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

June 27th

Zhang Chi 5740568

Supervisor:

Nail Ibrahimli

Nan Liangliang

TUDelft

Contents

1. Introduction
2. Research Question
3. Methodology & Result
4. Conclusion

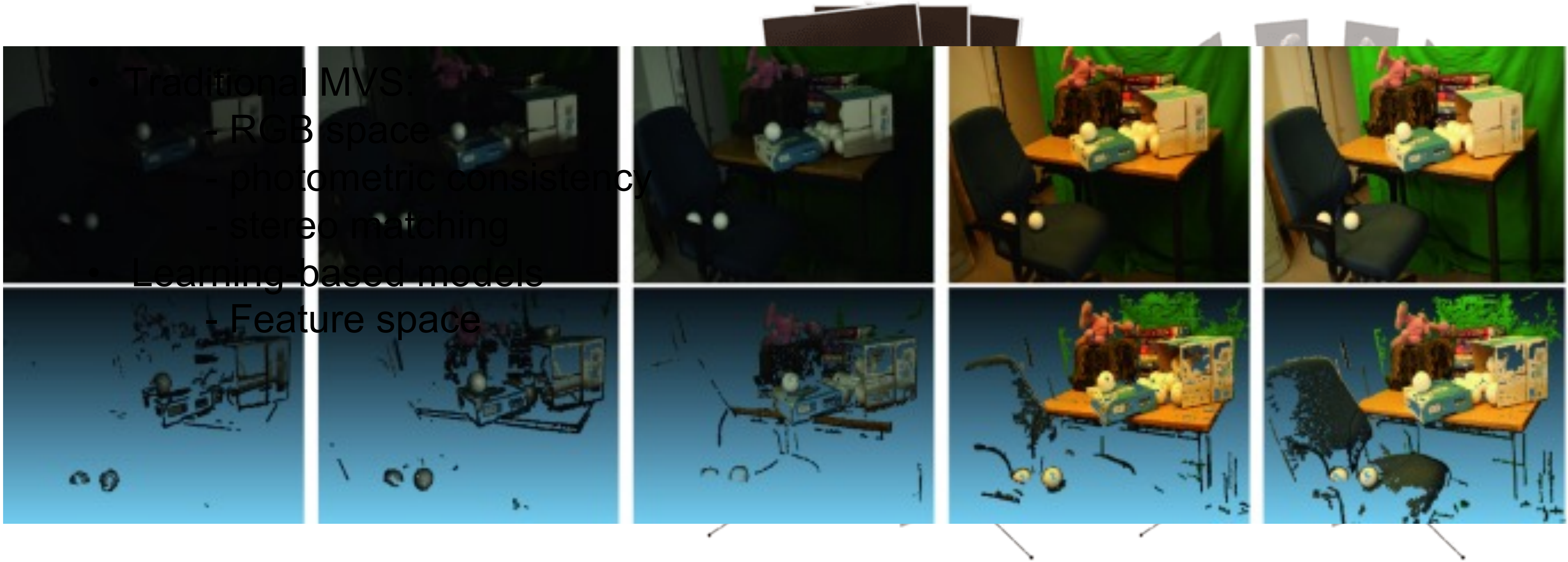
TU Delft

Introduction: Application

- Historical heritage protection
- Gaming and animation
- AR/VR
- BIM

Source: Kargas, A., Loumos, G., & Varoutas, D. (2019). Using different ways of 3D reconstruction of historical cities for gaming purposes: The case study of Nafplio. *Heritage*, *2*(3).

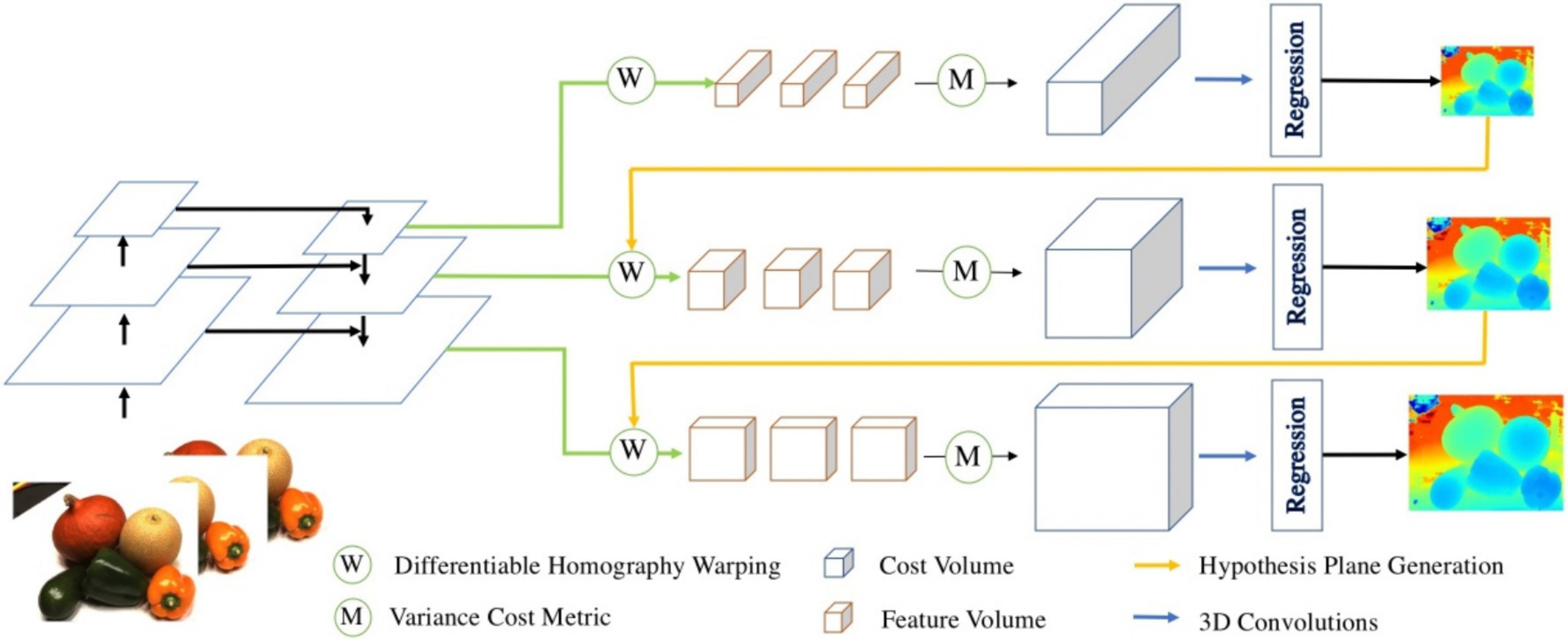
Introduction: MVS



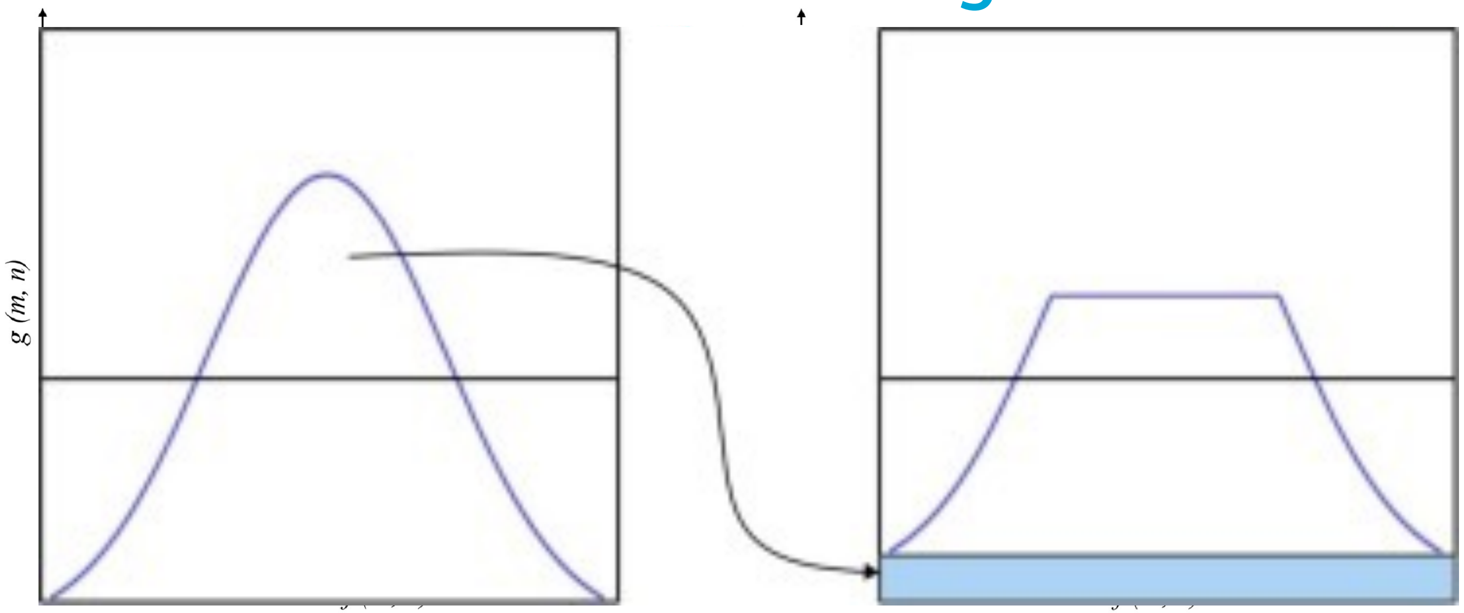
- Traditional MVS:
  - RGB space
  - photometric consistency
  - stereo matching
- Learning-based models
  - Feature space

Source: Weak Illumination Using Visibility-Enhanced LDR Imagery. In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1* (pp. 515-534). Springer International Publishing.
Furukawa, Y., & Hernández, C. (2015). Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, *9*(1-2), 1-148.

