DEEP LEARNING-BASED RESOLUTION-ENHANCED AUTOSTEREOSCOPIC THREE-DIMENSIONAL SURFACE METROLOGY

GAO SANSHAN

PhD

The Hong Kong Polytechnic University

2024

The Hong Kong Polytechnic University

Department of Industrial and Systems Engineering

Deep Learning-based Resolution-Enhanced Autostereoscopic Three-dimensional Surface Metrology

GAO Sanshan

A thesis submitted in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

December 2023

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

GAO Sanshan (Name of student)

Abstract

Precision measurement for micro-structured surfaces is important for the fabrication of micropatterns to guarantee form accuracy. Contact and non-contact measurement methods alike have been extensively used to inspect these surfaces, where the optical sensor used in the non-contact methods does not cause probable damage to the measured parts, and the speed of data acquisition is much faster. Among these non-contact measurement technologies, an autostereoscopic measurement system with a rapid data acquisition process is an effective method to achieve accurate on-machine measurement.

Autostereoscopy technology can provide a rapid and accurate three-dimensional (3D) measurement solution for micro-structured surfaces. The autostereoscopic 3D measuring system can record elemental images within one snapshot and the measurement accuracy can be quantified from the disparities existing in the 3D raw information. One of the primary challenges in improving the measurement resolution of autostereoscopic 3D systems is the natural compromise between spatial resolution resulting in finer details and angular resolution enabling accurate parallax restoration. This trade-off poses an obstacle to enhancing the resolution of the system.

Improving angular resolution is of utmost importance when considering the trade-off of light field data. Within the domain of enhancing angular resolution through deep learning methods, advancements have been made in both non-depth-based techniques and the techniques requiring depth estimation. However, non-depth-based methods usually generate image ghosting when the light field images have large disparity ranges. It is difficult to realize accurate disparity estimation directly obtained through light field

images for the depth-based methods, so image artefacts are usually fabricated in the novel views especially when scenes are complex. In addition, a lack of ground truth of the generated novel views makes the training on the finite data insufficient since the training data have to be split into inputs and their labels. In this thesis, a novel semi-supervised learning paradigm for light field angular super enhancement is presented without the need for ground truth. Following the learning paradigm, the learning models can be directly supervised by the input, and training data are not required to be paired as input and labels. Hence, more light field images with redundant parallax information can be used for the learning of deep light field reconstruction.

To take advantage of the learning paradigm, A convolutional network leveraging motion estimation is built to synthesize novel views via fusing adjacent views. The experiments demonstrate that the method, implemented under the proposed learning paradigm, achieves high-quality metrics for simulated and real-world light field data. This is particularly notable for scenes that include multi-depth targets, complex textures, and large baselines. More accurate parallax structures can be recovered based on the proposed learning paradigm, whilst over 69% of training data are saved compared with other methods. In addition, under the proposed learning paradigm, even a simple shallow network can synthesize high-quality novel views. The PSNR achieved by the baseline method is approximately improved by 2dB after the proposed semi-supervision. Hence, the proposed semi-supervised learning paradigm can be easily integrated with other learning models.

Regarding resolution enhancement of autostereoscopic measuring data, a self super-resolution algorithm driven by deep learning models has been designed. This algorithm is integrated into the measurement system, resulting in the creation of a self

super-resolution autostereoscopic 3D measuring system. The self super-resolution algorithm can generate novel perspectives between the neighbouring Elemental Images (EIs) so that the angular resolution is markedly enhanced several times over. The proposed algorithm has been embedded into an autostereoscopic 3D measuring system so that the system can achieve self super-resolution. To validate the feasibility and technical merit of the proposed self super-resolution 3D measuring system, a comprehensive comparison experiment was conducted between the traditional autostereoscopic measuring system without super-resolution and the proposed system. The results demonstrate that the self super-resolution system can significantly improve the resolution of the measuring data by around four-fold and enhance the measurement accuracy with lower standard deviations and biases.

To reduce the effect of vibration during on-machine measurement, multiple frames captured by the autostereoscopic measuring system are able to be used to eliminate the measurement errors induced by the vibration. Furthermore, essential information for achieving high spatial resolution in the measurement data can be extracted from the redundant subpixel-level information. As a result, the study introduces a multi-frame autostereoscopic system designed specifically for the on-machine measurement of three-dimensional surfaces, aimed at enhancing resolution. It leverages the vibrations produced by the machine tool during on-machine measurements to capture multiple frames of the target surface with offsets. This approach allows for resolution enhancement. A multi-frame resolution-enhanced deep learning model is developed, along with a supervised training process, to generate resolution-enhanced raw elemental images. This approach is pivotal to improving the measurement resolution. Through experiments, the system performance is assessed, and the results demonstrate a four-fold enhancement in spatial resolution along with improved measurement

accuracy.

In this study, learning-based techniques are applied to enhance LF resolution of measurement data gathered using the autostereoscopic 3D measuring system. Through experimental evaluation, the measurement resolution and accuracy for micro-structured surfaces are improved after artificial intelligence enhancement. In addition to the aforementioned advancements, the study also presents a generic semi-supervised learning paradigm specifically designed for deep learning models employed in angular resolution-enhancement tasks. This innovative paradigm allows for high data efficiency, ensuring superior performance in enhancing angular resolution using limited labelled data. The deep learning-based method results in an enhancement of angular resolution from 16 × 9 to 31 × 17, as well as an improvement in spatial resolution from 151 × 151 to 604 × 604. This enhancement in angular resolution leads to a reduction in error between measured and true values from over 1 μm to around 0.1 μm on average, along with a decrease in repeated measurement deviation by around 1 μm. Additionally, the spatial enhancement contributes to an increase in accuracy by 1 μm and a reduction in the deviation of repeated measurements from 1.533 μm to 1.388 μm. The research highlights the potential of combining autostereoscopy technology with deep learning technology for precise measurement.

Acknowledgements

The journey of my 3-year PhD study was like a dream that I had never imagined before I came to Hong Kong. It has been a turning point in my life, and I cherish the chance of studying at The Hong Kong Polytechnic University. I wish to convey my deepest appreciation to my chief supervisor, Prof. Benny C.F. Cheung, Chair Professor of Ultra-precision Machining and Metrology of the Department of Industrial and Systems Engineering and Director of the State Key Laboratory of Ultra-precision Machining Technology. I have learned a lot through the supervision by Prof. Cheung, including the desire to conduct research and passion for science. I appreciate his guidance, encouragement, and support during the years, and all of these made me more confident and eligible to gradually become a competent researcher.

I would also like to convey my thanks to Dr. Da Li for his invaluable technical support and guidance throughout my research. His guidance and counsel have been instrumental in facilitating my seamless entry into the PhD program. I am also deeply thankful to Dr. Chunjin Wang and Dr. Lesley Ho for their unwavering support and insightful guidance throughout my academic journey at the PolyU. Many thanks are equally given to all my colleagues and staff in the SKL. I am also deeply appreciative to PolyU's Research and Innovation Department for their financial support.

I am fortunate to have made many friends at PolyU, including Wu Songman, Wang Ruobing, Wang Yuxuan, Yang Yongqiang, and others. Memorable moments with them are the wealth of my life. We spent two unusual Spring Festivals together and I really appreciate the warmth they showed to me during the time of COVID-19.

I am also deeply grateful to my two closest friends, Mr Xiao Xie and Ms Siyu Yang. They have been my friends for over ten years and have become very important

persons in my life. They are good listeners regarding my emotional problems and always pulled me out of despondency. I hope the friendship will last forever.

Finally, I wish to express my heartfelt appreciation to my parents. I am grateful for their unwavering confidence in all the decisions I have made in the past. I will always keep a heart full of gratitude to all the people I will encounter during my life journey.

Research output arising from this study

Refereed Journal Papers

1 **Gao, Sanshan**, Cheung, Chi Fai, & Li, Da. (2022). Self super-resolution autostereoscopic 3D measuring system using deep convolutional neural networks. Optics Express, 30(10), 16313-16329.

2 **Gao, Sanshan** & Cheung, Chi Fai. (2023). Autostereoscopic 3D Measurement Based on Adaptive Focus Volume Aggregation. Sensors, 23(23): 9419.

3 **Gao, Sanshan**, Cheung, Chi Fai, & Li, Da. (2023). A semi-supervised angular super-resolution method for autostereoscopic 3D surface measurement. Optics Letters. 49(4), 858-861.

4 **Gao, Sanshan**, Li Da, & Cheung, Chi Fai. (2023). Multi-frame resolution-enhanced autostereoscopic system for on-machine three-dimensional surface metrology . IEEE Transactions on Instrumentation and Measurement. Accepted.

Refereed Conference Papers

1 **Gao, Sanshan**, Cheung, Chi Fai, & Li, Da. (2021). Autostereoscopic Measurement System for Rapid 3D Inspection of Wire Bonding. In 2021 International Conference of Optical Imaging and Measurement (ICOIM) (pp. 145-148). IEEE.

2 **Gao, Sanshan** and Cheung, Chi Fai (2024). Depth Estimation For Autostereoscopic 3D Surface Measurement Using A Deep Encoder-decoder Network, The 20th International Conference on Precision Engineering (ICPE2024), 23-27 October, Sendai, Japan. Accepted.

Awards

1 Best paper award in the 2021 International Conference of Optical Imaging and Measurement (ICOIM), Xi'an, China (2021).

Table of contents

List of figures

List of tables

Chapter 1 Introduction

1.1 Background of the study

Surfaces with micro-structures are demonstrated to realize various functions such as self-cleaning, drag reduction, anti-fouling, low friction, etc. These micropatterns significantly impact the physical properties of a material so that these micro-structured surfaces can be applied in many applications in various fields. The applications of the micro-structured surfaces include optics (Li & Allen, 2012; Zhang et al., 2018), energy (Bixler & Bhushan, 2013; Wang et al., 2015), robotics (Breckwoldt et al., 2015; Yao et al., 2020), etc. as shown in Figure 1.1.

Since the demand for micro-structured surfaces is growing quickly, great challenges are faced in manufacturing. As an important part of manufacturing, the accurate measurement of micro-structured surfaces has become vital in ultra-precision machining. Measurement aims to describe object properties using specific numbers to make quantitative descriptions of the observed objects. Measurement results are compared with other objects or the desired design value so that the difference between the two objects can be obtained. For ultra-precision manufacturing, form accuracy is a significant criterion since it directly determines the functional performance of the machined part.

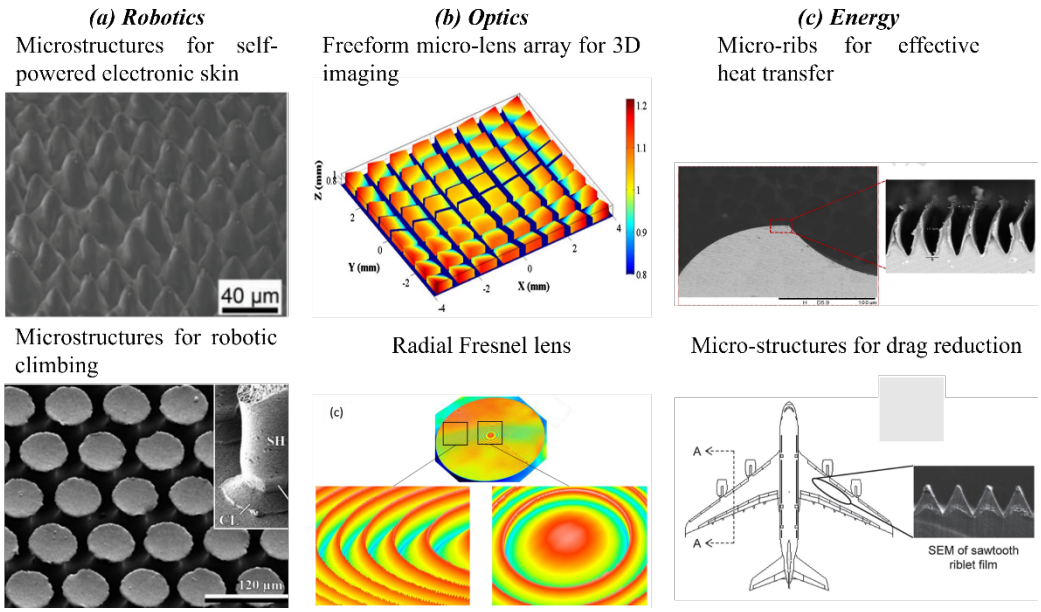

Figure 1.1 Applications of micro-structured surfaces in (a) robotics (Breckwoldt et al., 2015; Yao et al., 2020), (b) optics & imaging (Li & Allen, 2012; Zhang et al., 2018), and (c) energy (Bixler & Bhushan, 2013; Wang et al., 2015).

Surface measurement compares the machined surface form and the desired surface form so as to provide guidance on the post-machining process. Basically, dimensional measurement primarily concerns itself with assessing the dimensional and geometric accuracy of a workpiece's surface profile. The precision depends significantly on the cutting tool's movements relative to the workpiece. Different factors such as theoretical inaccuracy, geometric inaccuracy of machining tools and cutting tools, deformation of the machining system, and deformation of workpieces can contribute to the occurrence of form errors in the machined workpiece.

Widely used surface dimensional measurement techniques mainly include stylus profilometers, interferometry, deflectometry, confocal methods, structured light,

electron microscopy, autostereoscopy, etc. With the rapid advancements in computer science and artificial intelligence (AI) techniques, AI is identified as a promising tool in the field of measurement science and technology. It enables the exploration and extraction of implicit information from extensive measurement data, thereby improving the performance of existing measuring systems. This new approach brings a fresh perspective and opens up new possibilities for improving measurement accuracy and efficiency.

In the industry, contact profilometers are widely employed for surface measurement. These instruments utilize a contact stylus that moves in both vertical and lateral directions across the surface being measured. By doing so, the profilometer can accurately determine and record the distance between two measured points, as well as the contact force exerted during the movement. A contact profilometer requires no modelling, which makes it independent on the measured surfaces. It is also not sensitive to the specific properties of surfaces such as colour and reflectance that have a great influence on optical measurement instruments. Contact profilometers provide accurate, precise, and convincing measurement results, which contribute to their popularity in industry. However, two main drawbacks of contact profilometers are their time-consuming data acquisition process and the nature of contact on the measured surfaces. The interaction of the probe might cause damage to the surfaces, especially when the surfaces are soft and delicate.

Non-contact profilometers usually operate based on optical techniques. A vast amount of research on non-contact profilometers emerged over the decades, and various optical techniques have been employed in relation to measurement, including time of flight (ToF), laser triangulation, structure light, light field, etc. The most apparent merit of non-contact profilometers is that there are no additional force and effects interacting

with the measured surfaces. This suggests that non-contact instruments generally exhibit extended durability, while contact instruments that require interaction between a probe and a surface may cause wear on the probe tip over time.

Autostereoscopy technology, as an emerging optical technology, was first introduced in the measurement field by Li et al. (2014) who developed an autostereoscopic measuring instrument for on-machine/in-situ measurement based on integral imaging. In acquiring the optical information of measured surfaces, the instrument utilizes a micro-lens array (MLA). This MLA enables the recording of multiple perspectives, resulting in a collection of 2D elemental images that contain valuable light field information. These images are then utilized for digital refocusing and disparity extraction purposes. As a consequence, accurate on-machine measurement of micro-structured surfaces is achieved in a single snapshot. However, a primary constraint of the autostereoscopic system is the resolution quality of the obtained data. The compromise between spatial and angular resolution of autostereoscopic data restricts the effectiveness of measurements.

In light of the rapid advancement of artificial intelligence technology, the potential now exists to overcome the inherent resolution conflict in autostereoscopic measuring systems. This breakthrough can be achieved by leveraging artificial intelligence techniques to enhance the resolution capabilities of such systems. Following the proposal of the resolution-enhanced learning model (Dong et al., 2014), learning models have demonstrated their effectiveness in generating high-resolution data from low-quality inputs. The learning-based method can achieve automatic feature extraction and selection from the data without the requirement of expert experience and a priori knowledge. The complex mapping function from input to output, learned by learning models, is able to achieve a more accurate representation of the real world than

conventional methods. Various deep learning techniques including the residual architecture, the encoder-decoder network, the generative adversarial network (GAN), etc., have been designed and utilized to achieve a high-quality transformation from coarse information to detailed information.

Apart from 2D super-resolution techniques, learning models are also being used to enhance the quality of 4D light field data. The improvement in spatial resolution is akin to solving the single-image resolution-enhancement problem, wherein the objective is to interpolate pixels between two adjacent pixels in an elemental image. The redundant pixel-level information lying in the other elemental images captured from different perspectives contains high-resolution clues for the pixel reconstruction. In terms of the angular super-resolution problem, novel view elemental images from new perspectives are required to be generated based on the input featuring limited angular resolution. During angular super-resolution, pixel interpolation occurs in the epipolar plane rather than the 2D image plane. The new pixels are desired to be reconstructed between two corresponding points in two elemental images. However, the two corresponding points might not be adjacent in the epipolar plane when a large baseline exists in the light field data.

Depth-based and non-depth-based learning models have been presented to address the angular super-resolution problems. Non-depth-based methods usually can produce quite satisfying results when baselines are small, whereas image ghosting is generated for data with a large baseline. Depth-based learning models can avoid image ghosting by first performing preliminary disparity estimation. However, inaccurate estimation usually happens for data that record real-world scenes, especially those which contain various noises, occlusions, and complex illumination conditions. As a result, image artefacts are produced in the novel view images. Hence, the attainment of high-accuracy

reconstruction of high-resolution information remains a significant objective, ultimately leading to improved precision and accuracy in the measurement results acquired from the autostereoscopic 3D measuring system. As shown un Table 1.1, all the acronyms used in the thesis are listed for clarification.

Table 1.1 Acronyms Included in the Thesis.

Acronym	**Full Name**
CMM	Coordinate-measuring machine
CNN	Convolutional neural network
CP	Corresponding point
dB	Decibel
DEDI	Direct extraction of disparity information
EI	Elemental image
EPI	Epipolar-plane image
FOV	Field of view
GAN	Generative adversarial network
HR	High resolution
InI	Integral imaging
LF	Light field
LR	Low resolution
MAE	Mean absolute error
MLA	Micro-lens array
MSE	Mean squared error
PSNR	Peak signal-to-noise ratio
ReLU	Rectified linear unit
SAI	Sub-aperture image
SD	Standard deviation
SIFT	Scale invariant feature transform
SISR	Single image super resolution
SOTA	State-of-the-art
SR	Super-resolution
SSA	Self super-resolution autostereoscopic
SSIM	Structural similarity index measure
TAM	Traditional autostereoscopic measuring
Tanh	Hyperbolic tangent

The motivation is to leverage the high-speed capabilities of the autostereoscopic

system for data acquisition along with utilizing deep learning technologies to strengthen the quality of autostereoscopic data affected by insufficient illumination, various noise, and low resolution. By improving both the angular and spatial resolution through deep learning, achieving faster, highly accurate surface metrology is possible.

1.2 Research objectives

The research stated in the thesis aims to propose a high-resolution autostereoscopic 3D measuring system that can break through the inherent dilemma of the InI technology so as to achieve high precision measurement for micro-structured surfaces. The research objectives are outlined as follows:

(i) To propose a novel solution grounded in deep learning technologies to break through the resolution limitation of the autostereoscopic measuring system so that the measurement precision and accuracy can be improved.

(ii) To develop a novel generic learning paradigm to improve the sampling efficiency during the process of training angular SR learning models so as to improve the learning efficiency for small datasets

(iii) To develop a novel deep learning network that converts accurate depth estimation into a classification problem to achieve high-quality reconstruction of high-angular-resolution information with less image ghosting and image artefacts for both synthesis data and real-world data.

(iv) To develop a self super-resolution autostereoscopic measuring system that can realize self-enhancement solely based on the collected measurement data and achieve precision measurement for micro-structures.

(v) To develop a multi-frame super-resolution autostereoscopic system for on-

machine measurement so as to exploit the vibration of machine tools to collect multiple frames with subpixel displacement so as to achieve resolution enhancement of the measurement data.

1.3 Organization of the thesis

The thesis is divided into seven chapters.

Chapter 1 introduces the research background, highlights the current research gaps, and defines the research objectives. This chapter also outlines the structure of the thesis comprehensively.

Chapter 2 mainly focuses on a literature review of related research fields and investigates the latest development of the technologies. The review includes the research on precision surface metrology, contact and contactless measurement methods, autostereoscopic measurement methods based on InI, super-resolution techniques and reconstruction techniques applied for the InI systems, and deep learning approaches have been incorporated to boost the resolution of LF data collected by InI systems.

Chapter 3 discusses the fundamentals of the autostereoscopic 3D measuring system, including the data recording process and surface reconstruction process. Related reconstruction techniques such as digital refocusing, epipolar-plane images, and direct extraction of disparity information based on disparity patterns are discussed. The limitations and room for improvement of the autostereoscopic system are also presented in this chapter. An example of rapid 3D inspection of wire bonding based on autostereoscopic measurement is provided, and experimental analysis is presented to evaluate the capability of the autostereoscopic measurement system.

Chapter 4 introduces a deep learning model designed to enhance the angular

resolution of a LF captured using InI devices. A novel semi-supervised learning paradigm achieving high data efficiency is presented to train the deep learning model. Even a baseline method can be improved notably under the proposed semi-supervision. On the basis of the learning paradigm, a motion estimation network is proposed to achieve novel view synthesis through converting the regression problem of depth into a classification problem of motion. Experiments on public datasets are performed and the findings are discussed in the chapter.

In Chapter 5, a self super-resolution autostereoscopic measuring system is presented, which is able to achieve angular super-resolution solely relying on the data collected by the system. The integration of a learning-based algorithm is proposed for incorporation into the measuring system, with the goal of improving the resolution of the measurement data. The chapter elaborates on the composition of the learning-based algorithm and presents experimental results using various samples to assess the performance. Comparisons of the digital refocused images, the reconstruction point clouds, and the measurement results, based on the LR measurement data and the HR data enhanced by the proposed algorithm are provided to reveal the improvement of the autostereoscopic measuring system.

Chapter 6 provides a solution for on-machine measurement based on the autostereoscopic system and deep learning models. A multi-frame super-resolution solution is presented to make use of the vibration of the machine tools during on-machine measurement. Jitter analysis between different frames collected in various timespans during the on-machine process is provided to demonstrate the existence of subpixel displacement among the multiple frames. A deep learning SR model and its structure are elaborated in this chapter. Consequently, various experiments are presented to assess the presented learning model. A comparison of the measurement

results obtained using LR data and HR data concludes the chapter.

The final conclusion of the thesis is provided in Chapter 7, which outlines the key achievements of the research. Suggestions for further improvement are also made in this chapter to lead the way towards future endeavours.

Chapter 2 Literature review

2.1 Introduction

Measurement is essential for ultra-precision manufacturing, providing necessary feedback to machine tools. Both external and internal factors including vibration, thermal deformation, kinematic errors, etc. of machine tools result in form errors of measured parts during the ultra-precision machining process. The feedback obtained from measurements is utilized to identify and correct errors, thereby enhancing the quality of the machining process. In the last few decades, the instruments for offline measurement have been investigated and developed and have become a mature solution for precision measurement. However, removal and remounting of machined parts are unavoidable during offline measurement, which introduces extra errors. On-machine surface measurement is able to prevent remounting so that less transportation labour and time consumption are required.

Based on the nature of the probe used by a measurement instrument, the measurement is basically divided into contact and non-contact types. Contact measurement (Bauza et al., 2011; Lee et al., 2012; Lei et al., 2014; Yin et al., 2018), by definition, uses a stylus in contact with the measured surfaces under some scanning strategies so that groups of points are acquired to represent the profile of the measured surfaces. To reach higher resolution, finer tips are developed to measure micro-structured surfaces (Bauza et al., 2011; Lei et al., 2014). Contact measurement is mature and usually achieves high accuracy. However, the low scanning speed and the contacting nature make it inefficient in some situations, especially when the measured parts have soft surfaces. In terms of the limitations of contact measurement, non-contact

measurement that usually uses an optical probe for fast surface scanning without contact has emerged as a popular choice for ultra-precision measurement. Numerous optical probes based on various technologies such as interferometry, deflectometry, confocal, structured light, etc. have been developed to perform highly accurate metrology.

InI is an emerging technology to record light distribution that is known as the light field. Compared with traditional 2D imaging which solely records the 2D projection of light rays in the image plane, the directions of the rays are also contained in a light field. As a result, a 3D world can be described and reconstructed based on the abundant information in a light field. Albeit that InI was originally proposed to achieve photorealistic image-based rendering (Kim et al., 2009), the applications of InI have been extended to a wide range including 3D reconstruction, object detection, recognition, etc. InI systems position a MLA before the image sensor so that multiple 2D images (called elemental images) from different perspectives are recorded to acquire abundant 3D information. Plenoptic cameras which are also known as light field cameras are an alternative solution to record light field images, where an MLA is installed at the image plane of the main microscope lens (Howe et al., 2020) so that multiple micro-images named plenoptic frames are recorded by the image sensor behind the micro-lens array. Enormous amounts of research based on light field microscopy has been conducted in the biological sciences. In terms of ultra-precision measurement, pioneering work utilizing InI was introduced by Li et al. (2014) who presented an autostereoscopic 3D measuring system for micro-structured surfaces.

However, the deficiency of the autostereoscopic measurement system is the resolution of the acquired data. Because of the intrinsic conflict between the spatial and

the angular resolution of InI, it is challenging to enhance the two kinds of resolution simultaneously. In other words, a choice between more details of the target objects and more information from different perspectives must be decided. To tackle the issue of low resolution in light field images, researchers have developed projection-based methods and optimization-based methods. These techniques aim to enhance the resolution of images captured by InI systems or plenoptic systems.

Artificial intelligence has experienced significant advancements, leading to the extensive integration of deep learning in various industrial applications. Deep learning has seen outstanding success in multiple fields, including computer vision, natural language processing, and automated control. It has emerged into a benchmark approach for various tasks.. Deep learning has a strong capability to represent complex mapping functions, especially for image processing problems. Regarding super-resolution, deep CNN models with various network structures were developed to realize SISR that fully outperforms conventional methods. The emergence of generative adversarial networks further improves the resolution of super-resolved images, providing more realistic and natural textures. This allows possibilities to break through the bottlenecks of the autostereoscopy measurement system to achieve super-resolution of the recorded images. Through exploiting the powerful representation capability of the machine learning models, significant improvements to the effectiveness of the autostereoscopy measurement system are possible.

2.2 Three-dimensional precision measurement

Tactile profilometers are a kind of contact measurement instrument which are commonly used in the manufacturing industry. The tactile device used in profilometers is a stylus (also called a probe) that comes into contact with the actual measurement

surface directly. Contact measurement is a mature technique and can achieve accurate and precise measurement results. Another reason for the popularity of contact measurement is that the measurement data can be easily compatible with past accumulated data. Figure 2.1 illustrates the process of data acquisition. The measurement process is dependent on the movement of the stylus which moves vertically and laterally in contact with the measured workpiece and sequentially, and all the contact points are collected and stored, forming a point cloud representing the measured surface. With the vertical displacement of the stylus, the profilometer can measure small surface variations, achieving precision measurement.

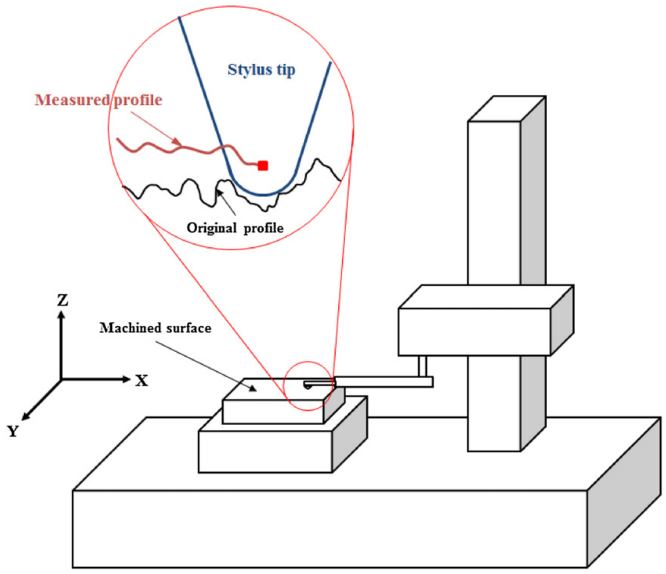

Figure 2.1 Acquiring profile data through a stylus-type profilometer (Lee et al., 2012).

To achieve measurement on micro-dimension surfaces, Bauza et al. (2011) presented a fine tactile sensing probe that was vibrated by a quartz crystal oscillator. The oscillator generated a standing wave so that the tip was able to move a larger distance than the other location of the rod. This guaranteed that the fine tip interacted with the measured surfaces before the rod did so. A diagram illustrating the system

setup is displayed in Figure 2.2, where the probe is only 7 µm in diameter and 3.5 mm in length. The probe was integrated into a scanning system and the interaction between the tip and the surface was defined as a function of the probe radius. The wave amplitude was highly controllable at the micrometre scale. The system was evaluated using microscale holes and a sample with surfaces that were difficult to access. As a result of the probe's high aspect ratio, the presented measurement system was able to measure deep and narrow features.

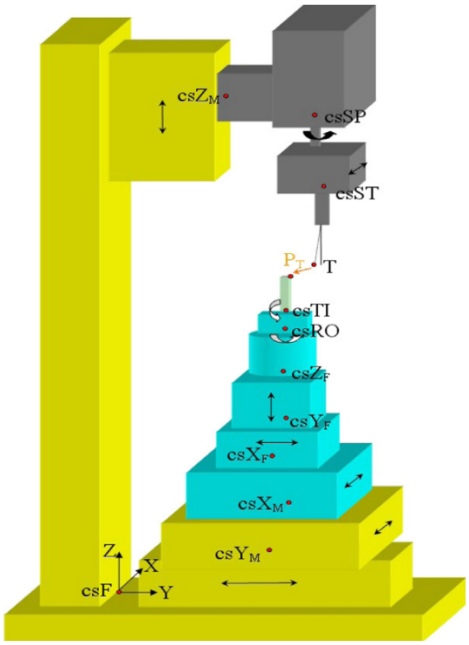

Figure 2.2 Diagram of a contact measurement instrument for micro-dimension metrology (Bauza et al., 2011).

Lei et al. (2014) presented a micro tactile sensor as a probe to be integrated with a nano measuring machine (NMM) to achieve nano-precise dimensional measurement of microstructures. Shown in Figure 2.3 is a diagram of the measurement system alongside the stylus. To attain a high measurement resolution, the stylus used for data acquisition had a length of 13 mm and a probing sphere with a diameter of 300 µm. The relationship

between the sensor response and the structure parameters was modelled so as to analyze the sensor design. To demonstrate the system's achievement of resolution, experiments were conducted in which a resolution of 5 nm was achieved in the z direction, and 10 nm in the x/y directions.

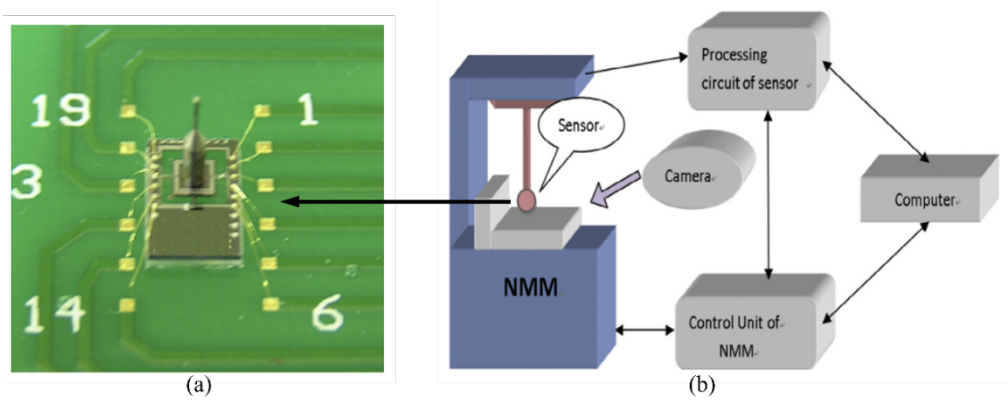

Figure 2.3 A tactile profilometer for the measurement of microstructures. (a) Photograph of the probe. (b) Diagram of the system setup (Lei et al., 2014).

Yin et al. (2018) suggested a contact profilometer for triangular microstructures and developed a compensation model to correct the errors resulting from the tilting of the sample plane and the dimension of the probe tip. A visual representation of the proposed system is included in Figure 2.4, where the measuring system is composed of a precision positioning stage, a stylus system with a diamond micro-stylus, and a vibration isolation table that was used to decrease the effects resulting in the perturbation of the measuring environment. Experiments showed that the contact profilometer achieved high accuracy for the measurement of triangular microstructures.

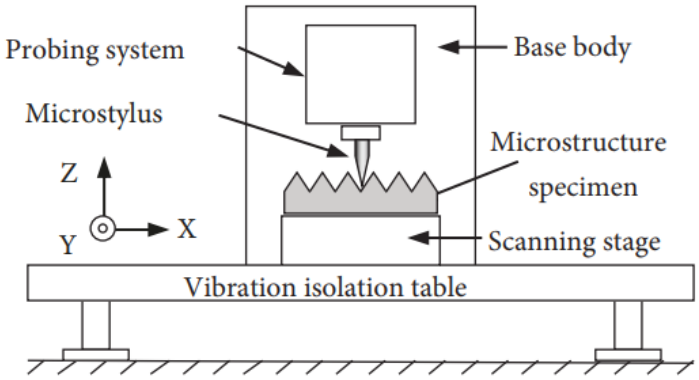

Figure 2.4 Diagram of a contact profilometer for measurement of triangular microstructures (Yin et al., 2018).

Although contact measurement is able to achieve high accuracy, the sampling process is usually slow, and the sampling strategy needs to be delicately designed. It is also obvious that the actual profile is different from the measured profile owing to the dimensions and form of the stylus tip, though many researchers have developed finer tips to reach a high resolution. To investigate the distortion resulting from the size of the stylus tip, Lee et al. (2012) constructed a simulation model to denote the real contact mechanism and provided suggestions for the proper selection of the stylus according to the characteristics and structure of the measured workpiece. Clark and Greivenkamp (2002) proposed an iteration algorithm to correct the stylus error particularly for smooth surfaces after analyzing the generation of the errors, and Ahn et al. (2019) compensated for the error by predicting the actual contact points using a least square fit and comparing the difference between actual and theoretical displacement to obtain the compensation values.

Another limitation of contact measurement is the nature of the stylus as inevitable damage and scratches may be caused by the contact, especially when the measured surfaces are soft (Li et al., 2019). This limits the application of profilometers in some

specific working situations. Hence, contactless measurement methods are gradually gaining significance in metrology..

Non-contact measurement methods basically depend on optical and computer technologies. Compared with contact methods, optical measurement methods require no touching of the surfaces, less time for measurement, less dependence on the conditions of the environment, and easier operation of deployment. This gives non-contact measurement methods more opportunities to be put under the spotlight.

One of the optical methodologies used for non-contact measurement is time of flight (ToF). It acquires and records the travel time of an object (e.g., particle, wave, ultrasonic signals, etc.) through a medium and conducts measurement based on the time. Robertson et al. (2002) made use of ultrasonic signals to realize surface metrology. As shown in Figure 2.5, an ultrasonic signal is focused onto the measured workpiece and then reflected from the surface through the air and returned to the transducer. To acquire all the features of the target surface, all of the waveform is acquired by the system.

Therefore, depth information for the surface is obtained using the signal's time of flight. As for laser signals, Tian et al. (2009) introduced a measurement system that utilizes pulsed ToF laser radar for the purpose of measuring hot forgings. In contrast to conventional methods that rely on XY guide rails, floating platforms, and theodolites to support the scanning device, the proposed system employed a scanning device based on a two-degrees-of-freedom spherical parallel mechanism. Compared with other devices, it produced more accurate results by avoiding accumulated errors resulting from parallel or serial driving.

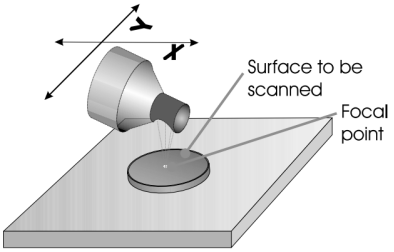

Figure 2.5 Surface metrology through the time of flight of ultrasonic signals (Robertson et al., 2002).

Interferometry is another important technology used in non-contact measurement. The interferometry of these measurement methods is often combined with microscopy so that both a high resolution and large vertical range can be realized. One example (Wyant, 2002) that uses a white light interferometer for measurement is demonstrated in Figure 2.6, where a two-beam Mirau interferometer was used at the microscope objective. Interference occurred due to the interaction between the ray reflected off the measured surface and the ray reflected off the reference surface. The detector array recorded the interference pattern as images, and through post-processing the measurement result was acquired.

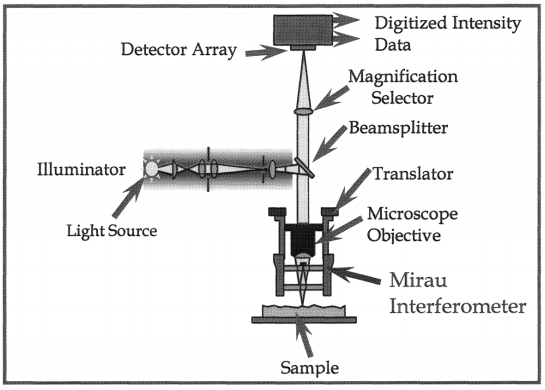

Figure 2.6 White light interferometer (Wyant, 2002).

Research on interferometers covers many fields, including using three-wavelength

light (red-green-blue) to inspect discontinuous structures (Pförtner & Schwider, 2003), increasing the speed of signal processing by using ellipse parameters (Dai et al., 2004), reducing the vibration of the measurement environment by active control so as to enhance the resolution of interference (Zhao & Burge, 2001), realizing high-resolution angle metrology using heterodyne interference (Hahn et al., 2010), etc.

In terms of the difficulty of the calibration of freeform surfaces, Hao et al. (2016) presented a digital moiré interferometric technique to correct the alignment errors for non-null interferometry. A schematic of the proposed technique is shown in Figure 2.7, where two components including a virtual interferometer and a real interferometer comprise the whole system. To identify the designed residual wavefront aberration of the interferometer, a theoretical model was developed to achieve accurate prediction of the aberration to reduce the alignment errors. Through the accurate simulation of the real interferometer, only coarse alignment was required to achieve good measurement repeatability, even when obvious alignment errors existed.

To realize adaptive interferometric null testing for unknown freeform surfaces, Huang et al. (2016) made use of a deformable mirror to realize the adaptive null measurement. The presented system is demonstrated in Figure 2.8, while the shape of the deformable mirror was refined employing the stochastic parallel gradient descent algorithm. Deformable mirror shape was precisely detected by an on-machine deflectometry system. In addition, a computer-generated hologram was integrated into the system to further compensate for the nominal wavefront deformation in case the deformation of the deformable mirror was insufficient. The final measurement profiles of the tested surfaces were acquired by combining the results of the deflectometry system and the interferometer data.

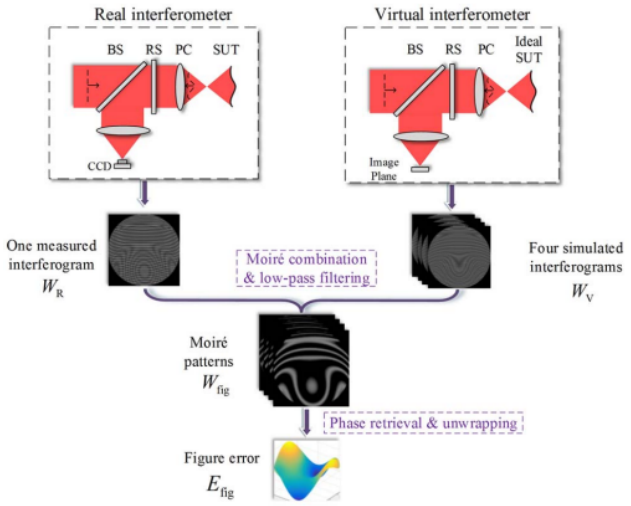

Figure 2.7 Schematic of a digital moiré interferometric technique presented in Hao et al. (2016).

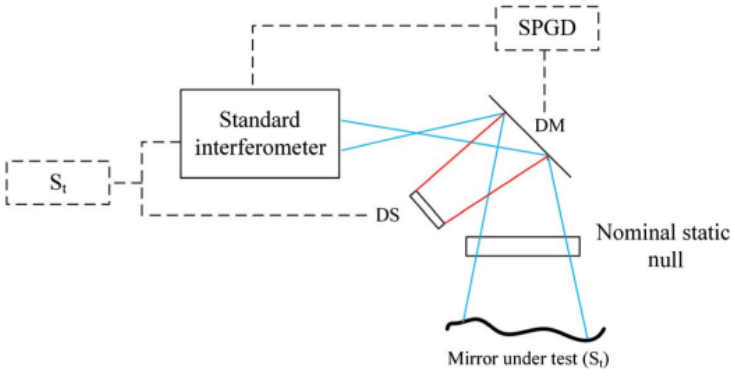

Figure 2.8 Diagram of an adaptive interferometer for the accurate evaluation of unknown freeform surfaces (Huang et al., 2016). DM, deformable mirror; DS, deflectometry system.

Lateral optical distortion is another challenge for coherence scanning interferometry. Systematic errors in the results of surface topography could be caused by the distortion. To reduce the errors, Ekberg et al. (2017) developed a correction model of optical distortion leveraging arbitrary surfaces to improve the measuring precision of coherence scanning interferometers. The correction process is shown in

Figure 2.9, where the calibration was executed in accordance with a subpixel image correlation method. Intensity maps were first extracted from a stack of data acquired by the coherence scanning interferometer through the detection of the best focus position of each pixel. Calibration grids were searched and determined using the subpixel image correlation method, and the final distortion and correction functions were obtained by a 2D self-calibration algorithm.

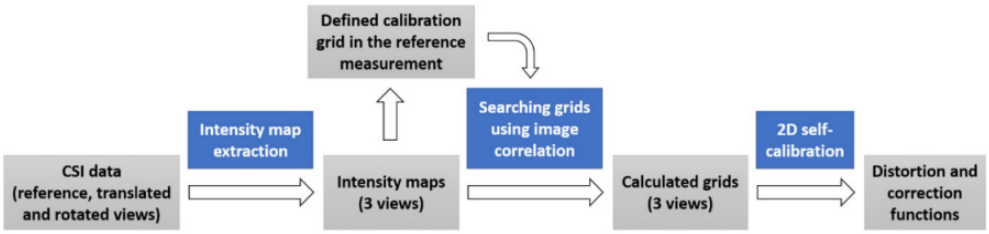

Figure 2.9 Correction model of optics distortion for coherence scanning interferometers (CSI) (Ekberg et al., 2017).

However, it is observed from the aforementioned research that a nominal static null component such as the computer-generated holograms are usually required during measurement using interferometers. In addition, the interferometers are usually sensitive to environmental factors including pressure, vibration, temperature, etc. (Faber et al., 2012). Deflectometry as an incoherent technique is used to develop surface measurement systems which are more tolerant towards environmental disturbance and require no extra null testing.

Maldonado et al. (2014) proposed A lightweight solution for HR surface measuremen based on the deflectometry technique, with the system setup as shown in Figure 2.10. The researchers constructed a reverse ray model on the basis of reflection principles. Deflected patterns were detected and determined by phase shifting methods. Consequently, The mapping between the camera pixels and the sample points on the

tested surfaces was determined and calibrated. The surface profiles were finally fitted using Noll Zernike polynomials. The deflectometry measurement system achieved high dynamic range measurements.

