Exploring the Point Feature Relation on Point Cloud for Multi-View Stereo (2023 TCSVT CCF-B)

2024.06.13

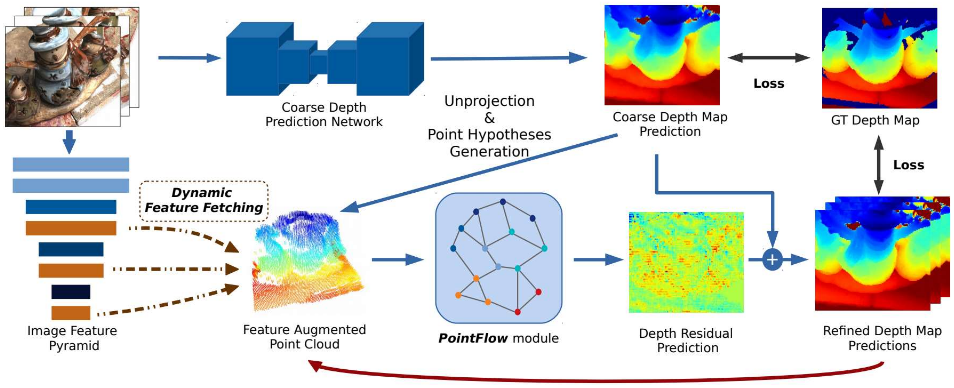
Existing Problems

1. Limited by the voxel structure, cost volume-based methods fail to establish the structure features among voxels inside the cost volume. Meanwhile, the cost volume is unable to dynamically learn the geometric properties among structures during the regularization. Therefore, the geometric features of the scene are lost in the process of network learning.

2. Cost volume-based networks cannot break through the spatial limitation of the 3D convolution kernel to freely perceive more reasonable feature representations outside the convolution kernel. Therefore, underlying similarity features cannot be learned during the cost volume regularization, which hinders the generalization of the network to unknown scenes.

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

Existing Problems



3. In point cloud-based methods, DGCNN fails to dynamically perceive the geometric properties implied in the scene based on the feature discrepancies among the structures. Moreover, point cloud-based methods only rely on kNN (k-Nearest Neighbor) to establish a local perception region. This weakens the positive effect of the intra-region features on the learning efficiency of the network.

基于点云的方法：PointMVSNet [55]提出在点云上推断深度图，它首先迭代地对前一阶段的深度图进行上采样并将其投影到点云。然后，DGCNN [22] 不断学习并聚合区域内的结构信息以形成点特征。最后，从点特征推断出高分辨率深度图

