DIG

DUSt3R: Geometric 3D Vision Made Easy (CVPR 2024)

2024.03.28

Motivation & Contribution

matching points

finding essential matrices

triangulating points

sparsely reconstructing the scene

estimating cameras

finally performing dense reconstruction

1. First holistic **end-to-end** 3D reconstruction pipeline from un-calibrated and un-posed images, that **unifies monocular and binocular 3D reconstruction.**

2. **Pointmap representation** for MVS applications

3. An optimization procedure to **globally align pointmaps** in the context of multi-view 3D reconstruction. Extract effortlessly all usual **intermediary outputs** of the classical SfM and MVS pipelines. Our approach **unifies all 3D vision tasks**

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

Pointmap

$$\textit{pointmap}\ X \in \mathbb{R}^{W \times H \times 3}$$

$$X_{i,j} = K^{-1}\left[iD_{i,j}, jD_{i,j}, D_{i,j}\right]^\top$$

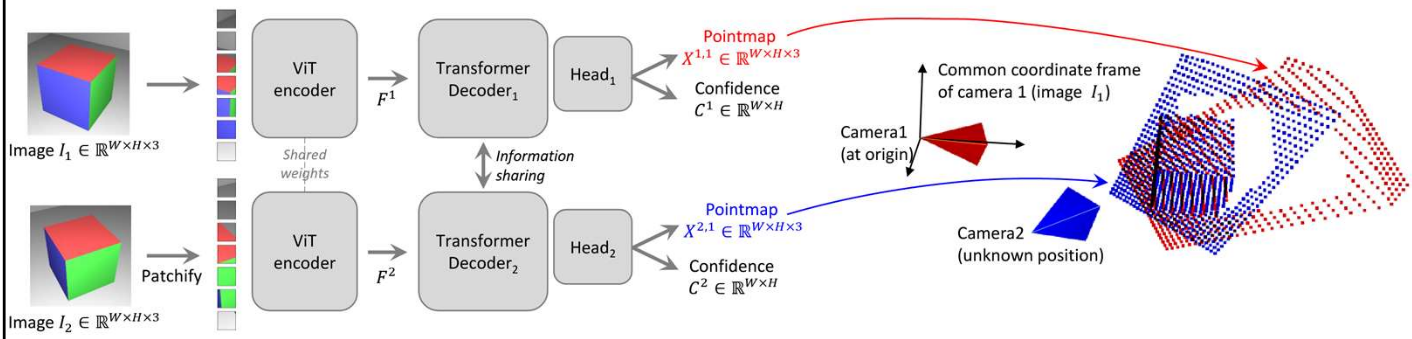
denote as $X^{n,m}$ the pointmap X^n from camera n expressed in camera m's coordinate frame:

$$X^{n,m} = P_m P_n^{-1} h\left(X^n\right) \tag{1}$$

with $P_m, P_n \in \mathbb{R}^{3\times 4}$ the world-to-camera poses for images n and m, and $h : (x, y, z) \rightarrow (x, y, z, 1)$ the homogeneous mapping.

optical flow: instance tracking

Method: Pipeline



Note that both pointmaps are expressed in the same coordinate frame of

$$\ell_{\text{regr}}(v,i) = \left\| \frac{1}{z} X_i^{v,1} - \frac{1}{\bar{z}} \bar{X}_i^{v,1} \right\|.$$

$$\text{norm}(X^1, X^2) = \frac{1}{|\mathcal{D}^1| + |\mathcal{D}^2|} \sum_{v\in\{1,2\}} \sum_{i\in\mathcal{D}^v} \|X_i^v\|.$$

$$z = \text{norm}(X^{1,1}, X^{2,1})$$

$$\bar{z} = \text{norm}(\bar{X}^{1,1}, \bar{X}^{2,1})$$

$$\mathcal{L}_{\text{conf}} = \sum_{v\in\{1,2\}} \sum_{i\in\mathcal{D}^v} C_i^{v,1} \ell_{\text{regr}}(v,i) - \alpha \log C_i^{v,1},$$

optical flow: instance tracking

Downstream Applications

$$f_1^* = \arg\min_{f_1} \sum_{i=0}^{W} \sum_{j=0}^{H} C_{i,j}^{1,1} \left\| (i', j') - f_1 \frac{(X_{i,j,0}^{1,1}, X_{i,j,1}^{1,1})}{X_{i,j,2}^{1,1}} \right\|,$$

$$i' = i - \frac{W}{2} \qquad j' = j - \frac{H}{2}$$

$$R^*, t^* = \arg\min_{\sigma,R,t} \sum_i C_i^{1,1} C_i^{1,2} \left\| \sigma(R X_i^{1,1} + t) - X_i^{1,2} \right\|^2,$$