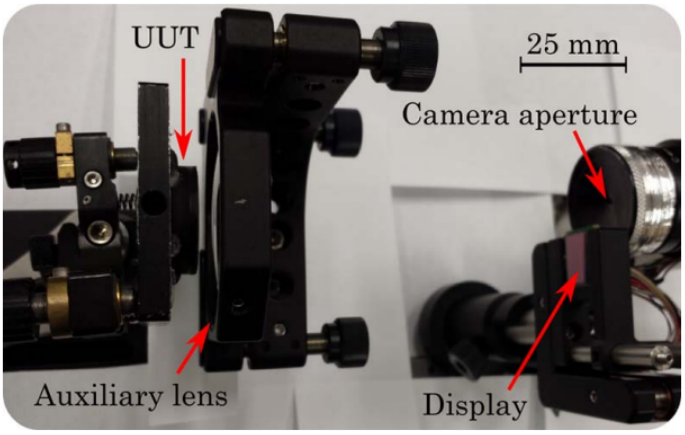

Figure 2.10 A portable deflectometry measurement system (Maldonado et al., 2014). UUT, unit under test.

To measure surfaces with steep slopes, Liu et al. (2017) presented a surface measuring system operating on the basis of direct phase-measuring deflectometry. Since classical phase-measuring deflectometry systems may have difficulty in measuring samples with multiple discontinuous surfaces effectively, direct phase-measuring deflectometry methods exploit two liquid crystal display (LCD) screens to form a parallel design which is more effective for the measurement of discontinuous surfaces. A diagram of the direct phase-measuring deflectometry method for 3D surfaces and its corresponding hardware setup are shown in Figure 2.11. The presented system constructed a mapping model between absolute phase maps and depth information of the tested surfaces through locating a liquid crystal display screen at two known positions. Through this configuration and operation, the desired depth was able to be directly determined from the phase map.

Structured light projection is a process to artificially generate patterns on recorded images to achieve matching. In general, the structure light systems widely make use of sequential-shot and single-shot schemes. Sequential 3D imaging systems adopt a series of patterns which are projected onto target surfaces. Multiple shots are required, and therefore the targets need to be static. In terms of dynamic targets, only single shot is acceptable. Hence, each pixel needs to be indexed uniquely to establish the correspondences.

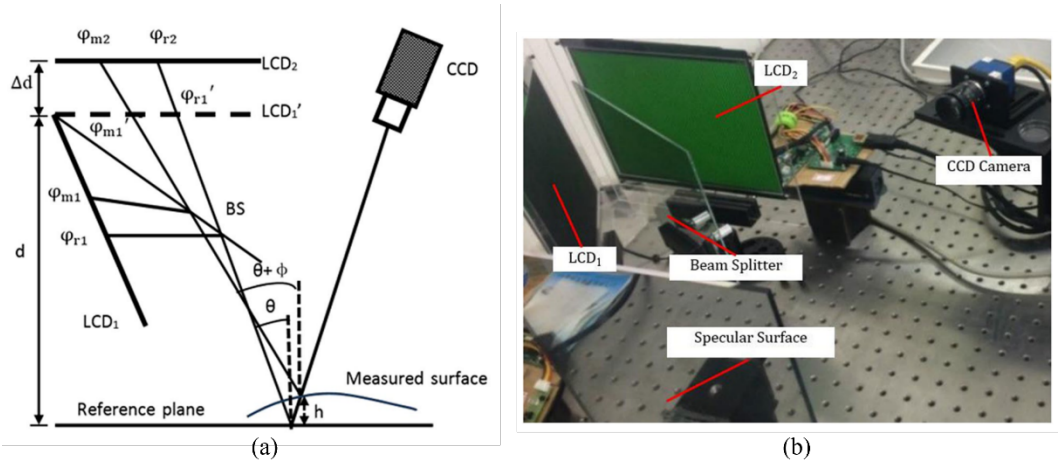

Figure 2.11 A direct phase-measuring deflectometry system for surface measurement proposed in Liu et al. (2017) (a) Schematic. (b) Hardware setup.

Li et al. (2021) presented a 3D measuring system that relies on structured light projection, making use of a divergent multi-line laser. The system was found to be simpler to implement compared to the traditional parallel multi-line laser system. Furthermore, this system achieved superior measurement accuracy and denser reconstruction in comparison to using a single-line laser. A diagram of the proposed structured light measurement system is illustrated in Figure 2.12. The pattern extraction was realized using the Steger algorithm based on a Hessian matrix and the fitting of the

light plane was achieved using the RAndom SAmple Consensus (RANSAC) algorithm. However, the experiments were only performed on macro-scale samples with simple surfaces.

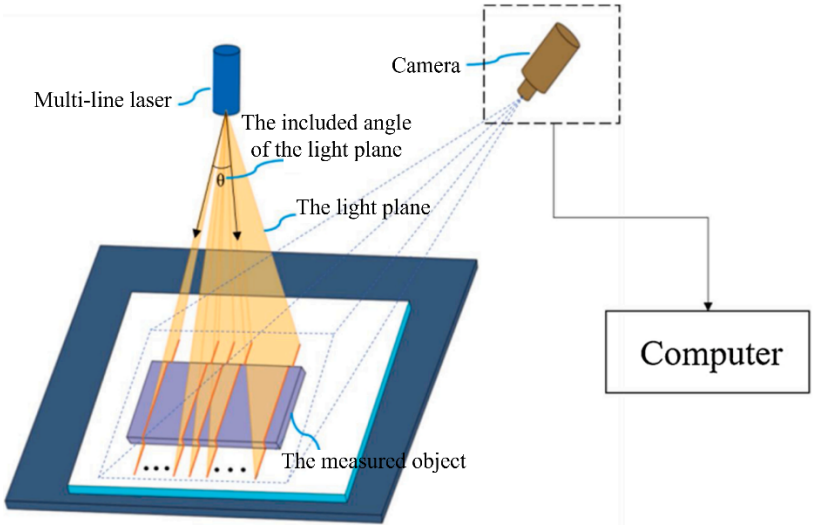

Figure 2.12 Diagram of a 3D measurement solution incorporating structured light projection (Li et al., 2021).

Over the years, great progress has been made in structured light techniques. However, challenges still exist. Multi-shot structured light systems usually produce high-accuracy imaging results, but the targets are required to be static. In addition, the measurement of micro-structured surfaces using structured light is still challenging.

In terms of the surfaces with large slopes that are difficult to access, the confocal technique is applicable to the measurement of complex micro-structured surfaces. Zou et al. (2017) investigated the application of the chromatic confocal technique in on-machine measurement to achieve nanometre-scale accuracy. Figure 2.13 showcases the system diagram and setup of a chromatic confocal probe incorporated into a high-precision diamond turning machine. Measurements were performed by mounting the confocal probe on the translation stage of the y-axis. A master sphere was mounted on

the vacuum chuck. Another reference sphere with a radius of 6.2 mm was mounted on a transition arm. The two spheres were used for accurate calibration of the measurement system. During measurement, the transition arm with the reference sphere and the confocal probe were required to be removed for component protection. Accurate reinstallation and recalibration were necessary for this system to guarantee the measurement accuracy.

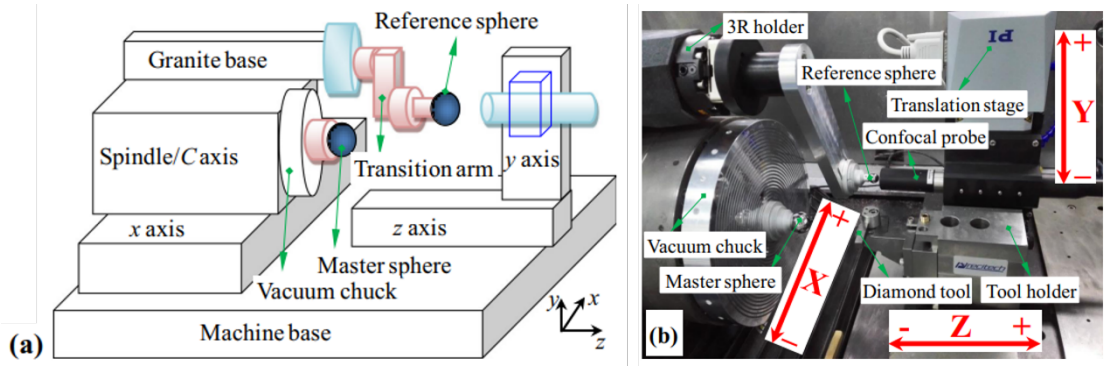

Figure 2.13 On-machine system method leveraging chromatic confocal (Zou et al., 2017). (a) System diagram. (b) System setup.

Fu et al. (2020) developed a confocal-based measuring system for surface roughness measurement. The system demonstrated effective inspection of surface quality in the mass finishing process. A diagram of the confocal-based measuring system and its system setup are shown in Figure 2.14, where a commercial chromatic confocal probe was mounted on a linear stage. In addition, an industrial robot arm was installed in the system to achieve highly accurate positioning. A reference specimen was used to evaluate the performance of surface inspection and two curved blades that possessed different degrees of roughness and were manufactured via 3D printing were tested using the presented confocal-based measuring system. The results showed that the relative errors were kept within 5%, which demonstrates the viability and efficiency

of the confocal-based system.

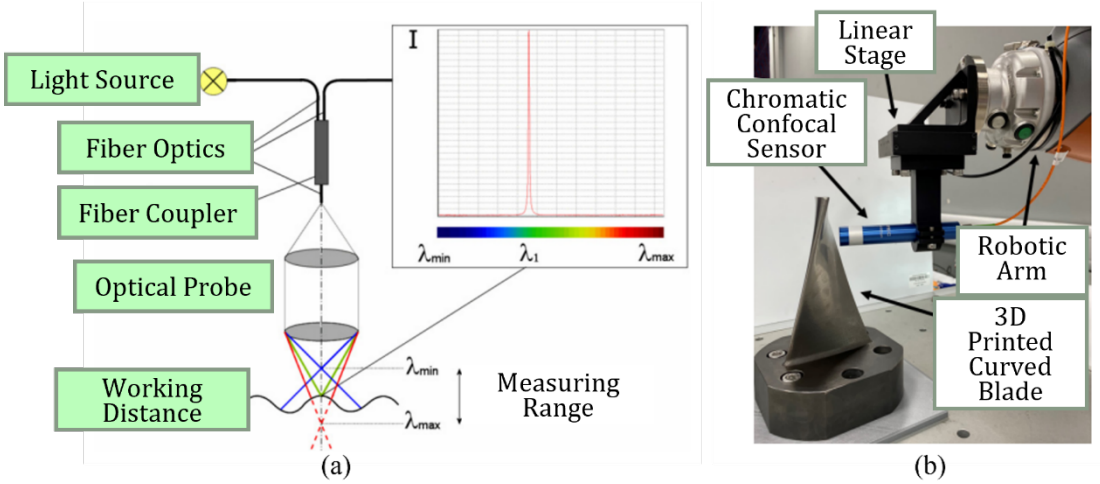

Figure 2.14 An on-machine confocal-based measuring system for surface roughness inspection (Fu et al., 2020). (a) System diagram. (b) System setup.

2.3 Autostereoscopic three-dimensional measurement

Autostereoscopy is an imaging and display technology which takes advantage of InI and LF techniques, and provides extra depth information compared with binocular systems. It is mainly used for glass-free 3D displays and virtual reality. Image integrating has been applied in many fields other than photorealistic rendering and display, such as 3D reconstruction, light field microscopy, biology, etc. The technology also plays a role in precision metrology. In this section, the development of InI technology is reviewed, and SOTA techniques are discussed. Additionally, the applications of InI in surface measurement are explored.

2.3.1 Integral imaging and plenoptic systems

The initial concept of InI was first presented in 1908 by Lippmann (1908) for real

3D display. Figure 2.15 demonstrates the recording process of the InI system. A set of images captured from various angles of a 3D scene are recorded by incorporating a MLA in front of a photographic film. These images with small lateral magnification are called elemental images. Overlapping usually happens in these elemental images following the initial concept. To further enhance the lateral resolution of elemental images, research on multi-camera systems (Lin et al., 2015) was investigated by directly recording high-resolution images from various perspectives using multiple synchronized cameras.

Winnek (1936) implemented the initial InI principle using a traditional camera to avoid the overlapping and much smaller images named microimages were acquired. Based on the system presented by Winnek (1936), Bergen and Adelson (1991) improved the system and proposed the initial plenoptic formalism through placing the micro-lens array right at the image plane. A diagram of the recording process of plenoptic cameras is shown in Figure 2.16. Similarly, the pixels in microimages are able to be rearranged to compose a series of elemental images from different perspectives. These images are also called sub-aperture images in the plenoptic field.

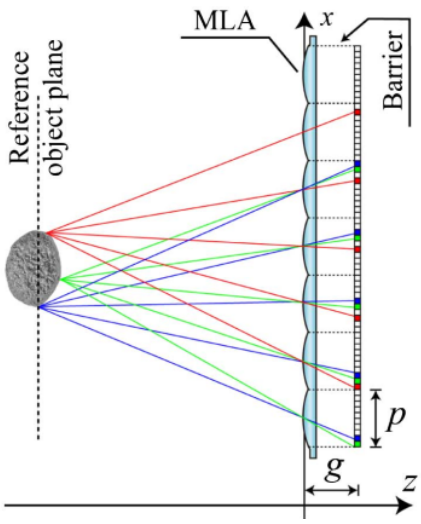

Figure 2.15 Scheme of InI. (Martínez-Corral et al., 2018).

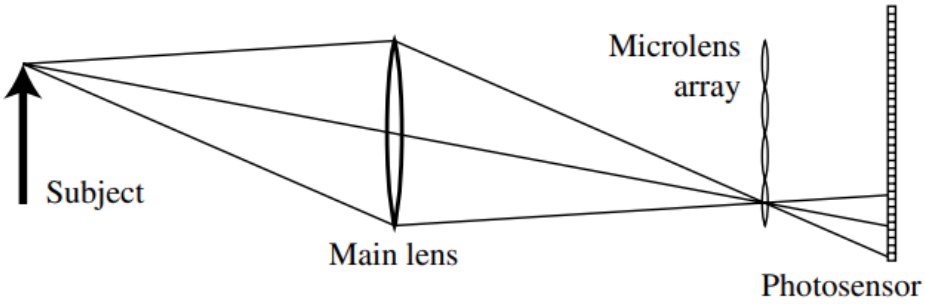

Figure 2.16 Scheme of plenoptic cameras. The MLA is situated at the focal plane (Ng et al., 2005).

Over the decades, both of these two development paths have been pursued. Although the configurations of InI systems and plenoptic cameras are different, the spatial–angular information lying in the captured light field is majorly similar. Many commercial products were also produced to have effects in many industrial fields, including biomedical applications, wavefront sensing, head-mounted display applications, etc.

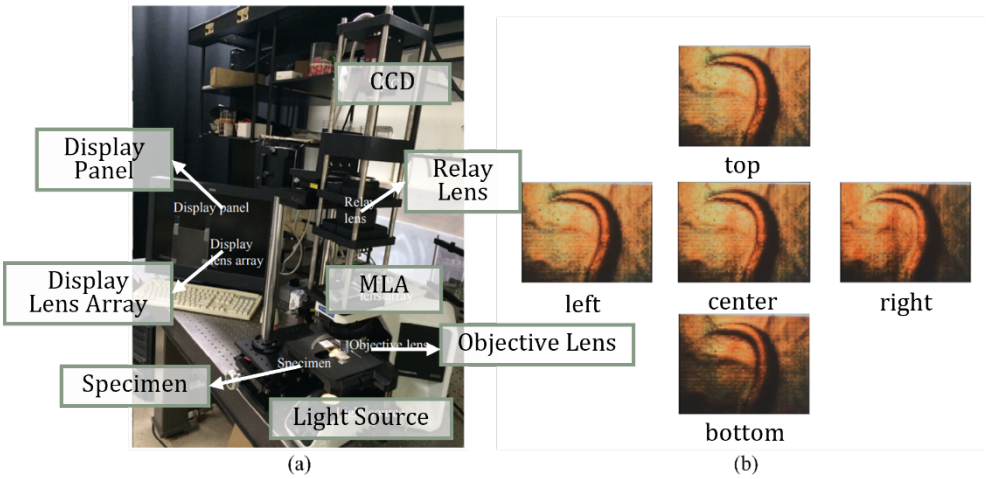

Figure 2.17 InI used for real-time LF microscopy (Kim et al., 2014). (a) InI system. (b) Organism observations from various perspectives using the real-time InI system.

Kim et al. (2014) presented a real-time InI system to achieve in-vivo observation of a living organism. The real-time system is shown in Figure 2.17 (a) and the observation results are exhibited in Figure 2.17 (b). An incoherent light source was used to illuminate the specimen and a relay lens passed the light field data to the image sensor. The capture rate reached 32 FPS and half of them were used for light field rendering. Similarly, Hua and Jia (2020) developed a Fourier light field microscopy system to realize high-resolution 3D live cell imaging. The lateral resolution and axial resolution reached 300~700 nm and 500~900 nm, respectively.

2.3.2 Three-dimensional reconstruction for light field data

Since the data collected by an InI system or a plenoptic system contains redundant stereo information in a light field, the 3D scenes are able to be reconstructed based on these 3D cues. The depth estimation methods for light field data are generally categorized as methods based on multi-view stereo (MVS), methods based on epipolar-plane image (EPI) technologies, and defocus-based methods. Among these, a vast literature has investigated the MVS-based methods (Heber & Pock, 2014; Jeon et al., 2015; Yu et al., 2013). MVS-based depth estimation methods usually perform a stereo matching process based on the multiple elemental images (also called sub-aperture images in plenoptic systems). Jeon et al. (2015) utilized the Fourier transform to convert the collected images from the space domain into the frequency domain so that the subpixel shifts among these images were able to be estimated in the frequency domain. The desired depth map was estimated by minimizing the matching cost between the shifted central sub-aperture image and the other sub-aperture images after the subpixel transformation. The estimation process and ultimate outcome is shown in Figure 2.18.

The matching cost included two components that are absolute difference and

gradient difference. The final depth was determined by solving the optimization problem. In addition, other constraints were used and aggregated to the objective function, forming a multi-objective optimization problem. The constraints were derived from confident matching correspondences, implying that the correspondences should also be matched at salient feature points. This would provide a strong constraint for a satisfactory optimization result. The SIFT algorithm was employed as the feature extractor during the matching process.

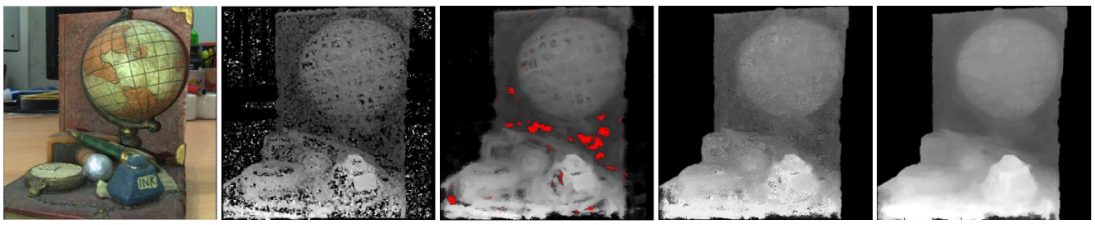

Figure 2.18 A method for LF depth estimation. (a) The central view image. (b) A disparity map estimated after minimizing the matching cost. (c) The map refined using a median filter. (d) Further optimization of the problem with constraints. (e) The final disparity map obtained by converting the discrete one into a continuous map.

Heber and Pock (2014) proposed another matching method for the depth estimation of light field data. Inspired by robust principal component analysis, they tried to find a warping method that mapped the sub-aperture images to a certain space so that the matrix merged by these new vectors had a low rank. A similar idea initiated by Yu et al. (2013) used light field triangulation as the matching method. On the whole, the key issue of these MVS-based depth estimation methods is choosing an appropriate projection space for the matching process and determining an effective way in which

the light field data are matched.

Epipolar-plane images are another widely used technology for the depth estimation of light field data. One example is illustrated in Figure 2.19 where the patterns in $X-U$ and $Y-V$ slices are epipolar-plane images. Figure 2.19 illustrates the generation of epipolar-plane images, which describes both the recording process of light field data and the extraction of epipolar-plane images. After obtaining the elemental images (or sub-aperture images), one row or one column of images are stacked and cut, and the cut section is an epipolar-plane image (the right top in Figure 2.19).

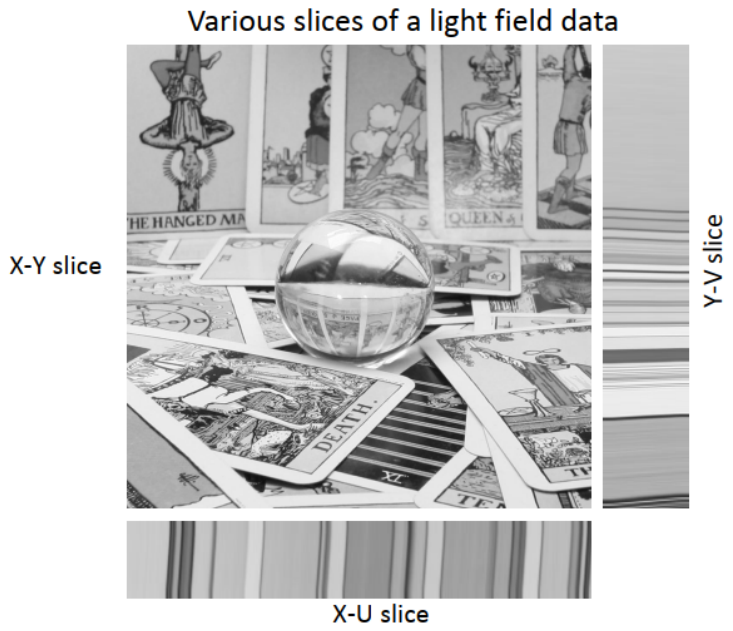

Figure 2.19 Various slices of light field data (Mitra & Veeraraghavan, 2012).

Research on EPI-based depth estimation methods (Johannsen et al., 2017; J. Li et al., 2015; Wanner & Goldluecke, 2014; Zhang et al., 2016) is in general based on the relation between EPI slopes and disparity. The generation of EPIs is shown in Figure 2.20. In an EPI, a larger slope corresponds to larger disparity. Hence, the key issue of

EPI-based methods is estimating the slopes of lines in EPI slices.

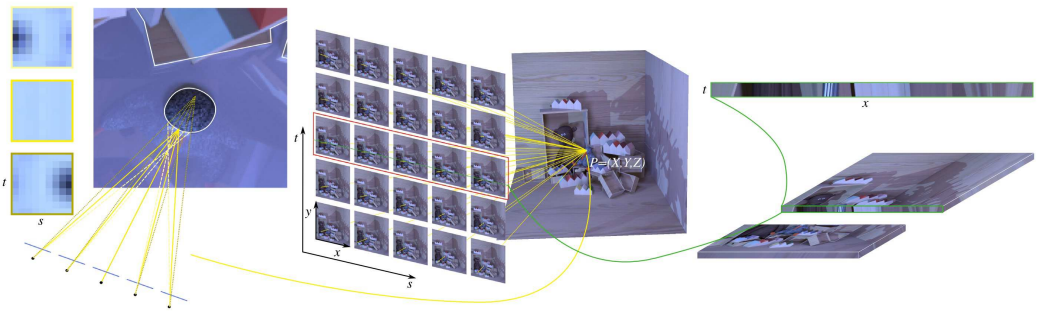

Figure 2.20 Generation of epipolar-plane images (Johannsen et al., 2017).

Wanner and Goldluecke (2014) used a Gaussian smoothing operator to perform estimation. A structure tensor was first estimated based on the epipolar-plane images, and depth estimation was determined using the structure tensor. The structure tensor was expressed as

$$J = \begin{bmatrix} G_\sigma \times (S_x S_x) & G_\sigma \times (S_x S_y) \\ G_\sigma \times (S_x S_y) & G_\sigma \times (S_y S_y) \end{bmatrix} = \begin{bmatrix} J_{xx} & J_{xy} \\ J_{xy} & J_{yy} \end{bmatrix} \tag{2.1}$$

where S is the epipolar-plane image, J is the structure tensor, G_σ is the Gaussian smoothing operator. S_x and S_y are the gradients. Then, the slopes n in S were obtained from

$$n = \begin{bmatrix} \Delta x \\ \Delta s \end{bmatrix} = \begin{bmatrix} \sin(\varphi) \\ \cos(\varphi) \end{bmatrix}, \varphi = \frac{1}{2} \arctan\left(\frac{J_{yy} - J_{xx}}{2J_{xy}} \right) \tag{2.2}$$

With the known focus distance f, the depth Z was estimated as

$$Z = -f \frac{\Delta s}{\Delta x} \tag{2.3}$$

Defocus-based depth estimation approaches (Tao et al., 2013; T.-C. Wang et al., 2015; Zhu et al., 2017) firstly defocus the light field data and then make use of the

defocus depth cues to reconstruct depth maps. A demonstration of using defocus cues to perform depth estimation is shown in Figure 2.21. Tao et al. (2013) made an appropriate combination of defocus cues and MVS cues to generate high-quality depth estimation. As a result, a two-stage refocus-based depth estimation method was presented, which firstly used these two cues to perform the initial estimation of disparity and then computed the confidence of each cue so as to combine them together using Markov random fields.

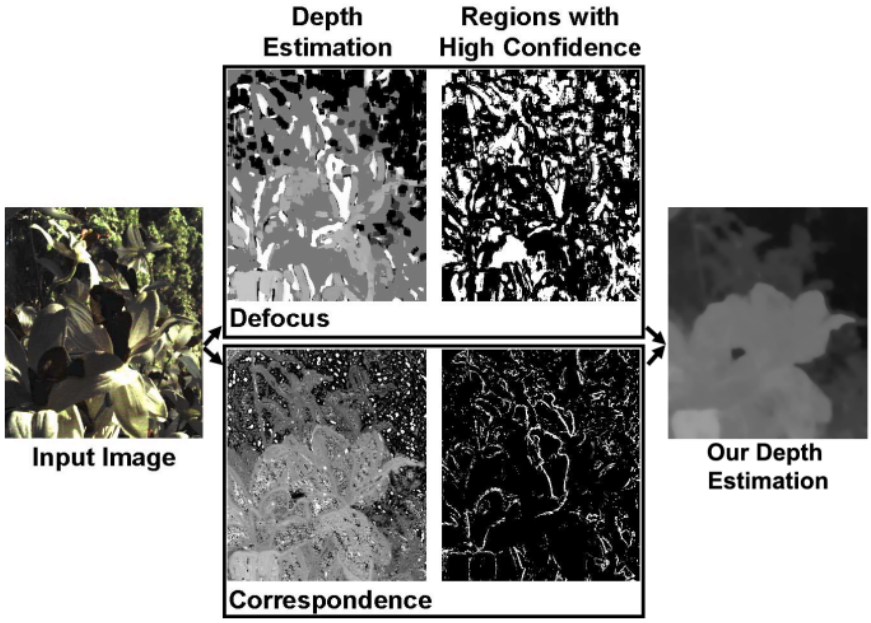

Figure 2.21 Depth estimation using both defocus cues and correspondence cues (Tao et al., 2013).

Wang et al. (2015) further improved the defocus-based method by putting occlusion into consideration. To determine which pixel point was occluded, the researchers modelled an occlusion predictor based on defocus cues, MVS cues, and depth cues from an initial depth map estimated by Tao et al. (2013). Finally, the prediction results of occlusion were combined with the initial depth map using a Markov random field to realize high-quality depth estimation. The method produced a

quite inspiring result as shown in Figure 2.22.

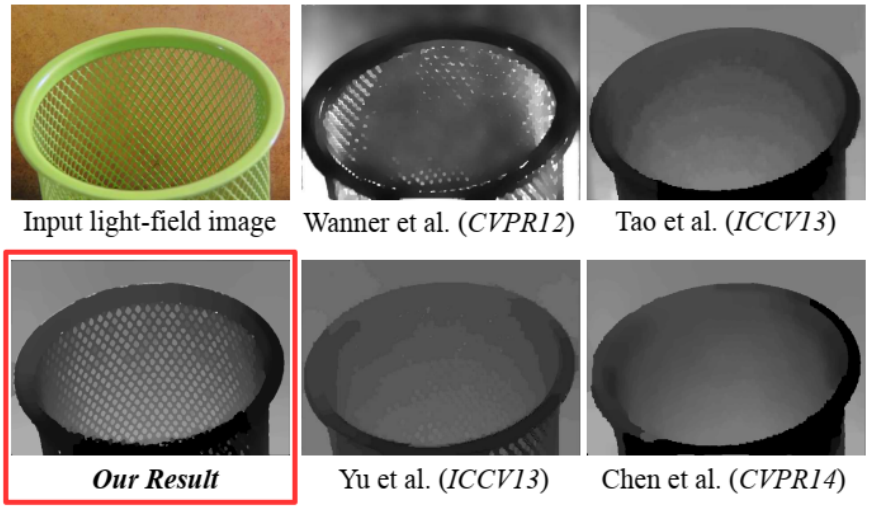

Figure 2.22 Comparison of the occlusion-ware defocus-based depth estimation method and other light field depth estimation methods (T.-C. Wang et al., 2015).

2.3.3 Autostereoscopic measurement systems

The pioneering research on the application of autostereoscopy in measurement systems was conducted by Li (2020), offering an innovative solution for on-machine micro-structured surface measurement. Figure 2.23 depicts a schematic diagram of the autostereoscopic measurement system. Taking a rectangular pyramid model as the measured workpiece, the measurement process consists of information capture and 3D reconstruction based on the recorded ray distribution. As shown in Figure 2.23, a MLA is positioned before an image sensor. The sensor plane is labelled as an elemental image plane to record the elemental images from multiple perspectives. At the information capture stage, a sequence of 2D elemental images, each corresponding to the number of micro-lenses, is captured. A slight difference among the captured images exists since the micro-lenses change the propagation direction of the optical rays emitted from the

measured object. A single point of the object recorded in different elemental images has different $X-Y$ coordinates, and these different pixel points originating from the same object point are called corresponding points (CPs). The reconstruction process reverses the information capture process, with a symmetrical framework. The elemental images are directly used to reconstruct 3D information according to the disparity determined by the corresponding points. The disparity information indicates the depth information so that the surface profile is able to be reconstructed. The reconstructed images with focused and defocused points are formed by rearranging the corresponding points captured in the elemental images. The reconstructed images from the InI system display focused points that indicate the depth of the corresponding object points.

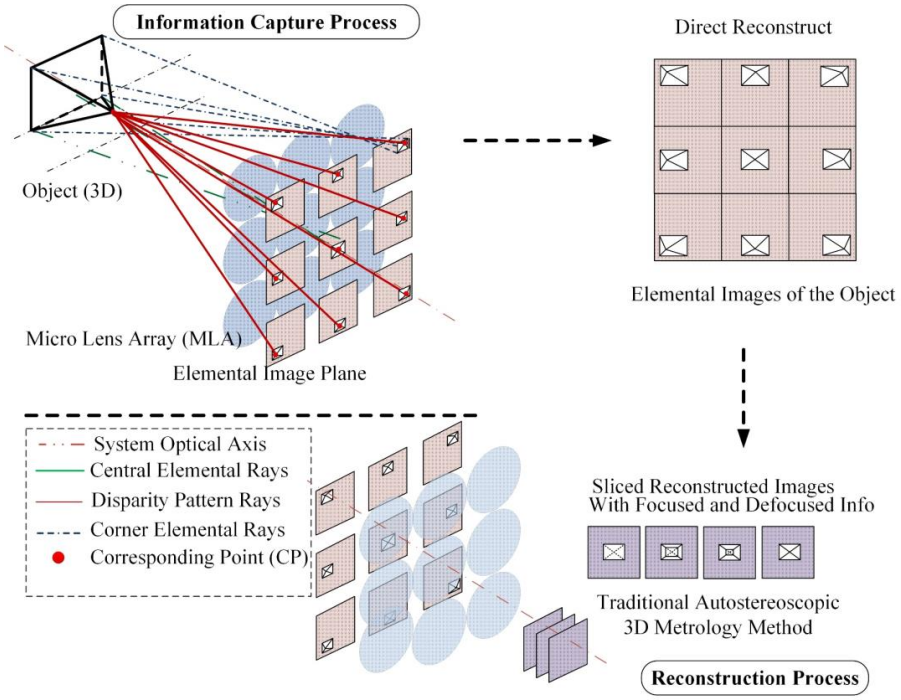

Figure 2.23 Diagram of the autostereoscopic measuring system (Li, 2020).

Based on the InI principle, Zhou et al. (2020) developed a 3D light field measuring system for specular surface measurements, with the system setup as shown in Figure

2.24. A diffuse light source was used to reduce the influence of the specular surfaces, and a polarizer was incorporated to further decrease the rays caused by specular reflection. A relay lens was employed to transmit the rays from the micro-lens array to the image sensor. The measured sample was a small tin wire stick as shown in Figure 2.24(a).

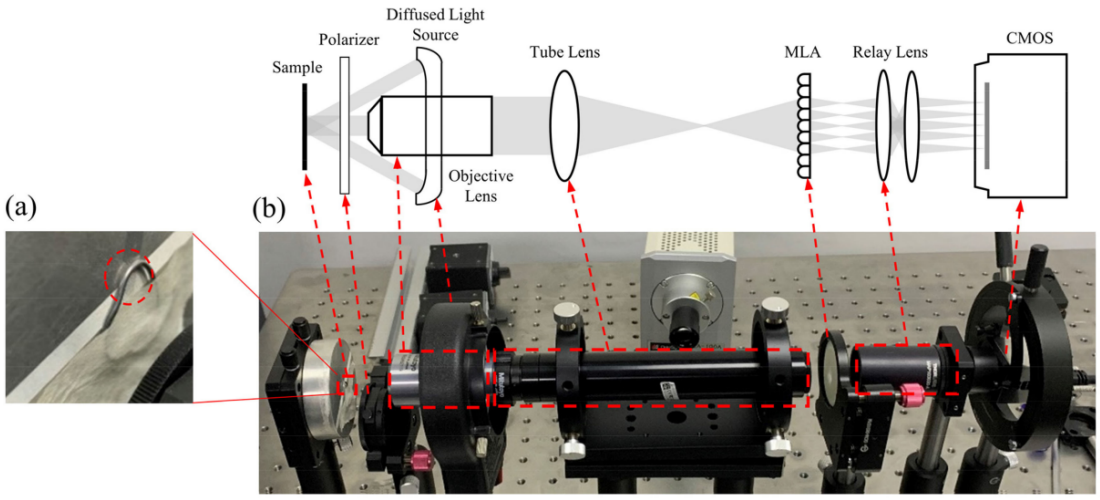

Figure 2.24 A 3D light field measurement system for the metrology of specular surfaces (Zhou et al., 2020). (a) Measured sample. (b) System setup.

Fundamental limitations are involved in the working principle of the autostereoscopy measurement system. The resolution and number of elemental images, namely the spatial and angular resolution of LF data, are influenced by the increase or decrease in micro lenses. This effect occurs because of the image sensor's fixed resolution. Therefore, there is a critical need to develop a method that can enhance both the spatial and angular resolution, allowing for improved measurement resolution and accuracy of the metrology system.

2.3.4 Deep learning for the super-resolution of light field data

Deep learning technologies recently have been developing rapidly and have drawn great attention in both academic and industrial fields. On the basis of the development of computer science and Internet of Things (IoT) technologies, a vast number of data are stored and can be processed quickly. The redundant data with enormous volumes of information make it possible to explore more complex relationships of different data in a much higher dimension. As a result, deep learning with powerful representation capability becomes a popular solution to mine the relationship of data, with less requirement of expert knowledge.

Deep learning was proposed in the 1960s (Ivakhnenko et al., 1967; Rosenblatt, 1961) but was limited by the capabilities of hardware. During the decades, deep learning has gone through several valleys, as well as reaching some peaks with the emergence of new techniques, and now is nearing maturity. Deep learning has gained immense traction with the advent of powerful devices like specialized graphic processing units (GPUs) and tensor processing units (TPUs), which possess high computing capabilities. These advancements have enabled deep learning to showcase its strength in utilizing much deeper neural networks. This ability enables the training of highly complex transformation functions to achieve superior performance in various domains. Living up to expectations, learning-based models have outperformed many conventional methods in a large number of fields, including vision and speech recognition (Amodei et al., 2016; Dong et al., 2015; He et al., 2016; Richardson et al., 2015), forecasting (Gensler et al., 2016; Shi et al., 2017), and control (Silver et al., 2017), even creating works that are similar to masterpieces (Briot et al., 2017).

To break through the inherit resolution limitation of light field data, researchers have investigated various methods to enhance the resolution of LF data. Conventional methods for super-resolution in light field data fall into two categories: projection techniques and optimization approaches. Projection-based methods employ the subpixel information inherent in the LF data to achieve resolution enhancement. Conversely, optimization-based approaches employ optimization models to produce HR images within the super-resolved domain.

In terms of projection-based super-resolution methods (Lim et al., 2009), the basic consideration behind the methods is the redundant radiance information recorded by light field systems. Different from traditional cameras which only record the position and wavelength of emitted rays, light field cameras are able to capture the directions of the rays. This is achieved by micro-lenses and their relative positions. Consequently, each point of the target object is recorded by different micro-lenses many times, with subpixel shifting existing among the corresponding pixel points in multi-perspective images. The subpixel shifting enables the inclusion of additional high-resolution information, allowing for the reconstruction of high-resolution images through accurate registration and merging of the raw data.

Other research on projection-based super-resolution methods for plenoptic cameras was conducted by Georgiev and Lumsdaine (2012), with the principle and results as shown in Figure 2.25(a) and Figure 2.25(b), respectively. As shown in Figure 2.25(a), through projecting the pixels of different microimages images onto a new plane at an appropriate angle, it is obvious that the pixel points in different microimages contribute extra resolution information to their neighbouring points. As a result, a high-resolution image is obtained by combining these relative points together. The combination is realized by convolutional operation using fixed kernels. A comparison

between this method and conventional demosaicing methods (Gotoh & Okutomi, 2004; Vandewalle et al., 2007) is shown in Figure 2.25(b). Conventional demosaicing methods have been developed for scenes involving SISR. The comparison revealed that the projection-based method not only enhances details significantly but also minimizes the occurrence of artefacts.

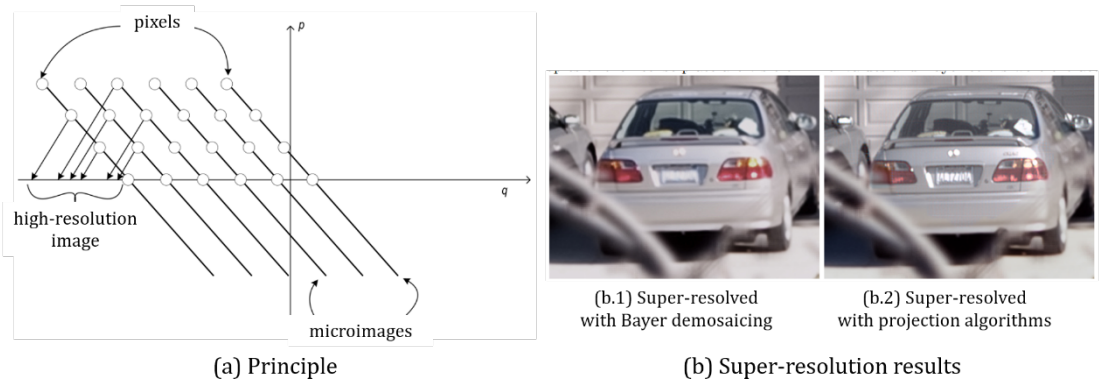

Figure 2.25 Projection-based super-resolution method (Georgiev & Lumsdaine, 2012). (a) simply illustrates the super-resolution principle by projecting. (b) compares the super-resolved images by the projection-based method and a traditional Bayer demosaicing method.

Liang and Ramamoorthi (2015) simulated the LF capturing and rendering process using a light transport framework. As depicted in Figure 2.26 (c), the illustrated process closely resembles the approach presented in Georgiev and Lumsdaine (2012), which shows that the resolved resolution is limited by the projection plane since the distribution density in the projection plane is higher than the micro-lens images. Through applying a prefiltering operation to the light field radiance, the limited resolution can be tackled, as shown in Figure 2.26(d). Similar work was done by Yu et al. (2012), where a frequency-domain filter was utilized to resample the 4D colour-

filtered radiance.

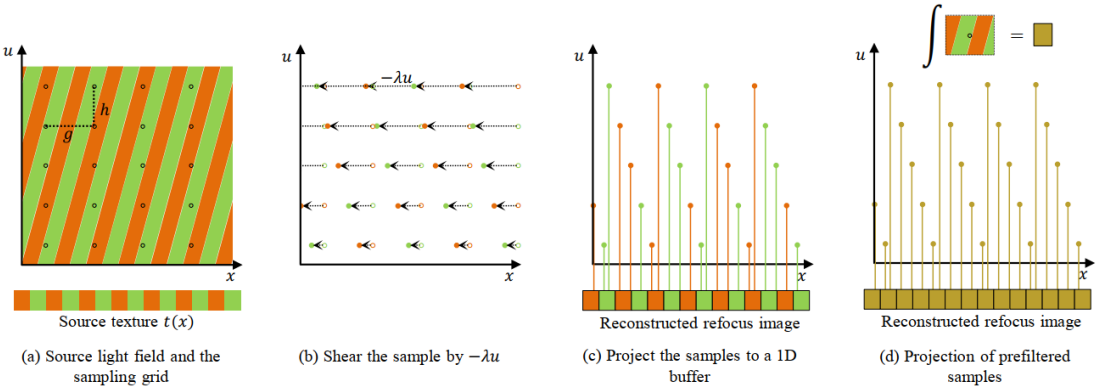

Figure 2.26 Illustration of the projection-based SR process (Liang & Ramamoorthi, 2015).

Optimization-based super-resolution methods for light field data also make use of the idea of subpixel information. Unlike directly projecting pixel points from different micro-lens images onto high-resolution image planes, the optimization-based method adopts a different approach. In this method, the initial step involves estimating the depth map of the scene using the light field representation. Subsequently, high-resolution images are obtained using a variational Bayesian framework. The author presented a point-spread function based on the Gaussian optics assumption for plenoptic cameras. The process can be formulated as

$$I = Hr + \omega \tag{2.4}$$

where I is the image captured by the plenoptic camera, I' is the unknown reflectance, i.e., the light field desired to be reconstructed, H represents the point-spread function of the plenoptic camera, and ω is Gaussian noise. The method uses conjugate gradient least squares for the optimization objective, trying to estimate the reflectance I' based on the observed I and an estimation of the depth map which indicates the point-spread function H. One super-resolved result is shown in Figure 2.27. On the left is the original

sub-aperture image, obtained by rearranging the pixels from the micro-lens images. On the right is the HR image generated through the SR method.

Figure 2.27 Comparison between traditional rendering methods and an optimization-based super-resolution method (Bishop et al., 2009). The two images are both the centre view extracted from a light field. The left was generated by traditional rendering methods and the right was generated by the super-resolution method.

Similar to Bishop et al. (2009), Mitra and Veeraraghavan (2012) also used Eq. (2.4) to establish a SR model to enhance the LF resolution. Compared with 2D images with only two dimensions x and y, light field data have two more dimensions u and v. As a result, the epipolar-plane images in $X-U$ and $Y-V$ planes can be obtained. The disparity information lies in these epipolar-plane images. To make use of the low dimension of epipolar-plane images, a Gaussian mixture model (GMM) was presented by Mitra and Veeraraghavan (2012). Gaussian patch priors were learned to determine the disparity so that the disparity maps were estimated first. The enhancement of the epipolar-plane images was performed by an optimization process using a linear minimum mean square estimator based on the previously estimated disparity values.

The high-resolution images were reconstructed using the enhanced epipolar-plane images that reach higher quality than images interpolated by bicubic interpolation as shown in Figure 2.28.

