JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

GeoMVSNet: Learning Multi-View Stereo with Geometry Perception

Zhe Zhang[1], Rui Peng[1], Yuxi Hu[2], Ronggang Wang[1*]

THU-PM-086

[1]Peking University

[2]CUHK, Shenzhen

Overview

Multi-View Stereo (MVS) aims to reconstruct the 3D model of objects or scenes from multiple overlapping images.
Autonomous Driving, Augmented Reality & Virtual Reality, Metaverse, Robotics, etc.

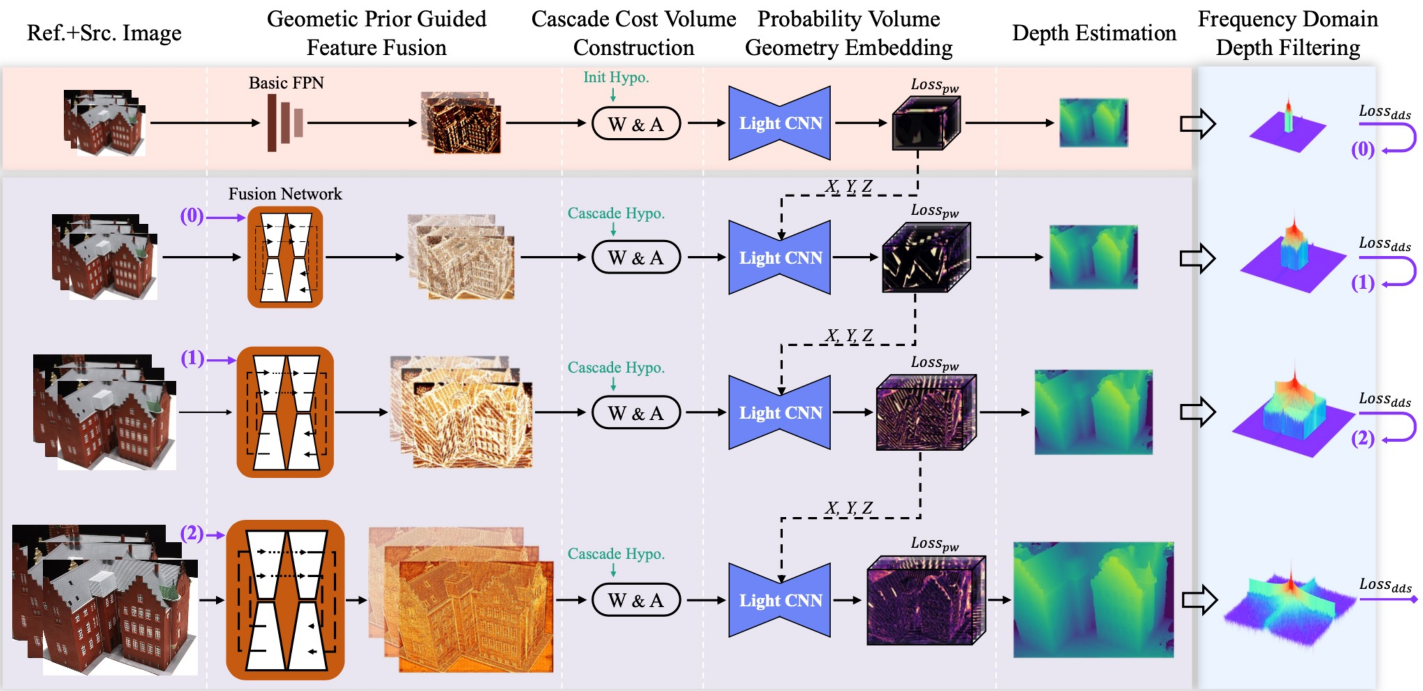


Figure 1. **Illustration of GeoMVSNet.** Structural features are extracted first by the geometry fusion network (Sec. 3.1) in finer stages, and W&A which denotes homography warping and aggregation is used to construct cascade cost volumes. The coarse probability volumes in coarse stages are embedded into the lightweight regularization network for geometry awareness (Sec. 3.1). And the frequency domain filtering equipped with curriculum learning strategy (Sec. 3.2) and depth distribution similarity loss (Sec. 3.4) based on Gaussian-Mixture Model (Sec. 3.3) are applied for full-scene geometry enhancement. The geometric prior output from the previous stage is used to guide the geometry perception for finer stages as shown by the numerical labels (0) ∼ (2).

① Geometric prior guided feature fusion
② Probability volume geometry embedding
③ Frequency domain filtering
+ Curriculum learning
④ Depth distribution similarity loss
+ Gaussian-Mixture Model

GeoMVSNet

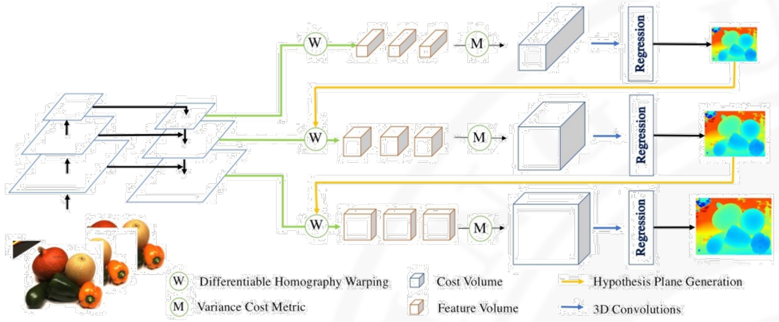
Table of Content

01 Motivation

Cascade-based architecture

- ✓ coarse-to-fine, reduce computational complexity
- ✗ valuable insight contained in early stage

