2022

Enhancing Multi-View 3D-Reconstruction Using Multi-Frame Super Resolution

Michael Lee
University of Windsor

Enhancing Multi-View 3D-Reconstruction Using Multi-Frame Super Resolution

by

Michael Lee

A Dissertation
Submitted to the Faculty of Graduate Studies
through the Department of Electrical and Computer Engineering
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy at the
University of Windsor

Windsor, Ontario, Canada
2022

Enhancing Multi-View 3D-Reconstruction Using Multi-Frame Super Resolution

by

Michael Lee

APPROVED BY:

S. Sfarra, External Examiner
University of LAquila

R. Caron,
Department of Mathematics and Statistics

J. Wu
Department of Electrical and Computer Engineering

M. Khalid
Department of Electrical and Computer Engineering

B. Shahrrava, Co-Advisor
Department of Electrical and Computer Engineering

R. Maev, Co-Advisor
Department of Physics

September 1, 2022

Author's Declaration of Originality

I hereby certify that I am the sole author of this dissertation and that no part of this dissertation has been published or submitted for publication.

I certify that, to the best of my knowledge, my dissertation does not infringe upon anyones copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my dissertation, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my dissertation and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my dissertation, including any final revisions, as approved by my dissertation committee and the Graduate Studies office, and that this dissertation has not been submitted for a higher degree to any other University or Institution.

Abstract

Multi-view stereo is a popular method for 3D-reconstruction. Super resolution is a technique used to produce high resolution output from low resolution input. Since the quality of 3D-reconstruction is directly dependent on the input, a simple path is to improve the resolution of the input.

In this dissertation, we explore the idea of using super resolution to improve 3D-reconstruction at the input stage of the multi-view stereo framework. In particular, we show that multi-view stereo when combined with multi-frame super resolution produces a more accurate 3D-reconstruction.

The proposed method utilizes images with sub-pixel camera movements to produce high resolution output. This enhanced output is fed through the multi-view stereo pipeline to produce an improved 3D-model. As a performance test, the improved 3D-model is compared to similarly generated 3D-reconstructions using bicubic and single image super resolution at the input stage of the multi-view stereo framework. This is done by comparing the point clouds of the generated models to a reference model using the metrics: average, median, and max distance. The model that has the metrics that are closest to the reference model is considered to be the better model.

The overall experimental results show that the generated models, using our technique, have point clouds with average mean, median, and max distances of 4.3%, 8.8%, and 6% closer to the reference model, respectively. This indicates an improvement in 3D-reconstruction using our technique. In addition, our technique has a significant speed advantage over the single image super resolution analogs being at

least 6.8x faster.

The use of multi-frame super resolution in conjunction with the multi-view stereo framework is a practical solution for enhancing the quality of 3D-reconstruction and shows promising results over single image up-sampling techniques.

Contents

List of Figures

List of Tables

1 Introduction

Computer vision is a technological problem which seeks to provide machines with the ability to see, interpret, and predict an understanding of the physical world through digital mediums such as video and images. These mediums can provide data for reconstructing 3D geometry for modeling real world objects and scenes. Stereo Vision was one of the first methods used to accomplish this by using two views, taken from slightly different overlapping perspectives, which provided the missing depth information. Later, improvements to 3D-reconstruction using more than two views was developed to overcome problems such as reduction of matching ambiguity, occlusions, and depth error [18, 30, 58, 94, 95].

The use of more than two views is called multi-view stereo (MVS). This framework builds 3D-reconstructions using a set of input images and calculates camera parameters from those images using structure from motion (SFM) [28]. One of the key features is that the process only requires images for reconstruction. However, the captured scene must be static. The steady improvement in the quality, resolution, and cost of digital cameras makes this a favourable cost-effective solution for quality 3D-reconstruction.

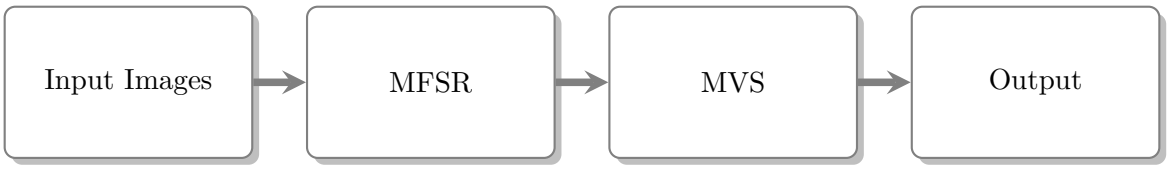
Since the quality of 3D-reconstruction is directly dependent on the input, a simple

path is to improve the input. Super resolution (SR) is a technique used to produce high resolution output from low resolution input. Previous work, concentrated on using super resolution to improve the input images [77]. Their work focused on using a single image super resolution (SISR) algorithm.

In this dissertation, we explore the idea of using multi-frame super resolution (MFSR) to improve 3D-reconstruction at the input stage of the multi-view stereo framework. In particular, we show that multi-view stereo when combined with MFSR produces a more accurate reconstruction. We address limitations of prior work by focusing on indoor small scale objects and using a simple MFSR based algorithm for super resolution, which does not require training and has inherent noise reduction.

Proposed Framework:

Multi-view Stereo with Multi-frame Super-Resolution



- MFSR applied to the input stage of the MVS framework will yield higher 3D-reconstruction accuracy with a particular focus on indoor small scale objects

This document is structured as follows:

Chapter 2 : Contains background and related work. Covers the following topics: Camera model, traditional stereo vision, multi-view stereo, super resolution, and related work combining vision frameworks with super resolution.

Chapter 3 : Method and hardware. This chapter details the experimental methods and techniques used, as well as, describes the hardware specifications of the vision system and its 3 major sub-systems: Image acquisition, linear motion, precision measurement.

Chapter 4 : Experiments. This chapter demonstrates some of the capabilities of the vision system and tests obtaining metrics from a scene, average distance per step and sub-pixel motion of the linear stage, super resolution comparisons, and 3D-model generation.

Chapter 5 : Conclusion and future work.

2 Background and Related Work

2.1 Camera Model

This subsection covers the basic camera model and summarizes the mathematical representation of intrinsic, extrinsic, and lens distortion parameters from [21, 44, 139].

2.1.1 The Pinhole Camera Model

A **Pinhole Camera** consists of a light proof chamber with a tiny aperture that projects an inverted image on the side opposite of the aperture.

The oldest known description of this device, also known as a Camera Obscura, dates back to the 4th century BCE in the writings of the Chinese philosopher Mozi(Mo-tzu) from the Han dynasty [39, 88, 108, 123].

The **Pinhole Camera model** is a representation of the projection process that traces light rays from the camera scene through the aperture to the image plane. The image produced is inverted both horizontally and vertically, that is the image will appear upside down and left to right.

This process can be viewed as a 3-Dimensional (3D) to 2-Dimensional (2D) transformation or as a projection from 3D-space onto 2D-space.

In Figure 1, we see light rays from world point $P(X, Y, Z)$ passing through the camera

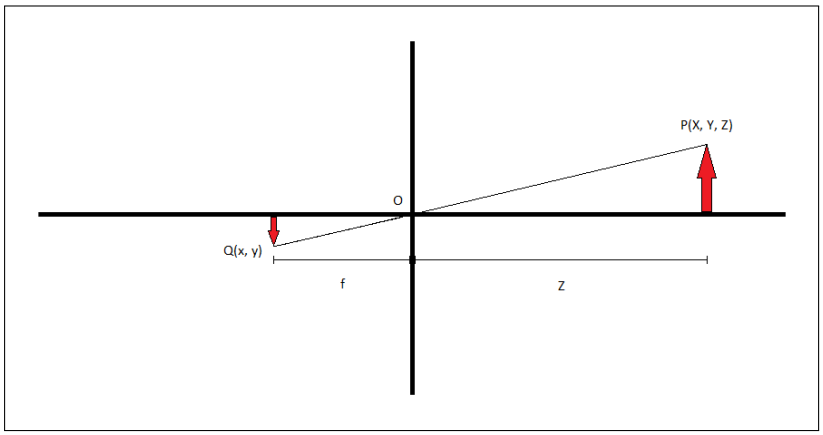


Figure 1: Pinhole Camera Model. 3D to 2D transformation a projection from 3-space into 2-space. $P(X, Y, Z)$ projected onto image plane at $Q(x, y)$.

aperture at O to the image plane at point $Q(x, y)$. The **depth**, Z, is the distance from the world point to the camera at O. The **focal length**, f, is the distance from the camera at O to the image plane point $Q(x, y)$.

Using similar triangles we get the following relationships for 3-space World point $P(X, Y, Z)$ and image plane point Q(x,y):

$$\frac{Z}{f} = \frac{X}{x} \Leftrightarrow x = f\frac{X}{Z}$$
$$\frac{Z}{f} = \frac{Y}{y} \Leftrightarrow y = f\frac{Y}{Z}.$$

Combining the two we get,

$$Q = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f\frac{X}{Z} \\ f\frac{Y}{Z} \end{bmatrix} \tag{1}$$

2.1.2 Camera Matrix Model (Intrinsic)

The Camera Matrix Model is the formal method for mapping camera coordinates into image coordinates. It contains all the parameters in matrix form accounting for parameters like focal length, translation, and skew. The following builds the **Camera Matrix Model** from the **Pinhole Camera model**.

The optical center of the camera needs to be aligned with the image plane coordinates since images have their origin, $(0, 0)$, in a corner region so it is necessary to translate. Let c_x and c_y be the translation offset. So, the projection map becomes:

$$Q = \begin{bmatrix} x \\ \\ y \end{bmatrix} = \begin{bmatrix} f\frac{X}{Z} + c_x \\ \\ f\frac{Y}{Z} + c_y \end{bmatrix} \tag{2}$$

Positions in camera coordinates are represented by units of measure such as millimeters whereas images use pixels. To convert from camera units into image coordinates we introduce variables k_x and k_y which are expressed in pixels per unit measure (eg. $\frac{pixels}{mm}$). Equal values for k_x and k_y indicate that the sensor has square pixels but there is no guarantee that the aspect ratio will be equal to one, so it is possible that the values for each are different.

$$Q = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f_x \frac{X}{Z} + c_x \\ f_y \frac{Y}{Z} + c_y \end{bmatrix}, \tag{3}$$

where $f_x = fk_x$ and $f_y = fk_y$.

We introduce a skew factor, α, that accounts for any shear present in the coordinate system possibly occurring when the optical axis is not orthogonal to the image plane which causes a translation in x of αY.

$$Q = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f_x \frac{X}{Z} + c_x + \alpha Y \\ f_y \frac{Y}{Z} + c_y \end{bmatrix} \tag{4}$$

A convenient way to express Equation (4), since it permits the use of matrix multiplication to represent the complete transformation.

$$Q_h = \begin{bmatrix} x \\ y \\ Z \end{bmatrix} = \begin{bmatrix} f_x X + c_x Z + \alpha Y Z \\ f_y Y + c_y Z \\ Z \end{bmatrix} = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \tag{5}$$

where $s = \alpha Z$.

The familiar intrinsic camera matrix, K, can be seen in Equation (5) which can be re-written as:

$$Q_h = \begin{bmatrix} x \\ y \\ Z \end{bmatrix} = \underbrace{\begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}}_{K} P_h \tag{6}$$

The **Intrinsic Camera Matrix** parameters consisting of focal length, skew/distortion/scale factor, and image center can be clearly seen. The following is written in it's recognizable 3x3-matrix form:

$$K = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{7}$$

The skew/shear distortion factor for modern cameras is normally very close to zero. So, it is customary to accept this value as zero. However, under unusual circumstances this factor could be non-zero in cases such as capturing an image of an image.

2.1.3 Extrinsic Parameters(Localization)

The intrinsic camera matrix maps points from the 3D-camera-space into the 2D-image-plane. We would like to be able to work with the world reference. So, we need to build a transformation from 3D-world coordinates into camera coordinates. The transformation is accomplished by rotation, R, and translation, T, from world coordinates into camera coordinates. Given some world point, P_w, we can apply the following transformation to move into camera coordinates:

$$P_h = \begin{bmatrix} R_{3x3} & T_{3x1} \\ 0_{1x3} & 1 \end{bmatrix} P_w \tag{8}$$

Combining both the intrinsic and extrinsic matrices, Equations (6) and (8), we get:

$$Q_h = \begin{bmatrix} x \\ y \\ Z \end{bmatrix} = \underbrace{\begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}}_{K} \begin{bmatrix} I_{3x3} & \vec{0}_{3x1} \end{bmatrix} \underbrace{\begin{bmatrix} R_{3x3} & T_{3x1} \\ 0_{1x3} & 1 \end{bmatrix}}_{Rot\_Trans} P_w \tag{9}$$

2.1.4 Perspective Projection Matrix M

The intrinsic and extrinsic relationship expressed in Equation (9) can be simplified using the following Projection matrix M_{3x4}:

$$Q_h = MP_w, \tag{10}$$

where

$$M = \underbrace{\begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}}_{K} \left[\underbrace{\begin{matrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{matrix}}_{R} \middle| \underbrace{\begin{matrix} t_x \\ t_y \\ t_z \end{matrix}}_{T} \right] = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \end{bmatrix}$$

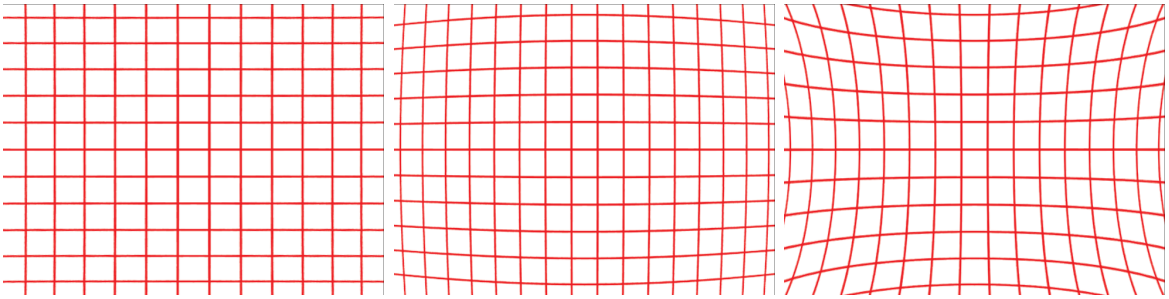
Camera Calibration, also known as Camera Pose Estimation, Resectioning, or Geometric Camera Calibration, is an operation that approximates intrinsic and extrinsic parameters of a camera.

2.1.5 Camera Lens Distortion

Pinhole cameras require the size of the aperture to be as small as possible to create a clear and sharp image. If the aperture is too large, incident rays of light cause the image to become blurry. But, if the aperture is too small, not enough light will

strike the sensor and will produce an image that is dark and grainy. The solution is to focus the light using a lens. This will produce an image that is bright and clear. However, using a lens will introduce distortion effects. The distortions can be modeled mathematically and to reduce image distortion caused by the lens. Most lens models accommodate for the following types of distortion:

- **Radial Distortion**: The unequal bending of light at the edges of the lens versus the center causing straight lines to appear curved. These curved lines can be classified as either **Barrel** or **Pincushion** distortion.
- **Tangential Distortion**: This type of distortion makes the image appear stretched or tilted and is caused by the angle of the lens with respect to the image sensor.



(a) No Distortion (b) Barrel (c) Pincushion

Figure 2: Lens Distortion.

The general model used for lens distortion is as follows:

$$\begin{aligned} x_d &= x + \hat{x}(1 + k_1 r^2 + k_2 r^4 + k_3 r^6 + \ldots) + [p_1(r^2 + 2\hat{x}^2) + 2p_2\hat{x}\hat{y}](1 + p_3 r^2 + \ldots) \\ y_d &= y + \underbrace{\hat{y}(1 + k_1 r^2 + k_2 r^4 + k_3 r^6 + \ldots)}_{\text{Radial Distortion}} + \underbrace{[p_2(r^2 + 2\hat{y}^2) + 2p_1\hat{x}\hat{y}](1 + p_3 r^2 + \ldots)}_{\text{Tangential Distortion}} \end{aligned}$$

which is normally simplified to:

$$x_d = x + \hat{x}(1 + k_1 r^2 + k_2 r^4 + k_3 r^6 + \ldots) + [p_1(r^2 + 2\hat{x}^2) + 2p_2\hat{x}\hat{y}] \quad (11)$$

$$y_d = y + \underbrace{\hat{y}(1 + k_1 r^2 + k_2 r^4 + k_3 r^6 + \ldots)}_{\text{Radial Distortion}} + \underbrace{[p_2(r^2 + 2\hat{y}^2) + 2p_1\hat{x}\hat{y}]}_{\text{Tangential Distortion}}, \quad (12)$$

where (x_d, y_d) is the distorted image coordinate, x and y are the projected undistorted image components from the world coordinates after rotation and translation, (x_c, y_c) is the radial optical center, $r^2 = \sqrt{\hat{x}^2 + \hat{y}^2}$, $\hat{x} = (x - x_c)$, and $\hat{y} = (y - y_c)$. (See [21, 71, 139] for more detail.)

2.2 Traditional Stereo Vision

Stereo Vision uses two cameras, with a known separation distance, to capture an image pair. The displacement between the two views is known as **disparity**.

Figure 3: Stereo Vision: Tea Cup

Intuitively, our brains naturally interpret the inverse relationship between **disparity** and **depth** (the distance from the subject to the camera plane). This is easily realized with Anaglyphs where two images with a slight perspective displacement are superimposed on top of each other, in different colours (red and blue), producing a stereo effect giving the perception of 3D.

Figure 4: Anaglyph: Tea Cup

2.2.1 Triangulation

Captured images can be considered as projections of a scene onto a plane, so they only contain 2D information and as a consequence depth information is lost. To recover this information we can use triangulation, this process matches identical points from one image to the other, to estimate the 3D positions.

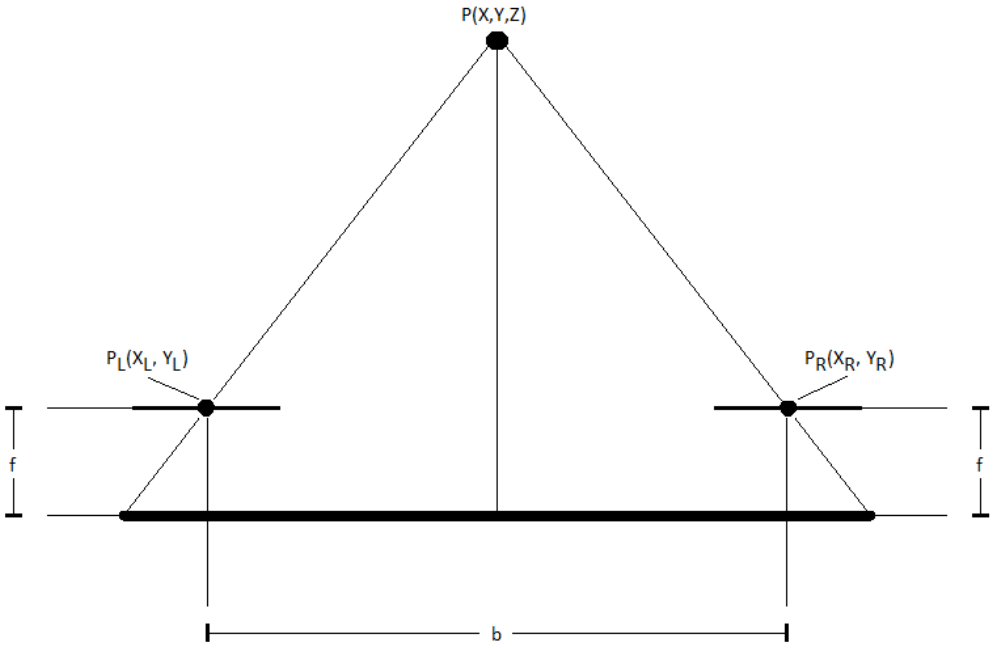


Figure 5: Scene to Stereo Image Plane, where $P(X, Y, Z)$ is the scene point and $\{P_L(X_L, Y_L),\ P_R(X_R, Y_R)\}$ are the corresponding scene left and right image plane points, respectively.

$$
\begin{aligned}
X &= \frac{b(X_L + X_R)}{2(X_L - X_R)} \\
Y &= \frac{b(Y_L + Y_R)}{2(X_L - X_R)} \\
Z &= \frac{bf}{(X_L - X_R)}
\end{aligned}
\tag{13}
$$

Figure (5) with equation set (13) illustrate the relationship for the reconstruction of the point $P(X, Y, Z)$ given corresponding Left and Right image points $P_L(X_L, Y_L)$, $P_R(X_R, Y_R)$, base length b, and focal length f(in pixels).

From equation set (13) we can clearly see the inverse relationship of **disparity** and **depth** by examining the formula for Z.

Sometimes it is useful to calculate the distance between two scene points, so, the well known formula (14) can be used.

$$
Dist = \sqrt{(\Delta X)^2 + (\Delta Y)^2 + (\Delta Z)^2} \,, \tag{14}
$$

where

$$\begin{aligned} \Delta X &= X_1 - X_2 \\ \Delta Y &= Y_1 - Y_2 \\ \Delta Z &= Z_1 - Z_2 \end{aligned} \tag{15}$$

To accommodate for multiple scene points we modify the notation of the equations from (13) to get the equation set (16).

$$\begin{aligned} X_n &= \frac{b(X_{L_n} + X_{R_n})}{2(X_{L_n} - X_{R_n})} \\ Y_n &= \frac{b(Y_{L_n} + Y_{R_n})}{2(X_{L_n} - X_{R_n})} \\ Z_n &= \frac{bf}{(X_{L_n} - X_{R_n})} \end{aligned} \tag{16}$$

We also define the 3-space point P_n in the following way:

$$P_n = (X_n, Y_n, Z_n) \ , \tag{17}$$

where $\{X_n, Y_n, Z_n\}$ come from the image point computations of equation set (16).