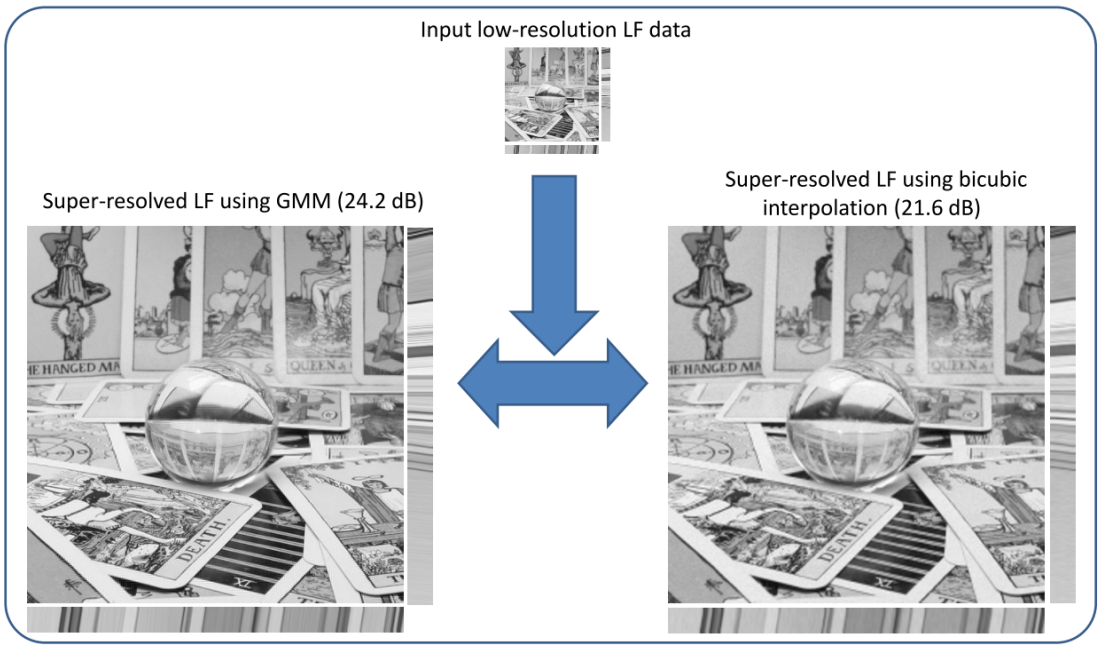

Figure 2.28 Spatial-resolution-enhanced light field data recorded by plenoptic cameras based on the method proposed in Mitra and Veeraraghavan (2012) .

So far, the mentioned research has mainly focused on improving the spatial resolution of LF images. Wanner and Goldluecke (2012) proposed a super-resolution method that focuses on enhancing both spatial and angular resolution, thus improving the overall angular resolution of the images. Similarly, the method realized the angular super-resolution through synthesizing novel view images based on depth estimation. Using the original light field data, the estimation of point positions in 3D space was conducted using the available disparity information. The novel views were then synthesized via remapping the points into the virtual image plane based on the geometrical relationship. The reconstruction process is depicted in Figure 2.29, where

τ_i is a mapping function between the image plane Ω_i and a new perspective plane Γ, the scene surface Σ is inferred based on depth estimation, and a mask is used to block out points that are invisible on plane Γ. After modelling the mapping function, novel views were synthesized by solving the inverse process using the fast iterative shrinkage and thresholding algorithm.

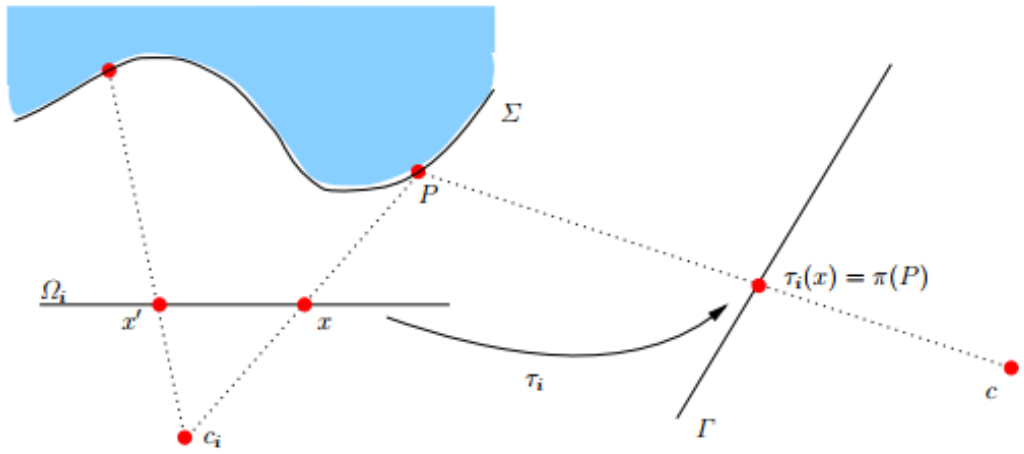

Figure 2.29 Reconstruction of novel view images based on depth estimation (Wanner & Goldluecke, 2012).

Regarding super-resolution, deep learning has made progress both in single 2D images and light field data. Dong et al. (2014) used a very simple CNN to perform SISR and acquired better results compared with conventional methods. The pioneering work inspired the development of learning-based models in super-resolution areas. Sequentially, learning-based models have also been developed to super-resolve LF images and produce satisfactory results.

A comprehensive comparison among different methods for LF super-resolution was conducted by Cheng et al. (2019), with the results shown in Figure 2.30 using the PSNR as a metric. Two datasets were used, where the HCI dataset (Honauer et al., 2016)

consists of synthetic light field images and the EPFL dataset (Rerabek & Ebrahimi, 2016) is composed of realistic LF images. In the research, only the quality of spatial SR was discussed, and BIC represents bicubic interpolation. PRO (Liang & Ramamoorthi, 2015), GB (Rossi & Frossard, 2017), and RR (Farrugia et al., 2017) are based on conventional theories. LFCNN (Yoon et al., 2017) is a fully learning-based model that utilizes convolutional neural networks. VDSR (Kim et al., 2016) is based on deep learning and specifically designed for SISR. It is obvious that for the synthetic data (the HCI dataset), although conventional methods outperformed the learning-based model in some situations, the learning-based model always produced a better result in the real-world dataset (the EPFL dataset). An additional discovery reveals that the learning-based SISR method consistently generates high-quality high-resolution images for both synthetic and realistic images. This underscores the significant potential of learning-based models for light field resolution enhancement.

The super-resolution CNN (Dong et al., 2014) laid the foundation as pioneering work that employed a learning-based model for reconstructing HR images. Figure 2.31 displays the framework of the method, which is a straightforward yet efficient architecture. The authors claimed that the process of patch extraction and representation in conventional super-resolution methods in fact is equivalent to convolutional operation.

Based on the statement, the network composed of three convolutional layers was proposed. The three layers contained 64, 32, and 1 convolutional kernel separately and the corresponding kernel sizes are 9×9, 1×1, and 5×5. The activation function chosen for each layer was the ReLU. In the image processing pipeline, the input images underwent a conversion from RGB to YCrCb colour space. Subsequently, only the luminance channel (Y channel) of the images was subjected to super-resolution by the

method. The frames of the other two channels were enhanced using traditional interpolation methods. The authors further improved the method of Dong et al. (2015) that was able to super-resolve all the frames of three channels, and high-resolution images were generated via merging the high-resolution frames.

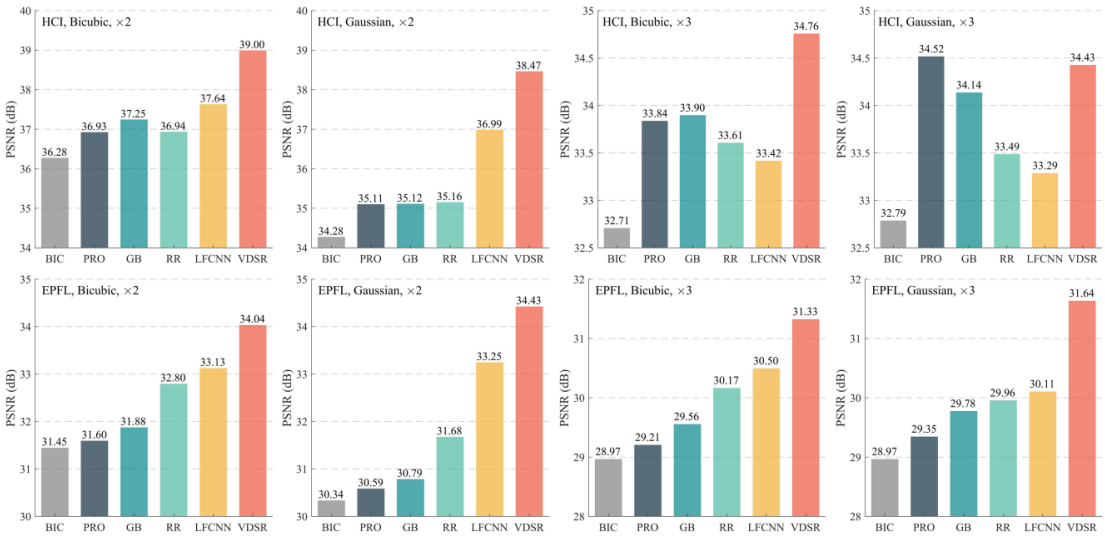

Figure 2.30 Evaluation of various SR techniques for LF (Cheng et al., 2019).

In contrast to conventional approaches, the initial layer functioned as a feature extractor, carrying out patch representation. To improve the representational capacity of the learning model, a non-linear mapping operation was conducted using the second layer. The high-resolution images were reconstructed by the third layer using the features generated by the second layer. The whole process was similar to the traditional super-resolution methods. However, the feature extraction and the reconstruction strategies were automatically learned by the learning model with no a priori knowledge.

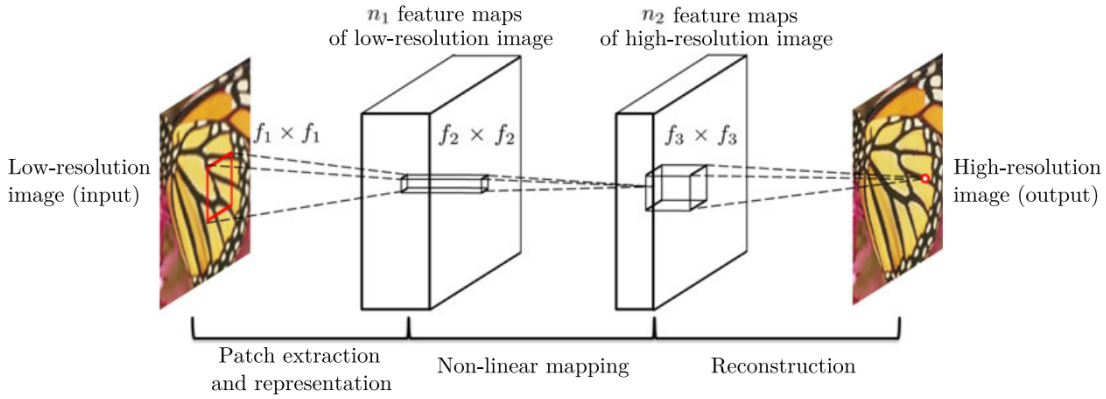

Figure 2.31 Framework of the resolution-enhanced learning model in Dong et al. (2014).

Another representative learning-based model named Very Deep Super-Resolution (VDSR) was proposed by Kim et al. (2016). VDSR made use of a residual architecture to reduce the information required to be learned so that the learning speed was improved. Figure 2.32 illustrates that the main distinction from the model in Dong et al. (2014) is the inclusion of a pixel-wise summation operator before the final output. The researchers held the belief that there exists similarity in low-frequency information between low- and high-resolution images. Consequently, the learning model can solely concentrate on capturing the dissimilarity in high-frequency information, which is known as the residual information.

The model was able to be trained at a faster speed with even better super-resolution performance. VDSR used a very deep architecture, totally consisting of 20 layers. The convolutional kernels in each layer were of a small size, only 3×3. Although the size was smaller, the receptive field was quite large due to the deep architecture. The authors also found that learning models with deep architectures always perform better.

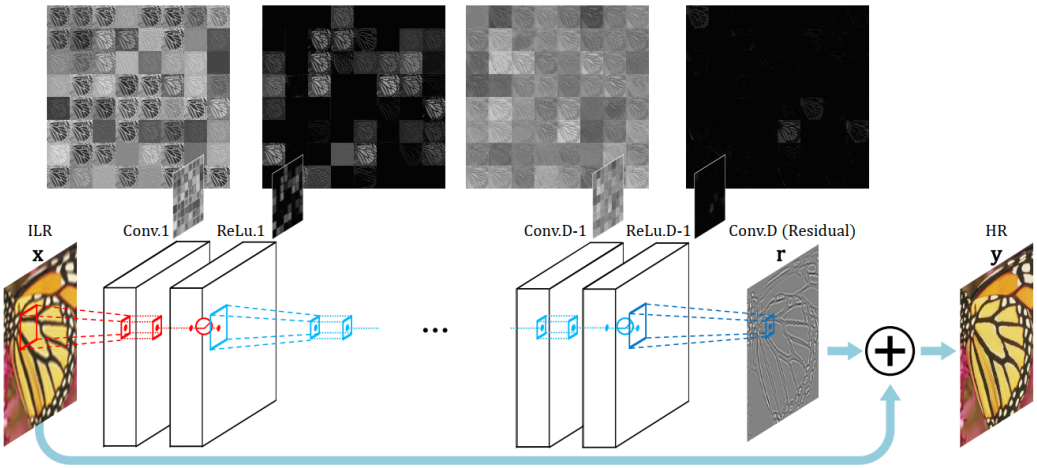

Figure 2.32 Framework of a residual model in Kim et al. (2016).

GAN proposed by I. Goodfellow et al. (2014) is an unsupervised learning method composed of two networks that engage in a zero-sum game, competing against each other. In a GAN model, a generative network is used to generate synthetic data from input such as noise data, low-resolution data, etc. A discriminative network is used to obtain the generator's output and judges whether the output of the generation is from a realistic dataset or a synthetic dataset. After continuous contests, the discriminator becomes cleverer at distinguishing a synthetic input from a real input, but meanwhile the generator is also able to output data with higher quality to fool the discriminator. It is obvious that the GAN models are very suitable for performing super-resolution. A GAN-based SR learning model named SRGAN (Ledig et al., 2017) was consequently developed, and its framework is shown in Figure 2.33. The generator used in SRGAN made use of many deep learning techniques including residual blocks, pixel shuffle, and batch normalization to enhance its representation ability, while the discriminator was a deep classification network to perform the distinguishment. The distinguishment results were only fed back to the generator while training, indicating the gap between the fake images and the real one.

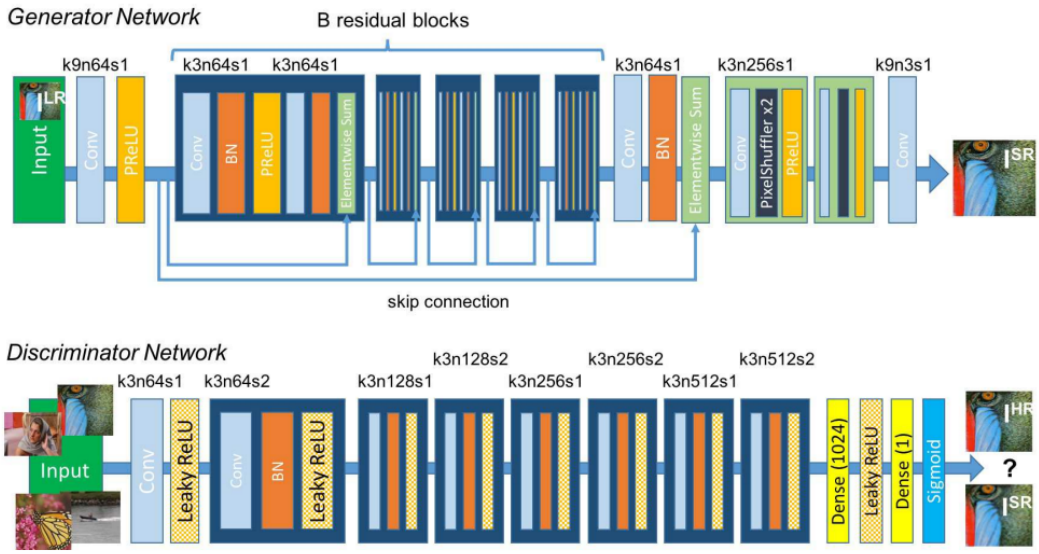

Figure 2.33 Framework of a GAN-based learning model proposed in Ledig et al. (2017).

Figure 2.34 illustrates the super-resolved outcomes of SRGAN, where the generator SRResNet is trained without receiving feedback from the discriminator. An interesting finding in the results is that SRGAN can produce high-resolution images with clearer edges and more elaborate details than the generator without the discriminator, but the artefacts are also more obvious (e.g. the necklaces in the third image compared to the original image have different patterns).

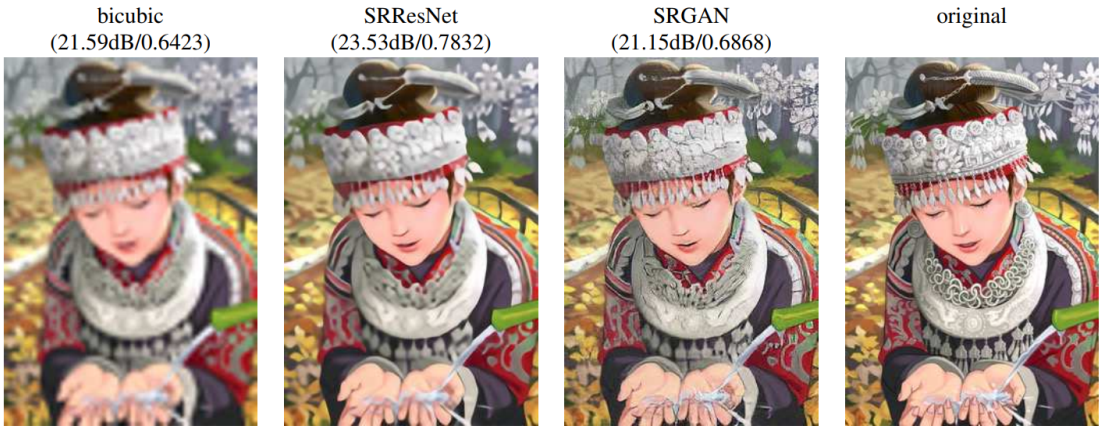

Figure 2.34 High-resolution images generated by bicubic

interpolation, SRResNet, and SRGAN with 4-time upscaling (Ledig et al., 2017).

Regarding LF super-resolution, the methods can be categorized into two types: spatial SR and angular SR. Spatial SR techniques are designed to enhance the resolution of each elemental or sub-aperture image through the interpolation of pixels that lie between adjacent pixels in the image. The methods for angular super-resolution are able to synthesize novel view images between adjacent elemental images or sub-aperture images. Since the experiment performed in Cheng et al. (2019) has shown that VDSR outperforms other conventional methods for spatial SR of LF data, learning-based SISR methods can serve as a more effective alternative for achieving spatial super-resolution in LF images.

In the context of angular SR, learning-based approaches include depth-based and non-depth-based models. Similar to the traditional angular SR methods, the learning models integrated with depth cues usually synthesize novel views with a higher quality especially when the disparities of LF data are large. However, accurate depth estimation is difficult so that image artefacts are always generated by the depth-based models, especially for real-world LF data.

The Light Field Convolutional Neural Network (LFCNN) was the pioneering deep convolutional neural network (Yoon et al., 2015, 2017). This network was specifically designed to enhance the resolution of light field data obtained from commercial plenoptic cameras. The framework is shown in Figure 2.35, where two stages exist in the super-resolving process. In the initial stage, a spatial SR net is employed to enhance the spatial resolution. The spatial network is totally the same as the network proposed in Dong et al. (2015), containing three convolutional layers.

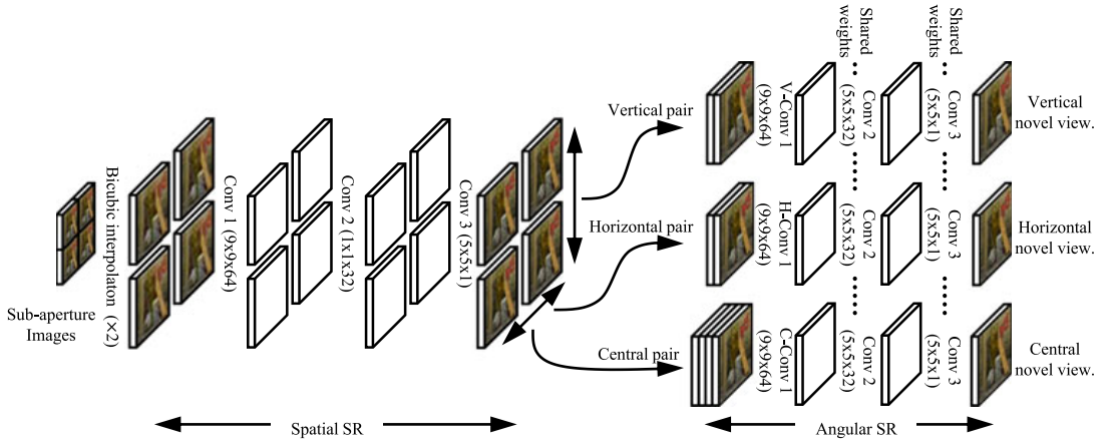

Figure 2.35 Framework of the pioneer learning model for the resolution improvement of LF data (Yoon et al., 2017).

After obtaining the high-resolution images, at the second stage, three angular super-resolution networks are used to synthesize new view images. The authors grouped the input elemental images into vertical, horizontal, and central pairs and the corresponding angular network generated a middle or a central new view according to the geometrical relationship of the image pairs. Every angular super-resolution network had three convolutional layers and the last two layers shared their weights. This was depicted as the first layer of each angular super-resolution network being used to extract features from different images recorded from perspectives and the last two layers being used to recover the high-resolution information. The authors compared LFCNN with the conventional method (Mitra & Veeraraghavan, 2012) using a real-world light field dataset, and the outcomes are displayed in Figure 2.36. The research demonstrates that even simple deep learning models can effectively achieve high-quality resolution enhancement.

To obtain the ground truth, the researchers down-sampled the angular resolution of the original dataset to train the model. This artificial sampling method became a

popular learning paradigm for the training of CNN-based resolution enhancement methods. Meng et al. (2020) integrated generative adversarial networks into the learning models for the LF super-resolution, and the framework is shown in Figure 2.37. High-dimensional convolution was utilized in the model to realize a 4D convolution so as to make full use of the spatial–angular redundance lying in the light field data. A pre-trained network was also integrated to provide perceptual loss. Through confronting the distinguishing of a discriminator, the super-resolution networks could achieve high-quality reconstruction.

Instead of directly processing the information in the spatial space, the models proposed in Wu et al. (2017, 2019) were trained to learn to reconstruct HR EPIs from the LR data. The authors (Wu et al., 2019) presented a sheared EPI concept that converts the depth estimation problem into an evaluation problem of candidates for the EPI shearing.

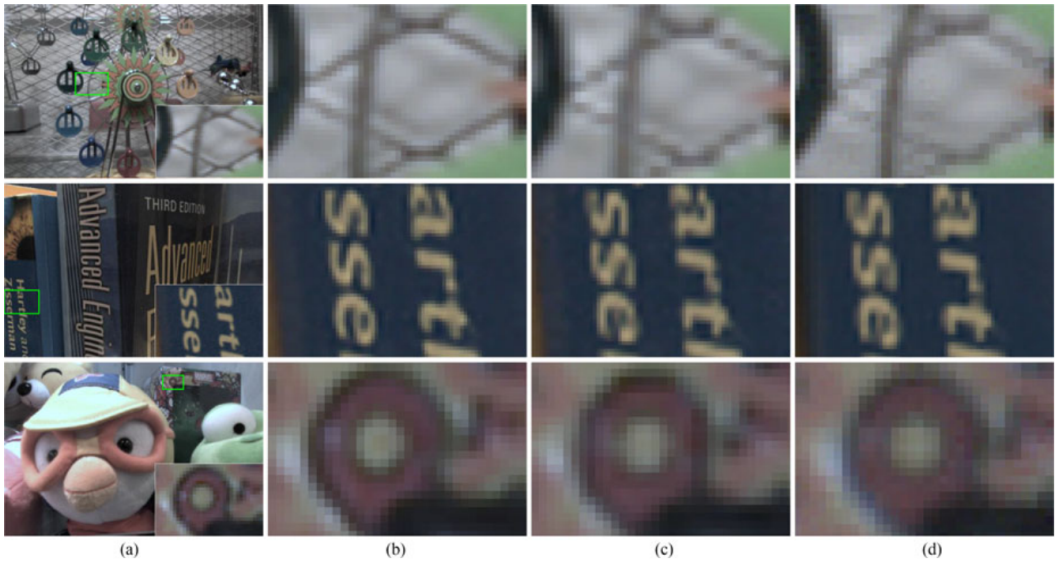

Figure 2.36 Evaluation of SR results obtained using conventional methods and a learning-based method. (a) Ground truth. (b) The learning-based method proposed in Yoon et al.

(2017). (c) Bicubic interpolation. (d) The conventional method proposed in Mitra and Veeraraghavan (2012).

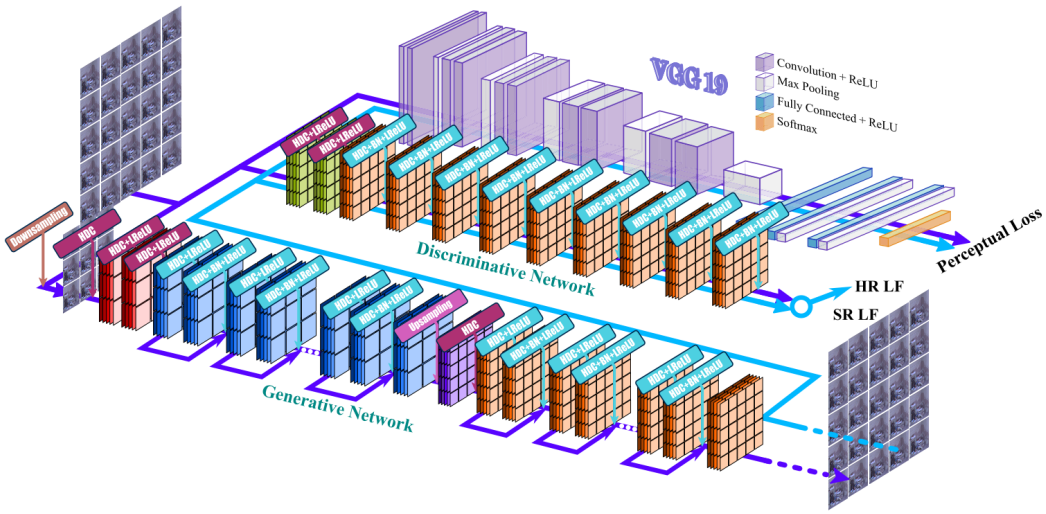

Figure 2.37 A GAN-based light field model (Meng et al., 2020).

The super-resolution process in Wu et al. (2019) is shown in Figure 2.38. In this study, a CNN was developed to quantify the resemblance between input sheared EPIs and the reference EPIs. Subsequently, this CNN generated evaluation scores for each pixel, thereby aiding in the super-resolution process. The volumes of score maps were used for the fusion tensor calculation. The final EPI reconstruction was performed using a pyramid decomposition–reconstruction technique. The method used a depth-free framework so that it was competent in handling the datasets without the ground truth of depth. The authors also showed its capability in relation to synthetic, real-world, and microscopy light field data.

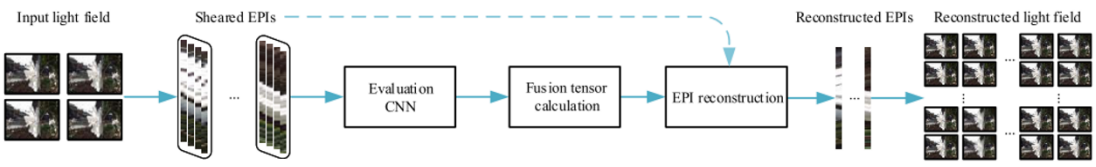

Figure 2.38 Angular super-resolution based on sheared EPIs (Wu et al., 2019).

Depth-based learning models do not regress target pixel values directly but use estimated depth maps as an intermediate transformation from the input low-angular-resolution LF data into HR data. Taking advantage of the CNN models, more accurate depth estimation can be achieved. Yeung et al. (2018) trained a CNN model utilizing spatial–angular separable filters that are a kind of 4D filter to process the 4D LF inputs. The network framework of Yeung et al. (2018) is shown in Figure 2.39, where two stages are required for the super-resolution. The first stage was performed by a view synthesis network that made use of pseudo 4D filters to extract spatial–angular clues existing in the input sparse light field data. A view refinement network was used to further refine the coarse intermediate synthesis views generated by the first stage. Although the method did not conduct depth estimation straightforwardly, the depth cues lying in the input light field data were implicitly exploited by the 4D convolution. The work generated good performance for real-world images with small disparity ranges.

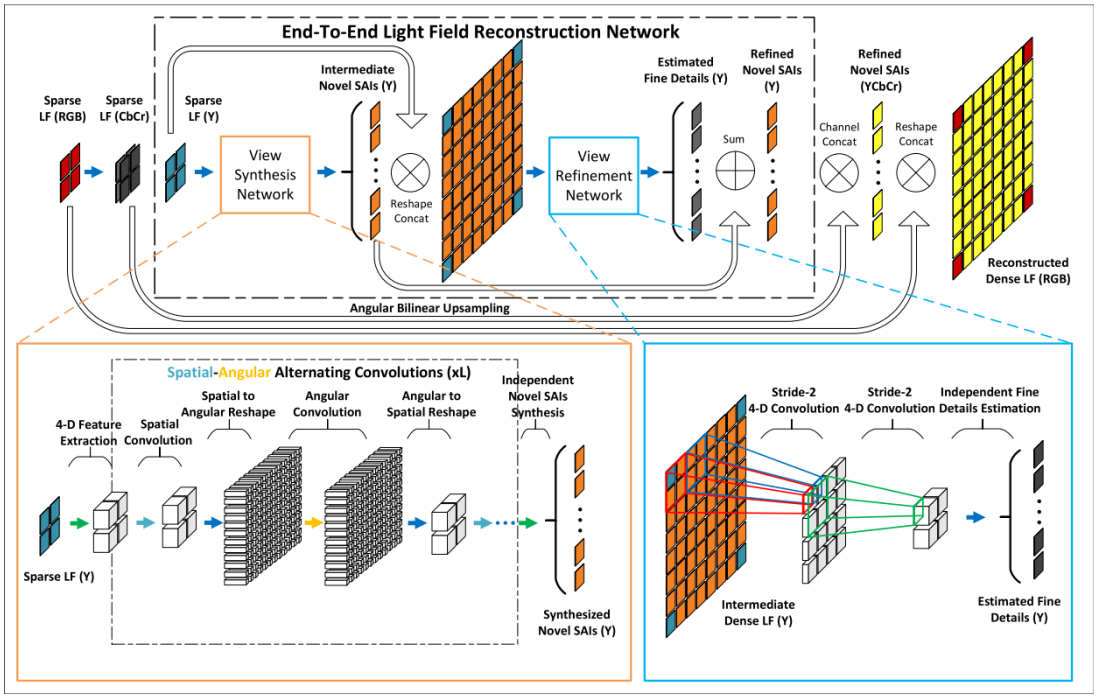

Figure 2.39 A learning-based angular SR network in Yeung et al. (2018).

Jin, Hou, Yuan, et al. (2020) used a more complicated and deeper network to realize the estimation and reconstruction processes, and the network framework is shown in Figure 2.40. A depth estimation module was used to perform direct depth estimation based on the input sparse light field data. After obtaining the depth maps, the input views were warped to new perspectives separately so that the novel views based on every input view were acquired. The final high-angular-resolution data were reconstructed by a light field blending module based on the warped light field data generated in the previous stage. A skip connection was used to form a residual learning relationship. This method achieved satisfactory reconstruction results, especially for the light field data with a large baseline. This was due to the accurate depth estimation performed by the estimation module. Nonetheless, the depth of real-world data is more difficult to predict. With incorrect estimation, the depth-based methods usually produce severe artefacts in the reconstruction results.

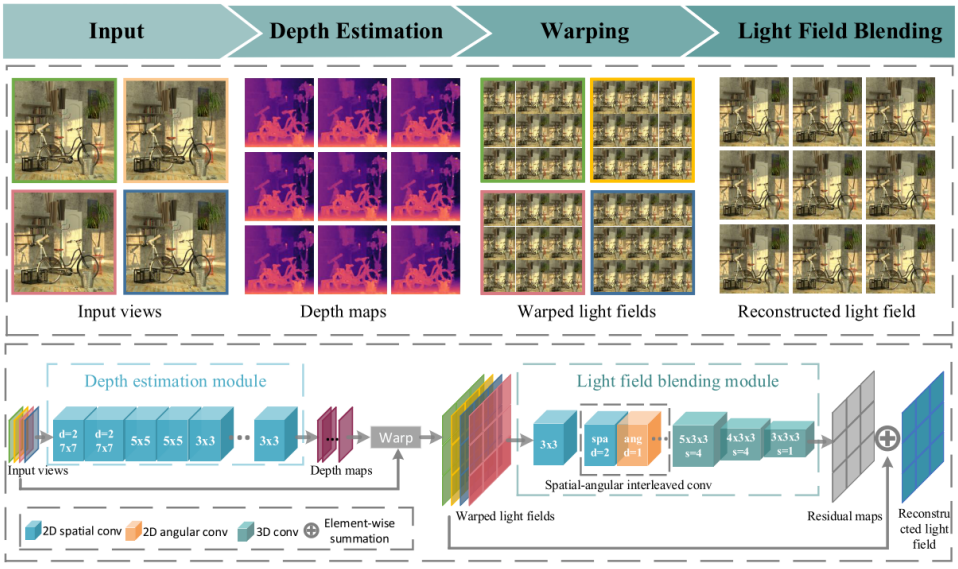

Figure 2.40 A learning-based angular SR network for large-baseline light field data (Jin, Hou, Yuan, et al., 2020).

Although these learning-based models have different architectures, the training

paradigm of them is very similar. Training is a significant process for learning-based methods, where the learning models adjust their parameters iteratively based on some evaluation metrics (i.e., loss functions) until the models converge to global minima. Abundant data are used during training, and ground truth (also called label data) is usually used as a reference in supervised learning. The training objective is to minimize the output difference from ground truth. Backpropagation is commonly employed in the training phase of learning models to address the optimization problem. In terms of unsupervised learning, learning models discover the underlying patterns and internal representations via self-organizing without expert experience, a priori knowledge, and manually labelling.

Regarding the training process of SISR models, MSE also referred to as L2 loss, and MAE also known as L1 loss, are commonly utilized as evaluation metrics in machine learning. A study (Zhao et al., 2016) compared the performance using different loss functions and found that L1 loss usually produces better results in image restoration. Part of the comparison outcomes are presented in Table 2.1, where ℓ_1 denotes the L1 loss and ℓ_2 denotes the L2 loss. The super-resolution results generated by the bilinear method, the learning model trained using L1 loss, and the same learning model trained using L2 loss are compared with the ground truth. It is observed that the model trained using L1 loss (MAE) tends to outperform the model trained with L2 loss (MSE) across all evaluation metrics. The researchers performed additional experiments and found that the model trained only using L2 loss usually converges to a local minimum. The model trained with L1 loss consistently achieves a superior minimum. Another interesting finding was that training a model with L2 loss initially and then refining it with L1 loss subsequently results in reaching an even better minimum. However, splotchy artefacts still exist in the reconstruction results of the model. The artefacts can

be avoided by training the model using L1 loss.

Perceptual loss is also usually utilized in image reconstruction. The perceptual loss quantifies the feature distance, where features are obtained using a pre-trained deep learning model. The pre-trained VGG (Simonyan & Zisserman, 2014) model is usually taken as the feature extractor, since the model was trained in ImageNet (Deng et al., 2009) which is a huge image dataset consisting of a vast number of natural images.

Table 2.1 Average scores of multiple image quality indicators under different loss function scenarios. Lower is better for ℓ_1 and ℓ_2, and higher is better for PSNR and SSIM. (Extracted from Zhao et al. (2016)).

Super-resolution	**Training loss function**		
Image quality metric	Bilinear	ℓ_2	ℓ_1
$1000 \cdot \ell_2$	2.5697	1.2407	1.1062
$1000 \cdot \ell_1$	28.7764	20.4730	19.0643
PSNR	27.16	30.66	31.26
SSIM	0.8632	0.9274	0.9322

Through integrating perceptual loss into the multi-loss function of model training, the reconstruction results usually contain more realistic textures. A study (Wu et al., 2020) compared the super-resolution results reconstructed by models trained only using L1 loss, low-level perceptual loss, high-level perceptual loss, and perceptual similarity loss. The results are shown in Figure 2.41, where L_{lp}, L_{hp}, and L_{ps} denote the low-level perceptual loss, high-level perceptual loss, and perceptual similarity loss, respectively. From a visual inspection perspective, it is evident that the model trained solely with perceptual loss can generate more realistic and intricate details.

2.4 Summary

With the increasing demand for micro-structured surfaces with high precision, on-machine measurement is playing an important role in advanced manufacturing. Albeit that offline measurement methods can produce high-accuracy measurement results, errors caused by remounting the measured parts are unavoidably introduced so that the precision of the re-machining process cannot be guaranteed.

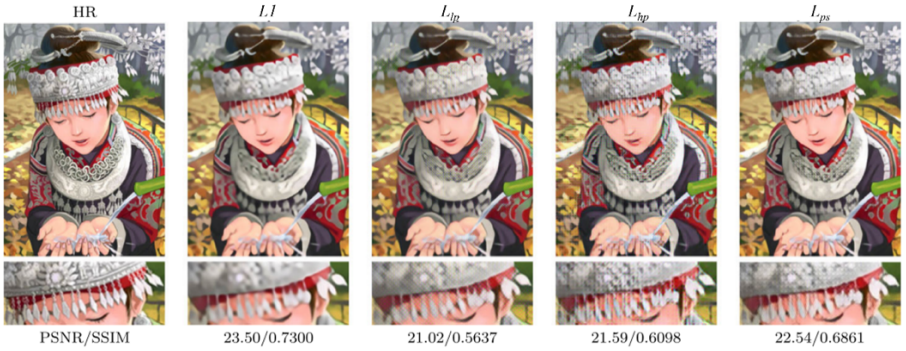

Figure 2.41 Evaluation of super-resolution results reconstructed by the learning models trained with only L1 loss, low-level perceptual loss, high-level perceptual loss, and perceptual similarity loss (Wu et al., 2020).

To acquire precision on-machine measurement results, both contact and contactless measurement methods have been developed over the decades. Contact measurement that uses a finer probe or stylus to make contact with the measured surfaces is easier to be implemented for accurate measurements, but the nature of the interaction with the surface may cause damage to the surfaces. Non-contact measurement methods usually incorporate an optical probe to perform inspection through analyzing the signals. Many techniques including interferometry, deflectometry, structured light, confocal, etc. have been integrated into an ultra-precision machine tool for on-machine measurement.

Among the optical techniques, autostereoscopy as a promising technique that can capture 3D information within one snapshot is able to perform on-machine measurements for micro-structured surfaces. The autostereoscopic 3D measurement system makes use of the InI technique to record spatial–angular information from which the 3D information of the measured surfaces can be extracted, and the profile can be reconstructed. However, an inherent trade-off of the autostereoscopy technique is the resolution of the recorded data. Given the inherent nature of the technique, it is difficult to simultaneously enhance the two types of LF resolution.

To this end, a vast number of studies have been reported to enhance the resolution of light field data based on conventional image processing techniques or artificial intelligence techniques. Deep learning, a highly powerful representation model for image processing shed light on the SR problems of LF data. However, the enhancement results generated by current deep learning methods usually contain severe image ghosting and artefacts that could induce additional errors during the depth estimation process.

In summary, the gaps existing in this research can be concluded as:

(i) The performance of the autostereoscopic 3D measuring system is limited by the inherent trade-off that arises from the principle of InI. A larger angular resolution provides more disparity information to the raw measurement data so that the matching accuracy can be improved during the depth estimation process. The spatial resolution is important to the details of the recorded 3D scene. With finer details in the elemental images, the resolution of the depth slices during the digital refocusing process will be improved. As a result, it is important to enhance LF resolution so as to improve the measurement accuracy of the autostereoscopic 3D measuring system.

(ii) The current deep learning methods for angular super-resolution can be categorized into non-depth-based and depth-based approaches. Methods without initial depth estimation often result in image ghosting in the novel views, while depth-based methods require accurate depth estimation prior to novel view reconstruction. Consequently, depth-based methods often produce image artefacts in the reconstruction results as a result of inaccurate estimation.

In addition, current deep learning methods for angular resolution require splitting the finite training data into input and the corresponding ground truth. This requires a sampling-inefficient learning paradigm and enormous data of various 3D scenes to be necessary for model training. In terms of the real-world light field data with a large baseline, most of the current models cannot produce high-quality reconstruction results since the real-world data usually contain severe noises and complex illumination conditions. Hence, it is necessary to develop a novel learning paradigm to improve the training efficiency for angular super-resolution and develop a learning-based approach to achieve high-quality super-resolution for real-world data with large baselines.

(iii) In terms of on-machine measurement, the vibration of the machine tools cannot be avoided and could introduce more measurement errors to the measuring system. To make full use of the vibration, multiple frames captured over various timespans can be used to eliminate the effects resulting from the vibration. In addition, the pixel-level information among the multiple frames will provide redundant information to reconstruct high-resolution patterns. By employing this approach, the spatial resolution of the data obtained from the autostereoscopic 3D measuring system can be enhanced. Hence, the development of a resolution enhancement method based on the multiple frames captured in an on-machine process will

benefit the inspection performance of the autostereoscopic system.

Chapter 3 Autostereoscopic three-dimensional measurement system

3.1 Introduction

An autostereoscopic measuring system is generally composed of a high-magnification zoom lens system, an objective lens, a micro-lens array, and an image sensor. Illumination devices are usually necessary for micro-structure measurement so that the system can receive the light rays from the target surfaces. A general system setup is shown in Figure 3.1, where a ring-type and a co-axial illumination device is utilized, and the target sample is mounted on a 3-axis displacement platform. By leveraging the InI principle, the MLA inserted in front of the image sensor captures the image occurring behind the zoom lenses from various perspectives. The separation between the object point and the objective lens determines the difference in pixel coordinates of the corresponding points captured by different micro-lenses. By analyzing the pixel difference, also known as disparity, among a group of image points originating from the same object point, it becomes possible to determine the separation from the object point to the object lens. This allows for the acquisition of axial information about the target surface. Since the single-image sensor is split into multiple regions to record the observations of every micro-lens, the resolution of the lateral information is decreased. This is an inevitable trade-off of the autostereoscopic 3D measuring system, limiting the resolution of the measurement results.

In this chapter, the measurement principle of the autostereoscopic system is discussed, and various depth recognition methods are presented. An experiment of rapid inspection of surface-mounted light-emitting diodes using the autostereoscopic 3D

inspection system was performed to evaluate the measurement performance. The main limitation of the current autostereoscopic measuring system is discussed.

