MVSFormer++: Revealing the Devil in TransFormer's Details for Multi-View Stereo



Chenjie Cao, Xinlin Ren Yanwei Fu

Fudan University {cjcao20, xlren20,yanweifu}@fudan.edu.cn

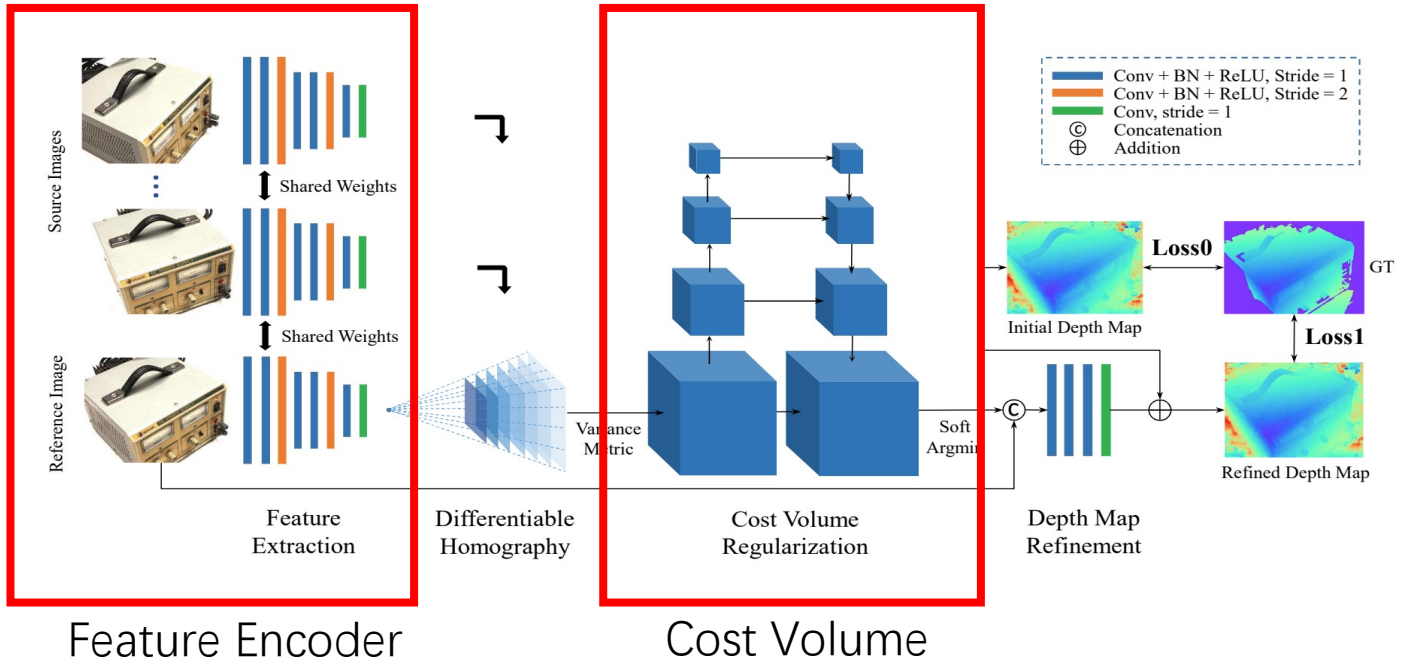
Project Page: https://github.com/maybeLx/MVSFormerPlusPlus

Motivation

Existing approaches have not thoroughly investigated the profound influence of **transformers** on different MVS modules.

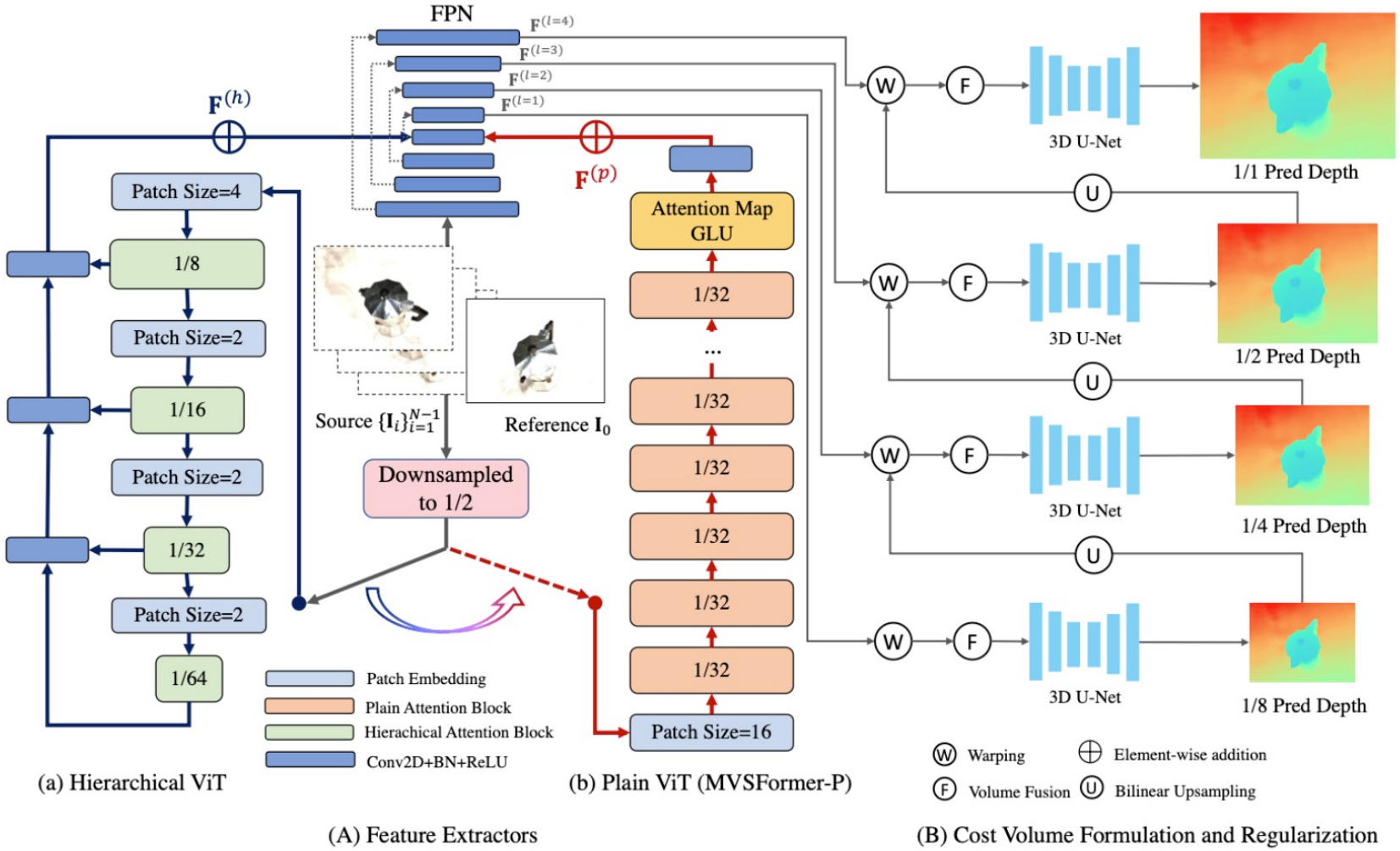
- Investigate tailored attention mechanisms for different MVS modules.
- Incorporating cross-view information into Pre-trained ViTs
- Enhancing Transformer's Length Extrapolation Capability in MVS

