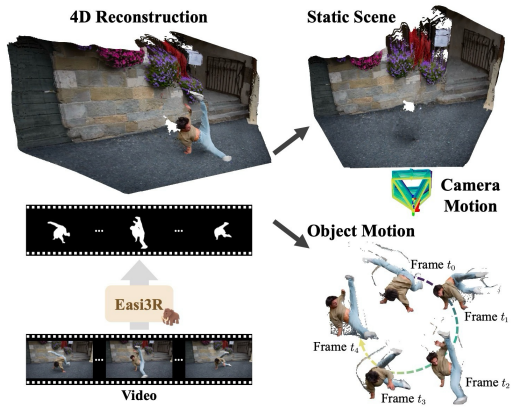
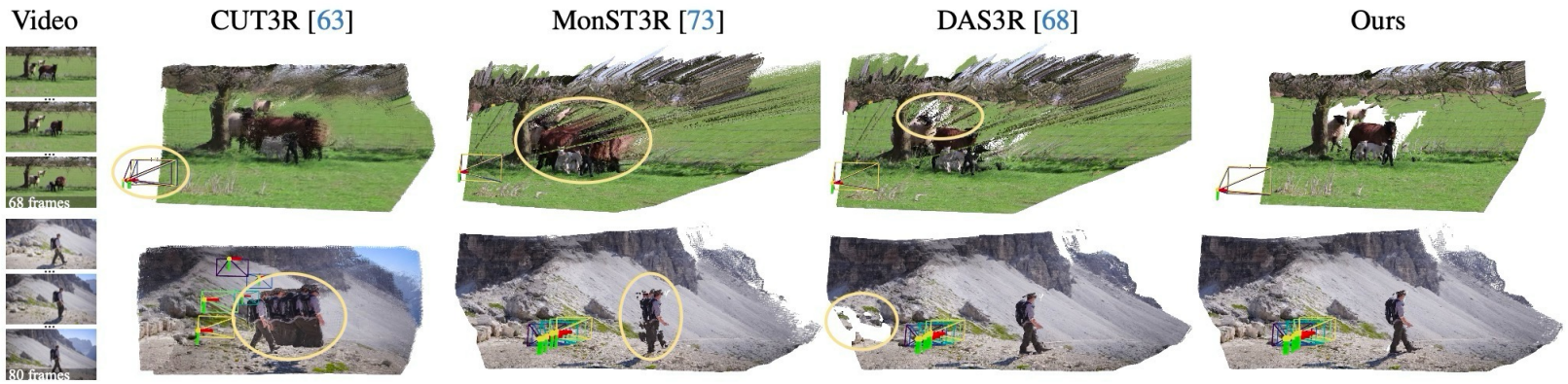




Easi3R: Estimating Disentangled Motion from DUSt3R Without Training

Xingyu Chen[1,2], Yue Chen[1,2], Yuliang Xiu[2,3], Andreas Geiger[4], Anpei Chen[2,4]

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

ICCV 2025

Presenter: 송우석

- **site:** https://easi3r.github.io/
- **github:** https://github.com/Inception3D/Easi3R
- **paper:** https://arxiv.org/abs/2503.24391

RAISE LAB

Index

- Problem & Motivation
- Method
- Experiments
- Conclusion

Recovery of geometry, motion

- foundation task of computer vision
  - downstream application: novel view synthesis, AR/VR, autonomous navigation, robotics etc
- literature commonly identifies this problem as ‘SfM’
- has been the core focus in 3D vision over decades
  - yielding mature algorithms that performs well under stationary conditions & wide baselines

⇒ but these algorithms often **fails on dynamic video input** (fail = degradation of accuracy & robustness)

- main reason: object dynamics (common component in real-world videos)
  - moving objects violate fundamental assumptions of homography & epipolar consistency in traditional SfM methods

- dynamic video
  - entangled camera & object motion → hard to disentangle these two motions
  - motion with rich texture → degrades camera pose estimate performance

Related works

- SfM & SLAM
  - long foundation for **3D structure** and **camera pose estimation**
  - refine structure and motion estimates by
    - associating 2D correspondences or minimizing photometric errors
    - bundle adjustment (BA)