2.3 Multi-View Stereo (MVS)

Traditional Stereo Vision is a good method for 3D-reconstruction but has limitations such as matching ambiguity, occlusions, and fixed working distance [18, 30, 58, 94, 95]. This is naturally improved by incorporating additional scene information through the use of more views [see Figure (6)]. The methodology of using more than two views is called: **Multi-View Stereo (MVS)**. This framework is a general description of techniques that use stereo correspondence from more than two images [28, 112, 120] and shares many principles with traditional Stereo Vision but differs with algorithms that can handle the larger variance in viewpoints [28]. It uses overlapping information from images taken from different viewpoints to aid in 3D-reconstruction.

Figure 6: Multi-View Stereo reconstruction from set of images (left) to 3D-model (right).

The noisy measurements of any given scene point can be made more robust by accounting for the overlapping information from the multi-view images through the use of redundant cues such as: texture, contours, shading, de-focus, and stereo correspondences.

The general pipeline for MVS is as follows (as illustrated in Figure 7): Image Acqui-

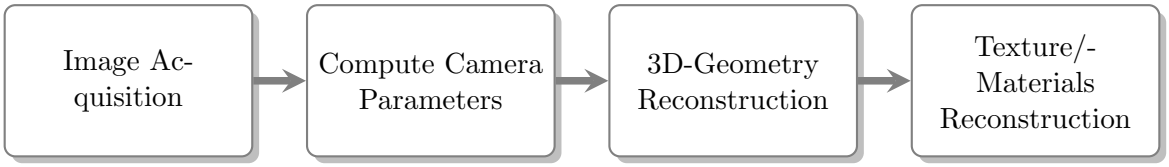


Figure 7: General Multi-View Stereo (MVS) Pipeline.

sition, Compute Camera Parameters, 3D-Geometry Reconstruction, Textures/Materials Reconstruction from the scene.

2.3.1 Image Acquisition

Multi-View Stereo utilizes a set of input images and estimates a reconstruction of the most likely 3D shape from the given information. The **Image Acquisition** process is a passive and fast method for accurately capturing content. Technical advances in this area, due to low-cost digital cameras with increasing image quality, have made it both an inexpensive and reliable method to generate 3D models. The acquisition of image data for MVS can be accomplished using a simple camera to complicated configurations using an automated turntable with high quality lighting. On a large scale, images can also be utilized from multiple cameras and sources such as images acquired from drones or crowd-sourcing from the Internet [36, 116].

2.3.2 Compute Camera Parameters

The advancement of image acquisition hardware for both quality and cost was not the sole factor that permitted the recent development of the MVS research field but

the success of progress can be partially attributed to the steady increase in computational power, which in turn aided the ability to process many images quickly and the development of the algorithms used to **Compute Camera Parameters**. Particularly, the rise and improvement of Reconstruction and Structure from Motion (SfM) algorithms used to compute these parameters played a significant role. This opened the possibility to process and calculate camera parameters for multi-terabytes of images for 3D-reconstruction [126].

Structure from Motion operates under the assumption that the scene is rigid and uses point correspondences as cues to compute the camera models for parameters. It is a similar technique to Visual Simultaneous Localization and Mapping (VSLAM) but differs in that it generates parameters, most often in non real-time, using unordered sets of images. Whereas, VSLAM computes locations and parameters of cameras from video streams in real-time [28]. Although, there exists works that utilize MVS with VLSAM techniques [90, 122], in this document we concentrate on MVS algorithms that use unordered sets of images in non real-time and save MVS with VSLAM for future research.

Early work in Photogrammetry, the field of study that obtains 3D measurements from photographs, used triangulation to solve for 3D locations in a scene from multiple photographs having different vantage points. This process utilizes a priori, the position and orientation, of where each photo was taken during the triangulation process. The converse of the problem, given 3D locations determine the position and orientation of where the photo was taken from (Camera Pose Estimation/Resection-

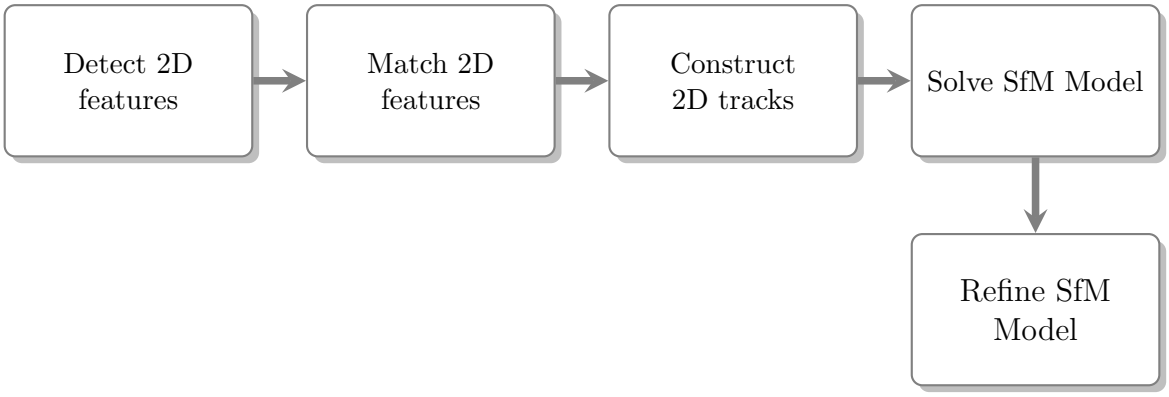


Figure 8: General SfM Pipeline.

ing). Traditionally, this problem is solved with known reference points in the photos, sometimes indicated by fiducial markers, which can be used to obtain Camera Pose Estimation. Structure from Motion simultaneously answers the problem of not knowing 3D locations or Camera Pose Estimation. This is a relaxation on the requirements of traditional Photogrammetry.

The purpose of Structure from Motion is to produce camera parameters of all input images and 3D points along with their corresponding 2D coordinates from subsets of the input images. The combination of a 3D point with its list of corresponding coordinates derived from a subset of input images is a called a track. The SfM pipeline can be generally outlined as (Figure 8): Detect 2D features, Match 2D features, Construct 2D tracks, Solve SfM Model, Refine SfM Model.

Much of the success for the development of the SfM methodology for unordered sets of images is due to the creation of high quality feature detectors [43, 78, 106] and descriptors [2,5,67,107], which permit stronger matching and higher quality tracks on

images where pose and illumination could be substantially different. Algorithms such as SIFT (Scale-Invariant Feature Transform) and ORB (Oriented FAST [Features from Accelerated Segment Test] and Rotated BRIEF [Binary Robust Independent Elementary Features]) are popular algorithms used for this process. Other areas of advancement that have attributed to the success in the development of SfM with unordered images are with efficient indexing [92], improved graph connectivity of tracks [117], and parallelization [1, 26] have improved the performance of matching features and descriptors. These improvements reduce the matching complexity of unordered images, that is every image has to be matched to all other images, which structured sequences of images do not have due to prior knowledge derived from nearby images.

The SfM process globally optimizes 3D points and the camera poses by minimizing the error, e_r, between the detected 2D points and the estimated re-projection of the 3D points from each camera (See Equation (18)).

$$Total\ Error \quad = \quad \sum_{i=0}^{k}\sum_{j=0}^{n} e_r(i,j)^2 = \sum_{i=0}^{k}\sum_{j=0}^{n} |\Pi_j(x_i^{3D}) - x_{ij}^{2D}|^2 \ , \tag{18}$$

where k is the number of points, n is the number of cameras/views, $\Pi_j(x_i^{3D})$ is the estimated re-projection of the i^{th} 3D point on the image plane of the j^{th} camera, x_{ij}^{2D} is the i^{th} point in the image of the j^{th} camera.

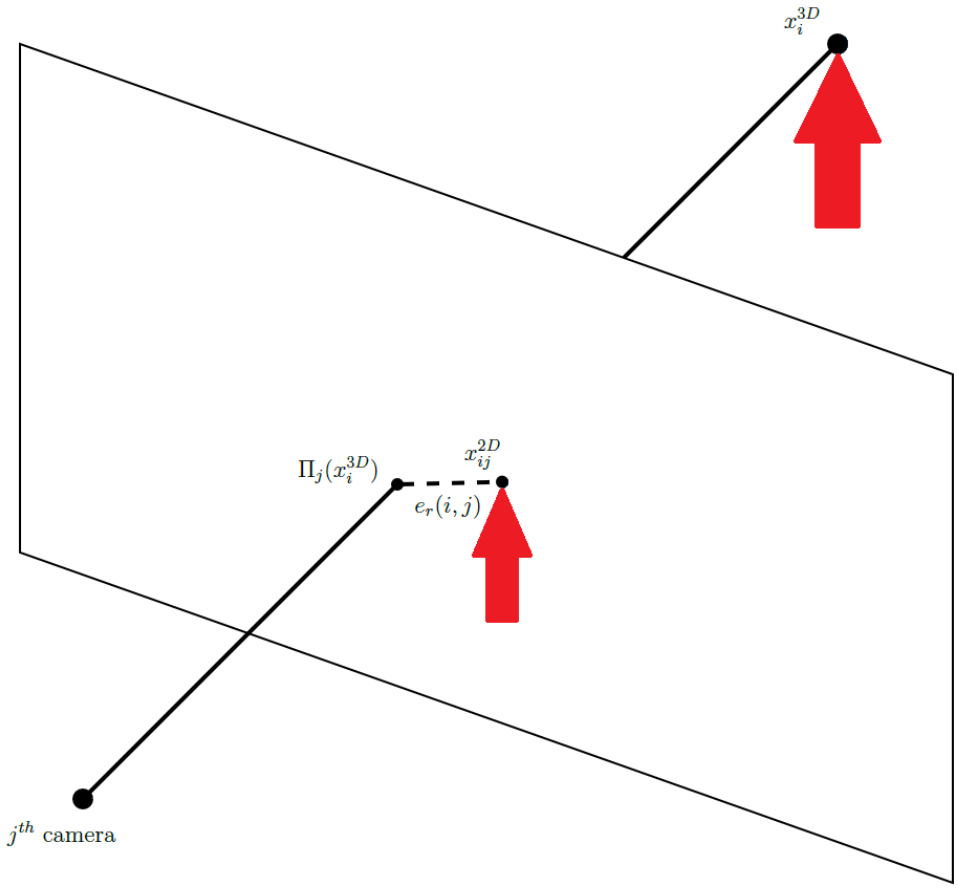


Figure 9: Bundle Adjustment: Re-projection Error, e_r, of point x_i^{3D} in the j^{th} camera plane.

The process of minimizing Equation (18)) is called **Bundle Adjustment**, which comes from the idea of having geometric bundles of light rays from each 3D point converging at a camera's optical center. The minimization of the function uses a nonlinear least-squares algorithm such as Gauss-Newton or Levenberg-Marquardt. Much success has been with the Levenberg-Marquardt Algorithm due to its simplicity by iteratively solving the non linear least squares problem by combing the gradient-descent method with Gauss-Newton. The reduction of error is accomplished by updating parameters in the steepest-descent direction for gradient-descent, and for

the Gauss-Newton method it assumes the parameters are locally quadratic thus reducing the problem to minimizing a quadratic. As a result, the Levenberg-Marquardt algorithm behaves like gradient-descent for parameters that are far from their optima, and behaves like Gauss-Newton for values close to their optimal value [34, 83].

2.3.3 3D-Geometry Reconstruction

Structure from Motion produces both camera parameters and a sparse 3D-reconstruction but the **3D-Geometry Reconstruction** process seeks to produce a dense reconstruction. So, this part of the process uses the information generated from the previous stage to create a dense reconstruction. This is accomplished by matching corresponding pixels across images. The goal of finding dense correspondences is similar to that of **Optical Flow**, where correspondences are typically only over two images (MVS uses more than two), camera calibration is not required (MVS assumes camera calibration is known), and the application is for interpolation and not 3D-reconstruction [4, 28]. For Optical Flow the matching search space is 2D since every pixel for an image can be matched against any other pixel but for MVS the search space is simplified since the camera parameters are known which reduces the search space to a 1D problem due to **Epipolar Geometry** (for more information on **Epipolar Geometry** the reader is directed to [44]).

The matching process evaluates correspondences between images using the concept of **photo consistency**, which measures similarity, coherence, and accuracy. The

general form of the **photo consistency** measure cost function for a pair of input images, I_i and I_j, and a 3D point p (seen by all images) is as follows (as taken from [28]):

$$C_{ij}(p) \quad = \quad s(I_i(\Omega(\pi_i(p))), I_j(\Omega(\pi_j(p)))) \ , \tag{19}$$

where $s(f, g)$ is a similarity measure that compares vectors f and g, $\pi_k(p)$ is the projection of p into the k^{th} image plane, $\Omega(x)$ defines a support domain around point x, $I_k(x)$ is the image intensity of the k^{th} image at position x.

Every **photo consistency** measure can be described as a particular choice of s and Ω [28]. Some examples of similarity measures used are: Sum of Squared Differences (SSD), Sum of Absolute Differences (SAD), Normalized Cross Correlation (NCC), Census, and Rank. The support domain Ω is used to encapsulate an area that defines the size of a unique region that has consistent illumination and viewpoint. The larger the defined region the more local uniqueness inside the domain thereby making it easier to match to other images but comes at the expense of loss of invariance with illumination and viewpoint due to issues such as reflectance, geometry assumptions, and boundaries. (For more information on **photo consistency** see [28].)

2.3.4 Texture/Materials Reconstruction

This process in the pipeline applies textures/materials to the 3D-Model to yield a realistic appearance. Sometimes it is considered an optional procedure during the MVS process depending on end-application. This is normally accomplished using a texture chart, obtained from the images of the acquisition stage and from the segmentation of the generated model mesh, which overlays and maps 2D surface images to various regions on the models mesh. Texture reconstruction is simply the registration between image and model geometry. Most approaches, for texture chart generation, use a Markov Random Fields (MRF) energy function to label each triangular face in the mesh. All triangles of identical labels are aggregated together into the texture chart [7, 64, 69, 141].

2.4 Super Resolution (SR)

Super Resolution (SR) is the reconstruction of Low Resolution (LR) input to produce High Resolution (HR) output. Throughout this document when referencing images we will interchangeably use Low Resolution (LR) for Low Quality (LQ), and High Resolution (HR) for High Quality (HQ), respectively. Super Resolution has been used in a wide variety of applications such satellite remote sensing, radar, and medical imaging [79, 118, 118, 128, 145, 145]. A simplified version of the taxonomy for Super Resolution, similar to [87], is presented in Figure (10). The diagram shows that Super Resolution can be achieved using frequency or spatial domains with both having two sub-categories: Fourier and Wavelet for frequency domain, and Single Image and Multi-Frame for spatial domain.

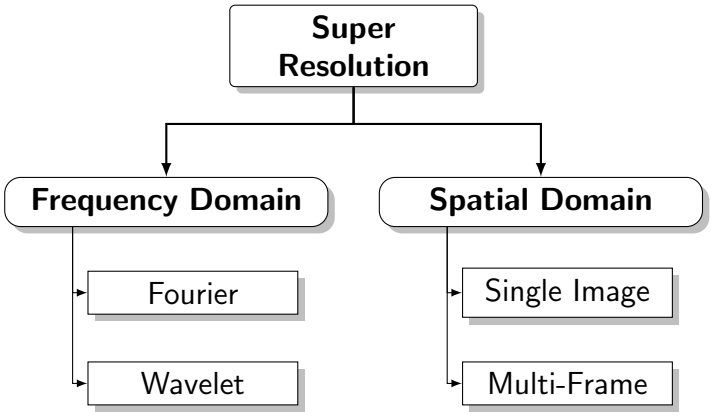


Figure 10: Taxonomy of Super Resolution.

Early work by Tsai were based on frequency domain and were easy to implement and computationally cheap but lacked the ability to add image priori and could only get good results for images without noise and degradation [75]. Although, this formulation through frequency domain gave excellent insights into the theory of Super

Resolution it is only effective for simple motion models (planar translation and rotation / rigid motion) and so cannot model for arbitrary displacement [60, 124, 129]. Further more, a tractable Linear Shift Invariant (LSI) kernel, by means of Fourier transform, for blur is necessary which restricts flexibility of the underlying image formation model [60]. So, for these reasons the spatial domain has been preferred and widely studied. The body of this research also uses Super Resolution that concentrates on the spatial domain with a particular interest in the Multi-Frame sub-category.

2.4.1 Super Resolution Mathematical / Generative / Observation Model

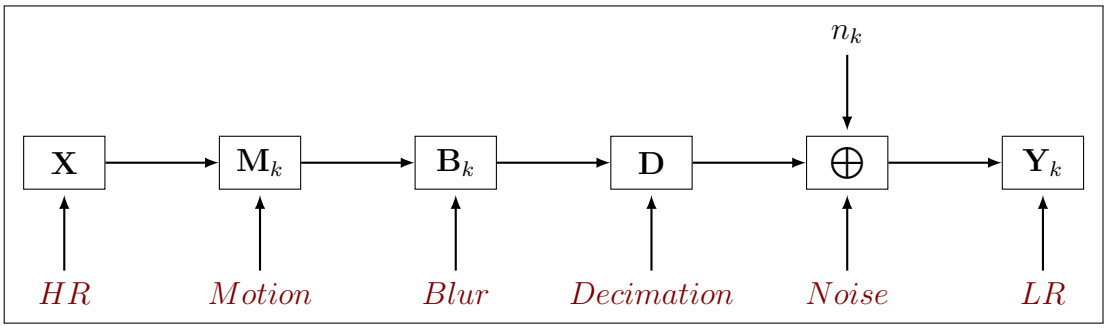


Figure 11: Super Resolution Mathematical Model Flow Chart.

The SR acquisition process has been modeled to accommodate for various degradations of the HR image to the observed LR image(s) using the following operations (Figure (11)):

- Warp / Motion
- Camera Blur
- Decimation / Down-sampling

- Noise

Warp / Motion. This operation describes the geometric transformation of an image with respect to a common coordinate reference, which encodes for camera motion and/or object motion.

Camera Blur. This operation describes various camera blurs such as motion and focus blurs. This process is normally accomplished using a low-pass filter, which is sometimes modeled from a Point Spread Function (PSF).

Decimation / Sampling. This describes the reduction in resolution/dimension from the HR image to the LR image(s).

Noise. This is the independent noise associated with each LR image, which is normally modeled using White Gaussian noise.

The widely accepted mathematical model (a.k.a Generative or Observation Model) for Super Resolution (SR) [22, 53, 87, 101] accounts for the degradation factors: motion, blur, and decimation (Down-sampling). Given by the following:

$$Y_k \;=\; DB_kM_kX + n_k \;, \tag{20}$$

where Y_k is the kth Low Resolution (LR) image for $k = 1, 2, \ldots, N$, X is the ideal High Resolution (HR), M_k is the geometric sub-pixel motion of the kth image, B_k is the blur

matrix of the kth image, D is the down sampling matrix, n_k is the nonhomogeneous additive Gaussian noise (uncorrelated between different measurements) of the kth LR image.

It is standard practice to simplify Equation (20) by combining D, B_k, and M_k as one matrix W_k [75, 101, 102, 124, 125] (sometimes denoted with H_k):

$$Y_k = W_k X + n_k \tag{21}$$

Sometimes the literature uses a functional notation to represent the model:

$$Y_k(m,n) = D(B_k(M_k(f(x,y)))) + n_k(m,n) \tag{22}$$

2.4.2 Single Image Super Resolution (SISR)

Single Image Super Resolution (SISR) uses a single LR input to produce HR output. Most approaches are based on priori and use an explicit distribution or energy function, or are implicit example-based [27, 104]. Learning based algorithms are most often used for implementing SISR [3, 57, 59, 65] which require a training step. The training step learns the relationship between LR input and the HR output in the

hopes that high resolution details can be recovered when presented with arbitrary LR input. There are two main approaches to defining low resolution to high resolution relations: mapping patches or mapping structures/features [59, 87]. It is well known that Learning approaches are only as good as their training and that generated output may not be a true representation [59, 91]. These imaginary details or hallucinations may cause undesirable effects on the output reducing the suitability for specific applications where true representations are preferred.

