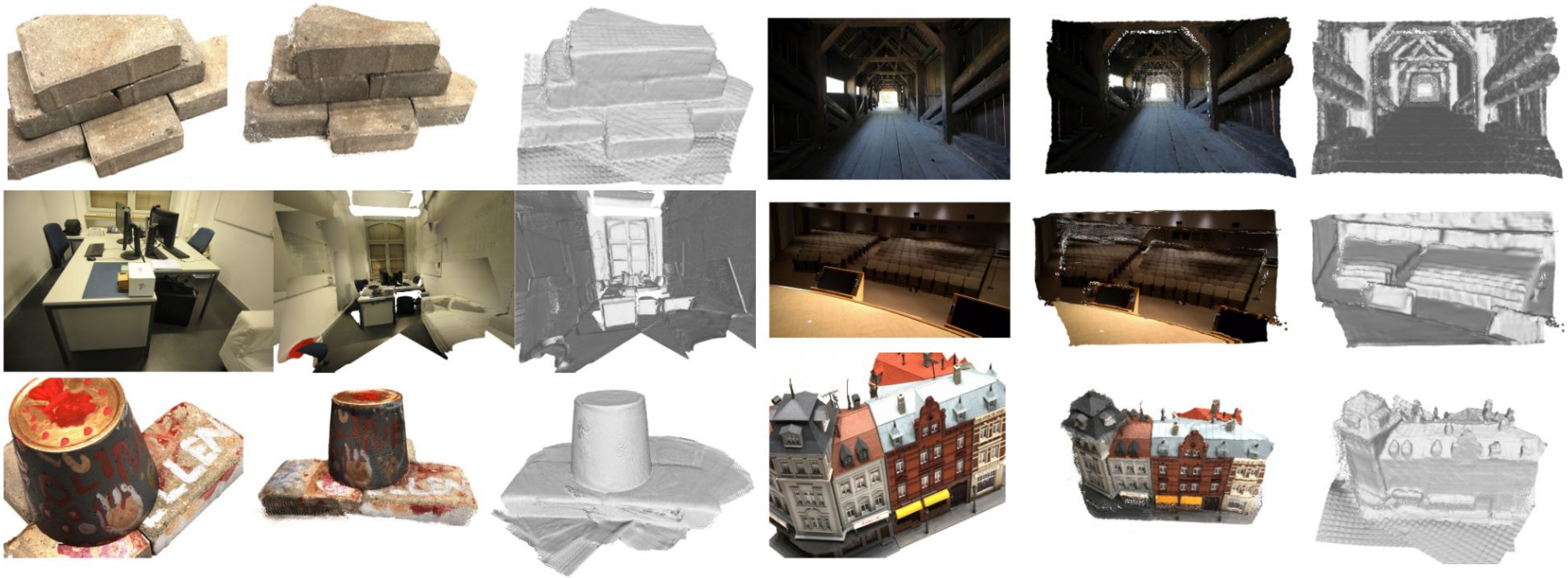
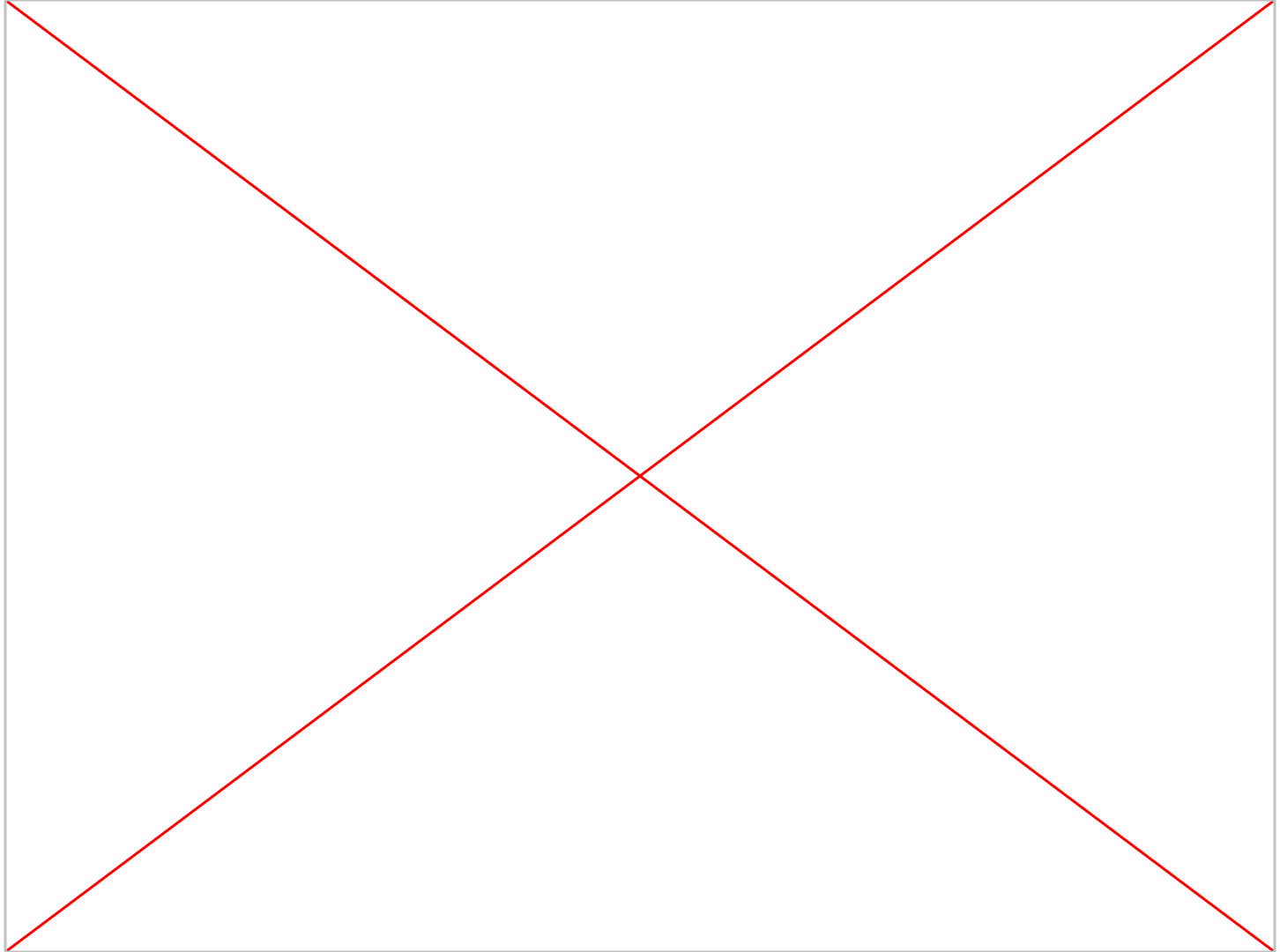
DUSt3R: Geometric 3D Vision Made Easy