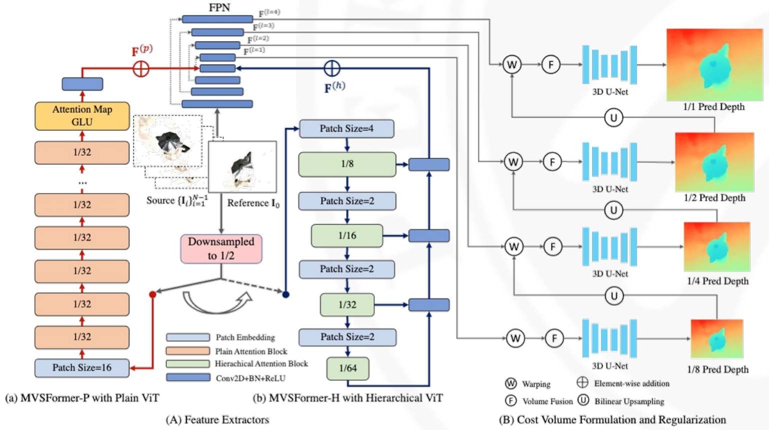
geometric information



CasMVSNet [1] (CVPR 2020)

Recent methods

- ✗ sophisticated external dependencies
- ✗ do not fully exploit geometric clues embedded in the scenarios



MVSFormer [2] (TMLR 2023)

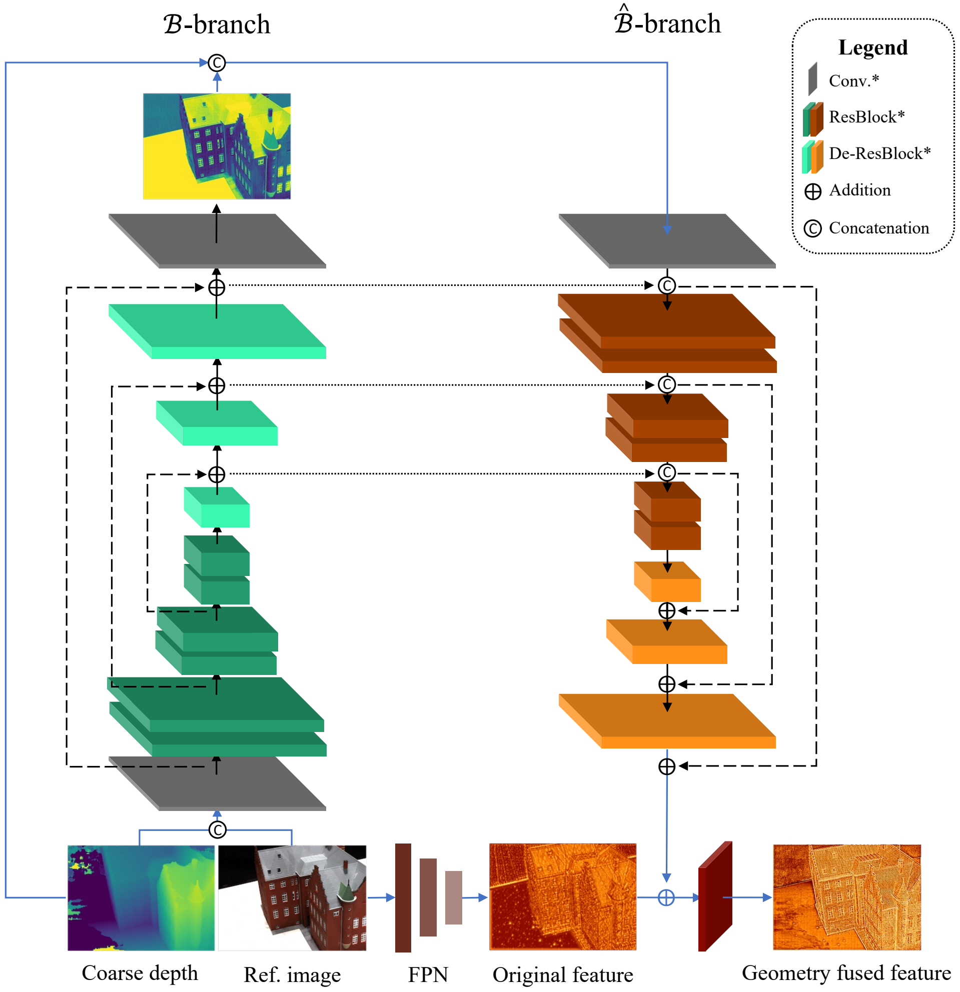
[1] Gu Xiaodong, et al. "Cascade cost volume for high-resolution multi-view stereo and stereo matching." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020).

[2] Cao Chenjie, et al. "Mvsformer: Learning robust image representations via transformers and temperature-based depth for multi-view stereo." Transactions of Machine Learning Research (2023).

