Hands-On AI Based 3D Vision
Summer 25

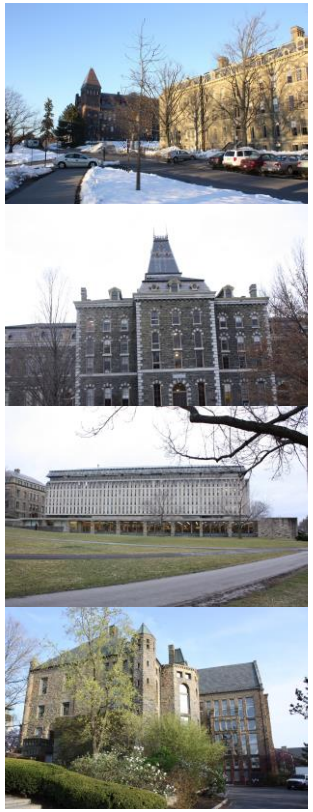
Lecture 09 – Learning-Based 3D Reconstruction

Prof. Dr. Gerard Pons-Moll

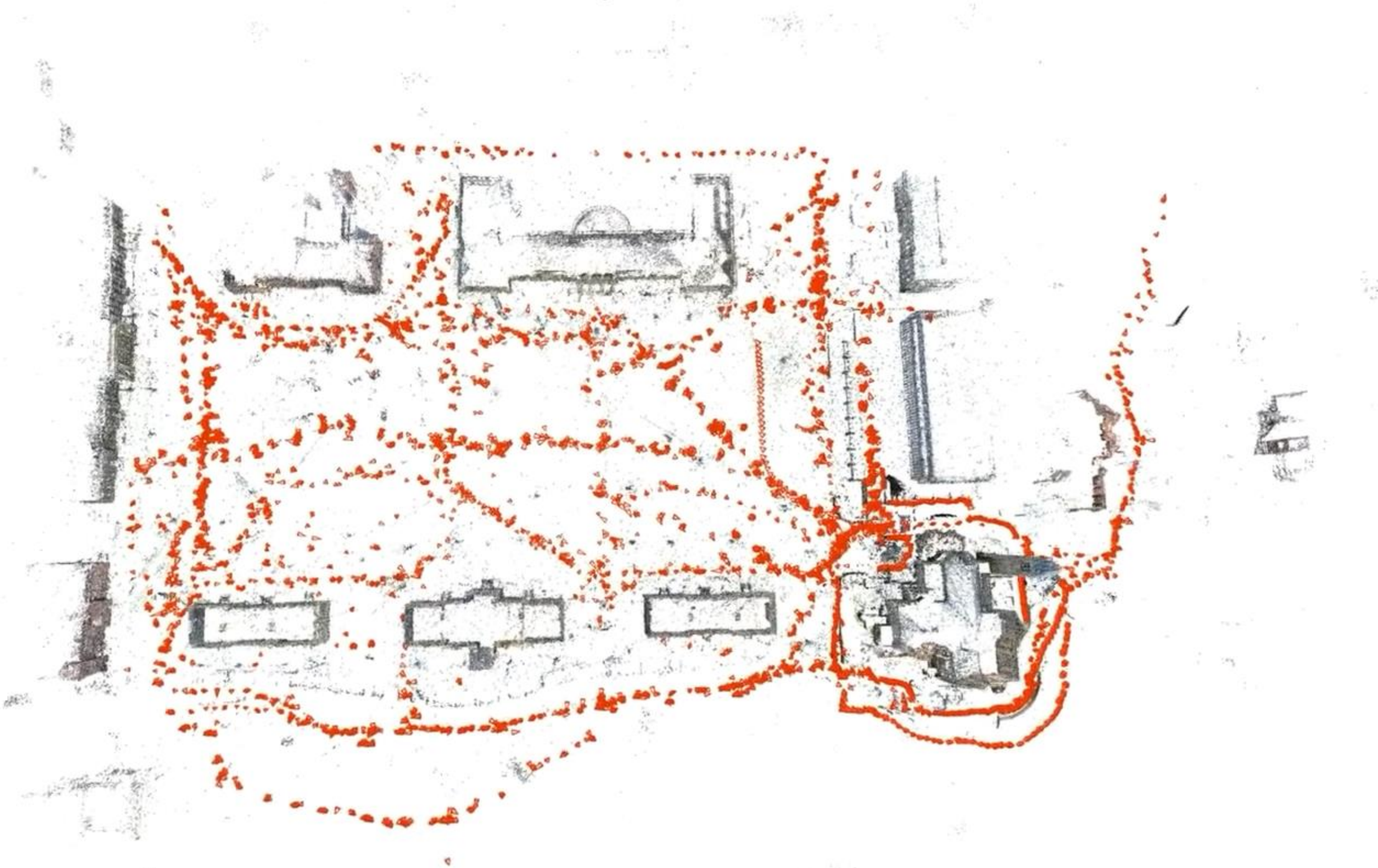
University of Tübingen / MPI-Informatics

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Reconstruction: Core of 3D

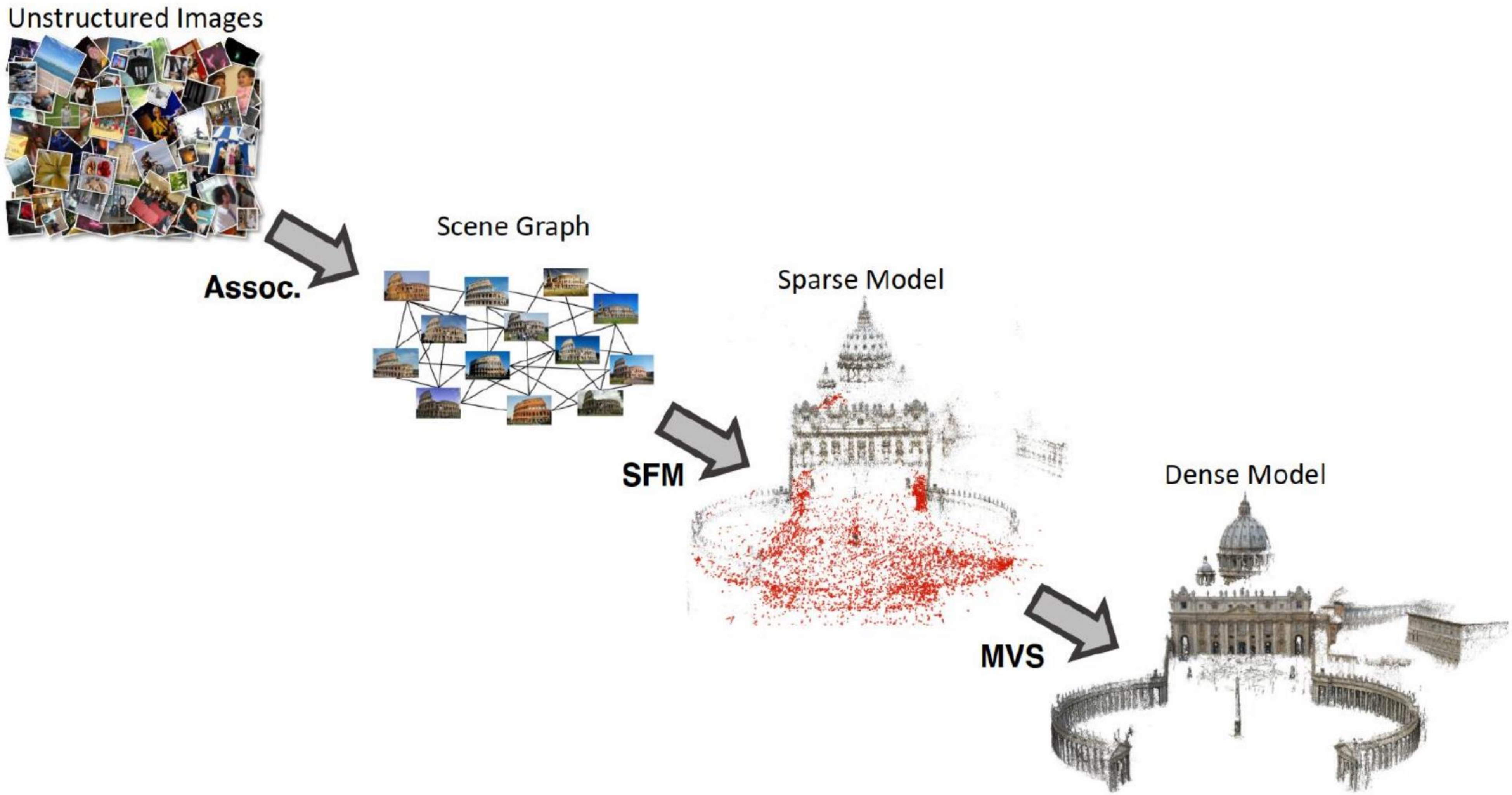


...

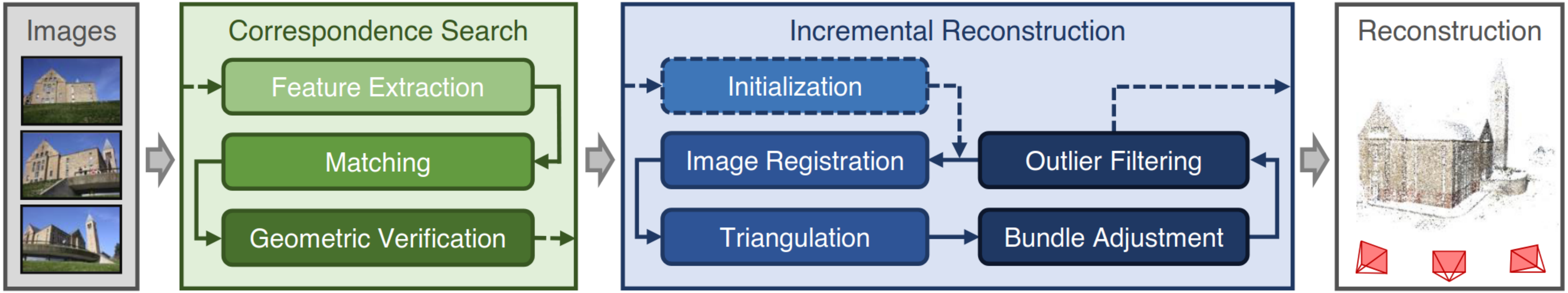


Video Source: COLMAP, Schönberger et al. 2016

Classical Reconstruction Pipeline



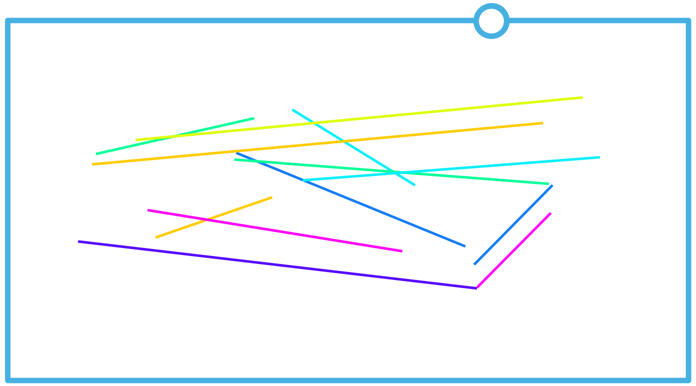
COLMAP: SotA Incremental SfM Pipeline



- **Requires enough images with overlaps**
- **Many subproblems**: Point Matching, Essential Matrix Estimation, Triangulation, Pose Estimation, …
- **No subproblem is solved perfectly**
- **No communication between components**
- **Brittle and prone to errors -> error propagation**
- **Slow (repeated BA)**

Bottleneck: Bundle Adjustment

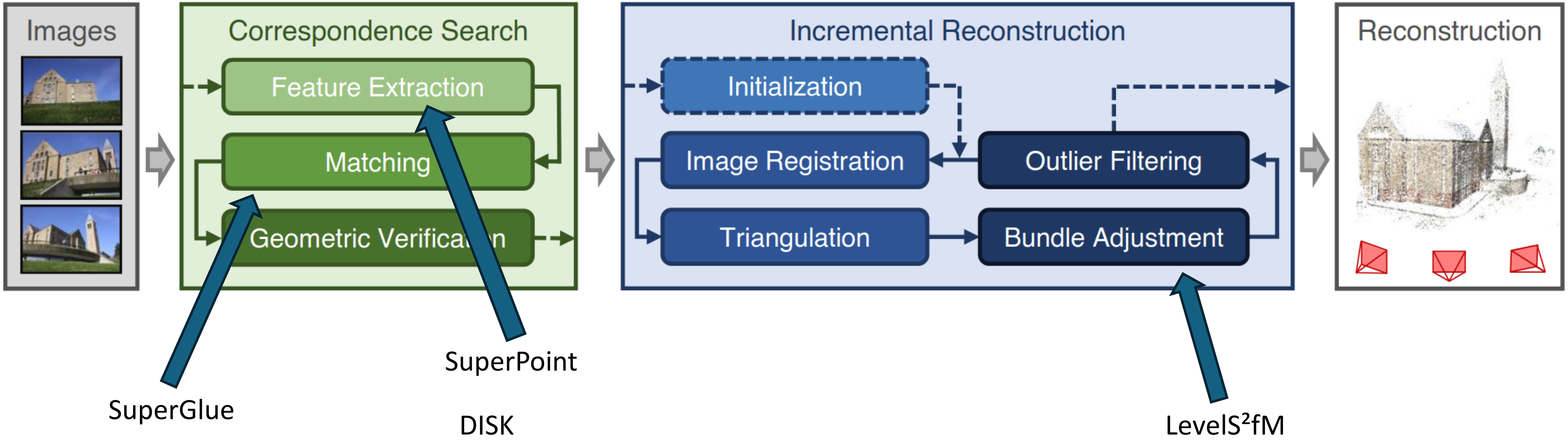
COLMAP



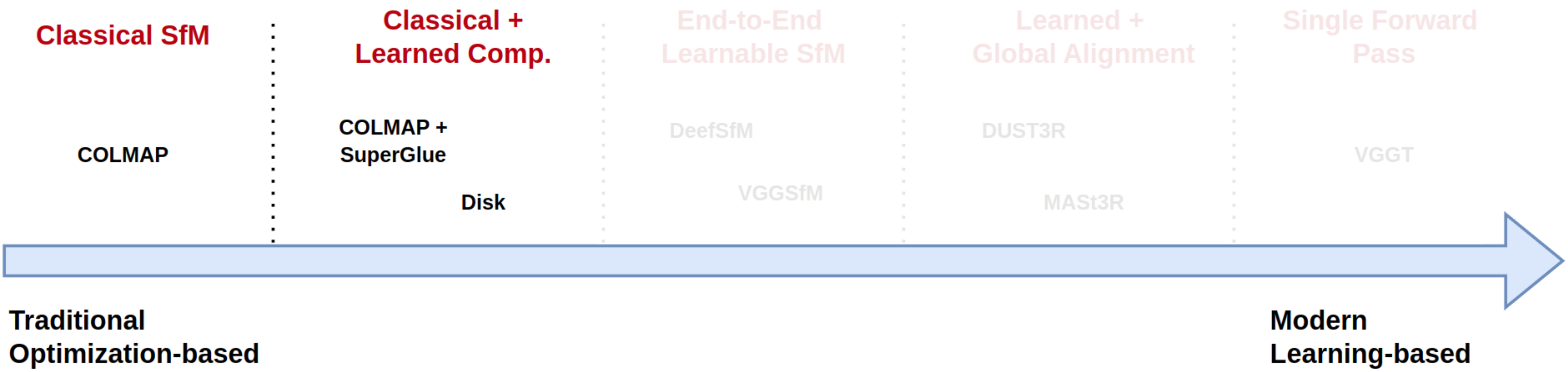
Bundle Adjustment

Traditional with Learned Components

Recent trend: Replace certain parts of SfM pipeline with learned modules



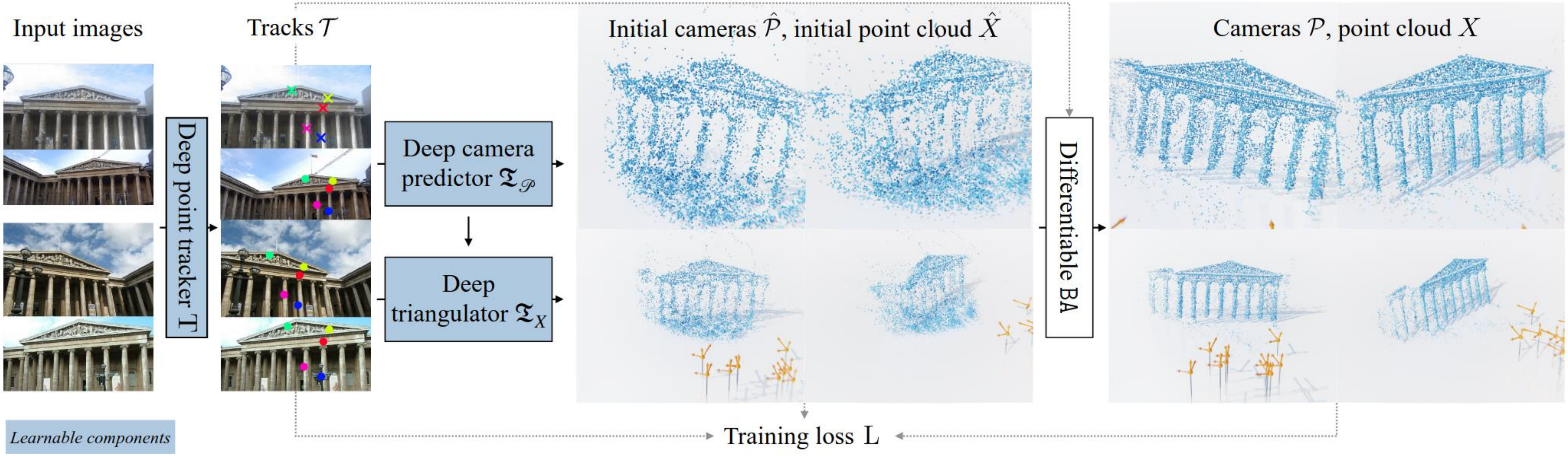
A Spectrum of Methods



- Epipolar Geometry
- Pose Estimation
- Triangulation
- Bundle Adjustment
- ...

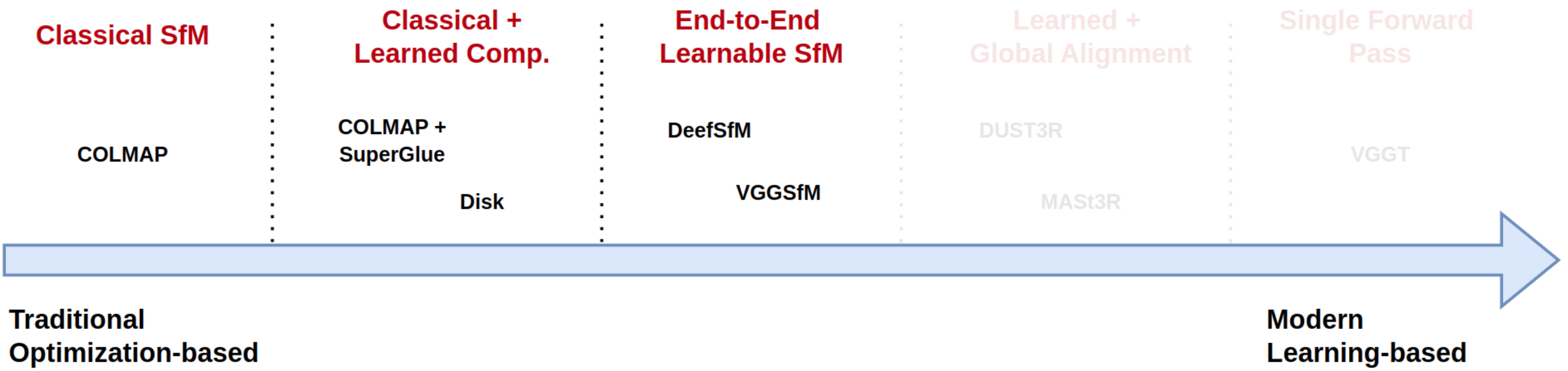
- Feature Extraction / Matching
- Monocular Depth Prediction
- Cross-View Attention
- ...