    ⇒ effective with dense input but struggle with limited camera parallax, ill-posed conditions
  - DUSt3R(+ follow-up methods) overcame above problems, but limited on static scenes
- Pose-free Dynamic Scene Reconstruction
  - SLAM + dynamic scenes
    - semantic segmentation, optical flows to enhance SLAM's resilence in dynamic scenes
  - Another line of work
    - stable video depth estimation (+ geometric constraints, generative priors)
    - optimization-based methods
      - **CasualSAM:** fine-tune depth network at test time using pre-computed otpical flow
      - **Robust-CVD:** refines pre-computed depth, camera pose by leveraging masked optical flow
      - **MegaSaM:** incorporate DROID-SLAM, optical flow, depth initializations
    - point-map-based methods
      - **MonST3R:** fine-tuned DUStR with dynamic datasets and incorporate optical flow
      - **DAS3R:** trains DPT on top of MonST3R and performs feedforward segmentation estimation
      - **CUT3R:** encode scene into persistent state and performs feedforward 3D reconstruction

    ⇒ but these methods needs costly training on diverse motion patterns to generalize well

Easi3R ✓

- training-free, plug-in-play adaption
- no fine-tuning, almost no additional cost
- scalable, efficient alternative for handling real-world dynamic videos

?

mation. CUT3R [63] fine-tunes MASt3R [24] on both static and dynamic datasets, achieving feedforward reconstruc-

Related works

- Motion Segmentation
  - Classical approaches
    - rely on optical flow estimation, point tracking
    - only trained in 2D datasets

      ⇒ struggle with occlusions, disentagling object, caemra motion
  - robustness enhanced approaches

    - **RoMo:** incorporate epipolar geometry, accurate calibration with COLMAP

      ⇒ focus primarily on removing dynamic objects and reconstructing static scene elements only
    - **SAM2:** disambiguate object, camera motion

  - complete 4D reconstruction

    - **MonST3R:** fine-tuned DUStR with dynamic datasets and incorporate optical flow
    - **DAS3R:** trains DPT on top of MonST3R and performs feedforward segmentation estimation

Key Insight

⇒ dynamic segmentation can be extracted from pre-trained 3D reconstruction models like DUSt3R

Easi3R ✓

- a simple, yet robust strategy to isolate this information from the **attention layers**
- no need for optical flow or pre-training on segmentation datasets

Easi3R

- Goal

input
- $\{\mathbf{I}^t\}_{t=1}^T$: video sequence

output
- $\mathbf{M}^t$: object motion
- $\mathbf{P}^t, \mathbf{K}^t$: camera movement

$+ \mathbf{X}^t$: point map

- Process

1. how the DUSt3R model handles videos
2. mechanisms of attention aggregation in spatial and temporal dimensions
3. how aggregated cross-attention maps can be leveraged to decompose dynamic object segmentation (re-weighting)

- DUSt3R

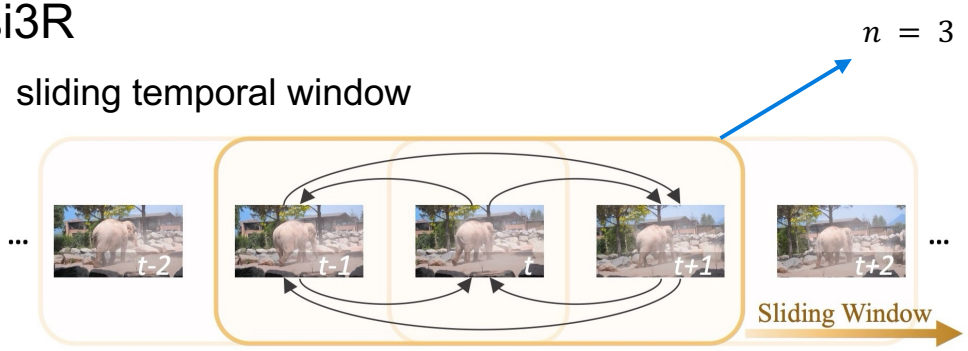
reference view

$$X^{a\to a}, X^{b\to a} = \mathrm{DUSt3R}(I^a, I^b) \qquad (1)$$

  - $X^{b\to a}$: point map of input I^b predicted in the view a coordinate space
  - In multi-views, DUSt3R globally aligns the pairwise predictions into a joint coordinate space using a connectivity graph

