MSc thesis in Geomatics

Semantically-Guided 3D Building Facade Reconstruction: A Learning-Based MVS Approach

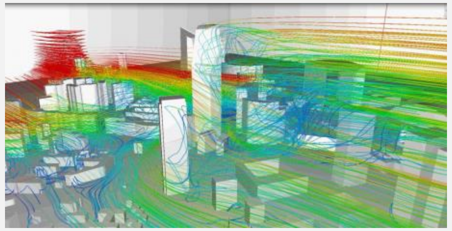
Author:
Ioanna Panagiotidou

Supervisors:
Nail Ibrahimli, Hugo Ledoux
Co-reader:
Shiming Wang

October 27, 2023

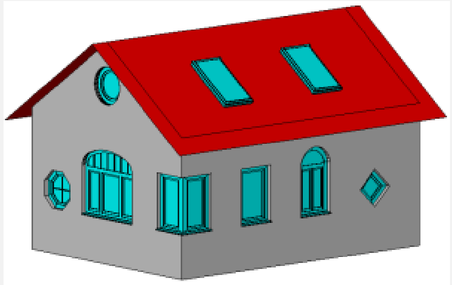
TU Delft

Introduction: 3D Building Models



Disaster Response

- simulate floods
- map water flows
- predict wind dispersion
- heat patterns



Urban Planning

- energy efficient buildings
- shadow estimation
- solar potential

Position of the IOF vehicles
Trajectory of shots
Position of the journalists before shooting begins

Source: https://forensic-architecture.org/

Forensics

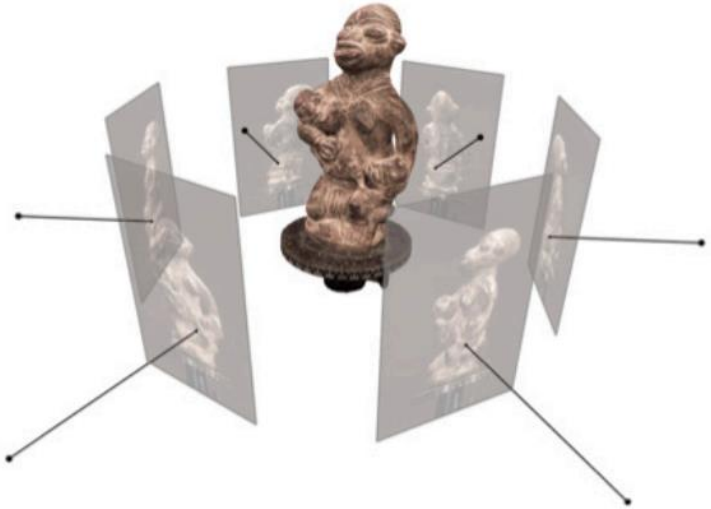
- work in tandem with other elements
- reconstruct crime scenes and unveil concealed evidence

Introduction: Point Clouds

Obtained via:

- Photogrammetry (Multi-View Stereo algorithm)
- LiDAR

Photogrammetry



LiDAR

Source: Carlos Hermandez (2015)

Introduction: MVS

- **reconstructs a 3D point cloud representation based on**
  - set of overlapping images and camera parameters

Introduction: MVS

- **reconstructs a 3D point cloud representation based on**
  - set of overlapping images and camera parameters

Bring back the depth!

Introduction: MVS

- **reconstructs a 3D point cloud representation based on**
  - set of overlapping images and camera parameters
- **Pipeline involves**

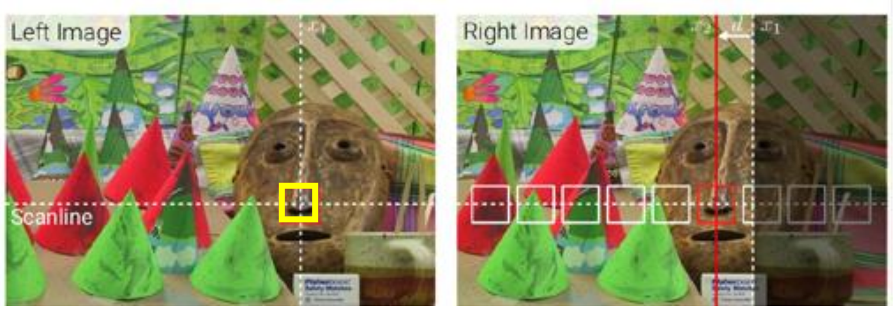
Introduction: MVS

- **reconstructs a 3D point cloud representation based on**
  - set of overlapping images and camera parameters

- **Pipeline involves**

  1. Locating **matching pixels** in overlapping images

Correspondences



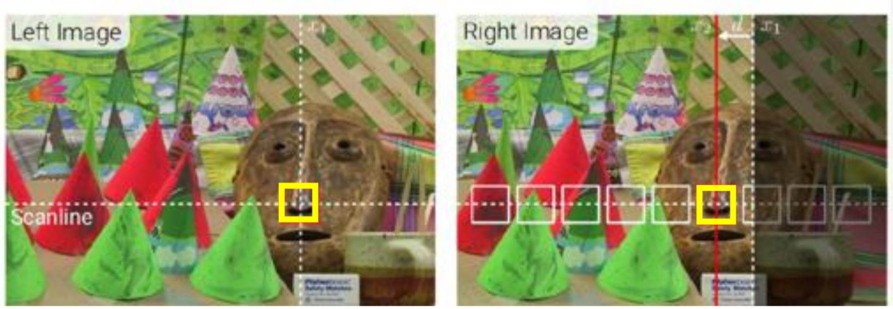
Introduction: MVS

- **reconstructs a 3D point cloud representation based on**
  - set of overlapping images and camera parameters

- **Pipeline involves**

  1. Locating **matching pixels** in overlapping images

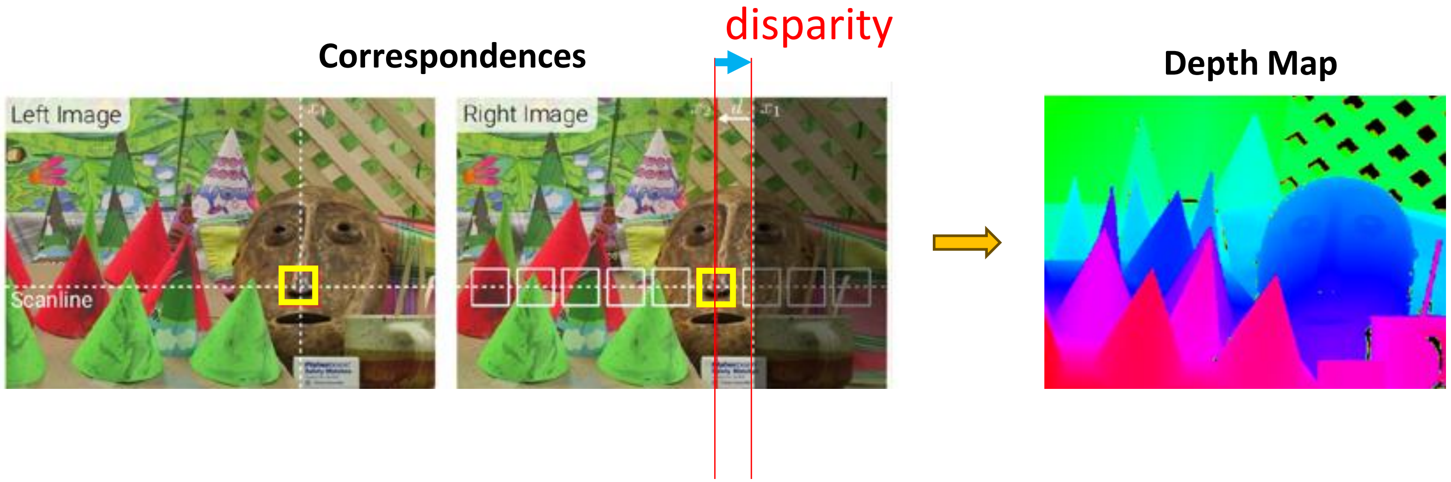
Correspondences



Introduction: MVS

- **reconstructs a 3D point cloud representation based on**
  - set of overlapping images and camera parameters

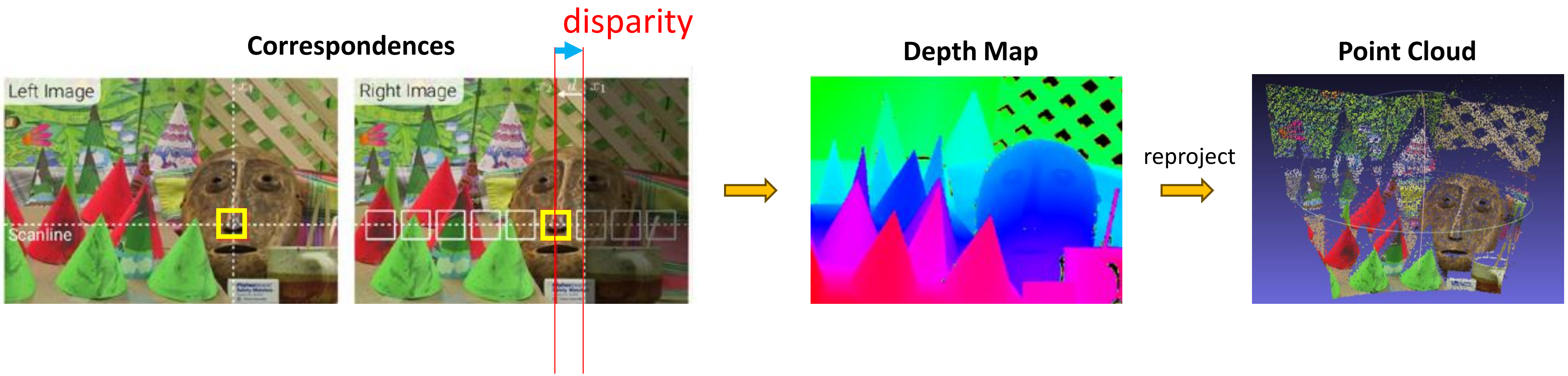
- **Pipeline involves**
  1. Locating **matching pixels** in overlapping images
  2. Deriving **depth** from disparities in pixel positions



Introduction: MVS

- **reconstructs a 3D point cloud representation based on**
  - set of overlapping images and camera parameters

- **Pipeline involves**
  1. Locating **matching pixels** in overlapping images
  2. Deriving **depth** from disparities in pixel positions
  3. Recovering **point cloud** (3D)



Correspondences

disparity

Depth Map

reproject

Point Cloud

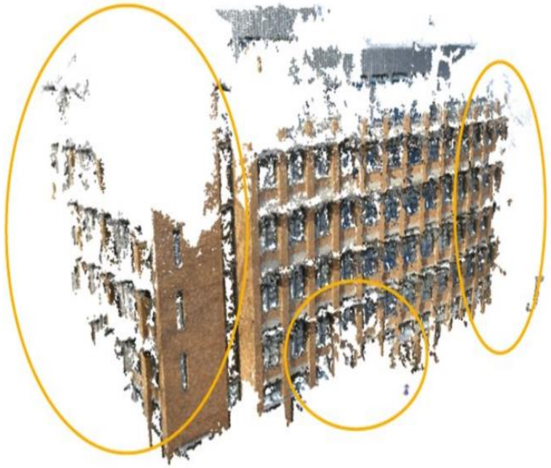
Introduction: Challenges with Traditional MVS

- Reconstruct **Accurate** but **Incomplete** models.
  - Rely on photo-consistent metrics (RGB) to locate the matching pixels
    - matching **impossible** in **reflective, low-textured** regions

Introduction: Challenges with Traditional MVS

- Reconstruct **Accurate** but **Incomplete** models.
  - Rely on photo-consistent metrics (RGB) to locate the matching pixels
    - matching **impossible** in **reflective, low-textured** regions

Incomplete Reconstruction

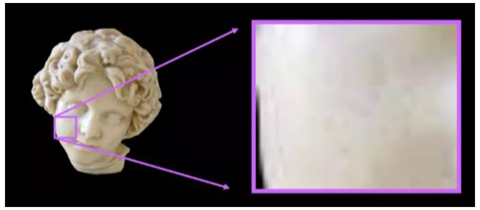


Introduction: Challenges with Traditional MVS

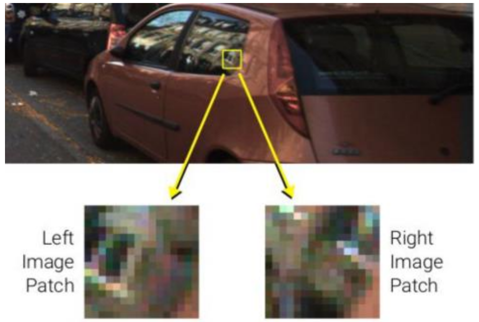
- Reconstruct **Accurate** but **Incomplete** models.
    - Rely on photo-consistent metrics (RGB) to locate the matching pixels
        - matching **impossible** in **reflective, low-textured** regions

RGB space

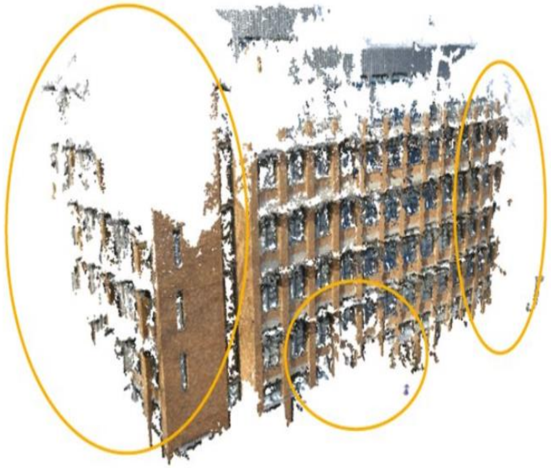
Low-textured



Reflective



Incomplete Reconstruction

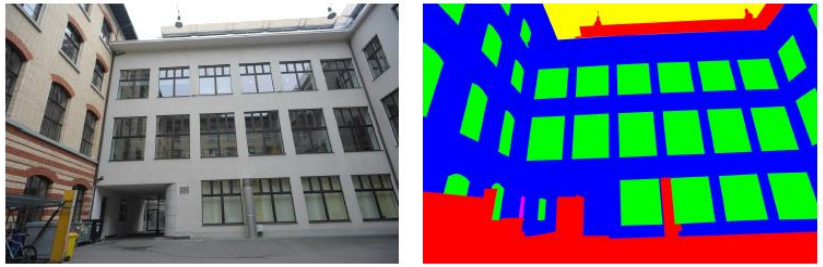


Background: Address limitation

- Traditional MVS with Semantic Priors
- Learning-based MVS

1. Semantic priors into Traditional MVS pipelines

- Semantics indicate the weak regions
- Guide class-specific geometric constraints in order to improve depth



RGB image and Semantic Map.

[Stathopoulou et al., 2021]

Background: Address limitation

- Traditional MVS with Semantic Priors
- Learning-based MVS

2. Learning-based MVS systems

- Outperform traditional MVS in these challenging regions

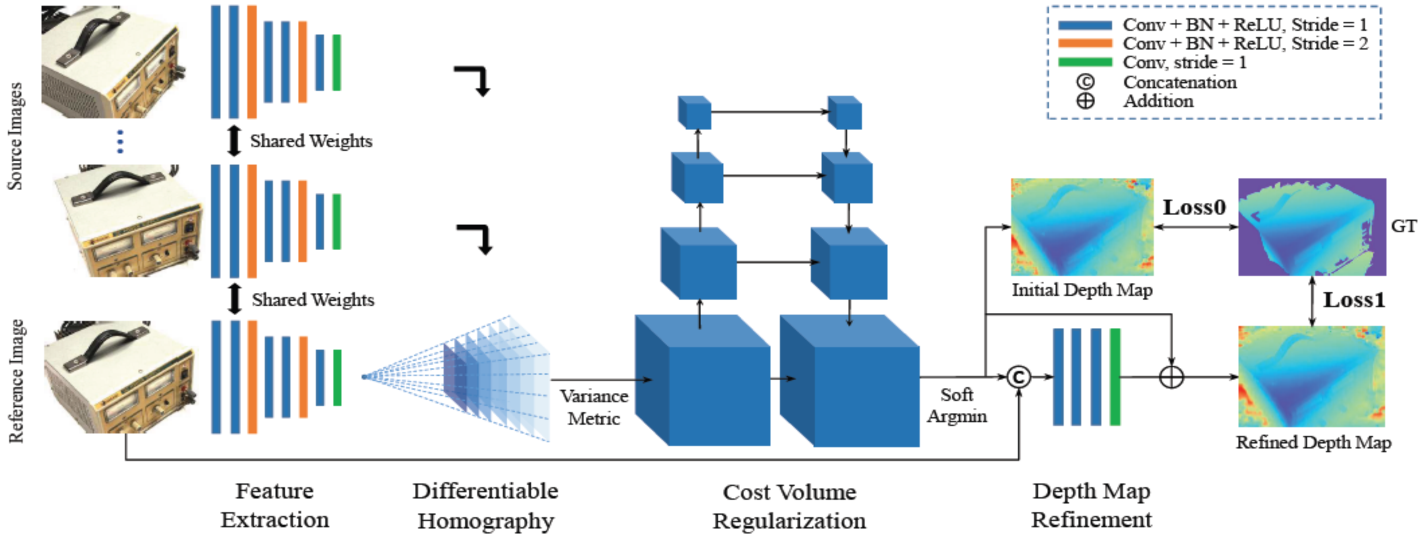


Figure: MVSNet. Source: Yao et al. (2018)

Background: Address limitation

- Traditional MVS with Semantic Priors
- Learning-based MVS

2. Learning-based MVS systems

- Outperform traditional MVS in these challenging regions

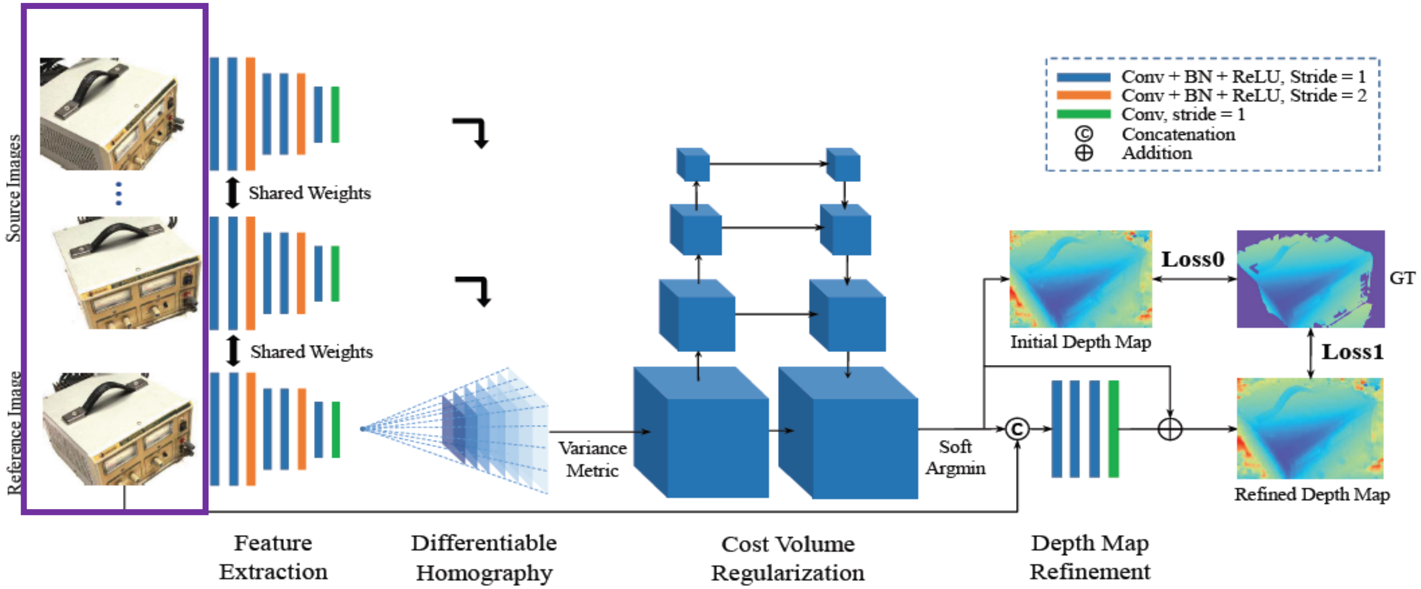


Figure: MVSNet. Source: Yao et al. (2018)

Background: Address limitation

- Traditional MVS with Semantic Priors
- Learning-based MVS

2. Learning-based MVS systems

- Outperform traditional MVS in these challenging regions

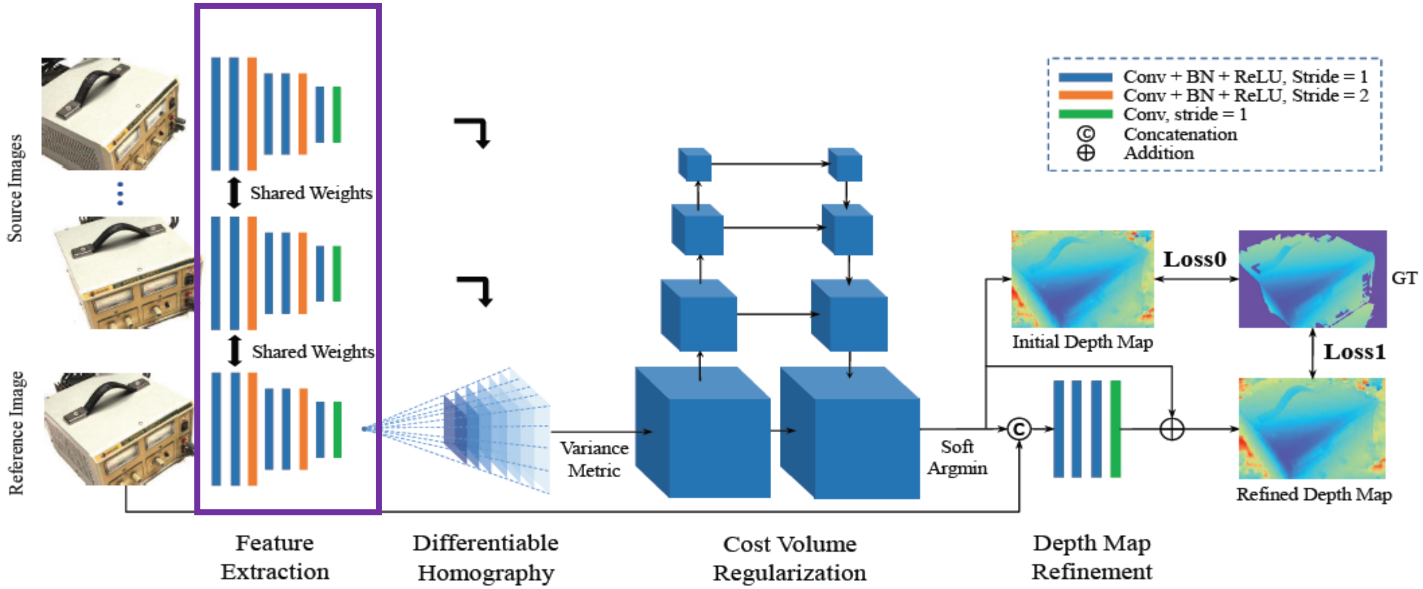


Figure: MVSNet. Source: Yao et al. (2018)

Bridging the Gap

Bridging the Gap

Semantic
guidance

Bridging the Gap

Semantic
guidance

Learning-based
MVS

Research Objective & Questions

Objective:

- **Refine the 3D reconstruction** of buildings using **semantic guidance** and **deep learning (DL)**.

Research Questions & Objective

Objective:

- **Refine the 3D reconstruction** of buildings using **semantic guidance** and **deep learning (DL)**.

Main Question:

- To what extent can **leveraging semantic priors within learned MVS techniques enhance** the accuracy and completeness of **3D models** of buildings?

Research Questions & Objective

Objective:

- **Refine the 3D reconstruction** of buildings using **semantic guidance** and **deep learning (DL)**.

Main Question:

- To what extent can **leveraging semantic priors within learned MVS techniques enhance** the accuracy and completeness of **3D models** of buildings?

Sub-questions:

- How can **semantic priors be effectively integrated into a DL framework** to facilitate the semantically-guided regularization of 3D models of buildings?

Research Questions & Objective

Objective:

- **Refine the 3D reconstruction** of buildings using **semantic guidance** and **deep learning (DL)**.

Main Question:

- To what extent can **leveraging semantic priors within learned MVS techniques enhance** the accuracy and completeness of **3D models** of buildings?

Sub-questions:

- How can **semantic priors be effectively integrated into a DL framework** to facilitate the semantically-guided regularization of 3D models of buildings?
- What is a **suitable refinement module architecture** for depth residual learning that can best contribute to the improvement of the 3D reconstruction of buildings?

Research Questions & Objective

Objective:

- **Refine the 3D reconstruction** of buildings using **semantic guidance** and **deep learning (DL)**.

Main Question:

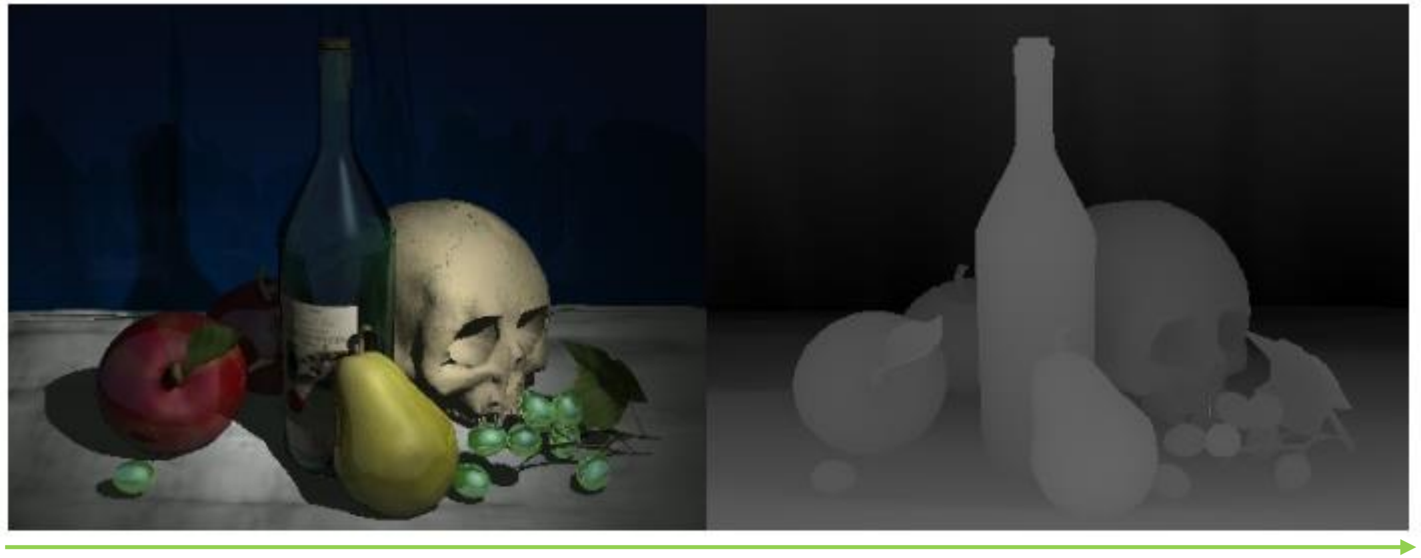
- To what extent can **leveraging semantic priors within learned MVS techniques enhance** the accuracy and completeness of **3D models** of buildings?

Sub-questions:

- How can **semantic priors be effectively integrated into a DL framework** to facilitate the semantically-guided regularization of 3D models of buildings?
- What is a **suitable refinement module architecture** for depth residual learning that can best contribute to the improvement of the 3D reconstruction of buildings?
- Which **deep learning architecture for semantic segmentation** demonstrates superior performance in detecting facade elements, such as walls, doors, and windows?

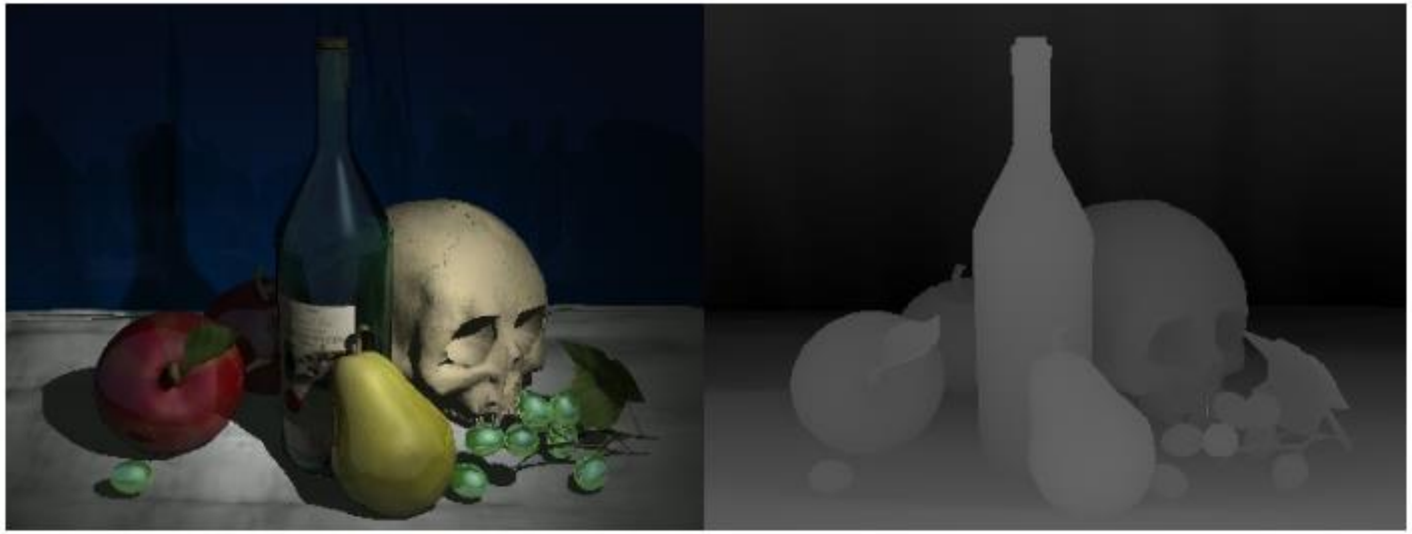
Related Work:

Convolutional Neural Networks



Related Work:

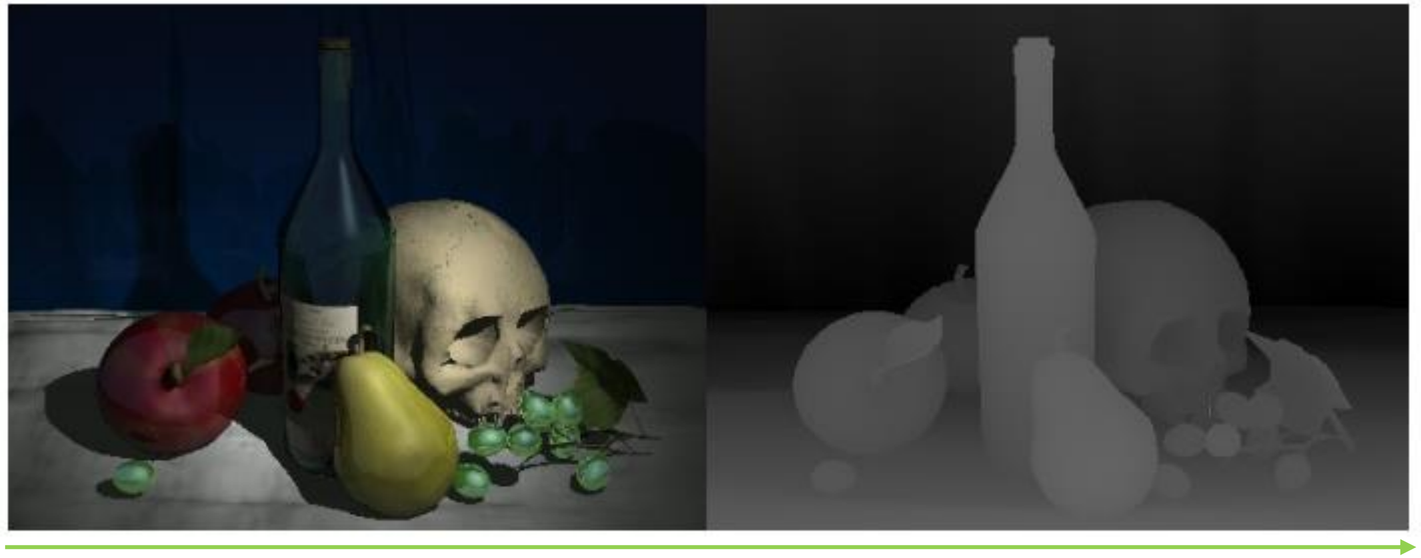
Convolutional Neural Networks

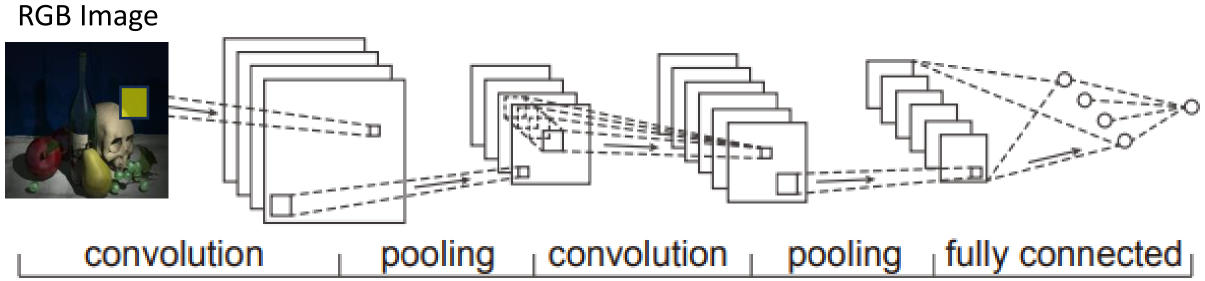


RGB Image

Related Work:

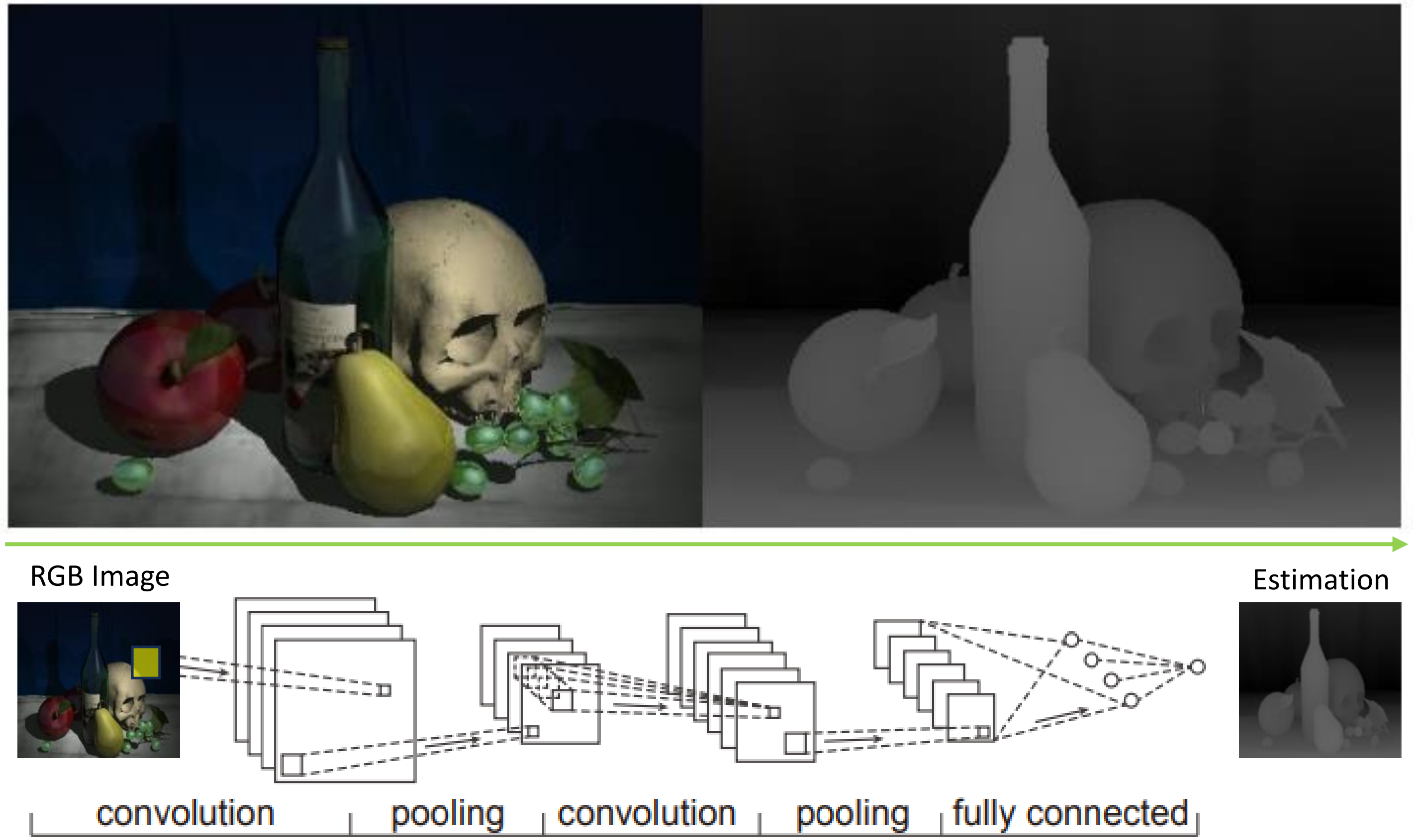
Convolutional Neural Networks





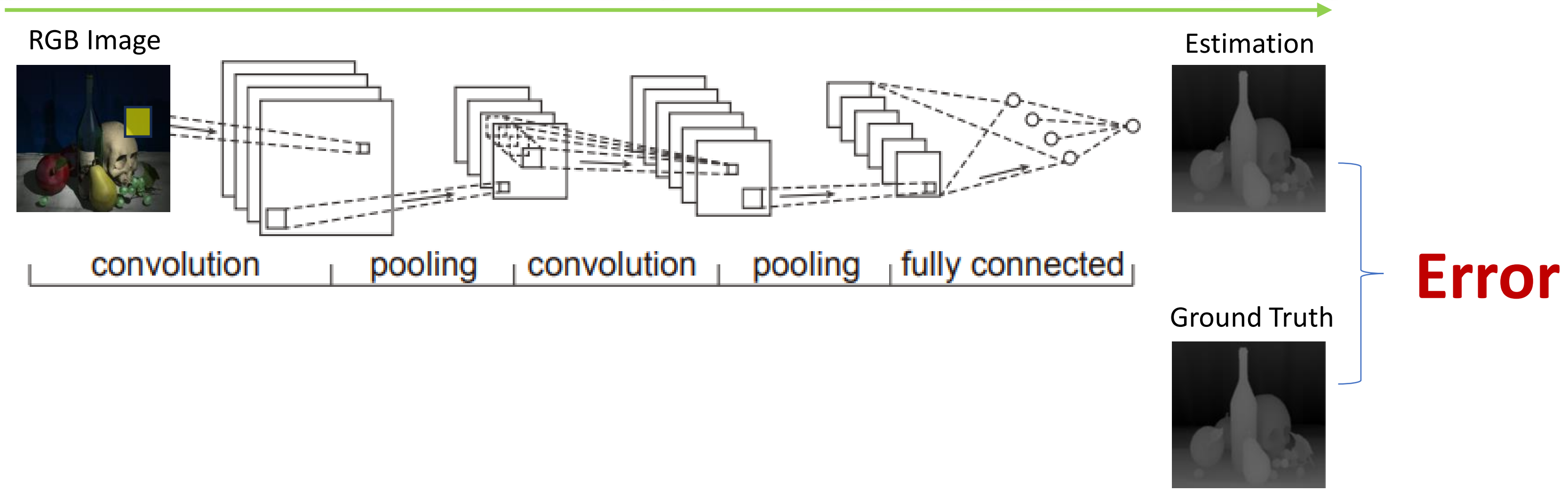
Related Work:

Convolutional Neural Networks



Related Work:

Convolutional Neural Networks



RGB Image

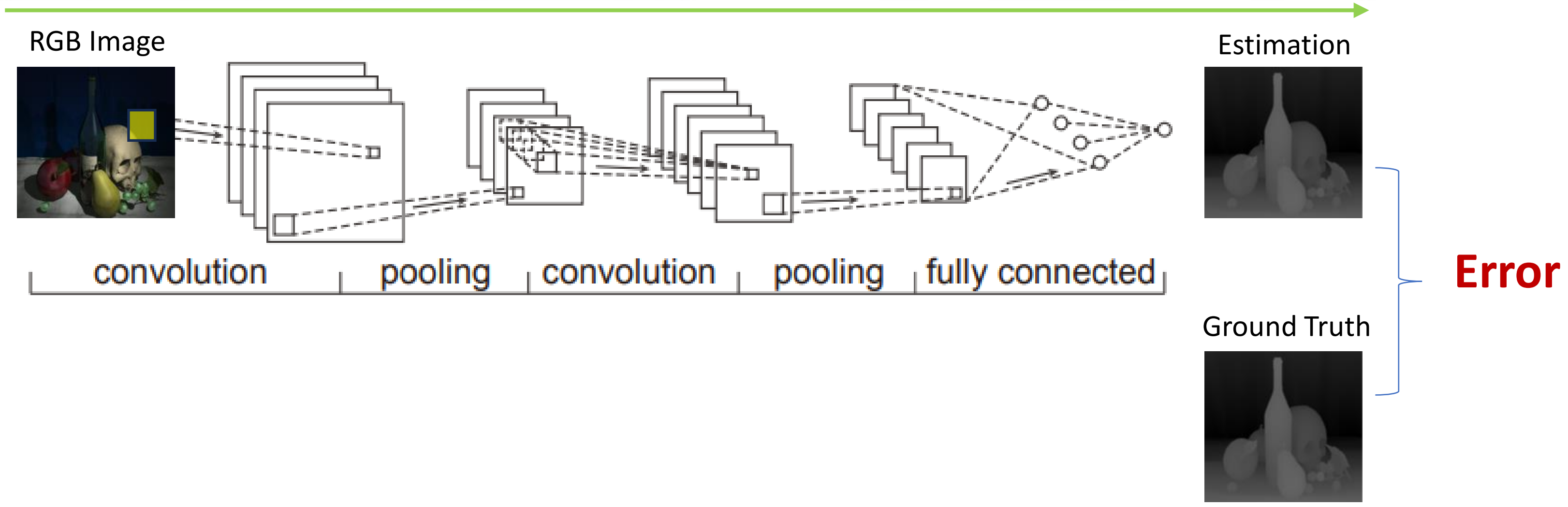
Estimation

Ground Truth

Error

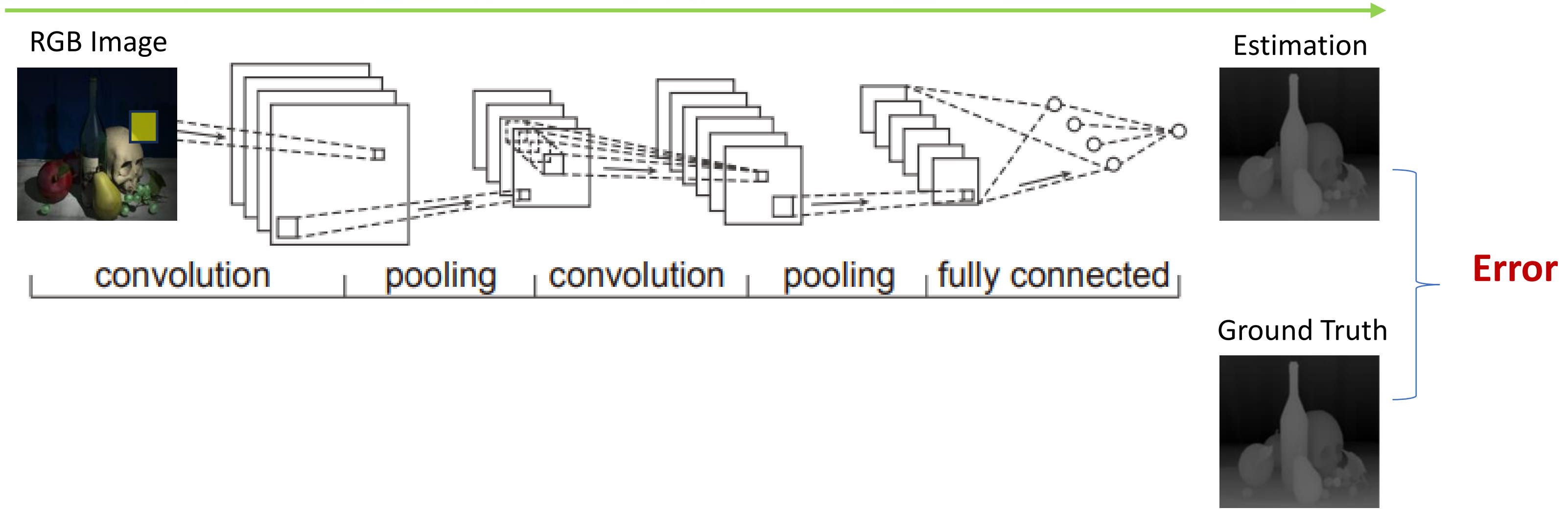
Related Work:

Convolutional Neural Networks



Related Work:

Convolutional Neural Networks



Related Work: Learning-based MVS

- MVS Network

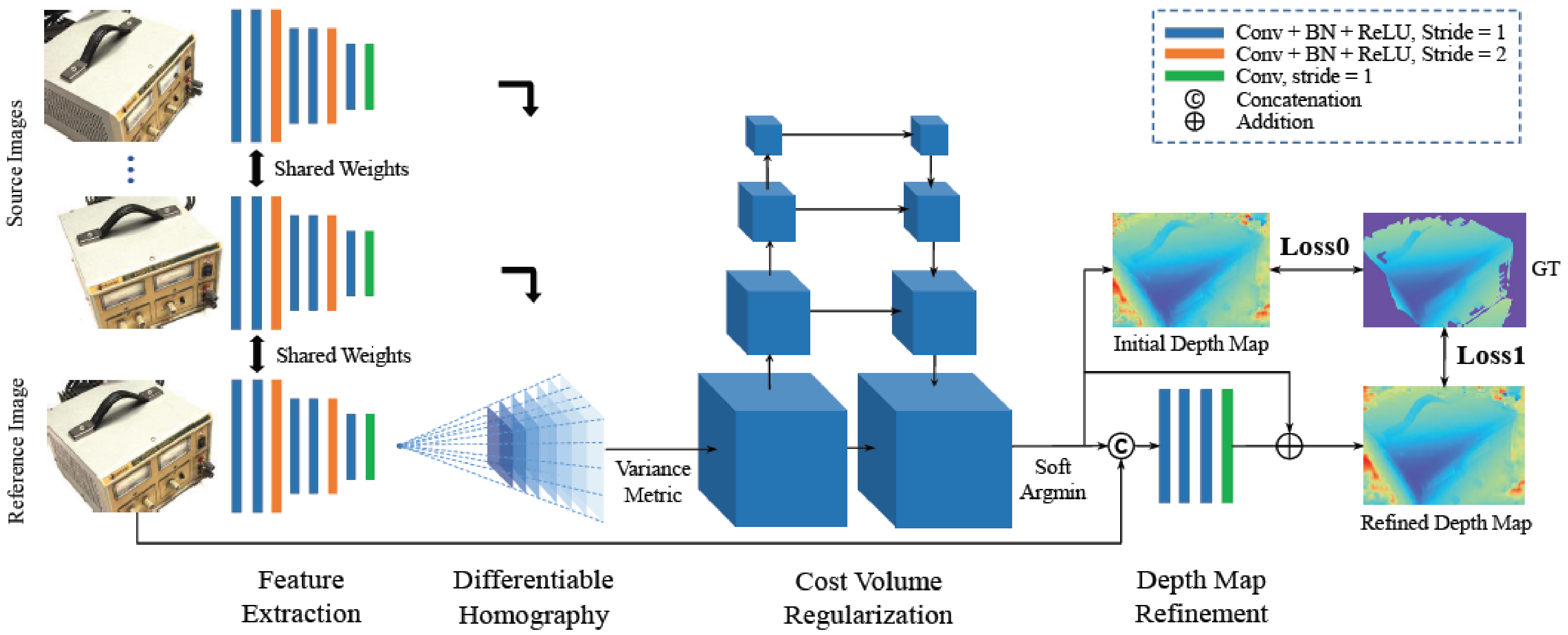


Figure: MVSNet. Source: Yao et al. (2018)

Related Work: Learning-based MVS

- MVS Network

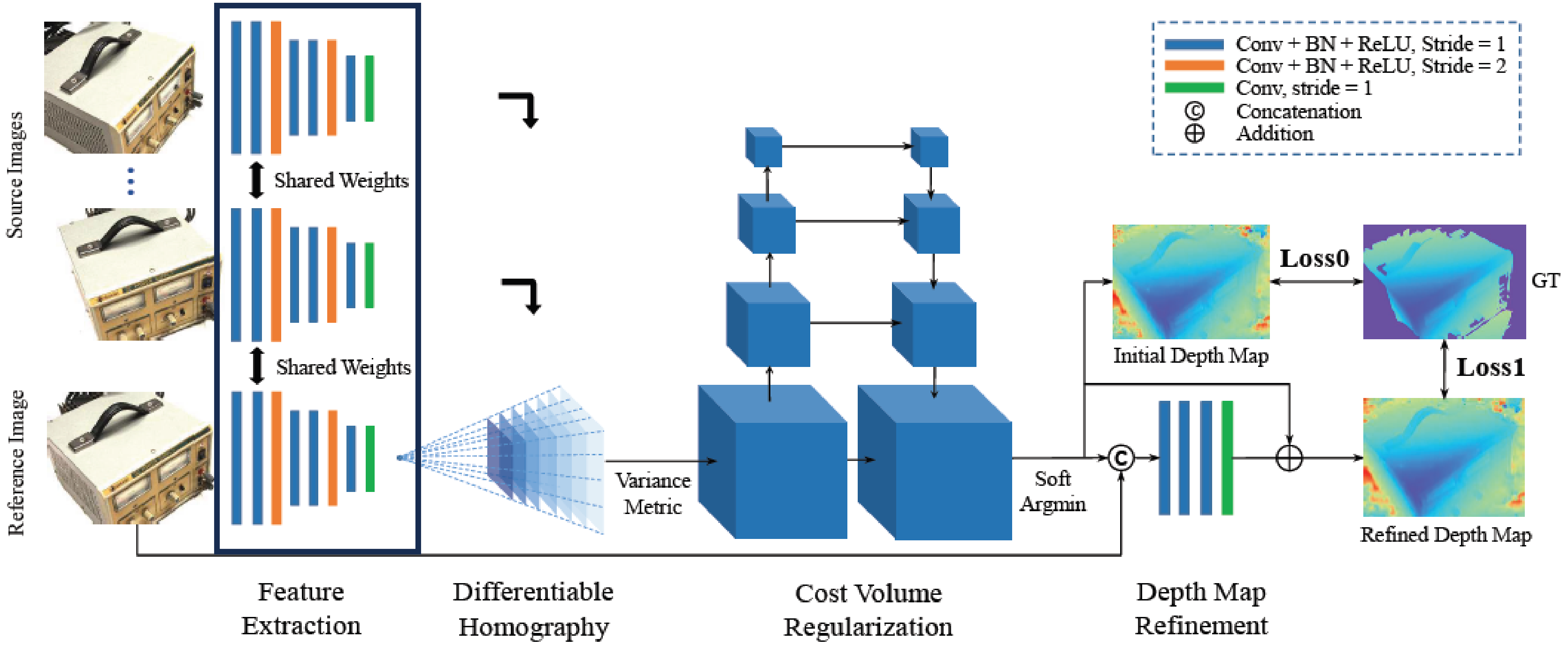


Figure: MVSNet. Source: Yao et al. (2018)

Related Work: Learning-based MVS

- MVS Network

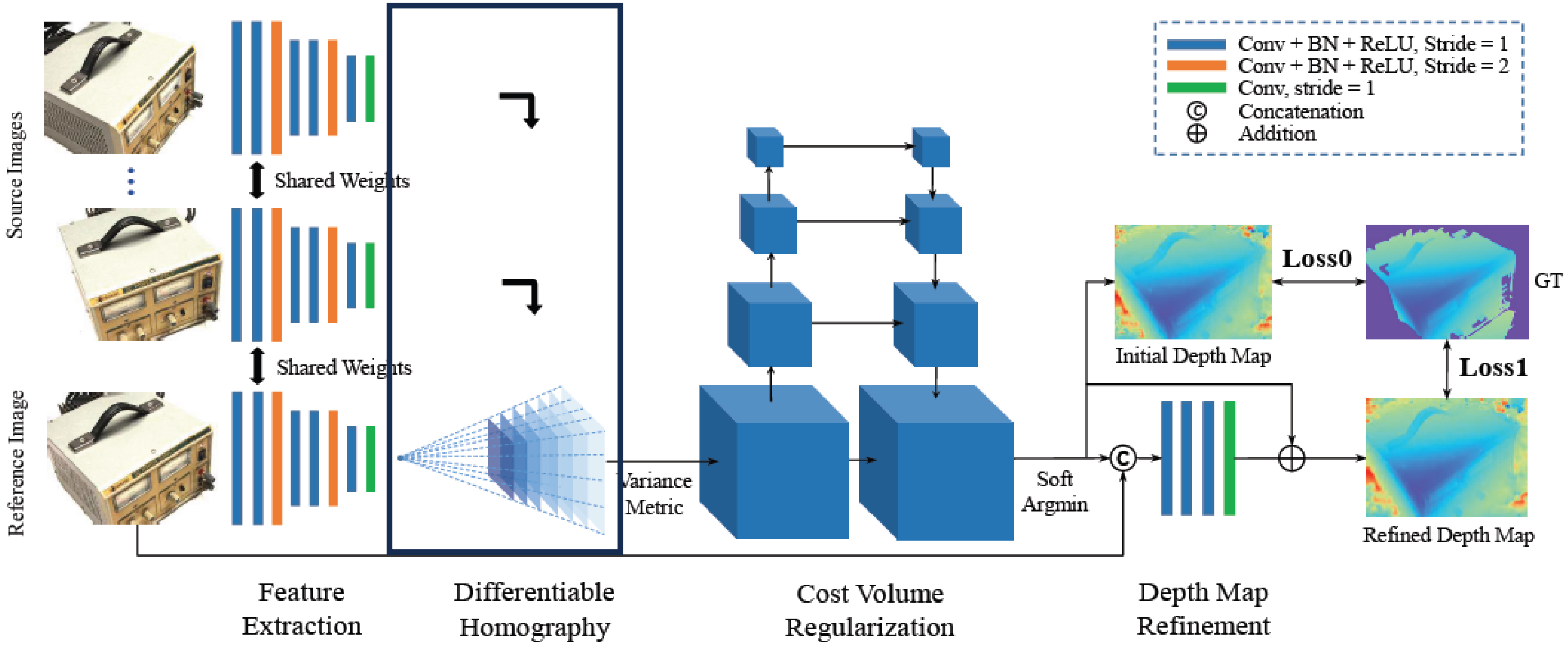


Figure: MVSNet. Source: Yao et al. (2018)

Related Work: Learning-based MVS

- MVS Network

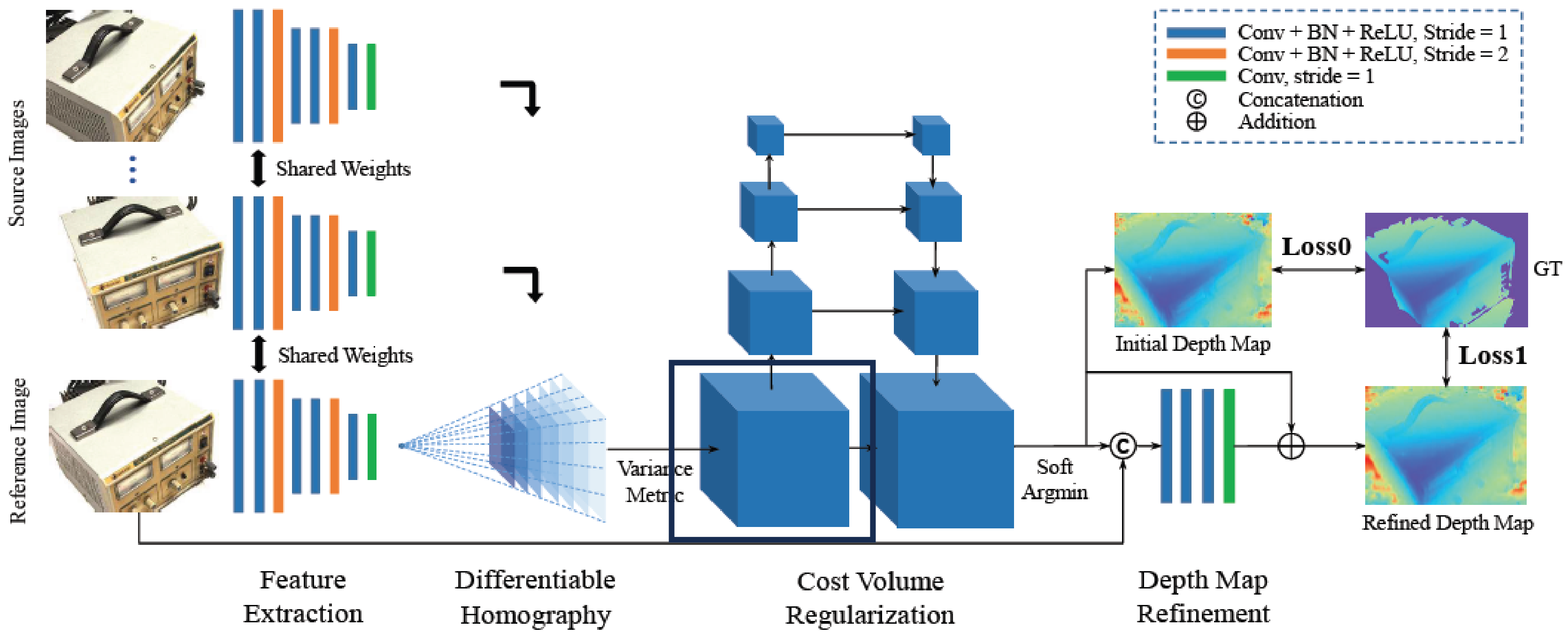
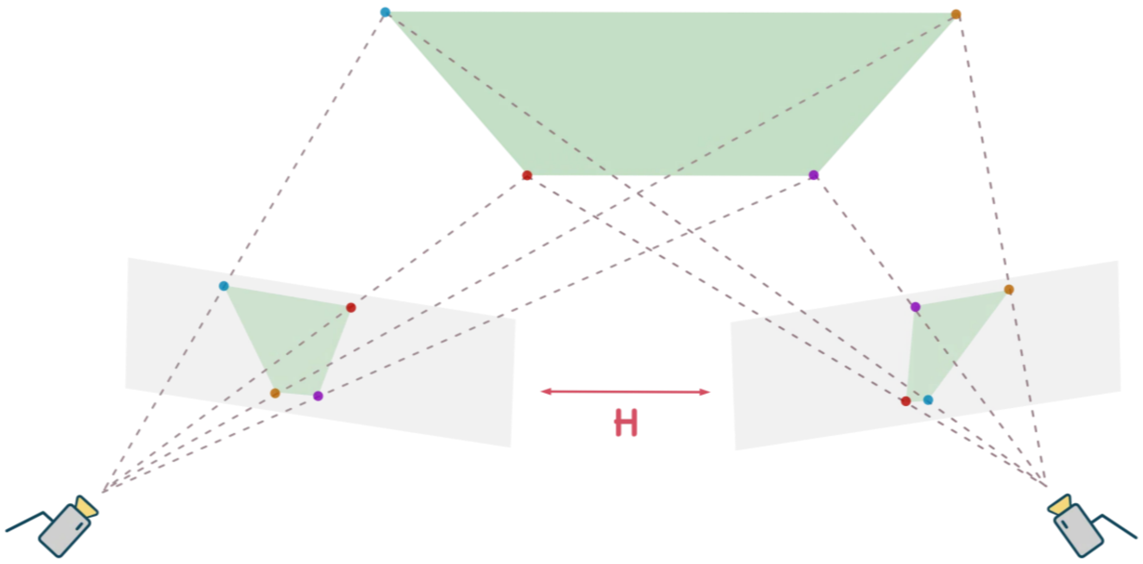
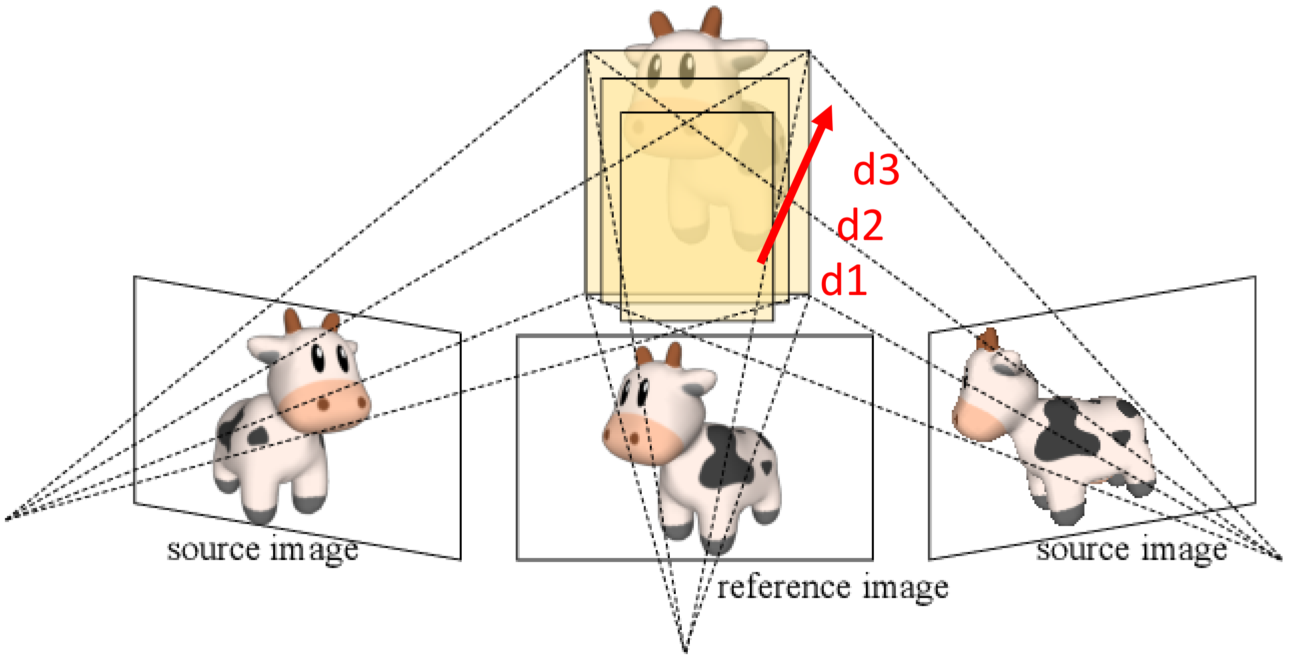


Figure: MVSNet. Source: Yao et al. (2018)

Related Work: Traditional MVS

- Differentiable Homography and the Plane Sweep Algorithm.





$$H_i(d) = K_i \cdot R_i \cdot (I - \frac{(t_1 - t_i) \cdot {n_1}^T}{d \; ?}) \cdot {R_1}^T \cdot {K_1}^{-1}$$

Source: https://towardsdatascience.com/estimating-a-homography-matrix-522c70ec4b2c Source: Deep Learning for Multi-View Stereo via Plane Sweep: A Survey

Related Work: Learning-based MVS

- MVS Network

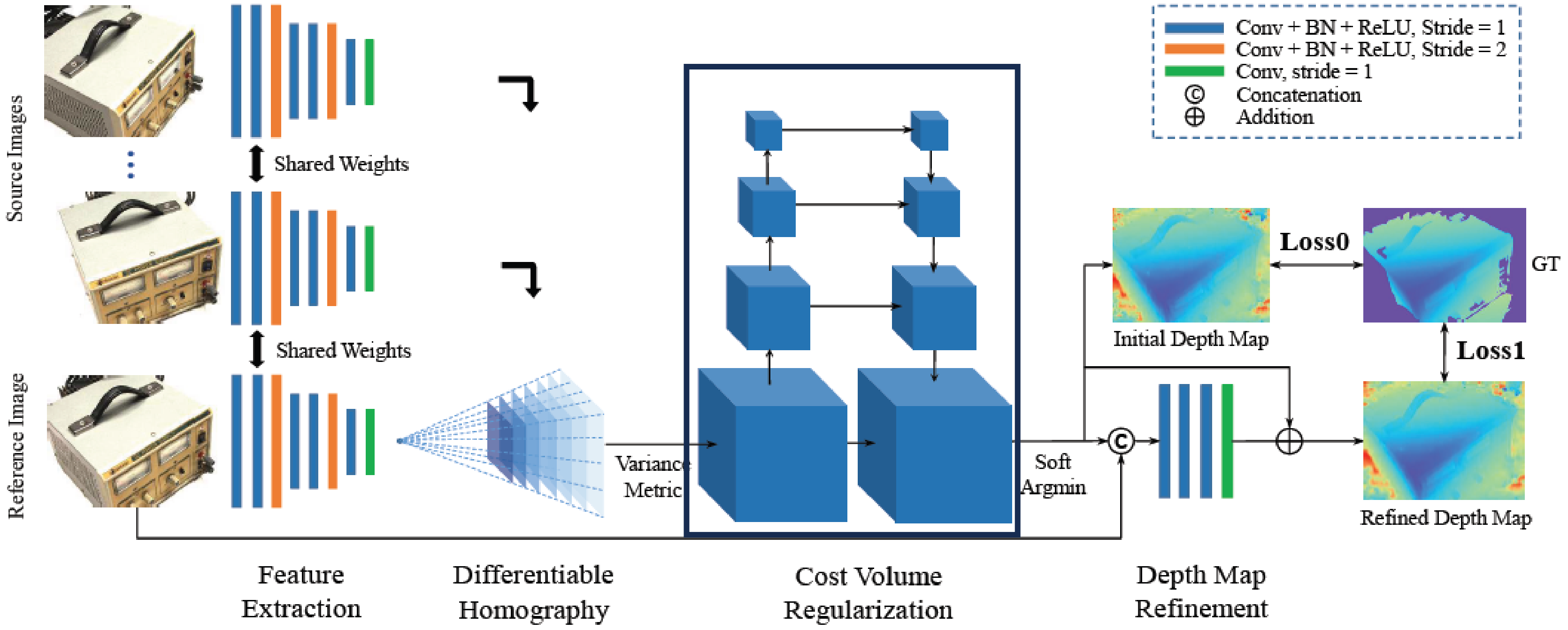
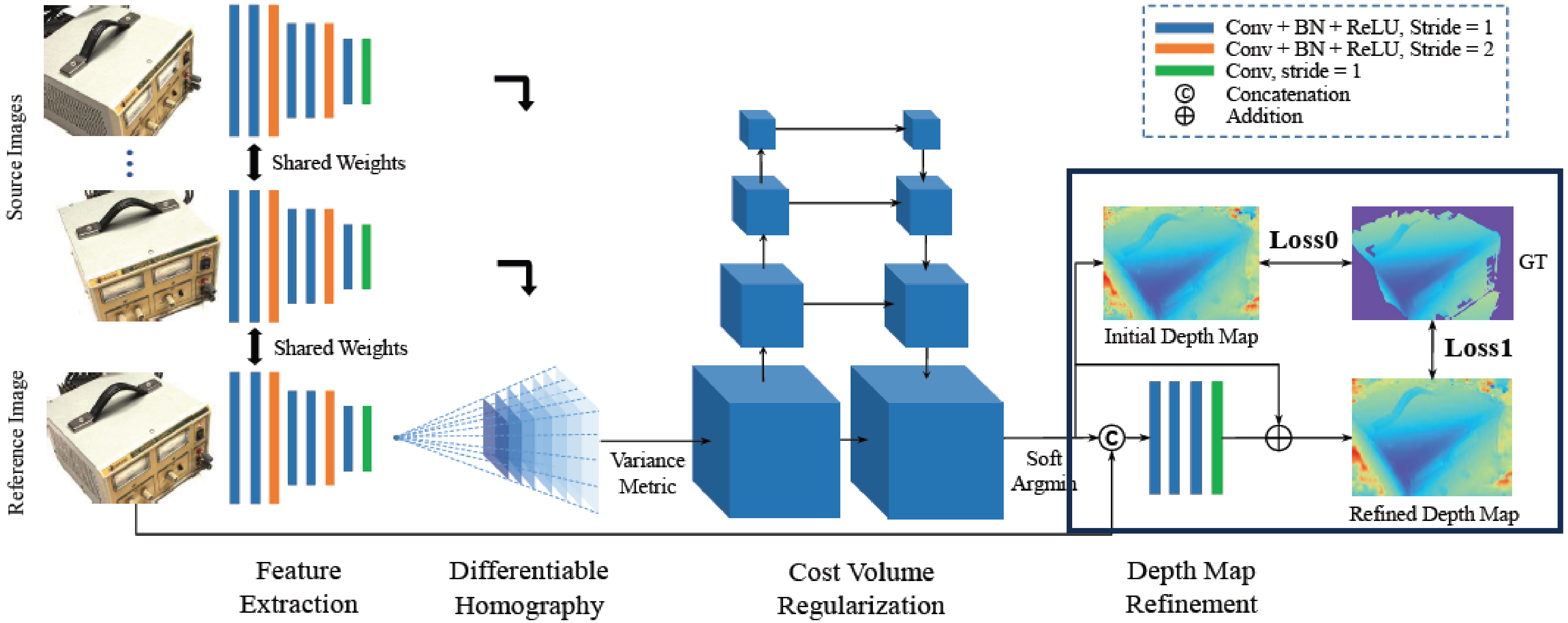


Figure: MVSNet. Source: Yao et al. (2018)

Related Work: Learning-based MVS

- MVS Network

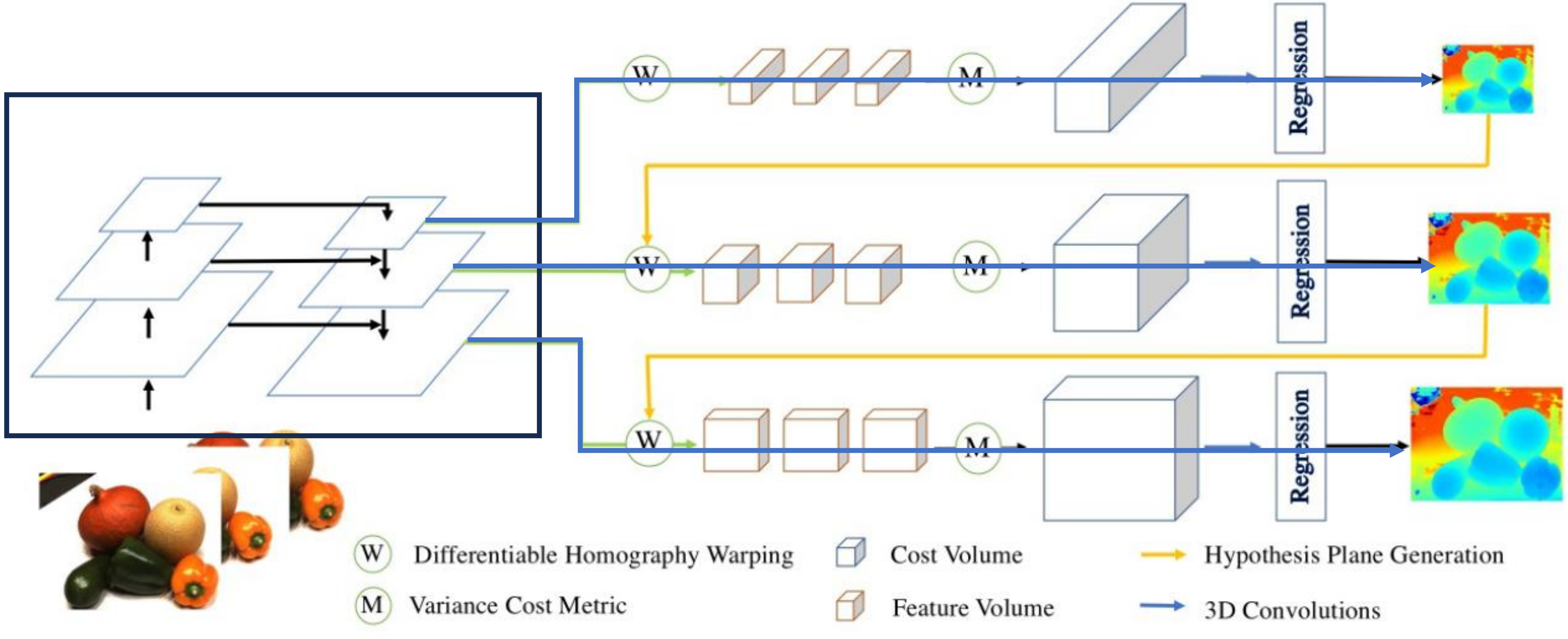


Error

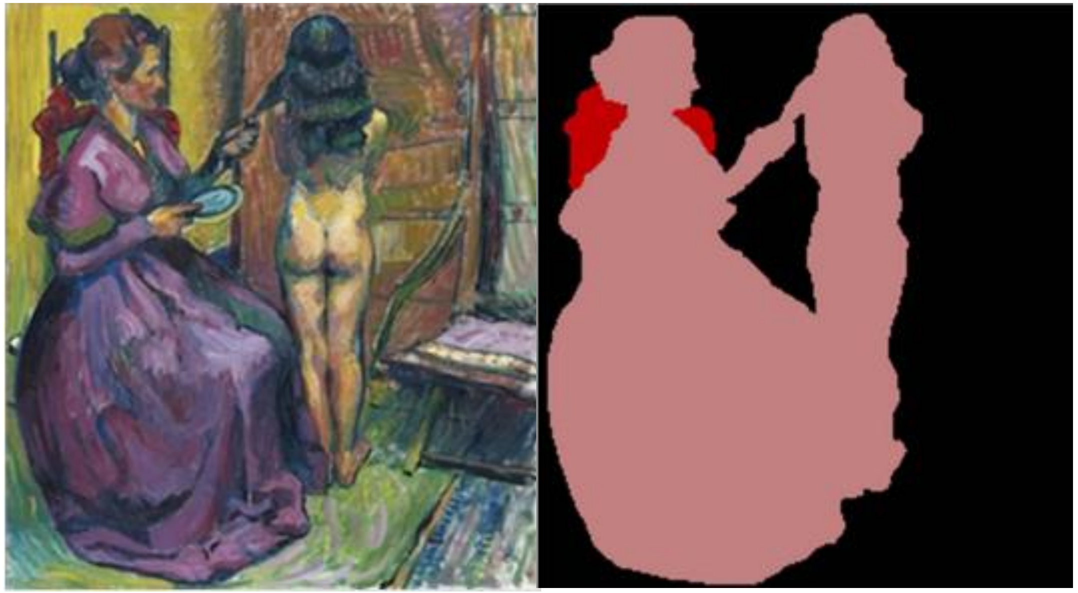
Figure: MVSNet. Source: Yao et al. (2018)

