3D Scanning & Spatial Learning

RGB Self-supervised MVS Reconstruction

Di Chang Halil Eralp Koçaş Ran Ding

Technical University of Munich

Feb. 11, 2022

Overview

- Motivation & Contribution
- Introduction
- Background & Related Work
- Method
- Quantitative and Qualitative Result
- Discussion

Motivation and Contribution

Motivation

- Utilizing NeRF to enhance feature extraction and matching
- Occlusion-aware Method
- More than photometric consistency: Cross-View Rendering Consistency

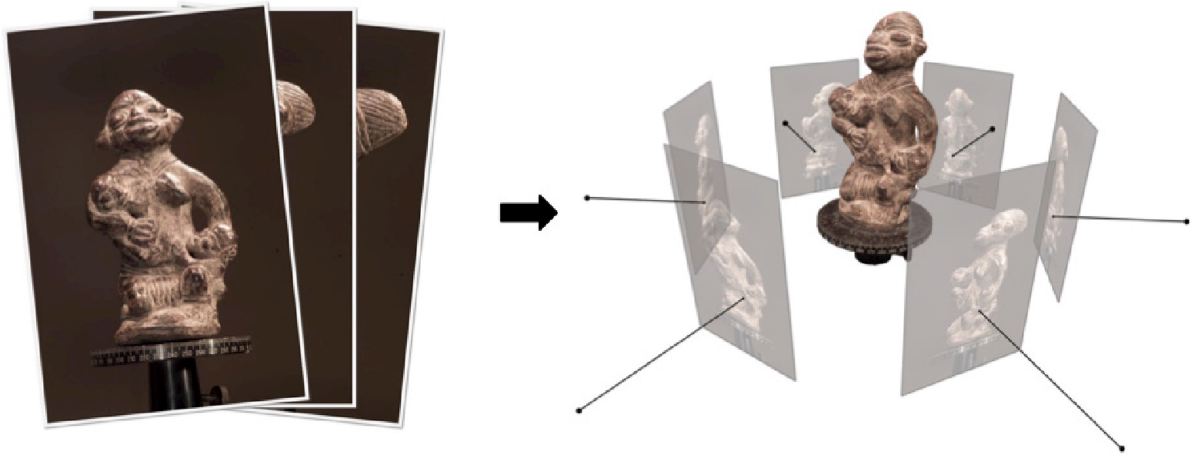
Contribution

- Novel end-to-end learning-based Self-Supervised MVS Depth Inference
- Propose Render Consistency loss
- State-of the art accuracy on challenging DTU Dataset
- Strong generalization ability: SOTA on Tanks&Temples without finetuning

Problem Formulation:

Input:

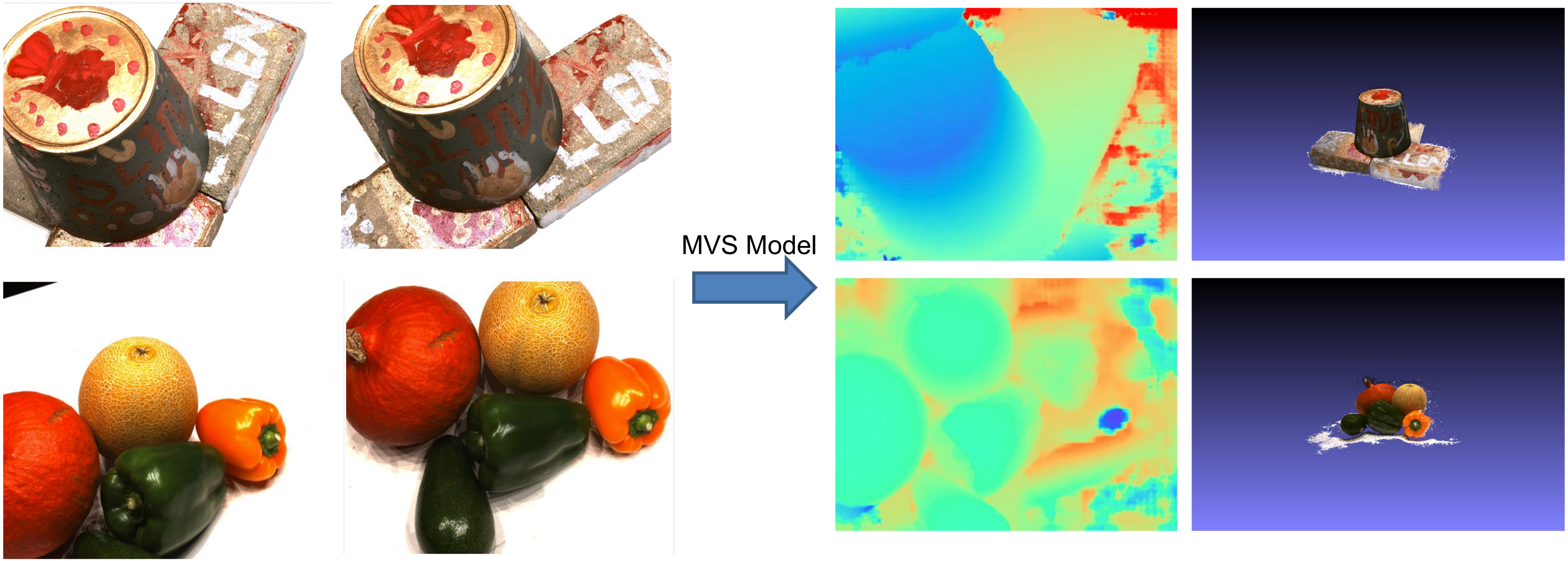
- Multi-View Images of the same scene
  - 1 refence view and several source
- Corresponding camera parameters



Goal:

- Reconstruct RGB Object using Depth Map and Point Clouds

Input & Output



Dataset for training and testing：DTU

- **128 different indoor scenes:**
  - 79 Training scenes
  - 18 Validation scenes
  - 31 Testing scenes

- **Within each scene:**
  - 49 RGB images from different views
  - Corresponding Camera Intrinsic and Extrinsic
  - A point cloud

Dataset for training and testing：DTU

- RGB Image Scan23 View001

- Ground Truth Depth Map Scan23 View001

Dataset for training and testing：DTU

- Predicted depth map Scan23

view 1

view 2

view 3

Fusion

Camera parameters

- Ground Truth Point Cloud Scan23

Dataset for testing：Tanks & Temples

Intermediate dataset:

- 8 outdoor scenes
- Includes: *Family, Francis, Horse, Lighthouse, M60, Panther, Playground, Train*

Advance dataset :

- 6 outdoor scenes
- Includes: *Auditorium, Ballroom, Courtroom, Museum, Palace, Temple*

- ***Training on DTU training set***
- ***Tanks & Temples dataset is only for testing!***

Dataset for testing：Tanks & Temples

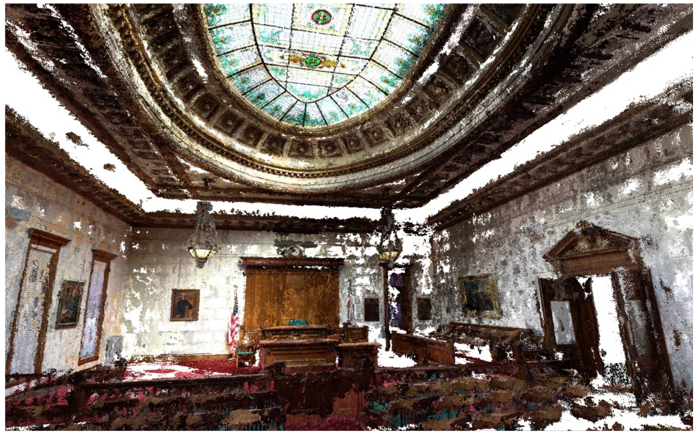
Intermediate dataset:

Family

Panther

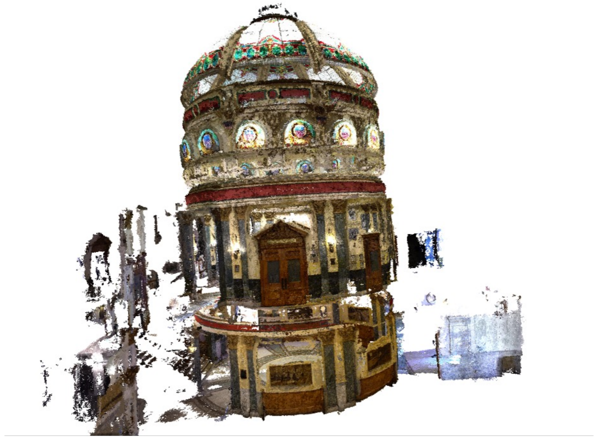
Train

Advance dataset :



Courtroom

Palace



Museum

Related Work: MVSNet

- End-to-end Depth Map inference network.

- MVSNet:
  - 2D Conv Network to feature extraction
  - Differentiable Homography Warping to Cost Volume generation.
  - 3D Conv U-Net to regularize Cost Volume
  - Soft Argmax and 2D Conv Network to obtain refined Depth Map

- SOTA in DTU and Tanks and Temples dataset as of 2018.