Mainstream Pipeline



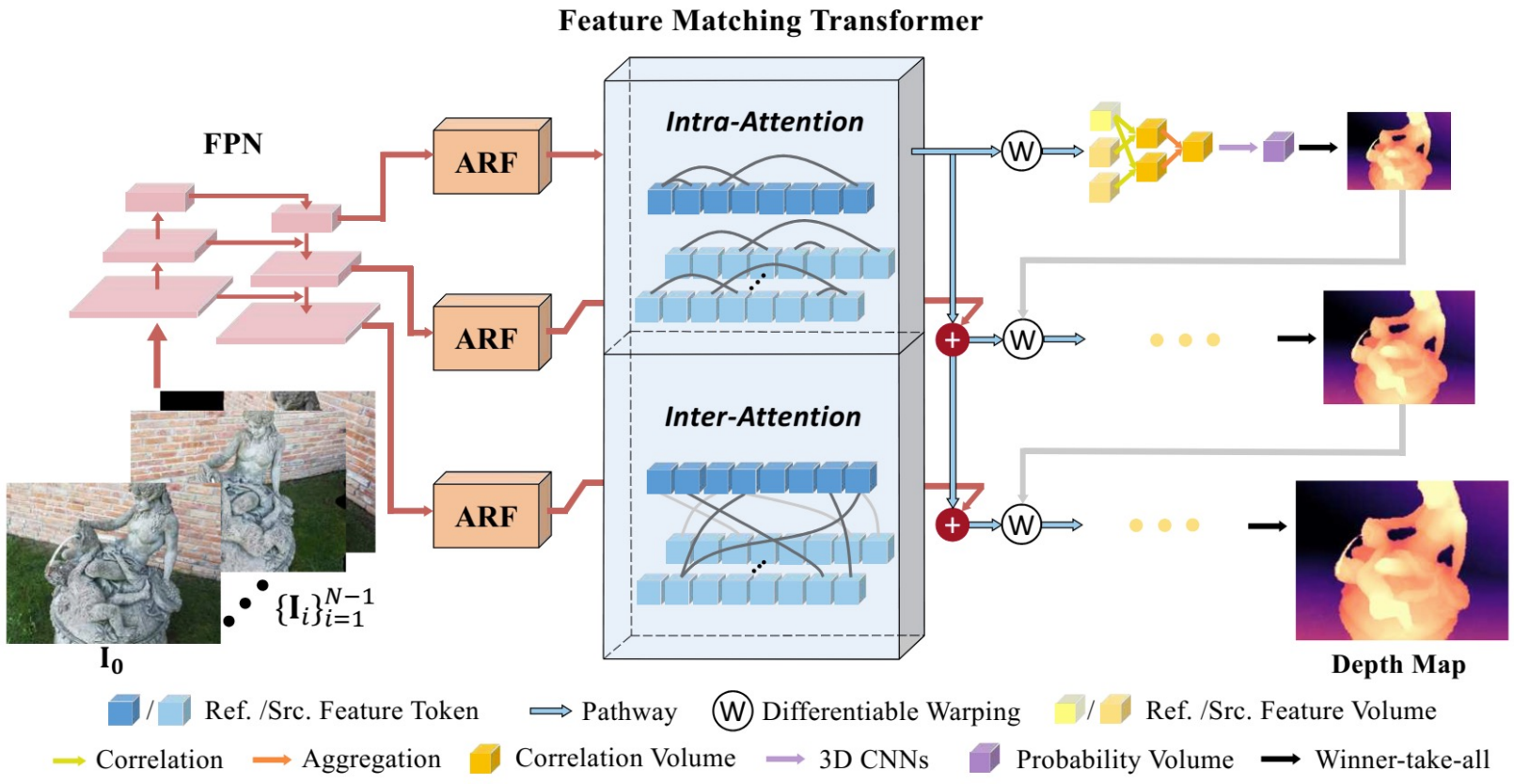
Related Work

Transformer in feature encoder, using pre-trained ViT prior for better feature representation