2.4.3 Multi-Frame Super Resolution (MFSR)

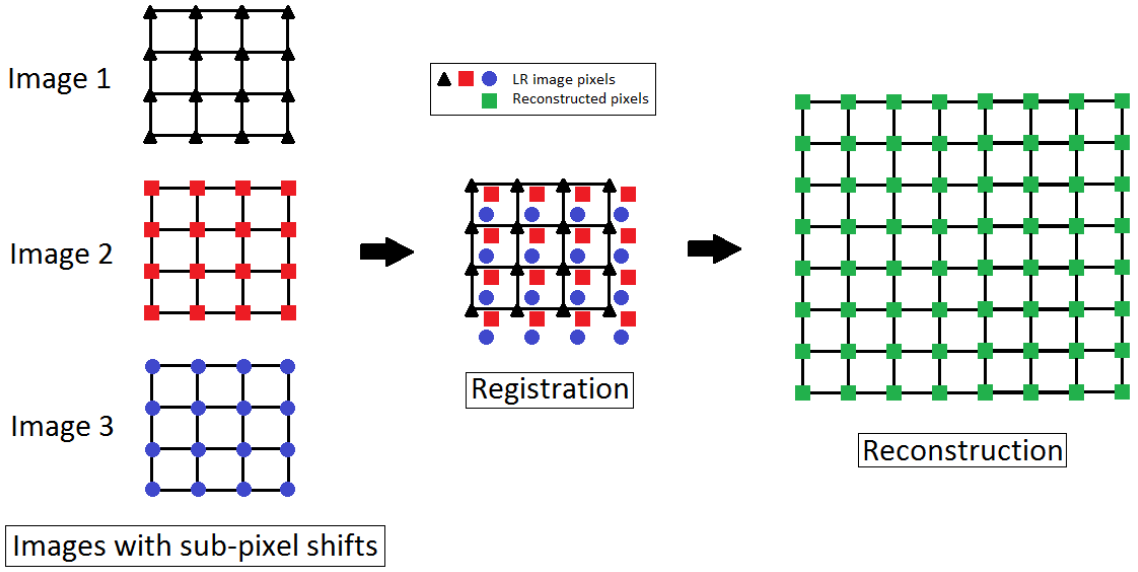


Figure 12: Multi-Frame Super Resolution overlapping sub-pixel example.

Multi-Frame Super Resolution (MFSR) uses multiple LR inputs to produce HR output. It works on the premise that multiple views capture independent information, of the same scene from slightly different perspectives, that can be combined to create output that has more information than its inputs alone (as illustrated in Figure (12)).

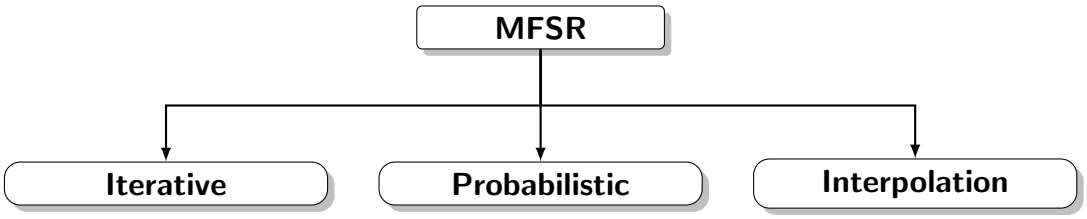


Figure 13: Taxonomy of MFSR over spatial domain.

Multi-Frame Super Resolution can be further broken down into three categories (as illustrated in Figure (13)) :

Iterative. Iterative methods most commonly use Iterative Back Projection (IBP) or Projection onto Convex Sets (POCS). Iterative Back Projection estimates the HR image iteratively as the projected sum of simulated LR images obtained through refined estimations of motion, blur, and noise in conjunction with the original LR images. This method is among the first algorithms developed for SR with notable works from [19, 53, 54, 96, 130, 148]. Projection onto Convex Sets produces solutions from the intersection of constraint sets that contain possible values for Super Resolution pixels such that every possible SR image can lead to each LR image. That is, each LR image imposes an a priori on the Super Resolution result. [100, 111]

Probabilistic. This method uses stochastic operations such as Bayesian inference, Markov/Gaussian Random Fields, and Total Variation to reconstruct HR images from LR images using probabilistic techniques that utilize methods such as Maximum Likelihood (ML), Maximum A Posteriori (MAP), and Inference models. (We refer the reader to [87] for a more detailed description.)

Interpolation. Interpolation-based methods, also known as Direct Methods, are the

Figure 14: Multi-Frame Super Resolution Flow Chart.

simplest and most straight forward procedures for producing super resolved HR output. They involve a three step process (Figure (14)): **Acquisition, Registration, and Reconstruction**.

The **Acquisition** process obtains LR images from slightly different viewpoints. The larger the overlap between all of the images the better since this would produce a result that has the greatest area and helps make the Registration process easier. Ideally, sub-pixel shifts are desired between the images.

The LR images are then geometrically aligned during the **Registration** process. Normally, one of the images is selected from the LR image set as reference for the alignment process since each of the LR images may have different sub-pixel displacements or rotations. The Registration process estimates motion parameters using methods like Optical Flow by matching pixel-to-pixel or feature-based methods [121]. Pixel-to-pixel approaches shift or warp images to the reference image and produce an alignment of pixels with the least discrepancy. Feature-based methods extract details from images such as corners, edges, ridges, and shapes. These details are matched to the reference image and a mapping is defined to transform the images to the reference space.

The **Reconstruction** stage, also known as **Restoration**, utilizes the aligned images and produces a High Resolution output image by scaling and filtering. Filters such as mean and median are commonly used for this process [25]. Other filters such as SVD-based [86] and Adaboost classifiers [114] have been used. Optionally, during this stage a de-blurring kernel kernel is applied to sharpen the result.

2.4.4 Image Registration Using Enhanced Correlation Coefficient

A good method for image registration, that achieves sub-pixel accuracy and is invariant to photometric illumination, is the enhanced correlation coefficient (ECC) maximization algorithm [23]. This method uses gradient descent with enhanced normalized cross correlation (ENCC) as the objective function. Although, this function is nonlinear the iterative scheme they proposed reduces the process to linear computational complexity.

The following is a brief overview of the ECC process (see [23] for more detail).

Let $I_r(x)$ and $I_w(y)$ be image intensity values of the reference and template/registration images with coordinates $x = (x_1, x_2)$ and $y = (y_1, y_2)$, respectively.

The performance for geometric registration is quantified by error metrics, having warping transformation parameters p, with the criterion represented by Equation (23).

$$E_{ECC}(p) \;=\; \left\| \frac{\bar{i}_r}{\|\bar{i}_r\|} - \frac{\bar{i}_w(p)}{\|\bar{i}_w(p)\|} \right\|^2 , \tag{23}$$

where $\bar{i}_r$ is the zero-mean of reference image I_r, $\bar{i}_w(p)$ is the zero-mean of registration image I_w warped by parameter p, and $\|\cdot\|$ is the standard Euclidean norm.

The minimization of the E_{ECC} criterion yields the optimal image alignment, which is equivalent to the maximization of the duality given by ENCC [103] represented by Equation (24).

$$\rho(p) \;=\; \frac{\bar{i}_r^T \bar{i}_w(p)}{\|\bar{i}_r\|\|\bar{i}_w(p)\|} = \hat{i}_r^T \frac{\bar{i}_w(p)}{\|\bar{i}_w(p)\|} , \tag{24}$$

where $\bar{i}_r$ and $\bar{i}_w(p)$ are the same as in Equation (23), and $\hat{i}_r^T = \frac{\bar{i}_r^T}{\|\bar{i}_r\|}$ is the normalized zero-mean reference image.

Gradient descent is applied to ENCC, $\rho(p)$, by updating $p = \tilde{p} + \Delta p$ (where $\tilde{p}$ is some nominal parameter close to p and Δp is a vector of perturbations) and approximating I_w by first order Taylor expansion represented by Equation (25).

$$I_w(y) \approx I_w(\tilde{y}) + [\nabla_y I_w(\tilde{y})]^T \frac{\delta\phi(x;\tilde{p})}{\delta p}\Delta p \ , \tag{25}$$

where $\tilde{y} = \phi(x;\tilde{p})$ are the warped coordinates under the nominal parameter vector, $y = \phi(x;p)$ under the perturbed vector, $\phi(\cdot\ ;\cdot)$ be a well-defined coordinate mapping for motion estimation between the reference and registration images, $\nabla_y I_w(\tilde{y})$ denotes the gradient vector of $I_w(y)$ of the warped image, and $\frac{\delta\phi(x;\tilde{p})}{\delta p}$ denotes the Jacobian matrix of the transform with respect to the parameters at the nominal values.

Applying Equation (25) to all coordinates will yield a linear version of the warped parameters p represented by Equation (26).

$$i_w(p) \approx i_w(\tilde{p}) + G(\tilde{p})\Delta p \ , \tag{26}$$

where $G(\tilde{p})$ denotes the Jacobian matrix of the warped intensity vector with respect to the parameters at $\tilde{p}$.

Using the estimation of $I_w(p)$ from Equation (refeqn:IwApproxJacobian) we can approximate Equation (refeqn:ENCC) under the nominal parameter and perturbation vectors shown in Equation (27).

$$\rho(p) \approx \rho(\Delta p|\tilde{p}) = \hat{i}_r^T \frac{\bar{i}_w(\tilde{p}) + \bar{G}(\tilde{p})\Delta p}{\|\bar{i}_w(\tilde{p}) + \bar{G}(\tilde{p})\Delta p\|} , \quad (27)$$

where $\bar{G}(\tilde{p})$ and $\bar{i}_w(\tilde{p})$ are the column-zero-mean versions of $G(\tilde{p})$ and $i_w(\tilde{p})$, respectively.

2.4.5 MFSR: Quality vs Quantity Tradeoff

The suggested minimum number of images required for good consistent super resolved output such that the tradeoff between number of images and quality, since there are diminishing returns after a certain number of images, is N^2, where N is the magnification factor [97, 111, 127].

2.4.6 Multi-Frame: Noise Reduction

It is well known that signals obtained with Charged Coupled Devices (CCDs) are inherently degraded by shot noise [99, 109] and that this noise can be modeled by a Poisson Distribution [55, 76]. This means that digitized images that are captured will have a certain degree of noise present in their result. The process of combining multiple image frames results in a noise reduction, which can be approximated using the following equation [47, 48]:

$$Percentage\ Noise\ Reduction \approx \left(1 - \frac{1}{\sqrt{N}}\right) \times 100\ , \tag{28}$$

where N is the number of images/frames.

By examining Figure 15 and Table 1, we can see that there is a diminishing return of noise reduction as the number of images increase. Using Equation (28) the noise reduction for 16 to 25 images will give an approximate reduction of 75% to 80%.

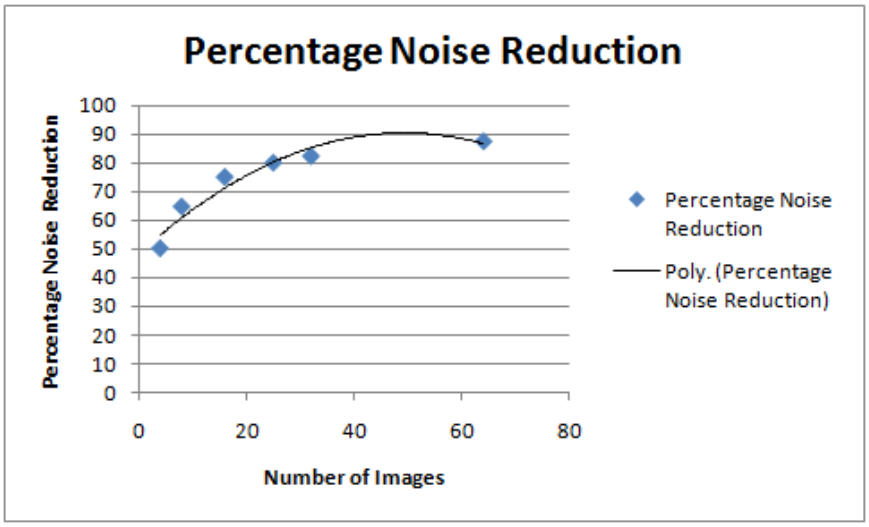


Figure 15: Multi-Frame Percentage Noise Reduction.

Number of Images	Noise Reduction
4	50%
8	65%
16	75%
25	80%
32	82%
64	88%

Table 1: Noise reduction from using multiple images.

2.4.7 Metrics for Image Quality Assessment

Determining the effects of processing on the quality of an image is an important operation used to evaluate resulting effects such as loss of information or degradation. Image quality assessment can be broken down into two categories [50, 110]:

- **Subjective**: A qualitative measure based on human perception and judgement without an explicit reference criteria.
- **Objective**: A quantitative measure that uses explicit numerical criteria for comparisons with references such as ground truth and statistical prior knowledge.

There are many methods to perform image quality assessment. In this document we use the following well known objective metrics: Mean Squared Error (MSE), Peak Signal to Noise Ratio (PSNR), and Structural SIMilarity (SSIM) (a.k.a. Structural Similarity Index Measure (SSIM)).

Mean Squared Error (MSE) is the average squared intensity difference between a test and reference image. MSE can be considered as a quadratic loss risk function based on the expected value of squared error loss [8] and is given by Equation (29). This metric, in simple terms, tells you how far apart two images are.

$$MSE = \frac{\sum_{n=0}^{M}\sum_{m=1}^{N}(I_1(n,m) - I_2(n,m))^2}{MN} \quad (29)$$

Peak Signal to Noise Ratio (PSNR) is the ratio, in decibels, between the peak (maximum) power of a signal and the noise floor as is given by Equation (30). It is often used as a measurement of quality between two images. The value for PSNR tends to infinity as MSE tends to zero, which can be interpreted as: The larger the PSNR value the better the quality of the signal or image. This means a smaller difference between the test and reference. Conversely, the smaller the PSNR value the larger the difference between the signal or image.

$$PSNR = 10log_{10}(peak^2)/MSE \quad (30)$$

Structural SIMilarity (SSIM) is a normalized measure of the similarity between two images and is given by Equation (31). This metric was first described by Wang et al. [138] and was developed with the idea of perceived human perception. It models image perception as distortions of luminance, contrast, and loss of correlation. The normalized result of SSIM is in $[0, 1]$ with 0 meaning the images are not similar or no correlation exits, and 1 meaning the images are completely similar and 100% correlated, that is the two images are equal.

$$SSIM = \frac{(2\mu_x\mu_y + C_1)(2\sigma_x\sigma_y + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (31)$$

2.5 Related Work

2.5.1 Stereo Vision with Super Resolution

There have been approaches that use Stereo Vision with Super Resolution some of which concentrate on view synthesis from 2D video to Stereoscopic 3D for use with multi-user 3D displays or auto-stereoscopic displays [62]. Other approaches work at improving the spatial resolution of Stereo images using parallax priori or cross view capture [15, 56, 135, 137, 144, 147] or improving disparity/depth map using Super Resolution [17, 143].

Super Resolution has also been applied to 3D face recognition using low resolution acquisition devices such as RGB-D (Red Green Blue - Depth) cameras. The low resolution 3D scans are super resolved to produce higher resolution 3D face models (superface models) [6, 51, 85].

2.5.2 Combining Multi-View Stereo with Super Resolution

Producing a Super Resolution image from a sparse 3D-reconstruction and Low Resolution input images.

The combination of using Multi-View Stereo with Super Resolution is a natural evolutionary path to improving either of the frameworks and has been accomplished in few different ways. One of the first works used multiple Low Resolution images to generate a sparse 3D-reconstruction, which was used to produce a super-resolved

image that exceeded the spatial resolution of the input images and could view the scene from an angle different from the observed input views [89]. The focus of their research was not the 3D-reconstruction to produce a 3D model but the production of a super resolved image, that can have a different viewpoint than the input images, from a scene with depth. Although, the work presented used an adaptation of a Bayesian SR algorithm from [131], the produced framework permits the usage of any 2D Super Resolution algorithm to generate super resolved output from a scene with depth variation.

Producing Super Resolution images from Multi-View images and depth maps.

The use of multi-view images in combination with depth maps have been used as a Super Resolution technique [31]. This technique uses high frequency details from adjacent views in conjunction with correspondences based on the related depth maps. Their work concentrates on a mixed resolution multi-view framework that yields a result with notable gain in Peak-to-Signal-Noise-Ratio (PSNR) over up-sampled and non-resolved images.

Enhancing model appearance by applying Super Resolution to textures.

Applying Super Resolution to the surface textures of a 3D model generated by Multi-View Stereo is another combination of the two frameworks used to improve the appearance of a model. The original motivation of this work came from the emergence of high quality 3D models and the importance to recover high resolution and high

quality texture maps from low resolution input images with the goal of estimating texture maps as precisely as possible.

One of the early works used a Partial Differential Equation (PDE) based gradient descent method to solve Euler-Lagrange equation of the texture surface [38]. In a later work, the authors extended and modernized their method by switching from a PDE based gradient descent to a convex optimization based method. This switch made their work easy to implement and consequently more computationally efficient [37].

Another method to enhance MVS model textures with SR used a direct keyframe-based Simultaneous Localization And Mapping (SLAM) frontend to estimate RedGreenBlue-Depth (RGB-D) camera motion followed by image alignment and volumetric fusion to produce a mesh. Low Resolution RGB-D images are deblurred and fused into Super Resolution keyframes which are texture mapped to the mesh and results in an improved texture quality as compared to simple volumetric blending alone [81]. The method uses an assumption of consistent 3D model geometry and camera poses between corresponding pixel values from neighbouring LR images and uses computes the weighted median of those values.

However, this procedure is subject to the same geometric inaccuracies as those of MVS reconstruction to which the median filter is supposed to correct for. The method from [12] overcomes this limitation of prior work by using an optical flow algorithm that corrects the initial geometric registration error directly in the image domain on a sub-pixel level. Another method for reconstructing 3D structure with high

resolution texture of a scene from multiple low resolution images, taken from different viewpoints, directly uses image intensity and iteratively estimates high resolution texture and structure forgoing the necessity for point correspondences among multiple images [93]. By simultaneously estimating high resolution textures and 3D structure they were able to produce an improved result that cannot be obtained alone by traditional methods.

More recently, the improvement in the appearance of 3D models by enhancing textures with Super Resolution using a Deep Learning based approach has been done by [70]. They address the limitation in the lack of multi-view data for a deep learning approach by introducing the data set 3D Appearance SR (3DASR) which is based on existing ETH3D, SyB3R, MiddleBury, and a few scenes of their own. Their method uses 2D Deep Learning SR techniques adapted to the texture domain using geometric information via normal maps which yeilds similar performance to that of model based methods.

Enhancing MVS by applying Super Resolution to depth maps.

Super Resolution has been applied to the depth map stage of the Multi-View Stereo framework to address the limitations between the resolution of captured depth information, using methods like Structured Light or Time-Of-Flight (TOF), to regular colour cameras.

One such implementation generated a Low Resolution (LR) 3D model, using a real-time depth sensor that captured RedGreenBlue-Depth (RGB-D) images, to guide

the acquisition process. Super Resolution methods were then applied to enhance the RGB-D images and merged with High Resolution images into a single mesh that was later textured using data from a high quality camera (Canon EOS 5D) [113]. This process generates an improved model appearance, due to the high quality textures captured by the high quality camera, in comparison to texture generation with the original LR colour images alone.

Model based Compressive Sensing (CS) is another technique that has been used as a reconstruction method to address the resolution differences between LR depth cameras and regular colour cameras [74]. This method transforms a LR depth map to a High Resolution (HR) depth map using CS depth map Super Resolution. Compressive Sensing (CS) is a signal processing technique for the acquisition of sparse compressible signals which permits signal reconstruction from a small set of random samples. That is, if a signal has a compressible representation then it can be represented by fewer samples than traditional Shannon/Nyquist representations [119]. The authors of [74] demonstrated that model based CS can be used effectively as a reconstruction method when applied as a SR technique to depth maps.

Super Resolution Depth maps have also been obtained using a Joint Bilateral Up-sampling (JBU) filter in a final refining step and an energy cost minimization of the similarity between the measured and estimated depth values, as well as, the smoothness of neighbouring pixels of those estimated depth values [16]. This method produced better HR depth maps than previous up-sampling methods (Bicubic up-sampling, Multi-lateral filtering [32], Pixel Weighted Average Strategy (PWAS) [33],

and Joint Bi-lateral Up-sampling (JBU) [63]).

In a similar work [66], two stages were applied to the depth map SR framework: Credibility and Synthesis. In the credibility based stage they perform a Multi-view Depth Map Fusion (MDMF) and in synthesis they performed a View Synthesis Quality - Trilateral Depth-map Up-sampling (VSQ-TDU). The fusion algorithm uses the credibility of depth map values by examining the disparity range from different view points that should fall within a certain error, and values of neigbouring views should be similar. The view synthesis quality algorithm uses the Joint Adaptive Bilateral Depth map Up-sampling (JABDU) filter from [61] since it is both simple, fast, and produces the desired results during up-sampling. Their method produces improved results since it considers both view synthesis and depth map quality in the depth map SR process whereas most algorithms only consider either or.

Another work combined Multi-View Stereo and Super Resolution in a unified framework [98]. Their work optimized a unified energy function between HR images and HR depth maps imposing consistency constraints between corresponding HR and LR images, as well as, employing a regularization constraint for depth map smoothness. They showed their formulation improves accuracy and eliminates mosaic artifacts from HR output.

Satellite imagery has been used to create 3D topographic surface maps of the earth, traditionally using Digital Elevation Model (DEM) generation which uses 2 or 3 images during the reconstruction process. This process can be enhanced by using more

view points and Super Resolution. Due to the advent of low cost micro-satellites there is a larger selection of available images from different view points, which permits the creation of many disparity maps that can be up-sampled, stacked, and transformed to produce a super resolved depth map [134]. The authors demonstrated their technique, which used Phase Correlation (PC) based sub-pixel stereo matching, to produce a super resolved topographic 3D-reconstruction of Usak (Western Turkey) from SkySat images. This lays the ground work for producing, on a large scale, super resolved topographic maps.

