RayMVSNet++: Learning Ray-based 1D Implicit Fields for Accurate Multi-View Stereo
(2023 TPAMI 2022 CVPR)

2024.05.30

Contribution

1. A novel formulation of deep MVS as learning ray-based 1D implicit fields.

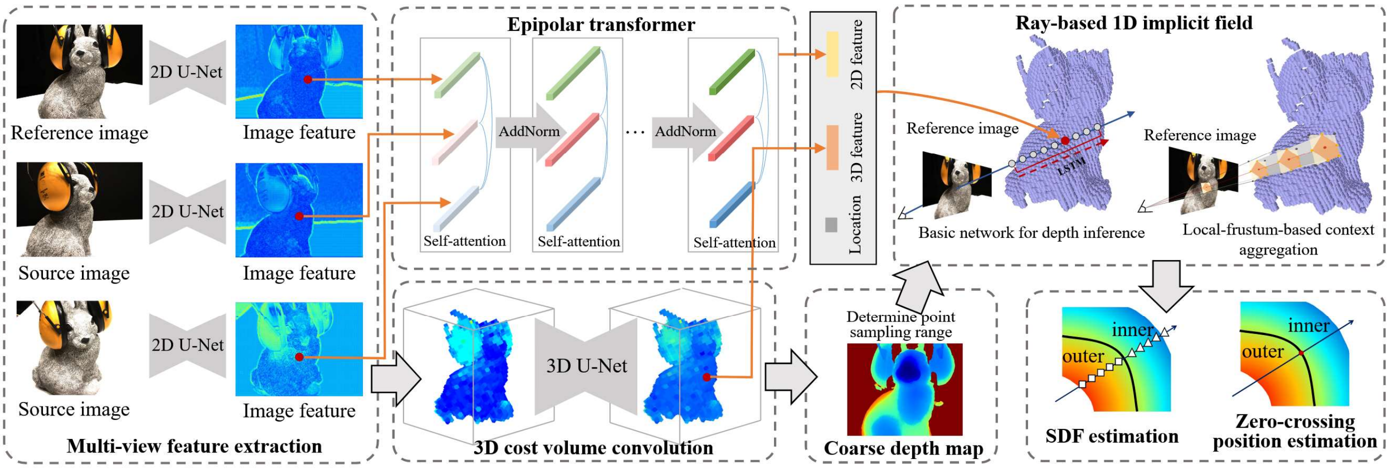
2. An epipolar transformer designed to learn cross-view feature correlation with attention mechanism.

3. A multi-task learning approach to sequential modeling and prediction of 1D implicit fields based on LSTM.

4. A local-frustum-based context aggregation that extends the receptive field of the ray-based model, leading to more accurate and robust predictions.

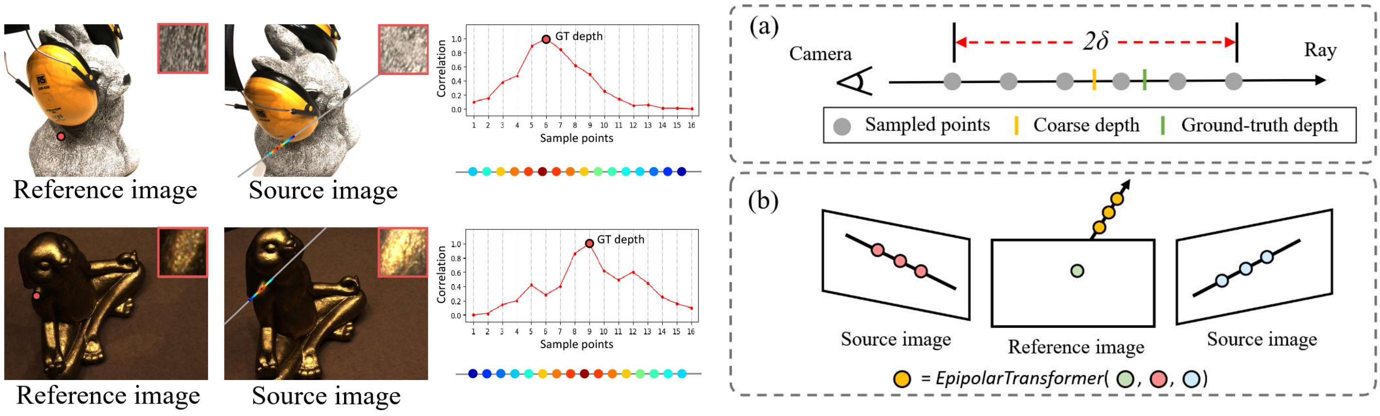
每个子问题都没有得到完美解决，并且给下一步增加了噪音，增加了管道整体工作所需的复杂性和工程工作量。
在这方面，每个子问题之间缺乏沟通就很能说明问题：如果它们互相帮助似乎更合理，即密集重建自然应该受益于为恢复相机姿势而构建的稀疏场景，反之亦然。
最重要的是，该流程中的关键步骤很脆弱