Edge: 当相邻像素有相同的全景标识符时（Iverson bracket为0），depth本身梯度越大，loss越高

Edge: 当相邻像素有不同的全景标识符时（Iverson bracket为1），这种损失会在全景边缘的视差图中强制出现梯度峰值（depth本身梯度越大，说明很正确，梯度越小）

通过学习，使得类间的距离要大于类内的距离。锚点位于面片的中心，正特征是与锚点具有相同全景类别的特征，负特征是具有不同全景类别的特征

Global Optimization

Given a set of images $\{I^1, I^2, \ldots, I^N\}$

$\mathcal{G}(\mathcal{V}, \mathcal{E})$ where N images form vertices $\mathcal{V}$ and each edge $e = (n, m) \in \mathcal{E}$ indicates that images I^n and I^m shares some visual content. To that aim, we either use existing off-the-shelf image retrieval methods, or we pass all pairs through network $\mathcal{F}$ (inference takes ≈40ms on a H100 GPU)

$$X^{n,e} := X^{n,n} \text{ and } X^{m,e} := X^{m,n}$$

$$\chi^* = \underset{\chi, P, \sigma}{\arg\min} \sum_{e \in \mathcal{E}} \sum_{v \in e} \sum_{i=1}^{HW} C_i^{v,e} \left\| \chi_i^v - \sigma_e P_e X_i^{v,e} \right\| .$$

DUSt3R: Geometric 3D Vision Made Easy

S. Wang [1], V. Leroy [2], Y. Cabon [2], B. Chidlovskii [2] and J. Revaud [2]

[1] Aalto University 2 Naver Labs Europe

The optimization is carried out using standard gradient descent and typically converges after a few hundred steps, requiring mere seconds on a standard GPU.

对于每张图，每条边，

Experiments

DIG USTC 1958

Relative Pose Estimation

Habitat
MegaDepth
ARKitScenes
MegaDepth
Static Scenes 3D
Blended MVS
ScanNet++
CO3D-v2
Waymo

CO3Dv2 (COLMAP)
RealEstate10k (SLAM with bundle adjustment)

Methods	Co3Dv2 [93]			RealEstate10K
	RRA@15	RTA@15	mAA(30)	mAA(30)
RelPose [176]	57.1	-	-	-
Colmap+SPSG [26, 99]	36.1	27.3	25.3	45.2
PixSfM [58]	33.7	32.9	30.1	49.4
PosReg [139]	53.2	49.1	45.0	-
PoseDiffusion [139]	80.5	79.8	66.5	48.0
DUSt3R 512 (w/ PnP)	94.3	**88.4**	**77.2**	61.2
DUSt3R 512 (w/ GA)	**96.2**	86.8	76.7	**67.7**

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

Methods	Train	Outdoor				Indoor					
		DDAD[40]		KITTI [35]		BONN [79]		NYUD-v2 [114]		TUM [118]	
		Rel↓	$\delta_{1.25}\uparrow$	Rel↓	$\delta_{1.25}\uparrow$	Rel↓	$\delta_{1.25}\uparrow$	Rel↓	$\delta_{1.25}\uparrow$	Rel↓	$\delta_{1.25}\uparrow$
DPT-BEiT[90]	D	10.70	**84.63**	9.45	89.27	-	-	**5.40**	**96.54**	**10.45**	**89.68**
NeWCRFs[173]	D	**9.59**	82.92	**5.43**	**91.54**	-	-	6.22	95.58	14.63	82.95
Monodepth2 [37]	SS	23.91	75.22	11.42	86.90	56.49	35.18	16.19	74.50	31.20	47.42
SC-SfM-Learners [6]	SS	16.92	77.28	11.83	86.61	21.11	71.40	13.79	79.57	22.29	64.30
SC-DepthV3 [120]	SS	**14.20**	**81.27**	11.79	86.39	**12.58**	**88.92**	**12.34**	**84.80**	**16.28**	**79.67**
MonoViT[181]	SS	-	-	**09.92**	**90.01**	-	-	-	-	-	
RobustMIX [91]	T	-	-	18.25	76.95	-	-	11.77	90.45	15.65	**86.59**
SlowTv [116]	T	**12.63**	79.34	(6.84)	(56.17)	-	-	11.59	87.23	15.02	80.86
DUSt3R 224-NoCroCo	T	19.63	70.03	20.10	71.21	14.44	86.00	14.51	81.06	22.14	66.26
DUSt3R 224	T	16.32	77.58	16.97	77.89	11.05	89.95	10.28	88.92	17.61	75.44
DUSt3R 512	T	13.88	81.17	**10.74**	**86.60**	**8.08**	**93.56**	**6.50**	94.09	**14.17**	79.89