End-to-End Learnable SfM Pipelines



Visual Geometry Grounded Deep Structure From Motion, Wang et al. CVPR '24
DeepSFM: Structure From Motion Via Deep Bundle Adjustment, Wei et al. ECCV '20

A Spectrum of Methods

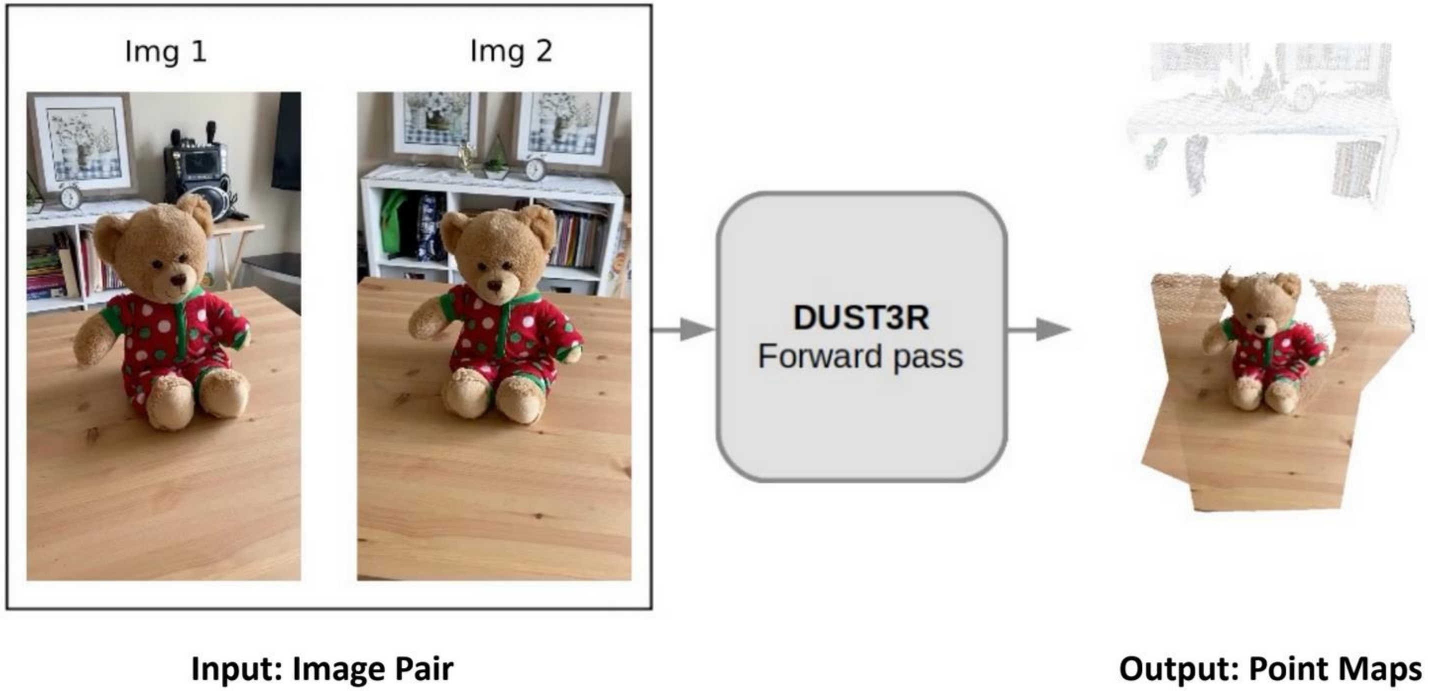


- Epipolar Geometry
- Pose Estimation
- Triangulation
- Bundle Adjustment
- ...

- Feature Extraction / Matching
- Monocular Depth Prediction
- Cross-View Attention
- ...

DUSt3R: Shifting The Paradigm

DUSt3R takes **unposed** images **without prior information about camera calibration** as input



First steps towards 3D foundation models?

DUSt3R, Wang et al. CVPR '24

Point Maps

- Dense, pixel-aligned 3D point cloud

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

- Forms a 1-to-1 mapping between image pixels and 3D scene points

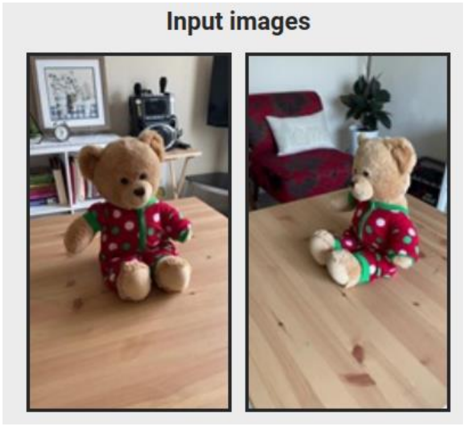
$$\mathbf{I}_{i,j} \leftrightarrow \mathbf{X}_{i,j}$$

- More structured than point cloud

- Easy conversion:

  Pointmap ↔ Depth Map

DUSt3R Task

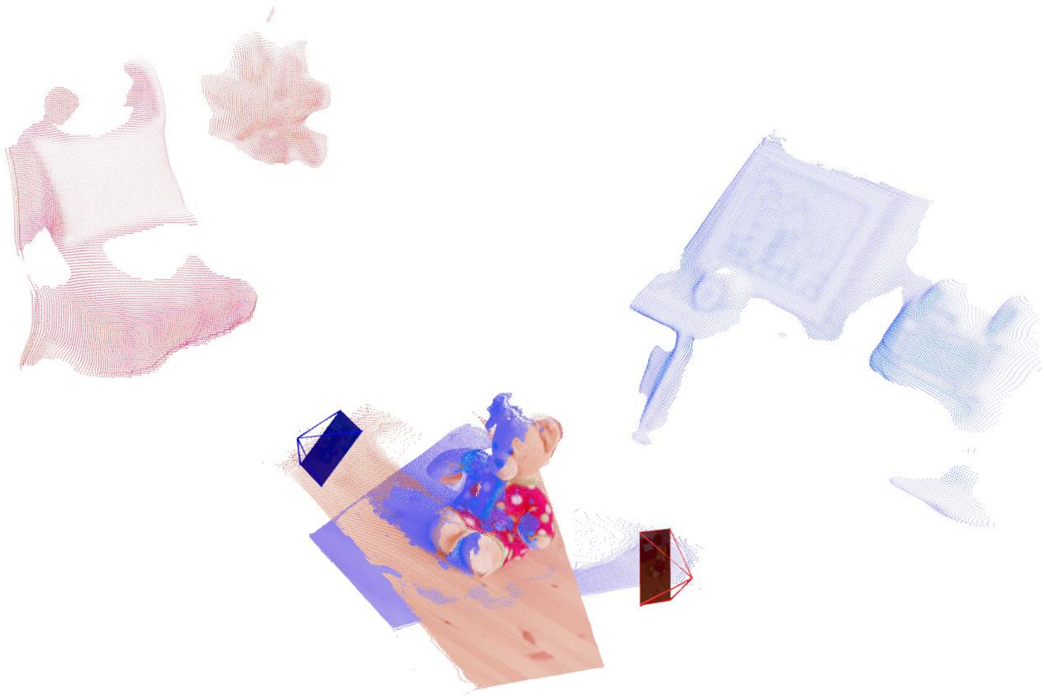


Input:

- Two images of a scene
- Different viewpoints

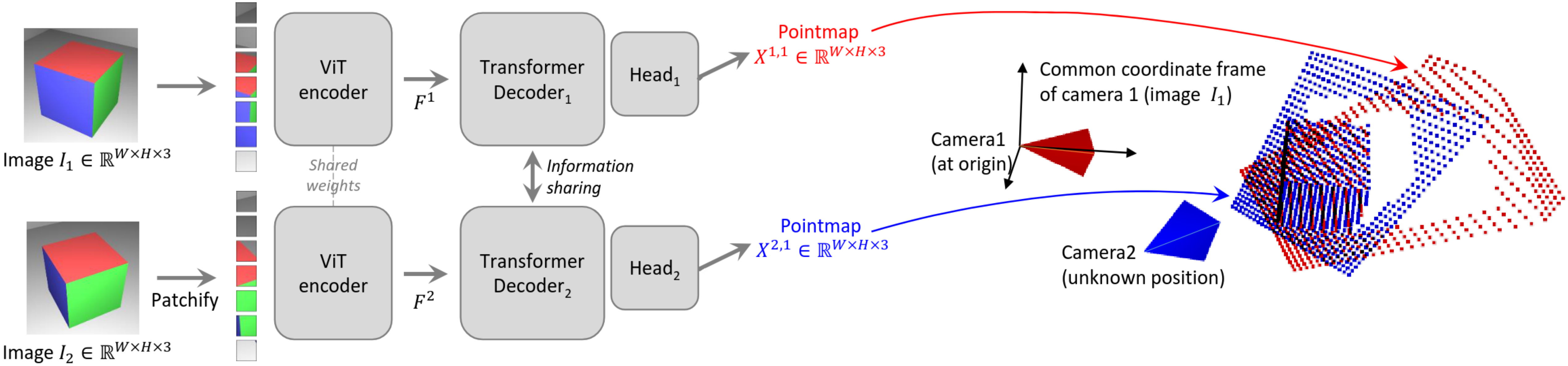
Output:

- Two pointmaps
- **Aligned** in C1's frame



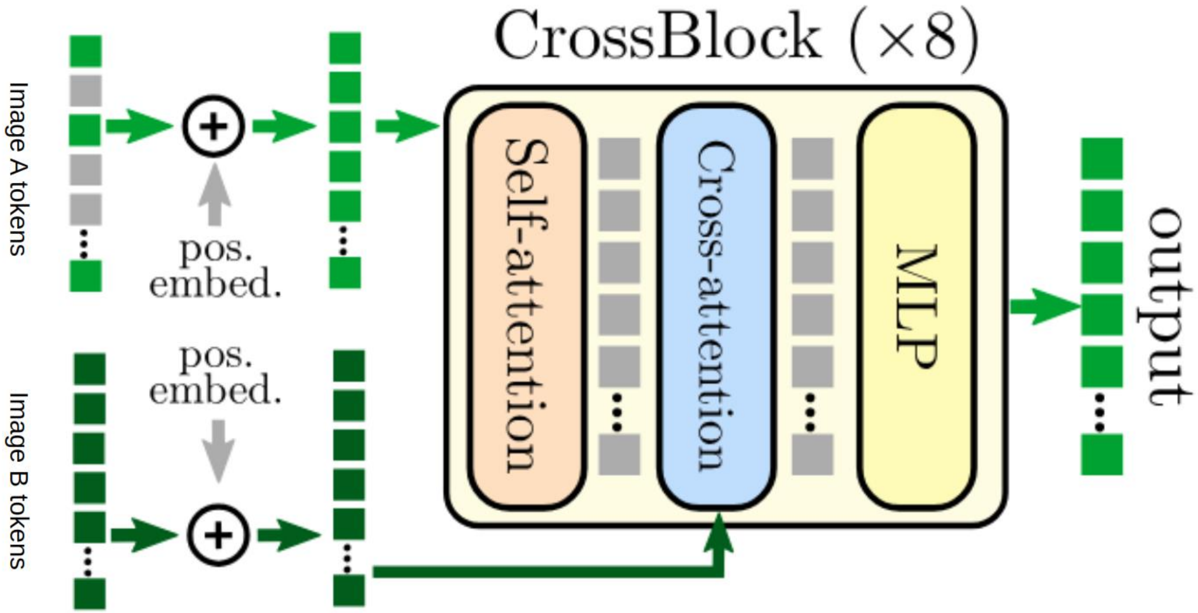
DUSt3R Architecture (without confidence)

- Siamese vision transformer encoder
- Multi-block transformer decoder that shares information between views via cross-attention
- Separate regression heads output pointmaps in **frame of cam 1**



Dust3R Decoder Blocks

- Self attention across all patches of an image
- Cross attention for information sharing between images
- Multiple blocks in series



CroCo, Weinzaepfel et al. NeurIPS '22

DUSt3R Training Objective

3D Regression Loss

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

View (1 or 2)

Image Pixel

Normalizing Factor

Predicted Point

Ground Truth Point

$$z = \frac{1}{|\mathcal{D}^1| + |\mathcal{D}^2|} \sum_{v \in \{1,2\}} \sum_{i \in \mathcal{D}^v} \|X_i^v\|$$

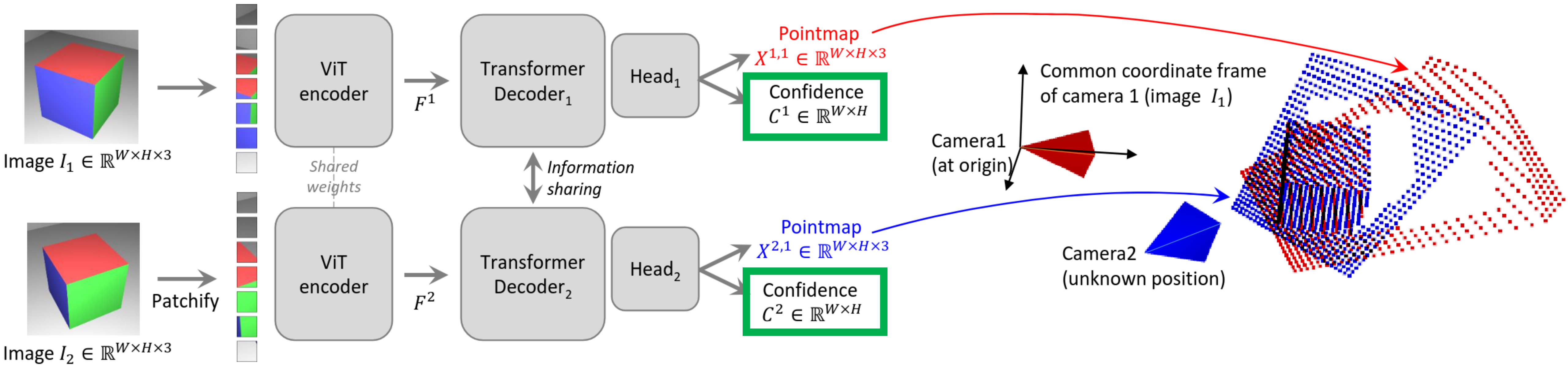
Dealing With Ambiguous Points

What are the ground truth positions for these points?

Even we humans are not confident.
Can we make a model with confidence scores?

Confidence-Aware Model and Training

- Head not only regresses pointmap but also gives confidence score
- How do we train this without ground truth?



Confidence-Aware Loss

Overall Training Loss:

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

Confidence of pixel i in view v

Regression Loss

Regularization weight /
Penalty for uncertainty

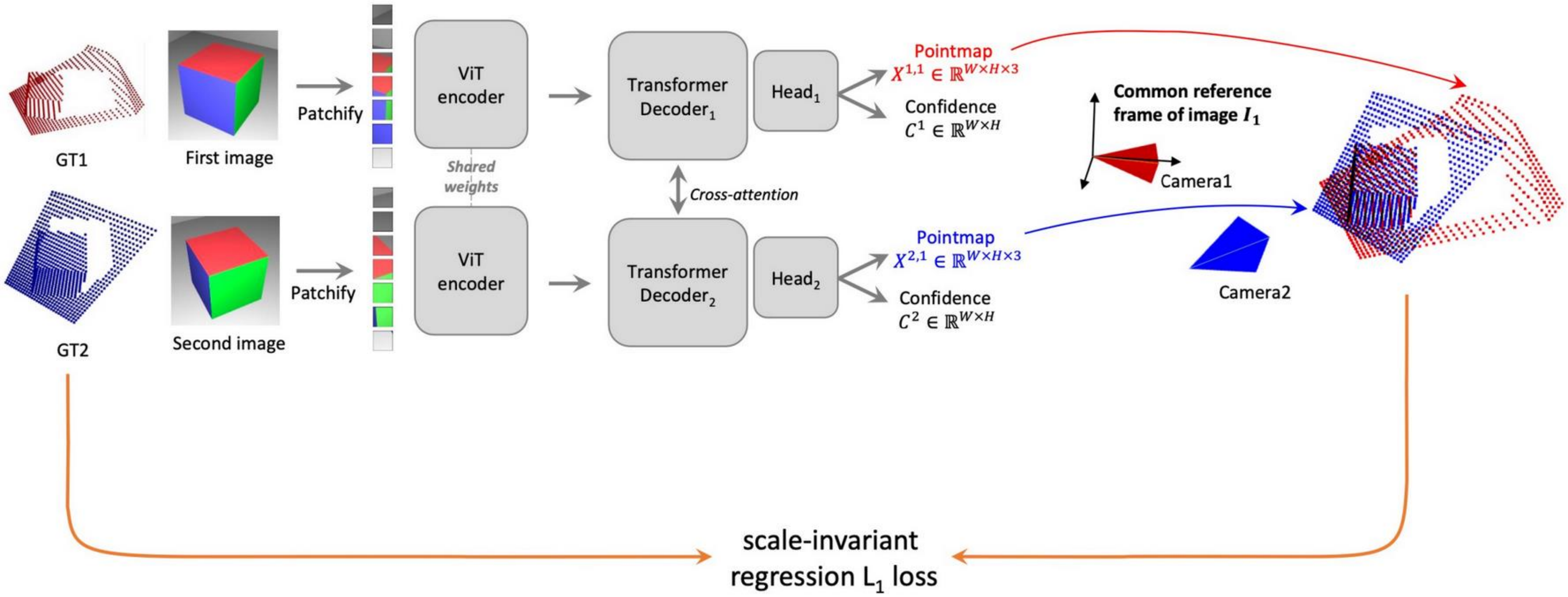
To force extrapolation in uncertain areas:

$$C_i^v = 1 + \exp \tilde{C}_i^v$$

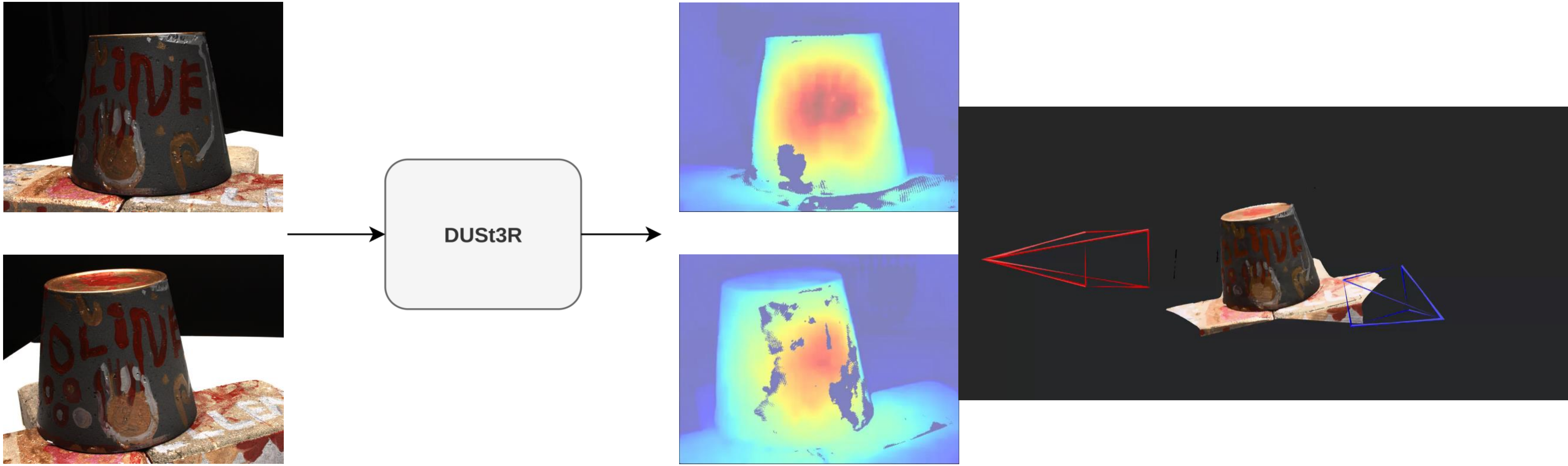
Network output

Confnet: Predict with confidence, Wan et al. ICASSP '18

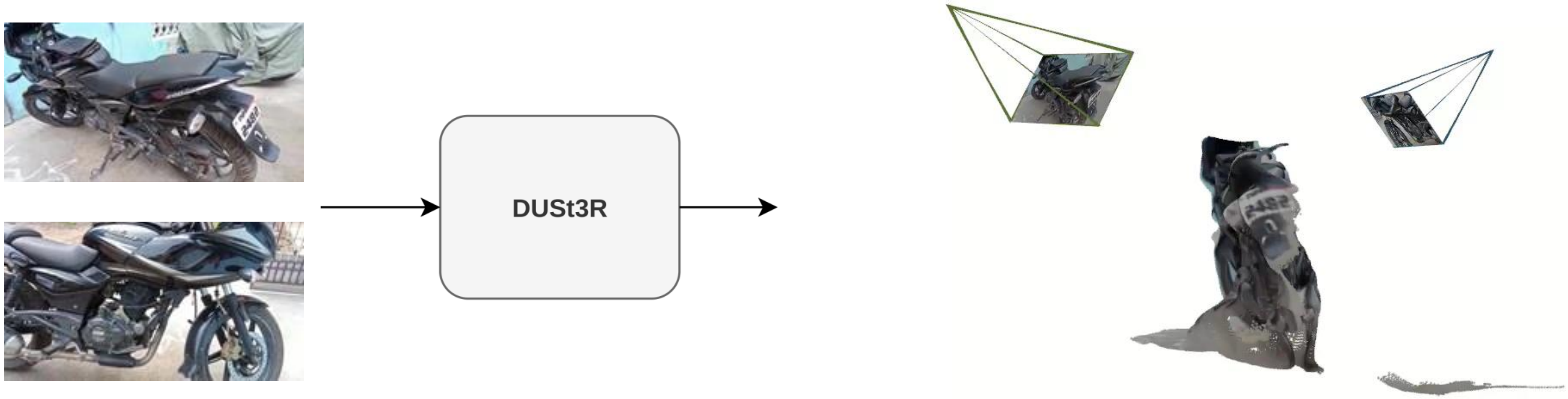
DUSt3R Final Pipeline



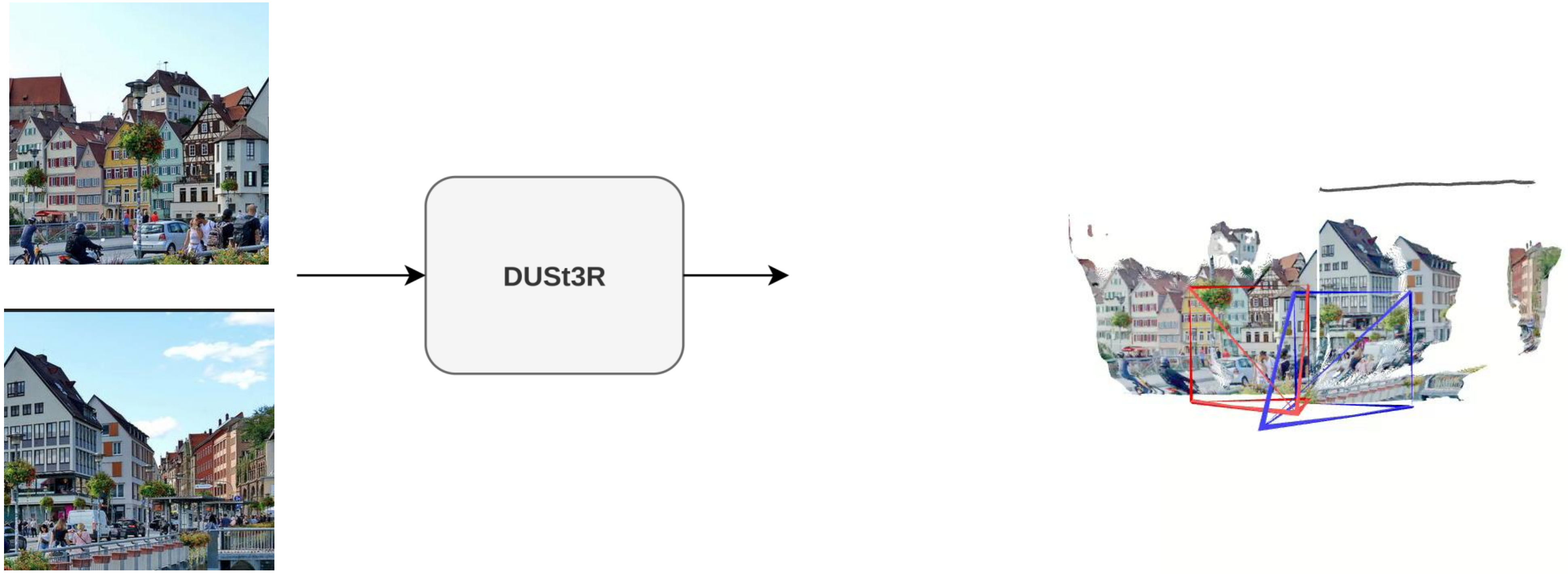
Example: Reconstruction + Confidence Maps



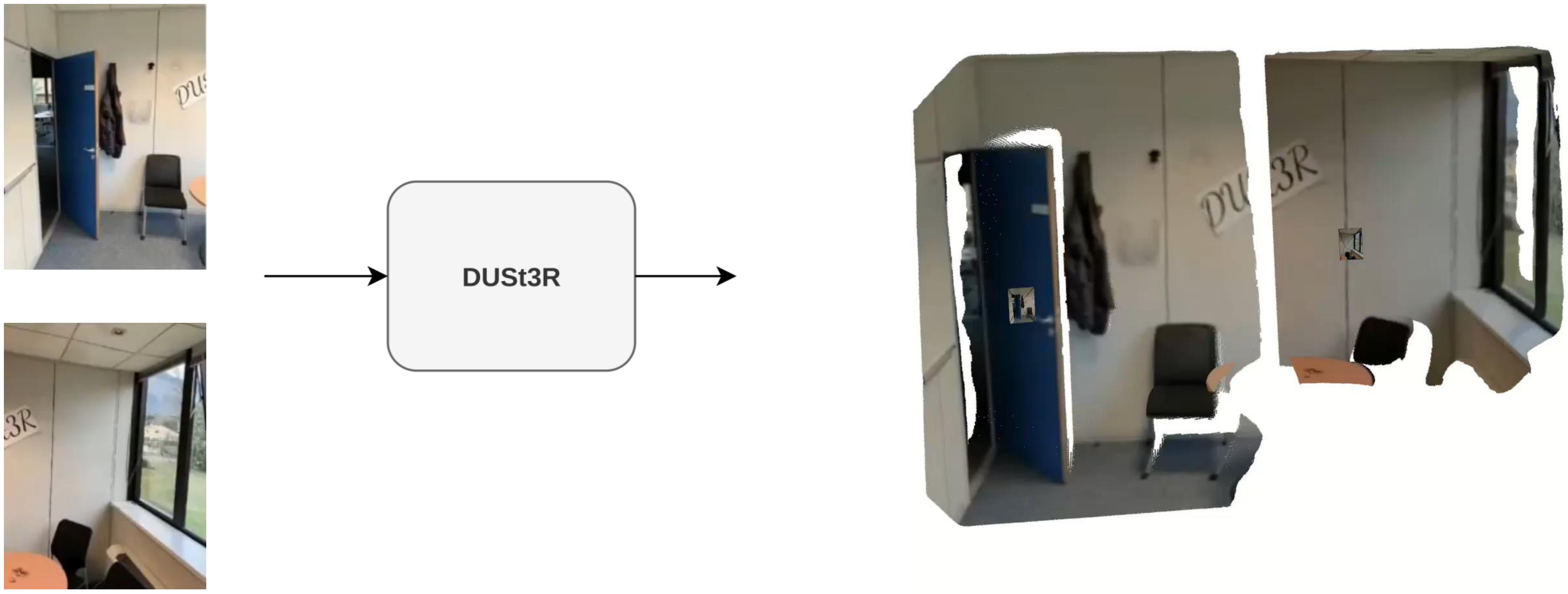
Heavy Viewpoint Changes



No Overlap



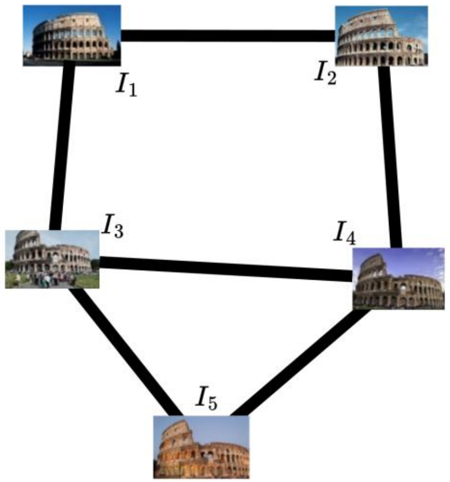
No Overlap



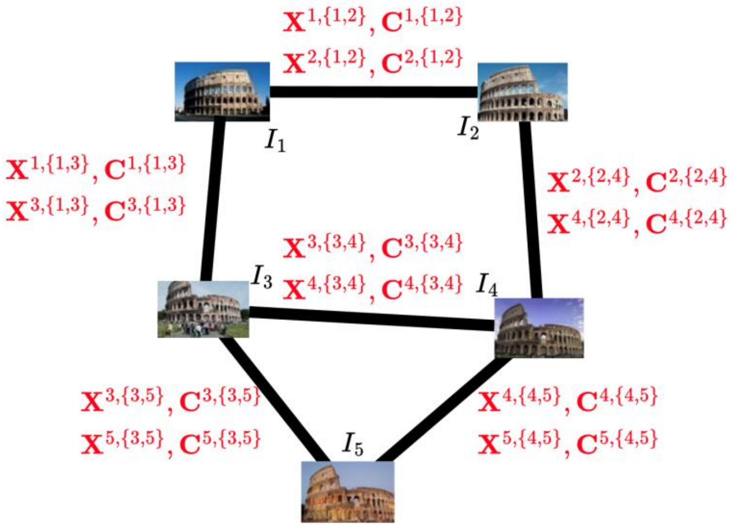
Multi-View Reconstruction

- DUSt3R only takes 2 views as input, what if we have more views?
- We are interested in *globally aligned pointmaps* $\{\chi^v \in \mathbb{R}^{W\times H\times 3}\}_{v\in V}$
- Requires rotating/scaling pairwise predictions into common world frame

1. Scene Graph



2 . Pairwise Reconstruction



3. Global Optimization

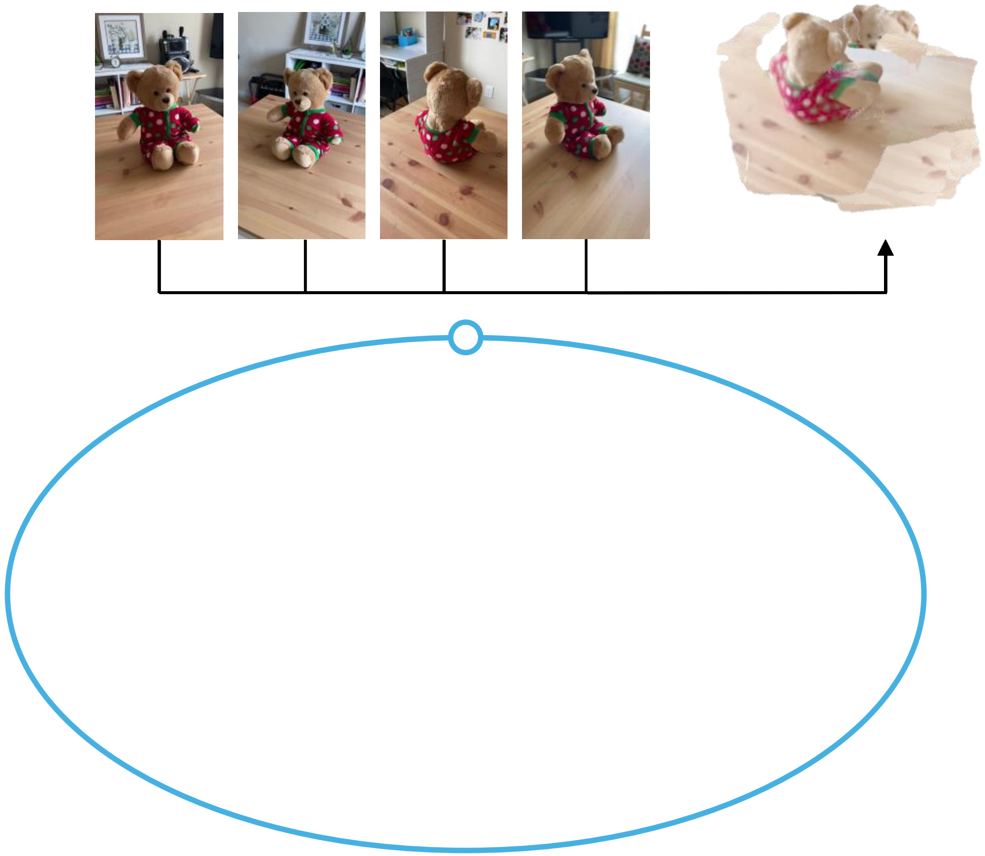
Optimize for

- Per-edge scale σ_e
- Per-edge rigid transform $P_e \in \mathbb{R}^{3\times 4}$
- Per-view global pointmap χ^n

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

Multi-View Alignment via Optimization

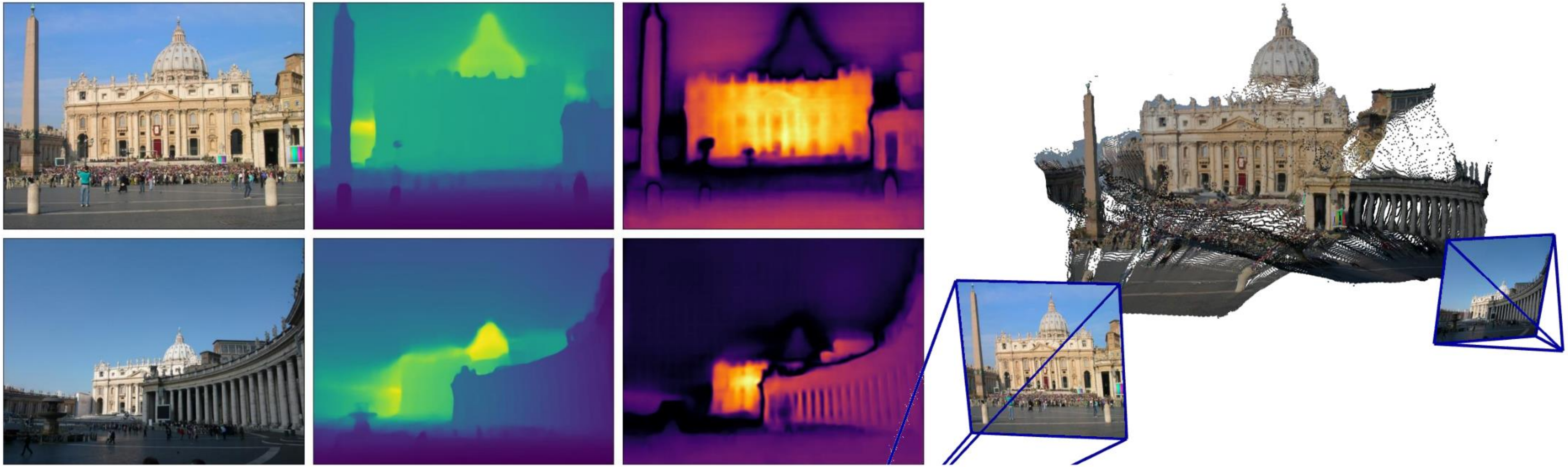
DUSt3R



Global Alignment

Multi-View Reconstruction Result

DUSt3R: Downstream Applications

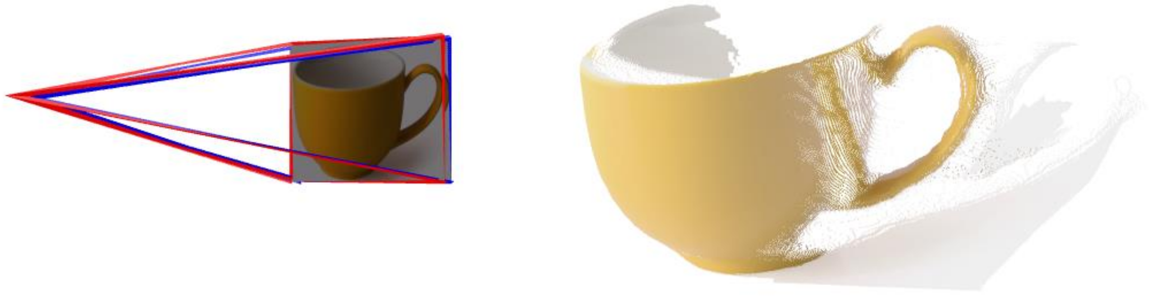


Input   Depth Map   Confidence Map   Reconstruction

Where do these depth maps come from?

Monocular Depth Estimation

- Feed same input image twice
- Depth = z coordinate of 3D point



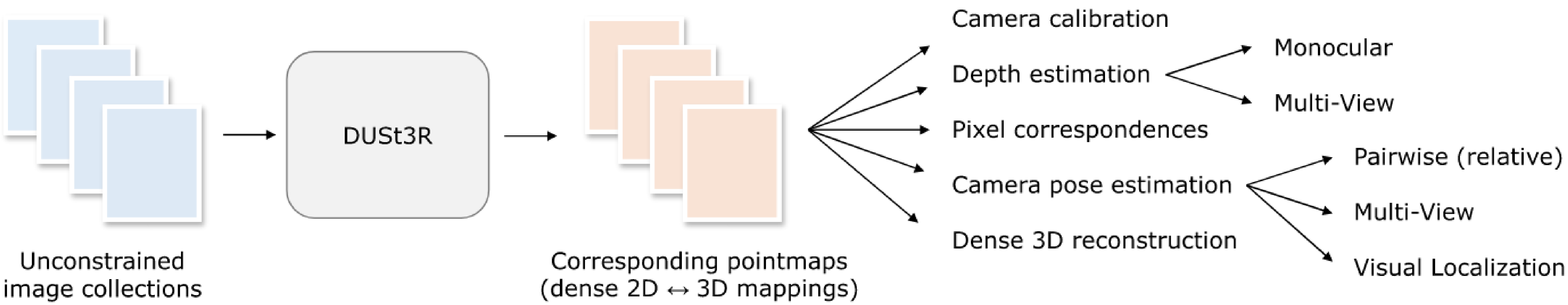
Monocular Depth Estimation

Methods	Train	Outdoor				Indoor					
		DDAD[41]		KITTI [35]		BONN [80]		NYUD-v2 [115]		TUM [119]	
		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[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

Towards 3D Foundation Models?

DUSt3R is trained for 2-view to 3D reconstruction task.