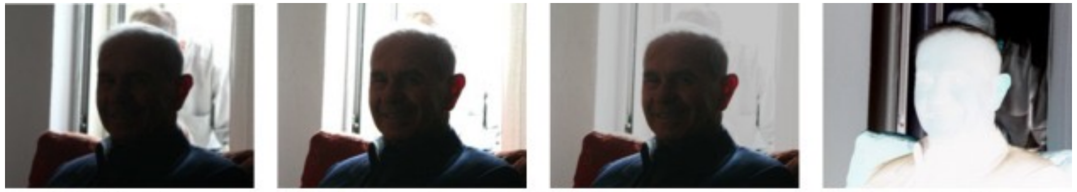
Introduction: Learning-based MVS



Related Work: Traditional Image Enhancement



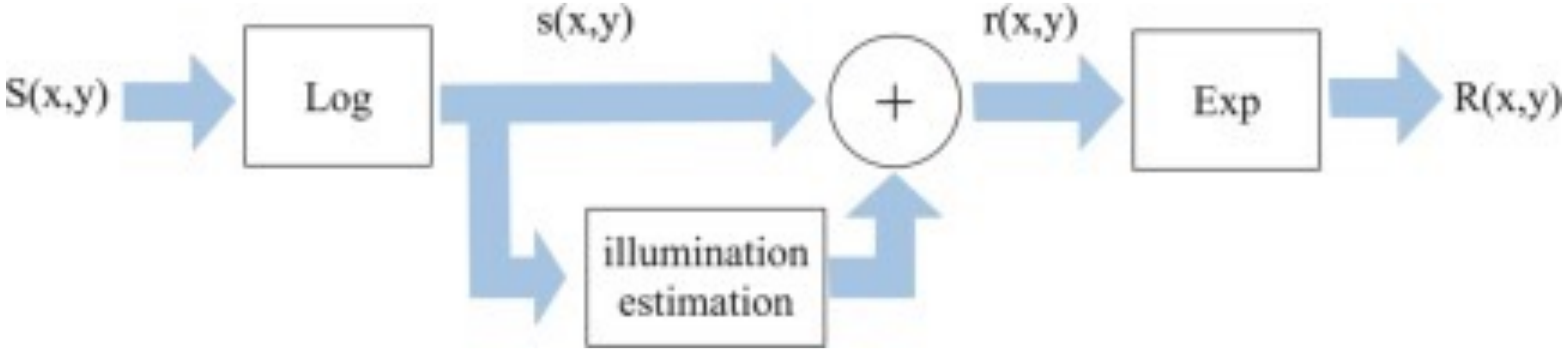
(a) Expand (a) Logarithmic transformation (b) Gamma transformation s
others static. the others.



(a) Original image (b) Linear transformation (c) Piece-wise transformation (d) Reverse transformation

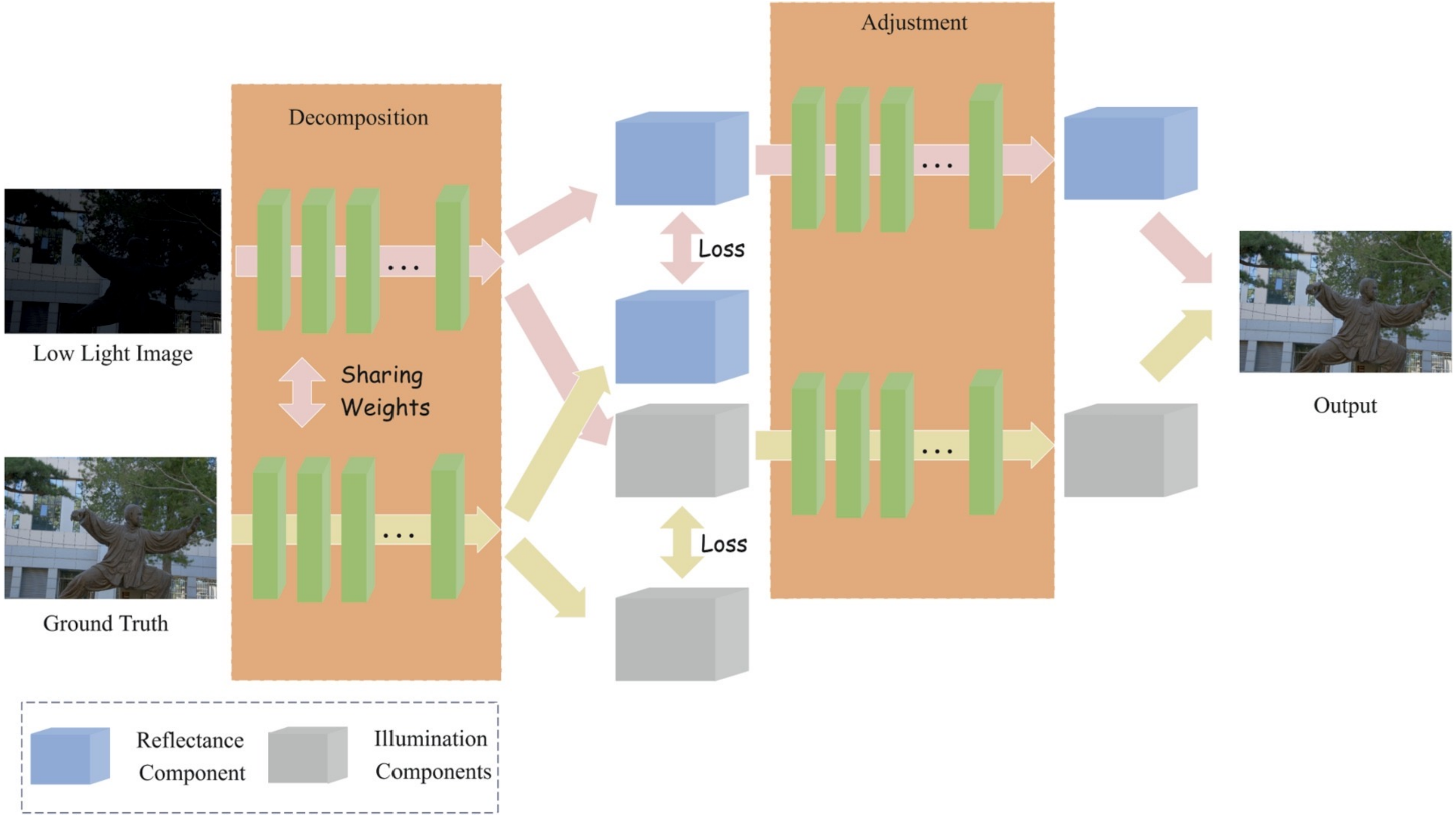
Gray Level Enhancement

Related Work: Traditional Image Enhancement



Retinex Theory

Related Work: Traditional Image Enhancement



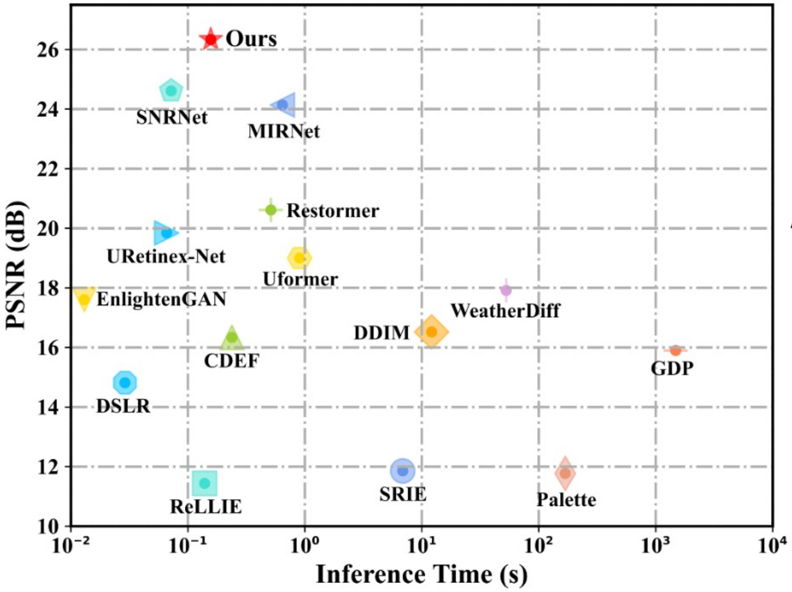
TU Delft

Related Work: Low-light Diffusion

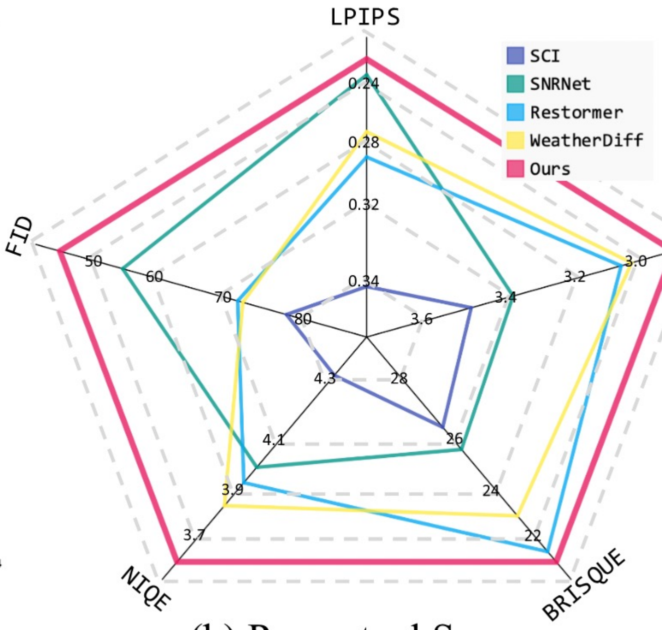
A_{high}^{k-1}

A_{low}^{k-1}

Source; Jiang, H., Luo, A., Fan,
(TOG), *42*(6), 1-14.



