NAVER LABS Europe

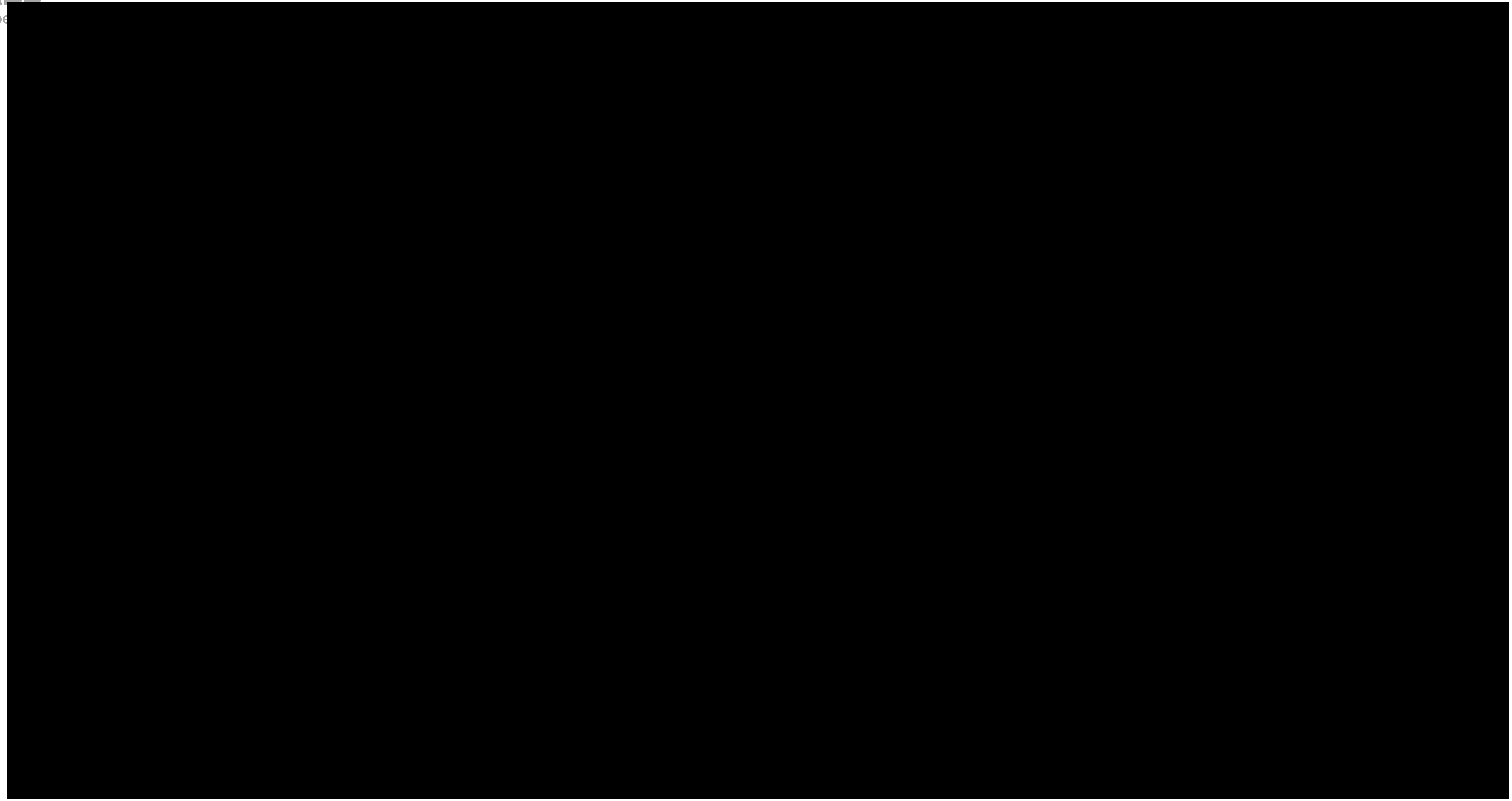
NAVER

Grounding Image Matching in 3D with

MAST3R

Vincent Leroy

Map-Free Workshop ECCV’24



Overview

Introduction

CroCo: Self-Supervised pretraining for 3DV

DUSt3R: towards a unified 3DV model

MASt3R: grounding matching in 3D

Conclusion

What is 3D vision?

What is 3D vision?

Monocular Depth estimation

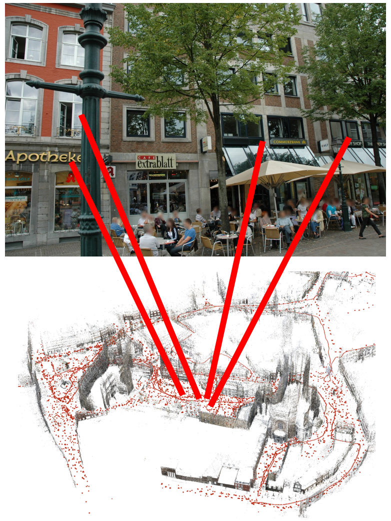
What is 3D vision?

Monocular Depth estimation

Large-scale 3D reconstruction

What is 3D vision?

Monocular Depth estimation

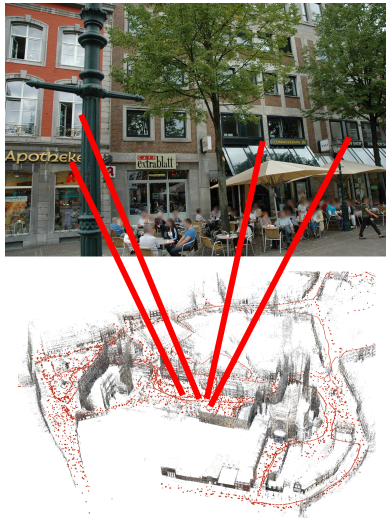


Point matching

Large-scale 3D reconstruction

What is 3D vision?

Monocular Depth estimation



Point matching

Large-scale 3D
reconstruction

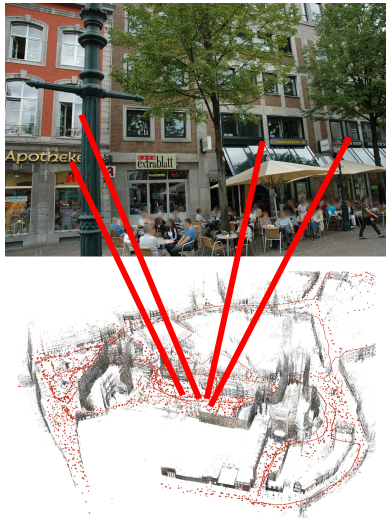
Visual
Localization

NAVER LABS Europe

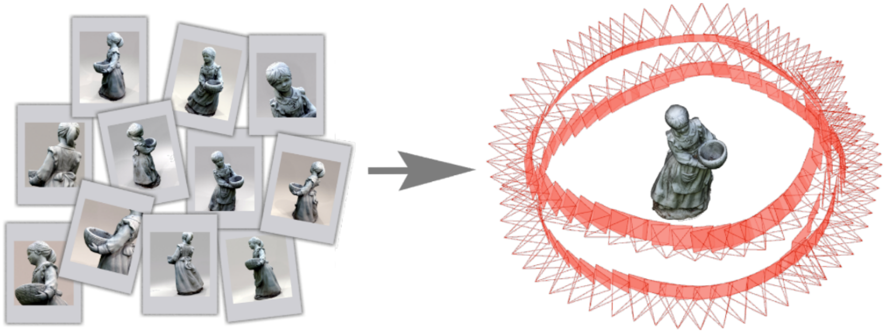
NAVER

What is 3D vision?

Monocular Depth estimation



Point matching



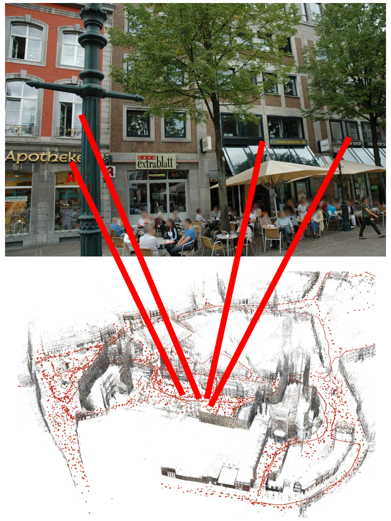
Multi-view pose estimation

Large-scale 3D reconstruction

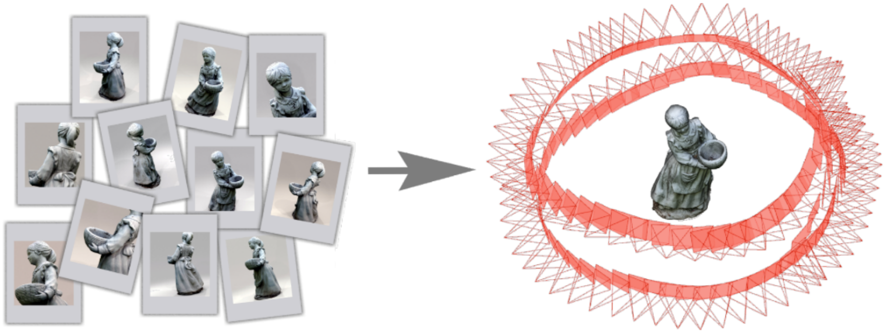
Visual Localization

What is 3D vision?

Monocular Depth estimation



Point matching



Multi-view pose estimation

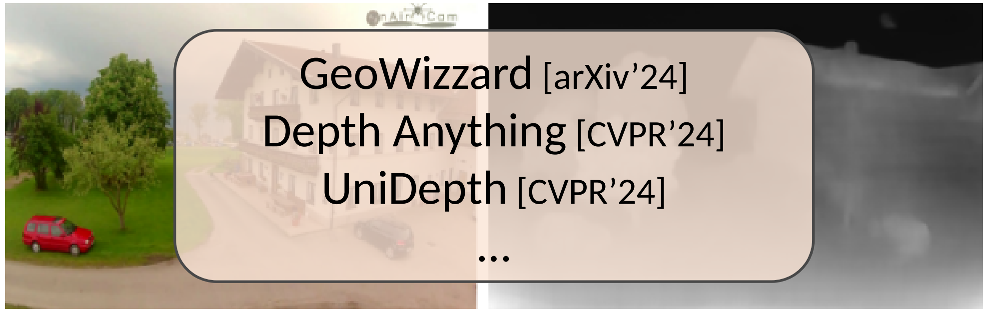
Large-scale 3D reconstruction

Visual Localization

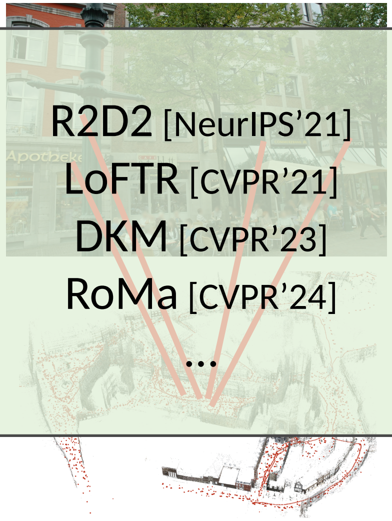
... and many more: SLAM, calibration, MVS, ...

What is 3D vision?

Monocular Depth estimation

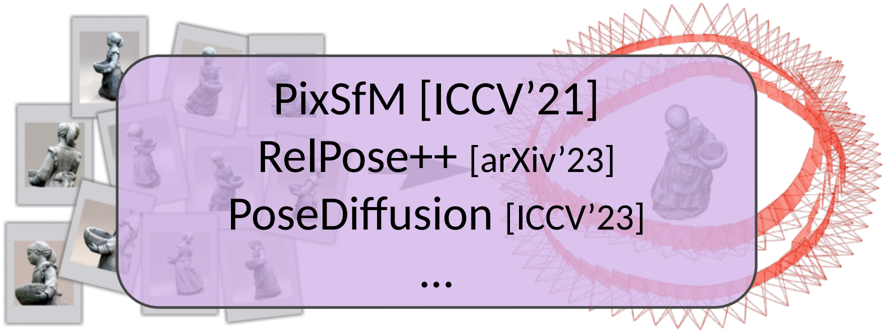


GeoWizzard [arXiv'24]
Depth Anything [CVPR'24]
UniDepth [CVPR'24]
...



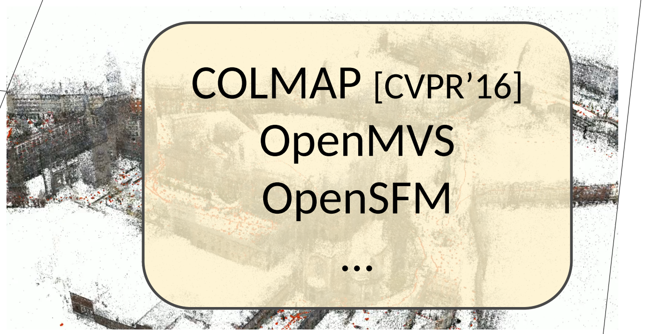
R2D2 [NeurIPS'21]
LoFTR [CVPR'21]
DKM [CVPR'23]
RoMa [CVPR'24]
...

Point matching



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

Multi-view pose estimation



COLMAP [CVPR'16]
OpenMVS
OpenSFM
...

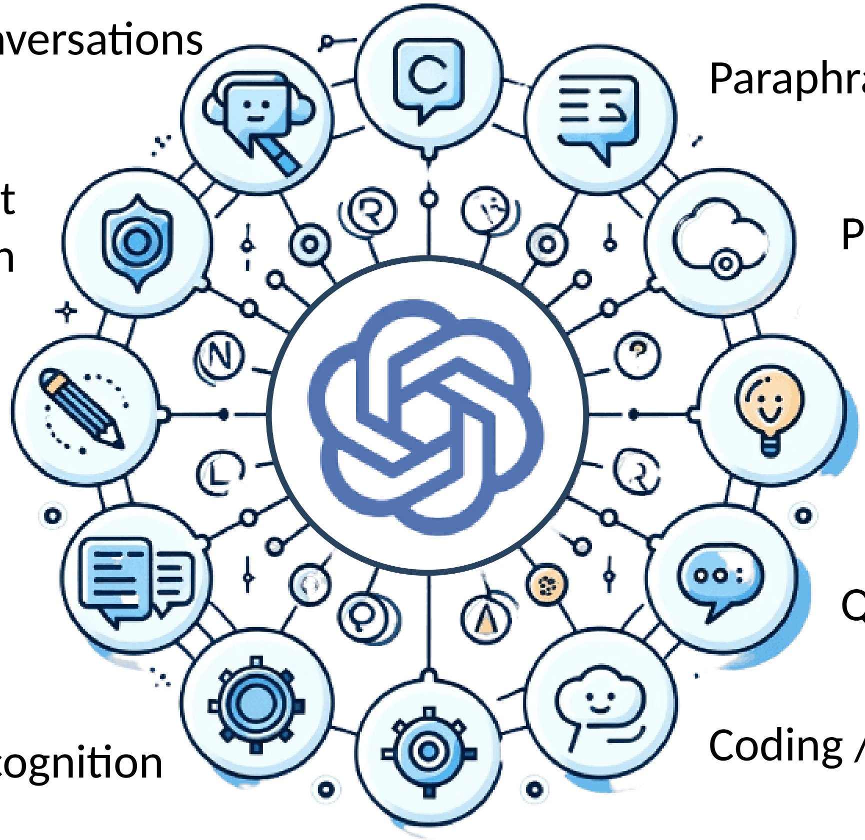
Large-scale 3D reconstruction

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 case of NLP



Why seek a unified model?

“Foundation models for 3DV”?

Weakly-supervised pretext task : **useful for many downstream tasks**

Many definitions, no consensus yet

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:

- establish correspondences between images (matching)
- infer 3D geometry

  from priors & from SfM
- infer relative pose (motion)
- decompose motion, lighting effects or long-term changes

NAVER LABS Europe

NAVER

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









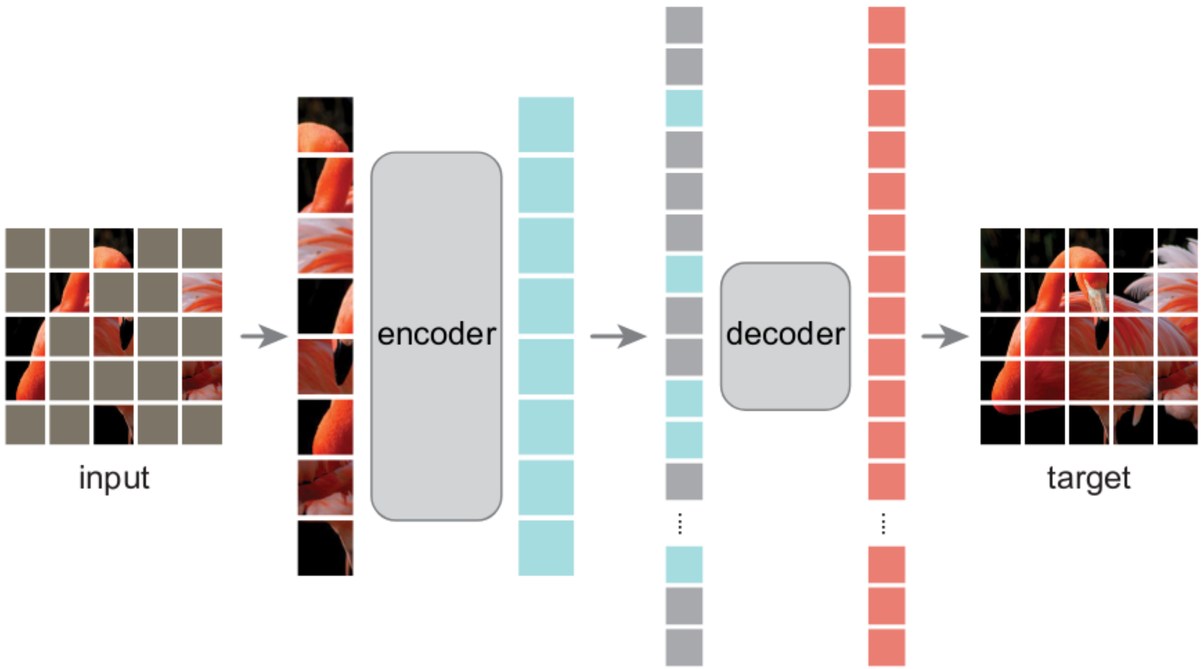

NAVER LABS Europe

NAVER

CroCo:
Self-supervised learning with Cross-View Completion

inspired by MAE

- self-supervised learning
- with masked modelling



Masked Autoencoders Are Scalable Vision Learners, Kaiming Het et. al. CVPR'22

NAVER LABS Europe

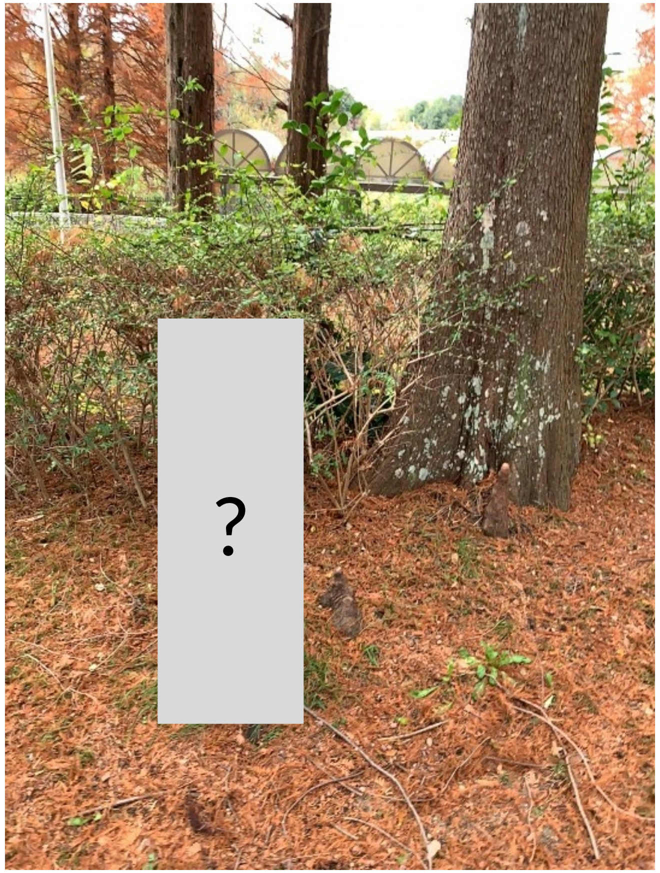
NAVER

CroCo: Self-supervised learning with Cross-View Completion

[NeurIPS'22] [ICCV'23]

A guessing game:

what's behind the mask?



