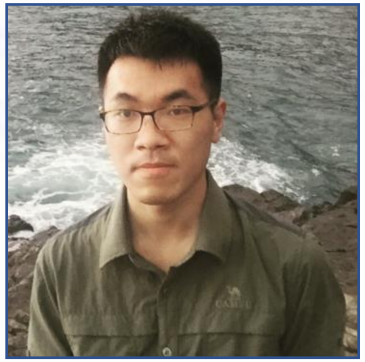
DUSt3R: Geometric 3D Vision Made Easy

CVPR 2024





Shuzhe Wang
Aalto University

Vincent Leroy
Naverlabs Europe

Yohann Cabon
Naverlabs Europe

Boris Chidlovskii
Naverlabs Europe

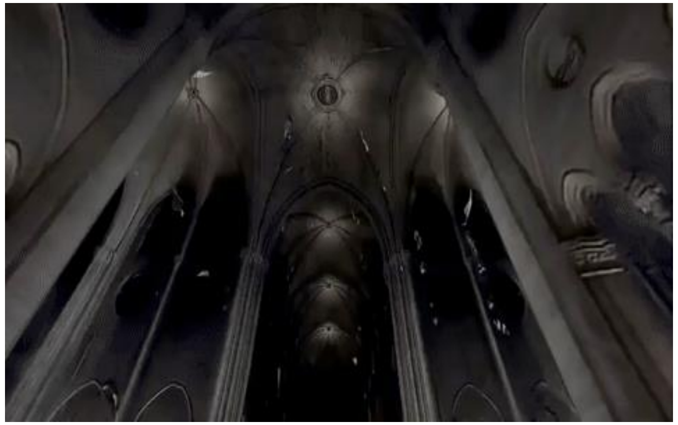
Jérome Revaud
Naverlabs Europe

Aalto University

NAVER LABS
Europe

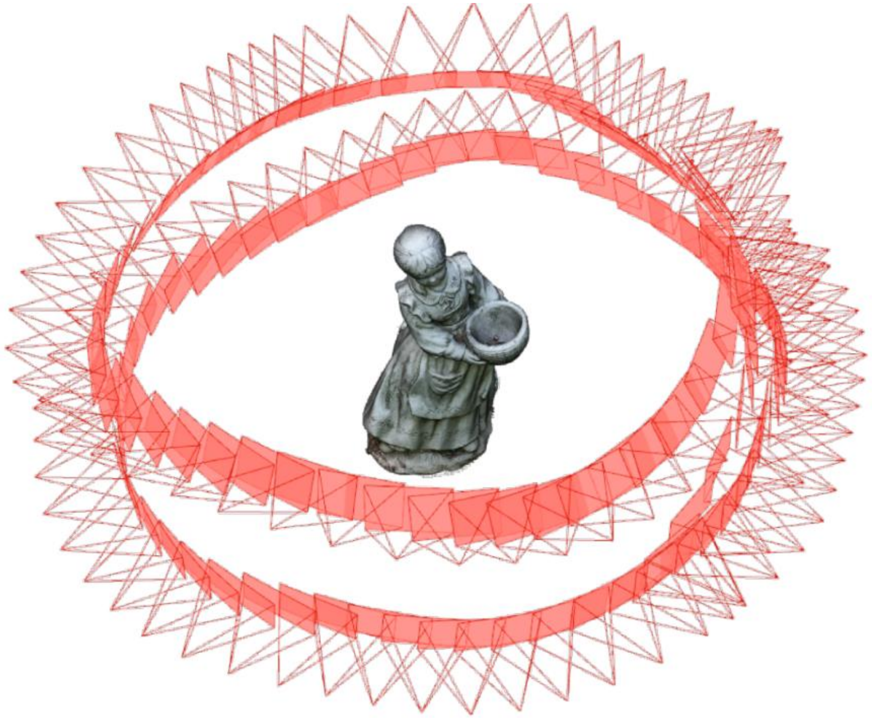
- **3D Dense Reconstruction**

  A key building block in many computer vision and real-world applications.



- **Conventional solution: Multi-view Stereo (MVS)**

  - Given a set of posed and calibrated images of a scene, the target is to reconstruct a dense 3D representation of the scene.