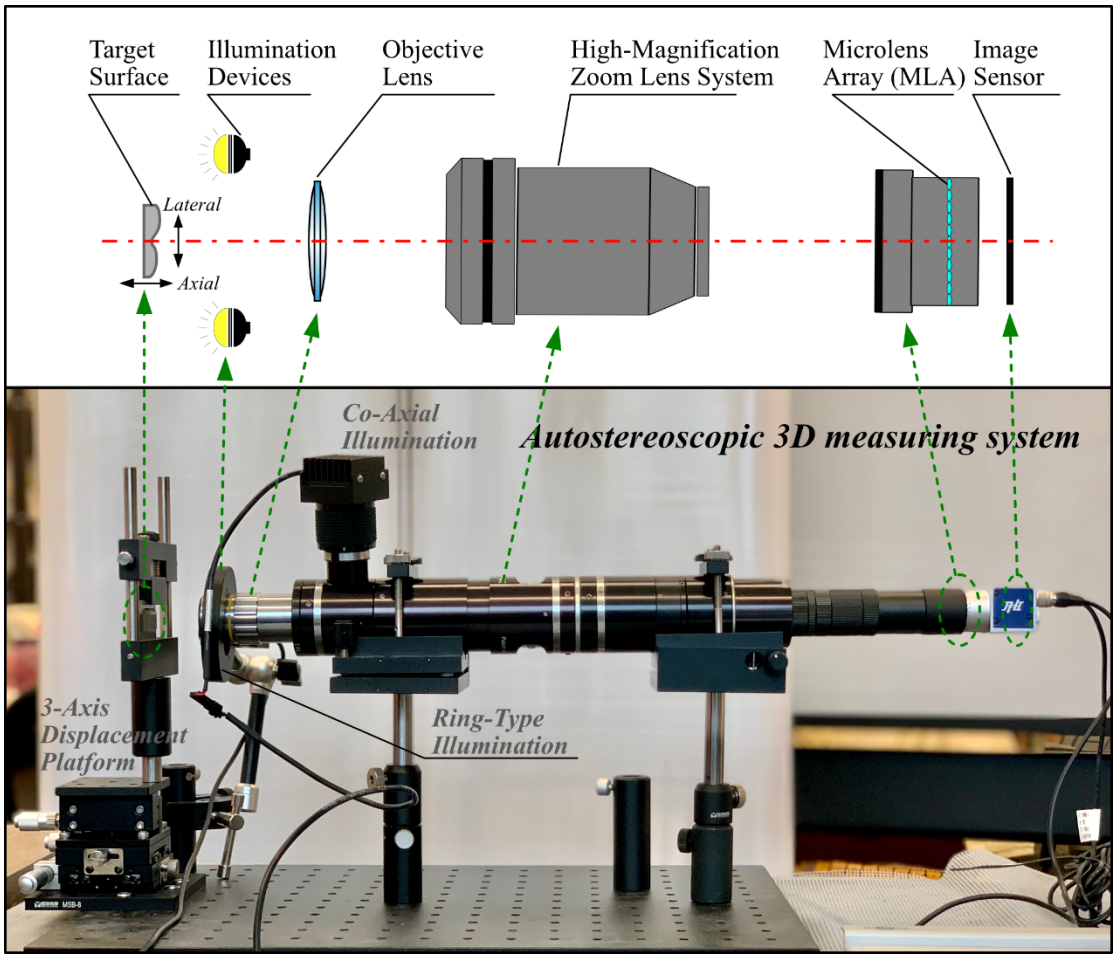

Figure 3.1 General system setup of an autostereoscopic 3D measuring system.

3.2 Autostereoscopic measurement principles

A 4D light field can be represented by parameterization (Levoy et al., 2006) as shown in Figure 3.2, where two planes are placed in a free 3D space so that rays that interact with the two planes are recorded with their directions available as well. The 4D light field function (also called plenoptic function) is represented as $L(u,v,x,y)$. By leveraging the spatial–angular cues, the 3D data of the scene can be extracted and

reconstructed from the 4D light field information. This is the fundamental approach of the autostereoscopy-based system to realize 3D inspection and measurement.

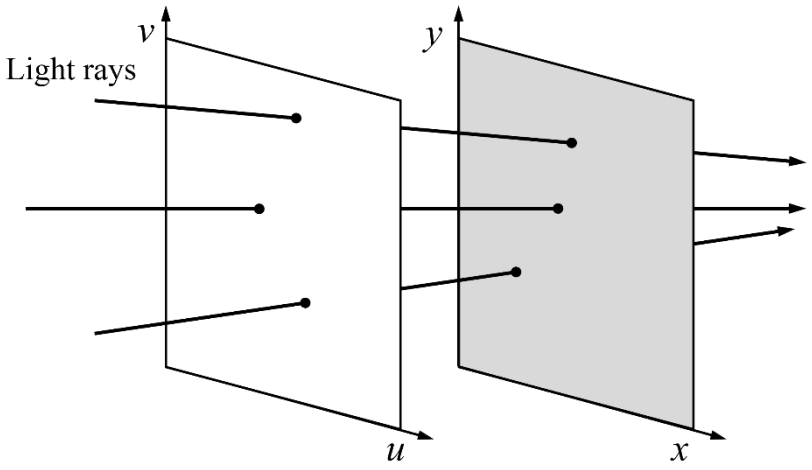

Figure 3.2 4D light field.

The measuring principle of the autostereoscopic 3D measuring system is shown in Figure 3.3, where the micro-lens array is regarded as a pinhole-like array of lenses for simplification. Under the assumption, light rays can only pass through the centre of the lenses. Three points A, B, and C are demonstrated in the diagram where point A and point C are at the same depth whereas point B is closer to the micro-lens array. It is obvious that multiple images (i.e., elemental images) are recorded by the image sensor behind every micro-lens from various perspectives. The number of recording perspectives is determined by the array size of the MLA. The separation between the MLA and the image sensor is g and the physical baseline distance of two adjacent micro-lenses is p. Obviously, the centre distance of two arbitrary micro-lenses in one row or column is $n \cdot p$ where $n \in N_+$.

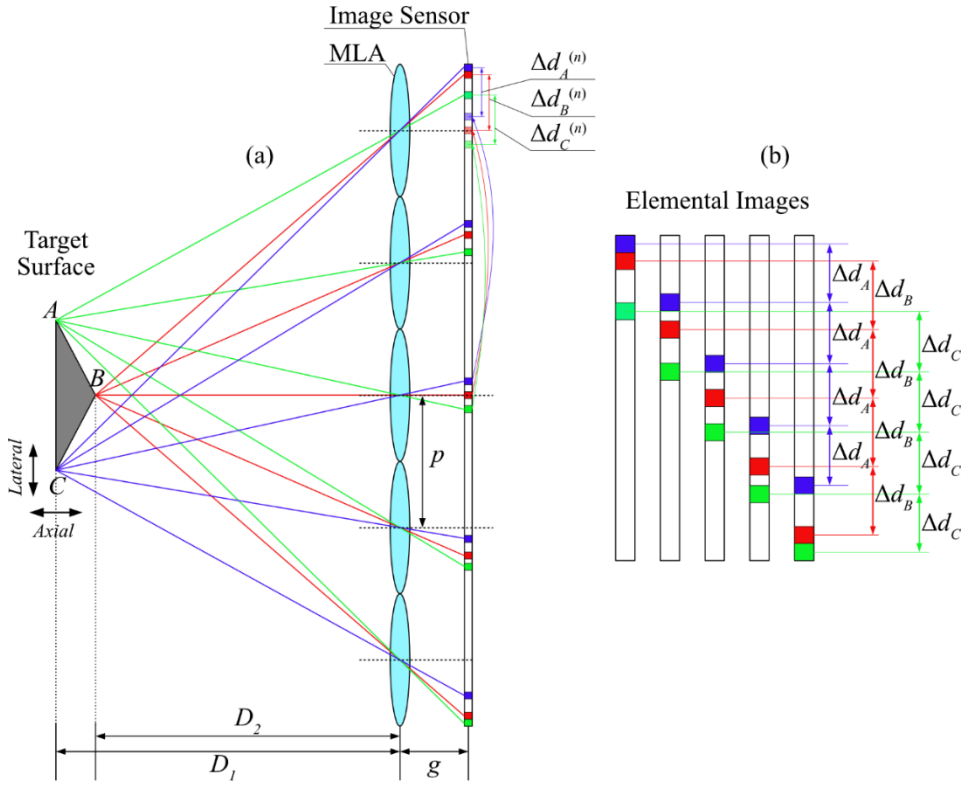

Figure 3.3 Measuring principle based on InI.

As shown in Fig. 3.3, each object point (A, B, and C) are recorded by each of the micro-lenses with different (x, y) coordinates. These pixel points from the same object point (e.g., the blue pixel points) are corresponding points whose coordinate difference is the disparity determined by the depth of the object point. After stacking the elemental images as Figure 3.3(b), the disparities of the corresponding points of A, B, and C are Δd_A, Δd_B, and Δd_C, respectively in the adjacent EIs. Based on the geometrical relationship shown in Figure 3.3(b), the relationship between the disparity and the depth can be established as shown in Eq. (3.1).

$$\frac{g}{D} = \frac{\Delta d \cdot \Delta_p}{p} \tag{3.1}$$

where D represents the gap from a point to the MLA and Δ_p is the pixel size. It is apparent that the depth for the object point can be acquired based on disparity information of the corresponding points. Regarding the non-adjacent EIs, the equation changes to

$$\frac{g}{D} = \frac{\Delta d^{(n)} \cdot \Delta_p}{n \cdot p}, \quad n \in N_+ \text{ and } n > 1 \tag{3.2}$$

where $\Delta d^{(n)}$ is the disparity of the corresponding points in non-adjacent EIs. Obviously, $\Delta d_A = \Delta d_C < \Delta d_B$ in the diagram. Finally, the axial distance ΔD of two object points can be acquired as

$$\Delta D = \left| D_1 - D_2 \right| = \frac{ngp}{\Delta_p} \left| \frac{1}{\Delta d_1^{(n)}} - \frac{1}{\Delta d_2^{(n)}} \right|, \quad n \in N_+ \tag{3.3}$$

3.3 Depth reconstruction

Based on the spatial–angular information recorded in the raw measurement data, it is possible to extract disparities directly or implicitly from the data to reconstruct the target surfaces. The digital refocusing method is used to rearrange the pixels of the recorded elemental images at various depth planes, and therefore, all the corresponding points from one single object point only focus at a deterministic depth. Through analyzing the focus level of the CPs, the depth of the corresponding object point is able to be determined. This method reconstructs the refocused images based on the reversibility of light rays and the disparities are not extracted directly but contribute to the focus level. The method based on epipolar-plane images is used to directly extract disparities from the elemental images, where one angular dimension and one spatial

dimension are fixed and the pixels in the remaining two directions form a map (which is called an epipolar-plane image). The CPs in one EPI form a diagonal whose slope is the disparity. In this section, various methods for depth reconstruction are discussed.

3.3.1 Digital refocusing

The digital refocusing method reprojects the pixels in EIs to various depth planes so as to compose multiple refocused images, with a vivid illustration shown in Figure 3.4. A virtual MLA that is assumed to be a pinhole-like array of lenses for simplification is placed behind the EI plane so that every pixel produces a backpropagated ray in a free space. Various depth planes can be placed in any position behind the virtual MLA at will, and the backpropagated rays are projected onto the planes to form a sequence of defocused images. In the diagram, obviously, point B' (in red) that is from point B is focused at the depth (b) and point A' (in blue) and C' (in green) are focused at the depth (d). At depth (a) and depth (c), all the corresponding points of points A, B, and C cluster in bokeh regions, and no focus information is detected.

Using ray transfer matrix analysis, it is possible to map the coordinates of each pixel from the EI plane to the depth plane, which is perpendicular to the EI plane and has a small ray angle denoted by α. Under the small-angle approximation, $\sin\alpha \approx \alpha$. For simplification, only two dimensions v and y are considered for the demonstration. A point P with $\left(v^P, y_e^P\right)$ coordinates in the EI plane is rearranged with y_{d1}^P and y_{d2}^P at the depth d_1^R and d_2^R, respectively. As a result, the difference between y_{d1}^P and y_{d2}^P is

$$diff_R = \left|y_{d1}^P - y_{d2}^P\right| = \left|\frac{\alpha}{\Delta_p}\left(d_1^R - d_2^R\right)\right| \tag{3.4}$$

where Δ_p is the pixel size of the image sensor. It is noted that the maximum value of $|\alpha|$ is determined by $p/2g$. Since the pixel number is finite in the refocused images, the axial resolution of the depth slices is determined by minimum $diff_R$. Obviously, the axial resolution of the depth slices is enhanced by decreasing the size of one single pixel.

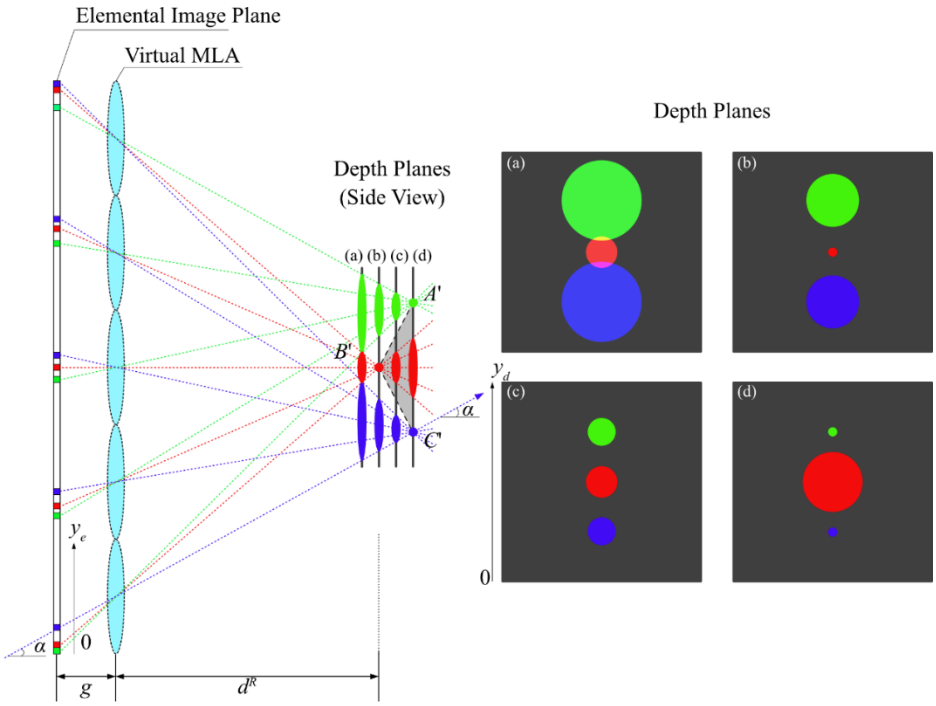

Figure 3.4 Principle of the digital refocusing method.

3.3.2 Epipolar-plane image analysis

An illustration of epipolar-plane image analysis is shown in Figure 3.5 where the central view of the elemental images is shown in (a), and one epipolar-plane image with its gradients along the *x* axis is shown in (c). The scene is from the *Stanford* light field dataset (Wilburn et al., 2005) and the angular resolution is 17×17. After fixing the *v* and *y* coordinates at (v_0, y_0), a 2D image (i.e., the epipolar-plane image) is acquired in

the $x-u$ plane. It is observed that the corresponding points form multiple diagonals in the EPI. The slopes $\Delta x/\Delta u$ of the diagonals are different and determined by the distance from the corresponding object point to the MLA.

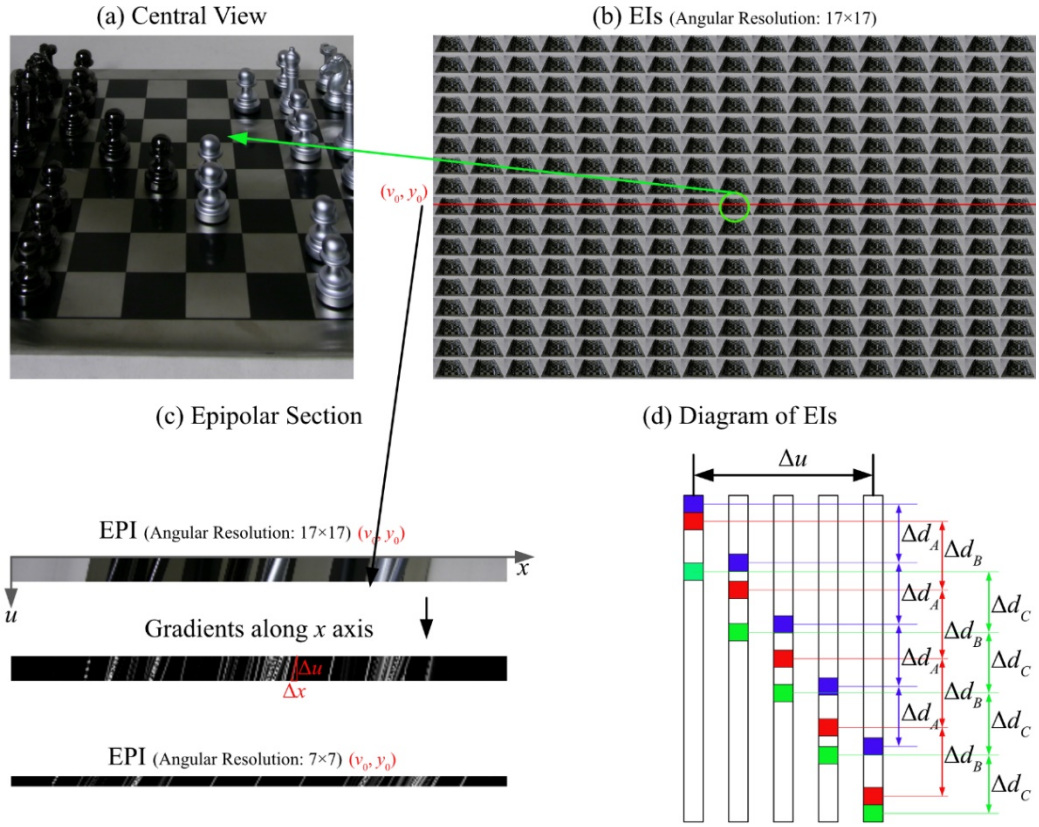

Figure 3.5 Illustration of EPI extracted from LF data.

Compare the EPI with the diagram shown in Figure 3.5 (d) where Δd_* denotes the disparity between two corresponding points in two adjacent EIs. It is found that the Δx of one slope in the EPI is the disparity which is determined by the object depth. As a result, the depth of the object points are obtained after the extraction of the slopes in the EPI based on Eq. (3.2). As shown in Figure 3.5 (c), the angular resolution determines the quantity of corresponding points and the length of the associated diagonals. Hence, the larger angular resolution increases the estimation accuracy of the slopes in the EPIs.

3.3.3 Disparity pattern-based autostereoscopic reconstruction

The disparity pattern-based reconstruction method was proposed by D. Li et al. (2015), which performs digital refocusing based on the disparity pattern to realize a reconstruction with only focused information. As shown in Figure 3.6, a point *S* moves axially and a group of disparity patterns is available at any depth based on the relationship between disparities and depth. In other words, if the point *S* is right at the depth d_R^*, the corresponding points of *S* must obey the same distribution as the disparity pattern. As a result, a group of points from the EIs are extracted based on the disparity patterns at an assigned depth and the points are matched as corresponding points. Evaluation of the matching is based on the greyscale of pixel points and the gradient of multiple directions. If the points do not correspond to the same object point, the points are filtered and not projected onto the depth planes for refocusing. Essentially, the method integrates the disparity patterns as a constraint to the matching process so as to improve the matching efficiency.

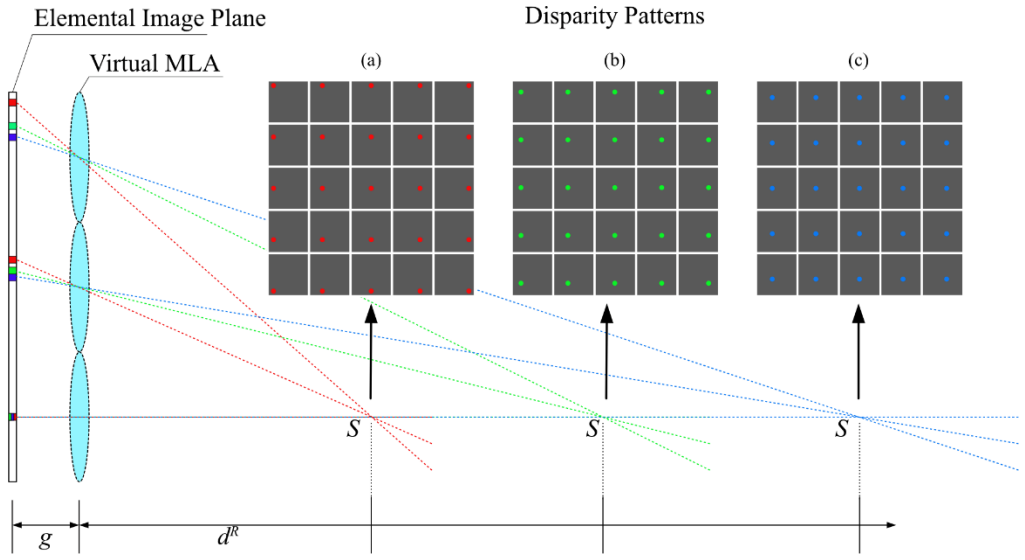

Figure 3.6 Illustration of disparity patterns at various depths.

3.4 Calibration process of autostereoscopic systems

The calibration process is illustrated in in Figure 3.7, where a standard target serves as the calibration reference. This reference target is mounted on an XY translation stage and moves axially within the depth of field of the autostereoscopic measuring system. As the target position changes, multiple calibration data points are acquired for disparity extraction through digital refocusing. The separation D_* between the target and the main lens determines the value of the disparity d_*, and their correlation can be expressed as $D_* = f(d_*)$. However, accurately measuring the exact distance is challenging, making it difficult to establish a precise relationship $f(\cdot)$.

It should be noted that the distance can be further represented as $D_* = \Delta + C_*$ where Δ is a constant. The value of C_* can be obtained from the translation stage, corresponding to the movement of the reference target. Obviously, it is possible to find a function $g(\cdot)$ that correlates the translation C_* with the disparity d_*, formulated as $C_* = g(d_*)$. Consequently, by using the multiple calibration data, a curve that maps the relationship between C_* and d_* can be fitted for the calibration of the autostereoscopic system.

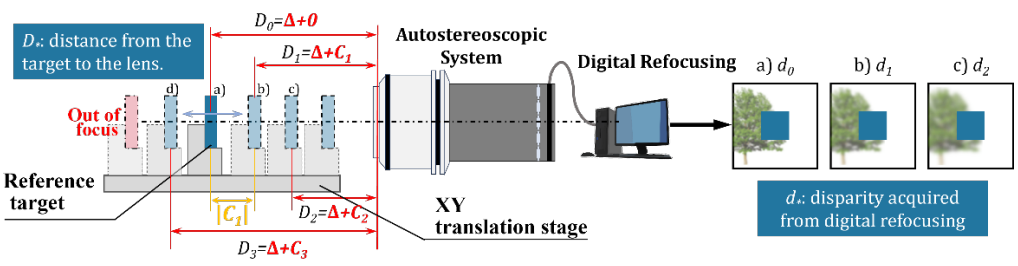

Figure 3.7 Calibration process of the autostereoscopic measuring system.

3.5 Rapid 3D inspection of wire bonding

Surface-mounted devices (SMD) such as light-emitting diodes (LED) are widely used electrical components, mounted or placed on the surface of a circuit board. They usually consist of resins, a chip, two pads, a gold wire, and a circuit pattern. SMD LEDs have been applied in car lights, street lamps, displays, projectors, general illumination, industrial illumination, decorative lights, etc. (Vieroth et al., 2009), because of their low power consumption and high luminance emission. An important manufacturing process of SMD LEDs is wire bonding during which the chip and the circuit are connected. Since the wire dimension is small, which is usually within the range of 15 μm to 50 μm in diameter, defects could occur during the wire bonding. The possible defect categories are shown in Figure 3.8.

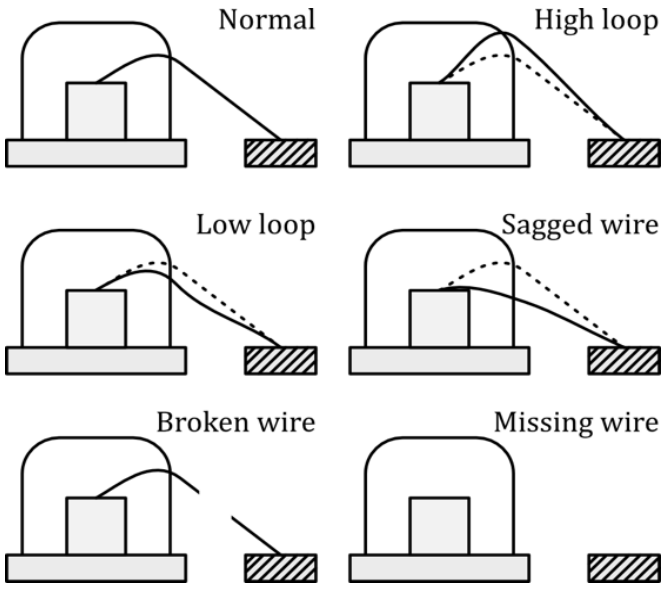

Figure 3.8 Defect categories of bonding wire.

The five types of defects include high loop, low loop, sagged wire, broken wire, and wire missing (Chen et al., 2021). To avoid the defects resulting in the failure of the LEDs, effective and efficient inspection of the LEDs to measure the 3D dimension

information of the bonding wire is necessary. Some of the current wire bonding inspection systems (Chen et al., 2021; Chen & Tsai, 2021) can only detect lateral defects by looking at the samples from above. However, the defects caused by the wrong height of the wire are not able to be recognized.

Traditional optical measurement systems (Perng et al., 2007) for wire bonding obtain 3D dimensional information by moving the lens several times along the axial axis or repositioning the sample stage to inspect different parts of the samples. Some stereo inspection systems (Ye et al., 2000) can obtain the height information from multiple angles, but the calibration could import further measurement errors. In addition, most of the current systems are either time-consuming or complicated to establish, which makes the practical implementation difficult for rapid and accurate 3D dimensional inspection. To this end, an autostereoscopic 3D measurement system for SML LED inspection has been developed to cater to the needs of the SMD LED manufacturing industry.

On the basis of the autostereoscopic principle, the proposed system for rapid inspection of SMD LEDs is illustrated in Figure 3.9, where Figure 3.9 (a) presents the system architecture. Figure 3.9 (b) demonstrates the measuring principle. One single super-zoom magnification objective lens directly receives light rays emitted by the measured SMD LEDs, an MLA is positioned between the objective lens and a high-resolution image sensor, and a series of EIs from slightly different view angles is recorded on the image sensor. Each small image region corresponding to the micro-lenses records a part of the measured object and one single object point is recorded on different image regions forming multiple pixel points which are called corresponding points. The depth information can be analyzed and extracted from these corresponding points.

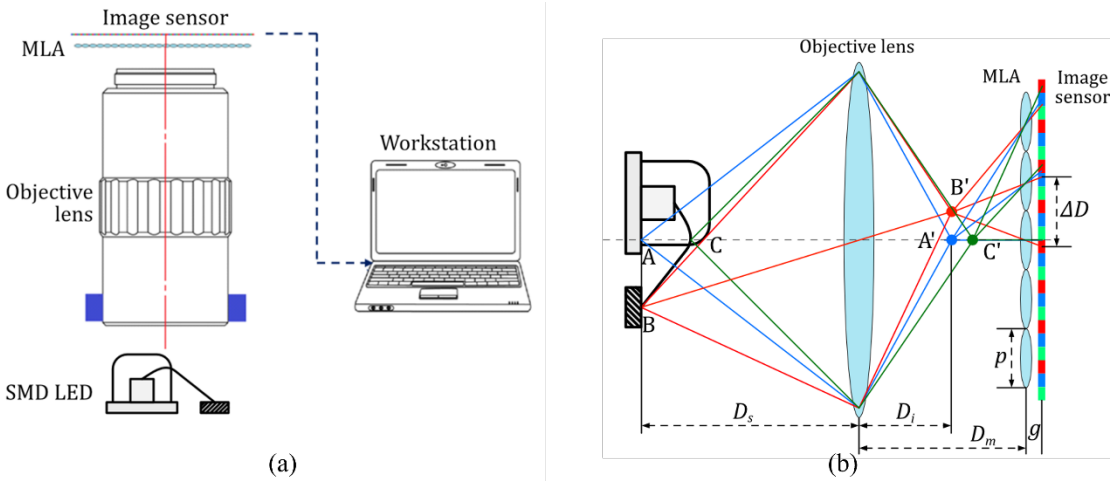

Figure 3.9 Rapid inspection system for SMD LEDs based on autostereoscopy. (a) System framework. (b) Recording process.

According to the imaging theory, the rays emitted by the measured object form an image behind the objective lens. Since the distances from the object points to the lens are different (e.g., point A, B, and C), the image distances are also different. The MLA adjacent to the image sensor receives the rays emitted by the image points and splits the rays into different directions. Each micro-lens from a specific angle observes the image points behind the objective lens, and the image sensor records the rays passing through the micro-lenses. As a result, a series of EIs with different view angles is obtained.

During measurement, the shooting distance D_s determines the recording positions of each object point on each EI. On the fundamental basis of the autostereoscopic principle, the recording process is formulated as Eq. (3.5) and Eq. (3.6).

$$\frac{1}{D_s}+\frac{1}{D_i}=\frac{1}{f} \tag{3.5}$$

$$\frac{\Delta D}{p} = \frac{g + D_m - D_i}{D_m - D_i} \tag{3.6}$$

where D_i is the imaging distance, D_m is the distance from the objective lens to the MLA, f represents the focal length of the main lens, p represents the individual micro-lens pitch, g is the distance from the MLA to the image sensor, and ΔD is the distance of the CPs of one point in the neighbouring elemental images. For the object points with the same depth (e.g., point A and B), they have the same ΔD. The points with different depths (e.g., point A and C) have different ΔD. Hence, the depth information can be straightforwardly extracted from the EIs obtained.

A reconstruction method for the rapid inspection of SMD LEDs is presented as illustrated in Figure 3.10. Since the depth information is directly related to the positions of the corresponding points of one object point, every group of the corresponding points is only focused on a specific depth. Through simulating the recording process using the virtual MLA, the refocusing is achieved by mapping all the pixel points in the EIs to a specific depth plane. After multiple refocusing processes on various depths, an image sequence of depth slices is acquired. The successful focused region in each depth slice indicates that the group of the corresponding points forming this focused region come from this depth. By stacking the depth slices, the focused slice can be detected to determine the 3D dimensions of the measured samples.

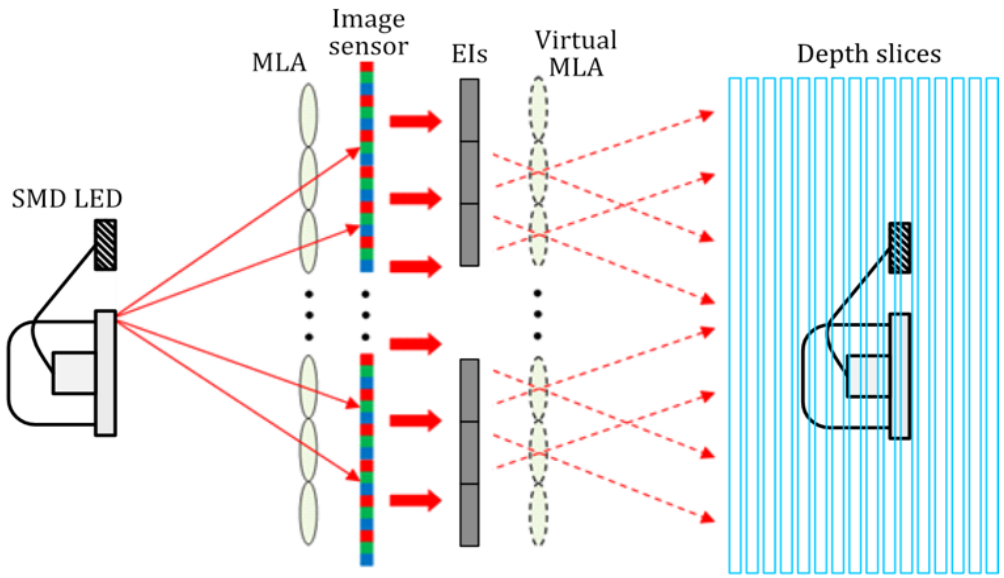

Figure 3.10 Refocusing process for the 3D inspection of SMD LEDs.

On the basis of the refocusing process, the depth extraction is realized by identifying the focused pixel regions in every depth slice. A sharpness function is defined to detect the focus level of a pixel point in one slice. The detected focus level is determined by the greyscales of the centre point and its surrounding points in a small region. Define a group of pixels in a 3×3 region as $I^{3\times3}$. The function is to compute the gradient value of the centre pixel point (i,j) using Eq. (3.7), where F_k is the sharpness index in the k-th depth slice. G_x and G_y are the gradient matrices in the x and y directions, respectively. After M depth slices are obtained, the peak gradient value of one pixel point among the slices is the desired focused value. Hence, the depth points can be acquired by Eq. (3.8), where $d(i,j)$ is the desired depth of the point (i,j).

$$G_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, \quad G_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix},$$

$$I^{3\times3} = \begin{bmatrix} I_{i-1,j-1} & I_{i,j-1} & I_{i+1,j-1} \\ I_{i-1,j} & I_{i,j} & I_{i+1,j} \\ I_{i-1,j+1} & I_{i,j+1} & I_{i+1,j+1} \end{bmatrix}, \tag{3.7}$$

$$F_k(i,j) = \sqrt{\left(G_x * I^{3\times3}\right)^2 + \left(G_y * I^{3\times3}\right)^2}$$

$$d(i,j) = \arg\max_k \left[F_1(i,j), \ldots, F_k(i,j), \ldots, F_M(i,j)\right] \tag{3.8}$$

An example of the image sequence of the depth slices is shown in Figure 3.11. The wire roof (red circle) and the welding spot (blue circle) are focused on slices (B) and (C), respectively. It is obvious that the pixel points of the wire roof have the maximum sharpness value in the depth represented by the slice (B), as shown by the peak of the red curve. Through the same analysis of the points of the wire spot, the span of the wire and the height of the wire are easy to calculate.

To achieve a fast inspection, the proposed system uses a local reconstruction based on the mentioned focus detection method, instead of reconstructing all the depth slices. The local reconstruction is performed on the regions of interest so that only finite pixels corresponding to the key points of target samples are remapped on the depth slices. This achieves a comparatively rapid inspection of the defects of the targets.

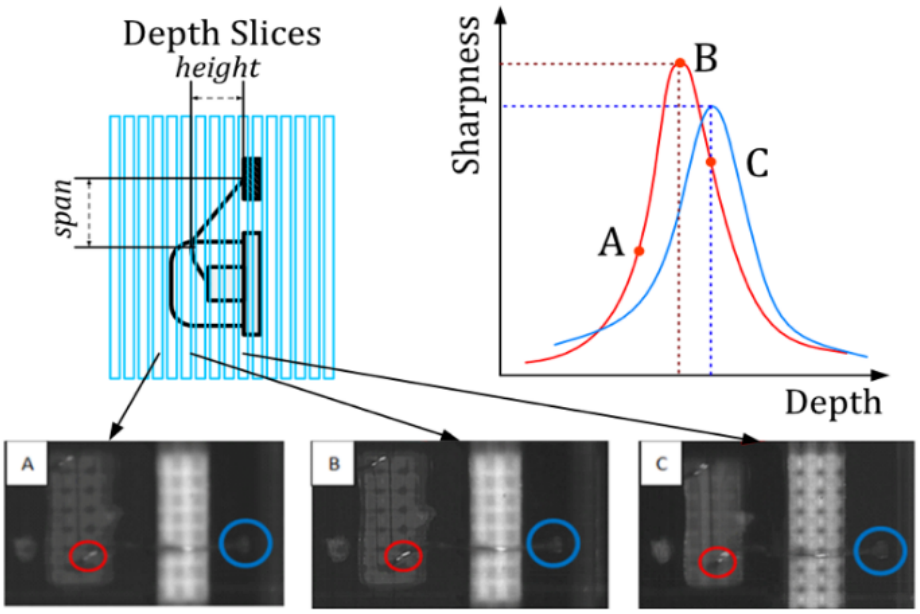

Figure 3.11 Depth detected via the focused planes.

The proposed 3D inspection system for SMD LEDs and the experiment platform were established as depicted in Figure 3.12, where a ring-shaped light device was placed above the package of SMD LEDs. An X-Y positioning stage was used to realize the lateral movement of the measured samples. A series of EIs of the samples were captured within one snapshot and were processed by the proposed refocusing method to determine the 3D information of the measured LEDs. A total of 10 experiments were performed to reduce the systematic error without changing the axial and longitudinal measurement region. The experimental outcomes are displayed in Table 3.1. When the sample contains defects, the height of the bonding deviates from the standard requirement according to the defect categories. Accurate measurement of the height is used to filter out defective samples. The results acquired by Alicona IntiniteFocus which is a mature commercial 3D measurement system were taken as a reference. It is shown that the measurement results obtained by the proposed system realize high accuracy and reliability by comparison with the reference value. In addition, the

proposed system only took approximately 1.8 seconds to obtain the results, which dramatically exceeds the processing speed of the compared commercial system.

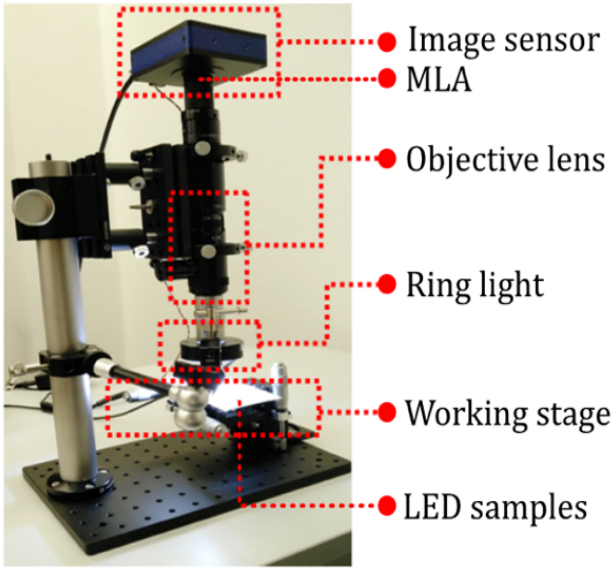

Figure 3.12 Setup of the proposed 3D inspection system for SMD LEDs based on autostereoscopy.

The proposed 3D inspection system is effective and efficient to conduct inspections of SMD LEDs. The system can reconstruct the desired 3D dimensions of the measured samples so that the quality of the bonding wire can be inspected during manufacturing. Since the system can record multiple viewpoint EIs within just one snapshot, the measurement speed is high. In addition, accurate height and lateral dimension information can be rapidly reconstructed using the proposed refocusing method. The proposed system is compact and easy to implement under various working conditions including external environments. The system has the potential for the improvement of the manufacturing efficiency of SMD LEDs and reduction of the failure rate resulting from the wire bonding defects.

Table 3.1 Experimental results for the 3D inspection of SME LEDs

Measurement		Time (s)	Height (μm)	Span (μm)
Alicona IntiniteFocus		> 300	53.146	348.475
Proposed	1	1.748	52.9	348.3
	2	1.764	52.7	348.1
	3	1.793	53.1	348.7
	4	1.802	53.1	348.7
	5	1.800	53.3	349.0
	6	1.811	53.1	348.9
	7	1.723	52.9	348.4
	8	1.746	52.8	348.2
	9	1.829	53.2	349.0
	10	1.834	53.3	349.1
	Avg.	**1.785**	**53.04**	**348.64**

3.6 Summary

In this chapter, the measuring principle of the autostereoscopic 3D measuring system based on InI is discussed. An autostereoscopic system typically consists of a main lens, a high-magnification zoom lens system, a micro-lens array, and an image sensor. Illumination devices are necessary for micro-structured surfaces to make the system able to inspect the micro-scale patterns. The reconstruction process is reversed with the recording process and the disparity information lying at the corresponding points which is directly determined by the distance between the observed point and the

micro-lens array. Based on the digital refocusing method, EPI-based method, and disparity patterns, the 3D information of the measured surfaces is detected so as to conduct the reconstruction. An experiment on rapid inspection of wire bonding was conducted to assess the feasibility and measurement performance of the autostereoscopic 3D measurement system. Importantly, the limitation of the autostereoscopy-based system is the resolution of the recorded elemental images. On the basis of the previous analysis, the spatial resolution contributes to the depth resolution in the digital refocusing process, whereas the angular resolution implicitly determines the matching accuracy during disparity extraction. This trade-off constrain the measurement accuracy of the autostereoscopic system. Hence, enhancing the resolution of the measurement data is crucial.

Chapter 4 Angular resolution enhancement for autostereoscopic measurement data using deep learning

4.1 Introduction

Light field techniques have been widely used since the techniques can provide rich 3D information of the real world within only one snapshot. Three-dimensional scenes can be easily reconstructed through extracting the disparity information in the LF images which are composed of a series of sub-aperture images (SAIs). This makes LF cameras draw a lot of attention in photography (Marwah et al., 2013), vehicle vision (Fürsich, 2019), virtual reality (Overbeck et al., 2018), etc. However, an inevitable obstacle of LF techniques is the trade-off between image detail and the range of viewing angles in LF images. Currently, a vast amount of research work has been conducted to restore spatial resolution from low-resolution SAIs (Jin, Hou, Chen, & Kwong, 2020; Zhang et al., 2019). The work is similar to SISR and many techniques have been developed to achieve spatial resolution enhancement. Enhancing angular resolution by interpolating novel views from low-angular-resolution SAIs is still challenging.

The methods applied for angular resolution improvement are basically non-depth-based (Meng et al., 2020; Yeung et al., 2018; Yoon et al., 2017) as well as depth-based (Jin, Hou, Chen, Zeng, et al., 2020; Jin, Hou, Yuan, et al., 2020; Wu et al., 2017, 2019). Some of the non-depth-based methods (Meng et al., 2020; Yoon et al., 2017) directly process low-resolution SAIs and generate novel views. Some researchers (Wu et al., 2017, 2019) used epipolar-plane images to achieve novel view interpolation. Nevertheless, when the disparity range of the LF images is large, these methods usually

produce severe ghosting on the interpolated novel view images. Depth-based methods usually produce high-quality images with sharper edges and less image ghosting. Correct depth estimation is vital to provide geometric transformation for the reconstruction of novel views. However, accurate depth estimation in real-world scenarios is more challenging especially when large baselines and multiple targets with various depths exist in a scene, and the target surfaces include complex textures or patterns.

One challenging issue related to angular resolution enhancement is the lack of ground truth. A popular supervised learning paradigm artificially down-samples the raw dataset first and makes use of the raw SAIs as ground truth. To reach a compromise on the traditional learning paradigm, the data are required to be paired as input and corresponding labels. This imposes a simple artificial down-sampling strategy to the training process of the learning models. According to Dansereau et al. (2013), the rays in a 4D light field captured by a plenoptic camera can be represented as

$$\phi^{A} = \boldsymbol{H}\boldsymbol{n} = \begin{bmatrix} H_{1,1} & 0 & H_{1,3} & 0 & H_{1,5} \\ 0 & H_{2,2} & 0 & H_{2,4} & H_{2,5} \\ H_{3,1} & 0 & H_{3,3} & 0 & H_{3,5} \\ 0 & H_{4,2} & 0 & H_{4,4} & H_{4,5} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} i \\ j \\ k \\ l \\ 1 \end{bmatrix} \tag{4.1}$$

where i, j are the pixel indices within each lenslet image, k, l are the indices of the micro-lens, and $\boldsymbol{H} \in^{\sim 5\times5}$ is a homogeneous intrinsic matrix. It is obvious that the count of rays in a captured light field is dependent on pixel density. The density of rays in a LF affects the precision of disparity estimation and the quality of novel view reconstruction. Hence, this down-sampling wastes the redundant light field information in the light field data. More density information is required to be complemented by the

learning models so that more challenges are imposed on the ill-posed problem. Another issue in current research is the performance for real-world images with large baselines. Although many learning models can produce excellent results and high PSNR/SSIM for synthetic light field datasets, it is still tricky to reconstruct clear novel views on real-world light field scenes with large baselines and muti-depth targets. More attention should be paid to the angular enhancement of the sparse real-world light field images.

This chapter introduces a new semi-supervised learning paradigm designed to fully exploit captured light field data. No ground truth is required by the proposed learning paradigm and the model is directly supervised by the input. As a result, all the data could be exploited thoroughly without the wastage of data. To adapt the proposed learning paradigm, a straightforward convolutional neural network is developed to synthesize novel views via fusing adjacent views in the raw dataset. The synthetic manner is similar to the pioneer baseline method (LFCNN) proposed in Yoon et al. (2017).

To achieve interpolation with a better view, inspired by the plane sweep volume technique (Im et al., 2019), a simplified equivalent of depth estimation is utilized in the proposed method which predicts the opposite motion values of local regions in two adjacent input views rather than directly predicting the disparity pixel-wisely. In addition, the motion estimation is converted to a classification problem (Peleg et al., 2019) to constrain the prediction of desired motion to a pre-defined range.

Evaluation experiments are conducted on real-world LF and autostereoscopic datasets to demonstrate the universal effectiveness of the presented algorithm and its superiority to autostereoscopic measurement. The proposed model is trained on 20

scenes of the *Heidelberg Collaboratory for Image Processing* (*HCI*) dataset (Honauer et al., 2016), following the proposed semi-supervised learning paradigm. To illustrate the data efficiency of the proposed learning model, the training data are assumed to have a low angular resolution.