Pointmap is expressive representation which can be used for a variety of downstream tasks:



Pixel Correspondences from DUSt3R

- Image correspondence search now boils down to 3D correspondence search
- Can be solved e.g. by mutual nearest neighbor matching

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

Estimating Focal Length From Pointmaps

Assuming centered principal point ($i' = i - W/2, \text{ and } j' = j - H/2$)

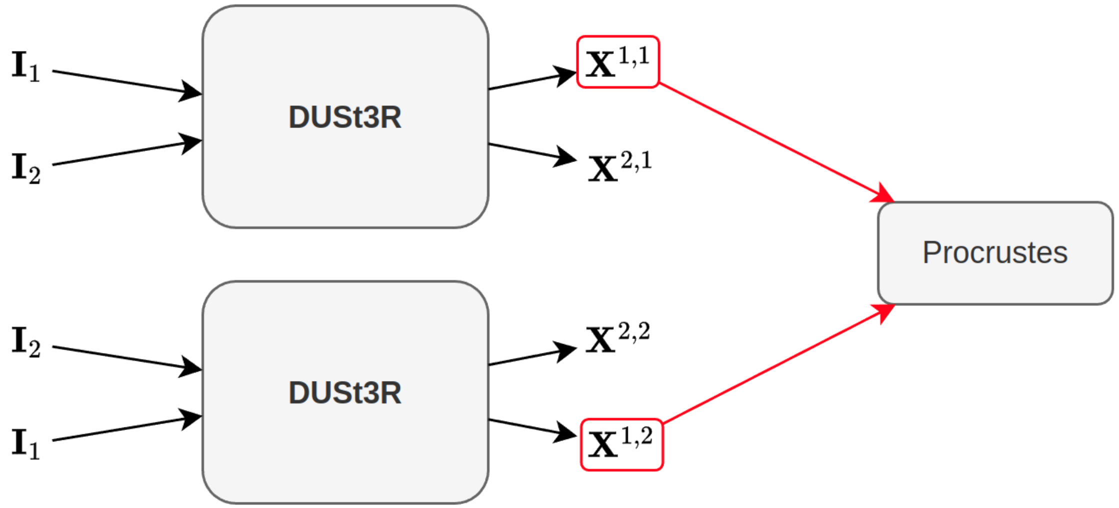
Then focal length can be estimated by minimizing confidence-aware reprojection loss

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

Can be solved by Weiszfeld-algorithm in a few iterations

Estimating Relative Camera Poses

1. Method (Procrustes):
   - Feed both ordered pairs
   - Compute optimal alignment via Procrustes
   - Derive relative camera poses



2. Method (PnP + RANSAC)
   - Procrustes not very robust
   - PnP possible because pointmap gives 2D-3D correspondences

MVS benchmark on DTU

		Methods	GT cams	Acc.↓	Comp.↓	Overall↓
Handcrafted	(a)	Camp [12]	✓	0.835	0.554	0.695
		Furu [33]	✓	0.613	0.941	0.777
		Tola [134]	✓	0.342	1.190	0.766
		Gipuma [34]	✓	**0.283**	0.873	0.578
Learning-based	(b)	MVSNet [161]	✓	0.396	0.527	0.462
		CVP-MVSNet [158]	✓	0.296	0.406	0.351
		UCS-Net [18]	✓	0.338	0.349	0.344
		CER-MVS [65]	✓	0.359	0.305	0.332
		CIDER [157]	✓	0.417	0.437	0.427
		CasMVSNet [40]	✓	0.325	0.385	0.355
		PatchmatchNet [139]	✓	0.427	0.277	0.352
		GeoMVSNet [180]	✓	0.331	**0.259**	**0.295**
		DUSt3R 512	×	2.677	0.805	1.741

All in mm

Acc = distance from reconstruction to closest ground truth point (averaged)

Comp = distance from ground truth to closest reconstruction point (averaged)

Overall = average of accuracy and completeness

Takeaways:

- Learning-based methods have overtaken handcrafted methods
- DUSt3R cannot compete for multiple reasons:
  1. Regression vs subpixel triangulation
  2. Does not leverage GT camera poses
  3. Zero-shot (other methods have trained on DTU train set)

DUSt3R Summary

- Very robust even to extreme view changes
- Simpler end-to-end learnable pipeline -> less prone to error accumulation
- Requires only 2 views
- For more views global alignment (GA) optimization procedure
  - Inefficient pairwise processing of O(N^2) pairs
  - Information sharing only between two images at a time
  - GA faster than BA but still not instant (couple of seconds to minutes)
  - Memory intensive (OOM on A100 with 80GB VRAM on 48 views)
- Cannot compete in 3D reconstruction accuracies
- Competitive in many other tasks such as depth, pose estimation

3R Models

DUST3R
Advanced Image-to-3D AI

Fast3R

Easi3R: Estimating Disentangled Motion from DUSt3R Without Training

Xingyu Chen[1] Yue Chen[1] Yuliang Xiu[1,2] Andreas Geiger[3] Anpei Chen[1,3]
[1]Westlake University [2]Max Planck Institute for Intelligent Systems [3]University of Tübingen, Tübingen AI Center

MAST3R
Advanced Image-to-3D AI

CUT3R

SLAM3R: Real-Time Dense Scene Reconstruction from Monocular RGB Videos

Yuzheng Liu[1*] Siyan Dong[2*†] Shuzhe Wang[3] Yingda Yin[1]
Yanchao Yang[2†] Qingnan Fan[4] Baoquan Chen[1†]
[1]Peking University [2]The University of Hong Kong [3]Aalto University [4]VIVO

MonST3R: A Simple Approach for Estimating Geometry in the Presence of Motion

Junyi Zhang[1] Charles Herrmann[2,+] Junhwa Hur[2] Varun Jampani[3] Trevor Darrell[1]
Forrester Cole[2] Deqing Sun[2,*] Ming-Hsuan Yang[2,4,*]
[1]UC Berkeley [2]Google DeepMind [3]Stability AI [4]UC Merced
(+: project lead, *: equal contribution)

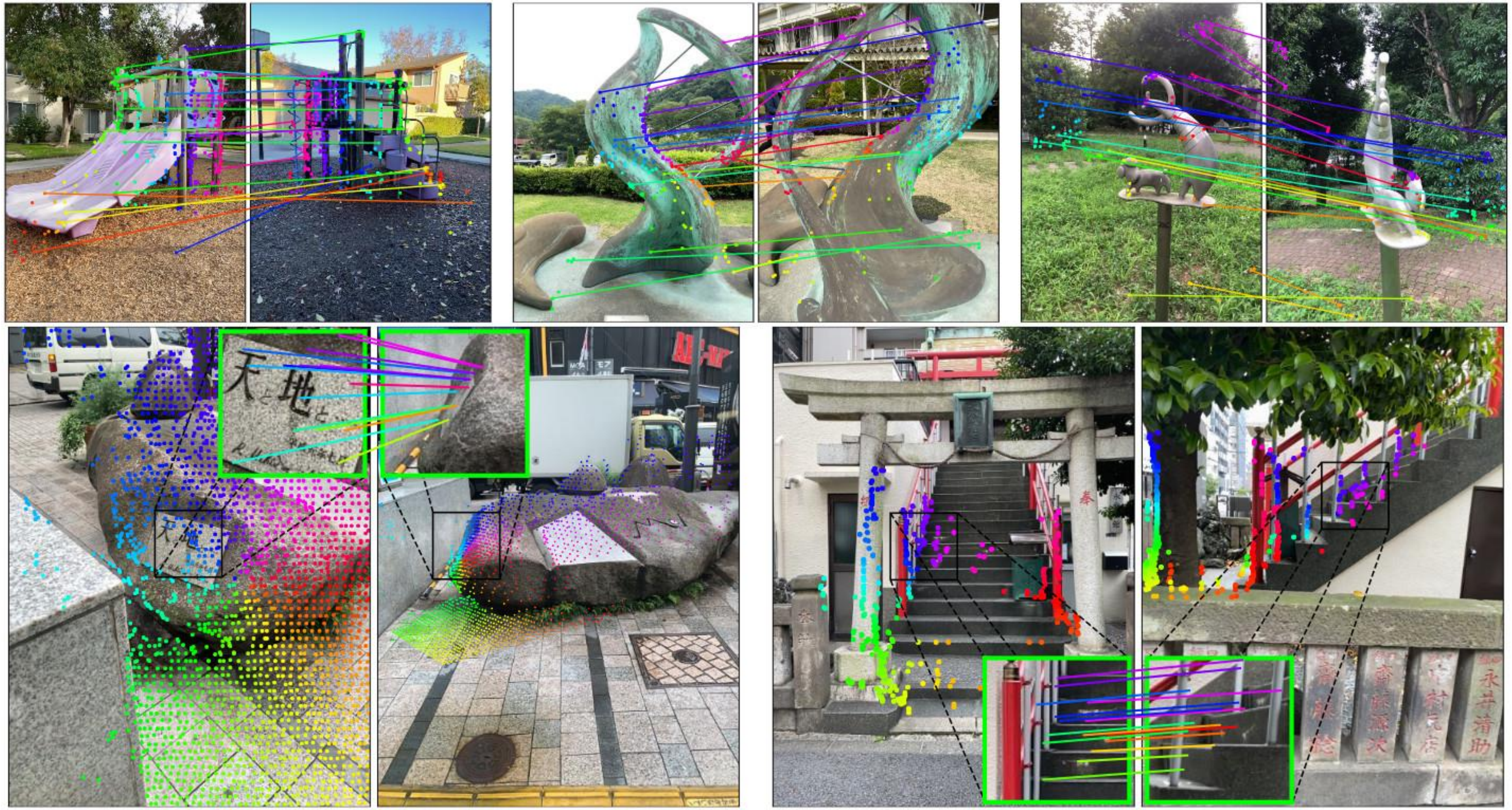
ICLR 2025 (Spotlight)

Spann3R
3D Reconstruction with Spatial Memory

Hengyi Wang, Lourdes Agapito
University College London

3DV 2025

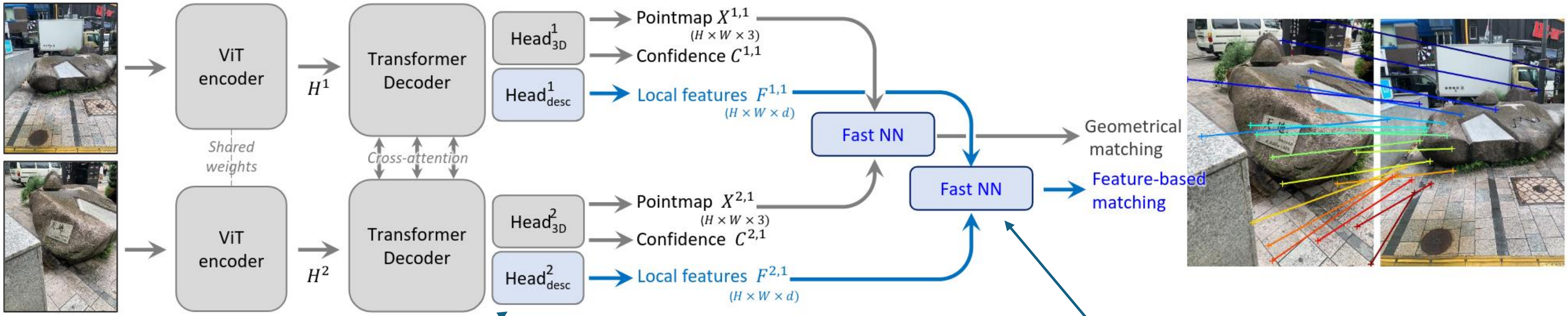
MASt3R: DUSt3R + Matching



MASt3R, Leroy et al. ECCV '24

MASt3R: Contributions

no longer scale-invariant, now metric



Trained using contrastive infoNCE loss

Iterative reciprocal nearest neighbor matching
-> O(kWH) instead of O(W^2H^2)

MASt3R, Leroy et al. ECCV '24

MVS Benchmark on DTU

		Methods	Acc.↓	Comp.↓	Overall↓
Handcrafted	(c)	Camp [13]	0.835	0.554	0.695
		Furu [31]	0.613	0.941	0.777
		Tola [90]	0.342	1.190	0.766
		Gipuma [32]	**0.283**	0.873	0.578
Learning-based	(d)	MVSNet [110]	0.396	0.527	0.462
		CVP-MVSNet [109]	0.296	0.406	0.351
		UCS-Net [17]	0.338	0.349	0.344
		CER-MVS [55]	0.359	0.305	0.332
		CIDER [107]	0.417	0.437	0.427
		PatchmatchNet [99]	0.427	0.277	0.352
		GeoMVSNet [119]	0.331	**0.259**	**0.295**
	(e)	DUSt3R [102]	2.677	0.805	1.741
		MASt3R	0.403	0.344	0.374

All in mm

MVS with MASt3R:

1. Forward passes to obtain 2D-2D correspondences
2. Triangulate matches in ground truth frame using gt camera parameters

No costly global alignment necessary!

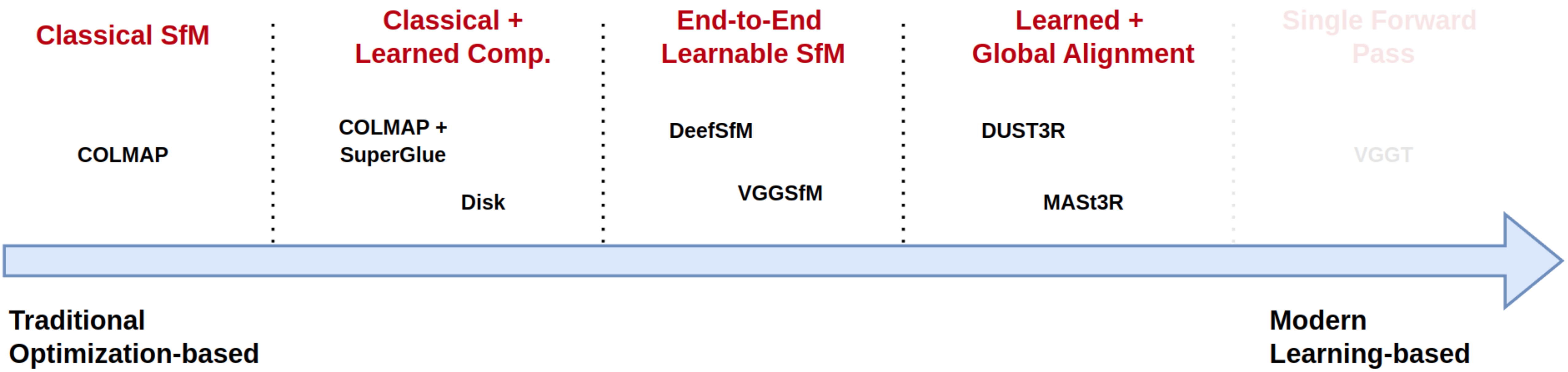
Takeaways:

- Triangulation outperforms regression
- MASt3R outperforms DUSt3R and is competitive with recent learning-based methods while:
  1. not using camera poses for matching
  2. not having seen DTU camera setup during training

MASt3R Summary

- Improved DUSt3R
- Regresses **metric** pointmaps
- Additional feature head for matching
- Fast reciprocal nearest neighbor matching procedure
- Retains robustness of DUSt3R and strengths of pixel matching
- Outperforms DUSt3R on many downstream tasks
- Still only pairwise images. For multiple images, global alignment of pointmaps still required -> memory intensive

A Spectrum of Methods



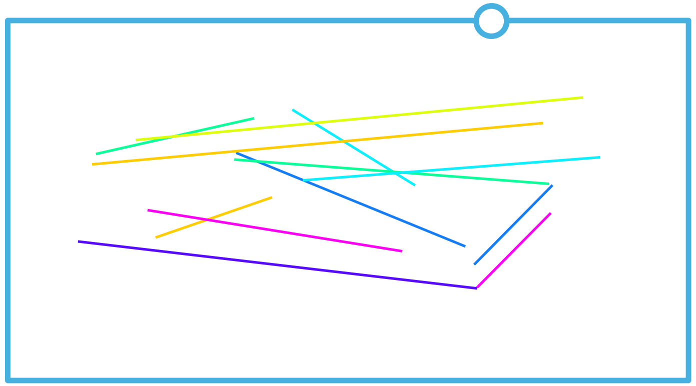
- Epipolar Geometry
- Pose Estimation
- Triangulation
- Bundle Adjustment
- ...

- Feature Extraction / Matching
- Monocular Depth Prediction
- Cross-View Attention
- ...

Efficiently Dealing with More Views

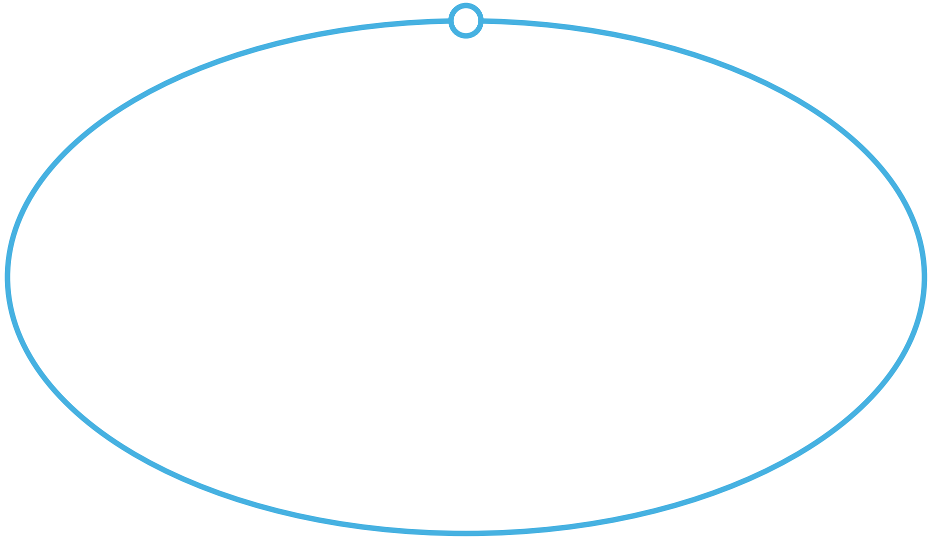
Multi-View Alignment via Optimization: Bottleneck for 3D

COLMAP



Bundle Adjustment

DUSt3R

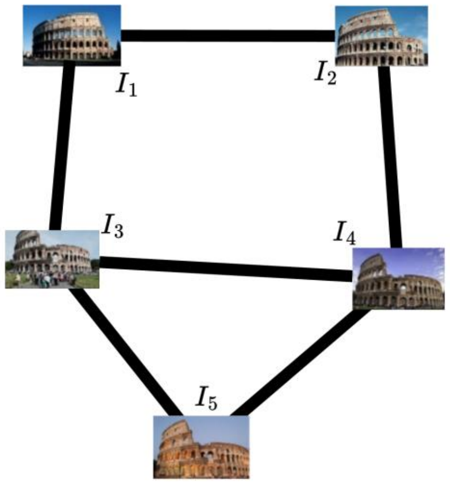


Global Alignment

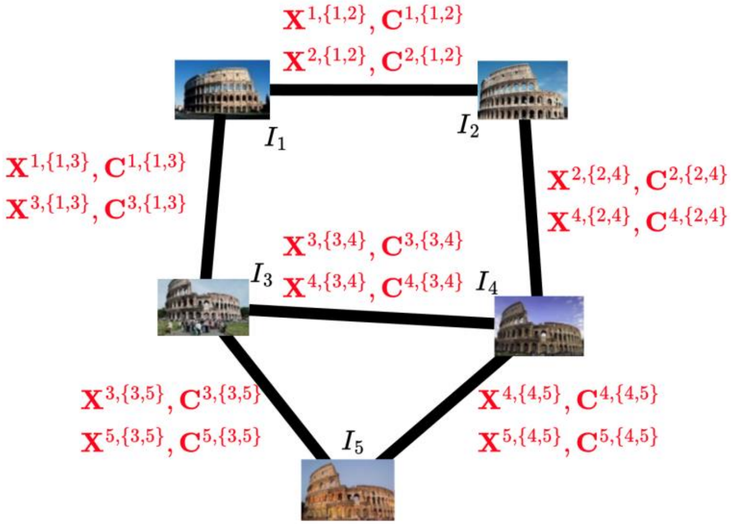
Multi-View Efficiency Problem of DUSt3R/MASt3R

- DUSt3R and MASt3R are 2-view models
- For multi-view, O(N^2) pointmaps need to be aligned with costly global alignment procedure -> infeasible for larger N

1. Scene Graph



2 . Pairwise Reconstruction



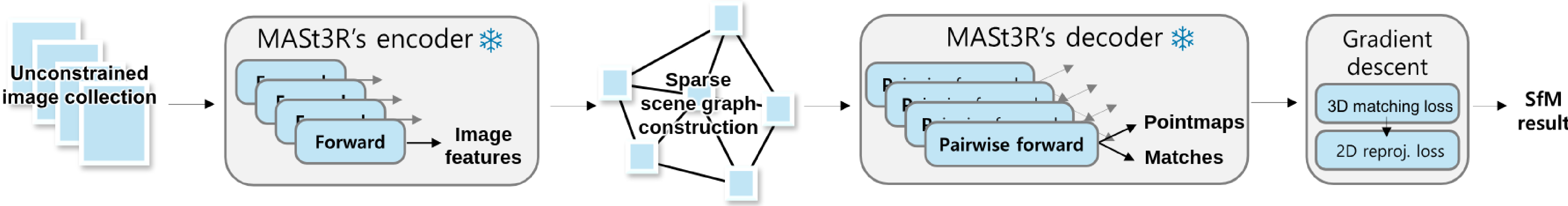
3. Global Optimization

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

MASt3R-SfM: Sparsification

Overview:

- Reduced number of pairwise forward passes via **sparse** scene graph with O(N) edges
- Coarse alignment: minimize 3D loss only for matching points
- Refinement: minimize 2D reprojection loss (BA)



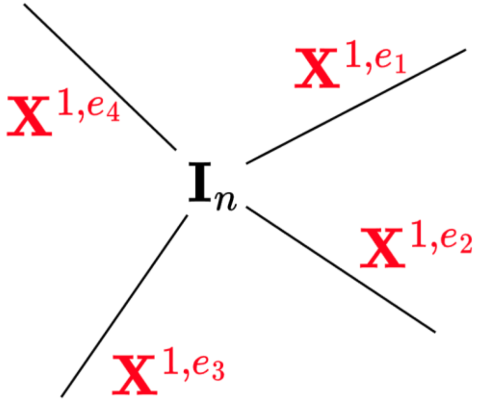
MASt3R-SfM, Duisterhof et al. arXiv '24

Coarse Refinement

- Canonicalize pointmaps

$$\tilde{X}^n_{i,j} = \frac{\sum_{e \in \mathcal{E}^n} C^{n,e}_{i,j} X^{n,e}_{i,j}}{\sum_{e \in \mathcal{E}^n} C^{n,e}_{i,j}}$$

- Estimate intrinsics $\mathbf{K}_n$ via focal length
- Find optimal rigid transforms and scales



$$\min_{\sigma, \mathbf{P}} \sum_{(n,m) \in \mathcal{E}} \sum_{\boxed{i \in \mathcal{M}^{n,m}}} C_i \| \chi^n_i - \chi^m_i \|^{1.5} \quad \text{where} \quad \chi^n_i = \frac{1}{\sigma_n} P_n^{-1} K_n^{-1} Z^n_i \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

Only applies to pixel correspondences

Multi-View Alignment via Optimization: Bottleneck for 3D

MASt3R-SfM **still** has some optimization for global alignment

In general, optimization is often the bottleneck for 3D Vision:

- Time-consuming
- Poor Compatibility with Deep Learning
  - Not inherently "plug-and-play"
  - Often non-differentiable
- Complexity
  - Scary for non-experts

Can we do without optimization, in one single forward pass?