    ⇒ computational redundancy problem (view connectivity is know for video sequence) (= waste)

- Easi3R
  - sliding temporal window



  - pair set

$$\varepsilon^t = \{(a,b) | a,b \in [t - \tfrac{n-1}{2}, \ldots, t + \tfrac{n-1}{2}], a \neq b\}$$

scale factor

$\mathbf{P}_i^t \in \mathbb{R}^{3\times 4}$ pair → global

$$\mathcal{X}^* = \arg\min_{\mathcal{X},\mathbf{P},\mathbf{s}} \sum_{t\in T} \sum_{i\in\varepsilon^t} \|\mathcal{X}^a - \mathbf{s}_i^t \mathbf{P}_i^t X^{a\to a}\|_1 + \|\mathcal{X}^b - \mathbf{s}_i^t \mathbf{P}_i^t X^{b\to a}\|_1 \qquad (2)$$

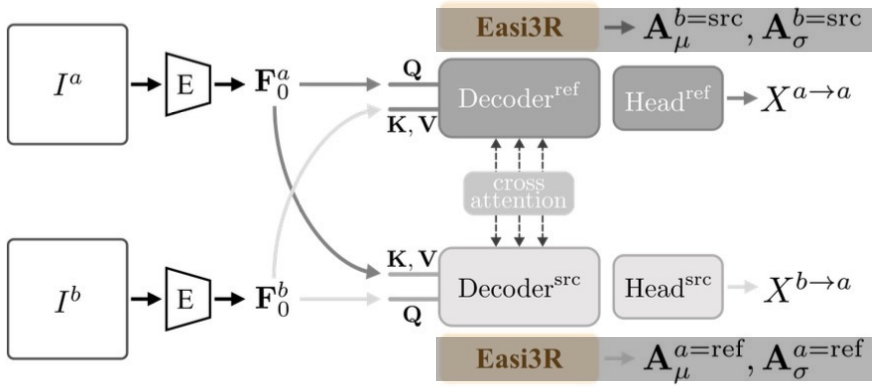
global alignment of pair-wise prediction

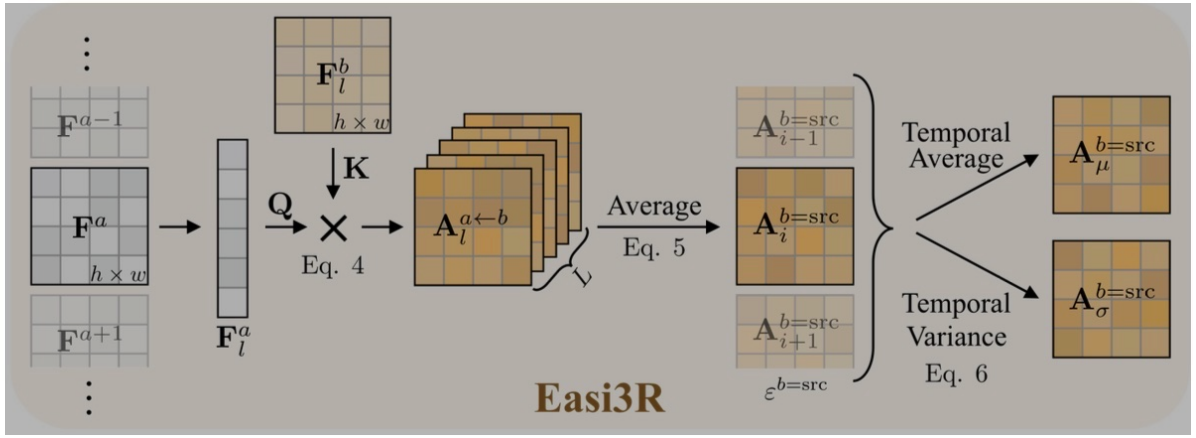
DUSt3R reconstruction

Video frame

DUSt3R can't handle frame with dynamic object

- Secret behind DUSt3R





- 2 branches
  - I^a: reference image
  - I^b: source image

- process
  1. encode image into token representation $\mathbf{F}$ with weight-sharing ViT encoder
  2. exchange information both within and between views (self-attention + cross-attention)
     - use previous block $(l-1)$
  3. point map prediction with 2 heads using feature tokens
  4. iterate minimizing Euclidean distance between predicted point map and GT point map

$$\mathbf{F}_0^a = \text{Encoder}(I^a)$$
$$\mathbf{F}_0^b = \text{Encoder}(I^b)$$