02 Methods

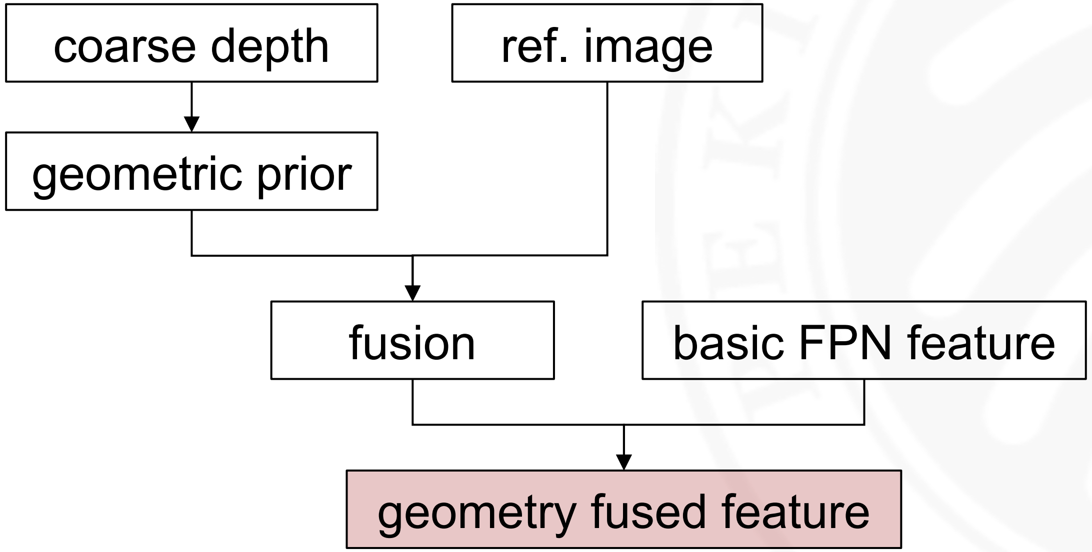
2.1 Geometry Awareness for Robust Cost Matching

Geometric prior guided feature fusion



$$Branch(z) = \hat{B}([D_{\uparrow}^{\ell}, B([I_0^{\ell+1}, D_{\uparrow}^{\ell}])])$$

$$F_0^{\ell+1}(z) = Fusion\{\bar{F}_0^{\ell+1}(z) \oplus Branch(z)\}$$

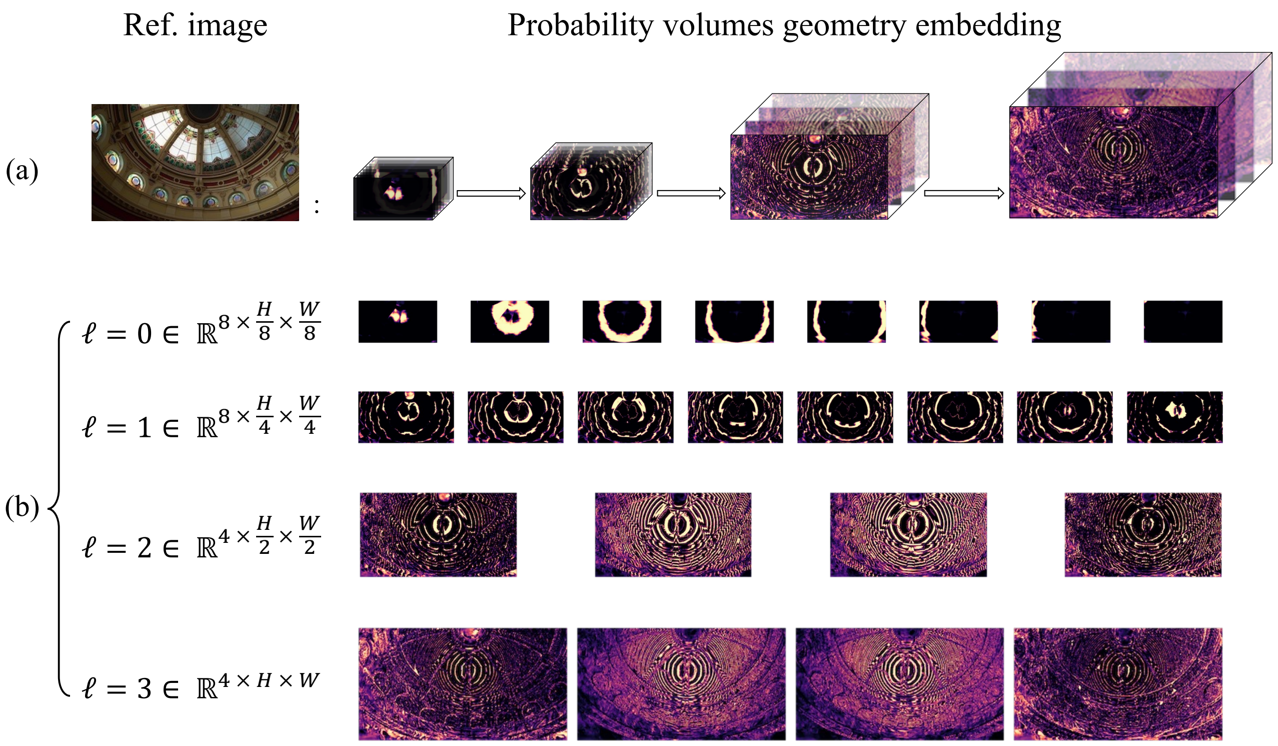


- ✓ strengthen the discrimination and structure of features
- ✓ solid foundation for robust aggregation