Pipeline



具体来说，为了预测某个像素位置的未来特征，F-Net 需要找到当前和之前时间步中可用的相应特征。这本质上使 F-Net 能够理解底层运动和多帧对应关系，以及较长上下文中的运动。

Method - Epipolar Transformer

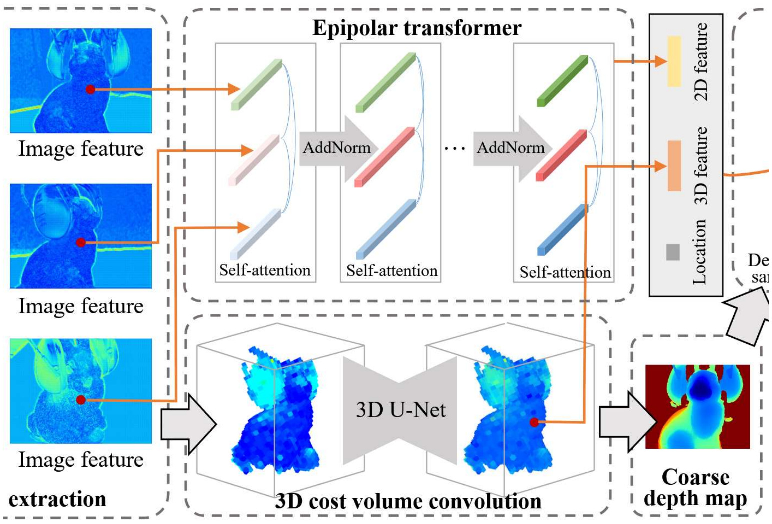


具体来说，为了预测某个像素位置的未来特征，F-Net 需要找到当前和之前时间步中可用的相应特征。这本质上使 F-Net 能够理解底层运动和多帧对应关系，以及较长上下文中的运动。

Method - Epipolar Transformer

$$\mathbf{X} = \mathrm{Concat}(\mathbf{F}^{I}_{1,p}, \ldots, \mathbf{F}^{I}_{N,p})$$

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{\mathbf{Q}},\ \mathbf{K} = \mathbf{X}\mathbf{W}^{\mathbf{K}},\ \mathbf{V} = \mathbf{X}\mathbf{W}^{\mathbf{V}}$$



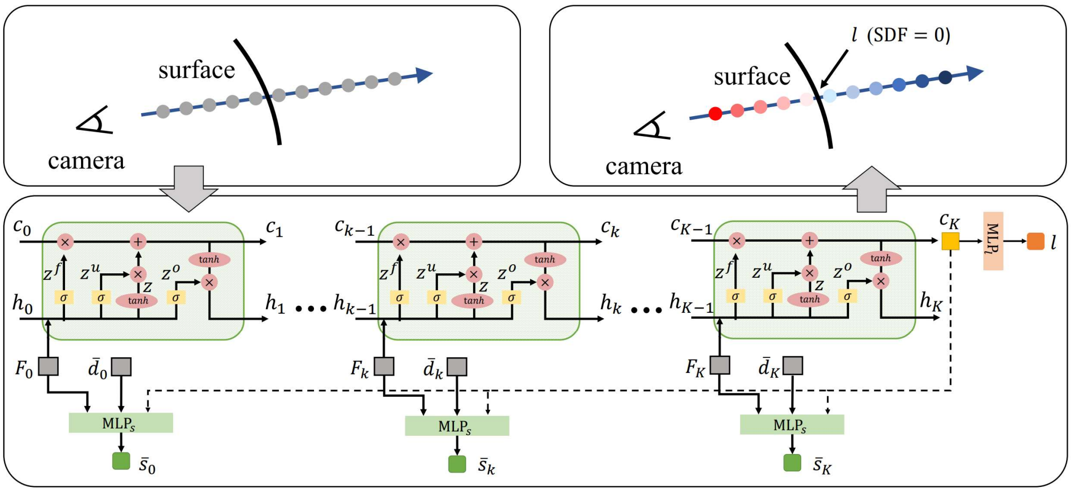
$$\mathbf{F}_p = \mathrm{Concat}(\mathbf{F}^{A}_{\mu,p}, \mathbf{F}^{A}_{\sigma,p}, \mathbf{F}^{A}_{1,p}, \mathbf{F}^{V}_{p}). \tag{3}$$