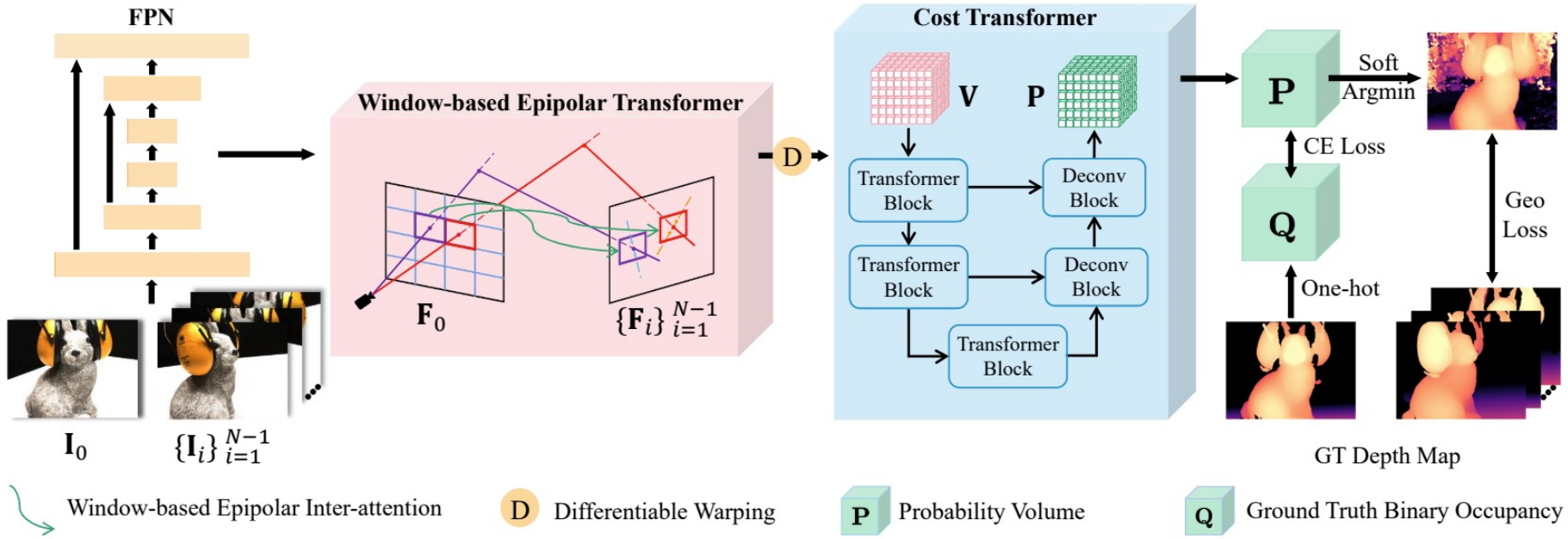
Cao C, Ren X, Fu Y. Mvsformer: Multi-view stereo by learning robust image features and temperature-based depth[J]. TMLR, 2023.

Transformer in feature encoder, adding cross view information to aggregate features



Ding Y, Yuan W, Zhu Q, et al. Transmvsnet: Global context-aware multi-view stereo network with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022

Transformer in cost volume



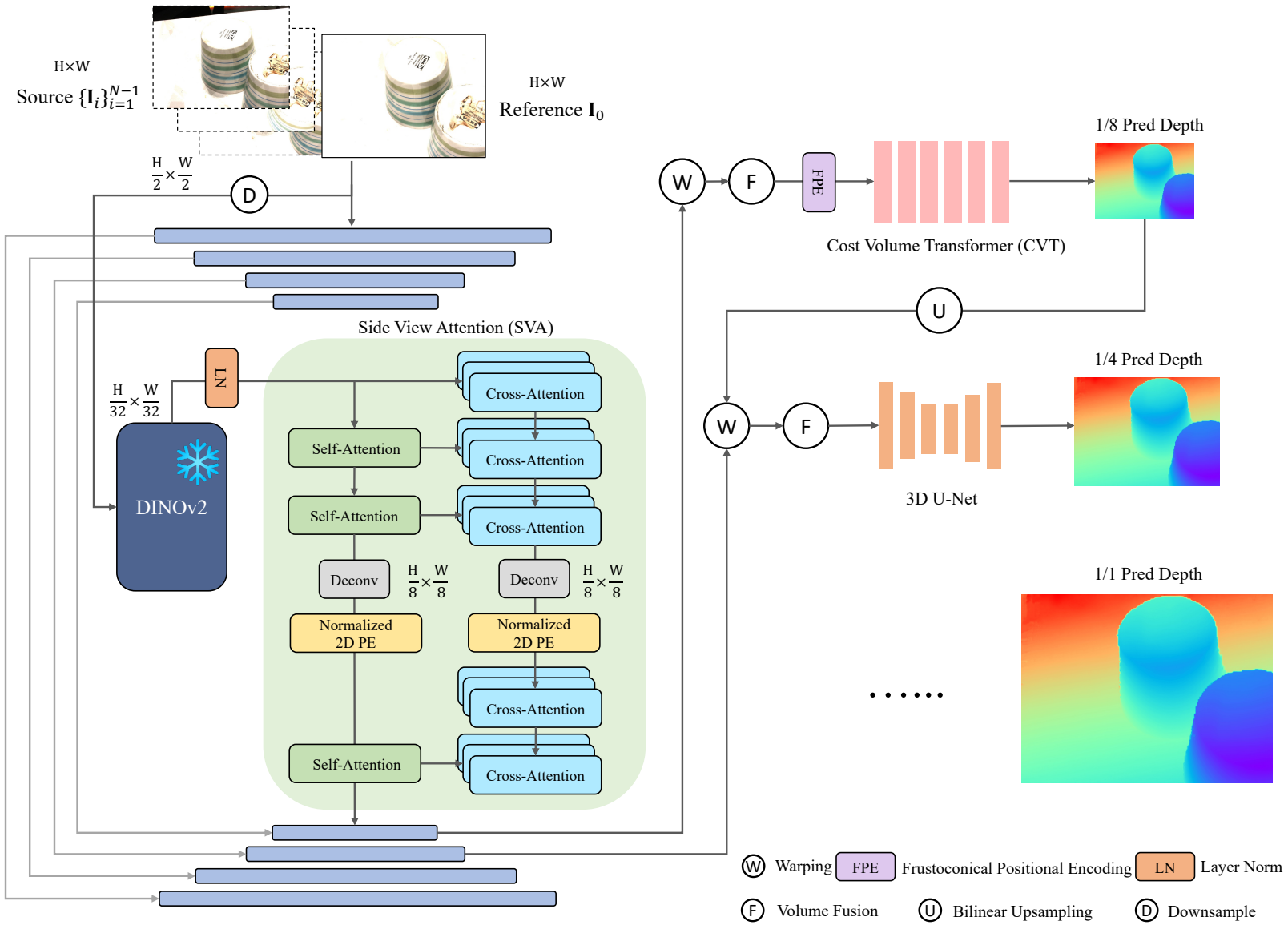
Liao J, Ding Y, Shavit Y, et al. Wt-mvsnet: window-based transformers for multi-view stereo[J]. Advances in Neural Information Processing Systems, 2022,

