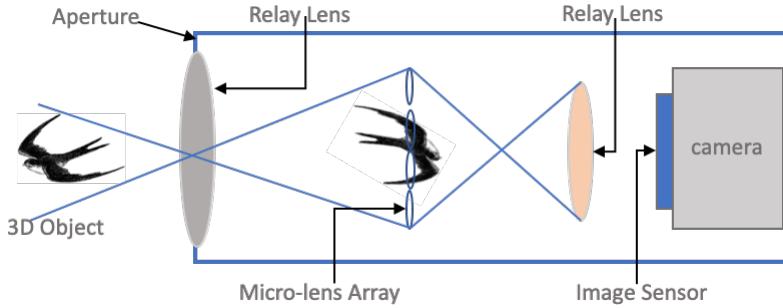


Fig. 2. Lens architecture of a Holoscopic 3D imaging technology

The first handheld prototype Holoscopic 3D camera was introduced by Adelson [25]. Fig. 3 shows a Schematic diagram of a Holoscopic 3D camera. A projected light ray with different angles and directions can be decoded into a set of Elemental Holoscopic images with different viewing angles. The output recorded images in the camera sensor can be represented in 4D data (See Fig. 4) [1, 20].

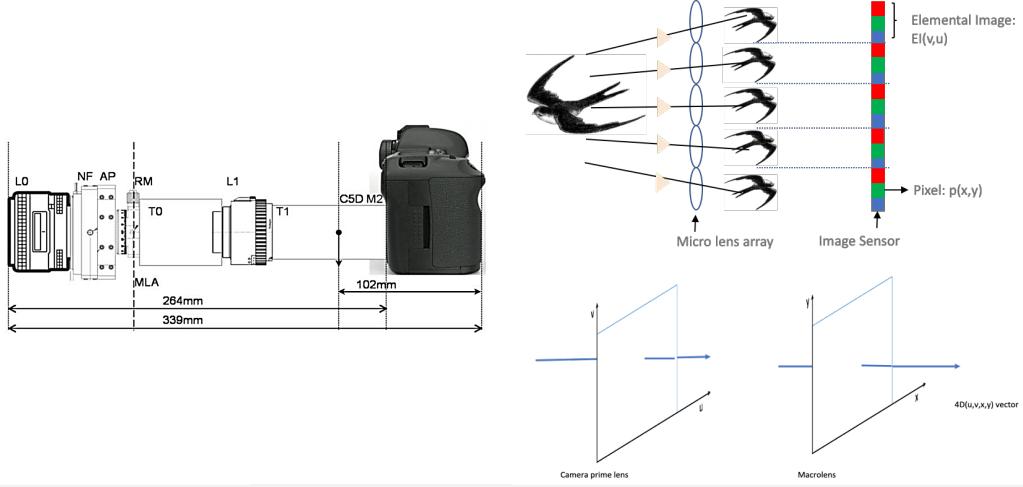


Fig. 3. Schematic diagram of a Holoscopic 3D camera

Fig. 4. Two planes Holoscopic data representation

2.2.2. Holoscopic Viewpoint extraction

One of the advantages of the 3D holoscopic technology is the ability to convert the H3D format to many 3D displays raw formats such as Multiview through a process called viewpoint extraction. Viewpoint extraction is one of the key stages in Holoscopic content adaptation for Multiview displays. The basic principle behind viewpoint extraction lays in the superimposing of pixels from all Elemental images as shown in Fig. 5. The Holoscopic image is defined as $H3DI = [H3DI(m,n)]$, where m and n are the horizontal and vertical positions of the H3DI pixels respectively.

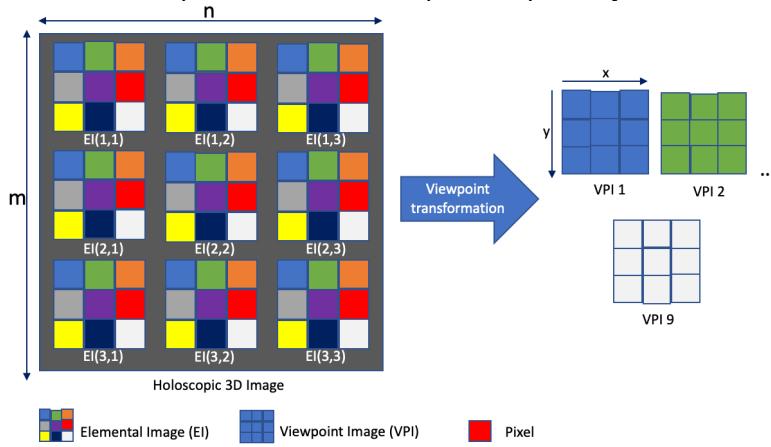


Fig. 5. Holoscopic Viewpoint extraction

Based on the Holoscopic 3D image viewpoint extraction principle presented in Fig. 5, it is clear that the default resolution of the extracted viewpoint images, VP1,

VP2...VP9 is directly proportional to the number of pixels within the elemental images that form the Holoscopic 3D image. As a result, an image interpolation technique to upsample the H3D image is required to improve the quality of viewpoint images. The next section elaborates on the full Holoscopic content adaptation process for Multiview displays.