$l = 1, \ldots, L$ block index

self-attention (within views) + cross-attnetion (between views)

$$\mathbf{F}_l^a = \text{DecoderBlock}_l^{\text{ref}}(\mathbf{F}_{l-1}^a, \mathbf{F}_{l-1}^b)$$
$$\mathbf{F}_l^b = \text{DecoderBlock}_l^{\text{src}}(\mathbf{F}_{l-1}^b, \mathbf{F}_{l-1}^a) \quad (3)$$

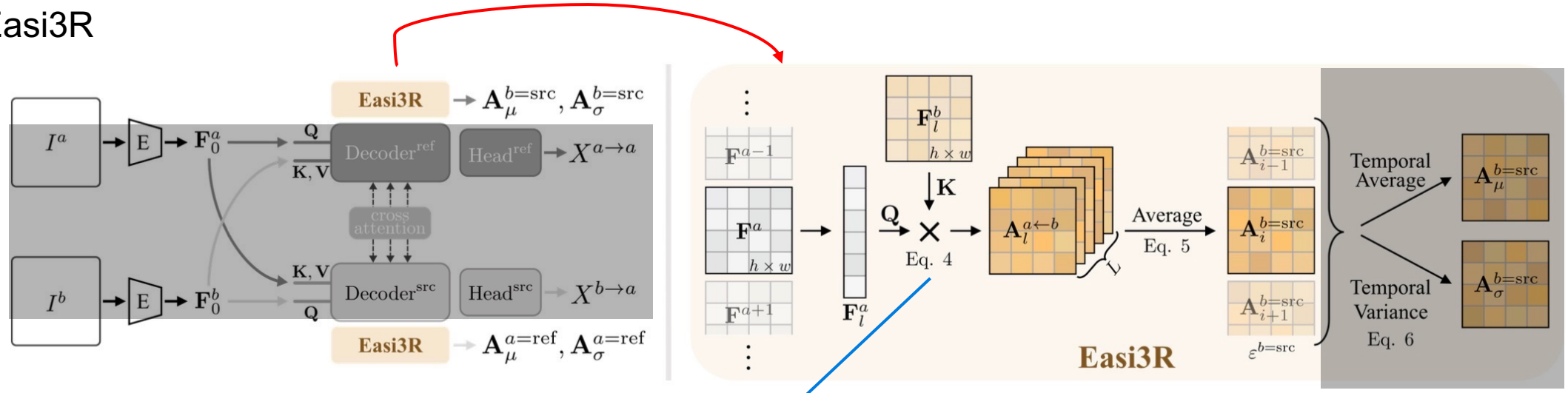
$$X^{a\to a} = \text{Head}^{\text{ref}}(\mathbf{F}_0^a, \ldots, \mathbf{F}_L^a)$$
$$X^{b\to a} = \text{Head}^{\text{src}}(\mathbf{F}_0^b, \ldots, \mathbf{F}_L^b)$$

→ point map prediction

Easi3R ✓

- DUSt3R implicitly learns rigid view transformations with cross-attention layer, assigning low attention values to tokens that violate epipolar geometry constraints (such as texture-less, under-observed, and dynamic regions)
- aggregating cross-attention outputs across spatial & temporal dimensions → enabling motion extraction

- Easi3R



- trainable linear function (projection function of $\mathbf{Q}$, $\mathbf{K}$)

$$\ell_{\mathbf{Q}}(\cdot), \ell_{\mathbf{K}}(\cdot) \qquad \mathbf{Q} = \ell_{\mathbf{Q}}(\mathbf{F}) \in \mathbb{R}^{(h\times w)\times c}$$