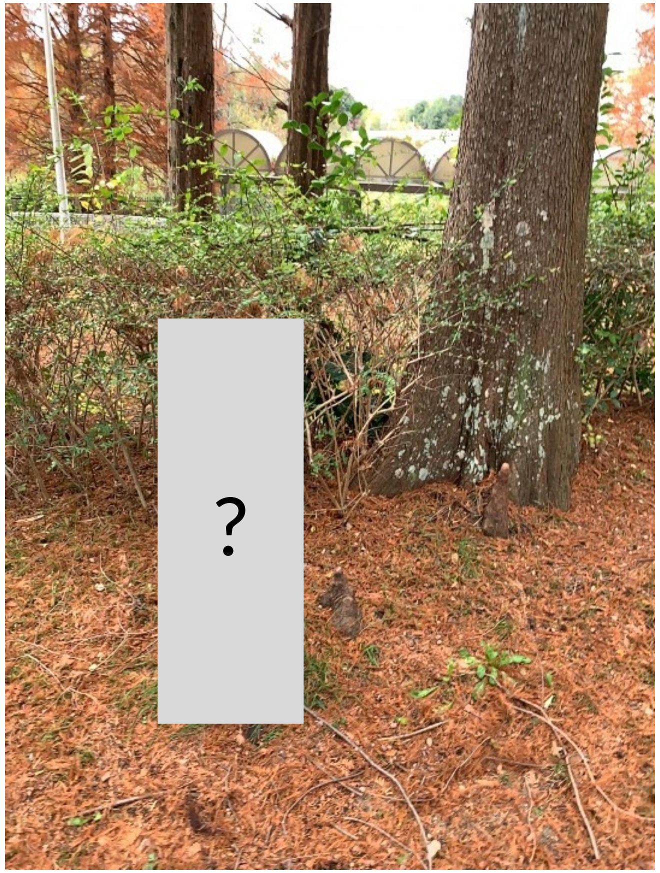
NAVER LABS Europe

NAVER

CroCo: Self-supervised learning with Cross-View Completion

[NeurIPS'22] [ICCV'23]

Reference view



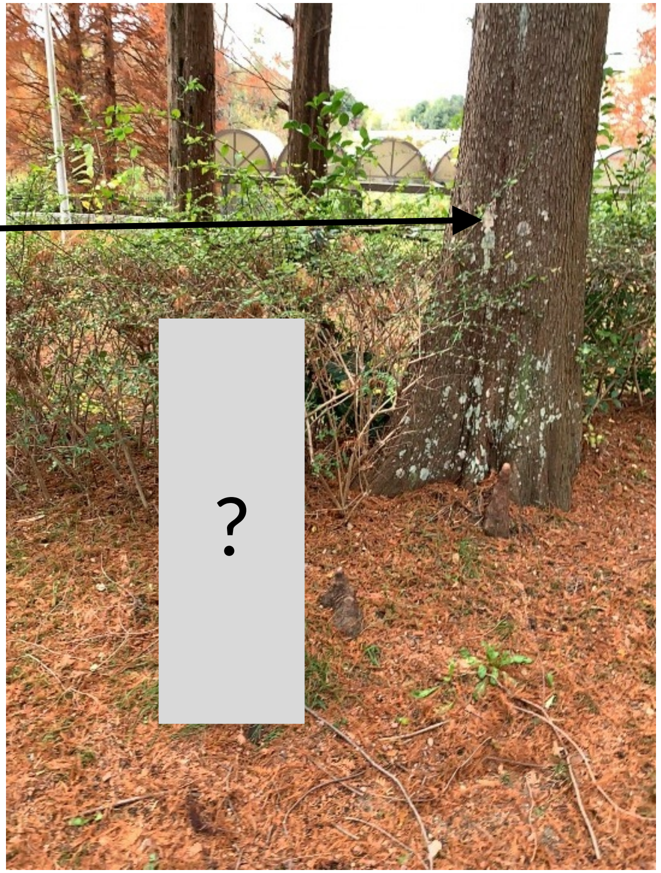
Query view

CroCo: Self-supervised learning with Cross-View Completion

[NeurIPS’22] [ICCV’23]

Image matching

Reference view



Query view

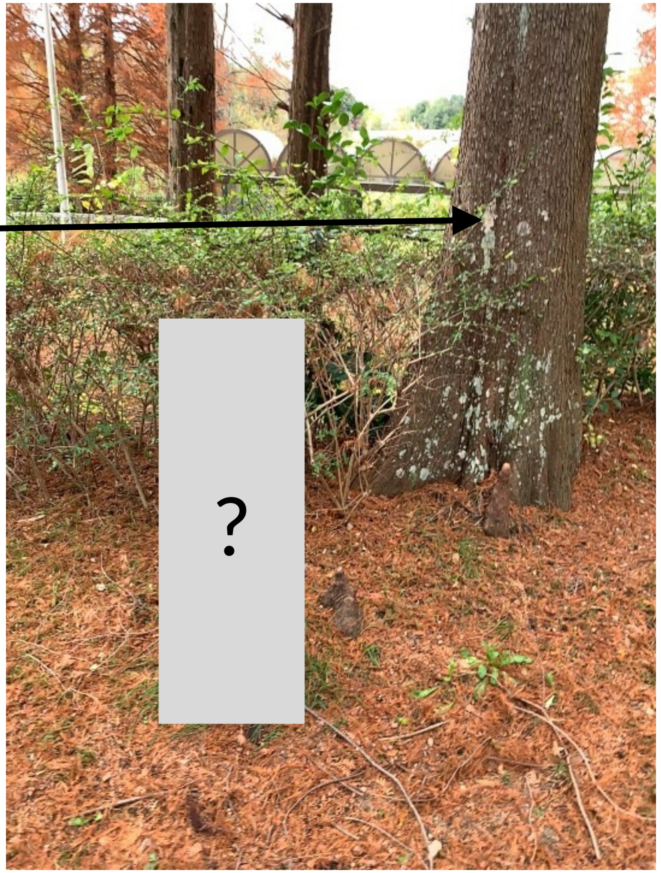
CroCo:
Self-supervised learning with Cross-View Completion

[NeurIPS'22] [ICCV'23]

Reference view

Image matching

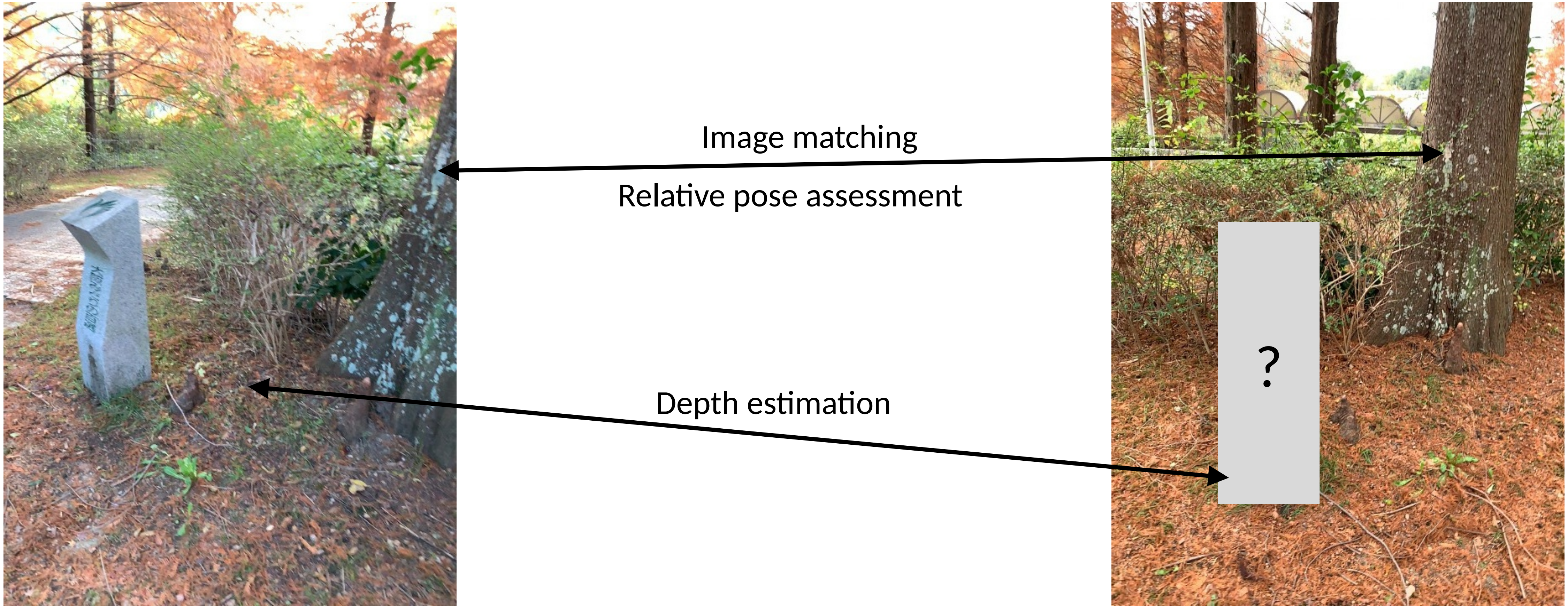
Relative pose assessment



Query view

CroCo: Self-supervised learning with Cross-View Completion

[NeurIPS'22] [ICCV'23]

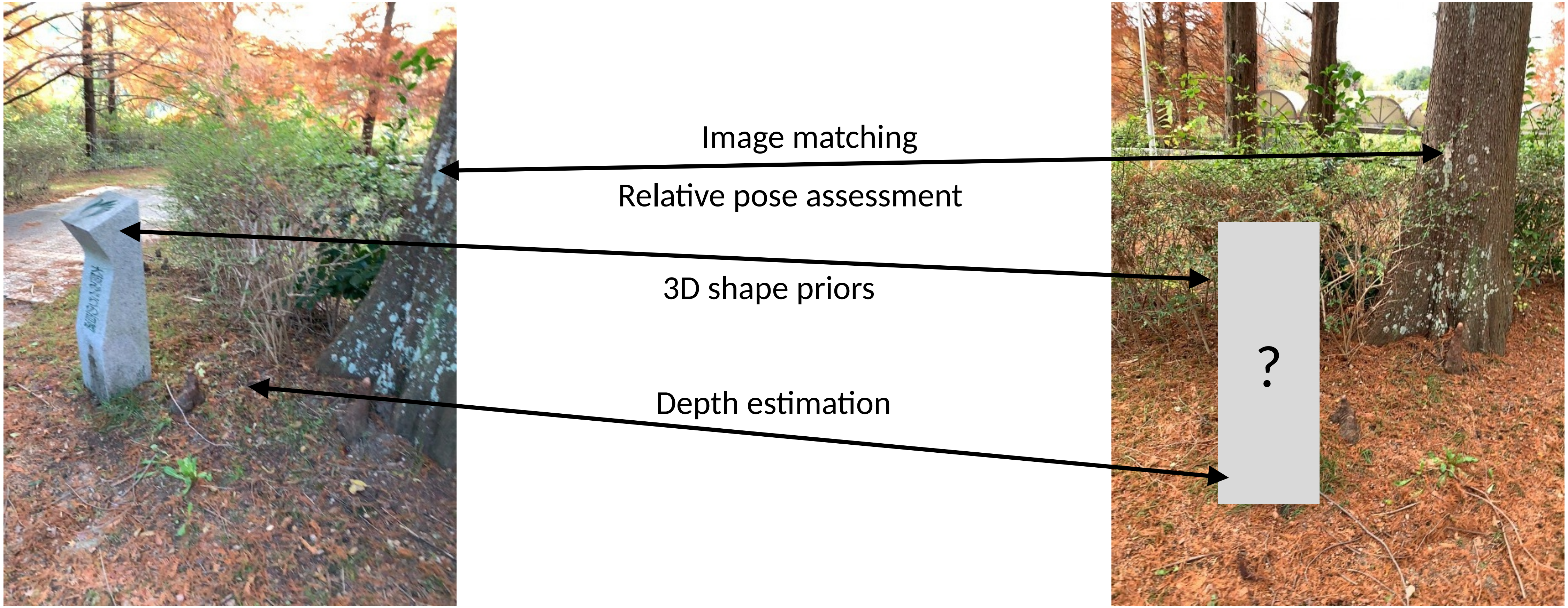


Reference view

Query view

CroCo:
Self-supervised learning with Cross-View Completion

[NeurIPS'22] [ICCV'23]

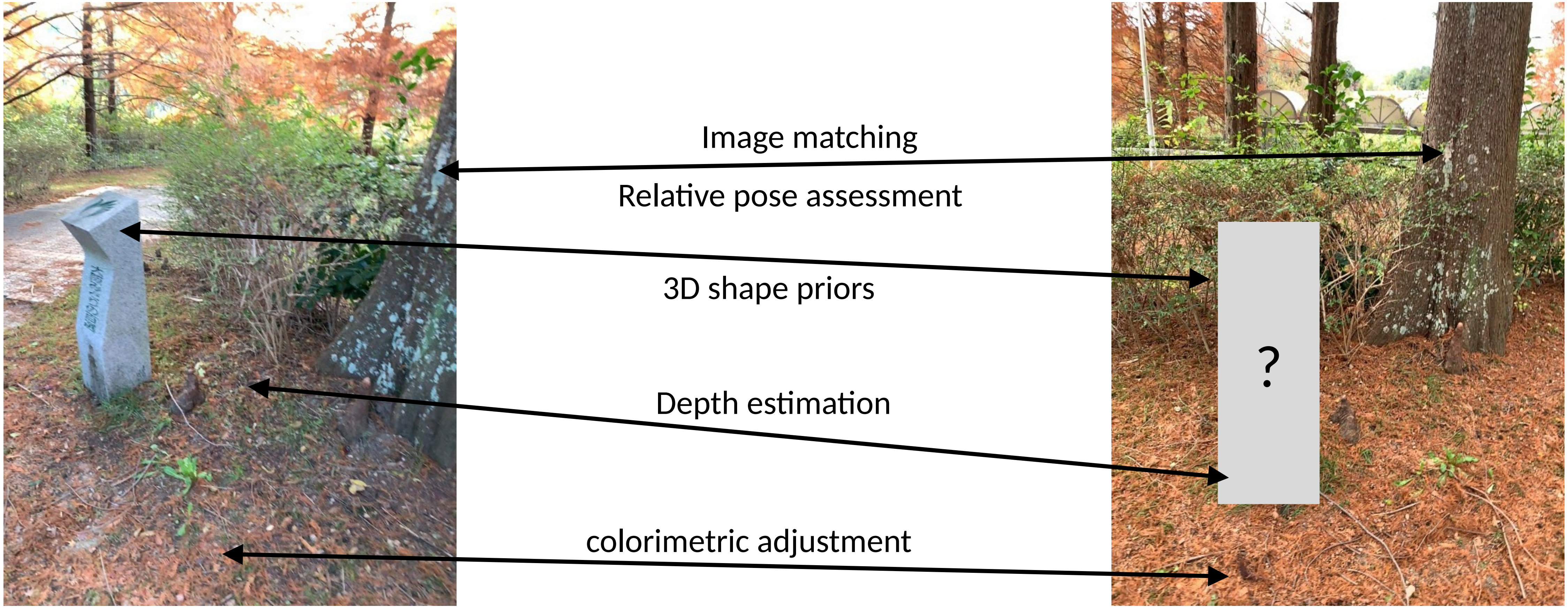


Reference view

Query view

CroCo: Self-supervised learning with Cross-View Completion

[NeurIPS’22] [ICCV’23]

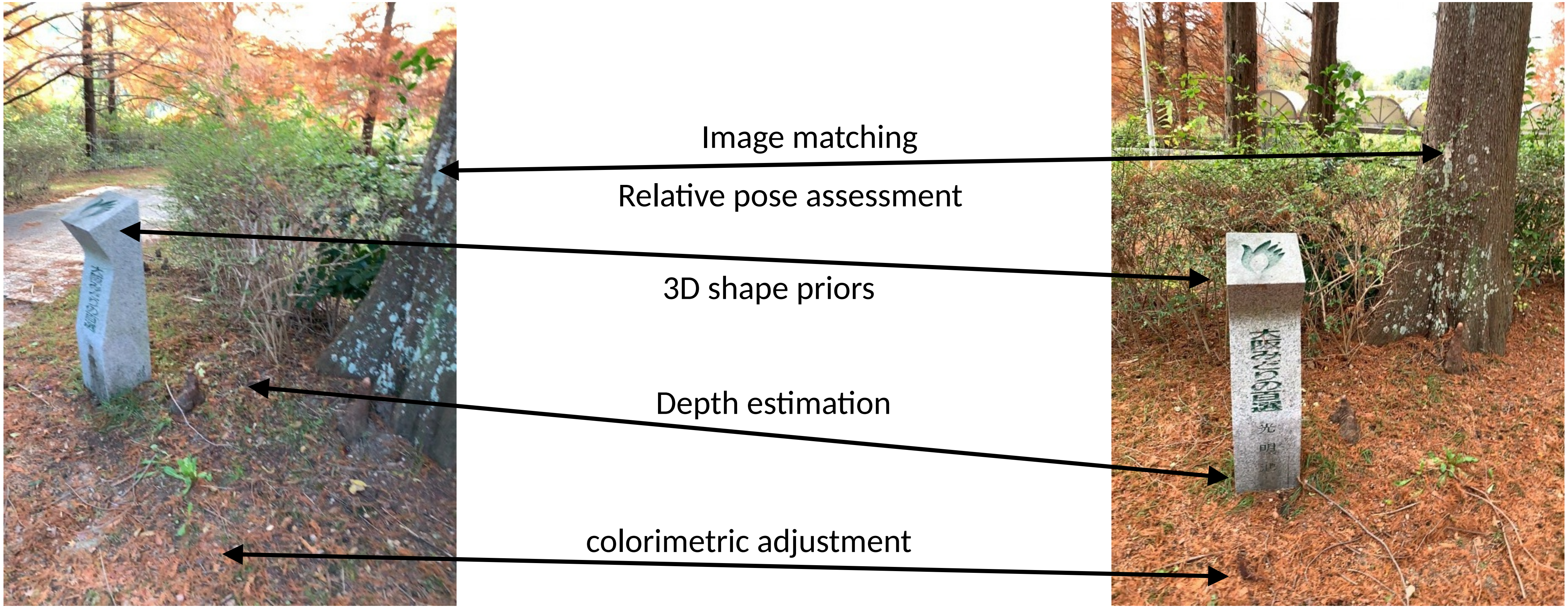


Reference view

Query view

CroCo: Self-supervised learning with Cross-View Completion

[NeurIPS’22] [ICCV’23]