Comparsion between other transformer-based MVS methods

Table 1: Comparison of transformer-based MVS methods, including TransMVSNet (Ding et al., 2022), WT-MVSNet (Liao et al., 2022), CostFormer (Chen et al., 2023), and MVSFormer (Cao et al., 2022). MVSFormer++ surpasses other competitors with a meticulously designed transformer architecture, including attention with global receptive fields, transformer learning for both feature encoder and cost volume, cross-view attention, adaptive scaling for different sequence lengths, and specifically proposed positional encoding for MVS.

Methods	Attention	Transformers work in		Cross-view	Adaptive scaling	Positional Encoding (PE)		
	global/window	Feature encoder	Cost volume			Abs./Rel.	Normalized	3D-PE
TransMVSNet	global	✓	×	✓	×	absolute	×	×
WT-MVSNet	window	✓	✓	✓	×	relative	×	×
CostFormer	window	×	✓	×	×	relative	×	×
MVSFormer	global	✓	×	×	×	absolute	✓	×
MVSFormer++	global	✓	✓	✓	✓	absolute	✓	✓

Overview of MVSFormer++.



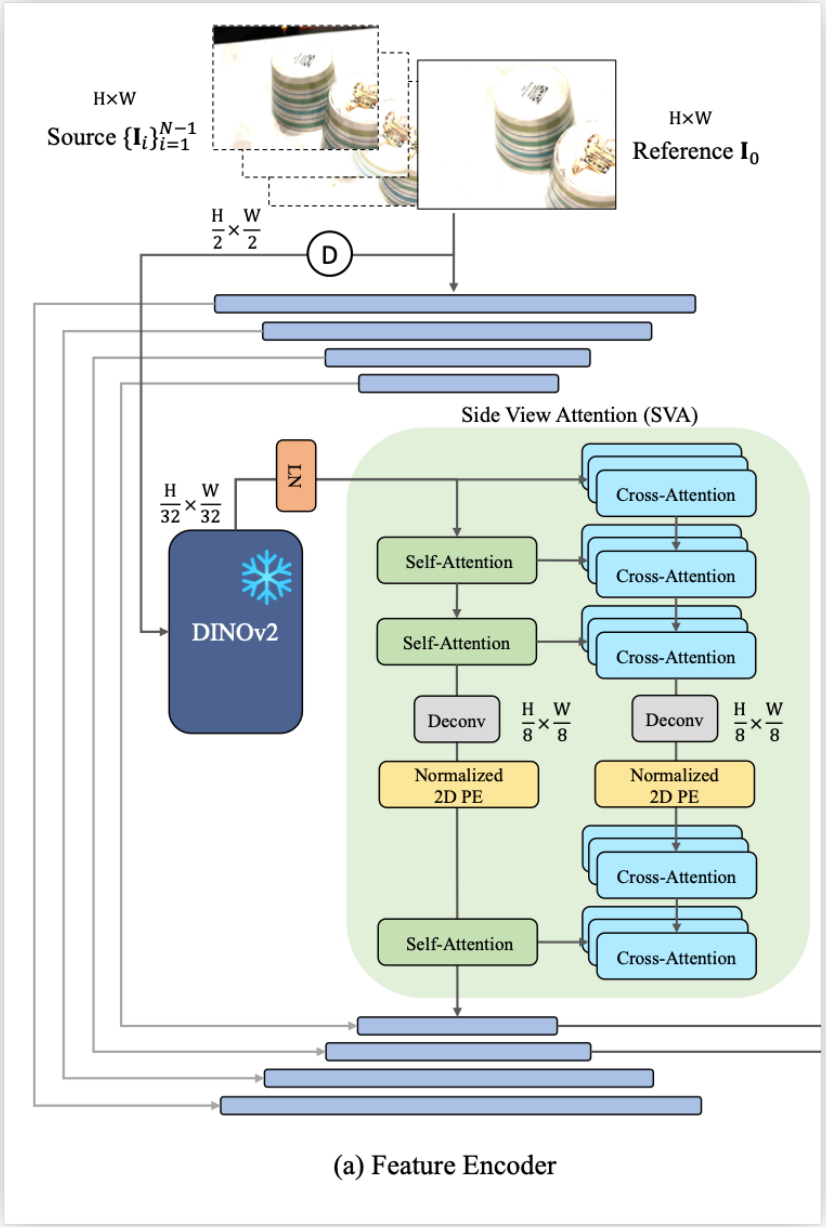
(a) Feature Encoder

(b) Cost Volume Formulation and Regularization

(A) Feature extraction enhanced with SVA module, normalized 2D-PE, and Norm&ALS.