- attention map

$$\mathbf{A}_l^{a\leftarrow b} = \mathbf{Q}_l^a \mathbf{K}_l^{bT}/\sqrt{c}, \quad \mathbf{A}_l^{b\leftarrow a} = \mathbf{Q}_l^b \mathbf{K}_l^{aT}/\sqrt{c} \tag{4}$$

$$\mathbf{A}_l^{a\leftarrow b},\ \mathbf{A}_l^{b\leftarrow a} \in \mathbb{R}^{(h\times w)\times h\times w}$$

$\mathbf{A}_l^{a\leftarrow b}$ determines how the information is aggregated from the view b into the view a in the lth decoder block

$$\mathbf{V} = \ell_{\mathbf{V}}(\mathbf{F}) \in \mathbb{R}^{(h\times w)\times c}$$

$$\mathrm{softmax}(\mathbf{A}_l^{a\leftarrow b})\mathbf{V}^b$$

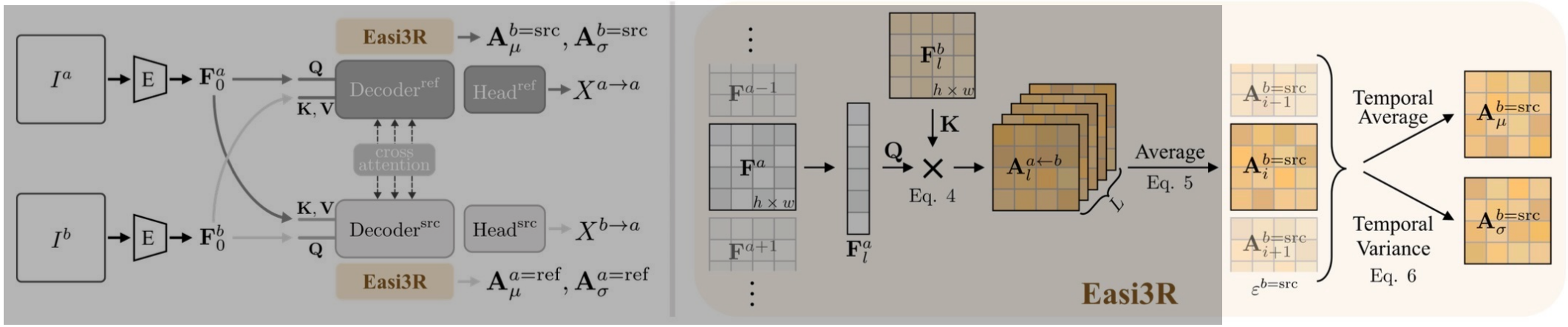
- average attention map

$$\mathbf{A}^{b=\mathrm{src}} = \sum_l \sum_x \mathbf{A}_l^{a\leftarrow b}(x,y,z)/(L\times h\times w)$$

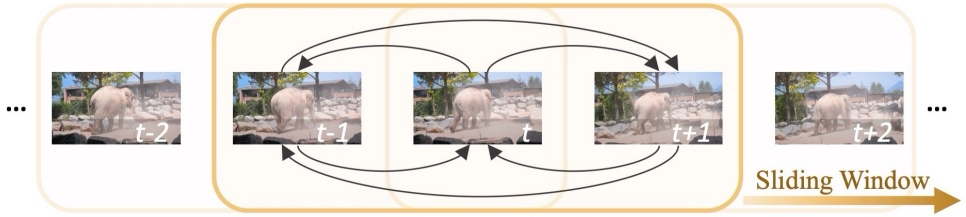
$$\mathbf{A}^{a=\mathrm{ref}} = \sum_l \sum_x \mathbf{A}_l^{b\leftarrow a}(x,y,z)/(L\times h\times w) \tag{5}$$

average attention map captures the overall influence of tokens from one view to another across all decoder layers

- Easi3R



- Temporal attention maps



- $2(n-1)$ attention maps per frame

$$\mathbf{A}_{\mu}^{b=\text{src}} = \text{Mean}(\mathbf{A}_{i}^{b=\text{src}}),\ \mathbf{A}_{\sigma}^{b=\text{src}} = \text{Std}(\mathbf{A}_{i}^{b=\text{src}}) \quad (6)$$