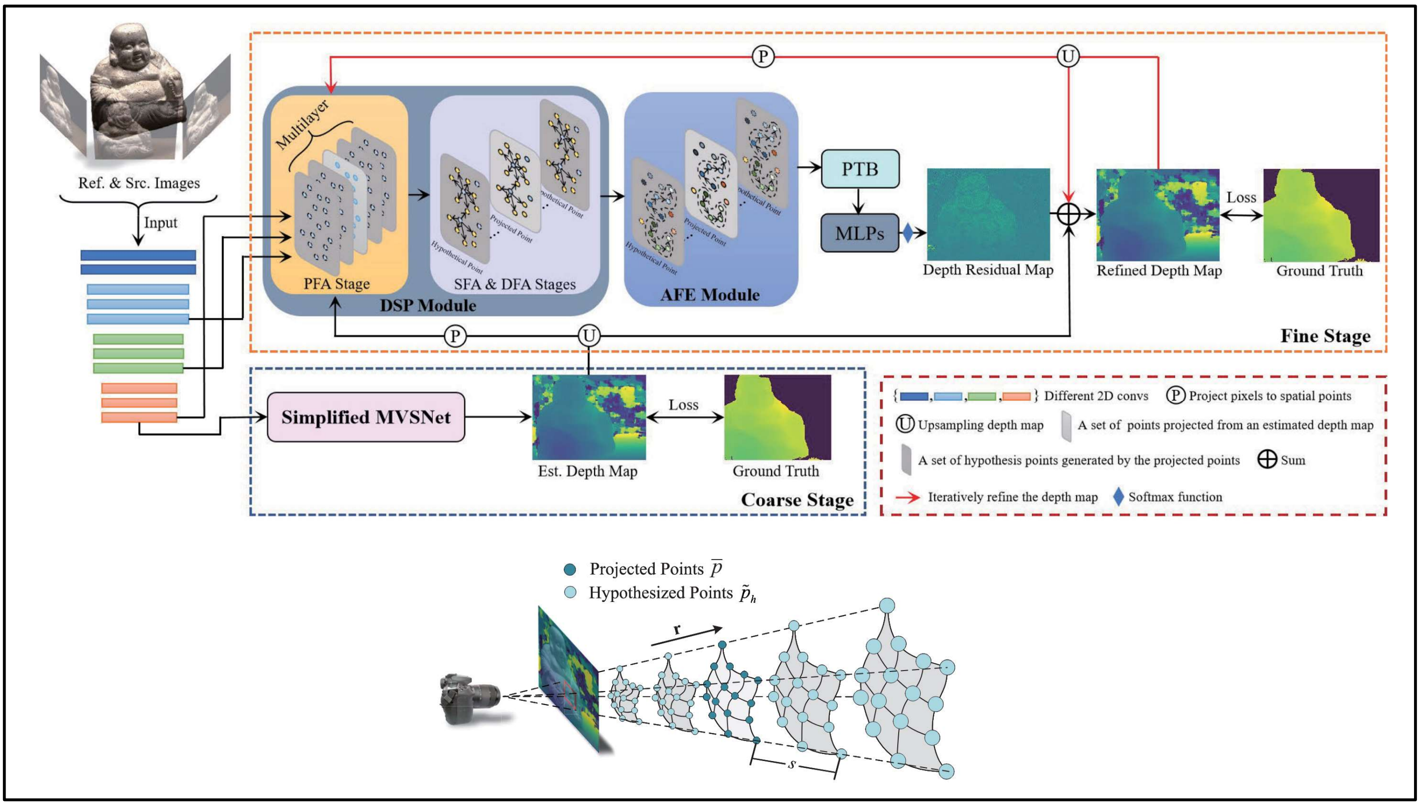
Contribution

The DSP module first augments the features of the 3D point cloud from the features of the multi-view 2D images and establishes the geometry of the scene, and dynamically aggregates local structure information to point features based on multi-dimensional structure similarity, guiding the feature representation of points to be more reasonable.

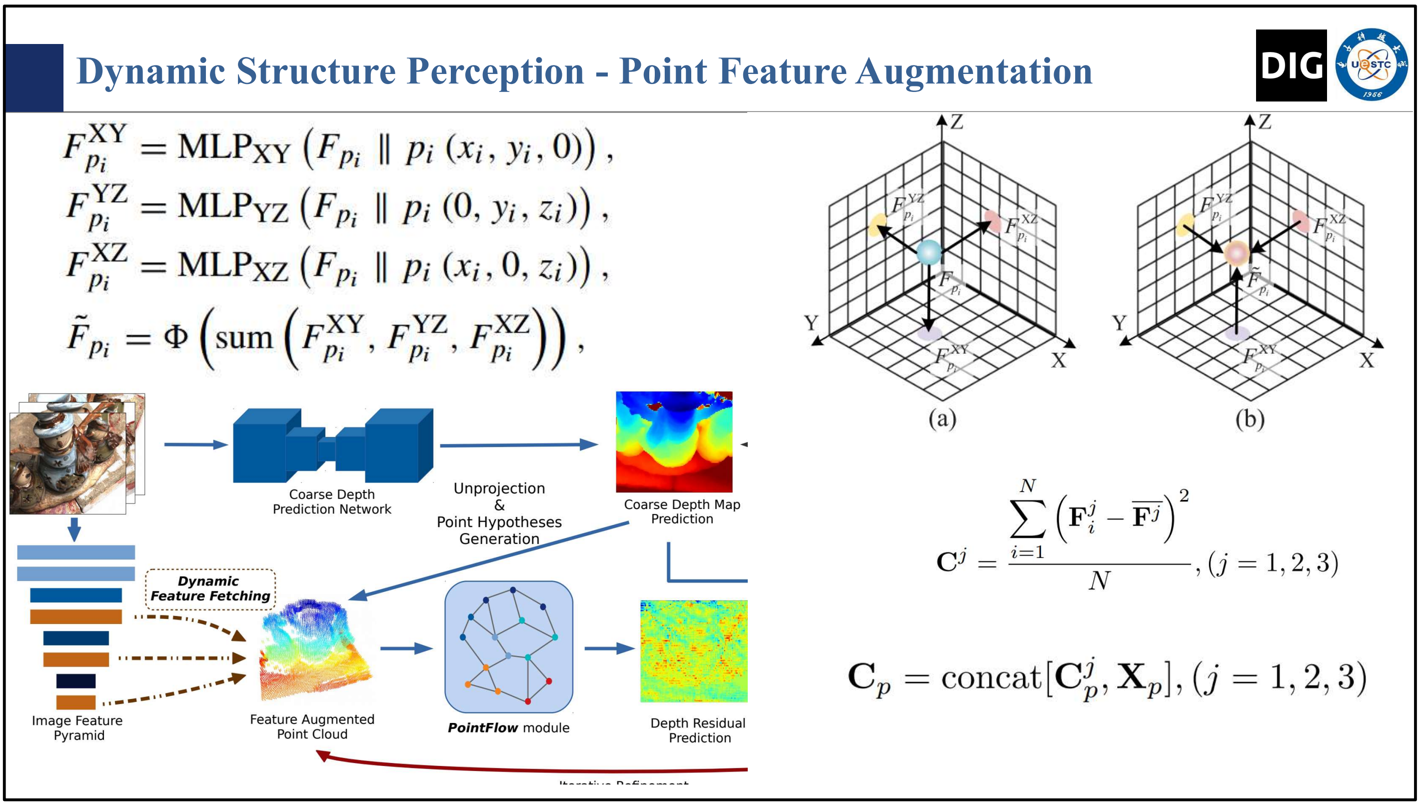
The AFE module adaptively explores the feature similarity region for each point with aggregated structure features.