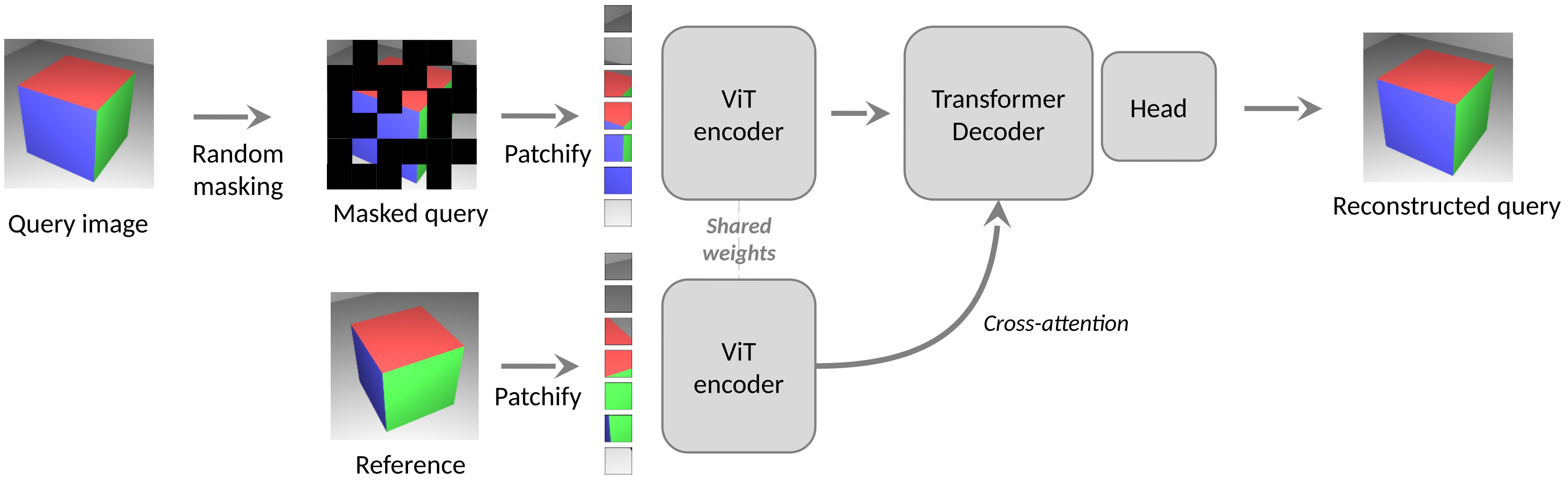
Reference view

Query view

NAVER LABS Europe

NAVER

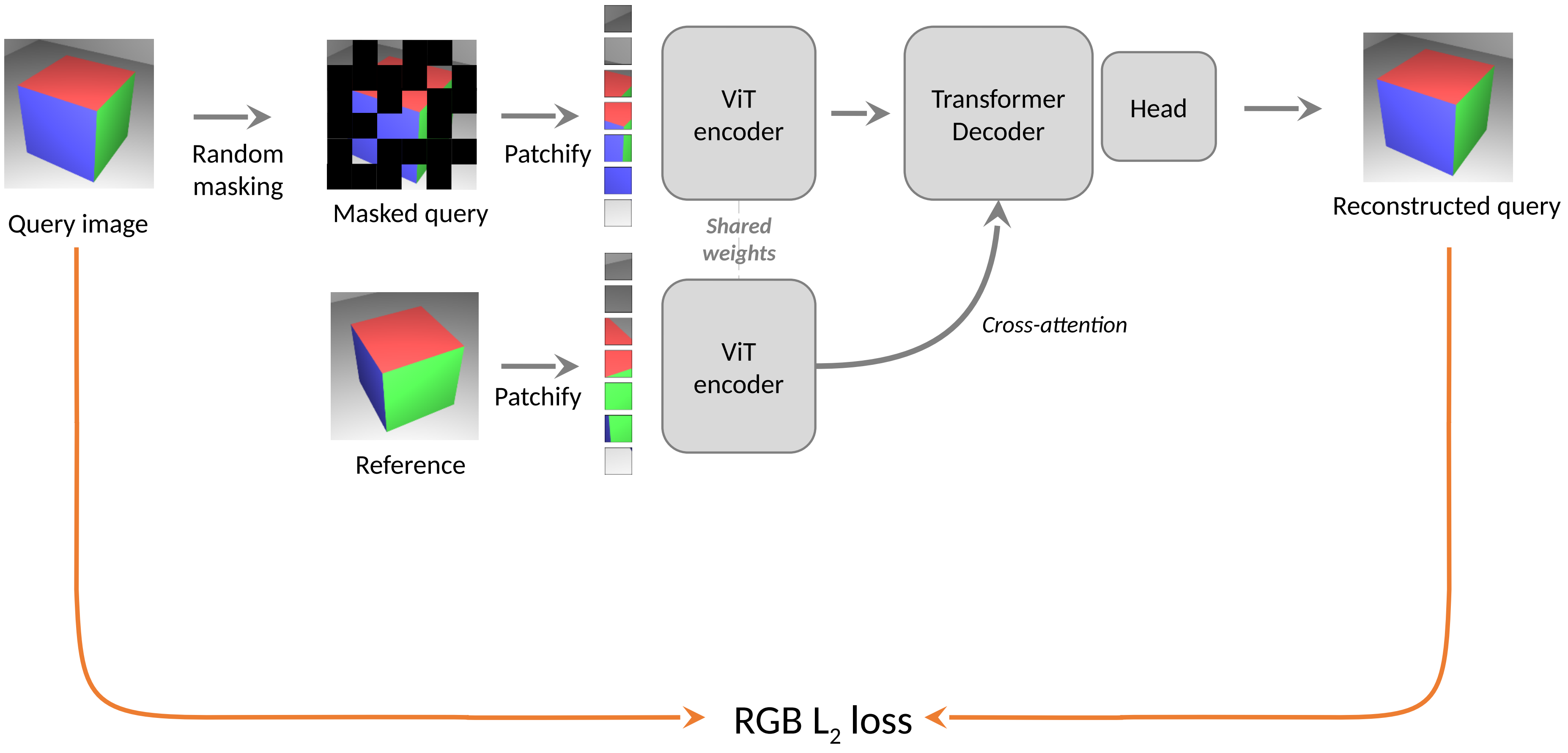
CroCo: Self-supervised learning with <u>Cro</u>ss-View <u>Co</u>mpletion



NAVER LABS Europe

NAVER

CroCo: Self-supervised learning with Cross-View Completion



NAVER LABS Europe

NAVER

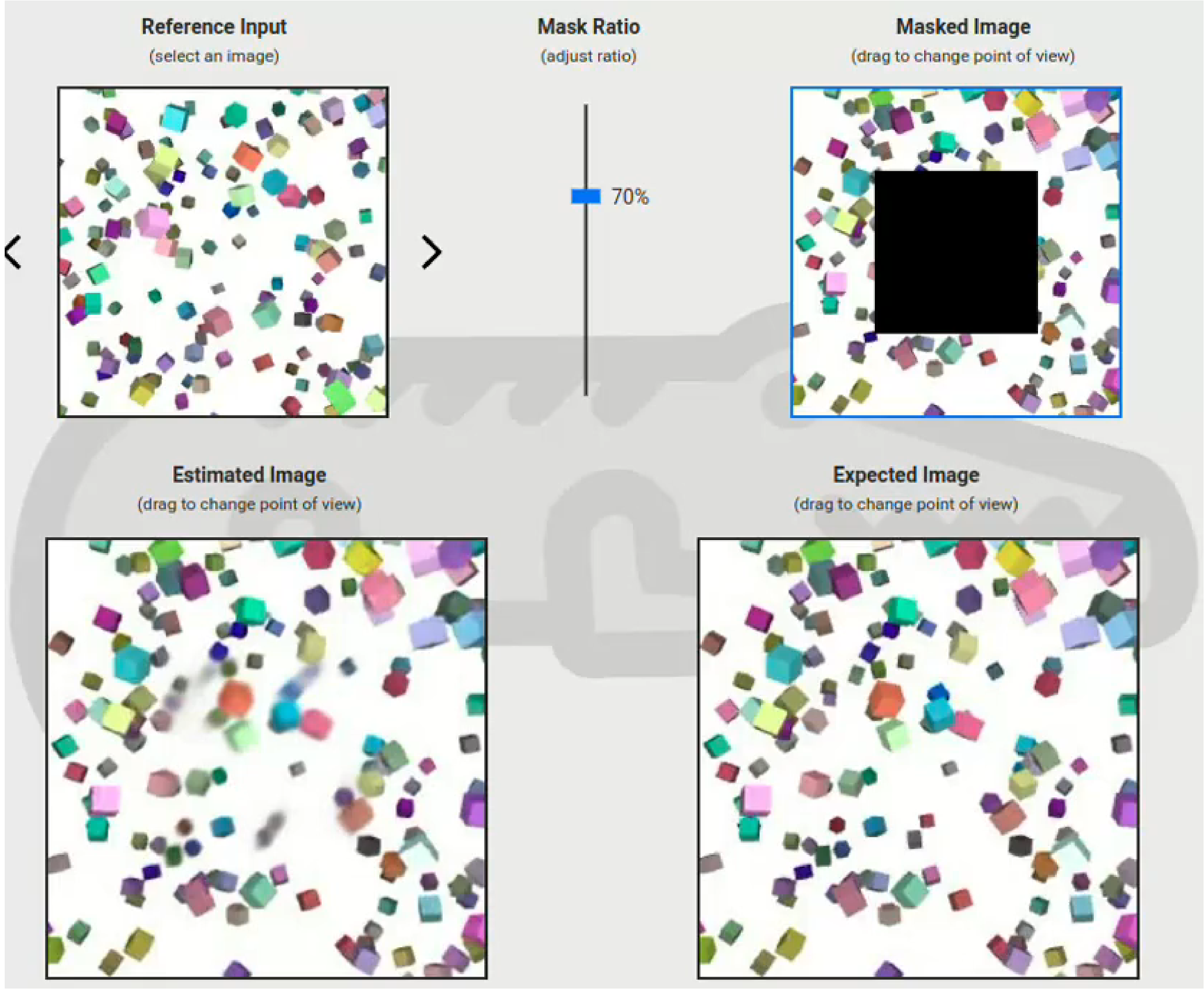
CroCo: Self-supervised learning with Cross-View Completion

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



NAVER LABS Europe

NAVER

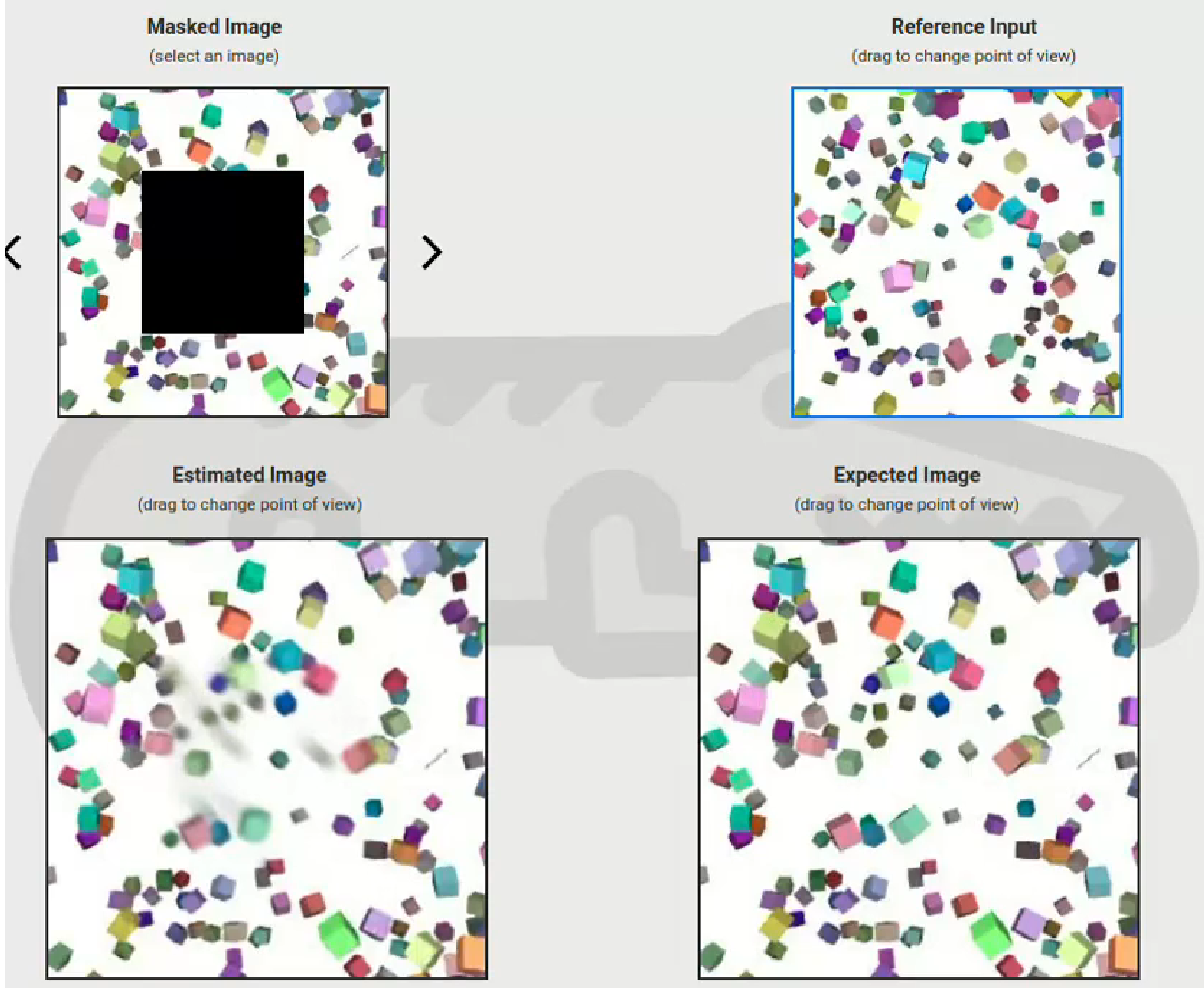
CroCo: Self-supervised learning with Cross-View Completion

Proof of concept:

- training with synthetic random scenes
- Test scene never seen before!

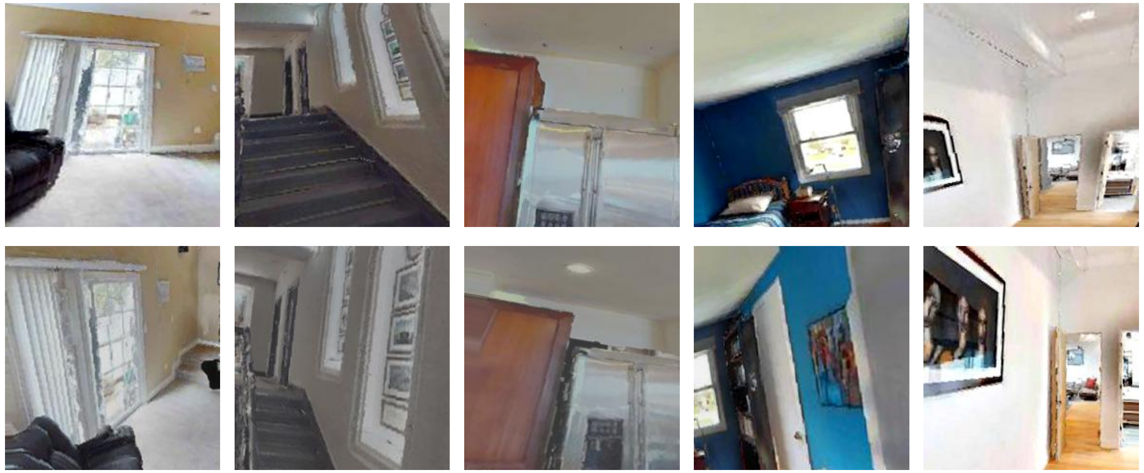
What solving this implies:

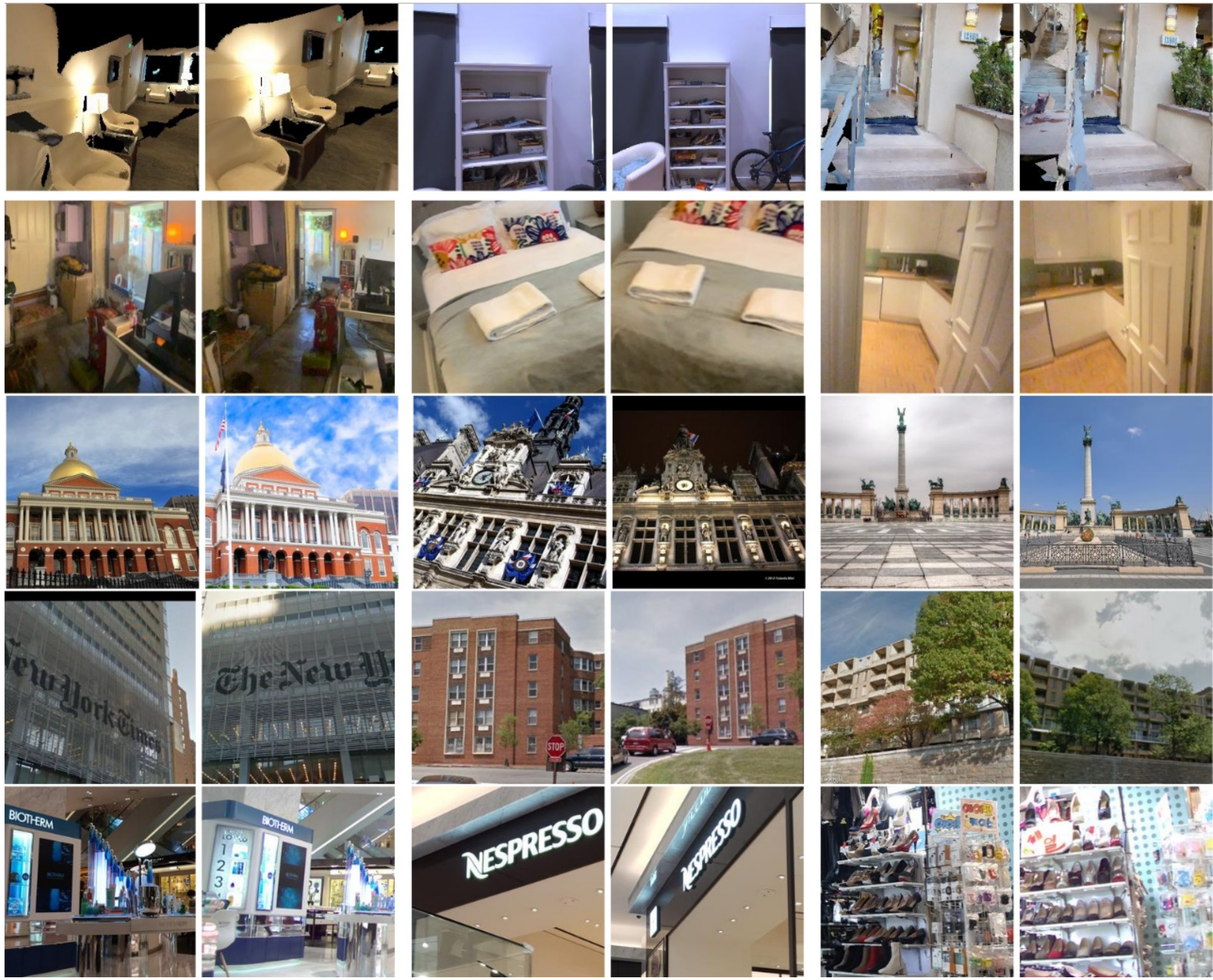
- Match the query and reference images
- Estimate the relative pose
- Infer an object-centric 3D reconstruction of the reference scene
- Align (rotate) the reference scene in 3D
- Render the reference scene based on imagined



Pre-training data

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

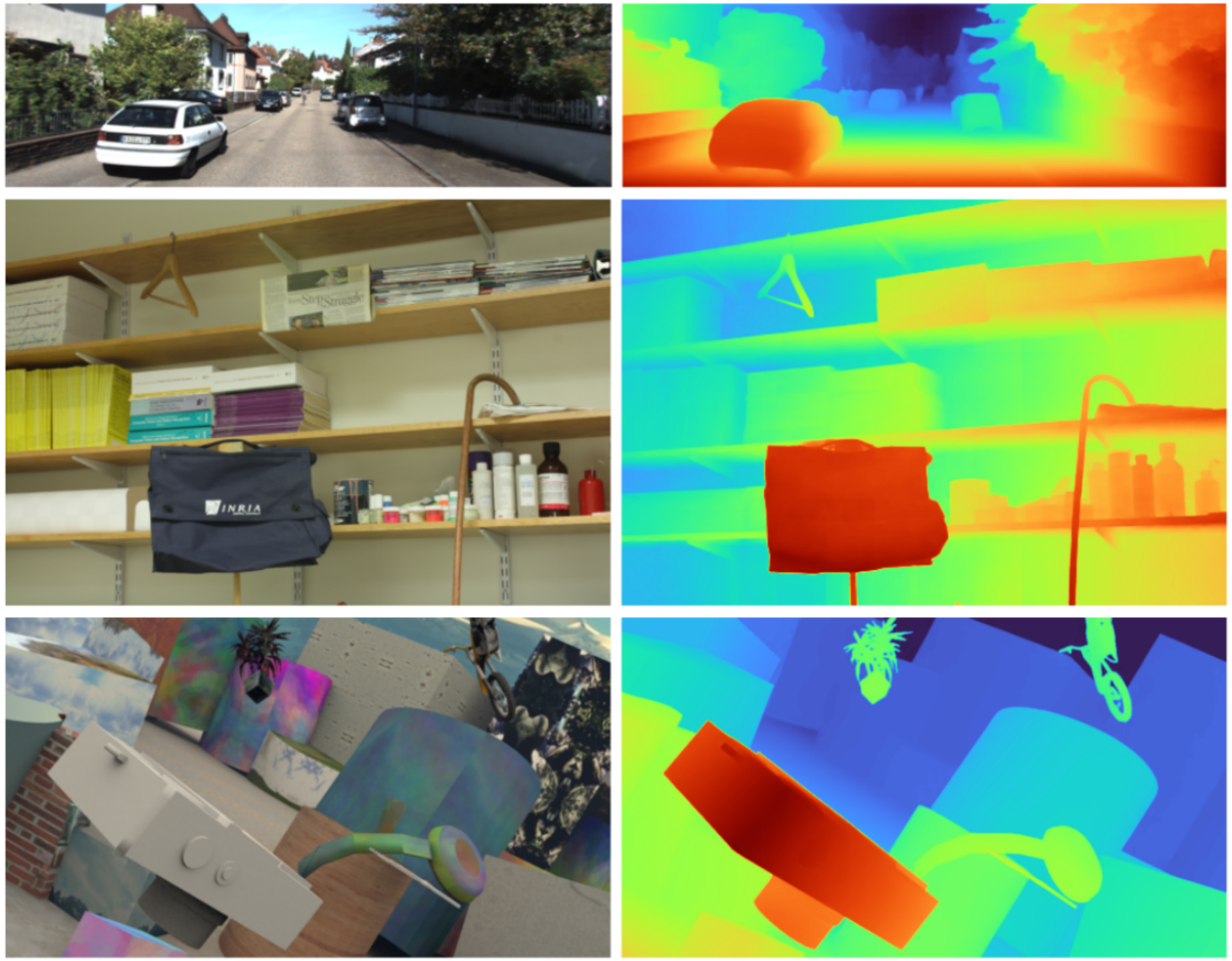




+ 5M training real image pairs

Binocular downstream tasks

CroCo encoder+decoder for stereo and optical flow



State-of-the-art results on Stereo Depth

NAVER LABS

1st rank on the ETH-3D benchmark
metric: <4% error

1st rank on the KITTI 2015 benchmark
Metric: <3px or <5% error on foreground objects

ETH3D   Home   Stereo

Low-res two-view benchmark

This table lists the benchmark results for the low-res two-view scenario. This benchmark evaluates the Middlebury stereo metrics (for all metrics, smaller is bet

- **bad 0.5, 1.0, 2.0, 4.0 [%]:** Fraction of pixels with errors larger than the given number of disparities.
- **Average error [px]:** The per-pixel average disparity error.
- **Root mean square error [px]:** The per-pixel root mean square disparity error.
- **50%, 90%, 95%, 99% error quantile [px]:** The highest disparity error within the given percentage of best pixels (for 50%, this is the median error).
- **Time [s]:** The runtime of the method.