Shuzhe Wang*, Vincent Leroy†, Yohann Cabon†, Boris Chidlovskii† and Jerome Revaud†
*Aalto University †Naver Labs Europe

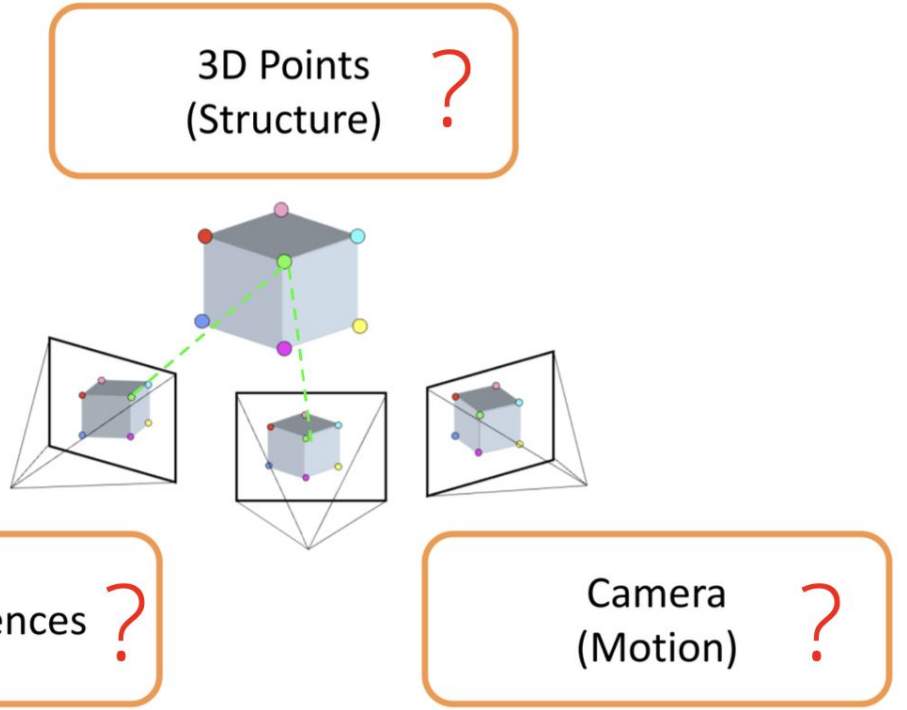




Camera Pose, Correspondence, and 3D Shape are Unknown

Solution: Learn to map each pixel to a 3D location among two views.