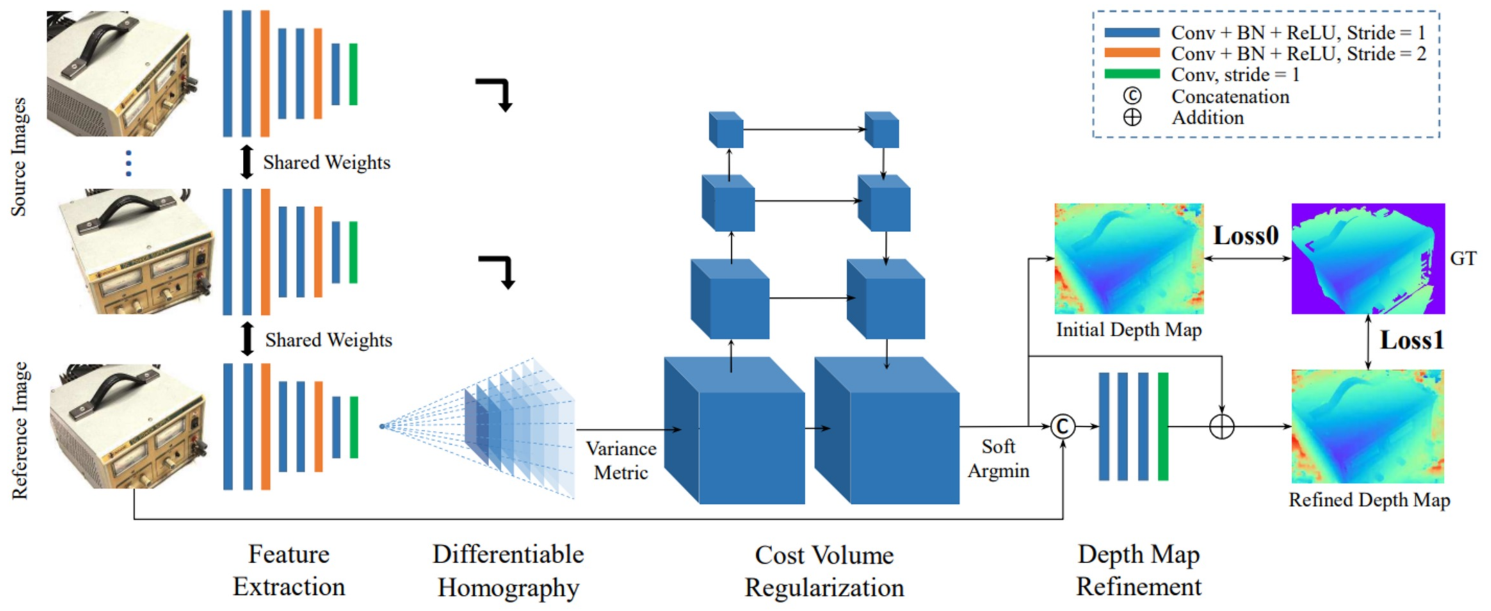


Fig. 1: The network design of MVSNet.

Related Work: MVSNeRF

- Neural rendering approach to reconstruct neural radiance fields for view synthesis.
- Generalizes well across scenes using only several multi-view input images.

- MVSNeRF:
  - 2D Conv Network to feature extraction
  - Differentiable Homography Warping to Cost Volume generation.
  - 3D Conv U-Net to obtain Neural Encoding Volume
  - MLPs and Ray Marching for Depth and RGB pixel rendering.

- Competitive results in DTU.

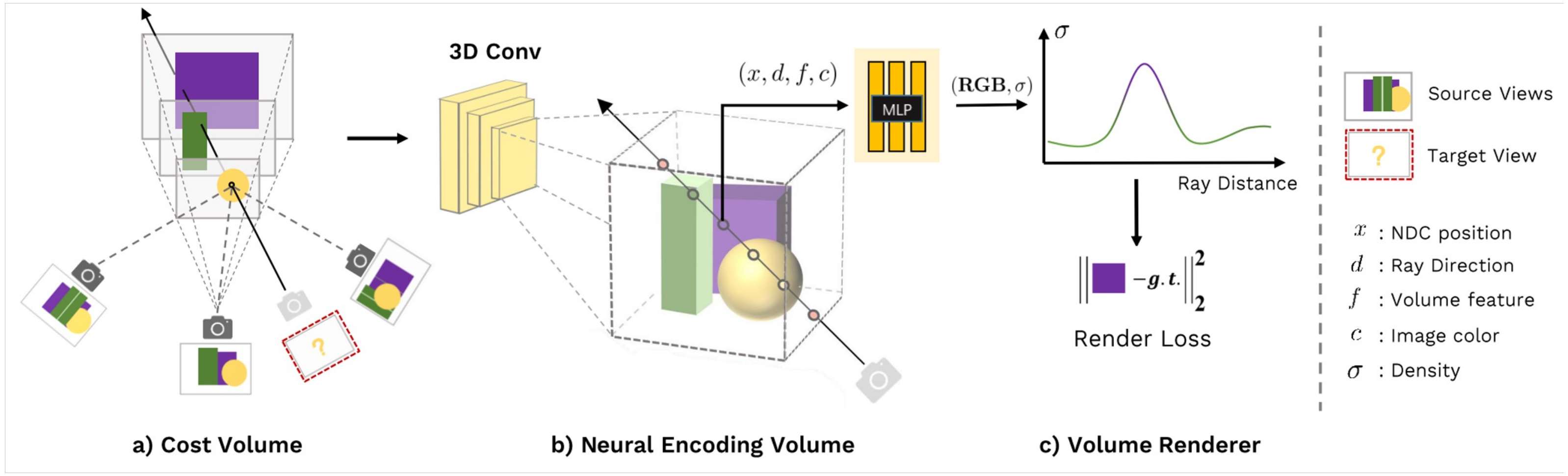


Figure 2. Overview of MVSNeRF.

Related Work: JDACS

- Self-supervised method.
- Addresses *color constancy ambiguity using:*
    - Prior semantic correspondence
    - Prior data augmentation consistency

- Depth Estimation Branch
    - MVSNet with Photometric Consistency Loss
- Data Augmentation Branch
    - Augmentation on reference view
    - MVSNet with Data Augmentation Consistency
- Co-Segmentation Branch
    - Localizes the foreground objects
    - Matrix Factorization to cluster the pixels
    - Semantic Consistency Loss between the reference and source view pixel labels

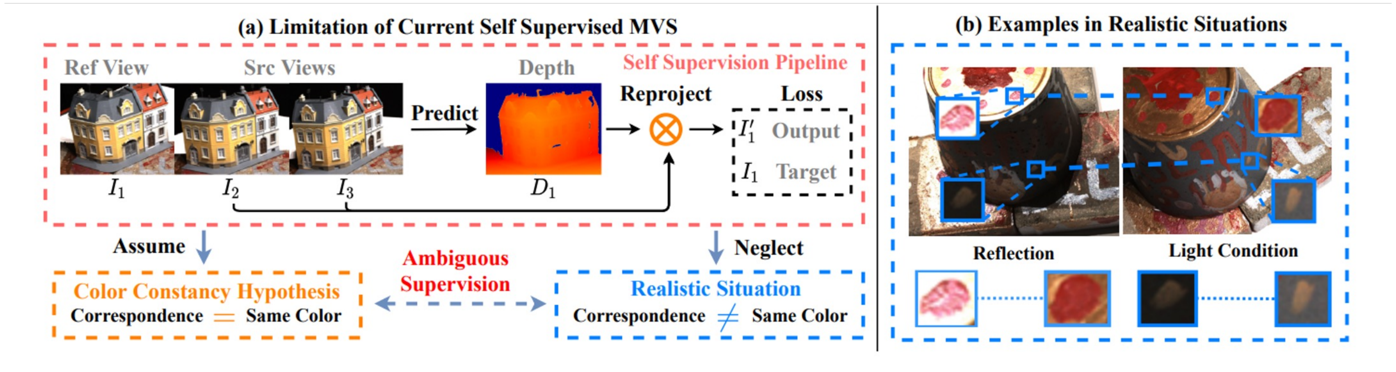


Figure 2: Illustration of the color constancy ambiguity problem in self-supervised MVS.