The mask determines whether the metric is evaluated for all pixels with ground truth, or only for pixels which are visible in both images (non-occluded). The coverage selector allows to limit the table to results for all pixels (dense), or a given minimum fraction of pixels.

Methods with suffix _ROB may participate in the Robust Vision Challenge.

Click one or more dataset result cells or column headers to show visualizations. Most visualizations are only available for training datasets. The visualizatio mobile browsers.

Coverage: dense | **Set**: Test | **Metric**: bad 4.0 [%] | **Mask**: non-occluded | Dov

Method	Info	all	lakes. 1l	lakes. 1s	sand box 1l	sand box 1s	stora. room 1l	stora. room 1s	stora. room 2l	stora. room 2s	stora. room 2 1l	stora. room 2 1s	stora. room 2 2l	stora. room 2 2s	stora. room 3l	stora. room 3s
sCroCo		**0.05** 1	0.06 8	0.86 12	**0.00** 1	**0.00** 1	**0.00** 1	**0.00** 1	0.03 2	**0.01** 1	**0.00** 1	**0.00** 1	0.02 5	0.01 22	**0.00** 1	**0.00** 1
XXStereo		0.08 2	0.54 108	0.90 18	**0.00** 1	**0.00** 1	0.03 3	**0.00** 1	0.04 3	0.03 2	**0.00** 1	**0.00** 1	**0.01** 1	0.02 30	**0.00** 1	0.03 15
CREStereo		0.10 3	0.06 8	**0.51** 1	**0.00** 1	**0.00** 1	1.06 103	**0.00** 1	0.20 13	0.08 5	**0.00** 1	**0.00** 1	0.02 5	**0.00** 1	**0.00** 1	**0.00** 1
Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, Shuaicheng Liu: Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation. CVPR 2022																
PMTNet		0.11 4	0.07 16	0.60 2	**0.00** 1	**0.00** 1	1.05 97	0.04 106	0.16 8	0.08 5	**0.00** 1	**0.00** 1	0.04 9	**0.00** 1	0.03 14	0.01 7
Gwc-CoAtRS		0.12 5	0.06 8	1.17 39	**0.00** 1	0.11 138	0.41 20	**0.00** 1	0.20 13	0.04 3	0.01 38	**0.00** 1	0.18 38	**0.00** 1	0.01 4	0.06 34

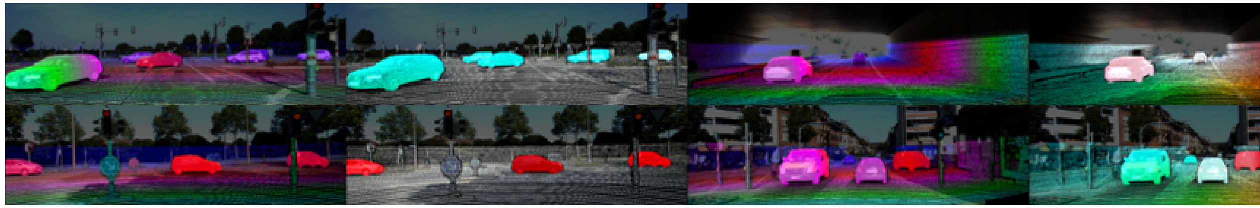
The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago

Karlsruhe Institute of Te

home setup stereo flow sceneflow depth odometry object tracking road semantics raw data submit resu

A. Geiger | P. Lenz | C. Stiller | R. Urtasun

Stereo Evaluation 2015



Evaluation ground truth All pixels   **Evaluation area** All pixels

	Method	Setting	Code	D1-bg	D1-fg	D1-all	Density	Runtime	Environment
1	sCroCo		code	1.73 %	**2.76 %**	1.90 %	100.00 %	1.1 s	1 core @ 2.5 Ghz (Python)
2	LaC+GANet		code	1.44 %	2.83 %	1.67 %	100.00 %	1.8 s	GPU @ 2.5 Ghz (Python)
ERROR: Wrong syntax in BIBTEX file.									
3	UPFNet			1.38 %	2.85 %	**1.62 %**	100.00 %	0.25 s	1 core @ 2.5 Ghz (Python)
4	CREStereo		code	1.45 %	2.86 %	1.69 %	100.00 %	0.41 s	GPU @ >3.5 Ghz (Python)
5	CSPN			1.51 %	2.88 %	1.74 %	100.00 %	1.0 s	GPU @ 2.5 Ghz (Python)
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. Advances in Neural Information Processing									
6	ACVNet		code	1.37 %	3.07 %	1.65 %	100.00 %	0.2 s	1 core @ 2.5 Ghz (Python)
G. Xu, J. Cheng, P. Guo and X. Yang: Attention Concatenation Volume for Accurate and Efficient Stereo Matching. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec									

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

💡 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)

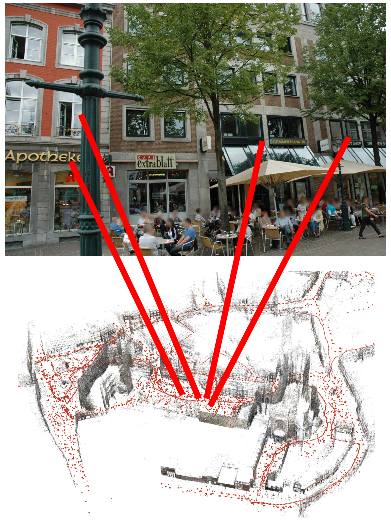
  → we are still seeking for a unified model …

NAVER LABS Europe

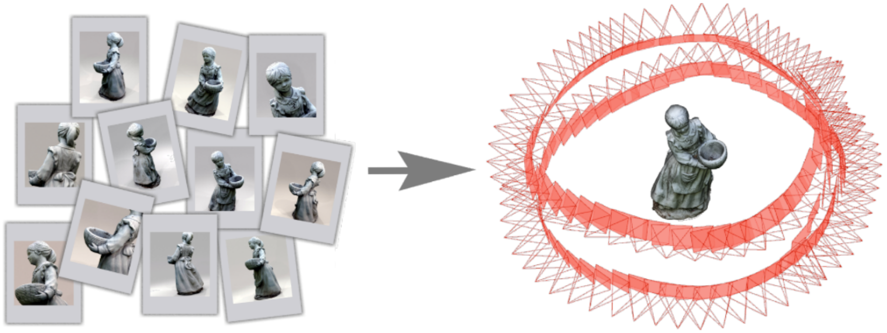
NAVER

What is 3D vision?

Monocular Depth estimation



Point matching



Multi-view pose estimation

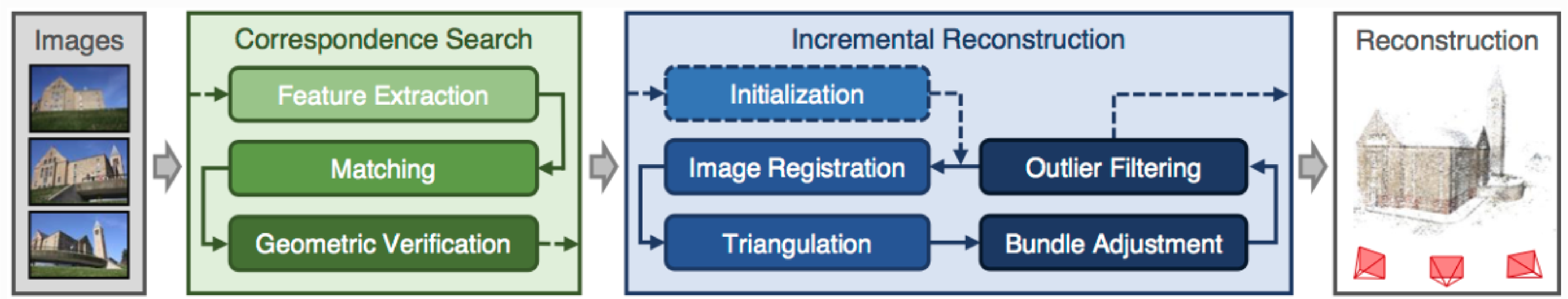
Dense 3D reconstruction

Visual Localization

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

Unifying all 3D vision tasks?

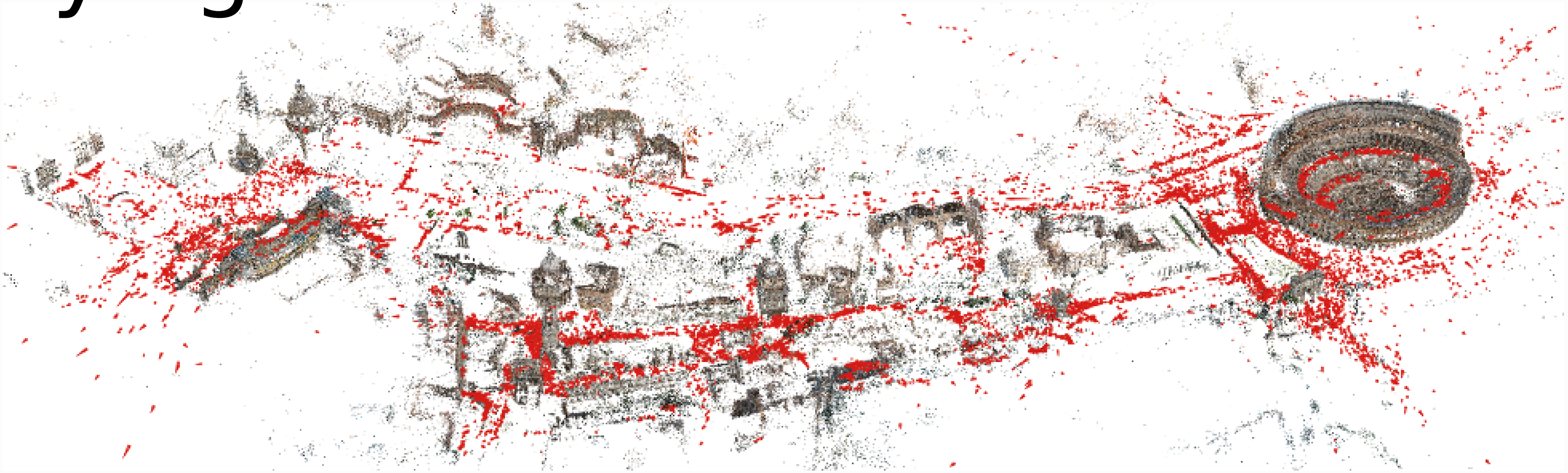
Could “dense 3D reconstruction” be a “super task” for 3DV?



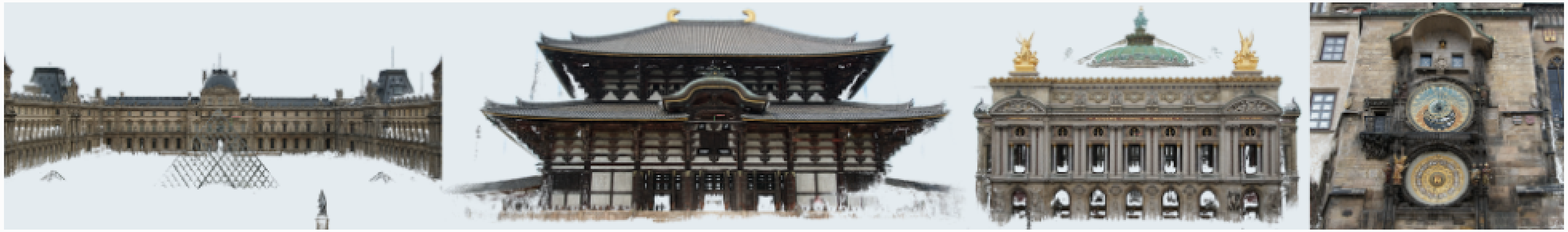
COLMAP’s incremental Structure-from-Motion pipeline.

“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”

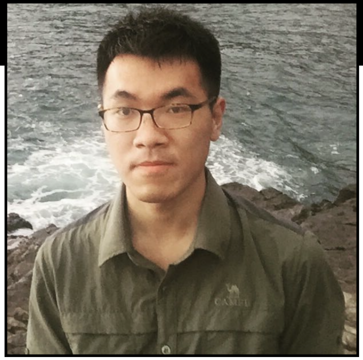
- intrinsically connected to all other 3DV tasks

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

DUST3R

Dense Unconstrained Stereo 3D Reconstruction





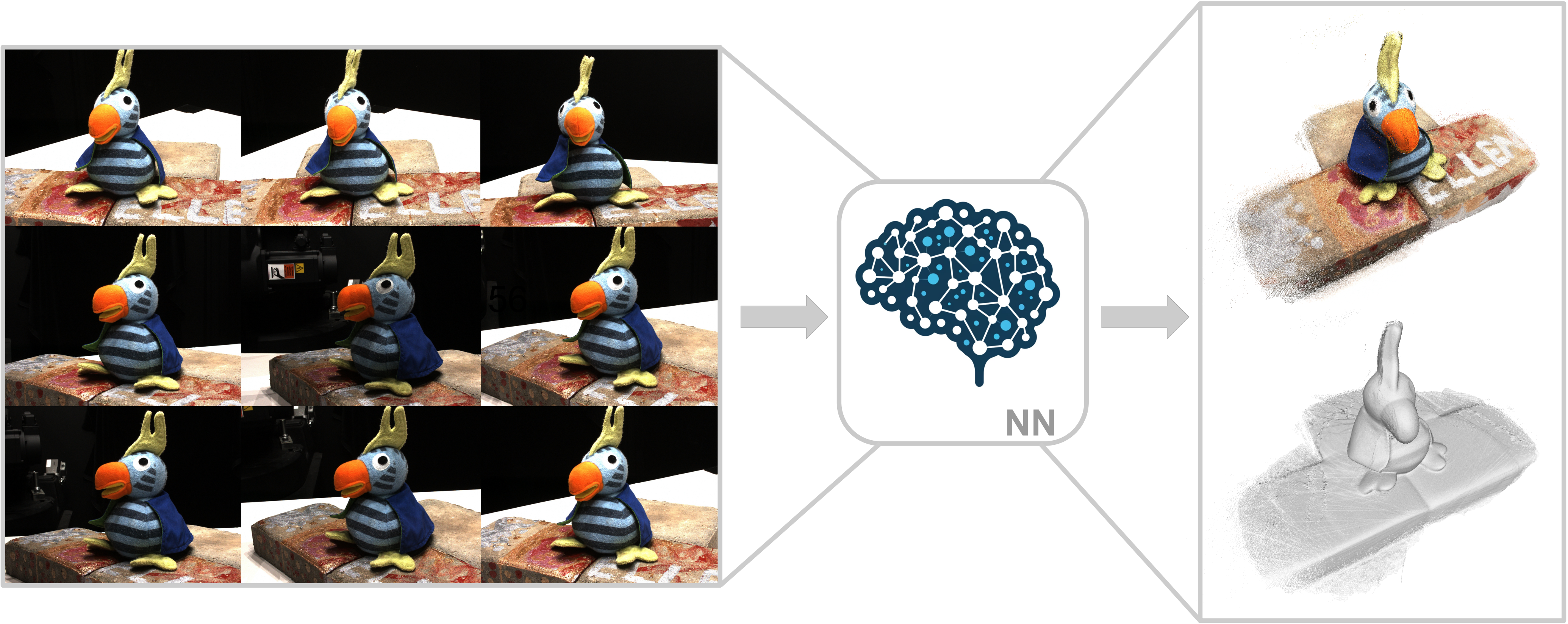
Shuzhe Wang	Vincent Leroy	Yohann Cabon	Boris Chidlovskii	Jérome Revaud
Aalto University	Naverlabs Europe	Naverlabs Europe	Naverlabs Europe	Naverlabs Europe

A! Aalto University

NAVER LABS Europe

Our Dream

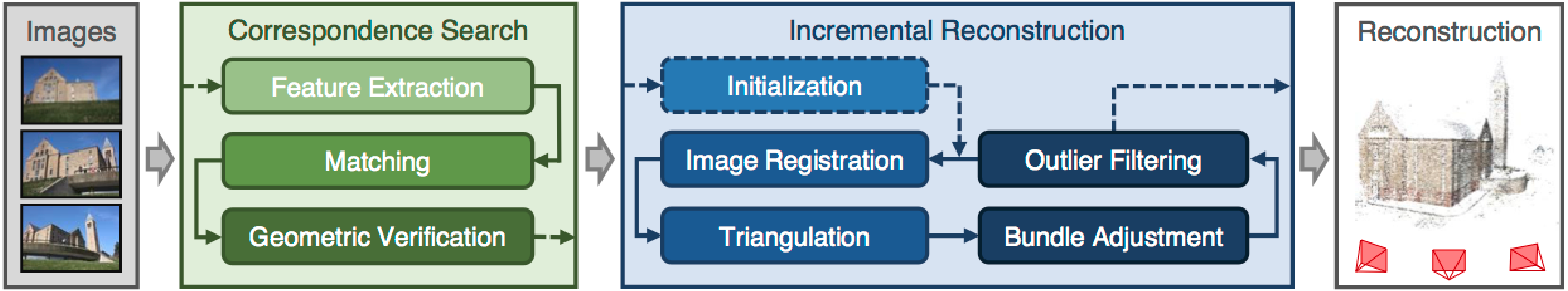
Dense Unconstrained Multi-View Stereo 3D Reconstruction (MVS)



Unconstrained = unknown cameras !