Let’s Reconstruct in One Go!

Images

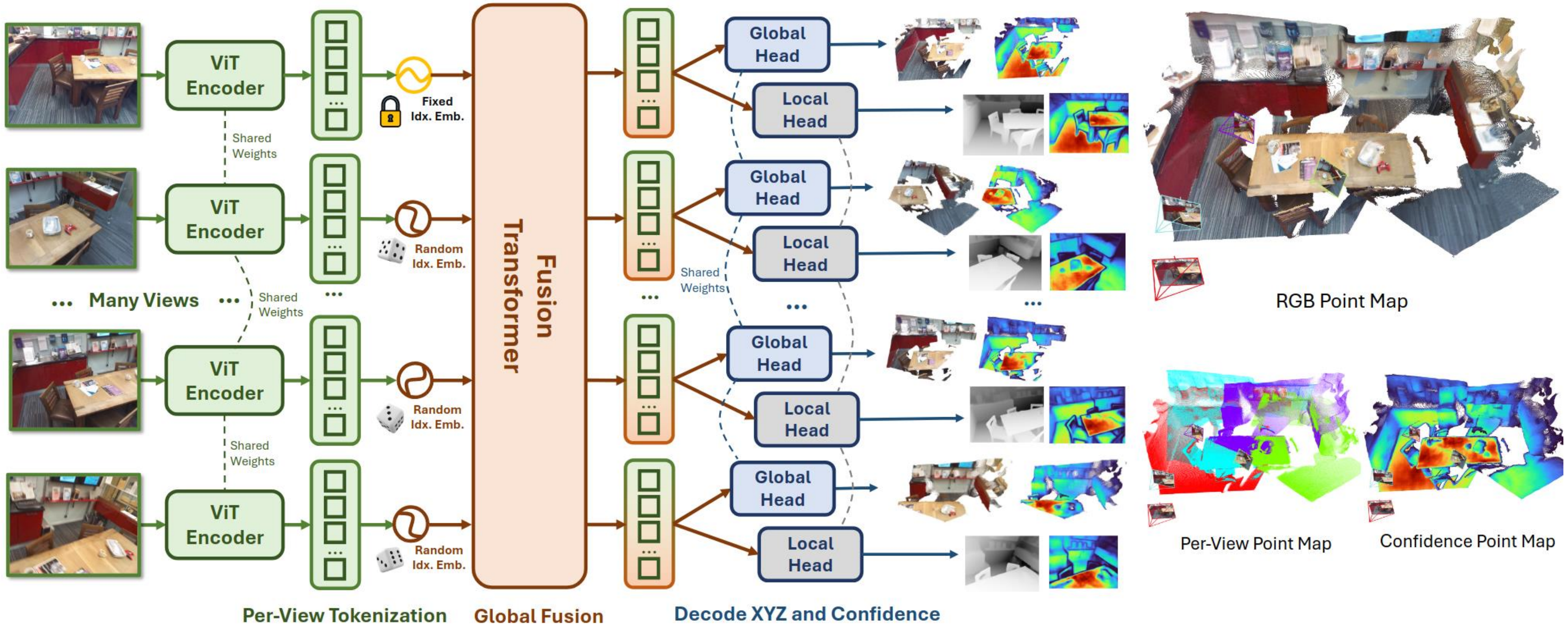
Neural Network

Reconstruction

Cameras, Depths, Points, and Correspondences

DUST3R Multi-View Extensions

- No longer two branches but fusion transformer which can handle arbitrary number of views
- All images can attend to each other
- No global alignment necessary



Fast3R, Yang et al. CVPR '25
MV-DUSt3R, Tang et al. CVPR '25

DUSt3R vs. Fast3R

Speed & Memory

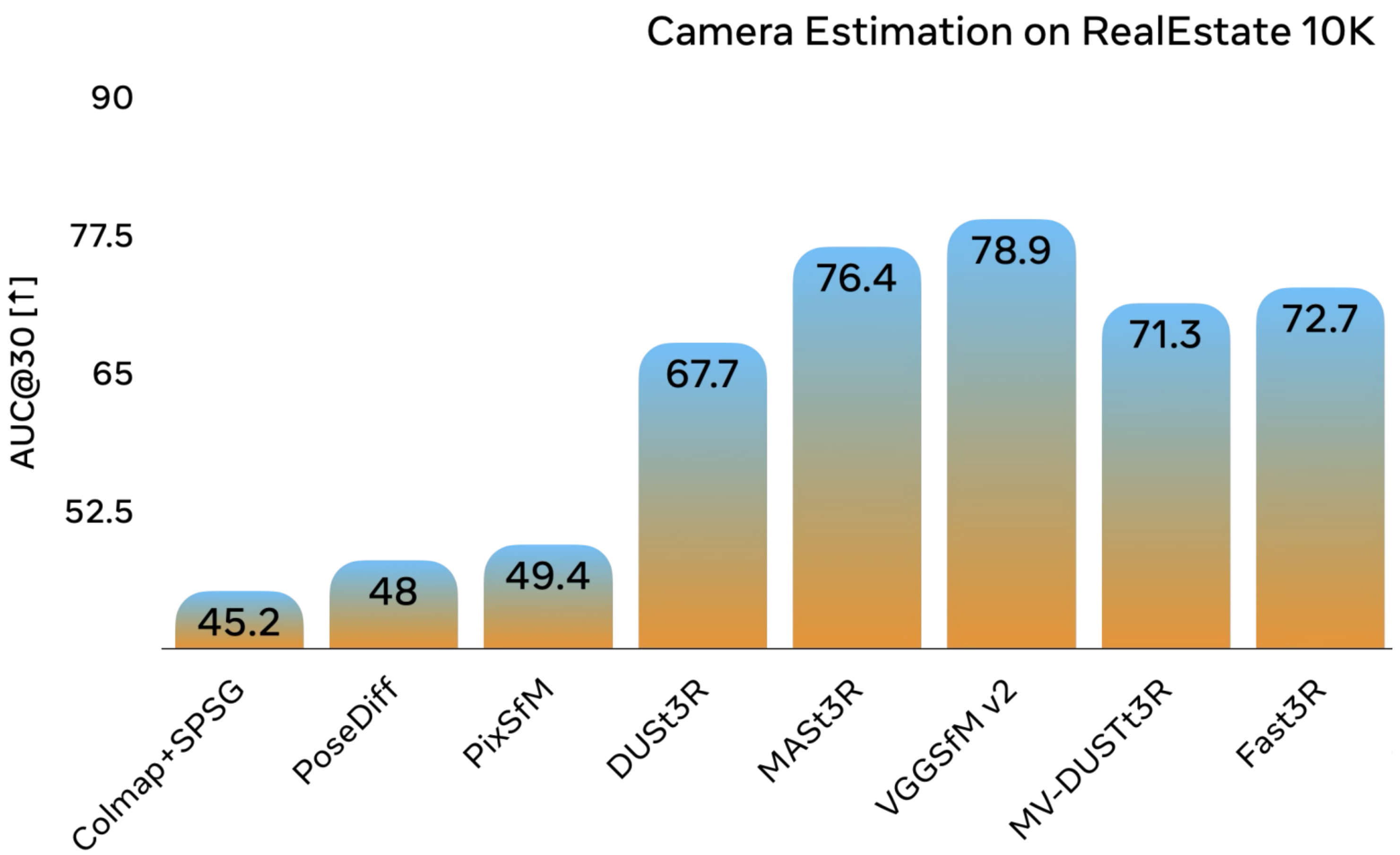
Comparison of computational efficiency between Fast3R and DUSt3R on a single A100 GPU. Each view has a 512×384 resolution.

# Views	Fast3R		DUSt3R	
	Time (s)	Peak GPU Mem (GiB)	Time (s)	Peak GPU Mem (GiB)
2	0.065	3.84	0.092	3.52
8	0.122	6.33	8.386	24.59
32	0.509	13.25	129.0	67.61
48	0.84	20.8	OOM	OOM
320	15.938	41.90	OOM	OOM
800	89.569	55.97	OOM	OOM
1000	137.62	63.01	OOM	OOM
1500	308.85	78.59	OOM	OOM

Note: "OOM" indicates Out of Memory. For DUSt3R, at 48 views the N^2 pairwise reconstructions consume all VRAM during global alignment.

- Much faster
- More memory efficient
- Information sharing between all views instead of pairwise

Fast3R vs MASt3R



Worse than MASt3R!

VGGT: Overparameterized Reconstruction in One GO

Images

Neural Network

Reconstruction

Cameras, Depths, Points, and Correspondences

Visual Geometry Grounded Transformer, Wang et al. CVPR '25 Best Paper Award

🏛️ VGGT: Visual Geometry Grounded Transformer

🌟 [GitHub Repository] | 🚀 [Project Page]

Upload a video or a set of images to create a 3D reconstruction of a scene or object. VGGT takes these images and generates all key 3D attributes, including extrinsic and intrinsic camera parameters, point maps, depth maps, and 3D point tracks.

Getting Started:

1. **Upload Your Data:** Use the "Upload Video" or "Upload Images" buttons on the left to provide your input. Videos will be automatically split into individual frames (one frame per second).
2. **Preview:** Your uploaded images will appear in the gallery on the left.
3. **Reconstruct:** Click the "Reconstruct" button to start the 3D reconstruction process.
4. **Visualize:** The 3D reconstruction will appear in the viewer on the right. You can rotate, pan, and zoom to explore the model, and download the GLB file. Note the visualization of 3D points may be slow for a large number of input images.
5. **Adjust Visualization (Optional):** After reconstruction, you can fine-tune the visualization using the options below (**click to expand**):

Please note: Our model itself usually only needs less than 1 second to reconstruct a scene. However, visualizing 3D points may take tens of seconds due to third-party rendering, which are independent of VGGT's processing time. Please be patient or, for faster visualization, use a local machine to run our demo from our GitHub repository.

Upload Video

Drop Video Here
- or -
Click to Upload

3D Reconstruction (Point Cloud and Camera Poses)

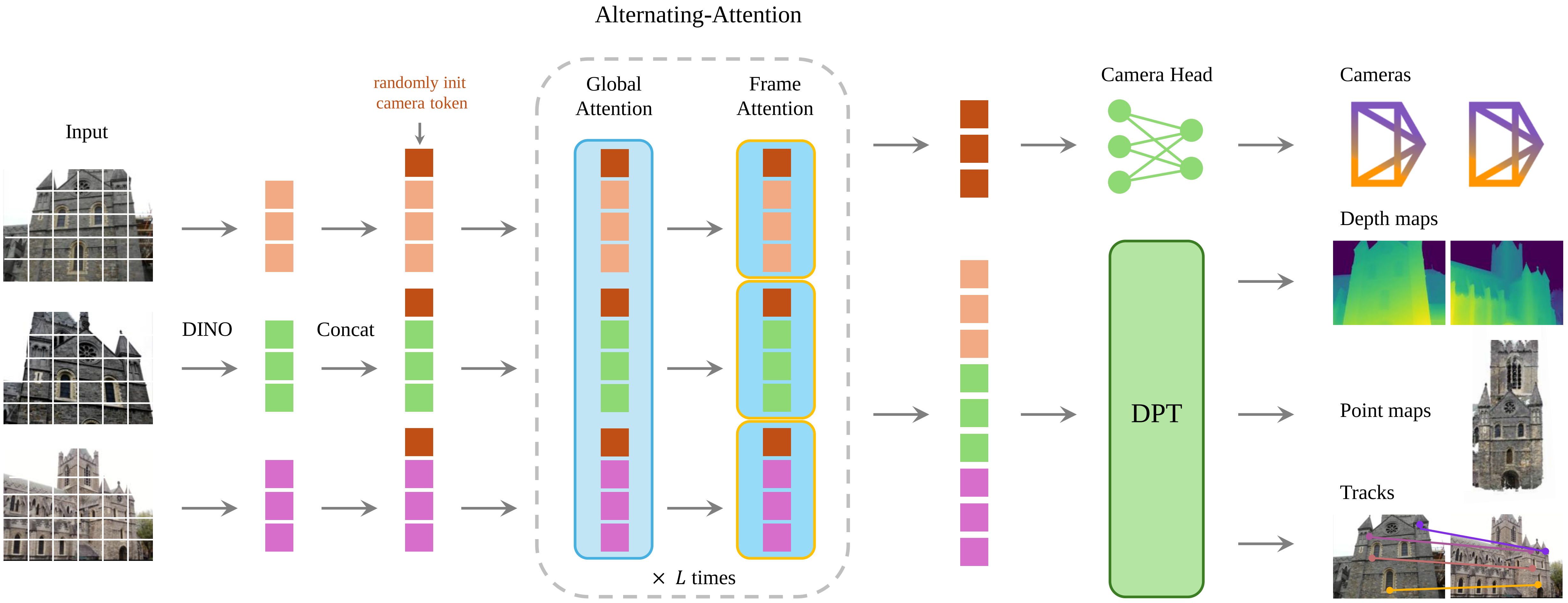
Please upload a video or images, then click Reconstruct.

3D Model

Upload Images

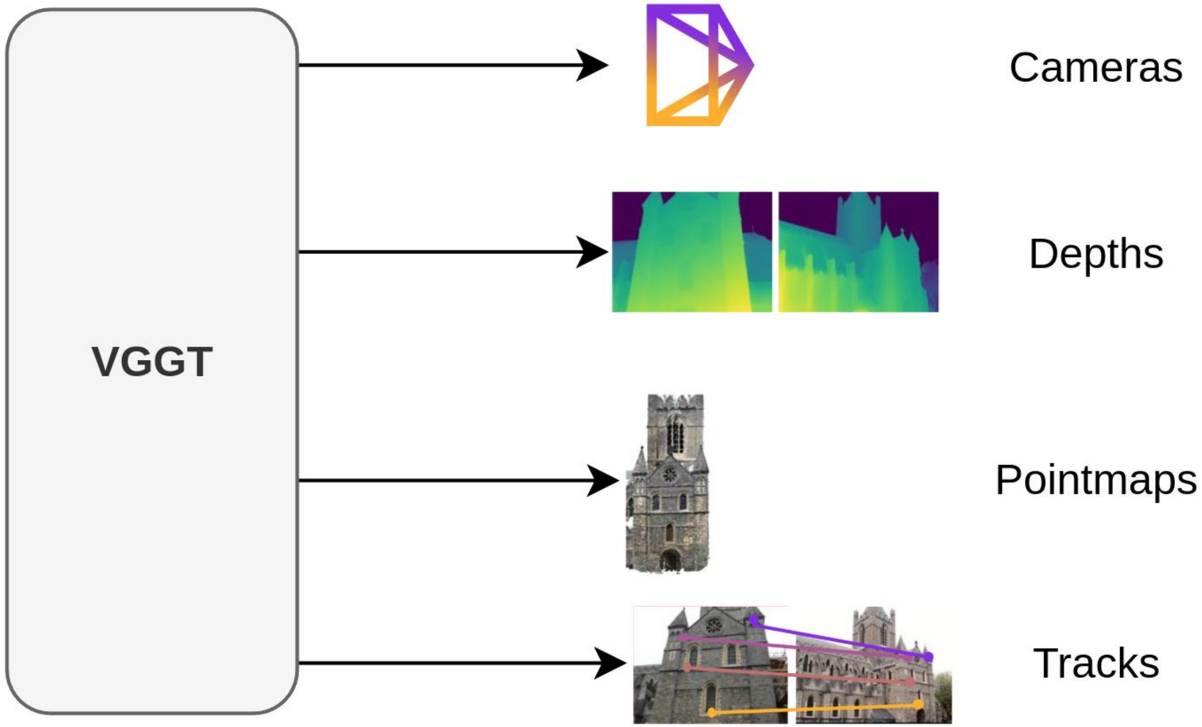
Drop File Here
- or -
Click to Upload

VGG Transformer



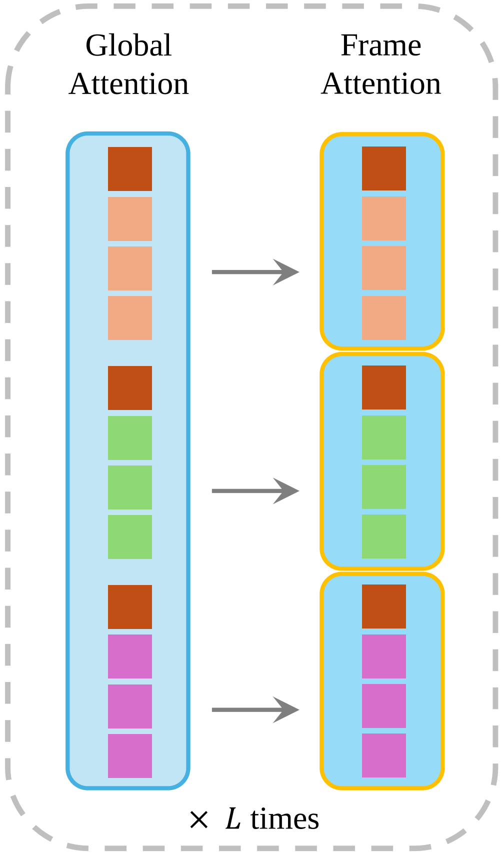
Why Overparameterized Output?

