From CroCo to MASt3R:

A paradigm change in 3D vision?

Jerome Revaud

April 2024

NAVER

NAVER LABS

NAVER LABS
Europe

NAVER LABS Europe

Making robots safe, robust and useful in our everyday spaces.

A multidisciplinary approach to AI.

Vision

Perception to help robots understand and interact with the environment.

Action

Enable embodied agents to efficiently execute tasks and to navigate in dynamic environments.

Interaction

Equip robots to interact safely with humans, other robots and systems.

Overview

Current limits in 3D vision

↓

CroCo: Self-Supervised pretraining for 3DV

↓

DUSt3R: towards a unified 3DV model

↓

MASt3R: grounding matching in 3D

↓

Conclusion & next steps

What is 3D vision?

What is 3D vision?

Monocular Depth estimation

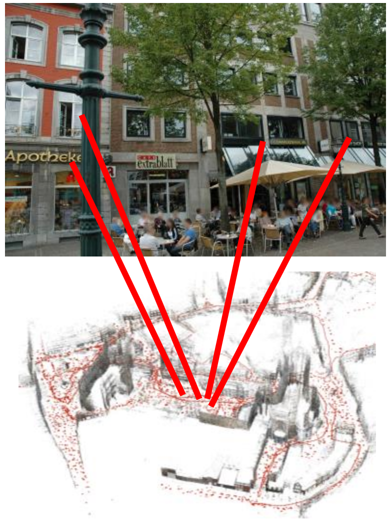
What is 3D vision?

Monocular Depth estimation

Large-scale 3D reconstruction

What is 3D vision?

Monocular Depth estimation

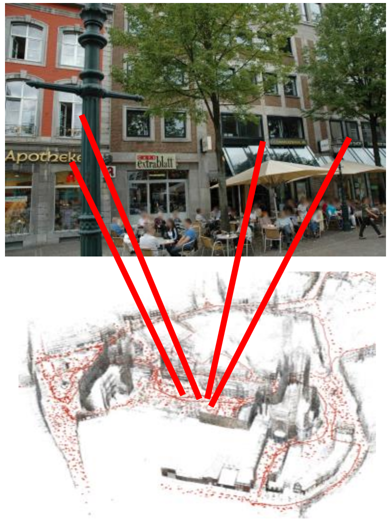


Point matching

Large-scale 3D
reconstruction

What is 3D vision?

Monocular Depth estimation



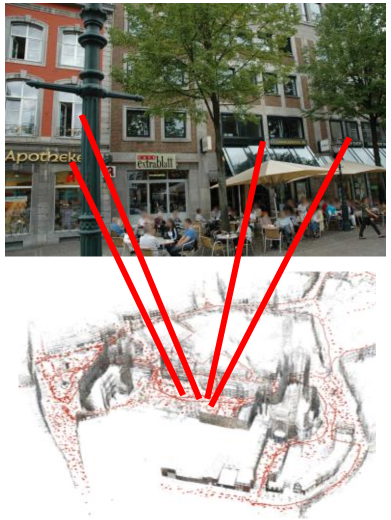
Point matching

Large-scale 3D reconstruction

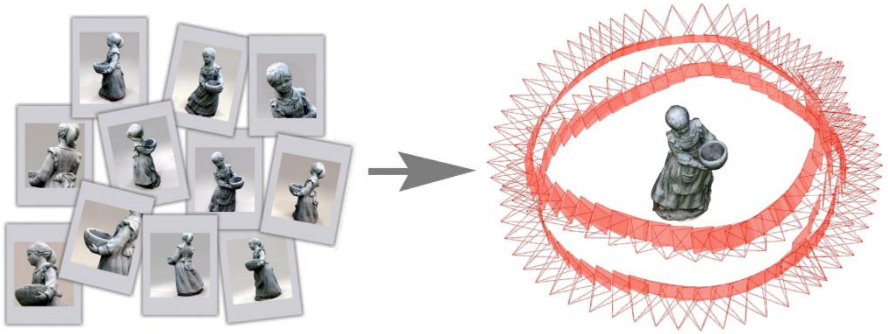
Visual Localization

What is 3D vision?

Monocular Depth estimation



Point matching

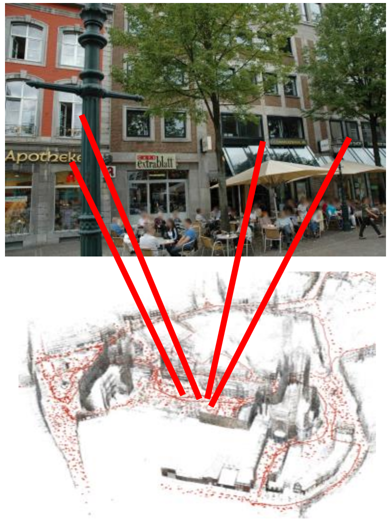


Multi-view pose estimation

Large-scale 3D reconstruction

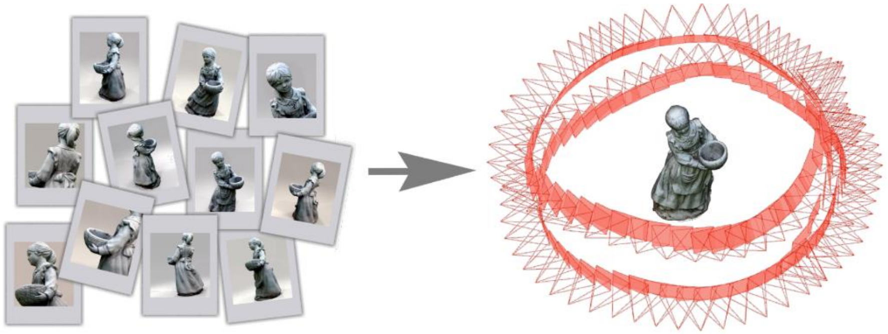
Visual Localization

What is 3D vision?



Point matching

Monocular Depth estimation



Multi-view pose estimation

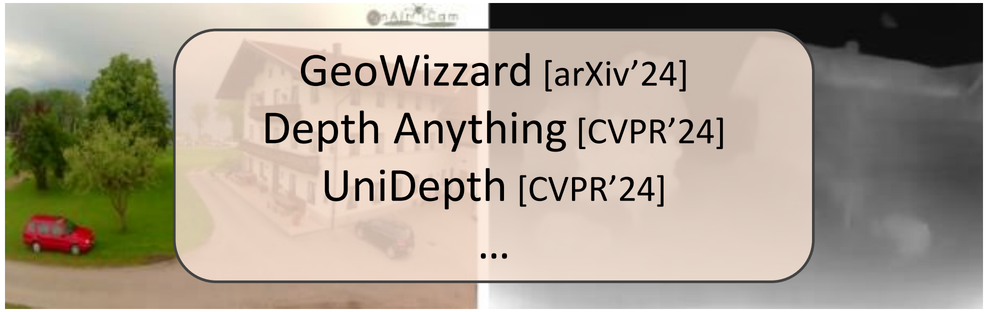
Large-scale 3D reconstruction

Visual Localization

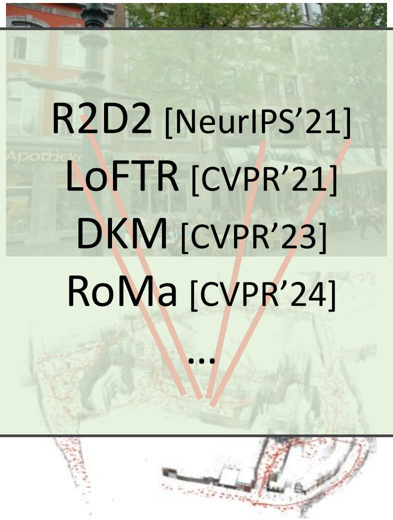
… and many more: SLAM, calibration, MVS, …

What is 3D vision?

Monocular Depth estimation

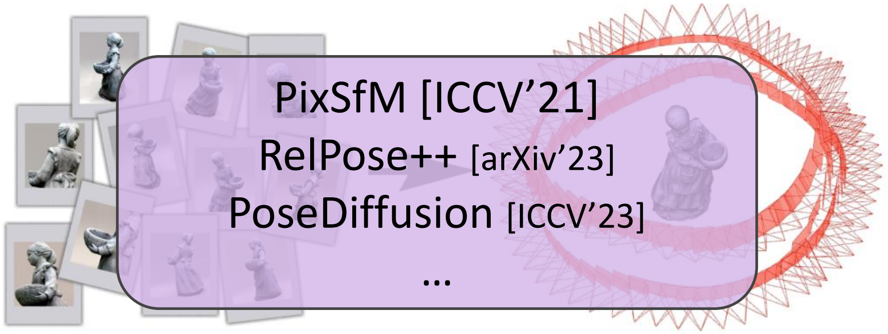


GeoWizzard [arXiv'24]
Depth Anything [CVPR'24]
UniDepth [CVPR'24]
…



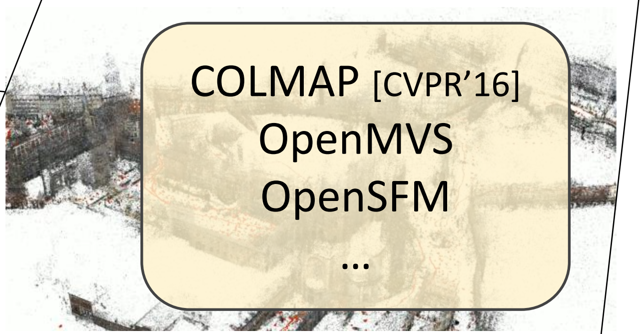
R2D2 [NeurIPS'21]
LoFTR [CVPR'21]
DKM [CVPR'23]
RoMa [CVPR'24]
…

Point matching



PixSfM [ICCV'21]
RelPose++ [arXiv'23]
PoseDiffusion [ICCV'23]
…

Multi-view pose estimation



COLMAP [CVPR'16]
OpenMVS
OpenSFM
…

Large-scale 3D reconstruction

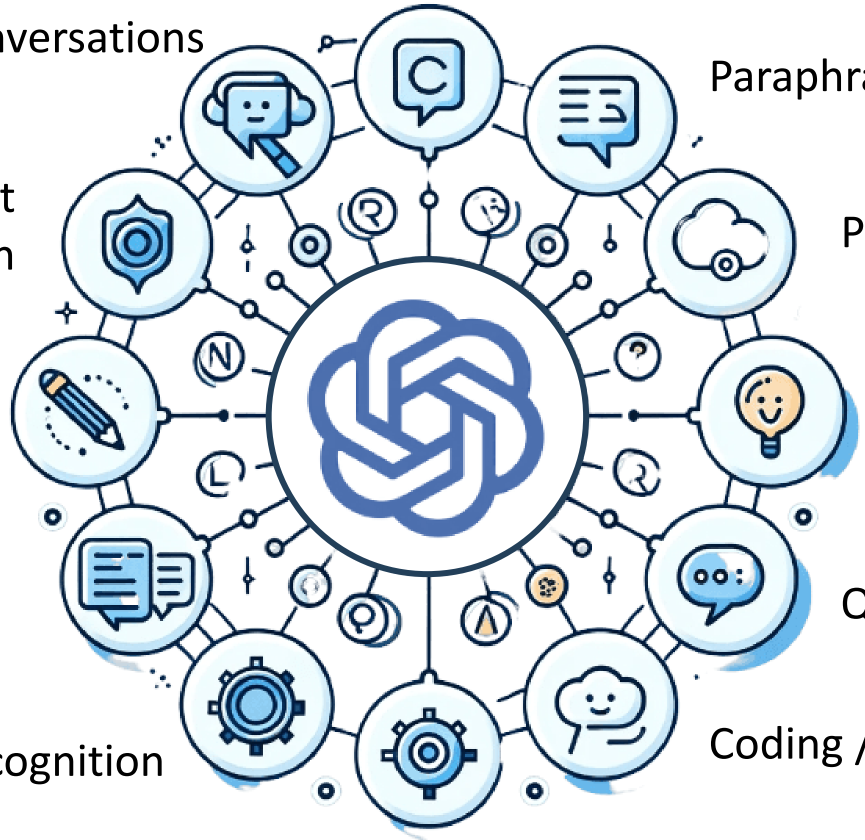
Hloc [CVPR'19]
Kapture [arXiv'20]
PixLoc [CVPR'21]
ACE [CVPR'23]
…

Visual Localization

… and many more: SLAM, calibration, MVS, …

Why seek a unified model?

The NLP case



Why seek a unified model?

- “Foundation models for 3DV”?
  - Weakly-supervised pretext task → useful for many downstream tasks
  - Many definitions, no consensus yet
  - Several CVPR workshops on the question

- Non-exhaustive listing of relevant works
  - “*Scene Representation Transformer: Geometry-Free Novel View Synthesis Through Set-Latent Scene Representations*” [CVPR’22]
  - “*FlowCam: Training Generalizable 3D Radiance Fields without Camera Poses via Pixel-Aligned Scene Flow*” [NeurIPS’23]
  - “*Where are we in the search for an Artificial Visual Cortex for Embodied Intelligence?”* [NeurIPS’23] → FM for robotics
  - “*PonderV2: Pave the Way for 3D Foundation Model with A Universal Pre-training Paradigm*”, [arXiv’23] → mostly semantic tasks
  - “*FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects*” [CVPR’24] → for object pose estimation and tracking
  - “Scalable Pre-training of Large Autoregressive Image Models” [arXiv’24] → LLM for images
  - “*FMGS: Foundation Model Embedded 3D Gaussian Splatting for Holistic 3D Scene Understanding*” [arXiv’24] → DINOv2 with 3DGS
  - “*Probing the 3D Awareness of Visual Foundation Models*” [arXiv’24] → only monocular models, DINOv2 & StableDiffusion work best

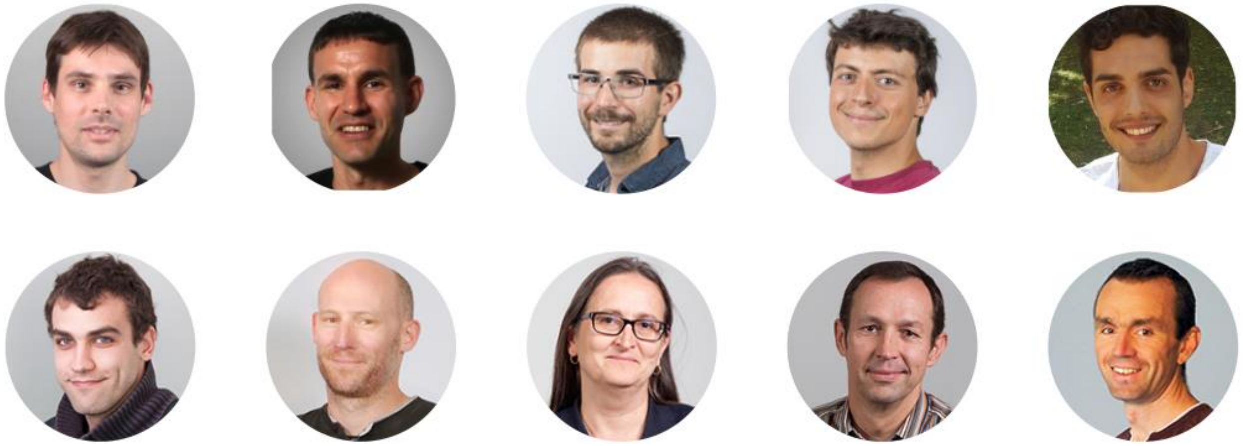
Foundation model for 3D vision

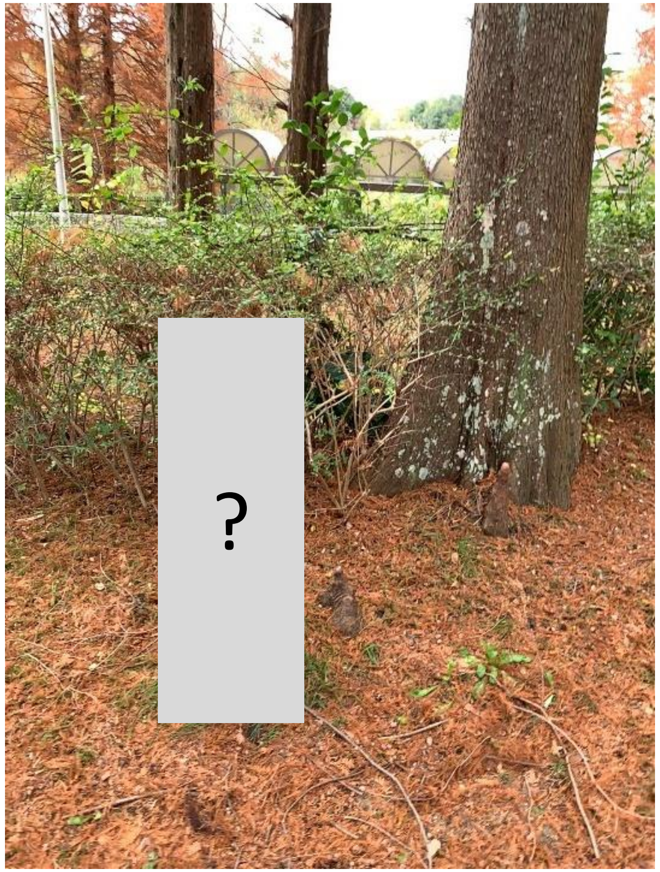
- Minimal model capabilities:
  - Ability to establish correspondences between images (matching)
  - Ability to infer 3D geometry
    - from priors & from SfM
  - Ability to infer relative pose (motion)
  - Ability to decompose motion and lighting effects or long-term changes

CroCo:
Self-supervised learning with Cross-View Completion

[NeurIPS'22] [ICCV'23]

Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon,
Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii , Gabriela Csurka, Jérôme Revaud





A guessing game:
what's masked?

CroCo:
Self-supervised learning with Cross-View Completion

[NeurIPS’22] [ICCV’23]

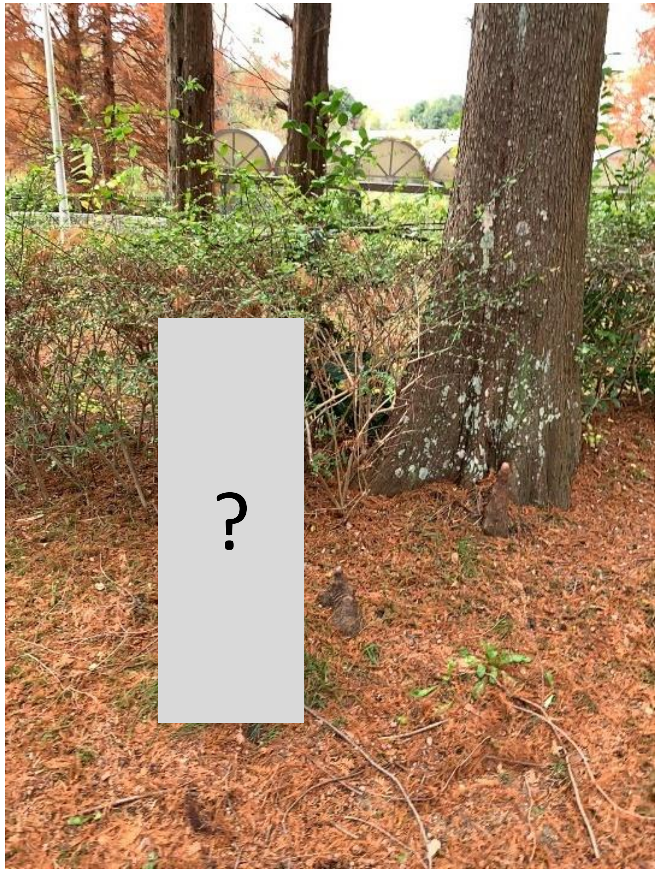
- Self-supervised learning (SSL)
  - Utilizes unlabelled data to learn useful features for downstream tasks
  - Can be used for various computer vision tasks such as object detection, segmentation, and image generation
  - Shows promising results and has been used in state-of-the-art models.

- **Masked Modelling** as a Key SSL Technique
  - Originally for text (MLM), in BERT, 2018
  - **Goal:** train a model to predict randomly-masked “parts” in the input
  - CroCo is strongly inspired by Masked Auto-Encoder (MAE) [He *et al.*, CVPR’22]

CroCo:
Self-supervised learning with Cross-View Completion

[NeurIPS’22] [ICCV’23]

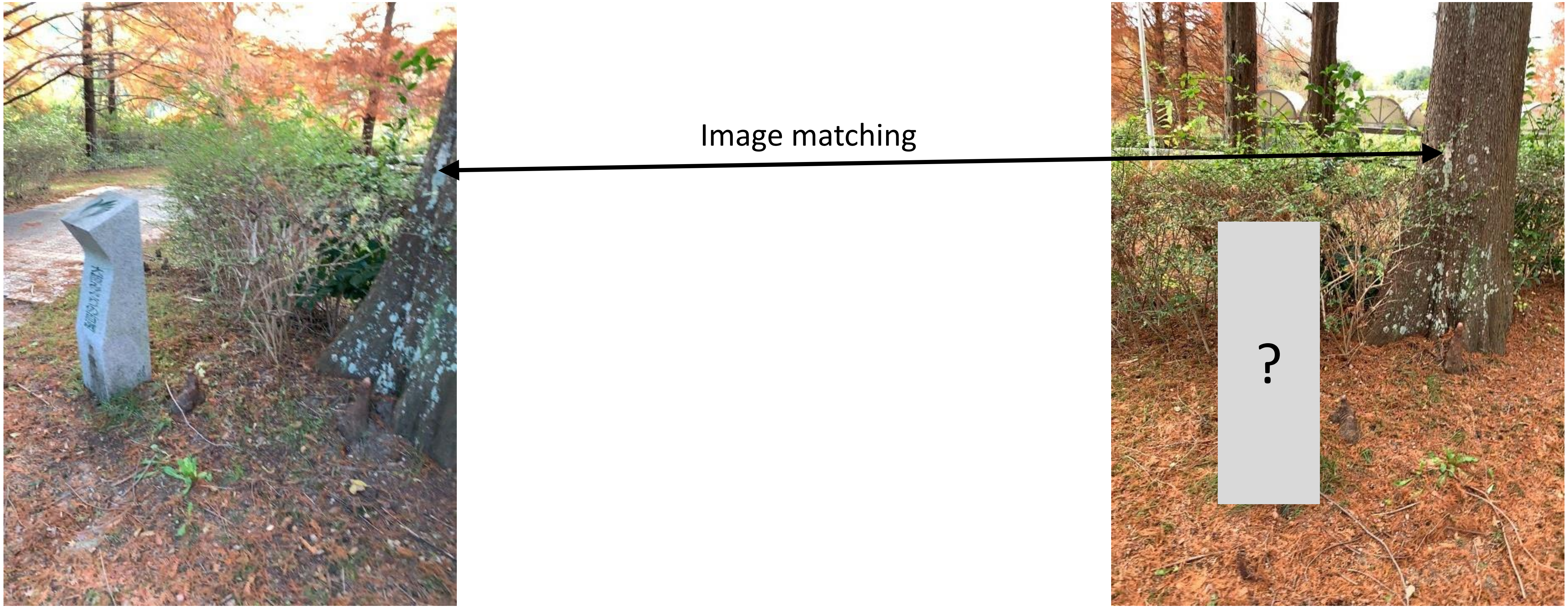
Reference view



Query view

CroCo:
Self-supervised learning with Cross-View Completion

[NeurIPS'22] [ICCV'23]



Reference view

Query view

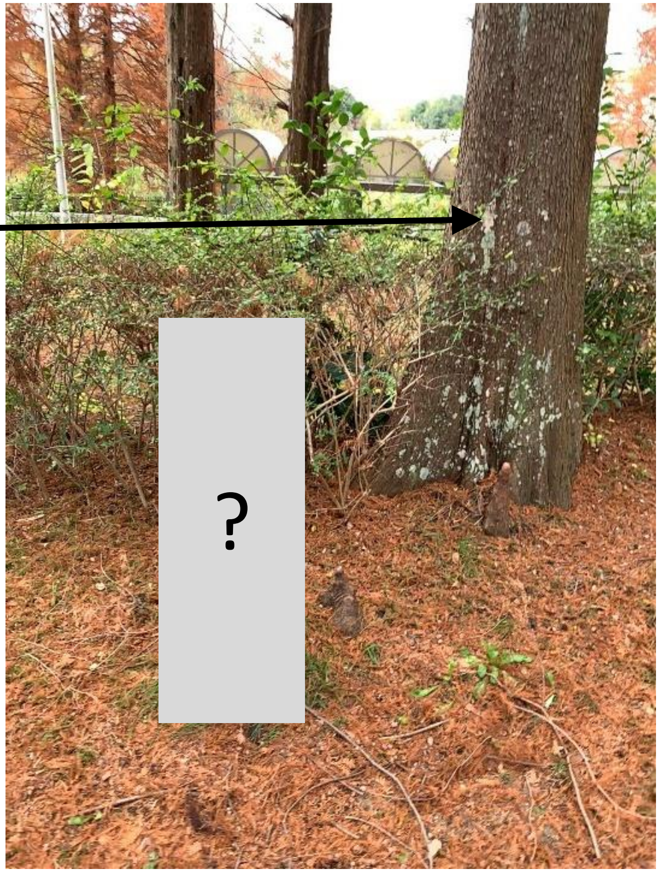
CroCo:
Self-supervised learning with Cross-View Completion