The authors of [105] applied a custom Super Resolution pipeline to depth map data to enhance the 3D reconstruction process. Their process used the following steps: Pre-process (Calculate 3D Volume Bounding Box), Registration (Enhanced Correlation Coefficient (ECC) Affine), Up-sampling (Nearest Neighbour), Warp (Registration of Up-sampled - Affine), Reconstruction (Mean filter with removal of invalid depths indicated by 0). The added pre-processing to calculate the volume bounding box, secondary registration, and elimination of invalid depth values during mean filtering are what's different from the traditional SR pipeline. Their work demonstrated an improved smoothness and reduction of holes in the generated models from Low Resolution low cost depth sensors on MAE/UFBA (Museum of Archaeology and Ethnology/Universidade Federal da Bahia) museum artifacts.

The development of learning based approaches for improving depth map resolution for 3D-reconstruction has been studied. One such work utilized a simple visual difference based loss function [133]. The loss function method yielded a significant im-

provement in 3D shapes when used with a simple deep prior or trained Convolutional Neural Network (CNN) and compared to standard metrics such as Structural Similarity (SSIM). Their method compared rendered images of the model surface which guided the depth map Super Resolution process. A similar method [68], improved the depth map resolution using a depth SR network based on gradient saliency which was further improved and guided by learning surface normals, occlusion boundaries, and blurriness of images from HR images. Another data driven learning approach used a Deep Residual Network to progressively up-sample LR depth images guided by HR intensity images on multiple scales [149]. The depth structure is recovered in a course to fine progression using multiple scale frequency synthesis which provides fast convergence and improved performance for both qualitative and quantitative. The authors of [45], provide a method that adaptively decomposes high frequency components from RGB images to guide depth map Super Resolution called Fast Depth Map Super Resolution (FDSR). Their work exploits contextual information from high frequency multi-scale structure which guides the depth map SR process.

Enhancing MVS by applying Super Resolution to the input stage.

Increasing the resolution of 3D geometry by enhancing the input stage using a combination of wavelet-bilinear interpolation based Super Resolution has been achieved in [136]. The authors improve on the stereo spatially enhanced 3D-reconstruction analog by using 3-views and super resolving each of those views using an adaptive interpolation between wavelet edge areas and bilinear up-sampling. Their work focuses on a theoretical basis for the combination of image enhancement and 3D model

reconstruction.

A similar work, generalizes the previous idea of of solely enhancing the input stage of Multi-View Stereo to improve 3D reconstruction by using Single Image Super Resolution (SISR) [77]. Their work improved completeness of 3D-reconstruction and is very effective for textured and outdoor scenes. They developed their framework around the Deep Back Projection Networks (DBPN) SISR algorithm which they directly apply to the LR input images before MVS reconstruction. The results of the models they obtained were improved in most cases by their methodology and had more robust and dense representations. Furthermore, they showed a strong correlation between quality of input images and quality of reconstructed models. However, their work performs well for large scale outdoor settings and concentrates on using a Deep Learning based SISR algorithm but has the following limitations:

- Does not perform as well with indoor compared to outdoor.
- Focuses only on large scale scenes and not small scale objects.
- SISR quality is dependent on network training [91] and may not be a true representation.
- Computationally intense for both training and usage (requires GPU for speed).
- Noise in the input data is propagated to the output.

3 Method and Hardware of the Vision System

This chapter covers the experimental method and techniques, as well as, the hardware specifications for the vision system and outlines its major subsystems (see Table 2 for more details).

3.1 Method

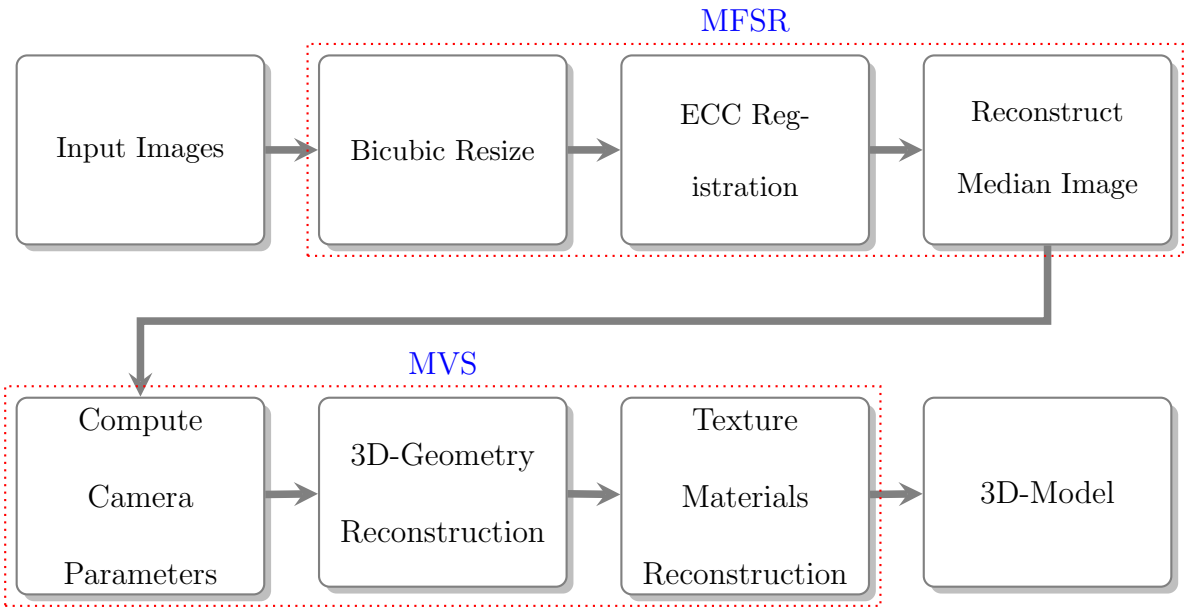


Figure 16: Proposed MFSR-MVS framework flow chart.

Our goal is to improve the output of the multi-view stereo framework by directly enhancing the input. We propose a method, Figure 16, that captures images from sub-pixel camera movements. These images can be recombined, using super resolution, to create a higher resolution image that contains more information or details than each image alone. For the super resolution process, we chose a multi-frame algorithm

because it only reconstructs information from details obtained from the scene, unlike SISR algorithms which use inferred details from learned priori and can potentially suffer from artifacts. We use a simple interpolation-based method, also known as a direct method, for MFSR. First, we bicubically up-scale the LR images and then use ECC to register the images, followed by a median filter to reconstruct and form the final super resolved image. Next, at the MVS stage camera parameters are computed and a sparse model is created. Using the generated parameters and the images, the sparse model is refined into a dense model, which is then textured to produce the final 3D-Model.

The proposed framework requires a comparison to a ground truth reference model. However, the availability of reference models and data sets, that are suitable for MFSR with MVS, present a difficulty due to the sub-pixel motion requirements. So, the popular data sets such as Middlebury, KITTI, Stretcha, and BlendedMVS are not suitable. One solution would be to generate reference models using a high quality laser scanner, but our research laboratory does not have access to such a device.

So, instead we utilize a method that is similar to the literature for super resolution [41, 72]. This method uses down-sampled HR images that are up-sampled back up to the original resolution by the super resolution algorithm and compared to the original HR image as reference. Our process generates models through the MVS pipeline using the original HR images and up-sampled LR images (generated from down-sampled HR images) as input for reference and comparison models, respectively. These models are compared by measuring the average, median, and max distance away from the

reference model. The model that has the closest metrics to the reference model is determined to be the best model.

In addition to the HQ and up-sampled models, we also generate a standard model solely from the LR images which represents a base starting point without enhancement and provides a reference to what a model would be if the starting point was at the lower resolution. This gives us a relative comparison for the other methods versus a standard starting point.

3.2 3D Vision System: Hardware Outline

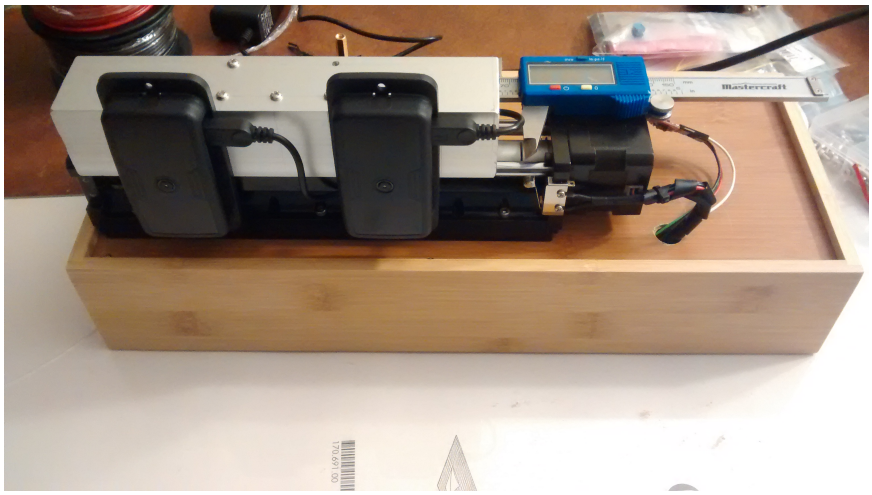


Figure 17: 3D Vision System: Image Acquisition System mounted on a linear rail with precision feedback.

Table 2: **Materials:** List of Major Hardware Components

System	Item
Image Acquisition	Raspberry Pi Zero W
	Raspberry Pi Camera Module V2
Linear Motion	Linear Stage w/ NEMA 17 Stepper
	Stepper Controller (TB6560)
Measurement	Mastercraft Vernier Digital Caliper
Power	12V 10A Switching Power Supply
	5V 2A DC Wall Adapter
Misc.	ESP8266 SOC (HUZ/ZAH)
	Logic Level Converter
	Wireless Router (Netis WF2411)

The vision system is composed of 3 major subsystems:

1. Image Acquisition

2. Linear Motion

3. Precision Measurment

3.2.1 Image Acquisition

The Image Acquisition system consists of two single board computers (Raspberry Pi Zero W) each attached with a camera module (RPiV2). The Raspberry Pi systems are running the standard LINUX Raspbian OS on a BCM2835 processor (ARM11 1GHz single-core) with 512MB RAM supporting wireless communication via WiFi (802.11n) and Bluetooth (v4.1). The RPiV2 camera modules utilize an 8-MegaPixel Sony IMX219 CMOS sensor (see Table 3 for more details) and are individually mounted in enclosures with their corresponding Raspberry Pi Zero W. The enclosures are attached to the Linear Stage by aluminum angle with regular screws and standoff mounting hardware. No additional effort was made in the alignment of the camera systems other than getting them close to a side by side parallel orientation by eye.

Table 3: Raspberry Pi V2 Camera Module with IMX219 CMOS Sensor Specifications.

Description	**Property**
Resolution	8 Megapixels
Sensor res.	3280 x 2464 px
Sensor dims.	3.68mm x 2.76mm
Pixel Size	1.12μm x 1.12μm
Optical Size	1/4”
Focal Length	3.04mm

The Raspberry Pi Zero W was selected because of its adequate computational power, low-cost, wireless capabilities, and highly functioning Operating System which has

a large support in the open-source community and a plethora of software and development tools, Application Programming Interfaces (APIs), and libraries such as support for C/C++, Python, and OpenCV.

3.2.2 Linear Motion

The Linear Motion system (FIG. 18) consists of a ball-screw driven stage with a NEMA 17 stepper motor and has an effective working distance of $120mm$ with an accuracy of $\pm 0.01mm$.

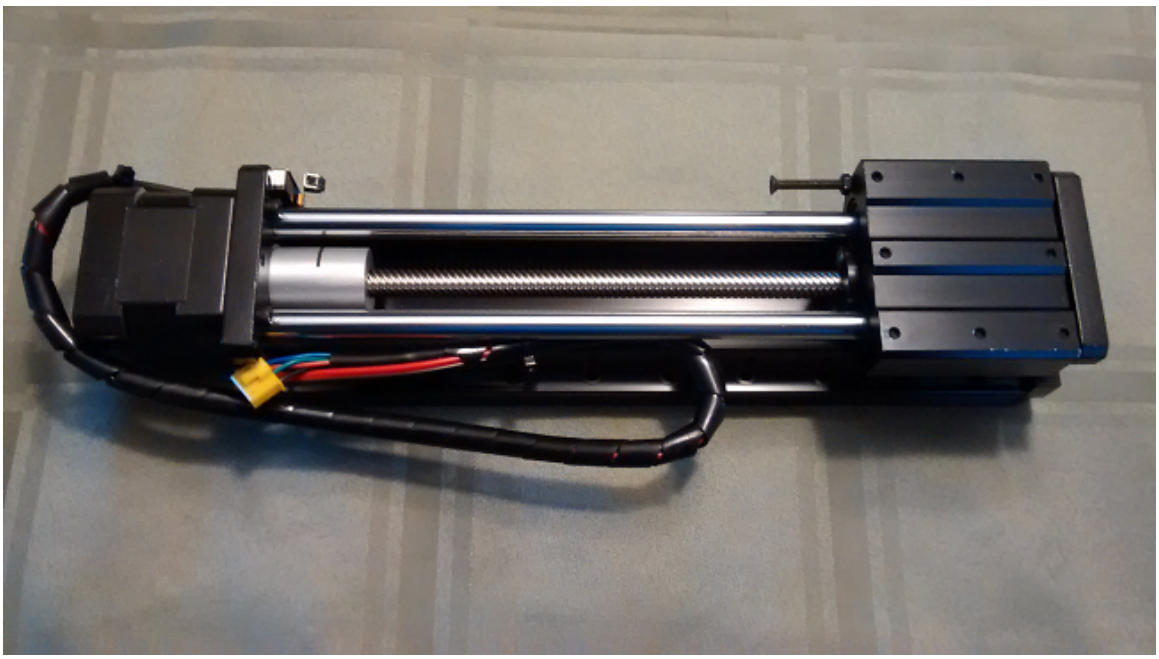


Figure 18: Linear Stage with NEMA 17 Stepper Motor.

The stepper motor is driven by a TB6560 controller attached to an ESP8266 SOC (System On Chip). The ESP8266 SOC was selected because of its low-cost and WiFi capability. It came pre-installed with the Lua interpreter by default which was replaced by from firmware from the Arduino IDE. The Linear Stage was selected

because of its low-cost, accuracy and precision, and rigid all metal construction.

3.2.3 Precision Measurement

The Precision Measurement system consists of digital Vernier calipers (Mastercraft) wired to the same ESP8266 SOC as the TB6560 stepper controller. The jaws of Mastercraft Vernier calipers are attached to the frame and moving platform of the Linear Stage such that one jaw moves with the platform and the other is stationary with the frame of the stage. The Mastercraft Vernier calipers have an accuracy of $\pm 0.01mm$ and measuring range of $150mm$.

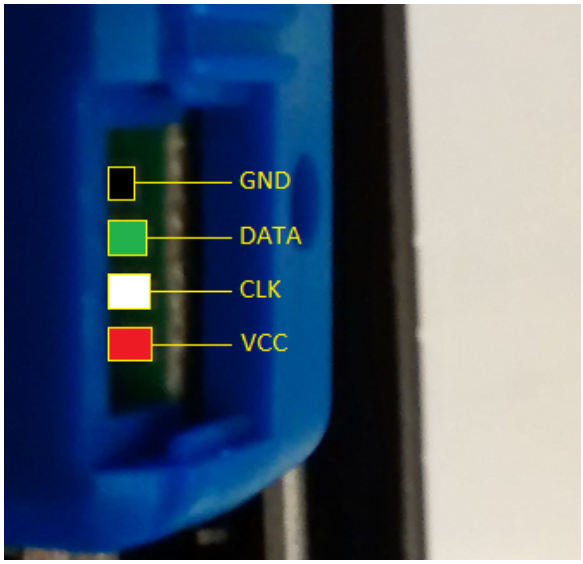


Figure 19: Pinout of Vernier Caliper port ascertained from probing.

There is a small undocumented proprietary data port, opposite to the side of the battery cover, which was carefully probed (See FIG. 19 for pinout) to ascertain it's functionality. The Mastercraft caliper uses a synchronous clock driven protocol for

data communication represented by a 19-bit payload. The 19-bit payload represents the measured value in 100ths of millimeters and is encoded using one's complement. The ESP8266 SOC is connected to the Clock (CLK) and Data lines of the digital calipers through a level-shifter to capture the measurement data from the caliper's undocumented port, since the data logic for the caliper runs at 1.5v and the ESP8266 at 3.3v. Also, the caliper's power requirement is very small and could be powered through an Input/Output (IO) pin on the ESP8266 which was attached through a simple voltage divider. This gives the functionality to turn the calipers on and off, and consequently gives the ability to reset the caliper to zero.

4 Experiments

4.1 Test: Obtaining Metrics From an Object

In this experiment, we demonstrate measuring the distance between two points on an object (teacup) by using the vision system's ability to move the cameras a precise distance. It is this unique ability that we harness to compute scene metrics from corresponding image points of the cameras. To illustrate this we examine Figures 20 and 21, and develop the scenario in the following way: we take a pair of pictures from the left and right cameras (Stereo Pair) of the vision system then move the mechanism 89.98mm where another picture pair is taken.

It is important that we differentiate between image sets from the Stereo Pair(STP) and image sets from the Single Camera Pair(SCP). Images from the Single Camera Pair are the ones from a single camera at precise positions on the vision system's linear stage, that is, the images taken from only either the left or right camera of the vision system. Images from the Stereo Pair (or Standard Pair) are images taken from both the left and right cameras of the vision system.

Using SCPs gives us the ability to accurately know the base distance between corresponding images or images at different camera positions of the same scene.

So, for this example both the left and right cameras move 89.98mm. This means that the base distance between the first left camera image and the second left camera image is 89.98mm. Similarly, true for the images of the right camera. So, images

Figure 20: Left Single Camera Pair(SCP) Coordinates for: Top(Left Image) [left point (2585, 1942) : right point (2811, 1924)], Bottom(Right Image) [left point (1504, 1942) : right point (1716, 1922)]. Here (Left Image) and (Right Image) refer to the SCP in terms of a stereo pair so the formulas from equation set (13) follow. The SCP for the Top and Bottom images have a base length of $89.98mm$

taken from the left camera make up the Left Single Camera Pair image set. Similarly, images taken from the right camera make up the Right Single Camera Pair image set. The image set generated by both the left and right cameras at the same position make up the Stereo Pair image sets.

By matching corresponding points from an image from the Left Single Camera Pair set to another, distinct, image of the Left Single Camera Pair set we can reconstruct the scene in 3-space by using equation set (13). The same can be done for the Right Single Camera Pair.

So, we can solve for the 3-space points P_1 and P_2, given base $b = 89.98$, focal length

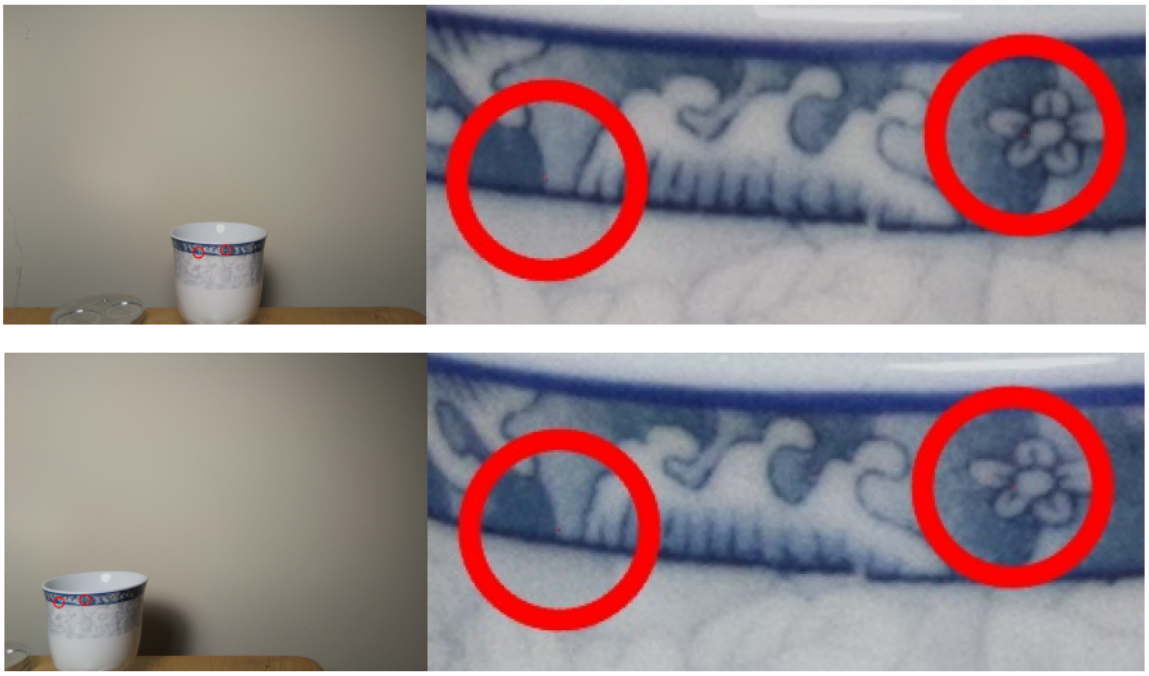


Figure 21: Right Single Camera Pair(SCP) Coordinates for: Top(Left Image) [left point (1517, 1919) : right point (1729, 1899)], Bottom(Right Image) [left point (425, 1933) : right point (626, 1914)]. Here (Left Image) and (Right Image) refer to the SCP in terms of a stereo pair so the formulas from equation set (13) follow. The SCP for the Top and Bottom images have a base length of $89.98mm$

$f = 3.04mm$, with the following points from Figure (20) and listed in Table 4.