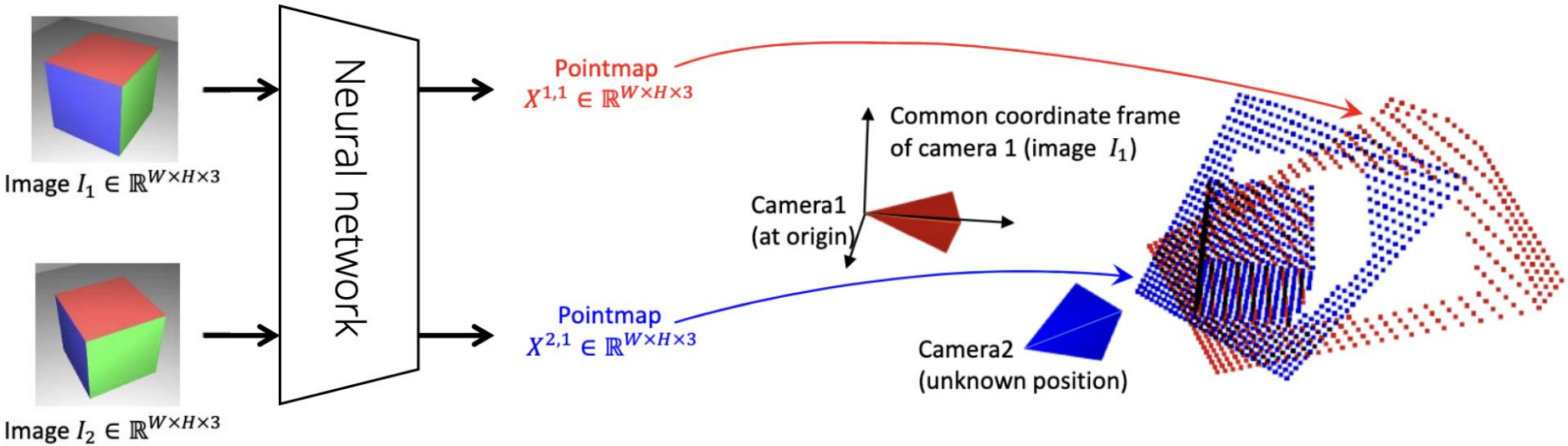
Correspondence emerges from the 3D location of every pixel!



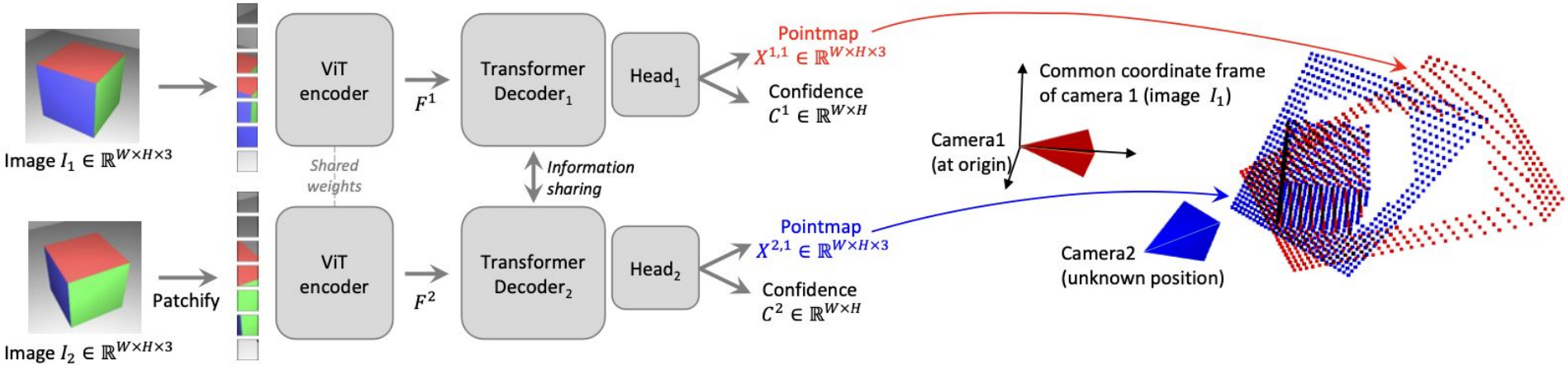
Problem Statement of DUSt3R

- Multi-View Stereo Reconstruction
  - Traditional MVS
    - Requires intrinsic and extrinsic parameters
    - Sensitive to feature matching
    - Optimizes with Bundle Adjustment
  - DUSt3R
    - No need for intrinsic and extrinsic parameters
    - Works with arbitrary image collections
    - Utilizes a Transformer-based architecture