The PTB module with multiple point transformer layers fully learns the feature correlations among intra-region points, producing more discriminative feature representations.

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

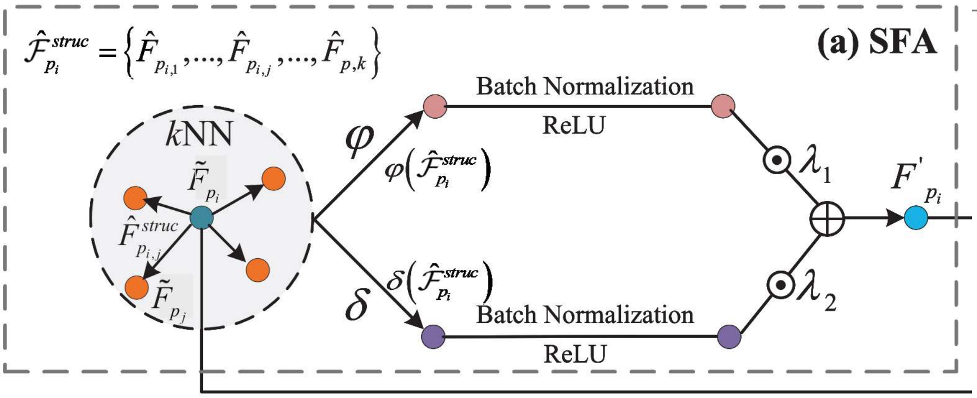


具体来说，为了预测某个像素位置的未来特征，F-Net 需要找到当前和之前时间步中可用的相应特征。这本质上使 F-Net 能够理解底层运动和多帧对应关系，以及较长上下文中的运动。