2.2.3. Holoscopic content adaptation for Multiview display

The 3D Holoscopic content adaptation stages for Multiview displays can be grouped into four steps (see Fig. 6) ; (i) Holoscopic data acquisition, where assets can be recorded with the Holoscopic 3D imaging technology with a linear or angular motion ii) Multiview frame extraction, during this stage the viewpoint images needed are extracted as well as the disparity range as to where viewers observation position is to be taken into account when extracting Multiview points. However, this information is used to make sure the right MLA parameters are set during capture. (iii) Viewpoint image up-sampling and (iv) Multiview pixel remapping.

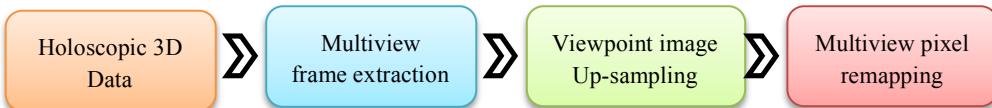


Fig. 6. H3D data processing stages

2.3 Deep learning and Single Image Super-resolution (SISR)

Deep learning represents a class of machine learning technologies mainly based on the concept of deep neural networks [26]. This class of algorithms was proven to work well in modeling complex functions and was also proven to deal with big data efficiently. Deep learning techniques are thus used in many domains such as natural language processing and computer vision. The advantage of deep learning in comparison with other machine learning techniques is mainly related to the generalization capabilities with unstructured raw data. Originally, these algorithms were used as classifiers and trained using pairs of (data, labels) [26]. The goal at the end is to efficiently train this classifier to generalize to unseen data samples. Since then, many deep learning algorithms were developed for other applications such as regression, time series prediction, super-resolution, etc. [27]. Generally, these neural network work well with high volumes of data.

Single Image Super-resolution (SISR) represents a class of image processing techniques mainly intended to increase the resolution of a low-resolution image (LR). The resulting high-resolution image (HR) needs to preserve the structural information and the high-frequency details of the original image. In fact, this task is very challenging as the possibilities are very large for the HR image (very wide search space) [28]. Legacy approaches for SR had several limitations related to the uncleanness regarding the LR-HR mapping, the inefficiency in dealing with larger amounts of data and the lack of generalization. Recently, deep learning-based techniques were proven to be efficient in

big data scenarios maintaining the ability to model and learn higher level abstractions from raw data [29].

There are mainly three categories of single image super-resolution approaches found in the literature. The first category of techniques are interpolation based methods such as bicubic interpolation which are considered among the fastest but unfortunately lack quality [28]. Reconstruction based methods are the second category which are mainly intended to solve the super-resolution problem of a certain category of images which induces a lack of generalization to other domains and categories [28]. The third category are learning based methods which are considered among the most robust and efficient techniques, and thus they are the most investigated in our study. Learning based super-resolution methods use advanced machine learning techniques to analyze visual features and learn a nonlinear mapping between the LR and the SR images [30]. Moreover, this class of techniques saw a real shift in interest due to the superior performance induced by deep learning-based methods. SRCNN and VDSR [30, 31] are without a doubt the most notable super-resolution contributions as they provided a concrete proof that deep learning-based methods are effective for super-resolution applications. The focus of our study is thus based on this category. A selection of super-resolution techniques based on deep learning and used within our study are detailed in section 3.3.

3 Methodology

In this section, we describe the material and methods used in order to design and implement a framework evaluating the single image super-resolution upsampling for H3D images and videos. This includes the dataset, the data preprocessing methodology, and the tested super-resolution techniques. A video motion interpolation prototype to increase the framerate of H3D videos is also presented.

3.1 Data collection and preprocessing

The data was one of the most critical parts of our research. Having a good quality dataset at hand is primordial in order to accurately evaluate the feasibility of our approach in a real-world scenario. This required setting up a professional studio environment in order to collect the data. The data we collected and captured consists of 13 cultural objects from the collection of the Museum of Islamic Art in Doha, Qatar. These assets vary in shape and size and were selected to provide real world samples as some of them were easy to capture and some others were more or less difficult due to their dark and light absorbing surfaces, tiny sizes, etc. For data collection, we relied on the H3D camera prototype developed in Brunel University London CMCR laboratory. The capturing scenarios were 360° H3D video, Linear Multiview, and direct captures.

The environmental variables such as lighting and distance were also recorded to provide additional details and metadata to tune the post-processing stage. In most cases, the same environmental setup (lighting, distance, etc.) was used to capture the assets. Each picture is saved in two formats: RAW ARW (uncompressed) and JPEG (compressed). For both setups, three scenarios were implemented.

- 360° using a turntable with varying turning steps (10° , 5°).
- Linear Multiview (8 views, 16 views).
- Direct captures.

Fig. 7 and Fig. 8 show H3D and 2D images for a statue and tombstone objects respectively.



Fig. 7. Statue object captured by Holoscopic 3D (H3D) camera on the left and traditional 2D camera on the right



Fig. 8. Tombstone object captured by Holoscopic 3D (H3D) camera on the left and traditional 2D camera on the right

The data collected will be used for testing the performance of the developed 3D holoscopic post-processing framework that is mainly intended to increase the output quality (perceived output).