where $\mathbf{F}^{A}_{\mu,p}$ and $\mathbf{F}^{A}_{\sigma,p}$ are the mean and variation of the elements in $\mathbf{F}^{A}_{p}$ [24], [74]. $\mathbf{F}^{A}_{1,p}$ is the attention-aware feature at 3D point p in the reference image.

```
outputs_stage = epipolar_feature(self, features_stage, proj_matrices_stage, patch_idx,
                                 depth_samps=depth_range_samples,
                                 depth_samps_2d=depth_range_samples_2d,
                                 cost_reg=self.coarse.forward_epipolar,
                                 is_training=self.training)

feature_2d = outputs_stage['feature_2d']
feature_3d = outputs_stage['feature_3d']
feature_3d=feature_3d.unsqueeze(3).repeat(1,1,1,2,1,1).view(1,1,-1,cur_h//2,cur_w//2)

set=feature_2d.reshape(1,feature_2d.shape[1],8,-1).squeeze(0).permute(2,0,1)

feature_new=self.tr1(set[:,:,:], set[:,:,:])
feature_new=self.tr2(feature_new[:,:,:], feature_new[:,:,:])
feature_new=self.tr3(feature_new[:,:,:], feature_new[:,:,:])
feature_new=self.tr4(feature_new[:,:,:], feature_new[:,:,:])
```

具体来说，为了预测某个像素位置的未来特征，F-Net 需要找到当前和之前时间步中可用的相应特征。这本质上使 F-Net 能够理解底层运动和多帧对应关系，以及较长上下文中的运动。

