Diffusion MVSNet: A Learning-based MVS Boosted by Diffusion-Based Image Enhancement Model

Zhang Chi

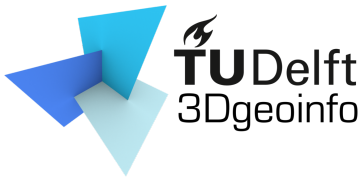
Master thesis submitted under the supervision of
Nail Ibrahimli

The co-supervision of
Liangliang Nan

Academic year
2022-2024

In order to be awarded the Master's Degree in Geomatics

The work in this thesis was carried out in the:



3D geoinformation group
Delft University of Technology

Supervisors: Nail Ibrahimli
Porf. Liangliang Nan
Co-reader: Casper van Engelenburg

Abstract

Keywords: 3D Reconstruction, Image Enhancement

Multiview Stereo (MVS) reconstruction techniques have made significant advancements with the development of deep learning. However, their performance often deteriorates in low-light conditions, where feature extraction and matching become challenging. Traditional image enhancement solutions are insufficient for MVS tasks in low illumination, relying on manual adjustments. We introduce an end-to-end MVS framework incorporating a diffusion-based image enhancement algorithm with MVS to build an end-to-end framework for improving the performance of MVS in low-light conditions. This integration improves color rendering and visualization of 3D reconstructions and slightly enhances geometric shapes. Our method uses a feature adapter to integrate the enhanced images from the Low-light Diffusion model into CasMVSNet, refining the feature maps in poorly lit environments. Validation on the DTU and Tanks and Temples datasets demonstrates our model's robustness and generalizability across various lighting conditions and MVS pipelines, including GeoMVSNet and MVSNet. Our approach simplifies the training process by requiring only the training of an adapter rather than a multi-view image enhancement model, underscoring the effectiveness of incorporating image enhancement into learning-based MVS frameworks for low-light conditions.

Zhang Chi
Academic Year 2024

Acknowledgments

I would like to extend my deepest gratitude to my supervisors, Nail Ibrahimli and Prof. Nan Liangliang. Your unwavering support, insightful guidance, and compassionate understanding during times of frustration have been invaluable. Your encouragement and wisdom have not only shaped this work but also helped me grow as a researcher and an individual.

A heartfelt thank you to my parents and friends. Your unconditional support and comforting presence have been my anchor. Without your belief in me and your steadfast support, the journey would have been much more daunting. You are the colors in my world, making every day brighter and more meaningful.

A special nod to the perpetually gloomy weather of the Netherlands. Your consistent drizzle and cloudy days have provided ample opportunity for me to hone my culinary skills. Every rainy day was a chance to practice cooking, turning what could have been a dreary day into a delightful culinary adventure.

Thank you all for being the pillars of my strength and the light in my life.

...

MVS Multiview Stereo

FPN Feature Pyramid Network

CNN Convolutional neural network

VAE Variational Auto Encoder

DWT Discrete Wavelet Transform

IDWT Inverse Discrete Wavelet Transform

HFRM High Frequency Reinforcement Module

RAM Random-Access Memory

GPU Graphics Processing Unit

SSR Single Scale Retinex

MSR Multi-Scale Retinex

GAN Generative Adversarial Network

Structure From Motion SfM

Contents

Chapter 1

Introduction

1.1 Background and Motivation

In the dynamic domains of computer science and information technology, the imperative for advanced 3D modeling techniques has intensified. Central to this development is 3D reconstruction technology, an integral aspect of contemporary computational vision. 3D reconstruction technology plays a critical role in domains such as the preservation of historical heritage, augmented and extended reality (AR/XR), video game, and animation development [45, 51]. Among the myriad techniques available for 3D reconstruction, Multiview Stereo (Multiview Stereo (MVS)) is an advantageous approach with high cost-efficiency and rapider processing capabilities [12].

There are two distinct branches of MVS algorithms: traditional MVS and learning-based MVS. Both traditional and learning-based MVS algorithms operate by matching correspondences in overlapped multi-view images to recover 3D representations from correspondences [66, 56, 5, 4]. The primary difference between traditional and learning-based MVS lies in the space where correspondences are matched. Traditional MVS matches correspondences directly in images' RGB space, whereas learning-based MVS techniques match correspondences in the feature space derived from neural networks [66, 73, 15, 33, 67, 72, 62, 21, 40, 69].

While traditional MVS can produce high-quality 3D reconstructions in ideal conditions, it struggles with textureless and non-Lambertian surfaces [46, 41, 6]. In contrast, learning-based MVS exhibits greater robustness and effectively addresses many challenges faced by traditional MVS [55]. The advancements in learning-based MVS have led to new benchmarks for MVS in completeness, resource efficiency, and robustness.

Despite significant advancements, current learning-based methods face challenges in low- illumination environments [46, 41, 6]. While effective under optimal lighting conditions, completeness and accuracy of learning-based MVS decreases a lot in poorly lit settings such as subterranean spaces or dimly-lit indoor areas like tunnels and mines [46, 41, 6].

Current solutions for low illumination problems in MVS involve using traditional image enhancement algorithms to manually adjust image contrastness and brightness or training an additional end-to-end model for low illumination, which requires an additional low illumination dataset.

Traditional image enhancement algorithms, such as histogram equalization [53], Retinex theory [30], gamma correction [3], and wavelet transformation [27], are not end-to-end and require manual parameter adjustments for each scene, which is highly empirical. Incorrect settings can reduce enhancement effects or negatively impact 3D reconstruction [3, 29]. Recent advancements in artificial intelligence have led to end-to-end image enhancement algorithms such as EnlightenGAN [22], RetinexNet [54], and Diff-Retinex [70], but these are designed for single-frame enhancement and have not been applied to MVS.

Another approach to address low illumination in MVS is to train a learning-based model specifically for these conditions. To our knowledge, three end-to-end studies have attempted this [46, 55].

DEMVSNet [18] has shown limited improvement on synthesized datasets. LOLIMVSNet by Wang et al. [55] requires paired low illumination multi-view datasets, which are challenging to obtain. Su et al. proposed zero-inference training methods [16, 46], but these often result in overexposure in normally lit images.

A significant research gap exists: no end-to-end framework directly integrates existing single-frame image enhancement models with MVS to improve 3D reconstruction under low illumination. Traditional techniques require manual adjustments, which are labor-intensive. Current end-to-end solutions necessitate training new models for low-illumination conditions, which is computationally demanding and requires comprehensive low-illumination multi-view datasets [46].

This thesis aims to bridge this gap by developing a method that seamlessly integrates existing image enhancement models with learning-based MVSNet. We aim to enhance 3D reconstruction performance under low illumination without imposing significant computational demands or requiring low-illumination multi-view training datasets.

1.2 Research Questions and Objectives

Given the challenges and research gaps identified, the central research question of this thesis is:
To what extent can existing single-frame image enhancement models be utilized to enhance the performance of MVS in low illumination conditions?

The primary question can be subdivided into three specific research objectives:

1. Which image enhancement model is suitable for MVS tasks under low illumination?
2. What architecture is suitable for integrating image enhancement models into MVS to boost 3D reconstruction?
3. How can image enhancement models, which typically contain millions of parameters, be efficiently utilized for 3D reconstruction tasks to minimize computational demands while maintaining good performance?

Research Objectives: This research includes the following objectives:

1. *Select the appropriate image enhancement model for MVS.*
   We will compare image enhancement results and select a model that can reveal features obscured by low illumination while retaining details to ensure multi-view consistency.
2. *Build a framework that effectively integrates the low light image enhancement model with the learning-based MVS model.*
   The framework should not only produce visually discernible images but also enhance the performance of the MVS model in both low light and normal light scenarios.
3. *Minimize computational resource requirements.*
   This step is crucial because many low-light illumination models use architectures that require significant computational resources. Since many 3D reconstruction models face Graphics Processing Unit (GPU) Random-Access Memory (RAM) shortages, we should reduce computational demands as much as possible while preserving performance.
4. *Evaluate integration methods and the impact of low illumination models on 3D reconstruction.*
   We aim to delineate how the different components of our designs contribute to improvement and explore the mechanisms behind these improvements.

1.3 Research Scope

We focus on developing a novel architecture to combine image enhancement models with learning-based MVS for a better 3D reconstruction process. Our goal extends beyond enhancing performance in low-light conditions; we also aim to preserve or improve performance under normal lighting conditions. The selection criteria for these methods will include accuracy, completeness, and computational resource demands of 3D reconstruction.

1.4 Thesis Outlines

This thesis is organized into five critical chapters, each addressing distinct aspects of the research domain. Below is a detailed outline of each chapter:

- **Chapter 2: Literature Review** This chapter extensively reviews the existing literature related to the field. It covers foundational theories, recent image enhancement techniques, and MVS reconstruction advancements. These studies establish the foundation of our work.
- **Chapter 3: Methodology** Here, the research methodology is thoroughly detailed. It outlines the proposed enhancements to existing MVS techniques and describes the novel integration of image enhancement algorithms into the MVS process. It also introduces the training strategy and how the evaluation is conducted.
- **Chapter 4: Experiment Results and Discussion** This chapter comprehensively explains the implementation details, including the tools and parameters utilized. It discusses the experimental procedures, training methodologies, and the quantitative results of 3D reconstruction. Qualitative results from experiments conducted on the DTU and 'Tanks and Temples' datasets are displayed, and ablation studies are presented to explore the efficacy of the proposed methodology.
- **Chapter 5: Conclusion and limitations** The final chapter synthesizes the study's findings and discusses the limitations and future work.
- **Appendices** - Supplementary materials, including additional data on image enhancement effects and detailed 3D reconstruction results, are provided in the appendices.

Chapter 2

Related Work

2.1 Low Illumination Enhancement

This section discusses traditional and learning-based approaches for low illumination image enhancement.

2.1.1 Traditional Image Enhancement Methods

Although the traditional image enhancement methods are still dominant in the relevant research to optimize the 3D reconstruction performance in a low-light environment [46]. The idea of traditional image enhancement algorithms is either modifying the distribution of pixel values or utilizing physical models to adjust gray levels [47, 45, 16].

Gray-level transformation and histogram equalization methods directly utilize mathematical functions to transform the gray values of images. Gray-level transformation adjusts the contrastness and brightness of dark regions by mapping the original gray value range to a broader range. Gray-Level Transformation techniques include linear transformations [37] and non-linear transformations such as logarithmic and gamma transformations [17, 50]. Unlike gray-level transformations, histogram equalization modifies the pixel distribution. Global histogram equalization redistributes pixel values across the entire image based on the cumulative distribution probability of original pixel values. However, global histogram equalization can potentially obscure necessary details in well-lit regions [36, 25]. Adaptive Histogram Equalization (AHE) and its variant Contrast Limited Adaptive Histogram Equalization (CLAHE) apply localized adjustments to prevent the loss of detail and reduce noise amplification in darker areas [74, 39].

Retinex Theory enhances images based on a physical assumption that color and brightness are determined by the interaction of light and objects. Based on this assumption, the color of the object's surface is composed of red, green, and blue primary colors, and the lightness is determined by surface reflectance and illumination strength. Applying this concept, Single Scale Retinex (SSR) and Multi-Scale Retinex (MSR) use Gaussian blurring at different scales to adjust illumination, enhancing image detail and color fidelity. SSR addresses single scale adjustments, potentially overlooking subtle details, whereas MSR combines effects from multiple scales to better balance visibility and color accuracy [23, 17].

Traditional image enhancement algorithms must adjust parameters manually and empirically and lack generalizability. This limitation underscores the need for innovative learning-based methods to enhance images across diverse conditions without extensive manual intervention.

2.1.2 Learning-based Low Illumination Enhancement

Traditional image enhancement methods often bring problems after adjustment, such as amplifying noise, losing details, and color distortion [46]. Many scholars have widely applied deep learning in image enhancement in recent years. The learning-based image enhancement methods allow images to be converted to normal light conditions automatically. The learning-based methods could be divided into three categories:

1. Encoder-decoder based methods
2. Decomposition-based methods
3. Generative-based methods.

2.1.2.1 Encoder-decoder based methods

Encoder-decoder methods consist of an encoder that compresses the input image into latent space and a decoder that reconstructs the enhanced image from latent space. Encoder-decoder structure allows for end-to-end training, effectively mapping low-light images to enhanced versions [17, 47].

Many scholars have applied encoder-decoder structure for image enhancement [32, 48, 17, 38]. The first notable application of this method was LLNet [32], which uses a variational autoencoder (VAE) [9] to enhance low illumination images. Chen et al. developed an end-to-end pipeline using UNet [43, 7]. Tao et al. introduced a Convolutional neural network (CNN)-based method incorporating an additional brightness channel to estimate transmission more accurately [48]. Ren et al. combined an encoder-decoder structure with Recurrent Neural Networks (RNNs) to enhance edge details and global information integration [38]. Zhang utilized attention layers [52] to focus the enhancement process on salient features of the image [71].

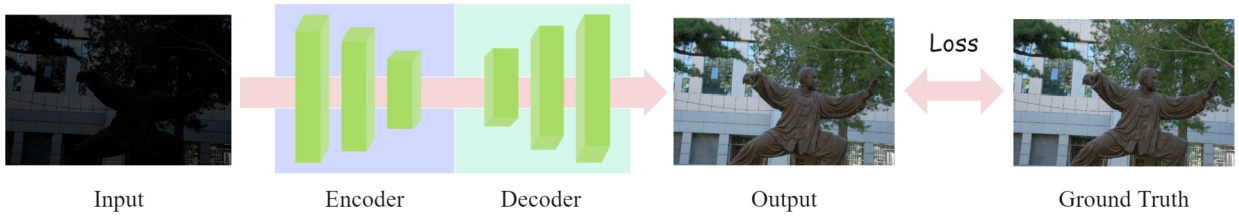


Figure 2.1: The pipeline of encoder-decoder based image enhancement [32]

2.1.2.2 Decomposition-based Methods

Motivated by the success of Retinex Theory [30]. Many studies combine image decomposition techniques with deep learning for image enhancement [17]. The basic idea of decomposition-based methods consists of three steps. Firstly, using a physical model decomposes an image into distinct components. Secondly, the different components are enhanced by neural networks. Lastly, all the components are recombined to form the final version of improved images.

A commonly used decomposition method is the Retinex Theory. Based on Retinex Theory, images can be decomposed into reflectance maps and illumination maps. RetinexNet decomposes images and uses dual parallel networks to predict the reflectance and illumination maps separately. These maps are then recombined to reconstruct the image [59]. URetinex-Net advances this concept by incorporating an unfolding optimization module to enhance noise suppression and detail preservation further [60].

Discrete Wavelet Transformation (DWT) is another effective decomposition technique for image enhancement (see appendix A.1) [10, 44]. DWT transforms images from the temporal domain to frequency domains. DWT is another commonly used technique to enhance low illumination images

combined with deep learning. Rahman et al. [35] train separate modules on each frequency domain, which are then synthesized into enhanced images through Inverse Discrete Wavelet Transform (IDWT). Xu et al. [61] introduces attention layers to process different frequency domains of Discrete Wavelet Transform (DWT), significantly improving detail retention in low illumination images [61].

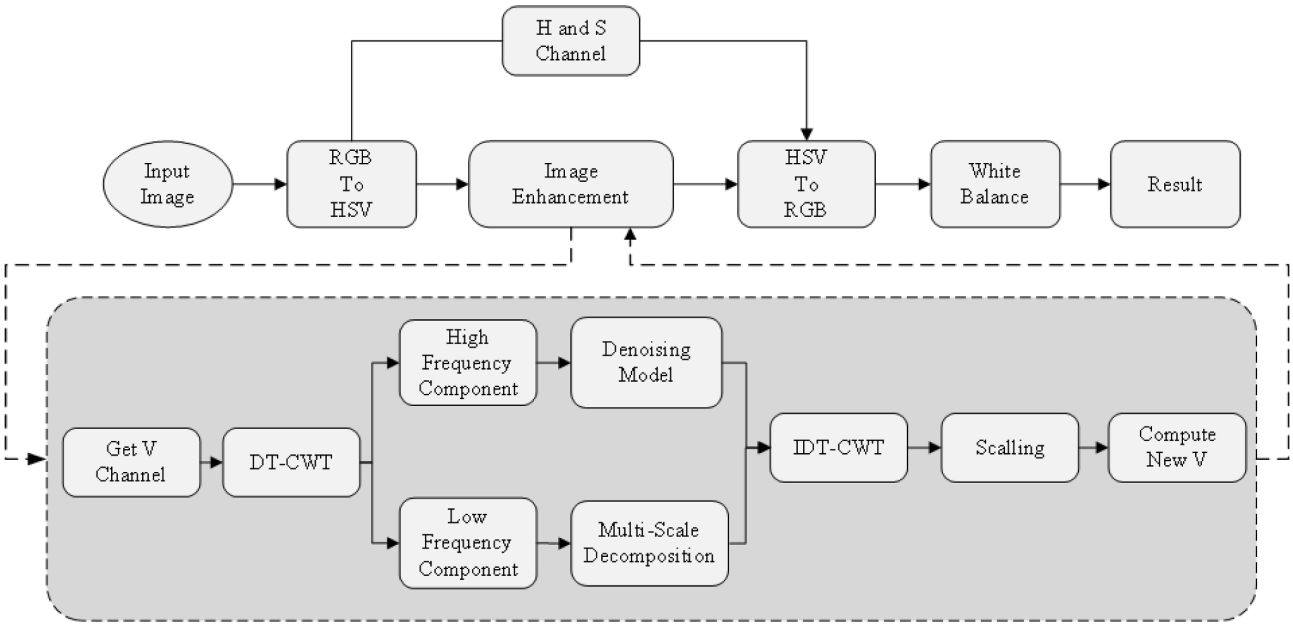


Figure 2.2: The structure of DWT multi-scale decomposition image enhancement model proposed by Rahman et al. [35]

2.1.2.3 Generative-based Methods

In recent years, generative models have also been applied to image enhancement tasks [17]. Generative models have demonstrated the strength of enhancing detail capture and adaptability to varied environments [17].

One of the architectures used in the generative-based image enhancement models is Generative Adversarial Networks (GAN). A key advantage of Generative Adversarial Network (GAN) is their ability to generate detailed enhancements without needing paired training data. EnlightGAN [22] is trained on unpaired datasets [14] for image enhancement. Compared with other models, applying EnlightenGAN shows a better ability to restore details and textures of images [22].

Another architecture used for image enhancement is diffusion models. The development of generative models has demonstrated that diffusion models offer superior performance compared to GAN [8]. The fundamental mechanism of diffusion models is to train a model predicting the noise added to images in an autoregressive process. The diffusion model can generate images from noise if the process is long enough. More explanation can be seen in appendix A.2

There are many studies combining the diffusion model with low illumination enhancement. ShadowDiffusion focuses on removing shadows while dynamically predicting masks during the denoising process. DR2 [58] leverages a pretrained diffusion model to restore obscured facial features by utilizing prior knowledge encoded in the diffusion model.

Diffusion-based methods can also be combined with image decomposition to achieve better image restoration effects. One of the decomposed diffusion methods is Retinex-Diffusion [70]. This approach integrates Retinex theory with diffusion processes, operating diffusion modules separately on reflectance and illumination maps to enhance images.