(B) Multi-scale cost volume formation and regularization, where CVT is strengthed by FPE and AAS resulting in solid depth estimation.

Transformer for Feature Encoder

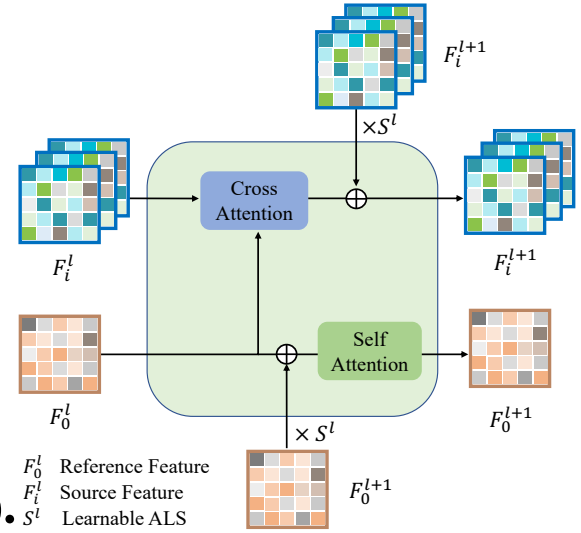


(a) Feature Encoder

- **Side View Attention (SVA)**

  1. Capture extensive global contextual information across different view

  2. Independently trained without any gradients passing from frozen DINOv2



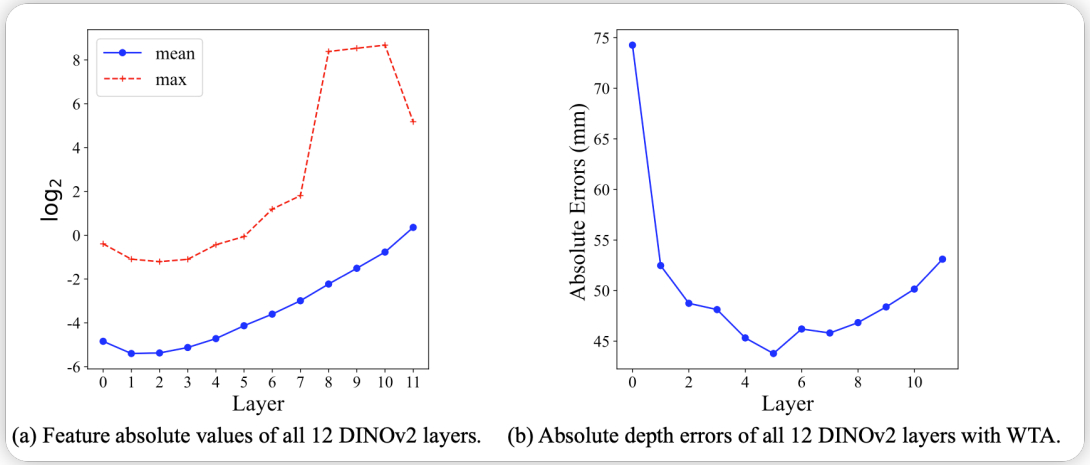
- **Normalized 2D Positional Encoding (PE).**

  Linear normalizing maximum values of height and width position to (128,128)