02 Methods

2.1 Geometry Awareness for Robust Cost Matching

Probability volume geometry embedding



P represent the possibility that the depth of a certain pixel attaches to a depth hypothesis (Plane Sweeping).

$$
\begin{cases}
X = \dfrac{(u - u_0)Z}{f_x} \\
Y = \dfrac{(v - v_0)Z}{f_y} \\
Z = Prob(\{m\} \leftarrow M)
\end{cases}
$$

3D ~~convolution~~ → 2D lightweight regularization network

+

3D "position maps"

- ✓ use depth-wise conv. instead of full 3D conv. to make the pipeline more efficient
- ✓ use geometric prior to compensate the reconstruction quality
- ✓ w/o external overload

--- from CVPR reviewers

02 Methods

2.2 Geometry Enhancement in Frequency Domain

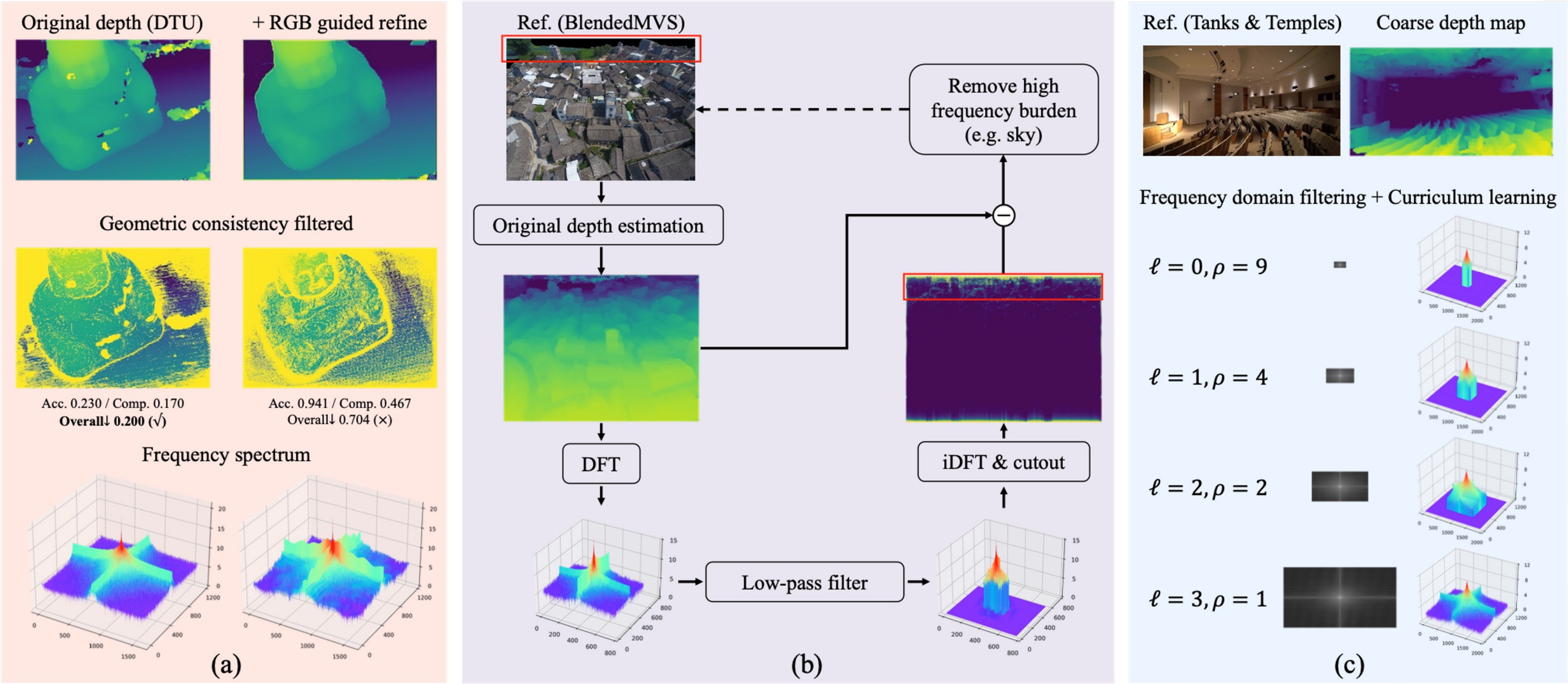
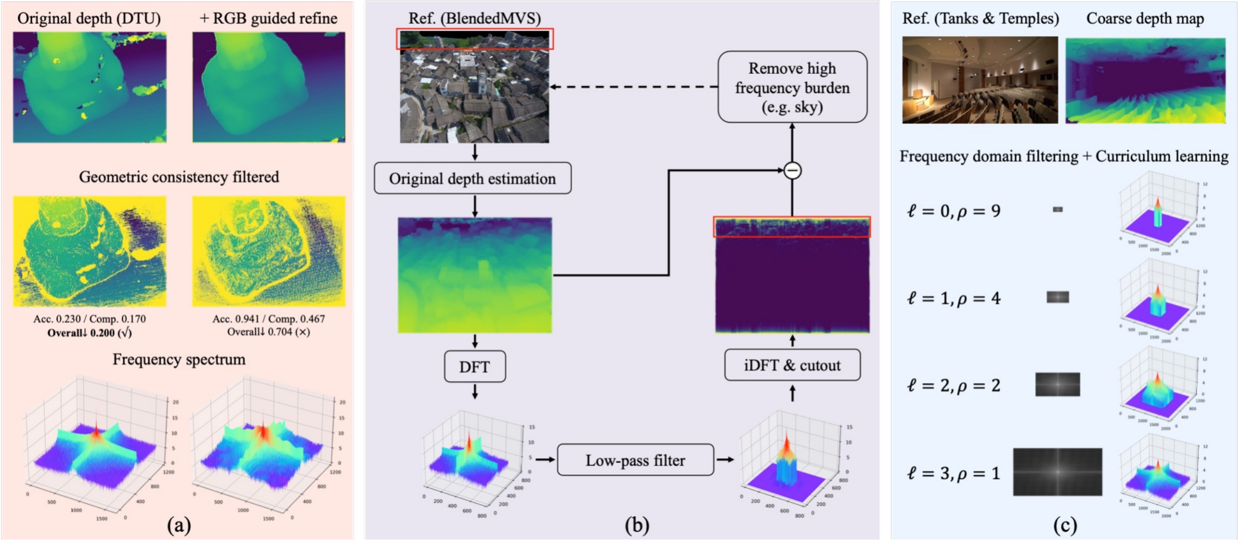