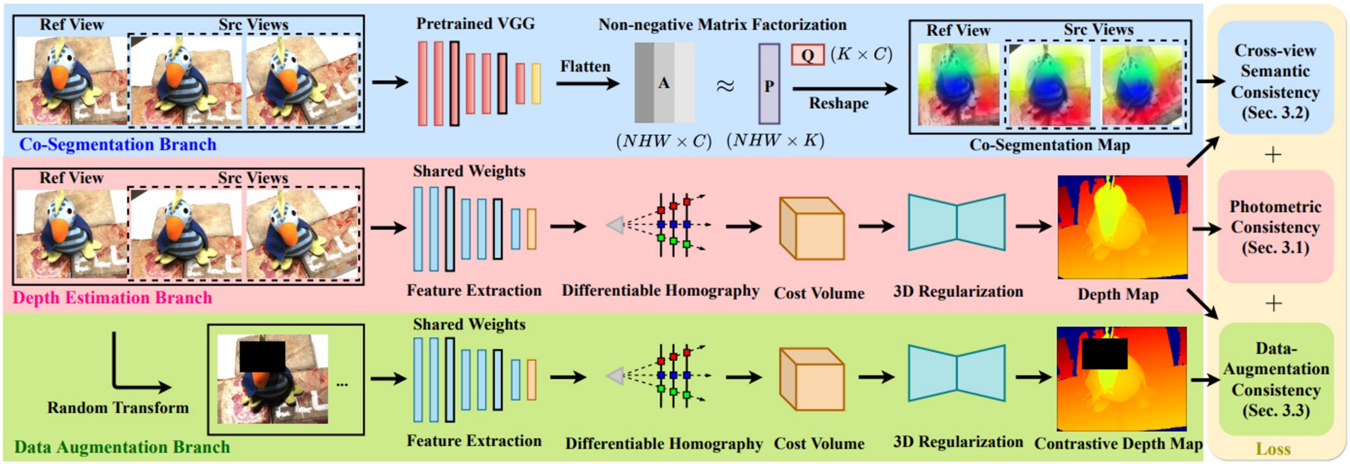
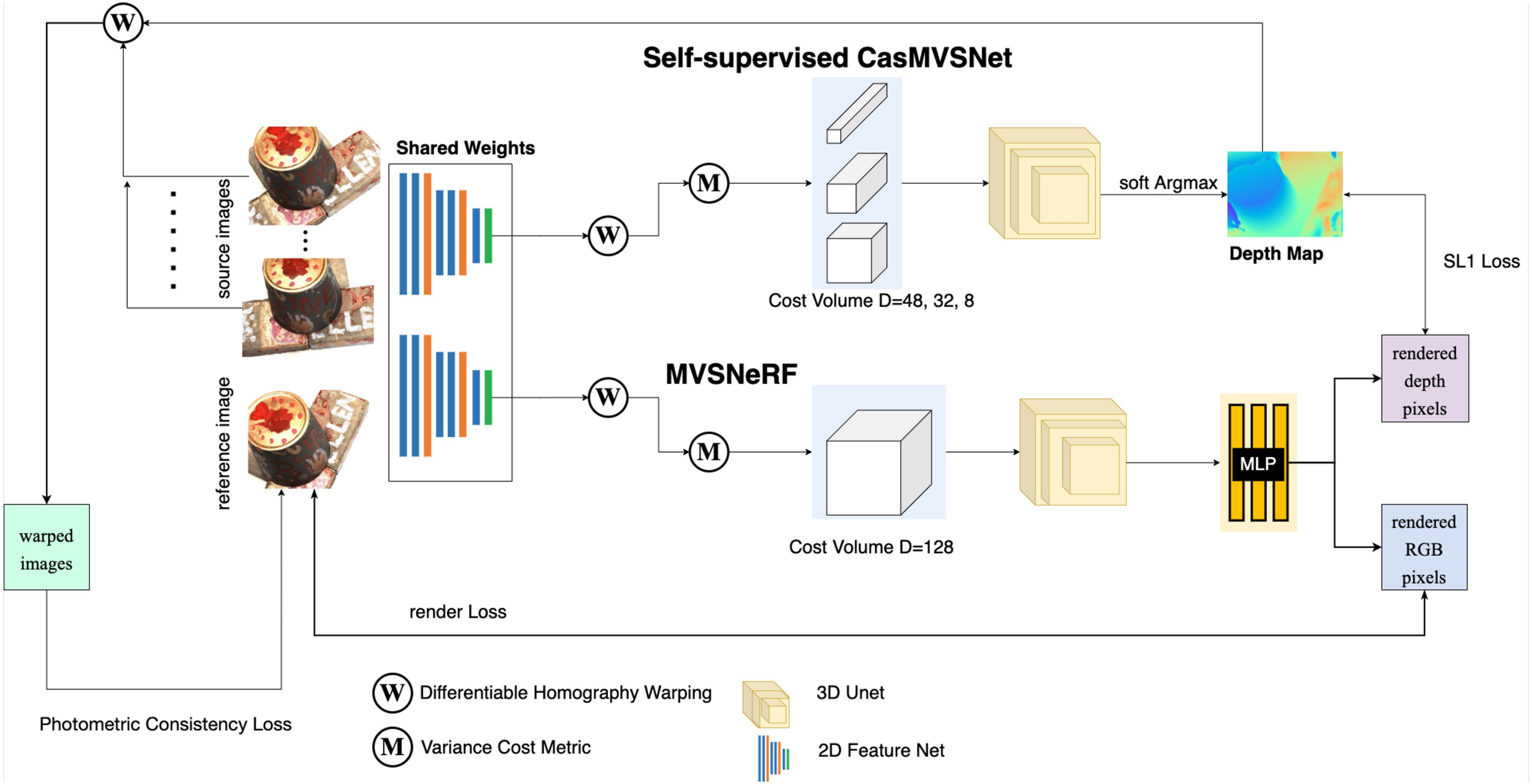
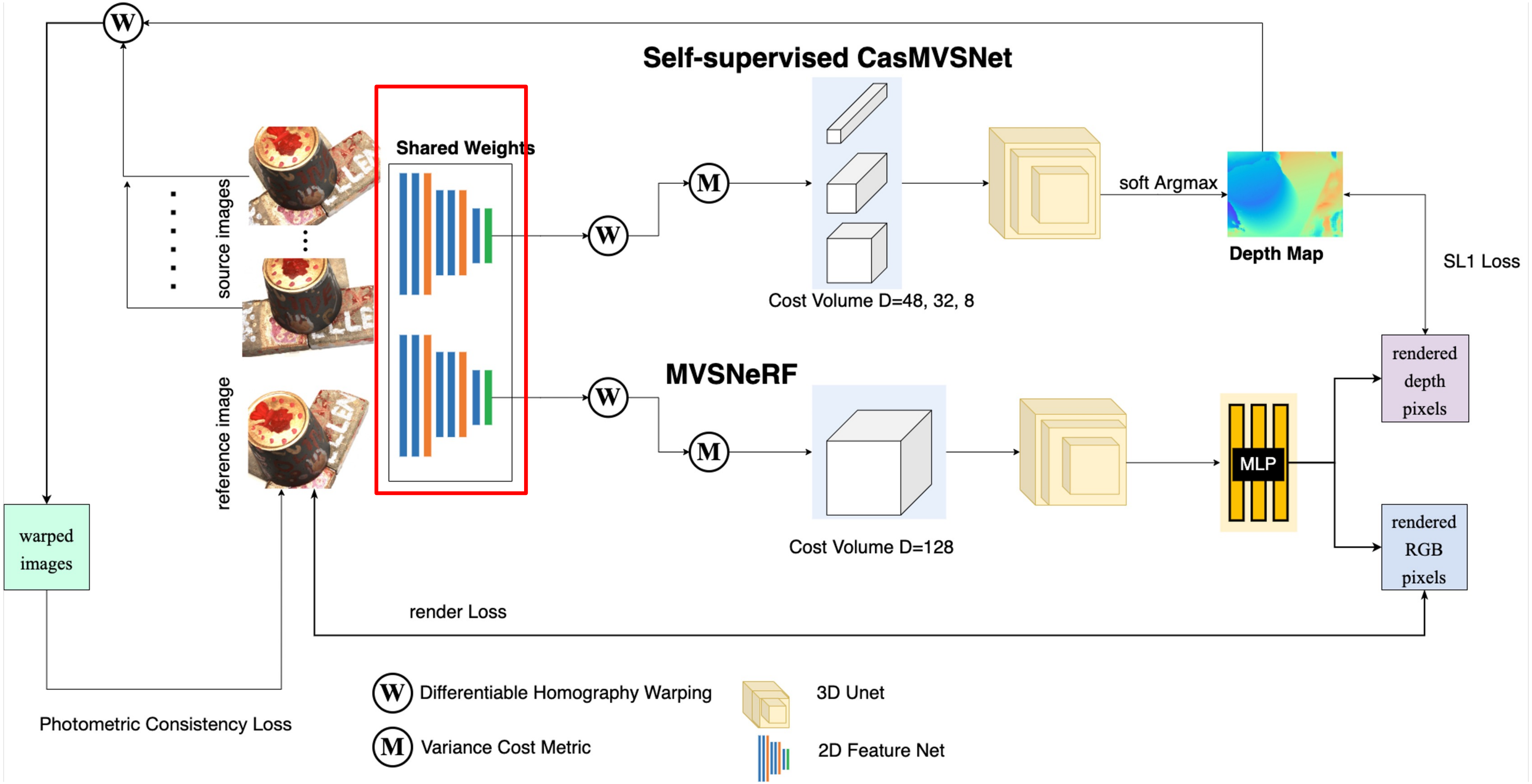


Figure 3: Illustration of our Joint Data-Augmentation and Co-Segmentation (JDACS) MVS framework.