The Unconstrained MVS Paradox

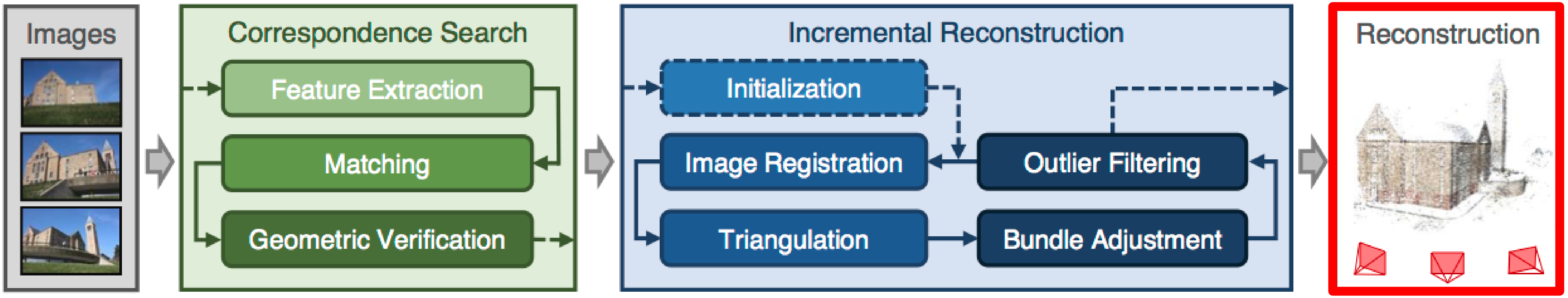
Obtaining Camera Parameters, e.g. COLMAP



Source: https://colmap.github.io/_images/incremental-sfm.png

The Unconstrained MVS Paradox

Obtaining Camera Parameters, e.g. COLMAP

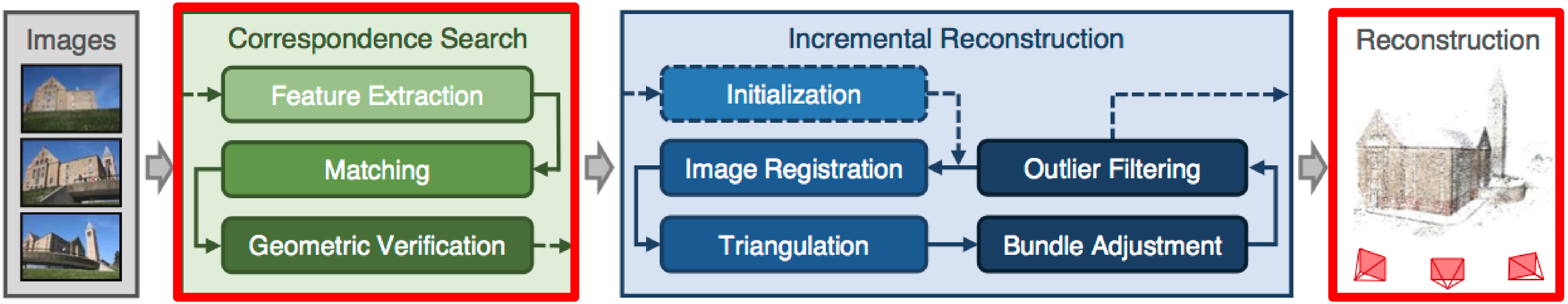


Need 3D scene for cameras

Source: https://colmap.github.io/_images/incremental-sfm.png

The Unconstrained MVS Paradox

Obtaining Camera Parameters, e.g. COLMAP



Matching is inherently 3D

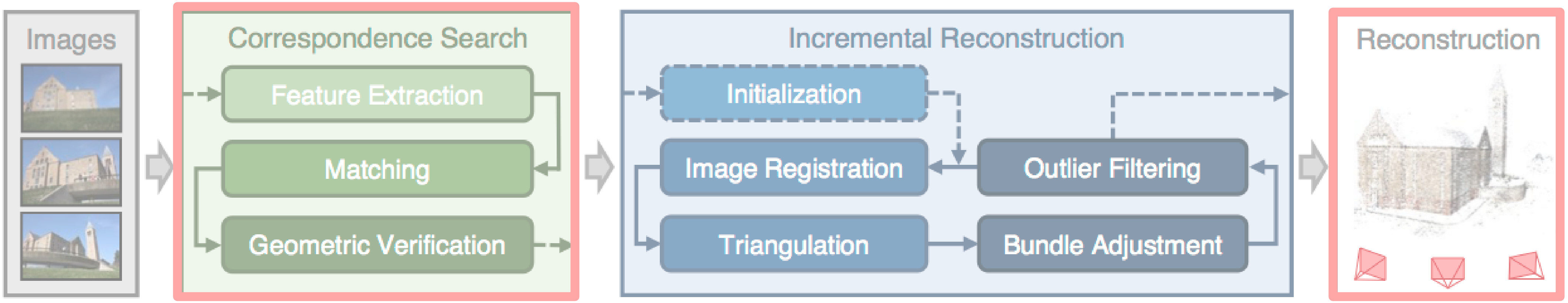
Need 3D scene for cameras

Source: https://colmap.github.io/_images/incremental-sfm.png

NAVER LABS Europe

NAVER

The Unconstrained MVS Paradox



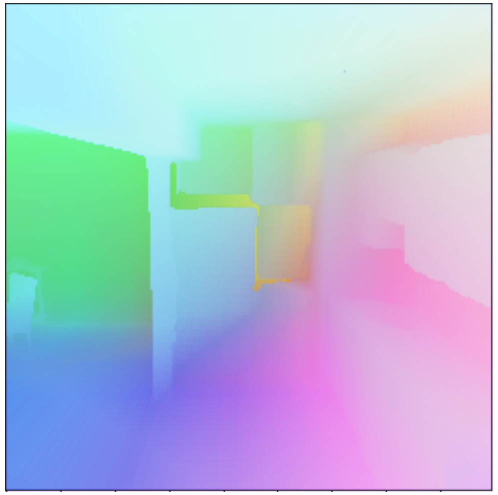
Source: https://colmap.github.io/_images/incremental-sfm.png

<u>We are looking for a Mapping between 2D image coordinates and 3D space</u>

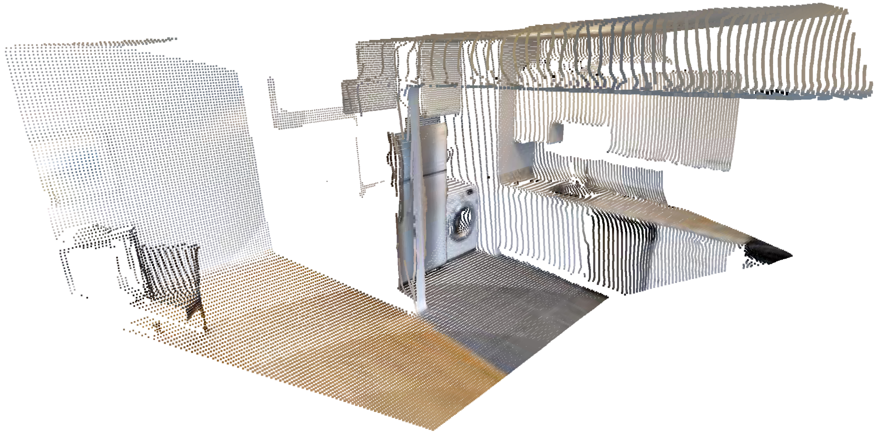
The Pointmap representation

Input Image

Corresponding *Pointmap*



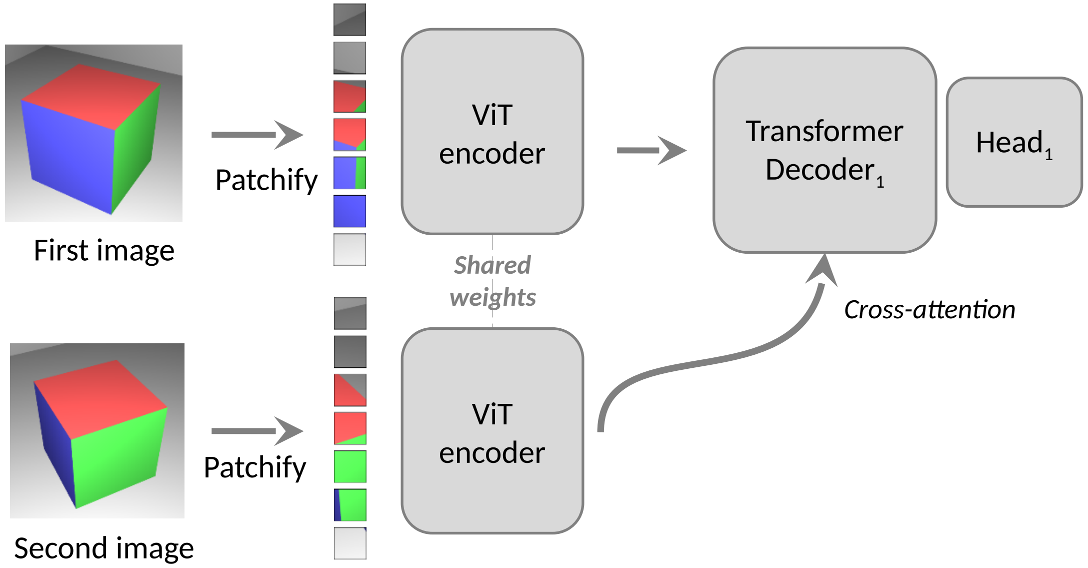
=



***Pointmaps* encode**

3D Scene geometry
2D pixels consistency
2D-3D relationships

DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction

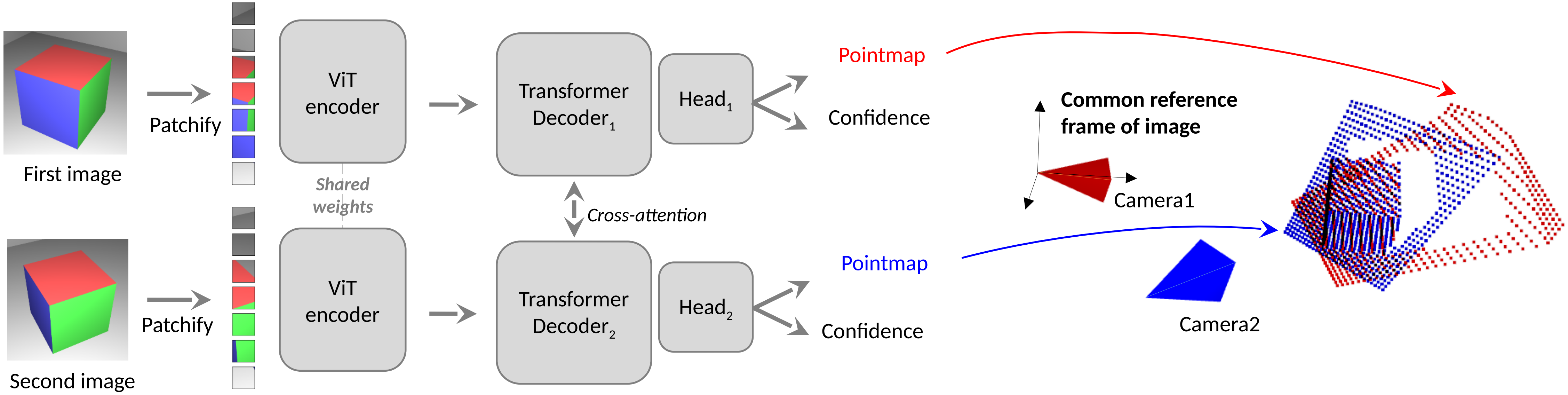


Start from CroCo …

NAVER LABS Europe

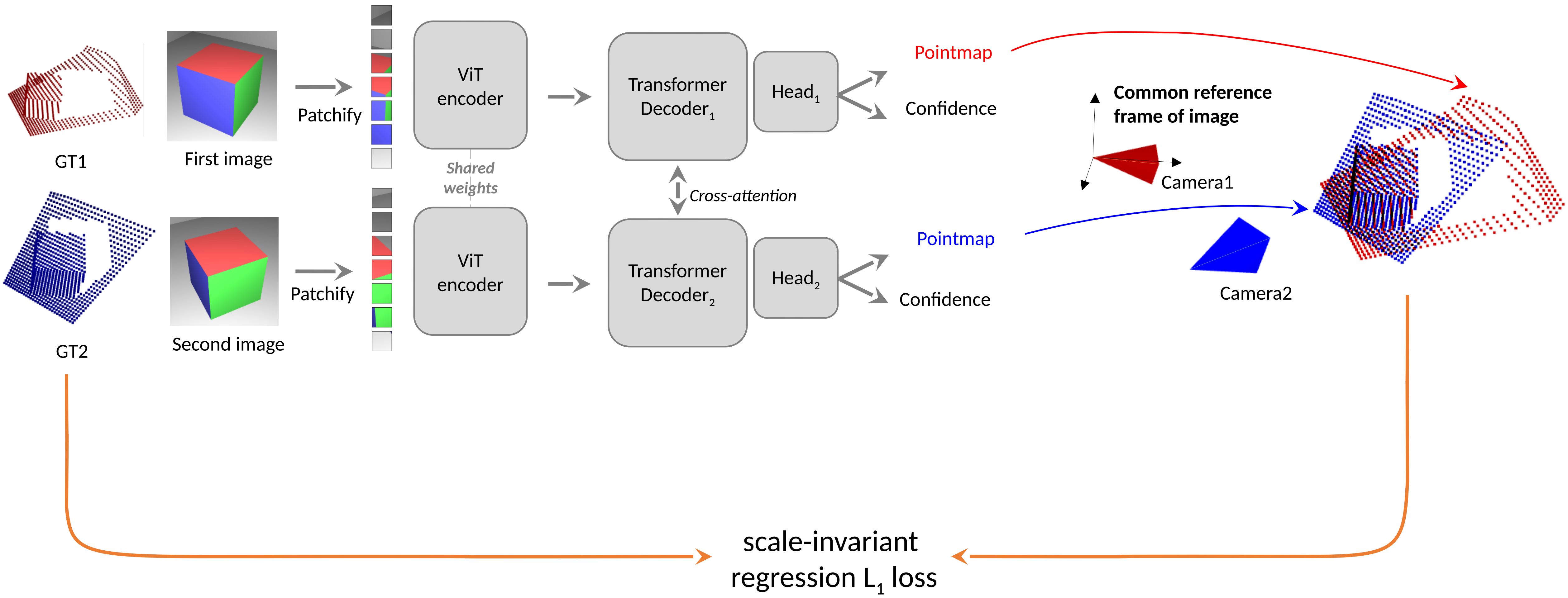
NAVER

DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction

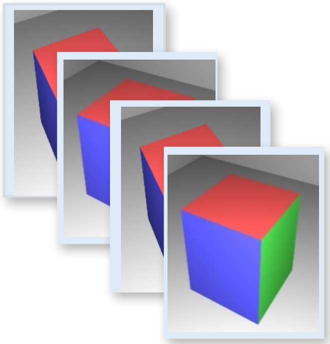


Start from CroCo and add a 2nd decoder

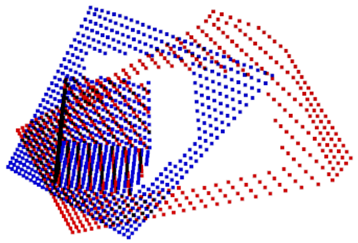
DUSt3R: Dense Unconstrained Stereo 3D Reconstruction



DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction



DUSt3R



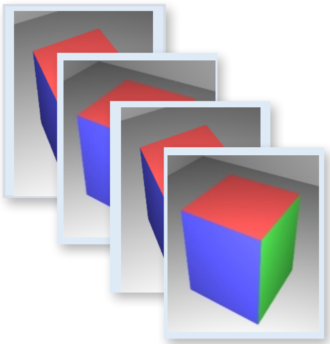
Unconstrained_
image collection

(no pose,
no intrinsics)

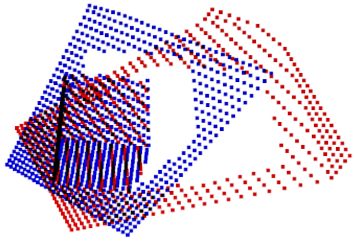
Output
pointmaps

(dense 2D ↔3D
mappings)

DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction



DUSt3R



Camera calibration

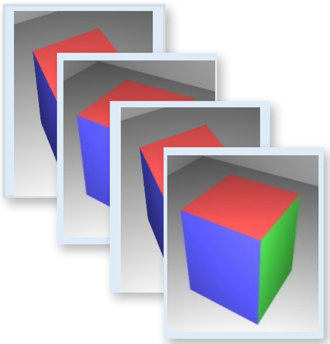
Unconstrained_
image collection

(no pose,
no intrinsics)

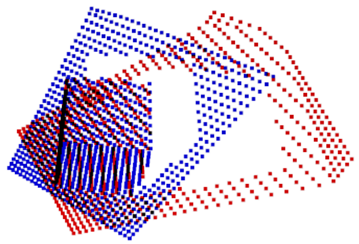
Output
pointmaps

(dense 2D ↔3D
mappings)

DUSt3R: Dense Unconstrained Stereo 3D Reconstruction



DUSt3R



Camera calibration

Depth estimation

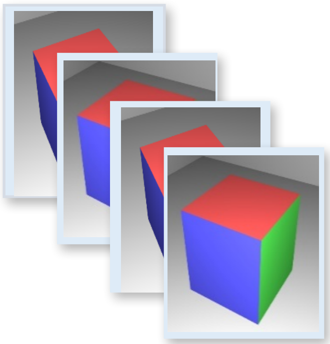
Unconstrained_
image collection

*(no pose,
no intrinsics)*

Output
pointmaps

*(dense 2D ↔3D
mappings)*

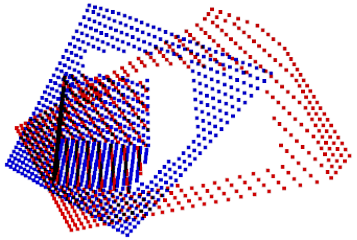
DUSt3R: Dense Unconstrained Stereo 3D Reconstruction



Unconstrained_
image collection

*(no pose,
no intrinsics)*

DUSt3R



Output
pointmaps

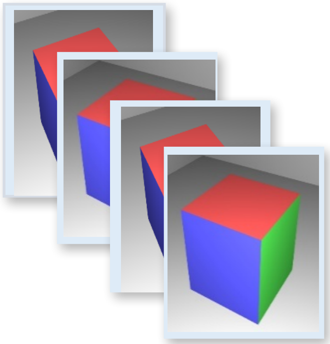
*(dense 2D ↔3D
mappings)*

Camera calibration

Depth estimation

Monocular

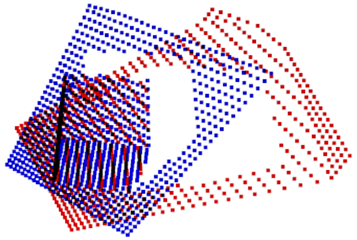
DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction



Unconstrained image collection

(no pose, no intrinsics)

DUSt3R



Output pointmaps

(dense 2D ↔3D mappings)

Camera calibration

Depth estimation

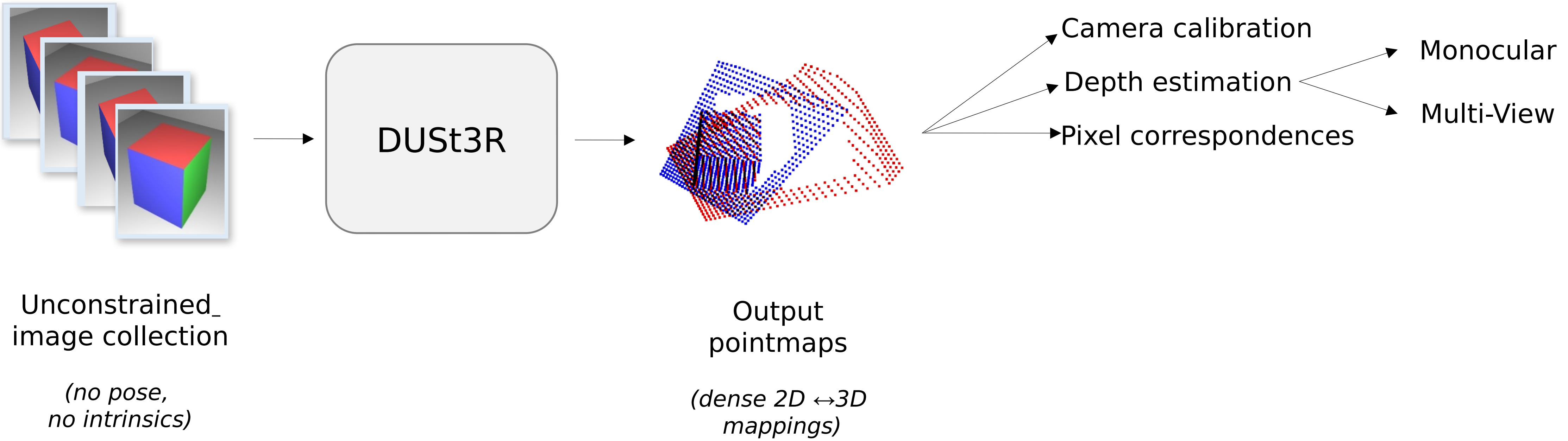
Monocular

Multi-View

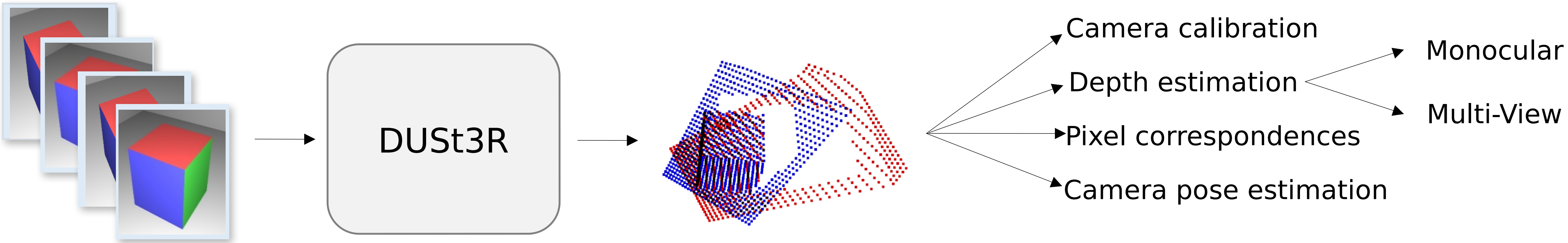
NAVER LABS Europe

NAVER

DUSt3R: Dense Unconstrained Stereo 3D Reconstruction



DUSt3R: Dense Unconstrained Stereo 3D Reconstruction



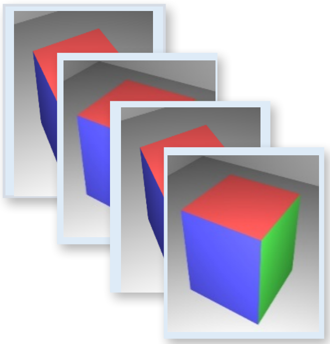
Unconstrained_
image collection

*(no pose,
no intrinsics)*

Output
pointmaps

*(dense 2D ↔3D
mappings)*

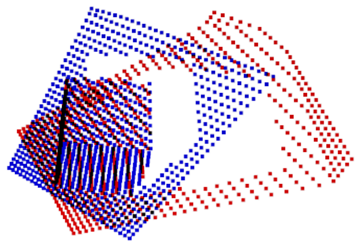
DUSt3R: Dense Unconstrained Stereo 3D Reconstruction