Figure 4. **Analysis of geometry enhancement in the frequency domain.** (a) The experiment results of the depth map refinement module on the DTU dataset [18]; (b) schematic flow chart of the frequency domain filtering on the BlendedMVS dataset [58]; (c) curriculum learning parameter configuration on the advanced set of Tanks and Temples dataset [19], coordinate spaces are unified for visualization.

02 Methods

2.2 Geometry Enhancement in Frequency Domain



✗ serve misestimations in coarse stages

e.g. infinite sky, areas near the edge of the image

✗ RGB-guided depth refinement

reduce the satisfaction of geometric consistency constraints

👓 a nearsighted person can still perceive a scene well without glasses, even if the texture details cannot be seen clearly

Frequency domain filtering

remove the complex and incomprehensible knowledge while avoiding producing more learning parameters

$$\mathcal{F}^{\ell}(u,v) = \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} D^{\ell}(x,y)\, e^{-j2\pi(\frac{ux}{W}+\frac{vy}{H})}$$

$$\tilde{D}^{\ell}(x,y) = \frac{1}{WH} \sum_{u=0}^{W-1} \sum_{v=0}^{H-1} \tilde{\mathcal{F}}^{\ell}(u,v)\, e^{j2\pi(\frac{ux}{W}+\frac{vy}{H})}$$