Pipeline: CasNeural-MVSNet



Pipeline: CasNeural-MVSNet

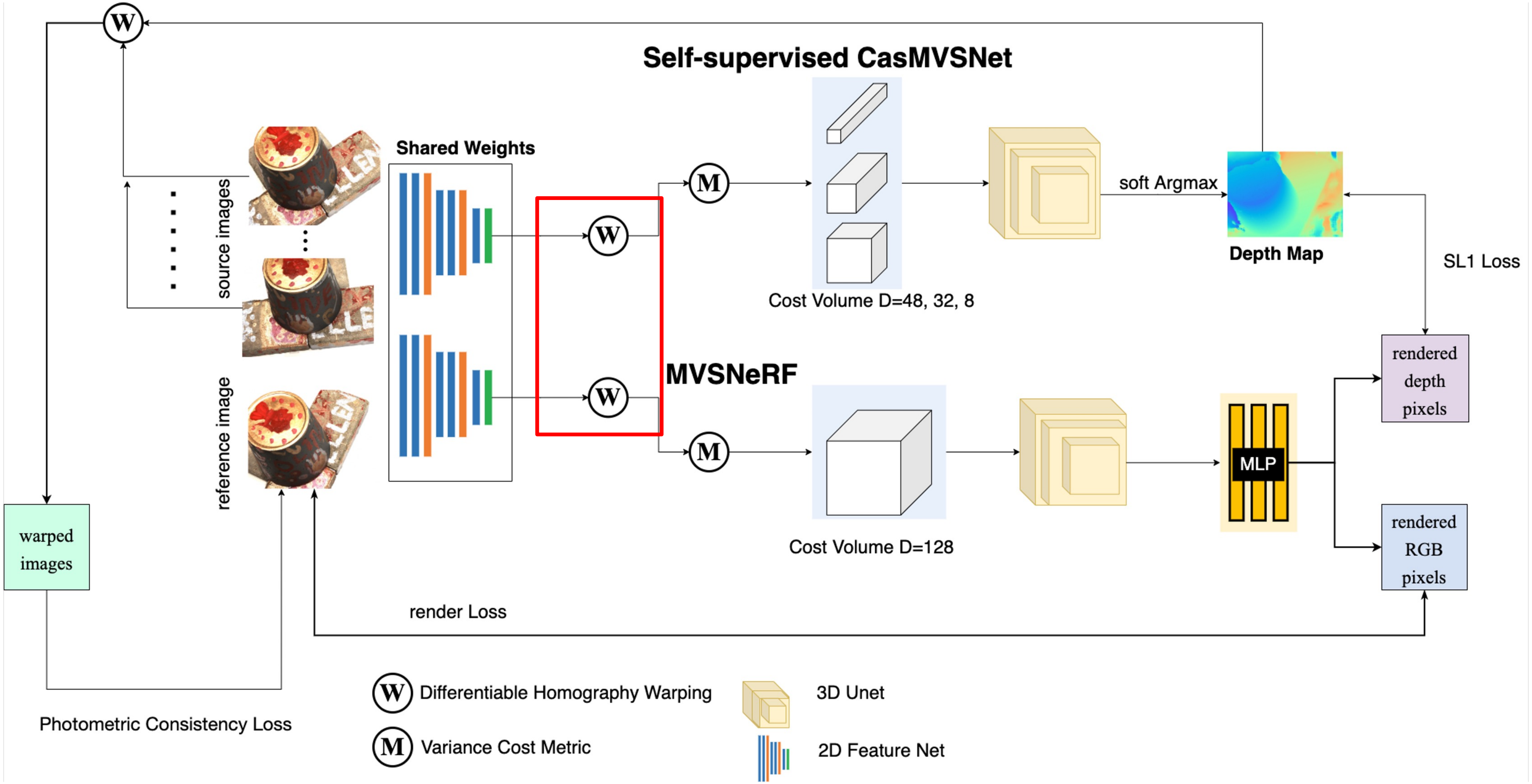


Feature Extraction

Shared-weight eight-layer 2D CNN

Input shape: B x 3 x H x W
Output shape: B x 32 x H x W

Pipeline: CasNeural-MVSNet



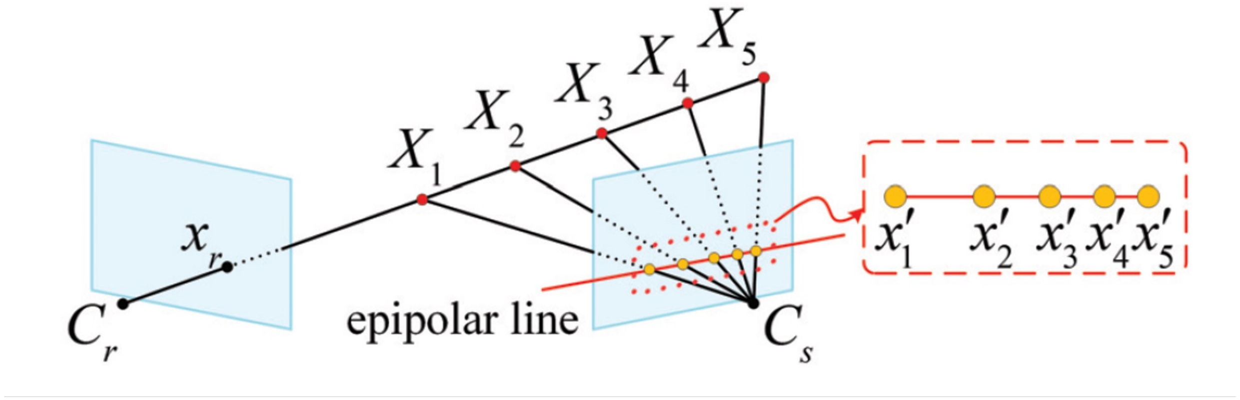
CasNeuralMVSNet: Self-supervised MVS with Neural Rendering