[NeurIPS'22] [ICCV'23]

Reference view

Image matching

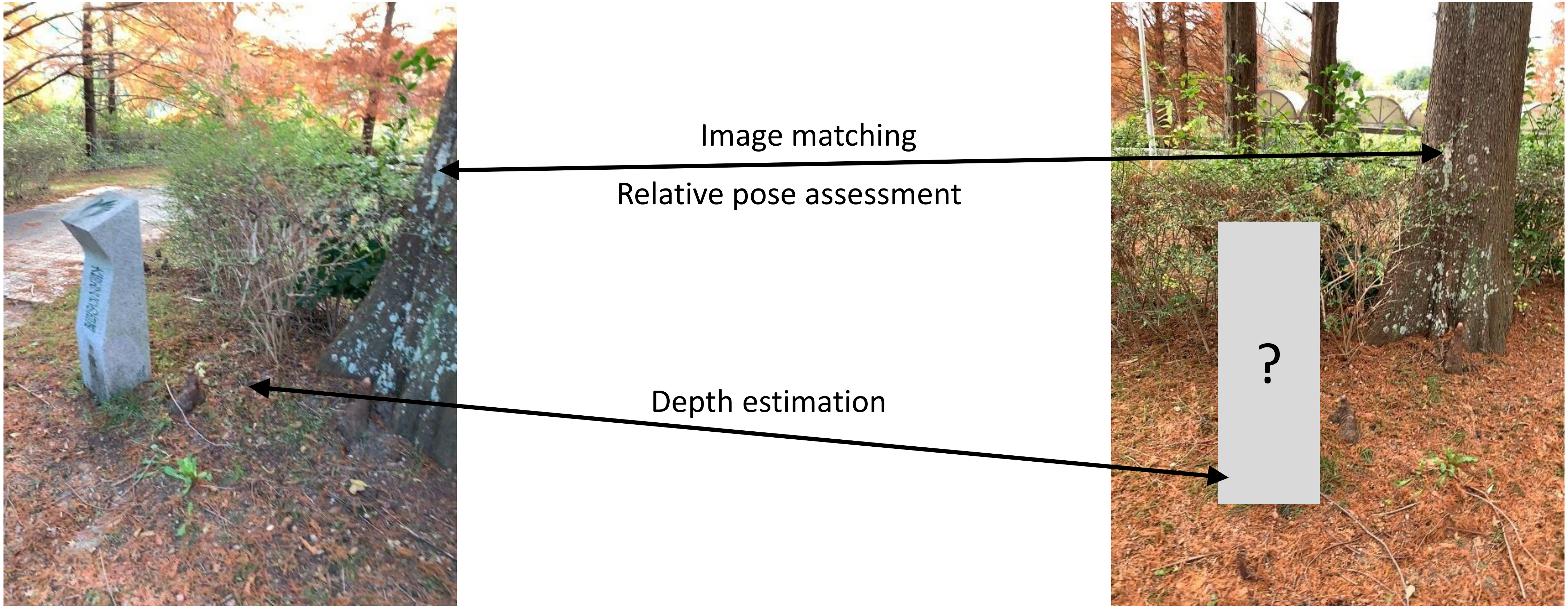
Relative pose assessment



Query view

CroCo:
Self-supervised learning with Cross-View Completion

[NeurIPS'22] [ICCV'23]

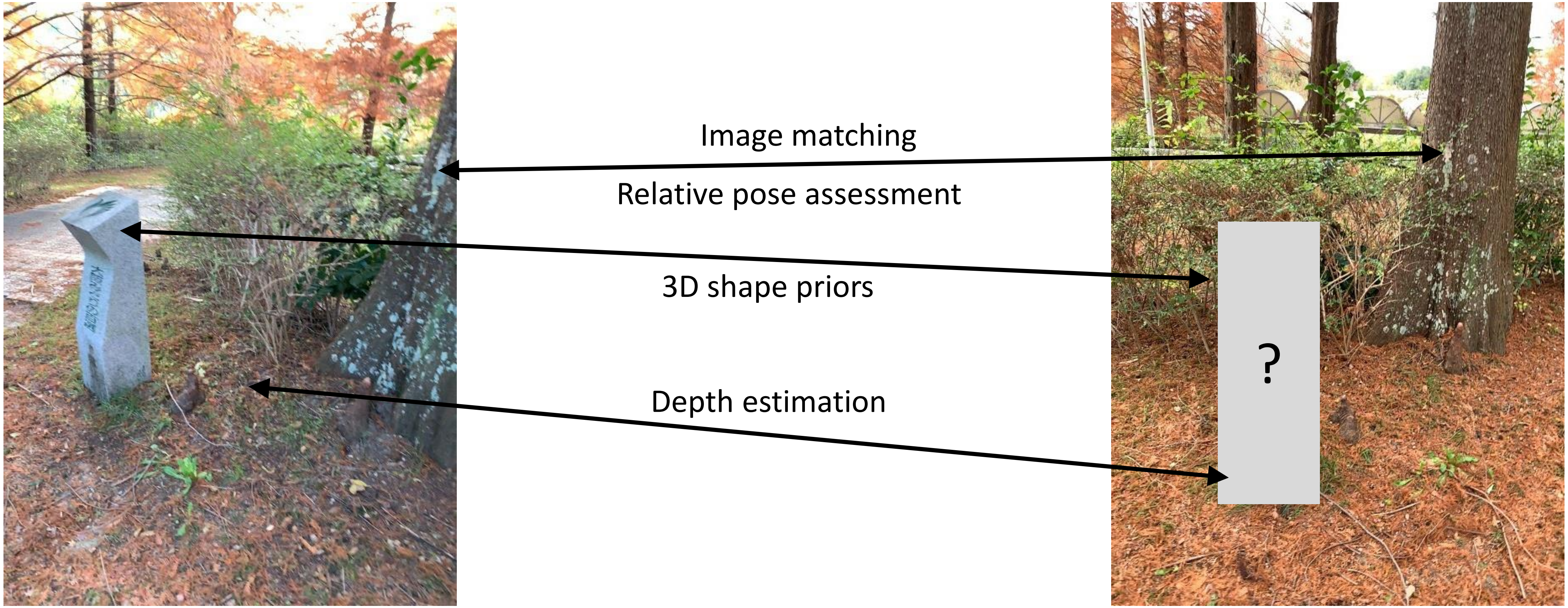


Reference view

Query view

CroCo:
Self-supervised learning with Cross-View Completion

[NeurIPS'22] [ICCV'23]

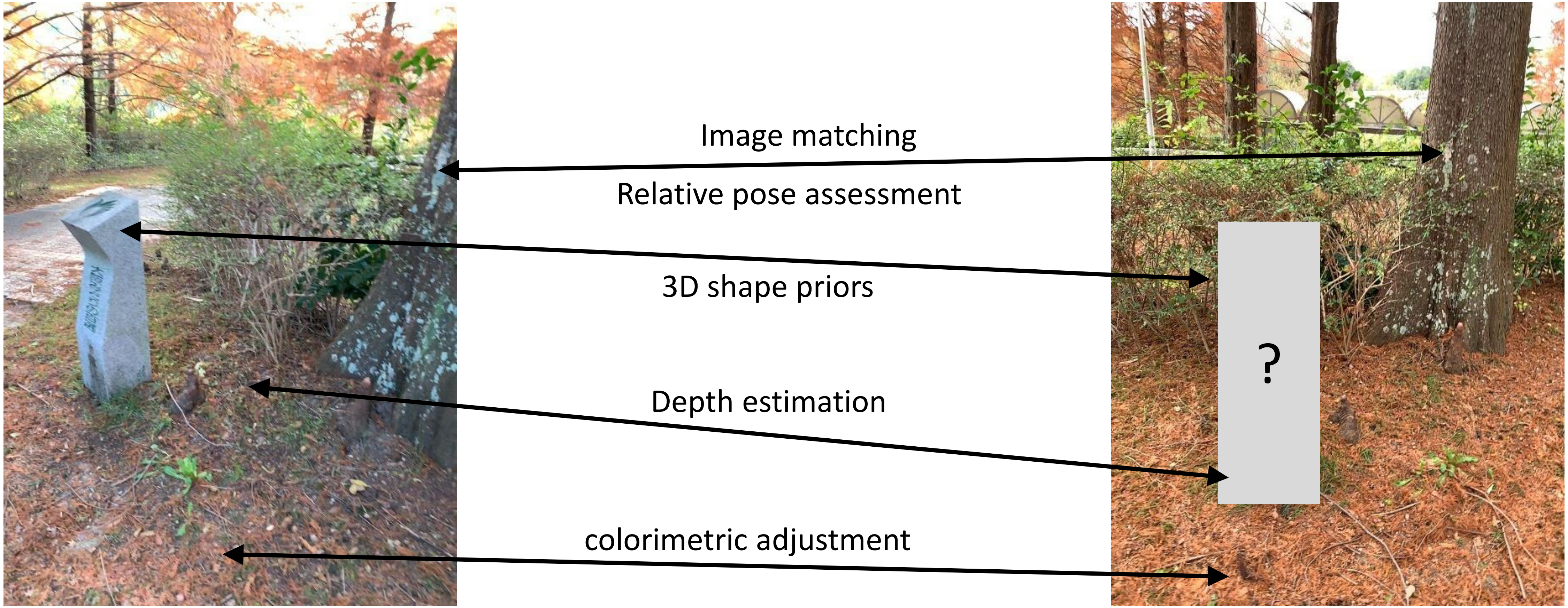


Reference view

Query view

CroCo:
Self-supervised learning with Cross-View Completion

[NeurIPS’22] [ICCV’23]

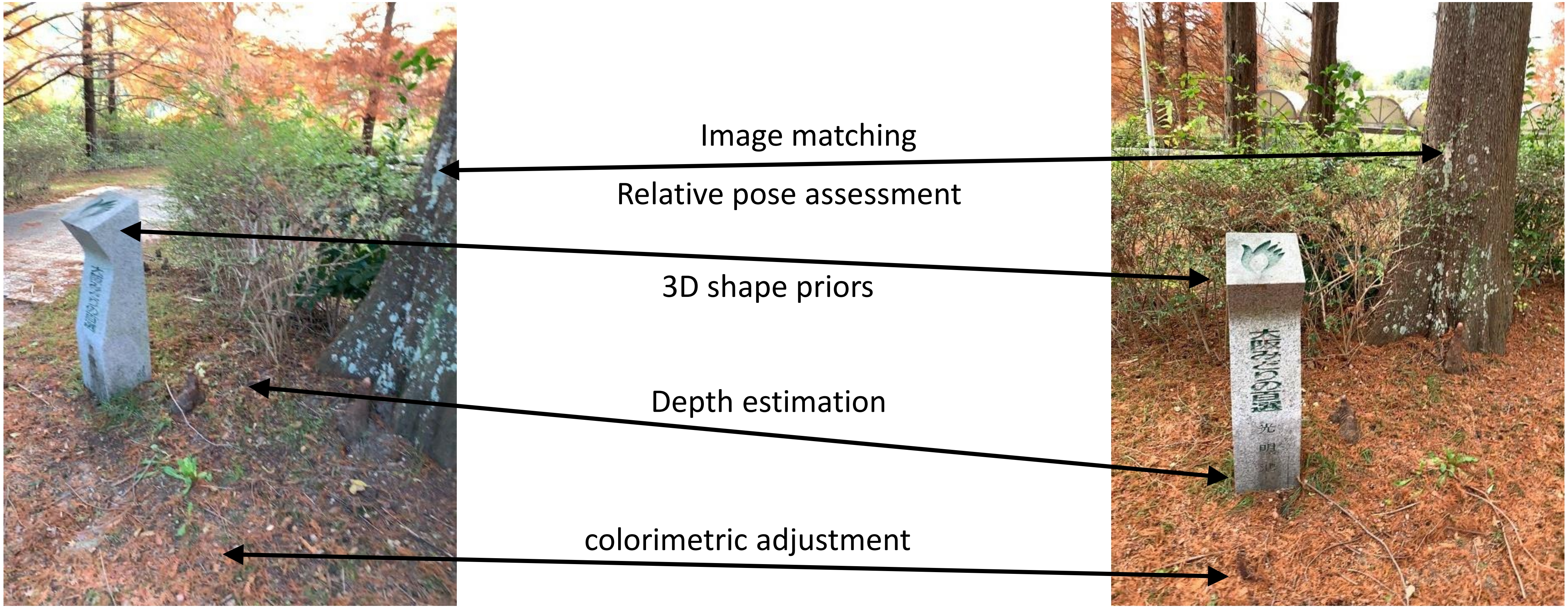


Reference view

Query view

CroCo: Self-supervised learning with Cross-View Completion

[NeurIPS'22] [ICCV'23]



Reference view

Query view

CroCo:
Self-supervised learning with Cross-View Completion

[NeurIPS'22] [ICCV'23]

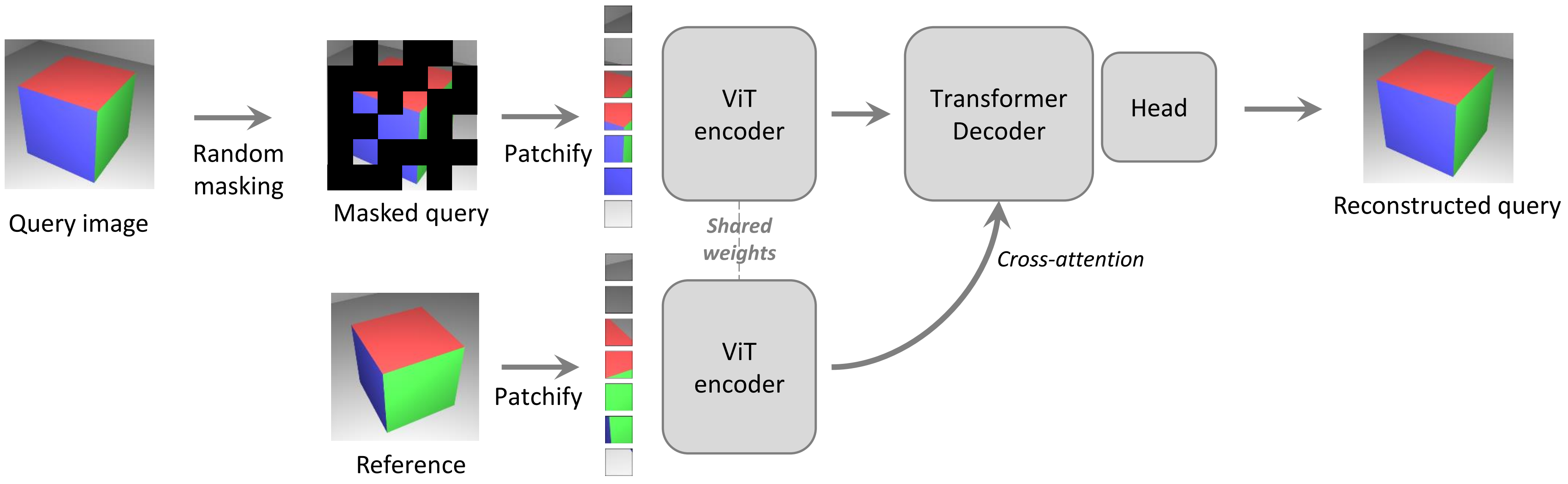
Reference view

More complex cases:

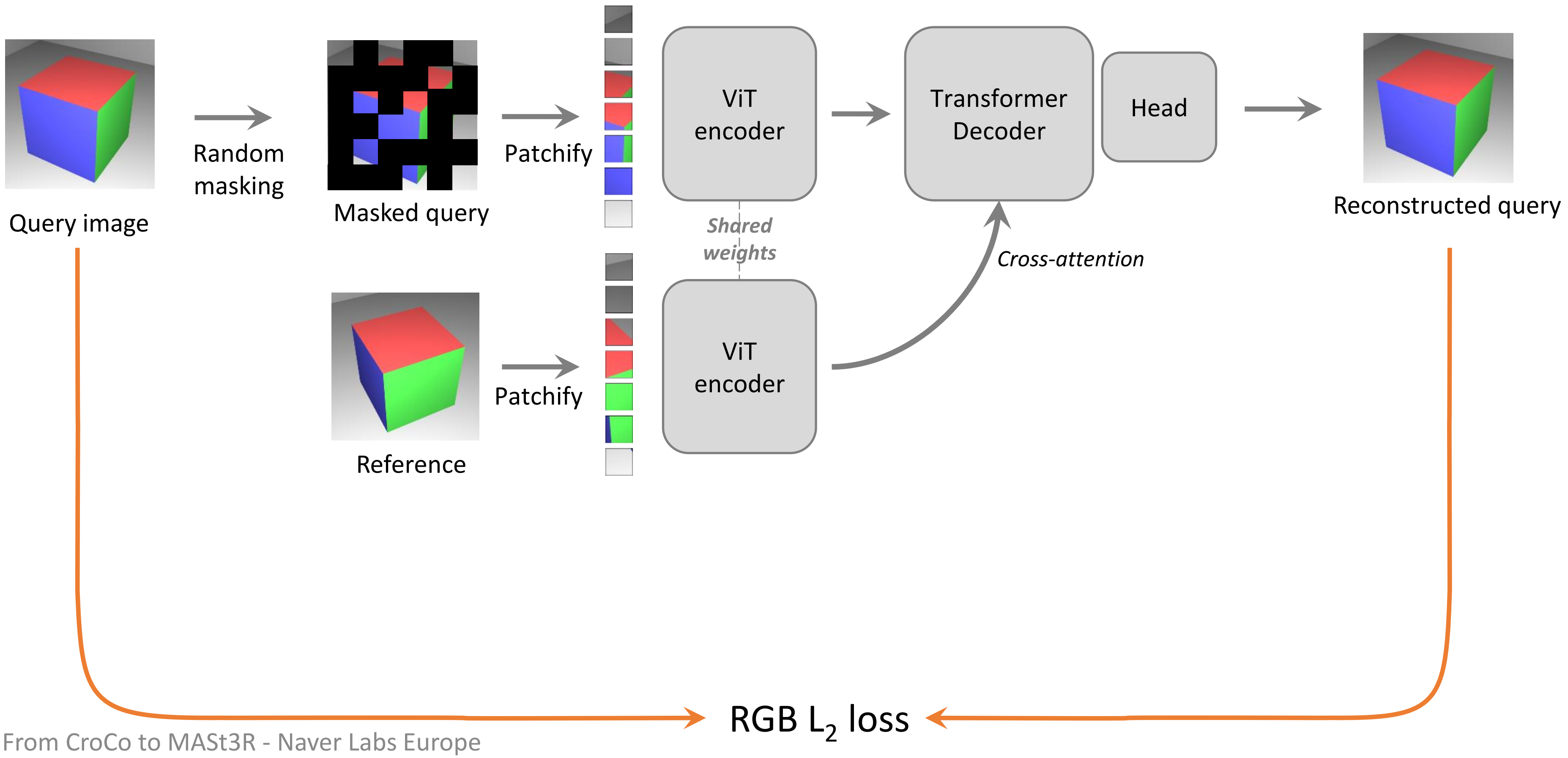
- intricate 3D shape,
- strong baseline,
- specularities,
- …

Query view

CroCo:
Self-supervised learning with Cross-View Completion

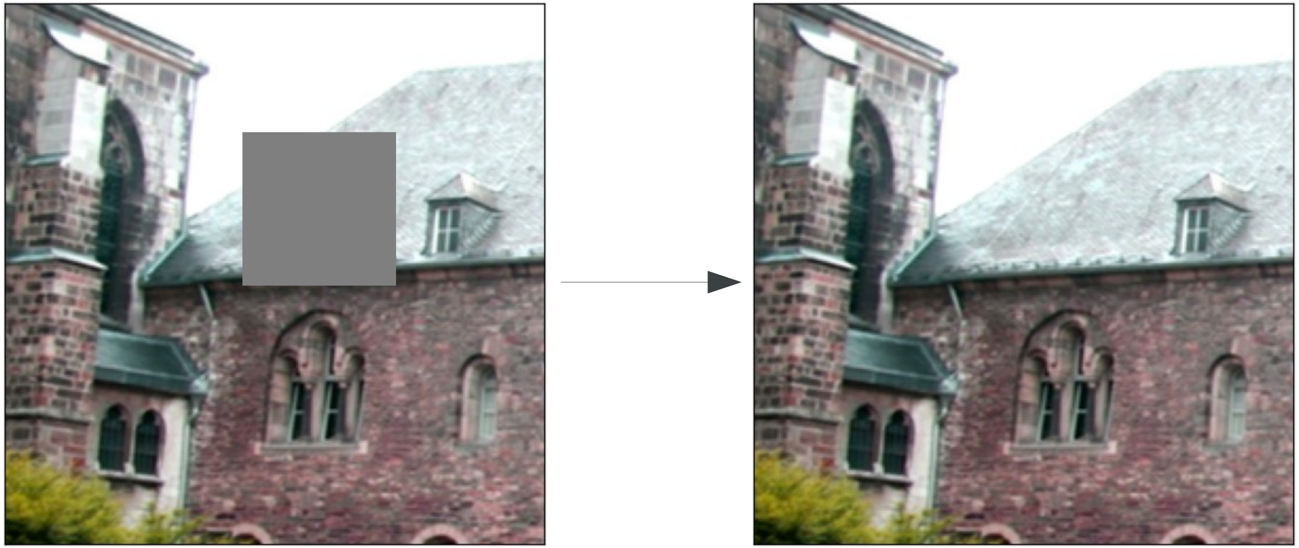


CroCo:
Self-supervised learning with Cross-View Completion



CroCo:
Self-supervised learning with Cross-View Completion

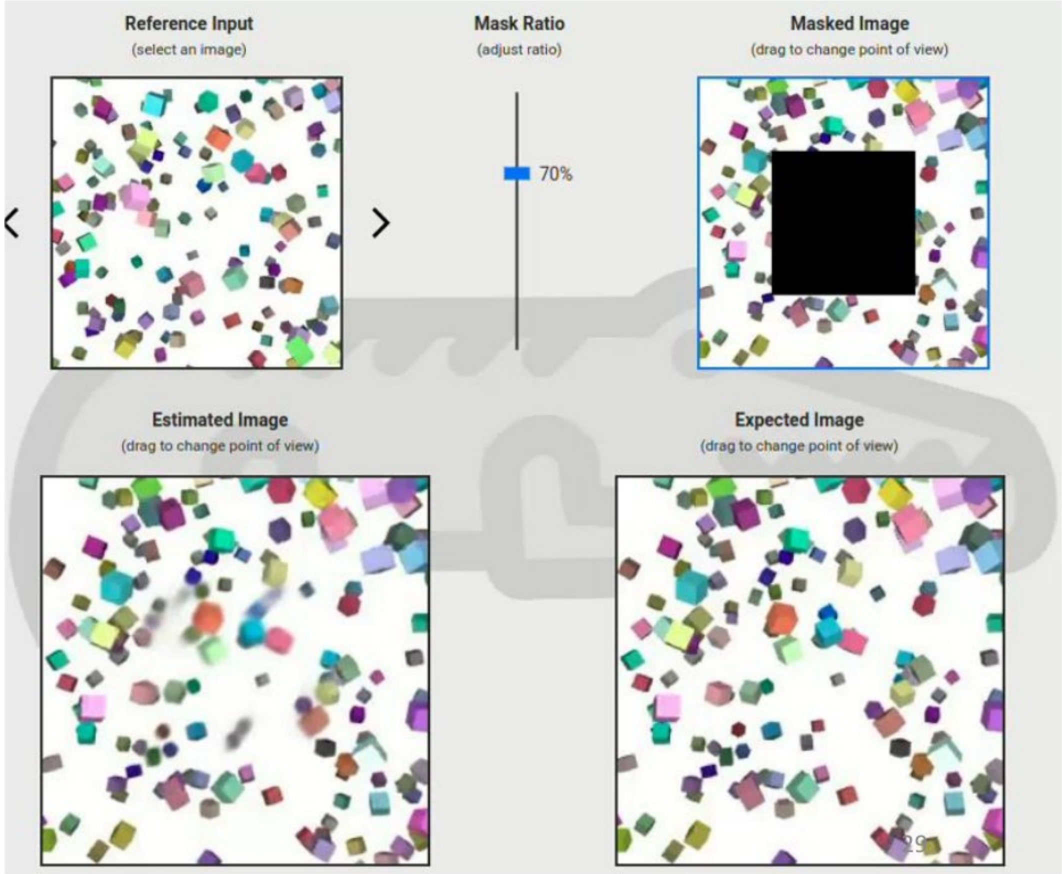
- What does it truly learn?
  - many cases can be resolved with good priors
  - To what extent is the reference view helpful?