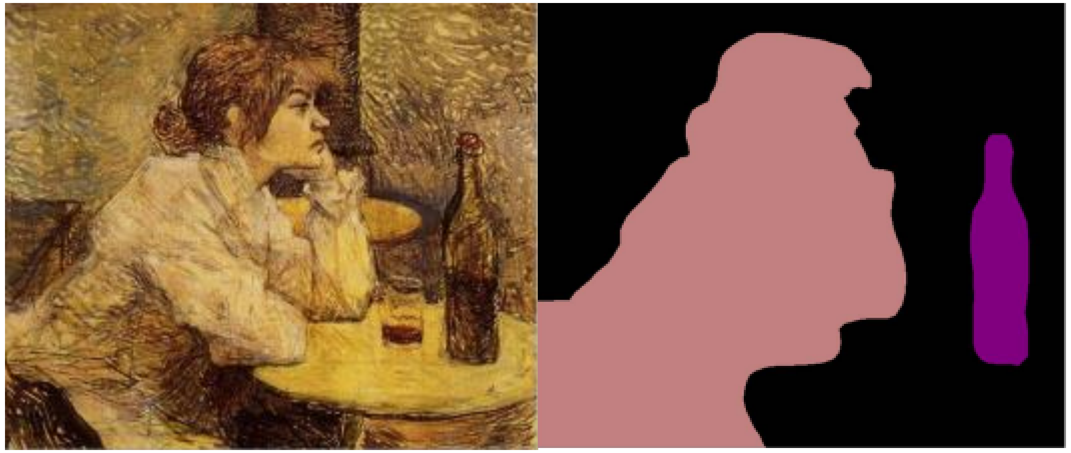
Related Work: Learning-based MVS

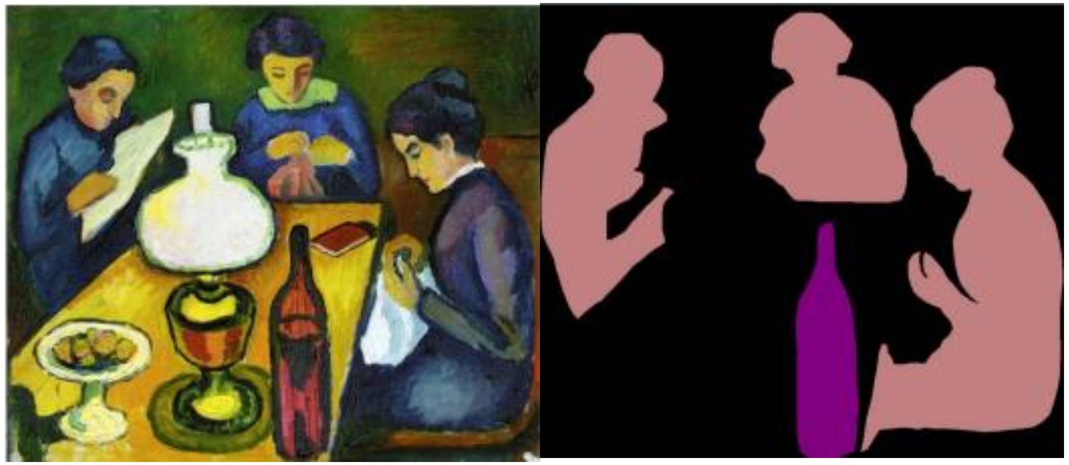
- Cascaded MVS Network



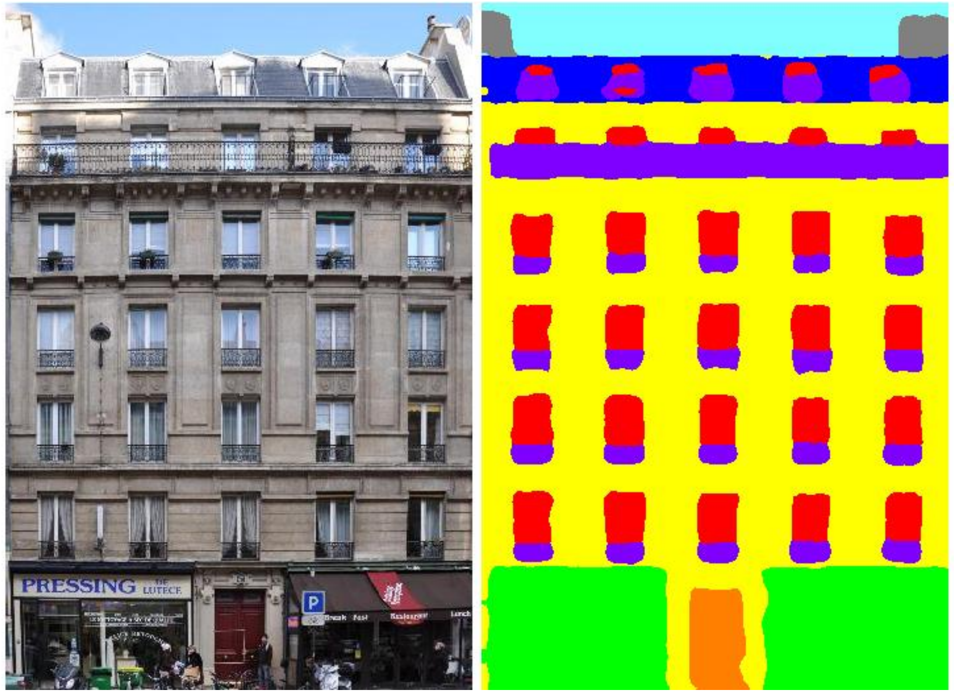
Related Work: Semantic Segmentation

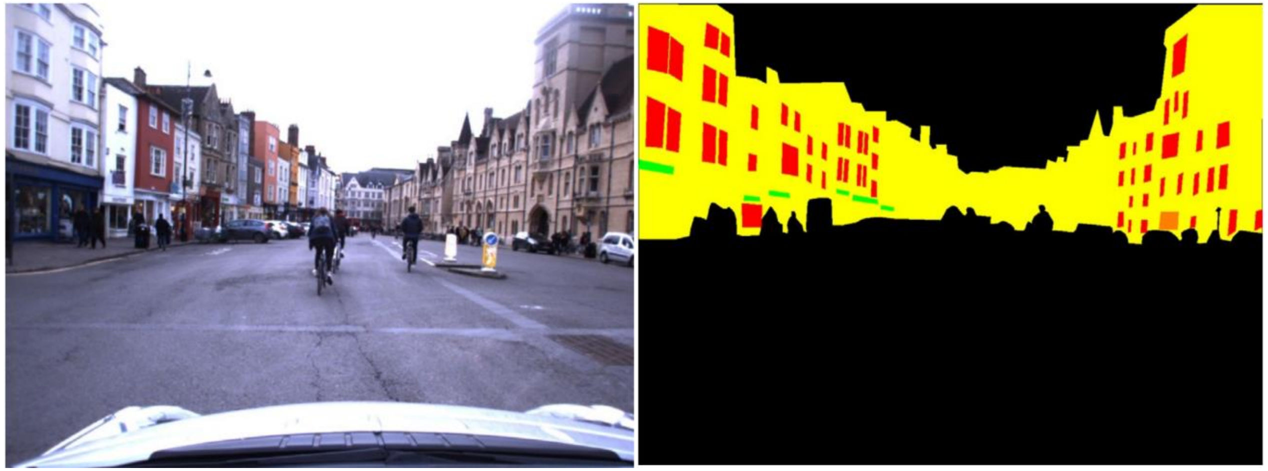






Related Work: Facade Parsing





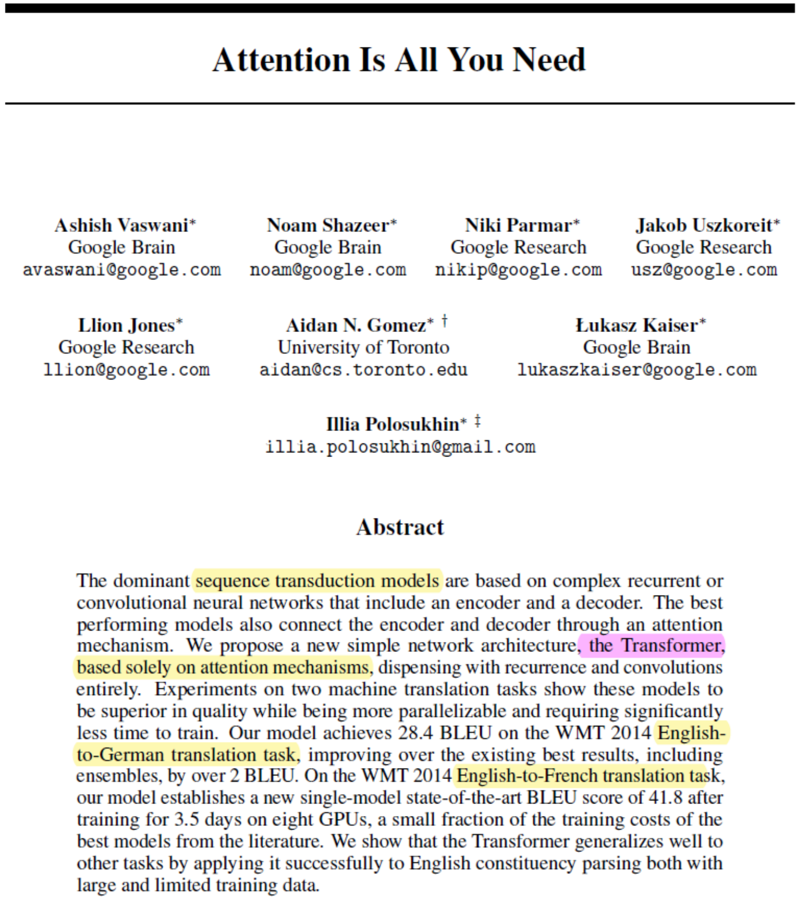
Related Work:
Transformers in Natural Language Processing

Self-attention mechanism enables to:

- Capture meaning
- Determine position in a sentence
- Analyse how each word interacts with other words in long sequences of text

*“**Meaning** is a result of **relationships** between things, and **self-attention** is a general **way of learning relationships**.”* (Vaswani)

Input sentence to translate:
*‘I poured water from the bottle into the **cup** until **it** was **full**.’*
*‘I poured water from the **bottle** into the cup until **it** was **empty**.’*



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

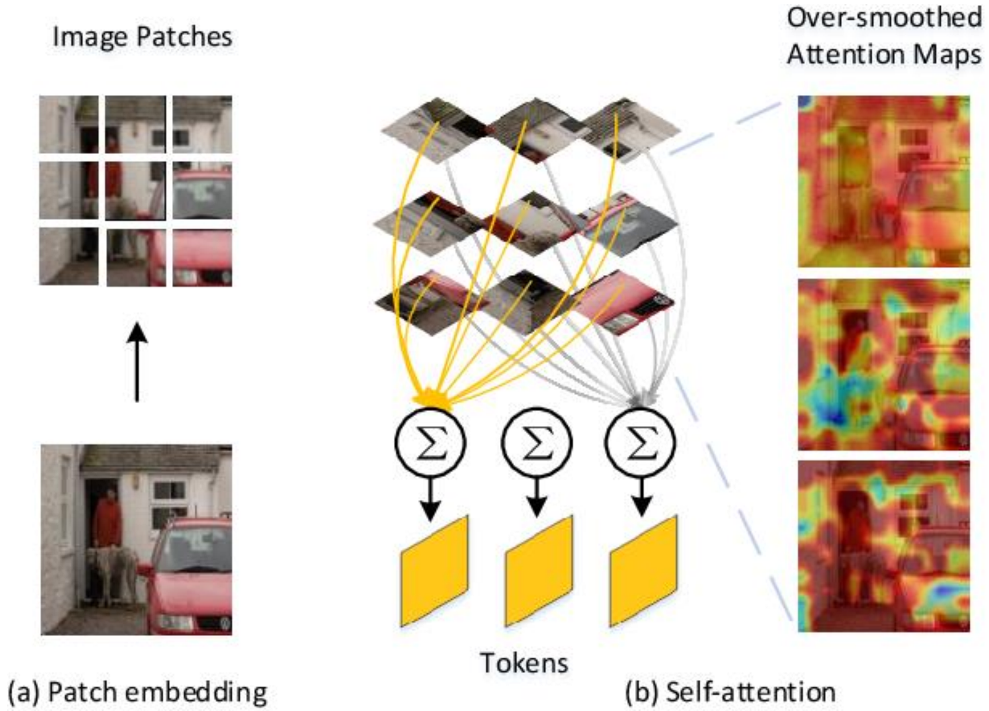
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Related Work:
Semantic Segmentation using Vision Transformers

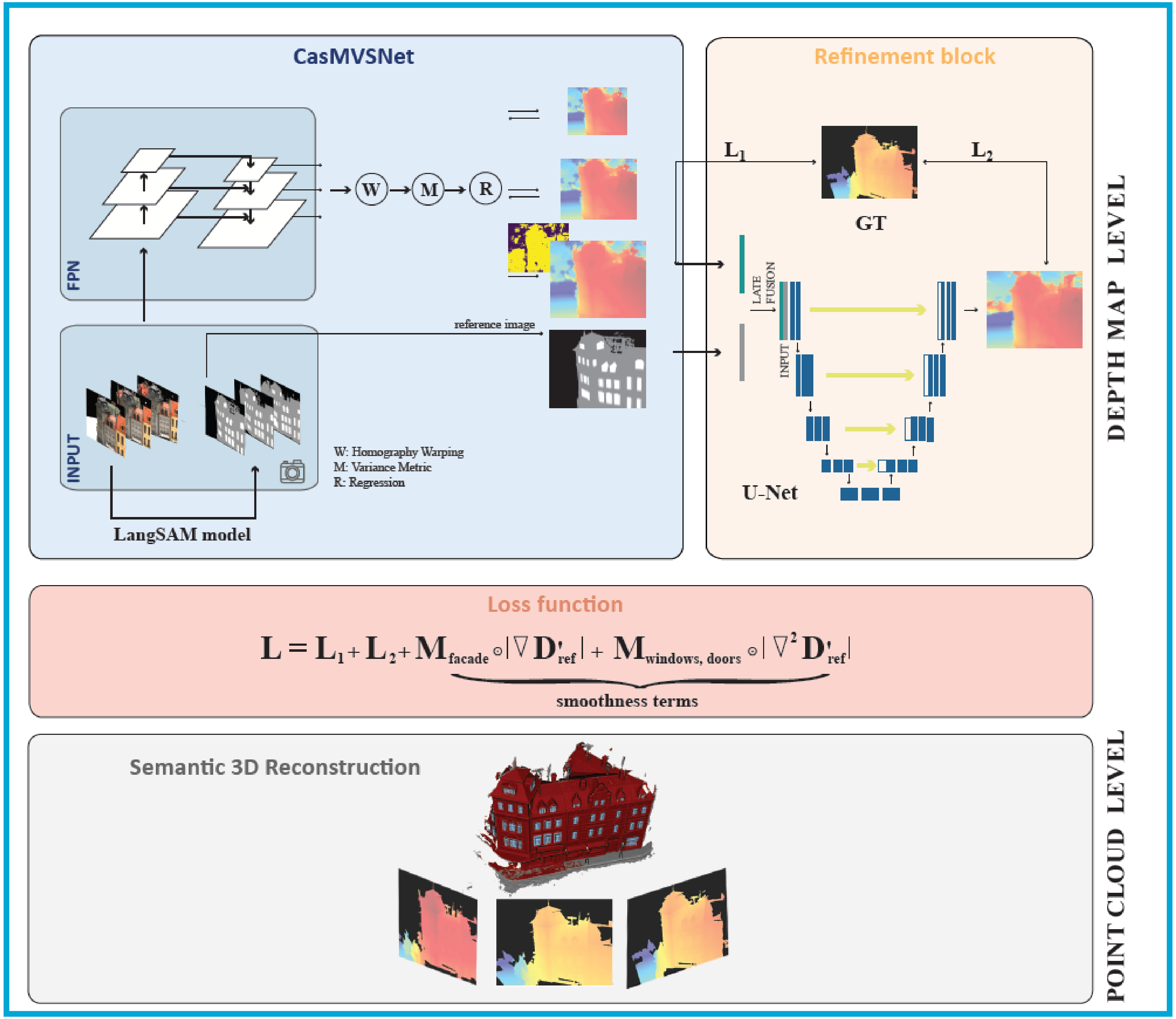
Vision Transformers:

- Transformers adapted for images
- **self-attention** mechanisms
  - Capture long-range dependencies



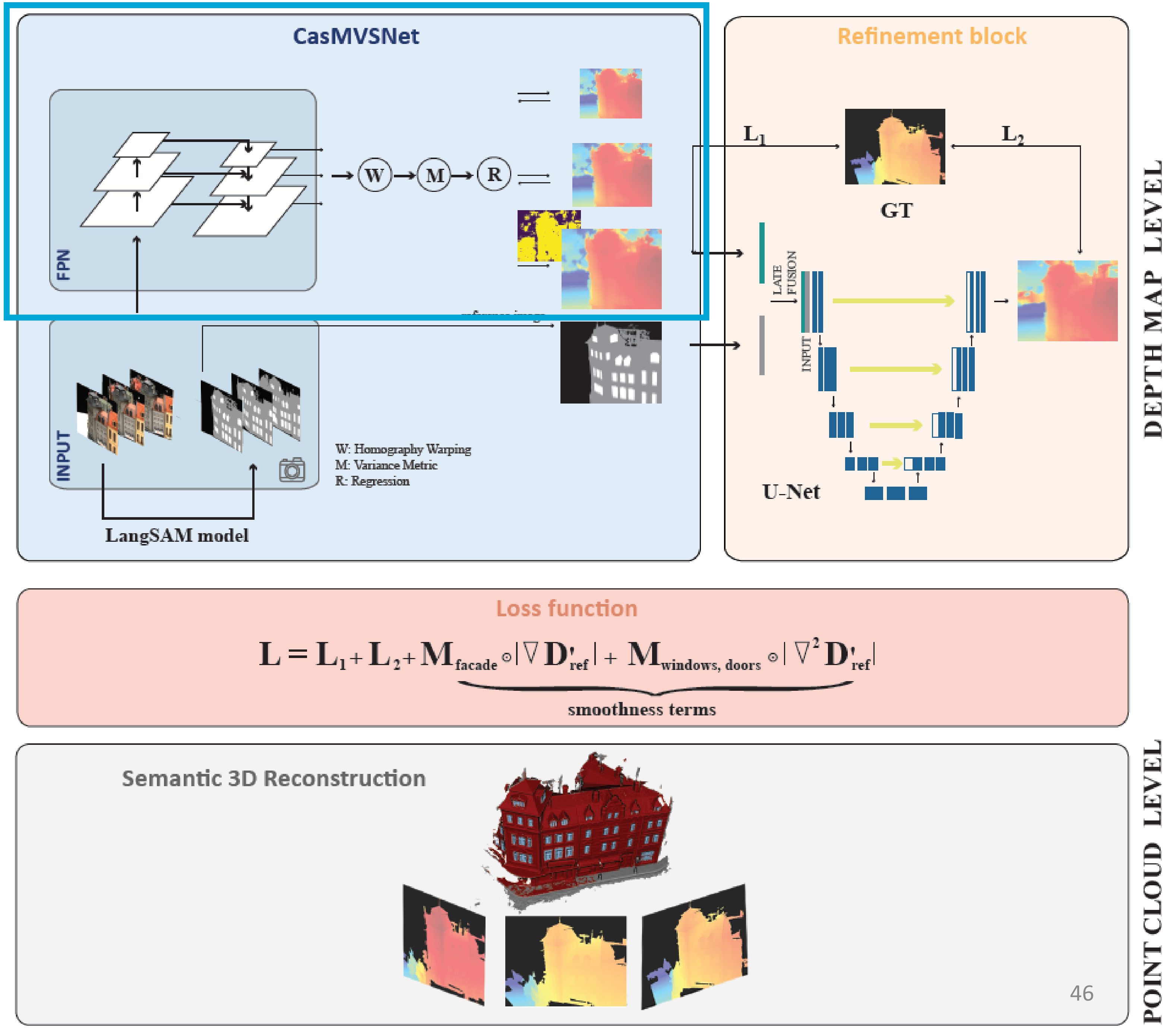
Methodology:

Semantic MVS



$$L = \underbrace{L_1 + L_2 + M_{facade} \odot |\nabla D'_{ref}| + M_{windows,\ doors} \odot |\nabla^2 D'_{ref}|}_{\text{smoothness terms}}$$

Methodology: Semantic MVS

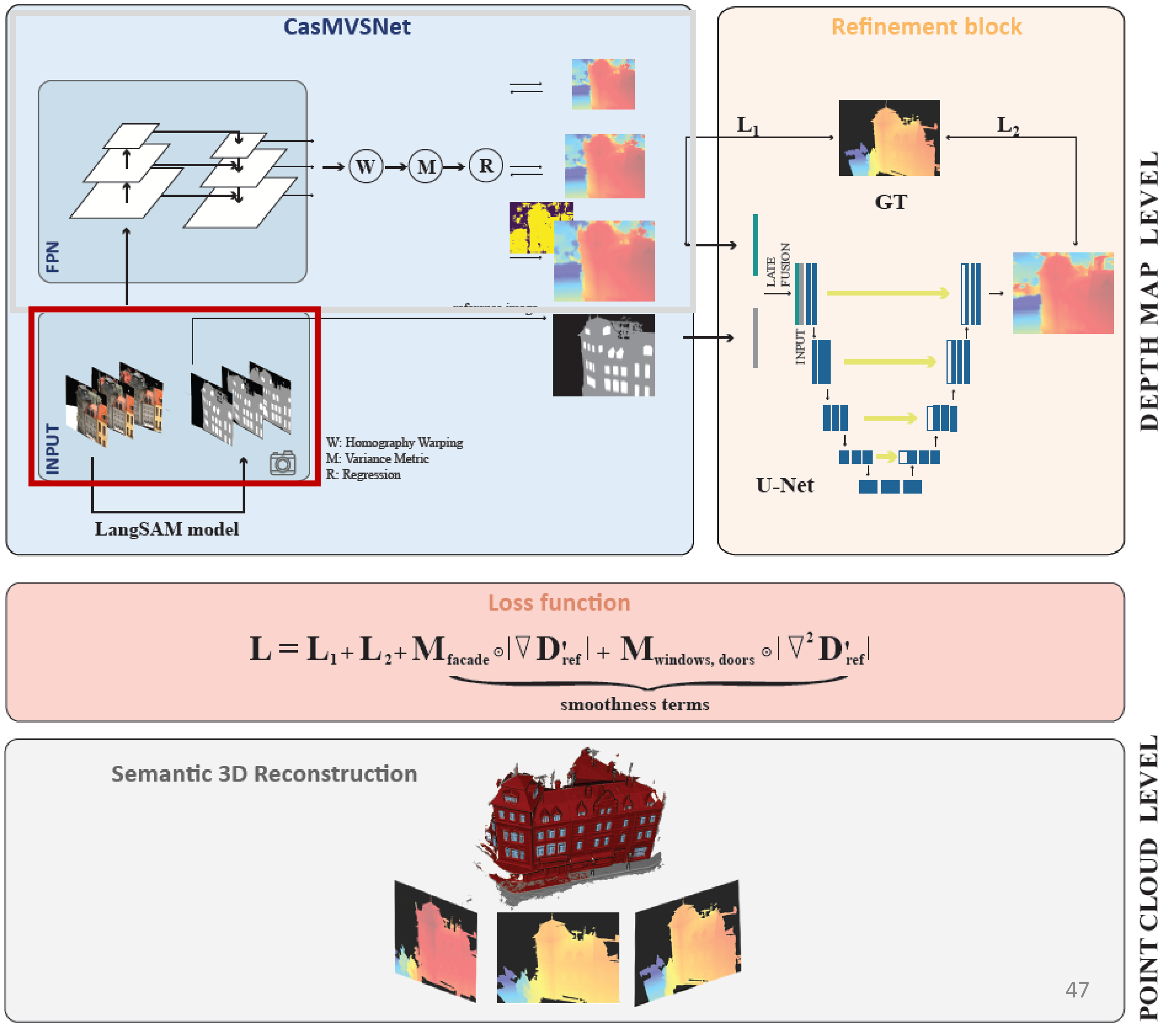


Loss function

$$\mathbf{L} = \mathbf{L}_1 + \mathbf{L}_2 + \underbrace{\mathbf{M}_{\text{facade}} \odot |\nabla \mathbf{D}'_{\text{ref}}| + \mathbf{M}_{\text{windows, doors}} \odot |\nabla^2 \mathbf{D}'_{\text{ref}}|}_{\text{smoothness terms}}$$

Methodology:
Semantic MVS

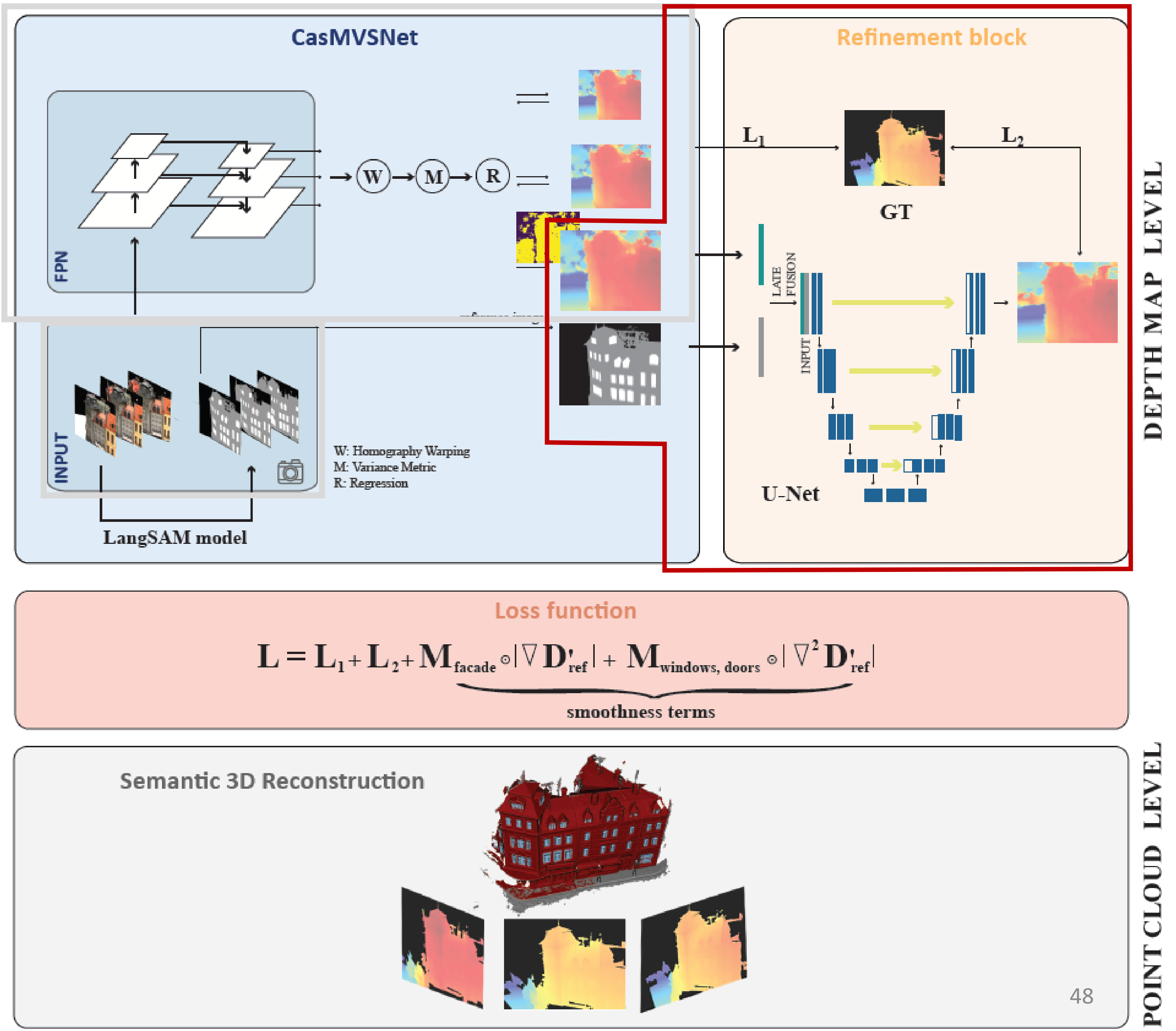
1. Semantics as input to MVS



$$L = L_1 + L_2 + \underbrace{M_{facade} \odot |\nabla D'_{ref}| + M_{windows,\ doors} \odot |\nabla^2 D'_{ref}|}_{\text{smoothness terms}}$$

Methodology: Semantic MVS

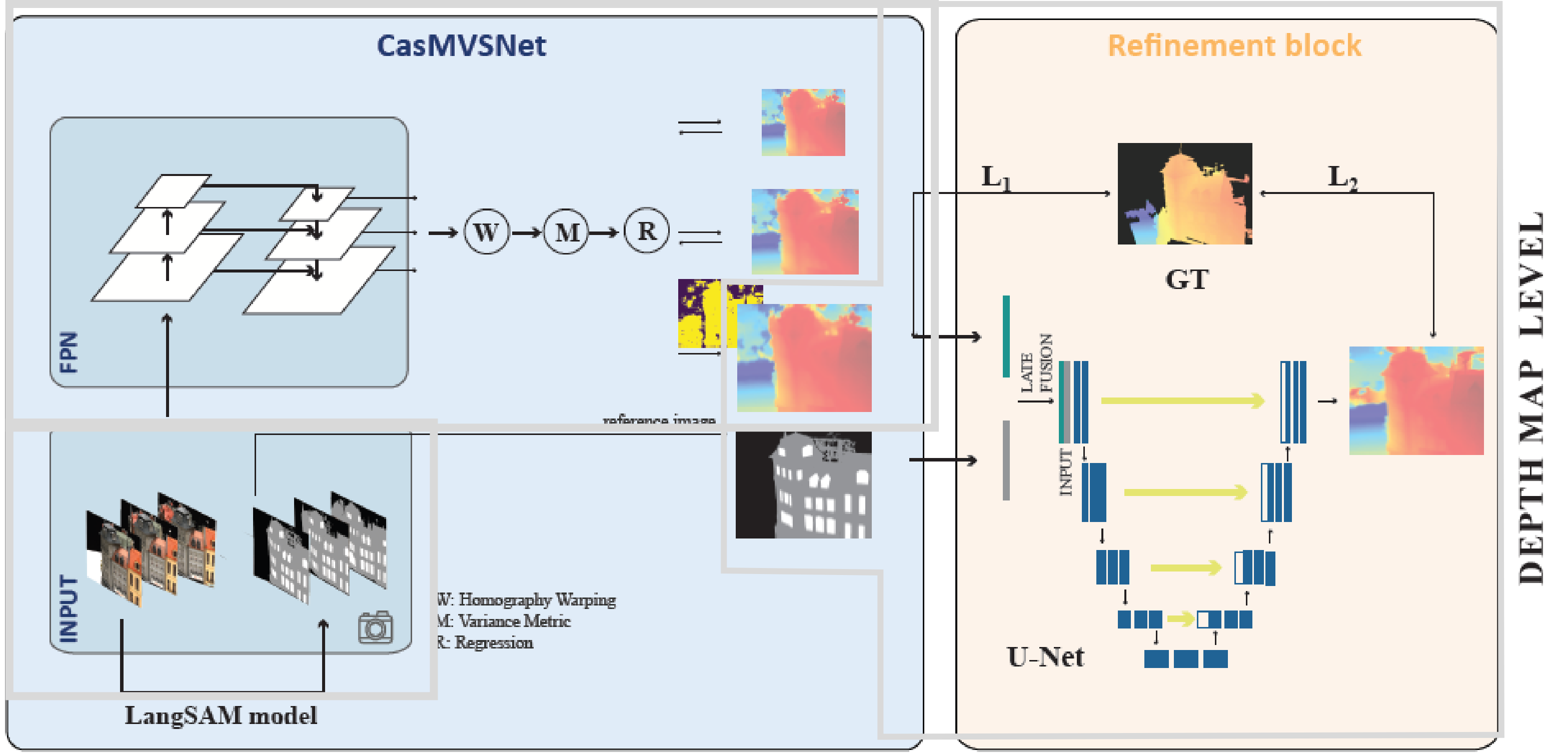
1. Semantics as input to MVS
2. Refinement Block



$$L = L_1 + L_2 + \underbrace{M_{facade} \odot |\nabla D'_{ref}| + M_{windows,\, doors} \odot |\nabla^2 D'_{ref}|}_{\text{smoothness terms}}$$

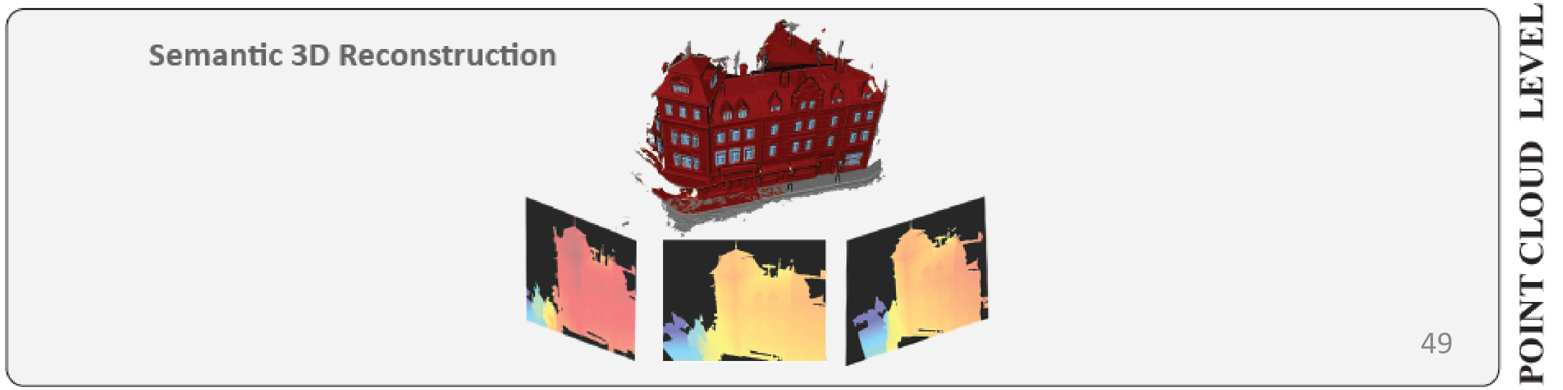
Methodology: Semantic MVS

1. Semantics as input to MVS
2. Refinement Block
3. Loss Function



Loss function

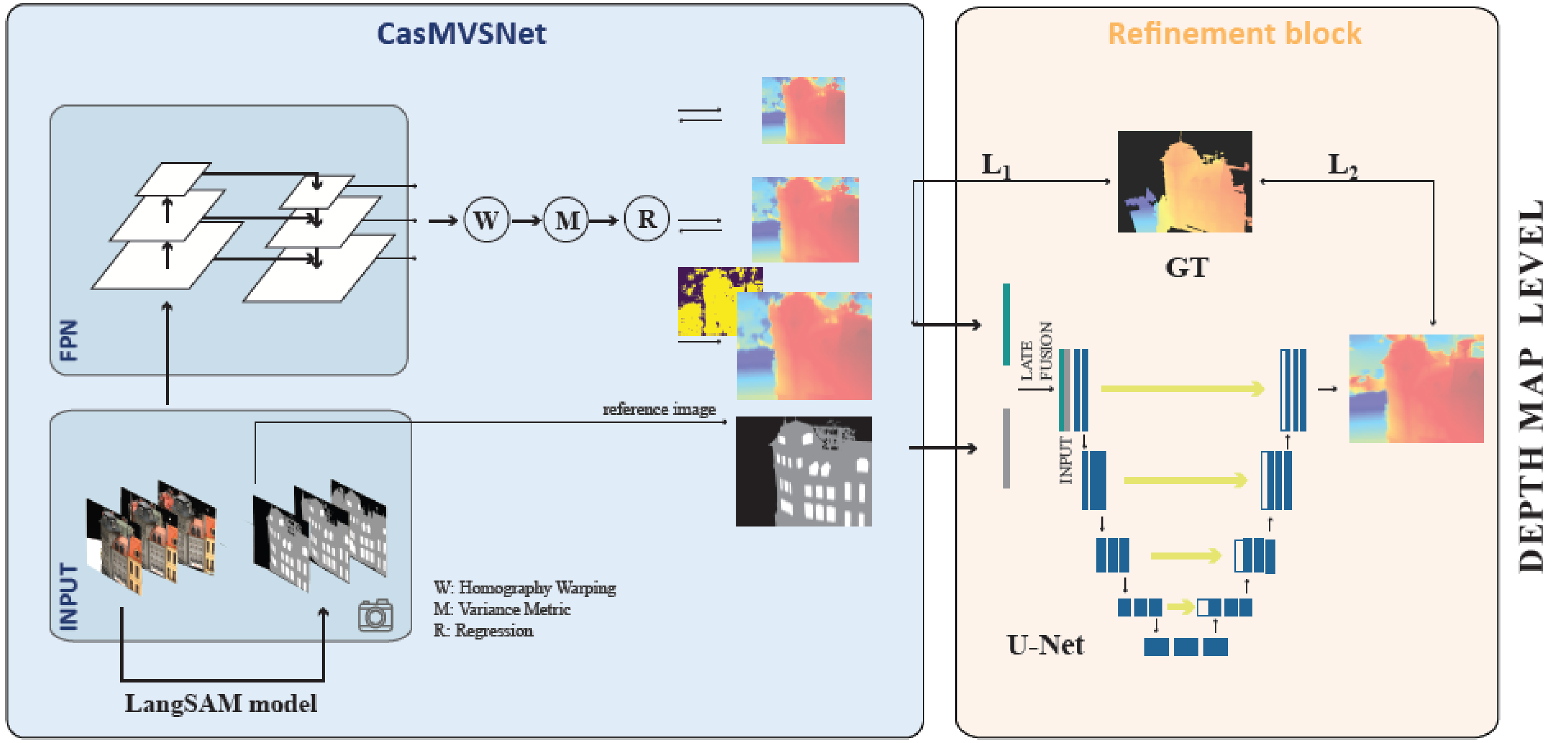
$$L = L_1 + L_2 + \underbrace{M_{facade} \odot |\nabla D'_{ref}| + M_{windows,\ doors} \odot |\nabla^2 D'_{ref}|}_{\text{smoothness terms}}$$