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

	Methods	GT Pose	GT Range	GT Intrinsics	Align	KITTI rel ↓	KITTI τ ↑	ScanNet rel ↓	ScanNet τ ↑	ETH3D rel ↓	ETH3D τ ↑	DTU rel ↓	DTU τ ↑	T&T rel↓	T&T τ ↑	Average rel↓	Average τ ↑	time (s)↓
(a)	COLMAP [105, 106]	✓	×	✓	×	**12.0**	**58.2**	**14.6**	**34.2**	**16.4**	**55.1**	**0.7**	**96.5**	**2.7**	95.0	**9.3**	**67.8**	≈ 3 min
	COLMAP Dense [105, 106]	✓	×	✓	×	26.9	52.7	38.0	22.5	89.8	23.2	20.8	69.3	25.7	76.4	40.2	48.8	≈ 3 min
(b)	MVSNet [160]	✓	✓	✓	×	22.7	36.1	24.6	20.4	35.4	31.4	(1.8)	(86.0)	8.3	73.0	18.6	49.4	0.07
	MVSNet Inv. Depth [160]	✓	✓	✓	×	18.6	30.7	22.7	20.9	21.6	35.6	(1.8)	(86.7)	6.5	74.6	14.2	49.7	0.32
	Vis-MVSSNet [175]	✓	✓	✓	×	**9.5**	**55.4**	8.9	33.5	**10.8**	**43.3**	**(1.8)**	**(87.4)**	**4.1**	**87.2**	**7.0**	**61.4**	0.70
	MVS2D ScanNet [159]	✓	✓	✓	×	21.2	8.7	(27.2)	(5.3)	27.4	4.8	17.2	9.8	29.2	4.4	24.4	6.6	**0.04**
	MVS2D DTU [159]	✓	✓	✓	×	226.6	0.7	32.3	11.1	99.0	11.6	(3.6)	(64.2)	25.8	28.0	77.5	23.1	0.05
(c)	DeMon [135]	✓	×	✓	×	16.7	13.4	75.0	0.0	19.0	16.2	23.7	11.5	17.6	18.3	30.4	11.9	0.08
	DeepV2D KITTI [130]	✓	×	✓	×	(20.4)	(16.3)	25.8	8.1	30.1	9.4	24.6	8.2	38.5	9.6	27.9	10.3	1.43
	DeepV2D ScanNet [130]	✓	×	✓	×	61.9	5.2	(3.8)	(60.2)	18.7	28.7	9.2	27.4	33.5	38.0	25.4	31.9	2.15
	MVSNet [160]	✓	×	✓	×	14.0	35.8	1568.0	5.7	507.7	8.3	(4429.1)	(0.1)	118.2	50.7	1327.4	20.1	0.15
	MVSNet Inv. Depth [160]	✓	×	✓	×	29.6	8.1	65.2	28.5	60.3	5.8	(28.7)	(48.9)	51.4	14.6	47.0	21.2	0.28
	Vis-MVSNet [175]	✓	×	✓	×	10.3	**54.4**	84.9	15.6	51.5	17.4	(374.2)	(1.7)	21.1	65.6	108.4	31.0	0.82
	MVS2D ScanNet [159]	✓	×	✓	×	73.4	0.0	(4.5)	(54.1)	30.7	14.4	5.0	57.9	56.4	11.1	34.0	27.5	**0.05**
	MVS2D DTU [159]	✓	×	✓	×	93.3	0.0	51.5	1.6	78.0	0.0	(1.6)	(92.3)	87.5	0.0	62.4	18.8	0.06
	Robust MVD Baseline [109]	✓	×	✓	×	**7.1**	41.9	**7.4**	**38.4**	**9.0**	**42.6**	**2.7**	**82.0**	**5.0**	**75.1**	**6.3**	**56.0**	0.06
(d)	DeMoN [135]	×	×	✓	$\|\mathbf{t}\|$	15.5	15.2	12.0	21.0	17.4	15.4	21.8	16.6	13.0	23.2	16.0	18.3	0.08
	DeepV2D KITTI [130]	×	×	✓	med	(3.1)	(74.9)	23.7	11.1	27.1	10.1	24.8	8.1	34.1	9.1	22.6	22.7	2.07
	DeepV2D ScanNet [130]	×	×	✓	med	10.0	36.2	**(4.4)**	(54.8)	11.8	29.3	7.7	33.0	8.9	46.4	8.6	39.9	3.57
	DUSt3R 224-NoCroCo	×	×	×	med	15.14	21.16	7.54	40.00	9.51	40.07	3.56	62.83	11.12	37.90	9.37	40.39	**0.05**
	DUSt3R 224	×	×	×	med	15.39	26.69	(5.86)	(50.84)	4.71	61.74	**2.76**	**77.32**	5.54	56.38	6.85	54.59	**0.05**
	DUSt3R 512	×	×	×	med	**9.11**	**39.49**	(4.93)	**(60.20)**	**2.91**	**76.91**	3.52	69.33	**3.17**	**76.68**	**4.73**	**64.52**	0.13

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