CroCo:
Self-supervised learning with Cross-View Completion

- Proof of concept:
  - training with synthetic random scenes
  - Test scene never seen before!

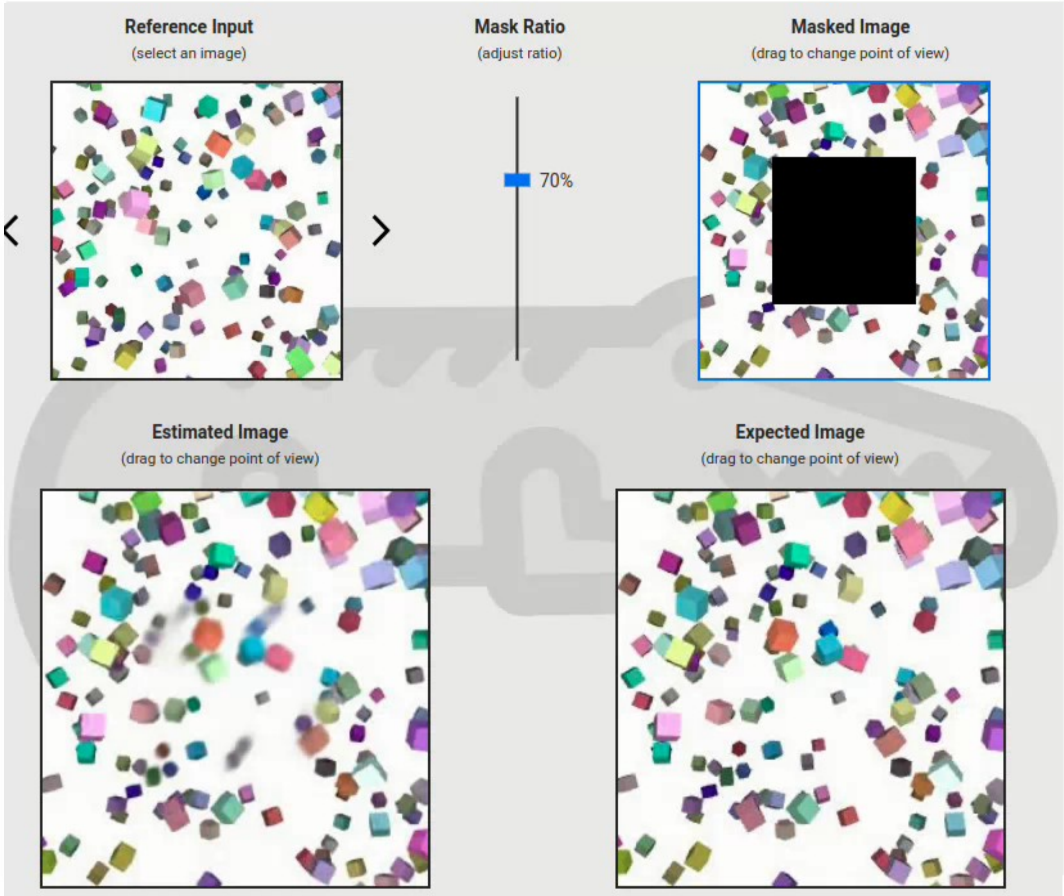
- What solving this implies:
  - Match the query and reference images
  - Estimate the relative pose
  - Infer an object-centric 3D reconstruction of the reference scene
  - Align (rotate) the reference scene in 3D
  - Render the reference scene based on imagined



CroCo:
Self-supervised learning with Cross-View Completion

- Proof of concept:
  - training with synthetic random scenes
  - Test scene never seen before!

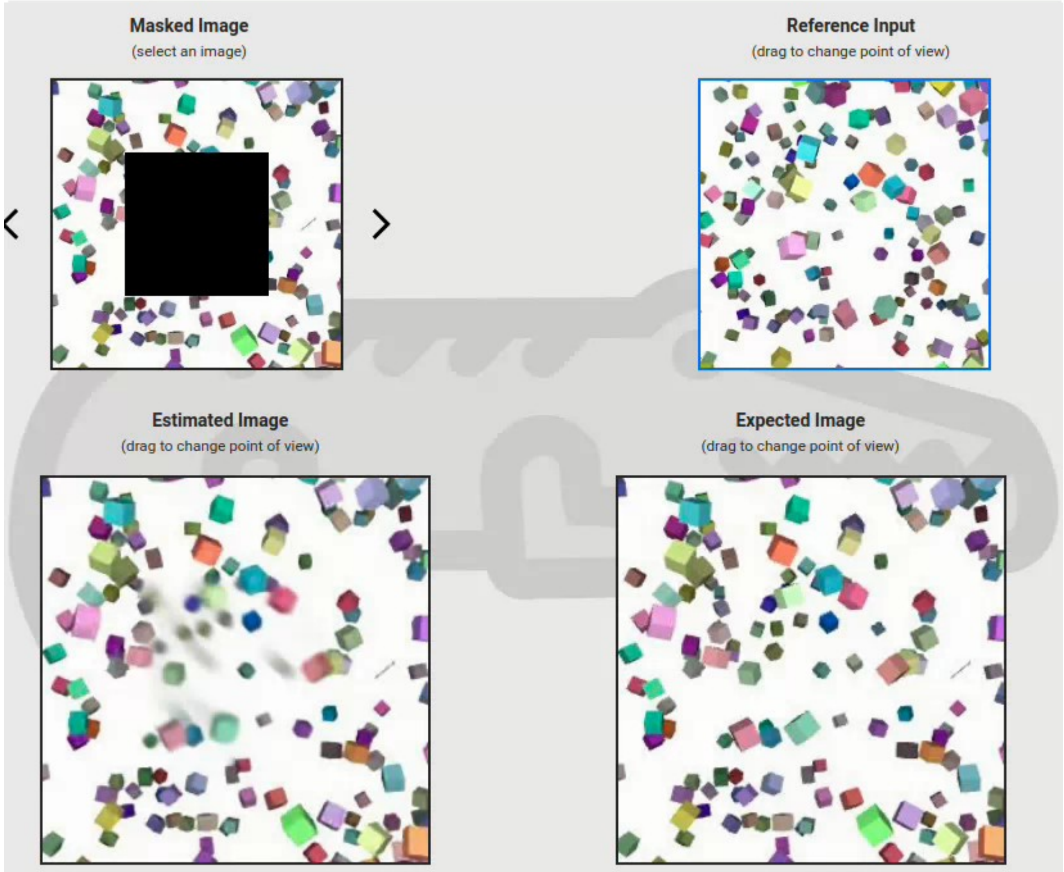
- What solving this implies:
  - Match the query and reference images
  - Estimate the relative pose
  - Infer an object-centric 3D reconstruction of the reference scene
  - Align (rotate) the reference scene in 3D
  - Render the reference scene based on imagined



CroCo:
Self-supervised learning with Cross-View Completion

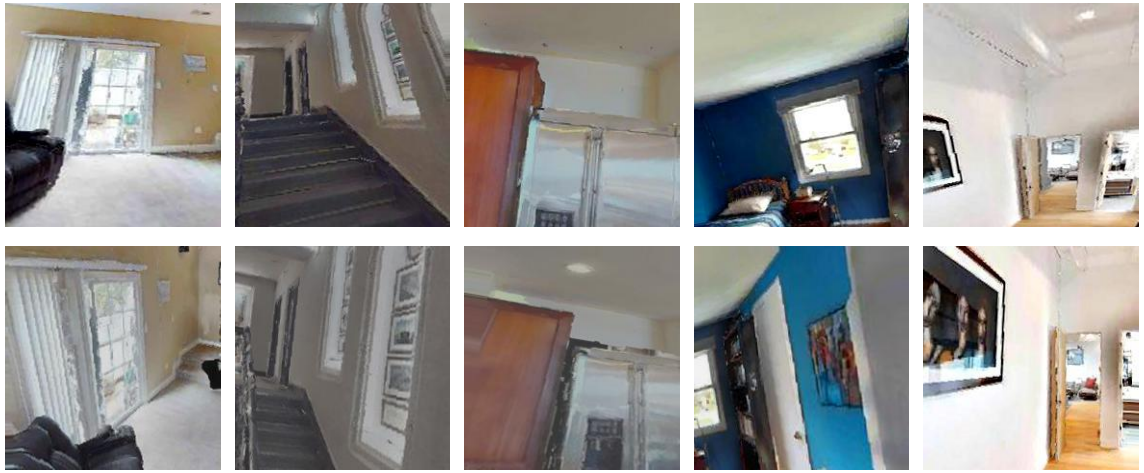
- Proof of concept:
  - training with synthetic random scenes
  - Test scene never seen before!

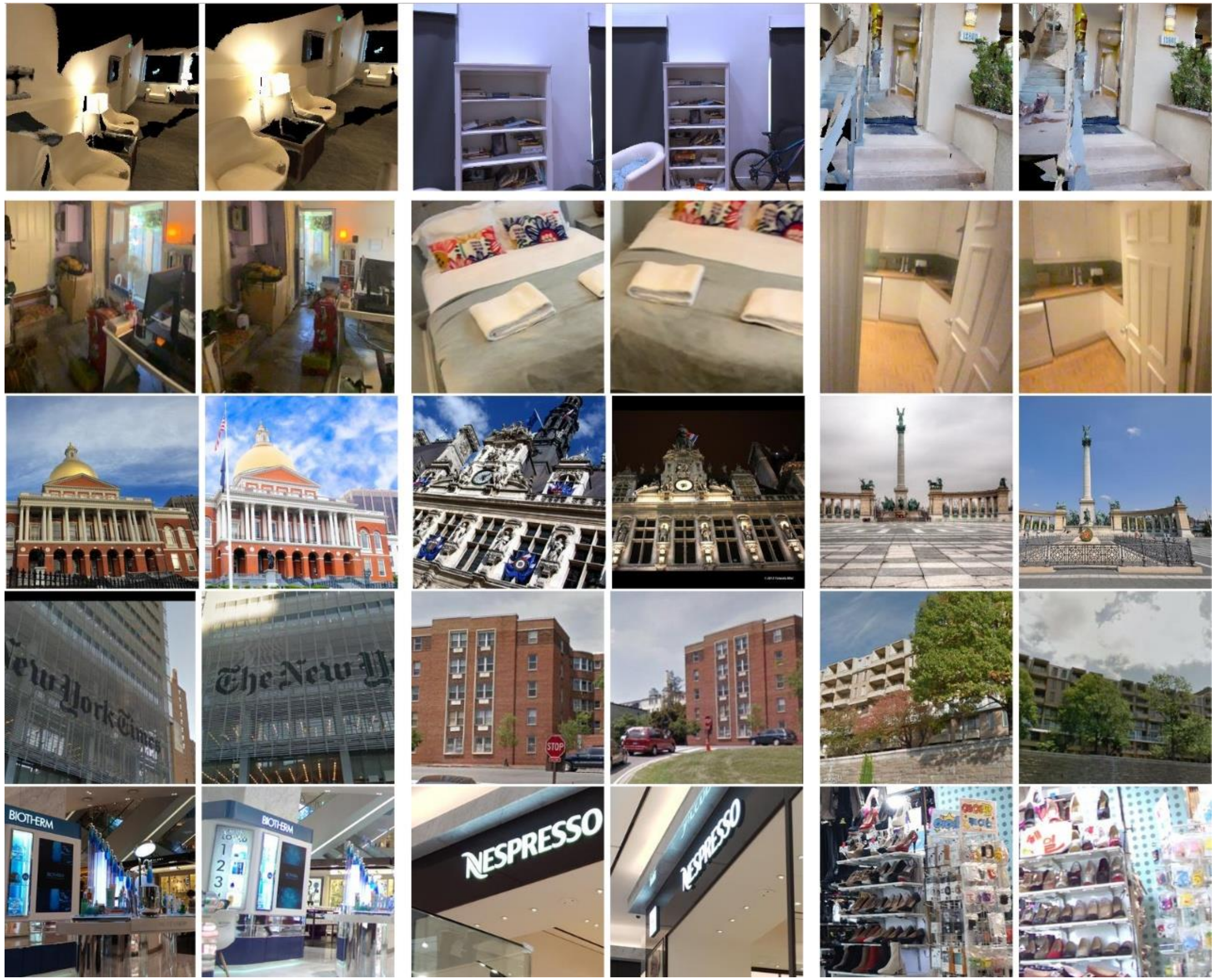
- What solving this implies:
  - Match the query and reference images
  - Estimate the relative pose
  - Infer an object-centric 3D reconstruction of the reference scene
  - Align (rotate) the reference scene in 3D
  - Render the reference scene based on imagined



Pre-training data

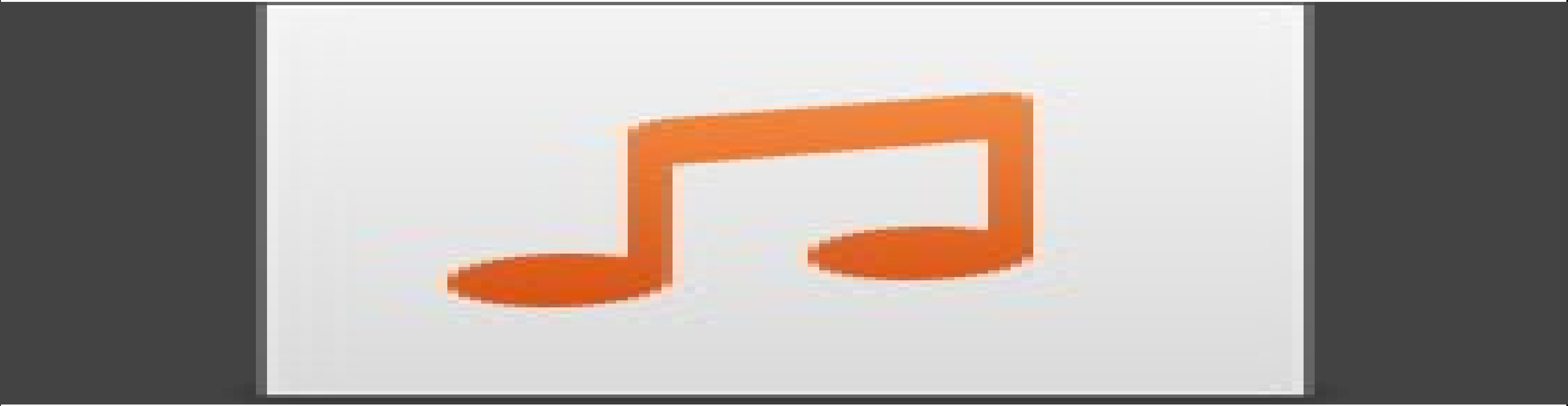
2M image pairs from the Habitat simulator
[Savva *et al.*, ICCV'19]





+ 5M training real image pairs

CroCo example



Reference input | Masked input | CroCo output | Target image

Monocular downstream tasks

CroCo encoder for monocular tasks

Semantic tasks

- Image classification
- Semantic segmentation

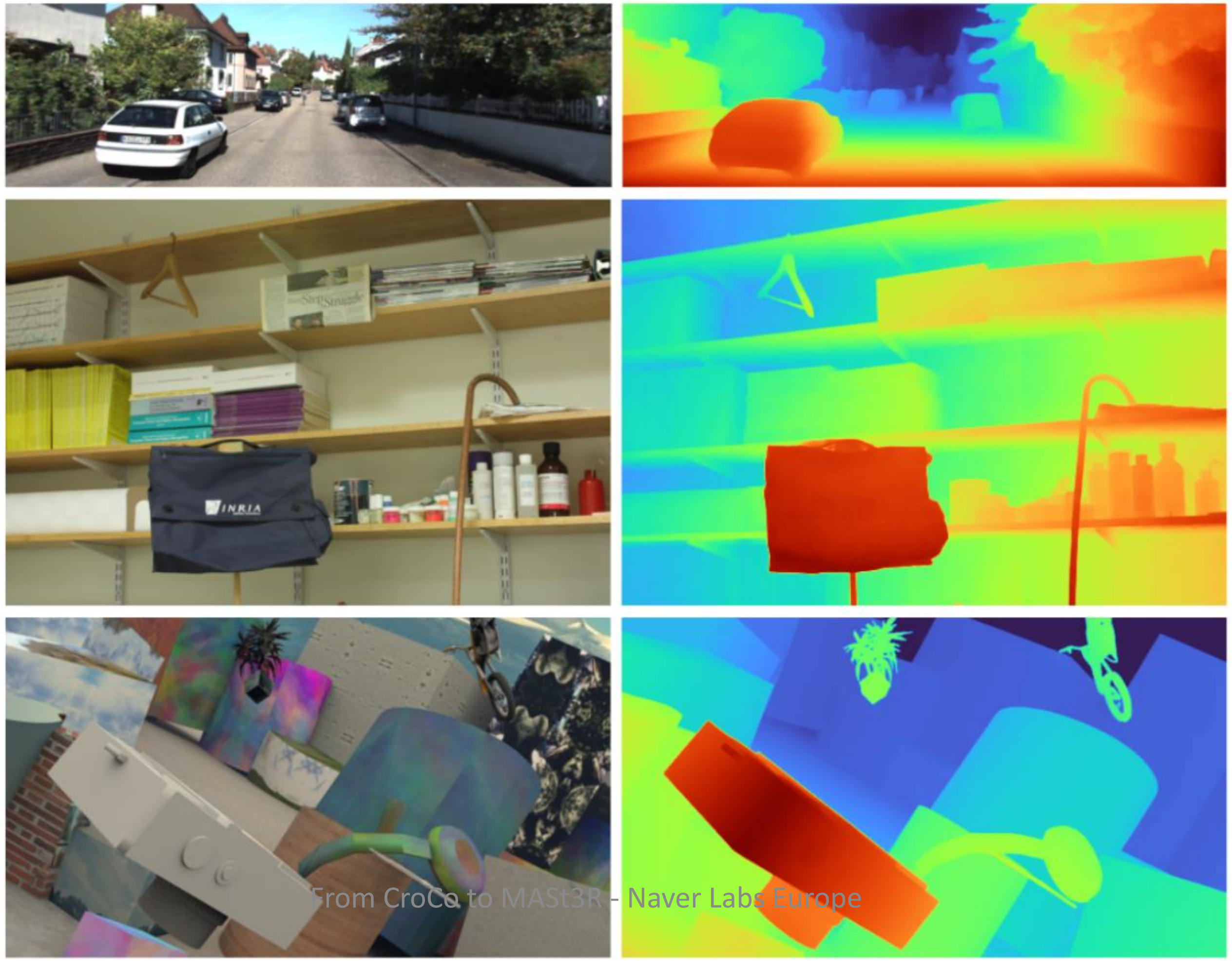
Geometric tasks

- Monocular depth (NYUv2)
- curvature, depth, edges, keypoints2d, keypoints3d, normal, occlusion, reshading

pre-training method (data)	IN1K ↑	ADE ↑	NYUv2 ↑	Taskonomy ↓									
	lin.	segm.	depth	curv.	depth	edges	kpts2d	kpts3d	normal	occl.	reshad.	avg.	rank.
DINO [14] (IN1K)	**78.2**	44.7	66.8	43.04	38.42	3.80	0.16	45.85	65.71	0.57	115.02	39.07	5.00
MAE [38] (IN1K)	75.1	46.1	79.6	41.59	35.83	**1.19**	0.08	44.18	59.20	**0.55**	106.08	36.09	2.13
MutliMAE [4] (IN1K)	60.2	**46.4**	83.0	41.42	35.38	2.17	**0.07**	44.03	60.35	0.56	105.25	36.17	2.75
MAE (Habitat)	32.5	40.3	79.0	42.06	33.63	1.79	0.08	44.81	59.76	0.56	102.54	35.65	2.88
CroCo (Habitat)	37.0	40.6	**85.6**	**40.91**	**31.34**	1.74	0.08	**41.69**	**54.13**	**0.55**	**93.58**	**33.00**	**1.25**

Binocular downstream tasks

CroCo encoder+decoder for stereo and optical flow



SPRING
Dataset & Benchmark

L. Mehl, J. Schmalfuss, A. Jahedi, Y. Nalivayko, A. Bruhn — University of Stuttgart

Download | Stereo | Optical Flow | Scene Flow | Submit | FAQ

Not logged in | Login

💡 Please note that methods marked *"submitted by spring team"* have not been finetuned on Spring.

	Name		1px ▲ total	1px low-detail	1px high-detail	1px matched	1px unmatched	1px not sky	1px sky	1px s0-10	1px s10-40	1px s40+	Abs
1	CroCo-Stereo	code	7.135	6.824	25.893	5.940	30.855	7.371	3.550	2.934	7.757	13.247	0.471
	CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow. Weinzaepfel et al. ICCV 2023.												
2	Ilnet		10.003	9.630	32.504	8.457	40.707	10.305	5.420	5.865	10.761	15.590	0.761
	Anonymous.												
3	ACVNet	code	14.772	14.432	35.273	12.600	57.894	11.163	69.621	18.386	11.346	18.145	1.516
	💡 submitted by spring team | G. Xu, J. Cheng, P. Guo, and X. Yang. "Attention Concatenation Volume for Accurate and Efficient Stereo Matching." In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.												
4	RAFT-Stereo	code	15.273	14.983	32.774	13.394	52.582	9.924	96.571	22.588	10.018	17.086	3.025
	💡 submitted by spring team | L. Lipson, Z. Teed, and J. Deng. "RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching." In International Conference on 3D Vision (3DV), 2021.												
5	PWOC-3D [SF]	code	18.226	17.831	42.067	16.020	62.014	15.946	52.877	18.279	12.716	34.570	1.343
	R. Saxena, R. Schuster, O. Wasenmuller, and D. Stricker. "PWOC-3D: Deep Occlusion-Aware End-to-End Scene Flow Estimation." In IEEE Intelligent Vehicles Symposium (IV), 2019.												
6	LEAStereo	code	19.888	19.547	40.396	17.611	65.086	16.735	67.805	19.076	13.861	39.412	3.884
	💡 submitted by spring team | X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drummond, and Z. Ge. "Hierarchical Neural Architecture Search for Deep Stereo Matching." In NeurIPS, 2020.												
7	M-FUSE (F) [SF]	code	19.888	19.547	40.396	17.611	65.086	16.735	67.805	19.076	13.861	39.412	3.884
	💡 submitted by spring team | L. Mehl, A. Jahedi, J. Schmalfuss, and A. Bruhn. "M-FUSE: Multi-frame Fusion for Scene Flow Estimation." In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023.												
8	SplatFlow3D (C+T) + LEAStereo (Things); Two-frame [SF]	code	19.888	19.547	40.396	17.611	65.086	16.735	67.805	19.076	13.861	39.412	3.884
9	GANet	code	23.225	22.912	42.064	20.976	67.878	18.418	96.274	24.286	16.427	41.499	4.594
	💡 submitted by spring team | F. Zhang, V. Prisacariu, R. Yang, and P. HS Torr. "GA-Net: Guided Aggregation Net for End-to-end Stereo Matching." In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.												
10	RAFT-3D (F) [SF]	code	23.225	22.912	42.064	20.976	67.878	18.418	96.274	24.286	16.427	41.499	4.594
	💡 submitted by spring team | Z. Teed, and J. Deng. "RAFT-3D: Scene Flow using Rigid-Motion Embeddings." In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.												
11	CamLiFlow (F) [SF]	code	23.225	22.912	42.064	20.976	67.878	18.418	96.274	24.286	16.427	41.499	4.594

https://github.com/naver/croco

CroCo: summary

- Self-supervised pretraining
  - Specifically designed for 3D vision, inherently multi-view
  - Arguably and provably learns important “bricks” of 3D vision
  - Generic architecture, easily adaptable for any 3DV downstream task

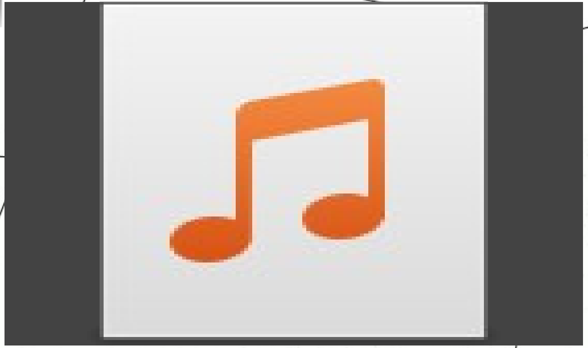
- CroCo lays the foundation for a unified model
  - But nothing is unified yet (each downstream task is finetuned separately)

    ➔ still seeking for a unified model …

What is 3D vision?