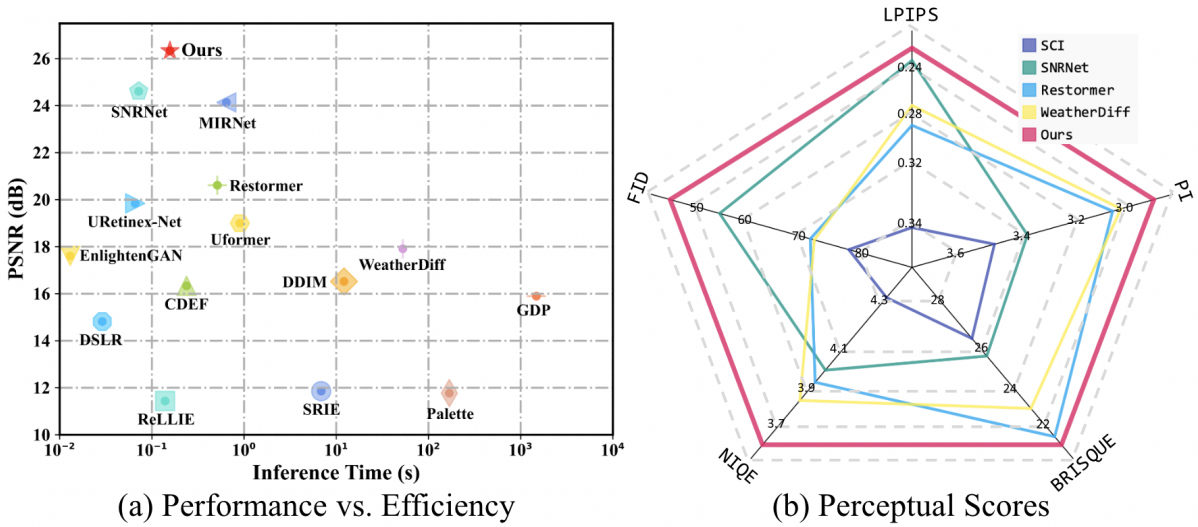
2.1.2.4 Low-light Diffusion

Currently, the state-of-the-art model for low illumination enhancement is low-light diffusion. The low-light diffusion combines DWT with the diffusion model. Firstly, the image is decomposed into high-frequency and low-frequency components via DWT. Secondly, low-light Diffusion employs two parallel modules: the conditional wavelet-based diffusion module and high-frequency reinforcement modules (HFRM) to independently enhance the low and high-frequency components. Lastly, the final output of low-light diffusion is the combination of all components via IDWT.

One of the modules in low-light diffusion is the wavelet-based diffusion module. This module is a conditional diffusion model that inputs DWT's low-frequency components and outputs enhanced low-frequency components.

The other module of low-light diffusion is the high-frequency reinforcement module (HFRM). HFRM aims to enhance the high-frequency components of DWT decomposition [20]. The purpose of HFRM is to restore the high-frequency parts of images, like details and textures. Unlike the wavelet-based diffusion module, HFRM directly enhances high-frequency components of DWT decomposition via attention blocks.

Low-light Diffusion was trained on the LOLv2 dataset [65]. This dataset includes 689 real-world and 1000 synthesized paired low-illumination images. Low-light Diffusion achieves state-of-the-art (SOTA) performance compared with other models like deep Retinex [59], EnlightenGAN [22] and LLNet [32]. The results can be seen in 2.3



(a) Performance vs. Efficiency (b) Perceptual Scores

Figure 2.3: Comparison of Low-light Diffusion with other methods [20]

Figure 2.4: Results of Low-light Diffusion on low illumination images [20]

2.2 Multi-view Stereo

The initial stage of image-based 3D reconstruction typically involves using the Structure From Motion (SfM) algorithm to extract feature points and calibrate camera matrices. SfM can only recover sparse point clouds from multi-view images. MVS aims to recover the dense point clouds from multi-view images given the calibrated camera matrices [42]. This section introduces traditional and learning-based MVS methods.

2.2.1 Traditional MVS

The basic principle of MVS is to find correspondences between multi-view images and extract 3D representations of objects from correspondences. The choice of 3D representations to extract from correspondences includes voxels, feature points, and depth maps. The depth map is the most predominant in MVS compared with other representations. Once correct depth maps for overlapped multi-view images are obtained, dense point clouds could be recovered.

The traditional methods in Multi-View Stereo (MVS) rely on the photometric consistency assumption, which posits that the same 3D point in multiple images will have similar features in RGB space. A classic approach is the plane sweeping algorithm introduced by Yang [64]. This technique involves assuming the existence of multiple parallel planes at different depths in front of the camera. Pixels are projected onto these planes across various views to determine the correct plane alignment. If the correct plane is obtained, the correspondent pixels on neighboring views should share similar RGB features.

Another classic algorithm is PatchMatch [4]. PatchMatch finds correspondences by dividing the image into small patches. It employs a randomized search to find correspondences between patches across stereo images quickly. This method iteratively refines the disparity map for each pixel, enhancing the accuracy of the depth estimation. PatchMatch leverages efficient propagation of good matches to neighboring pixels, significantly speeding up the computation compared to traditional exhaustive search methods [5].

While traditional MVS performs well under ideal conditions with consistent lighting and textures, it struggles with textureless surfaces, occlusions, and non-Lambertian materials. Conversely, learning-based MVS methods have shown greater robustness, inference speed, and completeness.

2.2.2 Learning-based MVS

Unlike traditional MVS searches for correspondences in the RGB space, learning-based MVS extracts correspondences in the feature space derived from neural networks. Learning-based MVS demonstrates superior performance on challenging objects, such as objects with non-Lambertian and textureless surfaces.

The most popular learning-based MVS method is the cost volume regularization network, firstly proposed in MVSNet by Yao et al. [66]. The MVSNet architecture comprises several key steps:

1. **Feature Extraction:** A feature extractor uses 2D CNN to derive feature space from the input images. The extracted feature space contains more informative cues for depth estimation than traditional RGB data.

2. **Cost Volume Construction:** Differentiable homography warping projects 2D feature space into a 3D feature volume across views. The Z axis is divided by the depth range of images. Then, MVSNet integrates multiple 3D feature volumes to create a cost volume.

3. **Depth Estimation:** Applies 3D convolution over the cost volume to predict the depth value for each pixel.

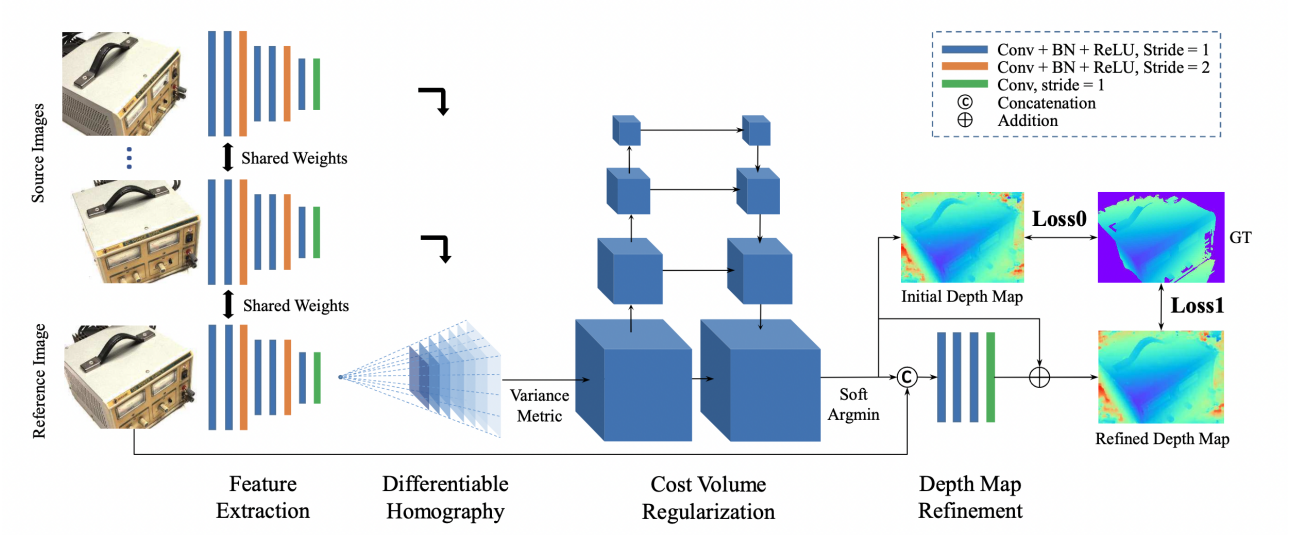


Figure 2.5: Architecture of MVSNet [66]

Among many followers [72, 67, 73, 21], CasMVSNet is one of the most influential improvements. CasMVSNet employs a coarse-to-fine strategy to enhance depth map estimation. This method systematically refines depth maps across multiple scales using a sequence of cascaded networks, each optimized for a specific resolution tier [15]. This hierarchical approach significantly improves depth accuracy by progressively refining the depth maps at increasingly finer resolutions. Additionally, using multiple scales allows the network to capture global context and fine details, leading to more precise and robust depth estimations [15].

GeoMVSNet is another significant advancement in the realm of multi-view stereo (MVS) networks. It also follows the coarse-to-fine strategy but strengthens the geometry before the feature

map and utilizes the confidence map during the cascading process to build probability volume embedding during the depth regression process [73]. The GeoMVSNet achieves SOTA and greatly improves MVS's performance.

Despite the great advancements achieved by learning-based MVS, the challenge of low illumination conditions remains largely unaddressed. Low illumination negatively affects traditional and learning-based MVS approaches [12].

2.2.3 MVS under low illumination

This section introduces both traditional methods and recent learning-based approaches developed to enhance MVS performance on low illumination images.

2.2.3.1 MVS Using Traditional Image Enhancement Methods

Traditional methods have long been employed in image preprocessing to enhance low-illumination images for 3D reconstruction and subsequently improve MVS performance. Substantial works utilize traditional image enhancement algorithms for MVS.

Yeh [68] employs an enhanced HDR system to boost 3D reconstruction, which requires RGB-D equipment and is not always accessible. Kanellakis [24] uses adaptive histogram techniques to modify the pixel value distribution, enhancing 3D reconstruction outcomes. Alasal [2] applies color balancing algorithms to improve image contrast for more effective 3D reconstruction. Ballabeni [3] proposed an image enhancement pipeline specifically tailored for MVS, which includes color balance, image denoising, gray level conversion, and image content enhancement.

Although traditional methods can largely reduce the effect of low illumination in MVS, traditional methods are not end-to-end solutions for low illumination images. The selection and application of these traditional methods are largely empirical. Incorrect choices and parameter settings may lead to marginal improvements or even degraded results [29], which is the biggest drawback of traditional methods.

2.2.3.2 Learning-Based Methods for 3D Reconstruction

To our knowledge, only three learning-based methods providing solutions to improve MVS in low illumination images are emerging. Different from traditional methods, they are end-to-end solutions for low illumination challenges.

Su et al. [46] developed an unsupervised enhancement model (ZDE3D) with seven convolution layers tailored for 3D reconstruction tasks in challenging environments such as tunnels and underground pipelines. Using the zero inference strategy [16], Su et al.'s model can be trained directly on unpaired data [46].

However, Su et al.'s model has some notable drawbacks. Firstly, their model tends to make images overexposed under normal lighting conditions. Secondly, ZDE3D may alter natural color tones. Additionally, ZDE3D has difficulty restoring high-quality textures.

DEMVSNet [18] introduces a joint training method for noisy images. DEMVSNet adds a parallel branch on existing learning-based MVS models to predict RGB images with depth values together. The loss function is the joint loss of RGB loss and depth loss [18]. Due to the absence of paired multi-view data, DEMVSNet utilizes synthesized multi-view datasets with the methods introduced in [57]. It should be noticed that DEMVSNet's improvement is not significant even on their synthesized dataset compared with CasMVSNet and MVSNet [18].

Different from DEMVSNet and ZDE3D, LOLIMVS proposed by Wang [55] assembled a paired multi-view low illumination dataset tailored for 3D reconstruction named LOLI100. This dataset underpins a state-of-the-art approach where an image enhancement model and an MVS system are trained consecutively on LOLI100. During depth estimation, their image enhancement model's

encoder is frozen, and its outputs are cascadedly added into the MVS decoder. This method significantly advances performance under low-light conditions, setting new benchmarks. However, the requirement for specialized data and the extensive training time needed for such dual models underscore the challenges of deploying this technology in broader applications.

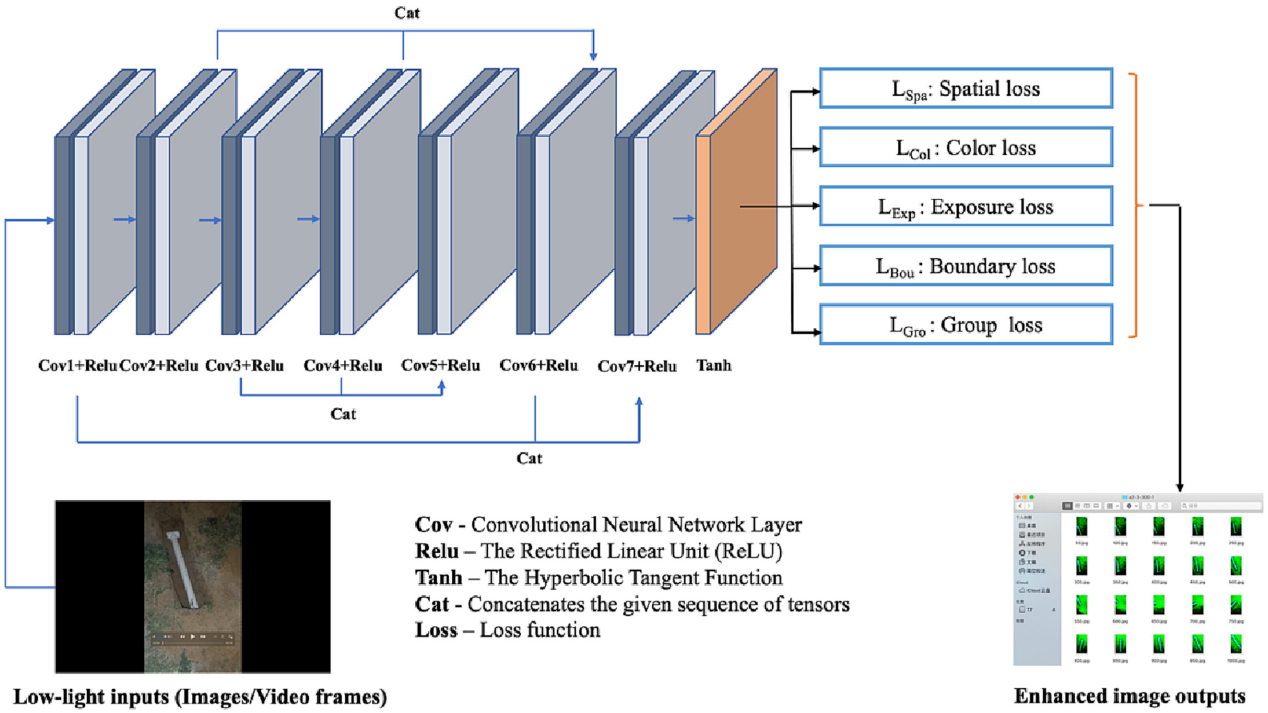


Figure 2.6: Architecture of ZDE3D [46]

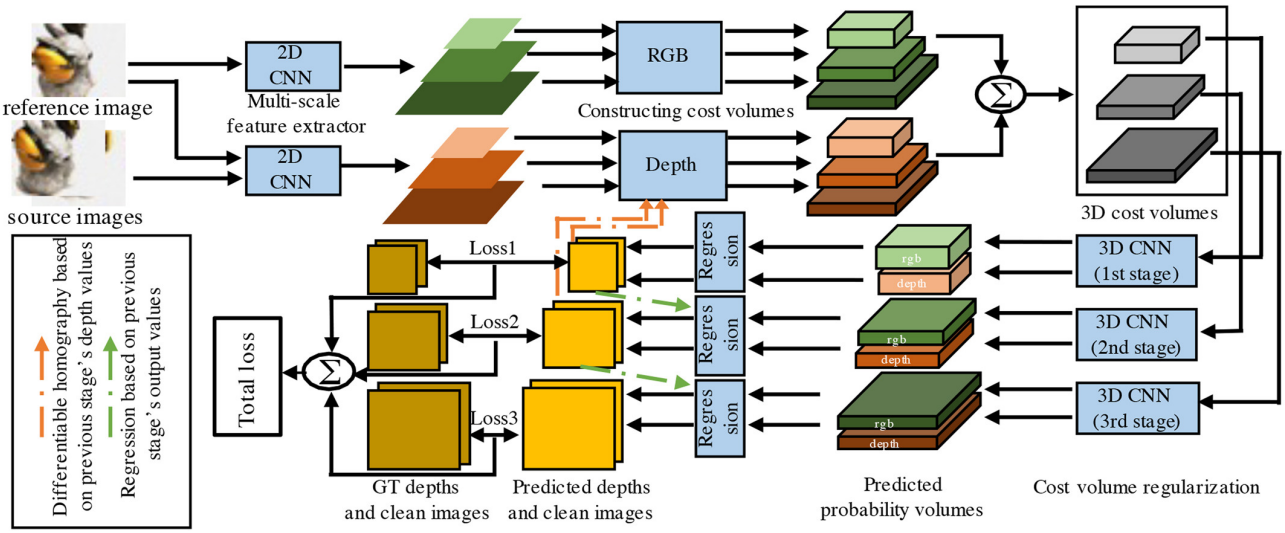


Figure 2.7: Architecture of DEMVSNet [18]

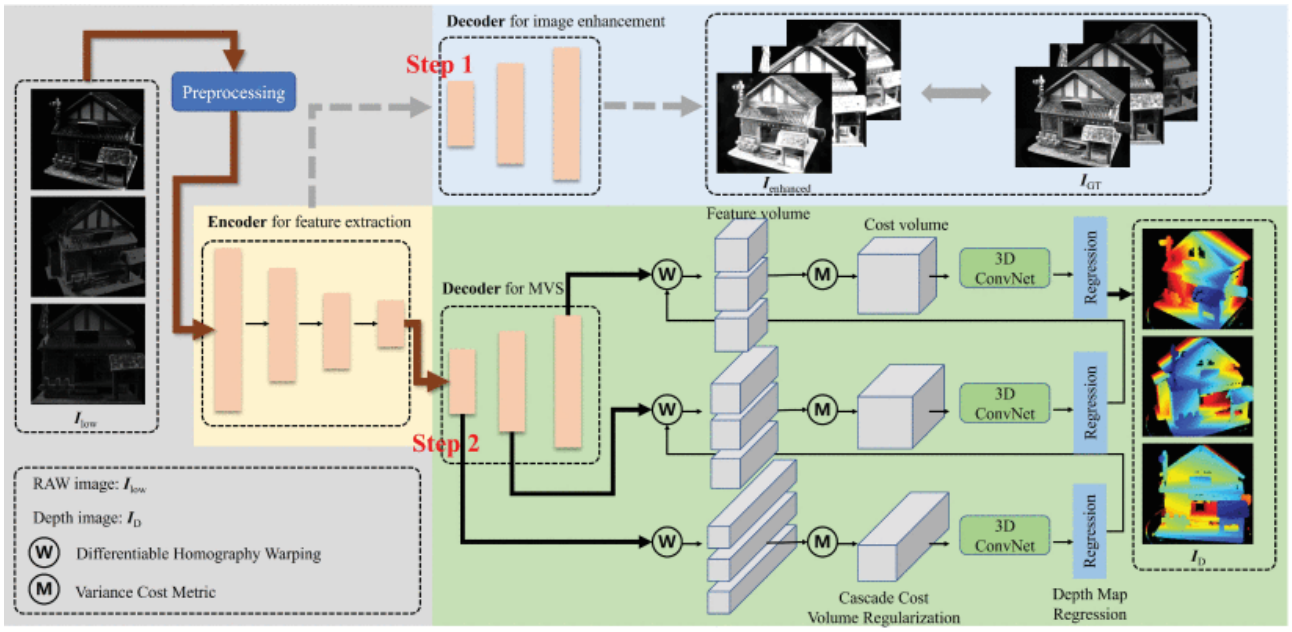


Figure 2.8: Architecture of LoliMVS [55]

Chapter 3

Methodologies

Section 3.1 introduces the overall structure of our model, outlining our architecture for integrating image enhancement with MVS. Section 3.2 describes the image enhancement module and introduces our original image enhancement model modification to preserve the multi-view consistency. Section 3.3 MVS model we used. Section 3.4 describes how we integrate the results of the image enhancement model with MVS to improve MVS performance. Section 3.5 explains the loss function of training, and section 3.6 shows how we train the pipeline. Sections 3.7 and 3.8 present our experiments' depth filtering, fusion methods, and evaluation metrics.