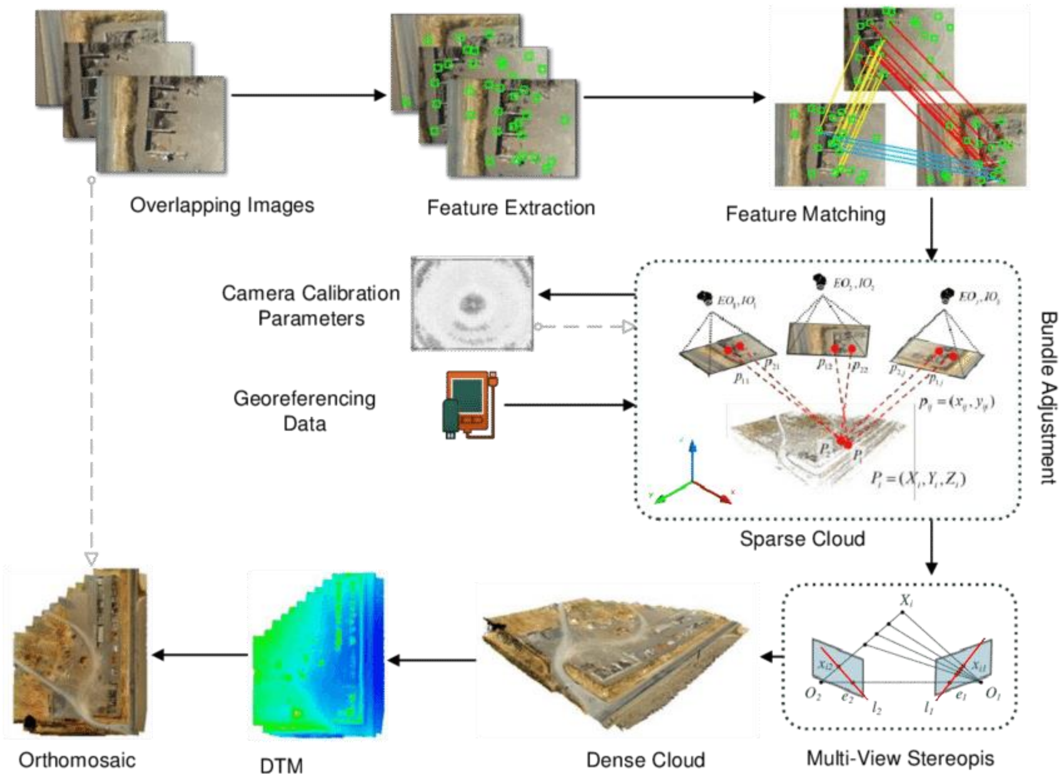


What if the camera parameters (intrinsics, extrinsics) are unknown or

MVS in the wild?

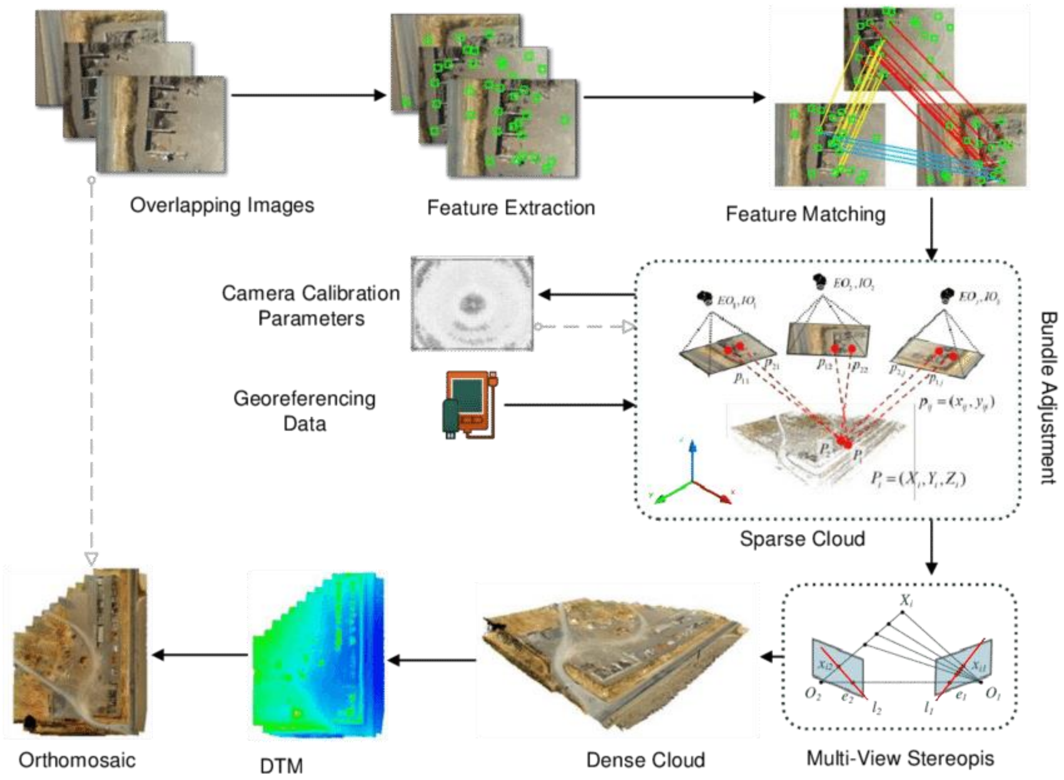
What if the camera parameters (intrinsics, extrinsics) are unknown or MVS in the wild?