3.2 Investigated state-of-the-art Super-Resolution techniques

In the following, we present some details about the super-resolution techniques we implemented and tested on 3D holoscopic images. The tested methods were either winners of super-resolution challenges or their performance was at some time the best in class.

- **SRCNN (Image Super-Resolution Using Deep Convolutional Networks)**

SRCNN [30] is one of the first CNN based super-resolution designs that were published. It provided a confirmation that deep learning upsampling approaches are far better than legacy approaches. The network original design upsamples only the luminance channel of the image to simplify the training and to optimize computations as it was found that the high frequency details in an image are mainly described by the luminance channel. Regarding its architecture, SRCNN has a 3-layer architecture which is relatively considered as simple in comparison with modern SR architecture. The network has for a goal to map the low-resolution input into a higher resolution output. The first layer is in charge of patch extraction, the second will non-linearly map the patches to higher resolution, and the final layer will reconstruct the input according to the mapping. This simple architecture, while currently considered inefficient, overpassed traditional upsampling methods and opened the ways for multiple contributions for deep learning based single image super-resolution. Some of the limitations of SRCNN that were discussed in many contributions [28] are related to the facts that it has a very basic architecture as theoretically, in deep learning, the deeper is the better. Other limitations were also reported such as the reliance of the network on a bicubic interpolated LR image and the very long training (slow convergence). The architecture of SRCNN is outlined in Fig. 9.

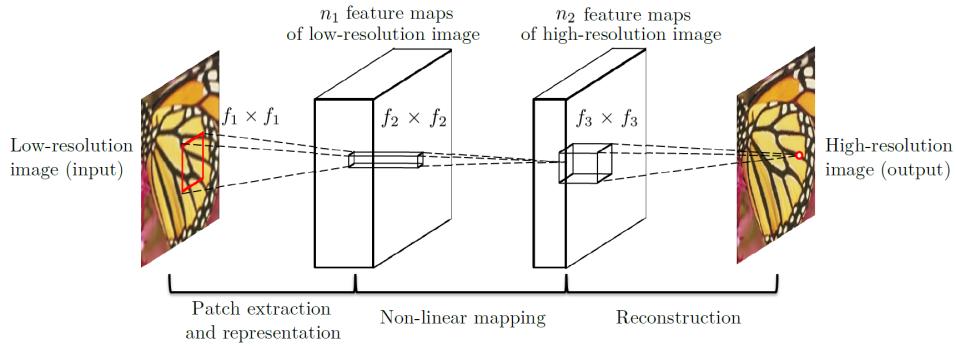


Fig. 9. Architecture of SRCNN

- **VDSR (Accurate Image Super-Resolution Using Very Deep Convolutional Networks)**

Theoretically, a neural network performance is tied to its depth. However, there are still several challenges related to the effectiveness of training which often is affected by problems such as vanishing and exploding gradients. A lot of research work was thus dedicated to finding techniques that can solve these issues. In this regard, the VDSR [31] network was the first to introduce a very deep architecture inspired by the

successful CNN VGG [32]. The network has 20 layer and small convolutional windows sizes (3×3). The network, similarly to SRCNN, tries to upsample the LR bicubic interpolated luminance channel. Its authors reported some difficulties in the training, and thus they introduced gradient clipping. One of the advantages of VDSR is that it can be trained for several scales at once. The output of the network is fixed to patches of 41×41 . In contrast with SRCNN, VDSR does not directly map the LR Y channel to the HR Y channel, but it maps what the authors call the *Residual* which is the difference between the Real HR and the Bicubic interpolated LR. Due to this, the authors claim that the computations are more lightweight and the network converges faster [28]. The architecture of VDSR is described by Fig. 10.

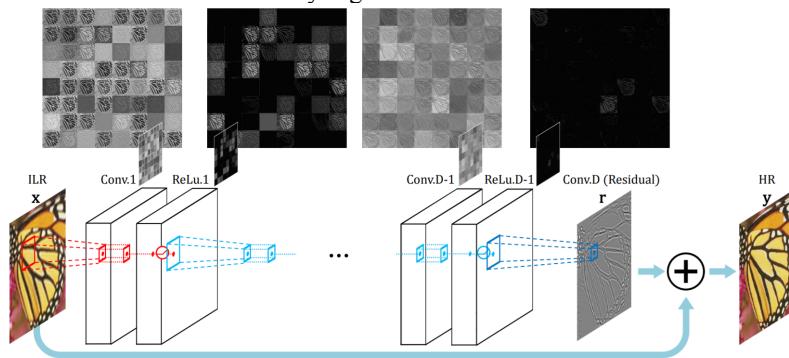


Fig. 10. VDSR architecture

- **ESPCN (Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network)**

The ESPCN network [33] was originally intended for real-time applications such as online video super-resolution. The authors of ESPCN tried to solve limitations found in using nearest neighbor interpolation when features get repeated in the adjacent positions. To mitigate this redundancy, the authors of ESPCN introduced a new layer called efficient subpixel convolution. They found that feature extraction directly in the LR image is more efficient than in the HR space. ESPCN achieved better performance than SRCNN with less computational complexity. The architecture of ESPCN is presented in Fig. 11.