3.1 Overview

Figure 3.1 illustrates the overall architecture of our model, which consists of three main components: the diffusion-based image enhancement block, CasMVSNet for depth prediction, and a feature adapter for refining CasMVSNet's feature map.

Diffusion-Based Image Enhancement Block: We utilize pretrained Low-light diffusion to enhance low illumination images [63]. This module enhances low illumination images, making them more suitable for MVS.

CasMVSNet: We use the CasMVSNet model [15] for depth estimation from multi-view images. Unlike the original CasMVSNet, which relies solely on extracted feature maps from input images, our pipeline incorporates feature maps refined by the diffusion-based image enhancement model.

Feature Adapter: This module connects the image enhancement model with CasMVSNet. It employs Feature Pyramid Network (FPN) [31] to integrate multi-scale outputs from the image enhancement model. In the top-down phase, CasMVSNet feature maps are refined by adding outputs from the adapter's corresponding level.

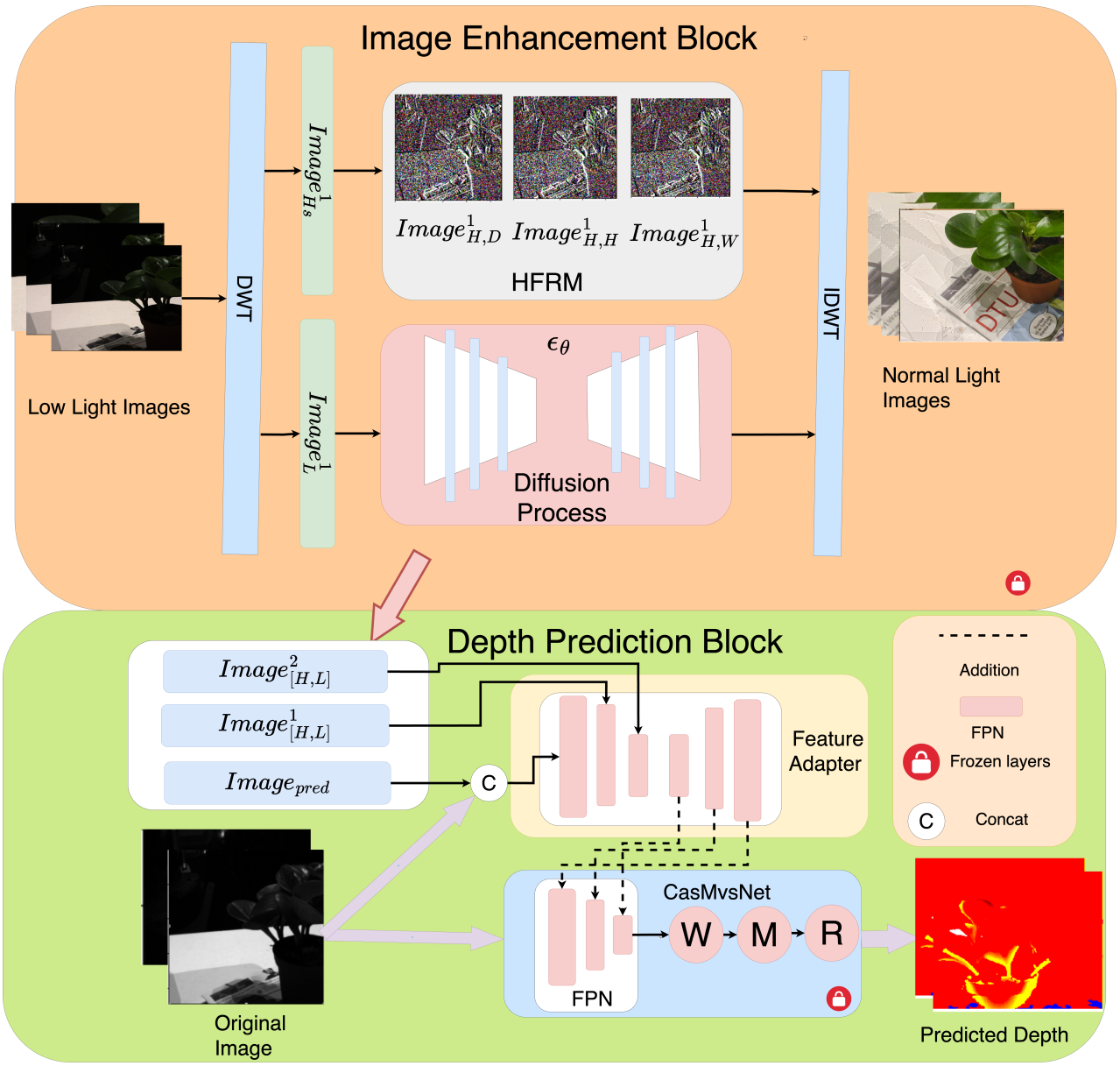


Figure 3.1: **Overview Architecture** This figure presents the overall architecture of our framework. The image enhancement block takes low-light images as input. It processes them through the DWT, decomposing them into high-frequency ($Image^k_H$) and low-frequency ($Image^k_L$) components after k times decompositions of DWT. The high-frequency components in diagonal ($Image^k_{H,D}$), horizontal ($Image^k_{H,H}$), and vertical ($Image^k_{H,W}$) directions are processed via a high-frequency reinforcement module (HFRM 2.1.2.4). The low-frequency components are enhanced via a diffusion model. Subsequently, the feature adapter takes components of DWT as input, where ($Image^k_{[H,L]}$) represents enhanced low-frequency and high-frequency components after the k times decomposition. The outputs in the top-down stage are added to the corresponding levels of the original feature map. The CasMVSNet uses the original structure, where "W" means differentiable homography warping, "M" means cost volume construction, and "R" represents depth regression on cost volume.

3.2 Image Enhancement Module

This module is primarily designed to enhance images captured in low illumination conditions, making it more for 3D reconstruction. We utilize Huang's Low-light Diffusion model for low illumination image enhancement [20]. According to Huang's experiments, Low-light Diffusion demonstrates

superior image enhancement capabilities and generates better image details [20]. For fairness, we compare the results of low-light diffusion with other image enhancement methods in section 4.2.

3.2.1 Cross-Frame Attention

In both wavelet-based conditional diffusion and the HFRM module for low-light diffusion, the enhancement is applied to individual frames only [20]. To improve the consistency of enhancement across multi-view images, we have adapted all single-frame attention blocks into cross-frame attention blocks, inspired by the method used in ViewDiff [19].

The original single-frame attention block in low-light diffusion is defined as:

$$\text{Attn}(Q_n, K_n, V_n) = \text{softmax}\left(\frac{Q_n K_n^T}{\sqrt{d_k}}\right) V_n \tag{3.1}$$

where the query, key, and value matrices Q_n, K_n, V_n are computed as follows, assuming W_q, W_k, W_v are the projection matrices:

$$Q_n = W_q \times \text{Image}_H^n \tag{3.2}$$
$$K_n = W_k \times \text{Image}_H^n \tag{3.3}$$
$$V_n = W_v \times \text{Image}_H^n \tag{3.4}$$

Here, Image_H^n represents the high and low-frequency components of a single-frame image.

In the cross-frame attention mechanism, Q_n remains unchanged, but K_n and V_n are modified to incorporate information from neighboring frames:

$$K_n = W_k \times [\text{Image}_H^{n-1}; \text{Image}_H^n; \text{Image}_H^{n+1}] \tag{3.5}$$
$$V_n = W_v \times [\text{Image}_H^{n-1}; \text{Image}_H^n; \text{Image}_H^{n+1}] \tag{3.6}$$

where $[\text{Image}_H^{n-1}; \text{Image}_H^n; \text{Image}_H^{n+1}]$ denotes the concatenation of the current image with its preceding and succeeding frames.

The weights W_q, W_k, W_v continue to use the pre-trained weights from the single-frame attention model, ensuring seamless integration with the existing framework.

3.3 Feature Adapter Module

We do not directly utilize the enhanced images as input into MVS. We refine the feature maps of learning-based MVS through the feature adapter module.

Our feature adapter employs an Feature Pyramid Network (FPN)-like structure with bottom-up and top-down blocks. The original and enhanced images are initially concatenated and fed into the bottom-up blocks. The outputs of the feature adapter are added to the original feature map at the corresponding levels during the top-down blocks.

Inspired by Res2Net [13], we integrate multi-scale inputs from the Discrete Wavelet Transform (DWT) decomposition into our feature adapter. Each component of the DWT decomposition is individually enhanced by the Low-light Diffusion model. By incorporating components from different frequency bands and scales, our feature adapter can utilize a broader range of useful features to refine the feature map of the MVS model. This approach ensures that the enhanced multi-scale features contribute effectively to improving the depth estimation process.

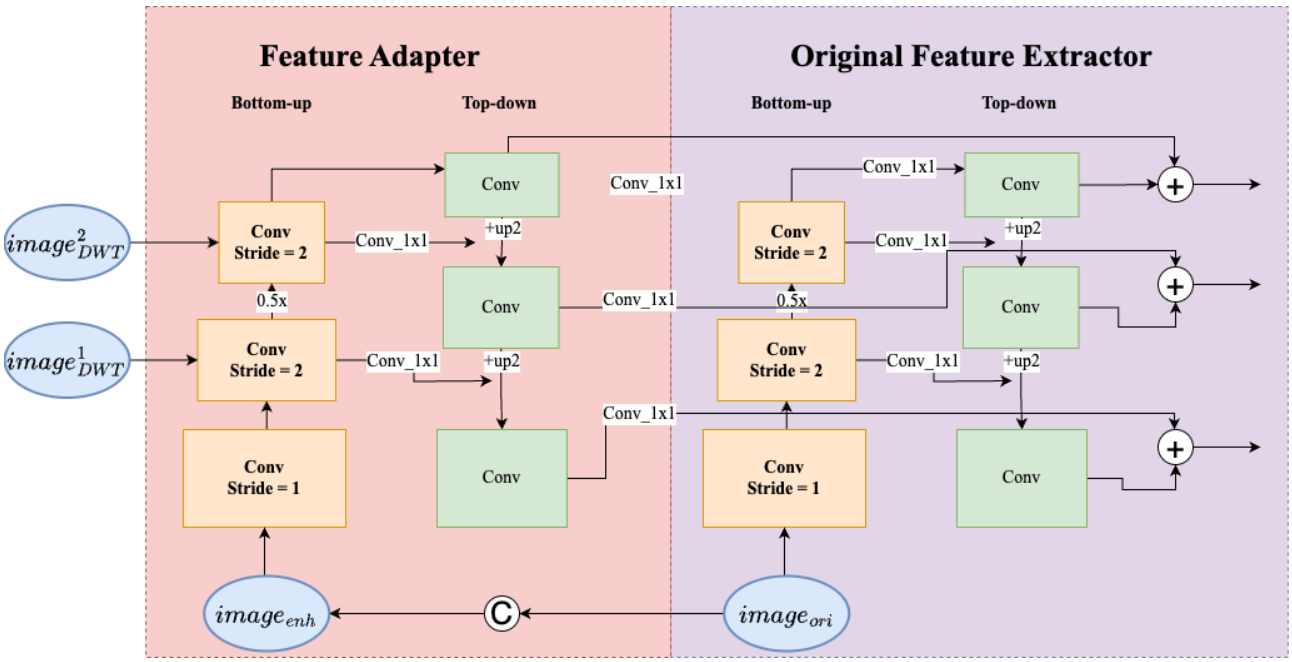


Figure 3.2: Architecture of the feature adapter module integrating enhanced image outputs ($image_{enc}$) with original CasMVSNet ($image_{ori}$) input features. This figure illustrates the comprehensive architecture of the feature adapter, detailing how enhanced image outputs are merged with the original input features. The $Image^k_{DWT}$ indicates the inputs at different DWT levels, where image size is scaled down to $1/2^k$ of the original size, "c" denotes concatenation, "+" indicates addition, and "Conv" represents convolution blocks employed at both the bottom-up and top-down stages of the FPN. "up2" means upsampling to 2 times of original size.

3.4 Learning-based MVS

Our framework builds upon CasMVSNet, employing the same architecture to estimate depth maps from multi-view images. Unlike CasMVSNet, the feature map is refined by a feature adapter and the Low-light Diffusion model.

3.5 Loss

The loss is the same as CasMVSNet [15], where we calculate the sum of mean absolute error (MSE) between the predicted depth map and ground true depth map on each stage of depth estimation:

$$\text{Loss} = \sum_{k=1}^{N} \lambda^k \cdot L^k \tag{3.7}$$

The l^k refers to the loss at the k stage, and λ^k refers to the loss weights at each stage. We apply the same loss function to train our model during the training process.

3.6 Training Strategy

Our methodology uses pre-trained models for the Image Enhancement module and CasMVSNet. Our framework freezes all the weights except for the feature adapter.

Such a training strategy saves computation resource demand. Also, it ensures that any enhancements in performance can be unequivocally attributed to the feature adapter in our framework rather than from extended training cycles on the established baseline models.

3.7 Depth Filtering and Fusion

Our methodology adopts the depth filtering and fusion approach implemented in the code provided by kwea123 [28].

3.7.1 Depth filtering

Dpeht filtering eliminates spurious or background points and ensures multi-view consistency of the depth estimates, following the procedure outlined by Yao et al. [66]. Trustworthy points are determined based on two criteria:

1. Geometric Consistencies

1. **Pixel Consistency:** By projecting the reference view pixels to the source view and back, the reprojected pixel coordinates (p_{reproj}) must lie within one pixel of the original coordinates (p_{ori}): $|p_{\text{reproj}} - p_{\text{ori}}| < 1$.

2. **Depth Consistency:** The original depth (d_{ori}) and the reprojected depth (d_{reproj}) should not differ by more than 1 mm: $|d_{\text{reproj}} - d_{\text{ori}}| < 1\text{mm}$.

2. Confidence Threshold

Points with a predicted depth confidence (d_{pred}) below a threshold, typically set at 0.999, are excluded. The confidence is the probability of the predicted depth relative to all depths post-softmax operation.

3.7.2 Depth fusion

Depth fusion fuses the depth value across different views and determines the color of points. The depth value is averaged across views, while the associated color is determined by the most frequent value among the views.

3.8 Evaluation Metrics

Our evaluation framework includes both depth and point cloud estimation, scrutinized through the following metrics:

3.8.1 Depth Estimation

1. **Mean Absolute Error (MAE)** We compute the MAE between the estimated depth map and the ground truth within the valid mask region as MAE $= \frac{1}{N}\sum_{i=1}^{N}|d_{\text{est},i} - d_{\text{gt},i}|$, where $d_{\text{est},i}$ and $d_{\text{gt},i}$ are the estimated and ground truth depths, respectively, and N is the number of valid pixels.

2. **Accuracy Threshold:** This is the ratio of estimated points within a defined threshold error distance from the ground truth.

3.8.2 Point Clouds

1. **Accuracy** The mean distance from the reconstructed points to the nearest ground truth points, gauging the precision of the reconstruction.

2. **Completeness** The mean distance from the ground truth points to the nearest reconstructed points, assessing the coverage of the reconstruction.

3. **Overall Metric** A combined metric taking the average of the accuracy and completeness to provide an aggregate performance measure.

Chapter 4

Experiment Results and Discussion

4.1 Dataset Preparation

To rigorously evaluate and compare our proposed model, we utilized two distinct datasets: the DTU dataset for baseline comparisons and the 'Tanks and Temples' dataset to assess model generalizability across varied environments.

DTU Dataset The DTU Dataset [1] comprises 119 scans featuring a variety of 80 toy objects. This dataset is primarily collected in a controlled laboratory setting. For our experiments, the dataset is divided into three subsets (several scans not used): 79 scans for training, 18 for validation, and 22 for testing, using the same scan split as CasMVSNet[15]. The DTU dataset is under controlled lighting conditions in 7 classes. It should be noted that the DTU dataset is collected in a laboratory environment. The split of the scan can be seen on: casmvsnet-pl.

Tanks and Temples 'Tanks and Temples' dataset [26] offers a collection of both outdoor and indoor scenes captured through laser scanning, presenting more complex and variable lighting and textural conditions. This dataset is classified into three categories: intermediate, advanced, and a training set. Notably, ground truth data is available only for the training set used for model fine-tuning. The performance of models can be further evaluated by uploading the processed results to the dataset's official website for benchmarking. In our experiment, we test the performance of our model on the intermediate set.

Including the 'Tanks and Temples' dataset allows us to test the robustness and adaptability of our model across different real-world settings, providing insights into its generalizability beyond indoor environments.

4.2 Implemention Details

This section delineates the rationale and process behind selecting specific models and strategies for constructing our pipeline. Critical decisions regarding the image enhancement algorithms, training strategies, and feature fusion methods are discussed.

4.2.1 Image Enhancement Model

The model selected for our study in image enhancement is the Low-light Diffusion model, as discussed by Huang et al. [20]. This model performs better in enhancing images in dark environments than other algorithms [22, 59, 20]. Compared to RetinexNet, the Low-light Diffusion model produces images with less haze, better texture, and superior detail preservation. It also avoids the overexposure effect commonly observed in RetinexNet. When compared to EnlightenGAN, the Low-light Diffusion model provides superior illumination, enhancing low-light areas more effectively.

We present the results of Low-light Diffusion with image enhancement models in Figure 4.1, with the comparison between Low-light diffusion and other image enhancement models in Appendix B.

4.2.2 Training Strategy

We chose four distinct training strategies to determine the optimal approach based on converged training loss and computational resource demands. Our model consists of three main components: a diffusion-based image enhancement block, CasMVSNet, and an adapter bridging the image enhancement block with CasMVSNet. Both the image enhancement module and CasMVSNet utilize pretrained models. The strategies are detailed in Table 4.1, 4.2.

1. **Adapter Only:** Focuses on training only the adapter that connects the image enhancement block to CasMVSNet. This optimizes the integration of enhanced images into the MVS pipeline.
2. **Adapter and MVS:** Involves training both the adapter and CasMVSNet while keeping the image enhancement block frozen.
3. **Full Model Training:** A comprehensive strategy that trains all blocks.
4. **Image Enhancement and Adapter:** Trains the image enhancement block alongside the adapter.

Our findings indicate that training solely the adapter is the most efficient approach. Fine-tuning the adapter and MVS converges at the lowest training errors. However, this improvement may not be solely due to the image enhancement module and adapter. It could result from extended training of CasMVSNe. Training the image enhancement model consumes considerable computational resources and yields marginal gains. Therefore, We focused exclusively on training the feature adapter. This strategy conserves resources and ensures that improvements primarily stem from our integration design.

Table 4.1: Computation resource consumption of different training strategies and minimal training loss after 5 epochs. The learning rate is 0.0001, and the optimizer is Adam.

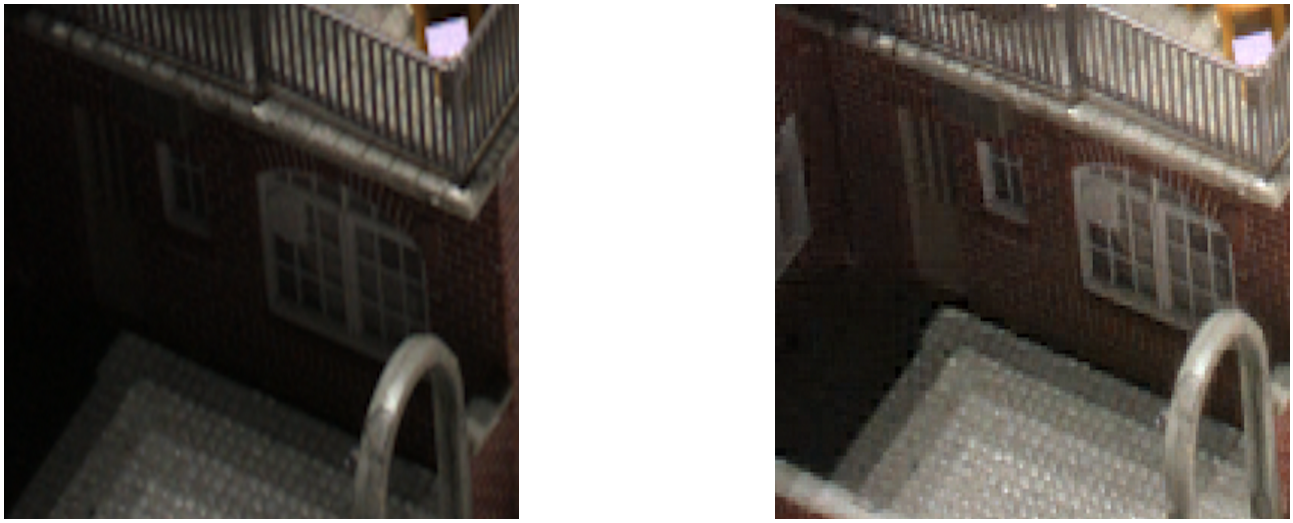
Training Strategy	Model Parameters	Hours/Epoch	GPU RAM	Minimal Loss
Adapter	104 K	50 mins	3.2 GB	4.08
Adapter+MVS	1.1 million	2 hours 10 mins	5.4 GB	3.89
Image Enhancement + MVS	23.2 million	7 hours 50 mins	34 GB	4.12
Image Enhancement + Adapter	22.4 million	6 hours 59 mins	34 GB	4.28

Table 4.2: Parameters and GPU RAM consumption for different model blocks.