Methodology: Semantic MVS

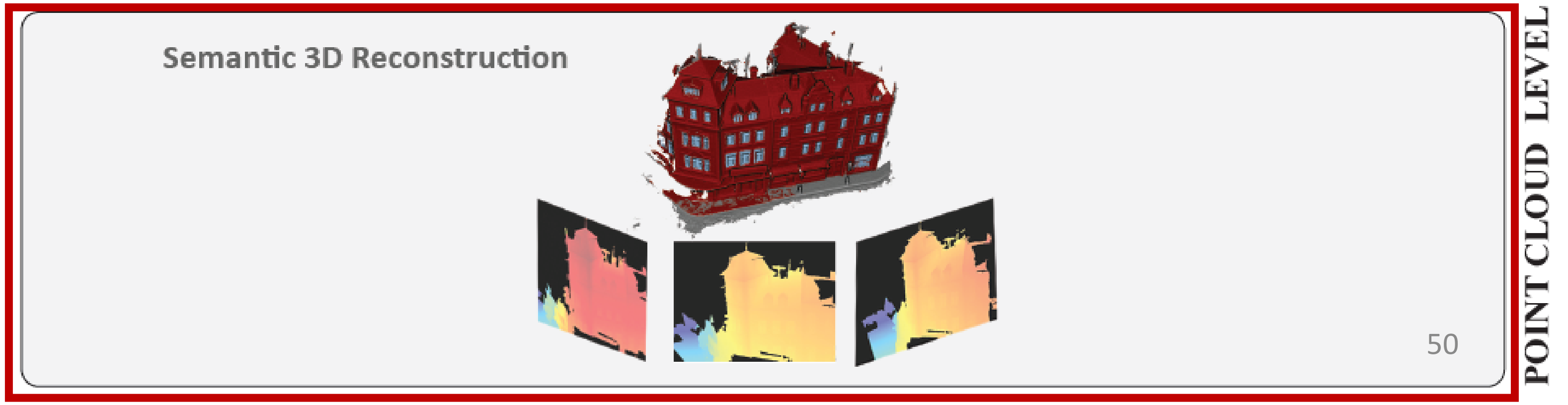
1. Semantics as input to MVS
2. Refinement Block
3. Loss Function

- Semantic Point Cloud Reconstruction



Loss function

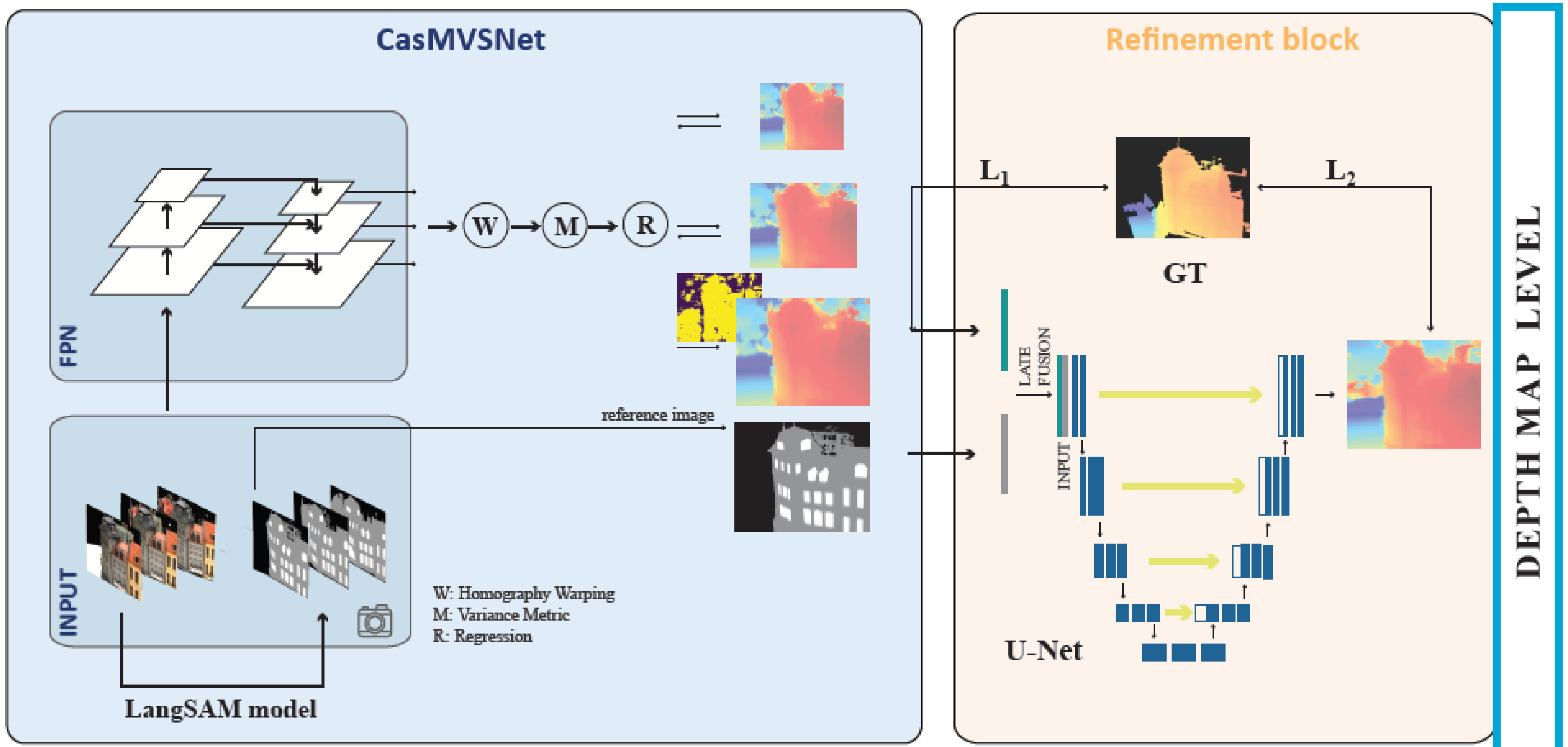
$$L = L_1 + L_2 + \underbrace{M_{facade} \odot |\nabla D'_{ref}| + M_{windows,\ doors} \odot |\nabla^2 D'_{ref}|}_{\text{smoothness terms}}$$



Methodology: Semantic MVS

1. Semantics as input to MVS
2. Refinement Block
3. Loss Function

- Semantic Point Cloud Reconstruction

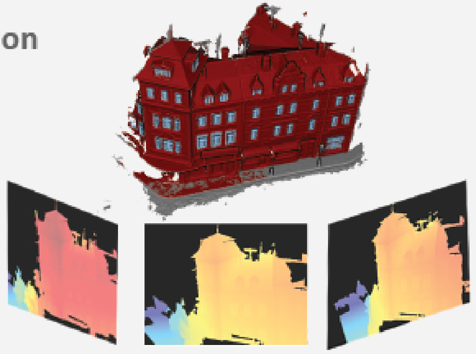


DEPTH MAP LEVEL

Loss function

$$L = L_1 + L_2 + \underbrace{M_{facade} \odot |\nabla D'_{ref}| + M_{windows,\ doors} \odot |\nabla^2 D'_{ref}|}_{\text{smoothness terms}}$$

Semantic 3D Reconstruction

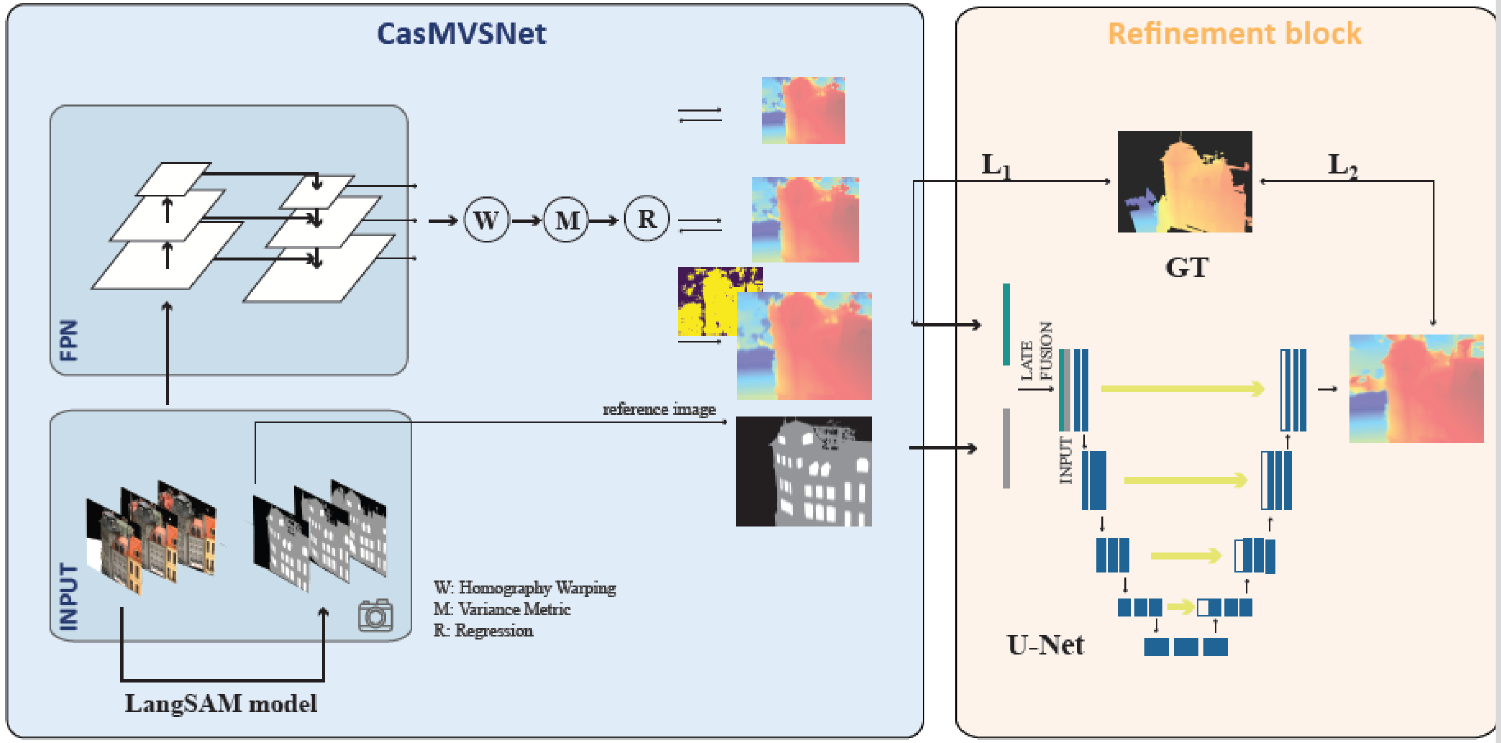


POINT CLOUD LEVEL

Methodology: Semantic MVS

1. Semantics as input to MVS
2. Refinement Block
3. Loss Function

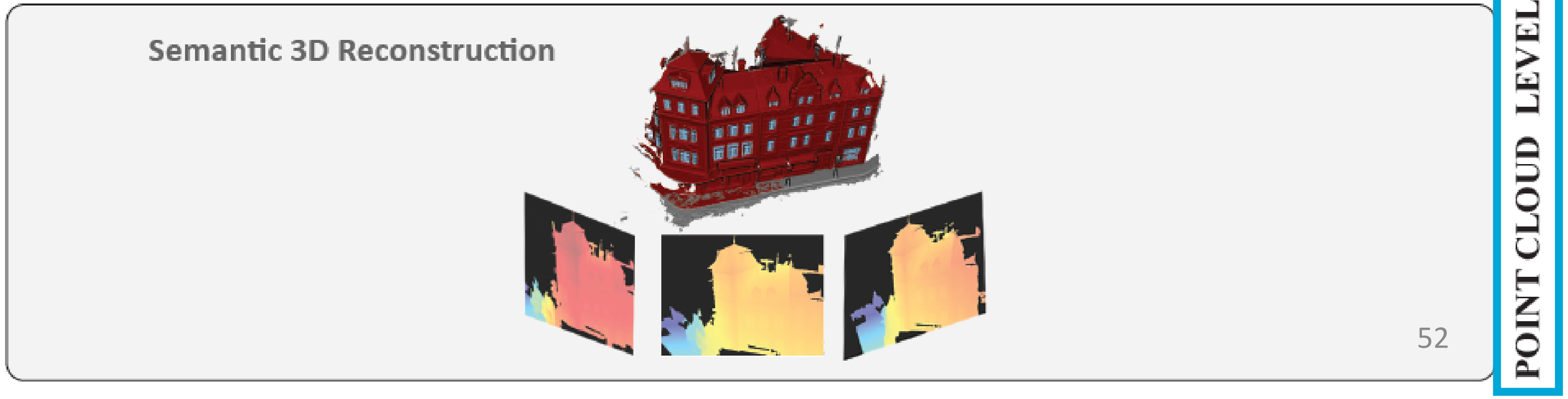
- Semantic Point Cloud Reconstruction



DEPTH MAP LEVEL

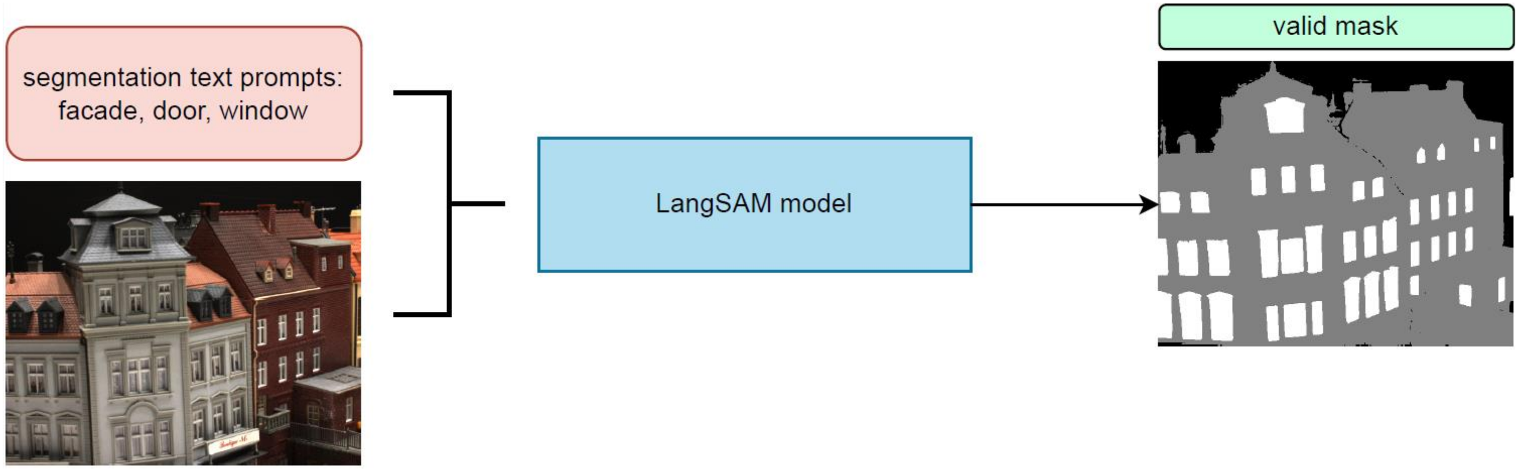
Loss function

$$L = L_1 + L_2 + \underbrace{M_{facade} \odot |\nabla D'_{ref}| + M_{windows,\ doors} \odot |\nabla^2 D'_{ref}|}_{\text{smoothness terms}}$$



POINT CLOUD LEVEL

Methodology: Semantic Segmentation



Implementation: Depth Datasets

- DTU Dataset
  - used for training and evaluation
  - **only subset pertaining to buildings !**
  - Small objects (<0.5m) shot in a laboratory setting

Source: https://roboimagedata.compute.dtu.dk/

- Facade ETH3D Dataset
  - Real-world outdoor data
  - used for generalization
  - Few meters to hundreds of meters

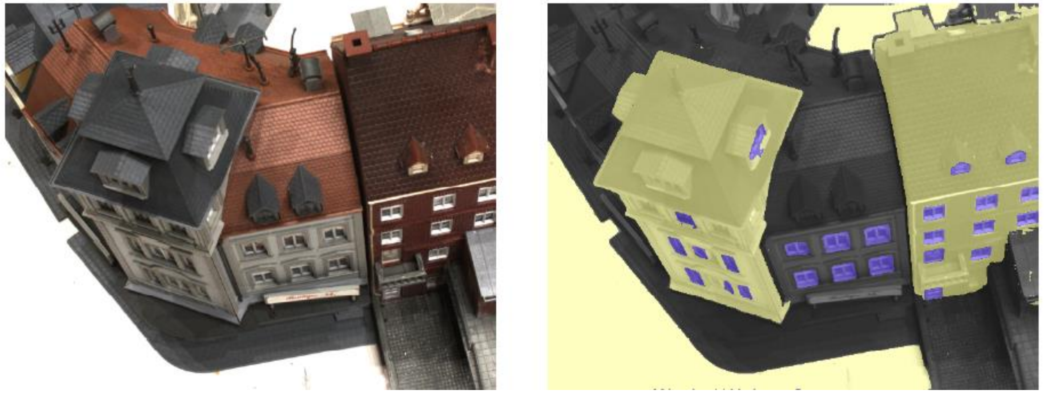
Source: https://www.eth3d.net/datasets

Results: Semantic Segmentation

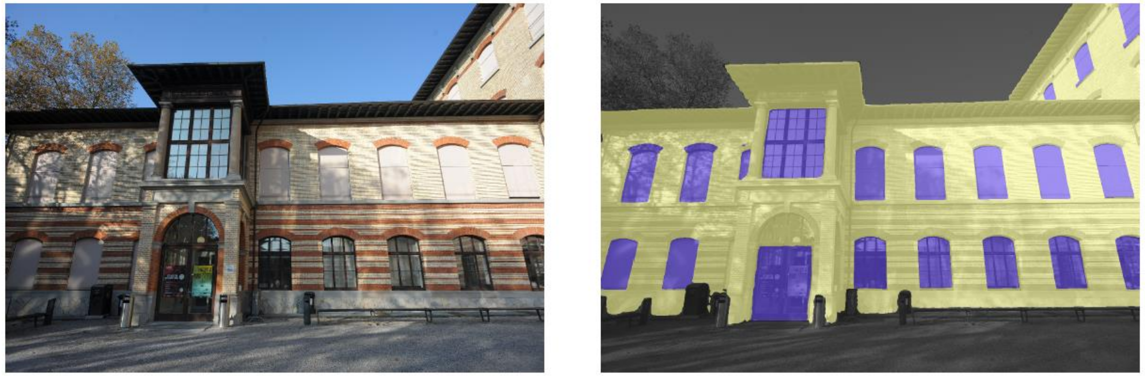
🟨 wall

🟪 window or door

DTU



ETH3D

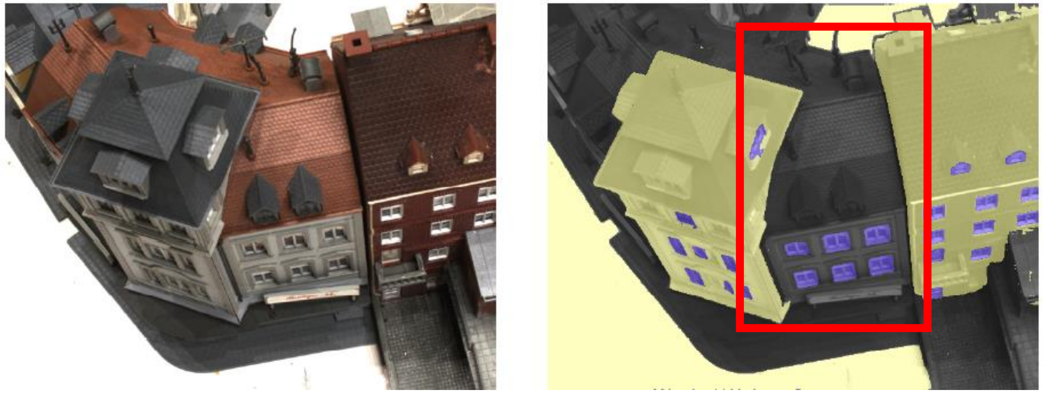


Results: Semantic Segmentation

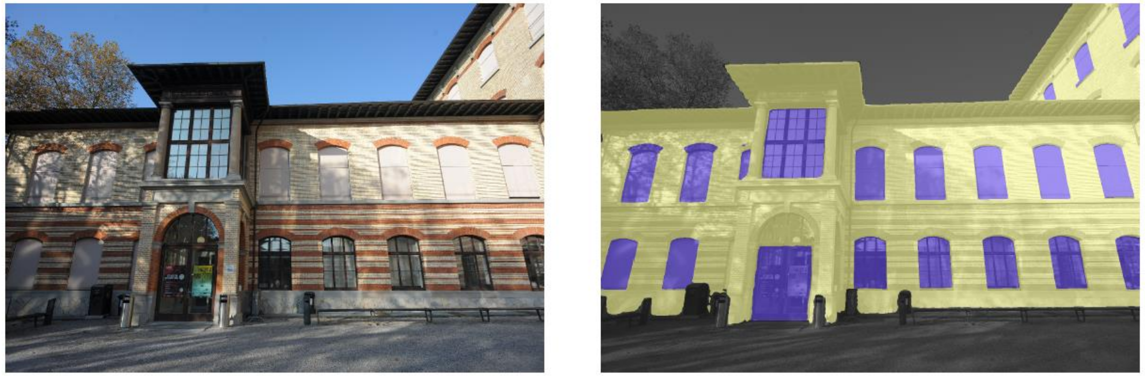
🟨 wall

🟪 window or door

DTU

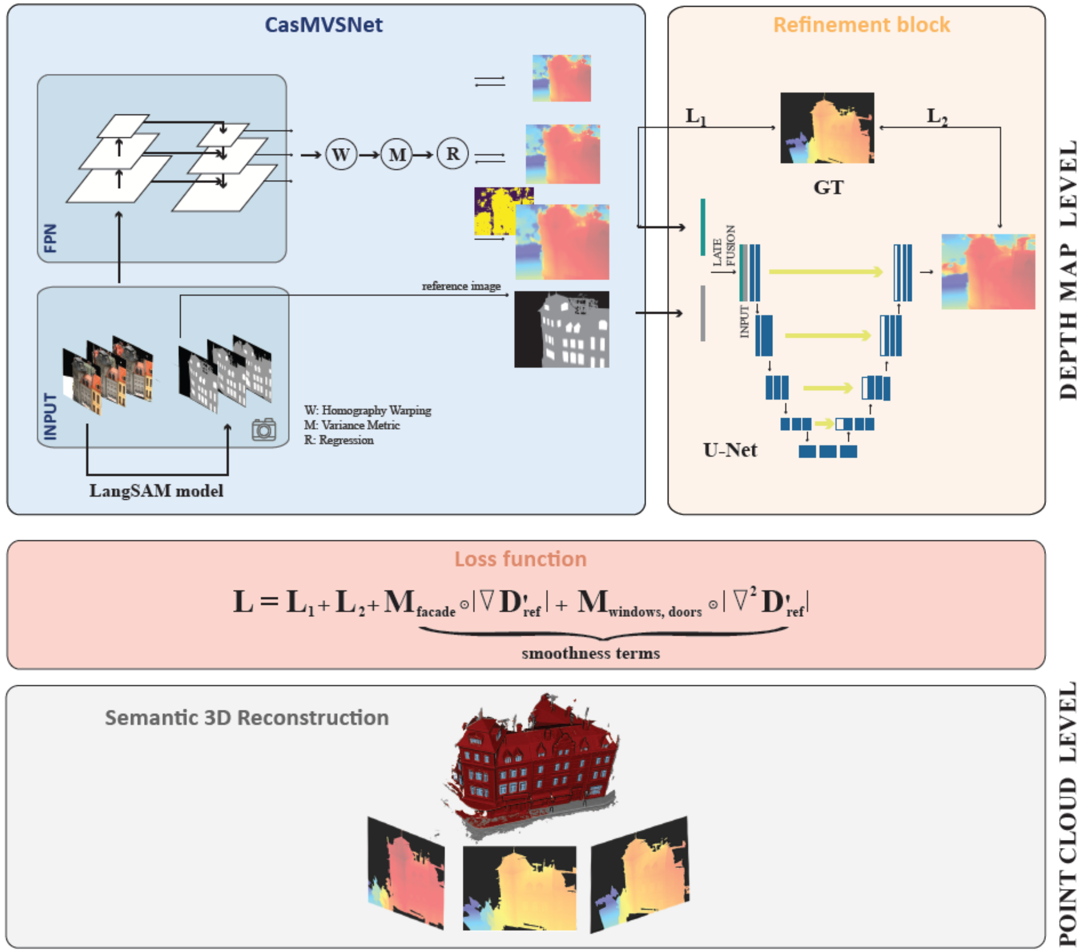


ETH3D



Experiments and Evaluation

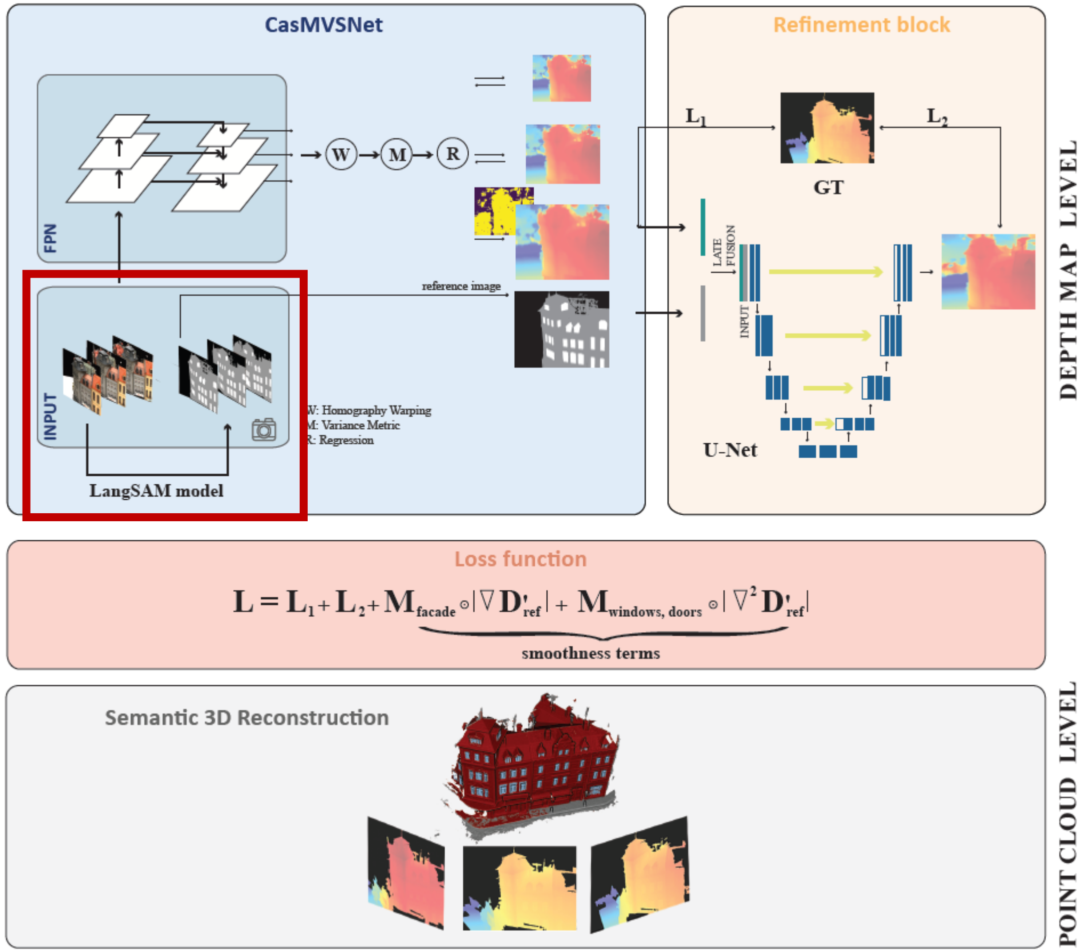
	Modules			
	FPN			Loss Function
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
rgb_FPN_RU-Net	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
rgb_AFPN_RU-Net	rgb	Attention-FPN	RU-Net	✓
srgb_AFPN_RU-Net	semantic + rgb	Attention-FPN	RU-Net	✓
rgb_AFPN_RAU-Net	rgb	Attention-FPN	RAU-Net	✓
rgb_AFPN_R2AU-Net	rgb	Attention-FPN	R2AU-Net	✓
rgb_FPN_CNN	rgb	FPN	CNN	✓
srgb_FPN_CNN	semantic + rgb	FPN	CNN	✓
Baseline Model (CasMVSNet)	rgb	FPN	No	No



$$L = L_1 + L_2 + \underbrace{M_{facade} \odot |\nabla D'_{ref}| + M_{windows,\ doors} \odot |\nabla^2 D'_{ref}|}_{\text{smoothness terms}}$$

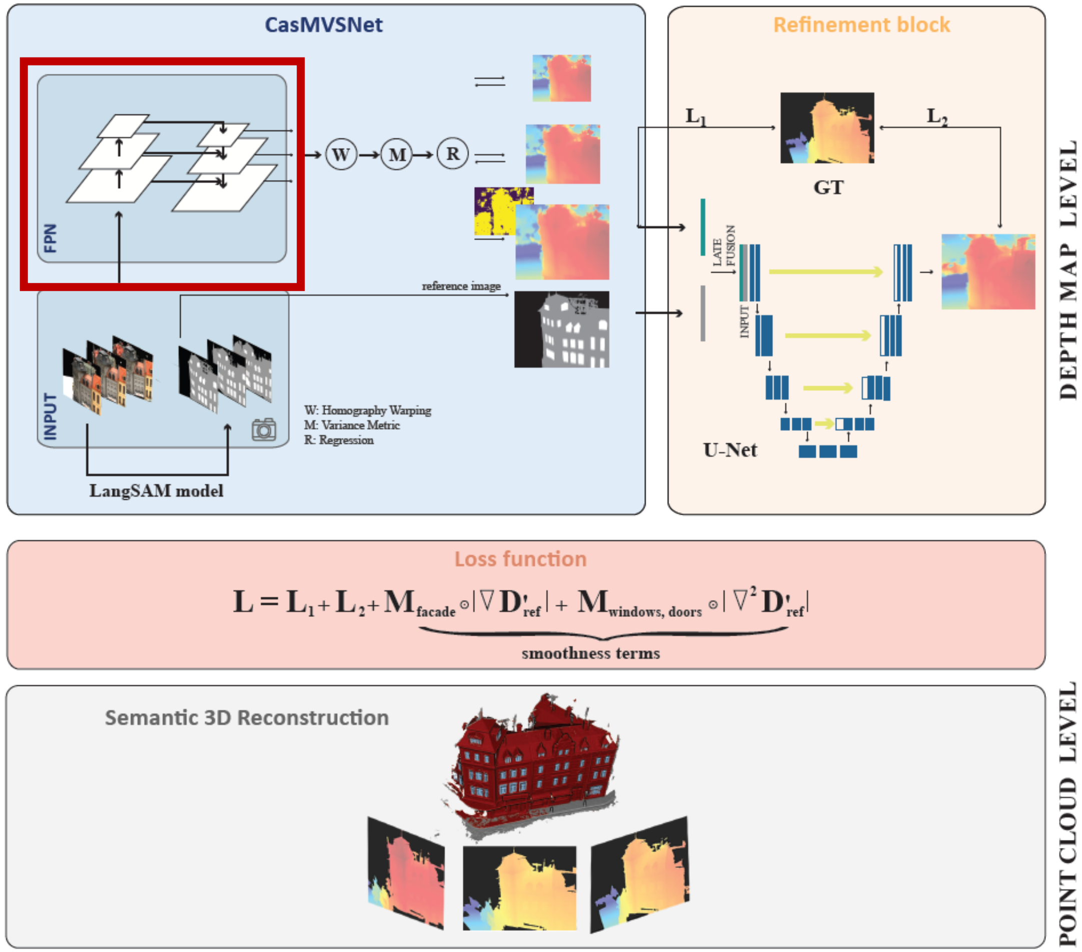
Experiments and Evaluation: Variations in the Input

	Modules			Loss Function
	FPN			
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
rgb_FPN_RU-Net	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
rgb_AFPN_RU-Net	rgb	Attention-FPN	RU-Net	✓
srgb_AFPN_RU-Net	semantic + rgb	Attention-FPN	RU-Net	✓
rgb_AFPN_RAU-Net	rgb	Attention-FPN	RAU-Net	✓
rgb_AFPN_R2AU-Net	rgb	Attention-FPN	R2AU-Net	✓
rgb_FPN_CNN	rgb	FPN	CNN	✓
srgb_FPN_CNN	semantic + rgb	FPN	CNN	✓
Baseline Model (CasMVSNet)	rgb	FPN	No	No



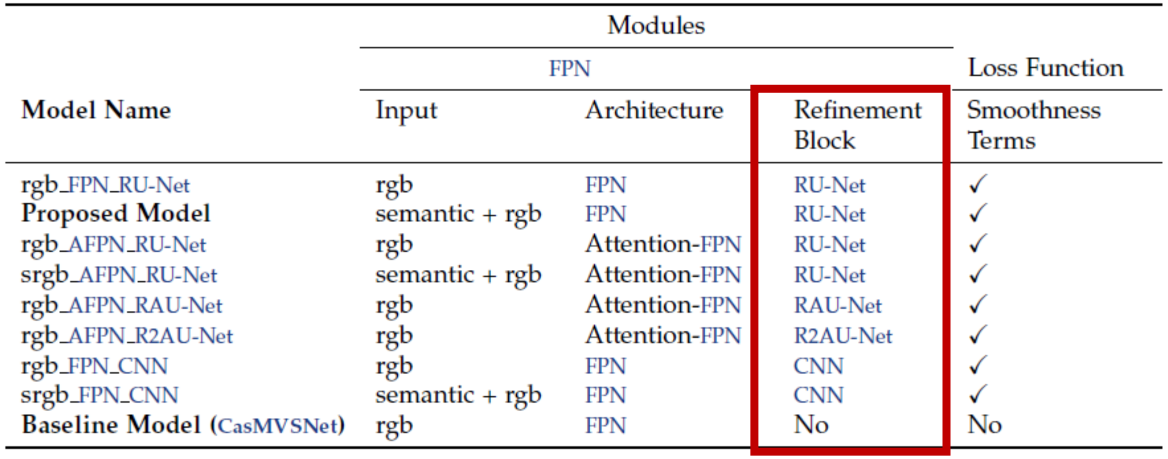
Experiments and Evaluation: Variations in the Feature Extraction

	Modules			Loss Function
	FPN			
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
rgb_FPN_RU-Net	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
rgb_AFPN_RU-Net	rgb	Attention-FPN	RU-Net	✓
srgb_AFPN_RU-Net	semantic + rgb	Attention-FPN	RU-Net	✓
rgb_AFPN_RAU-Net	rgb	Attention-FPN	RAU-Net	✓
rgb_AFPN_R2AU-Net	rgb	Attention-FPN	R2AU-Net	✓
rgb_FPN_CNN	rgb	FPN	CNN	✓
srgb_FPN_CNN	semantic + rgb	FPN	CNN	✓
Baseline Model (CasMVSNet)	rgb	FPN	No	No

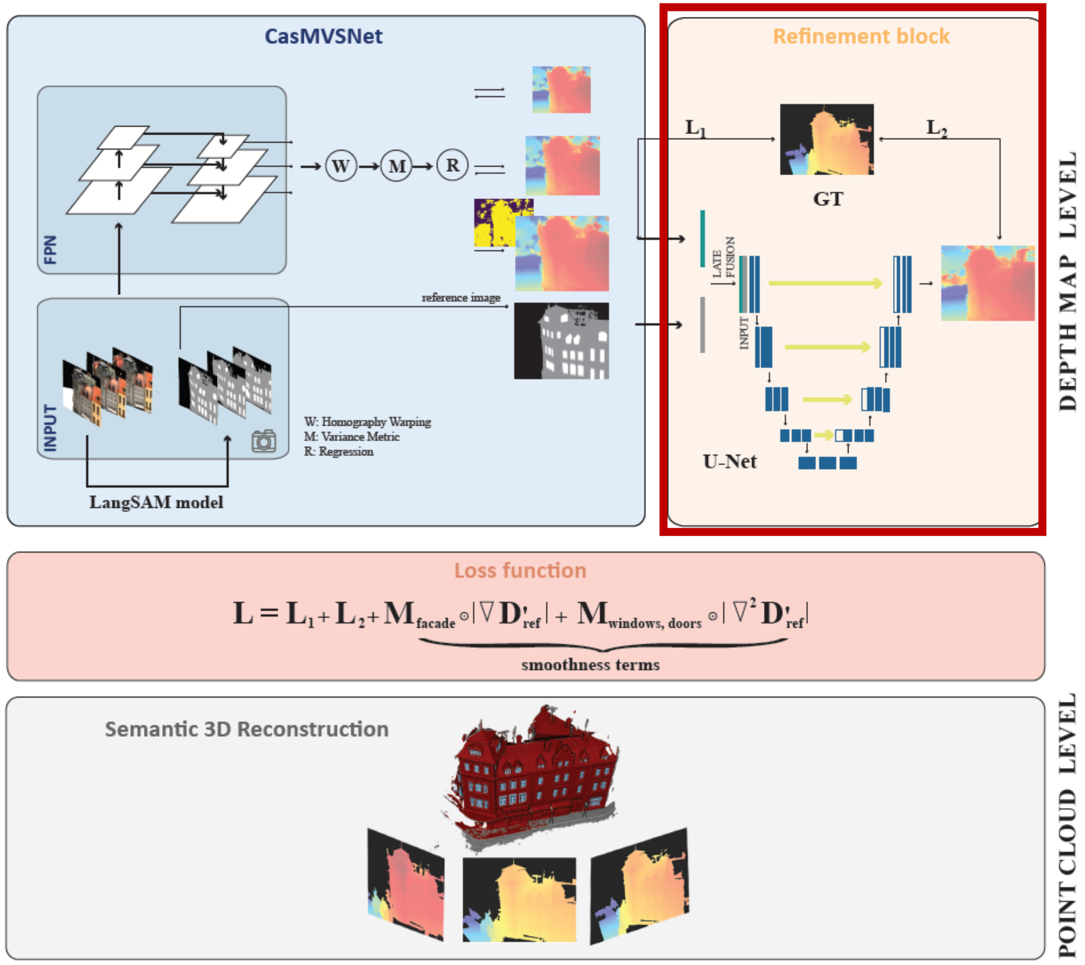


$$L = L_1 + L_2 + \underbrace{M_{facade} \odot |\nabla D'_{ref}| + M_{windows,\ doors} \odot |\nabla^2 D'_{ref}|}_{\text{smoothness terms}}$$

Experiments and Evaluation: Variations in the Refinement Block



	Modules			Loss Function
	FPN			
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
rgb_FPN_RU-Net	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
rgb_AFPN_RU-Net	rgb	Attention-FPN	RU-Net	✓
srgb_AFPN_RU-Net	semantic + rgb	Attention-FPN	RU-Net	✓
rgb_AFPN_RAU-Net	rgb	Attention-FPN	RAU-Net	✓
rgb_AFPN_R2AU-Net	rgb	Attention-FPN	R2AU-Net	✓
rgb_FPN_CNN	rgb	FPN	CNN	✓
srgb_FPN_CNN	semantic + rgb	FPN	CNN	✓
Baseline Model (CasMVSNet)	rgb	FPN	No	No