- **Normalization and Adaptive Layer Scaling (Norm&ALS)**

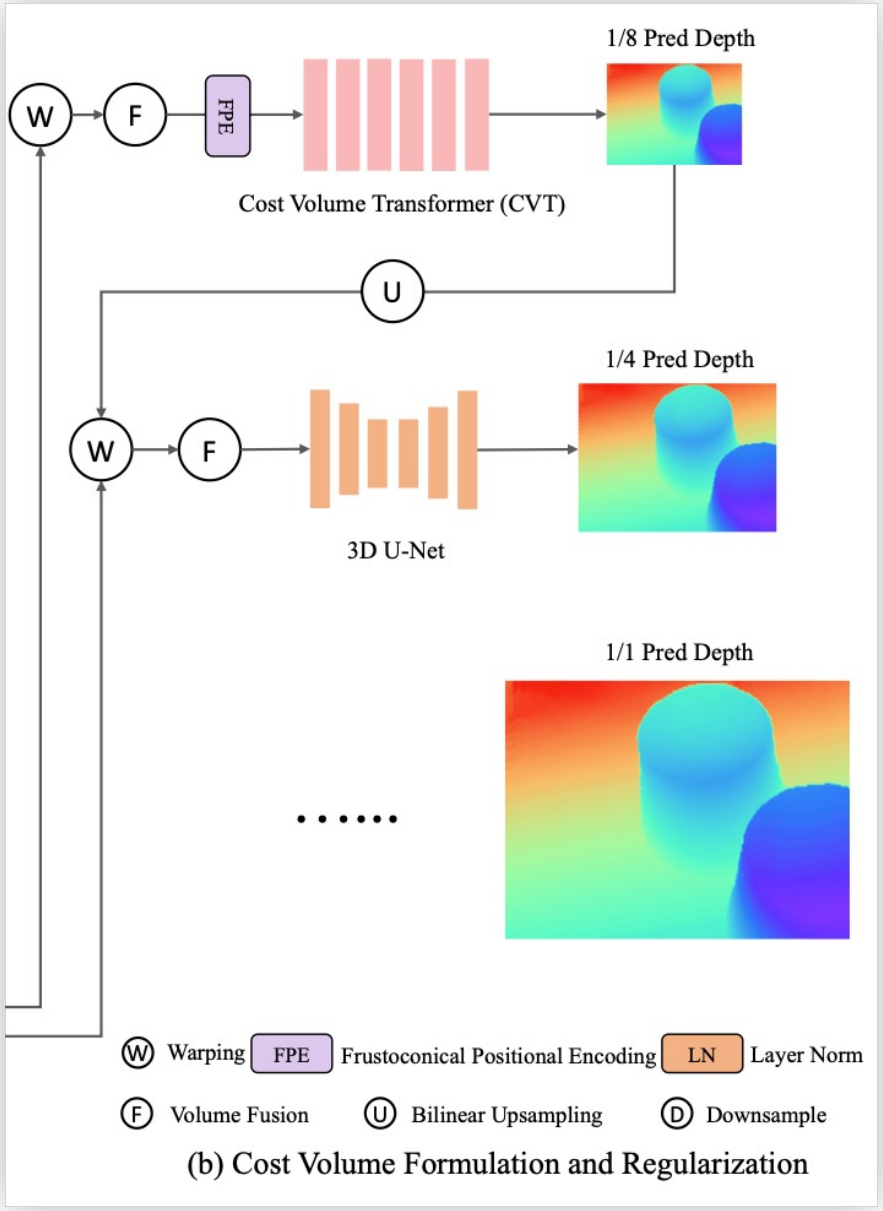
  1. Normalize all the DINOv2 features

  2. Adaptively adjust the significance of features from unstable frozen DINOv2 layers.



(a) Feature absolute values of all 12 DINOv2 layers. (b) Absolute depth errors of all 12 DINOv2 layers with WTA.

Transformer for Cost Volume Regularization



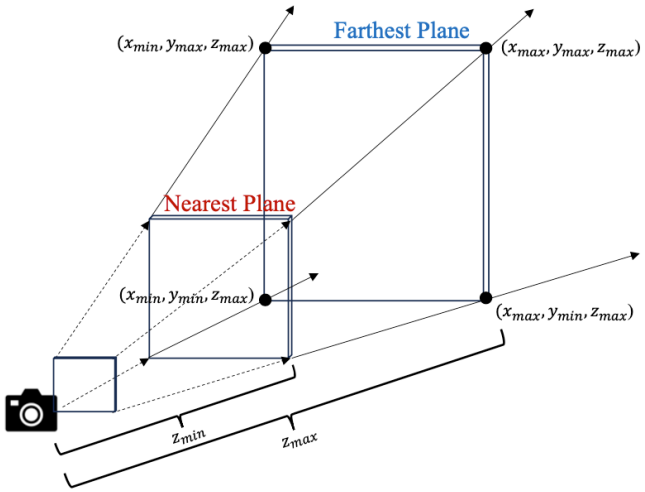
(b) Cost Volume Formulation and Regularization

- **Cost Volume Transformer (CVT)**

Pure transformer based on vanilla attention (FlashAttention). DHW can be seen as the global sequence learned by transformer blocks.

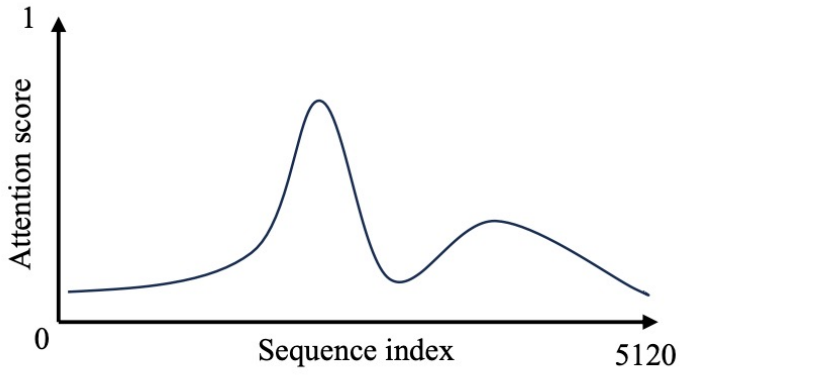
- **Frustoconical Positional Encoding (FPE)**
  - Normalize the 3D position P $\in \mathbb{R}^{3\times DHW}$ of the cost volume into the range $[0,1]^3$
  - Separately apply 1D sinusoidal PE along the x, y, z dimensions, then concatenate three PE into FPE ($3C \times DHW$) and project to $C \times DHW$

Crucial for improving CVT's depth estimation.

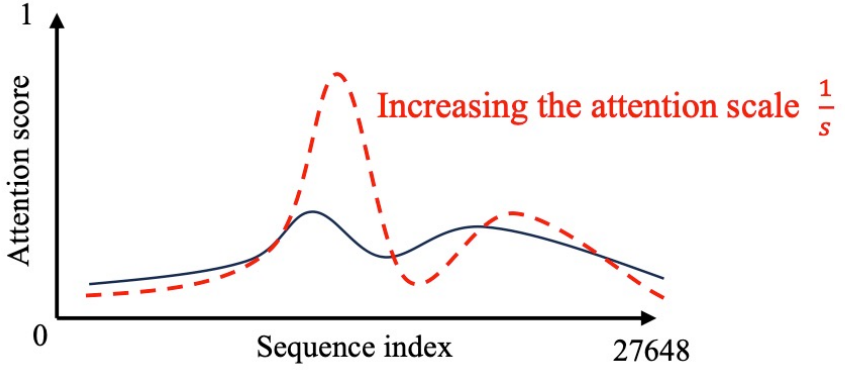


(a) 3D Frustoconical Positional Encoding (FPE)

Transformer for Cost Volume Regularization



$softmax(\frac{QK^T}{s})$ of normal attention sequence



$softmax(\frac{QK^T}{s})$ of long attention sequence

(b) Illustration of the attention dilution

The attention score would be diluted when the sequence increases, making it challenging to correctly focus on related target values.