- SfM: keypoint detection, description, matching, pose estimation, triangulation, bundle adjustment …
- MVS: per pixel depth, normal map, stereo image rectification …



What if the camera parameters (intrinsics, extrinsics) are unknown or MVS in the wild?

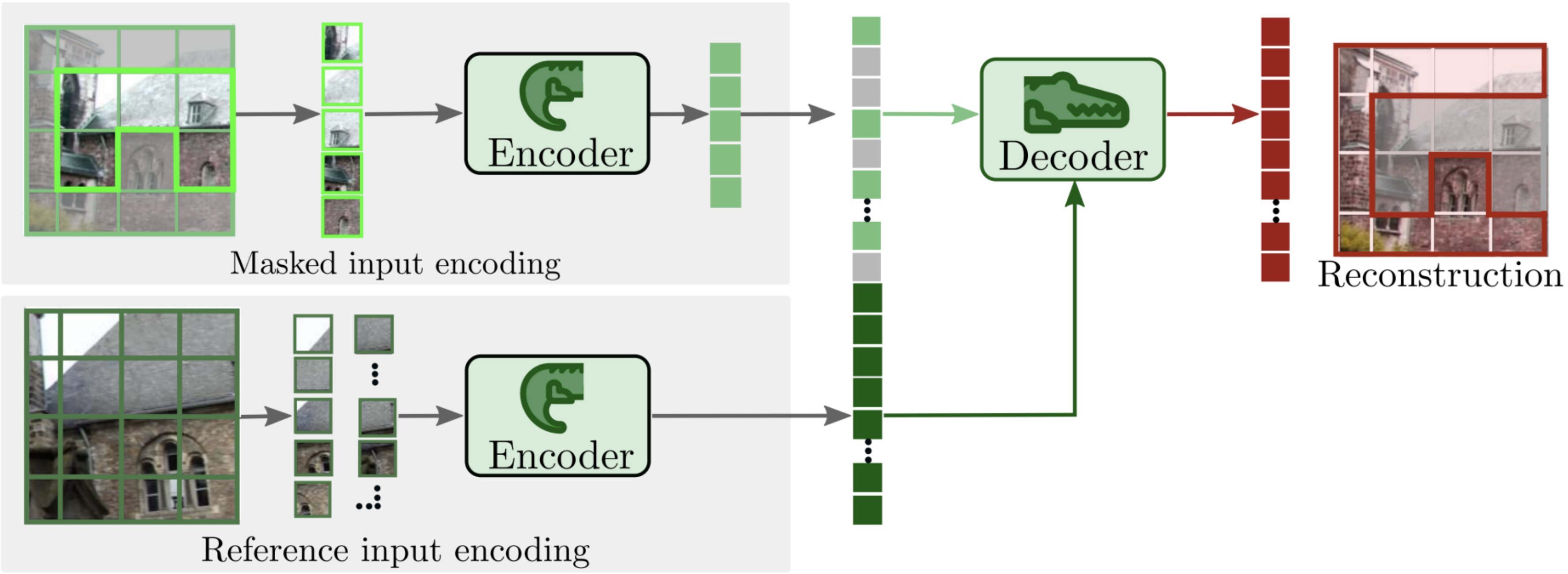
- SfM: keypoint detection, description, matching, pose estimation, triangulation, bundle adjustment …
- MVS: per pixel depth, normal map, stereo image rectification …



A viable solution, but not elegant.