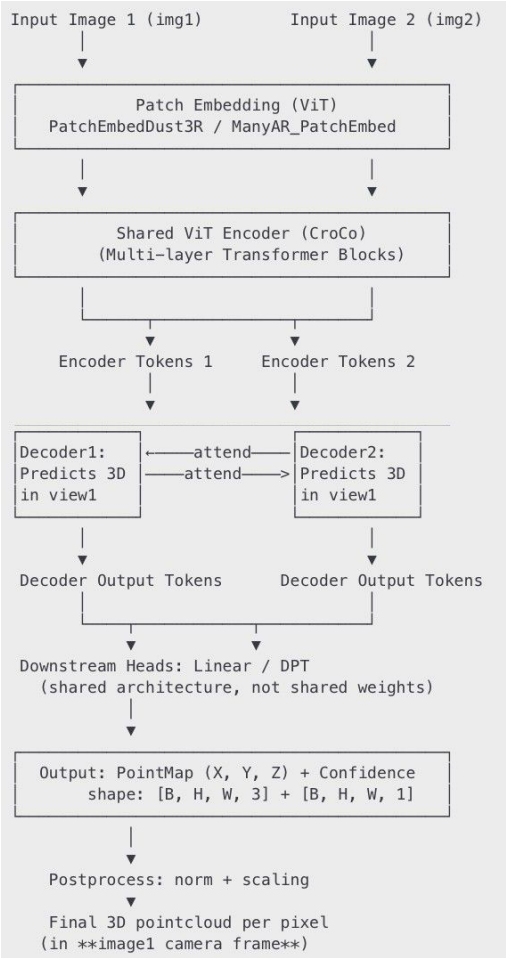
DUSt3R: Dense and Unconstrained Stereo 3D Reconstruction



Network Architecture



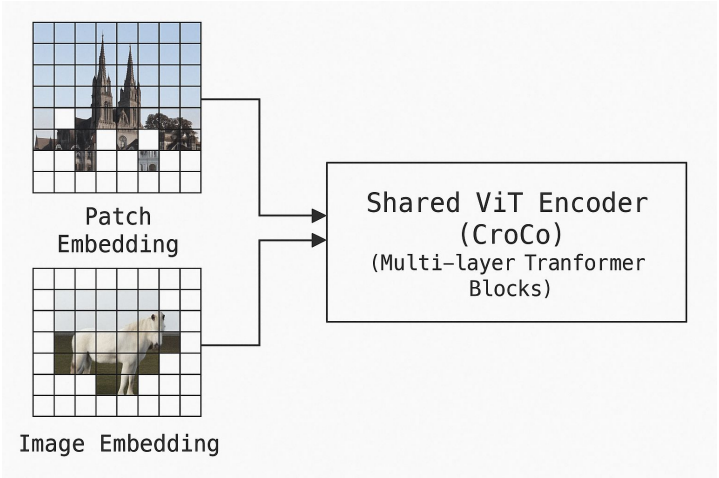
Pipeline

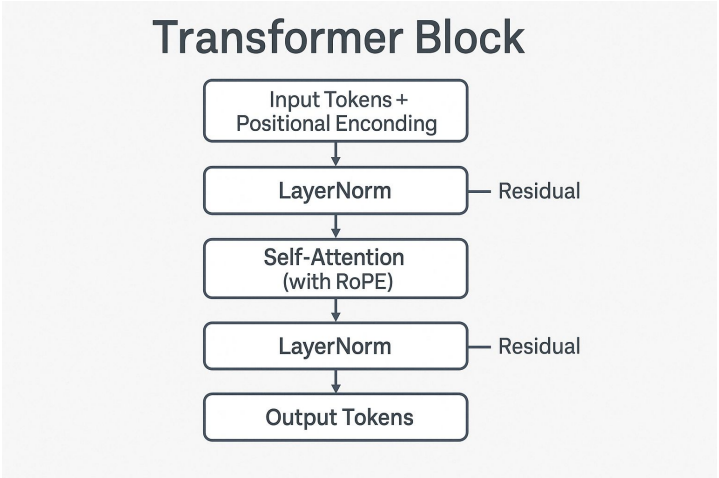

Input Image 1 (img1)
Input Image 2 (img2)
Patch Embedding (ViT)
PatchEmbedDust3R / ManyAR_PatchEmbed
Shared ViT Encoder (CroCo)
(Multi-layer Transformer Blocks)
Encoder Tokens 1
Encoder Tokens 2
Decoder1:
Predicts 3D
in view1
attend
attend
Decoder2:
Predicts 3D
in view1
Decoder Output Tokens
Decoder Output Tokens
Downstream Heads: Linear / DPT
(shared architecture, not shared weights)
Output: PointMap (X, Y, Z) + Confidence
shape: [B, H, W, 3] + [B, H, W, 1]
Postprocess: norm + scaling
Final 3D pointcloud per pixel
(in **image1 camera frame**)