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。

Dynamic Structure Perception - Structure-Based Feature Aggregation



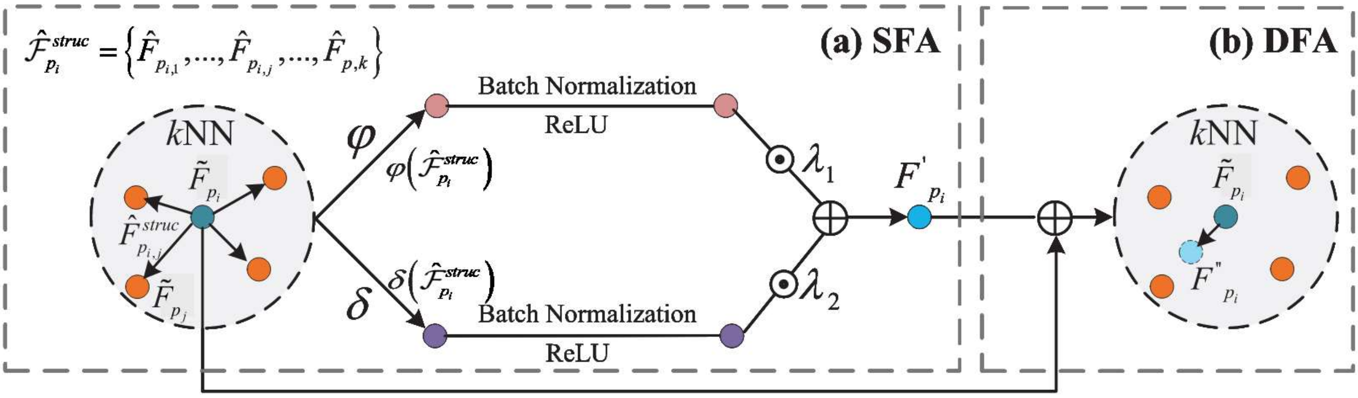
$$f\left(\hat{\mathcal{F}}_{p_i}^{struc}\right)=\lambda_1\cdot\varphi\left(\hat{\mathcal{F}}_{p_i}^{struc}\right)+\lambda_2\cdot\delta\left(\hat{\mathcal{F}}_{p_i}^{struc}\right),$$

$$\varphi\left(\hat{\mathcal{F}}_{p_i}^{struc}\right)=\sum_{j=1}^{k}\left(\mathcal{R}\left(\tilde{F}_{p_i}\right)\mathbf{W}_3\times\rho\left(\hat{\mathcal{F}}_{p_i}^{struc}\mathbf{W}_1\times\hat{\mathcal{F}}_{p_i}^{struc}\mathbf{W}_2\right)\right), \tag{6}$$

$$\delta\left(\hat{\mathcal{F}}_{p_i}^{struc}\right)=\sum_{j=1}^{k}\left(\mathcal{R}\left(\tilde{F}_{p_i}\right)\mathbf{W}_3\times\rho\left(\frac{\hat{\mathcal{F}}_{p_i}^{struc}\mathbf{W}_1\times\hat{\mathcal{F}}_{p_i}^{struc}\mathbf{W}_2}{\left|\hat{\mathcal{F}}_{p_i}^{struc}\mathbf{W}_1\right|\left|\hat{\mathcal{F}}_{p_i}^{struc}\mathbf{W}_2\right|}\right)\right), \tag{7}$$