Under this assumption, the data are down-sampled from 9×9 to 5×5 in angular resolution and the filtered data are removed from the training set. It is noted that the purpose of this down-sampling is different from that in previous research for the generation of ground truth. The down-sampling aims to create a training set with a limited angular resolution which usually happens in a self-built light field.

As a result, the proposed method is trained using only 500 SAIs, while other comparative methods utilize the complete dataset consisting of 1,620 SAIs for comparison. The evaluation is performed without any finetuning. Quantitative results show that the learning model following the proposed learning paradigm can achieve high PSNR/SSIM for both synthetic and real-world datasets in comparison with the SOTA methods (Jin, Hou, Chen, Zeng, et al., 2020; Jin, Hou, Yuan, et al., 2020; Wu et al., 2019), whereas fewer training data are used.

Qualitative comparisons reveal that the proposed method can interpolate high-quality novel views with sharper edges and clear contents, and the proposed method tends to reconstruct more accurate parallax structures under the semi-supervised learning paradigm. To further demonstrate the superiority of the proposed learning paradigm, the baseline method of LFCNN is trained under supervision and the proposed semi-supervision separately. Quantitative experiments are performed on multiple real-world datasets and the improvement is up to 2 dB in PSNR. It is also found that the

parallax structures recovered by the semi-supervised LFCNN contain more accurate details.

In terms of measurement data captured by the autostereoscopic 3D measuring system, finetuning is conducted for all the methods. However, the comparative methods which are effective for the LF images fail to produce accurate novel views, where severe image ghosts occur in the reconstruction. This is because the measurement data are small in regard to spatial resolution and texture-less. It is difficult for the comparative methods to achieve high-quality reconstruction by struggling for accurate depth estimation from the measurement data. The proposed method has the capability to reconstruct high-quality novel views under the training of limited measurement data. Digital refocusing is performed to show the improvement achieved by various methods. The proposed learning paradigm and method also achieve more accurate angular super-resolution by reconstructing more sharp edges and clear geometries compared with the other method.

4.2 Angular super-resolution definition

The proposed method aims to interpolate novel views among the low-resolution SAIs. Let $\Lambda_{u,v}(x,y)$ represent the intensity of a pixel located at coordinate (x,y) of the SAI and the SAI is at the angular coordinates (u,v) of the 4D LF. $(U\times V)$ and $(X\times Y)$ are the angular and the spatial resolution of the LF data, respectively. A group of SAIs is shown in Figure 4.1 and four neighbouring SAIs are circled. Novel-view SAIs can be interpolated based on the four neighbouring SAIs, which are located in the middle and centre of the four SAIs. A total of five novel-view images can be reconstructed from these four SAIs. As a result, the angular resolution is enhanced from

$U \times V$ to $(2U-1)\times(2V-1)$ after interpolation. To further enhance the angular resolution, the interpolation can be performed multiple times. For instance, two times of interpolation improves the angular resolution from 3×3 to 9×9.

Different from SISR that computes and interpolates new pixels between the neighbouring pixels, angular super-resolution methods make use of the pixels from the neighbouring SAIs and their neighbouring regions to determine the desired new pixel values. These input pixels are represented by the same object point but recorded from different view angles. Defining these different view-angle pixels as a group of corresponding points and the interpolation problem of angular super-resolution can be simply defined as

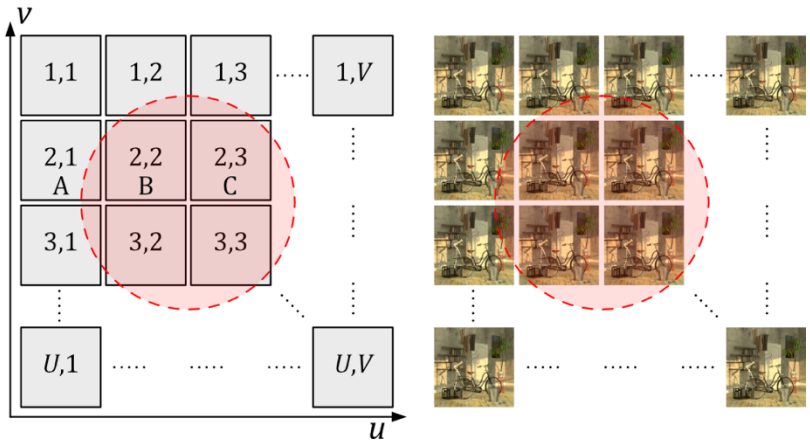

Figure 4.1 An example of 4D light field images.

$$\begin{aligned} \Lambda_m(x_m, y_m) &= f\left(\Lambda_1(D_1), \Lambda_2(D_2)\right) \\ D_* &= \left\{(x,y) \mid x_* - r \le x \le x_* + r, y_* - r \le y \le y_* + r\right\} \end{aligned} \tag{4.2}$$

where Λ_1 and Λ_2 are two neighbouring SAIs which could be in the horizontal direction or the vertical direction, and Λ_m is the desired novel-view SAI. (x_*, y_*) denotes the coordinates of the corresponding points in the SAI Λ_*. D_1 and D_2 are two

small pixel regions within a range $[-r,r](r>0)$, of which the centre pixels are the corresponding points. All the pixels in the two regions contribute to the reconstruction of the corresponding new points in the interpolated-view SAI. Obviously, the different view angles lead to the different centre coordinates (x_*, y_*) of the two corresponding regions. Direct convolution on the neighbouring SAIs (Meng et al., 2020; Yoon et al., 2017) usually holds the hypothesis that the (x_*, y_*) coordinates of the centres are approximately the same, i.e., $|x_1 - x_2| < \varepsilon$ and $|y_1 - y_2| < \varepsilon$, where ε is a negligible constant. As a result, the direct convolution usually generates ghosting interpolation results when parallax is large. For the data with large parallax, accurate depth estimation can guarantee that the two local regions have correspondence.

4.3 Semi-supervised learning paradigm

According to Eq. (4.1), the reconstruction process aims to extract the ray distribution in a light field based on the finite recorded pixels. In some research, only corner SAIs are input into the learning models and the remaining SAIs are used as ground truth. As a result, the learning models can only make use of a finite ray distribution of the captured light field and need to learn to fit the dense light field information under supervision. Albeit that the redundant light field information is not fully exploited, most of the models can achieve excellent estimation due to the powerful fitting capability of the learning models. To allow more efficient learning of a light field representation, a learning paradigm that can make full use of the collected data is necessary to improve the learning efficiency, especially when only limited training data are acquirable.

T Using three contiguous SAI regions, labelled (A), (B), and (C) in Figure 4.1 as

an illustration, a desired novel view (P) could be interpolated between (A) and (B), while another view (Q) is between (B) and (C). The pixels in the novel views (P) and (Q) can be interpolated by finding corresponding points in (A), (B), and (C). When ℓ is small, the corresponding points can be involved within one convolutional window so that the baseline method of LFCNN can even achieve satisfactory results for data with small baselines. In terms of a large baseline, explicit and implicit depth estimation is always necessary to determine the corresponding points in different views. The synthetic novel views (P) and (Q) are required to be supervised so that the models can be trained to detect the corresponding points. Unfortunately, there is no ground truth for (P) and (Q) if no artificial down-sampling is performed before the training. It is obvious that another novel view (Z) can be interpolated between (P) and (Q) and the view (Z) should be an equivalent to the original view (B). Therefore, the synthetic view (Z) can be supervised so (P) and (Q) can be constrained indirectly. It is noted that the interpolation can be performed infinitely and 'ground truth' for the novel views can always be found after every 2-step interpolation, though the 'ground truth' may not always be the original SAI but an interpolated view.

Motivated by the above analysis, a novel semi-supervised learning paradigm in a cycle-like fashion is proposed, which is illustrated in Figure 4.2. During each training iteration, totally nine neighbouring SAIs $\mathbf{\Lambda}^{ori}$ in a 3×3 grid are fed into the model. Horizontal and vertical interpolation happens in the middle of the adjacent SAI pairs, and another horizontal interpolation between the two novel views produces the centre view. In Step **1**, a total of 16 novel-view SAIs $\mathbf{\Lambda}^{sl}$ are interpolated horizontally, vertically, and centrally. This enhances the angular resolution of the original data from 3×3 to 5×5. The red SAIs in Figure 4.2 are the novel views interpolated in the first step.

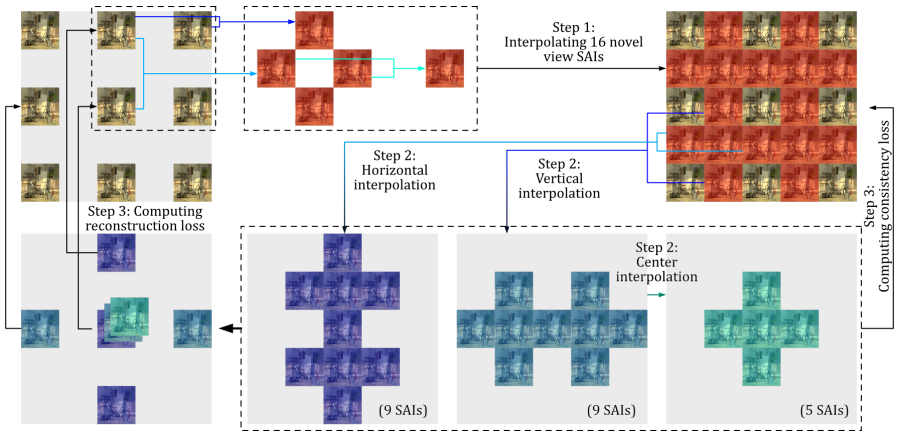

Figure 4.2 Flowchart of the proposed semi-supervised learning paradigm.

Since there is no ground truth for the 16 novel SAIs, in Step **2**, $\mathbf{\Lambda}^{s1}$ are directly fed into the network again and a new round of the three interpolation operations happens. It is noted that the values in the novel-view SAIs exceeding 1.0 are truncated for normalization. Based on the novel views horizontally interpolated in the first step, nine novel views (in blue) are produced by the second round of vertical interpolation. Similarly, another nine novel views (in purple) are produced by the second round of horizontal interpolation. Additional horizontal interpolation based on the blue views produced in the second step synthesizes five new views in green. As a result, a total of 23 novel-view SAIs $\mathbf{\Lambda}^{s2}$ (in purple, blue, and green) are produced in Step **2**. It is noted that no backpropagation happens during the first two steps. After the interpolation in the first two steps, redundant novel-view SAIs (a total of 39 views) are produced based on the nine input SAIs. Obviously, some views have the same angular coordinates (u,v) and can be constrained by each other.

According to the above analysis, the errors between the SAIs with the same angular coordinates are measured to provide the training loss in Step **3**. Firstly, a reconstruction loss between $\mathbf{\Lambda}^{ori}$ and $\mathbf{\Lambda}^{s2}$ is acquirable since part of the novel views interpolated in the

second step have the same (u,v) as the original input. These new perspectives can be directly supervised using the input. It is also found that a consistency loss between $\mathbf{\Lambda}^{s2}$ and $\mathbf{\Lambda}^{s1}$ is obtainable in a similar fashion.

In terms of the reconstruction loss, L1 (Least Absolute Deviations) and perceptual loss are used to obtain the errors. The L1 loss can constrain the pixel-wise errors between the different views at the same angular coordinate, and the perceptual loss measures the perceptual differences at feature levels. The reconstruction loss L_r is formulated as Eq. (4.3),

$$\begin{aligned}\mathrm{L}_r = \mathrm{E}\left(\sum_{i=1}^{23}\sum_{j=1}^{9}\delta_{\left(u_{2_i},v_{2_i}\right),\left(u_{0_j},v_{0_j}\right)} * \left|\mathbf{\Lambda}^{s2}_{u_{2_i},v_{2_i}} - \mathbf{\Lambda}^{ori}_{u_{0_j},v_{0_j}}\right|\right) \\ +\mu\cdot\mathrm{E}\left(\sum_{i=1}^{23}\sum_{j=1}^{9}\delta_{\left(u_{2_i},v_{2_i}\right),\left(u_{0_j},v_{0_j}\right)} * \left|\phi\left(\mathbf{\Lambda}^{s2}_{u_{2_i},v_{2_i}}\right) - \phi\left(\mathbf{\Lambda}^{ori}_{u_{0_j},v_{0_j}}\right)\right|^2\right)\end{aligned} \tag{4.3}$$

$$\delta_{(a,b),(c,d)} = \begin{cases} 1 & (a=c)\cap(b=d) \\ 0 & (a\neq c)\cup(b\neq d)\end{cases} \tag{4.4}$$

where $\mathrm{E}(\cdot)$ is the expectation, $\phi(\cdot)$ is the high-level features and μ is the penalty coefficient. A similar consideration is carried out for the consistency loss between $\mathbf{\Lambda}^{s1}$ and $\mathbf{\Lambda}^{s2}$. Only L1 loss is used to measure the errors to accelerate the training speed. Although the perceptual loss may provide a feature constraint between $\mathbf{\Lambda}^{s1}$ and $\mathbf{\Lambda}^{s2}$, the additional computation cost only yields a negligible improvement in our experiments. As a result, the consistency loss is formulated as Eq. (4.5). The integral loss function of the proposed learning paradigm is formulated as Eq. (4.6), where α is the penalty coefficient of the consistency loss.

During the training process under this paradigm, 39 novel views are interpolated. This

achieves considerable self-augmentation of the training data so that the learning efficiency can be improved significantly. In addition, the training paradigm is consistent with the recording of light fields where the interpolated views are the rays not recorded but that exist in the field, so that the light field structures can be learned during the training.

$$\mathrm{L}_c = \mathrm{E}\left(\sum_{i=1}^{23} \sum_{j=1}^{16} \delta_{\left(u_{2_i}, v_{2_i}\right),\left(u_{1_j}, v_{1_j}\right)} * \left| \mathbf{\Lambda}^{s2}_{u_{2_i}, v_{2_i}} - \mathbf{\Lambda}^{s1}_{u_{1_j}, v_{1_j}} \right| \right) \tag{4.5}$$

$$\mathrm{L} = \mathrm{L}_r + \alpha \mathrm{L}_c \tag{4.6}$$

4.4 Super-resolution model for angular resolution enhancement

4.4.1 Deep convolutional neural networks

Deep models are composed of many convolutional layers. Each layer has a certain number of convolutional kernels that slide on the input image data or feature maps with a given stride and use different weights to convert the input into a new feature map. Figure 4.3 illustrates a convolutional operation using one 2D kernel, where the stride size is (1×1), and the padding sizes in (a) and (b) are 0 and (1×1) respectively. The input data can be either images captured by sensors or feature maps output by the previous convolutional layer. The input has three dimensions, namely width w_{in}, height h_{in}, and the number of channels c_{in}. Generally, an RGB image has three colour channels, and a grey image only has one channel. One 2D kernel totally has $w_k \times h_k$ weights and transforms the input to a single channel output.

It is obvious that the shape (w_{out}, h_{out}) of the output is smaller than the input after convolution when no extra padding is used. To keep the output and input having the same shape, a padding operation is usually used in convolutional neural networks. As shown in Figure 4.3 (b), the yellow squares are zero pixels padded to the original input. Then, under the same convolutional operation as shown in Figure 4.3 (a), the output can have the same 2D dimension as the input. The relationship between w_{in} and w_{out} can be formulated as

$$\begin{aligned} w_{out} &= \frac{w_{in} - w_k + 2p_w}{s_w} + 1 \\ h_{out} &= \frac{h_{in} - h_k + 2p_h}{s_h} + 1 \end{aligned} \tag{4.7}$$

where (s_w, s_h) is the stride size and (p_w, p_h) is the padding size. In 2D convolution, the channel count of output is determined by the amount of kernels per layer.

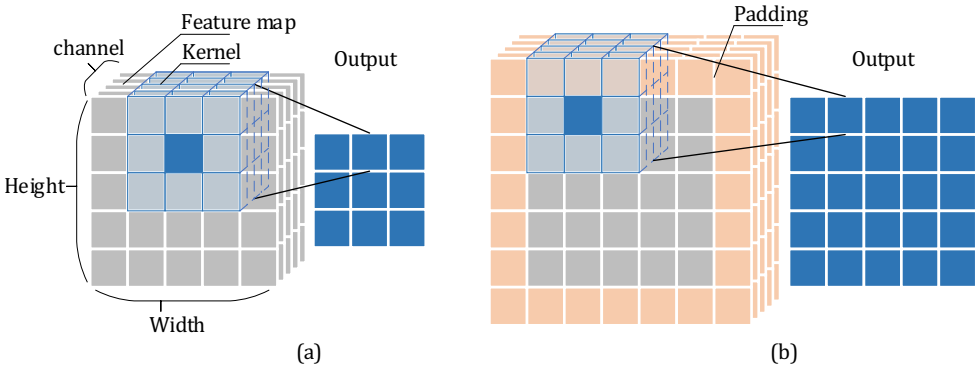

Figure 4.3 Convolution with 1×1 stride. Each square denotes a pixel point. (a) Using no padding. (b) Using zero-padding.

A pooling layer is a sampling process, following a convolutional layer as usual.

The pooling operation down-samples the contents in a region based on some rules including keeping the maximum value, stochastic selection, and computing average values. In general, pooling layers are used to achieve deformation invariance so that the deformation including shifts, rotation, etc. cannot affect the performance of learning models. However, a recent study (Ruderman et al., 2018) claimed that pooling is not necessary for a CNN, and deformation invariance is mainly realized via adequate training. However, pooling operation is still an effective method for dimension reduction to filter redundant information.

Activation functions are a key component for CNNs since the main non-linear transformation is achieved by them. Activation functions play a similar role as a switch in an electrical circuit, deciding which values should be passed to the next layers. In a CNN, an activation function node often follows a pooling layer or is directly placed after a convolutional layer. Commonly used activation functions include Logistic (Sigmoid and Softmax), TanH, Rectified linear unit, exceptional linear unit (ELU), etc. The logistic functions are usually used as the final activation node of classification networks. The Sigmoid function is mostly used for binary classification and the Softmax function is usually used for multi-classification problems. Since ReLU was proposed (Glorot et al., 2011), it has replaced TanH, becoming one of the most popular activation functions. Some ReLU-like functions such as leaky ReLU and PReLU (Parametric ReLU) performed better in some works. However, ReLU is still the first choice for the implementation of a deep model because other ReLU-like functions introduce more computational complexity to the learning models.

4.4.2 Angular super-resolution through motion estimation

A straightforward learning model inspired by plane sweep volume and motion estimation is developed to evaluate the proposed semi-supervised learning paradigm. In this model, the SAIs are assumed to be a group of scenes captured by a camera moving along the u and v axes. The difference of the u and v coordinates of two neighbouring SAIs causes a x-direction and a y-direction translation motion of the corresponding points in the two SAIs. The value of the translation motion of the corresponding points is different and depends on the disparity. Instead of direct estimation of the disparity pixel-wisely, the model directly predicts the motion value of small local regions of two views and synthesizes the new pixels in the novel view via fusing the two shifted local regions.

To improve the learning efficiency, the prediction problem is converted into a classification problem. Based on Eq. (4.2), the shifting of the local regions will reduce the ε so that the corresponding points in the different views are moved to one convolutional window, and pixel-level fusion can be achieved by the convolutional layers.

To simulate the motion, a series of translation SAI pairs with specific translation values are generated and stacked, forming a pyramid-like structure which is called a Motion Pyramid in this paper for a vivid demonstration. The layers in the Motion Pyramid are called *Levels* numbered with the specific translation value. The proposed method is shown in Figure 4.4, where a neighbouring SAI pair (*Level 0*) in the horizontal (u) direction is used for a demonstration. *Level* l is defined as all the pixel points in the left SAI moving along the x axis by $-l$ pixels and all the right SAI's points

moving by l pixels, where $l \in [-L, L]$ and $L > 0$. These *Levels* are inputted into a motion estimation network individually. The motion estimation network can output a confidence map. Every value in the map z_l at coordinates (x^l, y^l) represents the confidence of each point at (x^l, y^l) in the inputted *Level* l. The higher confidence indicates that the desired interpolated point at (x_m, y_m) of the novel-view SAI is more possibly acquired from this *Level*.

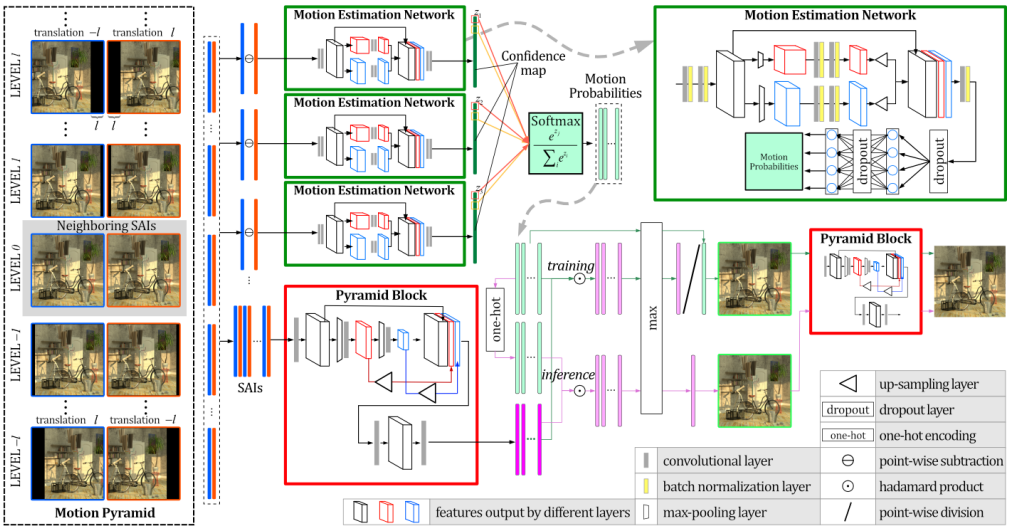

Figure 4.4 Framework of the proposed learning model based on motion estimation.

4.4.3 Motion estimation network

Before processing, the SAIs at different *Level*s are randomly shuffled first to improve the robustness of the estimation. In addition, the two SAIs at *Level* l first go through a point-wise subtraction. This is reasonable since the corresponding points are recorded from the same object point so that these points should have similar pixel intensity values. The point-wise subtraction operation can improve the learning

efficiency. The result from subtraction undergoes processing through a series of convolutional layers, succeeded by batch normalization. The output feature (in black) is shrunk by two max-pooling layers with different scales p_{d1} and p_{d2} simultaneously. The two low-dimension features are up-sampled by a bilinear process to the same size as the inputs for feature fusion. The three features (black, red, and blue) are then merged and concatenated into one global pyramid feature for the subsequent processing. The architecture is adapted from a pyramid pooling module that has been proved to be an effective global contextual prior (Zhao et al., 2017). It is noted that the number of channels vary among the three features during the concatenation process where the two low-dimension features (red and blue) only have half the channel count of the high-dimensional feature map (black).

Since the motion estimation is intended to perform for a local region, the pooling scales p_{d1} and p_{d2} determine the sizes of the local regions. It is noted that when $p_{d*}=1$, the estimation is equivalent to the depth estimation for each pixel. In the experiments, it is found that a large p_{d*} usually causes break areas on the borders of the local regions. Hence, p_{d1} and p_{d2} are set to 2 and 4 for multi-level feature fusion.

The final output of the motion confidence map is produced by two convolutional layers with 1×1 kernels to replace the operation of fully-connected neurons. This could reduce the number of weights imported by the fully-connected layers. Two dropout layers are used in the training process to avoid overfitting. The confidence maps generated by the motion estimation network are normalized in the channel direction using the SoftMax function. As a result, a motion probability map of each *Level* l is acquired. The entire motion estimation process can be formulated as Eq. (4.8) and Eq.

(4.9).

$$\tilde{\Lambda}_{u,vl}=\begin{bmatrix}\Lambda_{u,vl}^{-L}\\ \vdots\\ \Lambda_{u,vl}^{0}\\ \vdots\\ \Lambda_{u,vl}^{L}\end{bmatrix},\tilde{\Lambda}_{u,vr}=\begin{bmatrix}\Lambda_{u,vr}^{L}\\ \vdots\\ \Lambda_{u,vr}^{0}\\ \vdots\\ \Lambda_{u,vr}^{-L}\end{bmatrix},\boldsymbol{LMat}=\begin{bmatrix}\Lambda_{u,vl}^{-L} & \Lambda_{u,vr}^{L}\\ \vdots & \vdots\\ \Lambda_{u,vl}^{0} & \Lambda_{u,vr}^{0}\\ \vdots & \vdots\\ \Lambda_{u,vl}^{L} & \Lambda_{u,vr}^{-L}\end{bmatrix} \tag{4.8}$$

$$\boldsymbol{P}^{(j)}\left(x,y\right)=\frac{e^{f_P\left(\boldsymbol{LMat}^{(j)}\right)(x,y)}}{\sum_{i=1}^{2L/l_{in}+1}e^{f_P\left(\boldsymbol{LMat}^{(i)}\right)(x,y)}} \tag{4.9}$$

where $\boldsymbol{P}^{(j)}$ is the output motion probability map corresponding to the *Level* $\left(j\times l_{in}-L\right)$, i.e., $\boldsymbol{LMat}^{(j)}$ and l_{in} is the difference of the translation values between two adjacent *Levels*. $f_P\left(\cdot\right)$ denotes the mapping function of the motion estimation network. $\Lambda_{u,vl}$ and $\Lambda_{u,vr}$ are the neighbouring SAIs in the horizontal direction. Λ_*^k denotes that all the pixel points of a SAI Λ_* are translated by k pixels.

4.4.4 Novel-view reconstruction network

The next stage is the reconstruction of the novel-view SAI $\Lambda_{u,vm}$ between $\Lambda_{u,vl}$ and $\Lambda_{u,vr}$. In this stage, the two SAIs at each *Level* are inputted into a pyramid block. The block compresses the two SAIs into one single image, formulated as

$$\Lambda_{u,vm}^{l}=f_r\left(\Lambda_{u,vl}^{-l},\Lambda_{u,vr}^{l}\right) \tag{4.10}$$

$$\tilde{\Lambda}_{u,vm}=\begin{bmatrix}\Lambda_{u,vm}^{-L} & \cdots & \Lambda_{u,vm}^{0} & \cdots & \Lambda_{u,vm}^{L}\end{bmatrix}^T \tag{4.11}$$

where $f_r(\cdot)$ denotes the mapping function of the pyramid block which consists of two max-pooling layers and two up-sampling layers. This is a non-depth-based fusion of the two SAIs only based on pixel information. Similar to the motion estimation network, multi-dimensional and multi-level features extracted from the SAIs are concatenated and merged for the effective recognition of the multi-level feature information. Batch normalization is not used in the pyramid block since research has found that batch normalization can deteriorate the accuracy of image restoration (Fan et al., 2018). There are two different routes for the training and inference processes. As for the training process, a Hadamard product is directly conducted between the output $\Lambda_{u,vm}^{l}$ and its corresponding probability map $\boldsymbol{P}^{\left(\frac{l+L}{l_{in}}\right)}$. The maximum operation is then conducted at every point on $\tilde{\boldsymbol{\Lambda}}_{u,vm}$ to filter the points with fewer probabilities so that a single image $\hat{\Lambda}_{u,vm}^{max}$ is acquired, formulated as

$$\hat{\Lambda}_{u,vm}^{max}(x,y) = \max\left[\tilde{\Lambda}_{u,vm}(x,y) \odot \boldsymbol{P}(x,y)\right] \tag{4.12}$$

where $\odot$ is the Hadamard product. To restore the pixel intensities, the same maximum operation is also performed on the group of probability maps to acquire a maximum probability map $\boldsymbol{P}^{max}$.

$$P^{max}(x,y) = \max \boldsymbol{P}(x,y) \tag{4.13}$$

The novel view is obtained through dividing $\Lambda_{u,vm}^{max}$ by the maximum probability map pixel-wisely as shown in Eq. (4.14).

$$\hat{\Lambda}_{u,vm} = \frac{\hat{\Lambda}_{u,vm}^{max}}{P^{max}} \tag{4.14}$$

Since the values of the confidence map $f_P\left(\boldsymbol{LMat}^{(j)}\right)$ are arbitrary without being constrained by Sigmoid or other activation functions, the model tends to output very large values in the confidence map to achieve a high probability after the SoftMax function. Eq. (4.14) can help to avoid this instability during the training process since the maximum probability need not be close to 1.0 after this operation.

As for the inference process, the probability maps are encoded using the one-hot strategy so that only the maximum probability remains as 1 at each point (x, y) and the other probabilities are set to 0. Hence, the pixel intensities of the novel-view SAIs in the inference process are formulated as

$$\Lambda_{u,vm}(x,y) = \max\left[\tilde{\Lambda}_{u,vm}(x,y) \odot f_{o-h}\left(\boldsymbol{P}(x,y)\right)\right] \tag{4.15}$$

where $f_{o-h}(\cdot)$ denotes the one-hot encoding strategy. The produced SAIs both in the training and the inference processes are further refined by another pyramid block $f_l(\cdot)$ which generates the final output of the interpolation.

$$\Lambda_{u,vm}^{final}(x,y) = f_l\left(\Lambda_{u,vm}(x,y)\right) \tag{4.16}$$

In terms of the training of the model under the proposed learning paradigm, an additional L2 regularization is used to constrain the weights of the networks to prevent overflowing of the SoftMax. Hence, the integral loss function is formulated as Eq. (4.17), where θ_{f_p}, θ_{f_r}, and θ_{f_l} are the weights of the motion estimation network and the two pyramid network blocks. λ is the penalty coefficients of the L2 regularization.

$$\mathrm{L} = \mathrm{L}_r + \alpha \mathrm{L}_c + \lambda\left(\sum \theta_{f_p}^2 + \sum \theta_{f_r}^2 + \sum \theta_{f_l}^2\right) \tag{4.17}$$

4.4.5 Optimal maximum translation value

A rapid and simple method to identify the optimal maximum translation value is salient point matching. It is possible to detect and describe multiple salient points from the adjacent views based on mature feature detectors such as the SIFT method and the speeded up robust feature (SURF) method so that the point distance can be measured. The matching can provide a motion range for the motion estimation. In this paper, the SIFT detector and descriptor are used for the key point detection and matching. To cater for the vertical and the horizontal motions, the adjacent views of an input in the two directions are used for the matching. The point distances can be acquired based on the pixel coordinates of the matched points.

4.5 Experiments on LF and autostereoscopic measurement datasets

4.5.1 Training datasets

Similar to other light field super-resolution research, the proposed work uses the *HCI* dataset as the training dataset including 24 scenes. All the 24 scenes can be used for the training of the proposed method following the semi-supervised learning paradigm, because of no ground truth being required during training. The dataset is split for training and testing. The training set includes 20 scenes, and the test set contains 4 scenes (*bedroom*, *bicycle*, *dishes*, and *herbs*). To further simulate the real-world light field images which only have a limited angular resolution, the angular resolution of the *HCI* dataset is down-sampled from 9×9 to 5×5, and the original angular resolution is assumed to be 5×5.

As a result, only a total of 500 images are used for the training of the proposed learning model and the remaining 1,120 images are just dropped without being used. Compared with other approaches that used the 20 scenes of 9×9 light field images for training, the proposed learning paradigm saves over 69% of the training data. It is noted that the proposed method is not further trained and finetuned on other datasets for the evaluation.

4.5.2 Test datasets

Real-world LF datasets established with various recording devices are used for the real-world image evaluation. One dataset named *Lytro 1st* (Mousnier et al., 2015) is composed of 30 scenes recorded by the 1st-generation commercial LF camera, where the baseline between multiple perspectives is minor.

Another dataset named *SFLA-Lego* (*Stanford Light Field Archive: Light fields from the Lego gantry*) (Adams, 2008) is more challenging. Since the data were recorded by a self-built light field recording system which contains a camera on a moving platform, a large depth of field from a single perspective is achievable so that multiple objects from various distances can be recorded with sharp edges and clear details. In addition, the baseline is generally much larger and more flexible based on the movement of the platform. The illumination conditions and patterns of target objects are also more complicated in the *SLFA* dataset.

To collect testing autostereoscopic data, an autostereoscopic measuring system is established and its setup is shown in Figure 4.5. The system includes an objective lens and a zoom imaging system that allows for adjusting the magnification factor. Coaxial and ring-type illumination is installed to ensure the visibility of micro-structured

surfaces. The system utilizes a CCD sensor with a resolution of 2456×2058 and a pixel size of 3.45 μm. A MLA is incorporated, with a pitch size of 150 μm, a focal length of 5.6 mm, and a scale of 10 mm × 10 mm. The resolution of the raw data is 15×13×151×151.

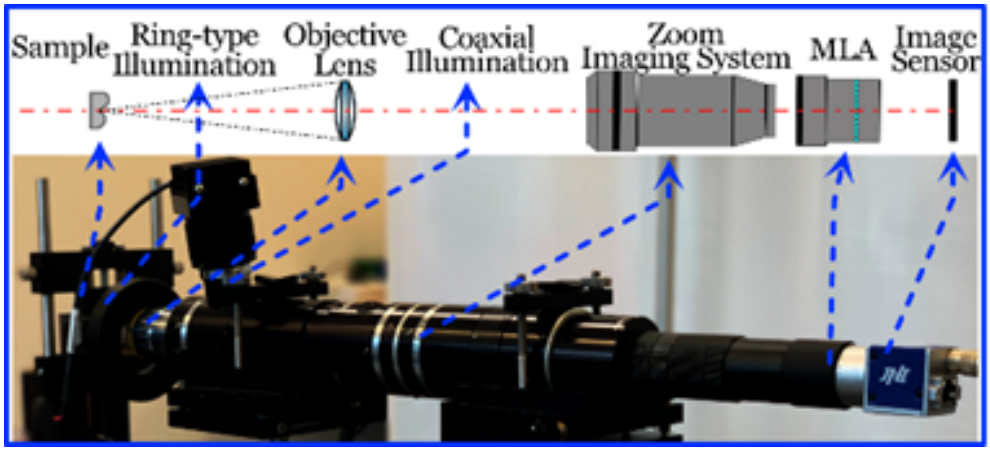

Figure 4.5 Autostereoscopic measuring system for data collection.

4.5.3 Experimental details

The ReLU is set as the non-linear activation function of the proposed network. To achieve higher accuracy of the view reconstruction, the penalty coefficient μ is set to 0.1 to constrain the contribution of the feature perceptual loss during training. Following previous research (Cheng et al., 2020; Rahim & Nadeem, 2018), the regularization parameter λ is set to $1e-4$ to avoid overfitting. A large penalty coefficient α usually resulted in difficulty in convergence during the beginning of the training in the experiments. Hence, α is determined through the grid search technique and finally set to $1e-3$. The Adam algorithm is used for weight optimization with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The input images are converted to the YCbCr space. Training is solely conducted on the Y channel. In the testing stage, the Cb and Cr

channels are reconstructed by the trained networks. Thorough data augmentation is conducted including flipping, rotation, and resizing with a pre-defined probability. The resizing scale factors are set to 1.0 and 0.5 randomly. The implementation of the learning models is implemented on the Pytorch platform using NVIDIA RTX 2080 GPUs.

During the training process, the loss curve is monitored for hyper-parameter adjustments. By appropriately adjusting the learning rate, the loss curve eventually converges and flattens, indicating no further decline. This can be considered a stop signal for the training. In the experiment, training is halted when the variations in L1 loss are less than 10^{-3} during the last 10 epochs. Checkpoints are also saved at different training epochs to identify the best model with greater generalization capability. The trained model is then evaluated to ensure the training is complete.

4.5.4 Evaluation of the proposed learning paradigm

An ablation study about the improvement made by the proposed semi-supervised learning paradigm is presented, where the baseline method of LFCNN is used and trained under the traditional supervised and the proposed semi-supervised learning paradigm. The training set is the same as stated in section 4.5.1. To only enhance the angular resolution, the spatial super-resolution layers of LFCNN are removed and the other layers are the same as described in Yoon et al. (2017). For clarification, LFCNN is trained in a supervised fashion and LFCNN-semi is the semi-supervised model. The two models both go through 50-epoch training. The learning rate is initialized to 1×10^{-4} initially and decays every 10 epochs. The same data augmentation is used, and the cropping size of the input patches is set to 96×96. To further eliminate the contribution

made by the perceptual loss, only L1 loss is employed in Eq. (4.3) during the proposed semi-supervised process.

Since LFCNN can barely solve the light field data with a large baseline, 5×5-to-9×9 enhancement is performed during the quantitative analysis. Similarly, the four scenes (*bedroom*, *bicycle*, *dishes*, and *herbs*) of the *HCI* dataset are used as the test set. Apart from the *Lytro 1st* dataset, another five real-world datasets including *Reflective* (Raj et al., 2016), *30 scenes* (Kalantari et al., 2016), and three categories (*ISO and colour charts*, *Light*, and *Mirrors and transparency*) of the *EPFL* (*École polytechnique fédérale de Lausanne*) (Rerabek & Ebrahimi, 2016) dataset are used for the real-world evaluation. All of the real-world data are recorded by commercial light field cameras and contain small baselines to cater for the requirements of the baseline method. The quantitative experimental results (PSNR/SSIM) are presented in Table 4.1, where LFCNN outperforms on the *HCI* test set, though LFCNN-semi achieves quite similar results.

Table 4.1 Evaluation of supervised baseline models against the introduced semi-supervised approach.

Datasets	LFCNN	LFCNN-semi
HCI	**29.95/0.855**	29.64/0.822
Lytro 1st	35.31/0.954	**38.07/0.973**
Reflective	36.70/0.950	**38.33/0.961**
30 scenes	36.70/0.958	**37.94/0.967**
EPFL-ISO	34.91/0.926	**36.72/0.941**
EPFL-Light	35.36/0.937	**37.66/0.960**
EPFL-Mirrors	35.10/0.929	**37.08/0.956**

However, regarding the six real-world datasets, LFCNN-semi improves the PSNR results by around 2 dB. The results show that the proposed semi-supervised learning paradigm can enhance the learning efficiency of the angular SR learning model by exploring more LF information among the data instead of splitting the finite data into input-label pairs.

4.5.5 Comparison with SOTA approaches

The proposed method is evaluated against several SOTA techniques, including Wu (2019) (Wu et al., 2019), LFASR (Jin, Hou, Chen, Zeng, et al., 2020; Jin, Hou, Yuan, et al., 2020), and LFASR-FS-GAF (Jin, Hou, Chen, Zeng, et al., 2020). The three models are trained on the same *HCI* training set but utilize the entire dataset consisting of 1,620 images. Only the proposed model is trained using only 500 images under the semi-supervised paradigm.

Table 4.2 Comparison using the synthetic dataset.

	Light field Scenes	Wu (2019)	LFASR	LFASR-FS-GAF	**Proposed**
HCI	*bedroom*	39.15/0.961	**41.98**/**0.975**	41.91/**0.975**	39.74/0.968
	bicycle	30.84/0.924	**34.03**/0.954	33.92/**0.959**	32.20/0.945
	herbs	30.80/0.831	32.76/0.882	**37.53**/**0.985**	34.31/0.945
	dishes	26.59/0.876	29.63/0.938	35.20/0.946	**37.77**/**0.984**
	Avg. over 4 scenes	31.84/0.898	34.60/0.937	**37.139**/**0.966**	36.01/0.961

The quantitative evaluation (PSNR/SSIM) using the *HCI* test set is shown in Table 4.2 where part of the results are acquirable in Jin, Hou, Yuan, et al. (2020). As the results demonstrate, LFASR-FS-GAF achieved relatively higher PSNR and SSIM, but

the proposed method obtained a higher PSNR in the *dishes* scene. The results are consistent with our previous assumption that the depth estimation for the regions containing complex patterns is challenging.

The quantitative comparison results (PSNR/SSIM) using the *Lytro 1st* dataset are presented in Table 4.3. The dataset consists of 30 scenes captured by a commercial light field camera in various indoor and outdoor environments. The SAIs are decoded from the raw lenslet images based on Dansereau et al. (2013). The results show that the developed model enhanced the metrics for the 30 scenes by around 1.5 dB on average. Note that the baselines of this dataset are quite small, so every model can produce quite satisfactory results. No finetuning is performed to these models to show the representation capabilities of different models for both the simulation and realistic light fields.

Table 4.3 Comparison using the *Lytro 1st* dataset.

Light field Scenes	Wu (2019)	LFASR	LFASR-FS-GAF	**Proposed**
Beers	32.45/0.961	34.57/0.966	33.46/0.959	**37.35/0.985**
BSNMom	33.19/0.951	27.98/0.868	30.52/0.912	**37.90/0.983**
Edelweiss	32.53/0.966	34.24/0.975	32.57/0.966	**36.27/0.988**
Street	33.55/0.970	35.04/0.976	33.62/0.973	**37.24/0.986**
Avg. over 30 scenes	36.68/0.970	38.28/0.970	36.85/0.967	**39.76/0.984**

The evaluation (PSNR/SSIM) for *SLFA-Lego* is presented in Table 4.4. Since the spatial resolution of the data is large and different in multiple scenes, the data for evaluation are resized and centrally cropped to 512×512 which is the same as the size of the *HCI* data. Since larger baselines are contained in some scenes, depth estimation

is necessary.

Table 4.4 Comparison using the *SLFA-Lego* dataset.

Light field Scenes	Wu (2019)	LFASR	LFASR-FS-GAF	**Proposed**
Bracelet	22.63/0.792	29.39/0.933	23.12/0.831	**36.17**/**0.982**
Jelly Beans	31.41/0.937	37.10/0.968	29.98/0.928	**41.60**/**0.977**
Lego Bulldozer	25.78/0.854	32.10/0.938	25.23/0.921	**35.16**/**0.974**
Lego Gantry Self Portrait	21.78/0.813	22.68/0.827	20.94/0.793	**24.83**/**0.919**
Avg. over 13 scenes	30.87/0.882	34.14/0.918	32.06/0.897	**34.88**/**0.960**

Regarding the evaluation using autostereoscopic data, LFASR, LFASR-FS-GAF, and the proposed method are compared. All the models are finetuned using the collected autostereoscopic dataset. Figure 4.6 displays the synthetic outputs of four-prism structures, frustums, wire bonding, and pyramid structures that are at a scale of a hundred micrometres. The first row displays the enhancement from low-resolution to high-resolution SAIs. The angular resolution of the system is improved from 15×13 to 29×25. The details are highlighted to show the accuracy of new-view synthesis. The results reveal that the LF-oriented methods are ineffective in generating novel views for autostereoscopic data. Compared with SSRAMS, the method produces sharper edges.

Another experiment on wire bonding structures is shown in Figure 4.7, where (a) exhibits the low-angular-resolution (low AR) data recorded by the measuring system, and (b) shows the high-angular-resolution (high AR) data. Columns (c) and (d) are the outcomes of LFASR and the proposed, respectively, where the four corners are the input and the other views are reconstructed for super-resolution. It is noted that LFASR

fails to produce clear edges and more artefacts occur in the novel views. This could have resulted from the depth estimation process which requires clearer details and textures to predict the depth for the novel-view reconstruction.