- DUST3R: Extract depthmap, cameras, and matches from pointmap
- VGGT: Predict all of them "independently"
- Overparameterized predictions brings substantial performance gains during training
- During inference, combining estimates often outperforms direct branch

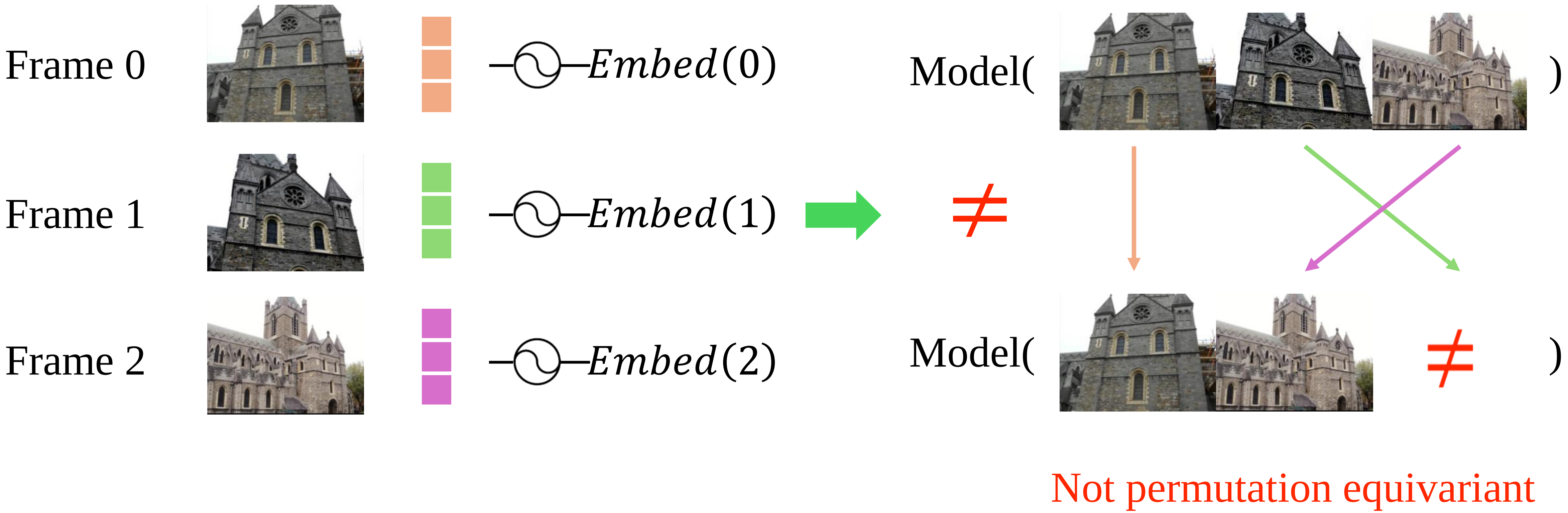


Why Alternating-Attention?

- Global Attention
  - Ensures scene-level coherence
- Frame-wise Attention
  - Eliminates **frame index embedding**
    - For permutation equivariance
    - For flexible input length



Why Alternating-Attention?



Why Alternating-Attention?

Frame 0  $Embed(0)$

Frame 1  $Embed(1)$

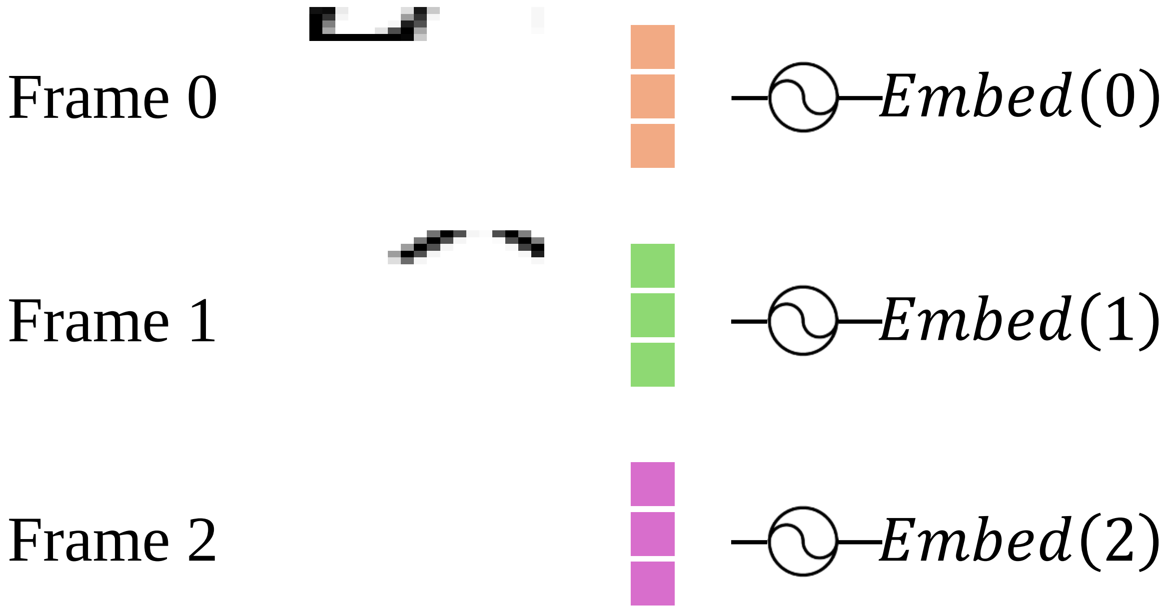
Frame 2  $Embed(2)$

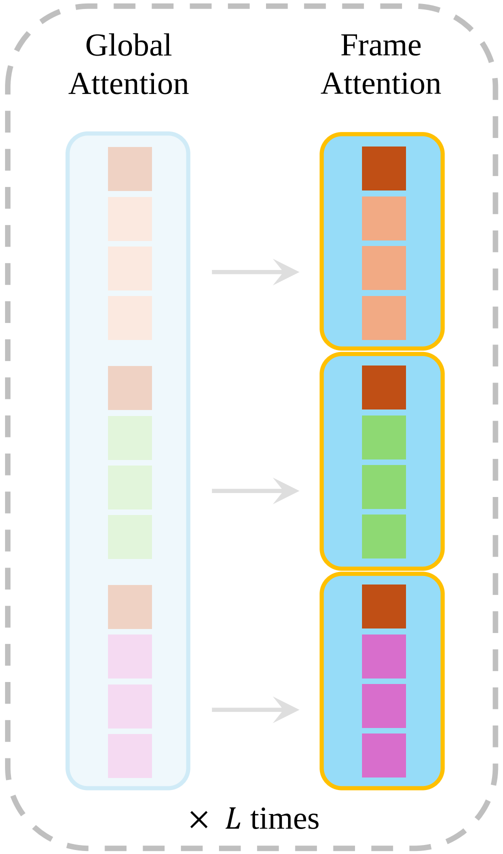
⋮

Frame 842

But model never sees $Embed(842)$ during training

Why Alternating-Attention?





Replaces frame index embedding by Frame-wise Attention

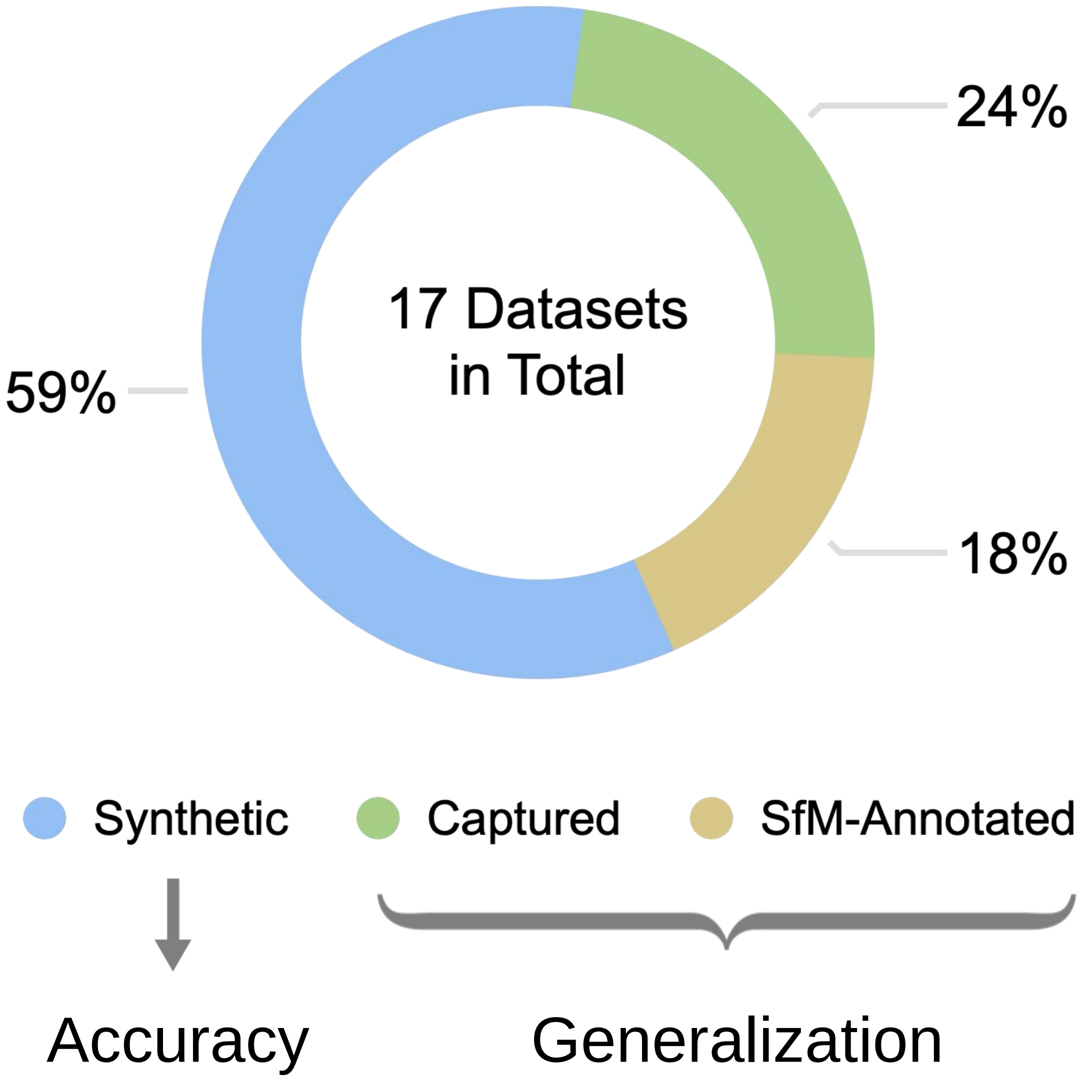
Training and Data

Training:

2 to 24 frames

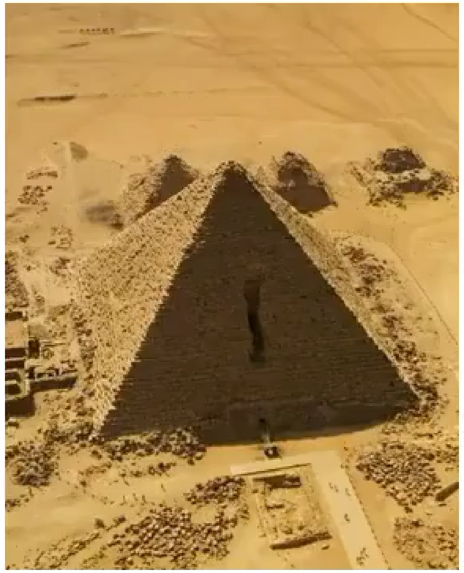
Inference:

1 to 300+ frames

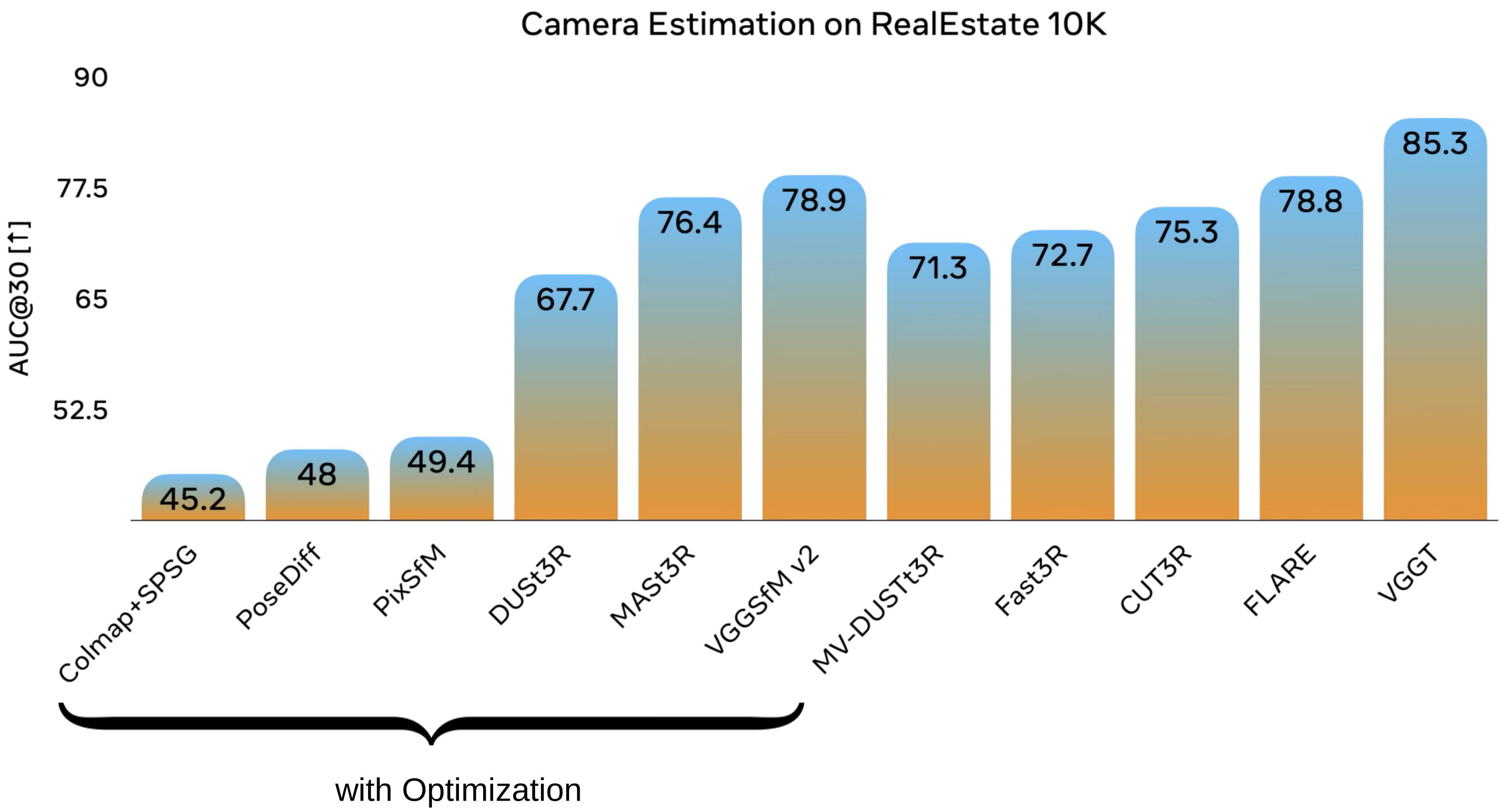


Qualitative

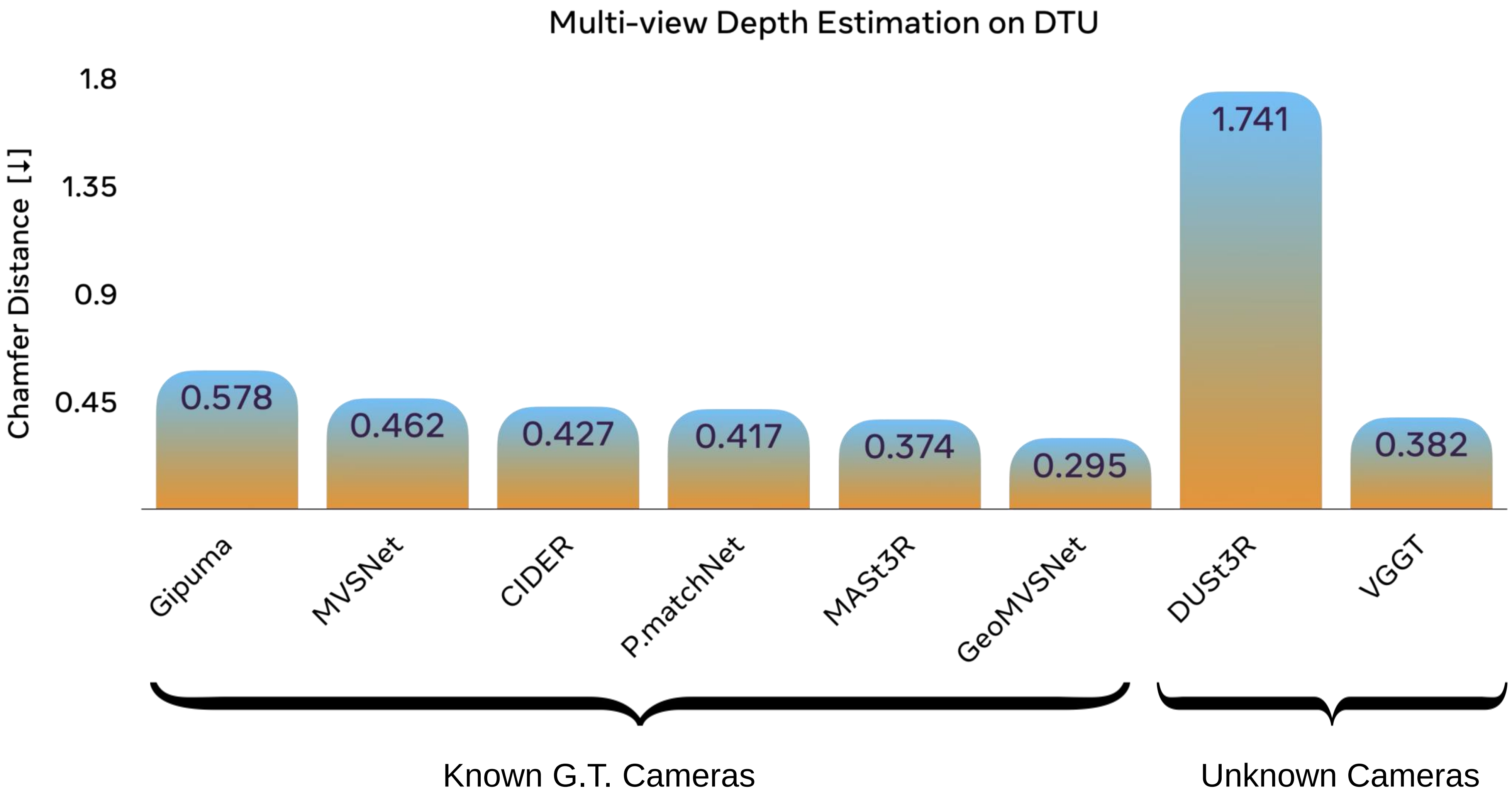
32 Views



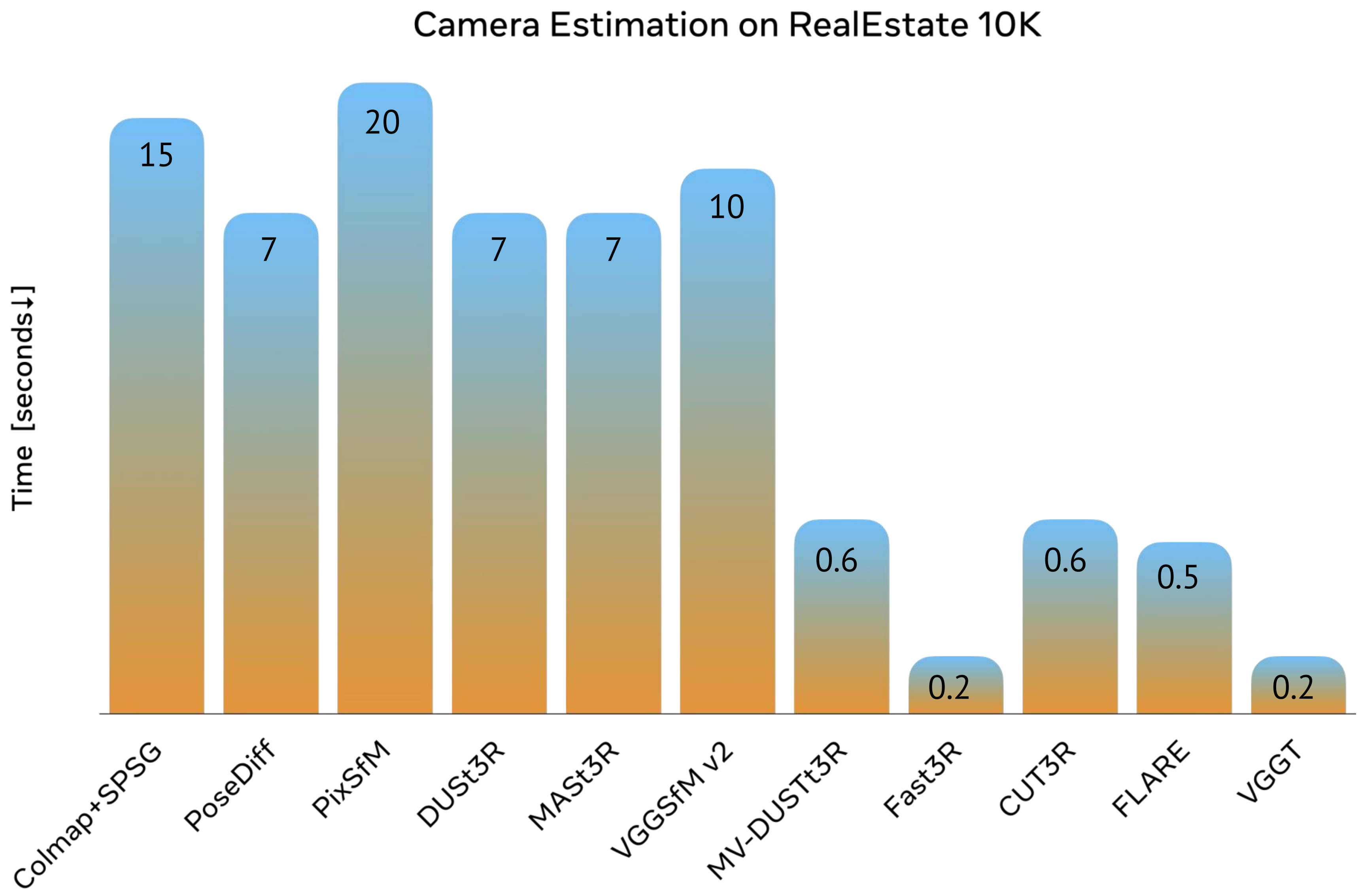
VGGT Is Accurate



VGGT Is Accurate

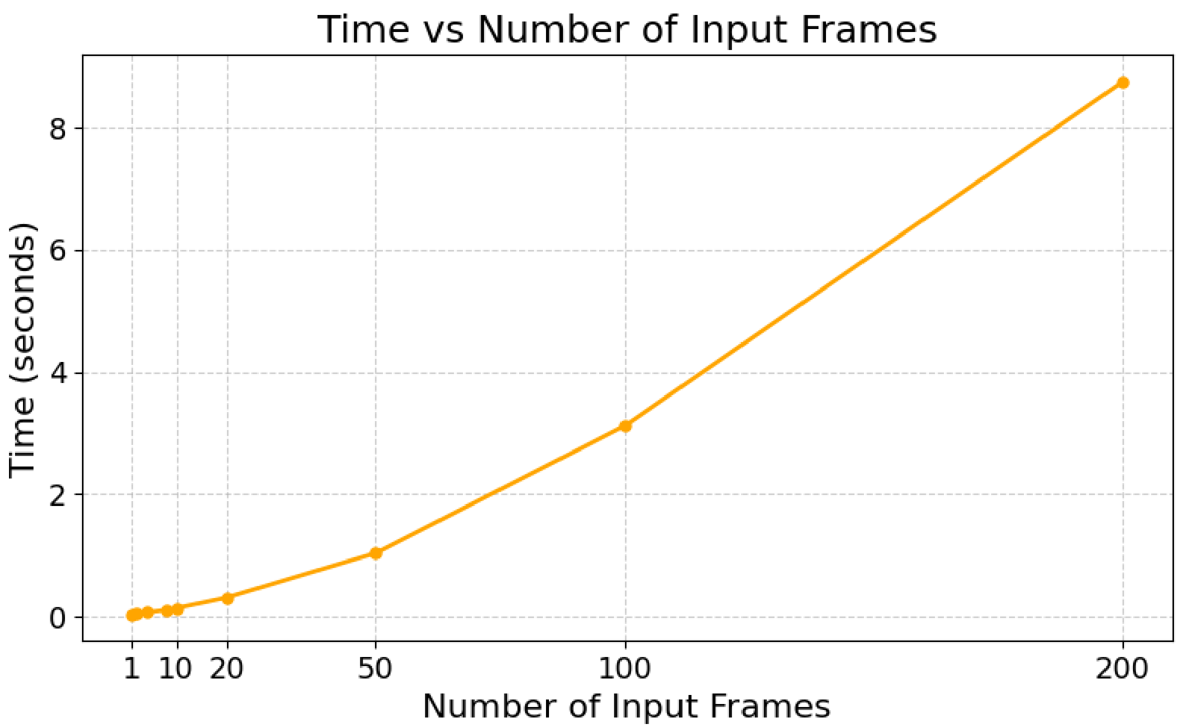


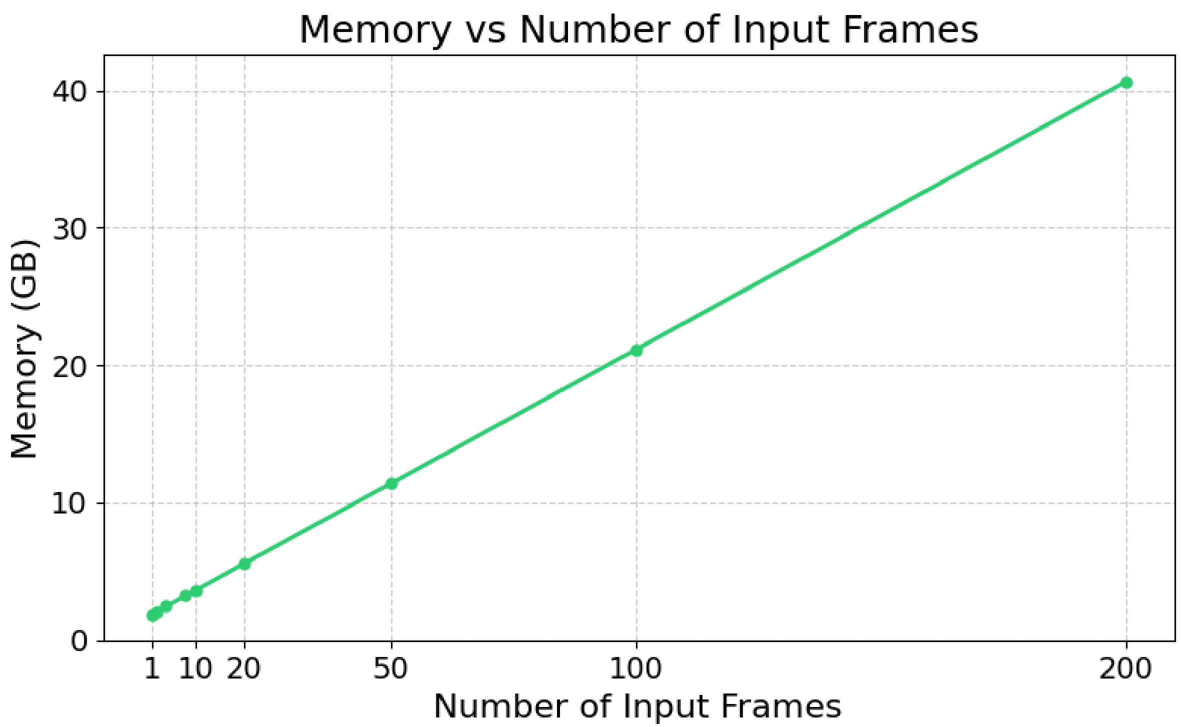
VGGT Is Fast



Camera Estimation on RealEstate 10K
Time [seconds↓]
15
7
20
7
7
10
0.6
0.2
0.6
0.5
0.2
Colmap+SPSG
PoseDiff
PixSfM
DUSt3R
MASt3R
VGGSfM v2
MV-DUSTt3R
Fast3R
CUT3R
FLARE
VGGT

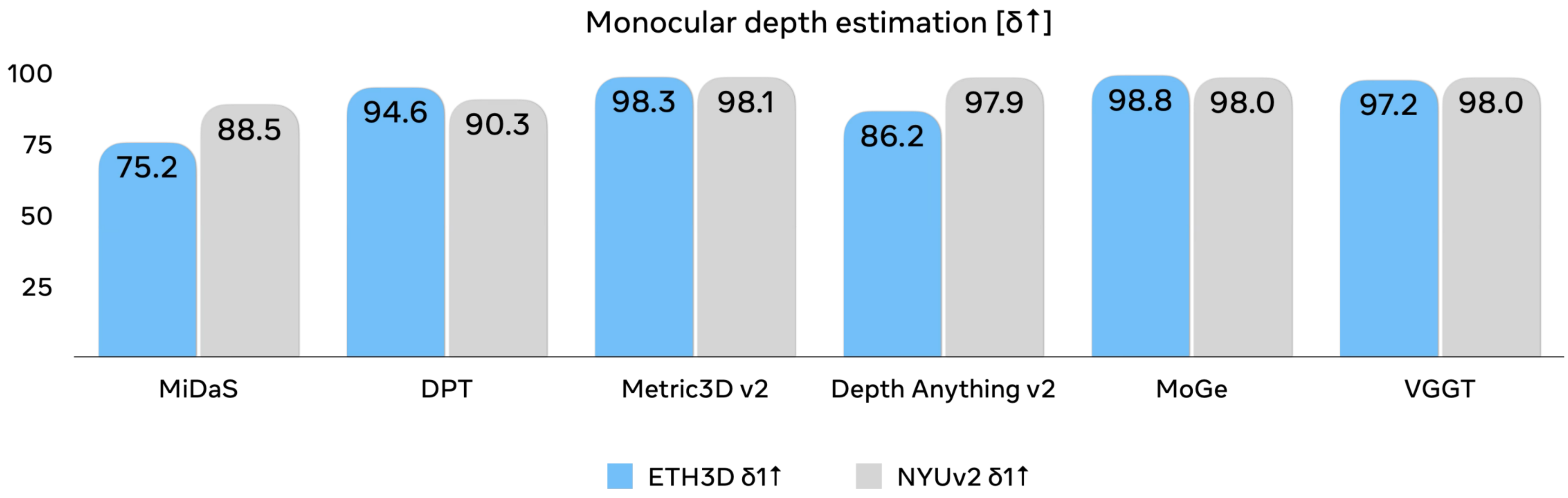
Runtime and Memory





- Memory usage scales roughly linearly with input frames
- The time usage is around $O(N^{1.5})$

Zero-shot Monocular Depth Estimation



As good as SoTA experts – but VGGT was never trained for monocular

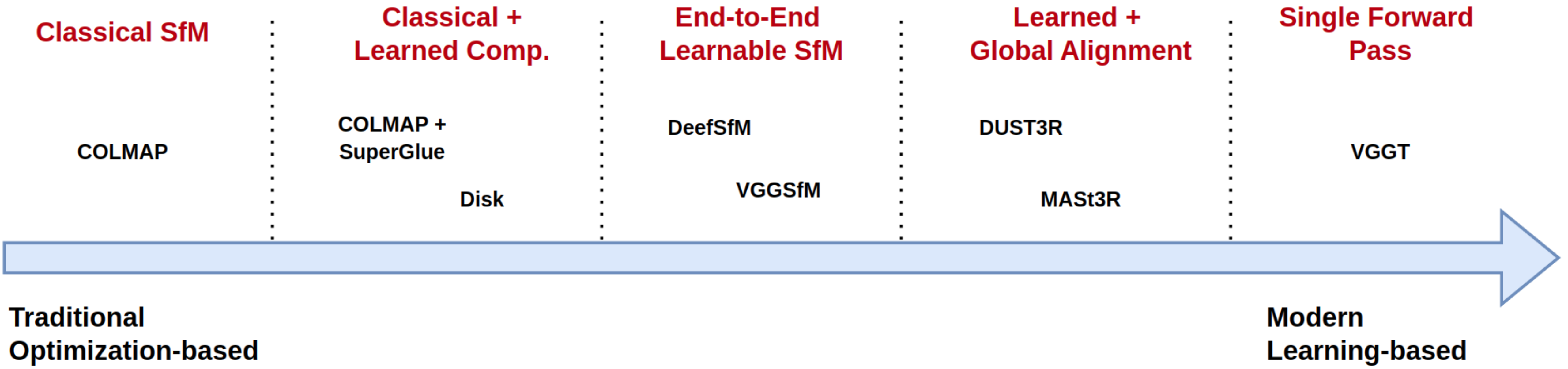
Zero-shot Monocular Depth Estimation

Single View

VGGT Is General, Seamless and Practical

General	Seamless	Practical
• Diverse images • Single to hundreds of views	• Just a neural network • Standard components	• Fast and accurate • Addresses all core 3D tasks

A Spectrum of Methods



- Epipolar Geometry
- Pose Estimation
- Triangulation
- Bundle Adjustment
- ...

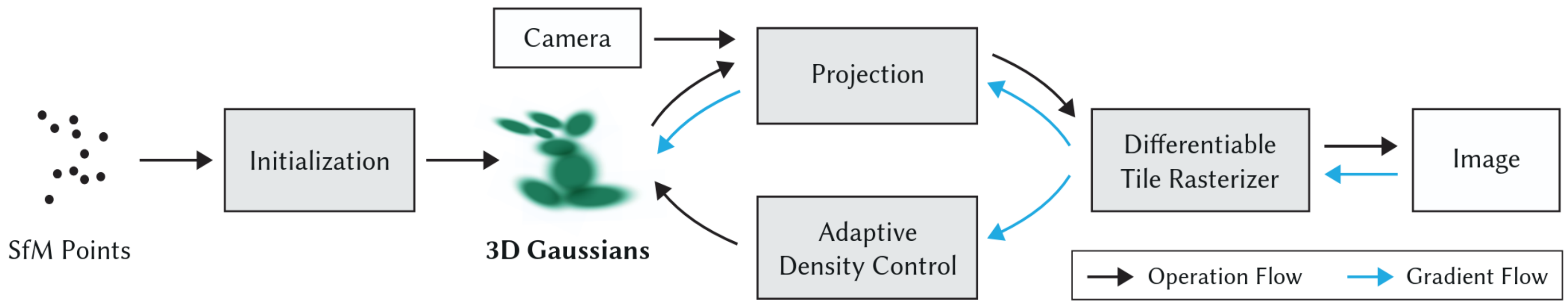
- Feature Extraction / Matching
- Monocular Depth Prediction
- Cross-View Attention
- ...

Novel View Synthesis from Sparse Images


Sora – Santorini – 3Views

Original 3DGS:

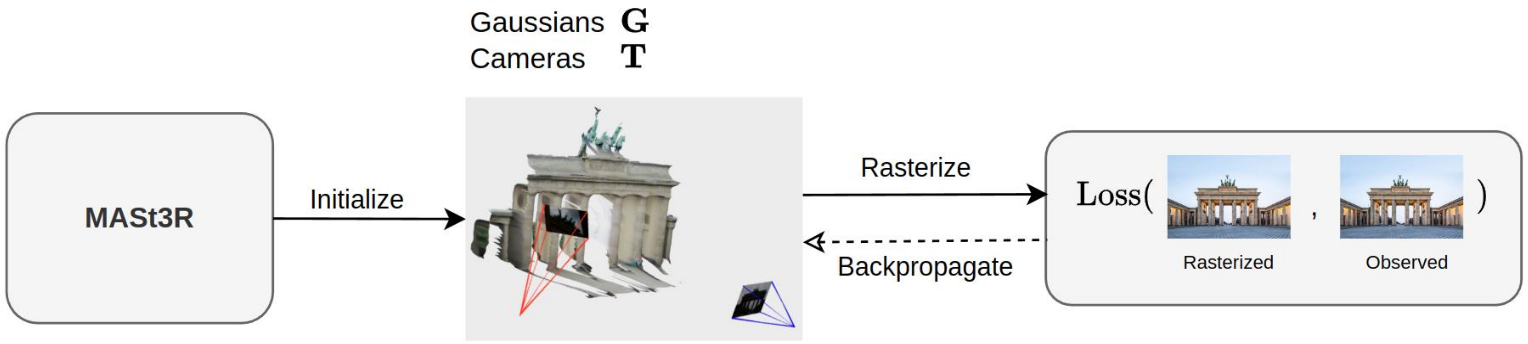
- Assumes known camera parameters
- Requires lots of views
- Sparse initialization from SfM
- Adaptive density control necessary for dense reconstruction



3D Gaussian Splatting for Real-Time Radiance Field Rendering, Kerbl et al. '23

InstantSplat:

- Unknown cameras
- Few views
- Dense initialization
- Joint pose and Gaussian optimization instead of adaptive density control



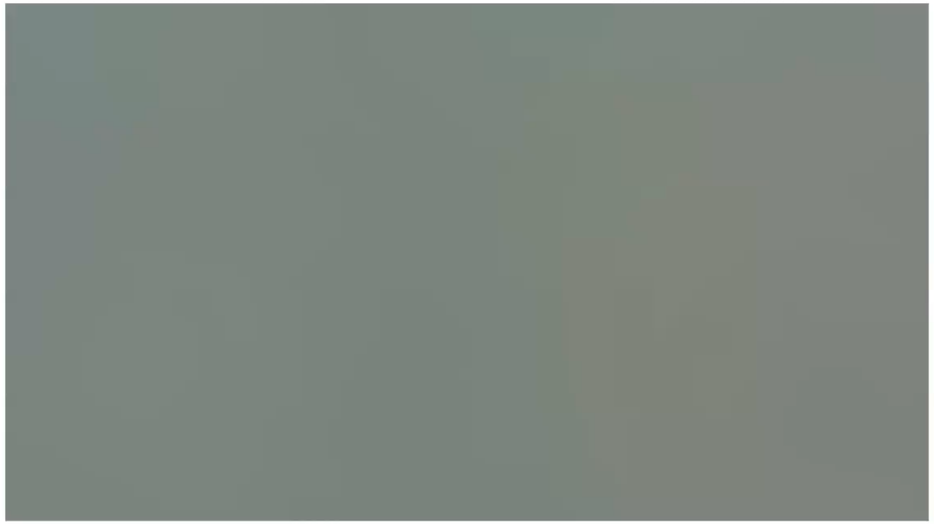
InstantSplat: Unbounded Sparse-view Pose-free Gaussian Splatting in 40 Seconds, Fan et al. '24

Comparison to Other Pose-Free Models

00:00

Minute Second

- NeRF without camera poses
- 50+ images

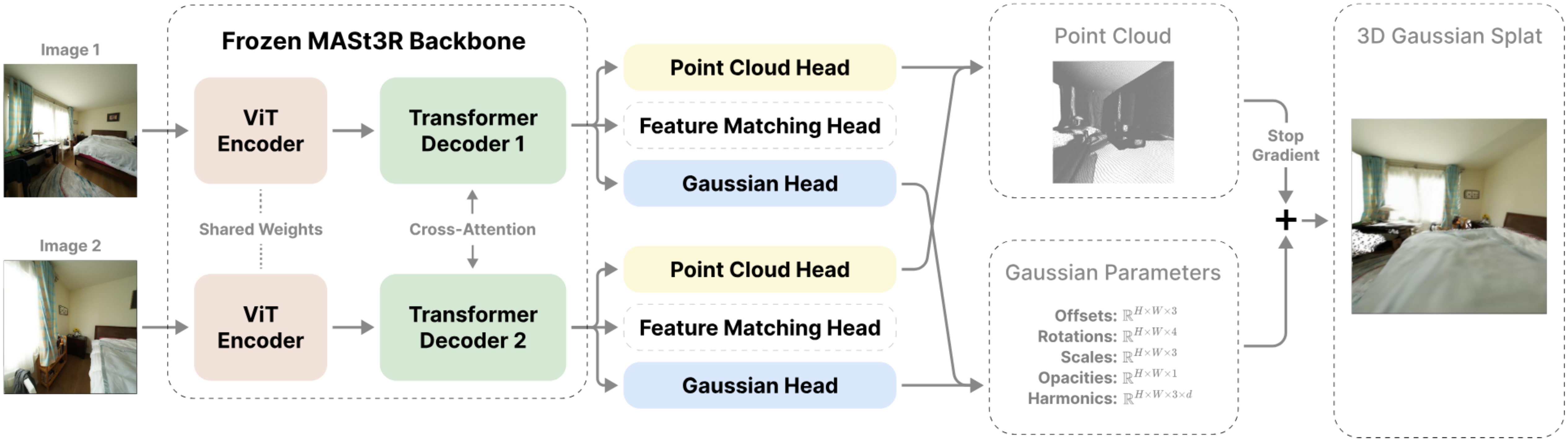


NoPe-NeRF

Ours (Dense Surface Point Initialization)