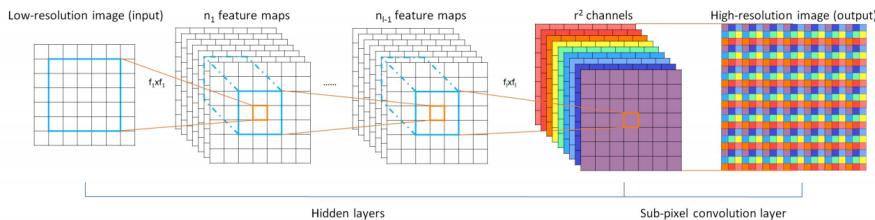


Fig. 11. Architecture of ESPCN

- **RDN** (Residual Dense Network for Image Super-Resolution)

A newer design based on the successful DenseNet [34] was proposed in [35]. It uses the concept of residual dense blocks (RDB) and residual learning exploiting hierarchical features. By doing so, the developed SR network can capture local features with densely connected convolutions. Each RDB is densely connected to the successive block achieving the concept of contiguous memory. Global features are then derived from local features hierarchically. The network demonstrated superior performance against state-of-the-art SR models. The network architecture is outlined in Fig. 12.

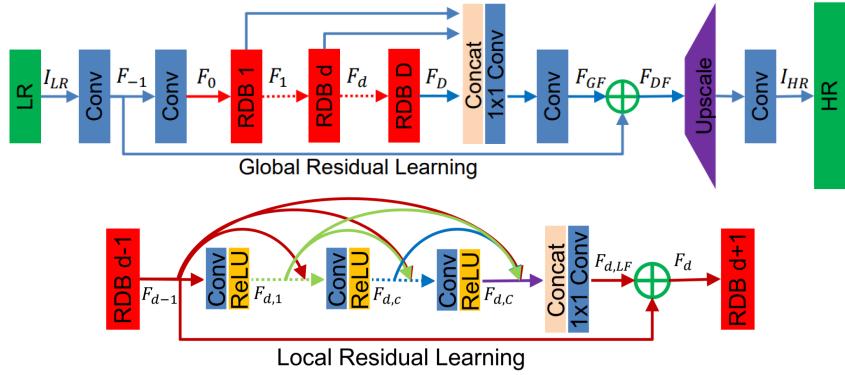


Fig. 12. Residual Dense Network (RDN) architecture

- **EDSR** (Enhanced Deep Residual Networks for Single Image Super-Resolution)

The EDSR super-resolution network was proposed in [36]. It introduced multiple new contributions and yielded state of the art performance in 2017. Its authors claimed that the Batch Normalization layer has no positive impact in SR as it was originally intended to classification CNNs. The authors leverage the fact that the visual features in different scales of upsampling are correlated and thus they relied on a transfer learning to train the network for higher scales. They mostly used the weights of the $\times 2$ scale to initialize the $\times 3$ and $\times 4$ scales which resulted in additional performance benefits. The network won the NTIRE 2017 super-resolution challenge [36]. The architecture of VDSR is outlined in Fig. 13.

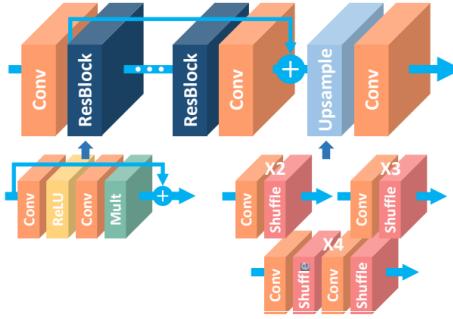


Fig. 13. Architecture of the single scale EDSR network

- **WDSR** (Wide Activation for Efficient and Accurate Image Super-Resolution)

The authors of WDSR further enhanced the EDSR network by designing and implementing two major modifications [37]. The first being Global residual pathways where the authors found that linearly stacked convolutions are computationally taxing. The authors' second modification is within the upsampling layer as in previous contributions, convolutional layers were stacked after the non-linear upscaling layer. The authors argue that extracting the low-resolution features does not reduce the accuracy in SR tasks, but it improves the training performance significantly. The network won the NTIRE 2018 super-resolution challenge. Its architecture is compared to EDSR in Fig. 14.

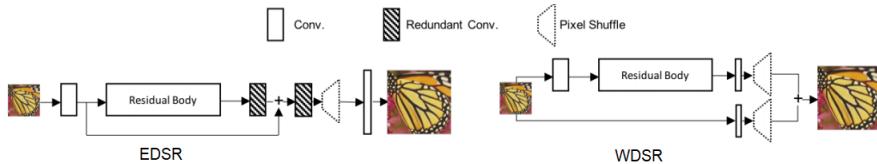


Fig. 14. Architecture of WDSR and EDSR

3.3 Video Motion Interpolation

Video motion interpolation is one of the most studied problems in computer vision as it plays a primordial role in many multimedia systems [38, 39]. The principle consists of increasing the framerate of a video sequence to achieve a smoother playback. There are multiple proposed approaches for this challenge such as motion interpolation using handcrafted techniques to compute correspondences between the frames. However, these techniques are inefficient when dealing with high-resolution inputs and massive amounts of data as they are computationally taxing [38]. Moreover, these approaches do not perform well in challenging scenarios such as sudden light changes from one frame to another and motion blur. Deep learning-based approaches try to learn the mapping between one frame to another which in fact may address these problems due to the fact that they are learning based techniques. For our tests, we tested two video motion interpolation approaches based on deep learning. The first one is based on CNNs where a network is trained to predict in-between frames with adaptive convolutions [39], the second one is called PhaseNet and is based on pixel phase-based motion to perform the motion interpolation in videos.