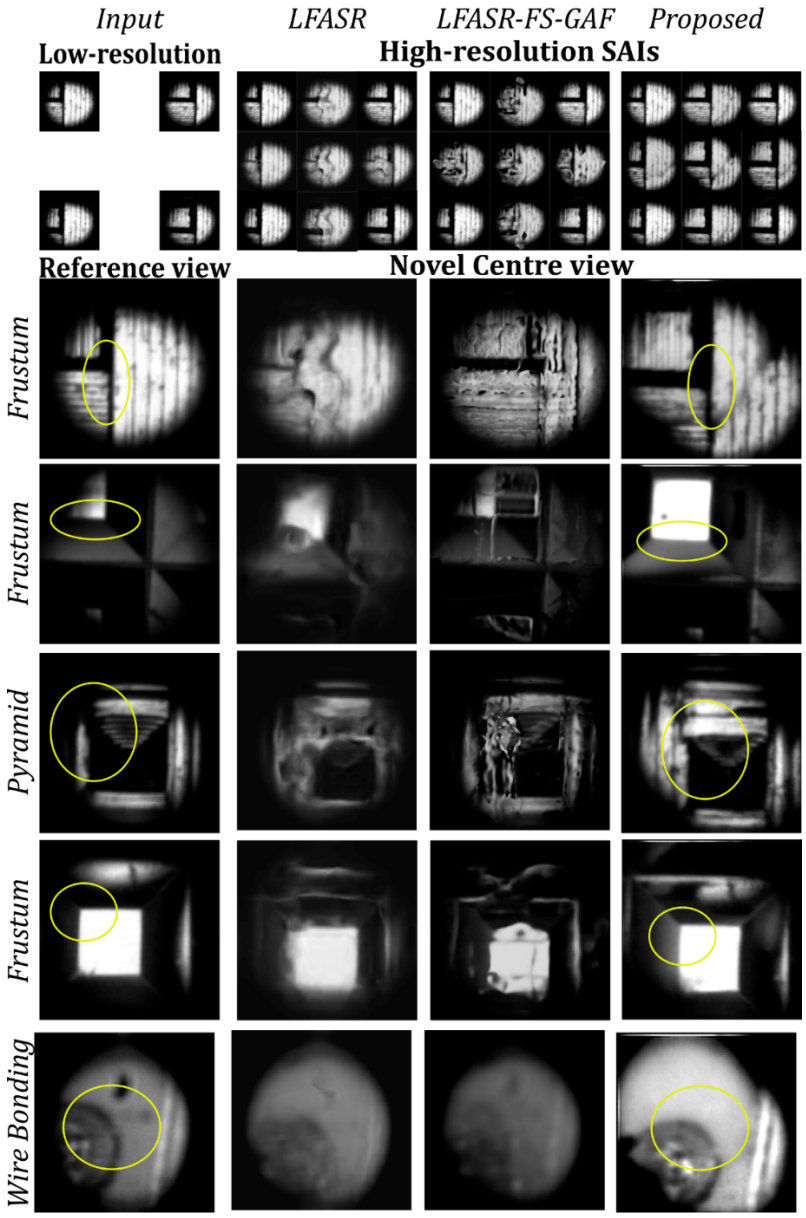

Figure 4.6 Evaluation using autostereoscopic measurement data.

Both the raw low-resolution data and the high-resolution data are processed via digital refocusing so that multiple refocused images at different depth planes are obtained for focus detection. The Laplacian filter is used for focus detection so that the

depth of a region of interest (ROI) can be determined. The detection results are shown in Figure 4.8. The three ROIs coloured red, green, and blue are focused at the bottom, the middle, and the top. It can be found that the checkerboard artefacts are eliminated by more perspectives of SAIs after super-resolution. The detection results retain more edge information and less high-frequency noise.

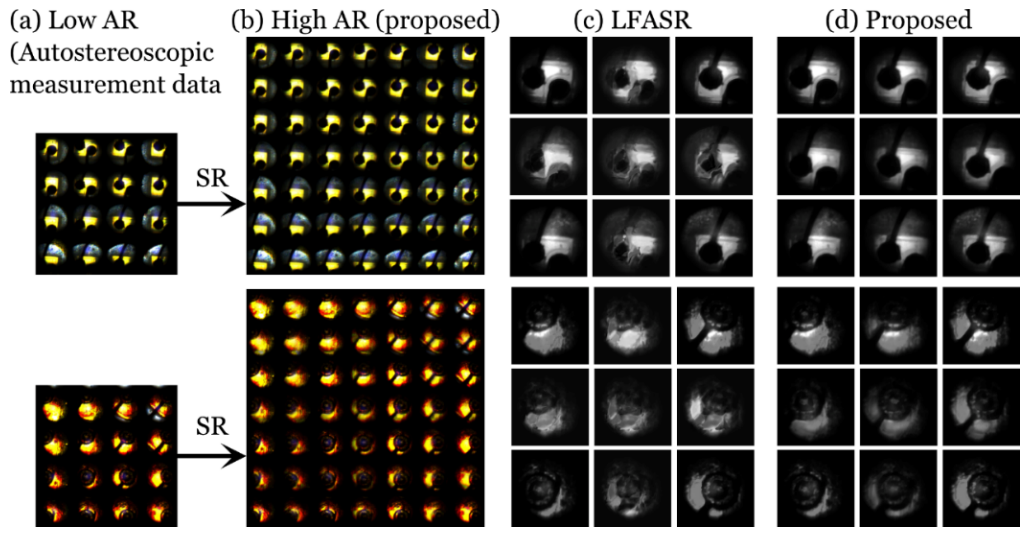

Figure 4.7 Qualitative comparison using the autostereoscopic measurement data (wire bonding).

4.6 Summary

In this chapter, a novel semi-supervised learning paradigm with no requirement of ground truth is presented for the training of angular resolution enhancement methods. A motion estimation model based on plane sweep volume is developed to cater for the behaviour of the proposed learning paradigm for performance evaluation. The improvement of the proposed semi-supervised learning paradigm is evaluated by supervising and semi-supervising a baseline method of LFCNN. The PSNR of the novel views produced by the semi-supervised LFCNN is enhanced by 2 dB compared with

the same model trained under traditional supervision. The evaluation of the motion estimation model under semi-supervised training is performed using both LF and autostereoscopic measurement datasets.

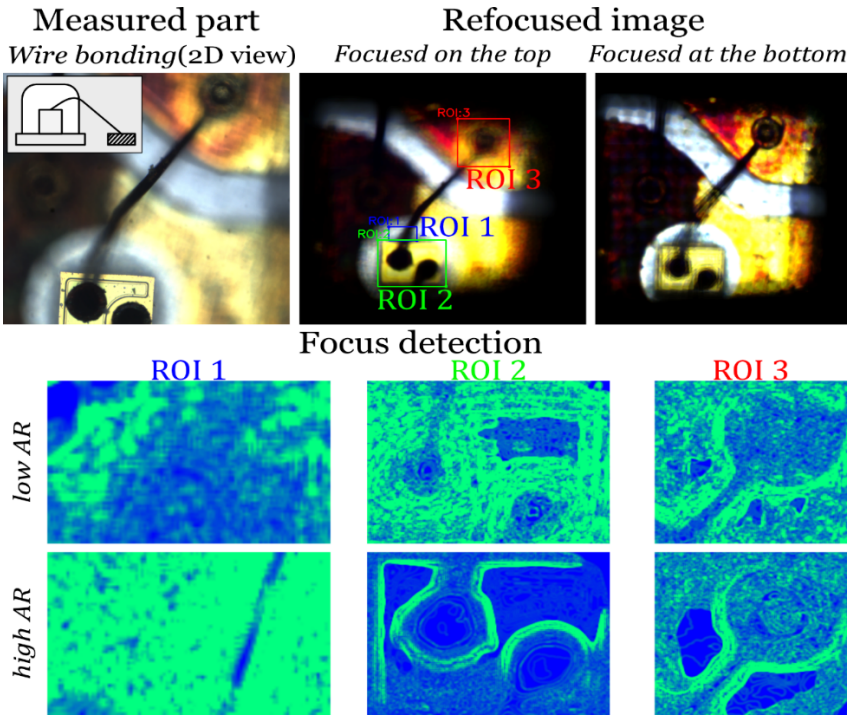

Figure 4.8 Focus detection of the digital refocusing results from the autostereoscopic measurement data.

The proposed model is trained only using 30% of data and compared with the other methods trained using the full data. For the scenes with large baselines, multi-depth objects, and complex textures, the PSNR achieved by the proposed method is improved by over 3 dB on average, whereas the SOTA methods always produced blurred images and artefacts in these scenes. The evaluation results using a real-world dataset with small baselines are also improved by around 1.5 dB on average compared with the SOTA methods. The experiments confirm the effectiveness of the proposed approach for the sparse LF data recorded by self-built devices, since the baselines can be much larger and the target textures can be more complex under various recording conditions.

In addition, the proposed learning paradigm is more data-efficient so that thorough training can be realized using a limited dataset. Experiments also indicate that the proposed learning paradigm is flexible and can be reformed for the training of other light field super-resolution models to improve the reconstruction performance.

As a way forward, there may be some room to further increase the precision of the detection of the maximum motion value, based on the SIFT matching in the paper. This detection can be integrated into the neural network to be performed more precisely in an end-to-end fashion. It is also anticipated that the powerful depth-based angular resolution models are benefited by the semi-supervised learning paradigm.

Hence, the proposed semi-supervised method proves its efficiency in angular super-resolution for LF images as well as autostereoscopic measurement data. Exploiting super-resolution enhancements, a high-angular-resolution autostereoscopic measuring system can be established, resulting in a more accurate 3D reconstruction of the target surface.

Chapter 5 Self super-resolution autostereoscopic measuring system using deep learning

5.1 Introduction

The manufacture of micro-structured freeform surfaces has become a key research issue with the rapidly increasing commercial demand for them. The applications of these complicated surfaces have been used in various industries, including biology and bionics science, space and astronautics science, advanced electronic products, etc. (Hornbuckle et al., 2020; Park et al., 2018). The research community has shown significant interest in the rapid and accurate measurement of products that possess micro-structured features. Coordinate measuring machines (CMMs) are widely used for complex structure measurement due to their stability and high accuracy (Mian & Al-Ahmari, 2014). A probing system is used in CMMs to scan the measured surfaces so that the surface geometry can be reconstructed.

Since a contact stylus could result in damage to the measured parts, optical scanners are being developed to perform non-contact measurement (Gapinski et al., 2014). However, a fact that cannot be neglected is the low efficiency of CMMs. The amount of sampling points acquired by the stylus or scanners directly affects the measurement performance. A large number of sampling points also contribute to a time-consuming measurement process. In addition, the risk of probe damage increases with a higher scanning speed (Bastas, 2020). To overcome the drawbacks, the autostereoscopic imaging technology is an alternative novel solution to realize fast and highly accurate 3D surface profile measurement.

Autostereoscopic three-dimensional imaging technology can obtain 3D

information from multiple view-angle elemental images that are captured in one snapshot. The measured surfaces can be reconstructed using the disparity information stored in the EIs and digital refocusing techniques. Compared with other non-contact measuring systems, the autostereoscopic 3D measurement system is relatively easy to implement, requires less restrictive machining conditions, and is able to record richer 3D information for disparity extraction. Nevertheless, an inherent conflict in autostereoscopic technology arises from the trade-off in the resolution.

The spatial resolution refers to the FOV of each individual EI, while the angular resolution indicates the number of views from which these EIs are captured. To obtain high-resolution EIs, some conventional methods (Mitra & Veeraraghavan, 2012; Wanner & Goldluecke, 2014) have been developed to super-resolve low-resolution 3D information. However, these research works mostly relied on accurate disparity estimation which imposed difficulty and severe errors. Moreover, the research studies on how to enhance the angular resolution are still receiving relatively little attention.

Machine learning, exceptionally deep CNNs, have witnessed significant advancements these years. These networks have been successfully employed for SR tasks, outperforming conventional methods and delivering superior performance (Dong et al., 2014; Kim et al., 2016). Some deep learning methods (Jin, Hou, Yuan, et al., 2020; Yoon et al., 2017) have also been presented to enhance angular resolution, that is generated via simulation or captured by commercial light field cameras. However, there are few noises and almost perfect illumination conditions in the simulated stereo images. The images recorded by the commercial light field cameras usually have a small baseline. As a result, these models cannot achieve high-quality enhancement of the measurement data with various noises, complex illumination effects, and a large

baseline. The advancement of a super-resolution method for measurement data can enhance the measurement performance of autostereoscopic measuring systems.

In this chapter, a self super-resolution autostereoscopic (SSA) 3D measuring system is presented. The objective of the study is to synthesize novel views between adjacent EIs so that more corresponding points originating from every object point are acquired. With more corresponding points, the matching errors can be reduced during the 3D reconstruction process and the digital refocusing process. As a result, the measuring results are improved due to the more accurate depth estimation. To this end, a self super-resolution algorithm for the EIs recorded by the optical measuring system has been developed using a deep CNN that can enhance the angular resolution of the EIs by nearly four-fold.

This enhancement also resulted in the enhancement of the spatial resolution of the refocused images which were two-fold larger, which contributed to a more delicate structure reconstruction in the axial direction and more accurate measurement by the autostereoscopic measuring system. The measurement results were greatly improved in the bias, standard deviation, and maximum absolute error dimensions compared with the traditional autostereoscopic measuring (TAM) system proposed in Li et al. (2014).

5.2 Autostereoscopic measurement for micro-structured surfaces

Autostereoscopy technology can obtain raw 3D information from one snapshot through embedding a micro-lens array into a traditional imaging system without any hardware aids. It is a rapid optical solution to acquire 3D information of the measured

parts. There are three steps during autostereoscopic 3D measurement, which are the recording process, 3D reconstruction, and disparity information extraction. During the recording process as shown in Figure 5.1, the MLA splits the rays emitted from the main objective lens so that multiple EIs at slightly different angles are recorded. The differences of these EIs indicate the disparities in the series of images. The disparities are directly related to the depth of the different target points on the measured surface, i.e., the desired measuring quantity.

As the whole measurement system is fixed and ensured, the disparities of the EIs are solely correlated to the depth information. The EIs and their disparities are used for the next-step 3D reconstruction process. The 3D reconstruction process is symmetrical with the recording process since optical rays are reversible. By utilizing the disparity information stored in the EIs, the depth can be determined by considering the distance between the MLA and the imaging sensor, as well as the spacing between two precisely focused points across different EIs. By establishing the relationship between image pixels and depth, it becomes possible to obtain 3D reconstruction surfaces and refocused images using the abundant information present in the EIs. The digital refocusing process involves rearranging the pixels layer by layer, and the refocused pixels are produced at different reconstructed planes. As a result, a sequence of 2D images, each with different focus on separate depth levels, is captured, enabling the identification of the precise depth where the image is sharpest.

In accordance with autostereoscopic theory, an essential factor influencing the measurement resolution and accuracy is the pitch size and the quantity of micro-lenses. When the dimensions of a MLA remain unchanged, increasing the pitch size leads to larger dimensions for each individual EI, but a smaller number of EIs. Figure 5.1

demonstrates that the resolution of each individual EI and the total number of EIs are determined jointly by the resolution of the image sensor and the array size of the MLA.

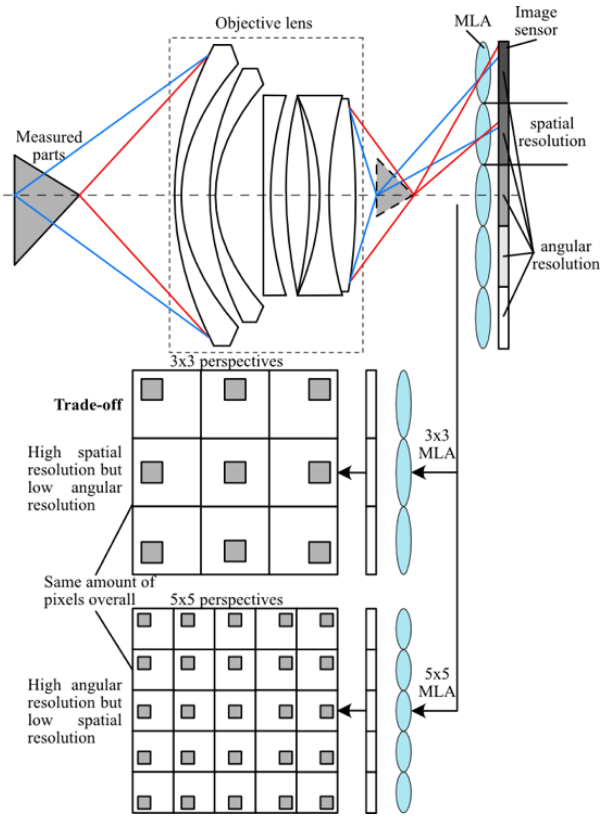

Figure 5.1 Autostereoscopic recording process.

As a result, an increase in spatial resolution leads to a decrease in angular resolution. Conversely, as the number of micro-lenses increases, the spatial resolution decreases. To break through the trade-off by avoiding changing the optical system, a SR algorithm that can enhance the angular resolution without reducing the size of EIs is essential.

5.3 Self super-resolution approach based on deep learning

The proposed self super-resolution approach based on deep CNNs consists of a registration network, a residual encoder–decoder, a refining network, and a discriminator, illustrated in Figure 5.2 and explained in sections 5.3.1, 5.3.2, 5.3.3, and 5.3.4, respectively. To enhance the angular resolution of the measurement data, novel-view EIs are interpolated between adjacent EIs. Taking a 2×2 neighbouring image grid as an example, novel views of the EIs are interpolated in the middle of the horizontal images, the vertical images, and the center of the four images. Hence, the input EIs of the proposed self super-resolution approach can be grouped as horizontal pairs, vertical pairs, and central groups.

In the network framework, the registration network applies affine transformation to the input image pairs. The residual encoder–decoder network extracts features from the registered images and reconstructs the features of the desired novel-view images. Lastly, the residual refining network refines the features and recovers the novel-view images. The generative network integrates three network elements: the registration network, the encoder–decoder, and the refining network. This generative process synthesizes the EIs that are desired to be interpolated in the low-angular-resolution measurement data. The synthetic images, which are the interpolation results, undergo additional constraints imposed by the discriminator network. This network distinguishes the synthetic images from real data. The discriminative results are provided as feedback to the generative network, enabling it to generate high-quality synthetic novel-view images that deceive the discriminator. Hence, the angular

resolution enhancement is achieved by high-quality interpolation. The specifics of the network components and training process are discussed in the following sections and two neighbouring horizontal EIs I_l and I_r are taken as a horizontal input pair for demonstration. The vertical input pairs and the central input groups are processed under the same rule.

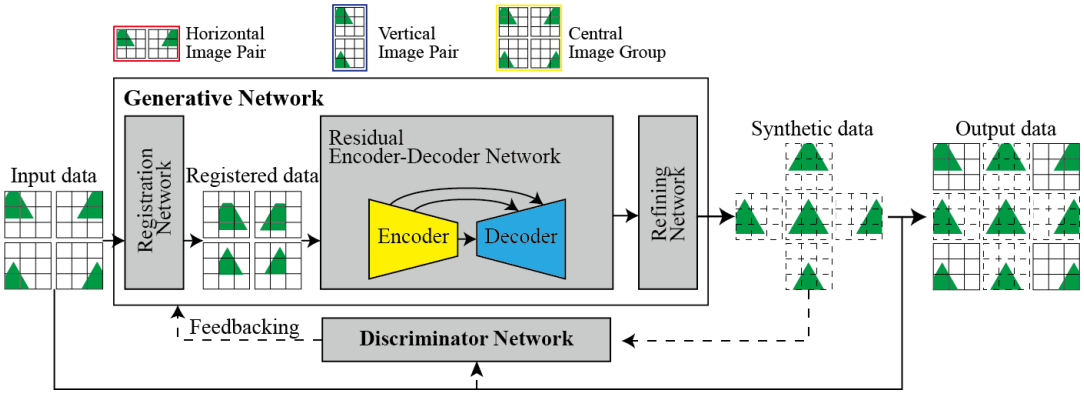

Figure 5.2 Framework of the proposed self super-resolution approach. The approach receives low-angular-resolution measurement data as input and enhances the angular resolution of the data by interpolating synthetic views. The final output high-resolution data are composed of the original measurement data and the synthetic data.

5.3.1 Registration network

In the autostereoscopic system, the three-dimensional information is reconstructed using the corresponding points in every EI. The axial dimension of the measured parts can be determined by calculating the disparity difference between two object points located in different depth planes. Δd is the disparity of a point which can be determined through $\Delta d = gu / D$. g is the distance from the MLA to the image sensor, u is the

baseline distance between two adjacent micro-lenses, and D is the shooting distance. The disparity difference between two object points located on the top surface and the bottom surface can be clearly expressed as

$$\left|\Delta d_t - \Delta d_b\right| = \frac{gu\left|D_t - D_b\right|}{D_t D_b}. \tag{5.1}$$

where the disparity and shooting distance of the top surface point are Δd_t and D_t, respectively. Δd_b and D_b are the disparity and shooting distance of the bottom surface point. Since the dimensions of the measured micro-structures are much smaller than the shooting distance (i.e., $\left|D_t - D_b\right| \ll \min\left(D_t, D_b\right)$), $\left|\Delta d_t - \Delta d_b\right|$ is much smaller than $\min\left(\Delta d_t, \Delta d_b\right)$. This reveals that the large baseline contributes less to the desired measuring values, and affine transformation to the adjacent EIs is able to reduce the redundant disparity information. In addition, the direct fusion of two adjacent EIs with the large baseline could result in severe image artefacts. To eliminate the impact of the large baseline, the proposed registration network is employed to align the neighbouring EIs. The registration process can be formulated in Eq. (5.2) where θ_* is the affine parameters. It is obvious that the neighbouring images in each input pair has their own affine parameters θ_x^* and θ_y^*, and the affine parameters are predicted by the registration network $f_R(\cdot)$.

$$\begin{bmatrix}\theta_x^l \\ \theta_y^l\end{bmatrix}, \begin{bmatrix}\theta_x^r \\ \theta_y^r\end{bmatrix} = f_R\left(I_l, I_r\right), \quad I_l' = \begin{bmatrix}1 & 0 & \theta_x^l \\ 0 & 1 & \theta_y^l \\ 0 & 0 & 1\end{bmatrix} I_l, \quad I_r' = \begin{bmatrix}1 & 0 & \theta_x^r \\ 0 & 1 & \theta_y^r \\ 0 & 0 & 1\end{bmatrix} I_r. \tag{5.2}$$

The registration network aims to predict the affine parameters from the input image

pairs, with the process as illustrated in Figure 5.3. The desired input images to be registered are firstly processed by four convolutional blocks and mapped to a feature space. Each block is comprised of two convolutional layers for feature extraction. The mapping from $\boldsymbol{I}$ to the features F happens in each layer. To reduce the feature dimension, a max-pooling layer is employed after each convolutional block. This layer applies a filter that strides on the input and only permits the maximum value to pass through. At the end of the registration network, there are three fully connected layers that flatten the 2D features outputted by the convolutional blocks into 1D neurons. Finally, the affine parameters are predicted by the last fully connected layer. The input image pairs are then registered using the affine parameters so that they can be involved in the same coordinate system.

It is noted that there are three types of input. To maintain consistency with different registration processes, a total of three registration networks are required for horizontal, vertical, and central registration. These registration networks are all in the same architecture but possess their own weights without sharing.

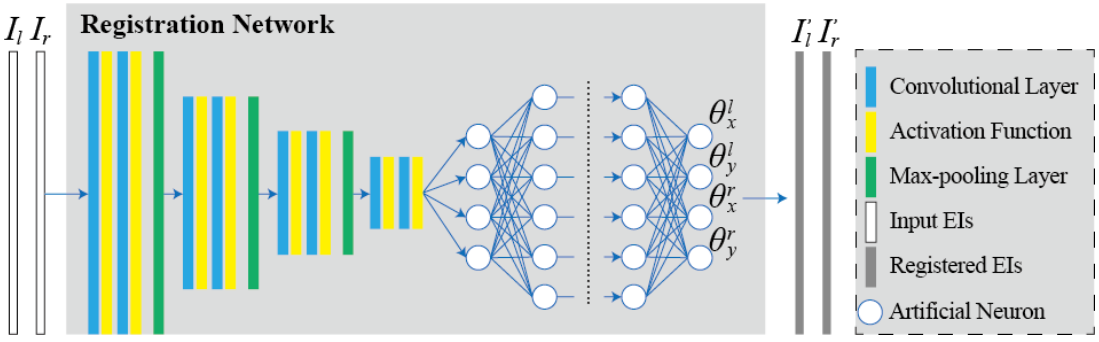

Figure 5.3 Registration network (horizontal). Vertical and central registration networks share an identical architecture but possess different weights.

5.3.2 Residual encoder–decoder network

The main component of the generative network is the residual encoder–decoder network, which is responsible for feature extraction and feature reconstruction. The output features are directly used for the synthesis of the novel view EIs. The framework of the residual encoder–decoder network, as depicted in Figure 5.4, enables feature extraction and view reconstruction. An encoder, comprising of three convolutional blocks, is employed to accomplish the feature extraction.

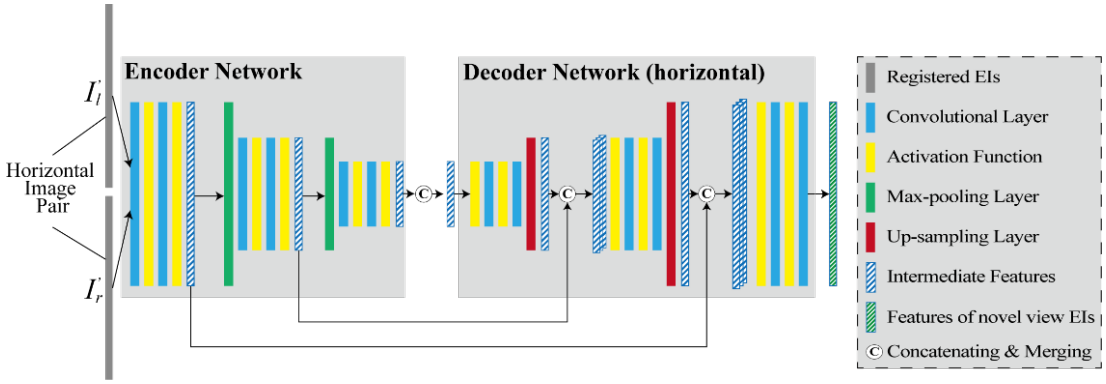

Figure 5.4 Framework of the residual encoder–decoder network. Vertical and central input pairs are processed under the same rule. All the input is processed by the encoder network separately.

Likewise, a max-pooling layer follows the first two blocks to reduce the feature dimensions. Direct convolution on the concatenation of the input EIs from different perspectives delivers image artefacts to the subsequential feature extraction and feature reconstruction, with a demonstration shown in Figure 5.5. To this end, the registered input is processed separately by the encoder and the fusion happens between the local and global features. The fusion on the feature level will reduce the artefacts in the final

output.

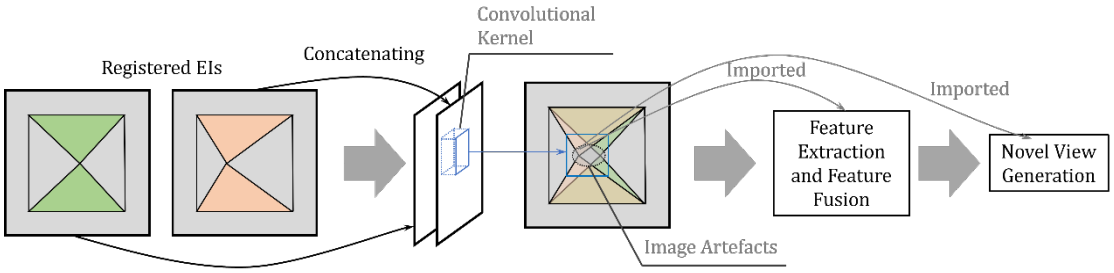

Figure 5.5 Depiction of the effect of separate processing of the encoder. The image artefacts are unavoidable after concatenation and are imported to the subsequential feature extraction, fusion, and novel view generation processes.

Taking the horizontal image pairs for explanatory purposes, the two images are separately processed by the encoder. Hence, two dimension-reduced features corresponding to the two input images are acquired. The two features are concatenated and merged as one feature and input to the horizontal decoder for feature reconstruction. To compensate for the reduction in feature dimensions during encoding, the decoder employs a symmetric architecture with the encoder. It replaces the max-pooling layers with up-sampling layers, which can restore the dimension of the features. It is worth noting that the two image features produced by each convolutional block of the encoder are shared with the decoder. These features are then concatenated with the output of each block of the decoder, resulting in a new feature that encompasses both local and global information. The concatenated feature is then inputted to the next block of the decoder. This forms a residual architecture and can dramatically accelerate the learning efficiency and reconstruction performance of the encoder–decoder network since multi-level features are taken into consideration during the feature reconstruction. With

regard to the three input types, there are also three types of decoders to process horizontal, vertical, and central paired images separately with the same concern of the setup of the registration network.

5.3.3 Refining network

To recover the novel view of EIs from the features reconstructed by different decoders, a refining network is built to refine features and recover images as shown in Figure 5.6, which is also in a residual architecture. It contains three residual convolutional blocks with their details shown in the middle of Figure 5.6. By learning the residual value between its input and output, the residual blocks can prevent gradient vanishing and enhance learning efficiency. Regardless of the input types, there is only one refining network that accepts all the outputs from the three different decoders to perform feature refining and novel view generation. This is helpful to maintain the consistency of the novel views corresponding to different input pairs since all the interpolated novel images are finally generated by the sole network.

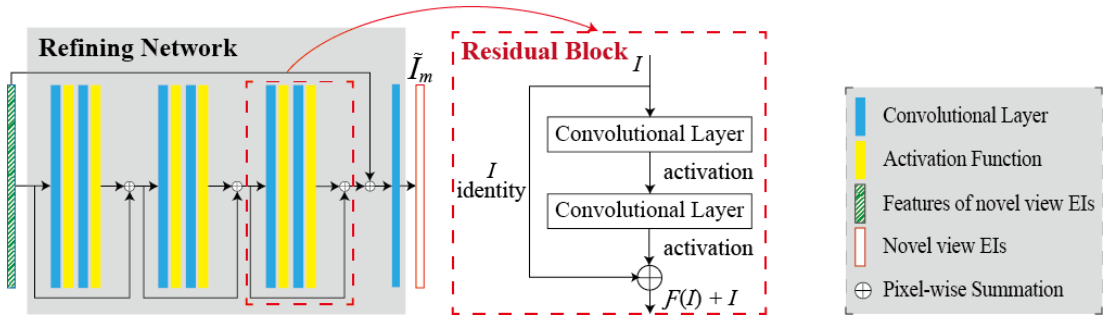

Figure 5.6 Framework of the residual refining network.

5.3.4 Generative adversarial network

The three networks, i.e., the registration network, the encoder–decoder network,

and the refining network form a generative network. The low-angular-resolution measurement data are inputted to the generative network and go through registration, encoding, decoding, and refining. Finally, high-angular-resolution measurement data are output. To further improve the quality of the synthetic interpolation data, a discriminator network is constructed to establish an adversarial relationship with the generative network. The GAN is an unsupervised learning framework, which can learn to generate data that follows a targeted distribution (I. J. Goodfellow et al., 2014).

In this work, the discriminator network is a classifier whose framework is shown in Figure 5.7. It is able to differentiate the real data obtained by the measuring system and the synthetic data interpolated by the generative network. During training, the differentiation results are provided as feedback to the generative network, enabling it to update its weights and generate data of higher quality. Consequently, the generative network has the capability to generate high-quality synthetic novel views that exhibit a similar distribution to the real measurement data, thereby enhancing the angular resolution. Eventually, the discriminator is unable to discern the distinction between real data and synthetic data. This adversarial game between the generative network and the discriminator can prevent the synthetic data far away from the real measurement data.

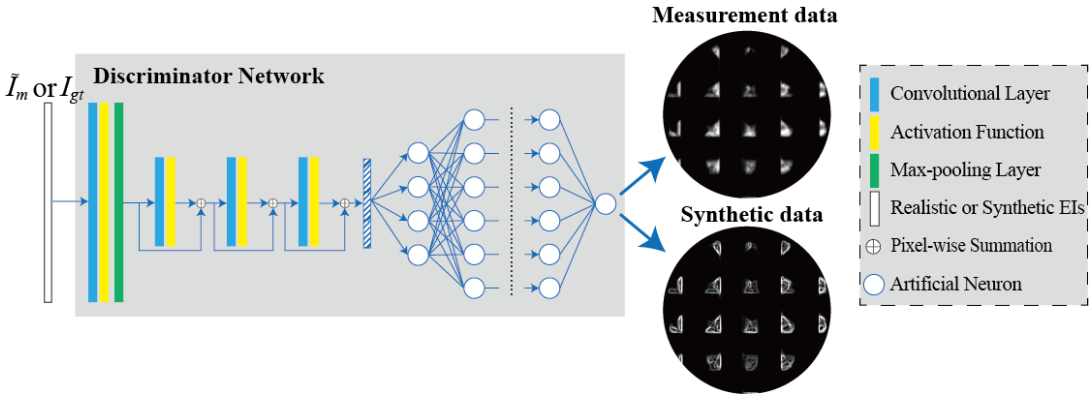

Figure 5.7 Framework of the discriminator network.

5.3.5 Network training

According to the working principle of the proposed self super-resolution approach, the data collected from previous experiments is split into different training pairs. Taking a 3×3 neighbouring grid of the EIs captured in one snapshot in Figure 5.8 as an example, the input data for the learning models consists of images captured at the four right angles, while the remaining views are utilized as ground truth. The four input images can be grouped as horizontal, vertical, and central input pairs. Following the previous discussion, the middle image in the first row can be regarded as the ground truth of the novel view generated from the horizontal pair. Similarly, the middle image in the left column and the central image can be regarded as the ground truth for vertical and central input pairs, respectively.

Hence, the training process aims to minimize the errors between the synthetic novel view images and the ground truth by iteratively updating the weights of the network. It is noted that the input data in the training process are not adjacent, and are different from those inputted in the super-resolution test process. The function of the registration network is to predict the distance between pixels from two EIs caused by the large

baseline $gu / D_m \left(D_t \le D_m \le D_b\right)$. Albeit that only non-adjacent EIs are able to be used for the supervised learning of the proposed model due to the unavailability of the ground truth of the novel views, the baseline between the non-adjacent EIs is still determined by the specifications of the MLA and the shooting distance.

The non-adjacent baseline is $g\left(i \cdot u\right) / D_m \left(D_t \le D_m \le D_b\right)$ where $i \cdot u$ is the centre distance of the two non-adjacent micro-lenses. It is possible for the registration network to predict the baseline with different values of i solely based on pixel information. For a generalization consideration, the input of the proposed algorithm is not only 3×3 neighbouring grids, but also 3×3 non-adjacent grids down-sampled from a $\left(2n+1\right) \times \left(2n+1\right)$ neighbouring grid. By undergoing end-to-end learning with the entire algorithm, the registration network becomes capable of predicting affine parameters using pixel information from the input EIs. This prediction helps eliminate the impact of the large baseline.

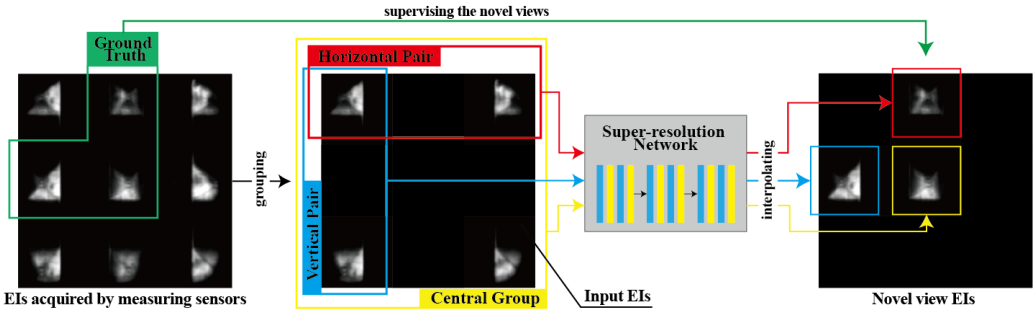

Figure 5.8 The training data and their corresponding ground truth.

To determine the errors, a mean absolute error loss, a perceptual loss, and an adversarial loss are used and incorporated as one composed multiple loss function for the network training. The MAE loss is used to directly compare the pixel error, and

formulated as

$$l_{MAE} = \frac{1}{N}\sum\left(\left|\tilde{I}_m - I_{gt}\right|\right). \tag{5.3}$$

where $\tilde{I}_m$ is the interpolation result , N is the batch size in one iteration during the network training, and I_{gt} is the ground truth corresponding to $\tilde{I}_m$. Perceptual loss was proposed by Johnson et al. (2016), which can compare the style differences of two images through determining the distance between the perceptual features extracted from the images.

Since the MAE loss only compares the difference between pixels, some image properties such as perceptual similarity and image styles, are overlooked. This neglection can contribute to a distortion and low-quality reconstruction of novel-view images. Hence, the perceptual loss serves to monitor the high-level differences, by comparing their features extracted by a fixed pretrained VGGNet. The VGGNet was proposed by the Visual Geometry Group of Oxford University (Simonyan & Zisserman, 2014) and trained on an enormous image recognition dataset. The perceptual loss is formulated as

$$l_{\phi} = \frac{1}{N}\sum\left(\left|\phi\left(\tilde{I}_m\right) - \phi\left(I_{gt}\right)\right|^2\right). \tag{5.4}$$

where $\phi(\cdot)$ denotes the feature extraction performed by the VGGNet whose parameters are frozen during the training process. In terms of the adversarial loss, the distinguishing results of the discriminator are used to constrain the output of the generative network, with the formulation as follows

$$l_a = -\frac{1}{N}\sum D_\varepsilon\left(\tilde{I}_m\right). \tag{5.5}$$

where D_ε is the distinguishing operation performed by the discriminator. The discriminator is trained simultaneously with the generative network through the following training loss

$$l_{dis} = \frac{1}{N}\sum\left(D_\varepsilon\left(\tilde{I}_m\right) - D_\varepsilon\left(I_{gt}\right)\right). \tag{5.6}$$

Hence, the multiple loss function of the generative network is

$$l = l_{MAE} + \mu_\phi l_\phi + \mu_a l_a = \frac{1}{N}\sum\left(\left|\tilde{I}_m - I_{gt}\right| + \mu_\phi\left|\phi\left(\tilde{I}_m\right) - \phi\left(I_{gt}\right)\right|^2 - \mu_a D_\varepsilon\left(\tilde{I}_m\right)\right). \tag{5.7}$$

where μ_ϕ and μ_a are the penalty coefficients for the perceptual loss and the adversarial loss, respectively. The data necessitates no additional processing, enabling seamless adaptation and transfer of the network for evolving datasets or new tasks without complications.

5.4 Depth reconstruction

To reconstruct the target surface, a depth reconstruction algorithm based on disparity patterns is developed. The method consists of four steps, including pixel-point description, matching, depth optimization based on disparity patterns, and coordinate mapping. First, all the pixel points are described using the local region information including greyscale values and gradient values. On the basis of the description, the points in each EI are matched to the central EI in the multi-view EI array separately.

One group of matched points are corresponding points in a determined 3D position in the reconstruction coordinate.

Due to the unavoidable measurement uncertainty and the matching errors, it is impossible for the group of matched points to focus accurately on one point in the 3D coordinate so that only the points in the centre EI are kept for the reconstruction. Finally, an optimization process is utilized to find the accurate depth of the matched points. From each resolvable depth, a group of disparity patterns can be determined based on the working principle of autostereoscopic technology. During the optimization process, the patterns should be the closest to the group of matched points. Finally, the spatial coordinates of the remaining points are mapped to the 3D coordinate and the point cloud can be obtained. In addition, since each EI acquired by the proposed system only contains part information of the target object, a sliding window technique is exploited to use a small image window sliding on the whole EI array with 1 stride. This can assure that in each sliding window the EIs required to be matched contain similar information so that the rate of correct matching can be improved.

5.5 Experiments on micro-structured surfaces

5.5.1 System setup

The SSA system was established as shown in Figure 5.9, where the schematic diagram of the system is shown in (a), the system implementation is shown in (b), and the measured sample is shown in (c). To demonstrate the advancement and improvement over the pioneering research (Li et al., 2014), the same sample was used for the evaluation to control the variables. The sample is a surface with pyramid micro-structures. Each pyramid has two edges named Edge A and Edge B in the lateral

direction and a height in the axial direction. The measured sample was mounted on a three-axis positioning stage for the lateral and longitudinal motion. An illumination device was used in the dark measurement environment. Multiple illumination conditions were performed to construct the dataset. The acquired measuring data were sent to a computing station for super-resolution, digital refocusing, and surface geometry reconstruction. Table 5.1 shows the specifications of the system.

Table 5.1 Specifications of the SSA system.

Item	Specification	
CCD Sensor	Pixel Size	5.86 μm
	Sensor Size	2/3 inch
MLA	Pitch	500 μm
	Focal Length	13.8 mm
	Scale	10×10
Objective Lens System	NA	0.28
	Overall Magnification	35X
	Adjustable Zoom	1-12

To train the proposed self super-resolution approach, a dataset was first built using the data collected from preliminary experiments. The dataset was composed of 80 scenes with various samples including the aforementioned micro-structured surface and 3D complex microstructures. The data were captured under different conditions, including optical system parameters, illumination conditions, exposure conditions, etc. to improve the generalization capability of the approach. The resolution of each scene was 16×9 and 105×105, in spatial and angular terms, respectively. As a result, there were 11,520 EIs in the dataset in total. Among the dataset, 72 scenes were used for the training of the proposed networks and eight scenes were used for testing.

It is noted that the proposed self super-resolution approach can only be trained using the measurement data obtained by the proposed system. This indicates that the trained networks can be easily finetuned on new scenes without modification of the networks and massive consumption of time. As pilot research, this work currently performs a series of measurements on one target surface to verify the proposed measurement method and system. Further research was conducted to validate the effectiveness of different workpieces. The approach was implemented using PyTorch. The multiple loss penalty coefficients μ_ϕ and μ_a were set to 0.01 and 0.01, respectively.

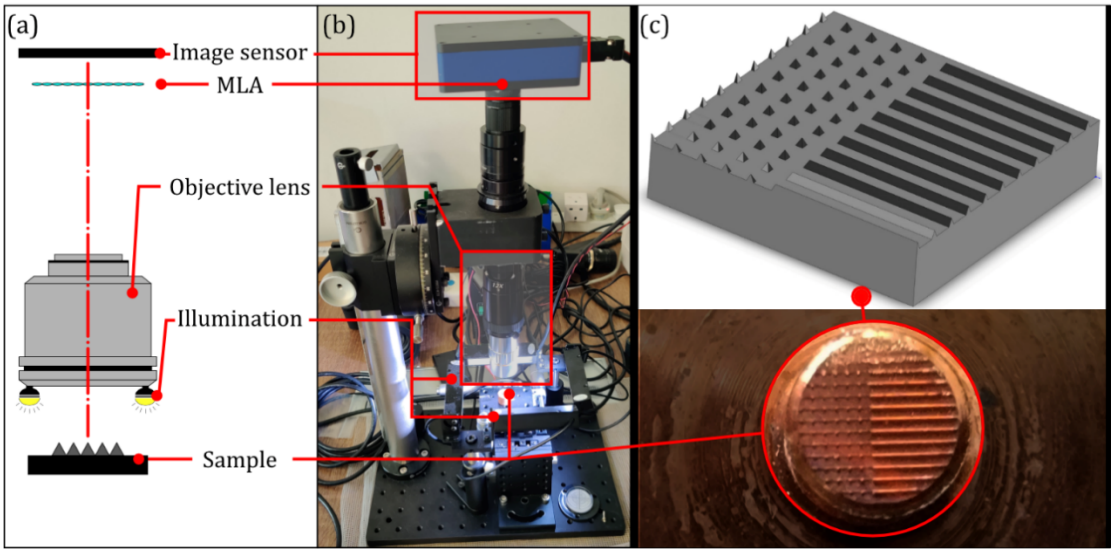

Figure 5.9 Setup of the SSA system. (a) Schematic diagram of the proposed system. (b) System implementation. (c) Measured samples.

The Leaky ReLU and ReLU were taken as the activation functions of the generative network and the discriminator. The Adam optimizer was used as the training optimizer. The computation platform was equipped using a Nvidia GeForce RTX 2080 graphics

card and an Intel Core i7-8700 central processing unit. The training was conducted over approximately 200 epochs, with continuous observation of the loss curve trends. Additionally, checkpoints were saved at various training epochs to identify the model that achieved the highest reconstruction score.

5.5.2 Experimental analysis

All the experimental results were acquired in accordance with the procedure as shown in Figure 5.1. The acquired measuring data, i.e., the Eis, went through a super-resolution process performed by the trained self super-resolution network, and then digital refocusing and reconstruction. The digital refocusing process was based on the method proposed in D. Li et al. (2015) and the reconstruction was conducted using the proposed depth reconstruction approach. As a result, the geometry and height information of the measured surfaces were obtained. The experimental results were compared with the measuring results acquired by the TAM system without super-resolution. Three different comparisons regarding angular resolution, spatial resolution of the refocused images, and 3D reconstruction results are presented in this experimental study to provide powerful illustration of the novelties of the proposed method.