Unconstrained image collection

(no pose, no intrinsics)

DUSt3R

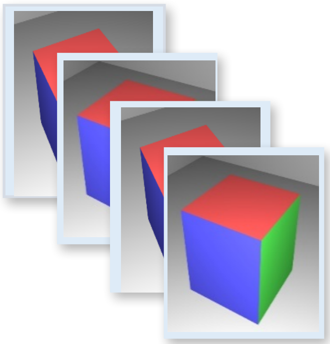


Output pointmaps

(dense 2D ↔3D mappings)

- Camera calibration
- Depth estimation
  - Monocular
  - Multi-View
- Pixel correspondences
- Camera pose estimation
  - Pairwise (relative)

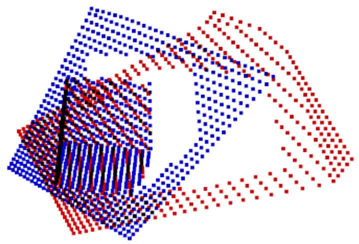
DUSt3R: Dense Unconstrained Stereo 3D Reconstruction



Unconstrained image collection

(no pose, no intrinsics)

→ DUSt3R →

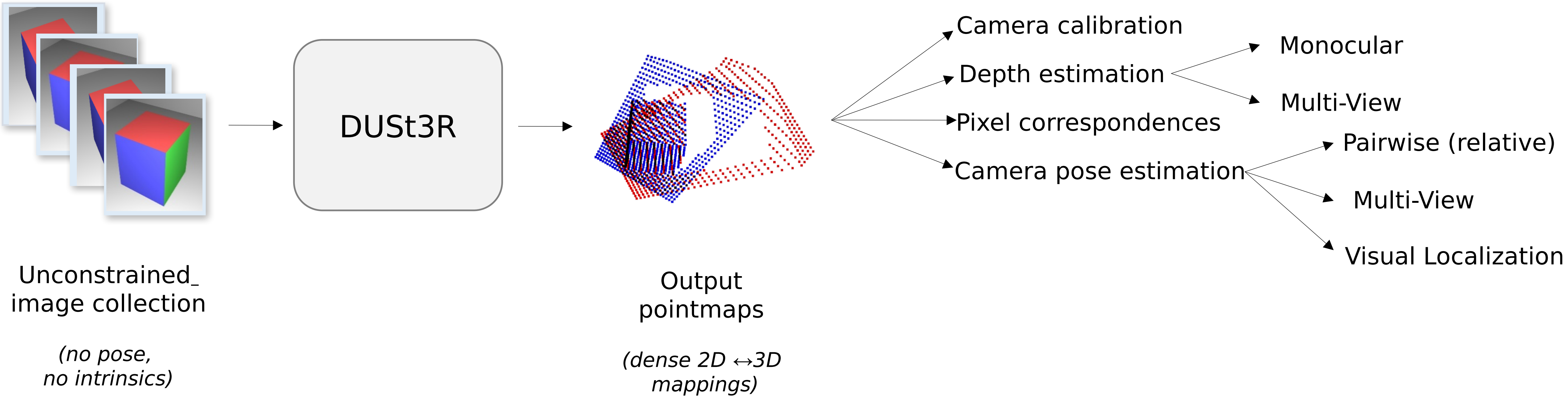


Output pointmaps

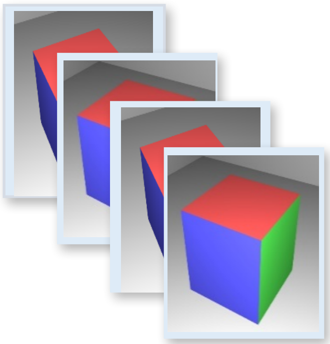
(dense 2D ↔3D mappings)

- Camera calibration
- Depth estimation
  - Monocular
  - Multi-View
- Pixel correspondences
- Camera pose estimation
  - Pairwise (relative)
  - Multi-View

DUSt3R: Dense Unconstrained Stereo 3D Reconstruction



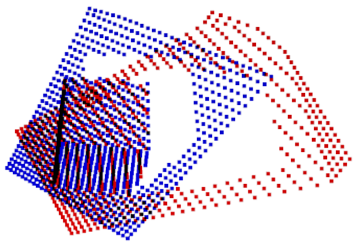
DUSt3R: Dense Unconstrained Stereo 3D Reconstruction



Unconstrained_ image collection

(no pose, no intrinsics)

→ DUSt3R →



Output pointmaps

(dense 2D ↔3D mappings)

- Camera calibration
- Depth estimation
  - Monocular
  - Multi-View
- Pixel correspondences
- Camera pose estimation
  - Pairwise (relative)
  - Multi-View
  - Visual Localization
- Dense 3D reconstruction

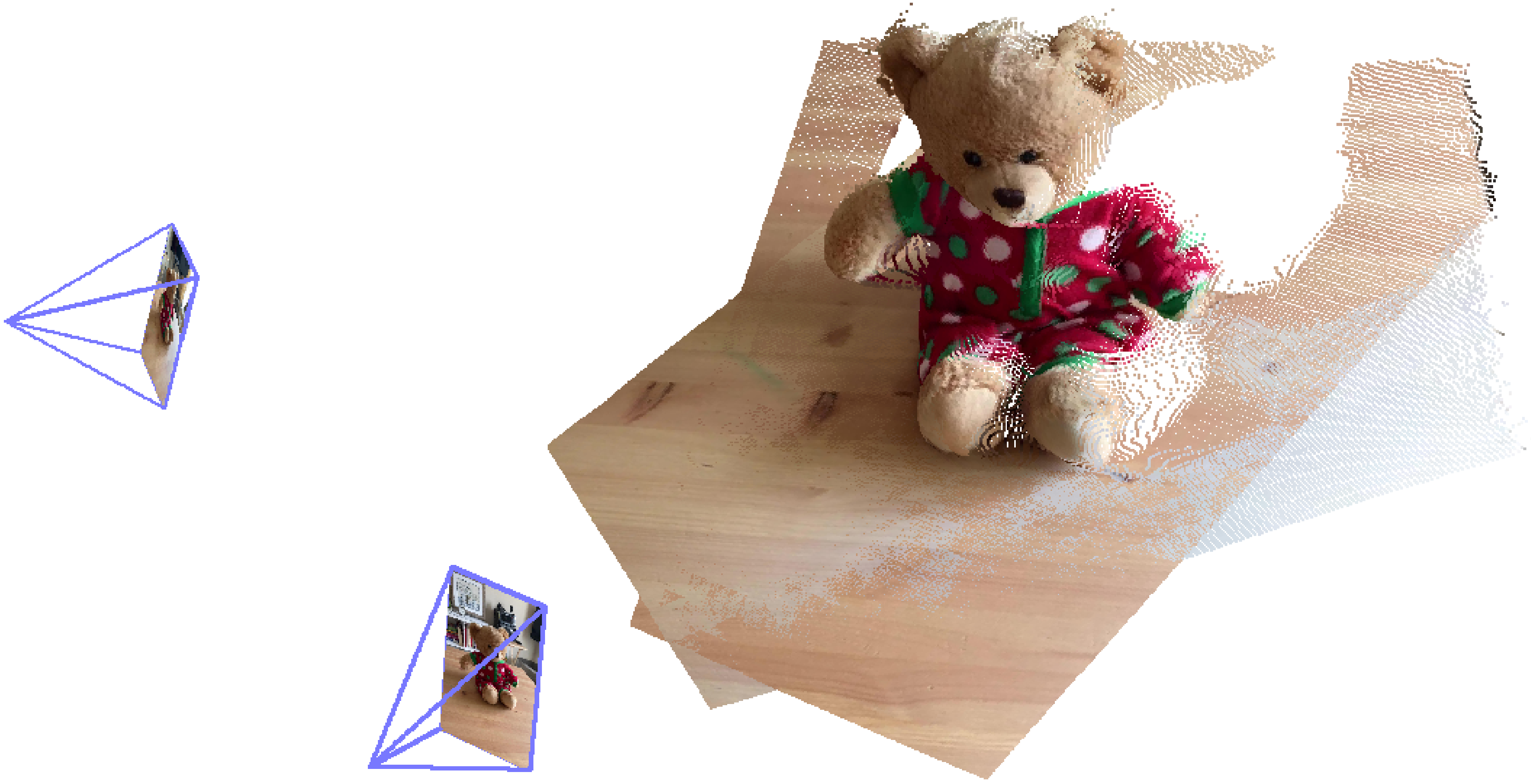
DUSt3R: Dense Unconstrained Stereo 3D Reconstruction

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

Jointly recovering cameras and scene

Jointly recovering cameras and scene



Monocular Input

Monocular Input

Feed same image twice

NAVER LABS
Europe

Monocular Input

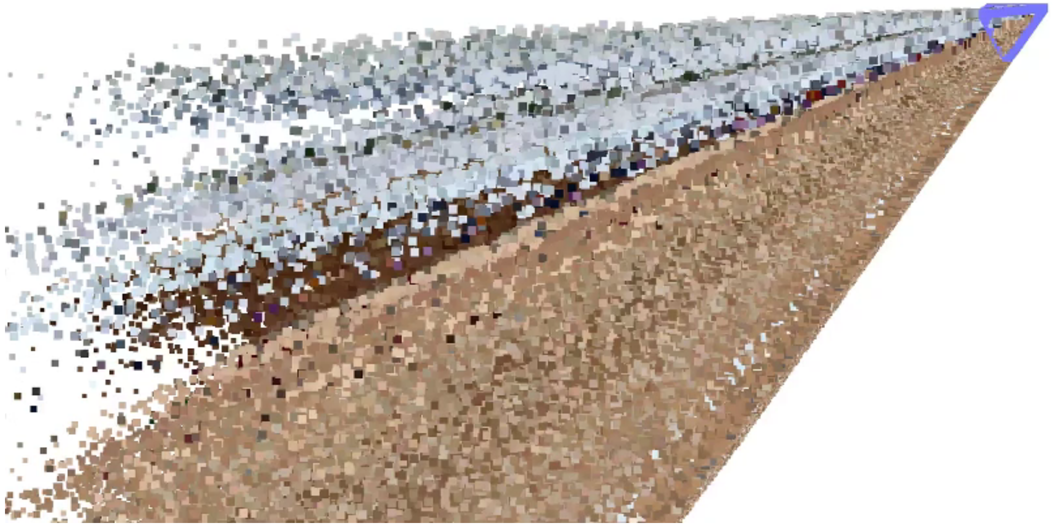
NAVER

DUSt3R

Global alignment

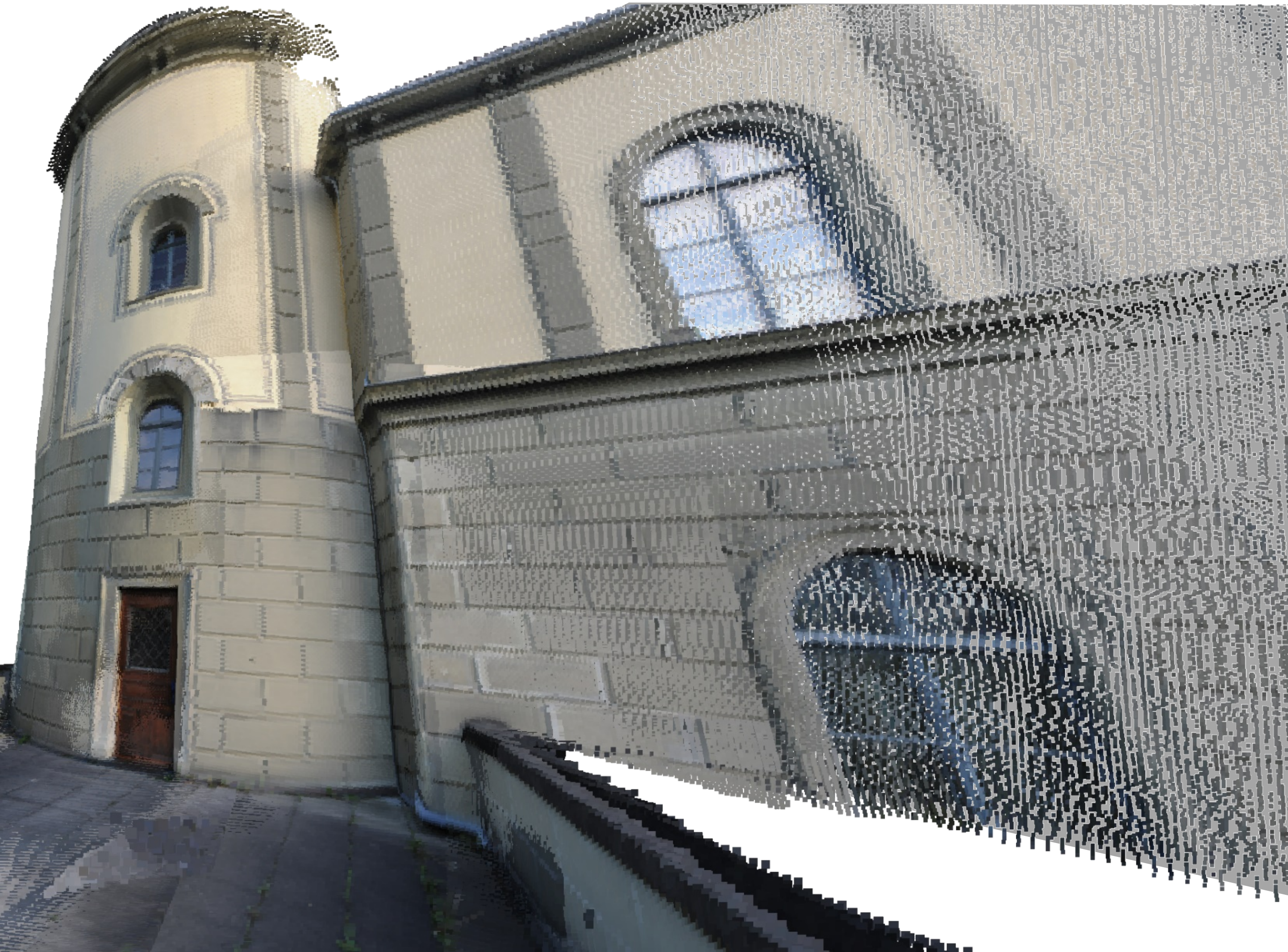
- A fast and simple post-processing optimization for multi-views (takes few seconds)

  = a well-behaved 3D version of bundle adjustment

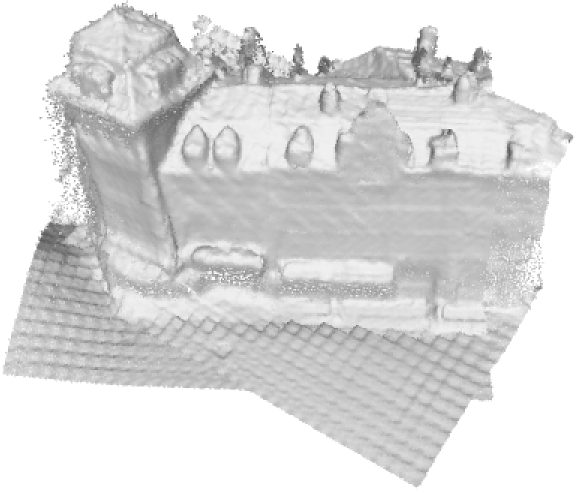


The same model works indoors ...

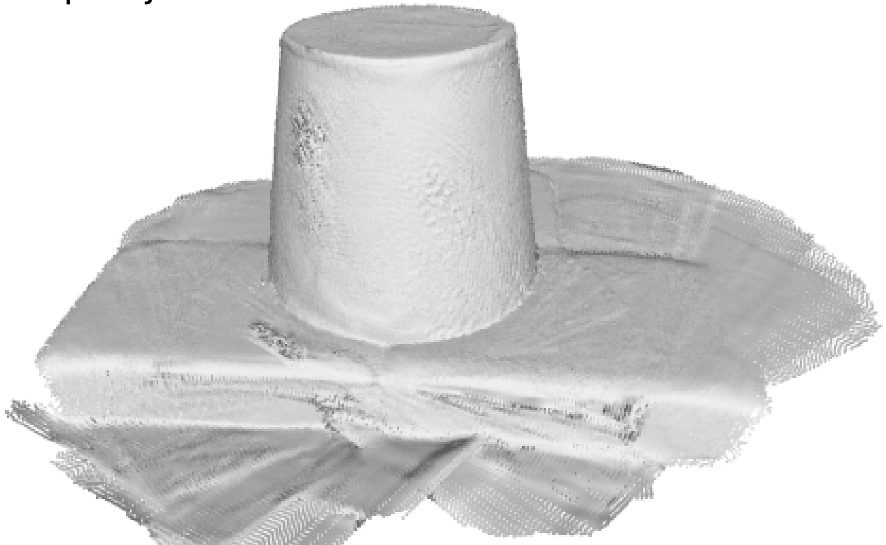
... and outdoors



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

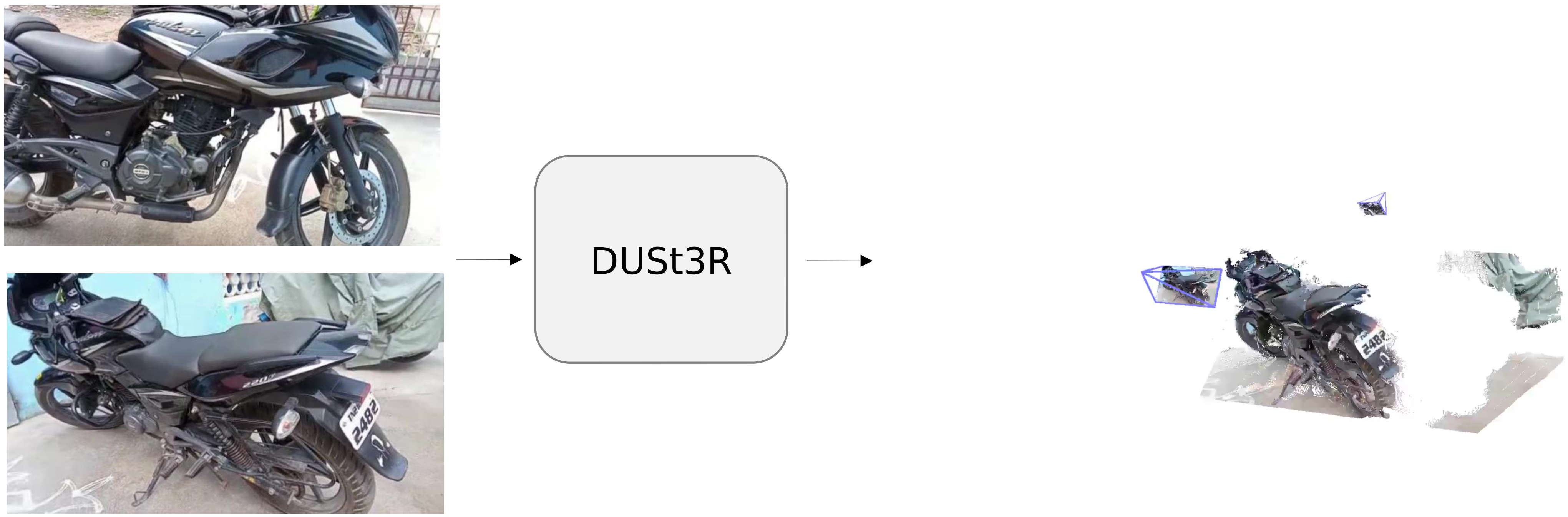


... and close-up objects ...