Shared ViT Encoder (CroCo)

- **CroCo Pretraining: Enabling Cross-View 3D Understanding**
  - 從一張圖預測另一視角的遮蔽部分
    - geometric consistency
    - object-aware spatial layout
    - cross-view mapping

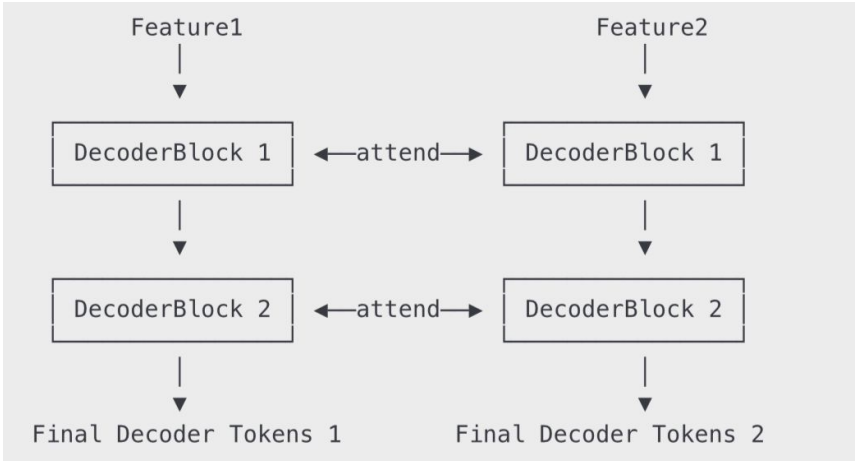
- **Why use CroCo in DUSt3R?**
  - 讓 DUSt3R 不需 fine-tune 就能做 3D matching
  - 對遮蔽區域也能建立有效的幾何推論





Decoder & Head

1. ViT decoder token layers（多層輸出）
2. 抽出多層 token（如 layer 4/8/12/16）
3. 轉為空間格式 → reshape [B, N, C] → [B, C, H, W]
4. 經過 UNet-style Refinement（RefineNet4 → 3 → 2 → 1）
5. 最後才進入 Output Head（1x1 conv → 輸出 XYZ+conf）



Training Objective

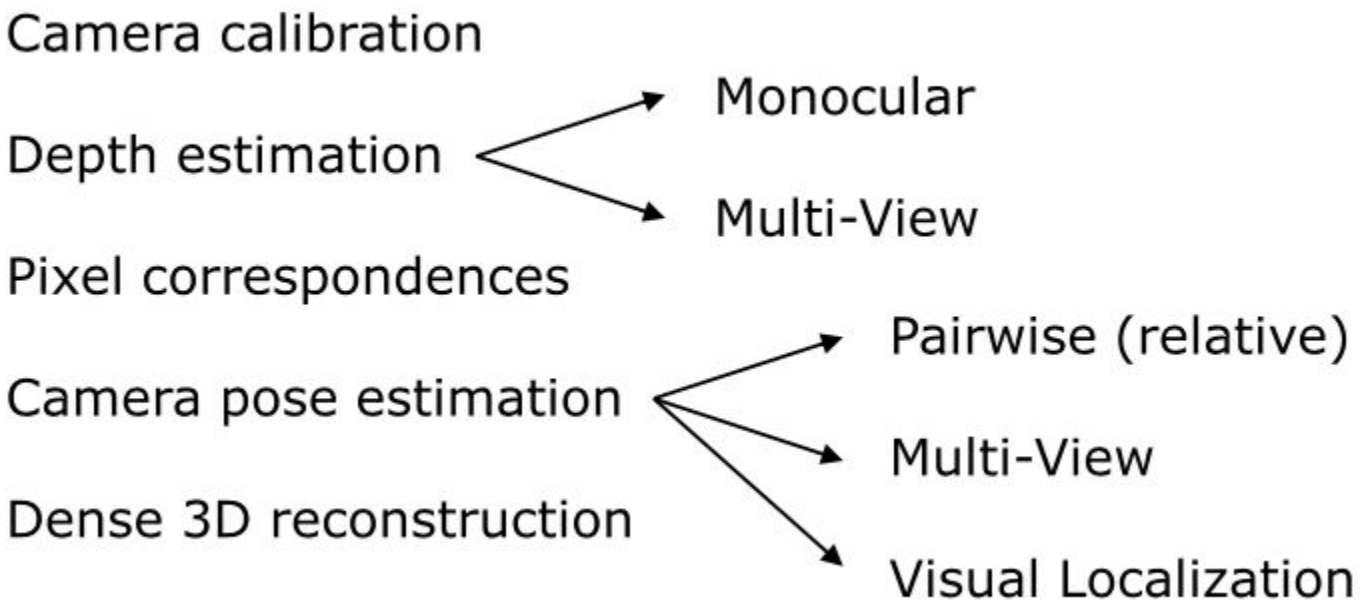
- $$\mathcal{L}_{\text{conf}} = \sum_{v \in \{1,2\}} \sum_{i \in \mathcal{D}^v} C_i^{v,1} \ell_{\text{regr}}(v,i) - \alpha \log C_i^{v,1}$$
  - $$\ell_{\text{regr}}(v,i) = \left\| \frac{1}{z} X_i^{v,1} - \frac{1}{\bar{z}} \bar{X}_i^{v,1} \right\|$$
  - $$\text{norm}(X^1, X^2) = \frac{1}{|\mathcal{D}^1| + |\mathcal{D}^2|} \sum_{v \in \{1,2\}} \sum_{i \in \mathcal{D}^v} \|X_i^v\|$$
  - $$C_i^{v,1} = 1 + \exp \widetilde{C_i^{v,1}} > 1$$