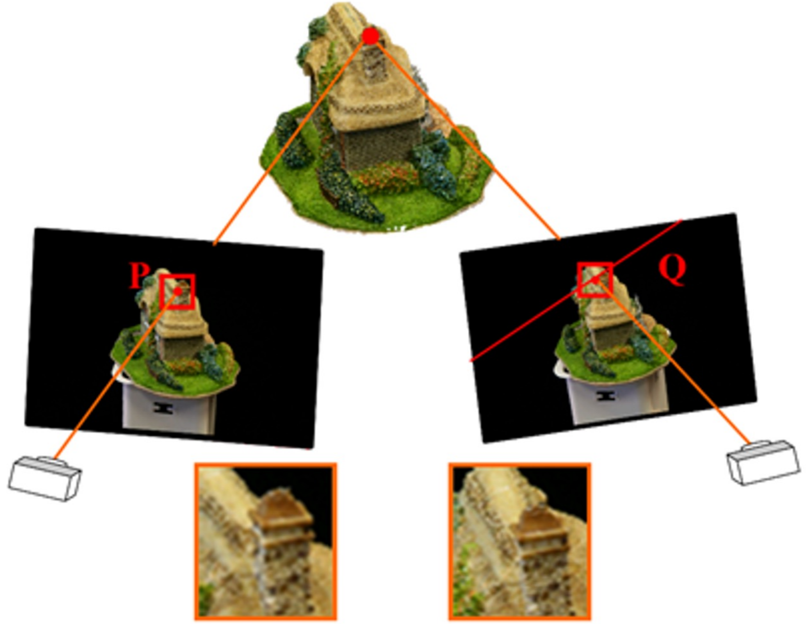
Group 2

Differentiable Homography Warping

Source View Image: Depth + Camera + RGB -> World Point 2D -> 3D

Reference View Image: World Point + Camera -> RGB 3D -> 2D

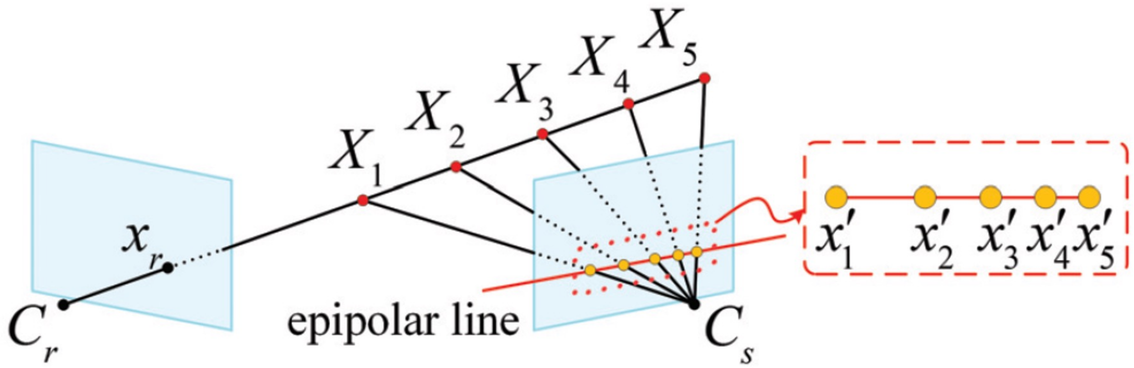


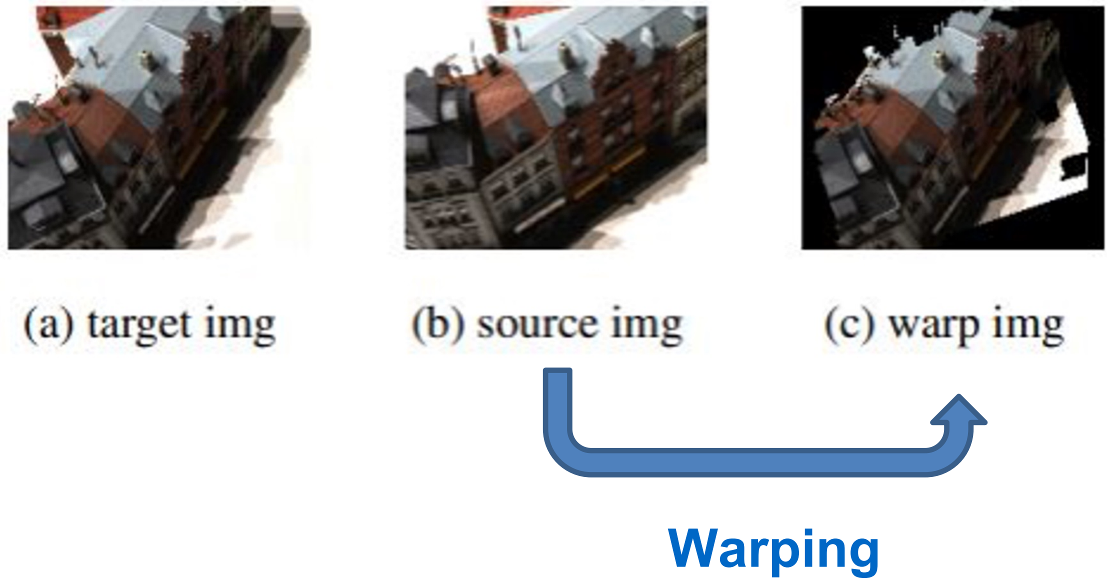


Differentiable Homography Warping

$$\mathbf{H}_i(d) = \mathbf{K}_i \cdot \mathbf{R}_i \cdot \left(\mathbf{I} - \frac{(\mathbf{t}_1 - \mathbf{t}_i) \cdot \mathbf{n}_1^T}{d}\right) \cdot \mathbf{R}_1^T \cdot \mathbf{K}_1^T.$$

All feature maps are warped into different front parallel planes of the reference camera to form feature volumes.
Input shape: B x 32 x H x W.
Output shape: B x 32 x 192 x H x W.





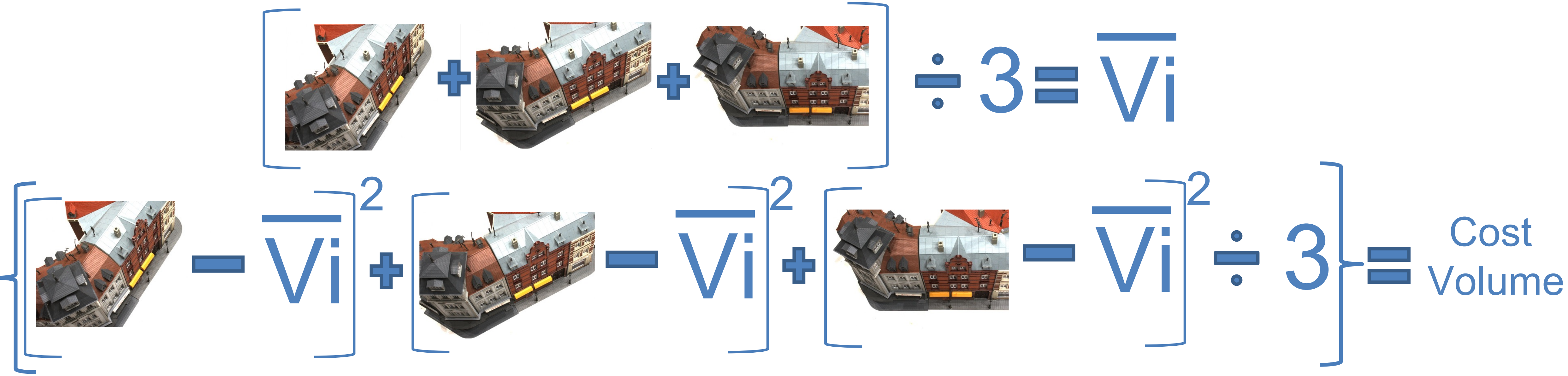
Feature aggregation

Variance Metric

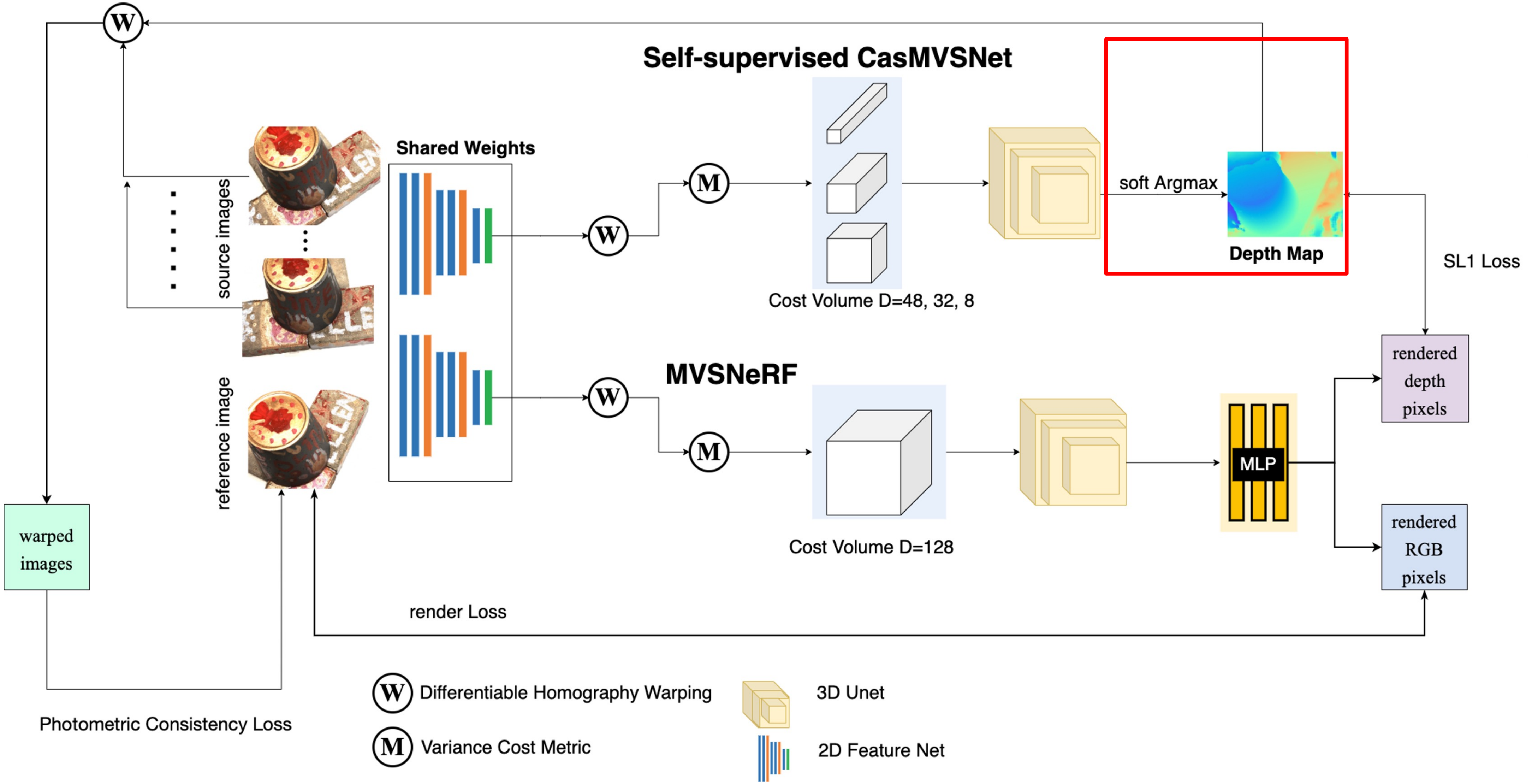
$$\mathbf{C} = \mathcal{M}(\mathbf{V}_1, \cdots, \mathbf{V}_N) = \frac{\sum_{i=1}^{N} (\mathbf{V}_i - \overline{\mathbf{V}_i})^2}{N}$$

Aggregate features from different views into one **cost volume**.

Input shape: B x 32 x 192 x H x W.
Output shape: B x 32 x 192 x H x W.