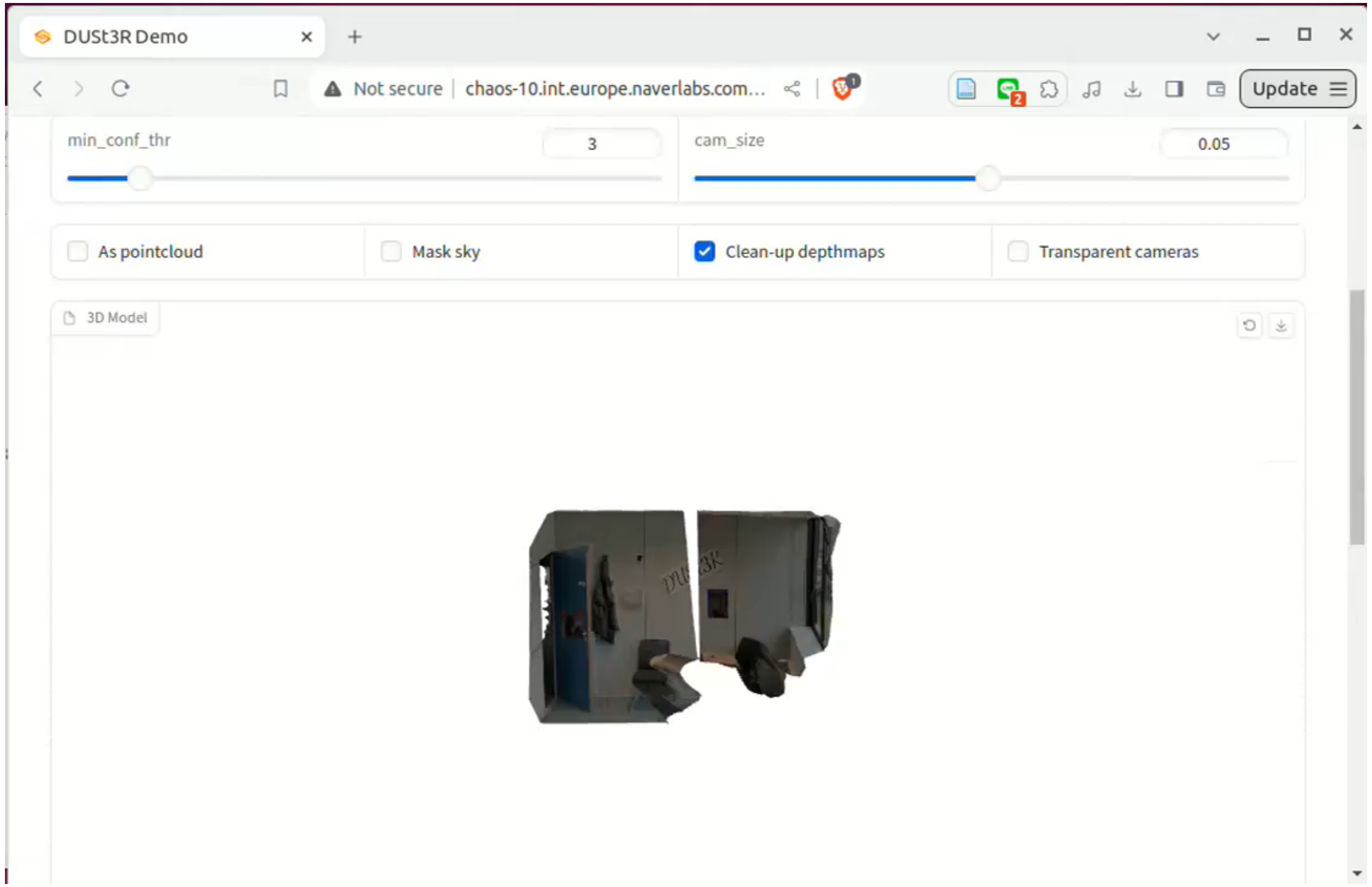
DUSt3R

Opposite view matching



DUSt3R

“impossible matching” = 3D reconstruction without any overlap!



NAVER LABS Europe

NAVER

DUSt3R: Dense Unconstrained Stereo 3D Reconstruction

https://github.com/naver/dust3r

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: limitations

DUSt3R is extremely robust but it lacks accuracy

		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

Not all routes leads to accurate visual localization

- Route 1: DUSt3R → NN in 3D space → pixel correspondences → PnP
- Route 2: DUSt3R → PnP

Methods	GT Focals	7Scenes (Indoor) [48]						
		Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs
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
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

Best results obtained from pixel correspondences

- but DUST3R is not trained explicitly for matching
- What if we did?

MASt3R

Matching And Stereo 3D Reconstruction



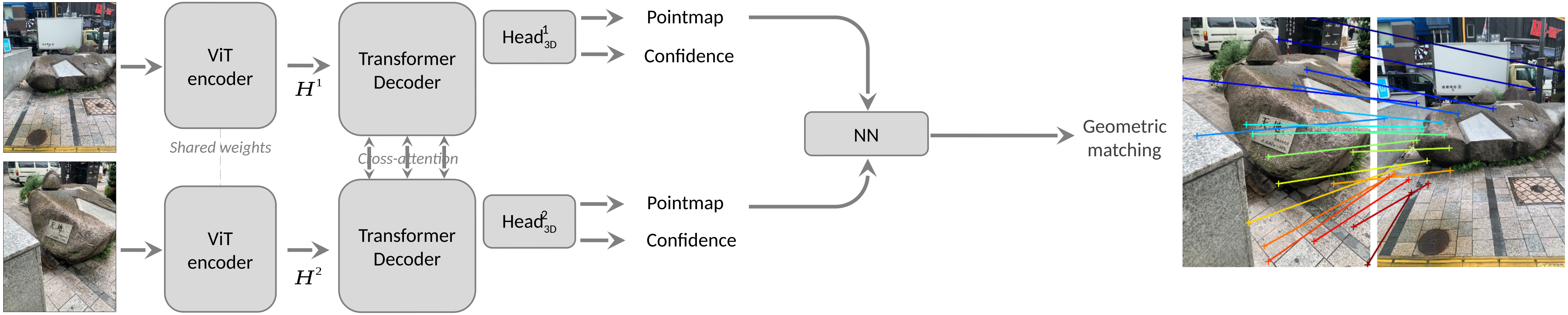
Vincent Leroy
Naverlabs Europe

Yohann Cabon
Naverlabs Europe

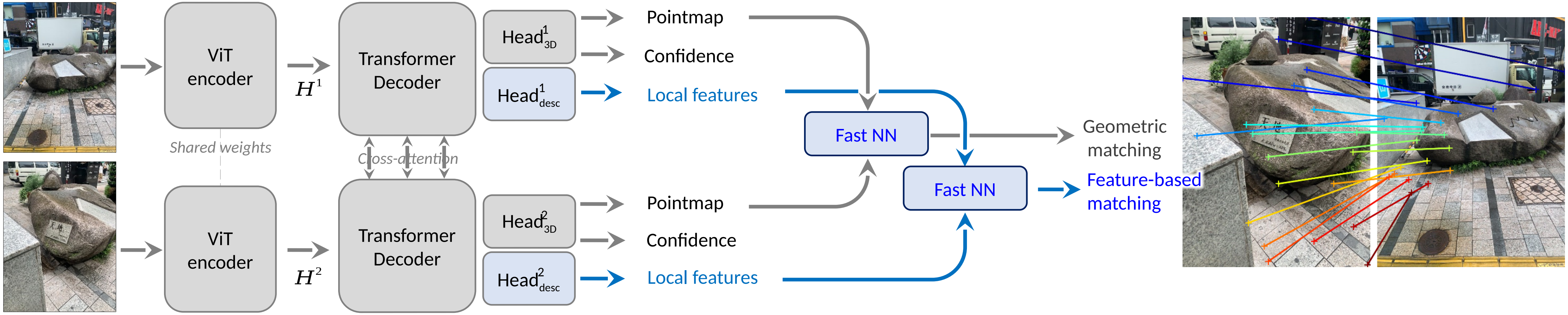
Jérome Revaud
Naverlabs Europe

NAVER LABS Europe

MASt3R: Matching And Stereo 3D Reconstruction



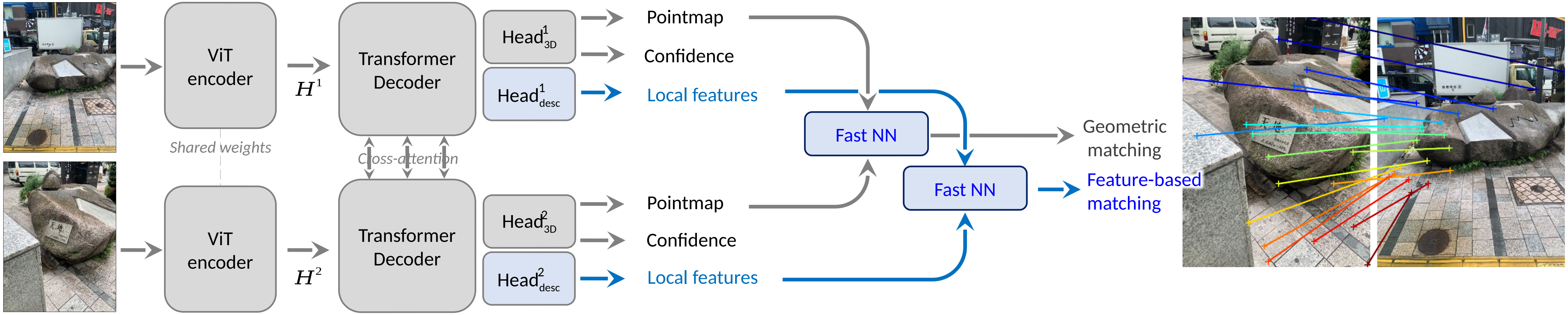
MASt3R:
Matching And Stereo 3D Reconstruction



NAVER LABS Europe

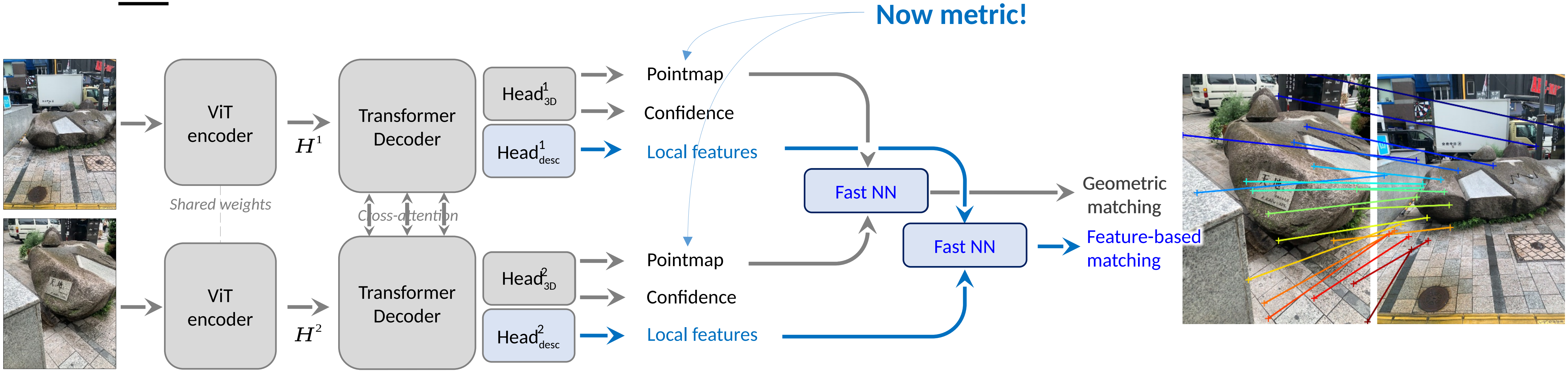
NAVER

MASt3R: Matching And Stereo 3D Reconstruction



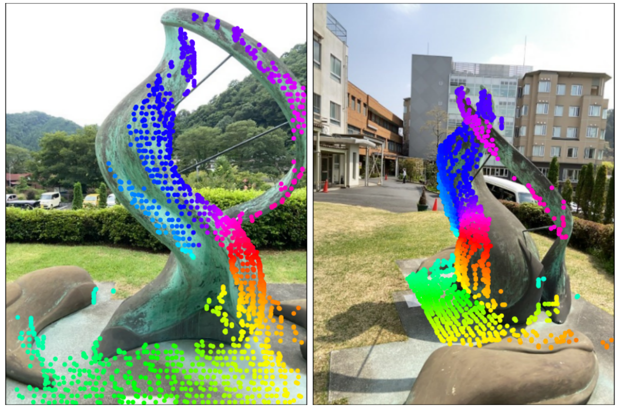
Local Features trained with an InfoNCE loss

MASt3R:
Matching And Stereo 3D Reconstruction



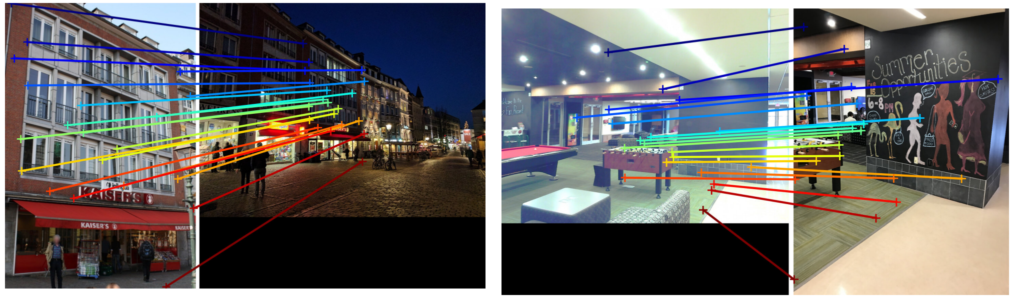
MASt3R

MapFree Relocalization



Relative Pose

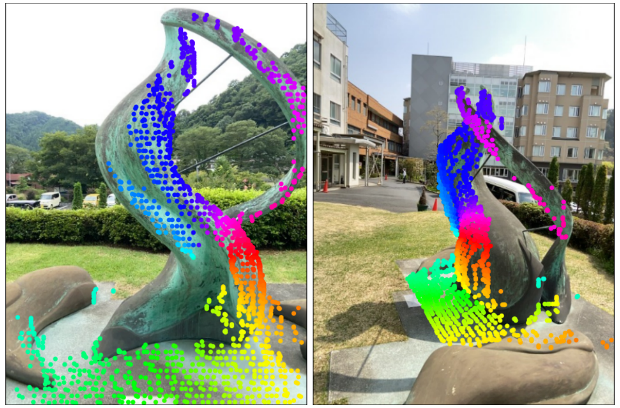
Visual Localization



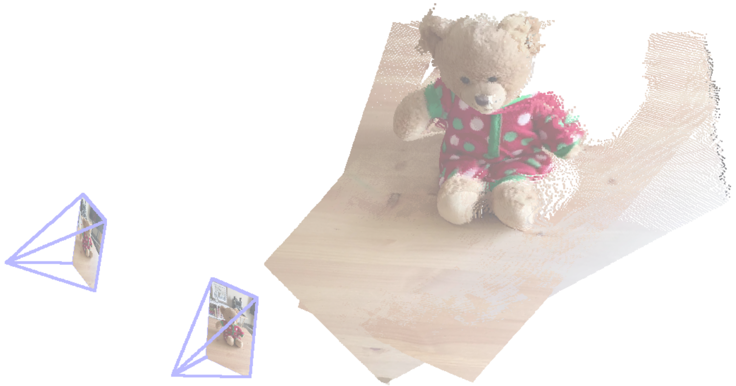
Multi-View Reconstruction

MASt3R

MapFree Relocalization



Relative Pose



Visual Localization

Multi-View Reconstruction

MASt3R: Map-Free Relocalization

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

Image 1

Image 2

Translation is metric $\rightarrow$ Pixel matching alone does not suffice

MASt3R: Map-Free Relocalization

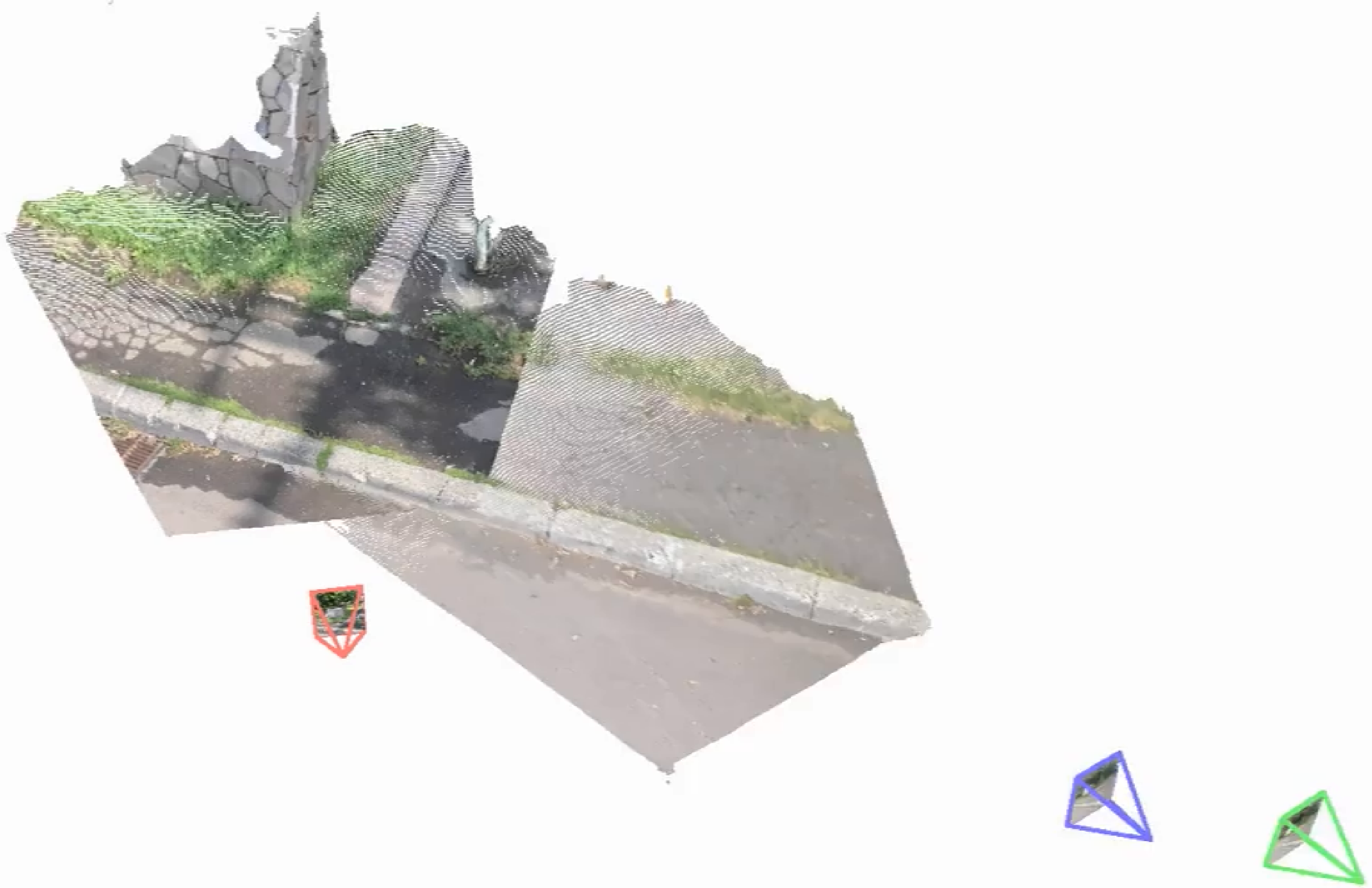
R^{12}, T^{12} ?

Image 1

Image 2

Almost no overlap $\rightarrow$ Pixel matching alone does not suffice

MASt3R: Map-Free Relocalization



MASt3R: Map-Free Relocalization

Evaluation Leaderboard: Single Frame

All Submissions

Method	AUC (VCRE < 45px)	AUC (VCRE < 90px)	Median Trans. Error (m)	Median Rot. Error (°)
MASt3R (Ess.Mat + D.Scale)	0.817	0.933	0.37	2.2
interp_metric3d_loftr_3d2d	0.681	0.796	1.75	31.2
Map-Free Visual Relocalization Enhanced by Instance Knowledge and Depth Knowledge	0.656	0.849	0.83	11.7
RoMa w/ MicKey depth maps	0.604	0.734	1.18	15.6
MicKey trained w/ Overlap Score	0.572	0.748	1.66	27.3
MicKey	0.558	0.741	1.59	26.0
SuperGlue w/ MicKey depth maps	0.556	0.711	1.70	26.1

MASt3R: MVS on DTU

MASt3R: MVS on DTU

Architecture is 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	$\boxed{0.403}$	$\boxed{0.344}$	$\boxed{0.374}$

Matching is far superior to regression

(in *mm* !)