(a) Performance vs. Efficiency



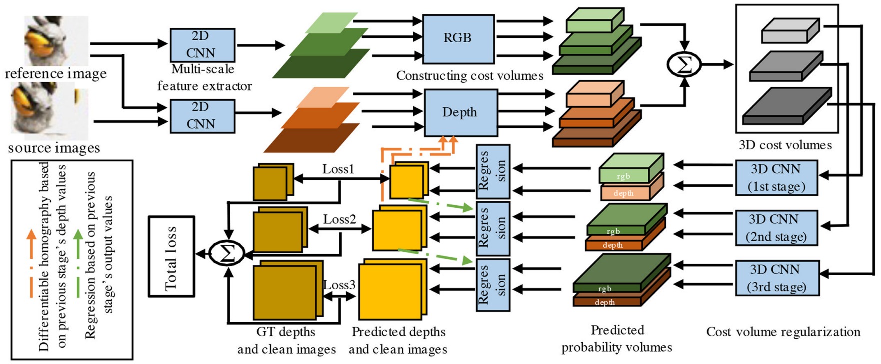
(b) Perceptual Scores

diffusion
denoising
nection

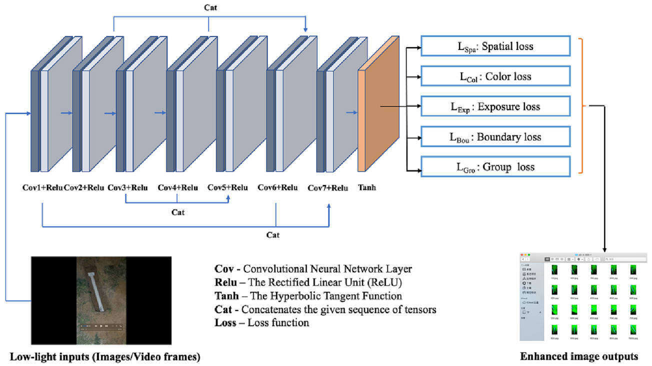
$\hat{A}_{low}^{k-1}$

Related Work: MVS in low illumination

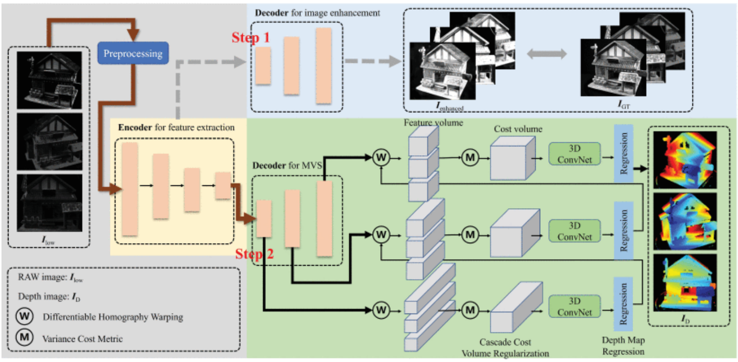
Method Name	Dataset	Image dnhancement model	MVS
DeMVS	Paired	Not pre-trained	MVSNet & CasMVSNet
LoliMVS	Paired	Not pre-trained	CasMVSNet
ZDE3D	Unpaired	Not pre-trained	Colmap



DeMVS



ZDE3D



LoliMVS

Sources: Han, J., Chen, X., Zhang, Y., Hou, W., & Hu, Z. (2022). DEMVSNet: Denoising and depth inference for unstructured multi-view stereo on noised images. *IET Computer Vision*, *16*(7), 570-580.
Wang, Y., & Jiang, Q. (2024).

Sources: Su, Y., Wang, J., Wang, X., Hu, L., Yao, Y., Shou, W., & Li, D. (2023). Zero-reference deep learning for low-light image enhancement of underground utilities 3d reconstruction. *Automation in Construction*, *152*, 104930.

Source: LoliMVS: an End-to-end Network for Multi-view Stereo with Low-light Images. *IEEE Transactions on Instrumentation and Measurement.*

TUDelft

Related Work: MVS in low illumination

Method Name	Dataset		Acc	Acc changes	Comp	Comp Changes
DeMVS	Paired	CasMVSNet	0.6140	-0.0349	0.4075	+0.0051
		DeMVS	0.5791		0.4024	
LoliMVS	Paired	CVP-MVSNet	0.365	- 0.164	0.787	+0.008
		LoliMVS	0.201		0.795	

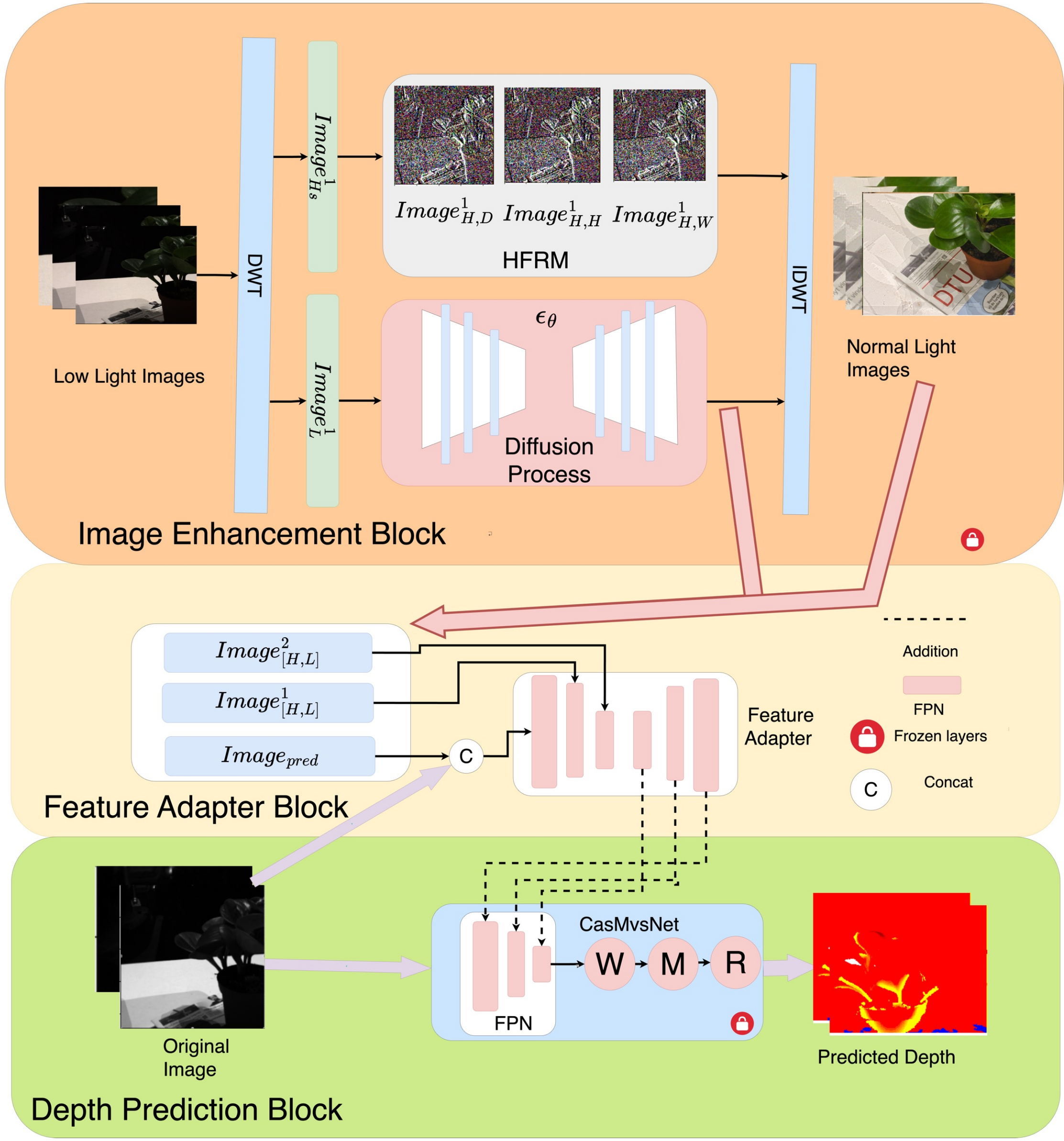
Research Question:

To what extent can existing single-frame image enhancement models be utilized to enhance the performance of MVS in low illumination conditions?