- **Adaptive Attention Scaling (AAS)**

Insight: Sequential lengths of cost volume of training and testing image are largely different. 6k vs 30k

we should keep the invariant entropy for the attention score :

$$\mathcal{H}_i = -\sum_j^n a_{i,j} \log a_{i,j}, \qquad a_{i,j} = \frac{e^{\lambda q_i \cdot k_j}}{\sum_j^n e^{\lambda q_i \cdot k_j}}$$

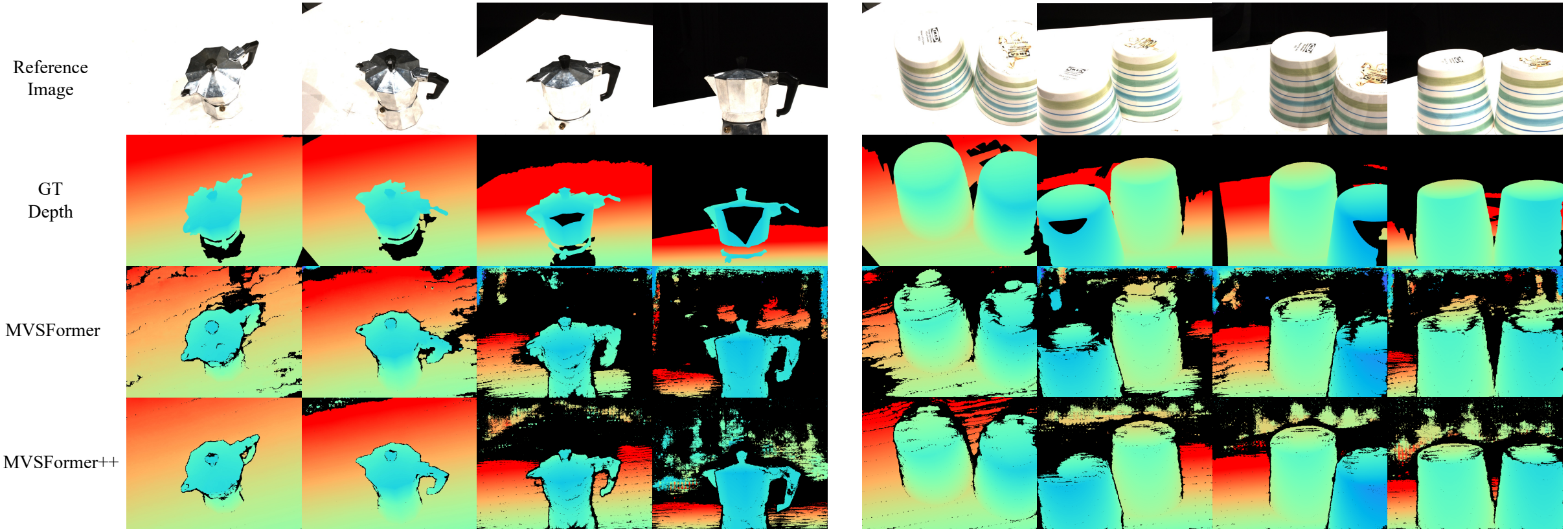
To make $\mathcal{H}_i$ independent of sequence length n , let $\lambda = \frac{\kappa \log n}{d}$

We formulate the attention as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\kappa \log n}{d}\mathbf{Q}\mathbf{K}^{\text{T}}\right)\mathbf{V}$$

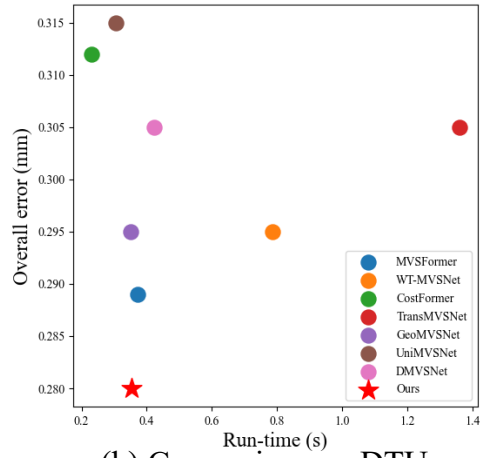
We empirically set $\kappa = \frac{\sqrt{d}}{\log \bar{n}}$, $\bar{n}$ is the mean sequential length during the multi-scale training.

MVSFormer++ vs MVSFormer

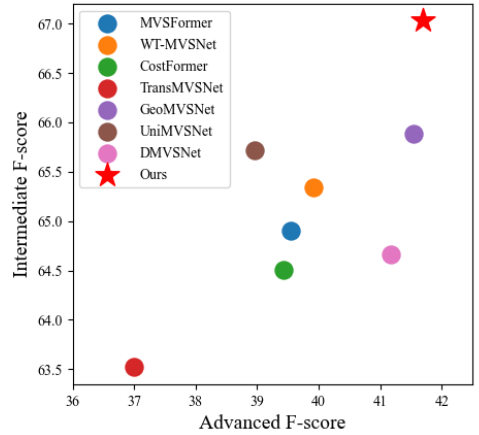


MVSFormer++ vs MVSFormer

(a) Point cloud results between MVSFormer and MVSFormer++ on DTU and Tanks-and-Temples.