Model Block	Model Parameters
Diffusion-based Image Enhancement blocks	22.3 million
Adapter	108 K
CasMVSNet	984 K

(a) A (b) B (c) C

(d) D (e) E (f) F



(g) G (h) H

Figure 4.1: Low-light diffusion output. The first row is low illumination images, and the second row is image enhancement of low illumination images. The last row is the details of enhancement, the left is original, and the right is enhanced details

4.2.3 Feature Adapter

The feature adapter aims to optimize feature extraction to enhance the quality and precision of 3D reconstruction. We chose from three feature adapter designs to refine the original feature map utilizing the enhancement model.

1. **Direct Concatenation:** This method integrates the output of the image enhancement model by directly concatenating it with the input of the primary feature extractor. This straightforward approach allows the primary feature extractor to immediately utilize the enhanced image details, enriching the initial stages of feature processing.
2. **FPN-based Integration:** The output from the image enhancement model is fed into an additional Feature Pyramid Network (FPN) [31]. The feature adapter enhances the original feature maps by adding adapter outputs on original feature maps.
3. **Multi-scale Feature Integration:** This approach modifies the FPN-based integration by using all enhanced components of DWT from low-light diffusion as introduced in section 2.1.2.4). These inputs are added to the bottom-up stage of the FPN. The resultant multi-scale features on the top-down stage are subsequently integrated at corresponding levels of the primary feature extractor within CasMVSNet.

Detailed comparisons of these methods are presented through ablation experiments 4.5.

4.2.4 Loss Function

We adopt the loss function used by the original CasMVSNet, as detailed in Equation 3.7.

4.2.5 Evaluation

The evaluation is conducted on scans with light classes 0,3, and 6 on the DTU dataset. Additionally, we conduct evaluations on the 'Tanks and Temples' dataset, which is captured under normal lighting conditions. The results and their detailed analysis are in section 4.4. The methods are introduced in section 3.8.

4.3 Implementation Details

This section introduces the details of the implementation of our experiments.

4.3.1 Pretrained Model

Our framework uses pretrained models for depth estimation and image enhancement. Specifically, we employ the pretrained CasMVSNet model, provided by Kwea [28], and the low-light diffusion model, developed by Huang [20]. The CasMVSNet model can be found at CasMVSNet_pl, and the low-light diffusion model is available at Diffusion-Low-Light.

4.3.2 Implementation Tools and Framework

The implementation is built upon Kwea's code [28], utilizing PyTorch [34] and PyTorch-Lightning [11] framework. The experiments were conducted using an NVIDIA A100 PCIE with 40 GB GPU RAM, hosted on the AutoDL platform [49] within an Ubuntu system. The configuration of loss weights and CasMVSNet parameters remained unchanged from the original. The only modification is a feature adapter that bridges the output from Low-light Diffusion to CasMVSNet, which is introduced in previous sections.

4.3.3 Training and Testing Settings

The specific settings for training and testing are detailed in Tables 4.3 and 4.4.

Table 4.3: Training Parameters and Hyperparameters.

Training Dataset	DTU
Training Resolution (H,W)	(512,640)
Input Views	3
Training Epochs	10
Feature Map Downsize Ratio	2
Cascaded Stages	5
Depth Interval	2.65mm
Ratio of Depth Interval	[1.0, 2.0, 4.0]
Number of depth Hypothesis	[8, 32, 48]
Batch Size	8
Learning Rate	0.0001
Optimizer	Adam
Learning Rate Scheduler	CosineAnnealingLR
Loss Function	Mean Absolute Error
Loss weights on each level	The same
Trainable Layer	Feature Adapter
Image enhancement Model	Low-light Diffusion

Table 4.4: Evaluation Parameters.

Testing Dataset	DTU
Evaluation Resolution (H, W)	(512, 640)
Input Views	3
Number of consistent Views	3
Confidence threshold	0.999
Generalization Dataset	Tanks and Temples
Evaluation Resolution (H, W)	(864, 1152)
Input Views	3
Number of consistent Views	3
Confidence threshold	0.999

4.4 Evaluation results

4.4.1 Evaluation on DTU

We present the results of our experiments conducted on the DTU dataset. Our evaluation spanned a broad spectrum of lighting conditions to thoroughly assess the performance of our model under both low illumination and normal lighting. Our evaluations included color rendering and balancing, geometric evaluation, and depth estimation.

- **Point Cloud Geometric:** Although there was a slight decrease in accuracy, we observed more improvement in completeness. This trade-off resulted in an overall performance enhancement across various lighting conditions, underscoring the effectiveness of our approach.
- **Depth Estimation Level:** We noted a reduction in the mean absolute error (MAE). However, the improvement in the ratio of pixels with errors under thresholds of 4mm, 2mm, and 1mm was modest, showing only about a 1% increase at various lighting conditions.
- **Color Rendering and Balancing:** Our method improves visualization by enhancing color rendering and balancing, effectively removing dark regions. This results in better visual representation across various lighting conditions.

Our algorithm improves the MVS low illumination images and could improve the performance of normal light. Moreover, we observed that the effectiveness of our methods diminishes as the brightness increases according to both the depth map and point cloud evaluation results. The discussion section will analyze these results (see Section 4.6).

4.4.1.1 Quantitative Test

The quantitative results are listed in table 4.5 and table 4.6. Table 4.5 presents the testing results for point clouds under three different lighting conditions: Light 0, Light 3, and Light 6. The metrics evaluated are accuracy (Acc.), completeness (Comp.), and overall performance. Our method shows a marginal improvement in overall performance compared to CasMVSNet. Specifically, our model achieves better completeness across all lighting conditions, though the accuracy slightly decreases.

Table 4.6 compares our method with CasMVSNet on depth map estimation. The metrics include mean absolute error (Abs Error) and the ratio of points with the mean absolute error below 1mm, 2mm, and 4mm (1mm Acc, 2mm Acc, 4mm Acc). Our method consistently shows a lower mean absolute error across all lighting conditions. The improvement in 1mm Acc is slight, while the 2mm and 4mm Acc remain comparable to CasMVSNet.

These results indicate that while our method provides notable enhancements in completeness and slight improvements in depth map accuracy, the overall gains in geometric accuracy are marginal.

4.4.1.2 Qualitative Test

The qualitative results are illustrated in Figures 4.2, 4.4, and 4.3. These results highlight two key advantages of our model over the original method:

1. Our model recovers more points, particularly in featureless areas, while maintaining the quality in normally lit areas. (See Figure 4.3, 4.4, and more results in Appendix C.1)
2. The appearance of point clouds is notably improved, with enhanced color and surface details, leading to a more visually appealing representation of objects. (See Figure 4.2 and more results in Appendix C.2)

Light: 0	Acc. ↓	Comp. ↓	Overall ↓
CasMVSNet	**0.328**	0.472	0.400
Ours	0.330	**0.460**	**0.395**
Light: 3	Acc. ↓	Comp. ↓	Overall ↓
CasMVSNet	**0.327**	0.464	0.395
Ours	0.328	**0.454**	**0.391**
Light: 6	Acc. ↓	Comp. ↓	Overall ↓
CasMVSNet	**0.315**	0.457	0.386
Ours	0.315	**0.452**	**0.384**

Table 4.5: The testing result of points cloud on DTU, in lighting condition 0,3,6

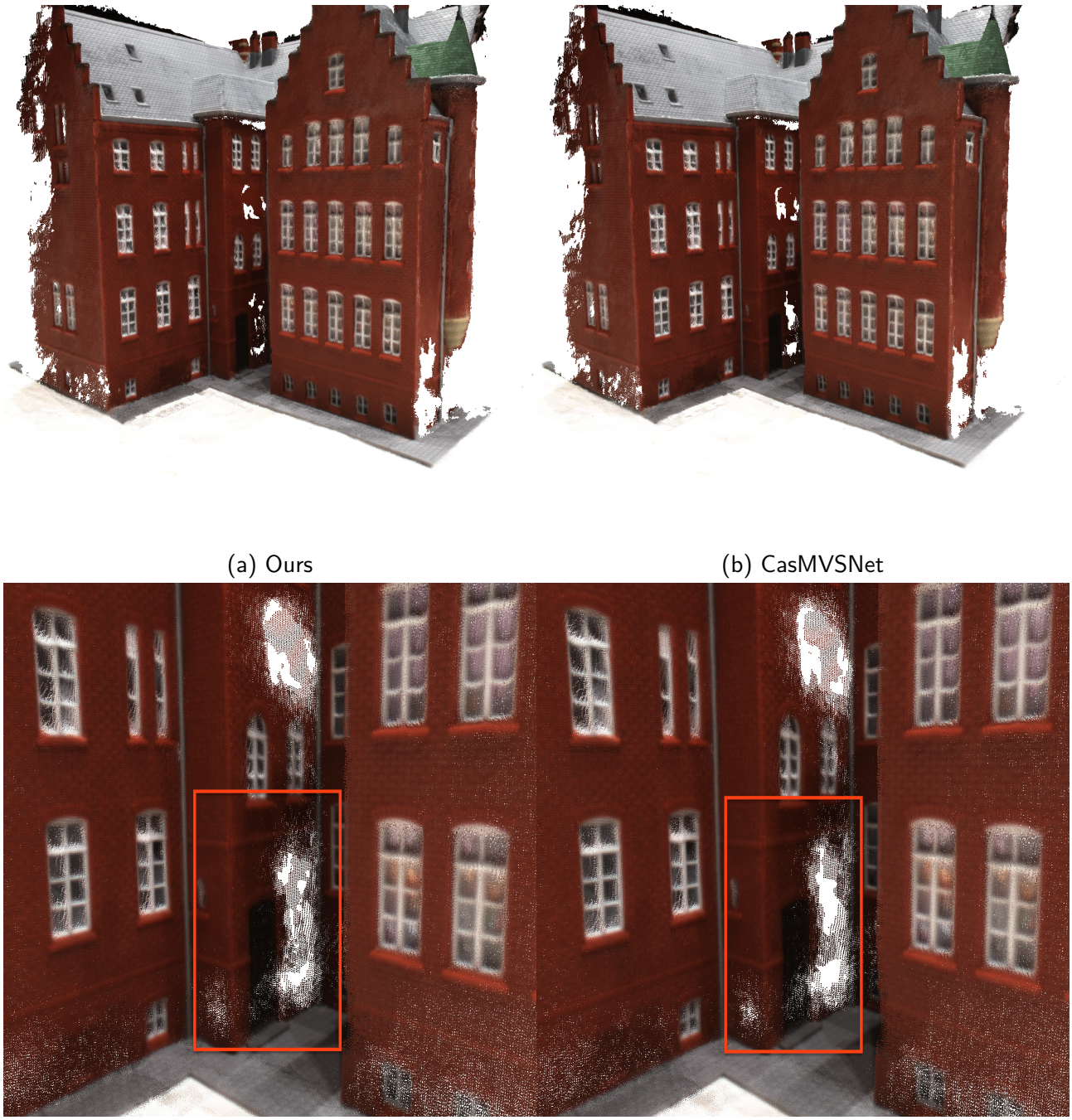
Light: 0	Abs Error	1mm Acc	2mm Acc	4mm Acc
Ours	6.39	70%	82%	90%
CasMVSNet	6.85	69%	82%	90%
Light: 3	Abs Error	1mm Acc	2mm Acc	4mm Acc
Ours	6.27	69%	83%	90%
CasMVSNet	6.70	69%	82%	90%
Light: 6	Abs Error	1mm Acc	2mm Acc	4mm Acc
Ours	6.23	70%	84%	91%
CasMVSNet	6.59	70%	84%	90%

Table 4.6: Comparison of Ours and CasMVSNet Methods on depth map estimation. Abs Error is the mean absolute error, and 1mm Acc, 2mm Acc, and 4mm Acc is the ratio of points with the mean absolute error below 1mm, 2mm, and 4mm

(a) CasMVSNet (b) Ours

(c) CasMVSNet (d) Ours

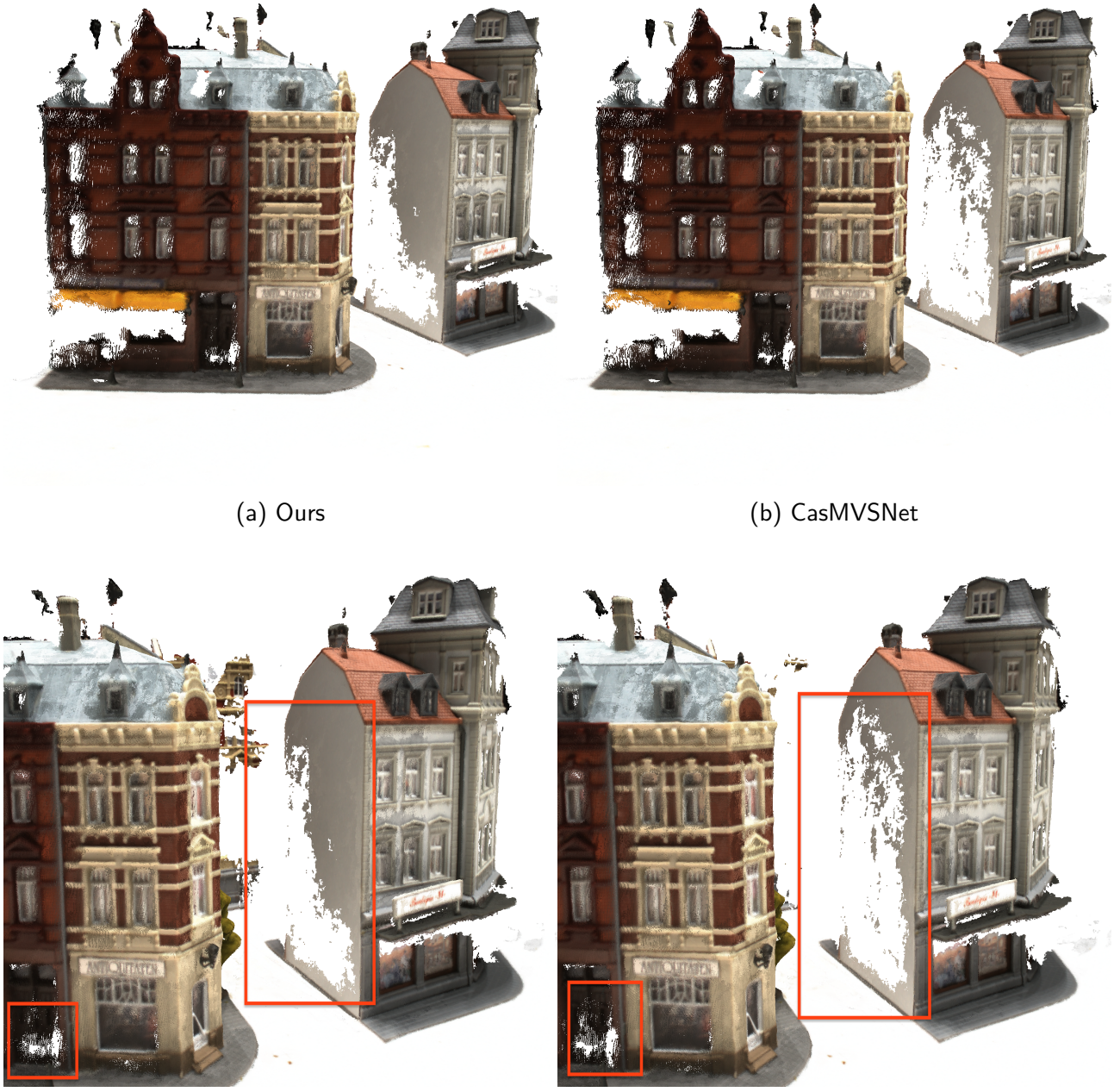
Figure 4.2: This is the comparison of color between the original CasMVSNet and our proposed model



(a) Ours (b) CasMVSNet

(c) Details on our model (d) Details on CasMVSNet

Figure 4.3: Qulitative evaluation of our methods with CasMVSNet on scan24



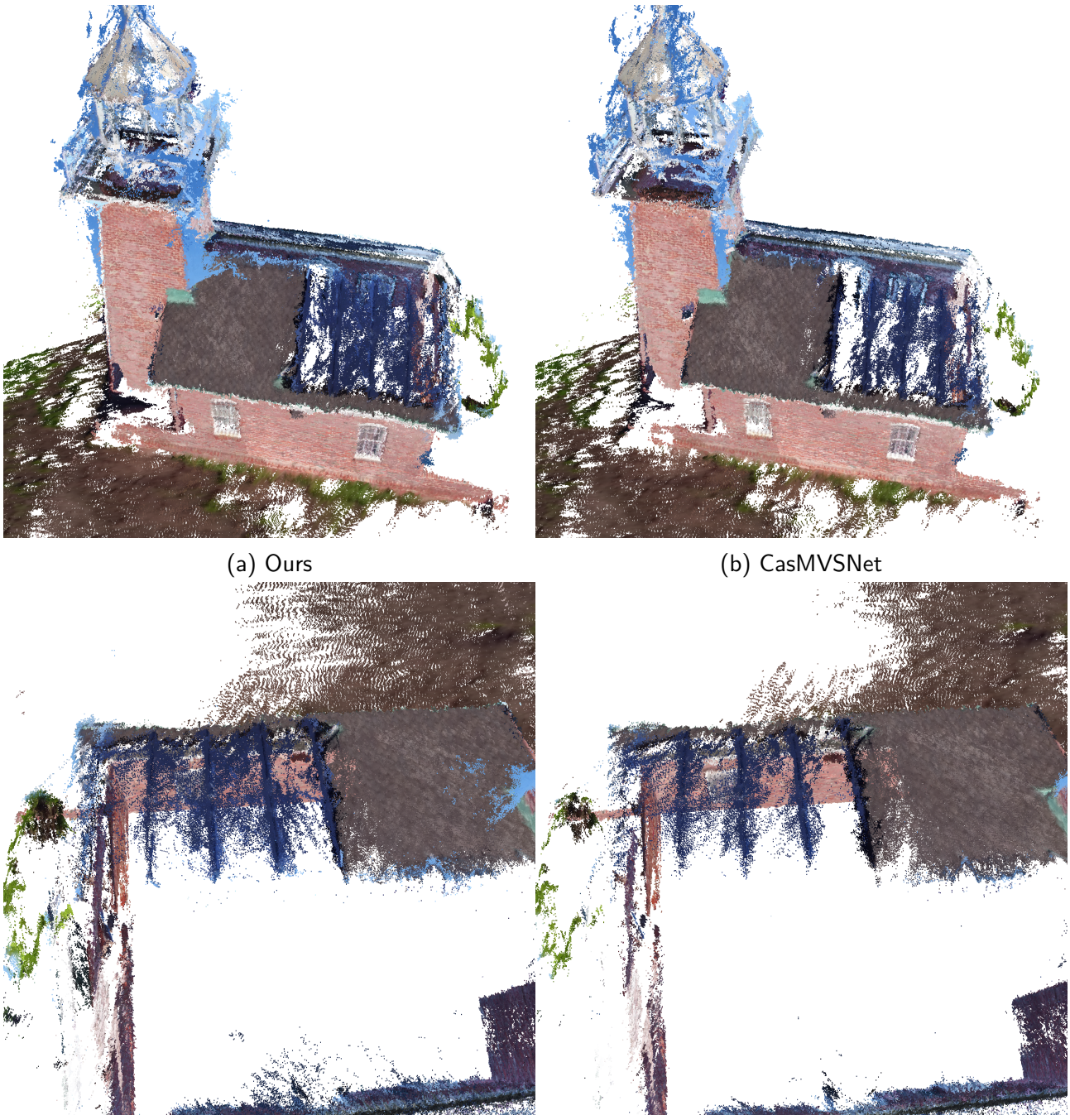
(a) Ours

(b) CasMVSNet

(c) Details on our model

(d) Details on CasMVSNet

Figure 4.4: Qulitative evaluation of our methods with CasMVSNet on scan28



(a) Ours (b) CasMVSNet

(c) Ours (d) CasMVSNet

Figure 4.5: Qulitative evaluation of on 'tanks and temples', the scan name: Lighthouse

(a) Ours (b) CasMVSNet

(c) Ours (d) CasMVSNet

Figure 4.6: Qulitative evaluation of on 'tanks and temples', the scan name: Family

4.4.2 Tanks and Temples

Our evaluation of the 'Tanks and Temples' dataset demonstrates an overall performance improvement, corroborating the generalizability of our methods across different datasets. Notably, qualitative assessments reveal enhanced completeness in featureless regions.