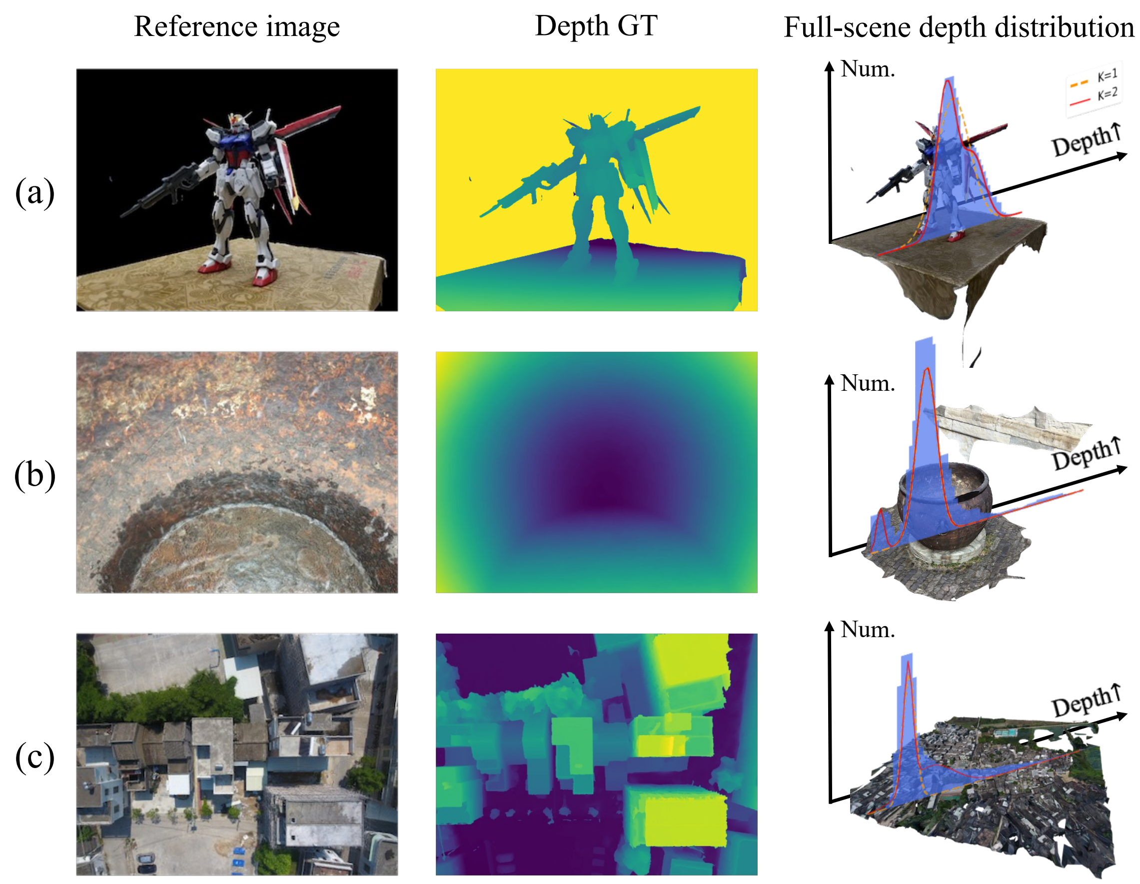
Curriculum learning strategy

embed coarse geometric priors into finer stages from easy to difficult

$$Q^{\ell}(d^{\ell}) \propto W^{\ell}(d^{\ell})\ \mathcal{N}(d^{\ell})\ ,\ d^{\ell} \in D^{\ell}$$

02 Methods

2.3 Mixed-Gaussian Depth Distribution



- uniform depth distribution
- inverse depth space
- multi-modal depth distribution

pixel-wise

$$Loss_{pw} = \sum_{z \in \Psi} (-P_{GT}(z)\ log[P(z)])$$

random variable depth value d

Gaussian-Mixture Model (GMM) $\mathcal{N}(d; \mu_i, \sigma_i^2)$

PauTa Criterion (3σ)

Full-scene depth distribution similarity loss

$$Loss_{dds} = \sum_{m=0, z \in \Upsilon}^{M'} \tilde{p}(z)\ (log\ \tilde{p}(z) - log\ \mathcal{N}_{GT}(z))$$

$$\Upsilon = \Psi\ \cap \bigcup_{i=1}^{K} \{(\mu_i - 3\sigma_i, \mu_i + 3\sigma_i)\}$$

03 Experiments
3.1 Benchmark Performance

DTU Dataset

GeoMVSNet estimates accurate depths and complete point clouds, especially for the geometry structures of the subject, and high-frequency clutter textures are well suppressed.