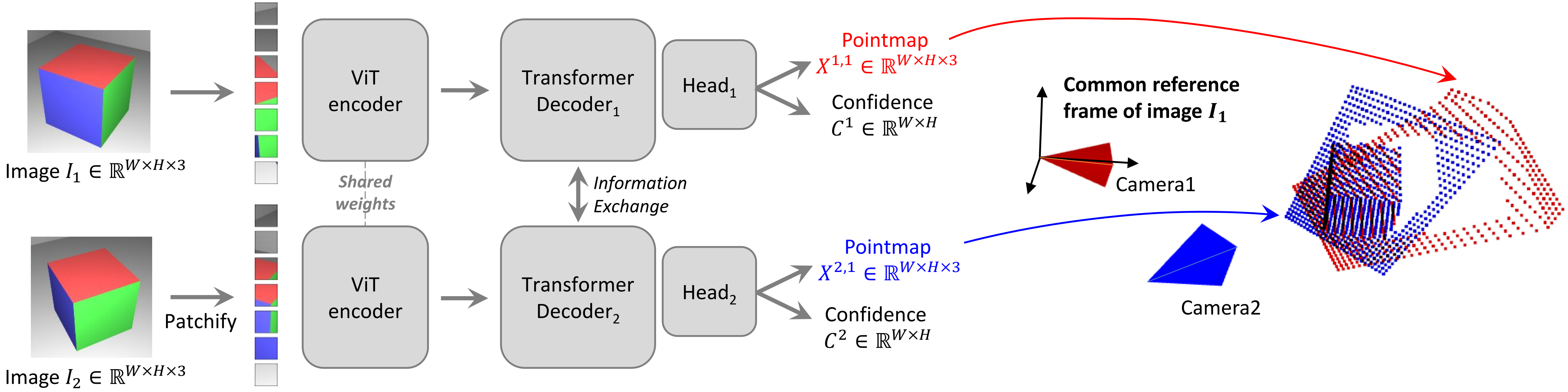
Network Architecture

The architecture is inspired by CroCo [1], a cross-view completion Pre-training pipeline that can understand the spatial relationship between the image pair.



[1] Weinzaepfel et al; CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion (NeurIPS 22)

Network Architecture



Dataset and Training Objective

Datasets	Type	N Pairs
Habitat [103]	Indoor / Synthetic	1000k
CO3Dv2 [93]	Object-centric	941k
ScanNet++ [165]	Indoor / Real	224k
ArkitScenes [25]	Indoor / Real	2040k
Static Thing 3D [68]	Object / Synthetic	337k
MegaDepth [55]	Outdoor / Real	1761k
BlendedMVS [161]	Outdoor / Synthetic	1062k
Waymo [121]	Outdoor / Real	1100k

Dataset and Training Objective:

Datasets	Type	N Pairs
Habitat [103]	Indoor / Synthetic	1000k
CO3Dv2 [93]	Object-centric	941k
ScanNet++ [165]	Indoor / Real	224k
ArkitScenes [25]	Indoor / Real	2040k
Static Thing 3D [68]	Object / Synthetic	337k
MegaDepth [55]	Outdoor / Real	1761k
BlendedMVS [161]	Outdoor / Synthetic	1062k
Waymo [121]	Outdoor / Real	1100k

➔ fully-supervised regression

- We utilize an off-the-shelf image retrieval and point matching algorithm to match and verify training image pairs
- Ground-truth pointmaps are obtained from the ground truth camera intrinsics, camera poses, and depthmap.

Dataset and Training Objective

The regression loss is defined as the Euclidean distance:

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

With $z = \text{norm}(X^{1,1}, X^{2,1})$ and $\bar{z} = \text{norm}(\bar{X}^{1,1}, \bar{X}^{2,1})$ to handle the scale ambiguity between prediction and ground-truth.

Dataset and Training Objective

The regression loss is defined as the Euclidean distance:

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

With $z = \text{norm}(X^{1,1}, X^{2,1})$ and $\bar{z} = \text{norm}(\bar{X}^{1,1}, \bar{X}^{2,1})$ to handle the scale ambiguity between prediction and ground-truth. To handle the ill-defined 3D points (e.g. sky), we extend the regression loss to confidence-aware loss [1].

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

Where $C_i^{v,1} = 1 + \exp \widetilde{C_i^{v,1}} > 1$ to force the network to extrapolate in harder areas.