3.4 H3D post-processing framework design

To solve the issues of low spatial density and the lack of higher framerate in H3D captures, we designed and implemented a post-processing framework for H3D content based on super-resolution and motion interpolation. The framework is mainly intended to improve the quality of the H3D technique while not inducing extra hardware costs. The framework is intended for both H3D still images and videos. However, despite yielding very promising results, the video motion interpolation is still at prototyping levels as the results must go through further tests. The framework architecture is presented in Fig. 15.

The data we used to validate the framework consists of some museum objects captured in a professional studio environment in order to replicate the best possible scenario. The H3D camera we used was the Sony α 7 II which outputs images in 40 Megapixels (MP) resolution. For video, the camera is limited to 4K resolution which is the fifth of its sensor recording capability. For the 360° video scenario, we designed a capturing framework which consisted of using a graduated turntable where we fitted the object. We then recorded the object using 72 pictures (a picture each 5°). The pictures were then compiled in a video with the framerate of 5 FPS. The resulted capture is a video of 40 MP that has the best possible quality. The frames then receive preprocessing which consists mainly of global parameters adjustments such as brightness and contrast, lens distortion corrections etc. as described in previous works [3].

The generated frames contain 4845 elemental images. Each of the frames is then split up into 64 patches. The patches are then forwarded to the super-resolution process to be upscaled in the desired scale (mostly $\times 2$). The resulting upscaled output will later be compiled to a 5 FPS lossless video (lossless codecs such as *Huffyuv* and *FFV1*). The video will at this point be forwarded to the motion interpolation neural network to increase the framerate and ensure a smoother playback for the user.

Regarding super-resolution, our goal is mainly to investigate the impact of applying such techniques on the output viewed by the end-user as well as the preservation of high-frequency details as legacy approaches lack output quality. Thus, we tested the previously mentioned super-resolution frameworks (SRCNN, VDSR, ESPCN, RDN, EDSR, WDSR) on H3D frames. For each of the super-resolution techniques, we relied mainly on the implementation specified by the authors: The kernel types and sizes, the preprocessing required, the datasets used to train and validate the networks, the training hyperparameters, and the other network parameters. We mainly focused on two scales in our tests ($\times 2$, $\times 4$). However, it is worth noting that networks such as VDSR could be trained for multiple scales at once.

At the highest pixel density of the camera, the elemental images have a resolution of 93×93 pixels which is in fact relatively low. Throughout this approach, we aimed at producing final output elemental images of 186×186 pixels in the $\times 2$ scale having the maximum quality possible and 372×372 (scale $\times 4$) with a medium to high quality.

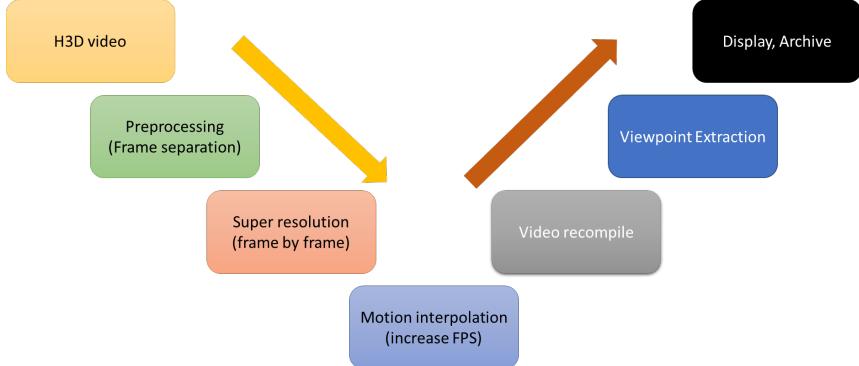


Fig. 15. H3D post-processing framework design

The goal of using video motion interpolation rather than recoding the assets at a higher framerate is mainly to increase the raw spatial resolution as in video captures, the camera we use is limited to 4K resolution (3840×2160). It is true that this resolution is very high for 2D captures, but for H3D, it results in tiny elemental pictures which induce a loss in detail and quality. The idea behind motion interpolation is to leverage the details and the similar visual features found in two successive frames in order to predict or compute the frames in between. Currently, we prototype a phase shift network based on deep learning which yielded good results, and this can be confirmed visually (higher output quality and smoothness).

4 Results and discussions

4.1 Experimental setup and implementation

For the hardware, we used a Sony α7R II mirrorless camera body (35mm CMOS sensor and 40 Megapixels pixel density) fitted with an H3D lens array prototype developed in the CMCR laboratory at Brunel University London. We have also used an Asus Laptop with i7-7700HQ, 16 GB of Ram and GTX 1070 GPU for our tests and software implementations.