表示使用自注意力机制聚合结构信息，δ (·) 表示使用余弦相似度聚合结构信息

Dynamic Structure Perception - Dynamic Feature Adjustment



$$F''_{p_i} = \tilde{F}_{p_i} + F'_{p_i}.$$

在SFA阶段，我们同时考虑局部区域结构特征之间的数值相似性和方向相似性，并通过相应的权重聚合局部结构信息。这种方法自然解决了特征空间中位移方向和大小的问题。

Method - Adaptive Feature Exploration

Algorithm 1 Region Construcτuion

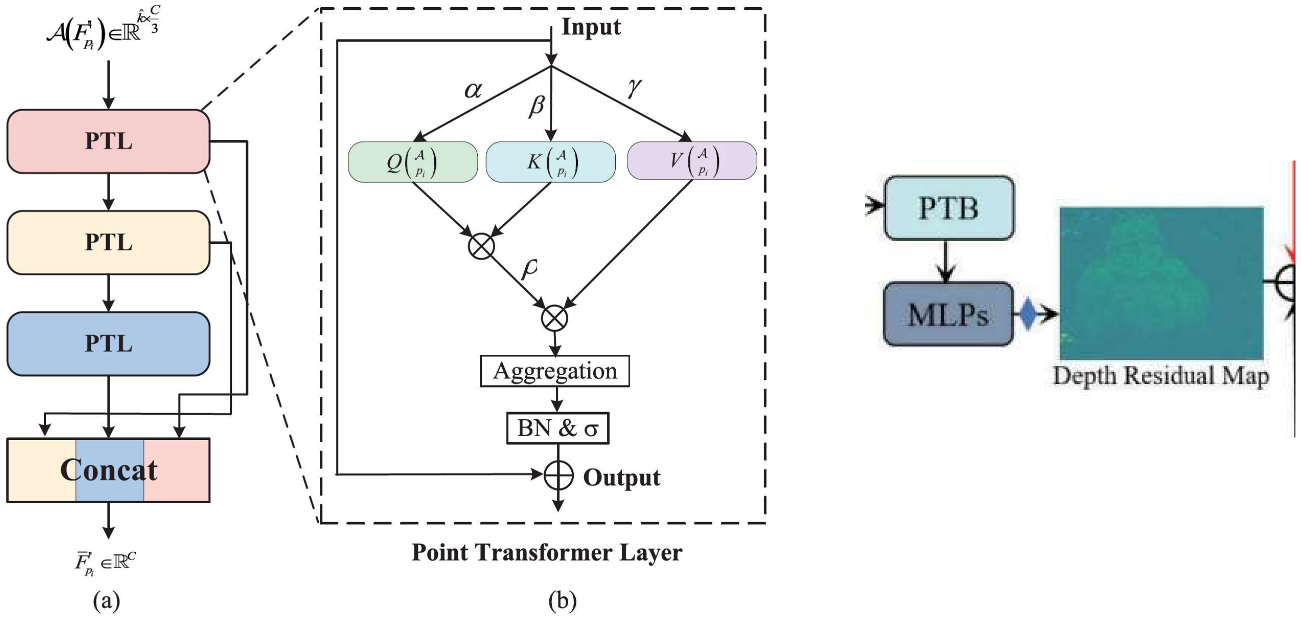
Input: Intra-region point feature $\mathcal{R}\left(F''_{p_i}\right) = \left\{F''_{p_{i,j}}\right\}_{j=1}^{k}$ in the kNN centered on a sampled point p_i.

Output: Intra - region point feature $\mathcal{A}\left(F''_{p_i}\right) = \left\{F''_{p_i^n}\right\}_{n=1}^{\hat{k}}$ in the adaptively partition region starting at a sampled point p_i.

```
Compute feature similarity weight
  {w_{i,j} | i = 1, ..., N; j = 1, ..., k} between p_i and neighbors
  {p_{i,j}}_{j=1}^{k} using Eq. (11)
for n = 0 to k̂ do
    if n == 0 then
        Select the point p_{i,j}^n with the second similar
        feature to p_i (p_i^n) in the weight {w_{i,j}} as p_i^{n+1}
        p_i^{n+1} is added to region A(p_i)
    else
        Select the point p_{i,j}^n with the second similar
        feature to p_i^n in the weight {w_{i,j}} as p_i^{n+1}
        p_i^{n+1} is added to region A(p_i)
    end
end
Gather corresponding point features A(F''_{p_i}) in
  region A(p_i)
return A(F''_{p_i})
```

AFE模块旨在通过自适应地探索与采样点具有相似特征的邻近点，为每个采样点构造一个新的局部感知区域。

Method - Point Transformer Block



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。

Experiments - DTU

Method	Acc.(mm)	Comp.(mm)	Overall.(mm)
Furu* [23]	0.613	0.941	0.777
Gipuma* [10]	**0.283**	0.873	0.578
COLMAP* [11]	0.400	**0.664**	**0.532**
SurfaceNet [34]	0.450	1.040	0.745
MVSNet [9]	0.396	0.527	0.462
R-MVSNet [14]	0.383	0.452	0.417
MVSCRF [38]	0.371	0.426	0.398
PointMVSNet [55]	0.342	0.411	0.376
VA-PointMVSNet[21]	0.359	0.358	0.359
CasMVSNet [18]	0.325	0.385	0.355
CVP-MVSNet [41]	0.296	0.406	0.351
PatchmatchNet [46]	0.427	0.277	0.352
AA-RMVSNet [16]	0.376	0.339	0.357
BH-RMVSNet [17]	0.368	0.303	0.335
UGNet [45]	0.334	0.330	0.332
Effi-MVS [56]	0.321	0.313	0.317
NP-MVSNet [20]	0.356	**0.275**	0.315
UniMVSNet [43]	0.352	0.278	0.315
TransMVSNet [59]	0.321	0.289	**0.305**
Ours	**0.289**	0.383	0.336