1. Which image enhancement model is suitable?
2. Which architecture is suitable for integrating the image enhancement model with MVS?
3. How to reduce the computation resource demands

Methodology & Result: Overview

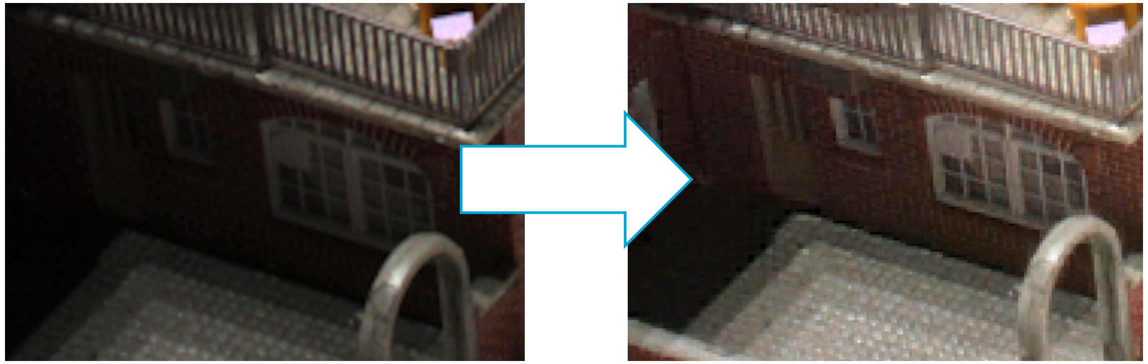
1. Image Enhancement Model
2. Feature Adapter
3. Learning-based MVS

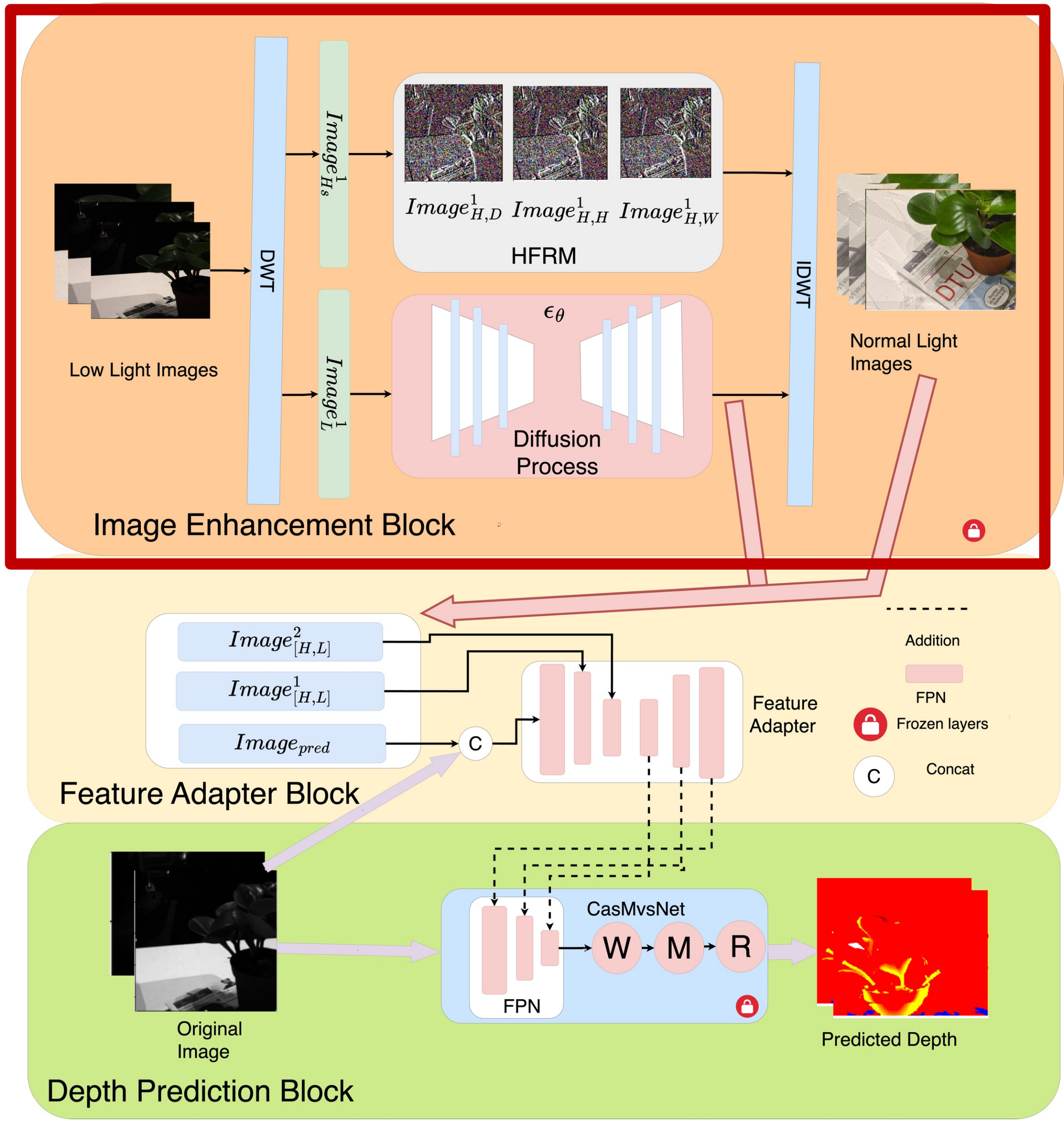


Methodology & Result : Diffusion-based Image Enhancement

Low-light Diffusion

- Use DWT to decompose image
- Diffusion on low-frequency parts
- Attention blocks to reinforce high-frequency parts





Methodology & Result : Multi-Frame Attention

Multi-view Consistency

- Similar to View-diffusion(2024)

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

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:

$$Q_n = W_q \times \text{Image}_H^n \tag{3.2}$$

$$K_n = W_k \times \text{Image}_H^n \tag{3.3}$$

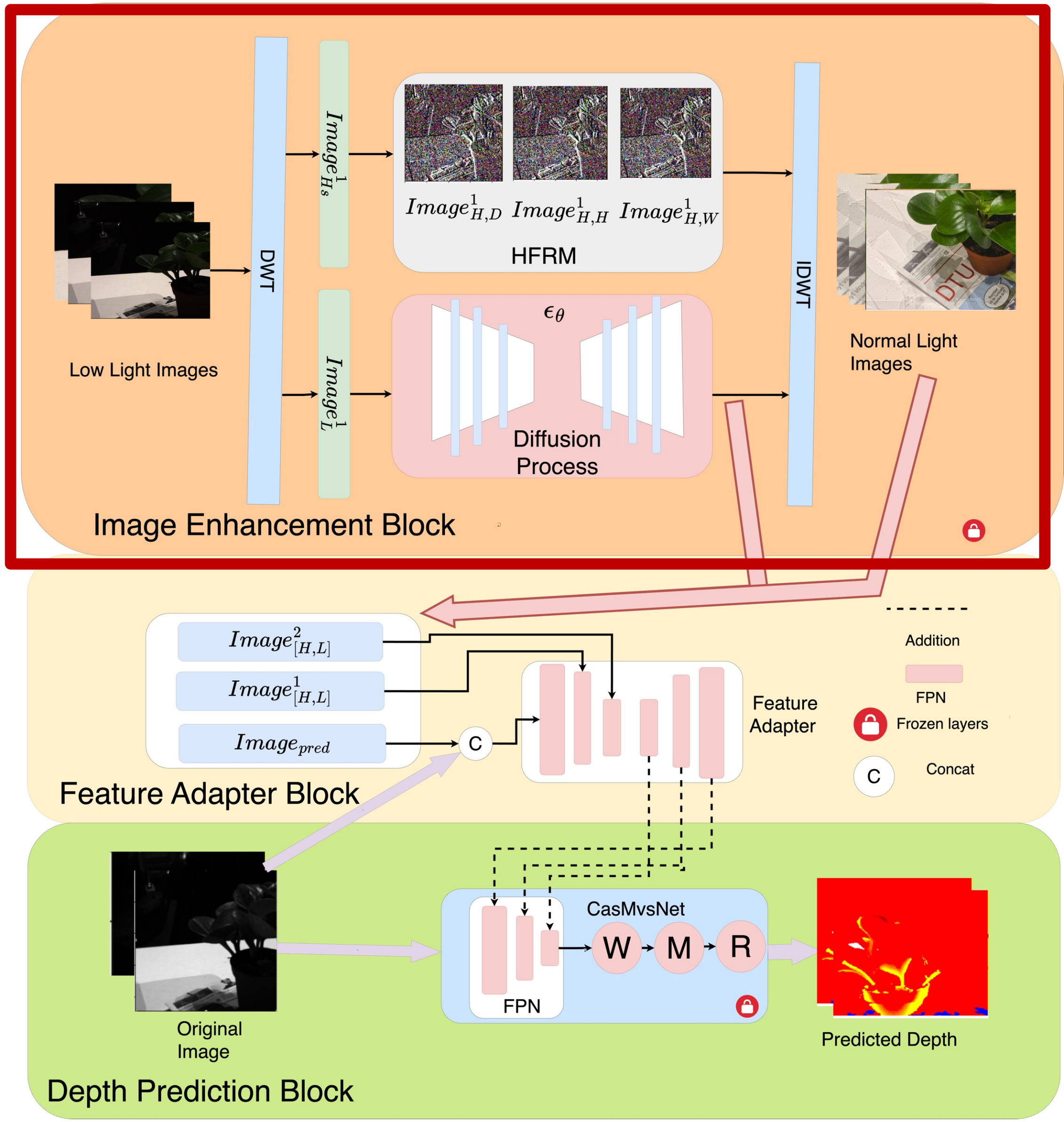
$$V_n = W_v \times \text{Image}_H^n \tag{3.4}$$

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:

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

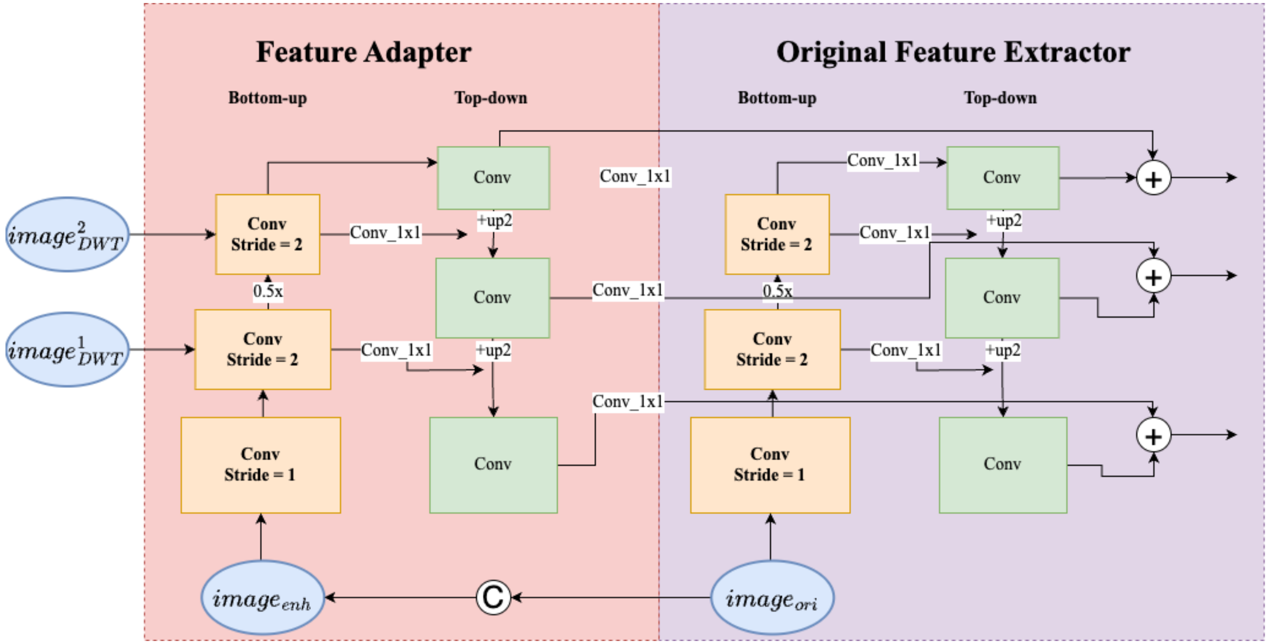
$$V_n = W_v \times [\text{Image}_H^{n-1}; \text{Image}_H^n; \text{Image}_H^{n+1}] \tag{3.6}$$

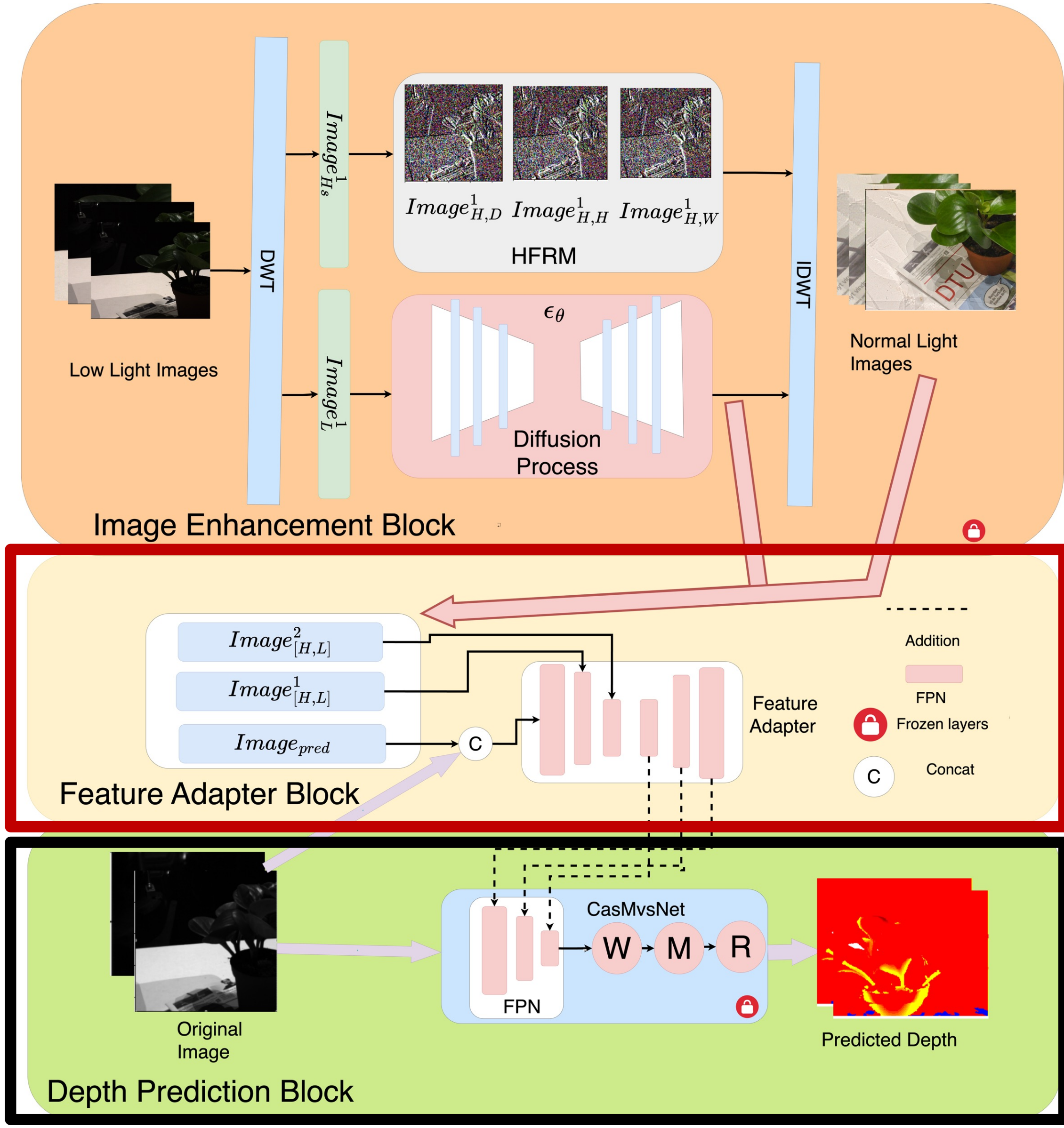


TU Delft

Methodology: Feature Adapter

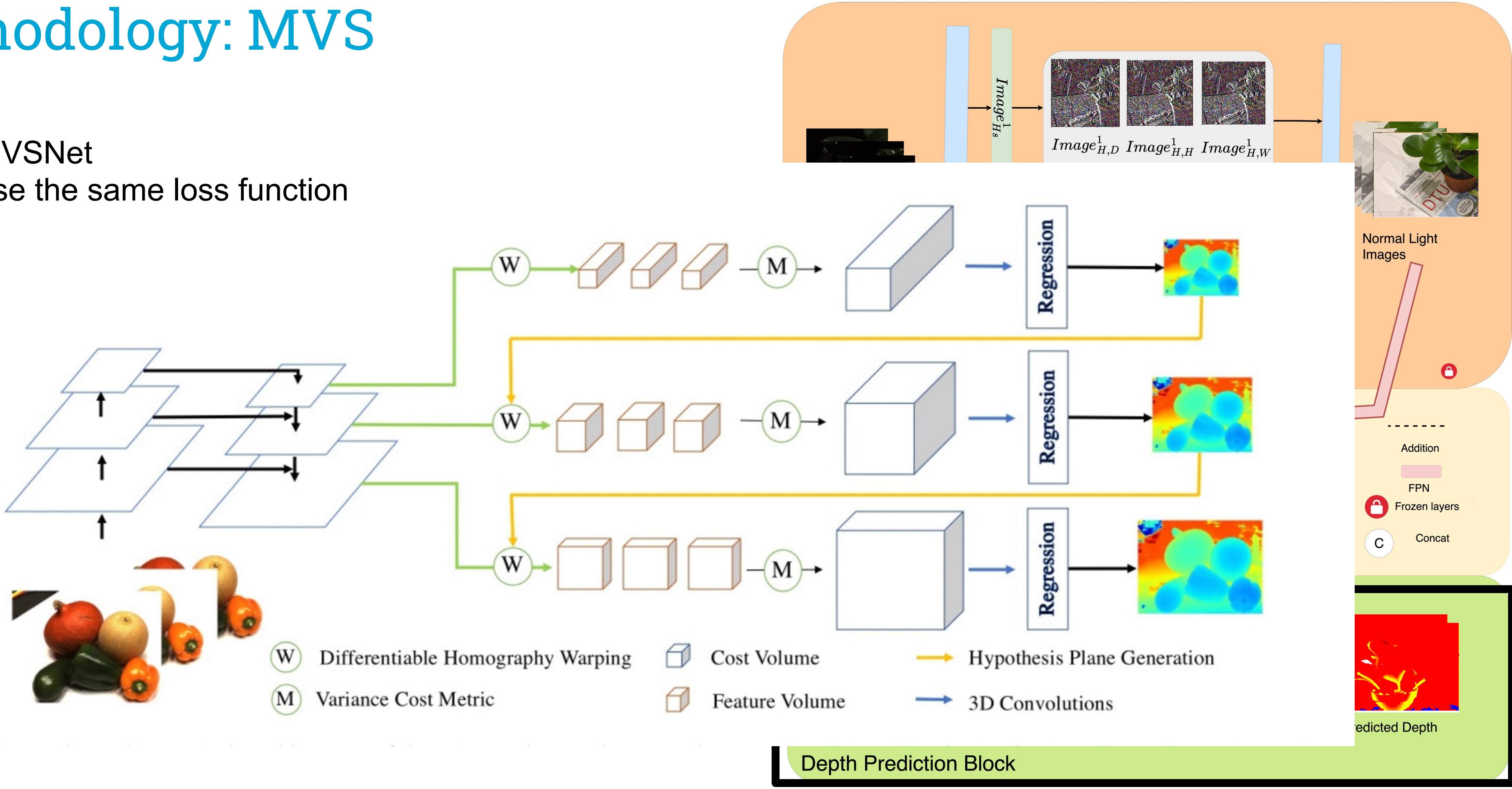
- Refine Feature Map
- Convert to feature space
- Improve efficiency and effectiveness





Methodology: MVS

- CasMVSNet
- We use the same loss function



Methodology : Depth filtering, depth fusion and evaluation metrics

Depth filtering:
1. Geometric Consistencies
   - Pixel Consistency
   - Depth Consistency
2. Confidence Threshold
   - > 0.999