Method: Ray-based 1D Implicit Field

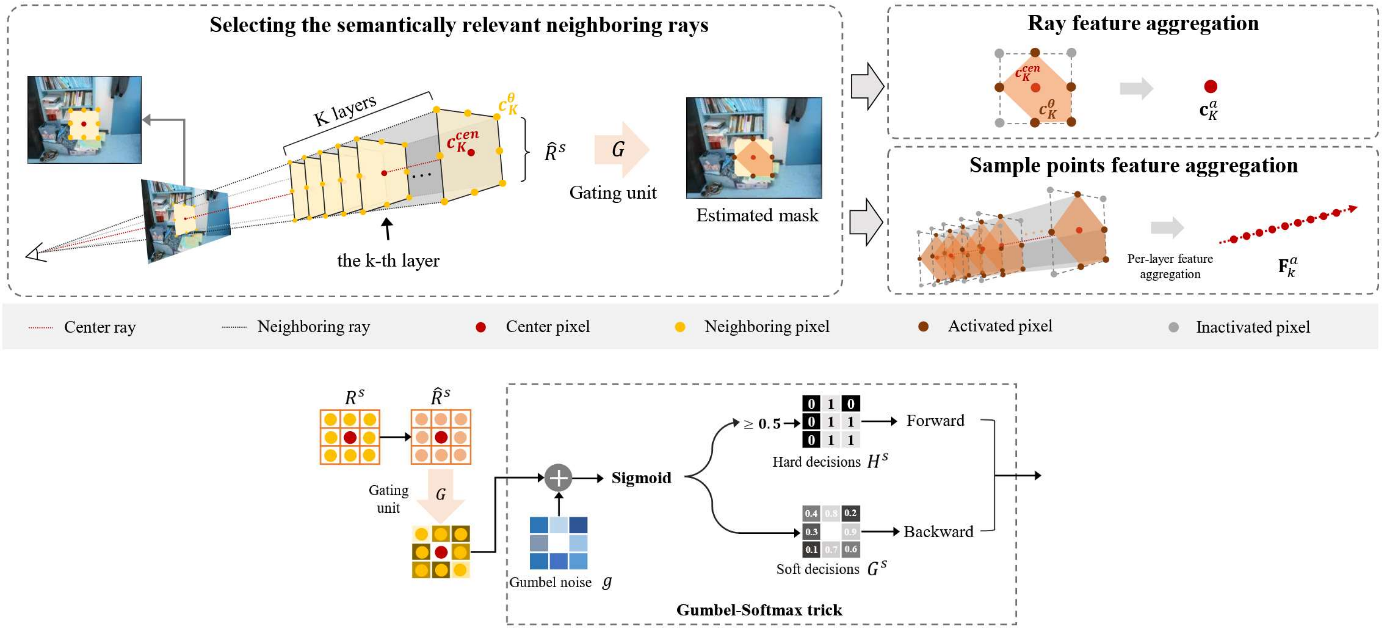


$$\bar{s}_k = \mathrm{MLP}_s([\mathbf{c}_K, \mathbf{F}_k, \bar{d}_k]).$$

$$l = \mathrm{MLP}_l(\mathbf{c}_K).$$

具体来说，为了预测某个像素位置的未来特征，F-Net 需要找到当前和之前时间步中可用的相应特征。这本质上使 F-Net 能够理解底层运动和多帧对应关系，以及较长上下文中的运动。

Method: Local-frustum-based context aggregation



具体来说，为了预测某个像素位置的未来特征，F-Net 需要找到当前和之前时间步中可用的相应特征。这本质上使 F-Net 能够理解底层运动和多帧对应关系，以及较长上下文中的运动。

Loss

$$\mathcal{L} = w_s \mathcal{L}_s + w_l \mathcal{L}_l + w_{sl} \mathcal{L}_{sl},$$

0.1, 0.8, 0.1,

$$\mathcal{L}_s = \sum_{k=1}^{K} L_1(s_k, \hat{s_k}),$$

$$\mathcal{L}_l = L_1(l, \hat{l}),$$

where $\hat{s_k}$ and $\hat{l}$ are the ground-truth, $L_1(\cdot)$ denotes the L1 loss function. $\mathcal{L}_{sl}$ is a relational loss that penalizes the inconsistency between the predicted SDFs and the predicted zero-crossing position:

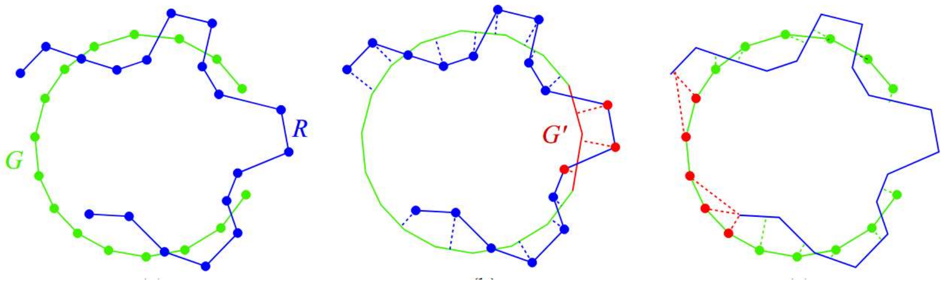
$$\mathcal{L}_{sl} = \begin{cases} 1, & s_l^a \times s_l^b > 0 \\ 0, & s_l^a \times s_l^b \leq 0, \end{cases} \tag{14}$$

where s_l^a and s_l^b are the predicted SDF of the closest two sampled points around the predicted zero-crossing position on the ray. w_s, w_l, w_{sl} are the pre-defined weights.