Table 4: Left Single Camera Points from Figure 20.

Left SCP	Left point (X_{L_1}, Y_{L_1})	Right point (X_{L_2}, Y_{L_2})
Set #1	$(2585, 1942)$	$(2811, 1924)$
Set #2	$(1504, 1942)$	$(1716, 1922)$

We note that the left points, in Table 4, of each set are corresponding(matching) and calculate P_1 and P_2 with the results shown in Table 5.

We calculate the distance between P_1 and P_2 using distance formula (14) to get:

Figure 22: Physical distance measured with Digital Vernier Calipers at 16.50mm.

Table 5: Reconstruction of Left Single Camera Pair points.

P_n	Left SCP #1 (X_{L_n}, Y_{L_n})	Left SCP #2 (X_{R_n}, Y_{R_n})	3D-Coord (X_n, Y_n, Z_n)
P_1	$(2585, 1942)$	$(1504, 1942)$	$(170.18, 161.65, 225.93)$
P_2	$(2811, 1924)$	$(1716, 1922)$	$(186.00, 158.02, 223.04)$

$$\begin{aligned} Left\ SCP\ Dist &= ||P_1 - P_2|| \\ &= 16.49mm \end{aligned}$$

Similarly, we do the same for the Right Single Camera Pair of Figure 21 listed in Table 6.

Table 6: Right Single Camera Points from Figure 21.

Right SCP	Left point (X_{R_1}, Y_{R_1})	Right point (X_{R_2}, Y_{R_2})
Set #1	$(1517, 1919)$	$(1729, 1899)$
Set #2	$(425, 1933)$	$(626, 1914)$

Table 7: Reconstruction of Right Single Camera Pair points.

P_n	Right SCP #1 (X_{L_n}, Y_{L_n})	Right SCP #2 (X_{R_n}, Y_{R_n})	3D-Coord (X_n, Y_n, Z_n)
P_1	$(1517, 1919)$	$(425, 1933)$	$(80.01, 158.70, 223.66)$
P_2	$(1729, 1899)$	$(626, 1914)$	$(96.06, 155.53, 221.42)$

We calculate the distance for the Right Single Camera Pair between P_1 and P_2 using distance formula (14).

$$\begin{aligned} Right\ SCP\ Dist &= ||P_1 - P_2|| \\ &= 16.51mm \end{aligned}$$

Taking the average of the Left and Right SCP distances we get $16.50mm$ and see that this value agrees with the physical measurement of $16.50mm$ as shown in Figure 22.

4.2 Determine Average Distance Per Step

In this experiment we determine the average distance, measured by the mounted digital calipers, moved per step of the linear stage. We do the following to accomplish this (illustrated in Figure 23):

- Move the linear stage to the home position
- Advance the mechanism 160 steps to remove backlash
- Read Caliper Start value
- Move the mechanism 14000 steps
- Read Caliper End value
- Repeat Process 10x.

The result of this experiment shows an average of about $\sim 0.9 \times 10^{-2}$ mm per step (as can be seen in Table 8).

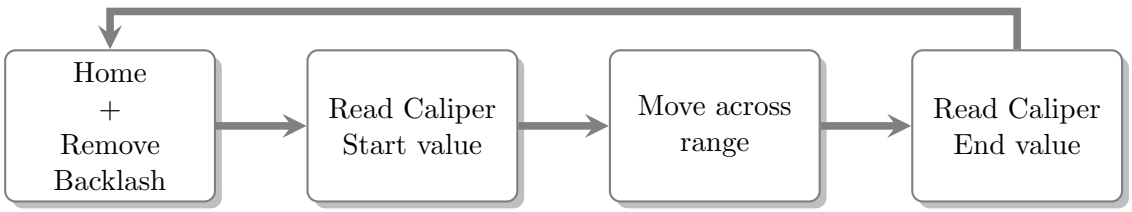


Figure 23: Determining average distance per step - Flow Chart.

Start (10^{-2} mm)	End (10^{-2} mm)	Net Distance (10^{-2} mm)	Distance/Step (10^{-2} mm)
130	12420	12290	0.87786
132	12422	12290	0.87786
130	12419	12289	0.87779
131	12419	12288	0.87771
129	12418	12289	0.87779
129	12418	12289	0.87779
129	12418	12289	0.87779
129	12418	12289	0.87779
129	12418	12289	0.87779
130	12419	12289	0.87779
		Average	**0.877796**

Table 8: Average distance per step (over 14000 steps).

4.3 Proof of Sub-Pixel Motion

In this experiment, we demonstrate the vision system's capability to move and capture sub-pixel information. This is accomplished by making slight movements (right) of the cameras mounted on the linear stage and capturing images at each position. For this experiment, we will be capturing images of a Chessboard pattern and detecting the positions of the corners.

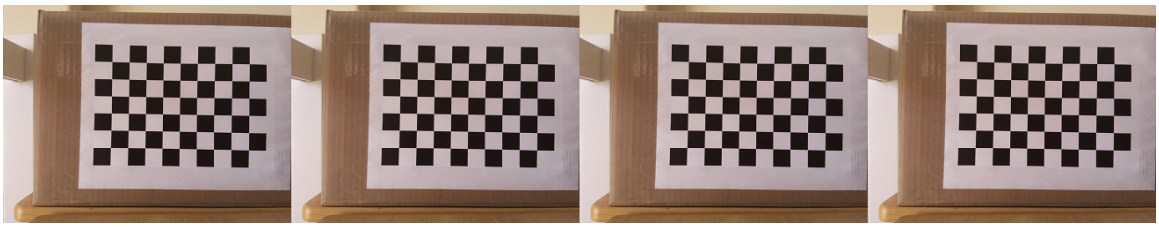


Figure 24: Images of Chessboard Pattern: L_0, L_1, L_2, and L_3, respectively.

After detection of the corners for every image at each position we calculate and compare the centroid values. The values in Table 9 show the results of the operation. The values of the calculated centroids indicate sub-pixel motion as seen by the decreasing x-coordinate value. The x-coordinate value decreases because the motion of the mechanism is to the right and results in the scene moving left in the captured images.

Image	(x, y) (px)
L_0	$(2017.422, 1143.328)$
L_1	$(2017.190, 1143.323)$
L_2	$(2017.129, 1143.335)$
L_3	$(2017.074, 1143.320)$

Table 9: Proof Sub-pixel Motion.

4.4 Sub-pixel Motion Per Pixel

This experiment answers the question: How many movements/steps of the linear stage mechanism can be made per pixel (@$\sim 30cm$)? Determining this information will let us know if we can capture an adequate number of frames to support Multi-Frame Super Resolution from a simple straight motion, since moving and being able to capture sub-pixel information is a necessary requirement for the process. We perform this process on a full resolution High Quality (HQ) (8 Mega-pixel) image and on a down-sampled, using bilinear interpolation, Low Quality (LQ) (2 Mega-pixel) image.

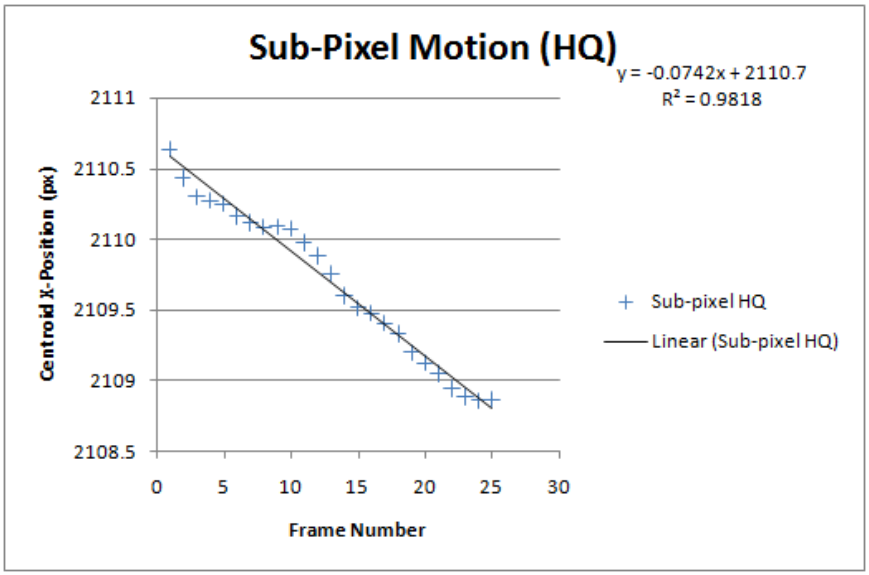


Figure 25: Sub-pixel motion mechanism movements per pixel (HQ).

The scatter plot graph in Figure 25 indicates that for the HQ image about 12 frames can be captured per pixel. This means we would need an equivalent of approximately 1.5 pixels of mechanism movements to achieve the suggested number of frames for Multi-Frame Super Resolution (MFSR) (Quality vs Quantity tradeoff) to magnify the

image by 4x. The graph indicates that there aren't enough available frames to meet the suggested number ($4^2 = 16$ frames) for simple straight motion of the mechanism within a pixel. However, if one were to relax the condition of being within 1 pixel to 2 pixels, there would certainly be enough room to capture frames for MFSR with the only loss of enhancement being a 1 pixel border around the resulting image. The other solution would be to relax the condition of simple straight motion and allow for moving the mechanism backwards and capturing over the same area in the opposite direction. Due to backlash, it is unlikely the mechanism would land in exactly the same spots as the first pass and thus would collect unique information. This introduces the problem of having to check the uniqueness of the reverse pass and is beyond the scope of this document, so it reserved for future research. For our current research, we are only concerned with restoring the down-sampled LQ images to HQ through the use of MFSR. Since, the resolution of the LQ image has dimensions with both half the height and width of the original HQ image, we expect that the number of frames within a pixel to be about double that of the HQ image.

Examining Figure 26 confirms our expectation and we see that we get about 25 image frames per pixel and is well within the suggest number of frames for MFSR.

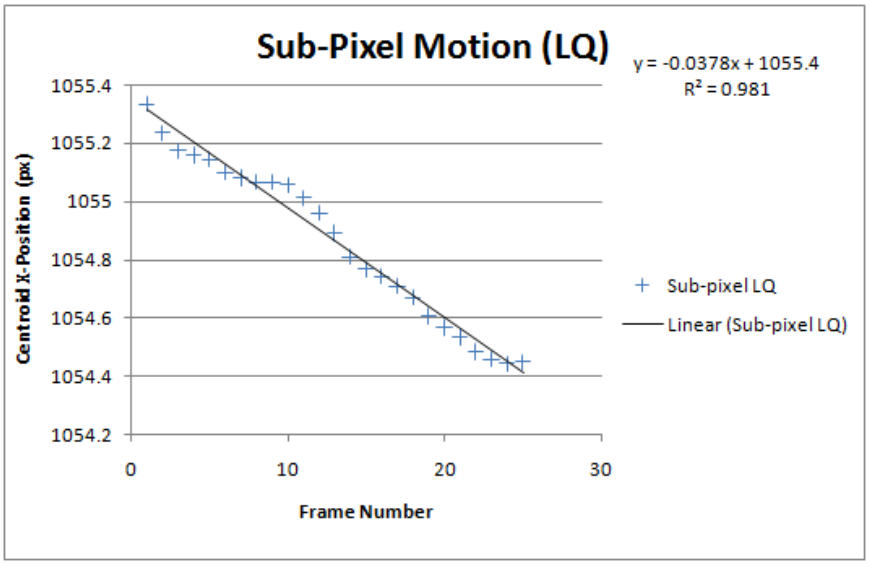


Figure 26: Sub-pixel motion mechanism movements per pixel (LQ).

4.5 Super Resolution Comparison Testing (SISR and MFSR)

In this experiment, we compare the up-sampling techniques Bicubic, Single Image Super Resolution (SISR), and Multi-Frame Super Resolution (MFSR) by applying them to down-sampled Low Resolution (LR) images to restore them to their original resolution. The up-sampled images are compared to the original images using the following metrics: Mean Squared Error (MSE), Peak Signal to Noise Ration (PSNR), and Structural SIMilarity (SSIM) (See Chapter 2.4.7 for more detail).

For this experiment, we will be using the following SISR algorithms:

- Enhanced Deep Super Resolution (EDSR) [40]
- Deep Back Projection Networks (DBPN) [72]

and for the MFSR Algorithm, we will use the following simple algorithm (illustrated

in Figure 27):

- Input images are bicubically resized / up-sampled
- Up-sampled images are registered with Enhanced Correlation Coefficient (ECC) Maximization Image Alignment [23]
- Median image is reconstructed from registered images

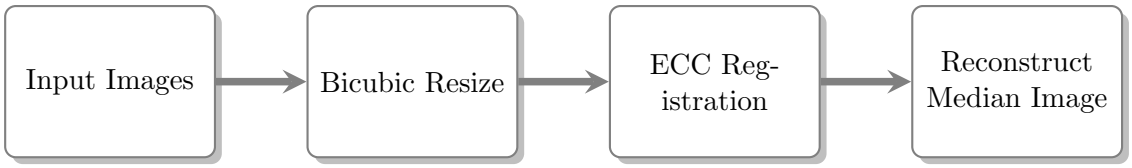


Figure 27: Simple Multi-Frame Super Resolution (MFSR) Algorithm - Flow Chart.

The first test will apply the up-sampling algorithms: Bicubic, SISR (EDSR, DBPN), and the simple MFSR algorithm outlined in Figure 27 to an image of a chessboard pattern (Figure 28) that has been down-sampled. In the case of MFSR, 25 LR images were used in the process which were obtained in a similarly fashion to Chapter 4.4. The goal was to use the up-sampling algorithms to restore the down-sampled image to the original resolution and determine which algorithm produced the closest output compared to the original image using the metrics (MSE, PSNR, SSIM).

Examining the results, shown in Figure 29 and Table 10, we see that EDSR outperforms all of the algorithms in terms of MSE, PSNR, and SSIM, with Bicubic in a close second place. DBPN comes in-front of MFSR for all metrics except SSIM. These results were slightly surprising and unexpected. Since, a closer examination of

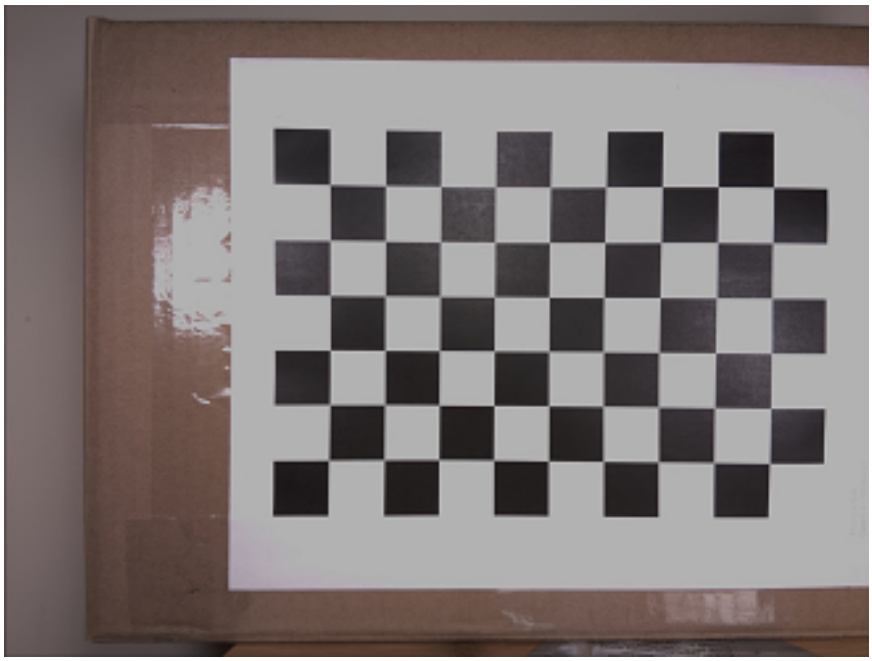


Figure 28: Image of Chessboard Pattern used for the comparison process.

the resulting output images of the algorithms revealed an intuitively different result than shown in the metrics.

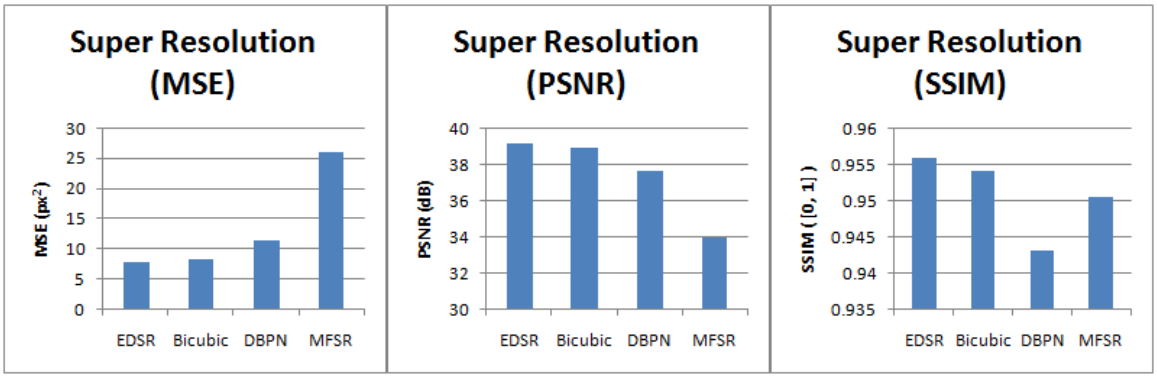


Figure 29: SR Comparison #1 Bar Graph (Chessboard).

By subjectively doing a visual comparison of the up-sampled images versus the original HQ image. We saw that noise from the original HQ image was propagated through, in varying amounts, for all of the algorithms with MFSR expressing the

Method	(px^2) MSE	(dB) PSNR	[0, 1] SSIM
EDSR	**7.88**	**39.17**	**0.9558**
Bicubic	8.35	38.91	0.9541
DBPN	11.32	37.59	0.9432
MFSR	25.85	34.01	0.9504

Table 10: SR Comparison #1 Testing (MSE, PSNR, SSIM) (Chessboard).

least (as shown in Figure 30). This made the results from Figure 29 and Table 10 understandable and coherent with intuitive reasoning.

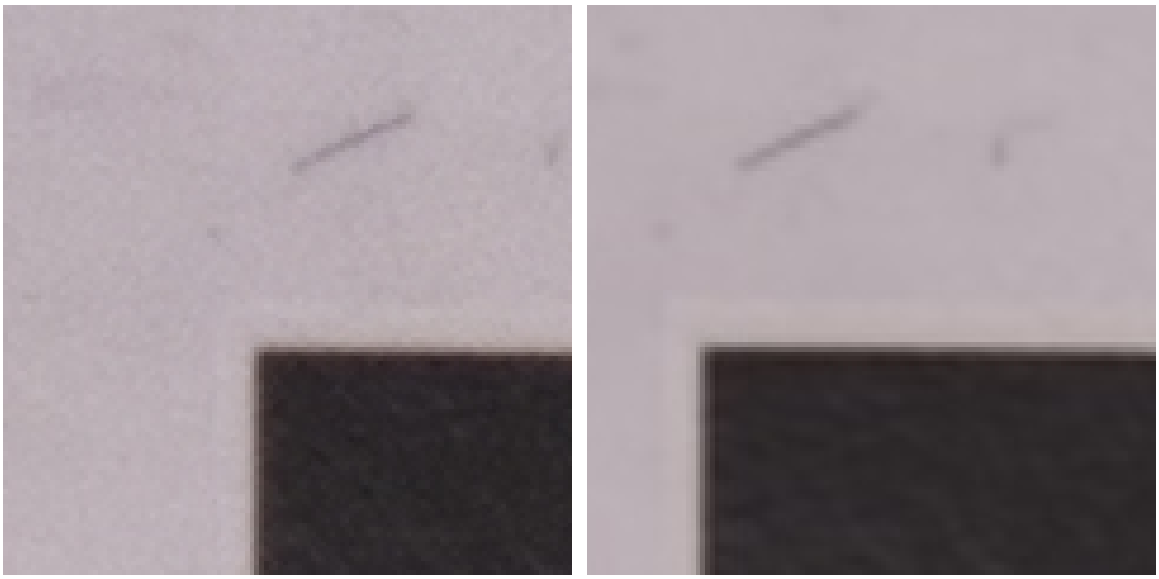


Figure 30: Noise comparison HQ (left) vs MFSR (right). The HQ image appears grainier than the MFSR image.

The noise reduction when combining multiple images, as outlined in Chapter 2.4.6, accounts for the MFSR image appearing less grainy than the HQ image. So, the MFSR image has less noise than the HQ image (ground truth) which gives the reason for the MFSR results. To reduce the noise in the HQ image we can take the 25 HQ images used to make the LQ image set for MFSR and combine (stack) them to produce a HQ image (stacked) with less noise.