Depth fusion
1. Depth Value
   - Average across multi-view images
2. RGB Value:
   - The most frequent value encountered

Depth:
1. Mean Absolute error
2. Accuracy Threshold
   - Ratio of MAE below threshold

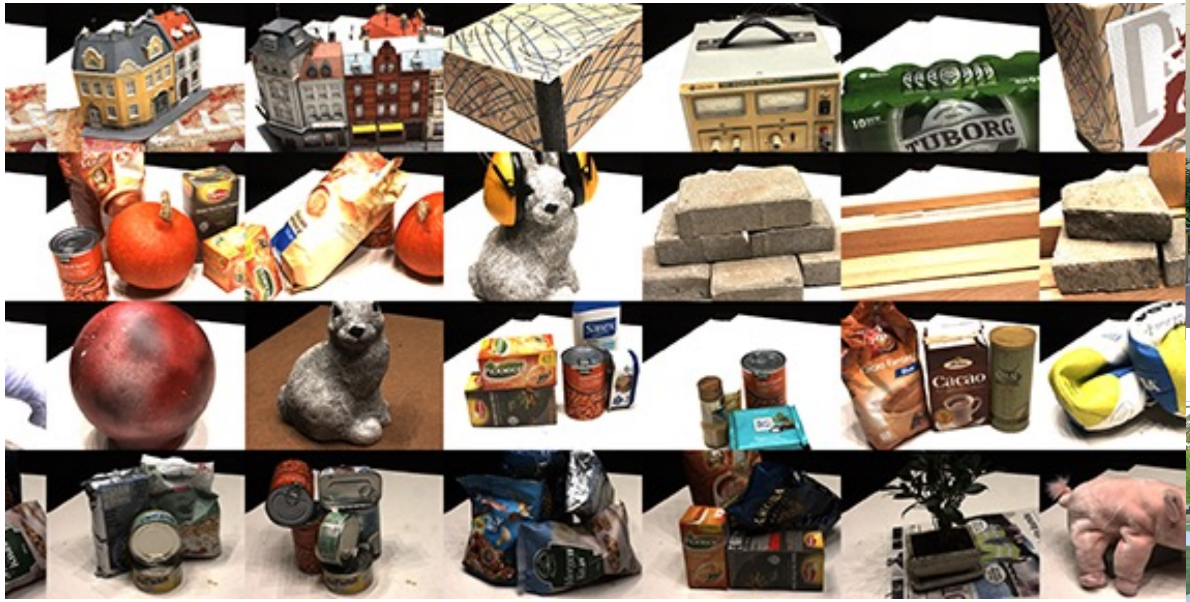
3D geometrics:
1. Accuracy
   - prediction to GT
2. Completeness
   - GT to prediction
3. Overall
   - Average

TU Delft

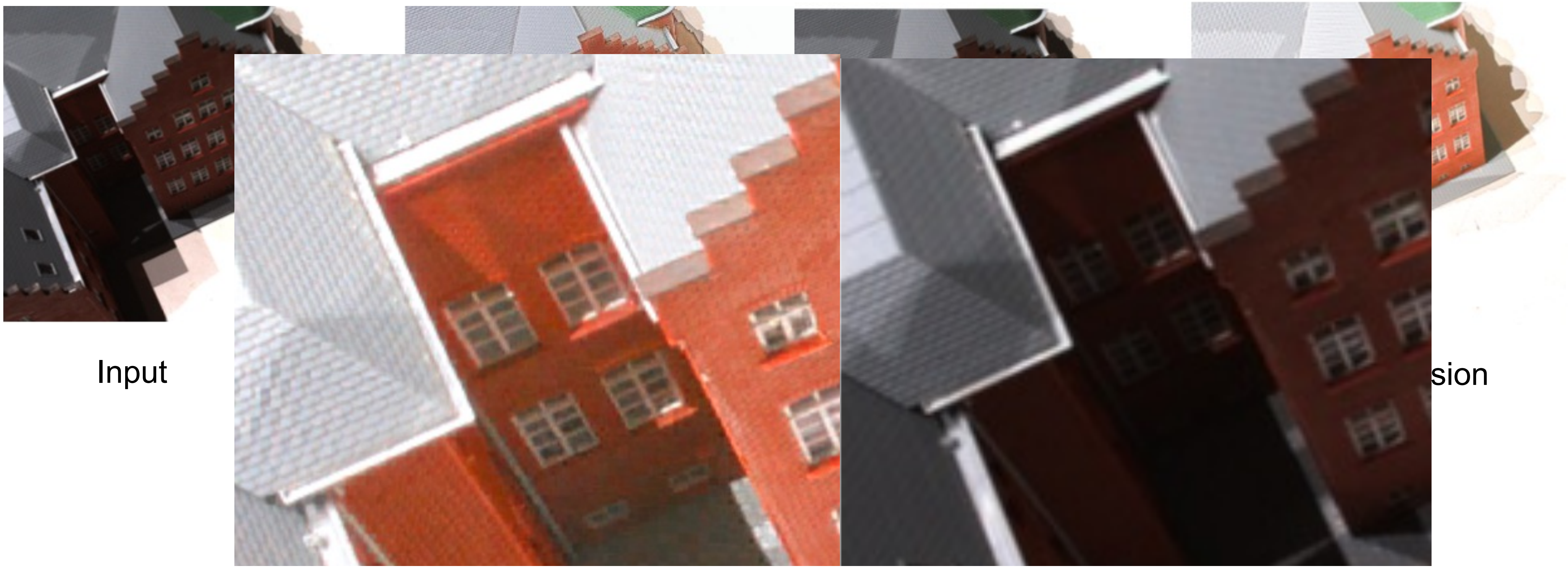
Experiment & Result: Dataset

DTU dataset: 119 scan, 79 for train, 18 for validation, 22 for test

Tanks and temples: intermediate, advanced and training



Experiment & Result: Image Enhancement



Input

sion

Experiment & Result: Training strategy

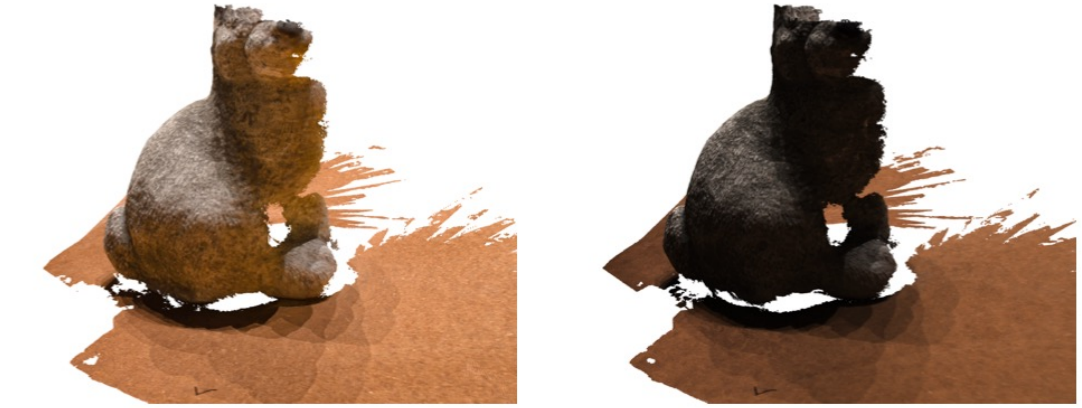
1. Strategy 1: MVS+Adapter
2. Strategy 2: Adapter
3. Strategy 3: Diffusion + Adapter
4. Strategy 3: Diffusion + MVS