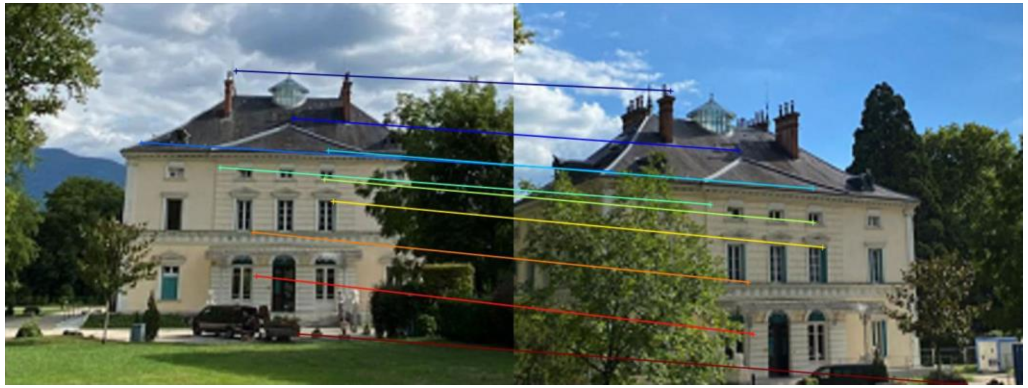
[1] "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics", Kendall et al. CVPR'18

Downstream Applications

1. Point Matching

Achieved by mutual nearest neighbor (MNN) search in the 3D pointmap space.

$$\mathcal{M}_{1,2} = \{(i,j) \mid i = \mathrm{NN}_1^{1,2}(j) \text{ and } j = \mathrm{NN}_1^{2,1}(i)\}$$
$$\text{with } \mathrm{NN}_k^{n,m}(i) = \underset{j \in \{0,\ldots,WH\}}{\arg\min} \left\| X_j^{n,k} - X_i^{m,k} \right\|.$$



Downstream Applications

2. Recovering intrinsics

The pointmap is expressed in the first image coordinate frame (Extrinsic as identical matrix), and we assume that the principal point is approximately centered. We only need to estimate the focal lengths by minimize:

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

Method	Habitat	BlendedMVS	CO3D
Monocular	4.13° / 98.3%	3.40° / 99.4%	1.88° / 97.8%
Binocular	2.09° / 95.2%	2.61° / 98.4%	1.62° / 97.7%

Left: Average absolute error of field-of-view (FoV) estimates.
Right: Average 2D reprojection accuracy (%) at the threshold of 1% of image diagonal.

Downstream Applications

3. Visual Localization

Given a **query image** and **retrieved database image**, the task can be achieved by :

(1) First build the pixel correspondences from point matching, which in turn yields 2D-3D correspondences. The camera pose is solved by the PnP-RANSAC with the estimated intrinsic.

(2) Estimate the relative pose by point matching, convert the pose to world coordinate by scaling (scale factor obtain from the predicted pointmap and ground truth pointmap of the database image)