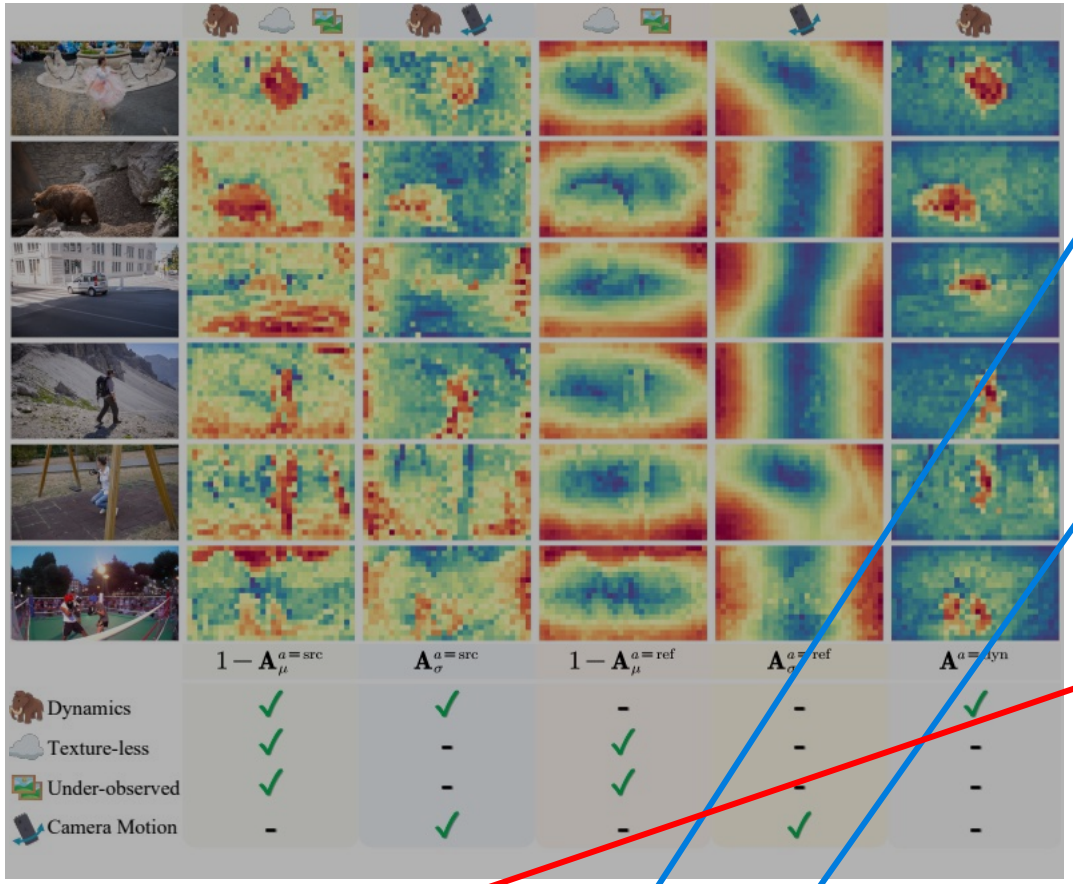
$$i \in \varepsilon^{b=\text{src}}$$

$$\varepsilon^{b=\text{src}} = \{(a,b)|\text{src} = b, a \in [t-n,\ldots,t+n], a \neq b\} \quad (7)$$

$$i \in \varepsilon^{b=\text{ref}}$$

$$\varepsilon^{b=\text{ref}} = \{(a,b)|\text{ref} = b, a \in [t-n,\ldots,t+n], a \neq b\} \quad (8)$$

- Easi3R



$$\mathbf{A}^{a=\text{dyn}} = (1-\mathbf{A}_{\mu}^{a=\text{src}}) \cdot \mathbf{A}_{\sigma}^{a=\text{src}} \cdot \mathbf{A}_{\mu}^{a=\text{ref}} \cdot (1-\mathbf{A}_{\sigma}^{a=\text{ref}}) \quad (9)$$

$$\mathbf{M}^t = [\mathbf{A}^{t=\text{dyn}} > \alpha]\ (\mathbf{M}^t \in \mathbb{R}^{h \times w})$$