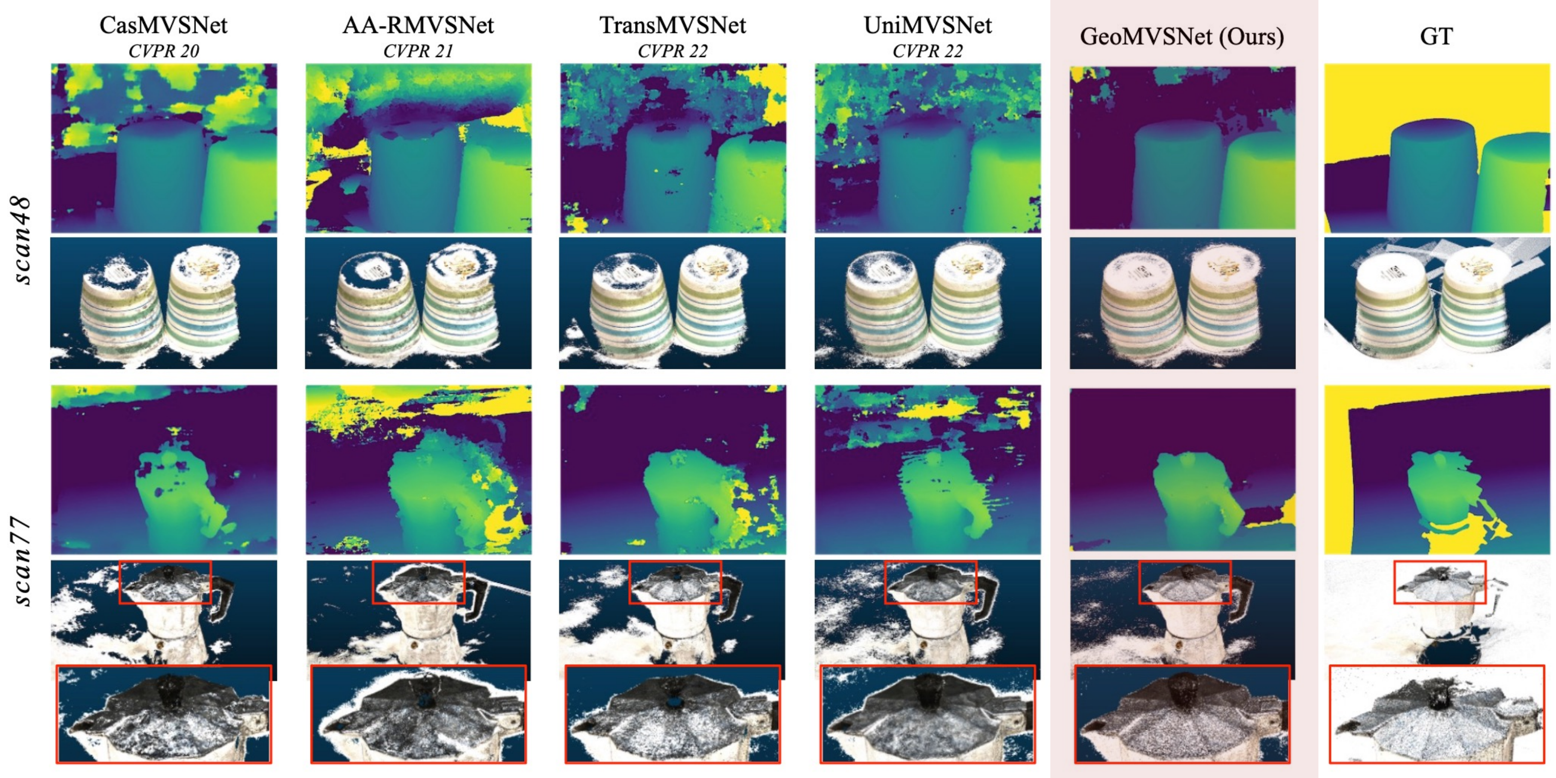


Figure 6. **Qualitative comparison of the most challenging *scan48* and *scan77* on the DTU evaluation dataset.** The first and third rows are estimated depth maps while others are point cloud reconstruction results. Our model produces remarkable accuracy and completeness.

	Method	Acc. (*mm*)	Comp. (*mm*)	Overall↓ (*mm*)
	Gipuma [12]	**0.283**	0.873	0.578
	COLMAP [36]	0.400	0.664	0.532
	R-MVSNet [57]	0.383	0.452	0.417
	CasMVSNet [14]	0.325	0.385	0.355
	CVP-MVSNet [54]	0.296	0.406	0.351
Post-pyramid Era	EPP-MVSNet [27]	0.413	0.296	0.355
	CER-MVS [28]	0.359	0.305	0.332
	RayMVSNet [48]	0.341	0.319	0.330
	Effi-MVSNet [45]	0.321	0.313	0.317
	CDS-MVSNet [13]	0.352	0.280	0.316
	NP-CVP-MVSNet [53]	0.356	0.275	0.315
	UniMVSNet [32]	0.352	0.278	0.315
	TransMVSNet [8]	0.321	0.289	0.305
	GBi-Net* [29]	0.312	0.293	0.303
	MVSTER* [46]	0.340	0.266	0.303
	GeoMVSNet (Ours)	0.331	**0.259**	**0.295**

03 Experiments
3.1 Benchmark Performance

Tanks and Temples Dataset
GeoMVSNet ranks 1st on the T&T-Advanced set (Oct. 2022 - PRESENT).

Method	Intermediate									Advanced						
	Mean↑	Family	Francis	Horse	L.H.	M60	Panther	P.G.	Train	Mean↑	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
COLMAP [36]	42.14	50.41	22.25	25.63	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 [14]	56.42	76.36	58.45	46.20	55.53	56.11	54.02	58.17	46.56	31.12	19.81	38.46	29.10	43.87	27.36	28.11
PatchmatchNet [44]	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81	32.31	23.69	37.73	30.04	41.80	28.31	32.29
CER-MVS [28]	64.82	81.16	64.21	50.43	**70.73**	63.85	63.99	**65.90**	58.25	40.19	25.95	45.75	39.65	51.75	35.08	42.97
Effi-MVSNet [45]	56.88	72.21	51.02	51.78	58.63	58.71	56.21	57.07	49.38	34.39	20.22	42.39	33.73	45.08	29.81	35.09
UniMVSNet [32]	64.36	81.20	66.43	53.11	63.46	**66.09**	64.84	62.23	57.53	38.96	28.33	44.36	39.74	52.89	33.80	34.63
TransMVSNet [8]	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
GBi-Net [29]	61.42	79.77	**67.69**	51.81	61.25	60.37	55.87	60.67	53.89	37.32	29.77	42.12	36.30	47.69	31.11	36.93
MVSTER [46]	60.92	80.21	63.51	52.30	61.38	61.47	58.16	58.98	51.38	37.53	26.68	42.14	35.65	49.37	32.16	39.19
GeoMVSNet (Ours)	**65.89**	**81.64**	67.53	55.78	68.02	65.49	**67.19**	63.27	58.22	**41.52**	**30.23**	**46.53**	**39.98**	**53.05**	**35.98**	**43.34**