4.4.2.1 Quantitative Test

Despite the intermediate set of 'Tanks and Temples' being captured in outdoor environments under normal lighting conditions, our model exhibits improvements in overall performance. Detailed results are presented in Table 4.7.

Model	F1 Score
Proposed Model	46.24
CasMVSNet	45.30

Table 4.7: F1 scores of test results on 'Tanks and Temples'

4.4.2.2 Qualitative Test

Figures 4.5 and 4.6 show the outcomes of our model on the 'Tanks and Temples' dataset, highlighting the enhanced details and improved model performance in various outdoor scenes. More samples can be found in Appendix C.3.

4.5 Ablation experiments

We conducted a series of ablation tests to assess the contributions of each component within our design. The goals were to determine (1) whether the image enhancement method contributes to the improvement of MVS, (2) whether our feature adapter designs effectively refine the feature map, and (3) whether the framework works for other MVS pipelines such as GeoMVSNet [73]. The quantitative results of different ablation experiments can be seen in Table 4.8.

Light: 0	Acc. ↓	Comp. ↓	Overall ↓
CasMVSNet	0.328	0.472	0.400
Ours	0.330	**0.460**	**0.395**
Only feature adapter	**0.326**	0.475	0.401
Without DWT input	0.329	0.462	0.396
Only image enhancement input	0.329	0.465	0.397
GeoMVSNet (original)	0.383	0.392	0.388
GeoMVSNet (ours)	0.343	0.338	0.340
MVSNet (original)	0.540	0.492	0.523
MVSNet (ours)	0.547	0.485	0.516
Light: 3	Acc. ↓	Comp. ↓	Overall ↓
CasMVSNet	0.327	0.464	0.396
Ours	0.328	**0.454**	**0.391**
Only feature adapter	**0.326**	0.463	0.395
Without DWT input	0.327	0.456	0.392
Only image enhancement input	0.328	0.457	0.393
GeoMVSNet (original)	0.342	0.344	0.343
GeoMVSNet (ours)	0.322	0.302	0.312
MVSNet (original)	0.538	0.501	0.520
MVSNet (ours)	0.542	0.493	0.518
Light: 6	Acc. ↓	Comp. ↓	Overall ↓
CasMVSNet	0.315	0.457	0.386
Ours	0.315	**0.452**	**0.384**
Only feature adapter	0.316	0.459	0.388
Without DWT input	**0.313**	0.456	0.385
Only image enhancement input	0.316	0.454	0.385
GeoMVSNet (original)	0.348	0.294	0.321
GeoMVSNet (ours)	0.325	0.282	0.304
MVSNet (original)	0.535	0.491	0.513
MVSNet (ours)	0.541	0.493	0.517

Table 4.8: The testing result of points cloud on DTU, in lighting condition 0,4,6

4.5.1 Ablation Experiment 1: only feature adapter

We evaluated the isolated impact of a feature adapter by integrating it into the original CasMVSNet architecture without the image enhancement model. This adapter operates in parallel to the original feature extractor and adds the outputs with original feature maps. The results indicate that adding the feature adapter without any image enhancement contents did not improve performance. On the contrary, it resulted in a slight decrease. This suggests that merely increasing the complexity of the feature extraction process without enhancing the input features does not contribute to better performance.

4.5.2 Ablation Experiment 2: only image enhancement input

This experiment involved directly concatenating the outputs from the image enhancement block with the input image before processing them using the original FPN structure within CasMVSNet. While this method showed improved performance, particularly under low illumination conditions, it did not surpass the enhancements observed with our integrated approach using the feature adapter. These results confirm that incorporating image enhancement boosts MVS performance, especially in challenging lighting conditions.

4.5.3 Ablation Experiment 3: Without Multi-scale Input

Our design takes the multi-scale output from the Low-light Diffusion model. To test the effectiveness of this multi-scale strategy, we conducted an ablation experiment by removing the multi-scale inputs. Instead, we only used the final output: the enhanced images from the inverse discrete wavelet transform (IDWT), which combines all enhanced components. This approach allows us to evaluate the contribution of multi-scale features to our overall method.

4.5.4 Ablation Experiment 4: Framework Applicability to Other MVS Pipelines

To evaluate the generalizability of our framework, we extended it to other MVS pipelines, specifically GeoMVSNet [73] and MVSNet [66]. We integrated the same feature adapter design into these models. The parallel feature adapters process the refined images from the enhancement model and add the adapter output with the original feature maps.

4.5.5 Summary

The results of our ablation experiments are presented in Tables 4.8. Experiment 1 shows that simply increasing network complexity by adding the feature adapter without image enhancement results in only marginal improvements. Experiment 2 demonstrates that incorporating the image enhancement model significantly improves the performance of our framework. Experiment 3 reveals that our design, which includes the parallel feature adapter and multi-scale input, provides only marginal improvements compared to simply concatenating the image enhancement output.

4.6 Discussion

4.6.1 Color Rendering and Balancing

The qualitative results reveal our model's enhanced color rendering and balancing effects. The model improves visualization by enhancing color details and illumination in low-light conditions. However,

although the illumination of the images is improved, the enhancement effects are still imbalanced. Shadows and occlusions remain problematic, especially in regions with occlusions that block light.

4.6.2 Geometric and Depth Estimation Evaluation

The experimental results indicate that although improvements are observed, they are relatively marginal and not very pronounced. This suggests that while our framework has the potential to enhance the process of 3D reconstruction, there is still significant room for improvement.

From the quantitative results in geometric evaluation: Table 4.5 and Table 4.7 and depth estimation 4.6, Our method shows a slight increase in completeness but a marginal decrease in accuracy. This trade-off results in an overall enhancement in performance across various lighting conditions but remains subtle.

Qualitative results from both the 'Tanks and Temples' and DTU datasets further illustrate our model's ability to recover more points in shadowed or otherwise featureless regions, as shown in Figures 4.4, 4.3, 4.5 and 4.6. However, the accuracy of these newly added points is relatively low compared to the original MVS points. The newly added points are not as dense and smooth as those in well-lit areas, indicating that while our model enhances completeness, it does not significantly improve the precision of 3D reconstructions.

The geometric and depth estimation evaluations show that our framework can potentially improve 3D reconstruction, particularly in filling void regions. However, the improvements are modest, and there is ample room for further enhancement in both accuracy and completeness.

4.6.3 Training Strategies

According to Tables 4.2 and 4.1, fine-tuning only the feature adapter achieves the second lowest training loss while demanding the least computational resources. Our method is highly efficient in terms of resource usage. However, incorporating the entire MVS model into the training process results in the lowest training loss. Despite this, including the MVS model in training can introduce risks such as overfitting and dataset bias since the improvement may stem from extended training epochs on the pre-trained MVS model. Additionally, this approach increases computational resource demands.

For our research, which aims to explore the contributions of newly added components, we decided not to include the MVS model in training. This ensures that improvements are solely attributed to the new components, adhering to the principle of control variables.

In practical applications, if computational resources are sufficient, incorporating more frozen layers and even the MVS model into training could lead to better performance. This approach might enhance both accuracy and completeness, making it a reasonable choice for achieving higher-quality results.

4.6.4 Robustness and Generalizability of Our Methods

Quantitative and qualitative evaluations of our approach maintain marginal improvements across varying lighting conditions, datasets, and MVS methods, demonstrating our model's adaptability and robustness.

The marginal improvement is consistent in different lighting conditions. However, as lighting intensity increases, the extent of these improvements diminishes, as evidenced by our experimental results on the DTU dataset. This trend can be attributed to the inherent characteristics of the low-light diffusion process, which introduces less additional content in already well-lit images, thus providing limited enhancement under such conditions.

The performance of our model on the 'Tanks and Temples' dataset further validates that it performs well in laboratory environments and shows commendable improvements in outdoor and brightly lit environments. This observation underscores our model's capability to enhance 3D reconstruction in varied real-world conditions.

Additionally, we tested our framework with different learning-based MVS methods, including CasMVSNet [15], MVSNet [66], and GeoMVSNet [73]. The results indicate that our framework is robust and generalizable across various MVS methods. We found that while CasMVSNet and MVSNet showed marginal improvements, the enhancement is relatively pronounced with GeoMVSNet. This indicates that our framework's effectiveness can vary based on the choice of the underlying MVS method.

Overall, our framework demonstrates robust performance and generalizability. It consistently improves across different scenarios, lighting conditions, datasets, and MVS methods.

4.6.5 Design of Feature Adapter

Our ablation experiments 1-3 are designed to dissect the individual contributions of each architectural component within our model, enhancing our understanding of their roles in improving 3D reconstruction performance under varied lighting conditions. These experiments yielded two key findings:

First, our tests reveal how each design component influences the MVS. We discovered that while multi-scale input offers only marginal improvements, the diffusion-based image enhancement model contributes most to performance enhancement. Although feature adapters alone provide modest benefits, their combination with the image enhancement model results in the best performance compared to using the image enhancement model in isolation.

Second, the ablation experiments' results validate the effectiveness of our architectural decisions and open pathways for further improvements. The observed synergy among the components confirms that our integrated approach is more effective in achieving more improvements than other methodologies.

4.6.6 pros and cons

4.6.6.1 Pros

Our method enhances the performance of Multi-View Stereo (MVS) in low illumination conditions, demonstrating robustness and generalizability across various scenarios. This extends the applicability of image-based 3D reconstruction.

Our approach is robust and adaptable to different lighting conditions, showing consistent improvements in low and normal light environments. This capability is validated by quantitative and qualitative evaluations on the DTU and Tanks and Temples datasets. Additionally, our framework is easy to transfer to other MVS pipelines, such as GeoMVSNet and MVSNet, highlighting its versatility.

Secondly, our method simplifies the MVS process by eliminating the need for individual preprocessing steps like manual brightness adjustments. This end-to-end framework is efficient and lightweight, requiring minimal computational resources. This makes it particularly useful in time-sensitive applications such as emergency response planning and rapid environment assessments.

Thirdly, our technique significantly improves the visual quality of 3D models. Enhancing the color visualization and rendering effects produces clearer and sharper images of objects captured in low light. This enhancement is crucial for accurate color and texture information and is vital for realistic 3D model rendering. Our method improves the geometric details, surface textures, and colors, making the 3D reconstructions more detailed and vibrant. This is essential for historical site documentation, augmented reality, gaming, and animation applications.

Lastly, the improvements observed in our methods are not significant for MVSNet and CasMVS-Net, with only very slight performance enhancements. This indicates that further optimization and studies are needed while our framework shows potential.

Our framework demonstrates potential improvements in low-light conditions and efficiency in the MVS process. However, significant areas require further investigation and refinement to realize their full potential. Although the improvements in geometric shape and depth estimation are marginal, they do indicate a positive trend and potential for further development.

4.6.6.2 Cons

However, there are several limitations and questions that remain unanswered in our research, indicating that the results are far from perfect:

Firstly, multi-view consistency in the original image enhancement model is a challenge. Although we used multi-frame attention and a feature adapter to fuse the output to mitigate this issue, multi-view inconsistencies still exist. Fine-tuning the image enhancement model for multi-view consistent enhancement is necessary but is also more expensive and complex.

The optimal type of image enhancement model for MVS remains an open question. We used a basic method to select the image enhancement model from three commonly used models by comparing their image enhancement results. More quantitative analysis and a broader selection from a larger pool of image enhancement models are necessary for better improvements. Ensuring

Secondly, our model's contribution to geometric accuracy lacks deep analysis. To evaluate the final contribution of our methods, we need at least a mask on the dark regions and well-lit regions from the original methods and compare the accuracy and completeness in these areas. In this way, we can quantify the contribution of our framework to dark regions and gain deep insight into the interaction between our framework and multi-view images.

Lastly, the improvements observed in our methods are not significant for MVSNet and Cas-MVSNet, with only very slight performance enhancements. This indicates that while our framework shows potential, its effectiveness can vary depending on the specific MVS pipeline used, and further studies and optimization are needed.

In summary, while our framework demonstrates potential improvements in low-light conditions and efficiency in the MVS process, significant areas require further investigation and refinement to realize their full potential.

Chapter 5

Conclusion and limitations

5.1 Conclusion

This thesis investigates enhancing 3D reconstruction in low illumination environments by leveraging the prior knowledge of diffusion processes embedded within an existing image enhancement model. We propose a streamlined and efficient framework integrating CasMVSNet with a diffusion-based image enhancement model, Low-light Diffusion. Our pipeline is composed of three primary components:

1. **Diffusion-based Image Enhancement Model:** This model processes images captured in low illumination conditions, enhancing them to resemble those taken under normal lighting conditions.

2. **Original CasMVSNet:** It performs depth estimation with a feature map refined by the feature adapter.

3. **Multi-scale Feature Adapter:** A novel adapter that merges the architectures of FPN and Res2Net. This adapter connects the diffusion-based image enhancement model with CasMVS-Net. It takes multi-scale inputs from Low-light diffusion outputs [20] during the bottom-up stage of the FPN. The adapter's outputs integrate with the original features at corresponding levels through addition during the top-down stage.

Training is confined to the feature adapter, enhancing the efficiency of our method. This targeted training approach conserves computational resources while preserving the improvements.

Our results answered the research questions:

1. **Which image enhancement model is suitable for MVS tasks under low illumination?** Our evaluation identified the Low-light Diffusion model proposed by [20] demonstrating superior performance to other image enhancement algorithms. Firstly, our method performs better enhancement effects, improving illumination and preserving details in poorly lit areas. Secondly, the framework exhibits improvements when using the Low-light Diffusion model, enhancing both color visualization and geometric shape reconstruction in MVS tasks.

2. **What architecture is suitable for integrating image enhancement models into MVS to boost 3D reconstruction?**
   We have developed a feature adapter that integrates multi-scale output from a low-light diffusion to refine the feature maps extracted by the FPN within CasMVSNet. The ablation experiments explored different integration architectures, including the image enhancement model and the multi-scale input design. Results indicate that our feature adapter design has provided more significant improvements than other methods, showing its effectiveness and versatility.

3. **How can image enhancement models, which typically contain millions of parameters, be efficiently utilized for 3D reconstruction tasks to minimize computational demands while maintaining good performance?**
   Our training strategy is only to train the feature adapter while all layers in the pretrained model are frozen. We tested different training strategies, and the proposed training strategy achieved the goal of training on fewer computation resources while the performance was not largely degraded. In addition, our methods save the computation resources and help measure the final study goal. Since all the weights of CasMVSNet and the image enhancement model are frozen, the training is only implemented on the adapter network, so we could ensure that the improvements only come from the adapter integrating the diffusion-based image enhancement model and CasMVSNet. This excludes the possible intervention from which the improvement may come, such as increased training on CasMVSNet or the image enhancement model.

Finally, the main research question is answered:
To what extent can we utilize the prior knowledge of the existing image enhancement model to improve the performance of MVS?

To enhance MVS performance in low-illumination, our model introduces an effective end-to-end framework that utilizes an image enhancement model to improve 3D reconstruction. Our experiments, conducted on both the DTU dataset and the 'Tanks and Temples' dataset under various lighting conditions, indicate that our architecture significantly enhances color rendering and visualization of 3D reconstructions. Additionally, we observed a slight improvement in geometric shape accuracy and a notable improvement in completeness.

Furthermore, our framework was tested on other MVS pipelines, including GeoMVSNet and MVSNet. The results show that our method is robust and adaptable across different MVS frameworks, maintaining its effectiveness in various scenarios. The improvements were relatively more significant in challenging lighting conditions, showcasing the robustness and generalizability of our approach. This adaptability highlights the potential for our framework to be applied broadly, improving MVS performance in diverse real-world applications.

5.2 Limitations

Our research, while contributing to the field of 3D reconstruction in low illumination environments, has certain limitations that should be noted:

1. **Accuracy vs. Completeness:** Our methods improve overall performance but at the expense of accuracy. This trade-off, favoring completeness, may reduce the perceived benefits, particularly in precision-critical applications.

2. **Fine-tuning on the DTU Dataset:** Our model is fine-tuned exclusively on the DTU dataset, characterized by controlled lighting conditions. The dataset's lowest lighting class does not adequately represent severe low-light environments. In addition, the DTU dataset is collected in a laboratory environment, which can only represent the indoor environment. This limitation may result in less robustness and generalizability of our proposed model. Expanding our training to include datasets with a broader range of illumination environments, especially those incorporating severe low-light effects and outdoor scenes, would enhance the robustness and applicability of our findings.

3. **Single-Frame Image Enhancement Model:** Our current image enhancement model is trained on single frames due to limited computational resources and the absence of a suitable multi-view low illumination dataset. Although we incorporate a cross-frame attention block

in the high frequency Reinforcement Module (HFRM), it is not enough to replace training on multi-view datasets. This results in a lack of multi-view consistency, potentially degrading the quality of the final output.

4. **Computational Resource Requirements:** Despite our efficient fine-tuning approach, which significantly reduces the demand for computational resources, the image enhancement model still requires up to 20 times more parameters than the original CasMVSNet. This requirement means that substantial computational resources are still necessary for processing, which could be a limitation in resource-constrained environments.

5.3 Future work

Several promising directions could be pursued to enhance the performance of our study. The first is to develop a multi-view low illumination dataset for MVS. Current datasets predominantly feature ideal lighting conditions. Even for the darkest class of DTU, there are still many well-lit regions. Only incorporating normal lighting conditions limits the study of MVS and reduces the robustness of pretrained MVS models, which are typically optimized for normal lighting conditions only.

Another direction is to expand the range of image enhancement methods to improve the robustness and overall performance of the image enhancement model. Enhancing images with challenges beyond low illumination, such as haze, fog, or overexposure, could also be addressed. These conditions could be improved by integrating sophisticated image enhancement algorithms with MVS using similar or more advanced architectures.

Additionally, generative-based methods hold great potential beyond image enhancement, extending to tasks like frame interpolation, occlusion removal, and texture addition in textureless areas. Particularly, there is a compelling case for developing a multi-view consistent diffusion model that enhances multi-view images in a way that is more conducive to 3D reconstruction. Such a model would ensure consistency across different viewpoints, significantly improving the quality and accuracy of the modeling results. Emphasizing a multi-view approach is crucial, as relying solely on single-frame enhancement could lead to inconsistencies in the 3D structure, underscoring the importance of adopting a multi-view consistent generative model to achieve superior and more reliable reconstruction outcomes.

Given the observed decrease in performance enhancement with increased brightness, future work could also focus on adapting the integration strategy to adjust based on dynamic ambient lighting conditions. Exploring more advanced methods for seamlessly integrating image enhancement outputs with MVS features could yield more significant improvements, particularly in varied environmental conditions.