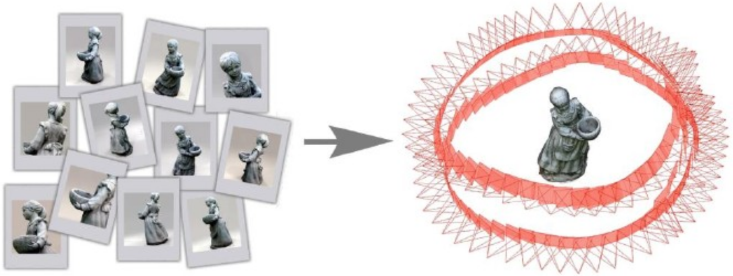
Monocular Depth estimation

Point matching



Dense 3D reconstruction

Visual Localization

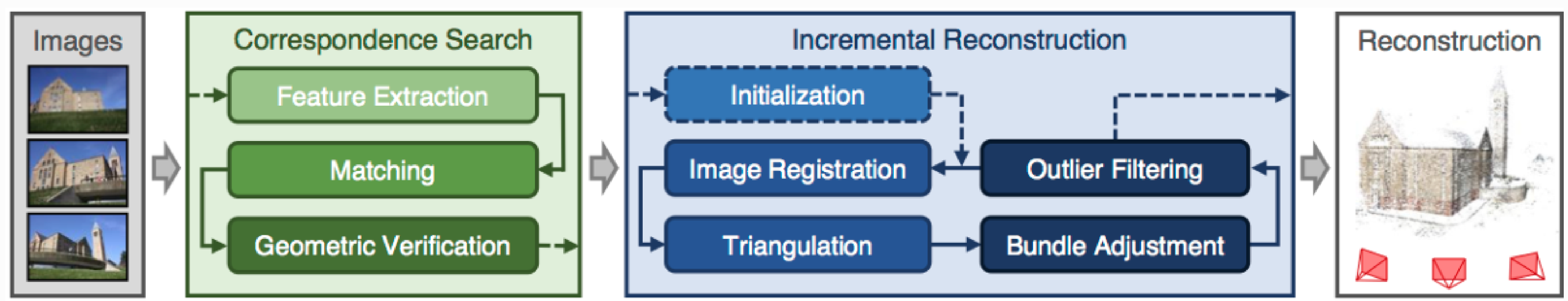


Multi-view pose estimation

… and many more: SLAM, calibration, MVS, …

Unifying all 3D vision tasks?

- Could “dense 3D reconstruction” be a “super task” for 3DV?



COLMAP’s incremental Structure-from-Motion pipeline.

“Structure-from-Motion Revisited”, “Pixelwise View Selection for Unstructured Multi-View Stereo”, Schonberger et al., in CVPR’16 & ECCV’16

Unifying all 3D vision tasks?

Sparse model of central Rome using 21K photos produced by COLMAP's SfM pipeline.

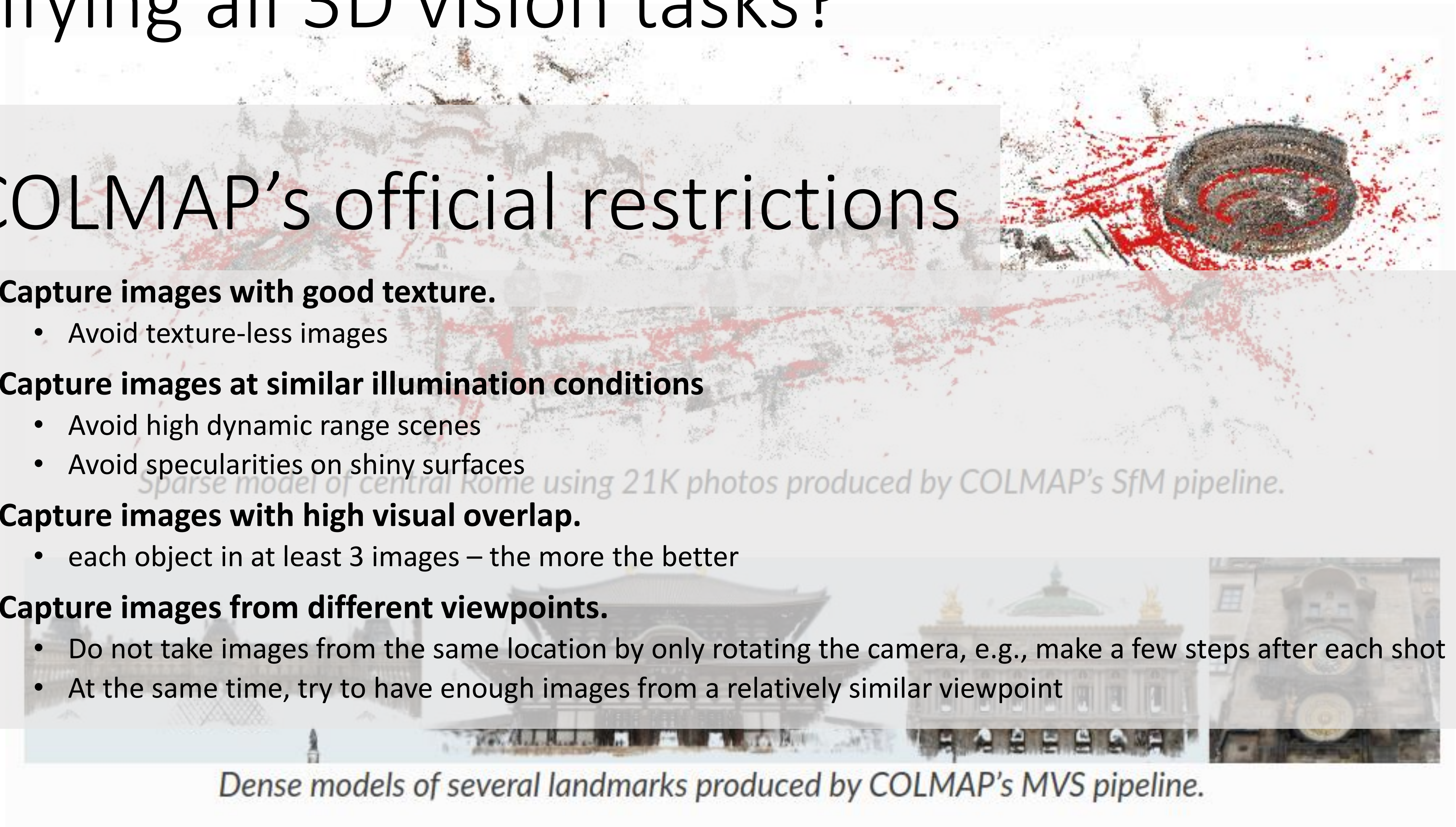
Dense models of several landmarks produced by COLMAP's MVS pipeline.

“Structure-from-Motion Revisited”, “Pixelwise View Selection for Unstructured Multi-View Stereo”, Schonberger et al., in CVPR’16 & ECCV’16

Unifying all 3D vision tasks?

COLMAP's official restrictions

- **Capture images with good texture.**
  - Avoid texture-less images
- **Capture images at similar illumination conditions**
  - Avoid high dynamic range scenes
  - Avoid specularities on shiny surfaces
- **Capture images with high visual overlap.**
  - each object in at least 3 images – the more the better
- **Capture images from different viewpoints.**
  - Do not take images from the same location by only rotating the camera, e.g., make a few steps after each shot
  - At the same time, try to have enough images from a relatively similar viewpoint



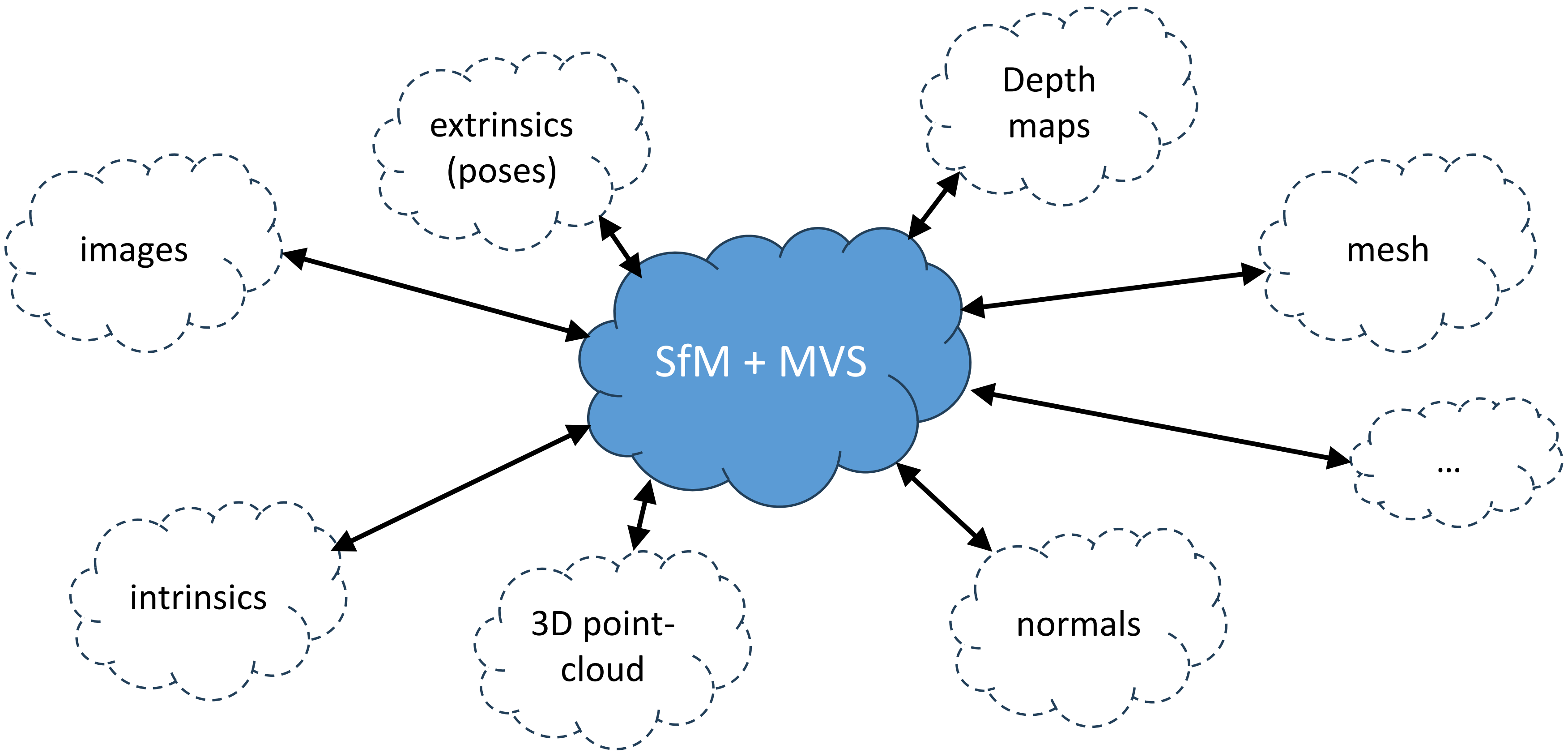
"Structure-from-Motion Revisited", "Pixelwise View Selection for Unstructured Multi-View Stereo", Schonberger et al., in CVPR'16 & ECCV'16

Unifying all 3D vision tasks?

- 3D reconstruction is a “super-task” ☺
  - intrinsically connected to all other 3DV tasks

- Current solution is problematic ☹
  - Brittle, requires enough *images* & *overlap* & *textures* & *viewpoints*
  - Heavily handcrafted at all levels
    - An engineering hell!
  - Multiple minimal problems solved sequentially
    - No internal collaboration between them
  - Slow

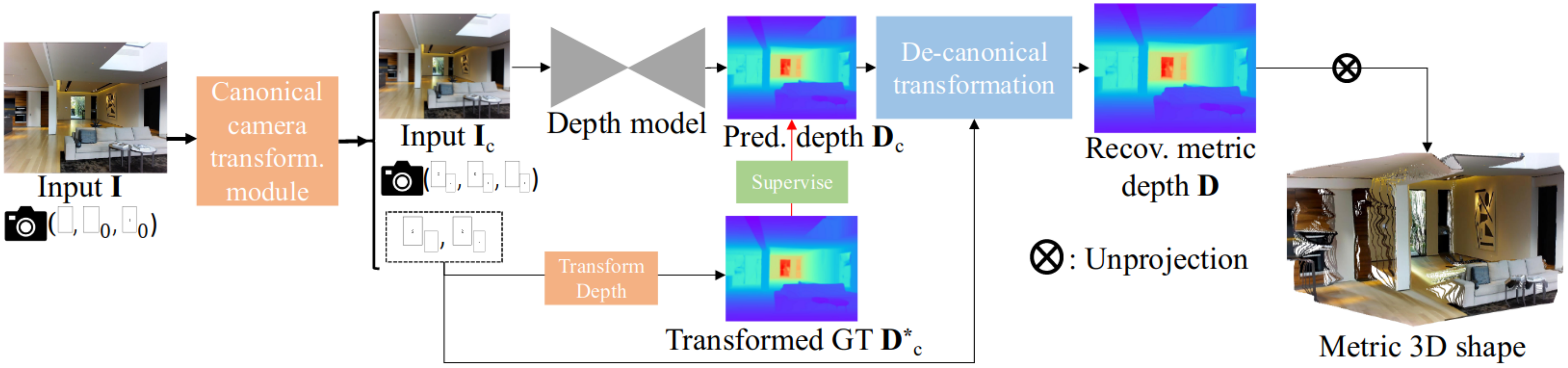
How to reconcile 3D reconstruction and DL?



How to reconcile 3D reconstruction and DL?

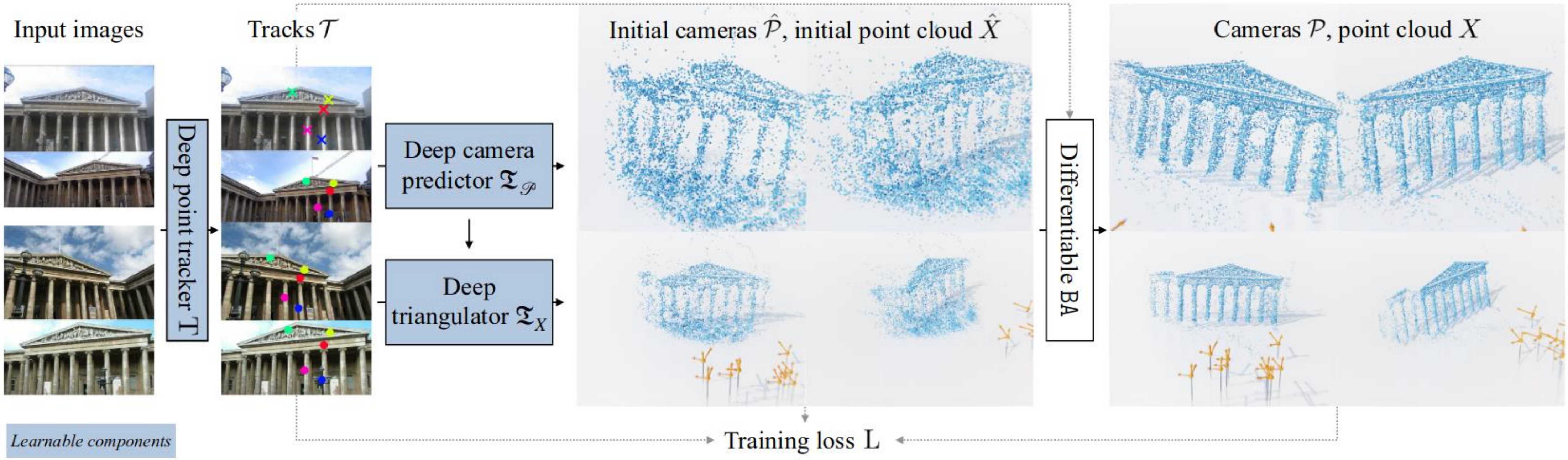
- Direct RGB-to-3D
  - Monocular depth estimation
  - Quite handcrafted

Metric3D, by Yin et. Al., [ICCV'23]



How to reconcile 3D reconstruction and DL?

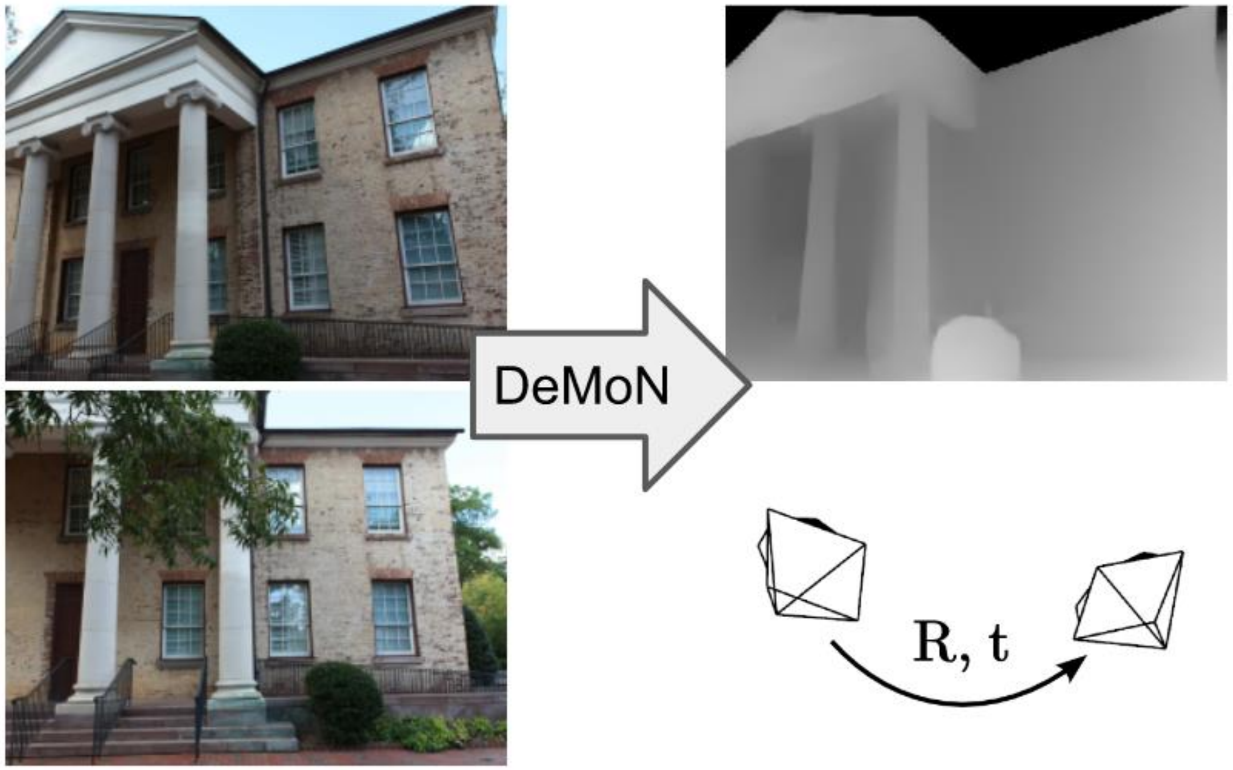
- End-to-end training of a differentiable version of the SfM pipeline



VGGSfM (Visual Geometry Grounded Deep Structure From Motion), by Wang et. Al., [CVPR'24]

How to reconcile 3D reconstruction and DL?

- 2-views SfM as a regression problem?
  - DeMoN [CVPR'17]
  - DeepTAM [IJCV'20]
  - DeepV2D [ICLR'20]



How to reconcile 3D reconstruction and DL?

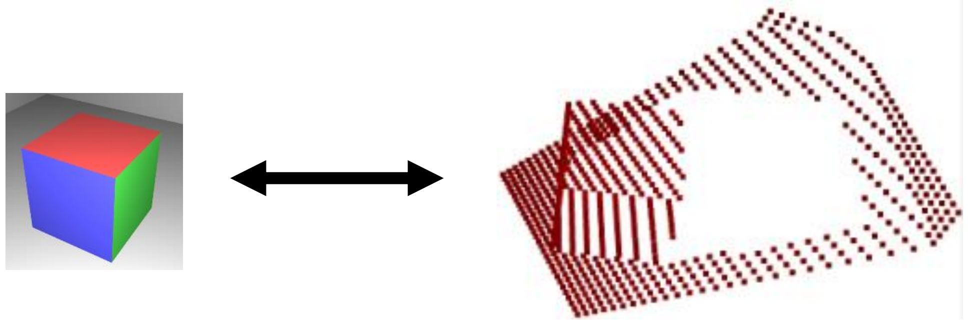
- 2-views SfM as a regression problem?
  - Does not generalize to new camera intrinsics or poses
  - Unstable output space

How to reconcile 3D reconstruction and DL?

- 2-views SfM as a regression problem?
  - What's a good parameterization of the output space?

➔ "Pointmap"

  - 1-to-1 mapping between pixels and their corresponding 3D points
  - *All geometric 3DV tasks* manipulate 2D-3D correspondences somehow!



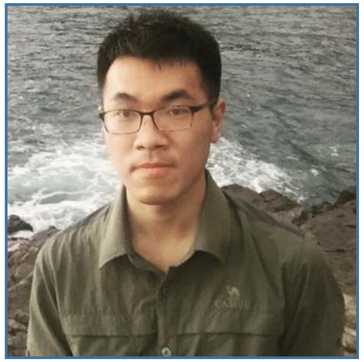
Note: similar considerations made by UniDepth [CVPR'24]

(Depth prediction should be conditioned on intrinsics, output space matters a lot)

How to reconcile 3D reconstruction and DL?