Global Alignment

- Pairwise graph
  - 每張image為一個vertices(v)
  - 每個edge(e) 代表2張image間有相同的visual content

- Global optimization
  - Recover globally aligned point maps
  - $$\chi^* = \arg\min_{\chi, P, \sigma} \sum_{e \in \mathcal{E}} \sum_{v \in e} \sum_{i=1}^{HW} C_i^{v,e} \left\| \chi_i^v - \sigma_e P_e X_i^{v,e} \right\|$$

Downstream Application



Experiments with DUSt3R

- Training Data: Habitat, MegaDepth, ARK- itScenes, MegaDepth, Static Scenes 3D, Blended MVS, ScanNet++ , CO3D-v2, Waymo
  - 8.5M image pairs

- Image size
  - 224 x 224 → 512 x 512

Result - Absolute camera pose

	Methods	7Scenes (Indoor) [114]							Cambridge (Outdoor) [49]				
		Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	S. Facade	O. Hospital	K. College	St.Mary's	G. Court
FM	AS [103]	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 [101]	**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 [55]	**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 [102]	**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 [152]	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 [125]	**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

Table 1. Absolute camera pose on 7Scenes [114] and Cambridge-Landmarks [49] datasets. We report the median translation and rotation errors ($cm/^\circ$) to feature matching (FM) based and end-to-end (E2E) learning-base methods. The best results at each category are in **bold.**

Result - Monocular depth estimation & Multi-view pose regression

Methods	Train	Outdoor						Indoor			
		DDAD[41]		KITTI [35]		BONN [80]		NYUD-v2 [115]		TUM [119]	
		Rel↓	$\delta_{1.25}$ ↑	Rel↓	$\delta_{1.25}$ ↑	Rel↓	$\delta_{1.25}$ ↑	Rel↓	$\delta_{1.25}$ ↑	Rel ↓	$\delta_{1.25}$ ↑
DPT-BEiT[91]	D	10.70	**84.63**	9.45	89.27	-	-	**5.40**	**96.54**	**10.45**	**89.68**
NeWCRFs[174]	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 [121]	SS	**14.20**	**81.27**	11.79	86.39	**12.58**	**88.92**	**12.34**	**84.80**	**16.28**	**79.67**
MonoViT[182]	SS	-	-	**09.92**	**90.01**	-	-	-	-	-	
RobustMIX [92]	T	-	-	18.25	76.95	-	-	11.77	90.45	15.65	**86.59**
SlowTv [117]	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

Methods	Co3Dv2 [94]			RealEstate10K
	RRA@15	RTA@15	mAA(30)	mAA(30)
RelPose [177]	57.1	-	-	-
Colmap+SPSG [26, 100]	36.1	27.3	25.3	45.2
PixSfM [59]	33.7	32.9	30.1	49.4
PosReg [140]	53.2	49.1	45.0	-
PoseDiffusion [140]	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**

Table 2. **Left**: Monocular depth estimation on multiple benchmarks. D-Supervised, SS-Self-supervised, T-transfer (zero-shot). (Parentheses) refers to training on the same set. **Right:** Multi-view pose regression on the CO3Dv2 [94] and RealEst10K [186] with 10 random frames.