(b) Comparison on DTU (Overall error↓)



(c) Comparison on Tanks-and-Temples (F-score↑)

Experiments

DTU dataset

Table 2: Quantitative point cloud results (mm) on DTU (lower is better). The best results are in bold, and the second ones are underlined. *All scenes share the same threshold for the post-processing.*

Methods	Accuracy↓	Completeness ↓	Overall↓
Gipuma (Galliani et al., 2015a)	**0.283**	0.873	0.578
COLMAP (Schönberger et al., 2016)	0.400	0.664	0.532
CasMVSNet (Gu et al., 2020)	0.325	0.385	0.355
AA-RMVSNet (Wei et al., 2021)	0.376	0.339	0.357
UniMVSNet (Peng et al., 2022)	0.352	0.278	0.315
TransMVSNet (Ding et al., 2022)	0.321	0.289	0.305
WT-MVSNet (Liao et al., 2022)	0.309	0.281	0.295
CostFormer (Chen et al., 2023)	<u>0.301</u>	0.322	0.312
RA-MVSNet (Zhang et al., 2023b)	0.326	0.268	0.297
GeoMVSNet (Zhang et al., 2023c)	0.331	0.259	0.295
MVSFormer (Cao et al., 2022)	0.327	**0.251**	<u>0.289</u>
MVSFormer++ (ours)	0.3090	<u>0.2521</u>	**0.2805**

Experiments

Tanks-and-Temples dataset

Table 3: Quantitative results of F-score on Tanks-and-Temples. A higher F-score means a better reconstruction quality. The best results are in bold, while the second ones are underlined.

Methods	Intermediate									Advanced						
	Mean	Fam.	Fra.	Hor.	Lig.	M60	Pan.	Pla.	Tra.	Mean	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
COLMAP (Schönberger et al., 2016)	42.14	50.41	22.25	26.63	56.43	44.83	46.97	48.53	42.04	27.24	16.02	25.23	34.70	41.51	18.05	27.94
CasMVSNet (Gu et al., 2020)	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51	31.12	19.81	38.46	29.10	43.87	27.36	28.11
CostFormer (Chen et al., 2023)	64.51	81.31	65.65	55.57	63.46	66.24	65.39	61.27	57.30	39.43	29.18	45.21	39.88	53.38	34.07	34.87
TransMVSNet (Ding et al., 2022)	63.52	80.92	65.83	56.94	62.54	63.06	60.00	60.20	58.67	37.00	24.84	44.59	34.77	46.49	34.69	36.62
WT-MVSNet (Liao et al., 2022)	65.34	81.87	67.33	57.76	64.77	65.68	64.61	62.35	58.38	39.91	29.20	44.48	39.55	53.49	34.57	38.15
RA-MVSNet (Zhang et al., 2023b)	65.72	82.44	66.61	58.40	64.78	**67.14**	65.60	**62.74**	58.08	39.93	29.17	46.05	40.23	53.22	34.62	36.30
D-MVSNet (Ye et al., 2023)	64.66	81.27	67.54	59.10	63.12	64.64	64.80	59.83	56.97	41.17	30.08	46.10	**40.65**	53.53	35.08	41.60
MVSFormer (Cao et al., 2022)	66.37	82.06	**69.34**	60.49	68.61	65.67	64.08	61.23	59.53	40.87	28.22	**46.75**	39.30	52.88	35.16	42.95
MVSFormer++ (ours)	**67.03**	**82.87**	68.90	**64.21**	**68.65**	65.00	**66.43**	60.07	**60.12**	**41.70**	**30.39**	45.85	39.35	**53.62**	**35.34**	**45.66**

Ablation Study

Attention in Cost Volume Regularization

Cost volume attention	$e_2 \downarrow$	$e_4 \downarrow$	$e_8 \downarrow$	Overall$\downarrow$
Shifted Window	15.03	9.93	6.90	0.2862
Linear	15.94	10.64	7.80	0.2980
Vanilla	13.89	8.91	6.34	0.2871
Vanilla + AAS	**13.76**	**8.71**	**6.17**	**0.2847**

- Linear attention suffers from terrible performance, primarily relying on group-wise feature dot product and variance, which lacks informative representations
- Capturing global contextual information in cost volume regularization is important. (Vanilla+AAS vs Shifted Window attention)

Attention in Feature Encoder

Feature encoder attention	$e_2 \downarrow$	$e_4 \downarrow$	$e_8 \downarrow$	Overall$\downarrow$
Shifted Window	12.80	8.05	5.64	0.2862
Top-K	13.04	8.43	6.12	0.2854
Vanilla	12.65	**7.88**	5.60	0.2835
Vanilla + AAS	**12.63**	**7.88**	5.60	0.2824
Linear	13.03	8.29	**5.35**	**0.2805**

- linear attention outperforms other attention mechanisms, which naturally robust for high-resolution images without attention dilution
- Top-K and shifted window-based attention lacking global receptive field fail to achieve proper results

Effect of Proposed Components based on other Baselines

Table 13: Quantitative ablation studies of CasMVSNet (Gu et al., 2020) and MVSFormer + DINOv1 (Caron et al., 2021) (MVSFormer-P) based on our proposed components including CVT and SVA. * indicates that CasMVSNet is re-trained with the 4-stage depth hypothesis setting (32-16-8-4) and cross-entropy loss as MVSFormer (Cao et al., 2022) and MVSFormer++.

Methods	$e_2 \downarrow$	$e_4 \downarrow$	$e_8 \downarrow$	Accuracy↓	Completeness↓	Overall↓
CasMVSNet	30.21	24.63	21.14	0.325	0.385	0.355
CasMVSNet*	23.15	18.68	15.35	0.353	0.286	0.320
CasMVSNet* + CVT	15.70	10.13	7.14	0.332	0.278	0.305
MVSFormer-P	17.18	11.96	8.53	0.327	0.265	0.296
MVSFormer-P + CVT	14.25	9.13	6.51	0.327	**0.261**	0.294
MVSFormer-P + CVT + SVA	**13.55**	**8.67**	**6.31**	**0.322**	0.254	**0.288**

CVT demonstrates substantial improvements for both CasMVSNet* and MVSFormer-P