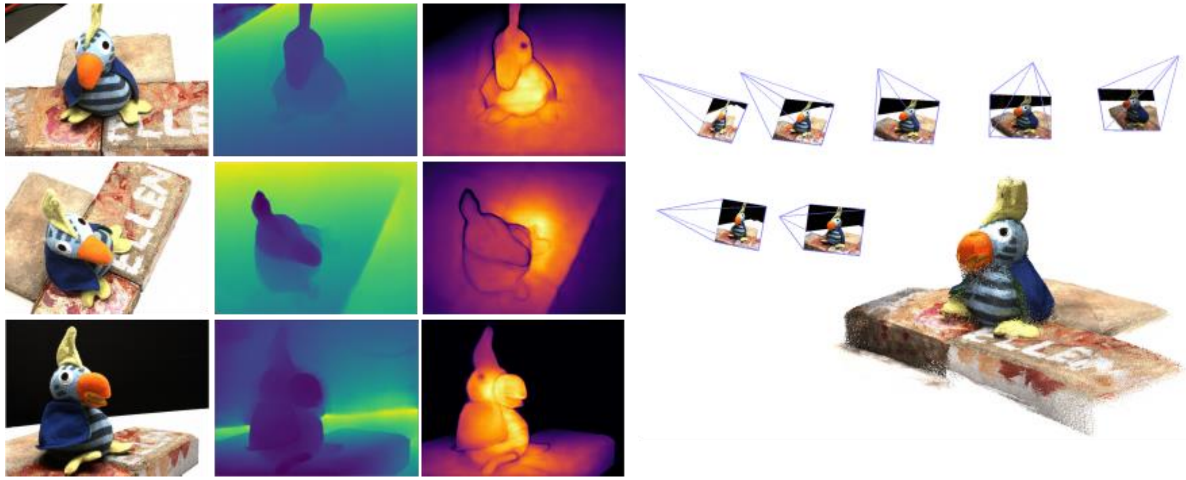
	Methods	7Scenes (Indoor) [113]							Cambridge (Outdoor) [48]				
		Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	S. Facade	O. Hospital	K. College	St.Mary's	G. Court
FM	AS [102]	4/1.96	3/1.53	2/1.45	9/3.61	8/3.10	7/3.37	**3**/2.22	**4**/0.21	20/0.36	13/0.22	8/0.25	24/**0.13**
FM	HLoc [100]	**2/0.79**	**2/0.87**	**2/0.92**	**3/0.91**	**5/1.12**	**4/1.25**	6/**1.62**	**4/0.2**	**15/0.3**	**12/0.20**	**7/0.21**	**11**/0.16
E2E	DSAC* [11]	**2**/1.10	**2**/1.24	**1**/1.82	**3**/1.15	**4**/1.34	4/1.68	**3**/1.16	**5**/0.3	**15/0.3**	15/0.3	13/0.4	49/0.3
E2E	HSCNet [54]	**2/0.7**	**2**/0.9	**1**/0.9	**3/0.8**	**4/1.0**	4/**1.2**	**3/0.8**	6/0.3	19/**0.3**	18/0.3	9/0.3	28/0.2
E2E	PixLoc [101]	**2**/0/80	**2/0.73**	**1/0.82**	**3**/0.82	**4**/1.21	**3/1.20**	5/1.30	**5/0.23**	16/0.32	14/0.24	10/0.34	30/0.14
E2E	SC-wLS [151]	3/0.76	5/1.09	3/1.92	6/0.86	8/1.27	9/1.43	12/2.80	11/0.7	42/1.7	14/0.6	39/1.3	164/0.9
E2E	NeuMaps [124]	**2**/0.81	3/1.11	2/1.17	**3**/0.98	**4**/1.11	4/1.33	4/1.12	6/0.25	19/0.36	14/0.19	17/0.53	**6/ 0.10**
	DUSt3R 224-NoCroCo	5/1.76	6/2.02	3/1.75	5/1.54	9/2.35	6/1.82	34/7.81	24/1.33	79/1.17	69/1.15	46/1.51	143/1.32
	DUSt3R 224	3/0.96	3/1.02	**1**/1.00	4/1.04	5/1.26	4/1.36	21/4.08	9/0.38	26/0.46	20/0.32	11/0.38	36/0.24
	DUSt3R 512	3/0.97	3/0.95	2/1.37	**3**/1.01	**4**/1.14	4/1.34	11/2.84	6/0.26	17/0.33	**11/0.20**	**7/0.24**	38/0.16

Downstream Applications

4. Multi-view Pose Estimation

(1) Obtained with relative pose estimation;

(2) Extract the pairwise camera poses from global alignment.



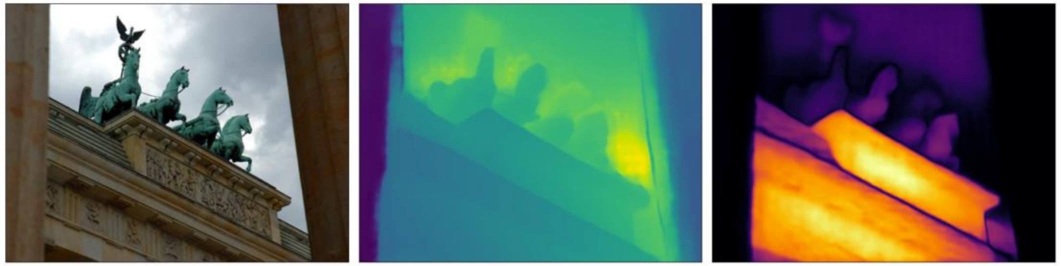
Methods	N Frames	Co3Dv2 [93] RRA@15	Co3Dv2 [93] RTA@15	Co3Dv2 [93] mAA(30)	RealEstate10K [185] mAA(30)
COLMAP+SPSG	3	∼22	∼14	∼15	∼23
PixSfM	3	∼18	∼8	∼10	∼17
Relpose	3	∼56	-	-	-
PoseDiffusion	3	∼75	∼75	∼61	- (∼77)
DUSt3R 512	3	95.3	**88.3**	**77.5**	**69.5**
COLMAP+SPSG	5	∼21	∼17	∼17	∼34
PixSfM	5	∼21	∼16	∼15	∼30
Relpose	5	∼56	-	-	-
PoseDiffusion	5	∼77	∼76	∼63	- (∼78)
DUSt3R 512	5	95.5	86.7	76.5	67.4
COLMAP+SPSG	10	31.6	27.3	25.3	45.2
PixSfM	10	33.7	32.9	30.1	49.4
Relpose	10	57.1	-	-	-
PoseDiffusion	10	80.5	79.8	66.5	48.0 (∼80)
DUSt3R 512	10	**96.2**	86.8	76.7	67.7