Result - Multi-view depth evaluation

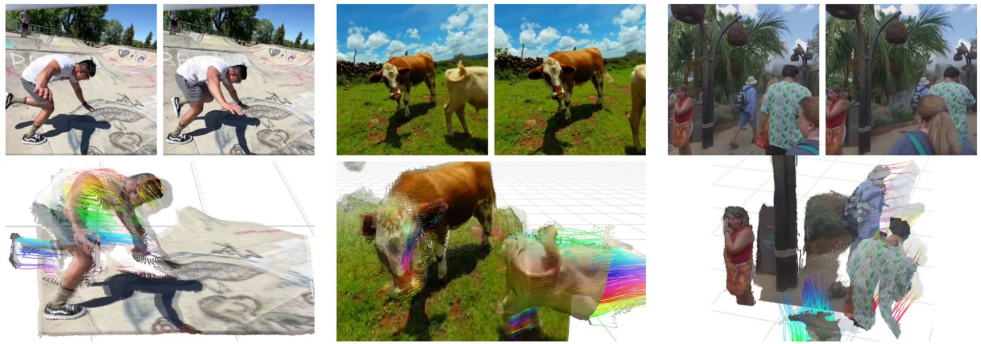
	Methods	GT Pose	GT Range	GT Intrinsics	Align	KITTI rel ↓	KITTI τ ↑	ScanNet rel ↓	ScanNet τ ↑	ETH3D rel ↓	ETH3D τ ↑	DTU rel ↓	DTU τ ↑	T&T rel↓	T&T τ ↑	Average rel↓	Average τ ↑	Average time (s)↓
(a)	COLMAP [106, 107]	✓	×	✓	×	**12.0**	**58.2**	**14.6**	**34.2**	**16.4**	**55.1**	**0.7**	**96.5**	**2.7**	95.0	**9.3**	**67.8**	≈ 3 min
	COLMAP Dense [106, 107]	✓	×	✓	×	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 [161]	✓	✓	✓	×	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 [161]	✓	✓	✓	×	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 [176]	✓	✓	✓	×	**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 [160]	✓	✓	✓	×	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 [160]	✓	✓	✓	×	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 [136]	✓	×	✓	×	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 [131]	✓	×	✓	×	(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 [131]	✓	×	✓	×	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 [161]	✓	×	✓	×	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 [161]	✓	×	✓	×	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 [176]	✓	×	✓	×	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 [160]	✓	×	✓	×	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 [160]	✓	×	✓	×	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 [110]	✓	×	✓	×	**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 [136]	×	×	✓	$\|\mathbf{t}\|$	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 [131]	×	×	✓	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 [131]	×	×	✓	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

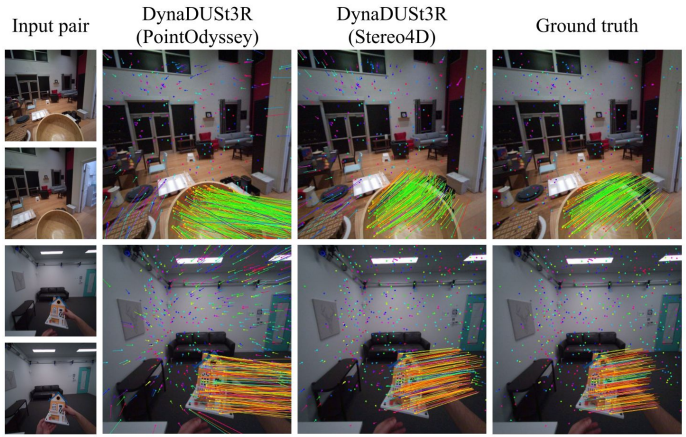
Table 3. **Multi-view depth evaluation** with different settings: a) Classical approaches; b) with poses and depth range, without alignment; c) absolute scale evaluation with poses, without depth range and alignment; d) without poses and depth range, but with alignment. (Parentheses) denote training on data from the same domain. The best results for each setting are in **bold**.

Conclusion

- Generalization without Assumptions
  - **No geometric priors**: DUSt3R **does not rely on camera poses, depth range, epipolar geometry**, or any hand-crafted geometric constraints.
  - 模型能夠直接應用在各種不同場景中，**zero-shot 運作依然穩定** 。
- Pretrained, Not Fine-tuned
  - **Single pretrained model** on diverse data (CroCo-style)
  - Monocular depth estimation, Multi-view stereo, Pose regression, ….
- Trustworthy Confidence
  - 信心調整 loss function，讓模型自己學會忽略難以預測的點