Experiment & Result : Enlighten Color

(a) Ours (b) CasMVSNet

(c) Ours (d) CasMVSNet

(a) Ours (b) CasMVSNet

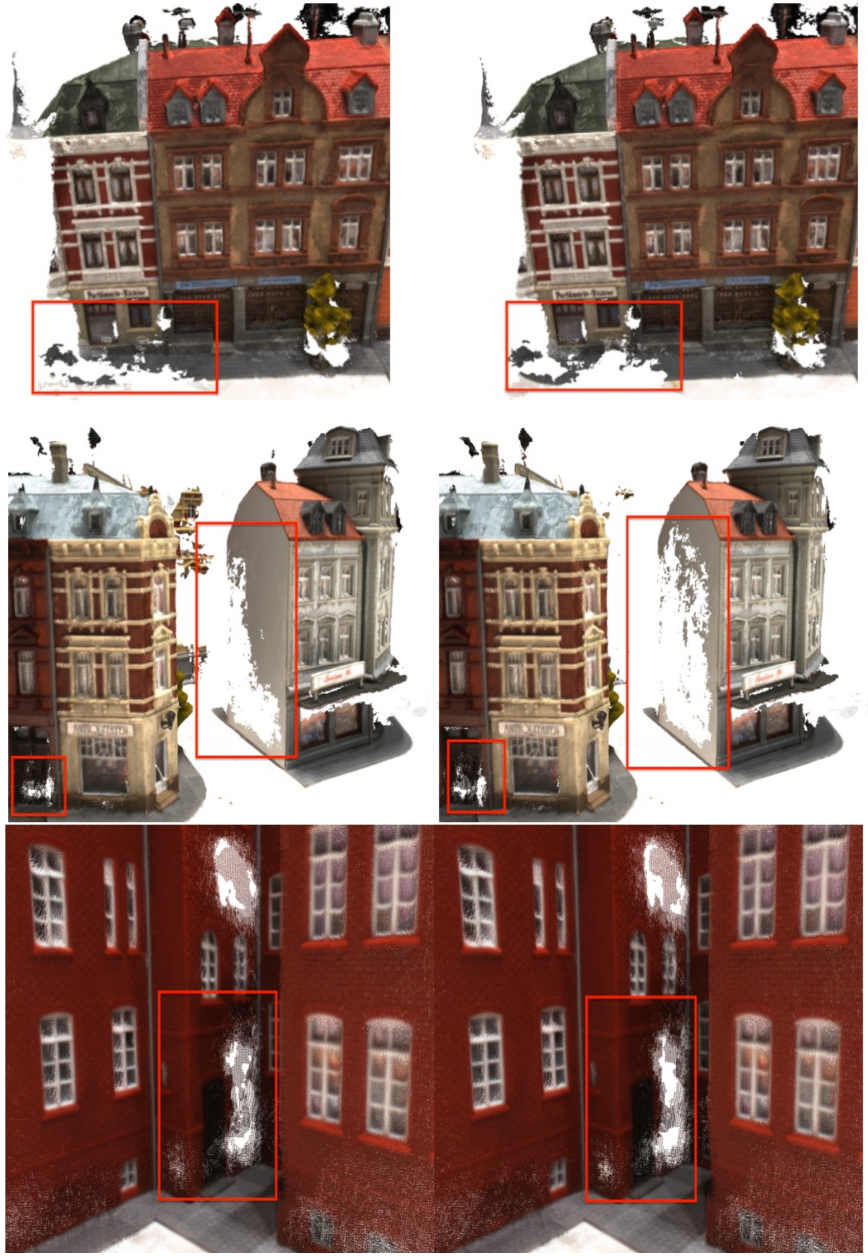


(c) Ours (d) CasMVSNet

Experiment & Result : Geometry

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**

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%



Methodology & Result: Tanks and temples

Model	F1 Score
Proposed Model	46.24
CasMVSNet	45.30

F1 scores of test results on 'Tanks and Temples'

Ours CasMVSNet

(a) Ours (b) CasMVSNet

Ours CasMVSNet

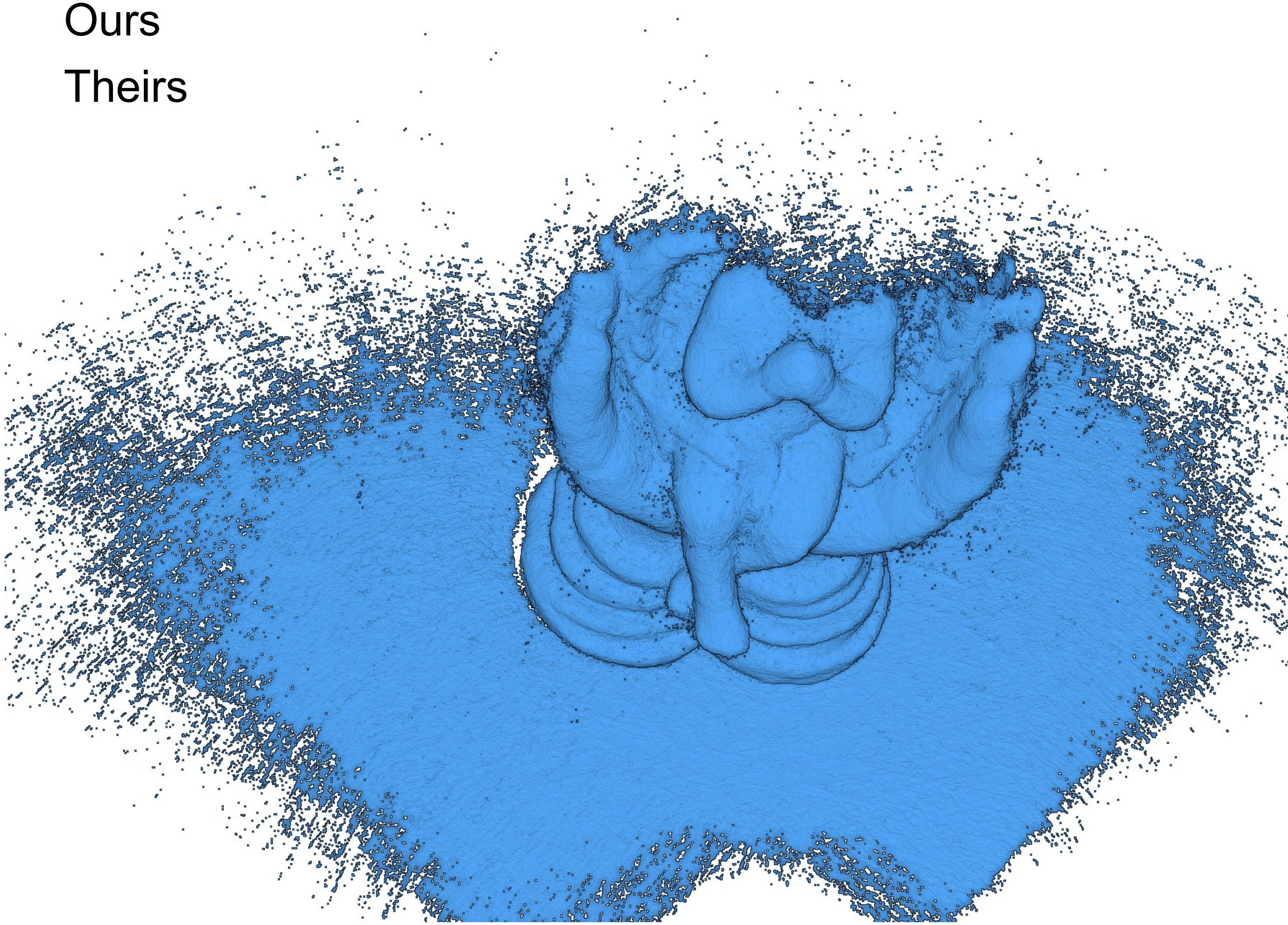
TUDelft

Methodology & Result: Integrate with other MVS

Light: 0	Acc. ↓	Comp. ↓	Overall ↓
GeoMVSNet (original)	0.342	0.344	0.343
GeoMVSNet (ours)	0.322	0.302	0.312
MVSNet (original)	0.540	0.492	0.523
MVSNet (ours)	0.547	0.485	0.516
Light: 3	**Acc. ↓**	**Comp. ↓**	**Overall ↓**
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
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