For this process, we use the following stacking algorithm (illustrated in Figure 31):

- Take HQ Input Images and Perform ECC Alignment / Registration
- Use Aligned / Registered Images to Reconstruct an Average Image

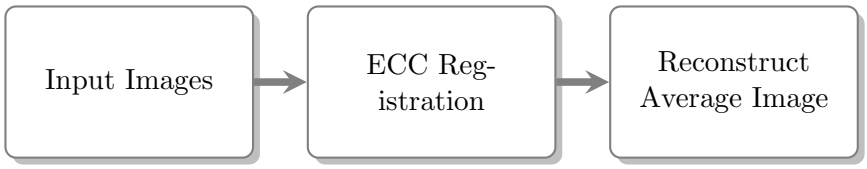


Figure 31: Noise Removing Stacking Algorithm - Flow Chart.

The 25 HQ images used to make the LQ image set for the MFSR process were used to produce a stacked HQ image, which has a noise reduction of approximately 80% less than the original HQ image, as calculated from Equation (28) or taken from Table 1.

The stacked HQ image was then down-sampled to produce a stacked LQ image, which was then used with the up-sampling algorithms: Bicubic, SISR (EDSR, DBPN) to restore the down-sampled image to the original resolution. These images along with the MFSR image were then used in the same way as the first comparison test to determine which algorithm produced the closest output compared to the original image using the metrics (MSE, PSNR, SSIM).

The results, shown in Figure 32 and Table 11, indicate that the MFSR algorithm is the top performer in terms of MSE, PSNR, and SSIM, with EDSR in second followed by Bicubic and then DBPN. All non-MFSR algorithms showed improved metric scores from the reduction of input noise and comparison with the stacked (noise reduced) HQ image.

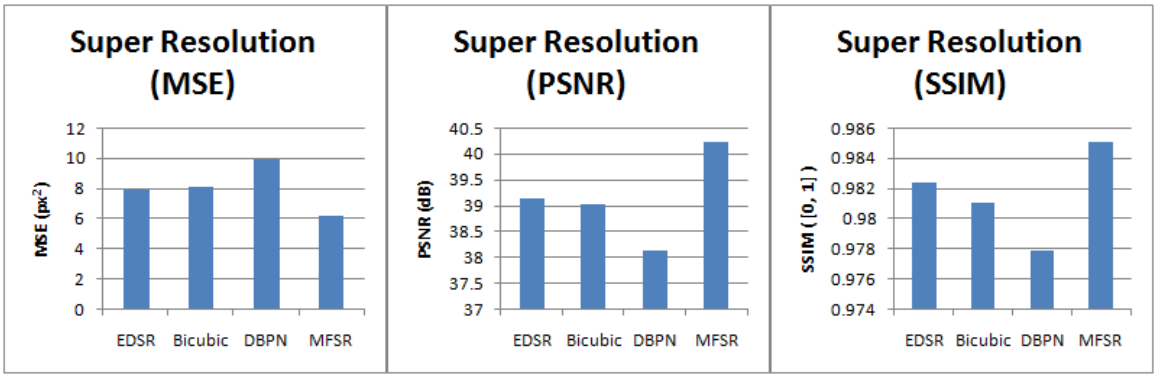


Figure 32: SR Comparison #2 Bar Graph (Chessboard w/o Noise).

Method	(px^2) MSE	(dB) PSNR	[0, 1] SSIM
EDSR	7.93	39.14	0.9824
Bicubic	8.13	39.03	0.9811
DBPN	9.95	38.15	0.9779
MFSR	**6.18**	**40.22**	**0.9851**

Table 11: SR Comparison #2 Testing (MSE, PSNR, SSIM) (Chessboard w/o Noise).

The stacked comparison process was then repeated again with an image of a feature rich rock (shown in Figure 33). Similar results were obtained, as indicated in Figure 34 and Table 12, showing the MFSR dominating the MSE, PSNR, and SSIM scores, followed by EDSR, Bicubic, and DBPN.

Method	(px^2) MSE	(dB) PSNR	[0, 1] SSIM
EDSR	4.47	41.63	0.9784
Bicubic	4.61	41.49	0.9779
DBPN	5.50	40.73	0.9734
MFSR	**3.00**	**43.36**	**0.9815**

Table 12: SR Comparison #3 Testing (MSE, PSNR, SSIM) (Rock).

Figure 33: Image of Rock used for the comparison process.

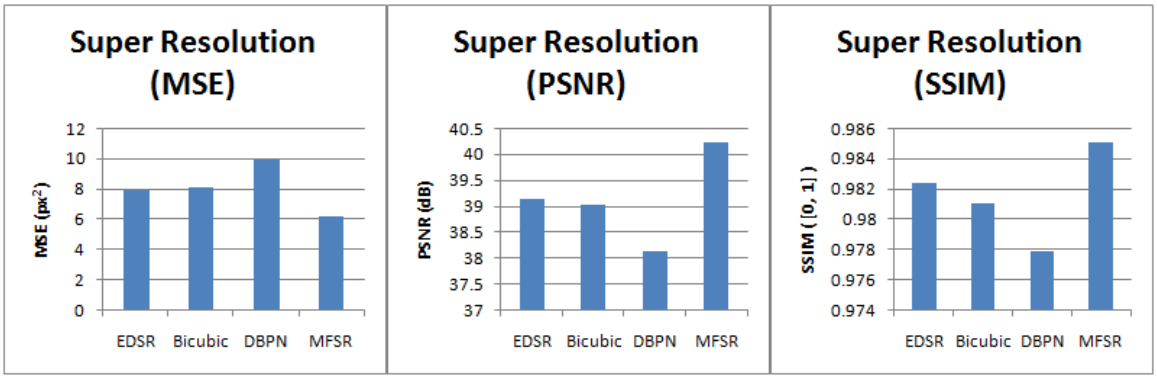


Figure 34: SR Comparison #3 Bar Graph (Rock).

4.6 3D-Model Comparisons

In this section, we develop reference and comparison models using multi-view stereo and the proposed architecture which enhances the input stage. In particular, we show that multi-view stereo when combined with multi-frame super resolution produces a more accurate 3D-reconstruction. The overall experimental results show that the generated models, using our technique, have point clouds with average mean, median, and max distances of 4.3%, 8.8%, and 6% closer to the reference model, respectively. This indicates an improvement in 3D-reconstruction using our technique. The use of multi-frame super resolution in conjunction with the multi-view stereo framework is a practical solution for enhancing the quality of 3D-reconstruction and shows promising results over single image up-sampling techniques.

4.6.1 Method (Model Comparison)

Our goal is to improve the output of the multi-view stereo framework by directly enhancing the input. We propose a method that captures images from sub-pixel camera movements. These images can be recombined, using super resolution, to create a higher resolution image that contains more information or details than each image alone. For the super resolution process, we chose a multi-frame algorithm because it only reconstructs information from details obtained from the scene, unlike SISR algorithms which use inferred details from learned priori and can potentially suffer from artifacts. We use a simple interpolation-based method, also known as a

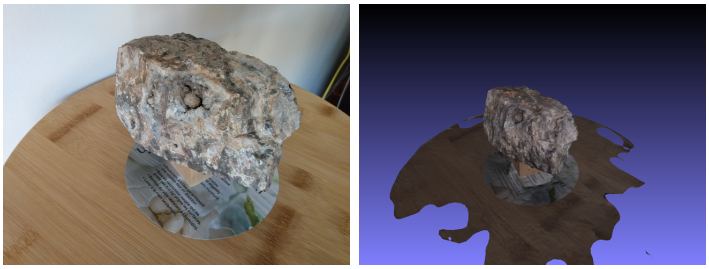
direct method, for MFSR. First, we bicubically up-scale the LR images and then use ECC to register the images, followed by a median filter to reconstruct and form the final super resolved image.

The proposed framework requires a comparison to a ground truth reference model. However, the availability of reference models and data sets, that are suitable for MFSR with MVS, present a difficulty due to the sub-pixel motion requirements. So, the popular data sets such as Middlebury, KITTI, Stretcha, and BlendedMVS are not suitable. One solution would be to generate reference models using a high quality laser scanner, but our research laboratory does not have access to such a device.

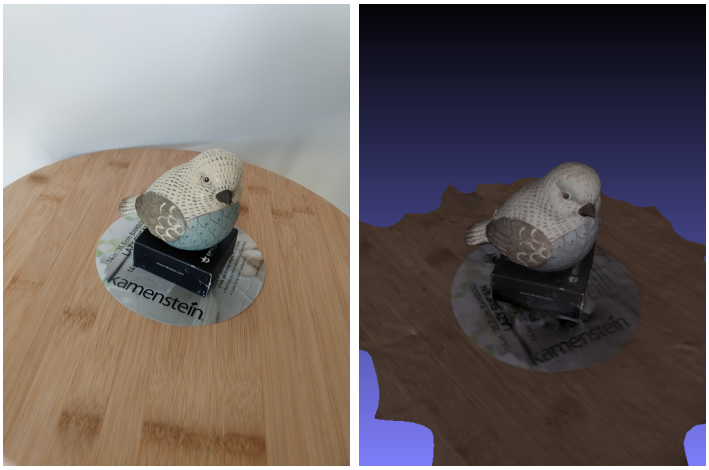
So, instead we utilize a method that is similar to the literature for super resolution [41, 72]. This method uses down-sampled HR images that are up-sampled back up to the original resolution by the super resolution algorithm and compared to the original HR image as reference. Our process generates models through the MVS pipeline using the original HR images and up-sampled LR images (generated from down-sampled HR images) as input for reference and comparison models, respectively. These models are compared by measuring the average, median, and max distance away from the reference model. The model that has the closest metrics to the reference model is determined to be the best model.

In addition to the HQ and up-sampled models, we also generate a standard model solely from the LR images which represents a base starting point without enhancement and provides a reference to what a model would be if the starting point was at the

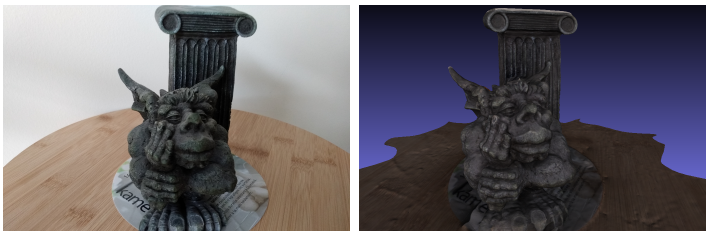
lower resolution. This gives us a relative comparison for the other methods versus a standard starting point.



(a) Photo of Rock (b) 3D-Model of Rock



(c) Photo of Bird (d) 3D-Model of Bird



(e) Photo of Gargoyle (f) 3D-Model of Gargoyle

Figure 35: Images and Models (Rock, Bird, Gargoyle).

In our experiment, we captured image data and developed 3D-model reconstructions of a rock, bird, and gargoyle [Figure 35]. The image capture process used a turntable, that rotated in 18 degree increments, with the object placed roughly in the center for a full 360 degree capture. For each rotational increment, the vision system

captured images from each camera at full resolution (3280 x 2464). This included 25 high resolution images (per camera) that contained sub-pixel motion from the step-movements of the linear stage. So, the total number of images captured was $1000 = (20\ incremental\ rotations)(25\ sub\text{-}pixel\ frames)(2\ cameras)$. We will refer to these full resolution images as high quality (HQ) images and the down-sampled versions of these images as low quality (LQ) images.

Next, we created HQ and standard reference models, [Figure 35(b,d,f)], by using the HQ and LQ images as input through the MVS framework [as illustrated in Figure 36(blue)], respectively. This step only utilized the first image of each rotational increment and did not use the images captured from successive movements of the linear stage.

Similarly, we created comparison models using up-sampled LQ images [Figure 36(red)]. The LQ images were up-sampled to the same resolution as the original HQ images using: bicubic, SISR (enhanced deep super resolution (EDSR), DBPN), and MFSR algorithms. For the up-sampling algorithms that used a single image, we took the first image of each rotational increment, like we did for the HQ reference model, and ignored the images captured during the step-movements of the linear stage. For the MFSR algorithm, we used the first image of each rotational increment as the primary image in the registration process for the rest of the images captured during the step-movements of the linear stage. After the registration process the images were combined to produced a SR image which was used as input to the MVS pipeline. (Each incremental rotation produced one super-resolved image for a total of 20 MFSR

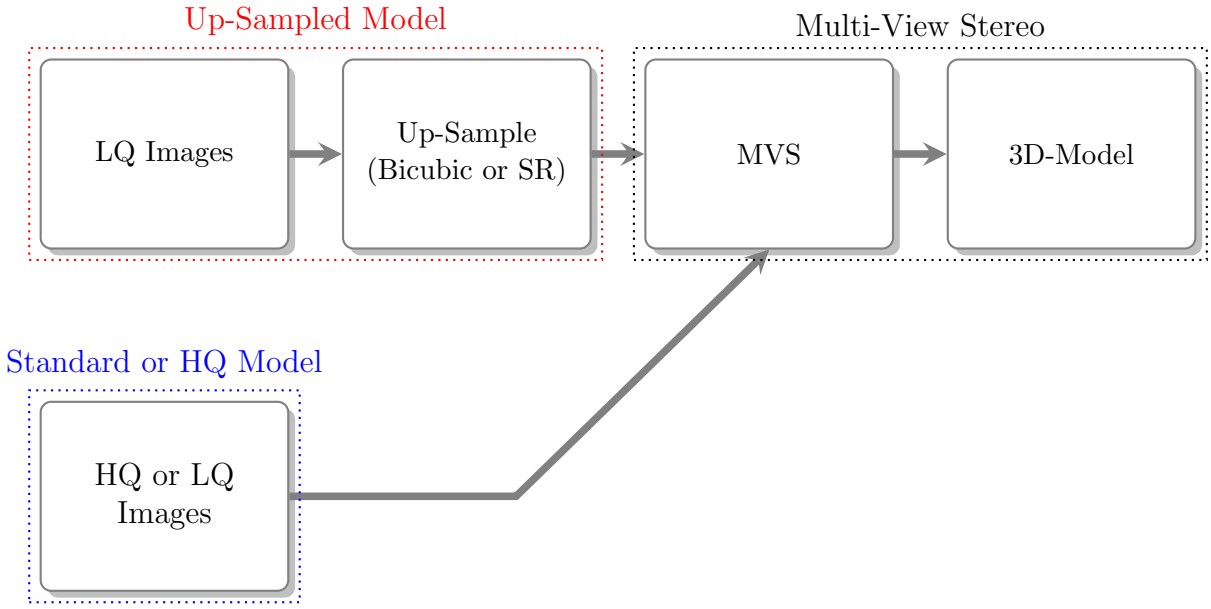


Figure 36: Up-sampled / high quality / standard - multi-view stereo framework - flow chart.

images per camera per full revolution.)

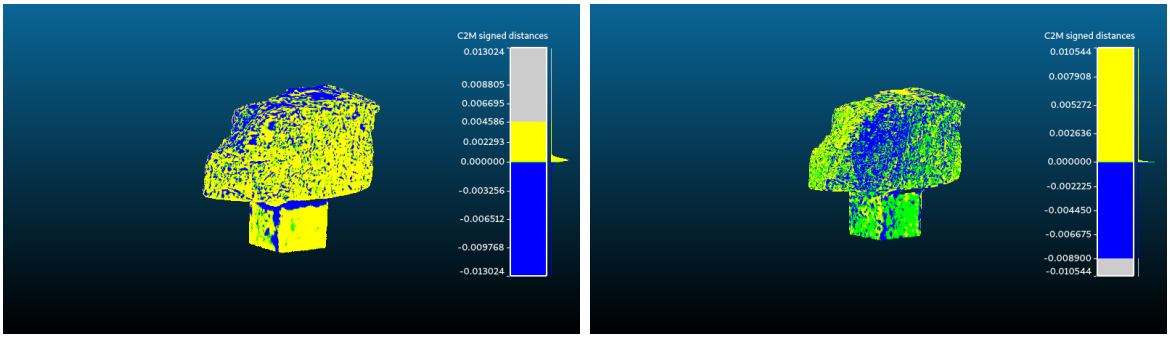
In total, 40 images were used during the input stage of MVS framework for the development of each model (HQ, bicubic, SISR, and MFSR). The models were evaluated and compared using the open source software CloudCompare (v2.11.3) [35]. This was accomplished by comparing the point clouds of the comparison models to the reference models using the metrics: average, median, max. The model that had metrics closest to the HQ reference model indicated the better 3D-reconstruction.

4.6.2 Results (Model Comparison)

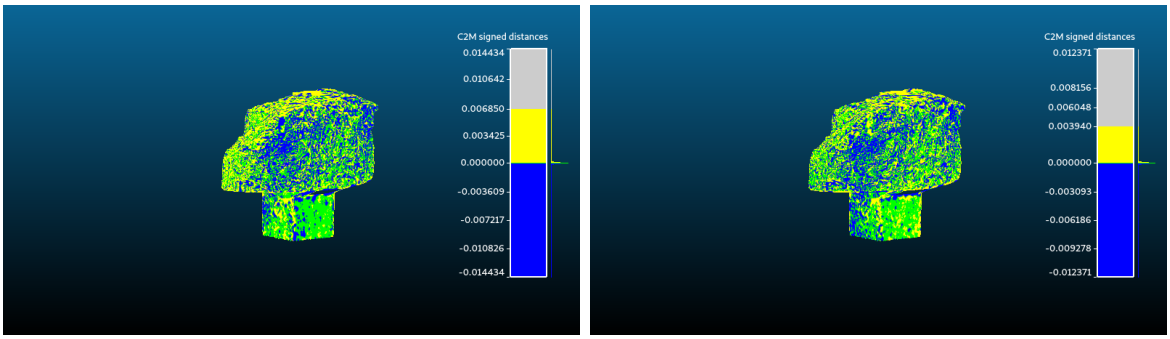
The comparison results of the point clouds to reference models have been represented in [Figures 37, 38, 39] for the rock, bird, and gargoyle models. These figures are colour coded and visually represent the difference relationship between the comparison models and reference models, such that green represents the zero distance (within 3.0×10^{-4} pixels (px) of the reference model), yellow represents above zero, and blue represents below zero. Having more green indicates that the comparison model is closer to the reference model, which means more points are at a zero distance.

The numerical results for the rock model, [Figure 40] and [Table 13], show that the MFSR model is closest to the reference model. Our technique generates a model that outperforms all of the comparison models with mean and max distances that are at least 2.0% and 11.6% closer to the reference model, respectively. The DBPN model has an equivalent median but performs worse in all other categories. However, DPBN is closely matched to the EDSR model but is slightly ahead due to the better median and max values. The bicubic model has the worst performance but has the second best max value that is only 11.6% farther from the reference model compared to MFSR.

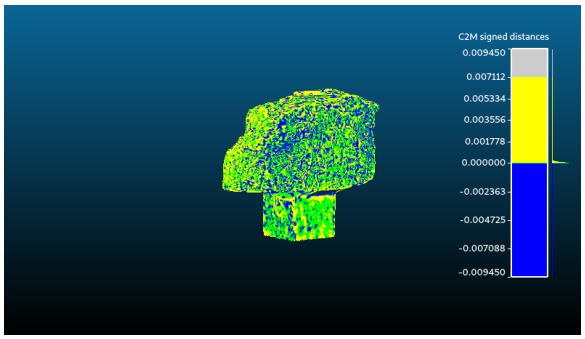
The numerical results for the bird model, [Figure 41] and [Table 14], indicate that the MFSR model performs better than the comparison models and has mean and median distances that are at least 0.5% and 11.6% closer to the reference model, respectively. The bicubic model has a smaller (better) max distance from the reference model, at



(a) Standard (b) Bicubic



(c) EDSR (d) DBPN

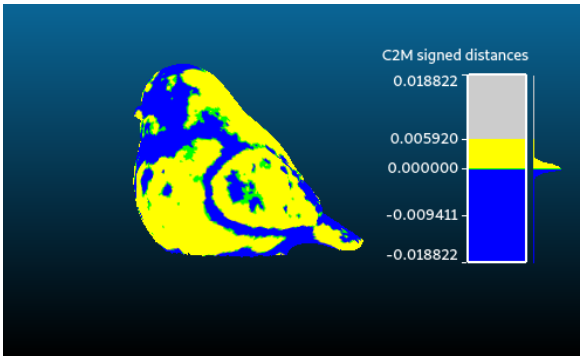


(e) MFSR

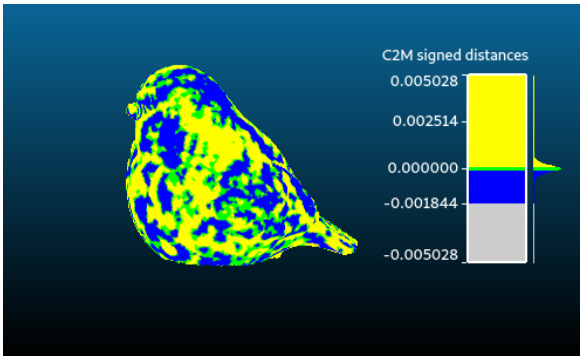
Figure 37: Visual cloud compare (Rock).