Extended Applications of DUSt3R – DynaDUSt3R





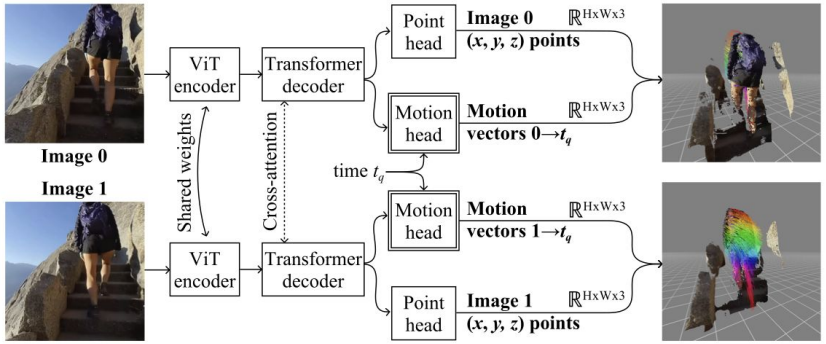
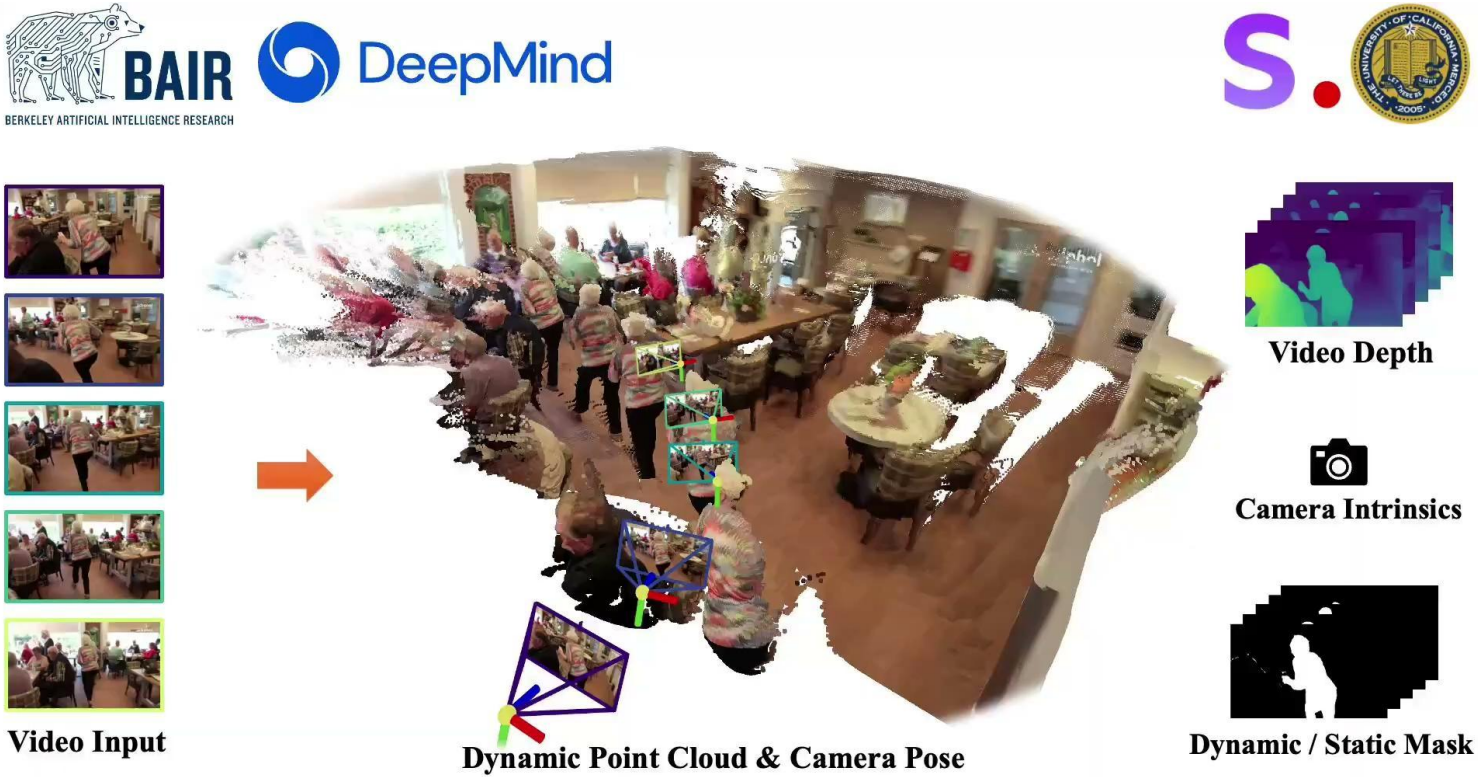


Figure 6. **DynaDUSt3R architecture.** Given two images $(\mathbf{I}_0, \mathbf{I}_1)$ of a dynamic scene and a desired target time t_q, the images are passed through a ViT encoder and transformer decoder. The resulting features are processed by (1) a pointmap head that predicts 3D points in the coordinate frame of $\mathbf{I}_0$, and (2) a 3D motion head that predicts the motion of all points to the target time t_q. A double outline indicates a new component compared to DUSt3R.

MonST3R


BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH
DeepMind
Video Input
Dynamic Point Cloud & Camera Pose
Video Depth
Camera Intrinsics
Dynamic / Static Mask