- 2-views SfM as a regression problem?
  - *Roughly invariant to camera parameters*
  - *But over-parameterized (no guarantee to get pinhole-consistent geometry)*

DUSt3R: Dense Unconstrained Stereo 3D Reconstruction


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

DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction

- Pointmaps as a proxy output that:
  - *capture 3D scene geometry (point-cloud)*
  - *connect pixels ↔ 3D points*
  - *spatially relate 2 viewpoints (relative pose)*

Unconstrained
image collection

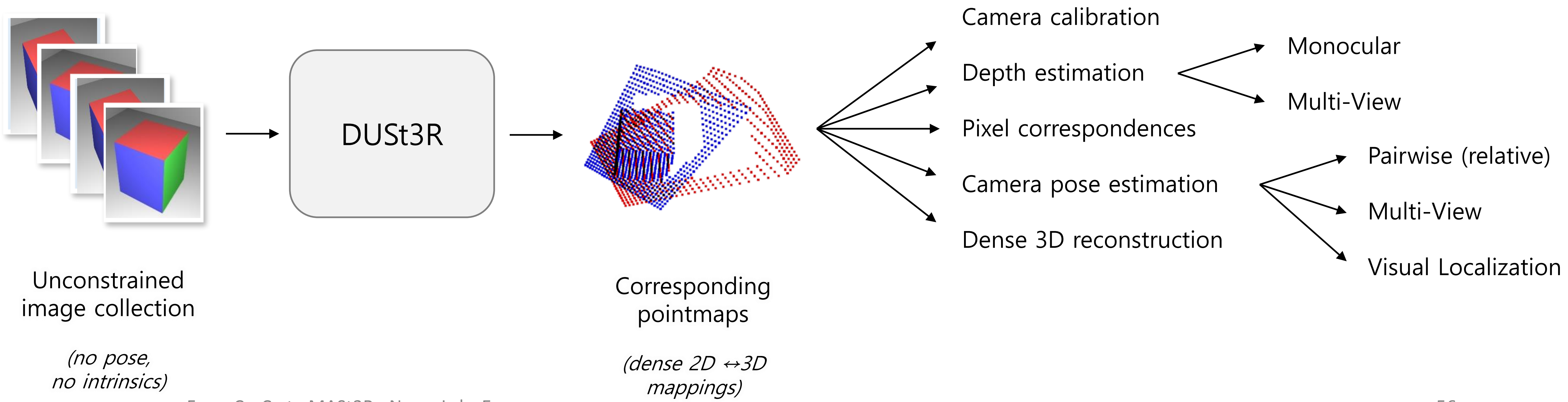
*(no pose,
no intrinsics)*

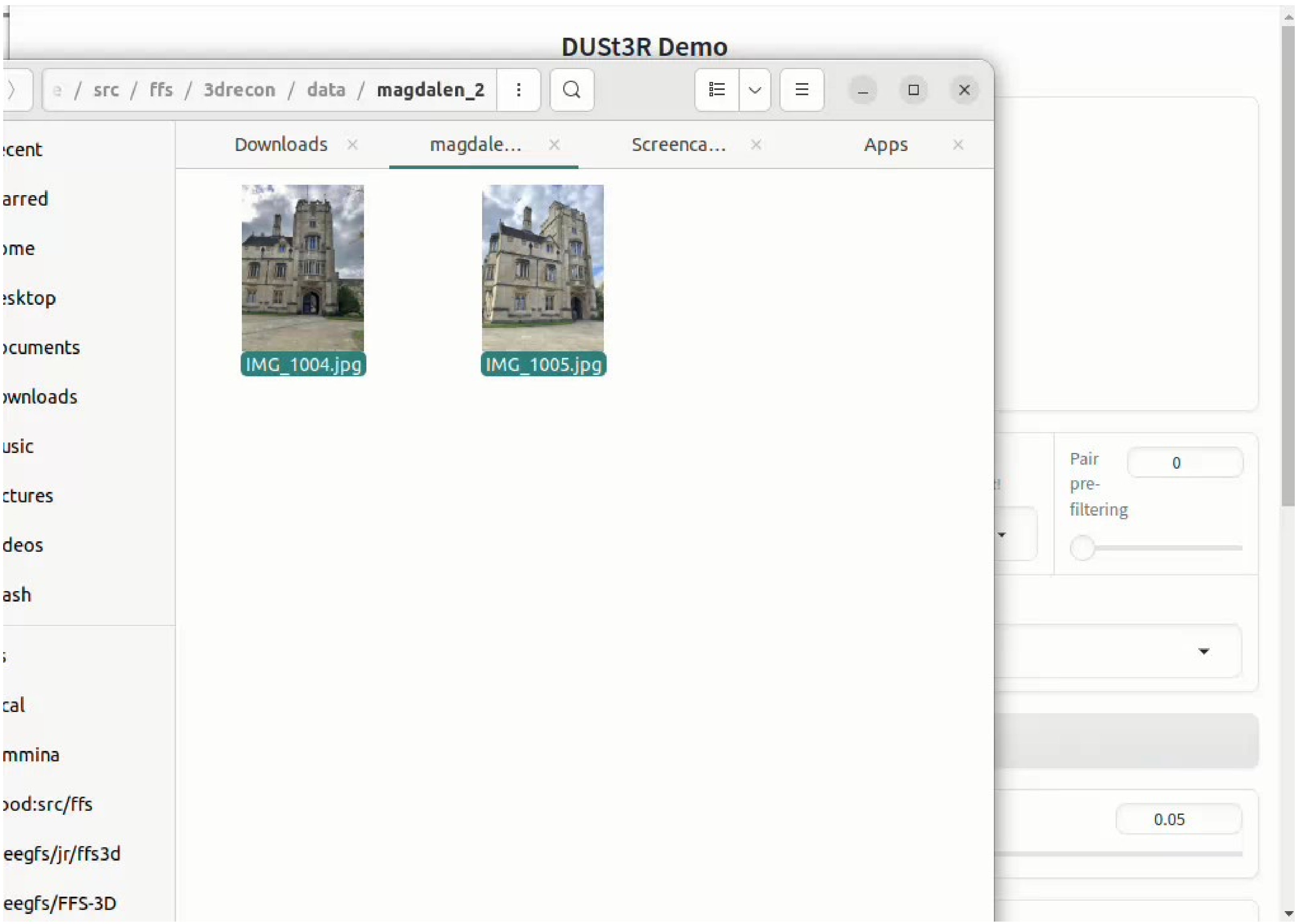
Corresponding
pointmaps

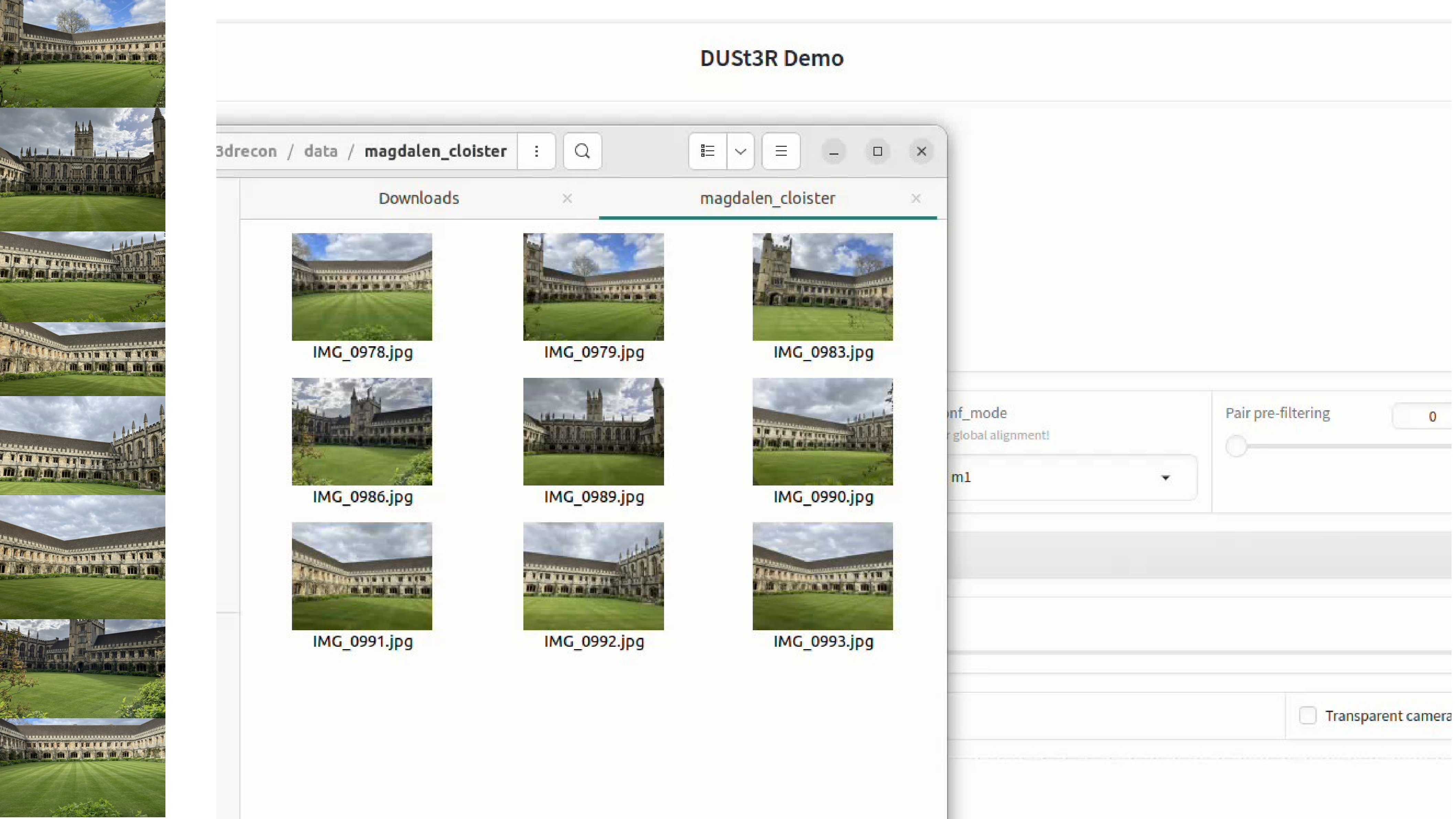
*(dense 2D ↔3D
mappings)*

DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction

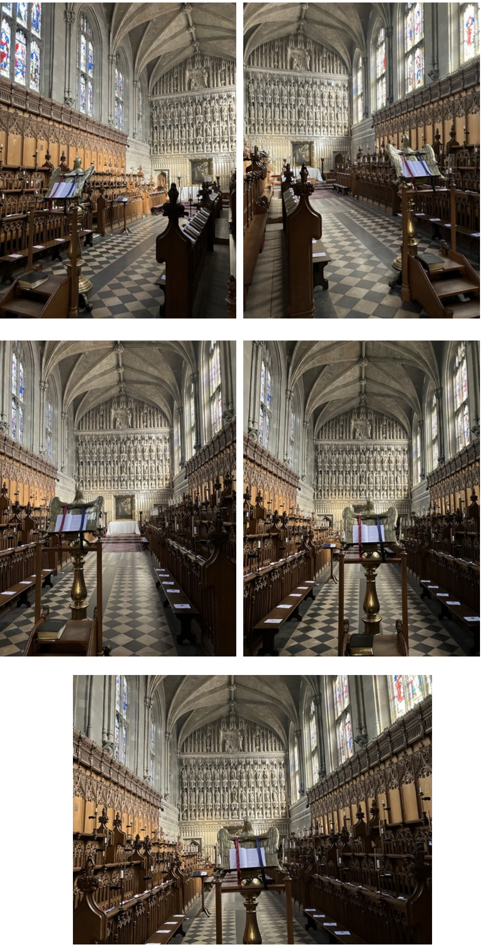
- Pointmaps as a proxy output that:
  - *capture 3D scene geometry (point-cloud)*
  - *connect pixels ↔ 3D points*
  - *spatially relate 2 viewpoints (relative pose)*




DUSt3R Demo
src / ffs / 3drecon / data / magdalen_2
Downloads
magdale...
Screenca...
Apps
IMG_1004.jpg
IMG_1005.jpg
Pair pre-filtering
0
0.05



DUSt3R Demo
3drecon / data / magdalen_cloister
Downloads
magdalen_cloister
IMG_0978.jpg
IMG_0979.jpg
IMG_0983.jpg
IMG_0986.jpg
IMG_0989.jpg
IMG_0990.jpg
IMG_0991.jpg
IMG_0992.jpg
IMG_0993.jpg
nf_mode
global alignment!
m1
Pair pre-filtering
0
Transparent camera



src / ffs /

ecent
tarred
lome
esktop
ocuments
ownloads
lusic
ictures
ideos
rash
fs
ocal
emmina
ood:src/ffs
beegfs/jr/ffs3d
beegfs/FFS-3D
ther Locations

47%

IMG_0968.jpg

Properties

Size	1000 × 1334 pixels
Type	JPEG image
File Size	491,2 kB
Folder	magdalen_3
Aperture	
Exposure	
Focal Length	
ISO	
Metering	
Camera	
Date	
Time	

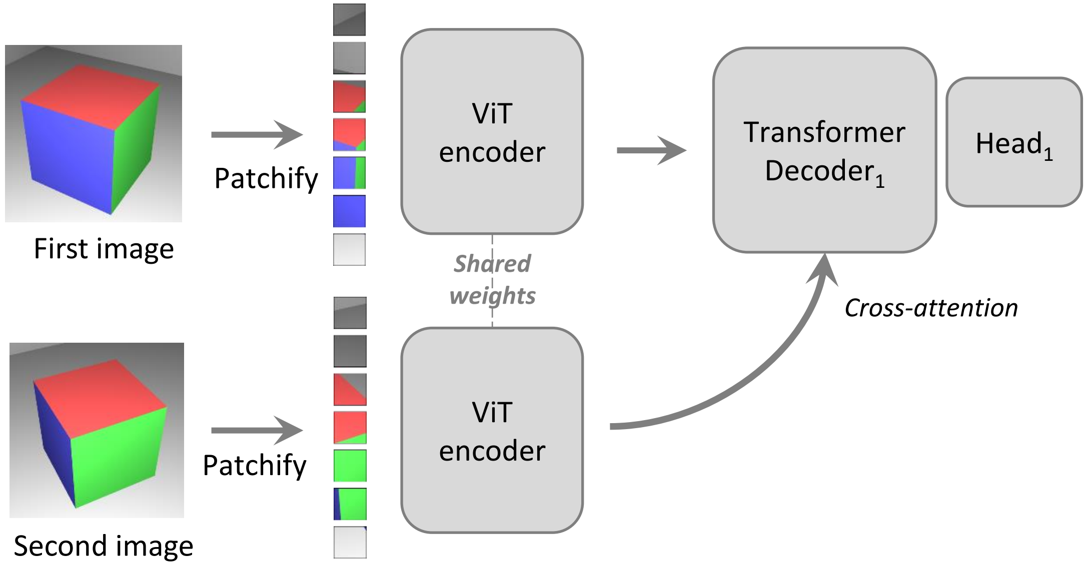
Show Details

0

ering

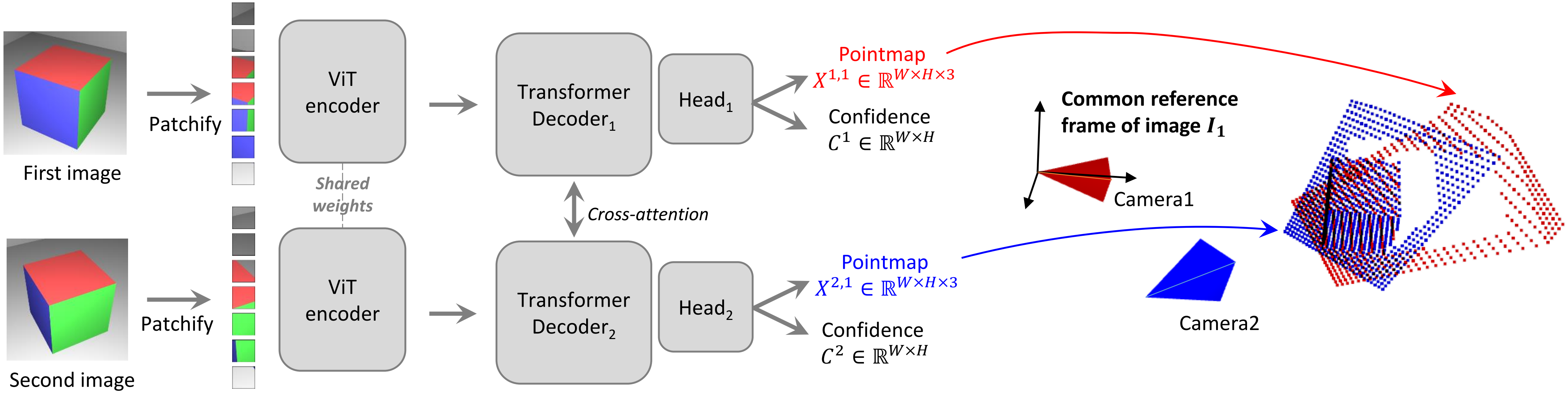
0.05

DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction



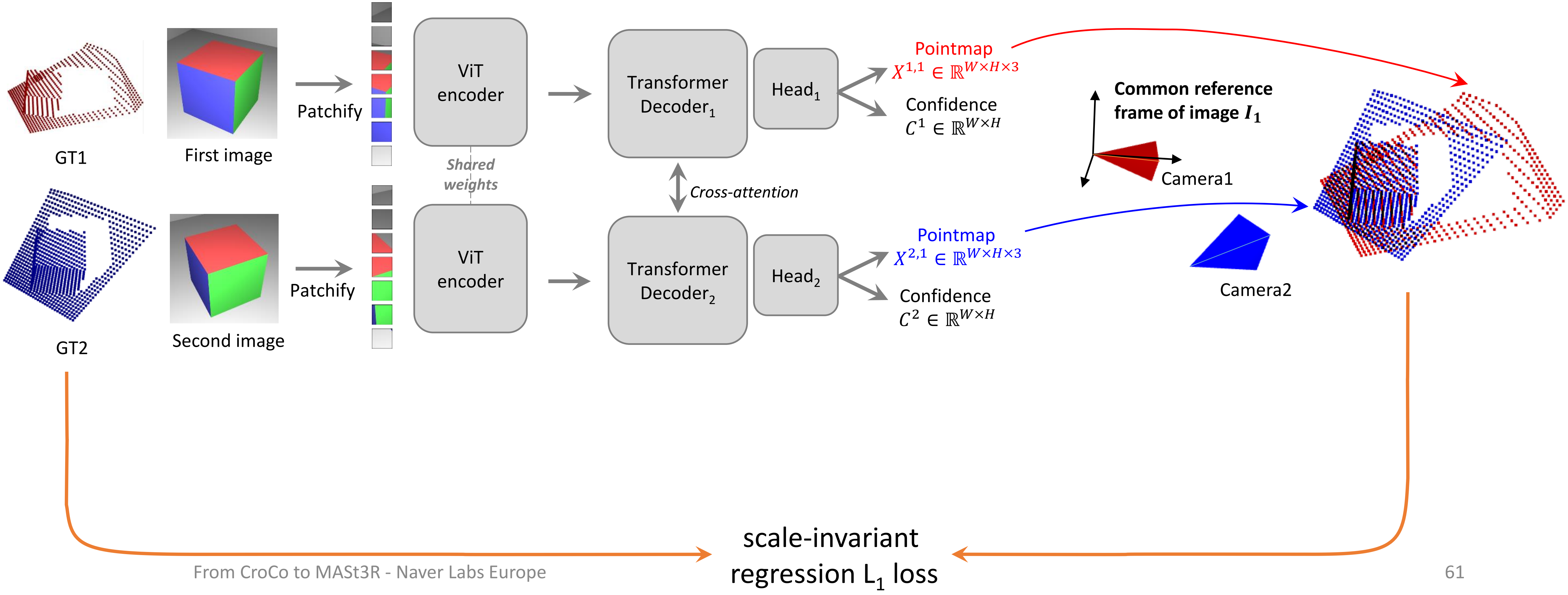
Start from CroCo …

DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction



Start from CroCo and add a 2nd decoder

DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction



DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction

- Training data

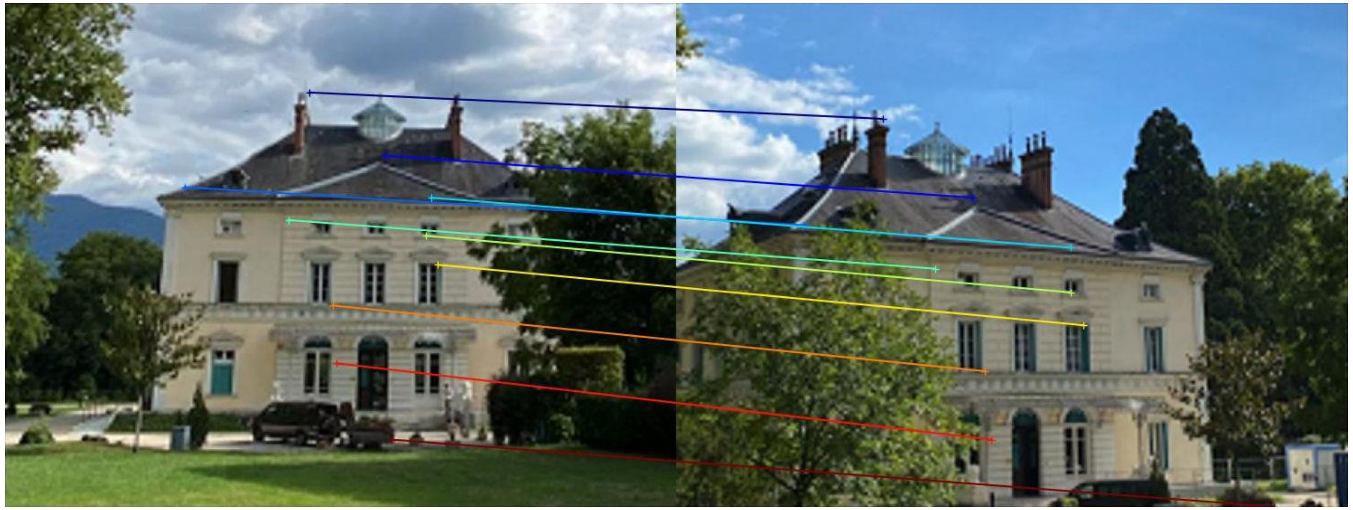
Datasets	Type	N Pairs
Habitat [103]	Indoor / Synthetic	1000k
CO3Dv2 [93]	Object-centric	941k
ScanNet++ [165]	Indoor / Real	224k
ArkitScenes [25]	Indoor / Real	2040k
Static Thing 3D [68]	Object / Synthetic	337k
MegaDepth [55]	Outdoor / Real	1761k
BlendedMVS [161]	Outdoor / Synthetic	1062k
Waymo [121]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

DUSt3R

Downstream Applications

1. **Point Matching**

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



DUSt3R
Downstream Applications

2. **Recovering intrinsics**

Assuming that the principal point is approximately centered, we can extract the focal lengths from the raw pointmaps.

<u>Despite the lack of any explicit priors, output pointmaps well respect the pinhole camera model!</u>

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%

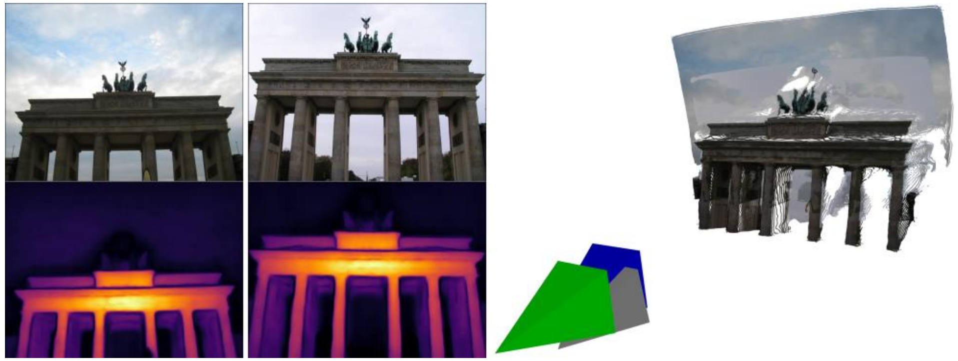
Left: Average absolute error of field-of-view (FoV) estimates.
Right: Average 2D reprojection accuracy (%) at the threshold of 1% of image diagonal.

DUSt3R

Downstream Applications

3. Visual Localization

Different ways of doing it, the most simple is from 2d correspondences + PnP



	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

DUSt3R
Global alignment

- A fast and simple post-processing optimization for multi-views (takes few seconds).