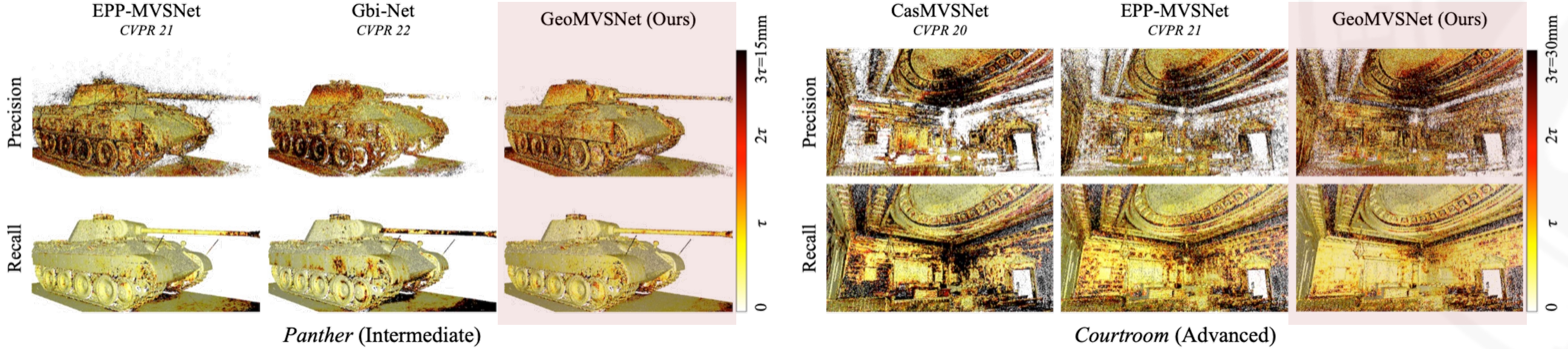


Figure 7. **Point clouds error comparison of state-of-the-art methods on the Tanks and Temples benchmark.** τ is the scene-relevant distance threshold determined officially and darker means larger error. The first row shows *Precision* and the second row shows *Recall*.

03 Experiments
3.2 Ablation Study

Table 3. **Ablation results on the DTU evaluation dataset.**

Method	Sec. 3.1 GFN	Sec. 3.1 PVE	Sec. 3.2 FDF	Sec. 3.2 CL	Sec. 3.4 $Loss_{pw}$	Sec. 3.4 $Loss_{dds}$	Acc.	Comp.	Overall↓
baseline (L=4, N=5)					✓		0.3629	0.3016	0.3323
+ geometry fusion network	✓				✓		0.3520	0.2893	0.3207
+ prob. volume embedding		✓			✓		0.3705	0.3053	0.3379
+ fusion & embedding	✓	✓			✓		0.3404	0.2922	0.3163
+ frequency domain filtering	✓		✓		✓		0.3663	0.2707	0.3185
+ curriculum learning	✓		✓	✓	✓		0.3650	0.2634	0.3142
+ distribution similarity loss	✓	✓			✓	✓	0.3346	0.2832	0.3089
proposed	✓	✓	✓	✓	✓	✓	**0.3309**	**0.2593**	**0.2951**

Table 4. **Ablation results of feature fusion on the DTU dataset.**

Method	Acc. (*mm*)	Comp.(*mm*)	Overall↓ (*mm*)
a) original feat.	0.3629	0.3016	0.3323
b) branch feat.	0.3577	0.3321	0.3449
c) original + branch	0.3520	0.2893	**0.3207**

Table 5. **Ablation results of geometry embedding on the intermediate set of the Tanks and Temples dataset.**

Method	Mean↑	Family	Francis	Horse	L.H.	M60	Panther	P.G.	Train
1) w/o embedding	62.12	80.96	**65.53**	46.91	63.87	61.78	60.90	60.49	**56.53**
2) X + Y	61.56	79.99	65.41	43.69	64.63	62.27	60.05	61.48	54.92
3) Z	62.89	80.27	64.61	51.67	64.29	**63.32**	**61.55**	61.42	55.98
4) X + Y + Z	**63.52**	**81.17**	65.48	**53.46**	**65.62**	62.85	61.26	**62.15**	56.14

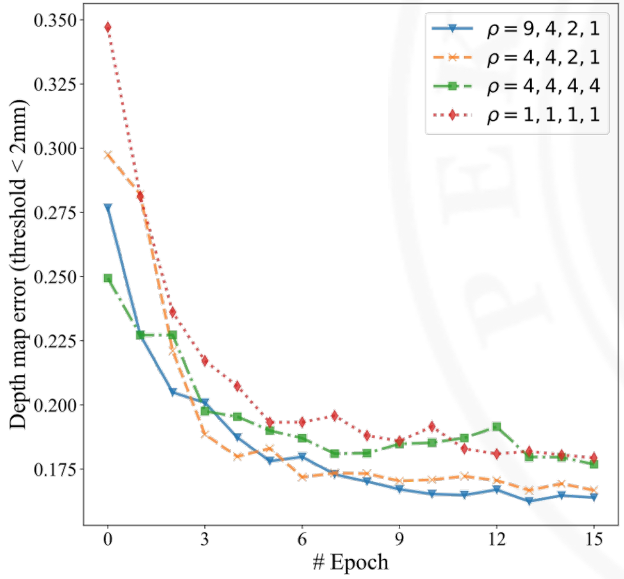


Figure 14. **Visualization of the evaluation depth map error (threshold $<$ 2mm) of the training process on the DTU dataset.**

04 Future Work

Limitation

- still increase the complexity of the cascade-based architecture (fusion network & prob. embedding)
- geometric clues for reference view only

Future Work

- explicitly modeled geometry extensions for un/self-supervised MVS frameworks
- skip (or replace) the complex intermediate cost volume
- multi-plane image/depth representation
- integrate with Nerf/Neus…

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thanks for Listening!

GeoMVSNet: Learning Multi-View Stereo with Geometry Perception

Zhe Zhang[1], Rui Peng[1], Yuxi Hu[2], Ronggang Wang[1*]

THU-PM-086

[1]Peking University

[2]CUHK, Shenzhen