具体来说，为了预测某个像素位置的未来特征，F-Net 需要找到当前和之前时间步中可用的相应特征。这本质上使 F-Net 能够理解底层运动和多帧对应关系，以及较长上下文中的运动。

Evaluation - DTU

- *Accuracy* is measured as the distance from the MVS reconstruction to the structured light reference, encapsulating the quality of the reconstructed MVS points.
- *Completeness* is measured as the distance from the reference to the MVS reconstruction, encapsulating how much of the surface is captured by the MVS reconstruction.



我们最终扩展了训练集（2,400 到 14,410 帧），这显着减少了误差，表明大数据集是自监督深度训练中非常重要的元素

Evaluation - Tanks and Temples

DIG

Measures. Let $\mathcal{G}$ be the ground truth and $\mathcal{R}$ a reconstructed point set being evaluated. For a reconstructed point $\mathbf{r} \in \mathcal{R}$, its distance to the ground truth is defined as

$$e_{\mathbf{r}\to\mathcal{G}} = \min_{\mathbf{g}\in\mathcal{G}} \|\mathbf{r} - \mathbf{g}\|. \tag{3}$$

These distances can be aggregated to define the *precision* of the reconstruction $\mathcal{R}$ for any distance threshold d:

$$P(d) = \frac{100}{|\mathcal{R}|} \sum_{\mathbf{r}\in\mathcal{R}} \left[e_{\mathbf{r}\to\mathcal{G}} < d\right], \tag{4}$$

where $[\cdot]$ is the Iverson bracket. $P(d)$ is defined to lie in the range $[0,100]$ for convenience and can be interpreted as a percentage.

Similarly, for a ground-truth point $\mathbf{g} \in \mathcal{G}$, its distance to the reconstruction is defined as

$$e_{\mathbf{g}\to\mathcal{R}} = \min_{\mathbf{r}\in\mathcal{R}} \|\mathbf{g} - \mathbf{r}\|. \tag{5}$$

The *recall* of the reconstruction $\mathcal{R}$ for a distance threshold d is defined as

$$R(d) = \frac{100}{|\mathcal{G}|} \sum_{\mathbf{g}\in\mathcal{G}} \left[e_{\mathbf{g}\to\mathcal{R}} < d\right]. \tag{6}$$

Precision and recall can be combined in a summary measure, the *F-score*:

$$F(d) = \frac{2P(d)R(d)}{P(d) + R(d)}. \tag{7}$$

at regular intervals. We sampled 150 frames for Family and Horse, 500 for Palace, and 300 for all other scenes.

		(m^2)	(m)	(mm)		(M)			(sec.)
Intermediate									
Family	S	5	2.1	3	4,395	5.5	640	f/3.2	1/160
Francis	S	81	15.2	5	7,830	19.3	Auto	f/7.1	1600
Horse	S	10	3.2	3	6,015	6.2	640	f/3.2	1/160
Lighthouse	D	108	11.1	10	8,322	8.2	200	f/4.0	Auto
M60	D	35	3.2	5	5,616	9.7	400	f/2.0	1/100
Panther	D	34	2.9	5	6,570	12.3	400	f/2.0	1/100
Playground	D	54	2.8	10	7,463	1.7	200	f/2.8	Auto
Train	S	35	5.6	5	12,630	21.7	Auto	f/5.6	1/1000
Advanced									
Auditorium	S	541	6.2	10	14,640	53.4	Auto	f/2.8	1/125
Ballroom	S	254	3.9	10	10,800	43.9	6000	f/3.2	1/160
Courtroom	S	206	7.8	10	7,049	43.4	1600	Auto	1/100
Museum	S	110	21.2	10	17,115	36.5	Auto	f/3.2	1/200
Palace	D	4,295	47.2	30	21,871	41.9	Auto	f/3.2	Auto
Temple	S	713	20.7	15	17,475	33.4	Auto	f/5.6	1/640