The SR models and the video motion interpolation approach were implemented with Python using the Keras deep learning library with Tensorflow GPU backend (Keras version 2.2.4, Tensorflow version 1.11.0) [40, 41]. The video frame extraction (decoding) and compiling (encoding) was performed using the FFmpeg 4.1 framework [42]. For the visual output evaluation, we relied on feedback collected from several users that evaluated the quality of the results in the CMCR laboratory at Brunel University London.

4.2 Experimental results

To measure and compare the performance of the models, we used the Peak Signal to Noise Ratio (PSNR) which is a mathematical measure that measures the image quality

based on pixel differences, and the grayscale structural similarity index (SSIM) which is a similarity measure between two images. However, these two measures are not reflecting at 100% the quality of a picture which can only be evaluated by human perception as the quality of a picture is a subjective result.

Regarding the tests, the super-resolution networks we tested were trained and tested for two scales ($\times 2$, $\times 4$). The networks were trained according to the structure, architecture, hyperparameters, and datasets specified by their original authors. Our dataset was thus downsampled using the two scales where the $\times 2$ networks will try to upsample a 3840×2160 input to 7680×4320 output, and the $\times 4$ networks will try to upsample a 1920×1080 input to 7680×4320 output.

The following table outlines the visual and numerical results of these networks on a selection of cultural assets sampled from the dataset we collected.

Table 1. Performance comparison of pretrained SR models on a selection of H3D capture of cultural assets. The used metrics are the PSNR and SSIM

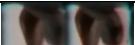
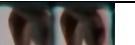
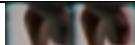
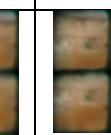
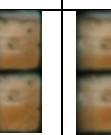
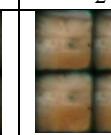
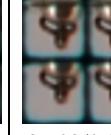
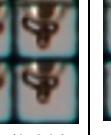
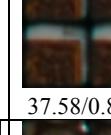
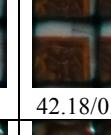
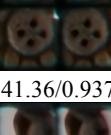
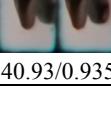
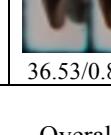
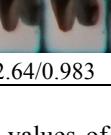
Scale	Bicubic		SRCNN		VDSR	
	$\times 2$	$\times 4$	$\times 2$	$\times 4$	$\times 2$	$\times 4$
	 38.87/0.891	 34.11/0.792	 39.12/0.901	 35.39/0.803	 40.09/0.917	 35.19/0.830
	 39.58/0.911	 35.10/0.817	 40.31/0.823	 35.67/0.793	 41.71/0.897	 36.21/0.819
	 39.15/0.893	 35.27/0.841	 40.12/0.857	 37.12/0.806	 40.02/0.936	 37.63/0.871
	 38.89/0.875	 33.73/0.748	 39.11/0.900	 35.12/0.734	 39.59/0.843	 36.66/0.793
	 39.01/0.918	 34.52/0.852	 40.03/0.910	 35.43/0.817	 40.19/0.918	 35.73/0.858
	 39.21/0.881	 34.09/0.858	 40.14/0.895	 34.57/0.785	 40.36/0.920	 36.01/0.829

Table 1. Performance comparison of pretrained SR models on a selection of H3D capture of cultural assets. The used metrics are the PSNR and SSIM

ESPCN		RDN		EDSR		WDSR	
$\times 2$	$\times 4$	$\times 2$	$\times 4$	$\times 2$	$\times 4$	$\times 2$	$\times 4$
							
41.02/0.926	35.98/0.889	42.19/0.946	37.86/0.910	42.76/0.982	38.74/0.924	42.92/0.973	38.84/0.936
							
41.12/0.911	37.03/0.879	41.83/0.967	37.75/0.903	42.33/0.981	38.43/0.932	42.50/0.972	38.73/0.930
							
40.93/0.939	36.87/0.873	41.87/0.943	37.13/0.911	42.04/0.980	38.40/0.936	42.07/0.980	37.97/0.928
							
40.12/0.936	37.58/0.824	40.80/0.913	37.17/0.887	42.15/0.982	38.33/0.915	42.18/0.982	38.50/0.912
							
41.36/0.937	37.13/0.883	40.76/0.930	37.19/0.915	42.29/0.982	38.57/0.913	42.34/0.972	38.65/0.919
							
40.93/0.935	36.53/0.843	40.89/0.917	37.03/0.904	42.64/0.983	38.77/0.918	42.69/0.969	39.30/0.929

Overall, we can see that in most cases the values of PSNR and SSIM reflect the feedback received by several people from the surveyed audience. A higher PSNR and an SSIM close to 1 will reflect that the upsampled image quality is closer to the original picture. We can observe that the results of the $\times 2$ upsampling are higher in comparison with the $\times 4$ upsampling across all the techniques. We can also observe that the frameworks EDSR and WDSR yield the best performance with the two scales which is also validated by the fact that they currently achieve superior performance across the SR frameworks found in the literature.

Regarding video motion interpolation, we tested two video motion interpolation approaches. The first one is based on adaptive convolutions CNNs, and the second one is based on pixel phase interpolation. Both of the frameworks yielded nearly similar results when presented to the surveyed audience although there were some glitches that need to be addressed. Overall, all the viewers confirm the increase in playback smoothness and quality despite the glitches observed.