= a well-behaved 3D version of bundle adjustment

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

DUSt3R

Downstream Applications

4. Multi-view Pose Estimation

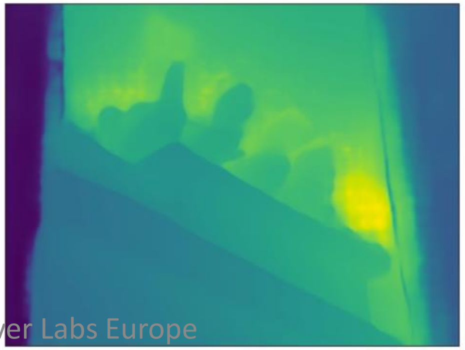
Methods	N Frames	Co3Dv2 [93]			RealEstate10K [185]
		RRA@15	RTA@15	mAA(30)	mAA(30)
COLMAP+SPSG	3	∼22	∼14	∼15	∼23
PixSfM	3	∼18	∼8	∼10	∼17
Relpose	3	∼56	-	-	-
PoseDiffusion	3	∼75	∼75	∼61	- (∼77)
DUSt3R 512	3	95.3	**88.3**	**77.5**	**69.5**
COLMAP+SPSG	5	∼21	∼17	∼17	∼34
PixSfM	5	∼21	∼16	∼15	∼30
Relpose	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

DUSt3R

Downstream Applications

5. Mono Depth Estimation

Methods	Train	Outdoor						Indoor			
		DDAD[33]		KITTI [29]		BONN [62]		NYUD-v2 [92]		TUM [94]	
		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[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

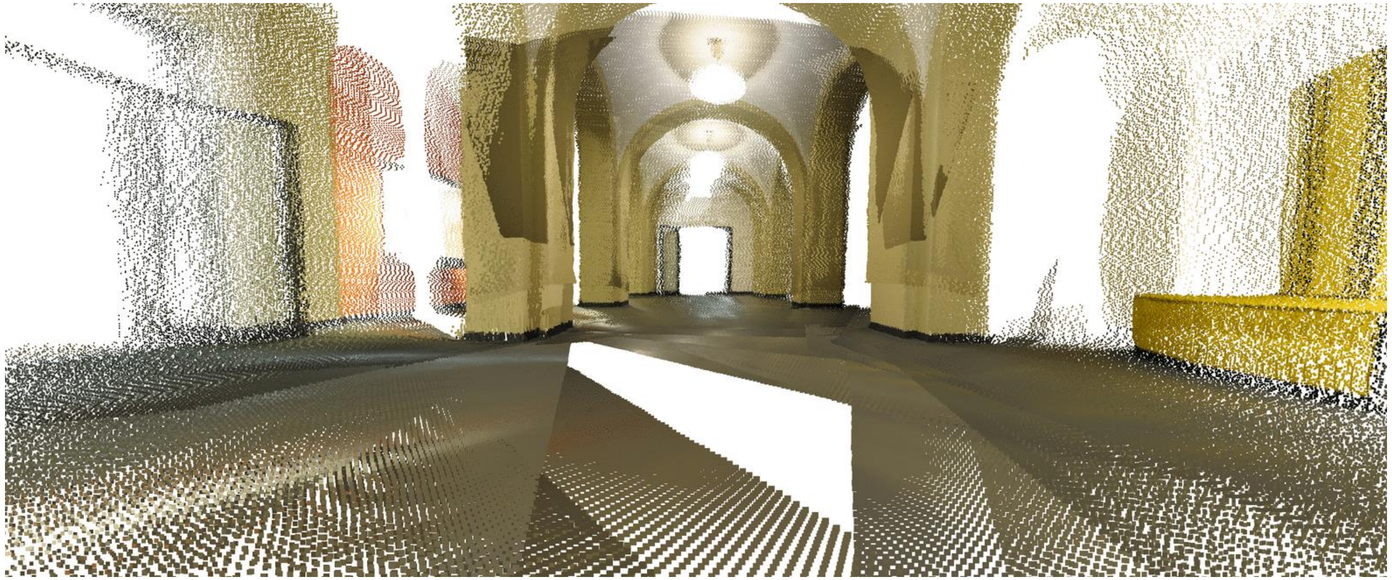
6. Multi-view Depth

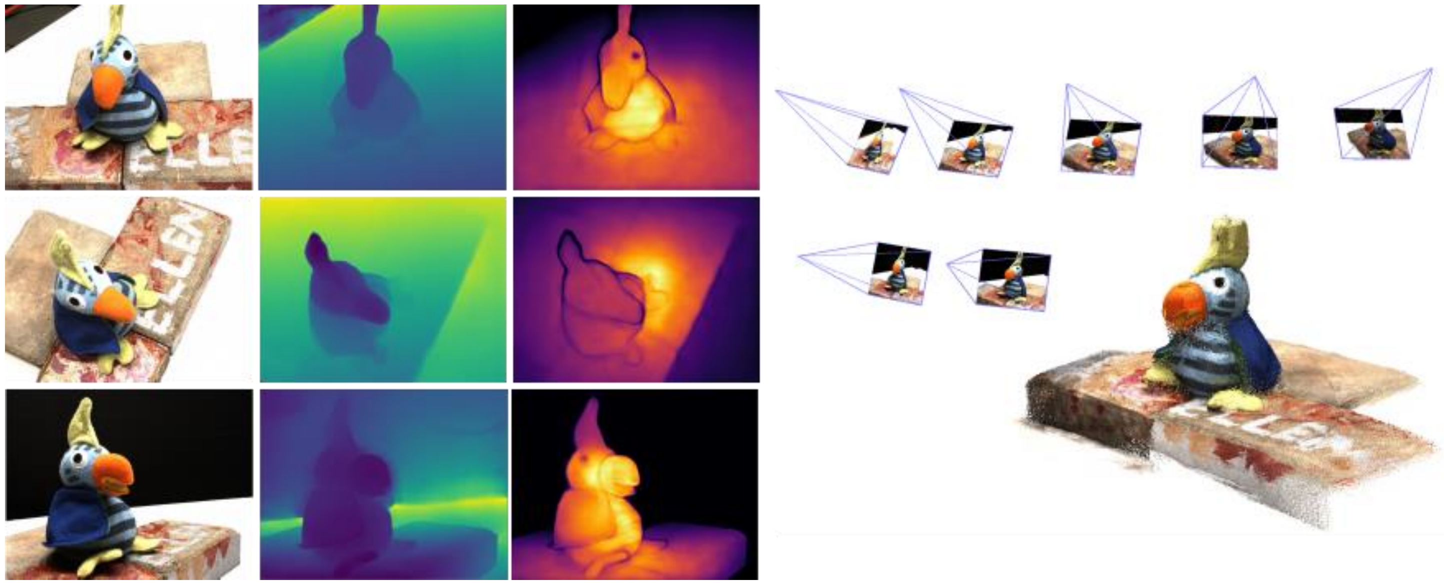
	Methods	GT Pose	GT Range	Align	KITTI		ScanNet		ETH3D		DTU		T&T		Average		
					rel ↓	$\tau\uparrow$	rel ↓	$\tau\uparrow$	rel ↓	$\tau\uparrow$	rel ↓	$\tau\uparrow$	rel↓	$\tau\uparrow$	rel↓	$\tau\uparrow$	time (s)↓
(a)	COLMAP [84, 85]	✓	×	×	**12.0**	**58.2**	**14.6**	**34.2**	**16.4**	**55.1**	**0.7**	**96.5**	**2.7**	95.0	**9.3**	**67.8**	≈ $3min$
	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	≈ $3min$
(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

DUSt3R

Downstream Applications

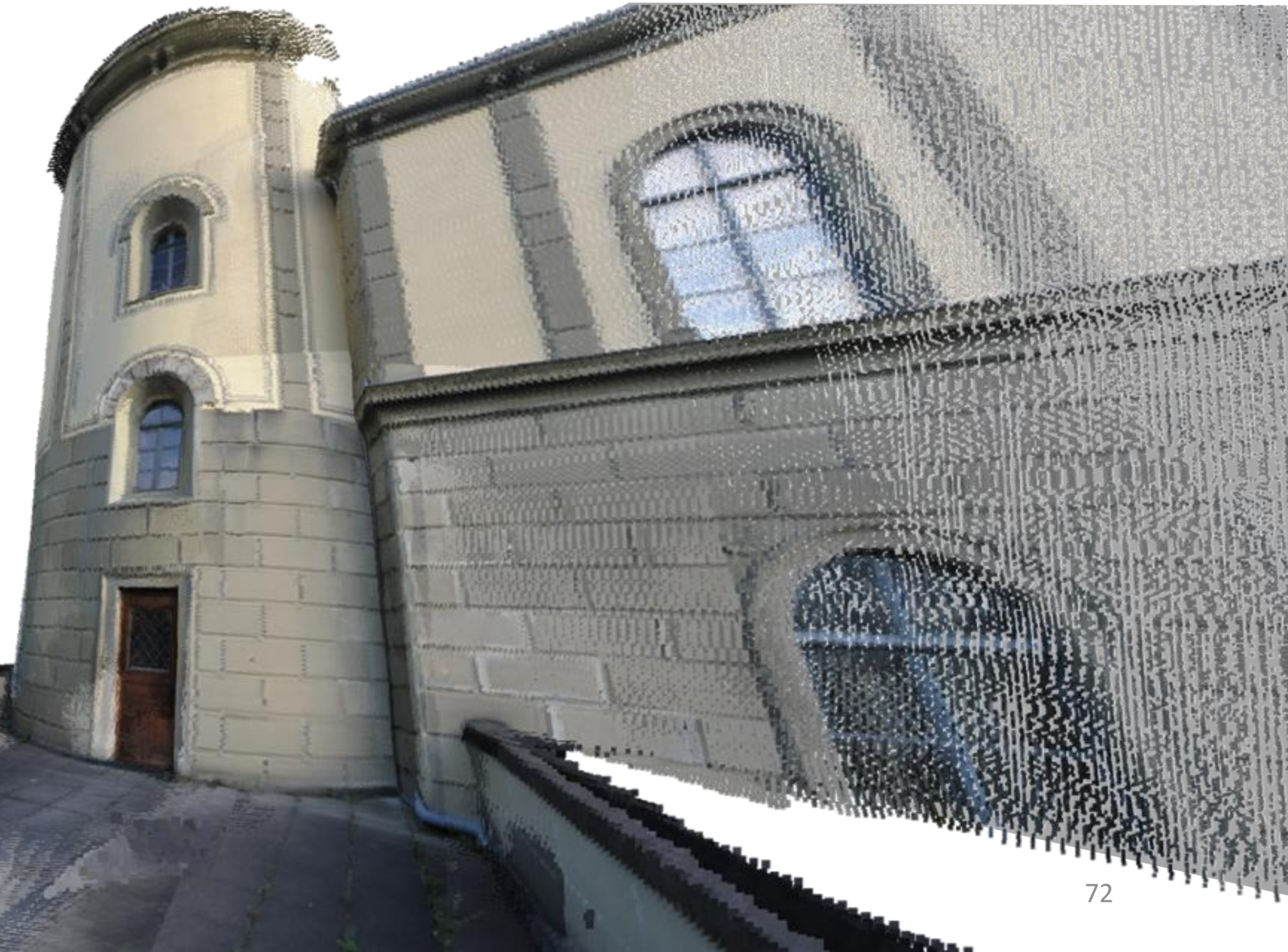
7. **3D Reconstruction**



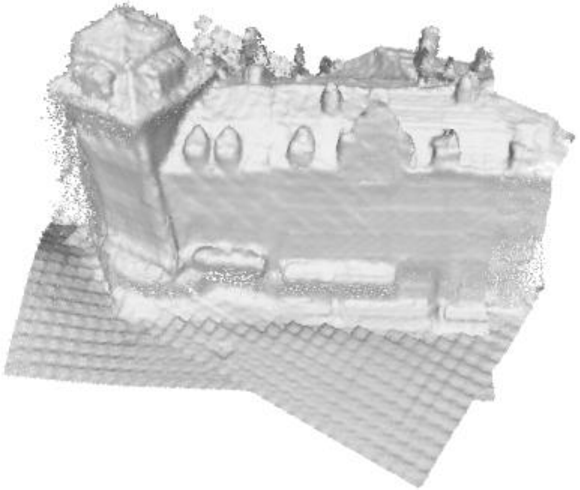


The same model works indoor …

… and outdoor



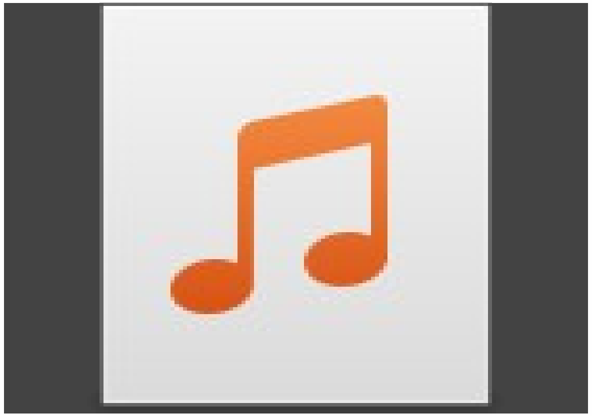
...on
large-scale outdoor scenes...

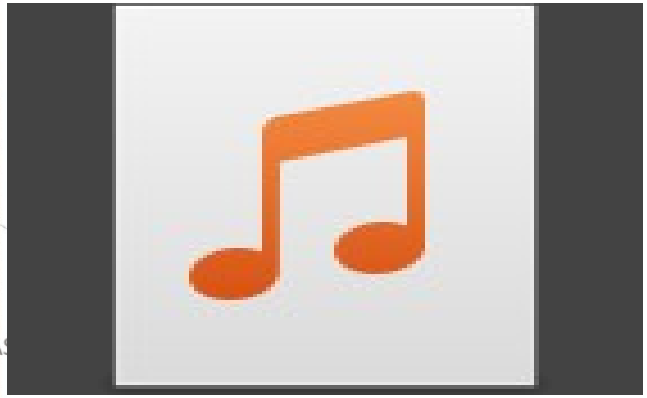


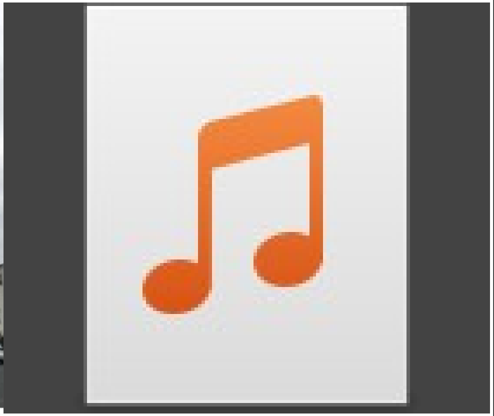
... and close-up objects ...

DUSt3R

Opposite view matching



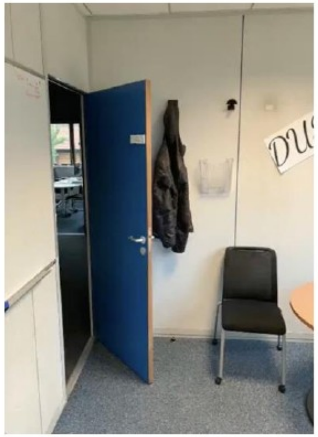




DUSt3R

“impossible matching”

3D reconstruction without any overlap!



https://github.com/naver/dust3r

DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction

Unifying all 3D vision tasks?

- **3D reconstruction is a "super-task"** ☺
  - intrinsically connected to all other 3DV tasks
- **Current solution is problematic** ☹
  - Brittle, requires enough *images* & *overlap* & *textures* & *viewpoints*
  - Heavily handcrafted at all levels
    - An engineering hell!
  - Multiple minimal problems solved sequentially
    - No internal collaboration between them
  - Slow

DUSt3R is:

Robust, works under any number of images, any overlap, any texture, any viewpoints

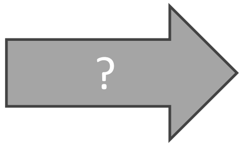
Simple, minimal handcrafting

Solves problems altogether

Fast! Takes a few seconds

DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction

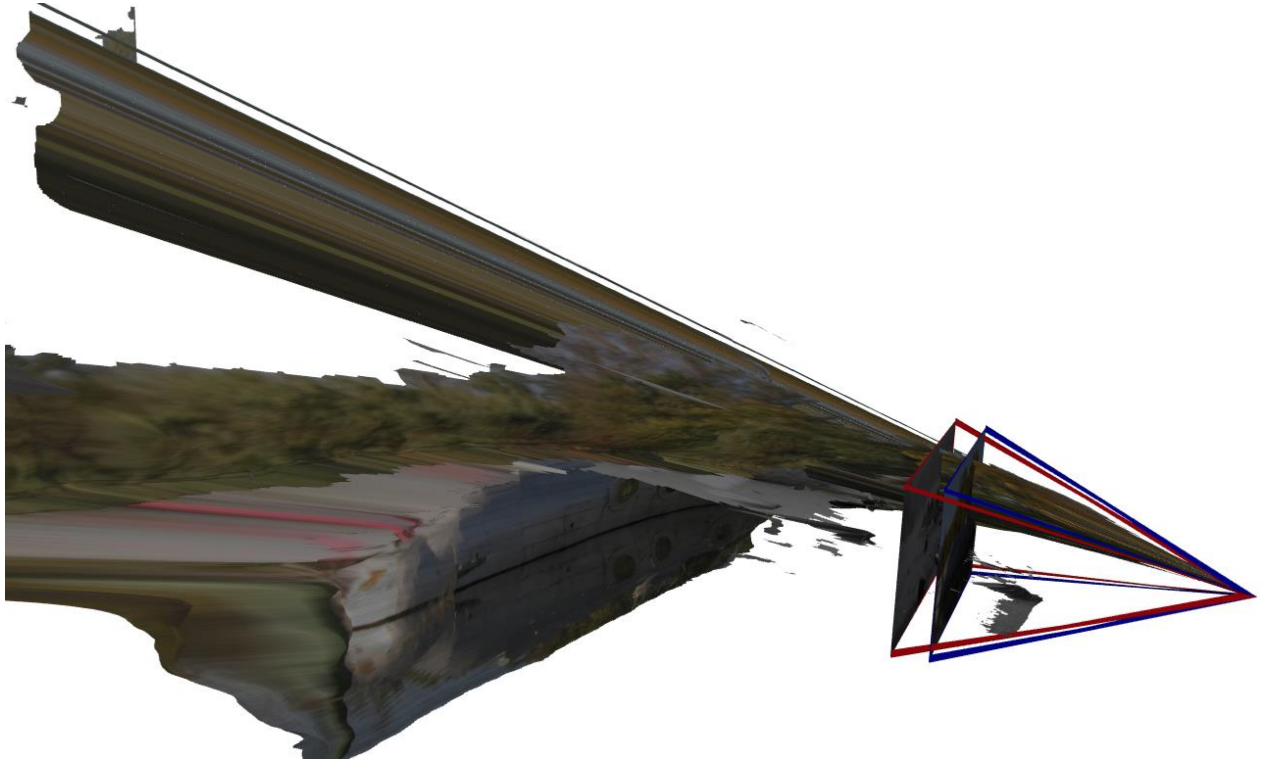
- Failure cases 1/3



DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction

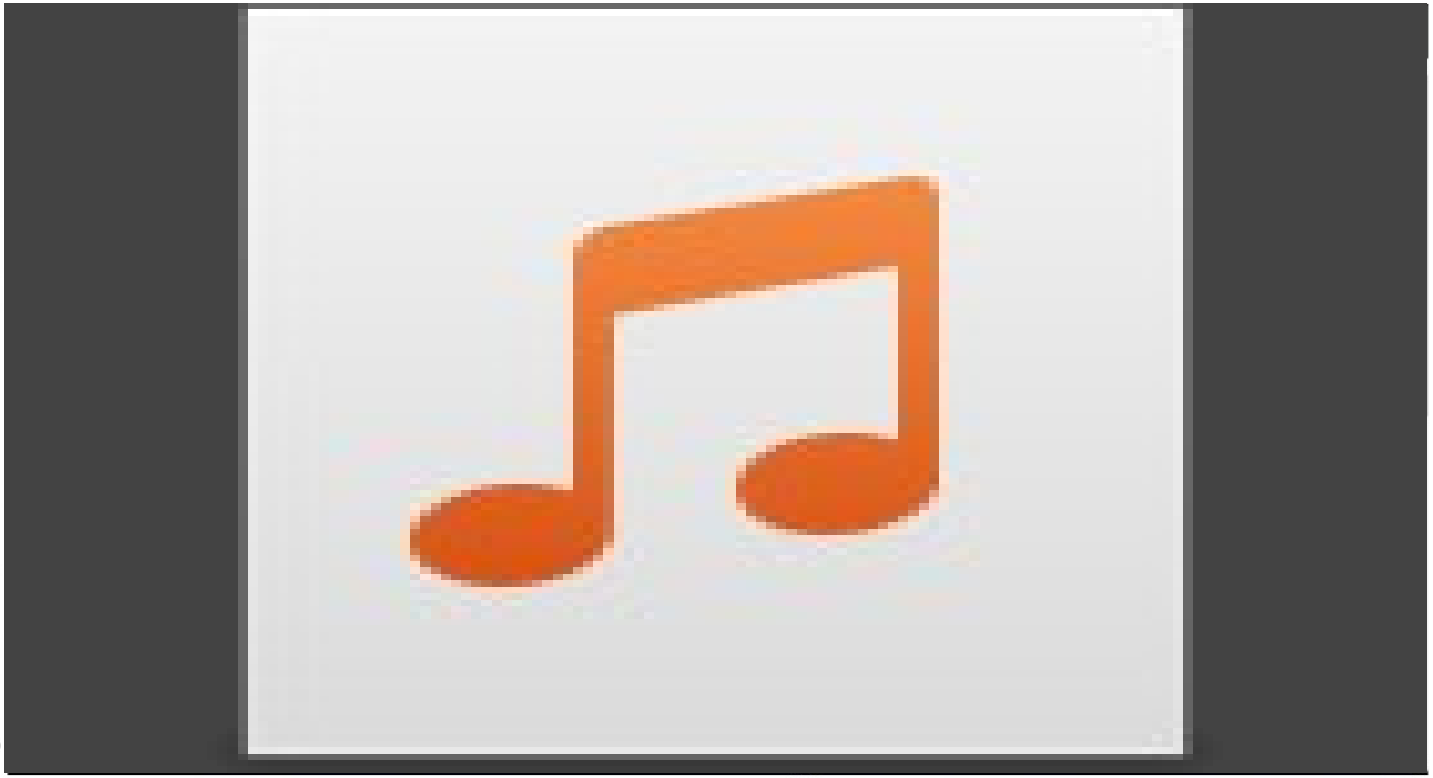
- Failure cases 2/3

?



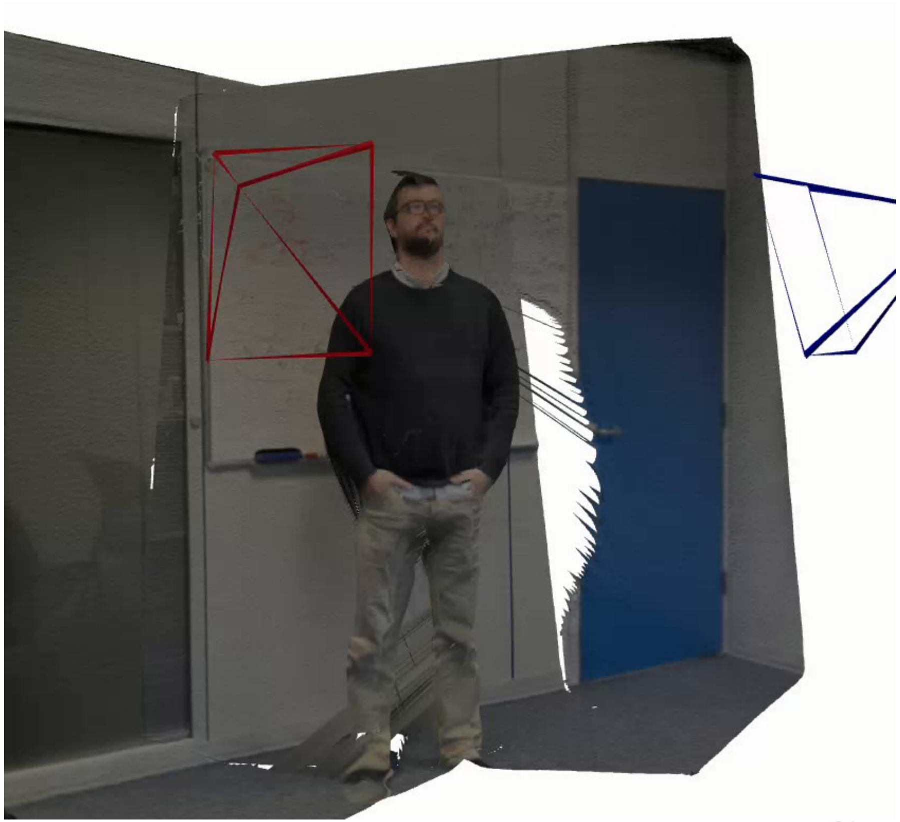
DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction

- Failure cases 3/3



DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction

- Generalization to OOD



DUSt3R: limitations

- Limitations of DUSt3R:
  - DUSt3R is extremely robust but lacks accuracy
    - 2 views only
    - intrinsic regression noise

	Methods	GT cams	Acc.↓	Comp.↓	Overall↓
Handcrafted (a)	Camp [11]	✓	0.835	0.554	0.695
	Furu [32]	✓	0.613	0.941	0.777
	Tola [100]	✓	0.342	1.190	0.766
	Gipuma [33]	✓	**0.283**	0.873	0.578
Learning Based (b)	MVSNet [121]	✓	0.396	0.527	0.462
	CVP-MVSNet [119]	✓	0.296	0.406	0.351
	UCS-Net [16]	✓	0.338	0.349	0.344
	CER-MVS [55]	✓	0.359	0.305	0.332
	CIDER [118]	✓	0.417	0.437	0.427
	PatchmatchNet [103]	✓	0.427	0.277	0.352
	GeoMVSNet [136]	✓	0.331	**0.259**	**0.295**
	DUSt3R 512	×	2.677	0.805	1.741

MVS benchmark on DTU

DUSt3R: limitations

- Best results often obtained from pixel correspondences

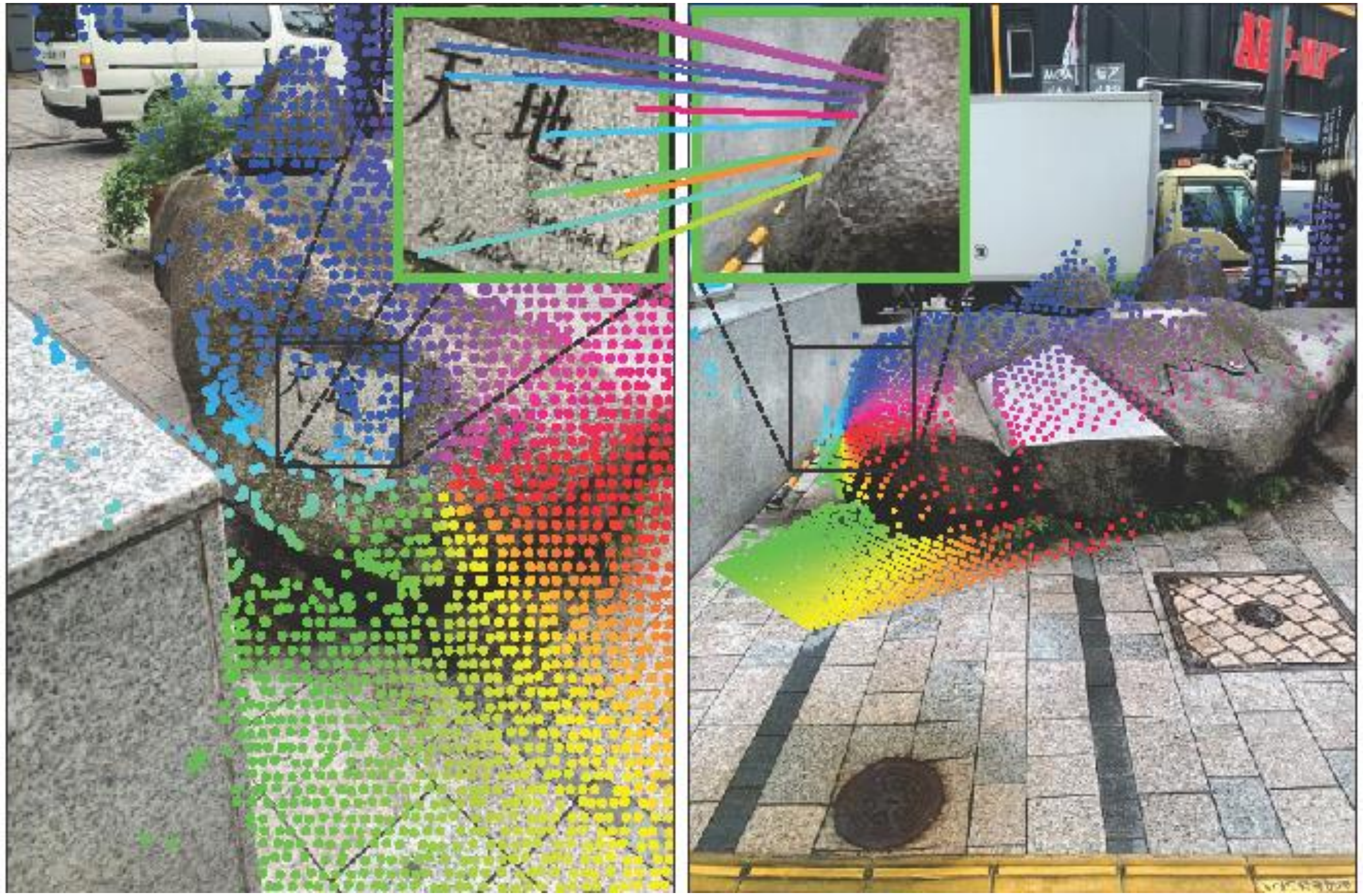
Methods	GT	7Scenes (Indoor) [48]							Cambridge (Outdoor) [14]				
	Focals	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	S. Facade	O. Hospital	K. College	St.Mary's	G. Court
DUSt3R 512 from 2D-matching	✓	3/0.97	3/0.95	2/1.37	**3**/1.01	**4**/1.14	4/1.34	11/2.84	6/0.26	17/0.33	**11/0.20**	**7/0.24**	38/0.16
DUSt3R 512 from scaled rel-pose	×	5/1.08	5/1.18	4/1.33	6/1.05	7/1.25	6/1.37	26/3.56	64/0.97	151/0.88	102/0.88	79/1.46	245/1.08

- But not trained explicitly for matching
  - What if we did?