Bibliography

[1] Henrik Aanæs et al. "Large-Scale Data for Multiple-View Stereopsis". In: *International Journal of Computer Vision* (2016), pp. 1–16.

[2] Sanaa Abu Alasal et al. "Improving passive 3D model reconstruction using image enhancement". In: (2018), pp. 1–7.

[3] Andrea Ballabeni et al. "Advances in image pre-processing to improve automated 3D reconstruction". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 40 (2015), pp. 315–323.

[4] Connelly Barnes et al. "PatchMatch: A randomized correspondence algorithm for structural image editing". In: *ACM Trans. Graph.* 28.3 (2009), p. 24.

[5] Michael Bleyer, Christoph Rhemann, and Carsten Rother. "Patchmatch stereo-stereo matching with slanted support windows." In: 11 (2011), pp. 1–11.

[6] Pawel Burdziakowski and Katarzyna Bobkowska. "UAV Photogrammetry under Poor Lighting ConditionsâAccuracy Considerations". In: *Sensors* 21.10 (2021).

[7] Chen Chen et al. "Learning to see in the dark". In: (2018), pp. 3291–3300.

[8] Prafulla Dhariwal and Alexander Nichol. "Diffusion models beat gans on image synthesis". In: *Advances in neural information processing systems* 34 (2021), pp. 8780–8794.

[9] Carl Doersch. "Tutorial on variational autoencoders". In: *arXiv preprint arXiv:1606.05908* (2016).

[10] Tim Edwards. "Discrete wavelet transforms: Theory and implementation". In: *Universidad de* 1991 (1991), pp. 28–35.

[11] William Falcon. *PyTorch Lightning*. `https://github.com/PyTorchLightning/pytorch-lightning`. Accessed: [Insert today's date here]. 2019.

[12] Yasutaka Furukawa, Carlos Hernández, et al. "Multi-view stereo: A tutorial". In: *Foundations and Trends® in Computer Graphics and Vision* 9.1-2 (2015), pp. 1–148.

[13] Shang-Hua Gao et al. "Res2net: A new multi-scale backbone architecture". In: *IEEE transactions on pattern analysis and machine intelligence* 43.2 (2019), pp. 652–662.

[14] Ian Goodfellow et al. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[15] Xiaodong Gu et al. "Cascade cost volume for high-resolution multi-view stereo and stereo matching". In: (2020), pp. 2495–2504.

[16] Chunle Guo et al. "Zero-reference deep curve estimation for low-light image enhancement". In: (2020), pp. 1780–1789.

[17] Jiawei Guo et al. "A survey on image enhancement for Low-light images". In: *Heliyon* (2023).

[18] Jiawei Han et al. "DEMVSNet: Denoising and depth inference for unstructured multi-view stereo on noised images". In: *IET Computer Vision* 16.7 (2022), pp. 570–580.

[19] Lukas Höllein et al. "Viewdiff: 3d-consistent image generation with text-to-image models". In: *arXiv preprint arXiv:2403.01807* (2024).

[20] Yourui Huang et al. "Low illumination soybean plant reconstruction and trait perception". In: *Agriculture* 12.12 (2022), p. 2067.

[21] Nail Ibrahimli et al. "DDL-MVS: Depth Discontinuity Learning for Multi-View Stereo Networks". In: *Remote Sensing* 15.12 (2023), p. 2970.

[22] Yifan Jiang et al. "Enlightengan: Deep light enhancement without paired supervision". In: *IEEE transactions on image processing* 30 (2021), pp. 2340–2349.

[23] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. "Properties and performance of a center/surround retinex". In: *IEEE transactions on image processing* 6.3 (1997), pp. 451–462.

[24] Christoforos Kanellakis, Petros Karvelis, and George Nikolakopoulos. "On image based enhancement for 3D dense reconstruction of low light aerial visual inspected environments". In: (2020), pp. 265–279.

[25] Manpreet Kaur, Jasdeep Kaur, and Jappreet Kaur. "Survey of contrast enhancement techniques based on histogram equalization". In: *International Journal of Advanced Computer Science and Applications* 2.7 (2011).

[26] Arno Knapitsch et al. "Tanks and temples: Benchmarking large-scale scene reconstruction". In: *ACM Transactions on Graphics (ToG)* 36.4 (2017), pp. 1–13.

[27] Anestis Koutsoudis et al. "Using noise function-based patterns to enhance photogrammetric 3D reconstruction performance of featureless surfaces". In: *Journal of Cultural Heritage* 16.5 (2015), pp. 664–670.

[28] kwea123. *CasMVSNet_pl*. `https://github.com/kwea123/CasMVSNet_pl`. Accessed: 2024-04-23. 2020.

[29] Piotr Łabędź et al. "Histogram adjustment of images for improving photogrammetric reconstruction". In: *Sensors* 21.14 (2021), p. 4654.

[30] Edwin H Land. "The retinex theory of color vision". In: *Scientific american* 237.6 (1977), pp. 108–129.

[31] Tsung-Yi Lin et al. "Feature pyramid networks for object detection". In: (2017), pp. 2117–2125.

[32] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. "LLNet: A deep autoencoder approach to natural low-light image enhancement". In: *Pattern Recognition* 61 (2017), pp. 650–662.

[33] Keyang Luo et al. "P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo". In: (2019), pp. 10452–10461.

[34] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[35] Ziaur Rahman et al. "Structure revealing of low-light images using wavelet transform based on fractional-order denoising and multiscale decomposition". In: *The Visual Computer* 37.5 (2021), pp. 865–880.

[36] V Rajamani, P Babu, and S Jaiganesh. "A Review of various global contrast enhancement techniques for still images using histogram Modification Framework". In: *International Journal of Engineering Trends and Technology* 4.4 (2013), pp. 1045–1048.

[37] A Raji et al. "A gray-level transformation-based method for image enhancement". In: *Pattern Recognition Letters* 19.13 (1998), pp. 1207–1212.

[38] Wenqi Ren et al. "Low-light image enhancement via a deep hybrid network". In: *IEEE Transactions on Image Processing* 28.9 (2019), pp. 4364–4375.

[39] Ali M Reza. "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement". In: *Journal of VLSI signal processing systems for signal, image and video technology* 38 (2004), pp. 35–44.

[40] Andrea Romanoni and Matteo Matteucci. "Tapa-mvs: Textureless-aware patchmatch multi-view stereo". In: (2019), pp. 10413–10422.

[41] Riccardo Roncella et al. "Photogrammetric Digital Surface Model Reconstruction in Extreme Low-Light Environments". In: *Remote Sensing* 13.7 (2021).

[42] Fengxiang Rong et al. "A survey of multi view stereo". In: (2021), pp. 129–135.

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: (2015), pp. 234–241.

[44] Ivan W Selesnick, Richard G Baraniuk, and Nick C Kingsbury. "The dual-tree complex wavelet transform". In: *IEEE signal processing magazine* 22.6 (2005), pp. 123–151.

[45] Hyojoo Son, Changmin Kim, and Changwan Kim. "3D reconstruction of as-built industrial instrumentation models from laser-scan data and a 3D CAD database based on prior knowledge". In: *Automation in Construction* 49 (2015), pp. 193–200.

[46] Yang Su et al. "Zero-reference deep learning for low-light image enhancement of underground utilities 3d reconstruction". In: *Automation in Construction* 152 (2023), p. 104930.

[47] Hao Tang et al. "Low-Illumination image enhancement based on deep learning techniques: a brief review". In: 10.2 (2023), p. 198.

[48] Li Tao et al. "Low-light image enhancement using CNN and bright channel prior". In: (2017), pp. 3215–3219.

[49] AutoDL Team. *AutoDL Computation Platform*. `https://www.autodl.com/home`. 2024.

[50] Chun-Ming Tsai and Zong-Mu Yeh. "Contrast enhancement by automatic and parameter-free piecewise linear transformation for color images". In: *IEEE transactions on Consumer Electronics* 54.2 (2008), pp. 213–219.

[51] Sinh Van Nguyen et al. "Reconstruction of 3D digital heritage objects for VR and AR applications". In: *Journal of Information and Telecommunication* 6.3 (2022), pp. 254–269.

[52] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[53] Komal Vij and Yaduvir Singh. "Enhancement of images using histogram processing techniques". In: *Int. J. Comp. Tech. Appl* 2.2 (2009), pp. 309–313.

[54] Junyi Wang et al. "RDGAN: Retinex decomposition based adversarial learning for low-light enhancement". In: (2019), pp. 1186–1191.

[55] Yangang Wang and Qingfang Jiang. "LoliMVS: an End-to-end Network for Multi-view Stereo with Low-light Images". In: *IEEE Transactions on Instrumentation and Measurement* (2024).

[56] Yuesong Wang et al. "Deepfusion: A simple way to improve traditional multi-view stereo methods using deep learning". In: *Knowledge-Based Systems* 221 (2021), p. 106968.

[57] Yuzhi Wang et al. "Practical deep raw image denoising on mobile devices". In: (2020), pp. 1–16.

[58] Zhixin Wang et al. "DR2: Diffusion-Based Robust Degradation Remover for Blind Face Restoration". In: (2023), pp. 1704–1713.

[59] Chen Wei et al. "Deep retinex decomposition for low-light enhancement". In: *arXiv preprint arXiv:1808.04560* (2018).

[60] Yicheng Wu et al. "How to train neural networks for flare removal". In: (2021), pp. 2239–2247.

[61] Jingzhao Xu et al. "Illumination guided attentive wavelet network for low-light image enhancement". In: *IEEE Transactions on Multimedia* (2022).

[62] Jiayu Yang et al. "Cost volume pyramid based depth inference for multi-view stereo". In: (2020), pp. 4877–4886.

[63] Ling Yang et al. "Diffusion models: A comprehensive survey of methods and applications". In: *ACM Computing Surveys* 56.4 (2023), pp. 1–39.

[64] Ruigang Yang and Marc Pollefeys. "Multi-resolution real-time stereo on commodity graphics hardware". In: 1 (2003), pp. I–I.

[65] Wenhan Yang et al. "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement". In: (2020), pp. 3063–3072.

[66] Yao Yao et al. "Mvsnet: Depth inference for unstructured multi-view stereo". In: (2018), pp. 767–783.

[67] Yao Yao et al. "Recurrent mvsnet for high-resolution multi-view stereo depth inference". In: (2019), pp. 5525–5534.

[68] Chia-Hung Yeh and Min-Hui Lin. "Robust 3D reconstruction using HDR-based SLAM". In: *IEEE Access* 9 (2021), pp. 16568–16581.

[69] Hongwei Yi et al. "Pyramid multi-view stereo net with self-adaptive view aggregation". In: (2020), pp. 766–782.

[70] Xunpeng Yi et al. "Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model". In: (2023), pp. 12302–12311.

[71] Cheng Zhang et al. "Attention-based network for low-light image enhancement". In: (2020), pp. 1–6.

[72] Jingyang Zhang et al. "Vis-mvsnet: Visibility-aware multi-view stereo network". In: *International Journal of Computer Vision* 131.1 (2023), pp. 199–214.

[73] Zhe Zhang et al. "GeoMVSNet: Learning multi-view stereo with geometry perception". In: (2023), pp. 21508–21518.

[74] Hui Zhu, Francis HY Chan, and Francis K Lam. "Image contrast enhancement by constrained local histogram equalization". In: *Computer vision and image understanding* 73.2 (1999), pp. 281–290.

Appendices

A Related works

A.1 Discrete Wavelet Transformation

Discrete Wavelet Transformation (DWT) is a signal processing technique that decomposes a signal into its constituent frequency components, providing both time and frequency information. This method is particularly useful for analyzing non-stationary signals where frequency components vary over time.

The fundamental concept of DWT involves passing the signal through a series of high-pass and low-pass filters to create detailed and approximated coefficients. These coefficients represent the signal's high-frequency and low-frequency components, respectively.

Mathematically, the DWT of a signal $x(t)$ can be expressed as:

$$x(t) = \sum_{k=-\infty}^{\infty} c_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k}(t)$$

where $\phi_{j_0,k}(t)$ are the scaling functions, $\psi_{j,k}(t)$ are the wavelet functions, $c_{j_0,k}$ are the approximation coefficients, and $d_{j,k}$ are the detail coefficients.

The DWT operates by performing the following steps:

1. **Decomposition:** The signal is decomposed into approximate and detail coefficients using a series of high-pass and low-pass filters. This process can be recursively applied to the approximate coefficients to create a multi-level decomposition.

2. **Thresholding:** Coefficients are often thresholded to remove noise and irrelevant information, retaining significant signal components.

3. **Reconstruction:** The signal is reconstructed by combining the thresholded coefficients, effectively denoising or compressing the signal.

DWT is widely used in various applications, including signal denoising, image compression, and feature extraction in machine learning. Its ability to capture both time and frequency information makes it a powerful tool for analyzing complex signals and enhancing image details in low-light conditions.

Inverse Discrete Wavelet Transformation (IDWT) is the process of reconstructing the original signal from its wavelet coefficients. This involves reversing the steps of the DWT to transform the frequency domain information back into the time domain, thus reconstructing the original signal.

The IDWT can be mathematically expressed as:

$$x(t) = \sum_{k=-\infty}^{\infty} c_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k}(t)$$

where $\phi_{j_0,k}(t)$ are the scaling functions, $\psi_{j,k}(t)$ are the wavelet functions, $c_{j_0,k}$ are the approximation coefficients, and $d_{j,k}$ are the detail coefficients.

The steps involved in IDWT are:

1. **Reconstruction of Approximations and Details:** Using the inverse of the high-pass and low-pass filters applied in DWT, the approximation and detail coefficients are combined to reconstruct the signal at each level.

2. **Combining Coefficients:** The combined coefficients at each level are then iteratively combined to reconstruct the original signal.

A.2 Diffusion Models

The fundamental mechanism of diffusion models is an autoregressive process of adding and reversing noise. The forward diffusion process begins by gradually adding Gaussian noise to an image across a series of steps. This can be represented mathematically as:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

where x_t is the image at step t, α_t is a variance schedule over time, and ϵ is the noise vector. In the reverse diffusion process, the model aims to reconstruct the original image by estimating and subtracting the added noise iteratively. This is modeled by a neural network, typically a UNet, which predicts the noise that was added at each step:

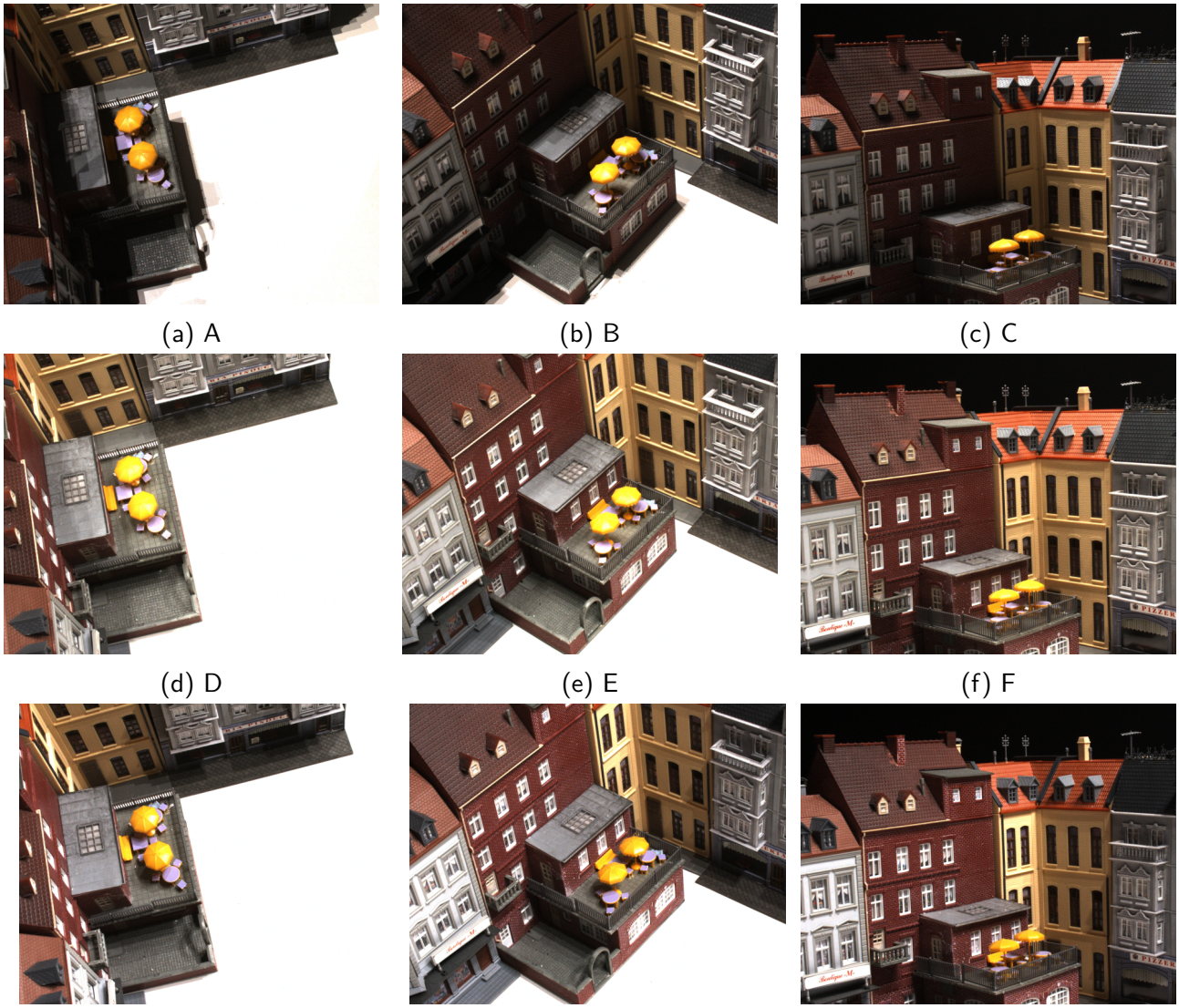
$$\hat{\epsilon} = f(x_t, t, \theta)$$

where f is the function approximated by the UNet, t is the timestep, and θ represents the model parameters. In reverse diffusion processes, the model generates images from noisy images at t step to $t-1$ step:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t^2}} \hat{\epsilon}(x_t, t) \right)$$

As long as timesteps are large enough, the image at timestep t is approximate to Gaussian noise. The diffusion model can generate images from noise.