Pipeline: CasNeural-MVSNet



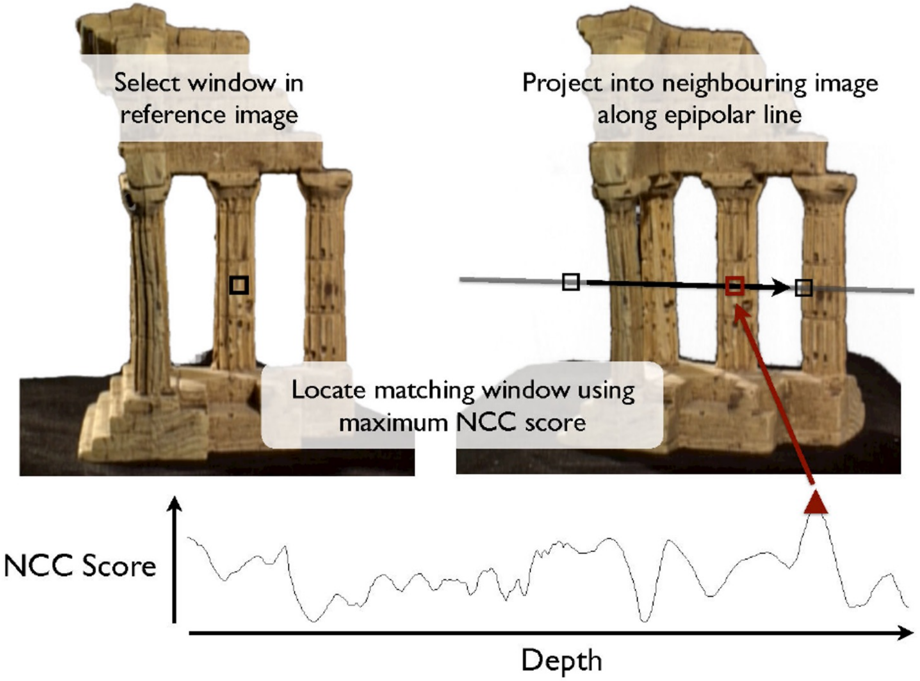
TUM VISUAL COMPUTING

Classification to regression

Classification: argmax along D dimension

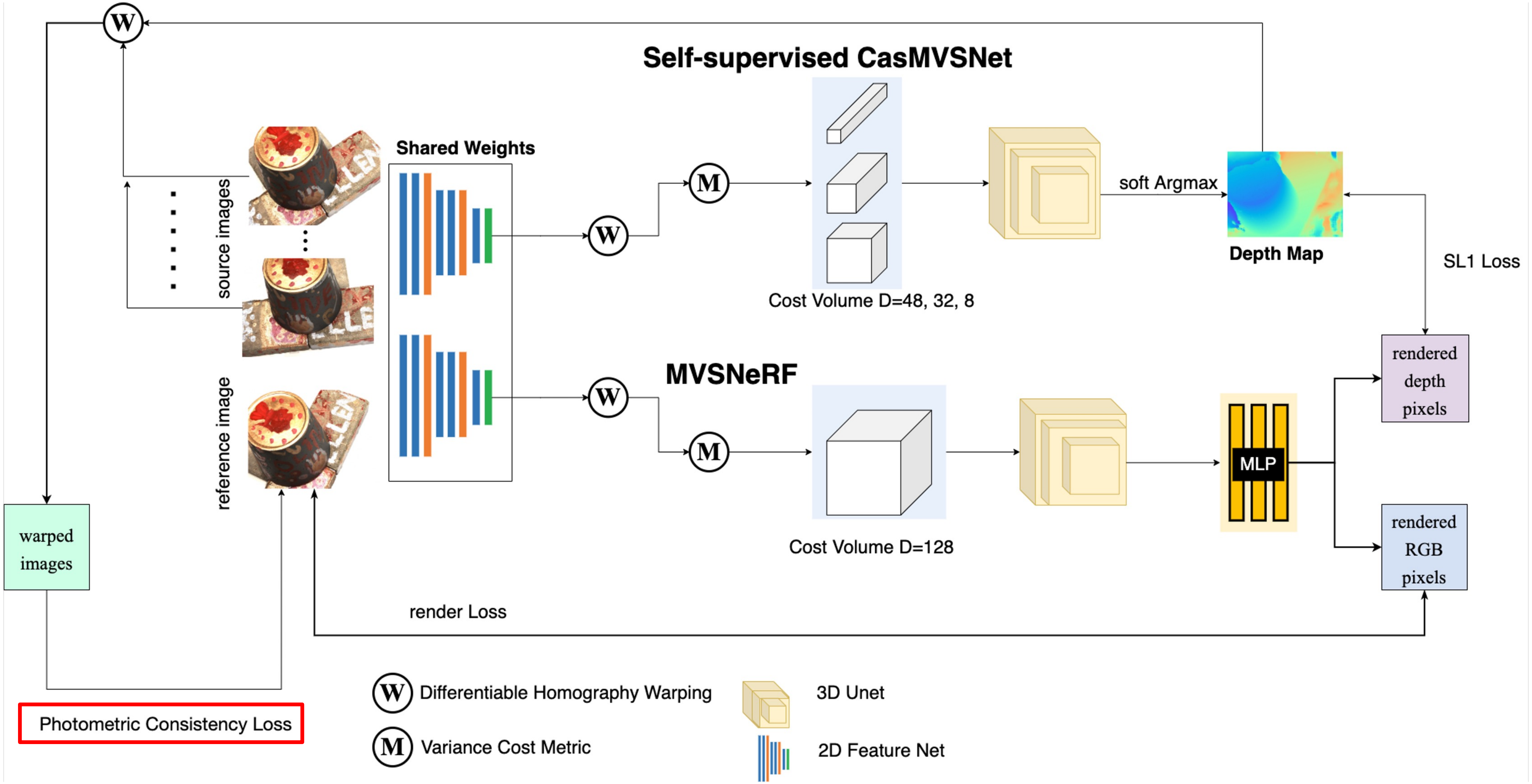
Regression: softmax along D dimension and calculate the **weighted sum** of depth values