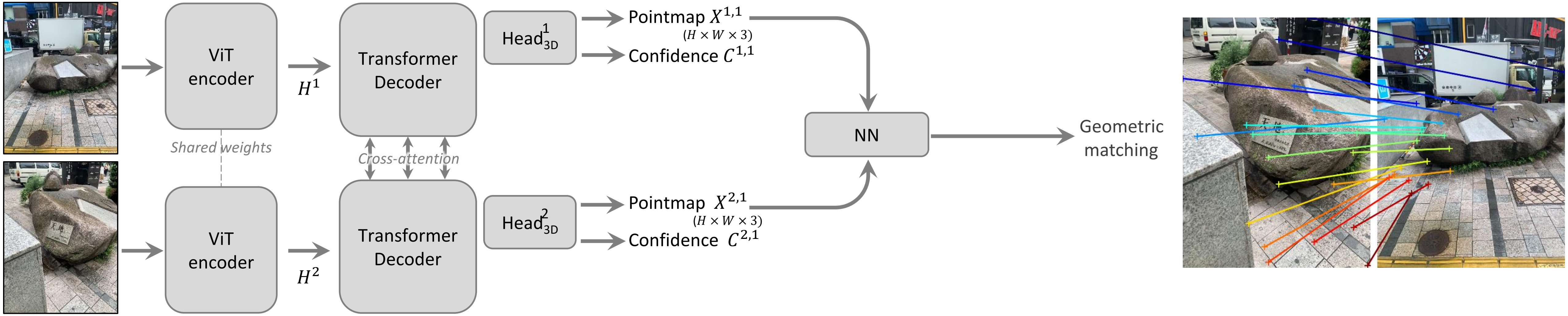
Grounding Image Matching in 3D with MASt3R



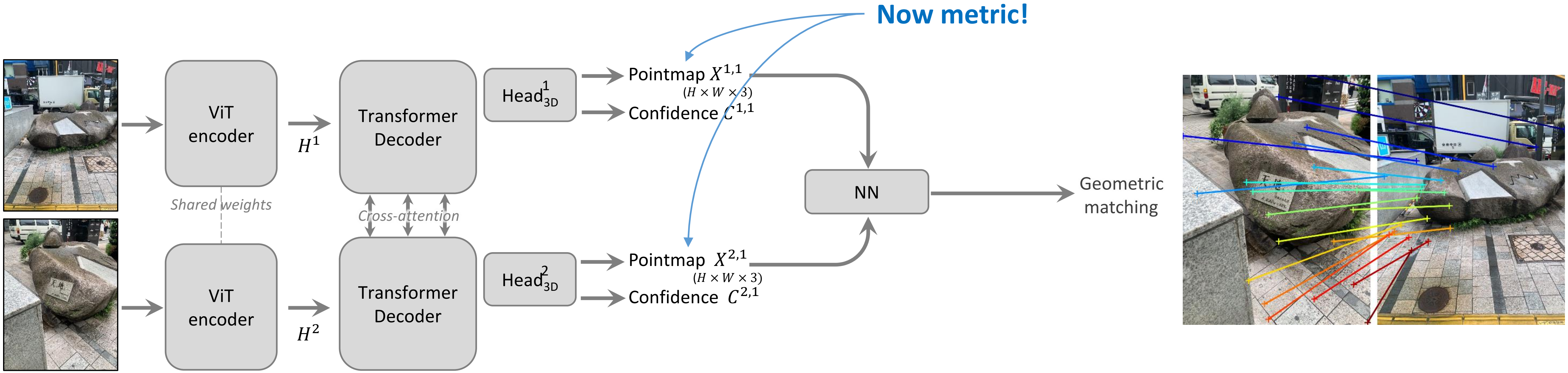
Vincent LEROY, Yohann CABON, Jerome REVAUD



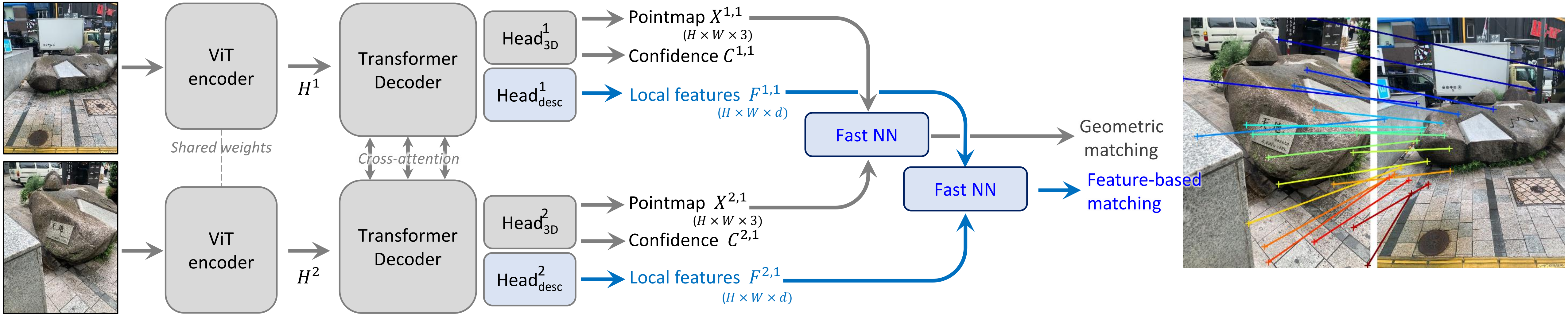
MASt3R:
Matching And Stereo 3D Reconstruction



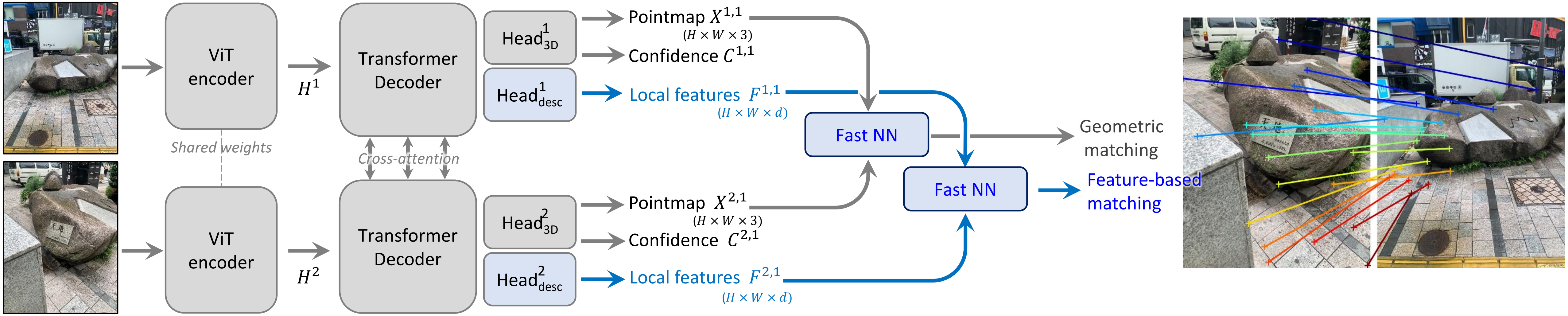
MASt3R: Matching And Stereo 3D Reconstruction



MASt3R:
Matching And Stereo 3D Reconstruction



MASt3R: Matching And Stereo 3D Reconstruction



Local Features trained with an InfoNCE loss

MASt3R: Matching And Stereo 3D Reconstruction

- **Training data:**
  - 14 datasets (Habitat, ARKitScenes, BlendedMVS, MegaDepth, Static Scenes 3D, ScanNet++, CO3D-v2, Waymo, Map-free, Wild-rgb, Virtual KITTI, Unreal4K, TartanAir, and internal data)
  - **50/50 synthetic and real scenes**
  - 10 datasets have metric GT

- Init from pretrained DUSt3R (same architecture)

MASt3R: Map-Free Relocalization

R^{12}, T^{12} ?

Image 1

Image 2

Translation is metric ➔ Pixel matching alone does not suffice

MASt3R: Map-Free Relocalization

Image 1

R^{12}, T^{12} ?

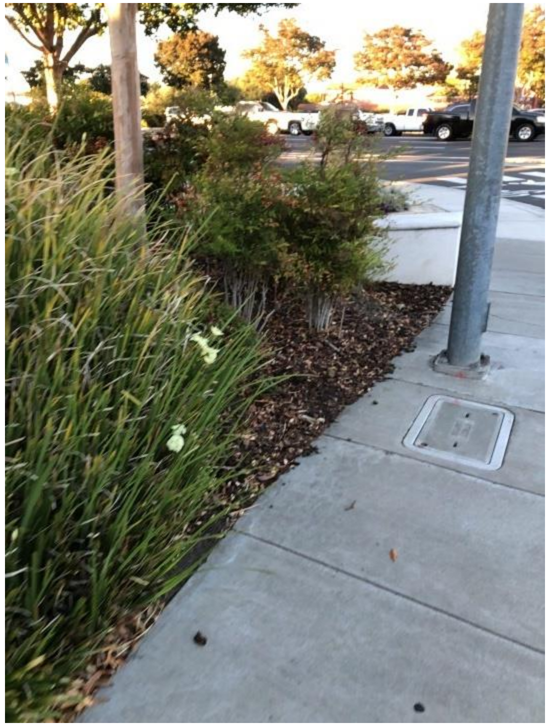
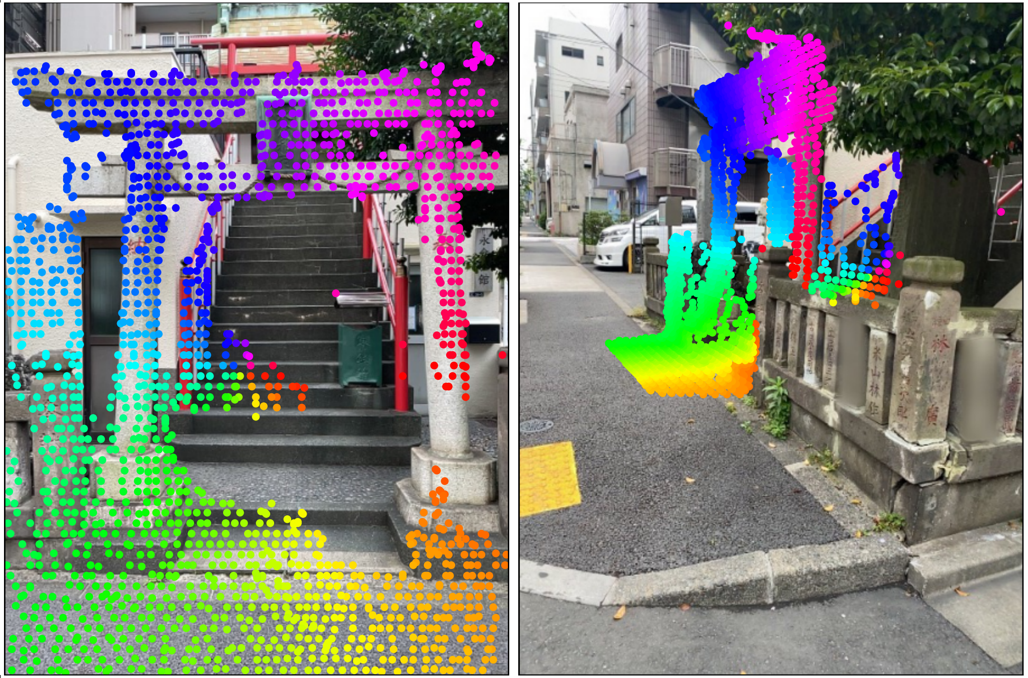