我们最终扩展了训练集（2,400 到 14,410 帧），这显着减少了误差，表明大数据集是自监督深度训练中非常重要的元素

Experiments - DTU

Method	Accuracy	Completeness	Overall
Gipuma [17]	0.283	0.873	0.578
MVSNet [74]	0.396	0.527	0.462
R-MVSNet [75]	0.383	0.452	0.417
CIDER [69]	0.417	0.437	0.427
P-MVSNet [37]	0.406	0.434	0.420
Point-MVSNet [7]	0.342	0.411	0.376
Fast-MVSNet [79]	0.336	0.403	0.370
Att-MVSNet [38]	0.383	0.329	0.356
CasMVSNet [19]	0.325	0.385	0.355
CVP-MVSNet [72]	0.296	0.406	0.351
PatchmatchNet [62]	0.427	0.277	0.352
UCS-Net [10]	0.338	0.349	0.344
AACVP-MVSNet [78]	0.357	0.326	0.341
U-MVS [68]	0.354	0.353	0.354
RayMVSNet	0.341	0.319	0.330
RayMVSNet++	0.344	0.312	**0.328**

Rank	Model	Overall↓	Acc	Comp	Paper	Code	Result	Year
1	MVSFormer++	0.2805	0.3090	0.2521	MVSFormer++: Revealing the Devil in Transformer's Details for Multi-View Stereo			2024
2	MVSFormer	0.289	0.327	0.251	MVSFormer: Multi-View Stereo by Learning Robust Image Features and Temperature-based Depth			2022
3	ET-MVSNet	0.291	0.329	0.253	When Epipolar Constraint Meets Non-local Operators in Multi-View Stereo			2023
4	GC-MVSNet	0.295	0.330	0.260	GC-MVSNet: Multi-View, Multi-Scale, Geometrically-Consistent Multi-View Stereo			2023
5	GeoMVSNet	0.295	0.331	0.259	GeoMVSNet: Learning Multi-View Stereo With Geometry Perception			2022
6	RA-MVSNet	0.297	0.326	0.268	Multi-View Stereo Representation Revisit: Region-Aware MVSNet			2023
7	GBi-Net	0.303	0.312	0.293	Generalized Binary Search Network for Highly-Efficient Multi-View Stereo			2021
8	TransMVSNet	0.305	0.321	0.289	TransMVSNet: Global Context-aware Multi-view Stereo Network with Transformers			2021
9	CDS-MVSNet	0.315	0.351	0.278	Curvature-guided dynamic scale networks for Multi-view Stereo			2021
10	UniMVSNet	0.315	0.352	0.278	Rethinking Depth Estimation for Multi-View Stereo: A Unified Representation			2022

我们最终扩展了训练集（2,400 到 14,410 帧），这显着减少了误差，表明大数据集是自监督深度训练中非常重要的元素

Experiments - Tanks and Temples

TABLE 3: Quantitative is better).

Method	Mean
MVSNet [74]	43.48
R-MVSNet [75]	48.40
PVA-MVSNet [77]	54.46
CVP-MVSNet [72]	54.03
CasMVSNet [19]	56.84
UCS-Net [10]	54.83
D2HC-RMVSNet [71]	59.20
U-MVS [68]	57.15
RayMVSNet	**59.48**
RayMVSNet++	58.47

Rank	Model	Mean F1 (Advanced)	Mean F1 (Intermediate)	Paper	Code	Result	Year	Tags
1	MVSFormer++	41.70	67.03	MVSFormer++: Revealing the Devil in Transformer's Details for Multi-View Stereo			2024	
2	MVSFormer	40.87	66.37	MVSFormer: Multi-View Stereo by Learning Robust Image Features and Temperature-based Depth			2022	
3	GeoMVSNet	41.52	65.89	GeoMVSNet: Learning Multi-View Stereo With Geometry Perception			2023	
4	RA-MVSNet	39.93	65.72	Multi-View Stereo Representation Revisit: Region-Aware MVSNet			2023	
5	ET-MVSNet	40.41	65.49	When Epipolar Constraint Meets Non-local Operators in Multi-View Stereo			2023	
6	APD-MVS	39.91	63.64	Adaptive Patch Deformation for Textureless-Resilient Multi-View Stereo			2023	
7	GC-MVSNet	38.74	62.74	GC-MVSNet: Multi-View, Multi-Scale, Geometrically-Consistent Multi-View Stereo			2023	
8	EPP-MVSNet	35.72	61.68	EPP-MVSNet: Epipolar-Assembling Based Depth Prediction for Multi-View Stereo			2021	
9	CDS-MVSNet		61.58	Curvature-guided dynamic scale networks for Multi-view Stereo			2021	
10	AA-RMVSNet		61.51	AA-RMVSNet: Adaptive Aggregation Recurrent Multi-view Stereo Network			2021	
11	GBi-Net		61.42	Generalized Binary Search Network for Highly-Efficient Multi-View Stereo			2021	
12	Vis-MVSNet		60.03	Visibility-aware Multi-view Stereo Network			2020	