NAVER LABS Europe

MASt3R Demo

NAVER

Hom

74%

IMG_1991.jpg

0

MB)

MASt3R Demo

File

/ ffs / 3drecon

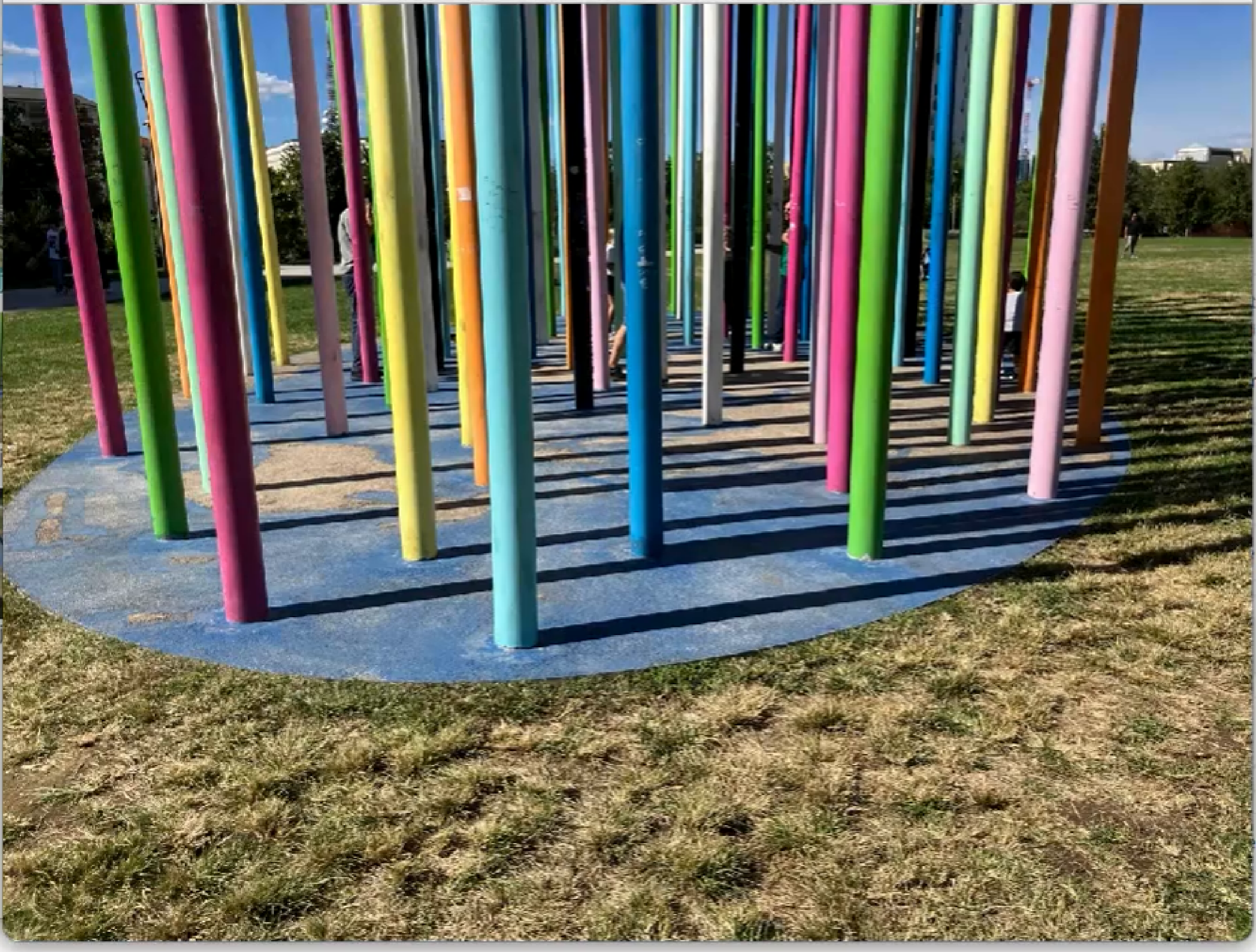
IMG_1976.j

IMG_1980.j

IMG_1984.j

IMG_1980.jpg

69%



tLevel

imization level

refine+depth

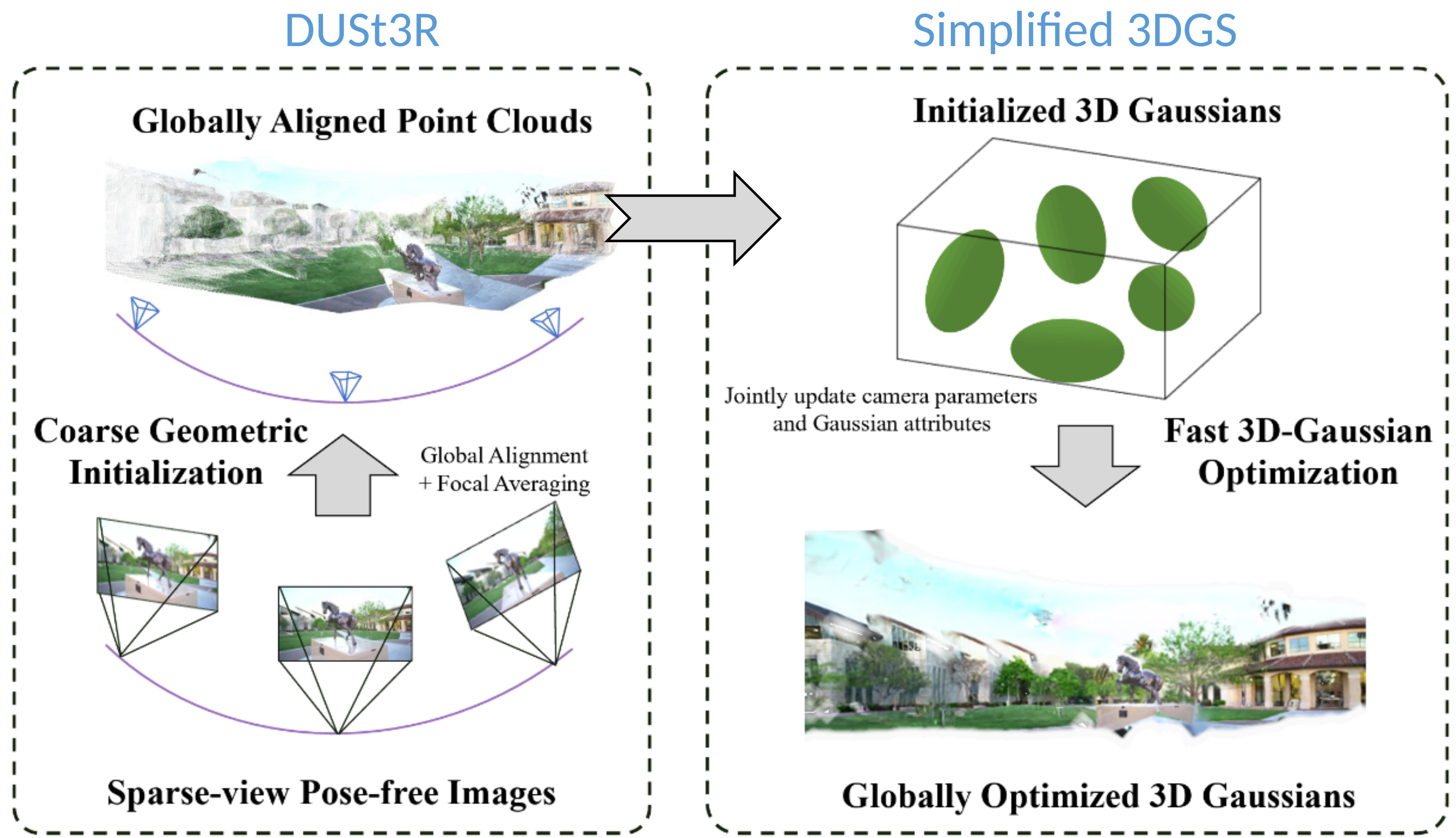
pairs

0

nt cameras

NAVER LABS Europe

InstantSplat: Novel View Rendering ***from scratch*** in seconds



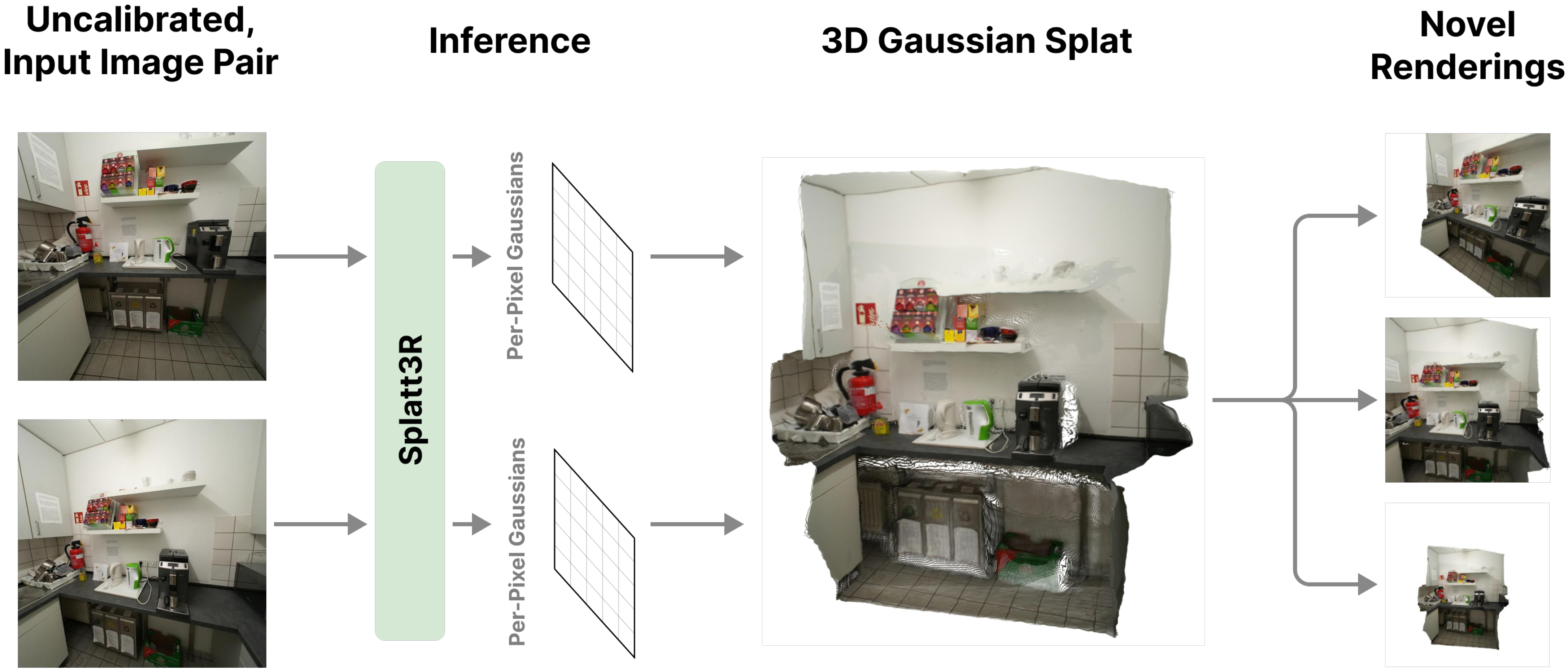
InstantSplat: Unbounded Sparse-view Pose-free Gaussian Splatting in 40 Seconds. Zhiwen Fan et. al., [arXiv'24]. https://instantsplat.github.io

InstantSplat

Result with only 3 input images in 20 seconds from scratch

Splatt3r

Brandon Smart · Chuanxia Zheng · Iro Laina · Victor Adrian Prisacariu, **University of Oxford**

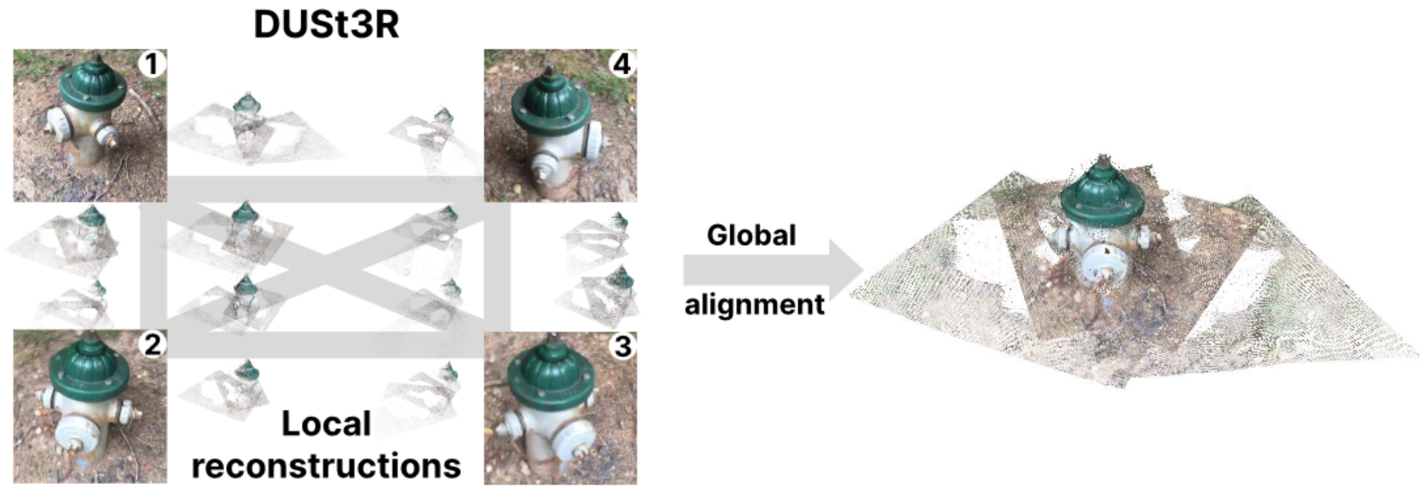


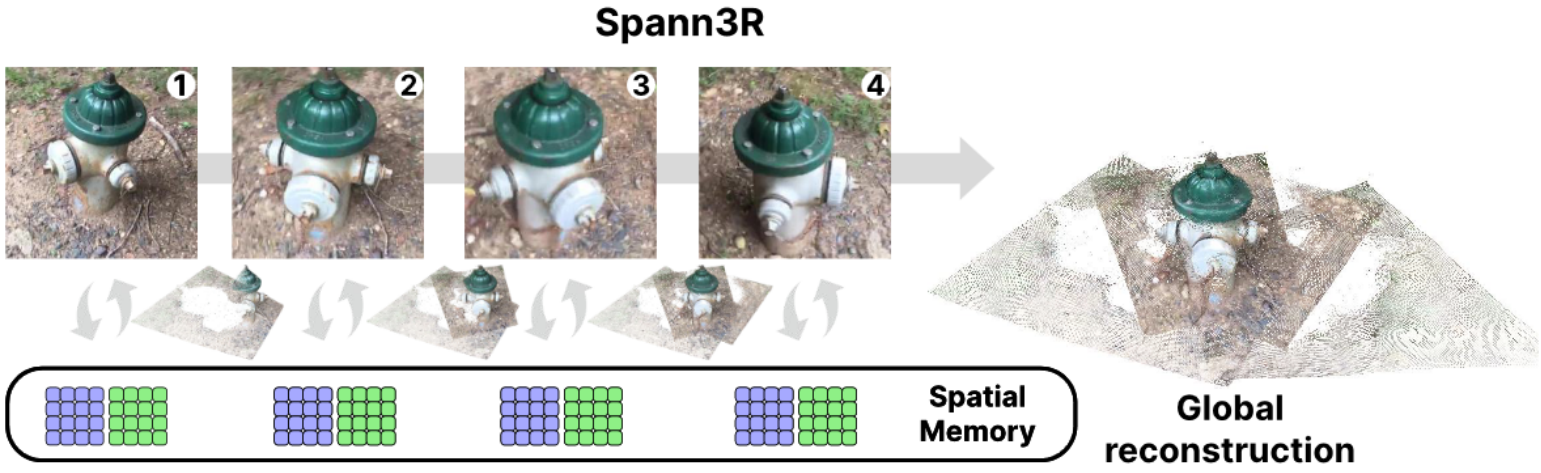
https://github.com/btsmart/splatt3r

NAVER LABS Europe

Spann3R

Hengyi Wang · Lourdes Agapito, University College London





https://github.com/HengyiWang/spann3r

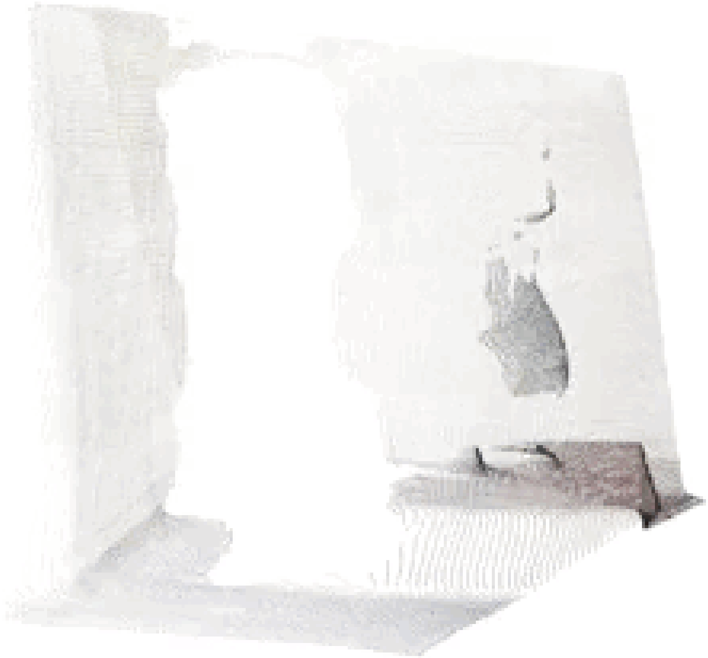
NAVER LABS Europe

Spann3R

Hengyi Wang, Lourdes Agapito University College London

RGB image collection
(w/o known camera params)

Spann3R

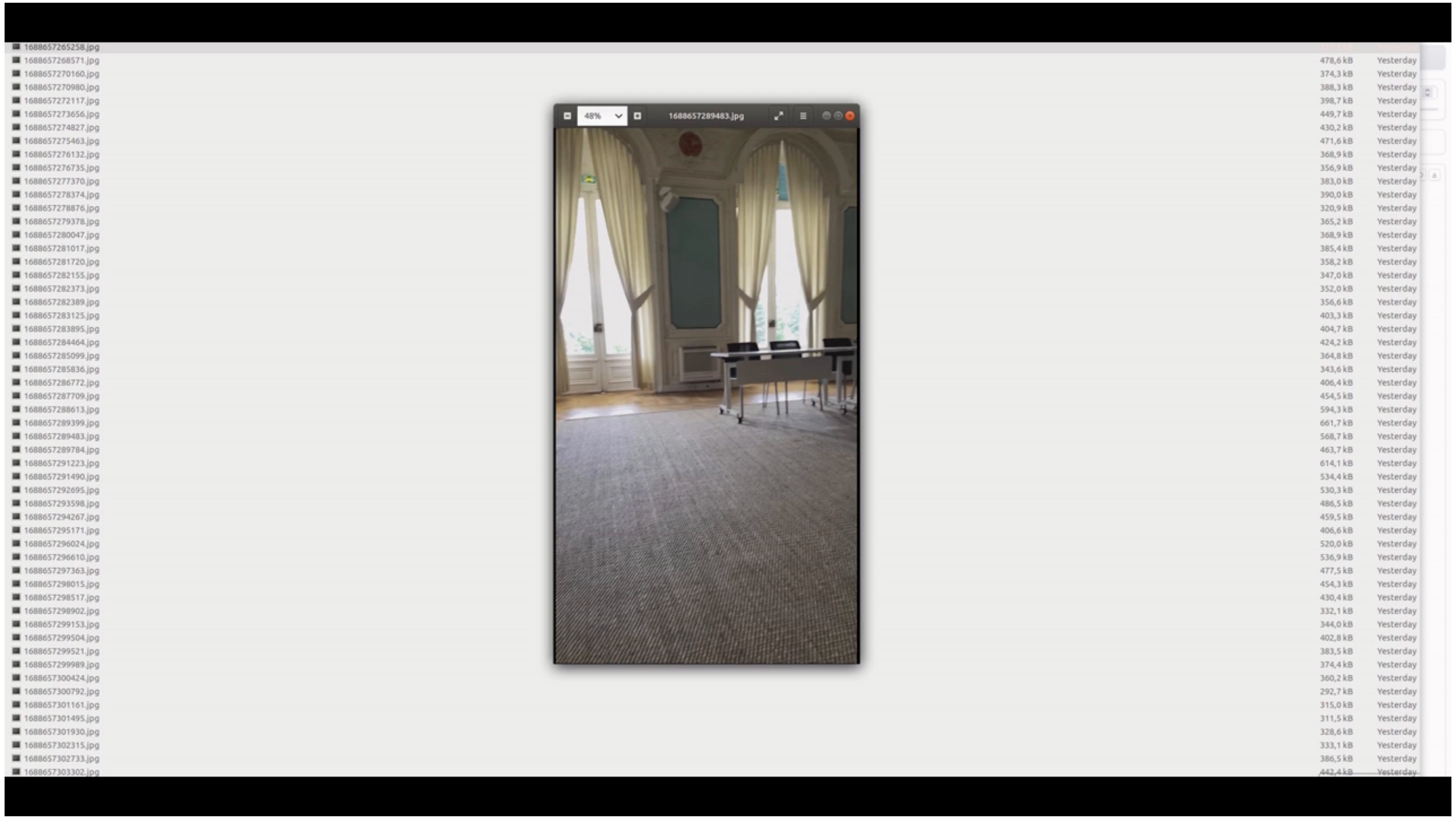


Incremental reconstruction
(>50 keyframes/sec)

https://github.com/HengyiWang/spann3r

Questions?

https://github.com/naver/croco **https://github.com/naver/dust3r** **https://github.com/naver/mast3r**



MASt3R is a network of the St3R series capable of robustly estimating:

- focal lengths
- metric camera poses
- metric geometry for large image collections (mapping)
- accurate correspondences even in extreme cases