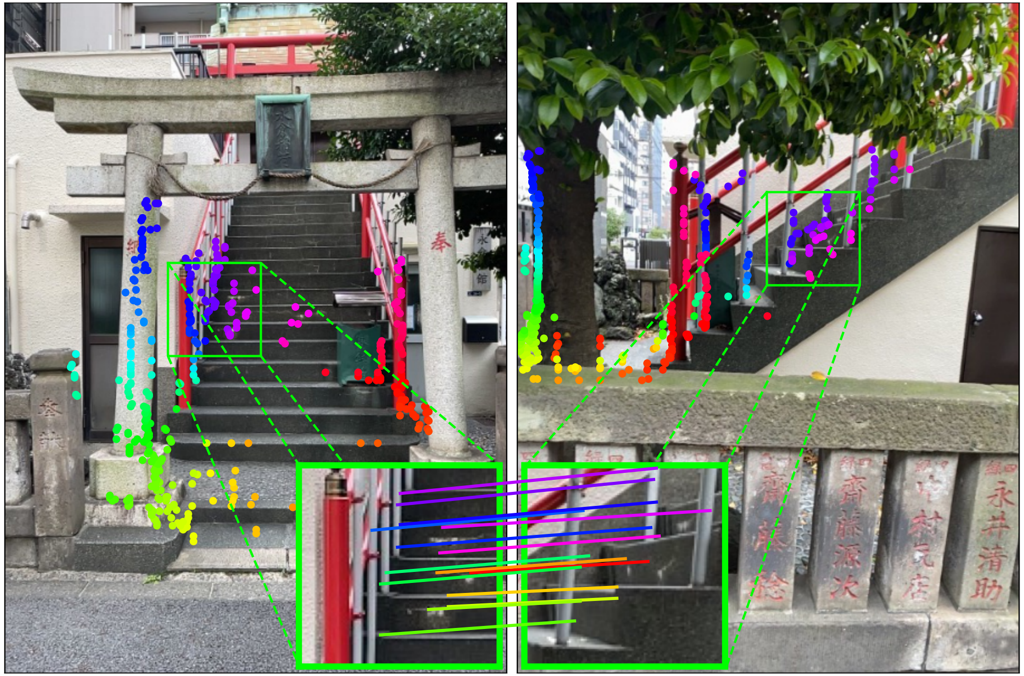
Image 2

Almost no overlap ➔ Pixel matching alone does not suffice

MASt3R: Map-Free Relocalization

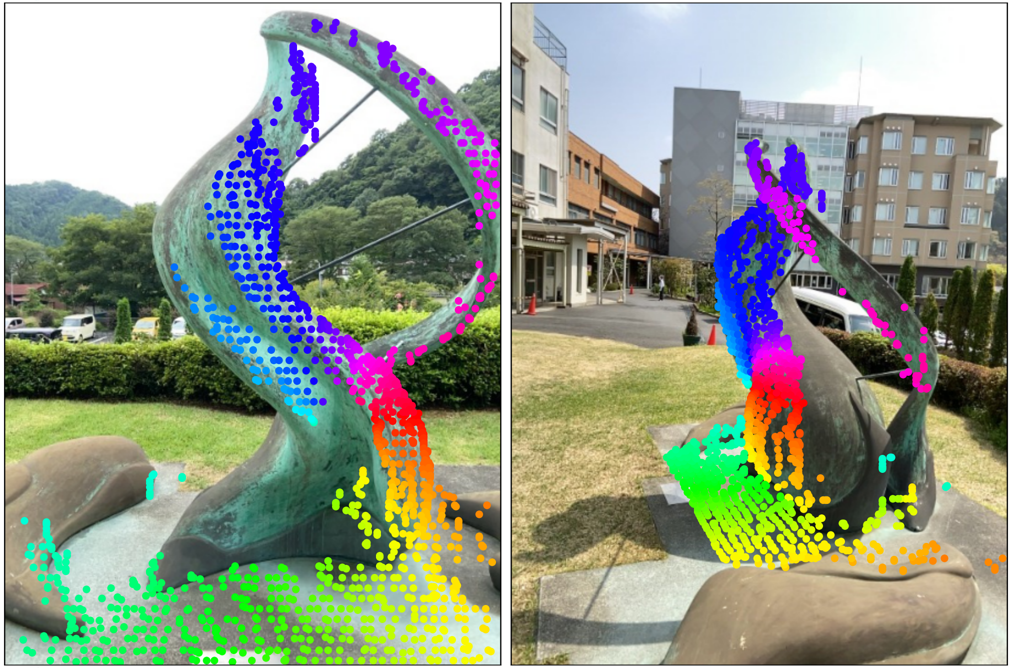


60° viewpoint change

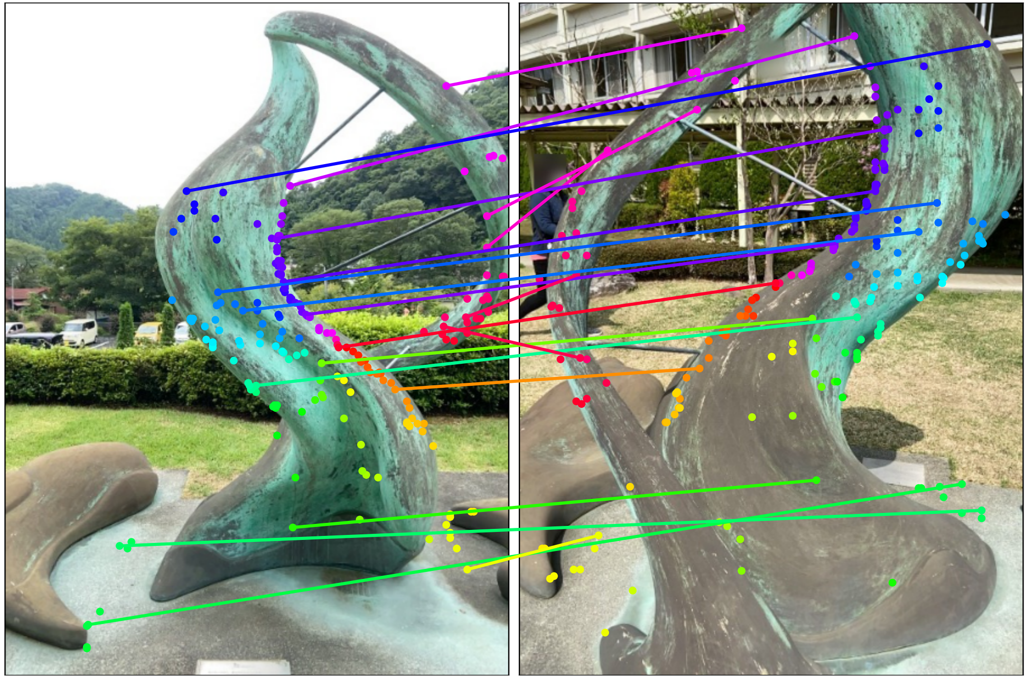


90° viewpoint change

MASt3R: Map-Free Relocalization



90° viewpoint change



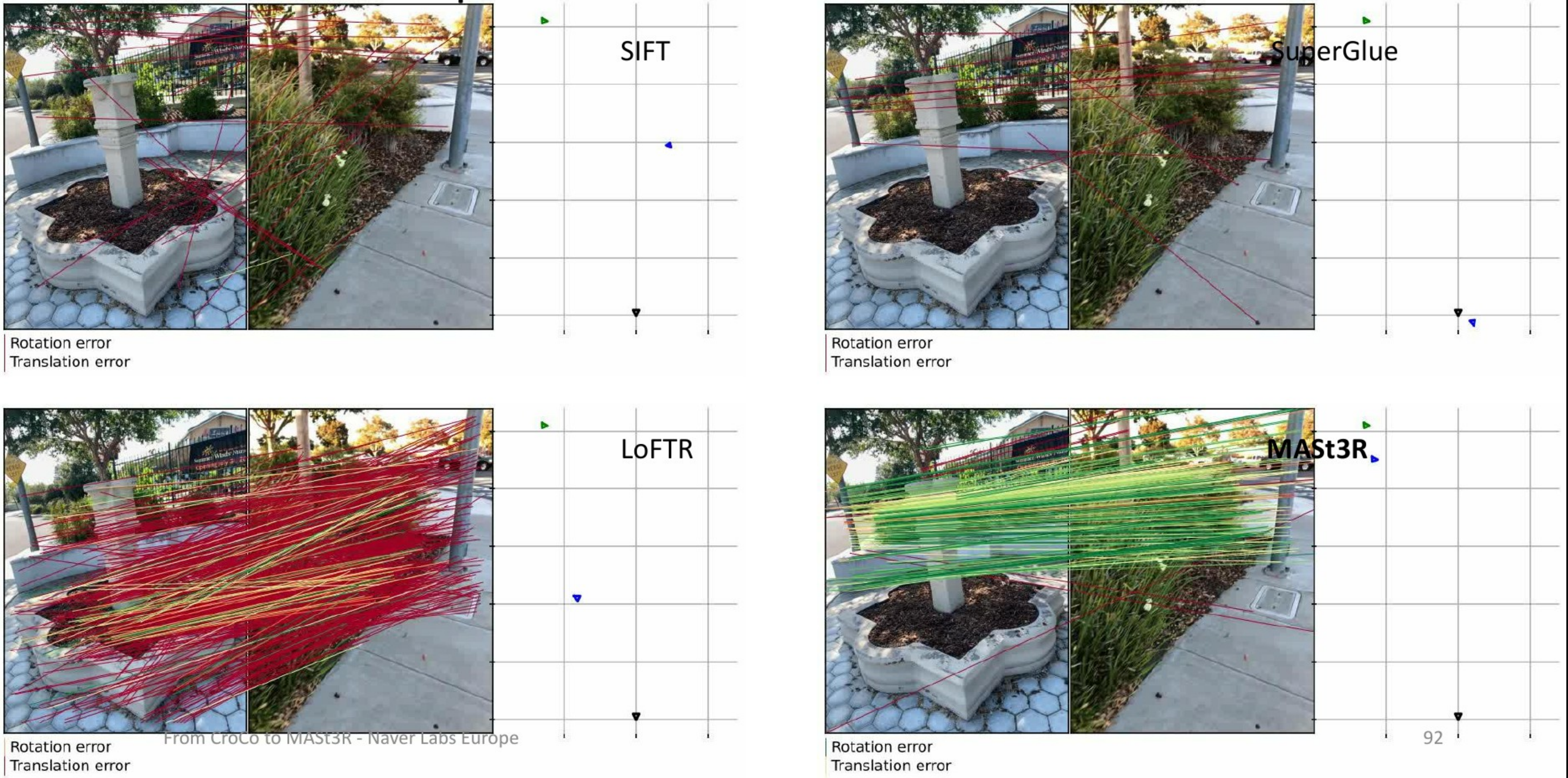
180° viewpoint change

MASt3R: Map-Free Relocalization

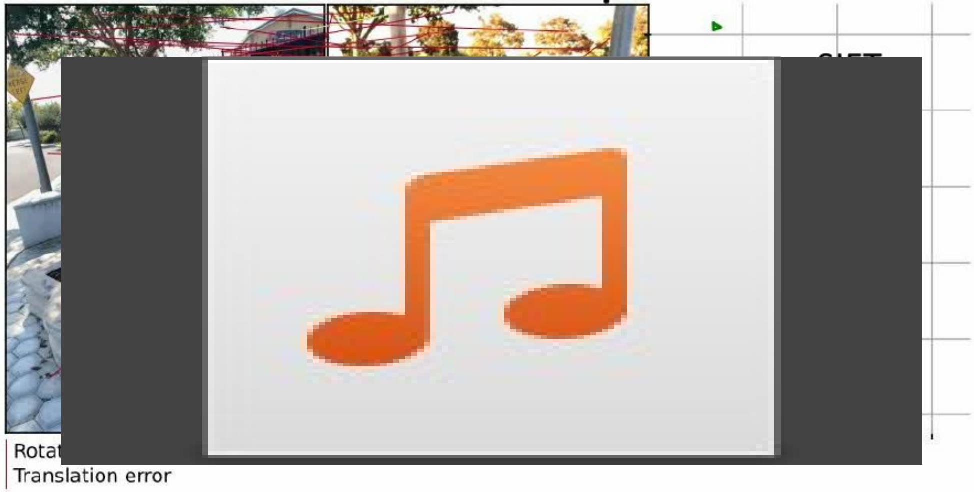
45° viewpoint change

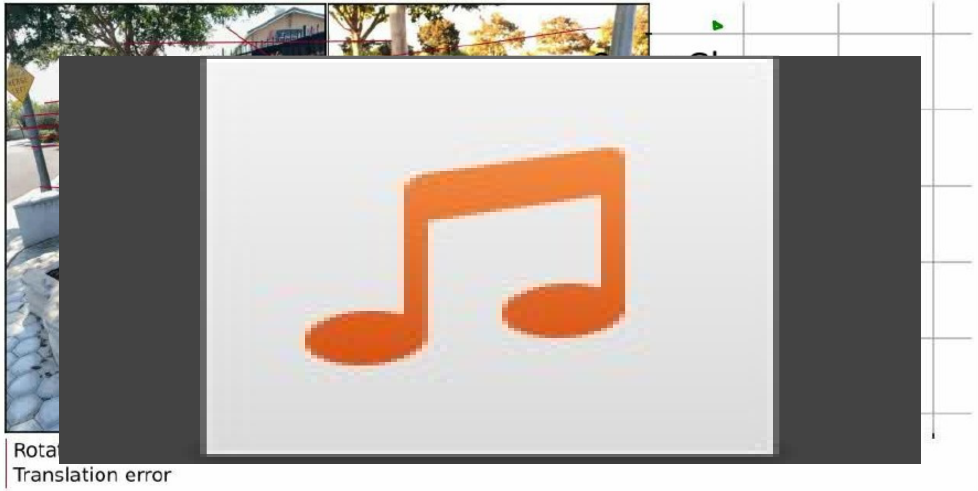
100° viewpoint change,
symmetrical object

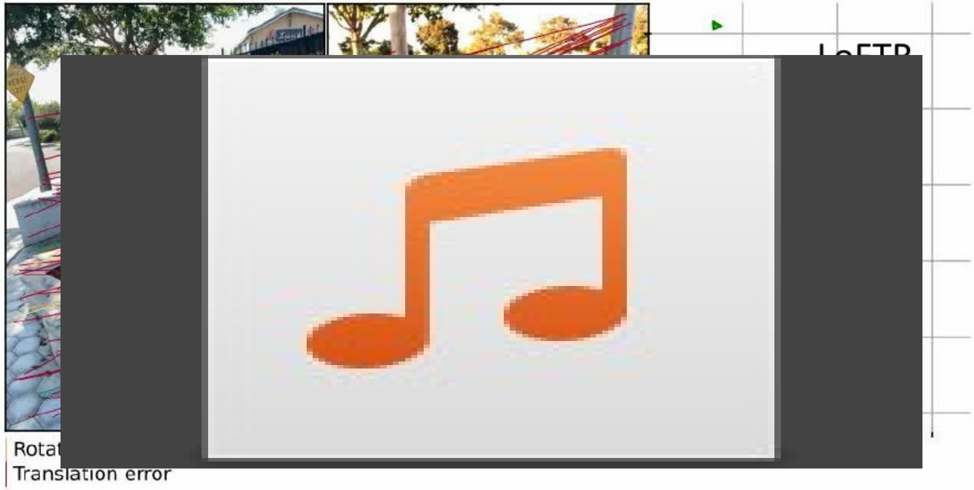
MASt3R: Map-Free Relocalization

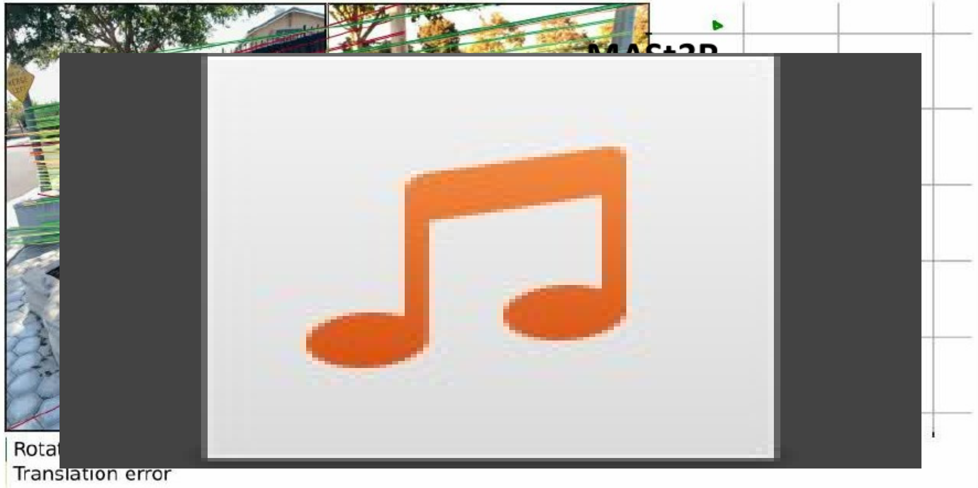


MASt3R: Map-Free Relocalization









MASt3R: Map-Free Relocalization

Evaluation Leaderboard

	AUC (VCRE < 90px)	Precision (VCRE < 90px)	Median Reproj. Error (px)	AUC (Err < 25cm, 5°)	Precision (Err < 25cm, 5°)	Median Trans. Error (m)	Median Rot. Error (°)
MASt3R	0.933	79.3% !!	48.7	0.740	54.7% !!	0.36 !!	2.2 !!
(CVPR'24) **Mickey w/ overlap**	0.747	49.2%	129.4	0.325	13.3%	1.65	27.2
(CVPR'24) **Mickey**	0.740	49.2%	126.9	0.283	12.0%	1.59	25.9
(CVPR'24) **DUSt3R**	0.697	50.3%	115.8	0.393	21.4%	0.98	7.1
(arXiv'24) FAR (LoFTR)	0.680	44.2%	137.0	0.392	17.7%	1.48	17.2
(CVPR'24) RoMa	0.669	45.6%	128.8	0.407	22.8%	1.23	11.1
(arXiv'24) FAR (SuperGlue)	0.668	44.1%	135.4	0.351	17.1%	1.49	17.2
KBR++ & LoFTR	0.634	34.3%	165.0	0.295	11.0%	2.23	37.8
KBR & LoFTR	0.631	34.0%	167.5	0.277	10.5%	2.32	39.5
(PnP) LoFTR	0.618	33.4%	166.7	0.269	9.8%	2.31	39.4
LoFTR	0.614	34.7%	167.6	0.346	15.4%	1.98	30.5
SuperGlue	0.602	36.1%	160.3	0.346	16.8%	1.88	25.4
(PnP) SuperGlue	0.598	36.0%	156.9	0.252	10.7%	2.10	32.5
SIFT	0.504	25.0%	222.8	0.252	10.3%	2.93	61.4
(PnP) SIFT	0.468	25.1%	192.0	0.190	7.8%	3.35	63.7

MASt3R: Relative Camera Pose

10 views

2 views

	Methods	Co3Dv2			RealEstate10K
		RRA@15	RTA@15	mAA(30)	mAA(30)
(a)	Colmap+SG [20, 73]	36.1	27.3	25.3	45.2
	PixSfM [49]	33.7	32.9	30.1	49.4
	RelPose [114]	57.1	-	-	-
	PosReg [98]	53.2	49.1	45.0	-
	PoseDiff [98]	80.5	79.8	66.5	48.0
	RelPose++ [48]	(85.5)	-	-	-
	RayDiff [113]	(93.3)	-	-	-
	DUSt3R-GA [100]	**96.2**	86.8	76.7	67.7
(b)	DUSt3R [100]	94.3	88.4	77.2	61.2
	MASt3R	94.6	**91.9**	**81.8**	**76.4**

High accuracy, even in a low-view regime!

MASt3R: Visual Localization

Methods	InLoc [83]	
	DUC1	DUC2
SP+SuperGlue [73]	49.0/68.7/80.8	53.4/77.1/82.4
SP+LightGlue [50]	49.0/68.2/79.3	55.0/74.8/79.4
LoFTR [81]	47.5/72.2/84.8	54.2/74.8/85.5
DKM [26]	51.5/75.3/86.9	63.4/82.4/87.8
DUSt3R top1 [100]	36.4/55.1/66.7	27.5/42.7/49.6
DUSt3R top20 [100]	53.0/74.2/89.9	61.8/77.1/84.0
MASt3R top1	41.9/64.1/73.2	38.9/55.7/62.6
MASt3R top20	55.1/77.8/90.4	**71.0**/84.7/89.3
MASt3R top40	**56.1/79.3/90.9**	**71.0/87.0/91.6**

Using precomputed maps, varying number of retrieved images.

MASt3R: MVS on DTU

MASt3R: MVS on DTU

Architecture and network are not task-specific: we simply triangulate matches in 3D

		Methods	Acc.↓	Comp.↓	Overall↓
Handcrafted	(c)	Camp [13]	0.835	0.554	0.695
		Furu [30]	0.613	0.941	0.777
		Tola [89]	0.342	1.190	0.766
		Gipuma [31]	**0.283**	0.873	0.578
In-domain Train on DTU	(d)	MVSNet [108]	0.396	0.527	0.462
		CVP-MVSNet [107]	0.296	0.406	0.351
		UCS-Net [17]	0.338	0.349	0.344
		CER-MVS [54]	0.359	0.305	0.332
		CIDER [105]	0.417	0.437	0.427
		PatchmatchNet [97]	0.427	0.277	0.352
		GeoMVSNet [116]	0.331	**0.259**	**0.295**
OOD Never seen before	(e)	DUSt3R [100]	2.677	0.805	1.741
		MASt3R	0.403	0.344	0.374

(in *mm* !)

➔ Matching is far superior to regression

MASt3R : MVS on DTU

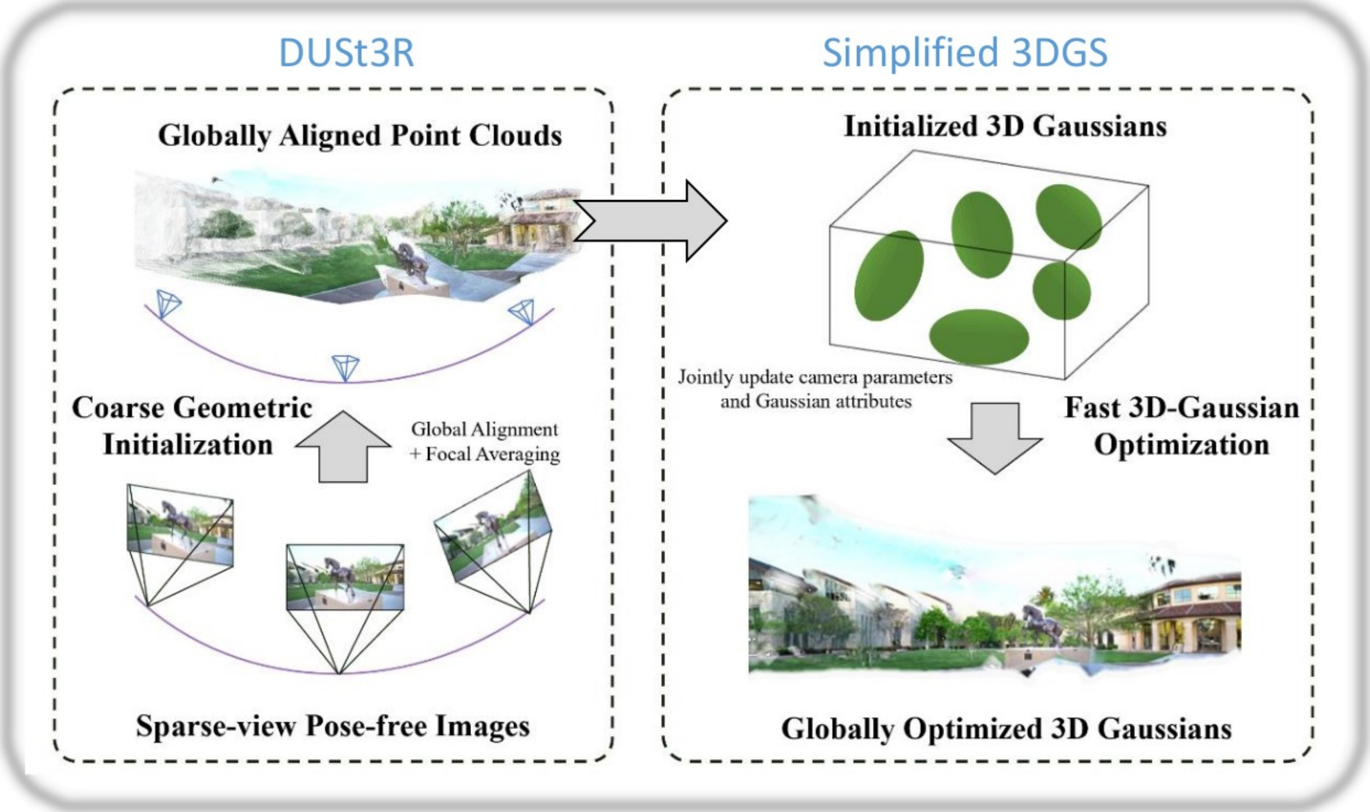
- Fun fact

MASt3R: summary

- Advantages over DUSt3R
  - accurate correspondences, even in extreme cases
  - accurate camera poses, possibly metric
  - coarse metric geometry

- Limitations
  - Data-driven ➔ only good on what it knows
  - Not good for long-term changes (e.g. snowy vs. sunny)
  - No semantic (yet)

InstantSplat:
Novel View Rendering *from scratch* in seconds

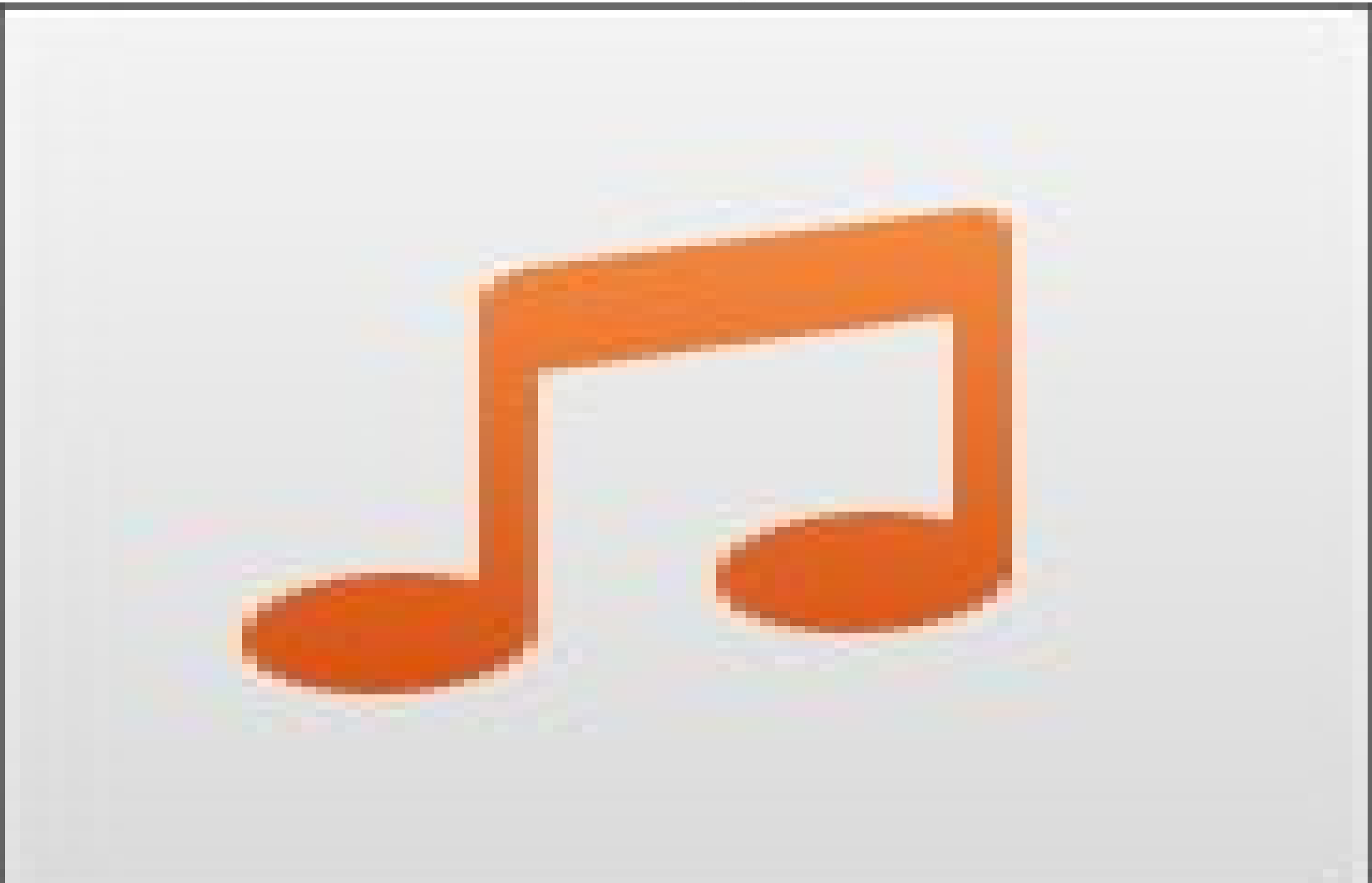


InstantSplat: Unbounded Sparse-view Pose-free Gaussian Splatting in 40 Seconds. Zhiwen Fan et. al., [arXiv'24]. https://instantsplat.github.io/

InstantSplat

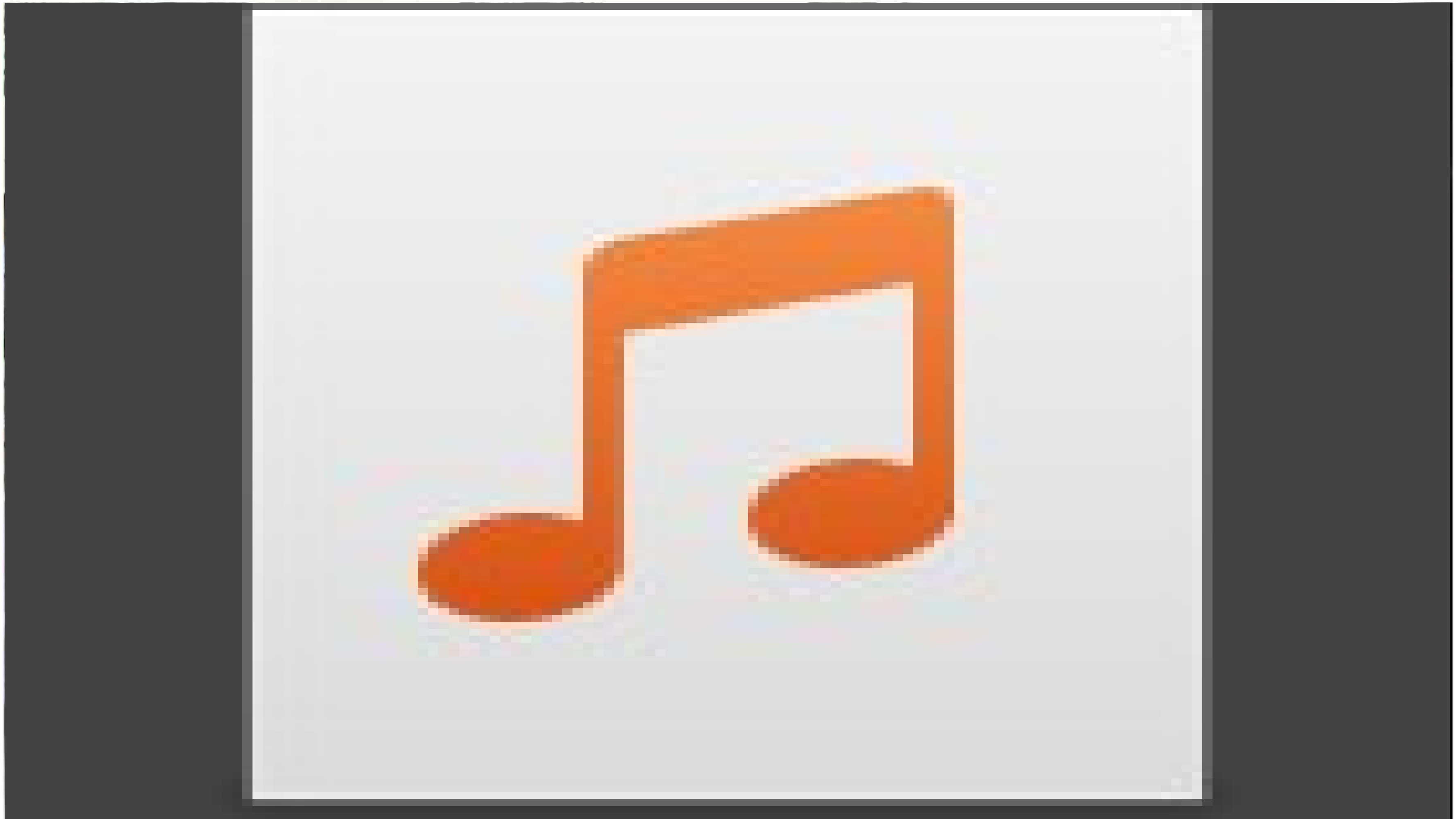
Global scene optimization via photometric loss and 3DGS

- Novel View Rendering **in 40 seconds!! (from scratch,** no pose, no intrinsics)
- Concurrent works require 200+ views and hours of optimization



InstantSplat

Result with only 3 input images in 20 seconds from scratch



Conclusion

- Dense 3D reconstruction
  - in a single step
  - without pose nor intrinsics

    ➔ indeed a paradigm change!

- DUSt3R/MASt3R is a universal model of 3D vision tasks
  - For the 1st time, unifying monocular & binocular depth estimation!
  - The pointmap representation looks obvious retrospectively ☺

- Simple, neat and fast!

What’s next?

- To-do list:
  - Large-scale
  - Handle lens distortion explicitly
  - Semantic MASt3R
  - Novel View rendering
  - SLAM
  - Dynamic scenes

- We tried, it works:
  - Symmetric decoder (instead of asymmetric right now)
  - DUSt3R with diffusion, but more costly

Thank you!

- Happy to take any question ☺