B Image Enhancement

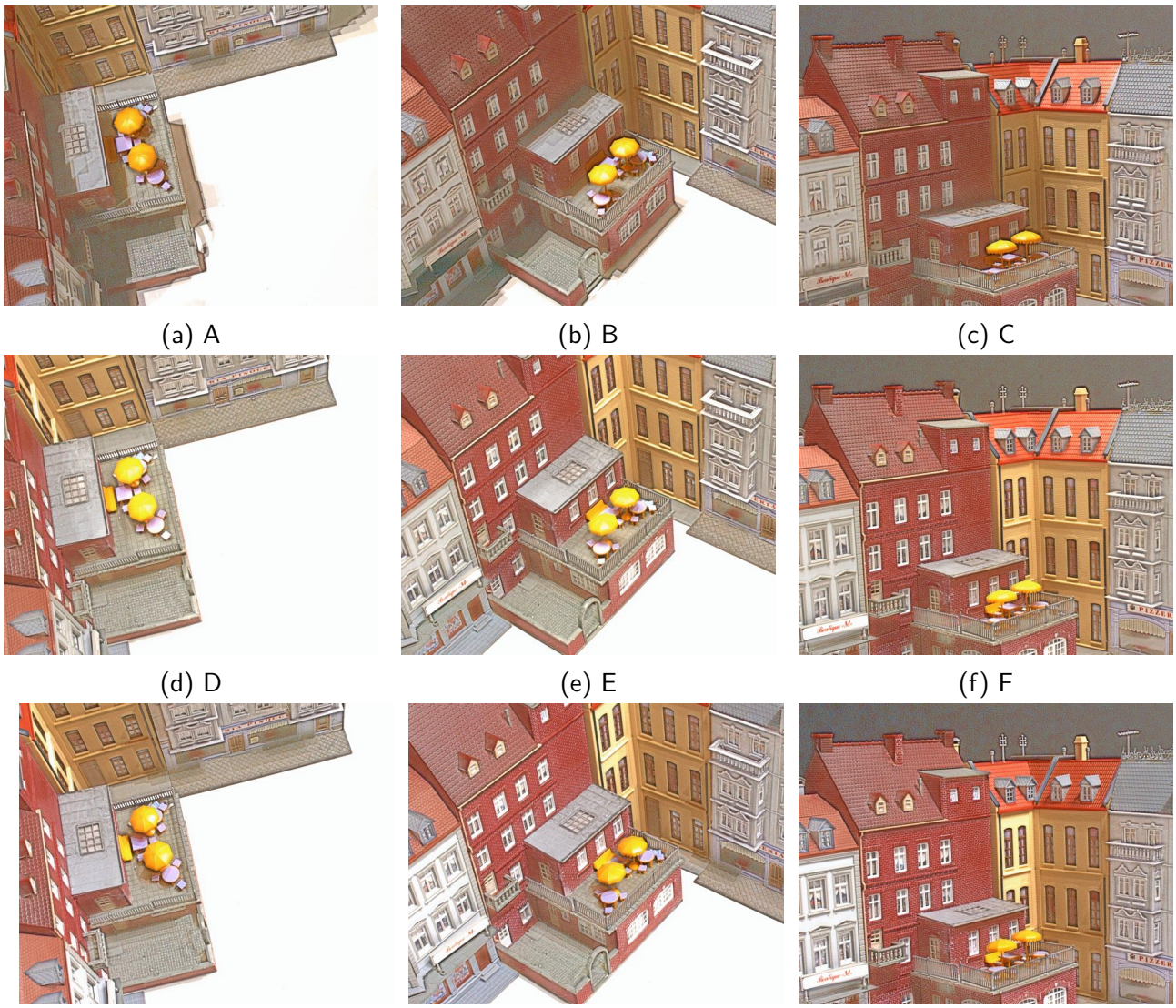


(a) A (b) B (c) C

(d) D (e) E (f) F

(g) D (h) E (i) F

Figure 1: The image input to image enhancement model in scan 23 of DTU, from top to bottom is image in lighting condition 0,3,6

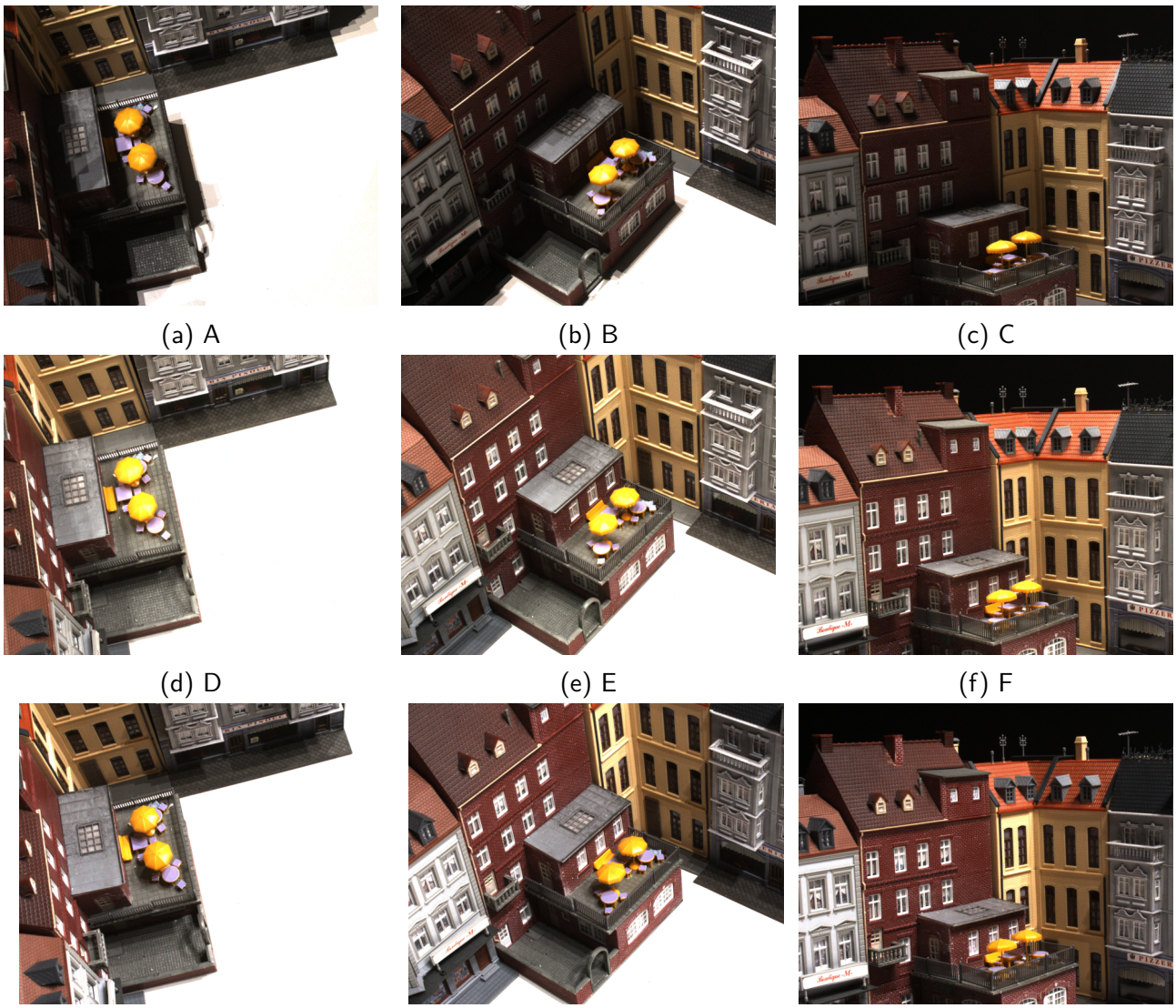


(a) A (b) B (c) C

(d) D (e) E (f) F

(g) G (h) H (i) I

Figure 2: Enhanced image using RetinexNet in scan 23 of DTU, from top to bottom is the image in lighting condition 0,3,6

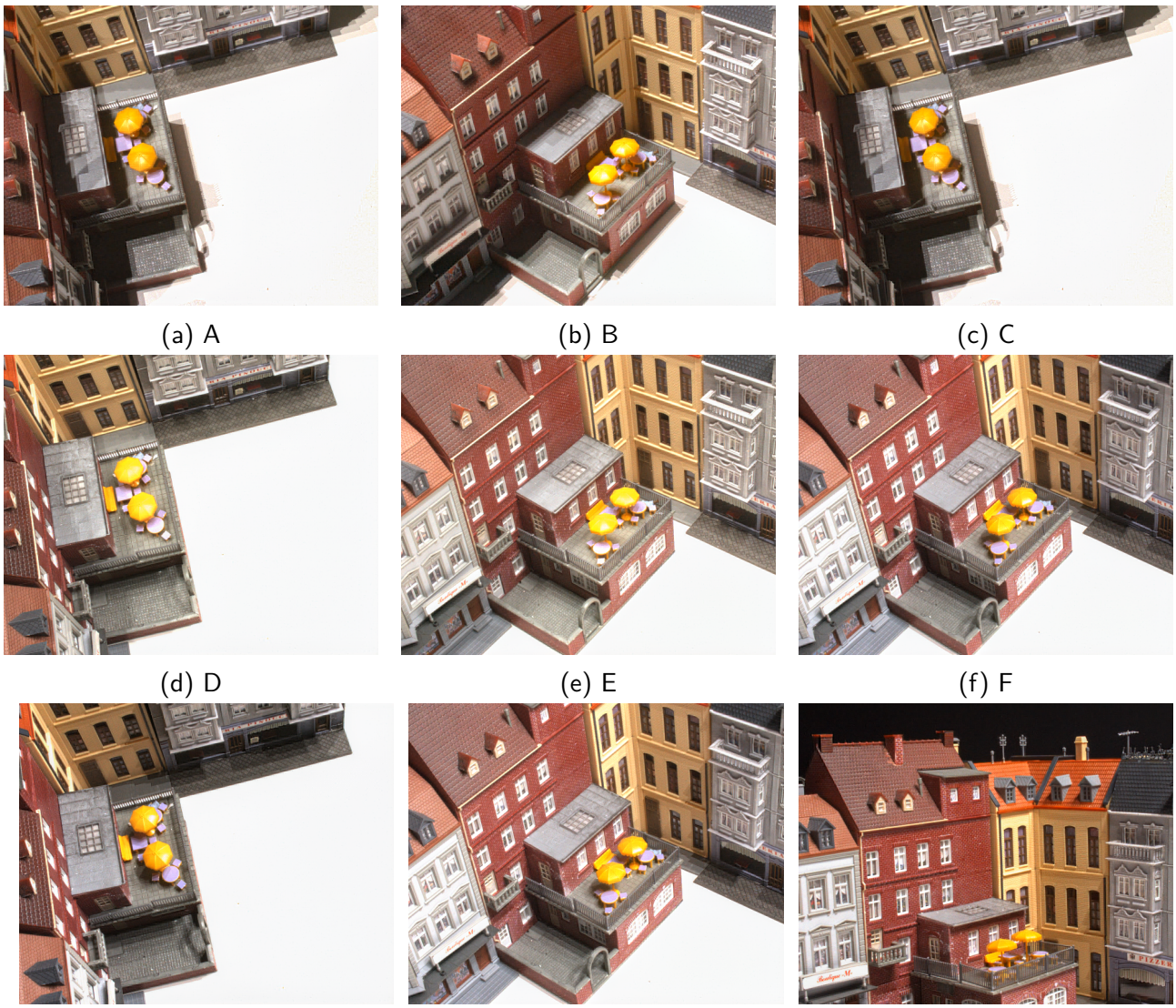


(a) A (b) B (c) C

(d) D (e) E (f) F

(g) G (h) H (i) I

Figure 3: Enhanced image using EnlightenGAN Diffusion in scan 23 of DTU, from top to bottom is the image in lighting condition 0,3,6

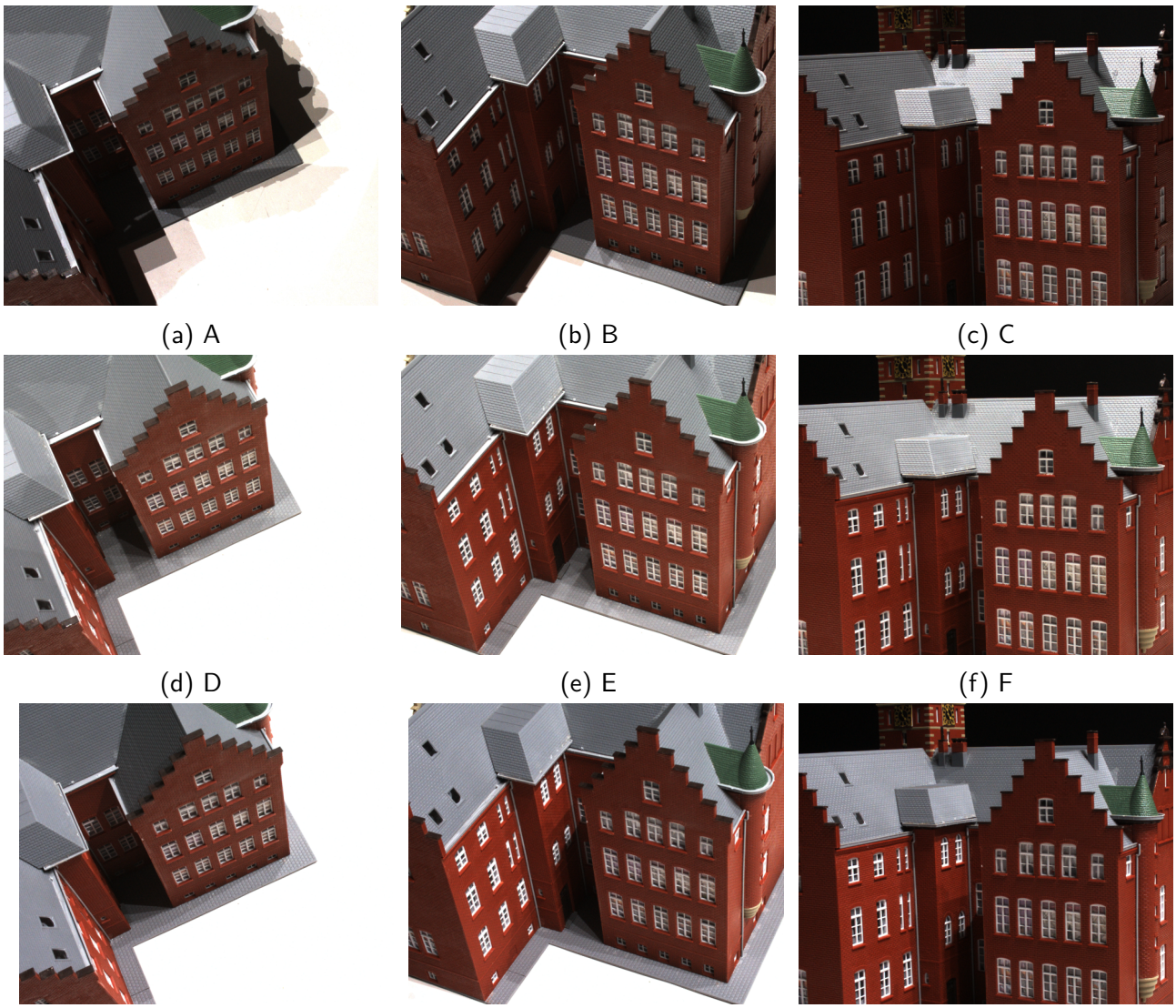


(a) A (b) B (c) C

(d) D (e) E (f) F

(g) G (h) H (i) I

Figure 4: Enhanced image using Low-light Diffusion in scan 23 of DTU, from top to bottom is the image in lighting condition 0,3,6

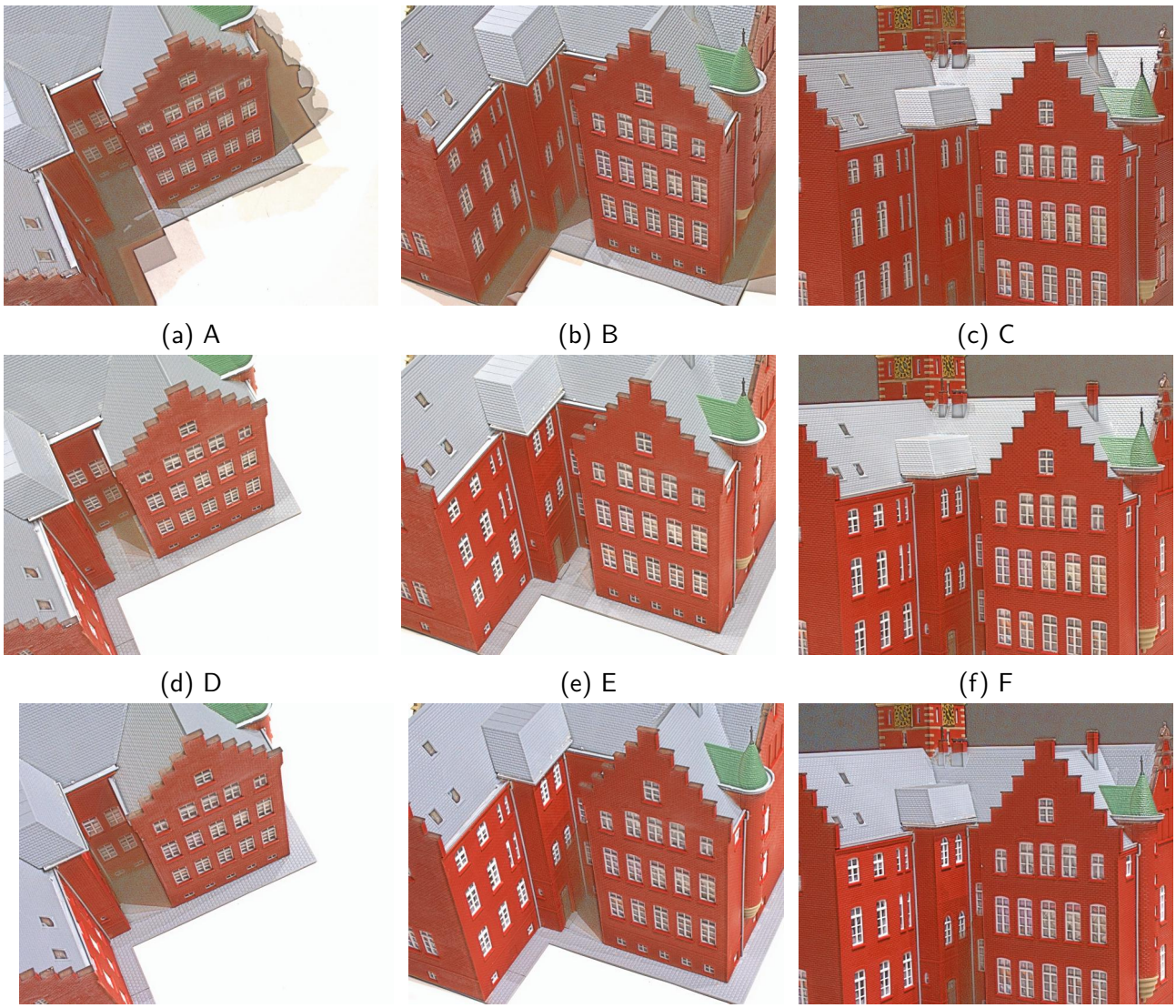


(a) A (b) B (c) C

(d) D (e) E (f) F

(g) D (h) E (i) F

Figure 5: The image input to the image enhancement model in scan 23 of DTU, from top to bottom, is the image in lighting conditions 0,3,6

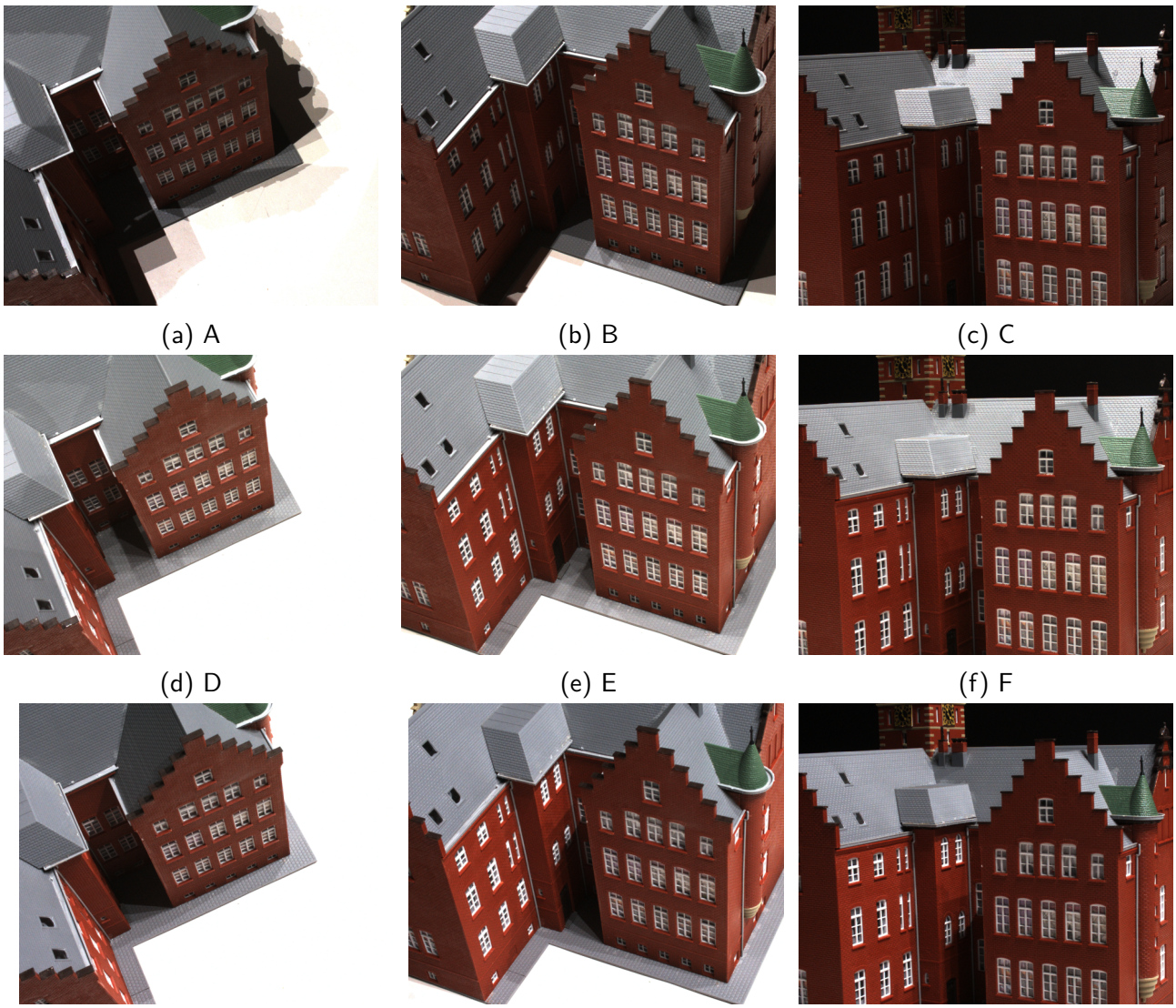


(a) A (b) B (c) C

(d) D (e) E (f) F

(g) G (h) H (i) I

Figure 6: Enhanced image using RetinexNet in scan 23 of DTU, from top to bottom is the image in lighting condition 0,3,6