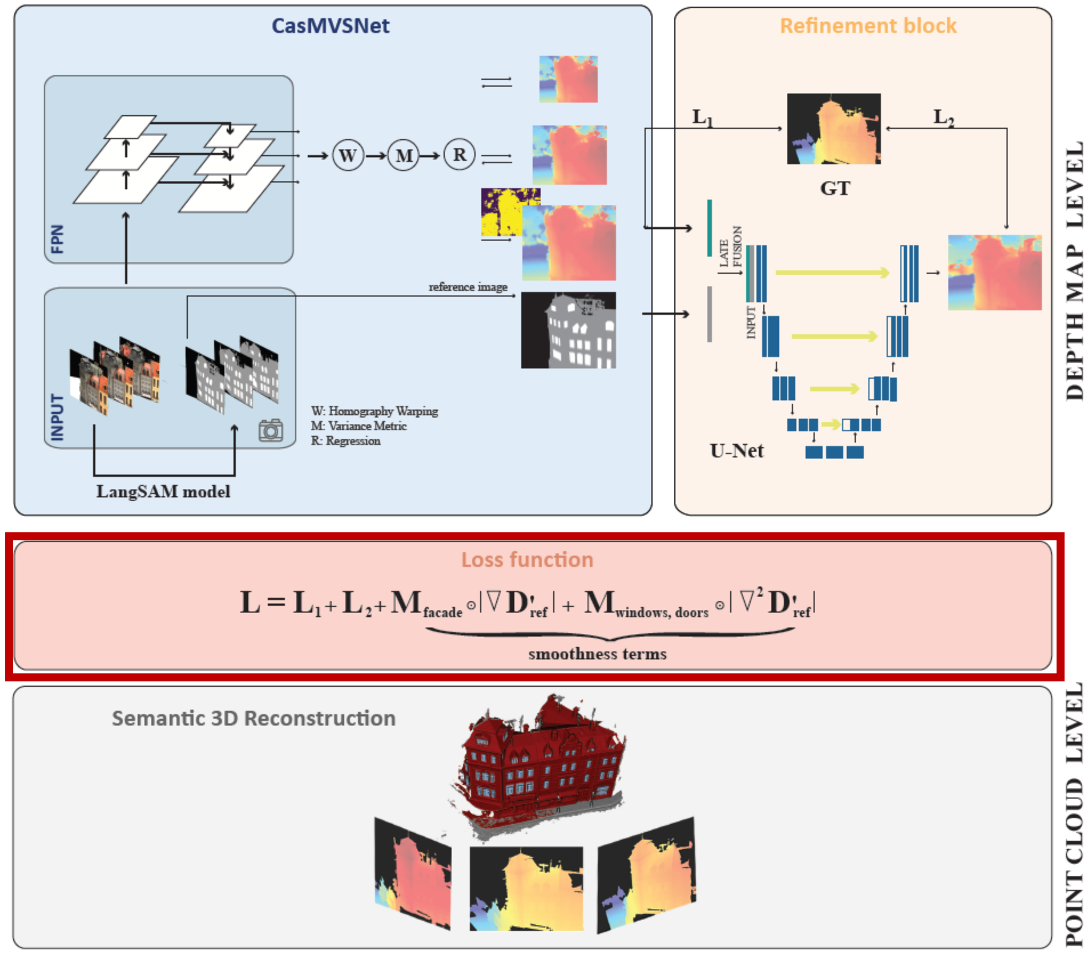


$$\mathbf{L} = \mathbf{L}_1 + \mathbf{L}_2 + \underbrace{\mathbf{M}_{facade} \odot |\nabla \mathbf{D}'_{ref}| + \mathbf{M}_{windows,\ doors} \odot |\nabla^2 \mathbf{D}'_{ref}|}_{\text{smoothness terms}}$$

Experiments and Evaluation: Smoothness terms

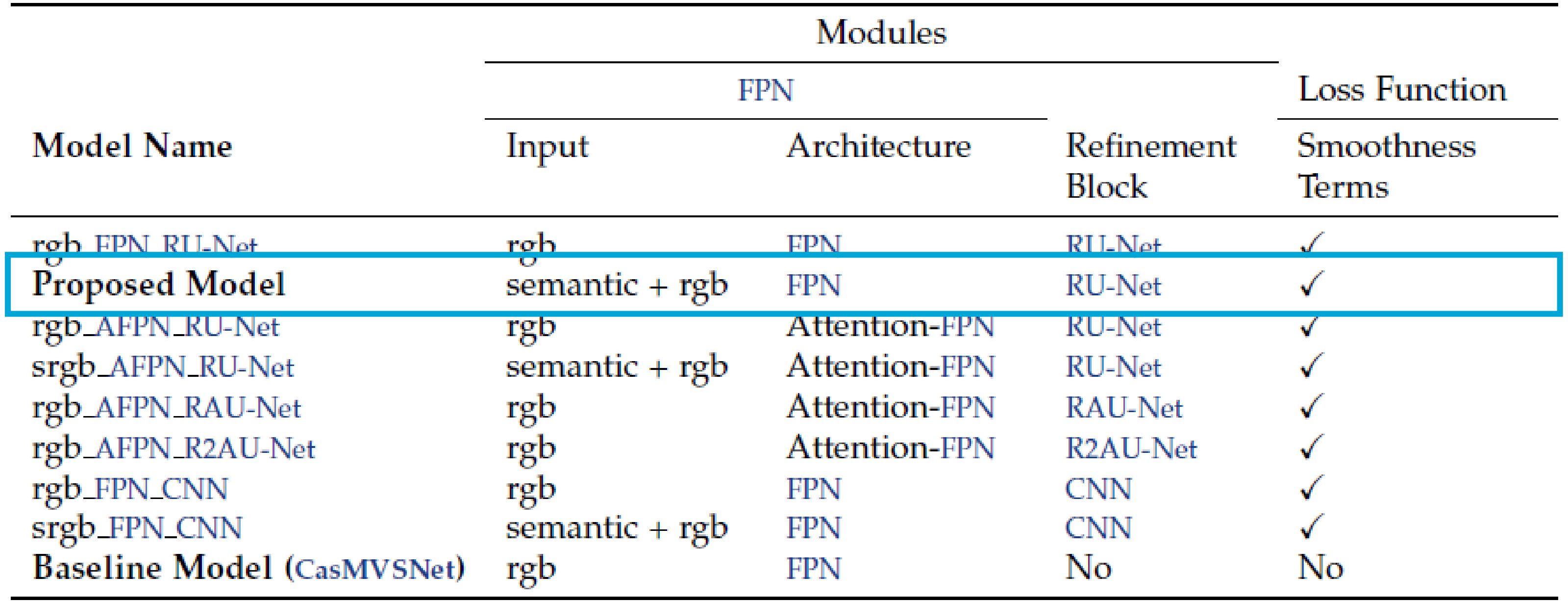
	Modules			
	FPN			Loss Function *
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
rgb_FPN_RU-Net	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
rgb_AFPN_RU-Net	rgb	Attention-FPN	RU-Net	✓
srgb_AFPN_RU-Net	semantic + rgb	Attention-FPN	RU-Net	✓
rgb_AFPN_RAU-Net	rgb	Attention-FPN	RAU-Net	✓
rgb_AFPN_R2AU-Net	rgb	Attention-FPN	R2AU-Net	✓
rgb_FPN_CNN	rgb	FPN	CNN	✓
srgb_FPN_CNN	semantic + rgb	FPN	CNN	✓
Baseline Model (CasMVSNet)	rgb	FPN	No	No

* Each experiment incorporated the two smoothness loss terms.



$$L = L_1 + L_2 + \underbrace{M_{facade} \odot |\nabla D'_{ref}| + M_{windows,\ doors} \odot |\nabla^2 D'_{ref}|}_{\text{smoothness terms}}$$

Experiments and Evaluation: Model Selection

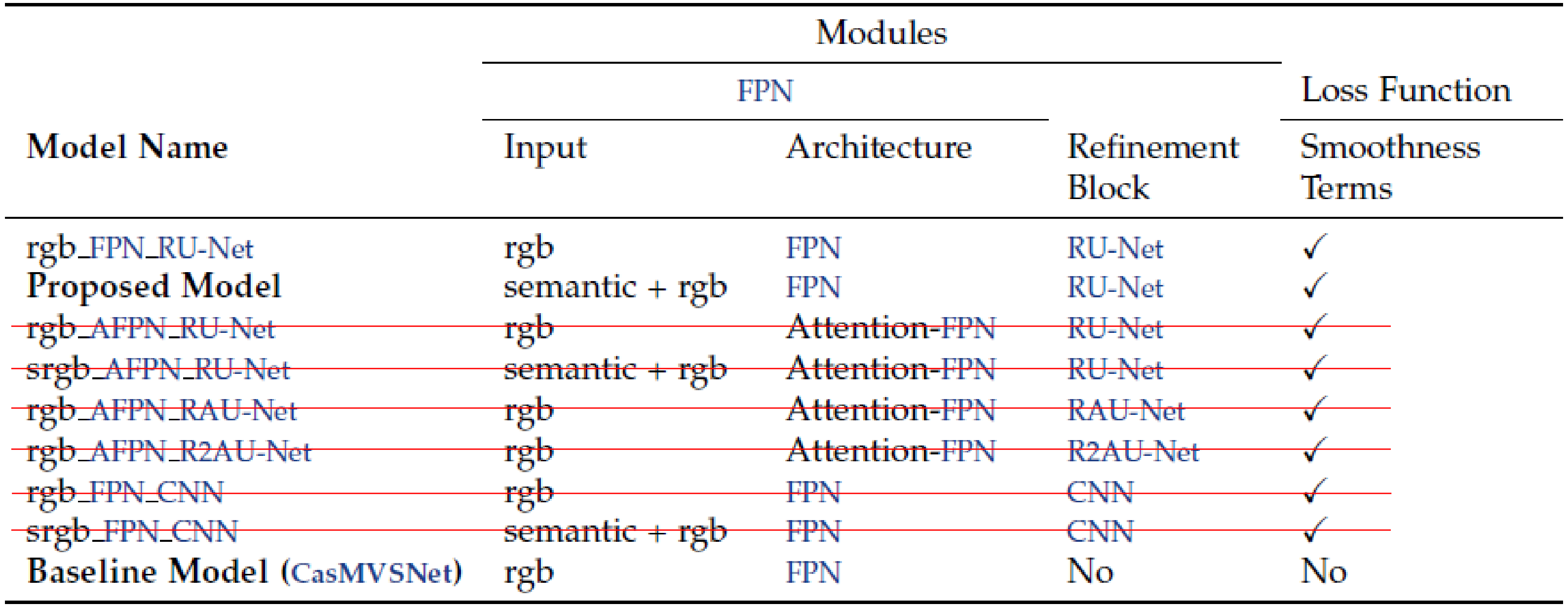


Model Name	Modules			Loss Function
	FPN			
	Input	Architecture	Refinement Block	Smoothness Terms
rgb_FPN_RU-Net	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
rgb_AFPN_RU-Net	rgb	Attention-FPN	RU-Net	✓
srgb_AFPN_RU-Net	semantic + rgb	Attention-FPN	RU-Net	✓
rgb_AFPN_RAU-Net	rgb	Attention-FPN	RAU-Net	✓
rgb_AFPN_R2AU-Net	rgb	Attention-FPN	R2AU-Net	✓
rgb_FPN_CNN	rgb	FPN	CNN	✓
srgb_FPN_CNN	semantic + rgb	FPN	CNN	✓
Baseline Model (CasMVSNet)	rgb	FPN	No	No

Model selection criteria:

- runtime efficiency
- complexity considerations
- performance at the depth map level *(% of pixels with a depth error less than 4mm)*

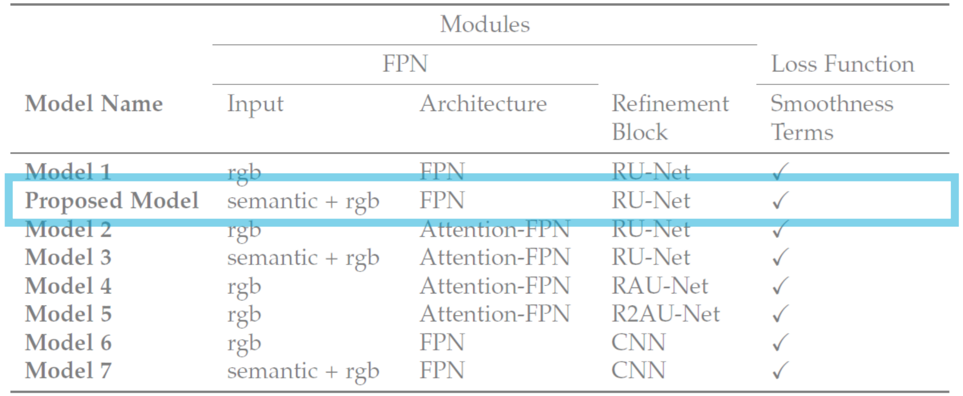
Experiments and Evaluation: Model Selection



	Modules			
	FPN			Loss Function
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
rgb_FPN_RU-Net	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
~~rgb_AFPN_RU-Net~~	~~rgb~~	~~Attention-FPN~~	~~RU-Net~~	~~✓~~
~~srgb_AFPN_RU-Net~~	~~semantic + rgb~~	~~Attention-FPN~~	~~RU-Net~~	~~✓~~
~~rgb_AFPN_RAU-Net~~	~~rgb~~	~~Attention-FPN~~	~~RAU-Net~~	~~✓~~
~~rgb_AFPN_R2AU-Net~~	~~rgb~~	~~Attention-FPN~~	~~R2AU-Net~~	~~✓~~
~~rgb_FPN_CNN~~	~~rgb~~	~~FPN~~	~~CNN~~	~~✓~~
~~srgb_FPN_CNN~~	~~semantic + rgb~~	~~FPN~~	~~CNN~~	~~✓~~
Baseline Model (CasMVSNet)	rgb	FPN	No	No

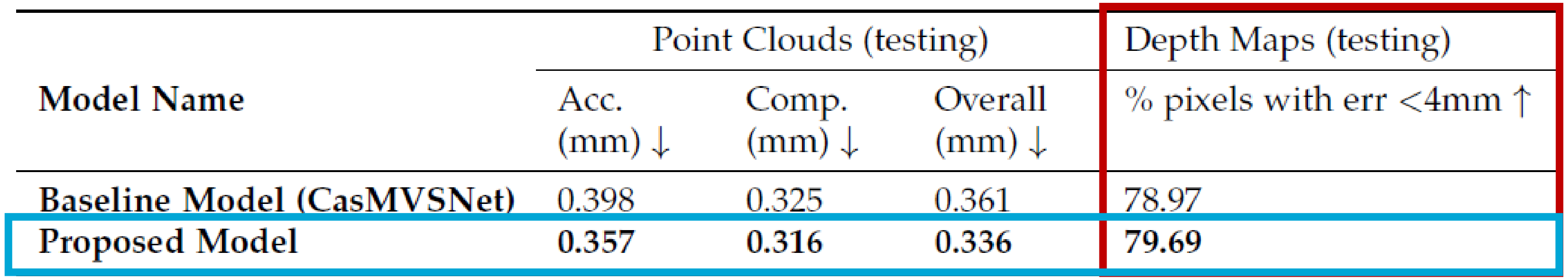
Proposed Model □ rgb_FPN_RU-Net

Experiments and Evaluation: Proposed Model



	Modules			
	FPN			Loss Function
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
Model 1	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
Model 2	rgb	Attention-FPN	RU-Net	✓
Model 3	semantic + rgb	Attention-FPN	RU-Net	✓
Model 4	rgb	Attention-FPN	RAU-Net	✓
Model 5	rgb	Attention-FPN	R2AU-Net	✓
Model 6	rgb	FPN	CNN	✓
Model 7	semantic + rgb	FPN	CNN	✓

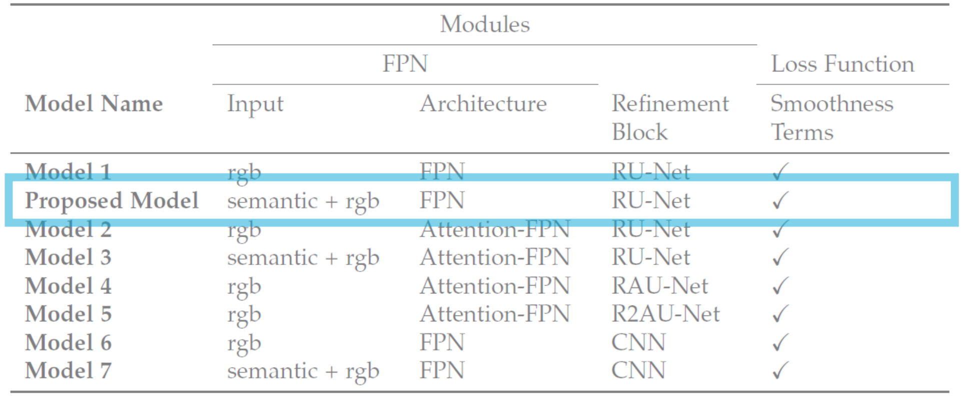
Evaluation on Point Cloud and Depth Map levels



	Point Clouds (testing)			Depth Maps (testing)
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓	% pixels with err <4mm ↑
Baseline Model (CasMVSNet)	0.398	0.325	0.361	78.97
Proposed Model	**0.357**	**0.316**	**0.336**	**79.69**

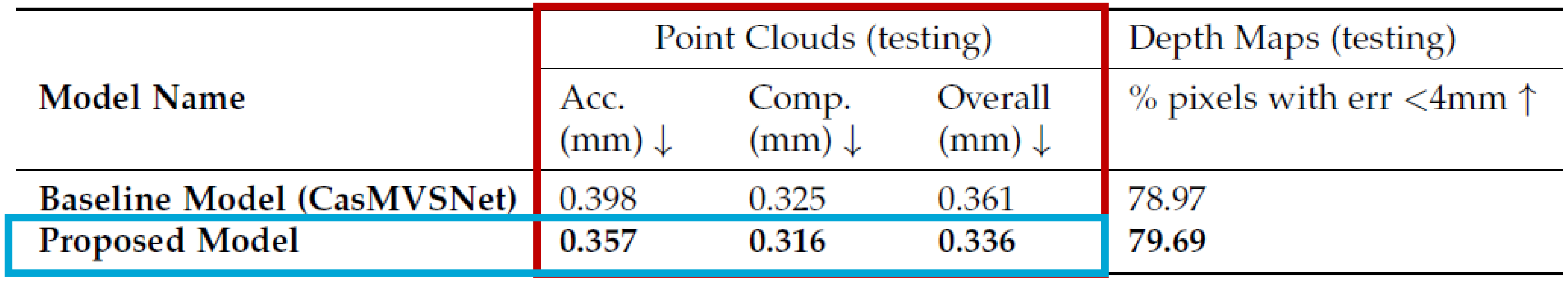
- The proposed model showed a 1% increase in accuracy at the depth map level.

Experiments and Evaluation: Proposed Model



	Modules			
	FPN			Loss Function
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
Model 1	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
Model 2	rgb	Attention-FPN	RU-Net	✓
Model 3	semantic + rgb	Attention-FPN	RU-Net	✓
Model 4	rgb	Attention-FPN	RAU-Net	✓
Model 5	rgb	Attention-FPN	R2AU-Net	✓
Model 6	rgb	FPN	CNN	✓
Model 7	semantic + rgb	FPN	CNN	✓

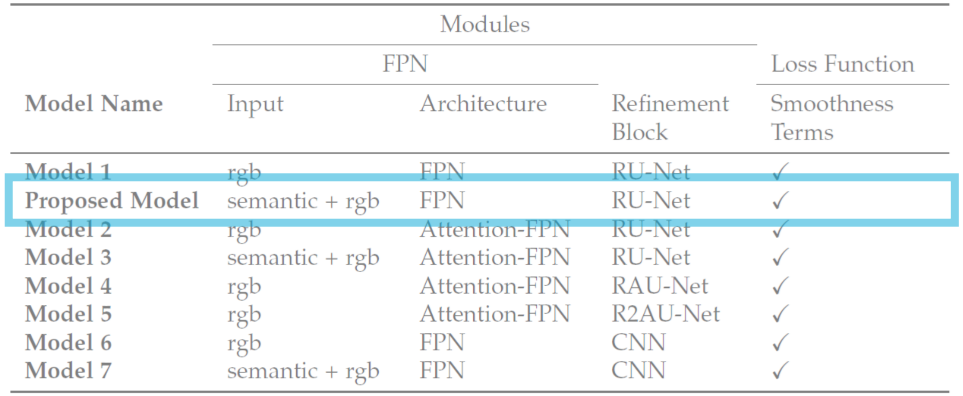
Evaluation on Point Cloud and Depth Map levels



	Point Clouds (testing)			Depth Maps (testing)
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓	% pixels with err <4mm ↑
Baseline Model (CasMVSNet)	0.398	0.325	0.361	78.97
Proposed Model	**0.357**	**0.316**	**0.336**	**79.69**

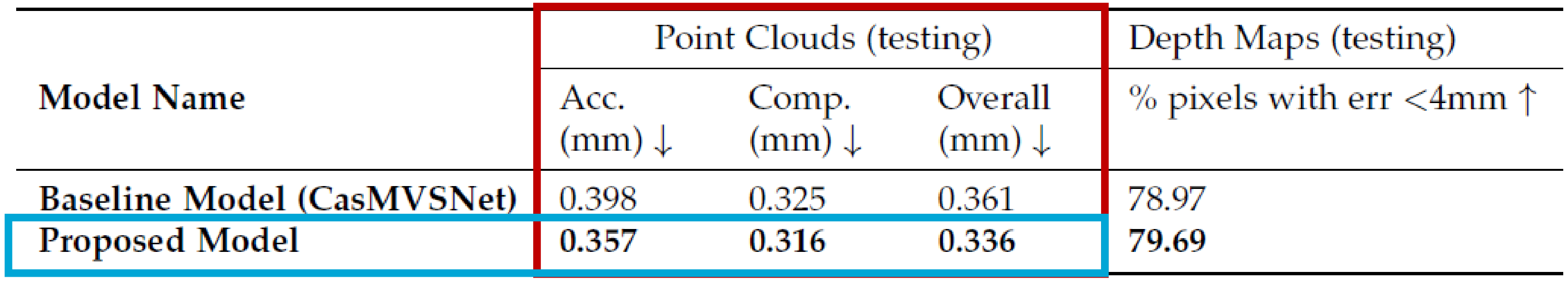
- Significant improvements in accuracy at the point cloud level

Experiments and Evaluation: Proposed Model



	Modules			
	FPN			Loss Function
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
Model 1	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
Model 2	rgb	Attention-FPN	RU-Net	✓
Model 3	semantic + rgb	Attention-FPN	RU-Net	✓
Model 4	rgb	Attention-FPN	RAU-Net	✓
Model 5	rgb	Attention-FPN	R2AU-Net	✓
Model 6	rgb	FPN	CNN	✓
Model 7	semantic + rgb	FPN	CNN	✓

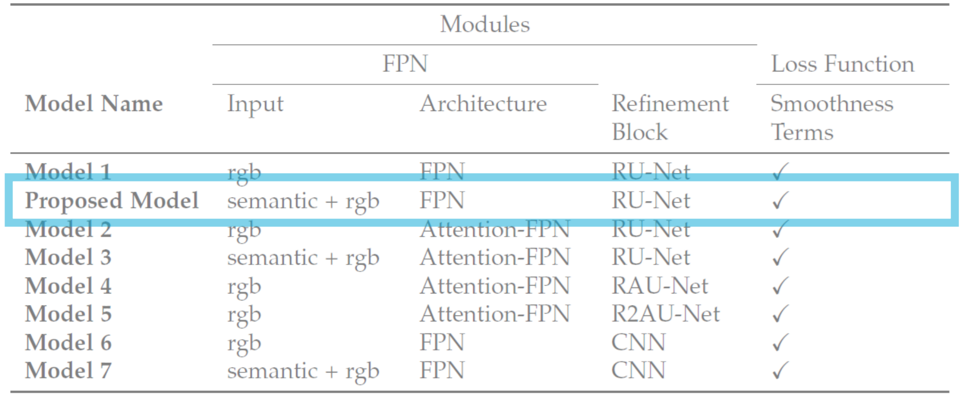
Evaluation on Point Cloud and Depth Map levels



	Point Clouds (testing)			Depth Maps (testing)
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓	% pixels with err <4mm ↑
Baseline Model (CasMVSNet)	0.398	0.325	0.361	78.97
Proposed Model	**0.357**	**0.316**	**0.336**	**79.69**

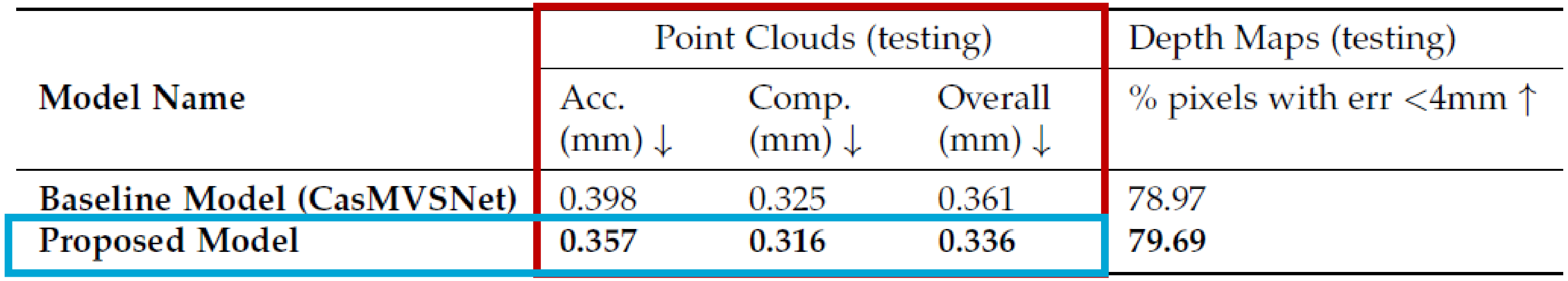
- Significant improvements in accuracy at the point cloud level **indicating a more precise reconstruction of the point cloud.**

Experiments and Evaluation: Proposed Model



	Modules			
	FPN			Loss Function
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
Model 1	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
Model 2	rgb	Attention-FPN	RU-Net	✓
Model 3	semantic + rgb	Attention-FPN	RU-Net	✓
Model 4	rgb	Attention-FPN	RAU-Net	✓
Model 5	rgb	Attention-FPN	R2AU-Net	✓
Model 6	rgb	FPN	CNN	✓
Model 7	semantic + rgb	FPN	CNN	✓

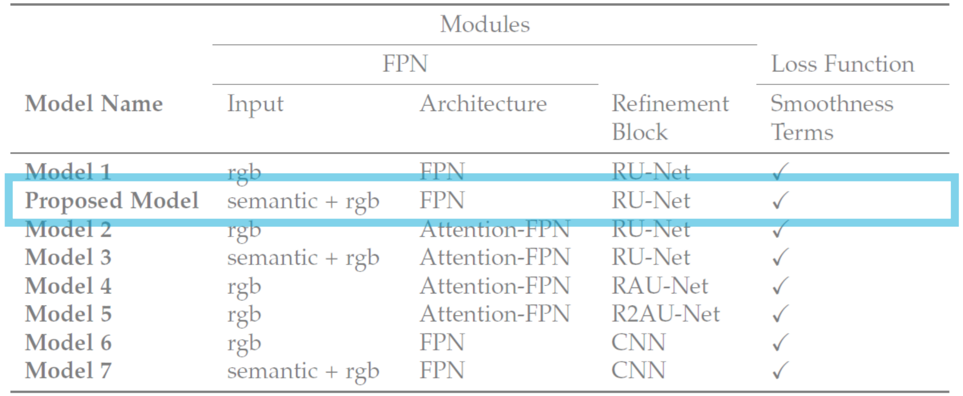
Evaluation on Point Cloud and Depth Map levels



	Point Clouds (testing)			Depth Maps (testing)
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓	% pixels with err <4mm ↑
Baseline Model (CasMVSNet)	0.398	0.325	0.361	78.97
Proposed Model	**0.357**	**0.316**	**0.336**	**79.69**

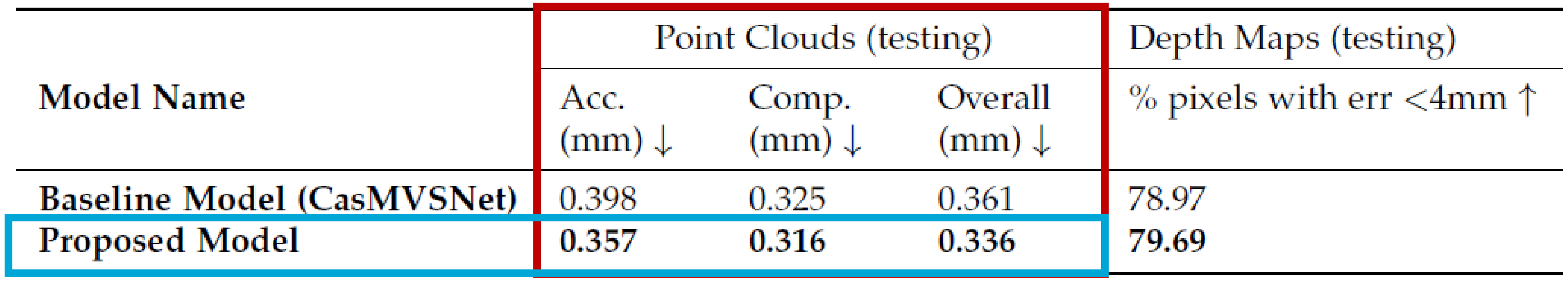
- Significant improvements in accuracy at the point cloud level
  indicating a more precise reconstruction of the point cloud.
- Improvement attributed to the **depth fusion algorithm:**

Experiments and Evaluation: Proposed Model



	Modules			
	FPN			Loss Function
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
Model 1	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
Model 2	rgb	Attention-FPN	RU-Net	✓
Model 3	semantic + rgb	Attention-FPN	RU-Net	✓
Model 4	rgb	Attention-FPN	RAU-Net	✓
Model 5	rgb	Attention-FPN	R2AU-Net	✓
Model 6	rgb	FPN	CNN	✓
Model 7	semantic + rgb	FPN	CNN	✓

Evaluation on Point Cloud and Depth Map levels



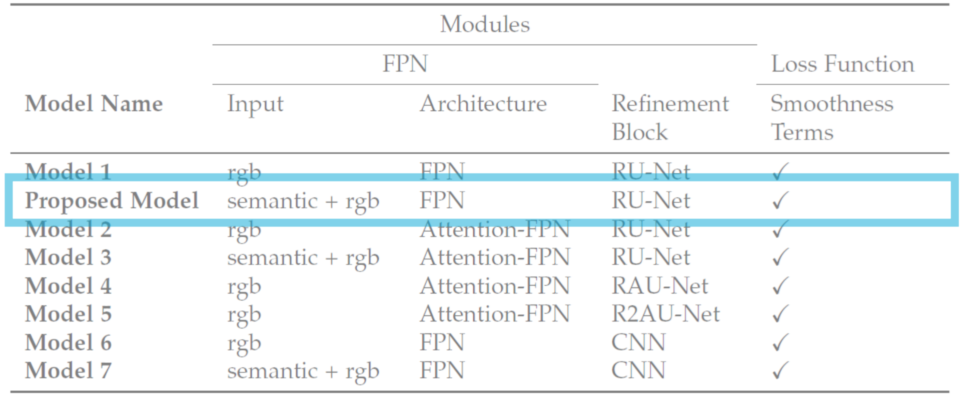
	Point Clouds (testing)			Depth Maps (testing)
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓	% pixels with err <4mm ↑
Baseline Model (CasMVSNet)	0.398	0.325	0.361	78.97
Proposed Model	**0.357**	**0.316**	**0.336**	**79.69**

- Significant improvements in accuracy at the point cloud level

  indicating a more precise reconstruction of the point cloud.

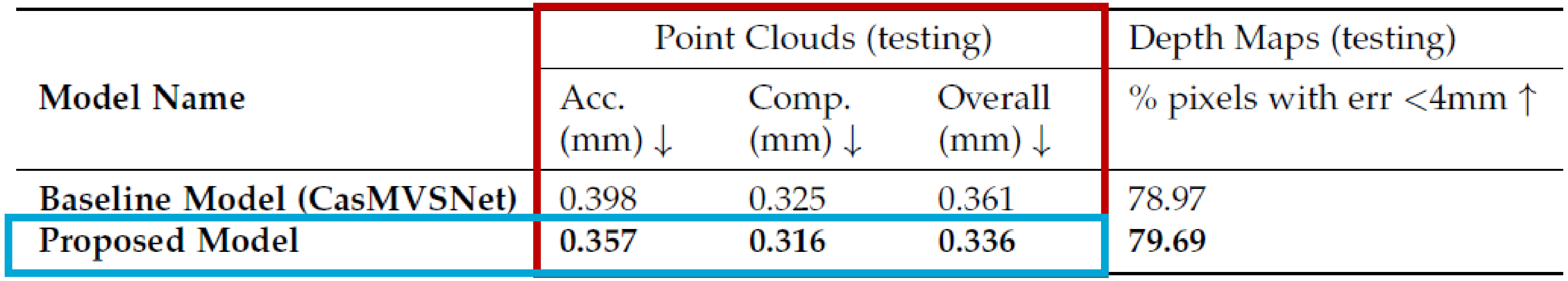
- Improvement attributed to the **depth fusion algorithm:**
  - **geometric + confidence tests**

Experiments and Evaluation: Proposed Model



	Modules			
	FPN			Loss Function
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
Model 1	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
Model 2	rgb	Attention-FPN	RU-Net	✓
Model 3	semantic + rgb	Attention-FPN	RU-Net	✓
Model 4	rgb	Attention-FPN	RAU-Net	✓
Model 5	rgb	Attention-FPN	R2AU-Net	✓
Model 6	rgb	FPN	CNN	✓
Model 7	semantic + rgb	FPN	CNN	✓

Evaluation on Point Cloud and Depth Map levels

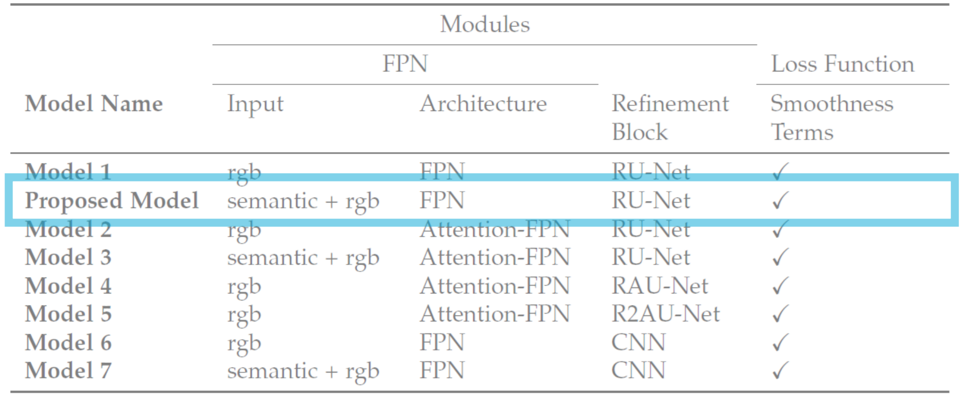


	Point Clouds (testing)			Depth Maps (testing)
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓	% pixels with err <4mm ↑
Baseline Model (CasMVSNet)	0.398	0.325	0.361	78.97
Proposed Model	**0.357**	**0.316**	**0.336**	**79.69**

- Significant improvements in accuracy at the point cloud level **indicating a more precise reconstruction of the point cloud.**

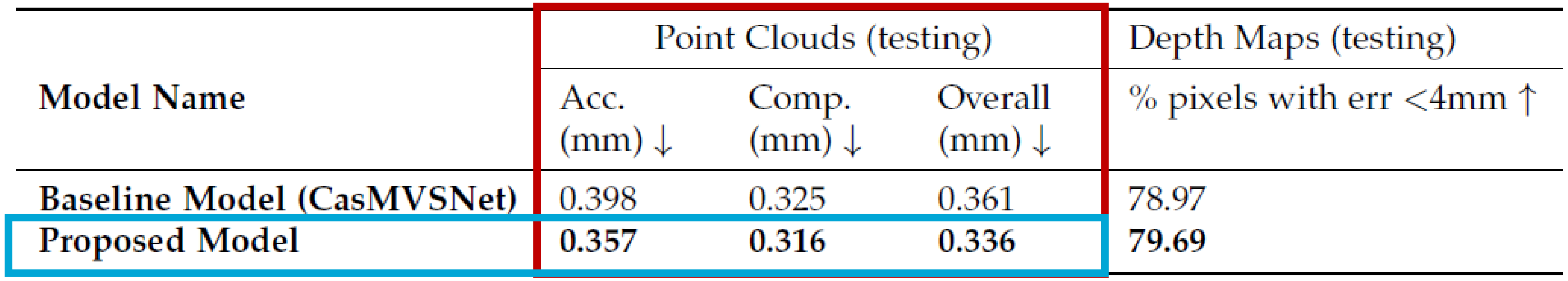
- Improvement attributed to the **depth fusion algorithm:**
  - **geometric + confidence tests**
    → reconstruction based on **multi-view consistent** and **confident** predictions

Experiments and Evaluation: Proposed Model



	Modules			
	FPN			Loss Function
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
Model 1	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
Model 2	rgb	Attention-FPN	RU-Net	✓
Model 3	semantic + rgb	Attention-FPN	RU-Net	✓
Model 4	rgb	Attention-FPN	RAU-Net	✓
Model 5	rgb	Attention-FPN	R2AU-Net	✓
Model 6	rgb	FPN	CNN	✓
Model 7	semantic + rgb	FPN	CNN	✓

Evaluation on Point Cloud and Depth Map levels



	Point Clouds (testing)			Depth Maps (testing)
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓	% pixels with err <4mm ↑
Baseline Model (CasMVSNet)	0.398	0.325	0.361	78.97
Proposed Model	**0.357**	**0.316**	**0.336**	**79.69**

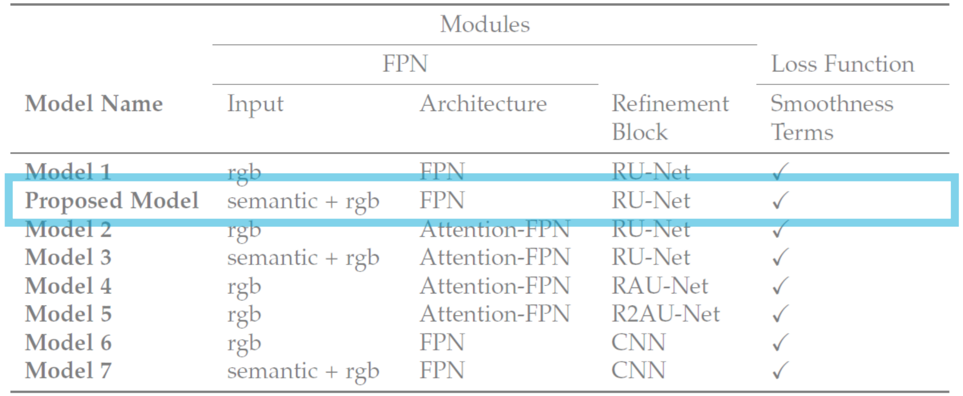
- Significant improvements in accuracy at the point cloud level

  indicating a more precise reconstruction of the point cloud.