Model	Mean (10^{-2} mm)	St.dev. (10^{-1} mm)	Median (10^{-2} mm)	Max (mm)
Standard	9.952(310%)	1.433	6.858(347%)	3.224(37.8%)
Bicubic	2.525(4.1%)	**0.436**	1.659(8.1%)	2.610(11.6%)
EDSR	2.476(2.0%)	0.453	1.609(4.8%)	3.573(52.7%)
DBPN	2.500(3.1%)	0.493	**1.535(0.0%)**	3.063(30.9%)
MFSR	**2.426**	0.456	**1.535**	**2.340**

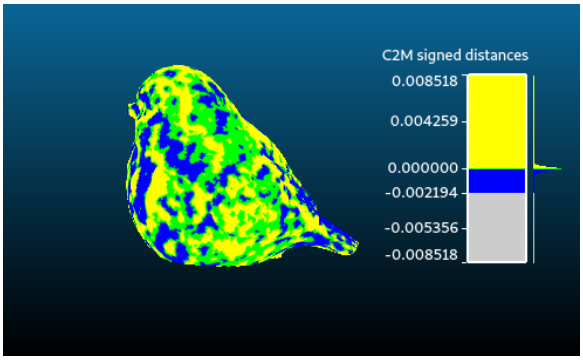
Table 13: Cloud compare (Rock): Standard, Bicubic, EDSR, DBPN, and MFSR.



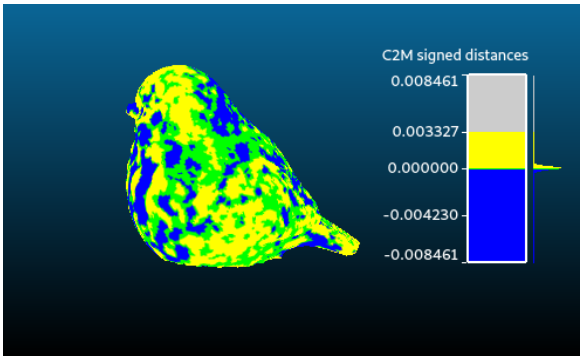
(a) Standard



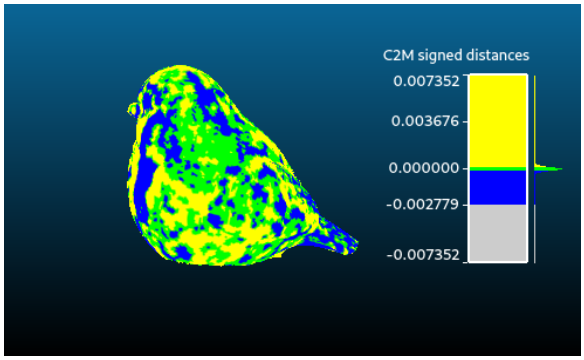
(b) Bicubic



(c) EDSR

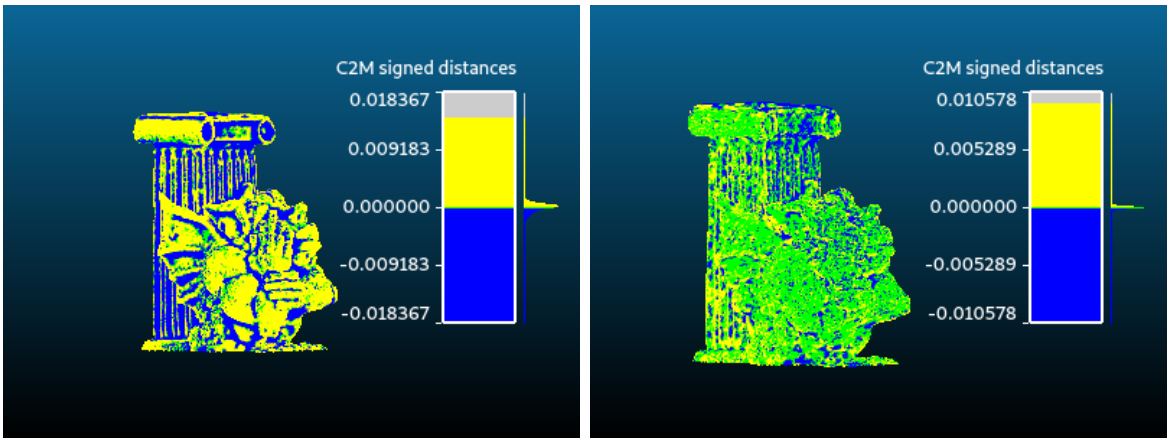


(d) DBPN

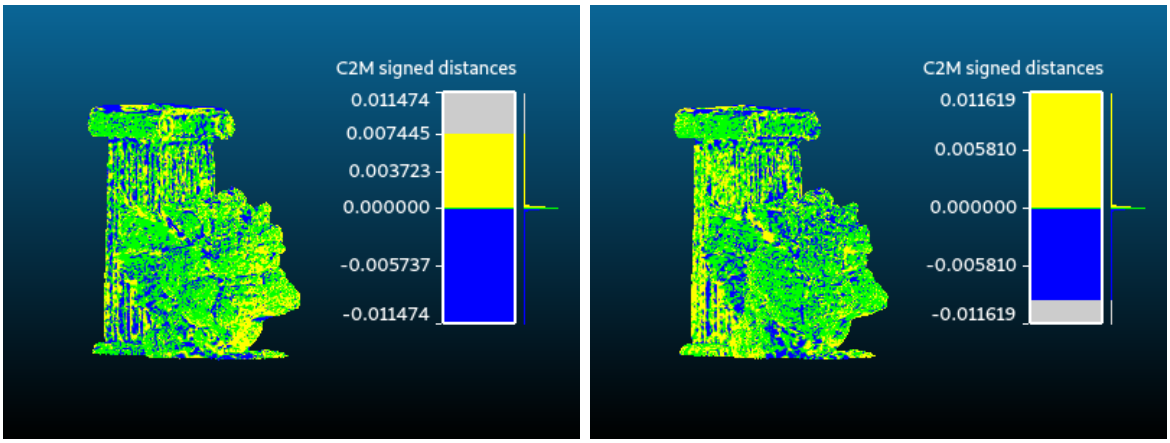


(e) MFSR

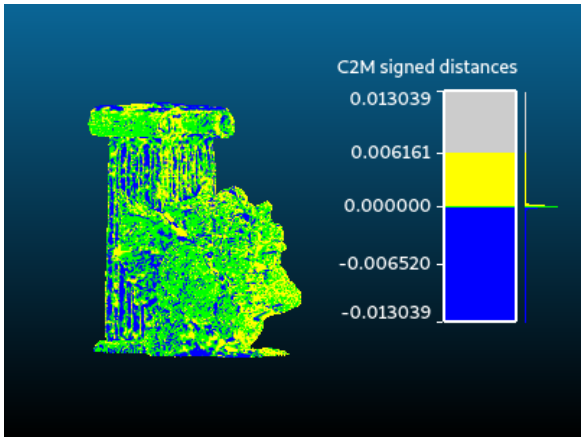
Figure 38: Visual cloud compare (Bird).



(a) Standard (b) Bicubic



(c) EDSR (d) DBPN



(e) MFSR

Figure 39: Visual cloud compare (Gargoyle).

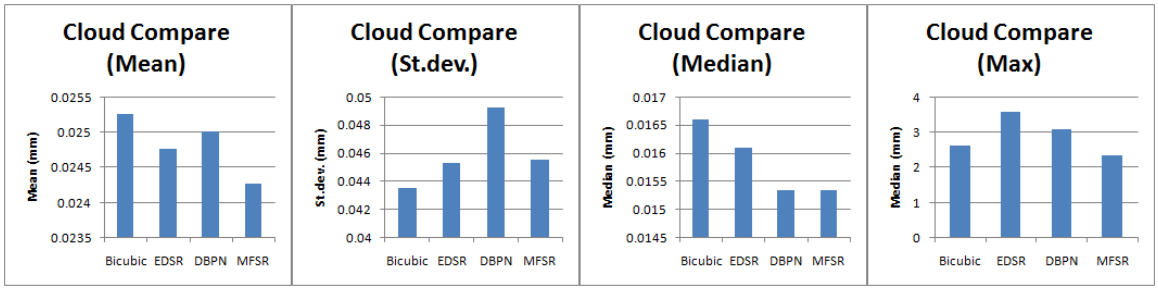


Figure 40: Cloud compare bar graph (Rock): Bicubic, EDSR, DBPN, and MFSR.

31.4% closer, but performs considerably worse in terms of mean and median at 22.6% and 38.8% farther from the reference model, respectively. The mean and median distances make the bicubic performance worse than both the second and third placed DBPN and EDSR, respectively, despite having a better max distance.

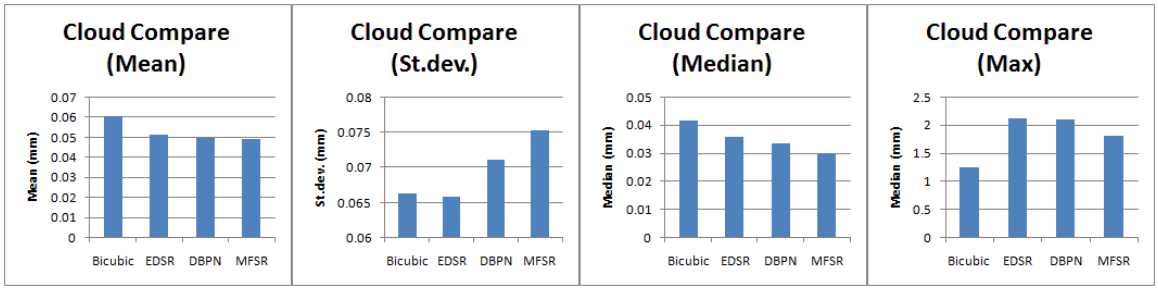


Figure 41: Cloud compare bar graph (Bird): Bicubic, EDSR, DBPN, and MFSR.

Model	Mean (10^{-1} mm)	St.dev. (10^{-1} mm)	Median (10^{-1} mm)	Max (mm)
Standard	1.966(299%)	2.117	1.483(395%)	4.660(157%)
Bicubic	0.604(22.6%)	0.663	0.416(38.8%)	**1.245(-31.4%)**
EDSR	0.510(3.5%)	**0.659**	0.359(19.8%)	2.108(16.1%)
DBPN	0.495(0.5%)	0.711	0.334(11.6%)	2.096(15.5%)
MFSR	**0.493**	0.753	**0.300**	1.815

Table 14: Cloud compare (Bird): Standard, Bicubic, EDSR, DBPN, and MFSR.

The numerical results for the gargoyle model, [Figure 42] and [Table 15], show that all models, excluding standard, performed relatively equivalent having mean and median

values within 0.001mm (at most 2.6%) of each other. For this model, bicubic had slightly better metrics and was followed by DBPN, EDSR, and MFSR. These results are in contrast to the previous rock and bird comparisons, where bicubic had the worse performance. This can be attributed to the dark and low contrast surfaces of the gargoyle, which camouflage and make details less evident. With less evident surface details the enhancement ability of SR algorithms are reduced. Thus, can cause problems with correspondence matching and can reduce the quality of 3D-reconstruction. This supports the reason why the models had relatively equivalent performance.

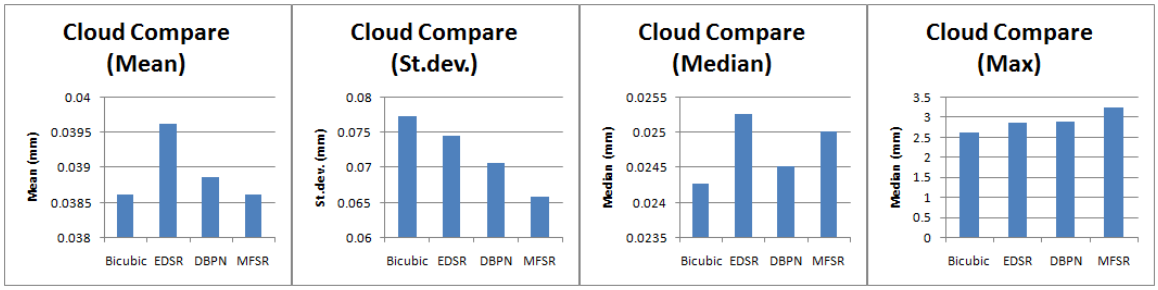


Figure 42: Cloud compare bar graph (Gargoyle): Bicubic, EDSR, DBPN, and MFSR.

Model	Mean (10^{-1} mm)	St.dev. (10^{-1} mm)	Median (10^{-1} mm)	Max (mm)
Standard	2.028(425%)	2.716	1.263(405%)	4.547(41%)
Bicubic	**0.386(0.0%)**	0.772	**0.243(-3.0%)**	**2.618(-18.9%)**
EDSR	0.396(2.6%)	0.745	0.253(1.0%)	2.841(-11.9%)
DBPN	0.389(0.6%)	0.706	0.245(-2.0%)	2.877(-10.8%)
MFSR	**0.386**	**0.659**	0.250	3.226

Table 15: Cloud compare (Gargoyle): Standard, Bicubic, EDSR, DBPN, and MFSR.

Summarizing the experimental results, we see that all up-sampling methods (bicubic, EDSR, DBPN, MFSR) improve MVS 3D-reconstruction, [Table 16]. Of particular interest, we see that the average percentage mean, median and max distances for the

Model	Mean	Median	Max
Standard	344.7%	382.3%	8.5%
Bicubic	8.9%	14.6%	-12.9%
EDSR	2.7%	8.6%	19.0%
DBPN	1.4%	3.2%	11.9%
Overall Avg.	4.3%	8.8%	6.0%

Table 16: Summary of average percentage distance to reference models relative to MFSR models: Standard, Mean, Median, and Max.

MFSR models are 344.7%, 382.3%, and 8.5% closer to the reference models than the standard models, respectively. MFSR performs better than bicubic in terms of mean and median with values of 8.9% and 14.6%, respectively. However, bicubic has an average maximum of 12.9% closer to the reference than MFSR, but the maximum values are outliers and do not represent the bulk of the data which is represented by the higher mean and median values. Our technique also outperforms the SISR methods having an average mean, median, and max distance that is at least 1.4%, 3.2%, and 11.9% closer to the reference models, respectively.

Combining the experimental results, we see that there is an improvement in 3D-reconstruction using our technique over the other up-sampling methods with an overall average mean, median, and max distances that are 4.3%, 8.8%, and 6% closer to the reference models, respectively.

4.7 Up-Sampling and 3D-Reconstruction Time Comparison

In this subsection, we examine and compare the average up-sampling and 3D-reconstruction times for the various methods (bicubic, EDSR, DBPN, MFSR) [Table 17]. These results are based on the average times for image up-sampling and 3D-reconstruction of the three models (Rock, Bird, and Gargoyle). All times are generated using a CPU (Intel Core i7).

The direct use of the LR images in the standard model performed the best total time, at 6.23 mins, but it generated less accurate models (farthest from the reference). We observe that the 3D-reconstruction times for the other methods are about 4 times longer in comparison to the standard model. This can be attributed to the up-sampled images having length and width dimensions that are twice the size, thus having 4 times the resolution.

Model	Up-Sampling (mins)	(mins/img)	Reconstruction (mins)
Standard	0.00	0.00	6.23
Bicubic	0.39	0.01	25.53
EDSR	357.80	8.94	27.33
DBPN	292.79	7.32	26.93
MFSR	21.51	0.54	25.41

Table 17: Up-sampling and 3D-reconstruction time comparisons (for sets of 40 images).

The results indicate that MFSR produces the best quality model, at an average time of 46.92 mins, which outperforms the other methods and is at least 6.8x faster than the SISR methods. The bicubic results are faster, at 25.92 mins, but produces a less

accurate model.

Overall, it can be seen that up-sampling the input images results in a more accurate 3D-reconstruction with MFSR yielding the best trade-off between model quality and reconstruction time.

5 Conclusions and Future Work

In this dissertation, we explored the idea of using multi-frame super resolution to improve 3D-reconstruction at the input stage of the multi-view stereo framework. We showed comparisons to bicubic and SISR (DBPN and EDSR) up-sampling as inputs to MVS for 3D-model generation versus the MFSR generated equivalent.

The overall experimental results show that the generated models, using our technique, have point clouds with average mean, median, and max distances of 4.3%, 8.8%, and 6% closer to the reference model when compared to the single image up-sampling analogs, respectively. This indicates an improvement in 3D-reconstruction using our technique. In addition, our technique has a significant speed advantage over the SISR analogs being at least 6.8x faster. We addressed the limitations of prior work by focusing on indoor small scale objects and using an MFSR based algorithm for super resolution, which only uses data from scene images, does not require training, and has built-in noise reduction. However, our technique is limited when applied to dark and low-contrast surfaces. These surfaces cause problems with correspondence matching and reduces the quality of reconstruction. This problem is not limited to our technique but in general is a common problem for all image based reconstruction. Under these conditions, we show that our technique performs equivalently to the other methods.

The use of MFSR in conjunction with the MVS pipeline is a practical solution for enhancing the quality of 3D-reconstruction by increasing the spatial resolution of the

input stage. It shows promising results over simple up-sampling and SISR analogs. Further improvements of this technique may be possible using different registration techniques, back propagation, or different reconstruction filters with the MFSR algorithm. In addition, expanding the acquisition process of the MFSR-MVS framework using burst photography from devices such as mobile phones and drones offer interesting possibilities for capturing high quality 3D-models with convenience. The possibility of capturing beautiful nature scenes or bird's eye view reconstruction are examples of practical applications.

Summary of Contributions:

- Constructed vision system with image acquisition mounted on a linear rail with precision feedback
- Up-sampling alone improves 3D-reconstruction with MFSR showing the most improvement
- Overall experimental results with mean (4.3%), median (8.8%), and max (6%) closer to the reference models averaged over all models
- At least mean (1.4%), median (3.2%), and max (11.9%) closer to the reference models than state of the art SISR analogs
- At least 6.8x faster combined up-sampling and 3D-reconstruction times than state of the art SISR analogs

References

[1] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. Building rome in a day. In *2009 IEEE 12th International Conference on Computer Vision*, pages 72–79, 2009.

[2] Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst. Freak: Fast retina keypoint. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 510–517, 2012.

[3] Furui Bai, Wen Lu, Lin Zha, Xiaopeng Sun, and Ruoxuan Guan. Non-local hierarchical residual network for single image super-resolution. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2821–2825, 2019.

[4] Simon Baker, Stefan Roth, Daniel Scharstein, Michael J. Black, J.P. Lewis, and Richard Szeliski. A database and evaluation methodology for optical flow. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[5] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008. Similarity Matching in Computer Vision and Multimedia.

[6] Stefano Berretti, Alberto Del Bimbo, and Pietro Pala. Superfaces: A super-resolution model for 3d faces. In Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, editors, *Computer Vision – ECCV 2012. Workshops and Demonstrations*, pages 73–82, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[7] Sai Bi, Nima Khademi Kalantari, and Ravi Ramamoorthi. Patch-based optimization for image-based texture mapping. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2017)*, 36(4), 2017.

[8] Peter .J. Bickel and Kjell A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics Volume I.* Chapman and Hall/CRC, New York, 1 edition, 2015.

[9] Wutthigrai Boonsuk. Investigating effects of stereo baseline distance on accuracy of 3d projection for industrial robotic applications. In *Proceedings of The 2016 IAJC-ISAM International Conference*, 2016.

[10] Amit Bracha, Noam Rotstein, David Bensad, Ron Slossberg, and Ron Kimmel. Depth refinement for improved stereo reconstruction, 12 2021.

[11] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[12] Calum Burns, Aurlien Plyer, and Frdric Champagnat. Texture super-resolution for 3d reconstruction. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pages 350–353, 2017.

[13] Minjie Cao, Zhe Liu, Xueting Huang, and Zhuoxuan Shen. Research for face image super-resolution reconstruction based on wavelet transform and srgan. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 5, pages 448–451, 2021.

[14] Dan Cernea. OpenMVS: Multi-view stereo reconstruction library. 2020.

[15] Canqiang Chen, Chunmei Qing, Xiangmin Xu, and Patrick Dickinson. Cross parallax attention network for stereo image super-resolution. *IEEE Transactions on Multimedia*, 24:202–216, 2022.

[16] Yu-Ping Chiu, Jin-Jang Leou, and Han-Hui Hsiao. Super-resolution reconstruction for kinect 3d data. In *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2712–2715, 2014.

[17] Qi Dai, Juncheng Li, Qiaosi Yi, Faming Fang, and Guixu Zhang. Feedback network for mutually boosted stereo image super-resolution and disparity estimation. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[18] Daniele De Gregorio, Matteo Poggi, Pierluigi Zama Ramirez, Gianluca Palli, Stefano Mattoccia, and Luigi Di Stefano. Beyond the baseline: 3d reconstruction of tiny objects with ¡underline¿si¡/underline¿ngle camera ¡underline¿ste¡/underline¿reo ¡underline¿r¡/underline¿obot. *IEEE Access*, 9:119755–119765, 2021.

[19] Anand Deshpande, Prashant P. Patavardhan, and D.H. Rao. Iterated back projection based super-resolution for iris feature extraction. *Procedia Computer Science*, 48:269–275, 2015. International Conference on Computer, Communication and Convergence (ICCC 2015).

[20] Arturo Donate, Xiuwen Liu, and Emmanuel G. Collins. Efficient path-based stereo matching with subpixel accuracy. *IEEE Transactions on Systems, Man,*

and Cybernetics, Part B (Cybernetics), 41(1):183–195, 2011.

[21] Pierre Drap and Julien Lefèvre. An Exact Formula for Calculating Inverse Radial Lens Distortions. *Sensors*, 16(6):807, 2016.

[22] M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE Transactions on Image Processing*, 6(12):1646–1658, 1997.