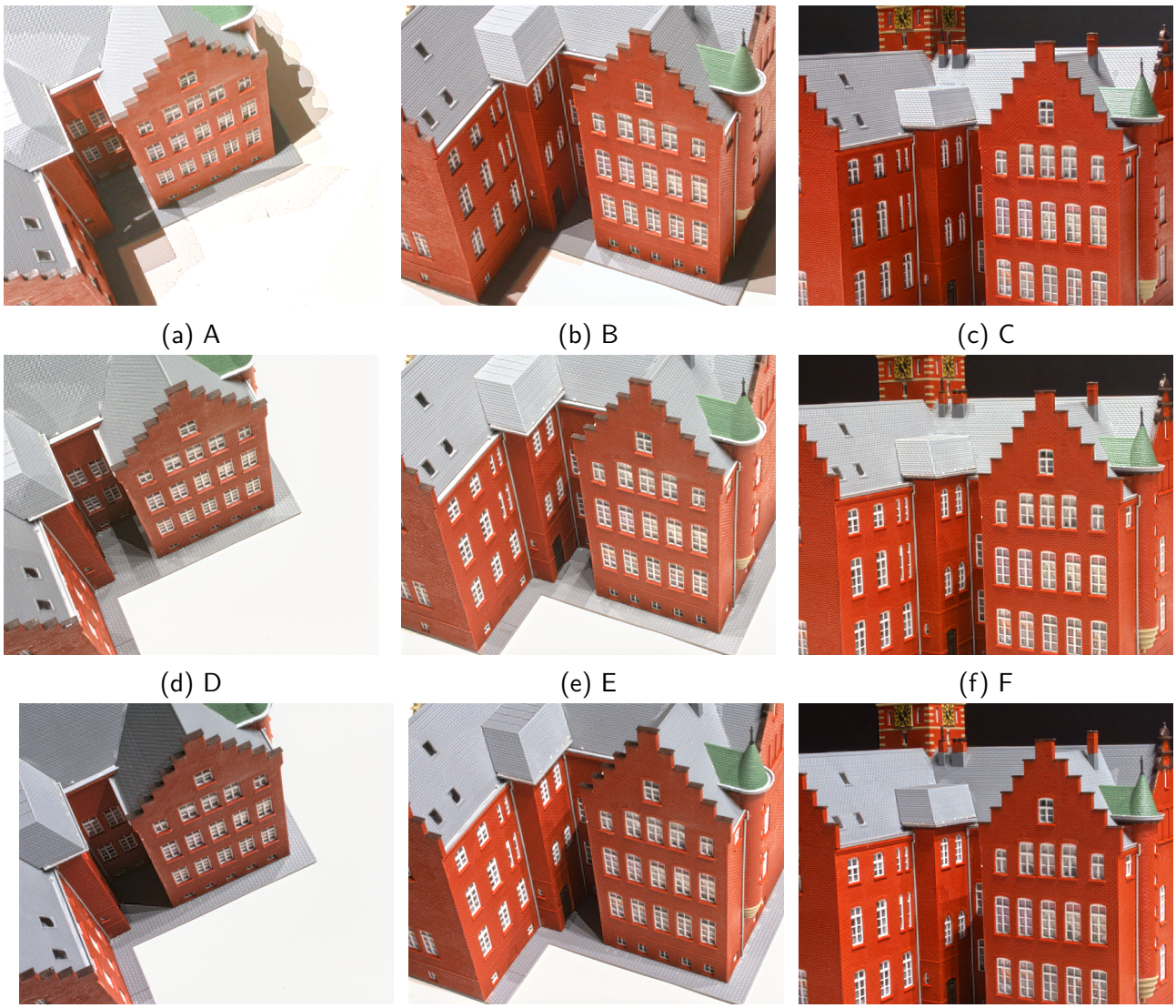


(a) A (b) B (c) C

(d) D (e) E (f) F

(g) G (h) H (i) I

Figure 7: Enhanced image using EnlightenGAN Diffusion in scan 23 of DTU, from top to bottom is the image in lighting condition 0,3,6



(a) A (b) B (c) C

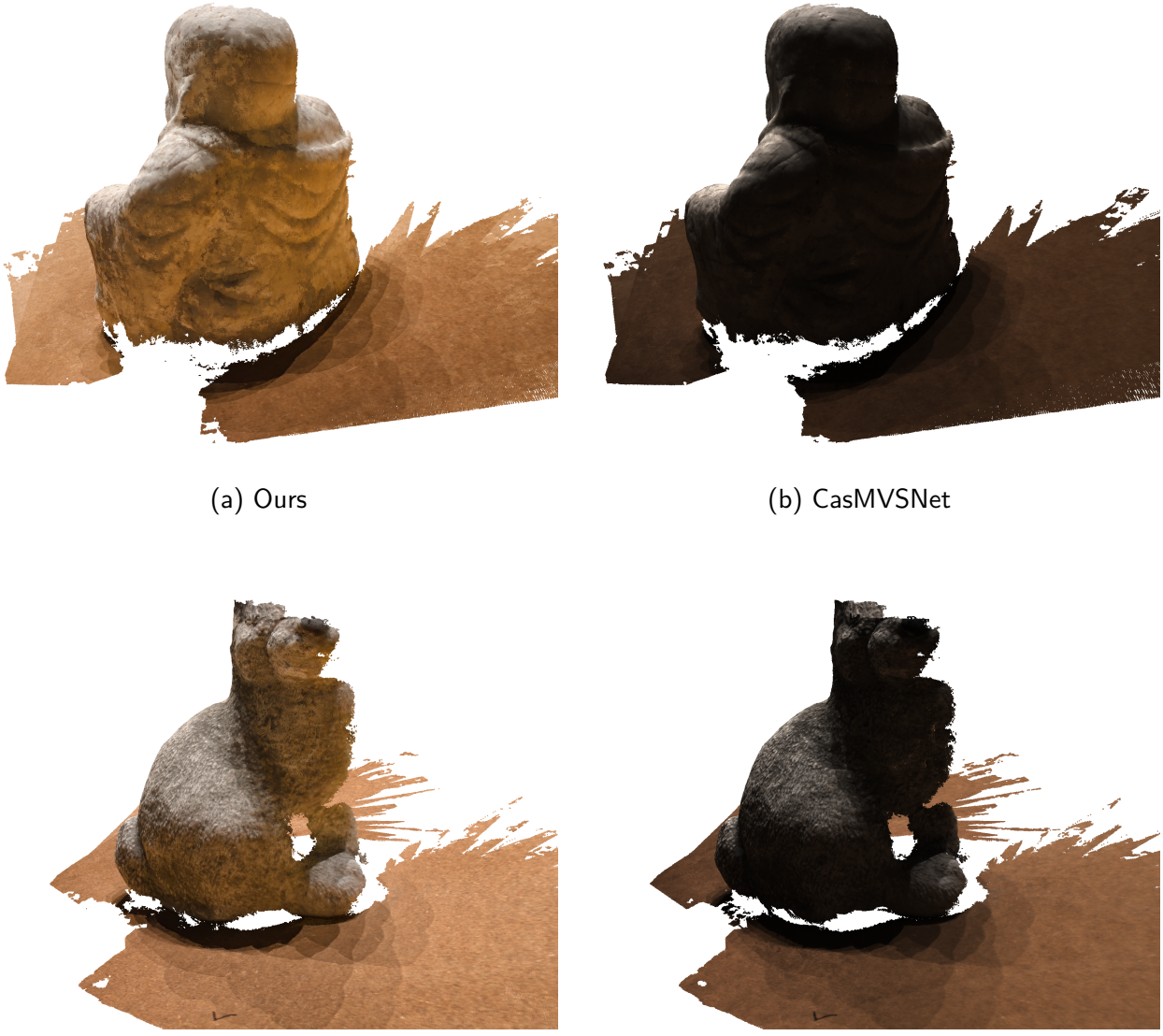
(d) D (e) E (f) F

(g) G (h) H (i) I

Figure 8: Enhanced image using Low-light Diffusion in scan 23 of DTU, from top to bottom is the image in lighting condition 0,3,6

C Qualitative result of 3D reconstruction

C.1 DTU: color



(a) Ours (b) CasMVSNet

(c) Ours (d) CasMVSNet

Figure 9: Color enhancement: scan 117 and scan 56

(a) Ours

(b) CasMVSNet

(c) Ours

(d) CasMVSNet

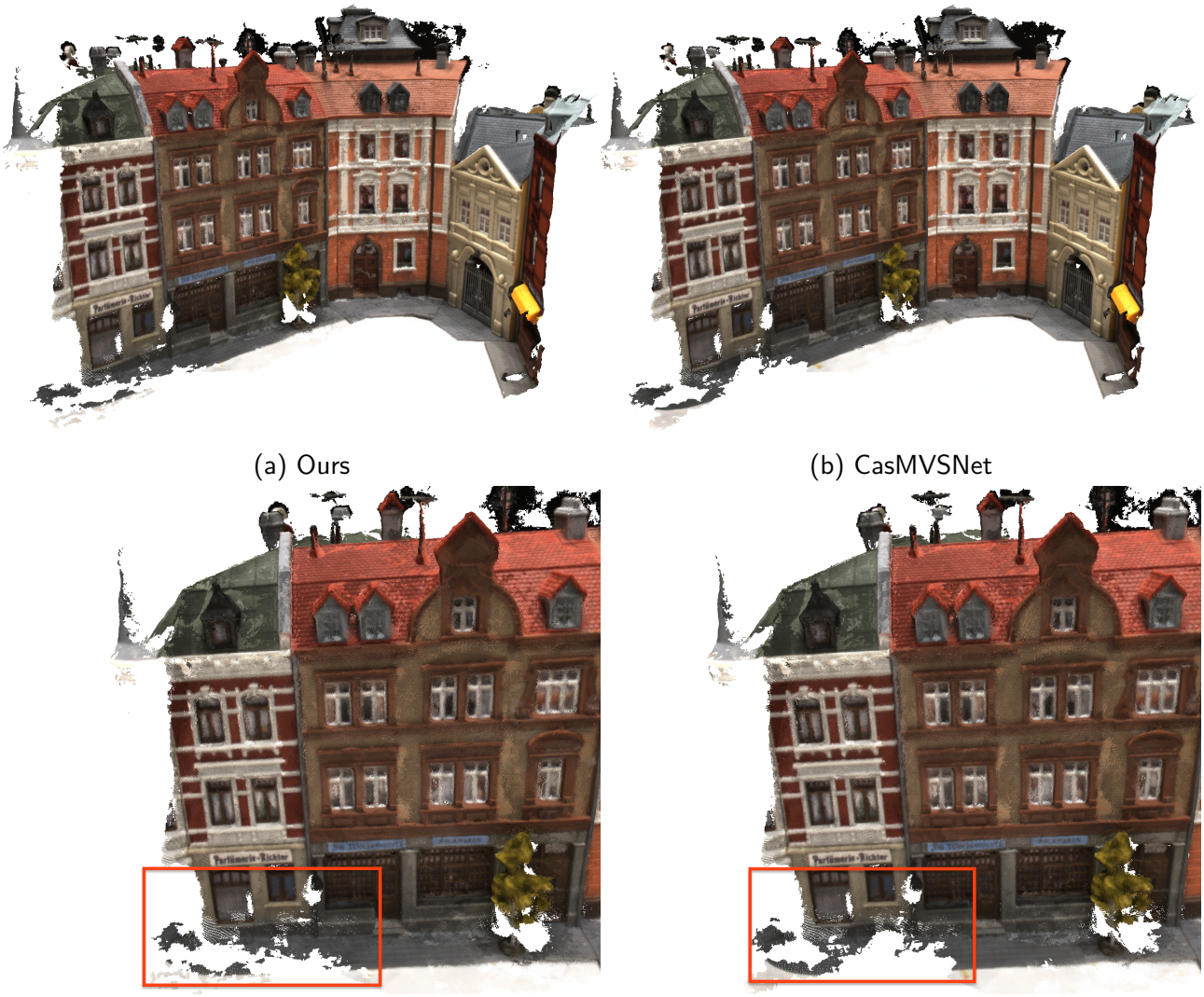
Figure 10: Color enhancement: scan 38 and scan 40

(a) Ours (b) CasMVSNet

(c) Ours (d) CasMVSNet

Figure 11: Color enhancement: scan 17 and scan 43

C.2 DTU: geometric



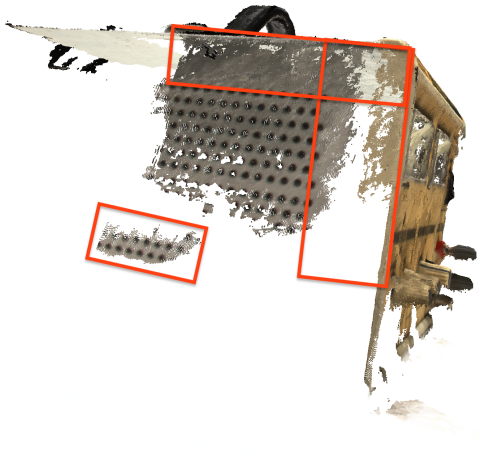
(a) Ours (b) CasMVSNet

(c) Ours (d) CasMVSNet

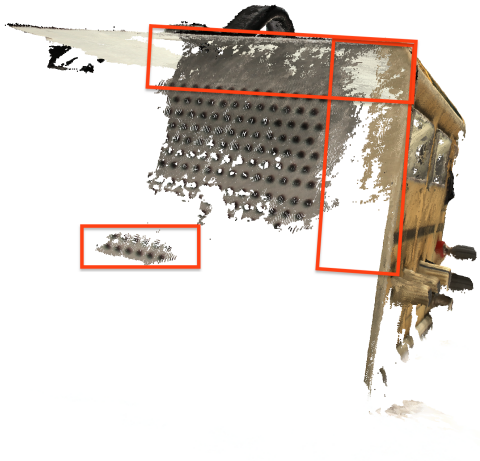
Figure 12: Qualitative results on scan 17

(a) Ours

(b) CasMVSNet



(c) Ours



(d) CasMVSNet

Figure 13: Qualitative results on scan 11

(a) Ours

(b) CasMVSNet

(c) Ours

(d) CasMVSNet

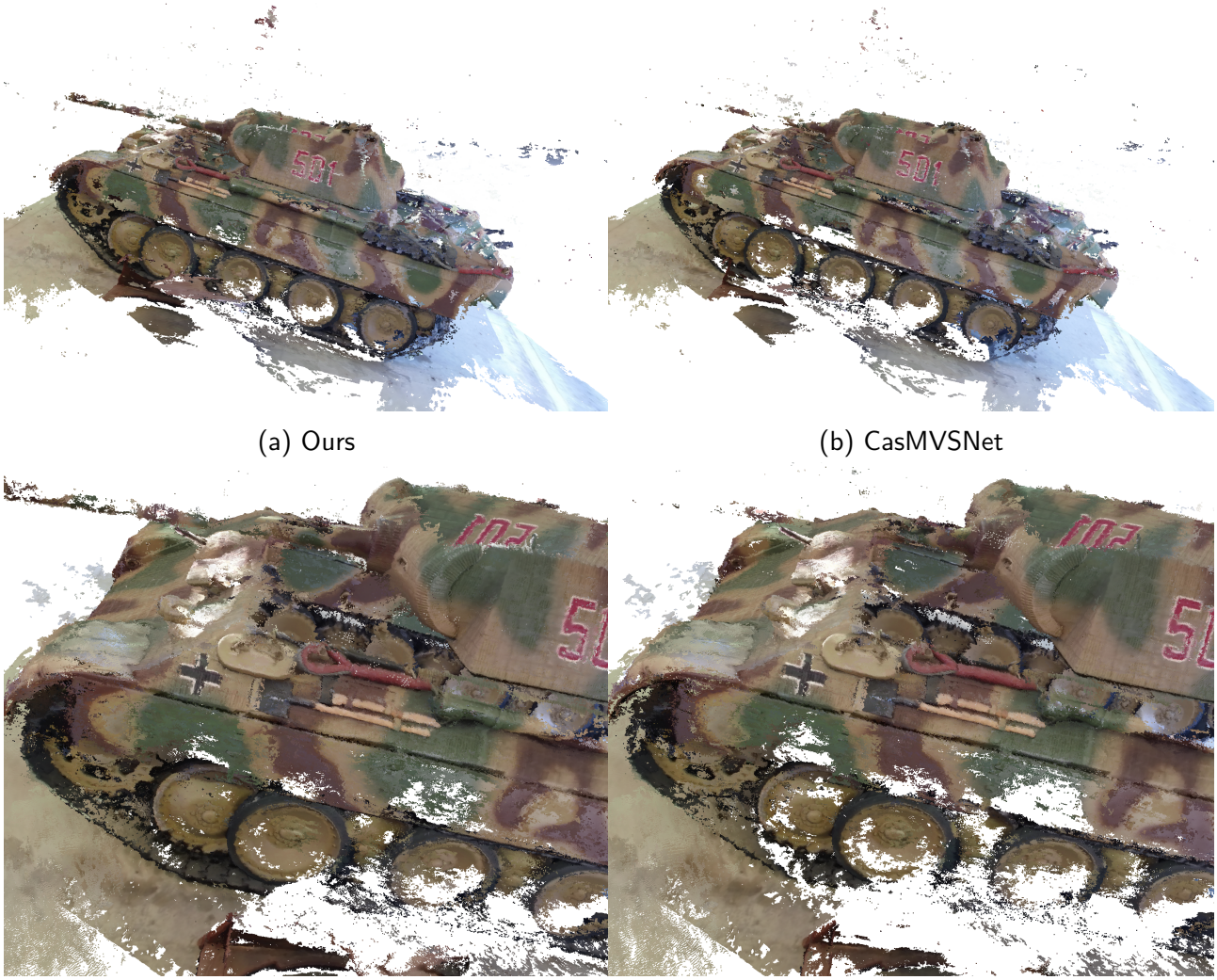
Figure 14: Qualitative results on scan 11

C.3 Tanks and Temples

(a) Ours (b) CasMVSNet

(c) Ours (d) CasMVSNet

Figure 15: Qulitative evaluation of on 'tanks and temples,' the scan name: M60



(a) Ours (b) CasMVSNet

(c) Ours (d) CasMVSNet

Figure 16: Qulitative evaluation of on 'tanks and temples', the scan name: Panther