4.3 Results discussion

Through this investigation, we wanted to study the impact of applying recent deep learning-based techniques and mainly super-resolution on the H3D imaging technology. H3D is a practical alternative to commercial solutions of 3D vision such as stereoscopic or Multiview as it captures and displays a true 3D representation of the scene. However, one of its limitations is related to the fact that the commercial camera sensors used do not have a high pixel density and thus, the elemental images resolution is small. Solving such issues by hardware (denser CMOS sensors) is theoretically the best way to solve such limitations, but unfortunately, the very high costs induced by these sensors will make the technique lose its main advantage (cost-effectiveness). As a potential solution, we designed a post-processing framework based on deep learning that aims at solving the low-density issues by software. For this purpose, we collected a dataset of several cultural objects in a professional studio environment in order to have the best possible conditions. Instead of taking video captures of the asset, we saw that the output quality of still images is far superior (5 times more). We thus captured the assets in H3D 360° by taking a picture every 5° and turning the asset accordingly. The final result was then 72 H3D images of 40 MP. The challenge was then to increase the spatial resolution of the Elemental Images and to smooth the video playback. For increasing the spatial density, we relied on pretrained SR neural networks such as SRCNN, VDSR, ESPCN, EDSR, RDN, and WDSR in order to investigate the added benefit of applying super-resolution to the output quality. Most of these networks were designed and trained according to the specifications set out by their authors (structure, hyperparameters, datasets, etc.). After that, we evaluated the quality of the super-resolution on our H3D dataset by comparing the SR performance of the above-mentioned models. The metrics we used are the PSNR and the SSIM but the most relevant metric is the human perception of the displayed images. According to the results and with the confirmation of the surveyed audience, the WDSR network had the best SR for H3D images which is also validated by the fact that the network was the winner of the NTIRE 2018 super-resolution challenge. Regarding the video motion interpolation, all the viewers felt the huge difference in the playback smoothness which is in fact very promising, but there are still challenges to increase the output quality and to get rid of occasional glitches.

5 Conclusion

Through this paper, we presented a post-processing framework to improve the quality of the 3D Holoscopic imaging technology. We mainly focused on deep learning

approaches for Single Image Super-resolution. A dataset of 13 cultural objects collected in the Museum of Islamic Art, Doha, Qatar was used to design, implement and test the framework using high-quality samples. We relied on state-of-the-art SISR models and compared their upsampling performance and quality on the H3D cultural content we collected. The results were both evaluated by standard metrics such as PSNR and SSIM in addition to feedback collected from a surveyed audience. The results showed a real improvement in terms of output quality according to both the technical results and the audience feedback. We have also tested two video motion interpolation approaches and reconstructed a 360° video of the collected samples. The results of these interpolation approaches show a real improvement in playback smoothness despite some glitches seen in some videos. In the future, we aim at enhancing the Super-resolution results through the use of a dataset collected with larger CMOS density camera, a new advanced holoscopic lens adapter and an evolutionary design of the SISR technique used. Regarding video motion interpolation, we aim at enhancing the approaches we used to provide a naturally looking flawless playback.

6 Acknowledgements

This publication was made possible by NPRP grant 9-181-1-036 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors (www.ceproqha.qa).

The authors would also like to thank Mr. Marc Pelletreau, the MIA Multimedia team, the Art Curators and the management staff of the Museum of Islamic art, Doha Qatar for their help and contribution in the data acquisition.

7 References

- [1] A. Aggoun, E. Tsekleves, M.R. Swash, D. Zarpalas, A. Dimou, P. Daras, P. Nunes, L.D. Soares, Immersive 3D holoscopic video system, *IEEE MultiMedia*, 20 (2013) 28-37.
- [2] J.P. McIntire, P.R. Havig, E.E. Geiselman, What is 3D good for? A review of human performance on stereoscopic 3D displays, in: Head-and Helmet-Mounted Displays XVII; and Display Technologies and Applications for Defense, Security, and Avionics VI, International Society for Optics and Photonics, 2012, pp. 83830X.
- [3] M. Swash, Holoscopic 3D imaging and display technology: Camera/processing/display, in, Brunel University London., 2013.
- [4] M.R. Swash, A. Aggoun, O. Abdulfatah, B. Li, J. Fernandez, E. Alazawi, E. Tsekleves, Pre-processing of holoscopic 3D image for autostereoscopic 3D displays, in: 3D Imaging (IC3D), 2013 International Conference on, IEEE, 2013, pp. 1-5.
- [5] C. Conti, L.D. Soares, P. Nunes, HEVC-based 3D holoscopic video coding using self-similarity compensated prediction, *Signal Processing: Image Communication*, 42 (2016) 59-78.
- [6] A. Belhi, A. Bouras, S. Foufou, Digitization and preservation of cultural heritage: The CEPROQHA approach, in: Software, Knowledge, Information Management and Applications (SKIMA), 2017 11th International Conference on, IEEE, 2017, pp. 1-7.