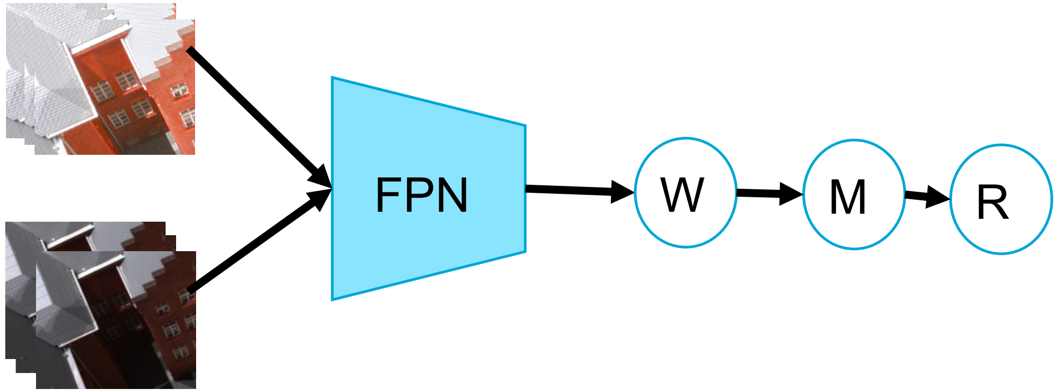
Ours

Theirs

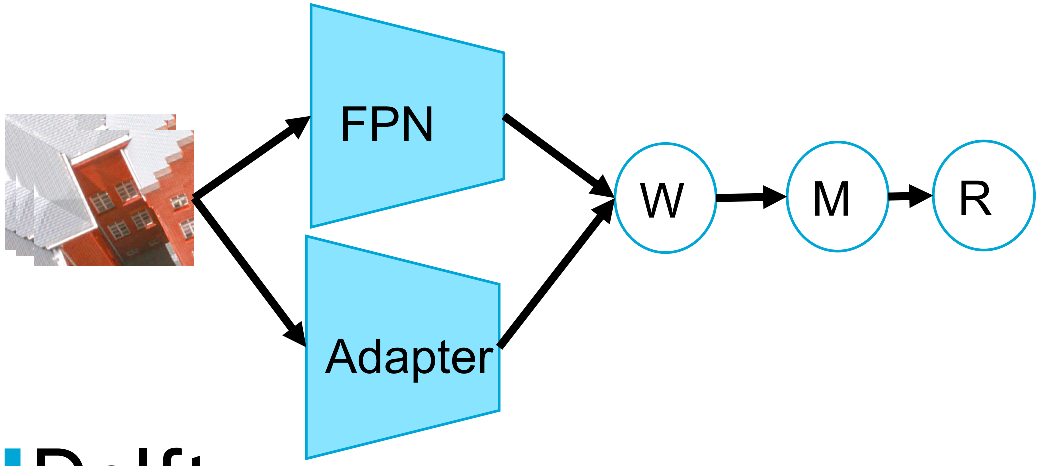


Experiment & Result: Feature Adapter

Only image enhancement input

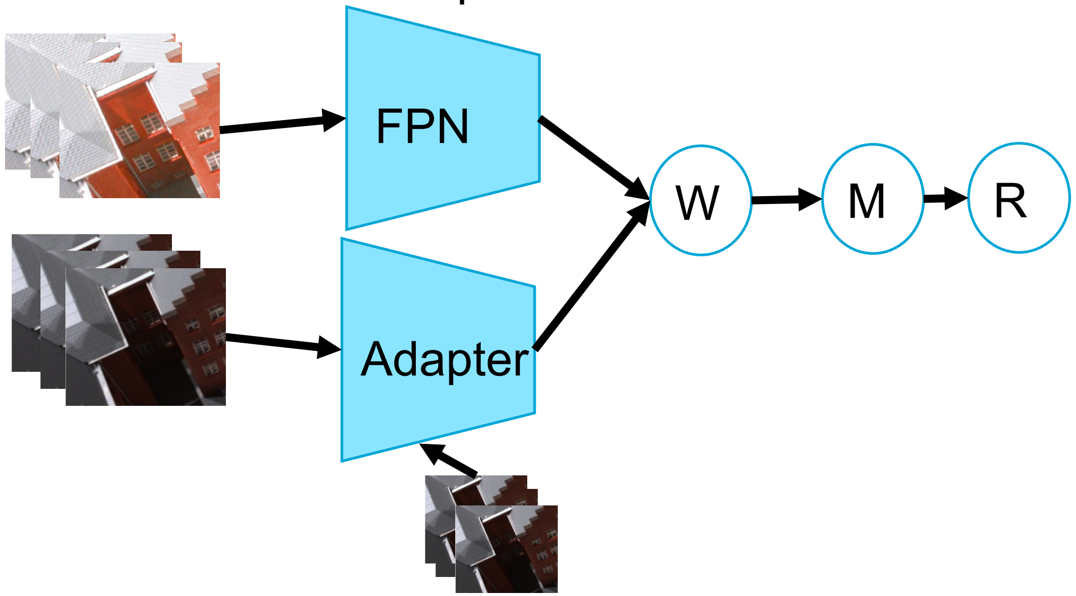


Only feature adapter



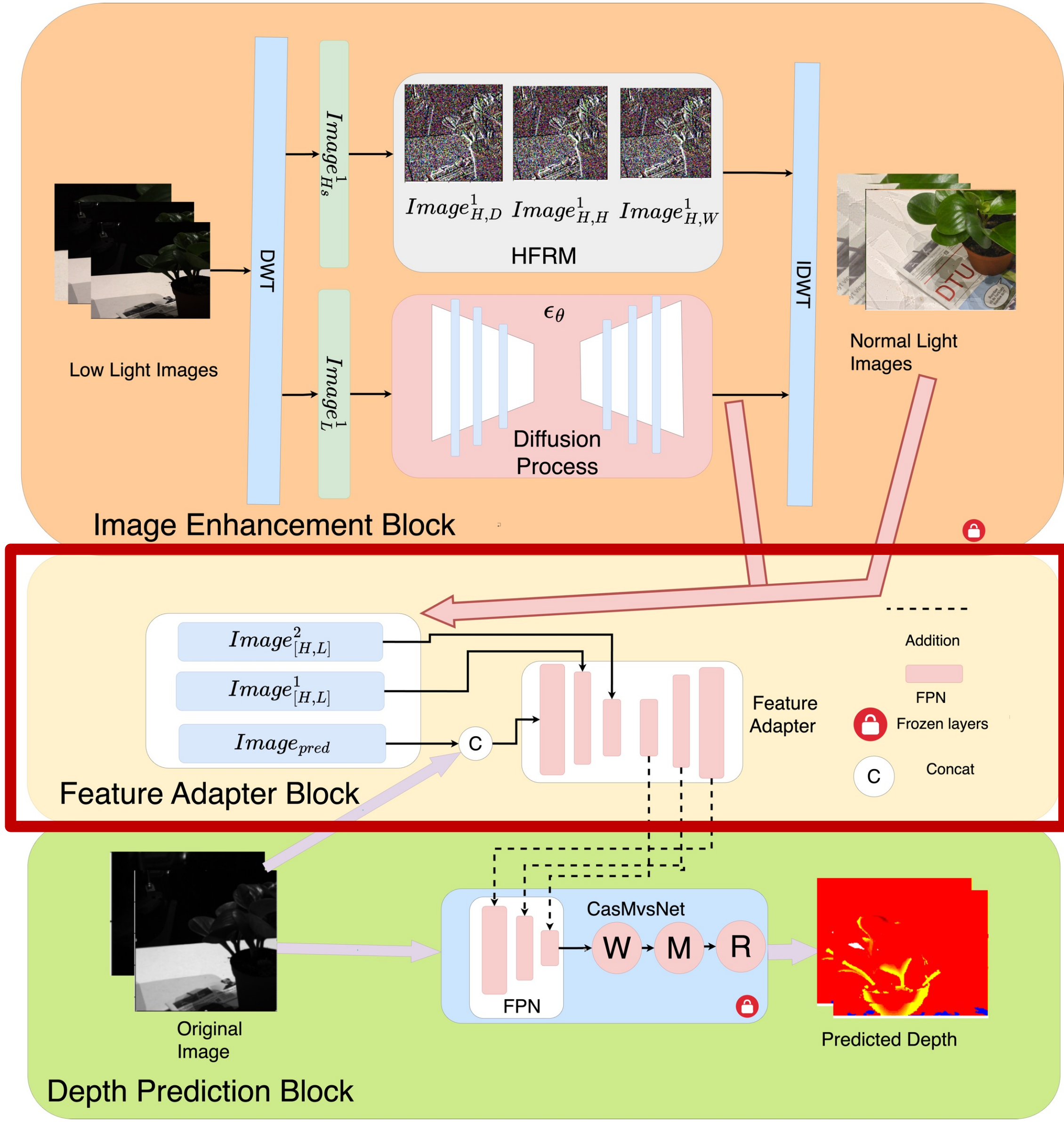
Ours

Without multi-scale input



Methodology & Result : Design of Feature Adapter

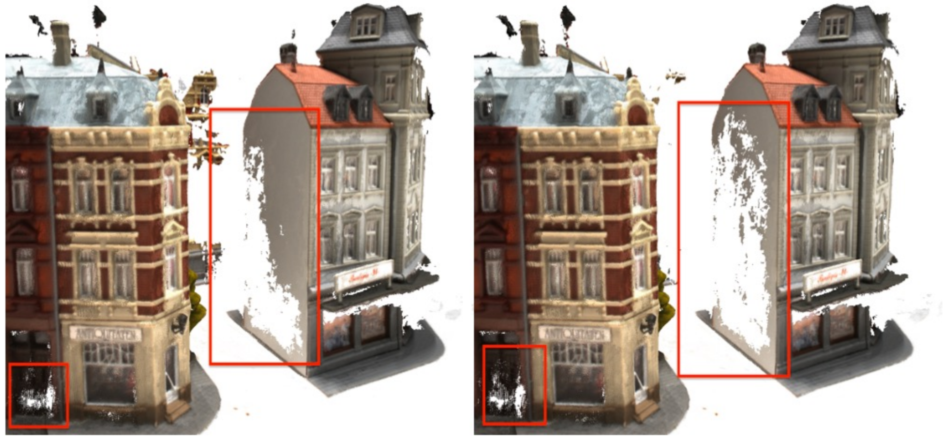
Light: 0	Acc. ↓	Comp. ↓	Overall ↓
CasMVSNet	0.328	0.472	0.400
Ours	0.330	**0.460**	**0.395**
Only feature adapter	**0.326**	0.475	0.401
Without DWT input	0.329	0.462	0.396
Only image enhancement input	0.329	0.465	0.397
Light: 3	Acc. ↓	Comp. ↓	Overall ↓
CasMVSNet	0.327	0.464	0.396
Ours	0.328	**0.454**	**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
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



Methodology & Result: Discussion

Pros:
1. Enhanced Performance in Low Light
2. Versatility
3. Efficiency
4. Improved Visual Quality

Cons:
1. Multi-View Consistency
2. Optimal Image Enhancement Model
3. Geometric Accuracy
4. Limited Improvement for Some Pipelines



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**

TU Delft

Conclusions

To what extent can existing single-frame image enhancement models be utilized to enhance the performance of MVS in low illumination conditions?

1. Which image enhancement model is suitable?
   - Low-light Diffusion
2. Which architecture is suitable for integrating the image enhancement model with MVS?
   - Feature adapter
3. How can we reduce the computation resource demands?
   - Only fine-tune feature Adapter

Conclusions

Contributions:

1. Evaluation of Image Enhancement Models
2. Innovative Feature Adapter Design
3. Efficient Training Framework
4. Ablation Studies and Mechanistic Insights

Limitations:

1. Loss of Accuracy
2. Dataset bias
3. Multi-view inconsistency exists
4. Computational Resource Requirements

TU Delft

Reference

1. Furukawa, Y., & Hernández, C. (2015). Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, *9*(1-2), 1-148.
2. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., & Tan, P. (2020). Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2495-2504).
3. Aldeeb, N. H., & Hellwich, O. (2020). 3D Reconstruction Under Weak Illumination Using Visibility-Enhanced LDR Imagery. In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1* (pp. 515-534). Springer International Publishing.
4. Kargas, A., Loumos, G., & Varoutas, D. (2019). Using different ways of 3D reconstruction of historical cities for gaming purposes: The case study of Nafplio. *Heritage*, *2*(3).
5. Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.
6. Han, J., Chen, X., Zhang, Y., Hou, W., & Hu, Z. (2022). DEMVSNet: Denoising and depth inference for unstructured multi-view stereo on noised images. *IET Computer Vision*, *16*(7), 570-580.
7. Wang, Y., & Jiang, Q. (2024). LoliMVS: an End-to-end Network for Multi-view Stereo with Low-light Images. *IEEE Transactions on Instrumentation and Measurement*.
8. Su, Y., Wang, J., Wang, X., Hu, L., Yao, Y., Shou, W., & Li, D. (2023). Zero-reference deep learning for low-light image enhancement of underground utilities 3d reconstruction. *Automation in Construction*, *152*, 104930.
9. Jiang, H., Luo, A., Fan, H., Han, S., & Liu, S. (2023). Low-light image enhancement with wavelet-based diffusion models. *ACM Transactions on Graphics (TOG)*, *42*(6), 1-14.

Q & A