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

Method	Mean
COLMAP* [11]	42.14
OpenMVS* [54]	55.11
ACMP* [13]	58.41
ACMMP* [12]	59.38
MVSNet [9]	43.48
R-MVSNet [14]	48.40
MVSCRF [38]	45.73
PointMVSNet [55]	48.27
VA-PointMVSnet [21]	48.70
CasMVSNet [18]	56.84
CVP-MVSNet [41]	54.03
PatchmatchNet [46]	53.15
Effi-MVS [56]	56.88
NP-MVSNet [20]	59.64
BH-RMVSNet [17]	61.96
UGNet [45]	63.12
UniMVSNet [43]	64.36
Trans-MVSNet [59]	63.52
Ours (without fine-tuning)	63.18
Ours (with fine-tuning)	**64.56**

Method	Mean
COLMAP* [11]	27.24
OpenMVS* [54]	34.43
ACMP* [13]	37.44
ACMMP* [12]	37.84
R-MVSNet [14]	24.91
CasMVSNet [18]	31.12
PatchmatchNet [46]	32.31
AA-RMVSNet [16]	33.53
Effi-MVS [56]	34.39
BH-RMVSNet [17]	32.72
UGNet [45]	37.12
UniMVSNet [43]	38.96
Trans-MVSNet [59]	37.00
Ours (without fine-tuning)	36.77
Ours (with fine-tuning)	**39.22**

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

Ablation

Point Hypotheses	Acc. (mm)	Comp. (mm)	Overall (mm)
b = 1	0.302	**0.380**	0.341
***b* = 2**	**0.289**	0.383	**0.336**

Model	DSP			AFE	PTB	Acc.(mm)	Comp.(mm)	Overall (mm)	F-score (%)
	PFA	SFA	DFA						
Baseline					√	0.320	0.400	0.360	53.51
+A	√				√	0.315	0.398	0.356	56.06
+B		√	√		√	0.309	0.394	0.351	57.53
+C	√	√	√		√	0.298	0.389	0.343	61.74
+D				√	√	0.317	0.396	0.356	55.90
+E	√	√	√	√	√	**0.289**	**0.383**	**0.336**	**63.18**

Strategy of Feature Encoding	Acc. (mm)	Comp. (mm)	Overall (mm)
Original Feature	0.306	0.394	0.350
Original Feature + Coordinate	0.293	0.388	0.340
Position Encoding	0.297	0.396	0.346
Tri-plane Feature (Ours)	**0.289**	**0.383**	**0.336**

k value	Acc. (mm)	Comp. (mm)	Overall (mm)	Mem.(GB)
k = 10	0.304	0.385	0.344	**13.91**
k = 12	0.300	0.384	0.342	14.29
***k* = 14**	**0.289**	0.383	**0.336**	14.71
k = 16	0.293	**0.381**	0.337	15.06
k = 18	0.301	0.384	0.343	15.34

Agg. Method	Acc. (mm)	Comp. (mm)	Overall (mm)	Runtime (s)
Sum	0.297	0.387	0.342	3.04
Max	0.299	0.387	0.343	3.05
Mean	0.295	0.385	0.340	**3.03**
Attention	0.293	0.384	0.338	3.39
SFA (Ours)	**0.289**	**0.383**	**0.336**	3.76

Iter.	Acc. (mm)	Comp. (mm)	O.A. (mm)	Depth Size	Depth interval (mm)
-	0.540	0.609	0.574	200 × 144	5.30
1	0.546	0.602	0.574	200 × 144	5.30
2	0.373	0.419	0.396	400 × 288	4.00
3	**0.289**	**0.383**	**0.336**	800 × 576	0.80
[55]	0.342	0.411	0.376	800 × 576	0.80

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

Limitation

Method	Depth Size	Mem. (MB)	Runtime (s)	Acc. (mm)	Comp. (mm)	Overall (mm)
CVP-MVSNet [41]	800×576	**2207**	0.49	0.340	0.418	0.379
TansMVSNet [59]	800×576	2381	0.48	0.377	**0.267**	**0.322**
UniMVSNet [43]	800×576	3931	**0.21**	0.385	0.296	0.341
PointMVSNet [55]	800×576	13127	3.08	0.342	0.411	0.376
Ours	800×576	14719	3.76	**0.289**	0.383	0.336

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