Splatt3R

Optimization-free method: Additional head predicting Gaussian parameters



Uncalibrated,
Input Image Pair

Inference

3D Gaussian Splat

Novel Renderings

Summary

- Traditional 3D reconstruction pipelines are overtaken by learning-based methods
- DUSt3R shifted paradigm towards direct pointmap regression from unposed images
- Very robust and versatile. Outperforms task-specific methods in classical 3D tasks
- Sparked a wave of 3R-methods for all kinds of applications
- VGGT: optimization-free feed-forward network outperforming state of the art

Further Resources and Slide Credit

We have not covered t3R models for videos / dynamic scenes:

- Spann3R
- MONSt3R
- DAS3R
- CUT3R
- Easi3R

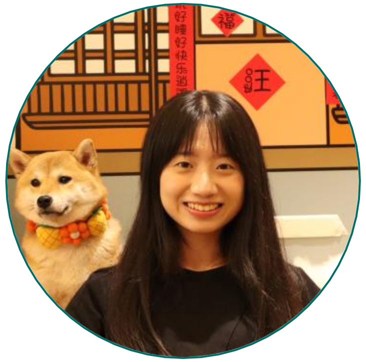
https://github.com/ruili3/awesome-dust3r for a list of DUSt3R-related works

Some slides were copied / adapted from the following sources:

- Vincent Leroy, "From CroCo to MASt3R: A Paradigm Change in 3D Vision"
- Jianyuan Wang, "VGGT" CVPR presentation

Feat2GS

Probing Visual Foundation Models with Gaussian Splatting


Yue Chen[1]


Xingyu Chen[1]


Anpei Chen[1,3]


Gerard Pons-Moll[3,4]


Yuliang Xiu[1,2]

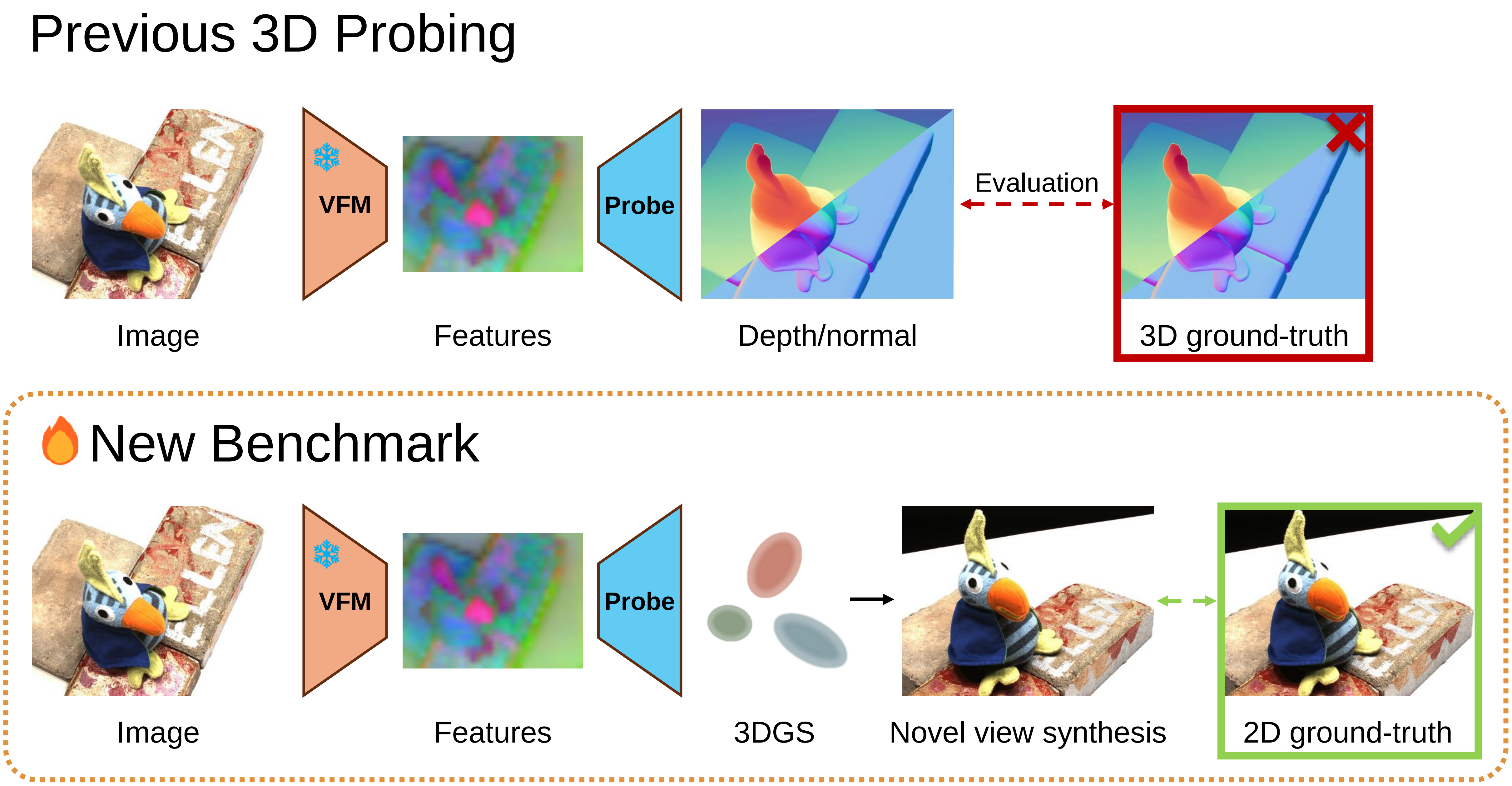
[1]Westlake University
[2]Max Planck Institute for Intelligent Systems
[3]University of Tübingen, Tübingen AI Center
[4]Max Planck Institute for Informatics

How well do they understand the **3D world**?

Visual Foundation Models

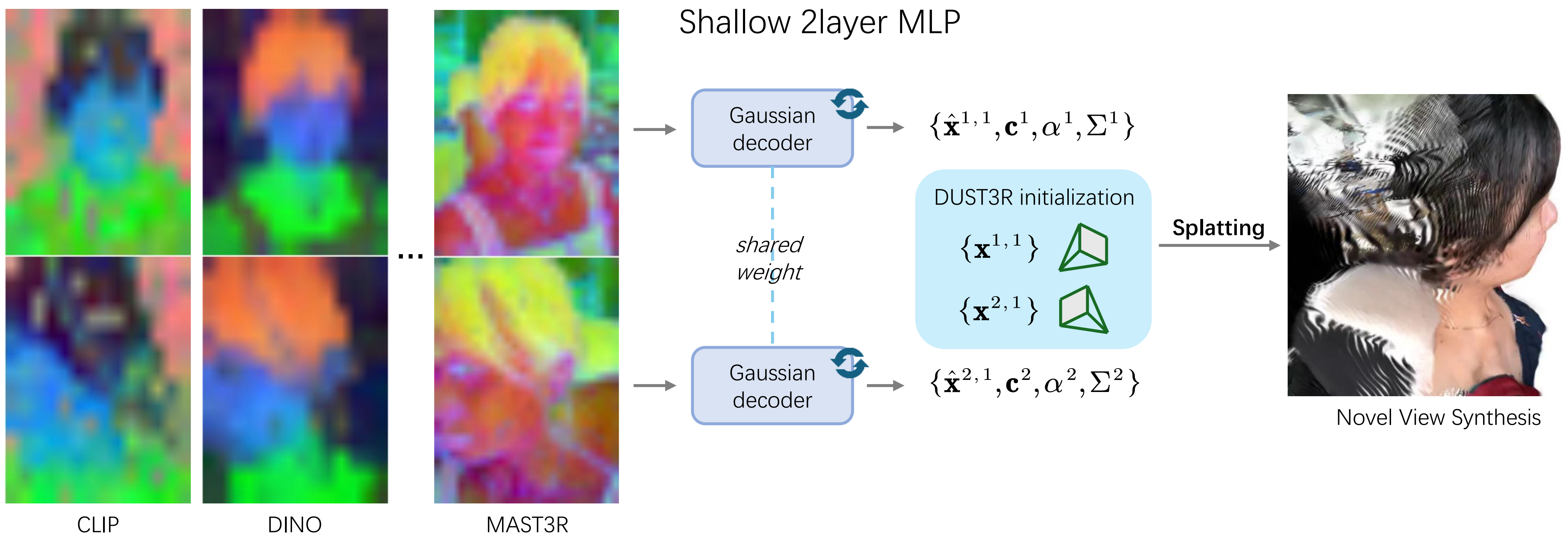
VFM	Arch.	Channel	Supervision	Dataset
DUSt3R [94]	ViT-L/16	1024	Point Regression	3D DUSt3R-Mix
MASt3R [49]	ViT-L/16	1024	Point Regression	3D MASt3R-Mix
MiDaS [70]	ViT-L/16	1024	Depth Regression	3D MiDaS-Mix
DINOv2 [64]	ViT-B/14	768	Self Distillation	2D LVD-142M
DINO [9]	ViT-B/16	768	Self Distillation	2D ImageNet-1k
SAM [44]	ViT-B/16	768	Segmentation	2D SA-1B
CLIP [69]	ViT-B/16	512	Contrastive VLM	2D WIT-400M
RADIO [72]	ViT-H/16	1280	Multi-teacher Distillation	2D DataComp-1B
MAE [33]	ViT-B/16	768	Image Reconstruction	2D ImageNet-1k
SD [75]	UNet	1280	Denoising VLM	2D LAION

We need 3D probing.

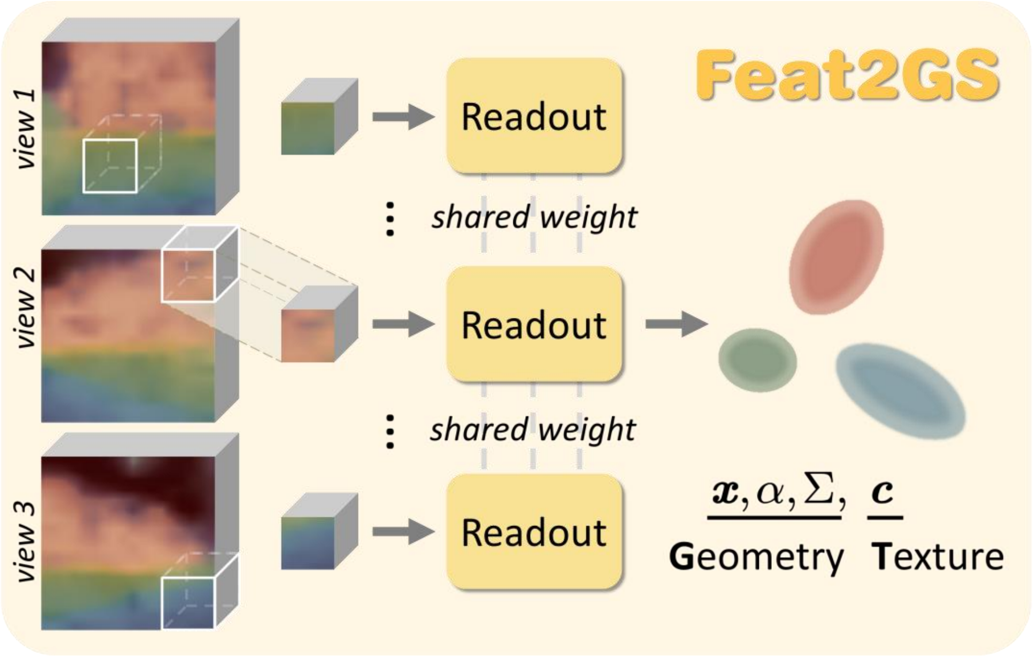


Previous 3D Probing
VFM
Probe
Evaluation
Image
Features
Depth/normal
3D ground-truth
New Benchmark
VFM
Probe
Image
Features
3DGS
Novel view synthesis
2D ground-truth

Feat2GS as Probe



Probing Geometry and Texture separately



Splatting

Novel View Synthesis

Probing Schemesc

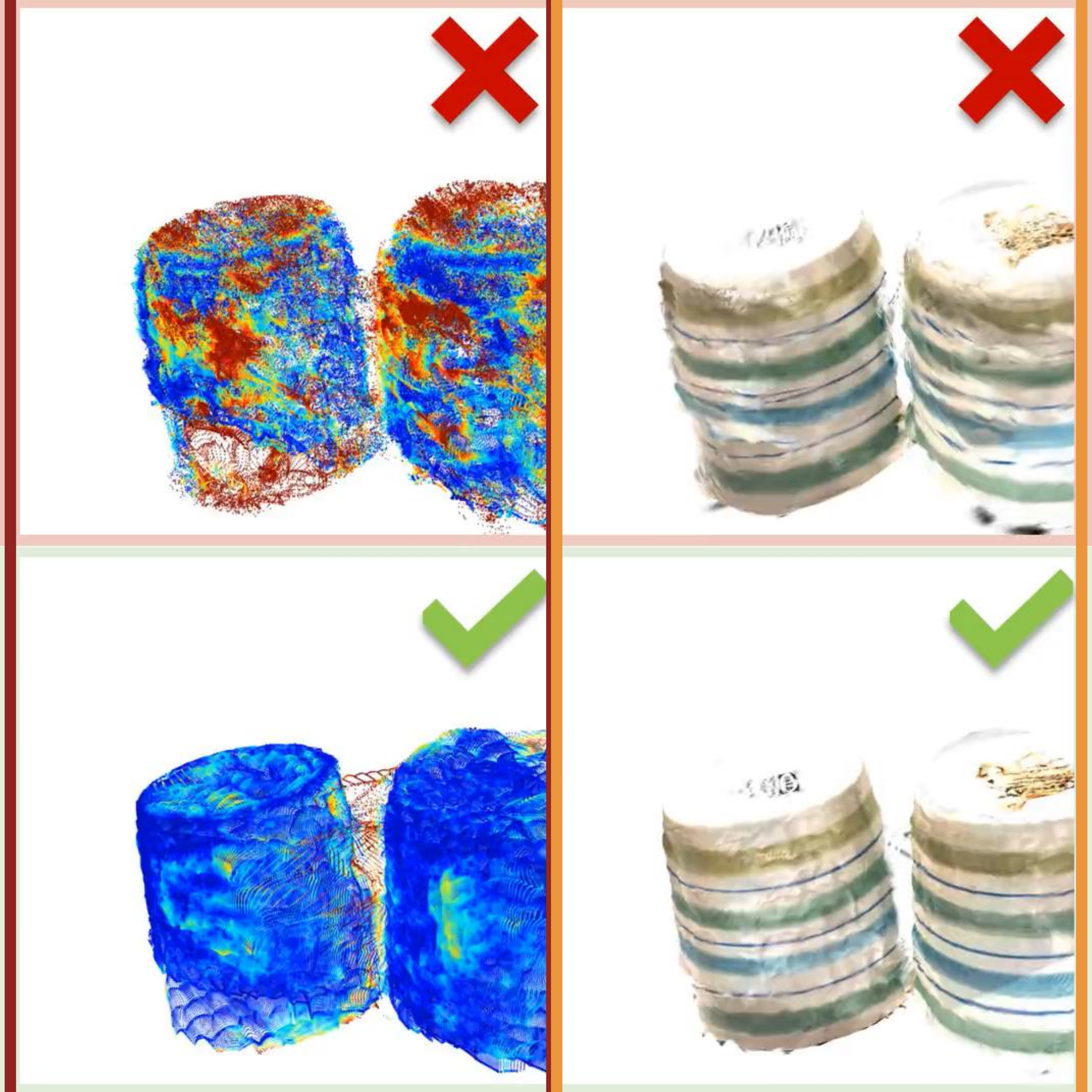
Probing	-Geometry	-Texture	-All
Feature-Readout	$\boldsymbol{x}, \alpha, \Sigma$	$\boldsymbol{c}$	$\boldsymbol{x}, \alpha, \Sigma, \ \boldsymbol{c}$
Free-Optimize	$\boldsymbol{c}$	$\boldsymbol{x}, \alpha, \Sigma$	/

Our Findings: 3D Metrics and 2D Metrics are well-aligned.

DINOv2

vs.

DUSt3R



Pointcloud Error Map | Novel View Synthesis

Accuracy↓: -2.448
Completeness ↓: -.277
Distance↓: -6.163

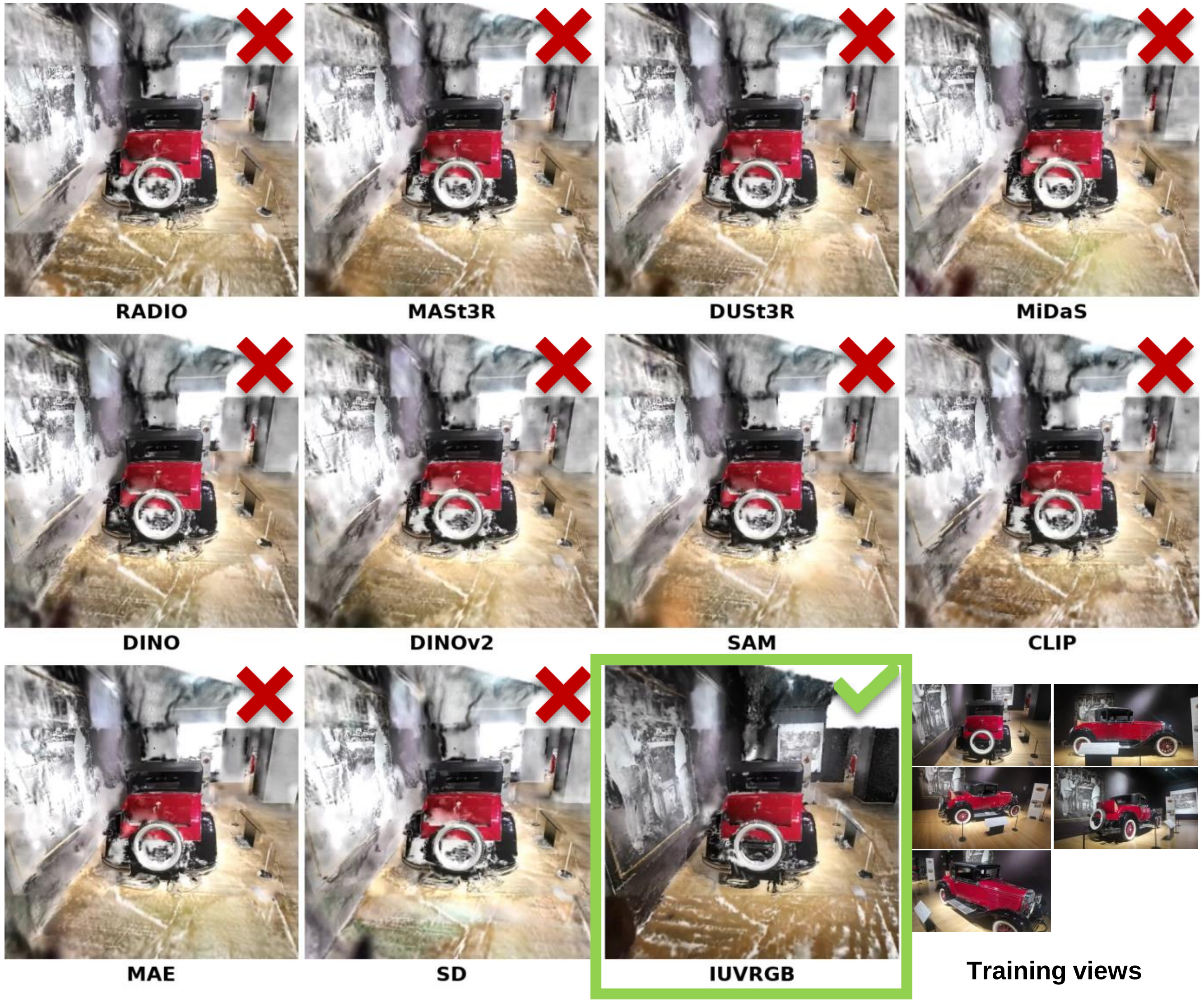
PSNR↑: +1.02
SSIM↑: +.0233
LPIPS↓: -.0301

Geometry Probing

RADIO

MASt3R

Texture Probing



Findings:

Foundation models
capture geometry well,
but struggle with texture.

Application

Input images

Novel View Synthesis / Normal Results