- [7] A. Belhi, A. Bouras, S. Foufou, Leveraging Known Data for Missing Label Prediction in Cultural Heritage Context, *Applied Sciences*, 8 (2018) 1768.
- [8] R. Parry, *Recoding the museum: Digital heritage and the technologies of change*, Routledge, 2007.
- [9] A. Belhi, S. Foufou, A. Bouras, A.H. Sadka, Digitization and Preservation of Cultural Heritage Products, in: IFIP International Conference on Product Lifecycle Management, Springer, 2017, pp. 241-253.
- [10] G. Singh, CultLab3D: digitizing cultural heritage, *IEEE Computer Graphics and Applications*, 34 (2014) 4-5.
- [11] G. Pavlidis, A. Koutsoudis, F. Arnaoutoglou, V. Tsioukas, C. Chamzas, Methods for 3D digitization of cultural heritage, *Journal of cultural heritage*, 8 (2007) 93-98.
- [12] T. Alfaqheri, S.A.E.M. Nasri, A.H. Sadka, 3D Holoscopic Imaging for Cultural Heritage Digitalisation, arXiv preprint arXiv:1810.03916, (2018).
- [13] G. Lippmann, Epreuves reversibles donnant la sensation du relief, *J. Phys. Theor. Appl.*, 7 (1908) 821-825.
- [14] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, P. Hanrahan, Light field photography with a hand-held plenoptic camera, *Computer Science Technical Report CSTR*, 2 (2005) 1-11.
- [15] M.S.K. Gul, B.K. Gunturk, Spatial and angular resolution enhancement of light fields using convolutional neural networks, *IEEE Transactions on Image Processing*, 27 (2018) 2146-2159.
- [16] D. Liu, P. An, R. Ma, C. Yang, L. Shen, 3D holoscopic image coding scheme using HEVC with Gaussian process regression, *Signal Processing: Image Communication*, 47 (2016) 438-451.
- [17] C. Conti, P. Nunes, L.D. Soares, Light field image coding with jointly estimated self-similarity bi-prediction, *Signal Processing: Image Communication*, 60 (2018) 144-159.
- [18] T.E. Bishop, S. Zanetti, P. Favaro, Light field superresolution, in: Computational Photography (ICCP), 2009 IEEE International Conference on, IEEE, 2009, pp. 1-9.
- [19] T.E. Bishop, P. Favaro, The light field camera: Extended depth of field, aliasing, and superresolution, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34 (2012) 972-986.
- [20] Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun, T. Tan, LFNet: A Novel Bidirectional Recurrent Convolutional Neural Network for Light-Field Image Super-Resolution, *IEEE Transactions on Image Processing*, 27 (2018) 4274-4286.
- [21] N.K. Kalantari, T.-C. Wang, R. Ramamoorthi, Learning-based view synthesis for light field cameras, *ACM Transactions on Graphics (TOG)*, 35 (2016) 193.
- [22] Y. Wang, G. Hou, Z. Sun, Z. Wang, T. Tan, A simple and robust super resolution method for light field images, in: Image Processing (ICIP), 2016 IEEE International Conference on, IEEE, 2016, pp. 1459-1463.
- [23] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, I. So Kweon, Learning a deep convolutional network for light-field image super-resolution, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 24-32.
- [24] Y. Yuan, Z. Cao, L. Su, Light-Field Image Superresolution Using a Combined Deep CNN Based on EPI, *IEEE Signal Processing Letters*, 25 (2018) 1359-1363.
- [25] E.H. Adelson, J.Y.A. Wang, Single lens stereo with a plenoptic camera, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1992) 99-106.

- [26] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep learning, MIT press Cambridge, 2016.
- [27] L. Deng, A tutorial survey of architectures, algorithms, and applications for deep learning, *APSIPA Transactions on Signal and Information Processing*, 3 (2014).
- [28] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, Deep learning for single image super-resolution: A brief review, *arXiv preprint arXiv:1808.03344*, (2018).
- [29] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *nature*, 521 (2015) 436.
- [30] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, *IEEE transactions on pattern analysis and machine intelligence*, 38 (2016) 295-307.
- [31] J. Kim, J. Kwon Lee, K. Mu Lee, Accurate image super-resolution using very deep convolutional networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646-1654.
- [32] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*, (2014).
- [33] W. Shi, J. Caballero, F. Huszár, J. Totz, A.P. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874-1883.
- [34] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: *CVPR*, 2017, pp. 3.
- [35] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [36] B. Lim, S. Son, H. Kim, S. Nah, K.M. Lee, Enhanced deep residual networks for single image super-resolution, in: *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*, 2017, pp. 4.
- [37] J. Yu, Y. Fan, J. Yang, N. Xu, Z. Wang, X. Wang, T. Huang, Wide Activation for Efficient and Accurate Image Super-Resolution, *arXiv preprint arXiv:1808.08718*, (2018).
- [38] S. Meyer, A. Djelouah, B. McWilliams, A. Sorkine-Hornung, M. Gross, C. Schroers, PhaseNet for Video Frame Interpolation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 498-507.
- [39] S. Niklaus, L. Mai, F. Liu, Video frame interpolation via adaptive convolution, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3.
- [40] F. Chollet, Keras: Deep Learning Library for Theano and Tensorflow, 2015. Available online: <https://keras.io> (accessed on 10 March 2019)
- [41] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, Tensorflow: a system for large-scale machine learning, in: *OSDI*, 2016, pp. 265-283.
- [42] S. Tomar, Converting video formats with FFmpeg, *Linux Journal*, 2006 (2006) 10.