Downstream Applications

4. Mono Depth Estimation

For the first frame, We have $D^1_{i,j} = \bar{X}^{1,1}_{i,j,2}$.

Methods	Train	Outdoor						Indoor			
		DDAD[33]		KITTI [29]		BONN [62]		NYUD-v2 [92]		TUM [94]	
		Rel↓	$\delta_{1.25}$ ↑	Rel↓	$\delta_{1.25}$ ↑	Rel↓	$\delta_{1.25}$ ↑	Rel↓	$\delta_{1.25}$ ↑	Rel ↓	$\delta_{1.25}$ ↑
DPT-BEiT[71]	D	10.70	**84.63**	9.45	89.27	-	-	**5.40**	**96.54**	**10.45**	**89.68**
NeWCRFs[139]	D	**9.59**	82.92	**5.43**	**91.54**	-	-	6.22	95.58	14.63	82.95
Monodepth2 [31]	SS	23.91	75.22	11.42	86.90	56.49	35.18	16.19	74.50	31.20	47.42
SC-SfM-Learners [5]	SS	16.92	77.28	11.83	86.61	21.11	71.40	13.79	79.57	22.29	64.30
SC-DepthV3 [96]	SS	**14.20**	**81.27**	11.79	86.39	**12.58**	**88.92**	**12.34**	**84.80**	**16.28**	**79.67**
MonoViT[145]	SS	-	-	**09.92**	**90.01**	-	-	-	-	-	
RobustMIX [72]	T	-	-	18.25	76.95	-	-	11.77	90.45	15.65	**86.59**
SlowTv [93]	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



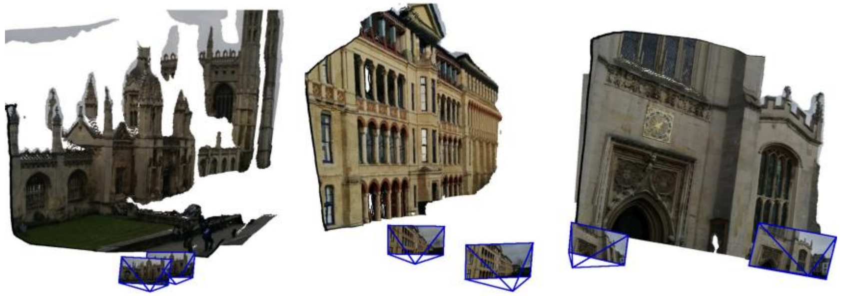
5. Multi-view Depth

	Methods	GT Pose	GT Range	Align	KITTI		ScanNet		ETH3D		DTU		T&T		Average		
					rel ↓	τ ↑	rel ↓	τ ↑	rel ↓	τ ↑	rel ↓	τ ↑	rel↓	τ ↑	rel↓	τ ↑	time (s)↓
(a)	COLMAP [84, 85]	✓	×	×	**12.0**	**58.2**	**14.6**	**34.2**	**16.4**	**55.1**	**0.7**	**96.5**	**2.7**	95.0	**9.3**	**67.8**	$\approx 3min$
	COLMAP Dense [84, 85]	✓	×	×	26.9	52.7	38.0	22.5	89.8	23.2	20.8	69.3	25.7	76.4	40.2	48.8	$\approx 3min$
(b)	MVSNet [129]	✓	✓	×	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 [129]	✓	✓	×	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 [141]	✓	✓	×	**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 [128]	✓	✓	×	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 [128]	✓	✓	×	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 [107]	✓	×	×	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 [103]	✓	×	×	(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 [103]	✓	×	×	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 [129]	✓	×	×	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 [129]	✓	×	×	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 [141]	✓	×	×	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 [128]	✓	×	×	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 [128]	✓	×	×	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 [88]	✓	×	×	**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 [107]	×	×	$\lVert \mathbf{t} \rVert$	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 [103]	×	×	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 [103]	×	×	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