- Improvement attributed to the **depth fusion algorithm**:
  - **geometric + confidence tests**

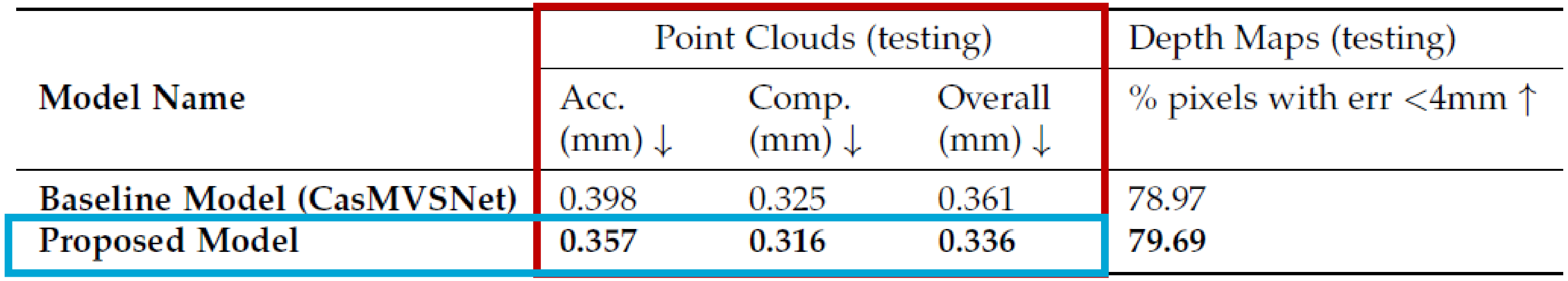
- Therefore, the higher accuracy suggests that the **Proposed Model** predicts **depth values that**:

Experiments and Evaluation: Proposed Model



	Modules			
	FPN			Loss Function
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
Model 1	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
Model 2	rgb	Attention-FPN	RU-Net	✓
Model 3	semantic + rgb	Attention-FPN	RU-Net	✓
Model 4	rgb	Attention-FPN	RAU-Net	✓
Model 5	rgb	Attention-FPN	R2AU-Net	✓
Model 6	rgb	FPN	CNN	✓
Model 7	semantic + rgb	FPN	CNN	✓

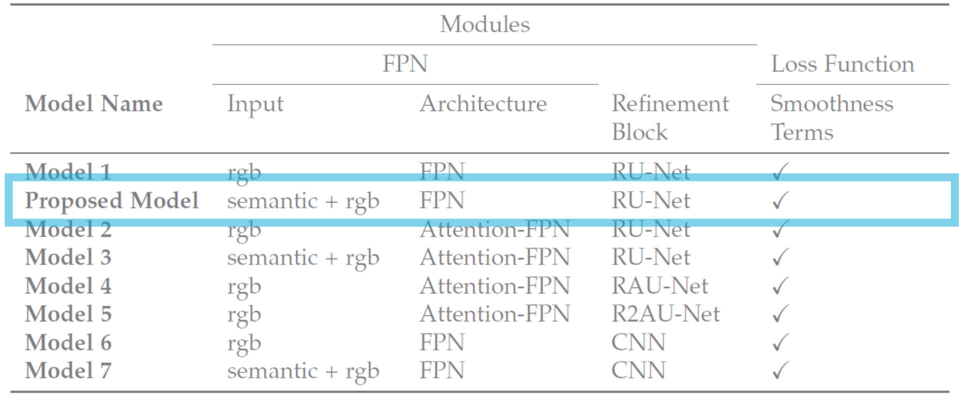
Evaluation on Point Cloud and Depth Map levels



Model Name	Point Clouds (testing)			Depth Maps (testing)
	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓	% pixels with err <4mm ↑
Baseline Model (CasMVSNet)	0.398	0.325	0.361	78.97
Proposed Model	**0.357**	**0.316**	**0.336**	**79.69**

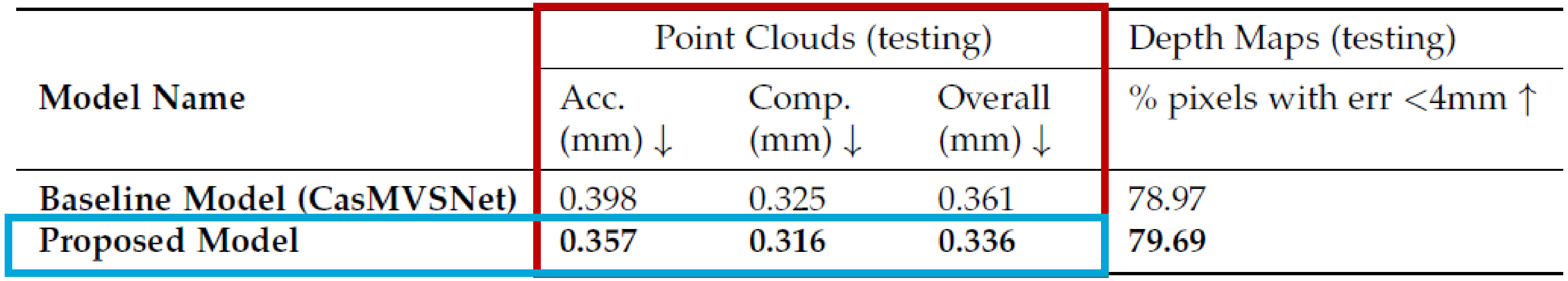
- Significant improvements in accuracy at the point cloud level **indicating a more precise reconstruction of the point cloud.**
- Improvement attributed to the **depth fusion algorithm**:
  - **geometric + confidence tests**
- Therefore, the higher accuracy suggests that the **Proposed Model** predicts **depth values that**:
  - **are more consistent** across multiple views

Experiments and Evaluation: Proposed Model



	Modules			
	FPN			Loss Function
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
Model 1	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
Model 2	rgb	Attention-FPN	RU-Net	✓
Model 3	semantic + rgb	Attention-FPN	RU-Net	✓
Model 4	rgb	Attention-FPN	RAU-Net	✓
Model 5	rgb	Attention-FPN	R2AU-Net	✓
Model 6	rgb	FPN	CNN	✓
Model 7	semantic + rgb	FPN	CNN	✓

Evaluation on Point Cloud and Depth Map levels



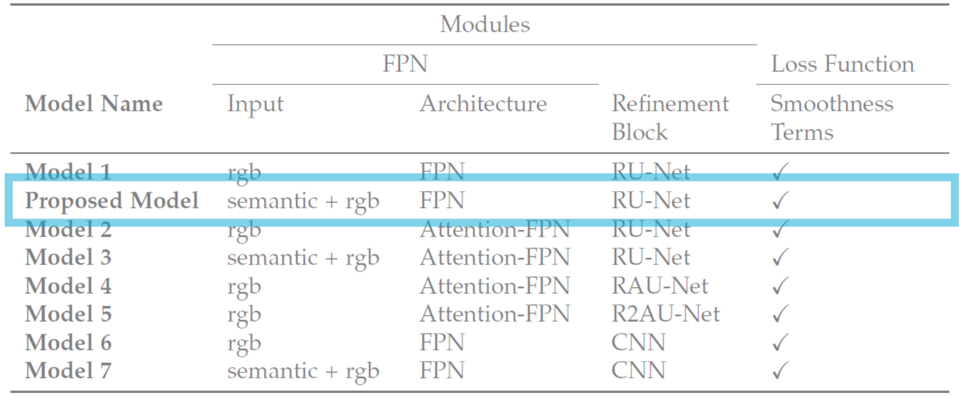
	Point Clouds (testing)			Depth Maps (testing)
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓	% pixels with err <4mm ↑
Baseline Model (CasMVSNet)	0.398	0.325	0.361	78.97
Proposed Model	**0.357**	**0.316**	**0.336**	**79.69**

- Significant improvements in accuracy at the point cloud level **indicating a more precise reconstruction of the point cloud.**

- Improvement attributed to the **depth fusion algorithm**:
  - **geometric + confidence tests**

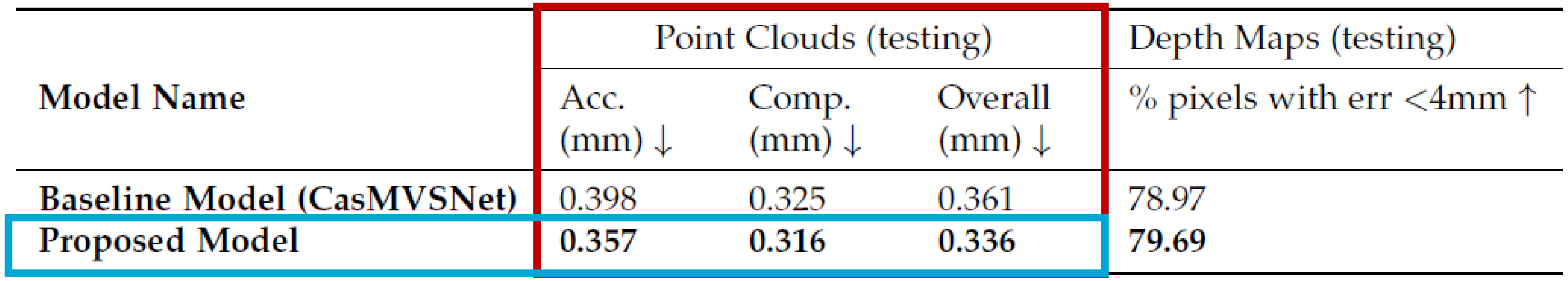
- Therefore, the higher accuracy suggests that the **Proposed Model** predicts **depth values that**:
  - **are more consistent** across multiple views
  - **more confidently**

Experiments and Evaluation: Proposed Model



	Modules			
	FPN			Loss Function
Model Name	Input	Architecture	Refinement Block	Smoothness Terms
Model 1	rgb	FPN	RU-Net	✓
Proposed Model	semantic + rgb	FPN	RU-Net	✓
Model 2	rgb	Attention-FPN	RU-Net	✓
Model 3	semantic + rgb	Attention-FPN	RU-Net	✓
Model 4	rgb	Attention-FPN	RAU-Net	✓
Model 5	rgb	Attention-FPN	R2AU-Net	✓
Model 6	rgb	FPN	CNN	✓
Model 7	semantic + rgb	FPN	CNN	✓

Evaluation on Point Cloud and Depth Map levels



	Point Clouds (testing)			Depth Maps (testing)
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓	% pixels with err <4mm ↑
Baseline Model (CasMVSNet)	0.398	0.325	0.361	78.97
Proposed Model	**0.357**	**0.316**	**0.336**	**79.69**

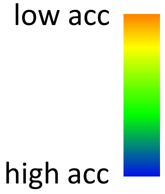
- Significant improvements in accuracy at the point cloud level **indicating a more precise reconstruction of the point cloud.**

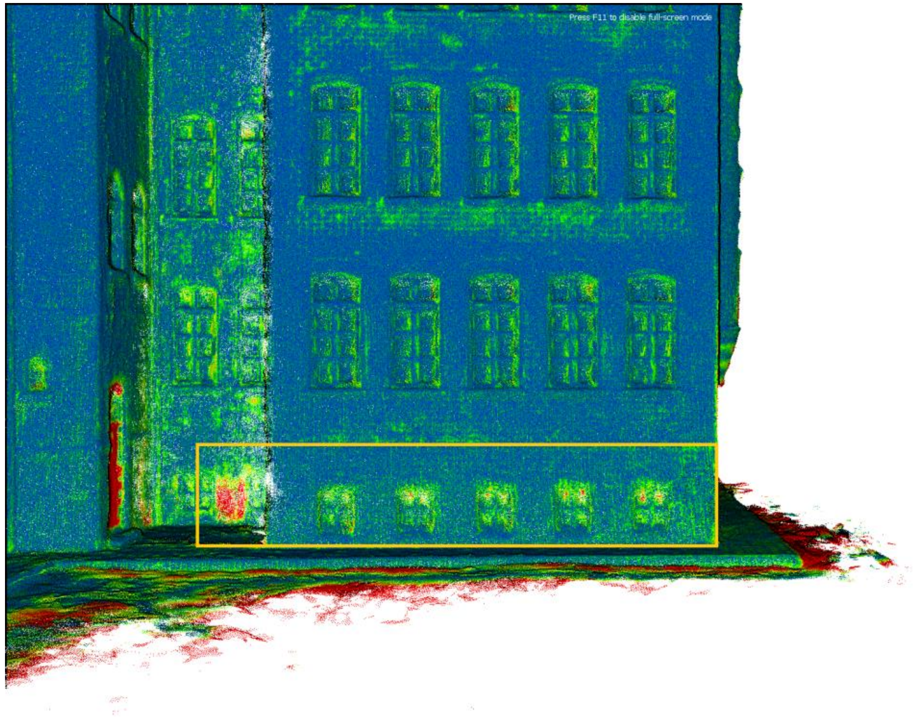
- Improvement attributed to the **depth fusion algorithm**:
  - **geometric + confidence tests**

- Therefore, the higher accuracy suggests that the **Proposed Model** predicts **depth values that**:
  - **are more consistent** across multiple views
  - **more confidently** (20.000 pixels more with a threshold of 0.999)

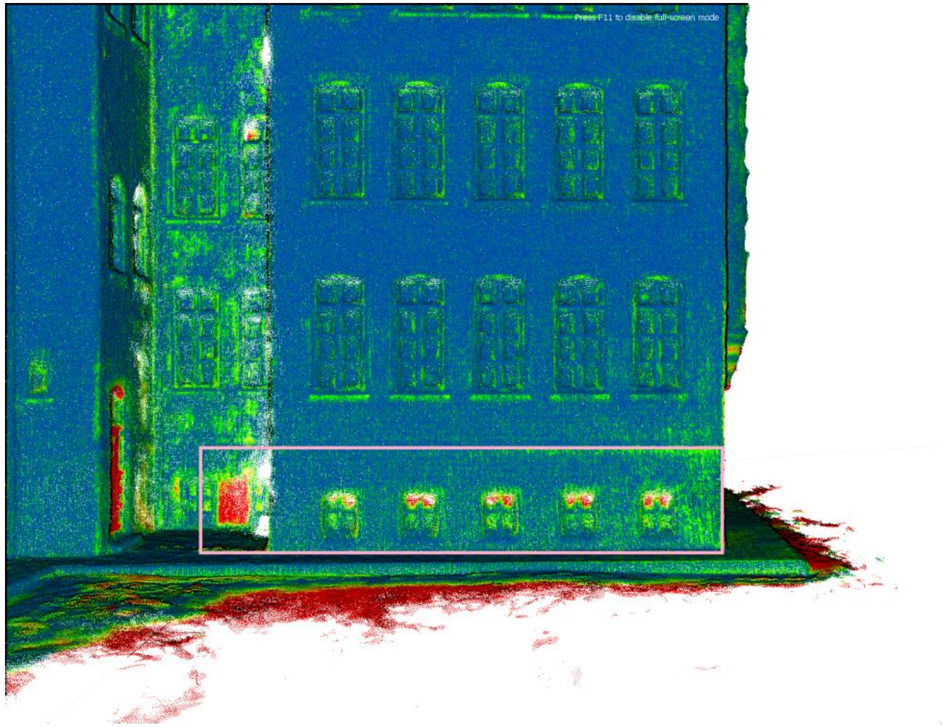
Experiments and Evaluation: Accuracy

- points color-coded based on their **proximity** to ground truth
- **bottom row** of windows in the **Proposed Model** **are closer** to the ground truth



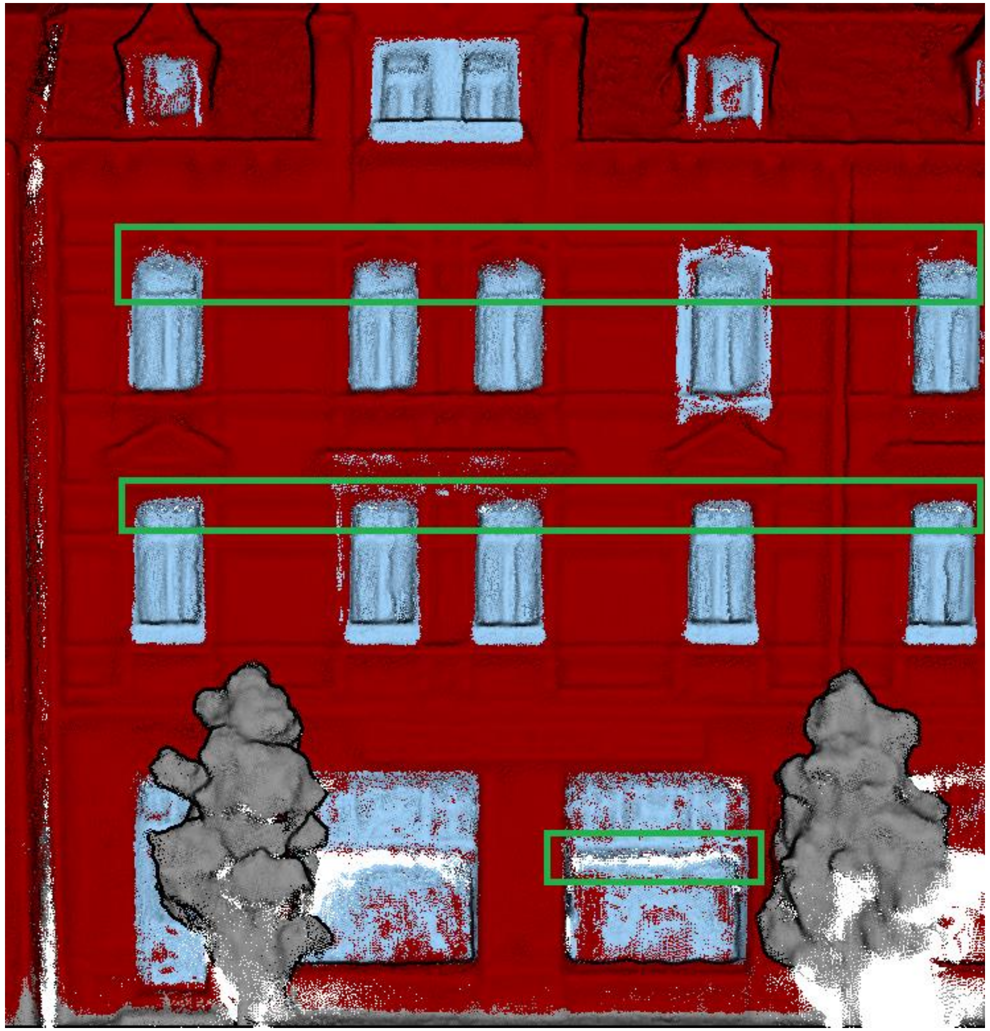


Proposed

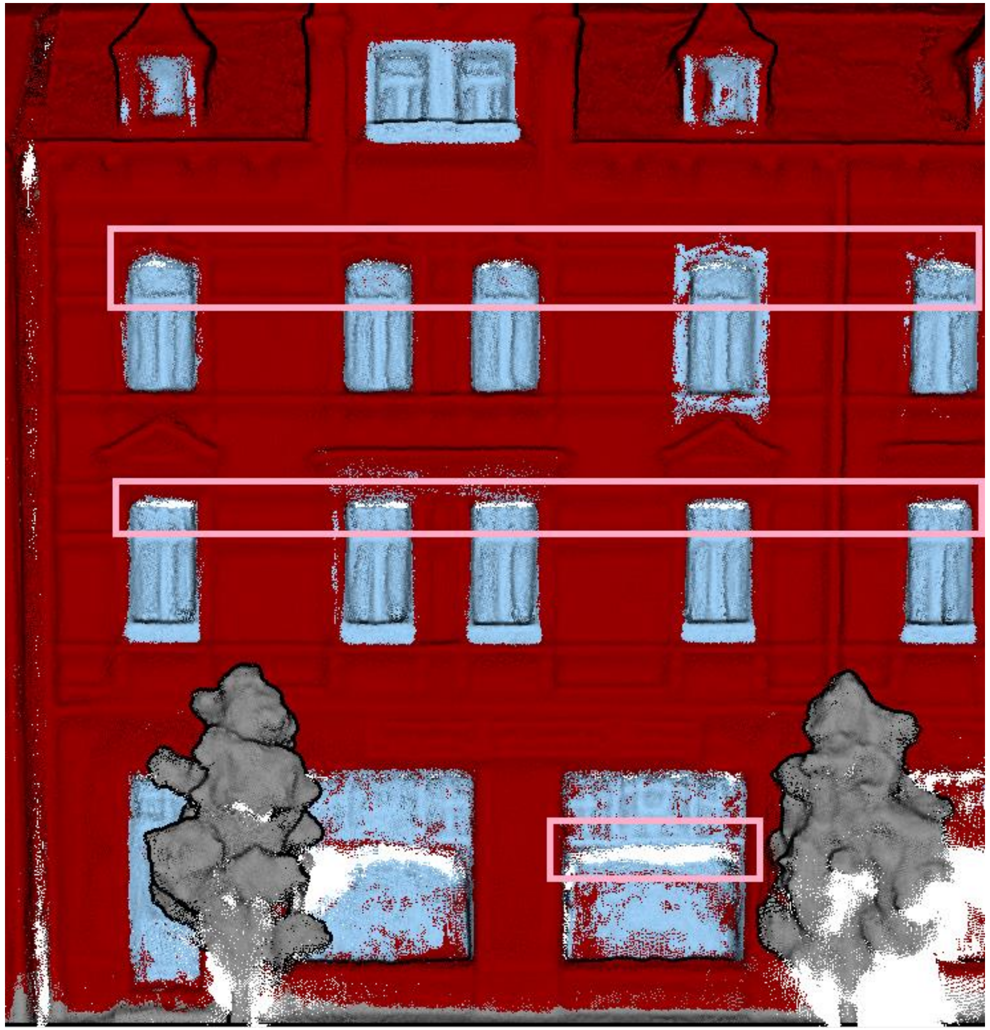


Baseline

Experiments and Evaluation: Completeness

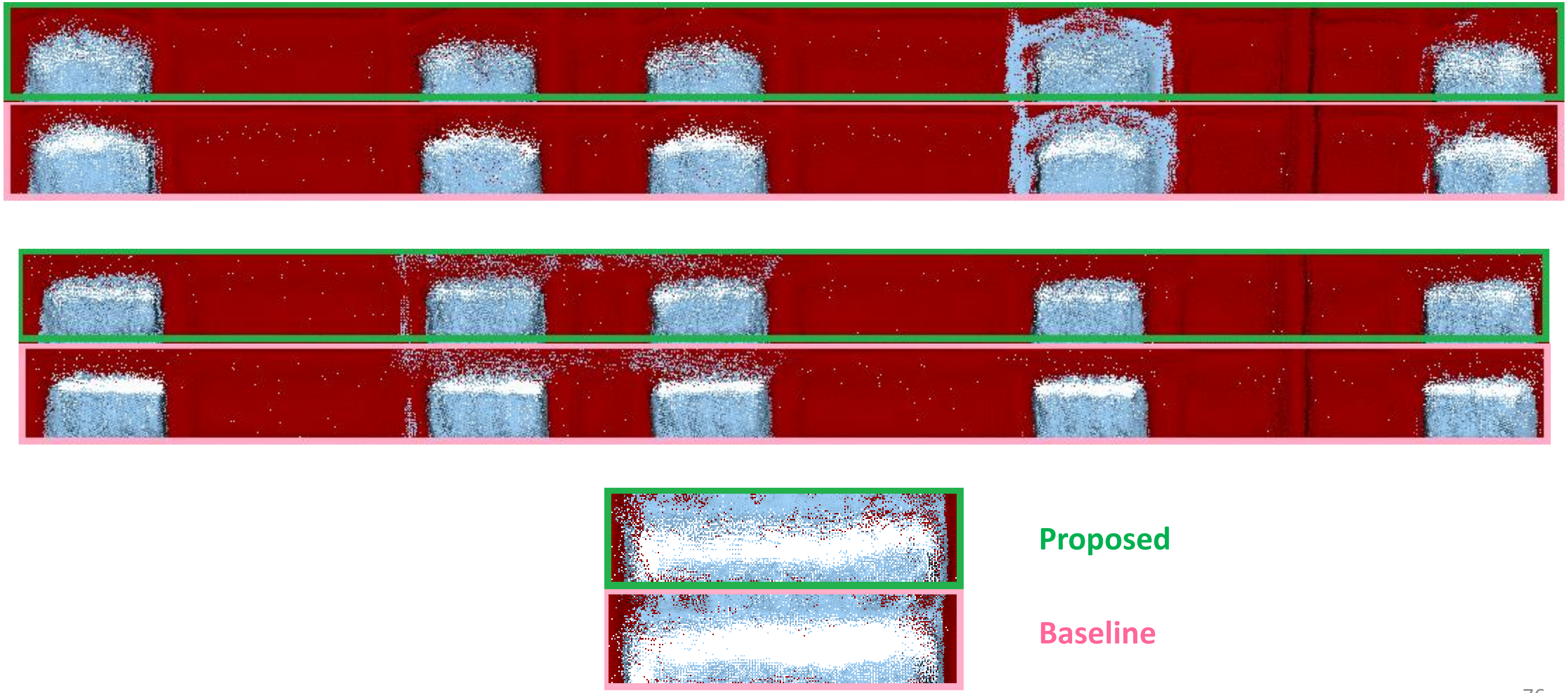


Proposed



Baseline

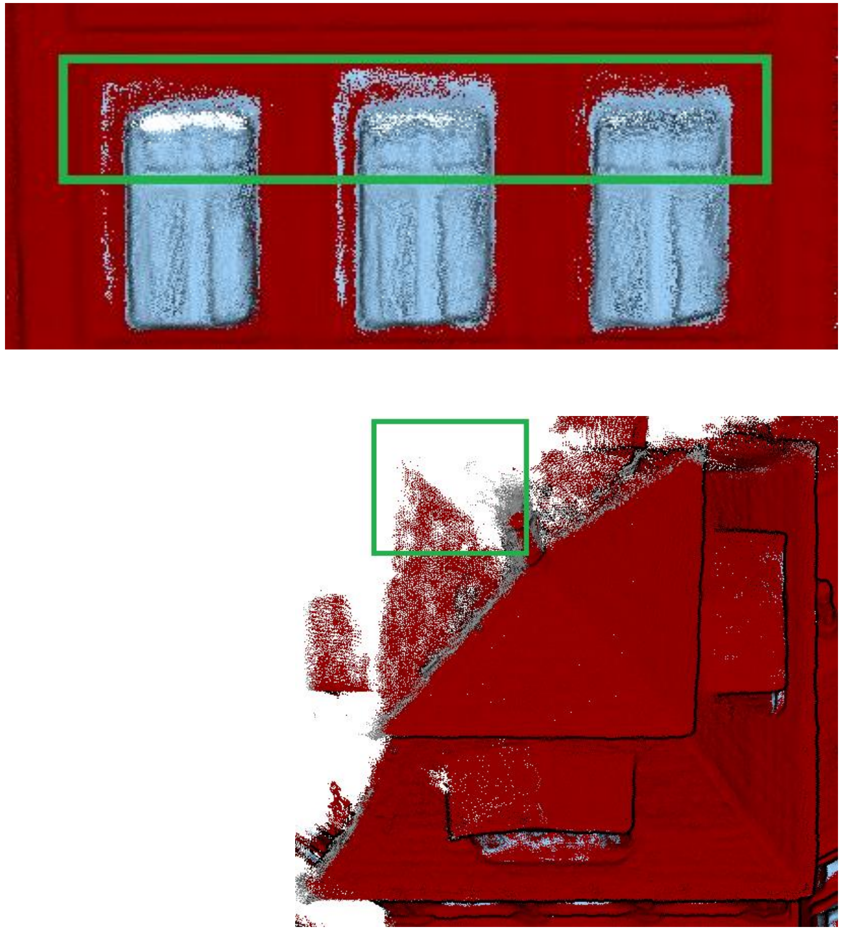
Experiments and Evaluation: Completeness



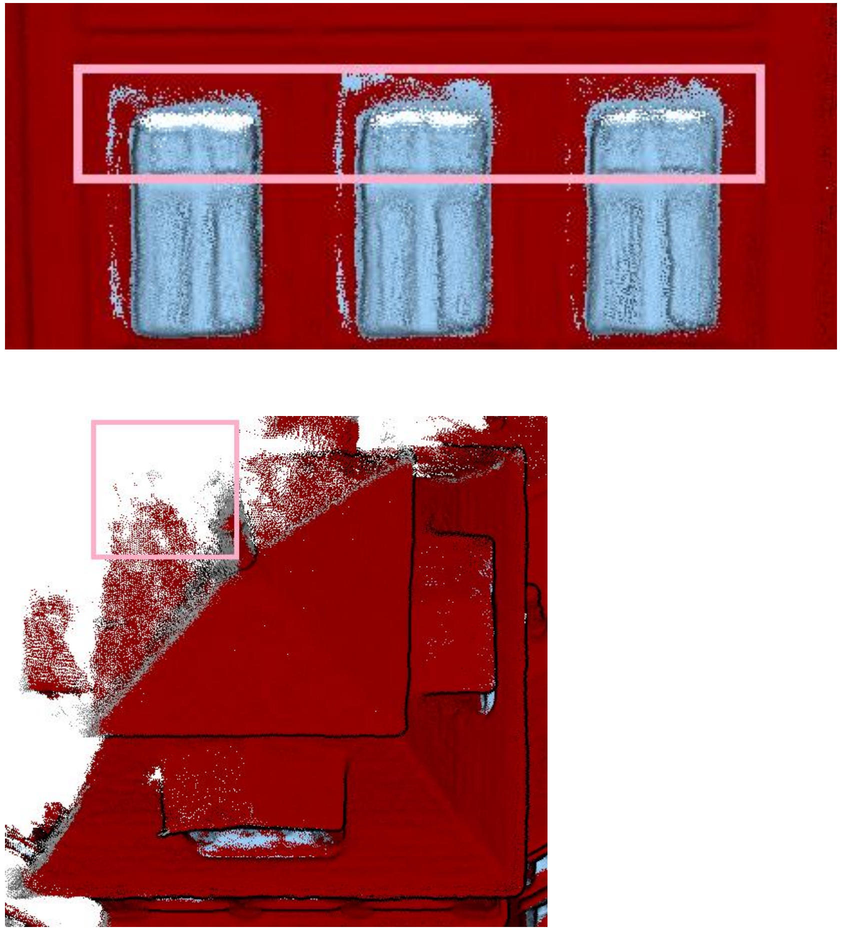
Proposed

Baseline

Experiments and Evaluation: Completeness



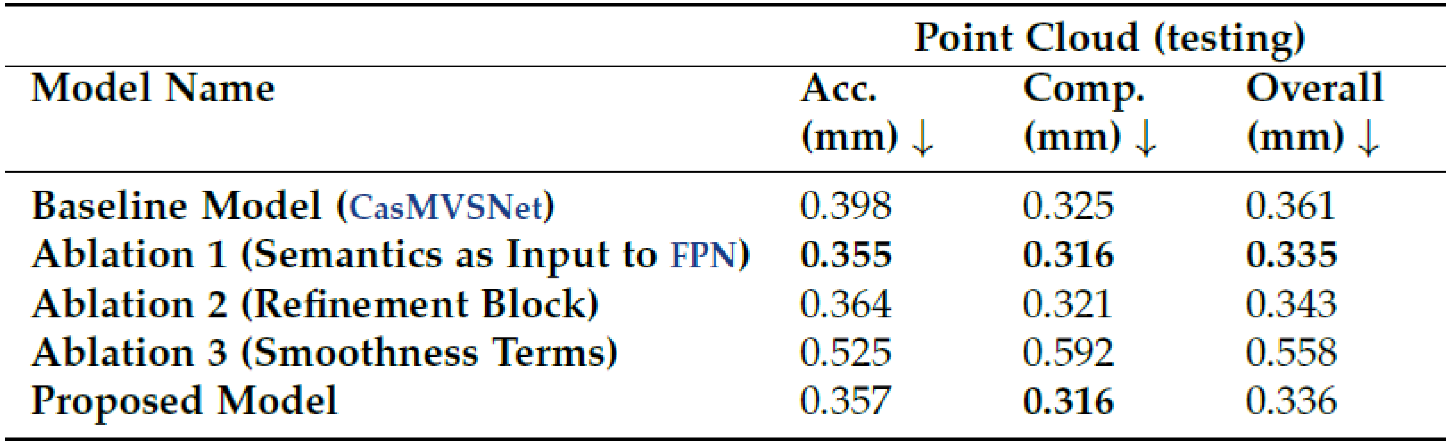
Proposed



Baseline

Ablation study

An ablation study **isolates components** of the approach and **assesses their individual contribution** to the overall performance.