The first comparison took place between the measuring data of the TAM system and the super-resolution measuring data of the proposed SSA system. The angular resolution of the EIs recorded in one snapshot by the TAM system were 16×9. After super-resolution, the angular resolution of those images recorded by the SSA system was expanded to 31×17, nearly a four-fold improvement. The low-angular-resolution images recorded by the TAM system are shown in Figure 5.10 (a). Figure 5.10 (c)

presents the high-resolution EIs produced by the SSA system and a 3×3 local region was enlarged as shown in Figure 5.10 (b) for a vivid comparison, where the low-angular-resolution EIs were bordered in colour and those without coloured borders were the novel views generated by the proposed self super-resolution approach. The comparison indicates that the proposed self super-resolution approach is able to interpolate high-quality novel-view EIs and the angular resolution of the autostereoscopic system is obviously enhanced.

A comparative experiment was conducted at this stage to appraise the effectiveness of the proposed SR approach. A 4D bilinear method was incorporated as a standard interpolation approach, which was taken as a baseline method, while a SOTA deep learning approach was used for comparison, as proposed in Jin, Hou, Yuan, et al., (2020), which has achieved high-quality angular SR for synthetic LF images.

In this experiment, the model of Jin, Hou, Yuan, et al. (2020) was retrained using the measurement data collected by the proposed autostereoscopic system under the same training conditions as the proposed approach. In addition, the pre-trained model of Jin, Hou, Yuan, et al. (2020) which was supervised by a public light field dataset was also finetuned using the measurement dataset for evaluation. The outcomes generated by the baseline method and the SOTA approach (Jin, Hou, Yuan, et al., 2020) are exhibited in Figure 5.11, where the baseline method produced some image artefacts in the novel views, and both the retrained and pre-trained model of Jin, Hou, Yuan, et al. (2020) failed to generate high-quality novel views since inevitable noises, imperfect illumination conditions, and the missing part in the large-baseline EIs limited the depth estimation premise of Jin, Hou, Yuan, et al. (2020).

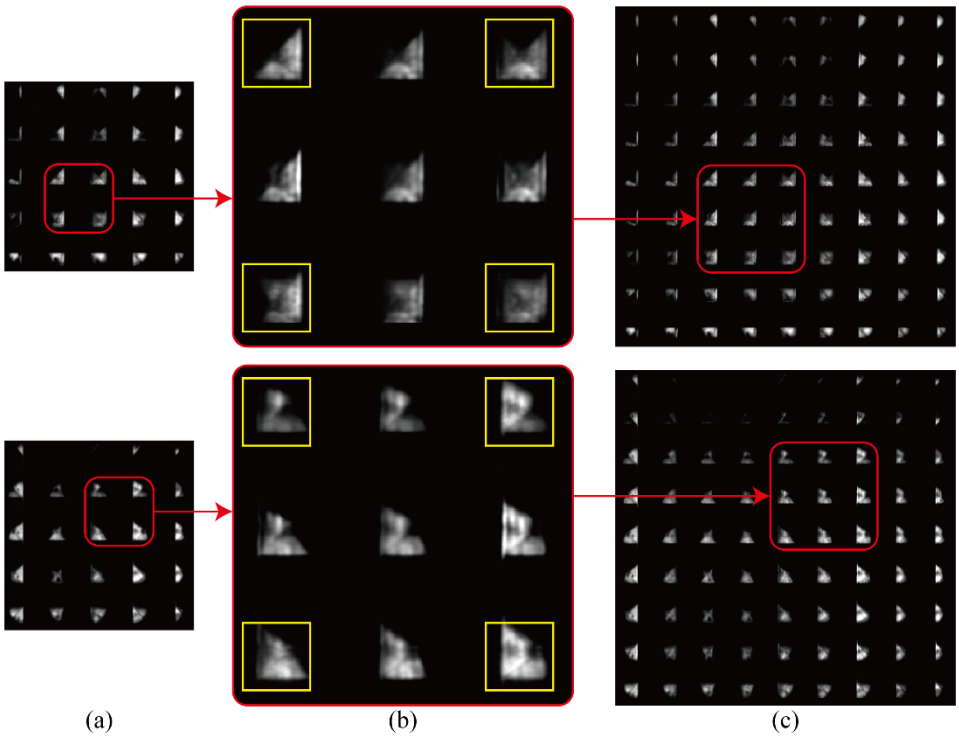

Figure 5.10 Comparison of the EIs between the proposed SSA system and the TAM system. (a) Low-resolution measurement EIs (TAM system). (b) Partial enlargement of the high-resolution EIs. (c) High-resolution EIs generated by the proposed self super-resolution approach (SSA system).

To further elaborate the advantage of the proposed high-resolution approach, the angular super-resolution results for other complex surfaces are shown in Figure 5.12, Figure 5.13, and Figure 5.14 which are the results of the pyramid sample under different illumination conditions, a wire bonding sample, and a pyramidal frustum structure, respectively.

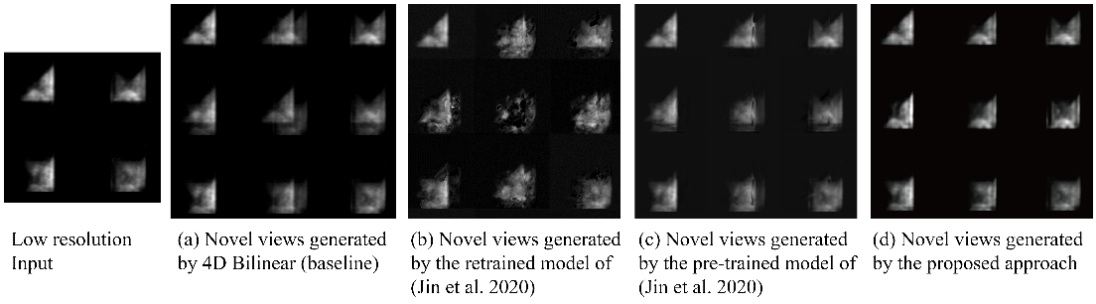

Figure 5.11 Comparison of the angular SR methods. The novel views are reconstructed using (a) the 4D Bilinear method as a baseline, (b) a SOTA deep learning model (Jin, Hou, Yuan, et al., 2020) trained on the measurement dataset, (c) the model (Jin, Hou, Yuan, et al., 2020) pre-trained on a public light field dataset and finetuned on the measurement dataset, and (d) the presented model, trained exclusively with the measurement set.

The second comparison was performed between the refocused images of the TAM system and the SSA system. With the proposed system, enhancing the angular resolution by roughly four times also led to a twofold improvement in the spatial resolution of the refocused images. Two refocused images obtained by the TAM system and the SSA system are shown in Figure 5.15, with the same local regions magnified to the same scale.

Moreover, it is noted that the smoothness of the refocused images is improved significantly since these novel-view EIs provide extra pixel information to fill the space between the two points in the digital refocusing process. According to the autostereoscopic theory and the refocusing principle, the corresponding points from the EIs are focused on different focal planes to form multiple refocused images. It is notable that the amount of the corresponding points determines the quantity of focal planes in

the stack. Increasing the layers of the focal stack improves the axial measurement precision. Hence, the novel views produced by the proposed SSA system result in the increase of the corresponding points and consequently the improvement of both lateral and axial measurement resolution.

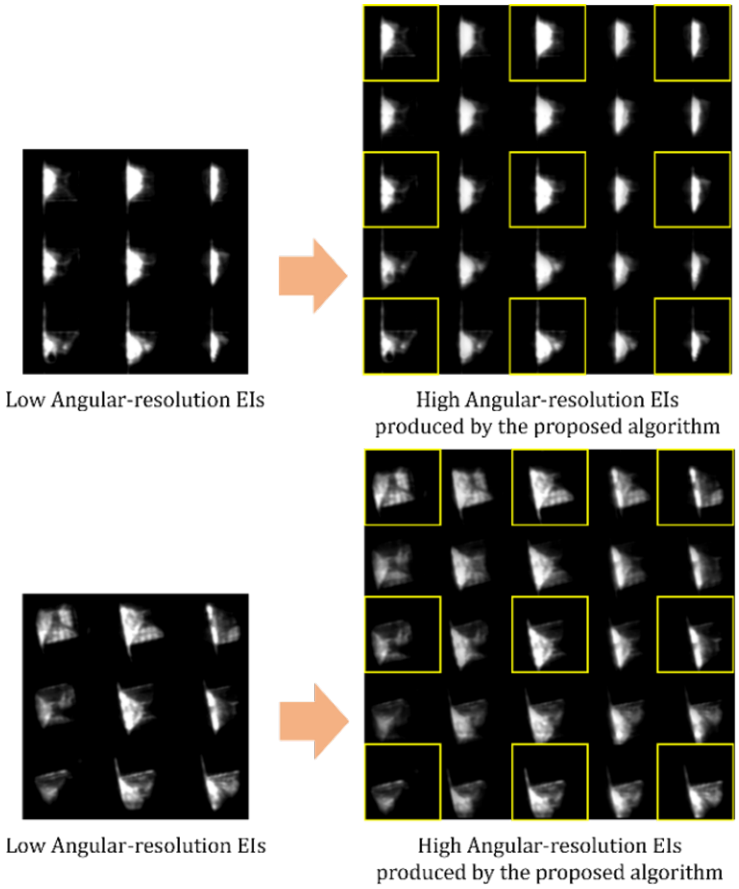

Figure 5.12 Angular super-resolution result of sample 1 (pyramid structures) under different illumination conditions.

The third comparison is about the 3D reconstruction results based on the measurement data. The reconstruction was performed using the low-resolution EIs recorded by the TAM system and our high-resolution EIs. The reconstruction results of the target surface are compared in Figure 5.16, where the point cloud generated by our high-resolution EIs is distinctly denser with more details kept.

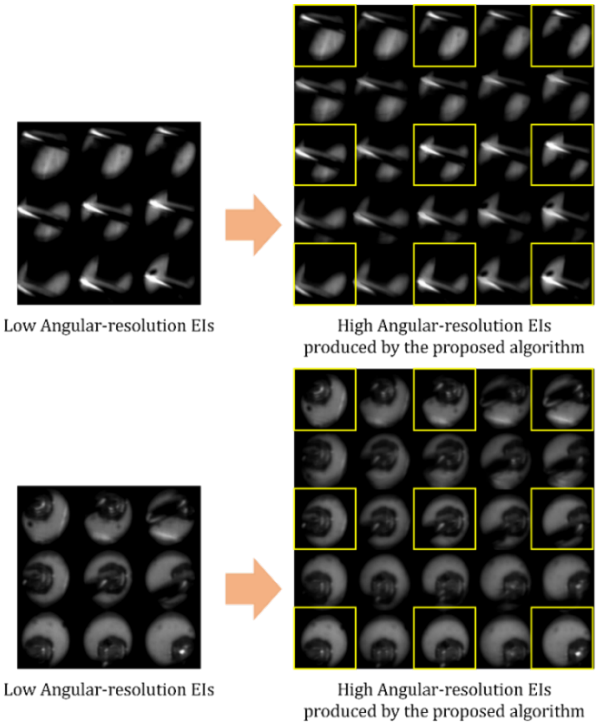

Figure 5.13 Angular super-resolution result of sample 2 (wire bonding).

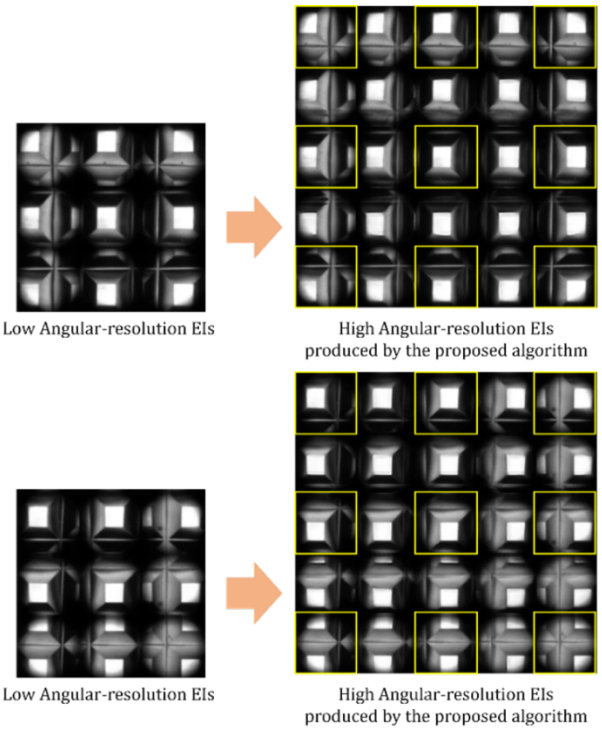

Figure 5.14 Angular super-resolution result of sample 3 (pyramidal frustum structures).

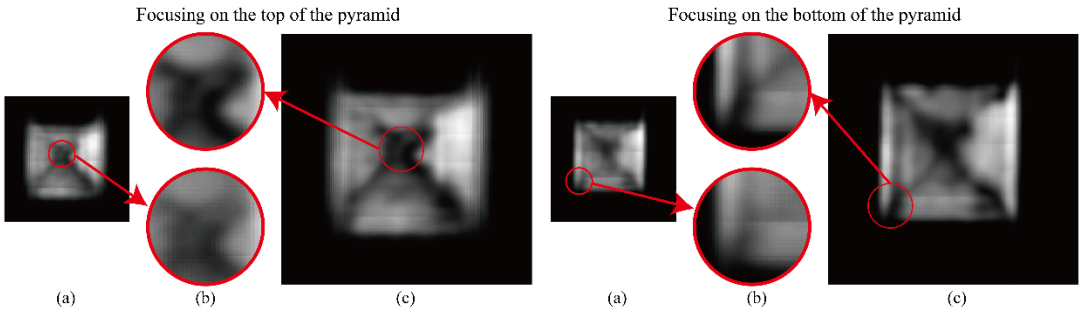

Figure 5.15 Comparison of the refocused images between the proposed SSA system and the TAM system. (a) Low-resolution refocused images with different focal length (TAM system). (b) Partial enlargement for comparison. (c) High-resolution refocused images (SSA system).

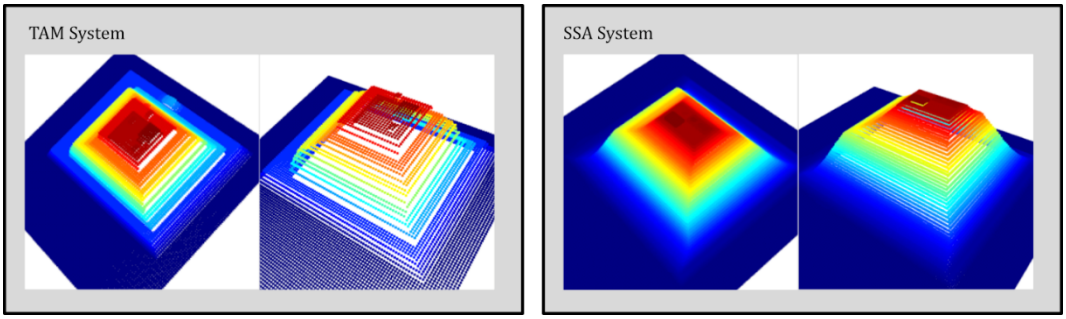

Figure 5.16 Reconstruction evaluation between the proposed SSA system and the TAM system

The measuring results obtained by the SSA system are shown in Table 5.2, which were determined using the disparity information extracted from the refocused image. Height is the axial dimension of the structure, while Edge A and Edge B are the two lateral dimensions. A total of 15 measurements were conducted and two targets were measured to verify the feasibility and measurement performance of the SSA system. The measurement result acquired by a commercial measurement product – Zygo Nexview Optical profiler – was used as the true value of the dimensions of the measured

pyramid structure. Statistical results including bias, standard deviation (SD), and maximum absolute error (MaxAE) are provided in Table 5.2, which indicates that the measurement data acquired by the SSA system is valid. A comparison between the SSA system and the TAM system is provided in both Table 5.2 and Figure 5.17. The numerical comparison shows the improvement of the measurement performance realized by the proposed system and the superiority of the system.

5.6 Summary

Autostereoscopic technology provides a rapid and accurate 3D measuring solution that can acquire the surface profile with only one snapshot. However, the dominant limitation of the autostereoscopic 3D measuring system is the trade-off regarding data resolution. In this chapter, a self super-resolution approach based on deep convolutional neural networks is embedded into the autostereoscopic measuring system, which helps the system to achieve self super-resolution during the measurement process and significantly enhances the angular resolution of the measuring data.

The self super-resolution approach was composed of a registration network, a residual encode–decoder network, and a refining network. All of these key components form a generative network that can interpolate novel views between the neighbouring EIs acquired by the proposed SSA measurement system. A discriminator network was implemented to distinguish the generative synthetic results from real measuring data. The distinguishing results were fed back to the generative network as an adversarial loss. Furthermore, a multi-loss function is used for training the proposed self super-resolution approach.

Table 5.2 Statistical comparison between the SSA system and the TAM system.

Measuring system	**SSA**	**TAM**	**SSA**	**TAM**	**SSA**	**TAM**
Pyramid 1	**Height**		**Edge A**		**Edge B**	
True value (μm)	55.2		67.4		67.1	
Bias (μm)	0.10	1.32	0.11	0.89	0.24	0.66
Standard deviation (μm)	0.27	1.14	0.37	0.85	0.40	0.92
Max. absolute error (μm)	0.45	4.05	0.78	2.12	0.84	1.96
Pyramid 2	**Height**		**Edge A**		**Edge B**	
True value (μm)	54.7		70.4		67.8	
Bias (μm)	0.12	1.40	0.15	0.90	0.44	0.77
Standard deviation (μm)	0.28	1.67	0.32	0.96	0.33	0.73
Max. absolute error (μm)	0.58	5.48	0.83	2.17	0.90	2.10

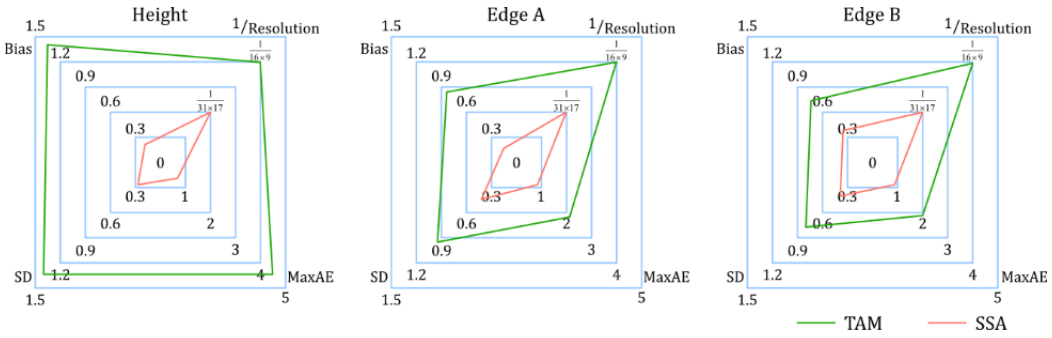

Figure 5.17 Accuracy evaluation between the proposed SSA system and the TAM system.

To showcase the effectiveness of the proposed self super-resolution approach, a series of experiments were conducted. Comparison was also conducted between the proposed SSA measurement system and the traditional autostereoscopic measurement system without super-resolution from many aspects including the resolution of measuring data, the quality of refocused images, and the statistics of their measurement results, which manifest in the better measurement performance of the proposed SSA

measurement system. This research reveals the potential and value of the SSA system to be used for rapid and accurate measurement on micro-structured surfaces.

Chapter 6 Multi-frame resolution-enhanced autostereoscopic measurement system

6.1 Introduction

The utilization of three-dimensional surfaces in product development has increasingly gained in popularity, enabling the realization of optical and mechanical functions that are specifically designed. Applications can be found in various industries such as biomedical (Tetsuka & Shin, 2020), optics (Hinman et al., 2017), aerospace (Civcisa & Leemet, 2015), energy, etc. Measuring the increasing geometrical complexity of 3D surfaces, especially for on-machine measurement, poses significant challenges. Although various high-precision on-machine measurement systems have been proposed, their performance is influenced by the machine kinematic errors, and they are susceptible to machine vibration (Gao et al., 2019). Although contact measurement methods generally achieve higher precision, noncontact methods are more flexible to implement and require less time consumption and system complexity, especially for small measured parts with microstructures.

Autostereoscopic 3D surface metrology is a noncontact surface detection technology that utilizes a single-lens imaging system integrated with a MLA. This setup enables the capture of raw 3D information of the measured surface in a single snapshot. This results in faster data acquisition for the measurement. A system-associated direct extraction of disparity information (DEDI) method (Li et al., 2015) provides 3D surface reconstruction. This solution offers a turnkey method for measuring 3D surfaces directly on the machine. Nevertheless, the resolution of this measurement system has been constrained due to the division of the image sensor's pixel count into multiple

smaller areas by the numerous small apertures of the MLA. Moreover, these segmented small areas of pixels need to go through a matching process and screening process before the final 3D point cloud of the target surface can be generated. Stated differently, the overall resolution of the measurement system is directly influenced by the resolution of each segmented small area. Undoubtedly, the segmentation caused by the MLA has a negative impact on the data resolution.

To boost the resolution of the autostereoscopic 3D surface measurement system, this study introduces a multi-frame resolution-enhanced autostereoscopic system. This system leverages the inherent vibrations generated by machine tools during the on-machine measurement process. By capturing multiple frames of the target surface with offsets caused by the vibrations over a short time span, it enables more precise measurements of 3D surfaces. The multi-frame resolution enhancement is realized by the subpixel information contained in different frames with slight displacement, a deep learning-based resolution-enhanced network and a training process. The processed resolution-enhanced image can reconstruct the 3D surface with significant improvements in both lateral and axial resolution. Experiments conducted on a sample with a micro-structured surface were employed to assess the efficacy of the presented approach and setup. The approach has been observed to effectively enhance spatial resolution and enhance measurement accuracy.

6.2 Multi-frame resolution-enhanced autostereoscopic measurement

Figure 6.1 is a schematic diagram of the system of multi-frame resolution-enhanced autostereoscopy for on-machine 3D surface measurement, including the recording and

reconstruction processes. Different spatial locations of the elemental lenses in a MLA cause small differences of viewing angle in the elemental images received on the image sensor (known as disparities). The disparity information can be utilized to infer the 3D information of the target surface, which is an essential step in the reconstruction process. The specific point's disparity can be quantitatively expressed by considering the parameters of the system setup, such as the size of each pixel on the image sensor, the pitch of the MLA, the gap (distance between the MLA and the image sensor), and the dimensional variation along the depth direction.

The quantitative disparity information, which encompasses both the lateral and depth directions, is transferred from the recording process to the reconstruction process. The corresponding points, represented by image points from different EIs, originate from a single object point in the object space (depicted as red points in Figure 6.1). These points adhere to a quantitative relationship between disparity information and system parameters. Corresponding points need to be accurately chosen based on the match of pixel information and its neighbourhood in EIs and form the 3D digital reconstruction at corresponding spatial locations. The reconstructed and object spaces are symmetrical both in the lateral and axial directions according to the reversibility of optical rays.

During the on-machine measurement process, vibration from the machine tool between the target surfaces and the measurement system is inevitable. This vibration results in a slight movement at the micrometre scale towards the image sensors. As a result, each of the measurement frames captures a combination of various optical signals, as depicted in Figure 6.1. In this sense, the pixel representation of the same scene and the same object is different. High-resolution (HR) images can be

reconstructed by analyzing and processing the different pixel representations of multiple frames.

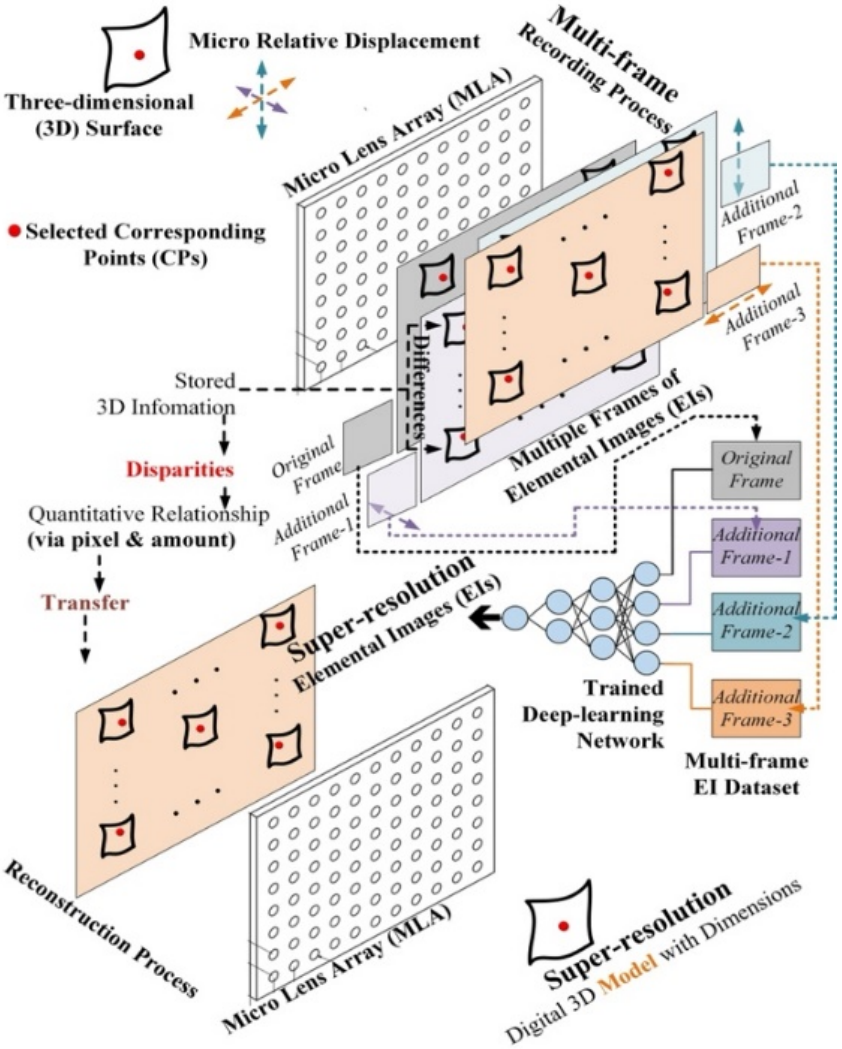

Figure 6.1 Working principle of multi-frame resolution-enhanced autostereoscopic metrology for on-machine 3D surface measurement.

As illustrated in Figure 6.2, the two different frames are recorded for the same object but have a slight displacement. The pixel distribution of the two frames, which refers to the value of the pixel points of the target surfaces, is different when the object

appears in a different position of the image sensor. After registration and fusion, the redundant pixel information is combined and forms a new pixel distribution in a subpixel space. The new pixel distribution is processed in the reconstruction process and a high-resolution frame with sharp edge information and details is generated from multiple frames. The fusion of multiple frames can be viewed as an optimization process aimed at finding an ideal distribution that closely approximates high-resolution data. Each frame is down-sampled from a high-resolution distribution, with sub-pixel differences generated by displacements resulting from vibrations. Since the super-resolution problem is ill-posed—meaning that a single low-resolution data does not correspond to a unique high-resolution data—more low-resolution data down-sampled from the high-resolution distribution can provide greater support for identifying the unique distribution of the high-resolution data. The proposed fusion method seeks to determine the ideal data distribution from the enriched multi-frame information.

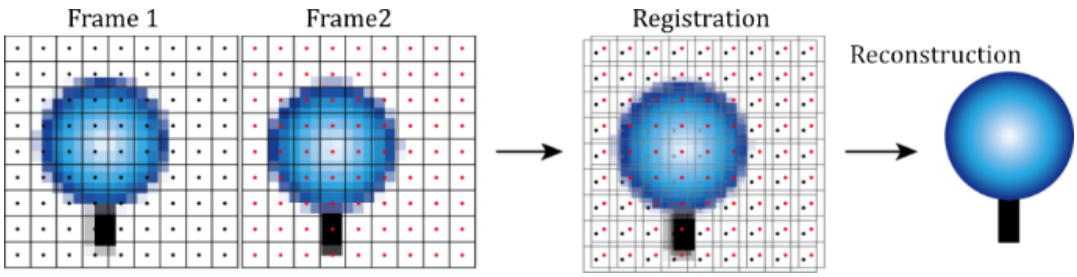

Figure 6.2 Illustration of the multi-frame resolution-enhancement process.

The two key issues to address in the multi-frame resolution-enhancement problem are registration and reconstruction. Conventional methods usually use a priori knowledge to extract features from the multiple frames and realize registration based on these features. Fusion is achieved by a series of designed kernels based on experiments or expert experience. However, conventional methods often struggle to extract effective features from images and reconstruct high-resolution images with

robustness due to the presence of various noises such as Gaussian noises, salt and pepper noises, smudge noises, etc. These noises are typically caused by factors like illumination, exposure, and lens conditions in the measurement images. It is inspiring to utilize deep learning to generate resolution-enhanced EIs through performing accurate registration and high-resolution reconstruction. A deep-learning network is employed to construct a multi-frame resolution-enhanced model. This model is designed to enhance the resolution of LR measurement image stacks, while also improving denoising and preserving clear details in the resulting HR images.

6.3 Multi-frame resolution-enhanced deep learning model

To generate resolution-enhanced EIs based on multiple frames of low-resolution EIs, a deep learning network is developed. A supervised training process is used to generate resolution-enhanced images based on image data collected under various conditions, light intensity, recording device, etc.

6.3.1 Model framework

As shown in Figure 6.3, the proposed multi-frame resolution-enhanced deep learning model consists of four components: a single-frame resolution-enhanced network, a registration network, an auxiliary-frame resolution-enhanced network, and a series of convolutional layers for post-processing. A schematic diagram of the model is depicted in Figure 6.3. The captured multiple frames are divided into a base frame and auxiliary frames. The high-resolution image output retains the same geometric position as the base frame, while the auxiliary frames are utilized to provide redundant

subpixel information. The base frame is first up-sampled using Bilinear so that the spatial dimension is increased to a desired value. The up-sampled base frame serves as the input to the single-frame resolution-enhanced network. Meanwhile, the input frame undergoes convolutional layers and activation functions to convert it into a stack of high-dimensional single-frame features. These features are then utilized for subsequent processing.

The registration network aligns all frames, including the base frame and auxiliary frames, by taking them as input. Although the displacement detection and registration are also able to be achieved by traditional methods such as SIFT (Lindeberg, 2012), the traditional methods are more sensitive to the noises which are unavoidable during on-machine measurement due to illumination, vibration, machining environment, etc. In addition, the registration network can realize an end-to-end training and inference fashion so that no extra pre-processing of the raw measurement data is required.

Since these frames only undergo slight displacement, it is assumed that only translation occurs. According to the affine transformation matrix, the registration process is formulated as

$$\begin{bmatrix} \tilde{x}_{Ai} \\ \tilde{y}_{Ai} \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \theta_x \\ 0 & 1 & \theta_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{Ai} \\ y_{Ai} \\ 1 \end{bmatrix} \tag{6.1}$$

where $I(x_{Ai}, y_{Ai})$ is one of the auxiliary frames and $I(\tilde{x}_{Ai}, \tilde{y}_{Ai})$ is the corresponding registered frame which has been aligned with the base frame. θ_* is the translation parameters. The translation parameters (θ_x, θ_y) are the output of the registration network.

To efficiently compress the dimensions of the input while retaining the essential information, the registration network employs max-pooling filters. Fully connected layers are followed to realize the prediction of the translation parameters. The Tanh (hyperbolic tangent) activate function is used to compress the output translation parameters in [-1, 1]. The output auxiliary frames are aligned with the base frame through registration. With a similar process to the aforementioned single-frame super-resolution route, the aligned auxiliary frames are up-sampled and input into the auxiliary-frame resolution-enhanced network.

As a result, these auxiliary frames are converted into a stack of multi-frame features. These two stacks of features, i.e., the single-frame features and the multi-frame features are merged and then input into the post-processing convolutional layers. After the post-processing layers, a high-resolution image is reconstructed. All these mentioned sub-networks use a residual connection architecture (ResB in Figure 6.3) to avoid gradient vanishing. Apart from the registration network using ReLU and Tanh, all the other sub-networks use Leaky ReLU as their activation functions since research (Lai et al., 2017; Xu et al., 2021) has shown the superiority of Leaky ReLU for super-resolution applications.

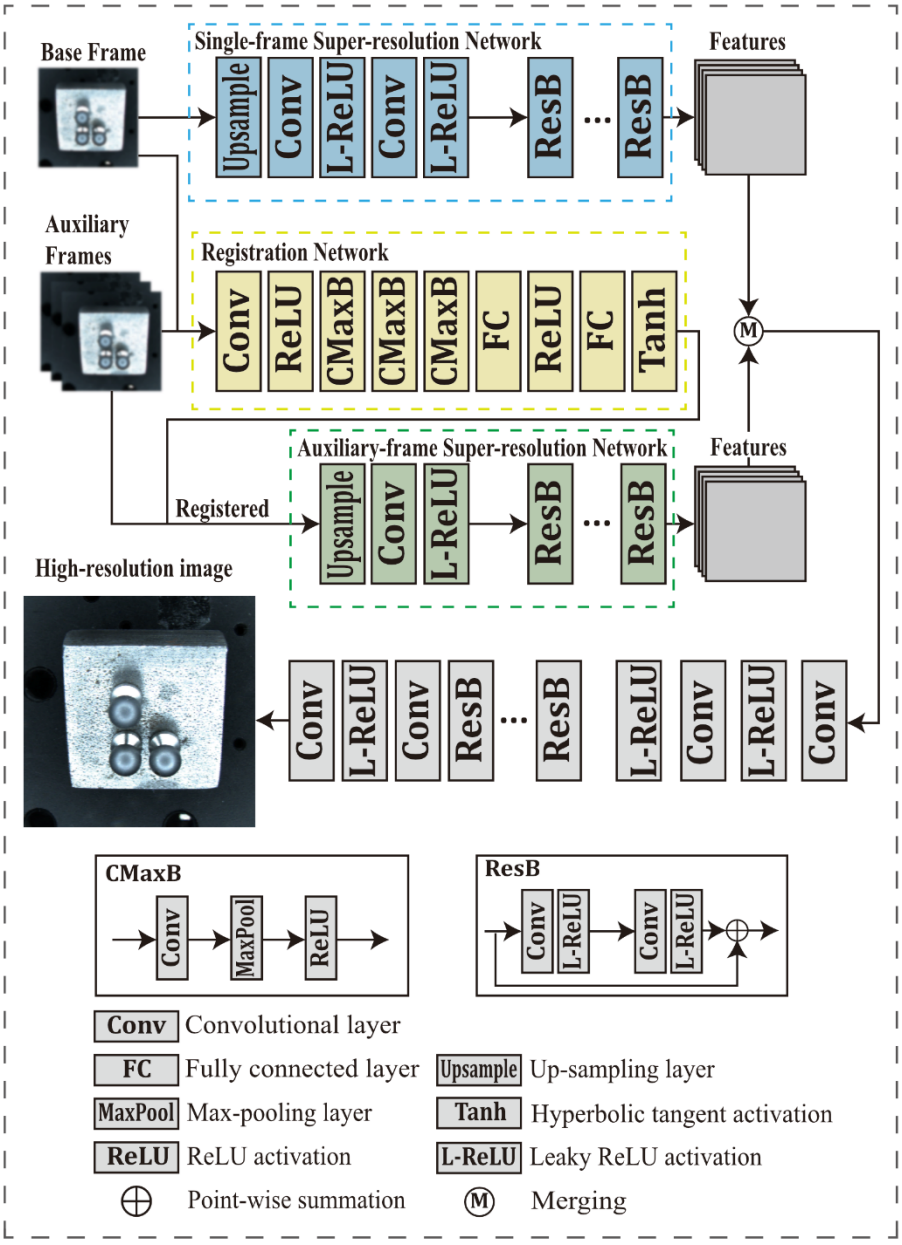

Figure 6.3 Multi-frame resolution-enhanced deep learning model.

6.3.2 Model training

To train the proposed model, a supervised training process is used based on image data collected under various conditions, light intensity, recording device, etc. During training, the input data are initially down-sampled, while the raw data are utilized as the ground truth. To realize a clearer reconstruction, Gaussian noises are added to the

input data to simulate the noises in realistic environments. Hence, the objective of the proposed multi-frame resolution-enhanced network is not only to recover the high-resolution information but also to achieve denoising.

The loss function is comprised of three parts which are reconstruction loss, gradient loss, and perceptual loss. The reconstruction loss is evaluated by computing the mean absolute errors between the HR synthetic images and the actual images, which serves as the metric for error assessment. To preserve the edge information of the reconstructed high-resolution images, the gradients of the ground truth and the reconstructed images are compared in both the horizontal and vertical directions. The resulting errors are then utilized as the gradient loss. The perceptual loss (Johnson et al., 2016) measures the feature distance. The features are acquired by a pre-trained network (Simonyan & Zisserman, 2014), which is a widely used trained network. Hence, the total loss is

$$\begin{aligned} L = \sum \Big(& \left| I^{HR} - \hat{I} \right| + \left| \nabla_x I^{HR} - \nabla_x \hat{I} \right| \\ & + \left| \nabla_y I^{HR} - \nabla_y \hat{I} \right| + \left| \phi\left(I^{HR} \right) - \phi\left(\hat{I} \right) \right|^2 \Big) \end{aligned} \tag{6.2}$$

where I^{HR} is the high-resolution images reconstructed by the proposed network, $\hat{I}$ is the ground truth, and $\phi(\cdot)$ denotes the VGG network. Furthermore, to ensure comprehensive training of the proposed network, the training data were augmented through techniques such as rotation, flipping, and random cropping.

6.3.3 Implementation details

In this work, the measurement images with 151×151 pixels are super-resolved and

up-scaled four-fold. The number of total input frames is 4. The single-image resolution-enhanced network and the auxiliary-frame resolution-enhanced network both contained two residual blocks and the post-processing layers contained two residual blocks.

The training data are collected by a 2D imaging system and a Lytro Illum commercial light field camera. Since a single EI captured by the proposed measurement system has limited pixels, the data obtained by the Lytro Illum camera and the 2D system with higher resolution can provide much richer pixels to achieve more effective training of the resolution-enhanced model that learns the mapping from the LR to HR images. Multiple scenes that include various samples such as sphere surfaces, machining parts, bonding wires, and other objects containing complex surfaces are captured under different illumination conditions for the construction of the training dataset. Each scene contains four frames with slight displacement. One of the four frames is used as the base frame and the other frames are used as the auxiliary frames.

For more efficient learning using the limited measurement data, the up-scale factor is set to 2 during training but changed to 4 after the model is well-trained so that the resolution of the autostereoscopic measurement data is improved by four-fold. Data augmentation is conducted by rotating the training data by 45, 90, 135, 180, 225, 270, and 315 degrees, flipping them from left to right or from up to down, and cropping the data into random-size patches. The input patch size is configured as 128. The model is realized using PyTorch and trained on NVIDIA RTX 2080 GPUs. The initial learning rate is 10^{-4}. The learning rate is decayed at intervals of 10 training epochs. In the experiment, learning is stopped when the L1 loss variation remains below 10^{-3} for the last 10 epochs. Furthermore, checkpoints are saved at different stages of training to determine the best-performing model.

6.4 Surface Reconstruction

The depth estimation process relies on the direct extraction of the disparity information method (Li et al., 2015), disparity patterns, and shape from focus via digital refocusing (Li et al., 2014), as shown in Figure 6.4. Digital refocusing is first performed using the recorded autostereoscopic data so that a stack of refocused images is acquirable. The corresponding points should focus at a specific depth plane that is equivalent to finding focus regions in the refocused image stack. A focus measure operator is used to detect focus points in every refocused image so as to obtain the focus volume. By applying smoothing and denoising techniques to the focus volume, an initial depth map is estimated using the winner-takes-all strategy. Utilizing the preliminary estimate, achieving a fully-focused image becomes feasible. To further refine the estimated depth, guided filtering is performed on the preliminary estimation based on the all-in-focus image. Outlier points caused by incorrect estimation such as small locally convex or concave regions in the depth map are further disposed of under pre-defined thresholds based on the assumption of continuity surfaces. As a result, desired depth maps, point clouds, and the corresponding all-in-focus images can be acquired from the low-resolution (LR) or high-resolution (HR) autostereoscopic measurement data.

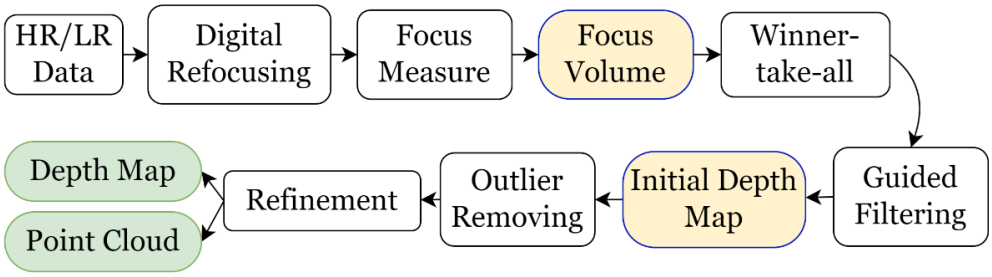

Figure 6.4 Framework of the surface reconstruction process from the autostereoscopic measurement data.

6.5 Experiments on micro-structured surfaces

6.5.1 System setup for the on-machine system

A prototype, depicted in Figure 6.5, of the multi-frame resolution-enhanced autostereoscopic 3D surface measurement system was built to perform on-machine 3D surface measurement. The whole system is mounted on the motion stage of a Moore Nanotech 350FG ultra-precision machine. Based on the offline calibration of the system, considering the overall magnification of the objective lens, zoom lens, and the size of the image sensor used, the measurement system has an overall field of view (FOV) of 625 μm diagonally.

To assess the accuracy and resolution of the autostereoscopic system, a series of measurement experiments are performed on a 3D micro-structured sample. The machine's air bearing work spindle is used to mount the sample. Multiple frames of the EIs of the sample with offsets of pixels are captured during the on-machine measurement process. The offsets among the multiple collected frames are analyzed based on pixel values and greyscale projection. The results show that the subpixel level offsets between the different frames happened during the on-machine measurement

which qualifies for the proposed multi-frame resolution-enhanced method.

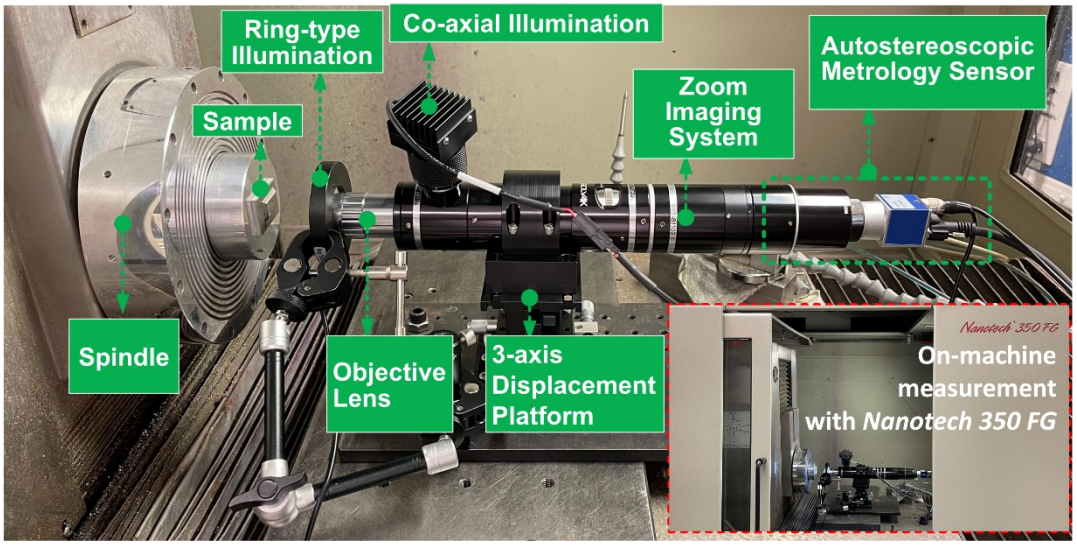

Figure 6.5 On-machine measurement through a multi-frame resolution-enhanced autostereoscopic 3D surface measurement system.