$$\mathbf{D} = \sum_{d=d_{min}}^{d_{max}} d \times \mathbf{P}(d)$$

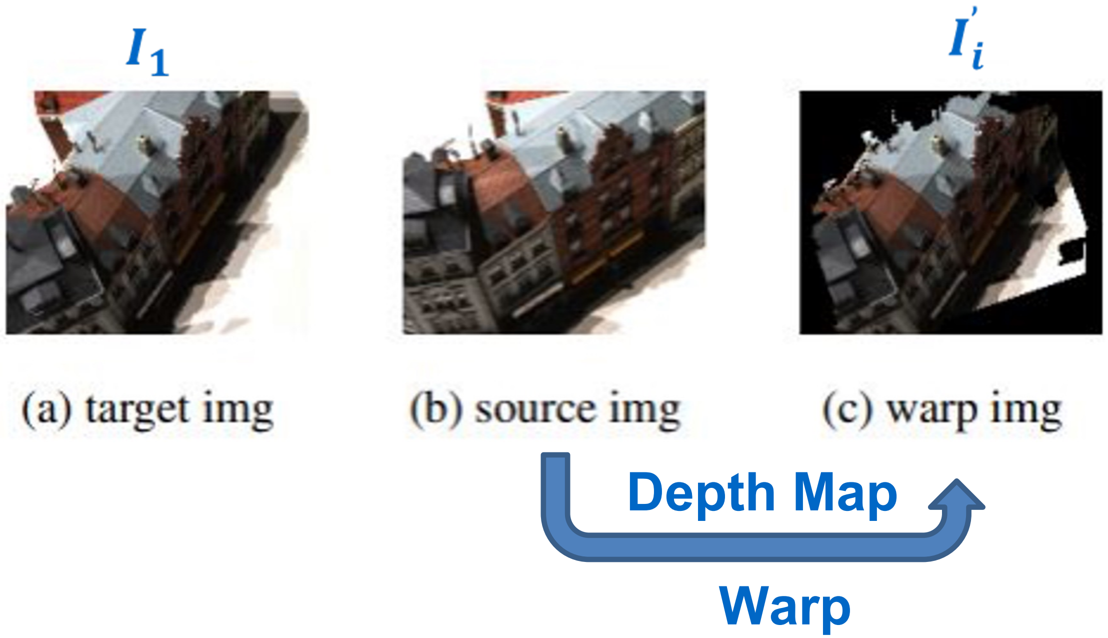


Input shape: B x1x 192 x H x W
Output shape: B x1x 1 x H x W

Pipeline: CasNeural-MVSNet



Loss: Photometric Consistency Loss



(a) target img (b) source img (c) warp img

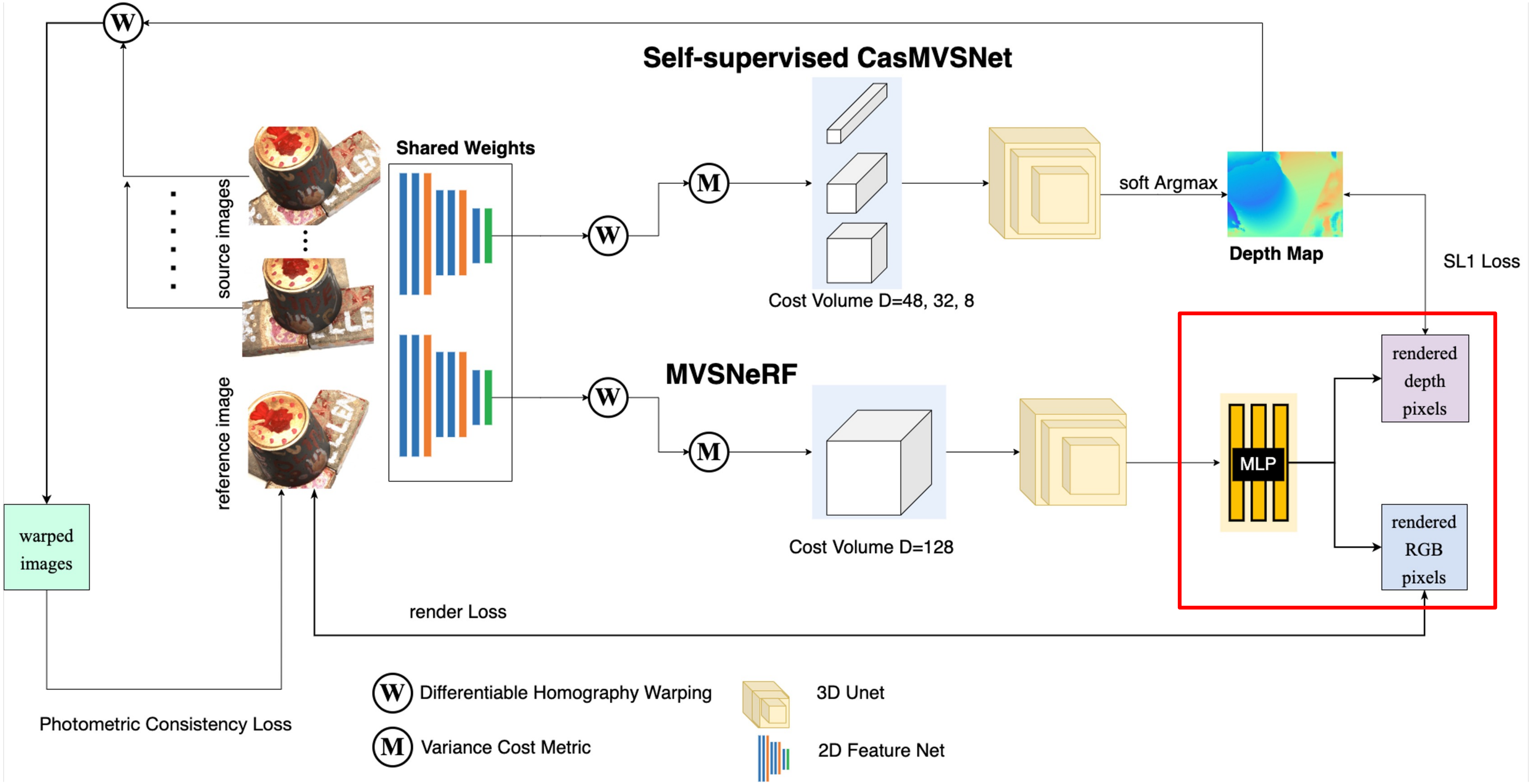
$$L_{PC} = \sum_{i=2}^{N} \frac{||(I_i' - I_1) \odot M_i||_2 + ||(\nabla I_i' - \nabla I_1) \odot M_i||_2}{||M_i||_1}$$

Total Loss

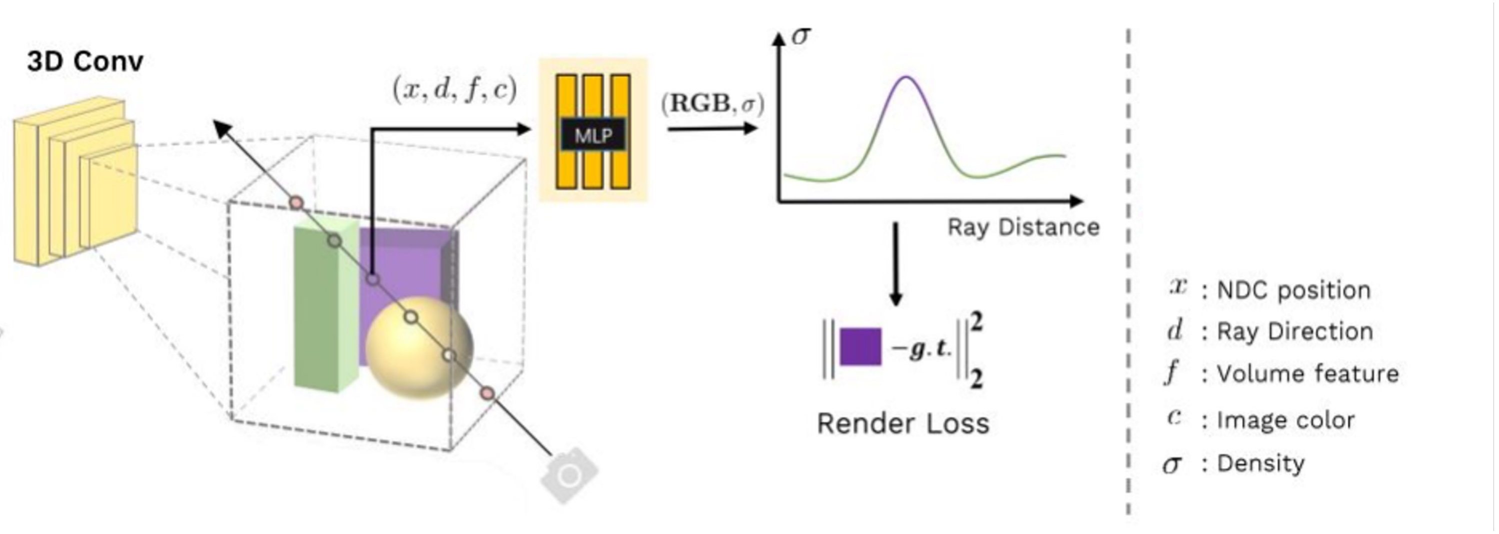
$$L = \sum \alpha L_{photo} + \beta L_{SSIM} + \gamma L_{Smooth}$$

$\alpha = 0.8$, $\beta = 0.2$ and $\gamma = 0.0067$

Pipeline: CasNeural-MVSNet



Neural Rendering



Depth

$$\hat{z}(\mathbf{r}) = \sum_{k=1}^{K} w_k t_k \,,$$

RGB

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{k=1}^{K} w_k \mathbf{c}_k \,, \tag{4}$$

$$\text{where} \quad w_k = T_k \left(1 - \exp(-\sigma_k \delta_k)\right) \,, \tag{5}$$

$$T_k = \exp\left(-\sum_{k'=1}^{k} \sigma_{k'} \delta_{k'}\right) \,, \tag{6}$$

$$\delta_k = t_{k+1} - t_k \,. \tag{7}$$

$$\mathcal{L}_{\theta_1} = \sum_{\mathbf{r} \in R} \left(\mathcal{L}_{\text{color}}(\mathbf{r}) + \lambda \mathcal{L}_{\text{depth}}(\mathbf{r}) \right)$$

Quantitative Results

- Point cloud evaluation results on DTU
  - The lower is better for Accuracy (Acc.), Completeness (Comp.), and Overall

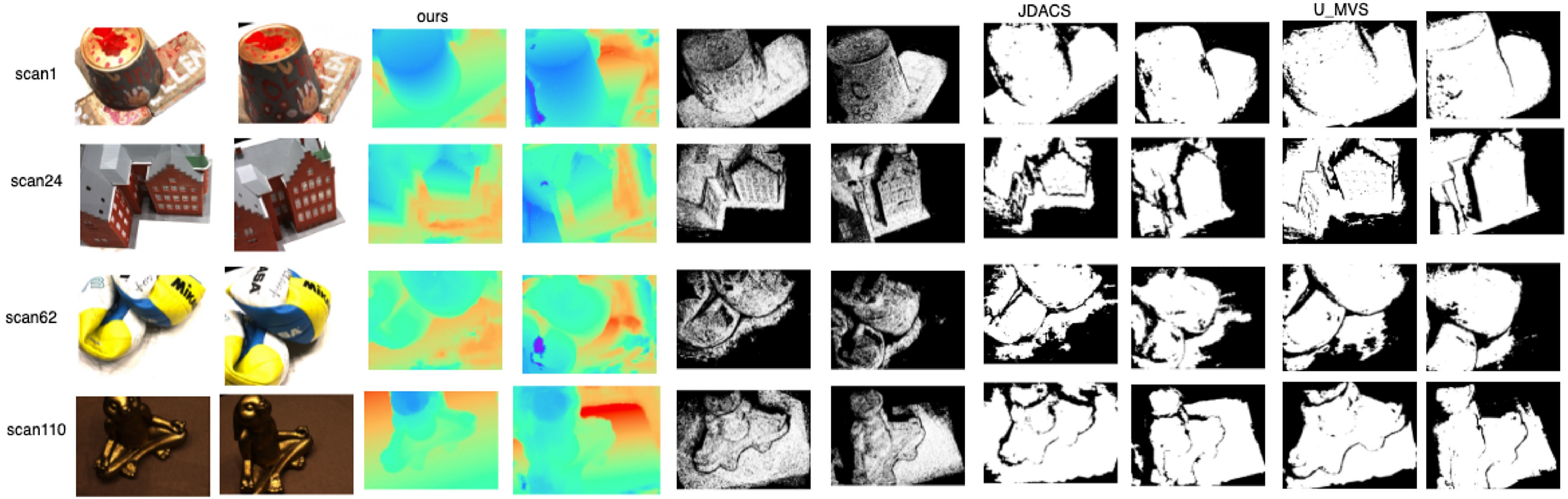
	Method	Acc.↓	Comp.↓	Overall.↓
Sup. and Geo.	Camp [3]	0.835	0.554	0.695
	Furu [8]	0.613	0.941	0.777
	Tola [22]	0.342	1.190	0.766
	Gipuma [9]	**0.283**	0.873	0.578
	SurfaceNet [12]	0.450	1.04	0.745
	MVSNet [29]	0.396	0.527	0.462
	R-MVSNet [30]	0.383	0.452	0.417
	CIDER [27]	0.417	0.437	0.427
	Point-MVSNet [5]	0.342	0.411	0.376
	GBi-Net [18]	***0.315***	**0.262**	**0.289**
Semi-Sup.	U-MVSNet [26]	0.354	0.3535	***0.3537***
UnSup.	Unsup_MVSNet [13]	0.881	1.073	0.977
	MVS2 [7]	0.76	0.515	0.637
	M3VSNet [11]	0.636	0.531	0.583
	JDACS [25]	0.398	0.318	0.358
	Ours	0.4209	***0.2927***	0.3568