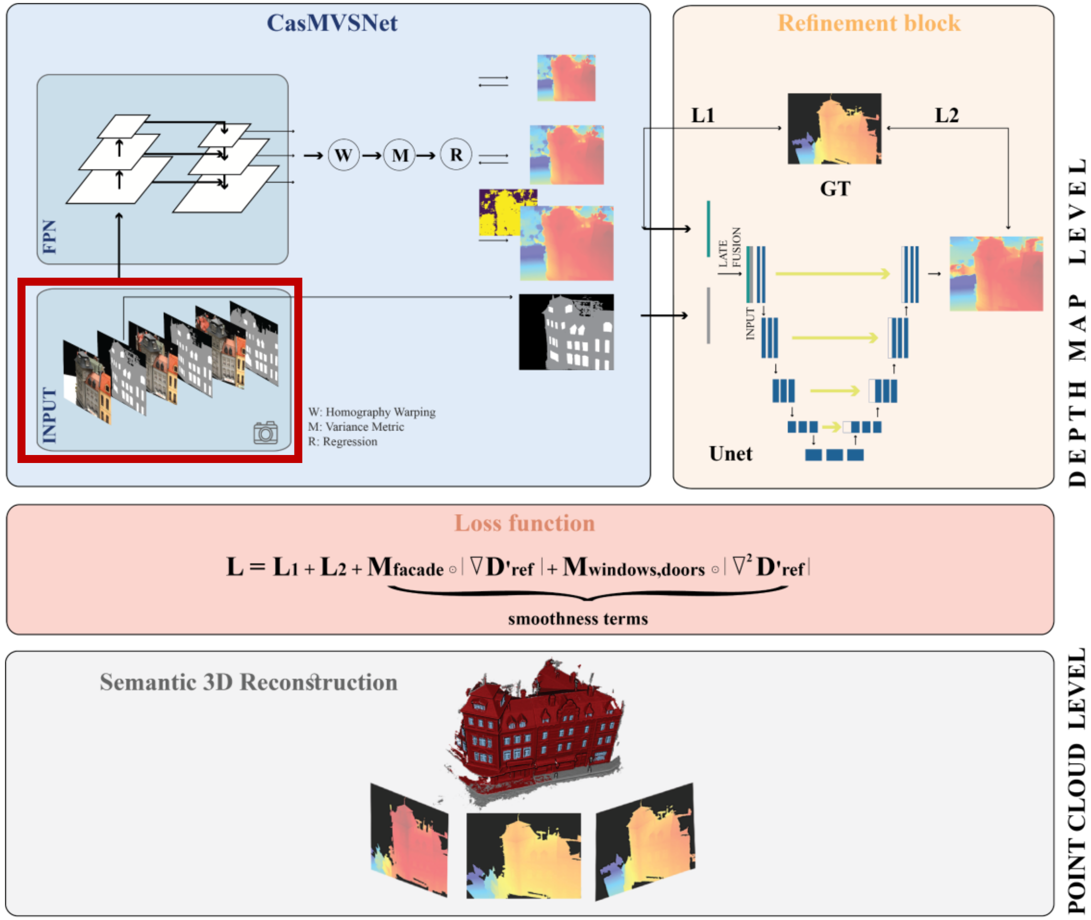
	Point Cloud (testing)		
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336

Ablation study 1

Model Name	Point Cloud (testing)		
	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336

Ablation 1:

network trained with the semantics as input to the FPN module

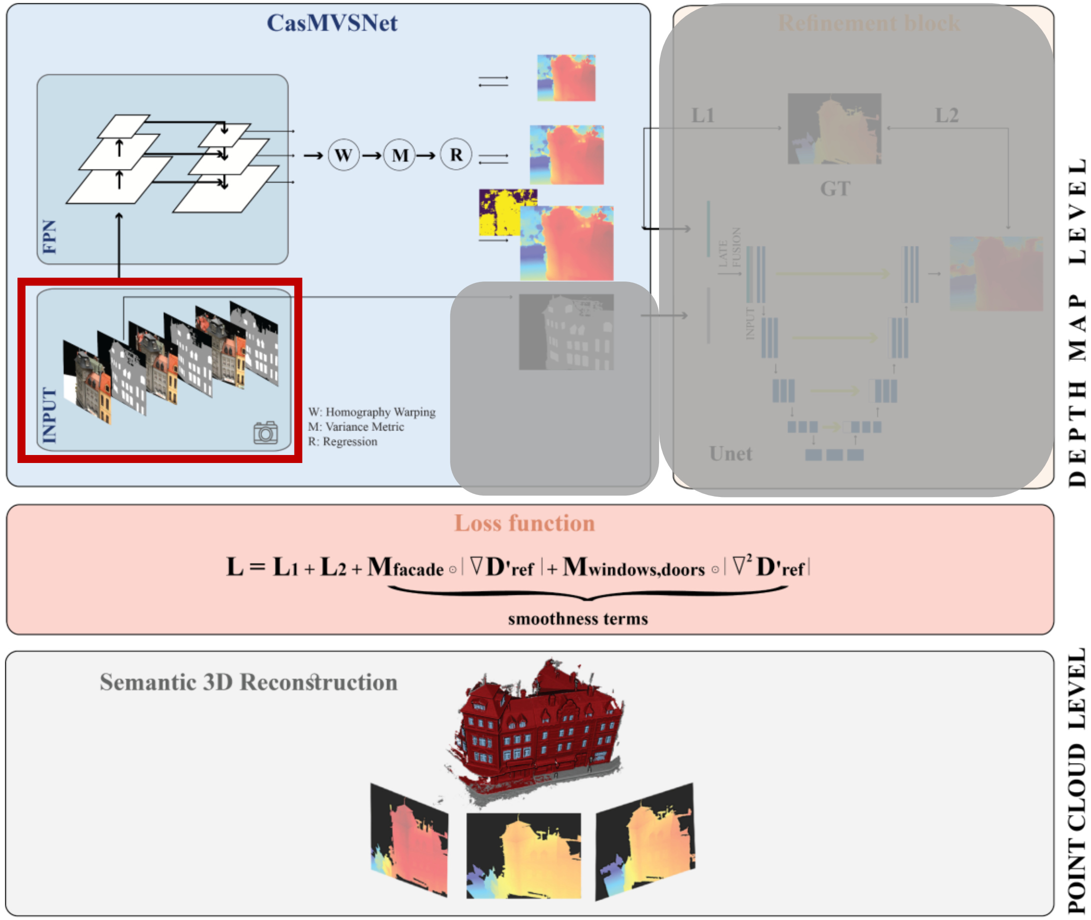

CasMVSNet
FPN
INPUT
W
M
R
W: Homography Warping
M: Variance Metric
R: Regression
Refinement block
L1
L2
GT
LATE FUSION
INPUT
Unet
DEPTH MAP LEVEL
Loss function
$L = L_1 + L_2 + \underbrace{M_{facade} \circ |\nabla D'_{ref}| + M_{windows,doors} \circ |\nabla^2 D'_{ref}|}_{\text{smoothness terms}}$
Semantic 3D Reconstruction
POINT CLOUD LEVEL

Ablation study 1

	Point Cloud (testing)		
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336

Ablation 1:

network trained with the semantics as input to the FPN module

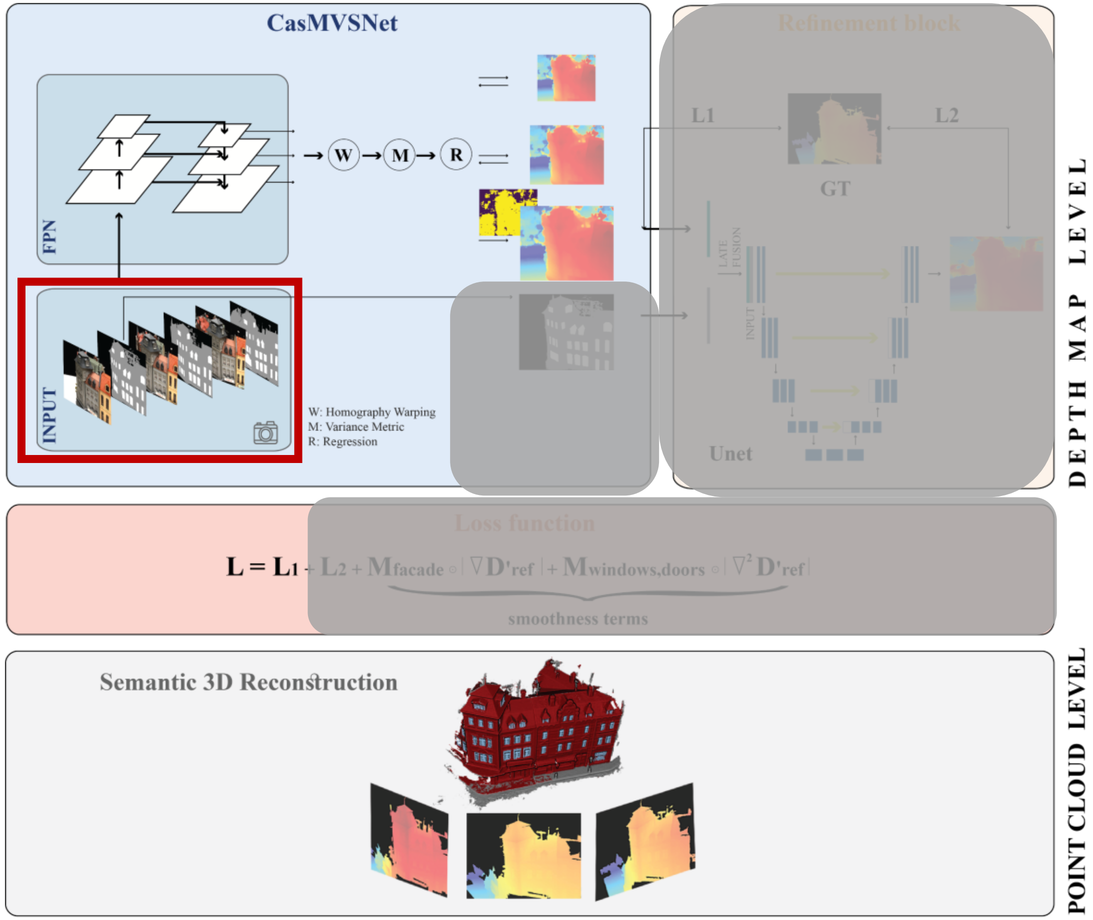

CasMVSNet
FPN
INPUT
W
M
R
W: Homography Warping
M: Variance Metric
R: Regression
Refinement block
L1
L2
GT
LATE FUSION
INPUT
Unet
DEPTH MAP LEVEL
Loss function
$L = L_1 + L_2 + \underbrace{M_{facade} \circ |\nabla D'_{ref}| + M_{windows,doors} \circ |\nabla^2 D'_{ref}|}_{\text{smoothness terms}}$
Semantic 3D Reconstruction
POINT CLOUD LEVEL

Ablation study 1

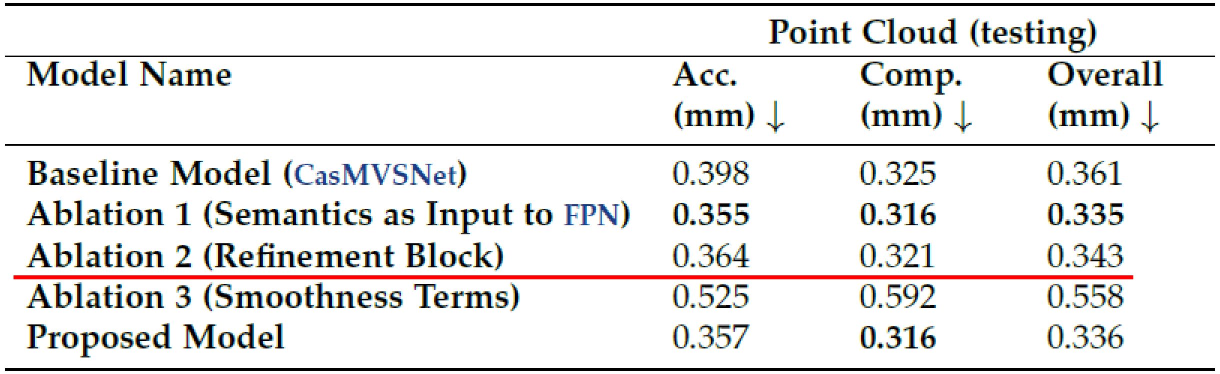
	Point Cloud (testing)		
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336

Ablation 1:

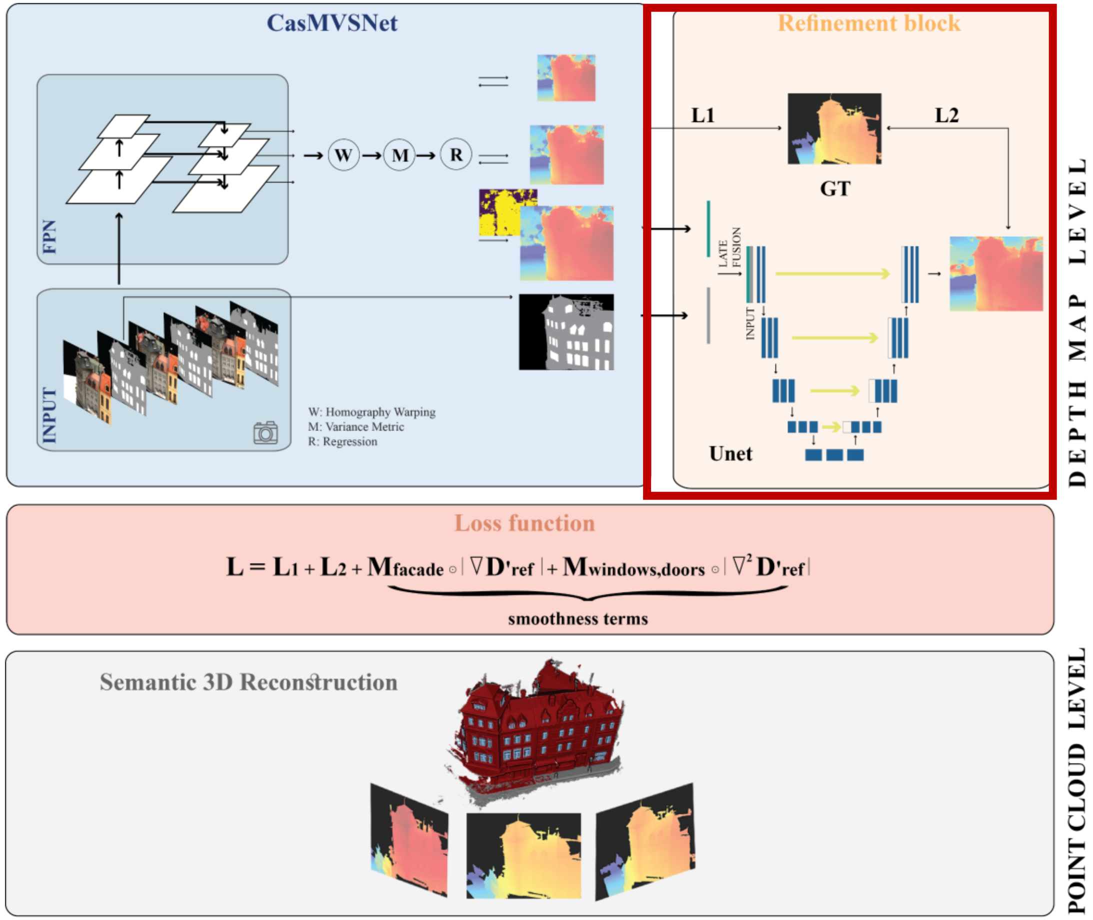
network trained with the semantics as input to the FPN module



Ablation study 2



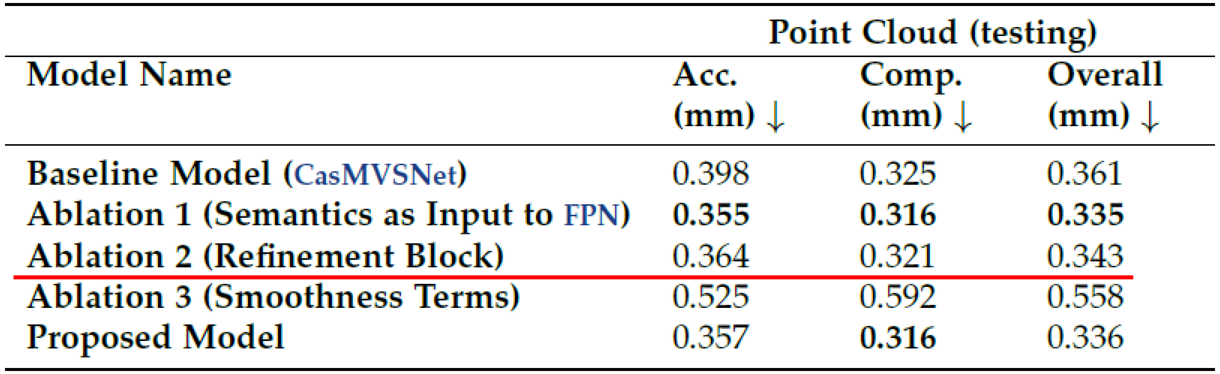
	Point Cloud (testing)		
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336


CasMVSNet
Refinement block
FPN
INPUT
W
M
R
W: Homography Warping
M: Variance Metric
R: Regression
L1
L2
GT
LATE FUSION
INPUT
Unet
DEPTH MAP LEVEL
Loss function
$L = L_1 + L_2 + \underbrace{M_{facade} \circ |\nabla D'_{ref}| + M_{windows,doors} \circ |\nabla^2 D'_{ref}|}_{\text{smoothness terms}}$
Semantic 3D Reconstruction
POINT CLOUD LEVEL

Ablation 2:

network trained solely with the refinement block

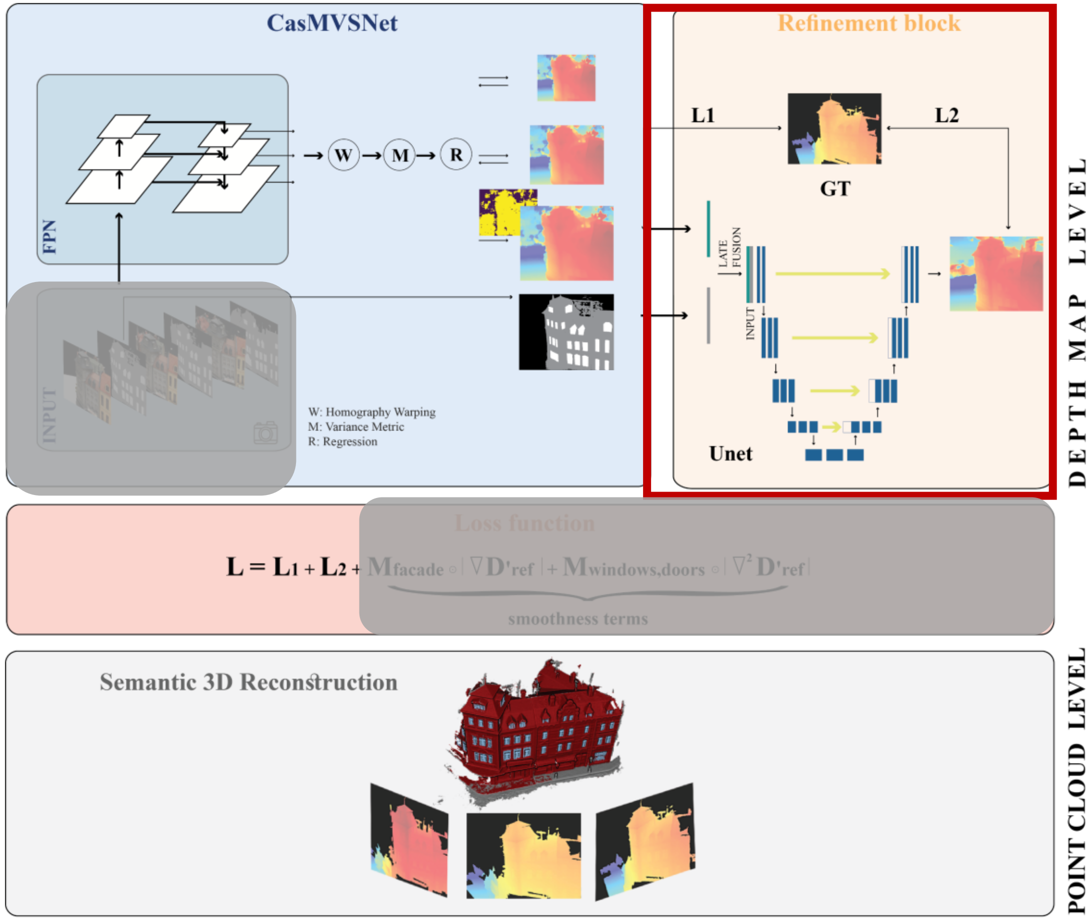
Ablation study 2



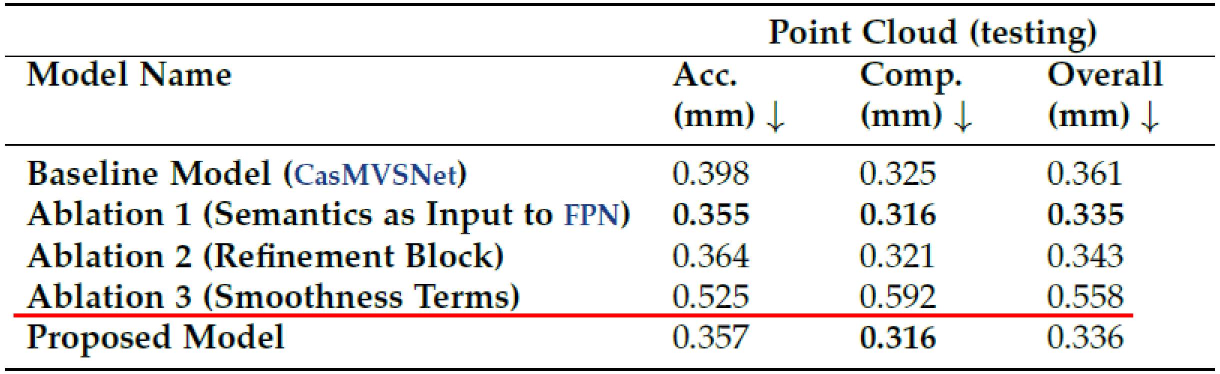
	Point Cloud (testing)		
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336

Ablation 2:

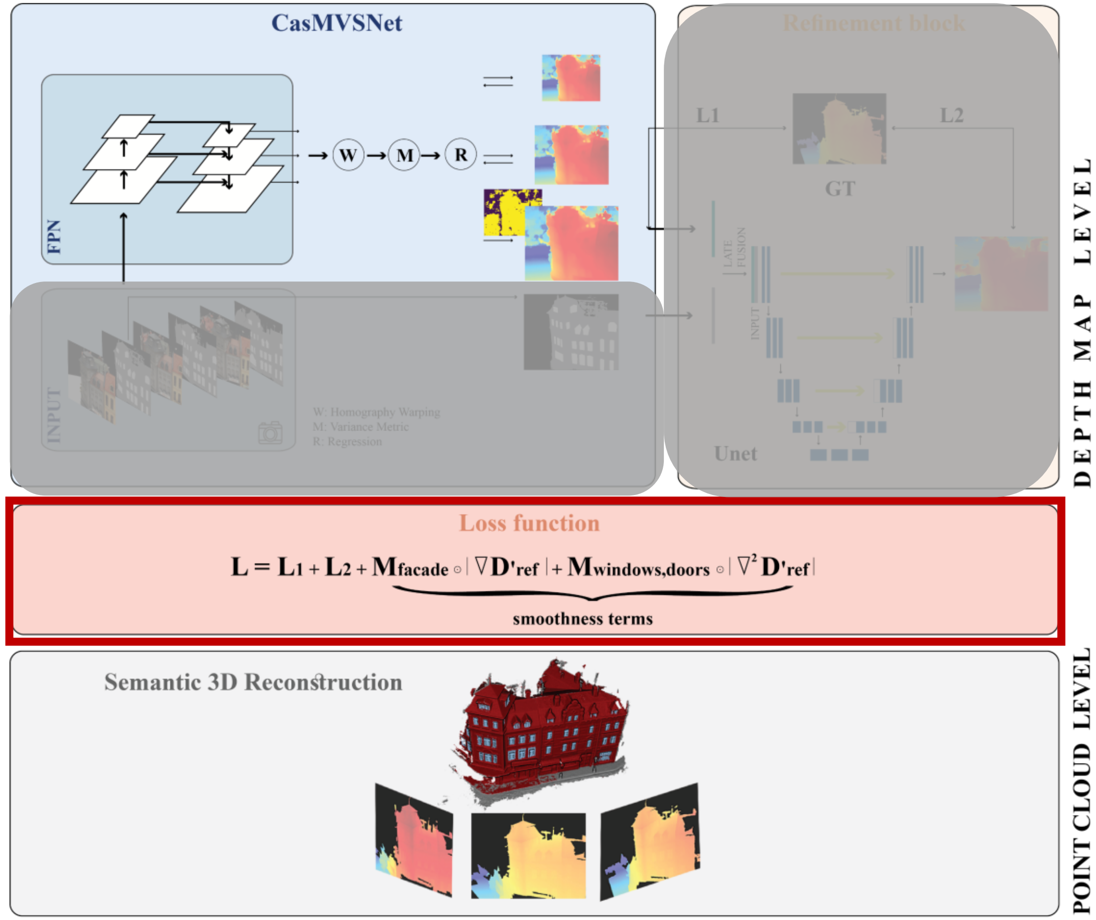
network trained solely with the refinement block



Ablation study 3



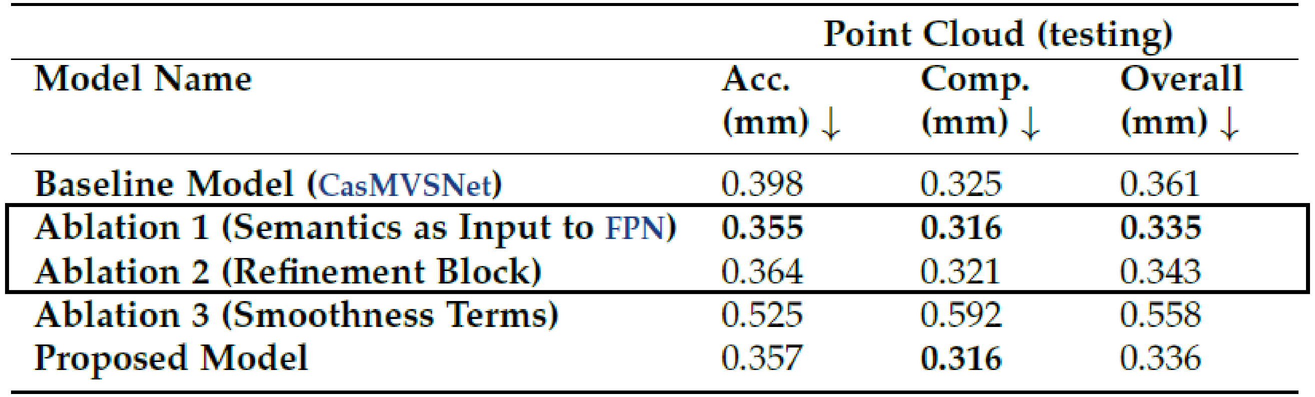
	Point Cloud (testing)		
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336



Ablation 3:

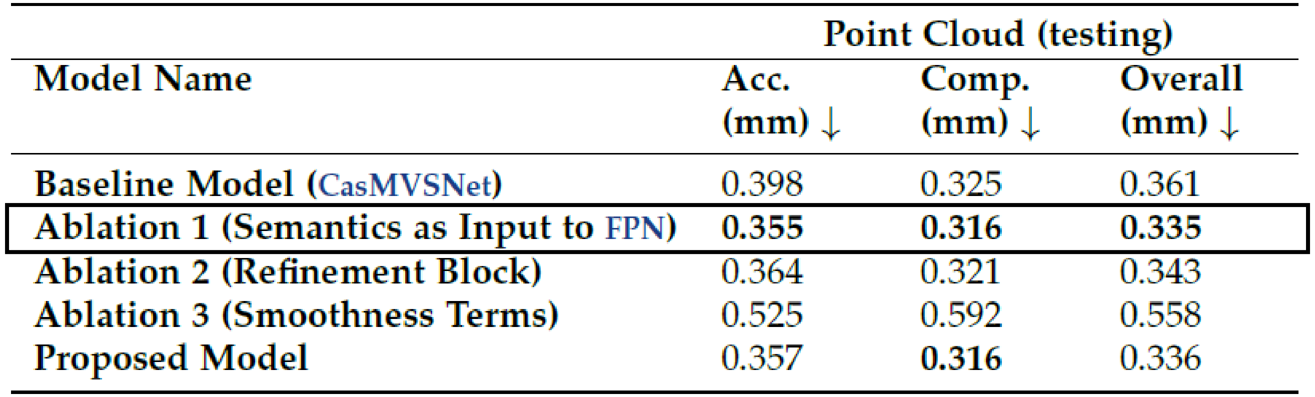
network trained separately with only the smoothness terms

Ablation study



	Point Cloud (testing)		
Model Name	**Acc. (mm) ↓**	**Comp. (mm) ↓**	**Overall (mm) ↓**
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336

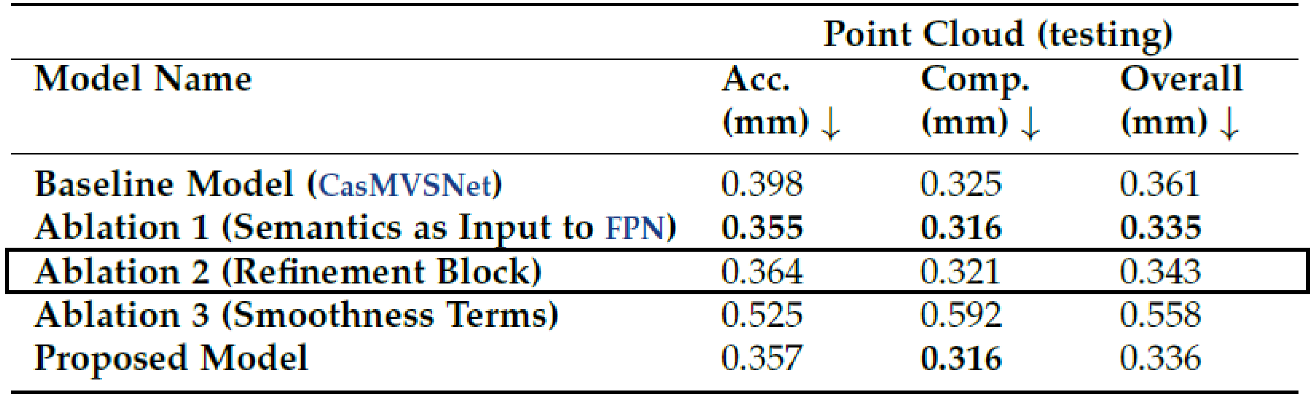
Ablation study



Model Name	Point Cloud (testing) Acc. (mm) ↓	 Comp. (mm) ↓	 Overall (mm) ↓
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336

- **Semantics as Input to FPN** (Ablation 1)

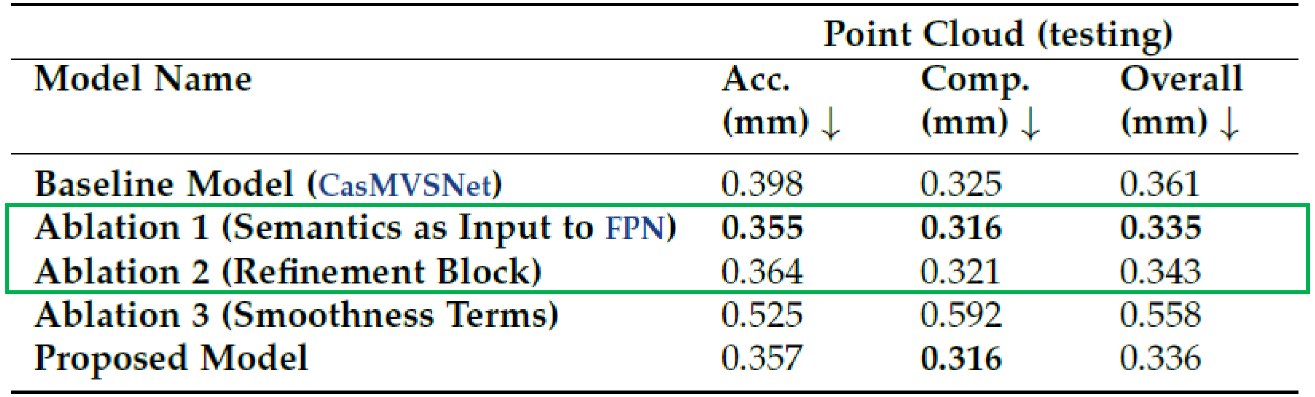
Ablation study



	Point Cloud (testing)		
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336

- **Semantics as Input to FPN** (Ablation 1)
- **Refinement Block** (Ablation 2)

Ablation study

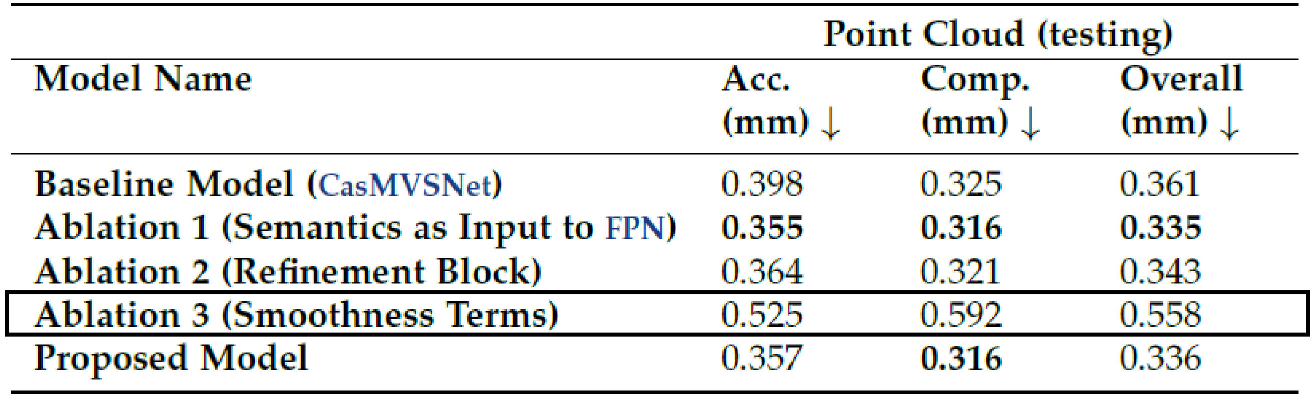


	Point Cloud (testing)		
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336

Proved effective:

- **Semantics as Input to FPN** (Ablation 1)
- **Refinement Block** (Ablation 2)

Ablation study



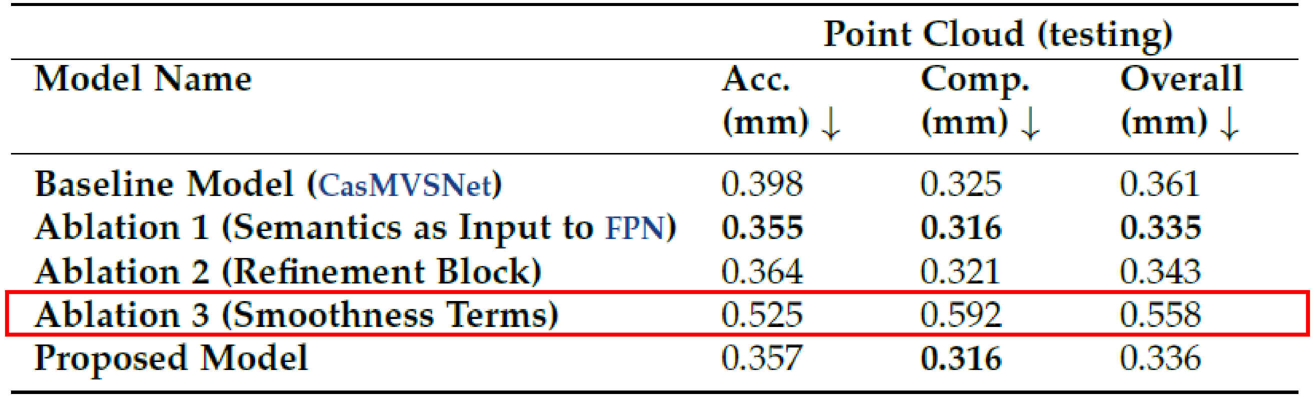
	Point Cloud (testing)		
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336

Proved effective:

- **Semantics as Input to FPN** (Ablation 1)
- **Refinement Block** (Ablation 2)

- **Smoothness Terms** (Ablation 3)

Ablation study



	Point Cloud (testing)		
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336

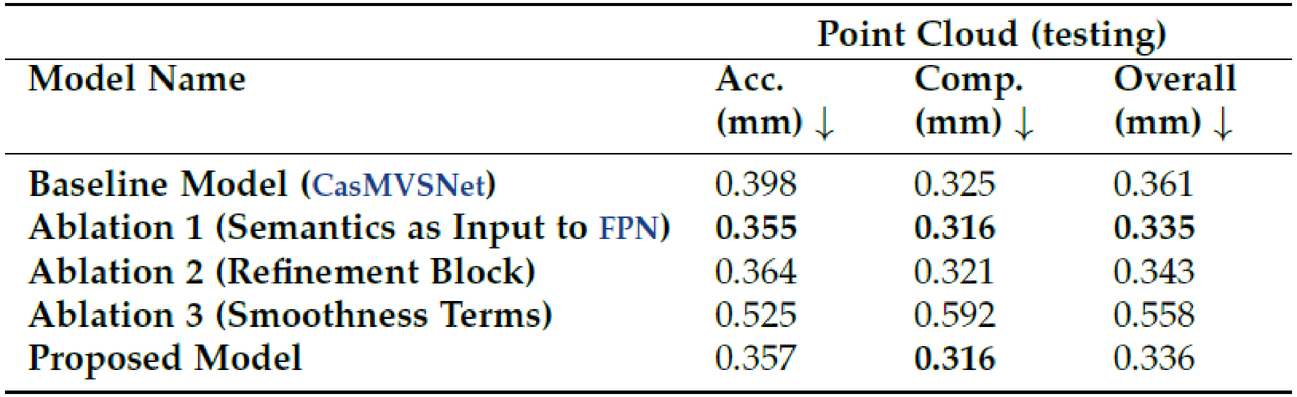
Proved effective:

- **Semantics as Input to FPN** (Ablation 1)
- **Refinement Block** (Ablation 2)

Negative impact:

- **Smoothness Terms** (Ablation 3)

Ablation study



	Point Cloud (testing)		
Model Name	**Acc. (mm) ↓**	**Comp. (mm) ↓**	**Overall (mm) ↓**
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336

Proved effective:

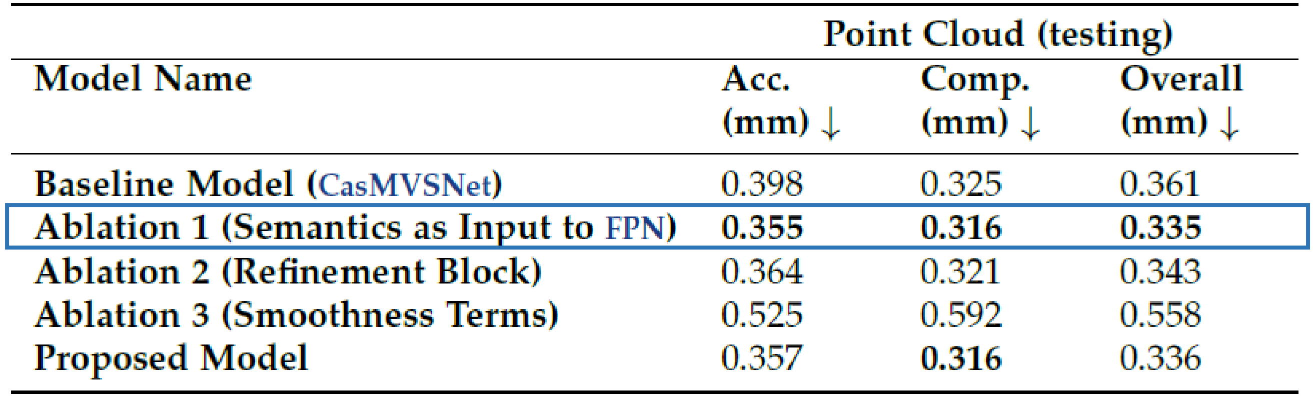
- **Semantics as Input to FPN** (Ablation 1)
- **Refinement Block** (Ablation 2)

Negative impact:

- **Smoothness Terms** (Ablation 3)

Interestingly,

Ablation study



	Point Cloud (testing)		
Model Name	**Acc. (mm) ↓**	**Comp. (mm) ↓**	**Overall (mm) ↓**
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336

Proved effective:

- **Semantics as Input to FPN** (Ablation 1)
- **Refinement Block** (Ablation 2)

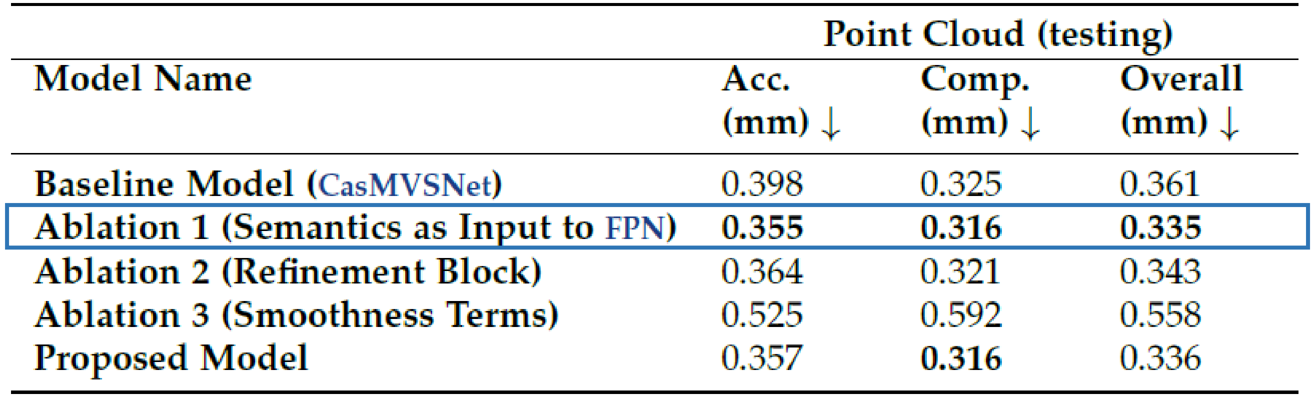
Negative impact:

- **Smoothness Terms** (Ablation 3)

Interestingly,

- solely the use of the **Semantics as Input to FPN**

Ablation study



	Point Cloud (testing)		
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336

Proved effective:

- **Semantics as Input to FPN** (Ablation 1)
- **Refinement Block** (Ablation 2)

Negative impact:

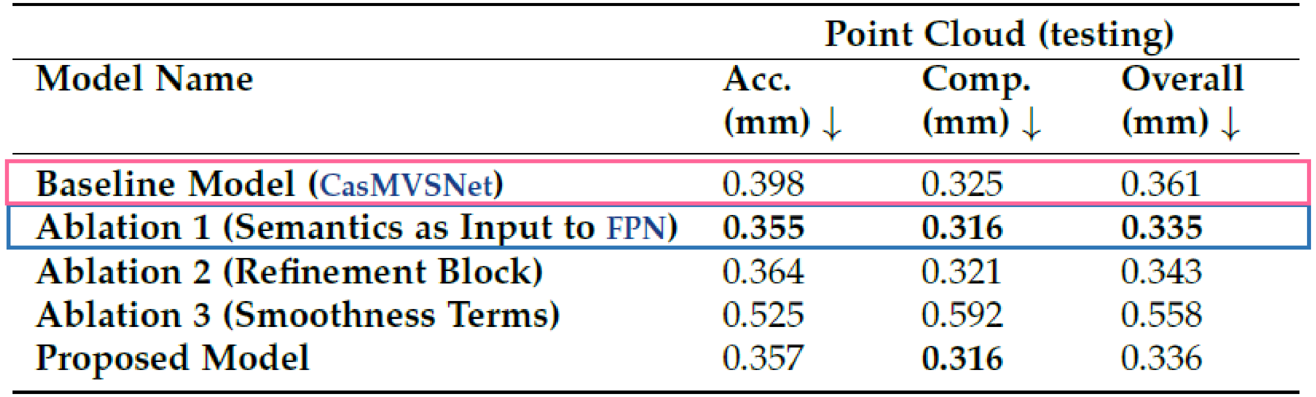
- **Smoothness Terms** (Ablation 3)

Interestingly,

- solely the use of the **Semantics as Input to FPN**

  PROVED SUFFICIENT to elevate the model's performance …

Ablation study



	Point Cloud (testing)		
Model Name	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓
Baseline Model (CasMVSNet)	0.398	0.325	0.361
Ablation 1 (Semantics as Input to FPN)	**0.355**	**0.316**	**0.335**
Ablation 2 (Refinement Block)	0.364	0.321	0.343
Ablation 3 (Smoothness Terms)	0.525	0.592	0.558
Proposed Model	0.357	**0.316**	0.336

Proved effective:

- **Semantics as Input to FPN** (Ablation 1)
- **Refinement Block** (Ablation 2)

Negative impact:

- **Smoothness Terms** (Ablation 3)

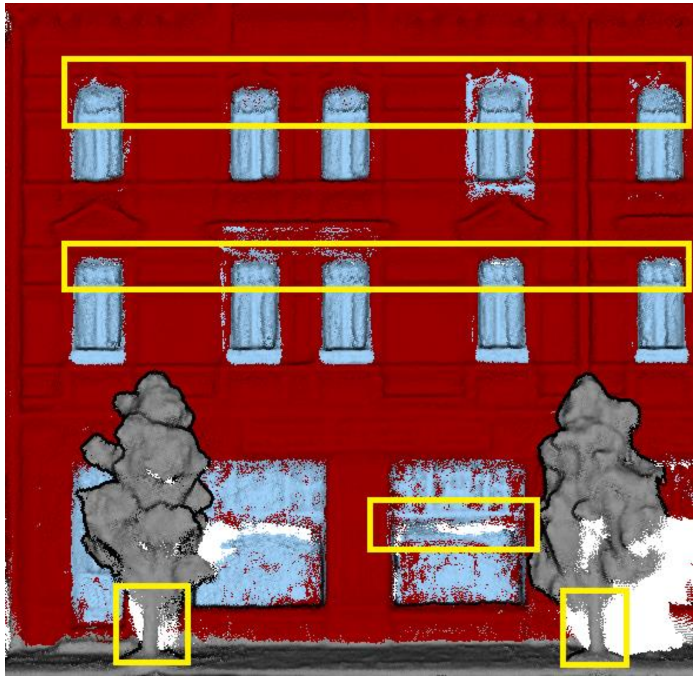
Interestingly,

- solely the use of the **Semantics as Input to FPN**

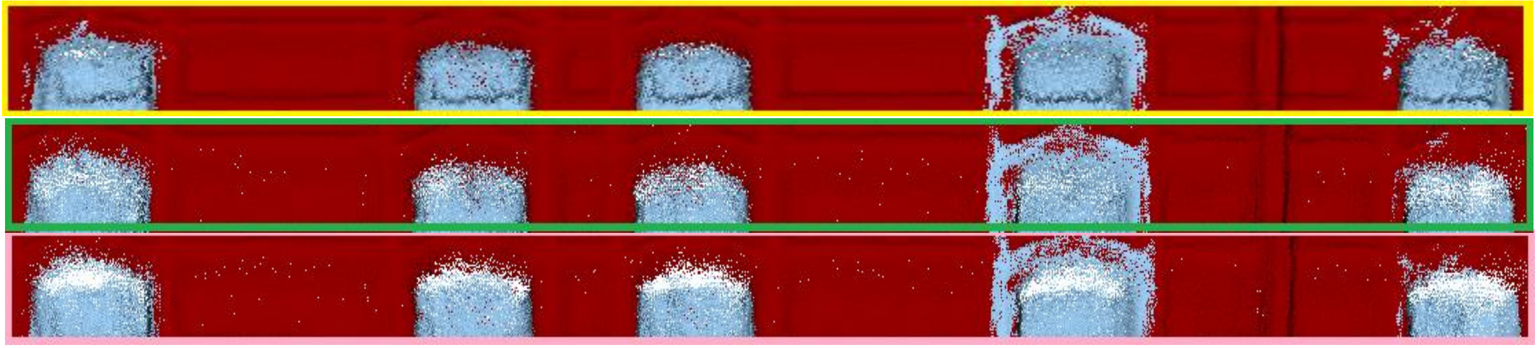
PROVED SUFFICIENT to elevate the model's performance …

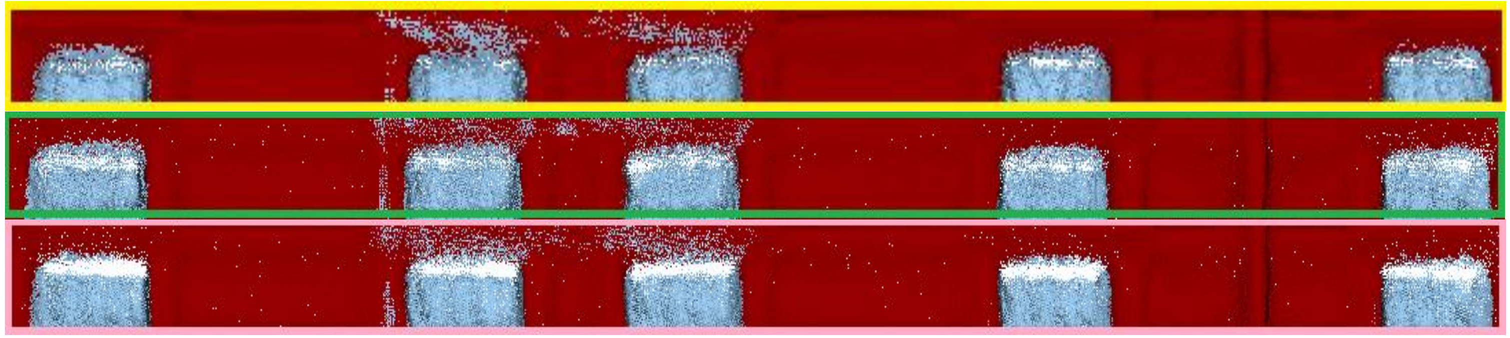
beyond the Baseline results.

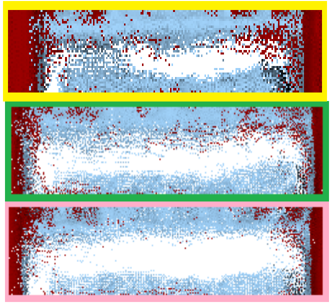
Ablation 1: Semantics as Input to FPN



Ablation 1





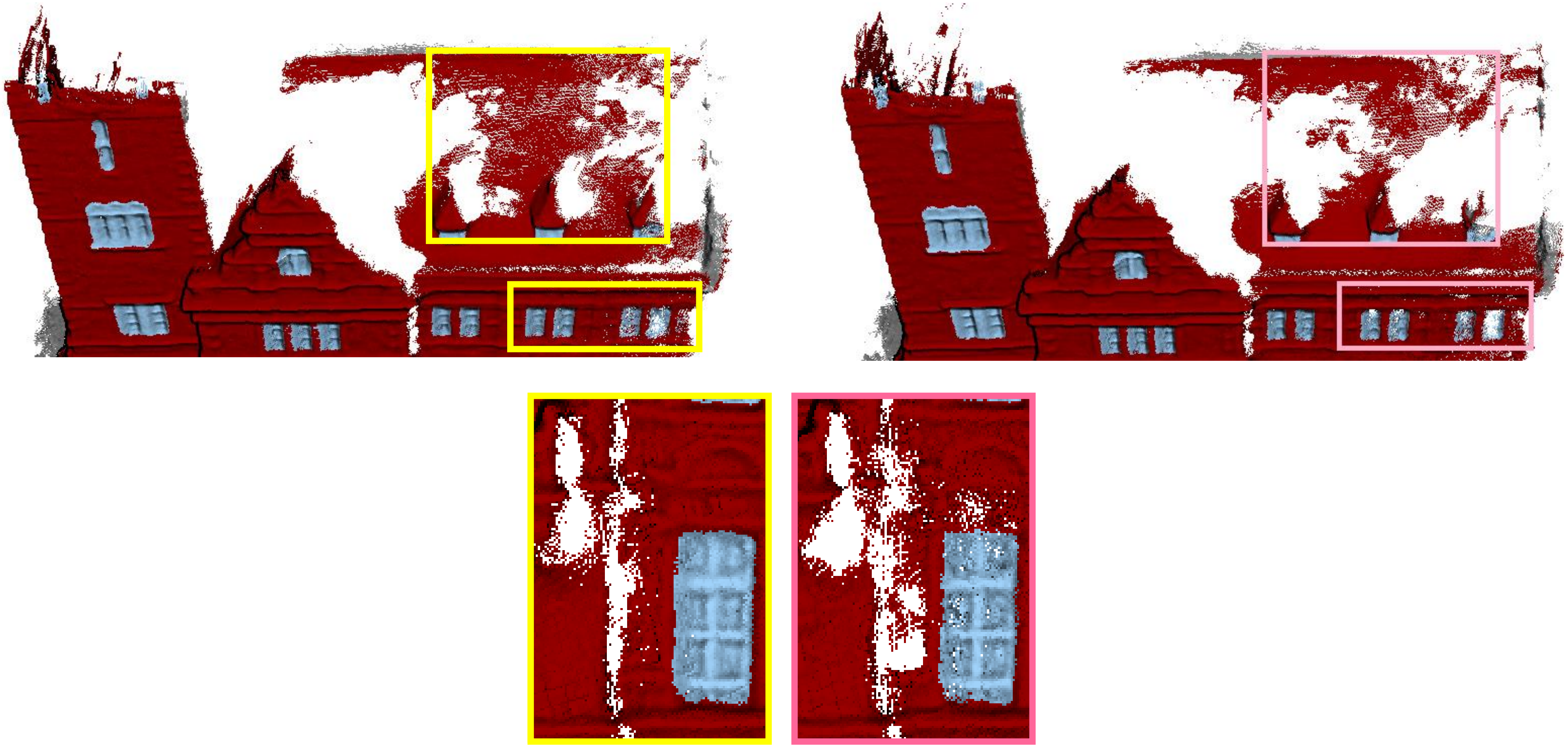


Ablation 1

Proposed

Baseline

Ablation 1: Semantics as Input to FPN

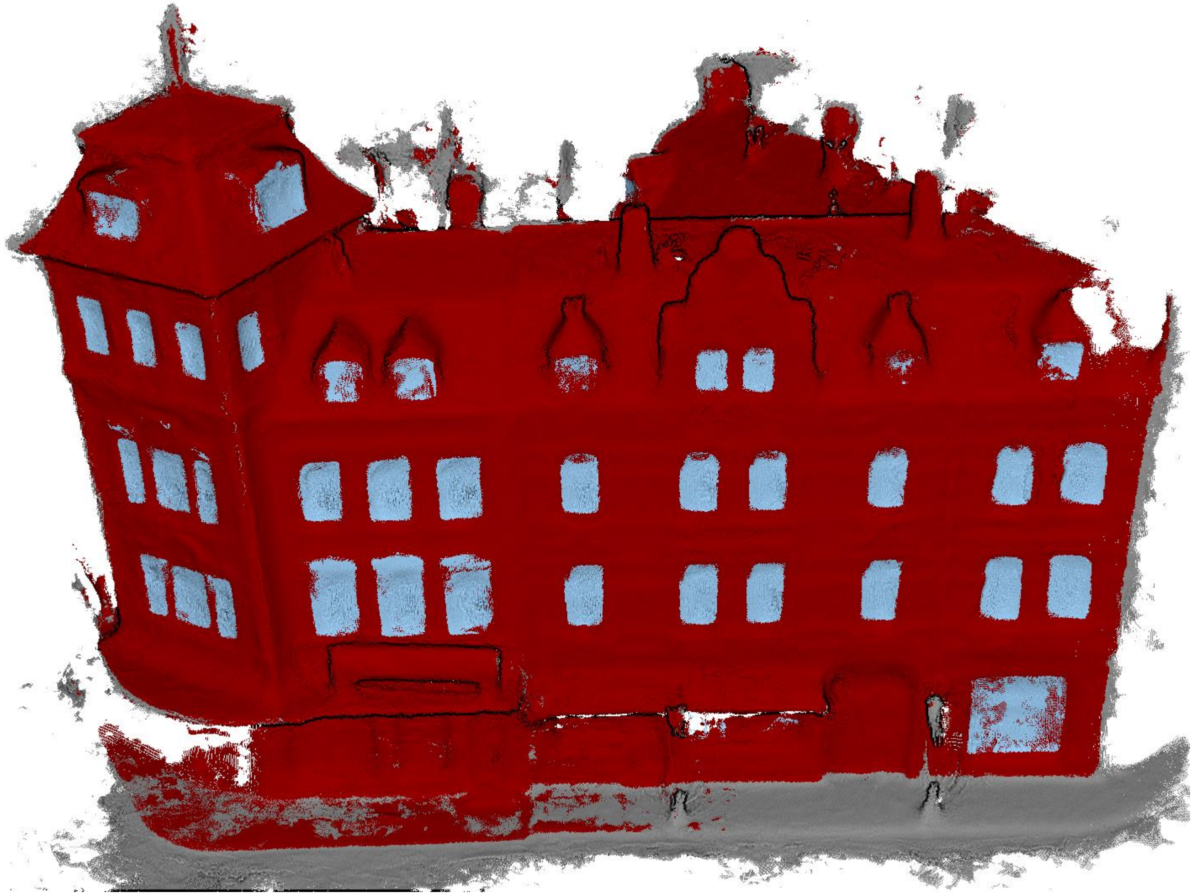


Ablation 1

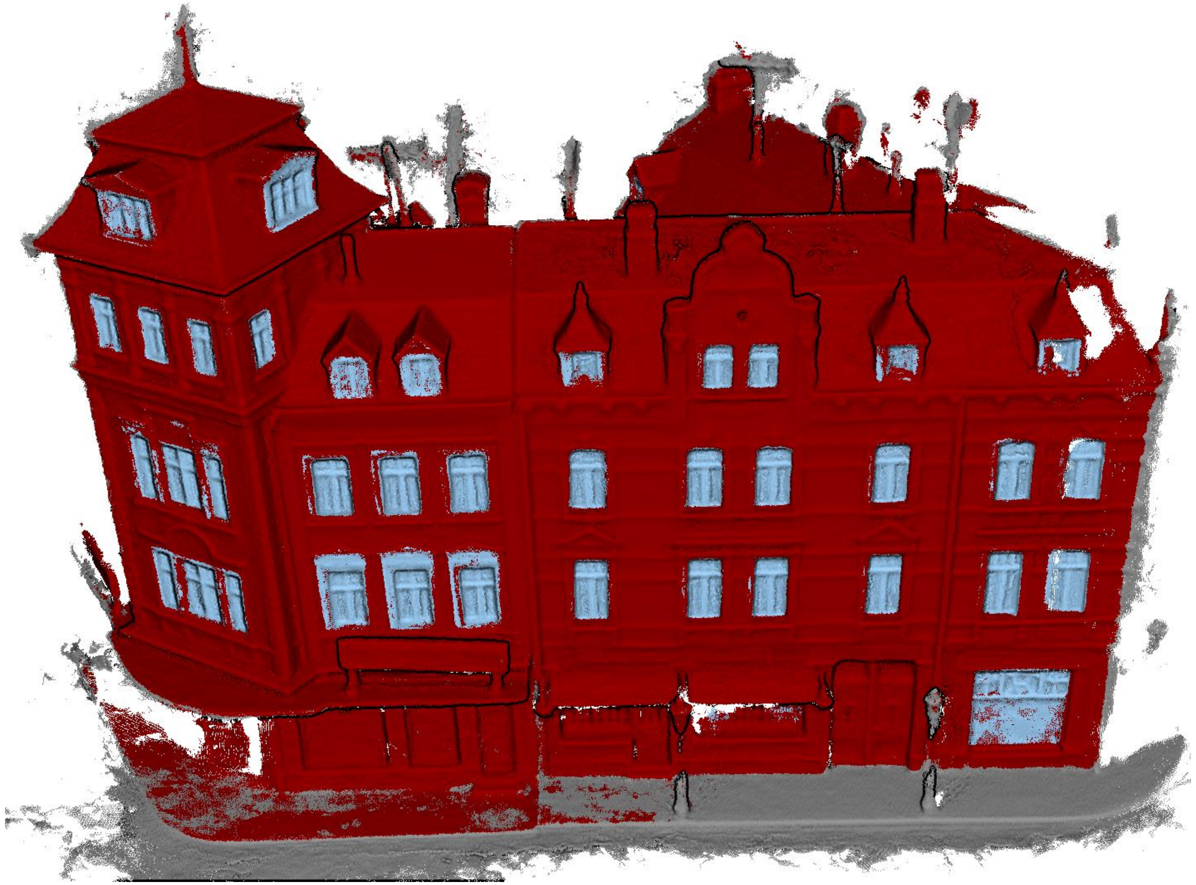
Baseline

Ablation 3: Smoothness Terms

Observation:
planar windows and **smoother facades**, at the **cost of detailed** reconstruction.

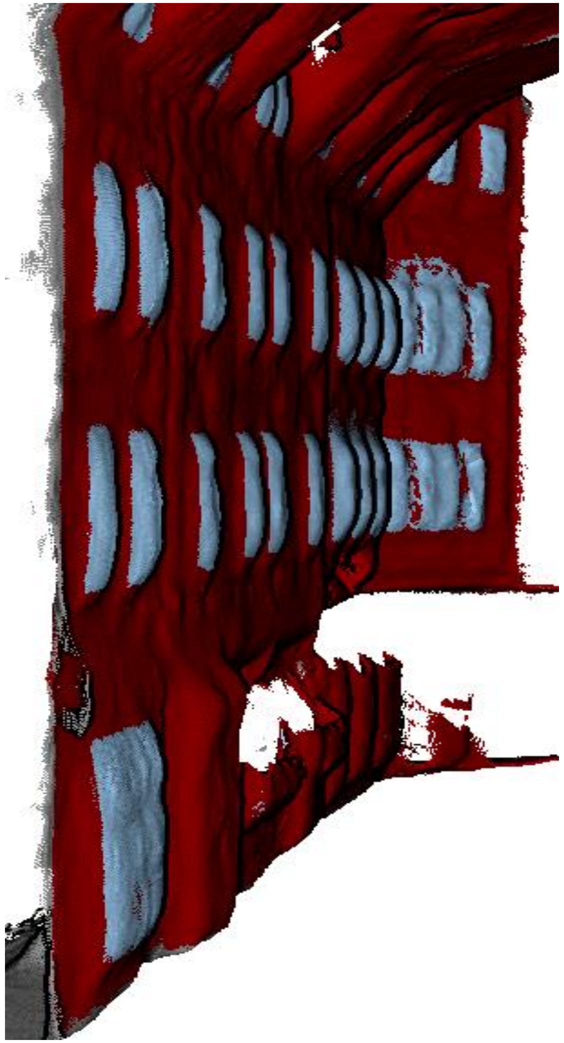


Ablation 3

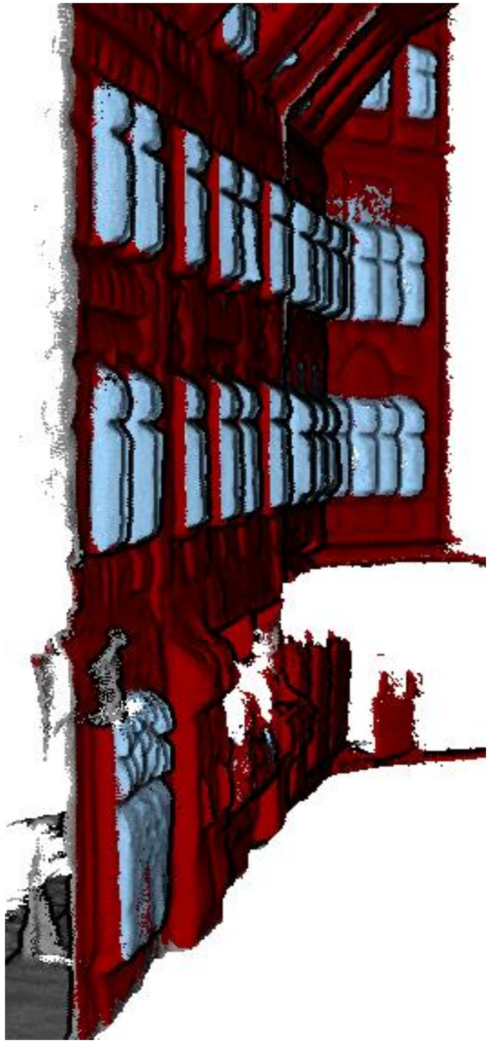


Baseline

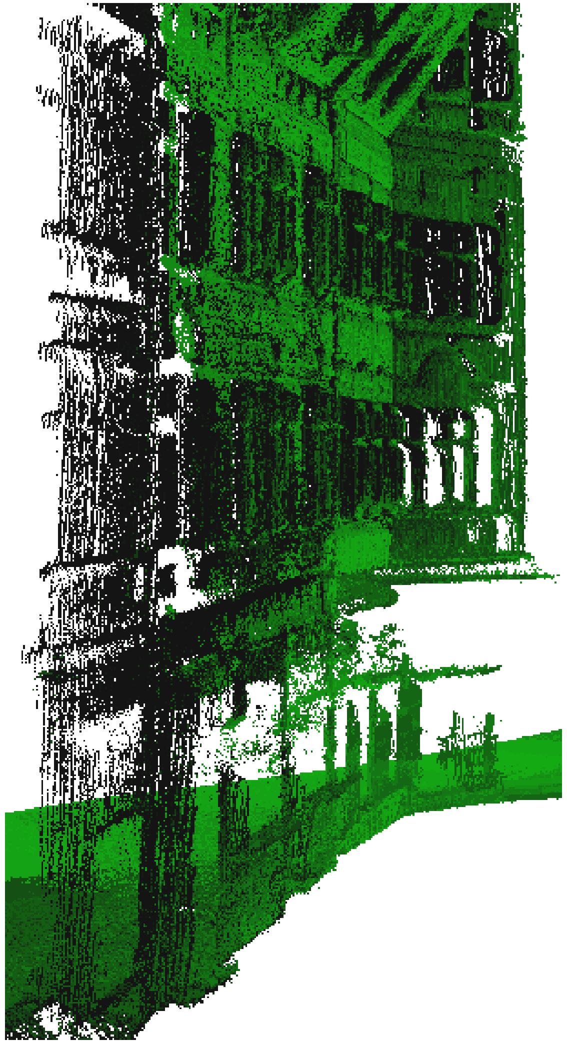
Ablation 3: Smoothness Terms



Ablation 3

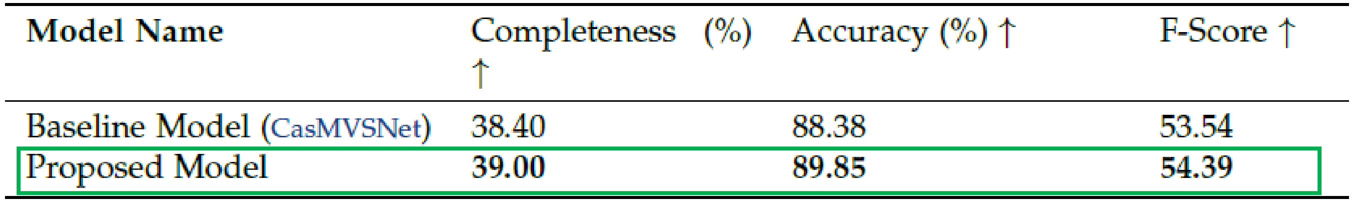


Baseline

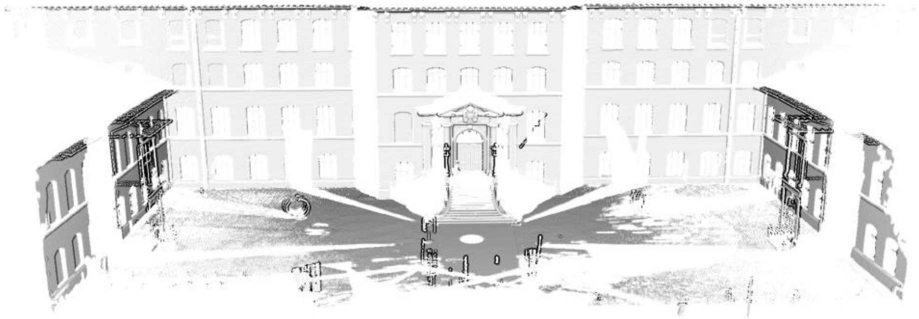


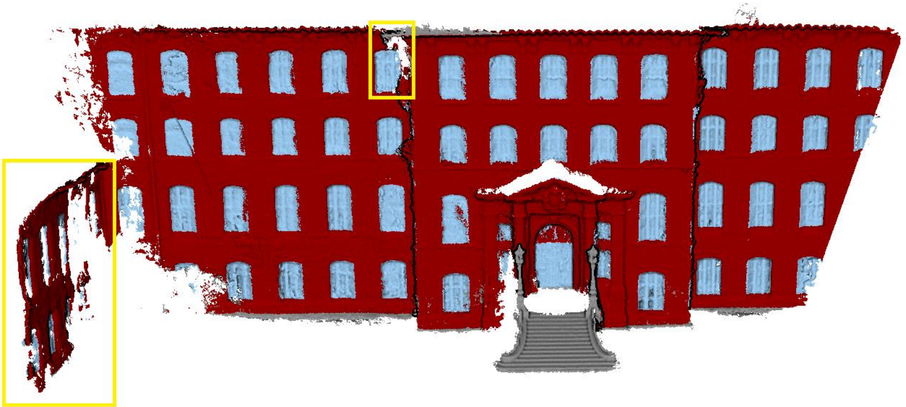
Ground Truth

Generalization to the ETH3D Dataset

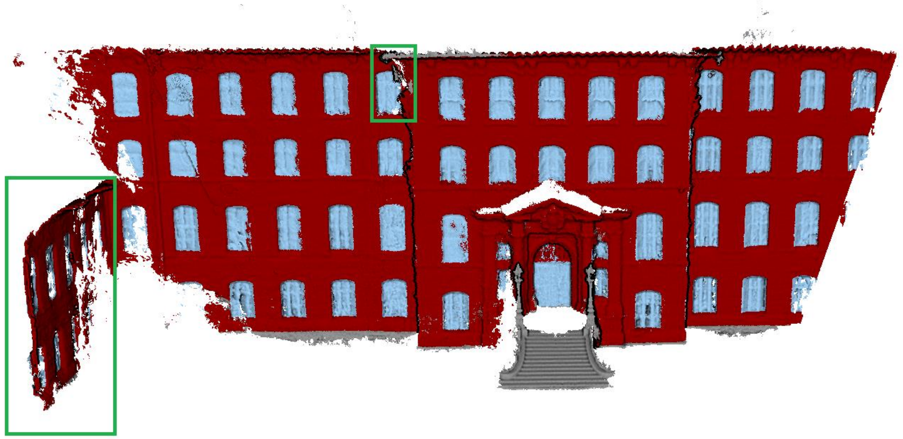


Model Name	Completeness (%) ↑	Accuracy (%) ↑	F-Score ↑
Baseline Model (CasMVSNet)	38.40	88.38	53.54
Proposed Model	**39.00**	**89.85**	**54.39**





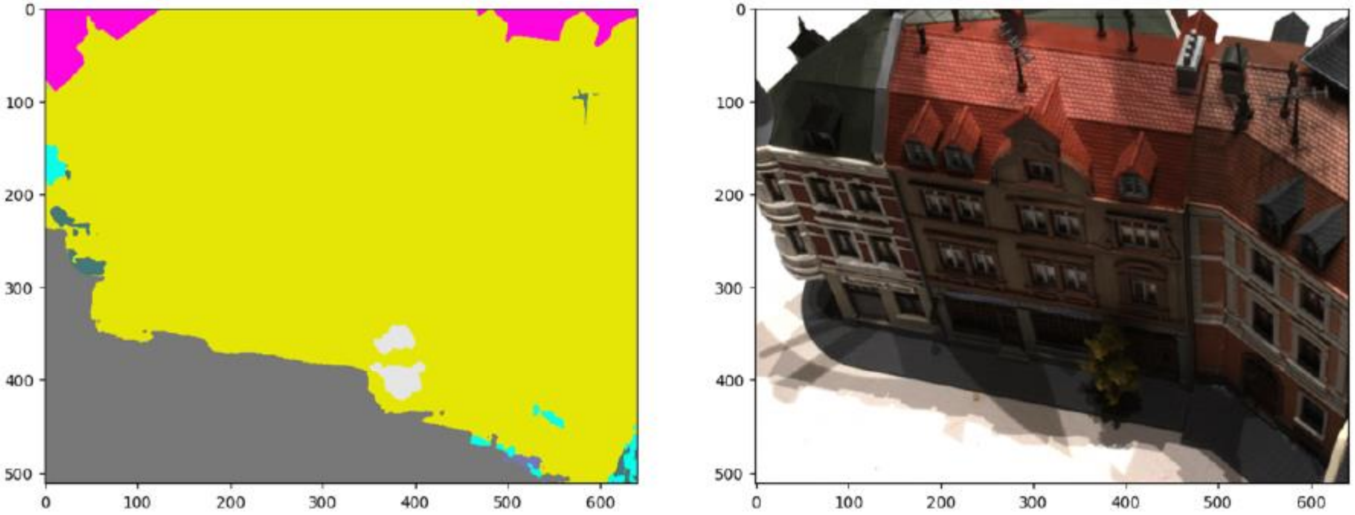
Baseline



Proposed

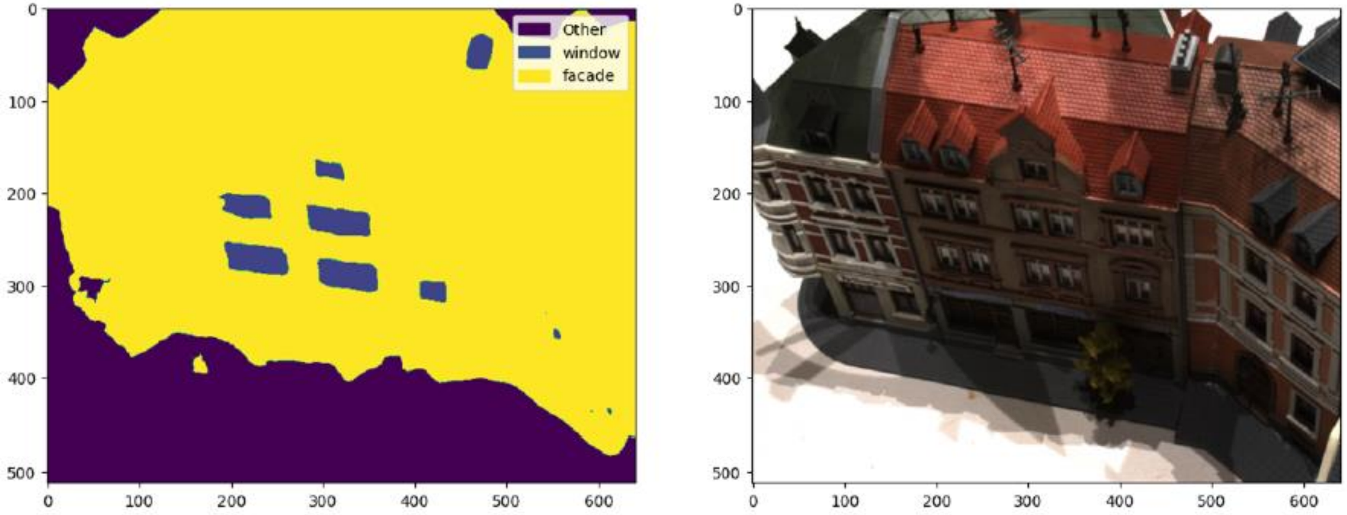
More Semantic Segmentation Results

More Semantic Segmentation Results



🟨 building

Pre-trained SegFormer



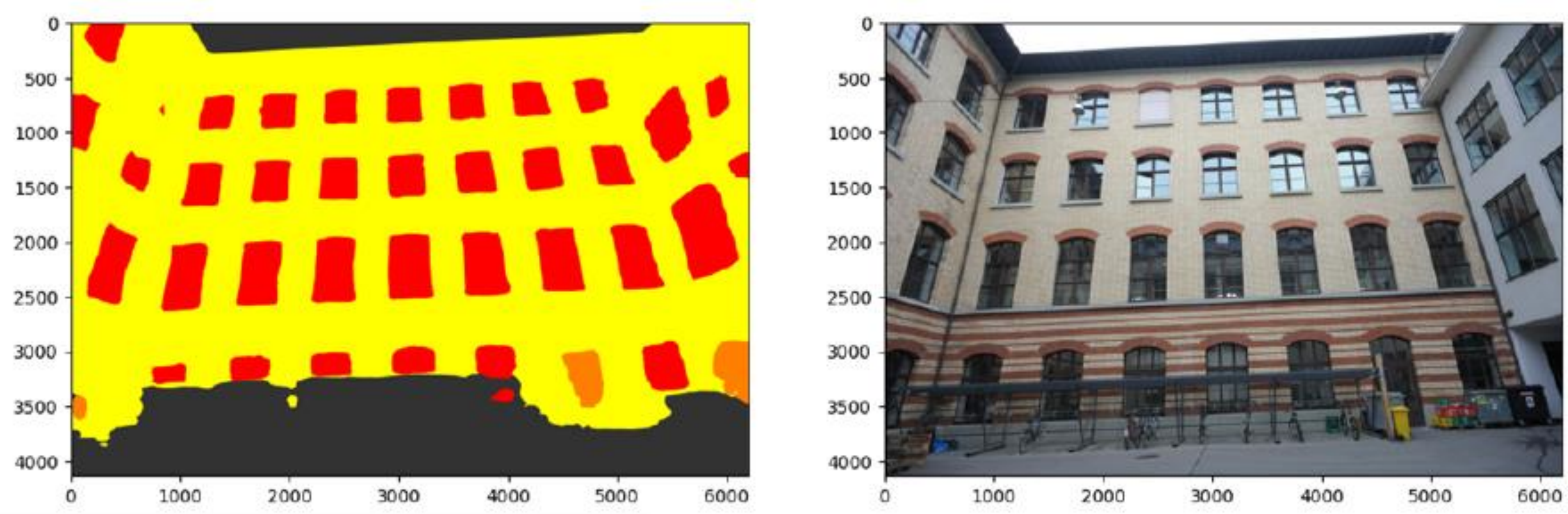
🟪 other

🟦 window

🟨 building

Fine-tuned SegFormer

More Semantic Segmentation Results



⬛ other

🟥 window

🟨 building

Fine-tuned SegFormer

Conclusions

- **Vision Transformer** models are powerful for semantic segmentation.
  - LangSAM, SegFormer (Fine-tuned) performed better on the real-world outdoor dataset

- 3D reconstruction **benefited** from **semantic information**:
  - semantics as input improved the reconstruction for both the DTU and ETH3D dataset

- 3D reconstruction **did not benefit** from **semantic guidance** under the current assumptions
  - Up to the user to prioritize whether the model should conform to the assumption made during its development or to the ground data and vice versa.

Thank you for your attention!

Discussion

References

[Lee et al., 2022] K. T. Lee, E. Liu, J. Yang, L. Hong. An image-guided network for depth edge enhancement (2022)

[Liu et al., 2017] Hantang Liu, Jialiang Zhang, Jianke Zhu, and Steven C. H. Hoi. Deepfacade: A deep learning approach to facade parsing. In Proc. Int. Joint Conf. Artif. Intell., pages 2301–2307, 2017.

[Schmitz and Mayer, 2016] M. Schmitz and H. Mayer. A convolutional network for semantic facade segmentation and interpretation. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLI-B3 (2016)

[Stathopoulou et al., 2021] Stathopoulou EK, Battisti R, Cernea D, Remondino F, Georgopoulos A. Semantically Derived Geometric Constraints for MVS Reconstruction of Textureless Areas. Remote Sens. 2021; 13: 1053.

[Wang et al., 2022] S. Wang, Q. Kang, R. She, W. P. Tay, D. N. Navarro, A. Hartmannsgruber. Building Facade Parsing R-CNN (2022)

[Zhu et al., 2020] J. Zhu, J. Zhang, Y. Cao and Z. Wang, “Image guided depth enhancement via deep fusion and local linear regularization”, in IEEE International Conference on Image Processing (ICIP), (2017)