To examine the frame jitter resulting from machine tool vibrations, the SIFT descriptor is employed to determine the pixel-level shifting among multiple consecutive frames captured during the measurement. The first frame is used as the reference, and the other frames are matched with the reference frame to detect the small displacement. The SIFT descriptor and detector are employed to detect and compute the key points of both the reference frame and the other frames. The matching is achieved using the FLANN (Fast Library for Approximate Nearest Neighbours) method.

Since the SIFT detector is able to achieve detection at subpixel scale, the subpixel-level distances between the matched points can be determined so that the frame jitter will be identified. To mitigate the impact of inaccurate matching, a total of 500 groups of matched points from the reference frame and the detected frame are utilized for the

analysis. To eliminate the effects resulting from noises, the same displacement detection is performed on noised data which are generated by adding extra Gaussian noises to the reference frame. To guarantee the reliability of the analysis, the average greyscale difference per pixel between the noised data and the reference data is measured at the same scale as the greyscale difference among the multi-frame data. The results show a difference of 0.950 per pixel between the noised data and the reference, and 0.931 per pixel among the multiple frames.

The results of the multi-frame data and the noised data are shown in Figure 6.6, where a total of five frames excluding the reference frame are involved. The pixel displacement of the noised data is represented by the red lines, while the pixel displacement of the multi-frames is depicted by the blue lines. The results demonstrate that the subpixel displacement happens during the on-machine measurement. Since the system utilizes a CCD sensor with a resolution of 2456x2058 and a pixel size of 3.45 μm, the vibration amplitude is around 1.725 μm based on the image point matching shown in Figure 6.6. Hence, the on-machine measurement data align with the aforementioned assumption of multi-frame super-resolution and enables resolution enhancement at a subpixel level.

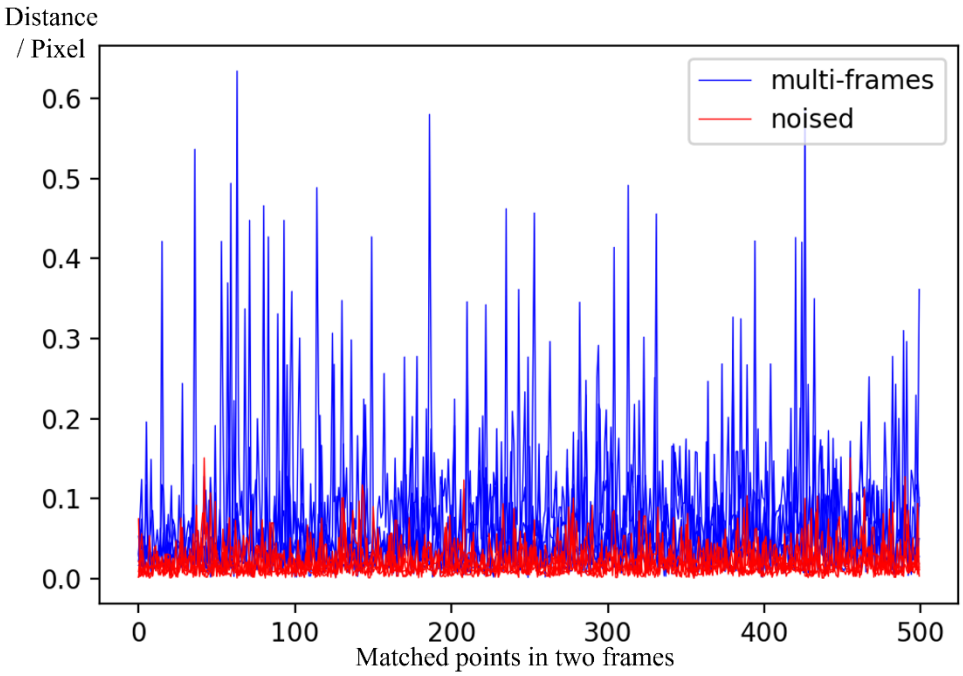

Figure 6.6 Jitter analysis of the multiple measurement frames captured during the on-machine process (blue lines) and the noised data generated by imposing Gaussian noises (red lines).

6.5.2 Experimental analysis

Figure 6.7 shows a comparison between high-resolution EIs acquired by the bilinear method and the proposed multi-frame resolution-enhanced deep-learning method. Upon comparing the details of the sectional zoom-out, as depicted in Figure 6.7, it is worth noting that the proposed method exhibits improved visual sharpness for high-frequency signals. Slightly different from computer image super-resolution tasks, the enhancement for measurement is invalid if more image artefact points are created for a huge contribution to the visualization.

The proposed method clearly enhances high-frequency signals, which typically

correspond to edges or key points in the measured sample. These signals are more important to depth estimation during the shape-from-focus process. The enhancement of these high-frequency signals contributes to the accuracy of focus measurement on the refocused image stack so that the corresponding points can be detected at the correct depth plane.

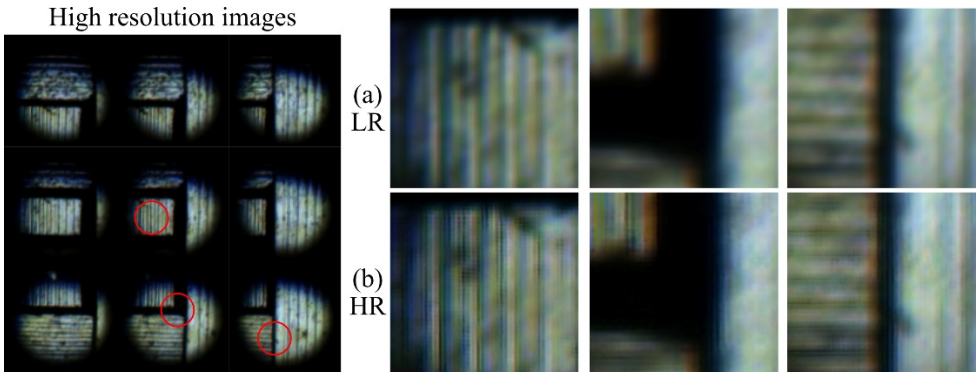

Figure 6.7 Comparison of experimental results obtained by the bilinear method (a) and the multi-frame resolution-enhanced deep-learning method (b)

For a convincing demonstration of the proposed resolution-enhancement model, multiple experiments on various samples both at micro scale and macro scale are performed. The scenes of several surfaces at various scales are captured by different systems and devices, with the resolution-enhancement results shown in Figure 6.8. The superiority of the proposed multi-frame resolution-enhancement model in recovering finer details from low-resolution inputs is evident when compared to traditional methods. In addition, the improvement is not limited to the proposed system; it is effective in various systems with a variety of fields of view. In the same way, the high-frequency signals in these scenes are further enhanced by the proposed method, and these sharp key points definitely benefit the corresponding point matching and the depth

estimation.

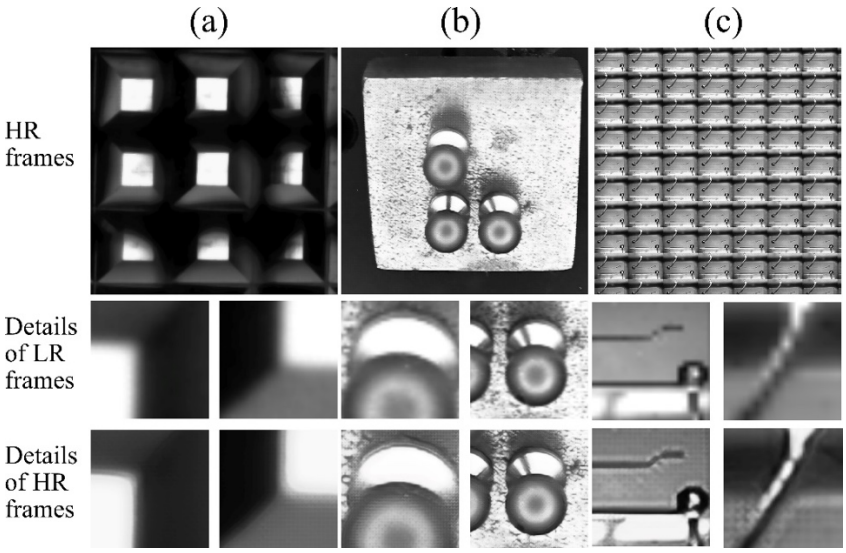

Figure 6.8 Multi-frame resolution-enhancement results of various surfaces. (a) Pyramidal frustums. (b) Sphere surfaces. (d) Wire bondings.

Based on the autostereoscopy theory, digital refocusing is able to reconstruct a series of image slices with various focus depths so that the height of the measured sample is able to be detected. In terms of the detection of the focus region so as to determine the desired depth information, a Sobel filter is used as the focus measure operator. A curve of the focus levels is fitted, and the peak value of the curve can be estimated. The peak value is the disparity, i.e., the depth of a target object point, and therefore the surface reconstruction can be achieved. Figure 6.9 presents a comparison between the conventional single-frame method and the proposed multi-frame resolution-enhanced deep-learning method. The comparison includes the all-in-focus image generated during shape from focus, the depth estimation results, and the point clouds. These results vividly demonstrate the resolution enhancement achieved by the proposed method. A measurement result from a commercial measurement product –

Zygo Nexview Optical profiler – is presented as the reference. The resolution of each EI both in lateral and axial directions has been enhanced four-fold, from 151×151 pixels to 604×604 pixels.

Regarding the all-in-focus results, it can be found that the focus measurement achieves more accurate detection for the high-resolution data acquired by the multi-frame system whereas the low-resolution data result in inaccurate focus detection at the edges, as highlighted in Figure 6.9 (a). The depth estimation derived from low-resolution data exhibits more noise and incorrect points, as evidenced in Figure 6.9 (b) and (c). Furthermore, the intensity of point clouds generated by the multi-frame system is significantly enhanced.

Figure 6.10 shows the error maps analyzed by the iterative closest point (ICP) method which compares measured data acquired via repeated measurements. The repeatability of the proposed displays better performance regarding the standard deviation of 10 repeated measurements as shown in Figure 6.10.

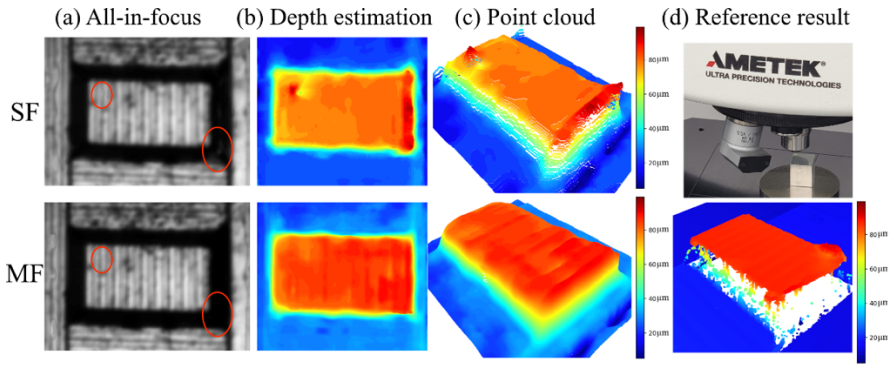

Figure 6.9 Comparison of experimental results from the single-frame (SF) system and multi-frame (MF) system. (a) All-in-focus image, (b) depth estimation, (c) point cloud, and (d) reference results.

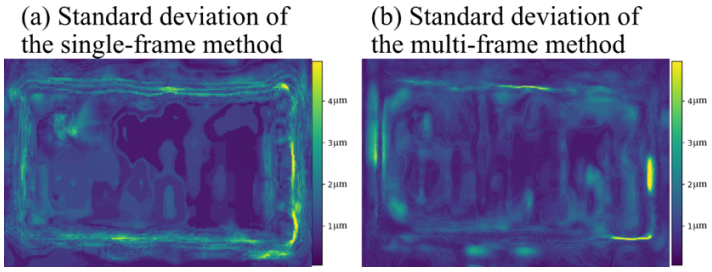

Figure 6.10 Standard deviation of repeated measurements using (a) the traditional single-frame method and (b) proposed multi-frame resolution-enhanced method.

However, it is still found that the deviation at the edges is much larger in both the single-frame and multi-frame systems. This could have resulted from the capability of the focus measurement operator which is sensitive to high-frequency signals which are not only edges and key points, but also could be noises. Hence, more investigations and research for robust and adaptive focus measurement operators can further benefit the improvement of depth estimation accuracy for the autostereoscopic measuring system.

6.6 Summary

This chapter presents the development of a multi-frame resolution-enhanced autostereoscopic system for on-machine 3D surface measurement. The system takes advantage of the machine vibration together with a multi-frame resolution-enhanced deep learning model to acquire multiple frames of the target surface profile with offsets to enhance the resolution and accuracy of on-machine 3D surface measurement. The performance evaluation results demonstrate that the proposed system outperforms the conventional single-frame system in terms of measurement accuracy in repeated measurements. The proposed method also provides around 16 times the total amount of point cloud data with the additional improvement of measurement accuracy and

robustness.

Chapter 7 Overall conclusion and future work

7.1 Overall conclusions

The autostereoscopic 3D measuring system achieves fast data collection within one snapshot, and the disparity information is extracted from the captured elemental images from multiple perspectives. Based on the disparity information and the recorded pixels, the axial and the lateral dimensions are inspected to realize surface reconstruction. The autostereoscopic measuring system typically consists of several components including an objective lens, a high-magnification zoom lens system, a MLA, and an image sensor. To enable observations from multiple perspectives, the MLA is positioned in front of the image sensor. These elemental images captured by the image sensor record a plenoptic map which contains redundant disparity information. Digital refocusing, epipolar-plane image analysis, and extraction of disparity information based on disparity patterns can be performed to implicitly or directly make use of the stereo clues for 3D reconstruction.

However, a main obstacle blocking the development of the autostereoscopy technology is the resolution of the recorded data. Since the micro-lenses split the image sensor into multiple regions to recode the multi-perspective views, the resolution of each view is limited by the finite pixels. Obviously, the increase of the resolution of each view results in a decrease in the number of perspectives. Enhancing the lateral and axial resolution simultaneously at the hardware level is a challenging trade-off for autostereoscopic systems.

To achieve this objective, the thesis contributes to an angular SR algorithm and a multi-frame spatial SR algorithm. These algorithms aim to improve the data resolution

of the autostereoscopic measuring system and are seamlessly integrated into the measuring system to augment its measuring capability. The notable contributions of this study are listed below:

(i) Firstly, the development of a novel algorithm based on deep learning becomes crucial for enhancing the angular resolution of LF data, since current learning models generally cannot produce high-quality interpolation results using real-world data with severe noise and large baselines that always happen in the measurement data. As a result, a learning model for angular SR is presented to perform a motion estimation so that the regression problem of coarse disparity estimation (which is used for the reconstruction of novel views) is converted into a classification problem of motion estimation.

A generic semi-supervised learning paradigm is presented for the learning models for angular super-resolution, which performs 2-step inference before every backpropagation and does not require splitting the training data into input and ground truth. The output is supervised directly by the input. To evaluate the effectiveness of the presented model, a series of extensive experiments have been performed using synthetic and realistic public datasets. These experiments aim to compare and assess the effectiveness of the proposed model in relation to other SOTA methods. In the scenes with large baselines and multi-depth objects, both PNSR and SSIM are improved by the proposed model compared with other methods. In contrast to other SOTA methods, the proposed model demonstrates the ability to generate images with fewer artefacts and reduced image ghosting. Moreover, it excels in achieving high-quality reconstruction using data with noise, large baselines, and

complex textures, where other methods often fall short. It is also inspiringly noted that the proposed model only takes around 30% of data for training to achieve better results.

(ii) To enhance the angular resolution and improve the measurement precision and accuracy of the autostereoscopic measuring system, a deep learning model for angular SR is integrated into the system. This integration enables the system to achieve self SR solely based on the measurement data collected by itself. Since a quite large baseline exists in the measurement data collected by the system, a registration network is built to estimate the displacement between the elemental images before the novel view reconstruction. Once the registration process is complete, an encoder–decoder network is utilized to extract features from the input data individually. The extracted features are subsequently employed to reconstruct the features of the novel views.

To further improve the quality, a refining network is employed which maps the features outputted by the encoder–decoder network onto the novel view plane. To further enhance the quality of the generated novel views, a discriminator network is established to distinguish the generated images. The differentiation results are incorporated into the generation process to enhance the quality of the output images. The super-resolution capability of the proposed deep learning model is evaluated using multiple samples. The experiments indicate that the outputs by the model exhibit comparable quality to realistic measurement data. To showcase the advancements achieved by the proposed self SR autostereoscopic measuring system, various evaluations

such as digital refocusing, surface reconstruction, and measurement results are performed using both the enhanced data and the raw LR data.

After applying the resolution enhancement method, the angular resolution is increased from 16 × 9 to 31 × 17. The discrepancy between the average measured values and the true values is reduced by approximately 1 micrometre, from over 1 μm to around 0.1 μm. Furthermore, the deviation of repeated measurements is also reduced by around 1 micrometre. As a result, the super-resolved high-resolution data lead to improved measurement accuracy.

(iii) In terms of the enhancement of spatial resolution, a multi-frame super-resolution algorithm is introduced. This algorithm utilizes the inherent vibrations of machine tools to enhance the accuracy of on-machine measurement. Since the vibration of machine tools is inevitable, the frames captured over various timespans should have a subpixel level of displacement between each other. Based on the subpixel level of displacement, the reconstruction of high-resolution information can exploit the redundant subpixel information in the multiple frames.

As a result, a multi-frame super-resolution model based on deep learning is proposed to enhance the spatial resolution of the elemental images recorded by the measuring system, thereby yielding significant improvements. Jitter analysis between different frames is performed to demonstrate the existence of subpixel displacement. In the proposed model, the initial step involves

reconstructing high-resolution features using a base frame. These features are then combined with the features extracted from other auxiliary frames to generate the final HR outcomes. The effectiveness is assessed by utilizing multiple multi-frame data captured from macro surfaces, microstructures, and complex scenes.

Experiments show that the learning model is capable of achieving spatial super-resolution based on multiple frames captured with minor vibration. The autostereoscopic measuring system integrated with the multi-frame super-resolution model is tested for on-machine measurement on a Moore Nanotech 350FG ultra-precision machine. Comparisons including the reconstruction point clouds and the measurement results are conducted and indicate that the measurement accuracy can be enhanced by the multi-frame super-resolution solution.

This method significantly improves the spatial resolution by 4 folds, from 151 × 151 to 604 × 604. This enhancement allows for the restoration of more detailed and precise edge information from the spatial data. Additionally, the average measurement bias towards the true values is reduced from approximately 1.4 μm to around 0.3 μm, resulting in a reduction of 1 μm. Furthermore, the deviation among repeated measurements is reduced from 1.533 μm to 1.388 μm.

7.2 Suggestions for future work

The proposed autostereoscopic measuring system in this thesis is still mainly

dependent on computer vision and image processing. There is still room to improve the measurement capability and suggestions are as follows:

(i) Spectrum analysis for super-resolution methods can provide a more straightforward evaluation of the enhanced results. The information in an image, categorized into low, medium, and high frequencies, offers different visual and physical insights for the final analysis. Generally, the main difference between hi gh-resolution and low-resolution data lies in the high-frequency information, which typically denotes edges and salient points in the images with significant pixel gradient changes. Although error maps between ground truth and the output of the methods implicitly indicate the accuracy of high-frequency information, spectrum analysis based on Fourier transform can provide more concrete evidence of the improvements made by the enhancement methods. Low- and medium-frequency information can also be analyzed similarly to investigate the performance of the learning models. Therefore, spectrum analysis can be a valuable approach for evaluating super-resolution methods.

(ii) A simulation model is essential to study the interaction between the light rays and the surfaces being measured. In addition, the simulation data are able to be used to train AI models to realize intelligent reconstruction directly from the measurement data. On the basis of the simulation, the ray propagation in the system can be described more precisely so that the noises and illumination can be adaptively controlled to cater to different samples and scenes. It is still a tricky issue to select and adjust illumination devices from different micro-structured surfaces since the rays reaching the small features of the surfaces are limited under improper illumination conditions. This can result in an unclear observation of some key features. In addition, the complex interaction between the rays and

the measured surfaces will affect the quality of measurement data significantly. The low quality of raw data could increase the difficulty during the matching process, even after enhancement. Hence, it is important to theoretically model the ray interaction in the whole system to achieve a more accurate measurement.

(iii) The main workflow of the autostereoscopic measuring system still relies on image processing based on human recognition. The learning models employed in the thesis serve to improve the image quality at a level that aligns with human perception. However, the interaction between rays and measured surfaces may be outside the scope of perception of human beings. The powerful representation capability of deep learning models enables the direct extraction of 3D information from the raw signal recorded by the sensor. Hence, it is inspiring to explore more possibilities for deep learning models to make a difference during autostereoscopic measurement instead of as a tool for image processing.

(iv) A more efficient depth estimation method is expected to be developed to achieve more accurate 3D inspection. Since the micro-structured surfaces usually have a very good finish, the features that are able to be captured for focus-level detection and stereo clue matching are limited. This will result in a sparse point cloud as the direct output from the input data. The dense point cloud is reconstructed primarily based on a priori knowledge. It is possible to get rid of a priori knowledge during the dense reconstruction by using machine learning models to recognize the measured surfaces based on the measurement data. It is possible to realize a more intelligent generation of a dense point cloud based on a sparse one.

References

Adams, A. The (New) Stanford Light Field Archive: Light Fields from the Lego Gantry. http://lightfield.stanford.edu/lfs.html. 2008

Ahn, H. K., Kang, H., Ghim, Y.-S., & Yang, H.-S. (2019). Touch Probe Tip Compensation Using a Novel Transformation Algorithm for Coordinate Measurements of Curved Surfaces. *International Journal of Precision Engineering and Manufacturing*, *20*(2), 193-199. https://doi.org/10.1007/s12541-019-00076-2

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., . . . Chen, G. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. International conference on machine learning,

Bastas, A. (2020). Comparing the probing systems of coordinate measurement machine: Scanning probe versus touch-trigger probe. *Measurement*, *156*, 107604.

Bauza, M., Woody, S., Woody, B., & Smith, S. J. W. (2011). Surface profilometry of high aspect ratio features. *271*(3-4), 519-522.

Bergen, J. R., & Adelson, E. H. J. C. m. o. v. p. (1991). The plenoptic function and the elements of early vision. *1*, 8.

Bishop, T. E., Zanetti, S., & Favaro, P. (2009, 16-17 April 2009). Light field superresolution. 2009 IEEE International Conference on Computational Photography (ICCP),

Bixler, G. D., & Bhushan, B. J. A. F. M. (2013). Fluid drag reduction with shark‐skin riblet inspired microstructured surfaces. *23*(36), 4507-4528.

Breckwoldt, W. A., Daltorio, K. A., Heepe, L., Horchler, A. D., Gorb, S. N., & Quinn, R. D. (2015). Walking inverted on ceilings with wheel-legs and micro-structured adhesives. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),

Briot, J.-P., Hadjeres, G., & Pachet, F.-D. J. a. p. a. (2017). Deep learning techniques for music generation--a survey.

Chen, J., Zhang, Z., & Wu, F. J. I. j. o. p. r. (2021). A data-driven method for enhancing the image-based automatic inspection of IC wire bonding defects. *59*(16), 4779-4793.

Chen, S.-H., & Tsai, C.-C. J. A. e. i. (2021). SMD LED chips defect detection using a YOLOv3-dense model. *47*, 101255.

Cheng, Q., Ihalage, A. A., Liu, Y., & Hao, Y. J. I. A. (2020). Compressive sensing radar imaging with convolutional neural networks. *8*, 212917-212926.

Cheng, Z., Xiong, Z., Chen, C., & Liu, D. (2019). Light Field Super-Resolution: A Benchmark. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,

Civcisa, G., & Leemet, T. (2015). 3D surface roughness parameters of nanostructured coatings with application in the aerospace industry. Applied Mechanics and Materials,

Clark, S. R., & Greivenkamp, J. E. J. P. e. (2002). Ball tip–stylus tilt correction for a stylus profilometer. *26*(4), 405-411.

Dai, G., Pohlenz, F., Danzebrink, H.-U., Hasche, K., Wilkening, G. J. M. S., & Technology. (2004). Improving the performance of interferometers in metrological scanning probe microscopes. *15*(2), 444.

Dansereau, D. G., Pizarro, O., & Williams, S. B. (2013). Decoding, calibration and rectification for lenselet-based plenoptic cameras. Proceedings of the IEEE conference on computer vision and pattern recognition,

Deng, J., Dong, W., Socher, R., Li, L. J., & Li, F. F. (2009). ImageNet: a Large-Scale Hierarchical Image Database. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA,

Dong, C., Loy, C. C., He, K., & Tang, X. (2014). Learning a deep convolutional network for image super-resolution. European conference on computer vision,

Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis machine intelligence*, *38*(2), 295-307.

Ekberg, P., Su, R., & Leach, R. J. O. E. (2017). High-precision lateral distortion measurement and correction in coherence scanning interferometry using an arbitrary surface. *25*(16), 18703-18712.

Faber, C., Olesch, E., Krobot, R., & Häusler, G. (2012). Deflectometry challenges interferometry: the competition gets tougher! Interferometry XVI: techniques and analysis,

Fan, Y., Yu, J., & Huang, T. S. (2018). Wide-activated deep residual networks based restoration for bpg-compressed images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,

Farrugia, R. A., Galea, C., & Guillemot, C. J. I. J. o. S. T. i. S. P. (2017). Super resolution of light field images using linear subspace projection of patch-volumes. *11*(7), 1058-1071.

Fu, S., Kor, W. S., Cheng, F., & Seah, L. K. J. P. C. (2020). In-situ measurement of surface roughness using chromatic confocal sensor. *94*, 780-784.

Fürsich, M. (2019). Vehicle vision system with light field monitor. In: Google Patents.

Gao, W., Haitjema, H., Fang, F. Z., Leach, R. K., Cheung, C. F., Savio, E., & Linares, J. M. (2019). On-machine and in-process surface metrology for precision manufacturing. *Cirp Annals-Manufacturing Technology*, *68*(2), 843-866. https://doi.org/10.1016/j.cirp.2019.05.005

Gapinski, B., Wieczorowski, M., Marciniak-Podsadna, L., Dybala, B., & Ziolkowski, G. (2014). Comparison of different method of measurement geometry using CMM, optical scanner and computed tomography 3D. *Procedia Engineering*, *69*, 255-262.

Gensler, A., Henze, J., Sick, B., & Raabe, N. (2016). Deep Learning for solar power forecasting—An approach using AutoEncoder and LSTM Neural Networks. 2016 IEEE international conference on systems, man, and cybernetics (SMC),

Georgiev, T. G., & Lumsdaine, A. (2012). Super-resolution with the focused plenoptic camera. In: Google Patents.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. Proceedings of the fourteenth international conference on artificial intelligence and statistics,

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems,

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative Adversarial Networks. *arXiv e-prints*, arXiv:1406.2661. https://ui.adsabs.harvard.edu/abs/2014arXiv1406.2661G

Gotoh, T., & Okutomi, M. (2004). Direct super-resolution and registration using raw CFA images. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.,

Hahn, I., Weilert, M., Wang, X., & Goullioud, R. J. R. o. S. I. (2010). A heterodyne interferometer for angle metrology. *81*(4), 045103.

Hao, Q., Wang, S., Hu, Y., Cheng, H., Chen, M., & Li, T. J. A. o. (2016). Virtual

interferometer calibration method of a non-null interferometer for freeform surface measurements. *55*(35), 9992-10001.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition,

Heber, S., & Pock, T. (2014). Shape from light field meets robust PCA. European Conference on Computer Vision,

Hinman, S. S., McKeating, K. S., & Cheng, Q. J. A. c. (2017). Plasmonic sensing with 3D printed optics. *89*(23), 12626-12630.

Honauer, K., Johannsen, O., Kondermann, D., & Goldluecke, B. (2016). A dataset and evaluation methodology for depth estimation on 4D light fields. Asian Conference on Computer Vision,

Hornbuckle, B. C., Williams, C. L., Dean, S. W., Zhou, X., Kale, C., Turnage, S. A., . . . Solanki, K. N. (2020). Stable microstructure in a nanocrystalline copper–tantalum alloy during shock loading. *Communications Materials*, *1*(1), 1-6.

Howe, C. L., Quicke, P., Song, P., Jadan, H. V., Dragotti, P. L., & Foust, A. J. J. b. (2020). Comparing volumetric reconstruction algorithms for light field imaging of high signal-to-noise ratio neuronal calcium transients.

Hua, X., & Jia, S. (2020). 3D Live Cell Imaging Using High-Resolution Fourier Light-Field Microscopy. Frontiers in Optics,

Huang, L., Choi, H., Zhao, W., Graves, L. R., & Kim, D. W. J. O. L. (2016). Adaptive interferometric null testing for unknown freeform optics metrology. *41*(23), 5539-5542.

Im, S., Jeon, H.-G., Lin, S., & Kweon, I. S. J. a. p. a. (2019). Dpsnet: End-to-end deep plane sweep stereo.

Ivakhnenko, A. G. e., Ivakhnenko, A. G., Lapa, V. G. e., & Lapa, V. G. (1967). *Cybernetics and forecasting techniques* (Vol. 8). American Elsevier Publishing Company.

Jeon, H.-G., Park, J., Choe, G., Park, J., Bok, Y., Tai, Y.-W., & So Kweon, I. (2015). Accurate depth map estimation from a lenslet light field camera. Proceedings of the IEEE conference on computer vision and pattern recognition,

Jin, J., Hou, J., Chen, J., & Kwong, S. (2020). Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. Proceedings of the IEEE/CVF Conference on Computer Vision

and Pattern Recognition,

Jin, J., Hou, J., Chen, J., Zeng, H., Kwong, S., Yu, J. J. I. T. o. P. A., & Intelligence, M. (2020). Deep coarse-to-fine dense light field reconstruction with flexible sampling and geometry-aware fusion.

Jin, J., Hou, J., Yuan, H., & Kwong, S. (2020). *Learning Light Field Angular Super-Resolution via a Geometry-Aware Network* AAAI,

Johannsen, O., Honauer, K., Goldluecke, B., Alperovich, A., Battisti, F., Bok, Y., . . . Diebold, M. (2017). A taxonomy and evaluation of dense light field depth estimation algorithms. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. European conference on computer vision,

Kalantari, N. K., Wang, T.-C., & Ramamoorthi, R. J. A. T. o. G. (2016). Learning-based view synthesis for light field cameras. *35*(6), 1-10.

Kim, J., Jung, J.-H., Jeong, Y., Hong, K., & Lee, B. J. O. e. (2014). Real-time integral imaging system for light field microscopy. *22*(9), 10210-10220.

Kim, J., Kwon Lee, J., & Mu Lee, K. (2016). Accurate image super-resolution using very deep convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition,

Kim, S. S., You, B. H., Choi, H., Berkeley, B. H., Kim, D. G., & Kim, N. D. (2009). 31.1: Invited Paper: World's First 240Hz TFT ‐ LCD Technology for Full ‐ HD LCD ‐ TV and Its Application to 3D Display. SID Symposium Digest of Technical Papers,

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv e-prints*, arXiv:1412.6980. https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6980K

Lai, W.-S., Huang, J.-B., Ahuja, N., & Yang, M.-H. (2017). Deep laplacian pyramid networks for fast and accurate super-resolution. Proceedings of the IEEE conference on computer vision and pattern recognition,

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., . . . Wang, Z. (2017). Photo-realistic single image super-resolution using a generative adversarial network. Proceedings of the IEEE conference on computer vision and pattern recognition,

Lee, D.-H., Cho, N.-G. J. M. s., & technology. (2012). Assessment of surface profile

data acquired by a stylus profilometer. *23*(10), 105601.

Lei, L., Deng, L., Fan, G., Cai, X., Li, Y., & Li, T. J. M. (2014). A 3D micro tactile sensor for dimensional metrology of micro structure with nanometer precision. *48*, 155-161.

Levoy, M., Ng, R., Adams, A., Footer, M., & Horowitz, M. (2006). Light field microscopy. In *ACM SIGGRAPH 2006 Papers* (pp. 924-934).

Li, D., Cheung, C. F., Ren, M., Whitehouse, D., & Zhao, X. J. O. l. (2015). Disparity pattern-based autostereoscopic 3D metrology system for in situ measurement of microstructured surfaces. *40*(22), 5271-5274.

Li, D., Cheung, C. F., Ren, M., Zhou, L., & Zhao, X. (2014). Autostereoscopy-based three-dimensional on-machine measuring system for micro-structured surfaces. *Optics Express*, *22*(21), 25635-25650.

Li, D., Wang, B., Tong, Z., Blunt, L., & Jiang, X. J. T. I. J. o. A. M. T. (2019). On-machine surface measurement and applications for ultra-precision machining: a state-of-the-art review. *104*(1), 831-847.

Li, J., Lu, M., & Li, Z.-N. J. I. T. o. I. P. (2015). Continuous depth map reconstruction from light fields. *24*(11), 3257-3265.

Li, L., & Allen, Y. Y. J. A. o. (2012). Design and fabrication of a freeform microlens array for a compact large-field-of-view compound-eye camera. *51*(12), 1843-1852.

Li, W., Hou, D., Luo, Z., & Mao, X. J. O. (2021). 3D measurement system based on divergent multi-line structured light projection, its accuracy analysis. *231*, 166396.

Liang, C.-K., & Ramamoorthi, R. J. A. T. o. G. (2015). A light transport framework for lenslet light field cameras. *34*(2), 1-19.

Lim, J., Ok, H., Park, B., Kang, J., & Lee, S. (2009). Improving the spatail resolution based on 4D light field data. 2009 16th IEEE International Conference on Image Processing (ICIP),

Lin, X., Wu, J., Zheng, G., & Dai, Q. J. B. o. e. (2015). Camera array based light field microscopy. *6*(9), 3179-3189.

Lindeberg, T. (2012). Scale invariant feature transform.

Lippmann, G. J. J. P. T. A. (1908). Epreuves reversibles donnant la sensation du relief. *7*(1), 821-825.

Liu, Y., Huang, S., Zhang, Z., Gao, N., Gao, F., & Jiang, X. J. S. r. (2017). Full-field 3D shape measurement of discontinuous specular objects by direct phase measuring deflectometry. *7*(1), 1-8.

Maldonado, A. V., Su, P., & Burge, J. H. J. A. o. (2014). Development of a portable deflectometry system for high spatial resolution surface measurements. *53*(18), 4023-4032.

Martínez-Corral, M., Javidi, B. J. A. i. O., & Photonics. (2018). Fundamentals of 3D imaging and displays: a tutorial on integral imaging, light-field, and plenoptic systems. *10*(3), 512-566.

Marwah, K., Wetzstein, G., Bando, Y., & Raskar, R. (2013). Compressive light field photography using overcomplete dictionaries and optimized projections. *32*(4), 1-12.

Meng, N., Ge, Z., Zeng, T., & Lam, E. Y. J. I. A. (2020). LightGAN: A Deep Generative Model for Light Field Reconstruction. *8*, 116052-116063.

Mian, S. H., & Al-Ahmari, A. (2014). New developments in coordinate measuring machines for manufacturing industries. *International Journal of Metrology Quality Engineering*, *5*(1), 101.

Mitra, K., & Veeraraghavan, A. (2012, 16-21 June 2012). Light field denoising, light field superresolution and stereo camera based refocussing using a GMM light field patch prior. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops,

Mousnier, A., Vural, E., & Guillemot, C. J. a. p. a. (2015). Partial light field tomographic reconstruction from a fixed-camera focal stack.

Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., & Hanrahan, P. (2005). *Light field photography with a hand-held plenoptic camera* Stanford University].

Overbeck, R. S., Erickson, D., Evangelakos, D., Pharr, M., & Debevec, P. J. A. T. o. G. (2018). A system for acquiring, processing, and rendering panoramic light field stills for virtual reality. *37*(6), 1-15.

Park, J., Kim, J., Hong, J., Lee, H., Lee, Y., Cho, S., . . . Ko, H. (2018). Tailoring force sensitivity and selectivity by microstructure engineering of multidirectional electronic skins. *NPG Asia Materials*, *10*(4), 163-176.

Peleg, T., Szekely, P., Sabo, D., & Sendik, O. (2019). Im-net for high resolution video frame interpolation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Perng, D.-B., Chou, C.-C., & Lee, S.-M. J. T. I. J. o. A. M. T. (2007). Design and development of a new machine vision wire bonding inspection system. *34*(3), 323-334.

Pförtner, A., & Schwider, J. J. A. o. (2003). Red-green-blue interferometer for the metrology of discontinuous structures. *42*(4), 667-673.

Rahim, R., & Nadeem, S. (2018). End-to-end trained CNN encoder-decoder networks for image steganography. Proceedings of the European Conference on Computer Vision (ECCV) Workshops,

Rerabek, M., & Ebrahimi, T. (2016). New light field image dataset. 8th International Conference on Quality of Multimedia Experience (QoMEX),

Richardson, F., Reynolds, D., & Dehak, N. J. I. s. p. l. (2015). Deep neural network approaches to speaker and language recognition. *22*(10), 1671-1675.

Robertson, T., Hutchins, D., Billson, D., Rakels, J., & Schindel, D. J. U. (2002). Surface metrology using reflected ultrasonic signals in air. *39*(7), 479-486.

Rosenblatt, F. (1961). *Principles of neurodynamics. perceptrons and the theory of brain mechanisms*.

Rossi, M., & Frossard, P. (2017). Graph-based light field super-resolution. 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP),

Ruderman, A., Rabinowitz, N. C., Morcos, A. S., & Zoran, D. J. a. p. a. (2018). Pooling is neither necessary nor sufficient for appropriate deformation stability in CNNs.

Shi, H., Xu, M., & Li, R. J. I. T. o. S. G. (2017). Deep learning for household load forecasting—A novel pooling deep RNN. *9*(5), 5271-5280.

Shi, J., Jiang, X., & Guillemot, C. J. I. T. o. I. P. (2019). A framework for learning depth from a flexible subset of dense and sparse light field views. *28*(12), 5867-5880.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., . . . Bolton, A. J. n. (2017). Mastering the game of go without human knowledge. *550*(7676), 354-359.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv e-prints*, arXiv:1409.1556.

Tao, M. W., Hadap, S., Malik, J., & Ramamoorthi, R. (2013). Depth from combining defocus and correspondence using light-field cameras. Proceedings of the IEEE International Conference on Computer Vision,

Tetsuka, H., & Shin, S. R. J. J. o. m. c. B. (2020). Materials and technical innovations in 3D printing in biomedical applications. *8*(15), 2930-2950.

Tian, Z., Gao, F., Jin, Z., & Zhao, X. J. T. I. J. o. A. M. T. (2009). Dimension measurement of hot large forgings with a novel time-of-flight system. *44*(1-2), 125-132.

Vaish, V., Levoy, M., Szeliski, R., Zitnick, C. L., & Kang, S. B. (2006). Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06),

Vandewalle, P., Krichane, K., Alleysson, D., & Süsstrunk, S. (2007). Joint demosaicing and super-resolution imaging from a set of unregistered aliased images. Digital Photography III,

Vieroth, R., Loher, T., Seckel, M., Dils, C., Kallmayer, C., Ostmann, A., & Reichl, H. (2009). Stretchable circuit board technology and application. 2009 International Symposium on Wearable Computers,

Wang, H., Lee, W. B., Chan, J., & To, S. J. A. T. E. (2015). Numerical and experimental analysis of heat transfer in turbulent flow channels with two-dimensional ribs. *75*, 623-634.

Wang, T.-C., Efros, A. A., & Ramamoorthi, R. (2015). Occlusion-aware depth estimation using light-field cameras. Proceedings of the IEEE International Conference on Computer Vision,

Wanner, S., & Goldluecke, B. (2012). Spatial and Angular Variational Super-Resolution of 4D Light Fields. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid, *Computer Vision – ECCV 2012* Berlin, Heidelberg.

Wanner, S., & Goldluecke, B. (2014). Variational Light Field Analysis for Disparity Estimation and Super-Resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(3), 606-619. https://doi.org/10.1109/TPAMI.2013.147

Wilburn, B., Joshi, N., Vaish, V., Talvala, E.-V., Antunez, E., Barth, A., . . . Levoy, M. (2005). High performance imaging using large camera arrays. In *ACM SIGGRAPH 2005 Papers* (pp. 765-776).

Winnek, C. D. F. (1936). Apparatus for making a composite stereograph. In: Google Patents.

Wu, G., Liu, Y., Dai, Q., & Chai, T. J. I. T. o. I. P. (2019). Learning sheared EPI structure for light field reconstruction. *28*(7), 3261-3273.

Wu, G., Zhao, M., Wang, L., Dai, Q., Chai, T., & Liu, Y. (2017). Light field reconstruction using deep convolutional network on EPI. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Wu, Q., Fan, C., Li, Y., Li, Y., Hu, J. J. M. T., & Applications. (2020). A novel perceptual loss function for single image super-resolution. *79*(29), 21265-21278.

Wyant, J. C. (2002). White light interferometry. Holography: A Tribute to Yuri Denisyuk and Emmett Leith,

Xu, Q., Zhu, Z., Ge, H., Zhang, Z., Zang, X. J. C., & Medicine, M. M. i. (2021). Effective Face Detector Based on YOLOv5 and Superresolution Reconstruction. *2021*.

Yao, G., Xu, L., Cheng, X., Li, Y., Huang, X., Guo, W., . . . Wu, H. J. A. F. M. (2020). Bioinspired triboelectric nanogenerators as self‐powered electronic skin for robotic tactile sensing. *30*(6), 1907312.

Ye, Q., Ong, S., Han, X. J. I. j. o. i. s., & technology. (2000). A stereo vision system for the inspection of IC bonding wires. *11*(4), 254-262.

Yeung, H. W. F., Hou, J., Chen, J., Chung, Y. Y., & Chen, X. (2018). Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues. Proceedings of the European Conference on Computer Vision (ECCV),

Yin, Q., Xu, B., Yin, G., Gui, P., Xu, W., & Tang, B. J. J. o. N. (2018). Surface profile measurement and error compensation of triangular microstructures employing a stylus scanning system. *2018*.

Yoon, Y., Jeon, H.-G., Yoo, D., Lee, J.-Y., & Kweon, I. S. J. I. S. P. L. (2017). Light-field image super-resolution using convolutional neural network. *24*(6), 848-852.

Yoon, Y., Jeon, H.-G., Yoo, D., Lee, J.-Y., & So Kweon, I. (2015). Learning a deep convolutional network for light-field image super-resolution. Proceedings of the IEEE international conference on computer vision workshops,

Yu, Z., Guo, X., Lin, H., Lumsdaine, A., & Yu, J. (2013). Line assisted light field triangulation and stereo matching. Proceedings of the IEEE International Conference on Computer Vision,

Yu, Z., Yu, J., Lumsdaine, A., & Georgiev, T. (2012). An analysis of color demosaicing in plenoptic cameras. 2012 IEEE Conference on Computer Vision and Pattern Recognition,

Zhang, S., Lin, Y., & Sheng, H. (2019). Residual networks for light field image super-

resolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Zhang, S., Sheng, H., Li, C., Zhang, J., Xiong, Z. J. C. V., & Understanding, I. (2016). Robust depth estimation for light field via spinning parallelogram operator. *145*, 148-159.

Zhang, X., Huang, R., Liu, K., Kumar, A. S., & Shan, X. J. P. E. (2018). Rotating-tool diamond turning of Fresnel lenses on a roller mold for manufacturing of functional optical film. *51*, 445-457.

Zhao, C., & Burge, J. H. J. A. o. (2001). Vibration-compensated interferometer for surface metrology. *40*(34), 6215-6222.

Zhao, H., Gallo, O., Frosio, I., & Kautz, J. J. I. T. o. c. i. (2016). Loss functions for image restoration with neural networks. *3*(1), 47-57.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. Proceedings of the IEEE conference on computer vision and pattern recognition,

Zhou, P., Kong, L., Sun, X., & Xu, M. J. I. P. J. (2020). Three-dimensional measurement of specular surfaces based on the light field. *12*(5), 1-13.

Zhu, H., Wang, Q., & Yu, J. J. I. J. o. S. T. i. S. P. (2017). Occlusion-model guided antiocclusion depth estimation in light field. *11*(7), 965-978.

Zou, X., Zhao, X., Li, G., Li, Z., & Sun, T. J. T. I. J. o. A. M. T. (2017). Non-contact on-machine measurement using a chromatic confocal probe for an ultra-precision turning machine. *90*(5), 2163-2172.