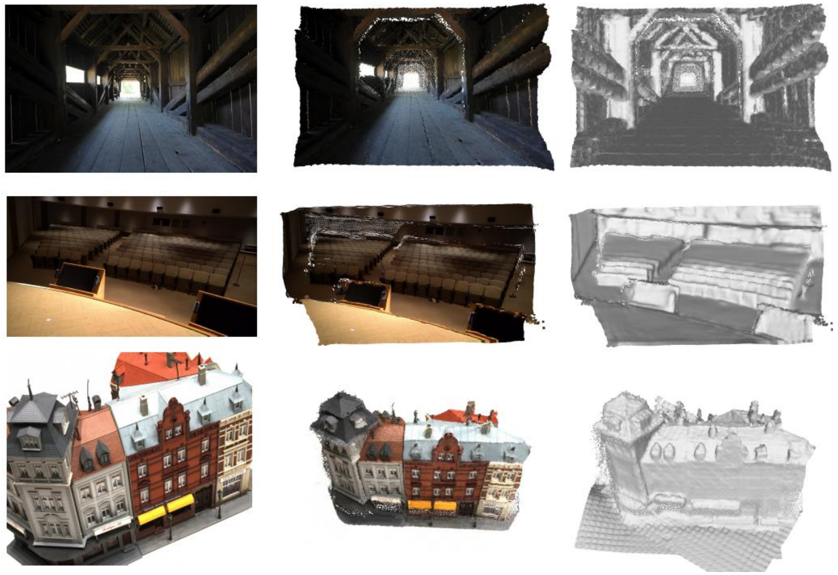
Downstream Applications

6. 3D Reconstruction

Two views

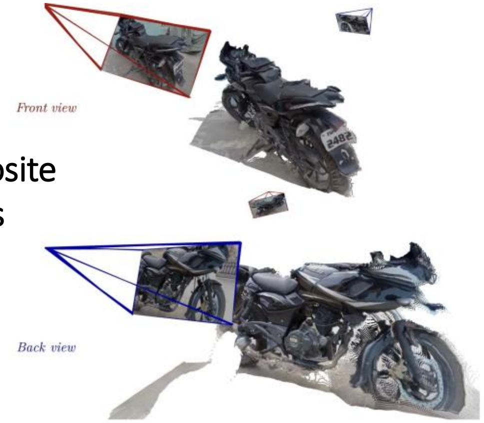


Dense Reconstruction



	Methods	GT cams	Acc.↓	Comp.↓	Overall↓
(a)	Camp [10]	✓	0.835	0.554	0.695
	Furu [27]	✓	0.613	0.941	0.777
	Tola [105]	✓	0.342	1.190	0.766
	Gipuma [28]	✓	**0.283**	0.873	0.578
(b)	MVSNet [129]	✓	0.396	0.527	0.462
	CVP-MVSNet [126]	✓	0.296	0.406	0.351
	UCS-Net [15]	✓	0.338	0.349	0.344
	CER-MVS [52]	✓	0.359	0.305	0.332
	CIDER [125]	✓	0.417	0.437	0.427
	PatchmatchNet [109]	✓	0.427	0.277	0.352
	GeoMVSNet [143]	✓	0.331	**0.259**	**0.295**
	DUSt3R 512	×	2.677	0.805	1.741

Opposite views



No overlap

More visualization

Conclusions

1. We present the first holistic end-to-end 3D reconstruction pipeline from un-calibrated and un-posed images.

2. We introduce the pointmap representation for MVS applications, that enables the network to predict the 3D points, while preserving the implicit relationship between pixels and the scene.

3. We introduce an optimization procedure to globally align pointmaps in the context of multi-view 3D reconstruction. Our procedure can extract effortlessly all usual intermediary outputs of the classical SfM and MVS pipelines.

4. We demonstrate promising performance on a range of 3D vision tasks In particular, our all-in-one model achieves state-of-the-art results on monocular and multi-view depth benchmarks, as well as multi-view camera pose estimation.

Thank You!

Code and model are available!