[23] Georgios D. Evangelidis and Emmanouil Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865, 2008.

[24] Gabriele Facciolo, Carlo De Franchis, and Enric Meinhardt-Llopis. Automatic 3d reconstruction from multi-date satellite images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1542–1551, 2017.

[25] S. Farsiu, M.D. Robinson, M. Elad, and P. Milanfar. Fast and robust multiframe super resolution. *IEEE Transactions on Image Processing*, 13(10):1327–1344, 2004.

[26] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, and Marc Pollefeys. Building rome on a cloudless day. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vi-*

sion – ECCV 2010, pages 368–381, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[27] W.T. Freeman, T.R. Jones, and E.C. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65, 2002.

[28] Yasutaka Furukawa and Carlos Hernndez. 2015.

[29] David Gallup. *Multi-baseline Stereo*, pages 498–501. Springer US, Boston, MA, 2014.

[30] David Gallup, Jan-Michael Frahm, Philippos Mordohai, and Marc Pollefeys. Variable baseline/resolution stereo. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[31] Diogo C. Garcia, Camilo Drea, and Ricardo L. de Queiroz. Super-resolution for multiview images using depth information. In *2010 IEEE International Conference on Image Processing*, pages 1793–1796, 2010.

[32] Frederic Garcia, Djamila Aouada, Bruno Mirbach, Thomas Solignac, and Bjrn Ottersten. A new multi-lateral filter for real-time depth enhancement. In *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 42–47, 2011.

[33] Frederic Garcia, Bruno Mirbach, Bjorn Ottersten, Frdric Grandidier, and ngel Cuesta. Pixel weighted average strategy for depth sensor data fusion. In *2010 IEEE International Conference on Image Processing*, pages 2805–2808, 2010.

[34] Henri P. Gavin. The levenberg-marquardt algorithm for nonlinear least squares curve-fitting problems. Unpublished Lecture: https://people.duke.edu/ hpgavin/ce281/lm.pdf, 2020.

[35] Daniel Girardeau-Montaut. Cloudcompare v2.11.3 anoia. 2020.

[36] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M. Seitz. Multi-view stereo for community photo collections. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[37] Bastian Goldluecke, Mathieu Aubry, Kalin Kolev, and Daniel Cremers. A Super-resolution Framework for High-Accuracy Multiview Reconstruction. *International Journal of Computer Vision*, 106(2):172–191, January 2014.

[38] Bastian Goldluecke and Daniel Cremers. Superresolution texture maps for multiview reconstruction. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1677–1684, 2009.

[39] Massimo Guarnieri. An historical survey on light technologies. *IEEE Access*, 6:25881–25897, 2018.

[40] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[41] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for single imaage super-resolution. *arXiv preprint arXiv:1904.05677*, 2019.

[42] Hannes Harms, Johannes Beck, Julius Ziegler, and Christoph Stiller. Accuracy analysis of surface normal reconstruction in stereo vision. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pages 730–736, 2014.

[43] Christopher G. Harris and M. J. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, 1988.

[44] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, USA, 2 edition, 2003.

[45] Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, and Yao Zhao. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9225–9234, 2021.

[46] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008.

[47] Markus Hoffmann. Measurements, statistics, and errors. Unpublished Presentation CERN Accelerator School 2008: https://cas.web.cern.ch/sites/default/files/lectures/dourdan-2008/hoffmann.pdf, 2008.

[48] Markus Hoffmann. Measurements, statistics, and errors. *CAS - CERN Accelerator School: Beam Diagnostics*, page 157, 2009. Paper: https://cds.cern.ch/record/1213276/files/p157.pdf.

[49] Dominik Honegger, Torsten Sattler, and Marc Pollefeys. Embedded real-time multi-baseline stereo. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5245–5250, 2017.

[50] Alain Hor and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369, 2010.

[51] Xiaobin Hu, Wenqi Ren, John LaMaster, Xiaochun Cao, Xiaoming Li, Zechao Li, Bjoern Menze, and Wei Liu. Face super-resolution guided by 3d facial priors. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 763–780, Cham, 2020. Springer International Publishing.

[52] Marius Huber, Timo Hinzmann, Roland Siegwart, and Larry H. Matthies. Cubic range error model for stereo vision with illuminators. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 842–848, 2018.

[53] M. Irani and S. Peleg. Super resolution from image sequences. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, volume ii, pages 115–120 vol.2, 1990.

[54] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP: Graphical Models and Image Processing*, 53(3):231–239, 1991.

[55] James R. Janesick. *Photon Transfer Noise Sources Chapter 3*. Spiedigitallibrary, Bellingham, WA USA, 1 edition, 2007.

[56] Daniel S. Jeon, Seung-Hwan Baek, Inchang Choi, and Min H. Kim. Enhancing the spatial resolution of stereo images using a parallax prior. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1721–1730, 2018.

[57] Linfu Jiang, Minzhi Zhong, and Fangchi Qiu. Single-image super-resolution based on a self-attention deep neural network. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 387–391, 2020.

[58] Sing Bing Kang, R. Szeliski, and Jinxiang Chai. Handling occlusions in dense multi-view stereo. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001.

[59] Herman Kelder. Multi-frame super-resolution image enhancement for autonomous landing of unmanned aerial vehicles. Unpublished Thesis, 2021.

[60] Thomas Khler. Multi-frame super-resolution reconstruction with applications to medical imaging. Unpublished Dissertation, 2017.

[61] Joohyeok Kim, Gwanggil Jeon, and Jechang Jeong. Joint-adaptive bilateral depth map upsampling. *Signal Processing: Image Communication*, 29(4):506–513, 2014.

[62] Sebastian Knorr, Matthias Kunter, and Thomas Sikora. Stereoscopic 3d from 2d video with super-resolution capability. *Signal Processing: Image Communication*, 23(9):665–676, 2008.

[63] Johannes Kopf, Michael F. Cohen, Dani Lischinski, and Matt Uyttendaele. Joint bilateral upsampling. *ACM Trans. Graph.*, 26(3):96es, jul 2007.

[64] Jiing-Yih Lai, Tsung-Chien Wu, Watchama Phothong, Douglas Wang, Chao-Yaug Liao, and Ju-Yi Lee. A high-resolution texture mapping technique for 3d textured model. *Applied Sciences*, 8:2228, 11 2018.

[65] Tae Bok Lee and Yong Seok Heo. Single image super resolution using convolutional neural networks for noisy images. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 195–199, 2020.

[66] Jianjun Lei, Lele Li, Huanjing Yue, Feng Wu, Nam Ling, and Chunping Hou. Depth map super-resolution considering view synthesis quality. *IEEE Transactions on Image Processing*, 26(4):1732–1745, 2017.

[67] Stefan Leutenegger, Margarita Chli, and Roland Y. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *2011 International Conference on Computer Vision*, pages 2548–2555, 2011.

[68] Jianwei Li, Wei Gao, and Yihong Wu. High-quality 3d reconstruction with depth super-resolution and completion. *IEEE Access*, 7:19370–19381, 2019.

[69] Shenhong Li, Xiongwu Xiao, Bingxuan Guo, and Lin Zhang. A novel openmvs-based texture reconstruction method based on the fully automatic plane segmentation for 3d mesh models. *Remote Sensing*, 12(23), 2020.

[70] Yawei Li, Vagia Tsiminaki, Radu Timofte, Marc Pollefeys, and Luc Van Gool. 3d appearance super-resolution with deep learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9663–9672, 2019.

[71] Yunting Li, Jun Zhang, Wenwen Hu, and Jinwen Tian. Laboratory calibration of star sensor with installation error using a nonlinear distortion model. *Applied Physics B: Lasers and Optics*, 115:561–570, 2014, 09 2013.

[72] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. *CoRR*, abs/1707.02921, 2017.

[73] Huei-Yung Lin, Chun-Lung Tsai, and Van Luan Tran. Depth measurement based on stereo vision with integrated camera rotation. *IEEE Transactions on Instrumentation and Measurement*, 70:1–10, 2021.

[74] Li-Wei Liu, Yang Li, Liang-Hao Wang, Dong-Xiao Li, and Ming Zhang. Tof depth map super-resolution using compressive sensing. In *2013 Seventh International Conference on Image and Graphics*, pages 135–138, 2013.

[75] Shanshan Liu, Minghui Wang, Qingbin Huang, and Xia Liu. Robust multi-frame super-resolution based on adaptive half-quadratic function and local structure tensor weighted btv. *Sensors*, 21(16), 2021.

[76] Zhaoyan Liu, William Hunt, Mark Vaughan, Chris Hostetler, Matthew McGill, Kathleen Powell, David Winker, and Yongxiang Hu. Estimating random errors due to shot noise in backscatter lidar observations. *Appl. Opt.*, 45(18):4437–4447, Jun 2006.

[77] Eugenio Lomurno, Andrea Romanoni, and Matteo Matteucci. Improving multi-view stereo via super-resolution. *CoRR*, abs/2107.13261, 2021.

[78] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, 1999.

[79] Yang-Yi Luo, Hui-Guo Lu, and Ning Jia. Super-resolution algorithm of satellite cloud image based on wgan-gp. In *2019 International Conference on Meteorology Observations (ICMO)*, pages 1–4, 2019.

[80] Manimala Mahato, Shirishkumar Gedam, Jyoti Joglekar, and Krishna Mohan Buddhiraju. Dense stereo matching based on multiobjective fitness functiona genetic algorithm optimization approach for stereo correspondence. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6):3341–3353, 2019.

[81] Robert Maier, Jrg Stckler, and Daniel Cremers. Super-resolution keyframe fusion for 3d modeling with high-quality textures. In *2015 International Conference on 3D Vision*, pages 536–544, 2015.

[82] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2434–2442, 2020.

[83] Jorge J. Moré. The levenberg-marquardt algorithm: Implementation and theory. In G. A. Watson, editor, *Numerical Analysis*, pages 105–116, Berlin, Heidelberg, 1978. Springer Berlin Heidelberg.

[84] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. OpenMVG: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*, pages 60–74. Springer, 2016.

[85] Fang Nan, Wei Jing, Feng Tian, Jizhong Zhang, Kuo-Ming Chao, Zhenxin Hong, and Qinghua Zheng. Feature super-resolution based facial expression recognition for multi-scale low-resolution images. *Knowledge-Based Systems*, 236:107678, 2022.

[86] Haidawati Nasir, Vladimir Stankovi, and Stephen Marshall. Singular value decomposition based fusion for super-resolution image reconstruction. *Signal Processing: Image Communication*, 27(2):180–191, 2012.

[87] Kamal Nasrollahi and Thomas Baltzer Moeslund. Super-resolution: a comprehensive survey. *Machine Vision and Applications*, 25:1423–1468, 2014.

[88] JOSEPH NEEDHAM, WANG LING, and KENNETH GIRDWOOD ROBINSON. *Science and Civilisation in China*. Cambridge University Press, The Edinburgh Building, Cambridge CB2 2RU, UK, 6 edition, 1962.

[89] Kyle Nelson, Asim Bhatti, and Saeid Nahavandi. Super-resolution of a 3-dimensional scene from novel viewpoints. In *2012 12th International Conference on Control Automation Robotics Vision (ICARCV)*, pages 1380–1385, 2012.

[90] Richard A. Newcombe, Steven J. Lovegrove, and Andrew J. Davison. Dtam: Dense tracking and mapping in real-time. In *2011 International Conference on Computer Vision*, pages 2320–2327, 2011.

[91] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015.

[92] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2161–2168, 2006.

[93] Tomoaki Nonome., Fumihiko Sakaue., and Jun Sato. Super-resolution 3d reconstruction from multiple cameras. In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory*

and Applications - Volume 5: VISAPP,, pages 481–486. INSTICC, SciTePress, 2018.

[94] M. Okutomi and T. Kanade. A multiple-baseline stereo. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 63–69, 1991.

[95] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4):353–363, 1993.

[96] A. Papoulis. A new algorithm in spectral analysis and band-limited extrapolation. *IEEE Transactions on Circuits and Systems*, 22(9):735–742, 1975.

[97] A. Papoulis. Generalized sampling expansion. *IEEE Transactions on Circuits and Systems*, 24(11):652–654, 1977.

[98] Haesol Park, Kyoung Mu Lee, and Sang Uk Lee. Combining multi-view stereo and super resolution in a unified framework. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4, 2012.

[99] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, 20(3):21–36, 2003.

[100] Lyndsey Pickup. Machine learning in multi-frame image super-resolution. Unpublished Dissertation, 2007.

[101] Lyndsey C. Pickup, David P. Capel, Stephen J. Roberts, and Andrew Zisserman. Overcoming registration uncertainty in image super-resolution: Maximize or marginalize? *EURASIP Journal on Advances in Signal Processing*, 023565:1687–6180, 2007.

[102] Lyndsey C. Pickup, David P. Capel, Stephen J. Roberts, and Andrew Zisserman. Bayesian methods for image super-resolution. *Comput. J.*, 52:101–113, 2009.

[103] E.Z. Psarakis and G.D. Evangelidis. An enhanced correlation-based method for stereo correspondence with subpixel accuracy. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 907–912 Vol. 1, 2005.

[104] Xie Qinlan, Chen Hong, and Cao Huimin. Improved example-based single-image super-resolution. In *2010 3rd International Congress on Image and Signal Processing*, volume 3, pages 1204–1207, 2010.

[105] Pedro Raimundo and Karl Apaza-Agero. Improved point clouds from a heritage artifact depth low-cost acquisition. *Revista Brasileira de Computao Aplicada*, 12:84–94, 02 2020.

[106] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In Aleš Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision – ECCV 2006*, pages 430–443, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[107] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, 2011.

[108] James M. Russell. *Plato's Alarm Clock*. Michael O'Mara Books Limited, 9 Lion Yard, Tremadoc Road London, SW4 7NQ, 1 edition, 2018.

[109] T. Saito, T. Komatsu, and K. Aizawa. A signal-processing based method for acquiring very high resolution images with multiple cameras and its theoretical analysis. In *1992 International Conference on Image Processing and its Applications*, pages 365–368, 1992.

[110] U. Sara, M. Akter, and M. Uddin. Image quality assessment through fsim, ssim, mse and psnra comparative study. *Journal of Computer and Communications*, 7(3):8–18, 2019.

[111] Hilario Seibel, Siome Goldenstein, and Anderson Rocha. Fast and effective geometric k-nearest neighbors multi-frame super-resolution. In *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 103–110, 2015.

[112] S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 519–528, 2006.

[113] Jong Wan Silva, Leonardo Gomes, Karl Apaza Agero, Olga R. P. Bellon, and Luciano Silva. Real-time acquisition and super-resolution techniques on 3d

reconstruction. In *2013 IEEE International Conference on Image Processing*, pages 2135–2139, 2013.

[114] K. Simonyan, S. Grishin, D. Vatolin, and D. Popov. Fast video super-resolution via classification. In *2008 15th IEEE International Conference on Image Processing*, pages 349–352, 2008.

[115] Sudipta N. Sinha. *Multiview Stereo*, pages 516–522. Springer US, Boston, MA, 2014.

[116] Noah Snavely. Scene reconstruction and visualization from internet photo collections: A survey. *IPSJ Trans. Comput. Vis. Appl.*, 3:44–66, 2011.

[117] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Skeletal graphs for efficient structure from motion. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[118] Rewa Sood and Mirabela Rusu. Anisotropic super resolution in prostate mri using super resolution generative adversarial networks. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1688–1691, 2019.

[119] Srdjan Stankovic. Compressive sensing: Theory, algorithms and applications. In *2015 4th Mediterranean Conference on Embedded Computing (MECO)*, pages 4–6, 2015.

[120] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution im-

agery. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[121] Richard Szeliski. 2006.

[122] Petri Tanskanen, Kalin Kolev, Lorenz Meier, Federico Camposeco, Olivier Saurer, and Marc Pollefeys. Live metric 3d reconstruction on mobile phones. In *2013 IEEE International Conference on Computer Vision*, pages 65–72, 2013.

[123] Helmut H. Telle and ngel Gonzlez Urea. *Laser Spectroscopy and Laser Imaging: An Introduction*. CRC Press Taylor & Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742, 1 edition, 2018.

[124] Damber Thapa, Kaamran Raahemifar, William R. Bobier, and Vasudevan Lakshminarayanan. Comparison of super-resolution algorithms applied to retinal images. *Journal of Biomedical Optics*, 19(5):1 – 16, 2014.

[125] Jing Tian and Kai-Kuang Ma. A survey on super-resolution imaging. *Springer: Signal, Image and Video Processing*, 5:329342, 2011.

[126] Carlo Tomasi. Technical perspective: Visual reconstruction. *Communications of the ACM*, 54(10):104, 2011.

[127] Yann Traonmilin, Said Ladjal, and Andrs Almansa. On the amount of regularization for super-resolution interpolation. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 380–384, 2012.

[128] Xingyu Tuo, Yu Xia, Yin Zhang, Junyu Zhu, Yongchao Zhang, Yulin Huang, and Jianyu Yang. Super-resolution imaging for real aperture radar by two-dimensional deconvolution. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 6630–6633, 2021.

[129] Patrick Vandewalle, Sabine Ssstrunk, and Martin Vetterli. A frequency domain approach to registration of aliased images with application to super-resolution. *Eurasip Journal on Applied Signal Processing*, 2006, 12 2006.

[130] Patrick Vandewalle, Sabine E. Susstrunk, and Martin Vetterli. Superresolution images reconstructed from aliased images. In Touradj Ebrahimi and Thomas Sikora, editors, *Visual Communications and Image Processing 2003*, volume 5150, pages 1398 – 1405. International Society for Optics and Photonics, SPIE, 2003.

[131] Salvador Villena Morales. Superresolucin y reconstruccin bayesiana de imgenes a partir de imgenes de baja resolucin rotadas y desplazadas. combinacin de modelos. Unpublished Dissertation: https://digibug.ugr.es/handle/10481/20549, 2011.

[132] George Vogiatzis, Carlos Hernandez Esteban, Philip H.S. Torr, and Roberto Cipolla. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2241–2246, 2007.

[133] Oleg Voynov, Alexey Artemov, Vage Egiazarian, Alexandr Notchenko, Gleb Bobrovskikh, Evgeny Burnaev, and Denis Zorin. Perceptual deep depth super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5652–5662, 2019.

[134] Xue Wan, Jianguo Liu, Hongshi Yan, Gareth L.K. Morgan, and Tao Sun. 3d super resolution scene depth reconstruction based on skysat video image sequences. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 6653–6656, 2016.

[135] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12242–12251, 2019.

[136] Senhua Wang, Xiangzhong Li, Ping Wang, Yu Li, and Qin Chen. The study of 3d super-resolution geometric modeling based on wavelet analysis. In *2016 13th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 24–27, 2016.

[137] Yingqian Wang, Xinyi Ying, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Symmetric parallax attention for stereo image super-resolution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 766–775, 2021.

[138] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[139] J. Weng, P. Cohen, and M. Herniou. Camera calibration with distortion models and accuracy evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(10):965–980, 1992.

[140] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics*, 38(4):118, Aug 2019.

[141] Lin Xu, Eric Li, Jianguo Li, Yurong Chen, and Yimin Zhang. A general texture mapping framework for image-based 3d modeling. In *2010 IEEE International Conference on Image Processing*, pages 2713–2716, 2010.

[142] Koki Yamashita and Konstantin Markov. Medical image enhancement using super resolution methods. In Valeria V. Krzhizhanovskaya, Gábor Závodszky, Michael H. Lees, Jack J. Dongarra, Peter M. A. Sloot, Sérgio Brissos, and João Teixeira, editors, *Computational Science – ICCS 2020*, pages 496–508, Cham, 2020. Springer International Publishing.

[143] Yuxiang Yang, Junjie Cai, Zhengjun Zha, Mingyu Gao, and Qi Tian. A stereo-vision-assisted model for depth map super-resolution. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2014.

[144] Tianyi Zhang, Yun Gu, Xiaolin Huang, Enmei Tu, and Jie Yang. Stereo endoscopic image super-resolution using disparity-constrained parallel attention. *ArXiv*, abs/2003.08539, 2020.

[145] Can Zhao, Aaron Carass, Blake E. Dewey, and Jerry L. Prince. Self super-resolution for magnetic resonance images using deep networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 365–368, 2018.

[146] Yang Zhou, Yangyang Xu, Yong Du, Qiang Wen, and Shengfeng He. Pro-pulse: Learning progressive encoders of latent semantics in gans for photo upsampling. *IEEE Transactions on Image Processing*, 31:1230–1242, 2022.

[147] Xiangyuan Zhu, Kehua Guo, Hui Fang, Liang Chen, Sheng Ren, and Bin Hu. Cross view capture for stereo image super-resolution. *IEEE Transactions on Multimedia*, pages 1–1, 2021.

[148] A. Zomet, A. Rav-Acha, and S. Peleg. Robust super-resolution. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001.

[149] Yifan Zuo, Qiang Wu, Yuming Fang, Ping An, Liqin Huang, and Zhifeng Chen. Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2):297–306, 2020.

Vita Auctoris

Name:	Michael Lee
Place of Birth:	Windsor, Ontario
Education:	University of Windsor
	Windsor, Ontario
	Ph.D. Electrical and Computer Engineering
	University of Windsor
	Windsor, Ontario
	M.Sc. Mathematics and Statistics
	University of Windsor
	Windsor, Ontario
	B. Mathematics and Computer Science