Quantitative Results

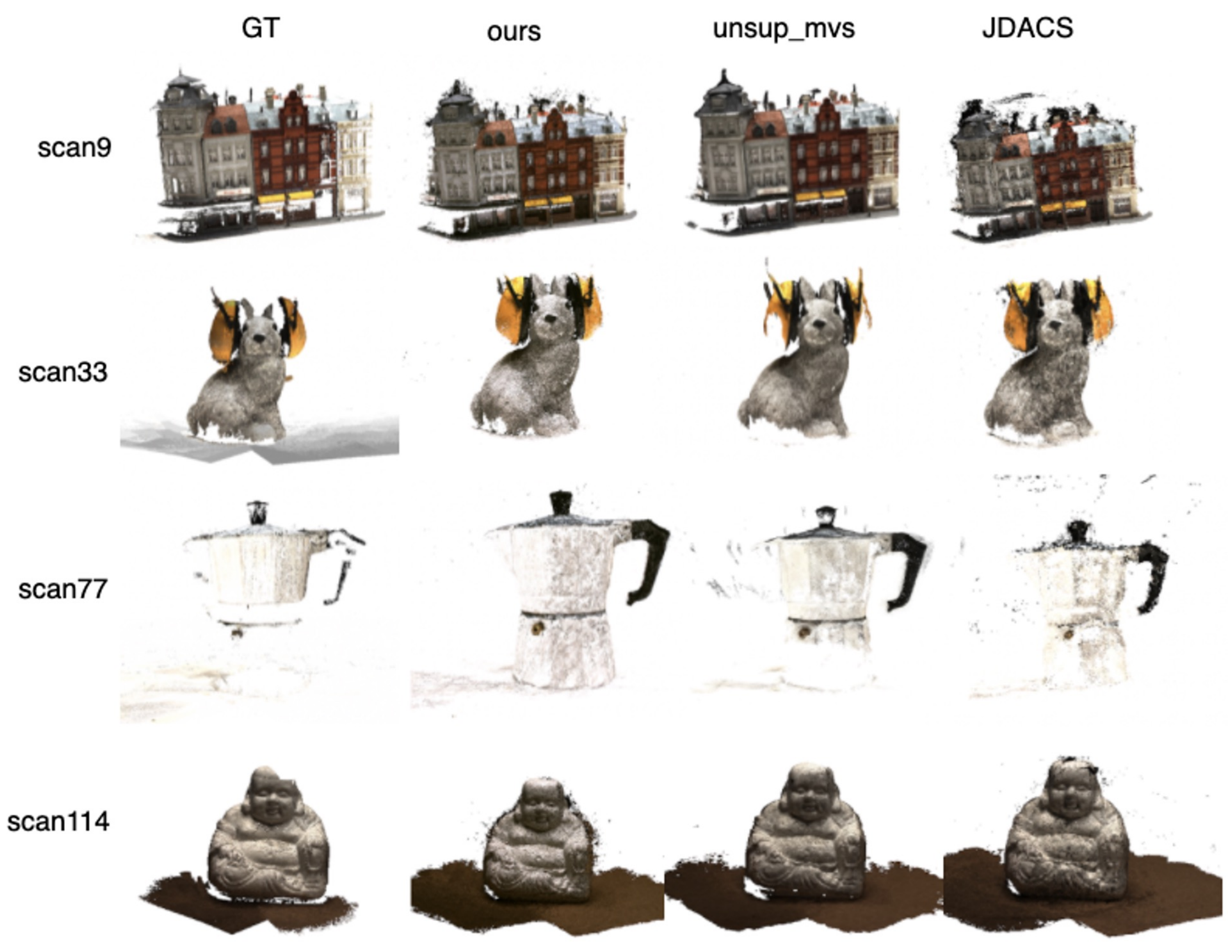
- Point cloud evaluation results on the Advanced and Intermediate subsets of Tanks and Temples dataset
  - Higher scores are better. The Mean is the average score of all scenes

	Advanced							Intermediate								
Method	Mean	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.	Mean	Fam.	Fra.	Hor.	Lig.	M60	Pan.	Pla.	Tra.
MVSNet [29]	-	-	-	-	-	-	-	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69
Point-MVSNet [5]	-	-	-	-	-	-	-	48.27	61.79	41.15	34.20	50.79	51.97	50.85	52.38	43.06
UCSNet [6]	-	-	-	-	-	-	-	54.83	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89
CasMVSNet [10]	31.12	19.81	38.46	29.10	43.87	27.36	28.11	56.42	76.36	58.45	46.20	55.53	56.11	54.02	58.17	46.56
PatchmatchNet [23]	32.31	23.69	37.73	30.04	41.80	28.31	32.29	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81
GBi-Net [18]	**37.32**	**29.77**	**42.12**	**36.30**	**47.69**	**31.11**	**36.93**	**61.42**	**79.77**	**67.69**	**51.81**	**61.25**	**60.37**	**55.87**	**60.67**	**53.89**
U-MVSNet [26]	30.97	22.79	35.39	28.90	36.70	28.77	33.25	57.15	76.49	60.04	49.20	55.52	55.33	51.22	56.77	52.63
MVS2 [7]	-	-	-	-	-	-	-	37.21	47.74	21.55	19.50	44.54	44.86	46.32	43.38	29.72
M3VSNet [11]	-	-	-	-	-	-	-	37.67	47.74	24.38	18.74	44.42	43.45	44.95	47.39	30.31
JDACS [25]	-	-	-	-	-	-	-	45.48	66.62	38.25	36.11	46.12	46.66	45.25	47.69	37.16
Ours	29.46	20.87	34.3	27.46	36.55	26.78	30.81	53.61	73.53	50.3	44.89	52.66	52.18	49.76	54.55	51

Qualitative Results



Qualitative Results



Qualitative Results

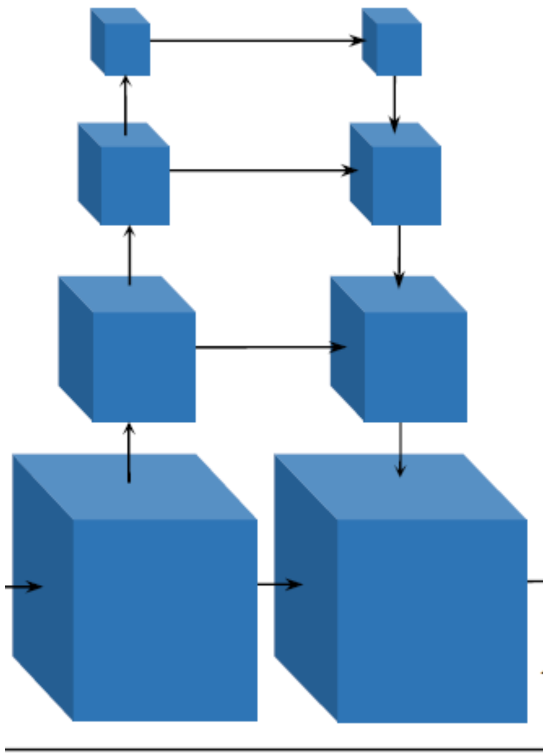
Discussion: Pros and Cons

Pros: Perfect Accuracy and Generalization Ability

Cons: Memory cost and test time efficiency

Possible Solution for future works:

- Introduce efficient design for 3D U-Net e.g. Binary Search
- Introduce coarse-to-fine structure for MLPs



Cost Volume Regularization

Zhenxing Mi* , Di Chang* and Dan Xu. Generalized Binary Search Network for Highly-Efficient Multi-View Stereo. Under review at CVPR 2022. https://arxiv.org/abs/2112.02338