我们最终扩展了训练集（2,400 到 14,410 帧），这显着减少了误差，表明大数据集是自监督深度训练中非常重要的元素

Experiments - Tanks and Temples

DIG

Intermediate ▾ Advanced ▾

Intermediate F-score

method	rank	mean	runtime*	Family	rain
Steuart Systems	**2.50**	**71.97**	N.A.	**83.02**	**2.99**
ETV-MVS	15.38	67.21	N.A.	80.14	0.56
MVSFormer_plusplus	17.62	67.18	N.A.	82.69	0.12

Steuart Systems 历史

2024 年：在 Tanks and Temples 2D 到 3D 基准测试中，在近 500 个游戏中排名第一

2022 年：在 ETH 摄影测量 3D 基准测试中，从近 200 个参赛作品中排名第一，两年后该作品排名第五

2020 年：构建并测试 32 摄像头原型

2015 年：发现被动式 3D 扫描。开始使用 DJI 无人机和 Agisoft Metashape 制作 3D 模型。

2011 年：开始迁移到 Linux、OpenCL 和 AMD APU（APU 在同一芯片上具有 CPU 和 GPU，并且可以有效共享内存）

2008 年：编程团队开始使用 CUDA 1.0 将图像处理速度提高约 10 倍

2006 年：设计/建造带图案投影的 8 相机扫描仪，并在第四届 3D 数字成像和建模国际会议上与 UVA 发表论文

2006 年：开始培养 HDR 摄影技能

2005 年至 2018 年：建造了几个带有摄像头、灯光、伺服器和激光器的原始相机阵列原型

2003 年：申请 3D-360 专利，解决全景摄影中的问题（2D-360s）

1996 年：开始使用 Apple 的 QuickTimeVR 1.0 制作全景图（2D-360s）

1987 年至 2002 年：为多达 1,000 名员工的公司设计和管理可靠计算机和网络系统的部分系统。

1984 年：采用摩托罗拉 6800 系列 8 位 CPU 和汇编代码构建第一个计算机系统（带有传感器、伺服器

我们最终扩展了训练集（2,400 到 14,410 帧），这显着减少了误差，表明大数据集是自监督深度训练中非常重要的元素

Ablation

TABLE 7: Ablation studies of RayMVSNet. The performance under distance metric is reported (lower is better).

Method	Accuracy	Completeness	Overall
w/o epipolar transformer	0.347	0.339	0.343
w/o 2D image feature	0.345	0.352	0.348
w/o 3D volume feature	0.434	0.322	0.378
vis-max feature aggregation	0.345	0.331	0.338
w/o ray-based inference	0.573	0.642	0.608
Ray with Transformer	**0.339**	0.343	0.341
Ray with average pooling	0.356	0.406	0.381
Ray with max pooling	0.466	0.383	0.424
w/o SDF prediction	0.354	0.330	0.342
Visibility-aware view aggregation	0.345	0.331	0.338
RayMVSNet	0.341	**0.319**	**0.330**

TABLE 8: Ablation studies of RayMVSNet++. p@x represents Percentage@x.

Method	RMSE(m)↓	p@0.2↑	p@0.4↑	p@0.6 ↑
w/o frustum	0.211	0.794	0.918	0.963
w/o gating unit	0.193	0.807	0.925	0.966
w/o Gumbel-Softmax	0.176	0.838	0.950	0.980
RayMVSNet++	**0.158**	**0.861**	**0.957**	**0.982**

我们最终扩展了训练集（2,400 到 14,410 帧），这显着减少了误差，表明大数据集是自监督深度训练中非常重要的元素

Thanks