→ : inverted

⇒ no worries for o.o.d input (e.g. black pixels) resulting performance degradation

- Temporal attention maps

$1-\mathbf{A}_{\mu}^{a=\text{ref}}$

$\mathbf{A}_{\mu}^{a=\text{ref}}$
- smooth attention values
- low attention values with texture-less regions, under-observed areas
  ⇒ DUSt3R belives that they are less useful for registration
  ⇒ can be extracted using $(1-\mathbf{A}_{\mu}^{a=\text{ref}})$

$\mathbf{A}_{\sigma}^{a=\text{ref}}$
- changes of token contribution in image coordinate space
- pixels perpendicular to the direction of motion generally share similar pixel flow speeds
  ⇒ consistent deviation that allow to infer camera motion

$1-\mathbf{A}_{\mu}^{a=\text{src}}$

$\mathbf{A}_{\mu}^{a=\text{src}}$
- indicates low texture & under-observed area + dynamic object
  ⇒ they violate the rigid body transformation prior that DUSt3R has learned from the 3D dataset
  ⇒ can be extracted using $(1-\mathbf{A}_{\mu}^{a=\text{src}})$

$\mathbf{A}_{\sigma}^{a=\text{src}}$
- highlights camera & object motion
  ⇒ these areas continously changes over time, leading high deviation

(supplementary material)

- Clustering Fusion

$$\bar{\mathbf{F}} = [\mathbf{F}_0^1; \mathbf{F}_0^2; \ldots; \mathbf{F}_0^T] \in \mathbb{R}^{(T \times h \times w) \times c} \quad (12)$$

number of cluster ($k = 64$)

$$C = \mathrm{KMeans}(\bar{\mathbf{F}}, k), \quad C^t(x, y) \in \{1, \ldots, k\}, \ \forall t, x, y \quad (13)$$

dynamic score

indicator function

$$s_c = \frac{\sum_t \sum_{i,j} \mathbb{1}[C^t(x, y) = c] \cdot \mathbf{A}^{t=\mathrm{dyn}}(x, y)}{\sum_t \sum_{x,y} \mathbb{1}[C^t(x, y) = c]} \quad (14)$$

cluster fused dynamic attention map

$$\mathbf{A}_{\mathrm{fuse}}^{t=\mathrm{dyn}}(x, y) = s_{C^t(x,y)} \quad (15)$$

$$\mathbf{M}^t(x, y) = \mathbb{1}[\mathbf{A}_{\mathrm{fuse}}^{t=\mathrm{dyn}}(x, y) > \alpha] \quad (16)$$

- α is automatic image thresholding using Otsu's method

Backbone	Ablation	DAVIS-16 JM↑	DAVIS-16 JR↑	DAVIS-17 JM↑	DAVIS-17 JR↑	DAVIS-all JM↑	DAVIS-all JR↑
DUSt3R	w/o $\mathbf{A}_{\mu}^{a=\mathrm{src}}$	45.1	45.2	42.8	39.9	42.2	38.5
	w/o $\mathbf{A}_{\sigma}^{a=\mathrm{src}}$	42.3	50.0	35.0	37.0	30.9	28.3
	w/o $\mathbf{A}_{\mu}^{a=\mathrm{ref}}$	33.3	28.4	31.5	27.9	32.5	29.7
	w/o $\mathbf{A}_{\sigma}^{a=\mathrm{ref}}$	47.7	54.1	46.2	54.3	43.7	48.6
	w/o Clustering	40.0	38.5	38.3	38.3	34.3	30.5
	Full	**53.1**	**60.4**	**49.0**	**56.4**	**44.5**	**49.6**
MonST3R	w/o $\mathbf{A}_{\mu}^{a=\mathrm{src}}$	47.2	51.5	44.4	46.7	40.9	41.5
	w/o $\mathbf{A}_{\sigma}^{a=\mathrm{src}}$	49.7	60.1	48.7	57.8	44.9	49.6
	w/o $\mathbf{A}_{\mu}^{a=\mathrm{ref}}$	46.4	54.0	47.4	55.9	45.3	50.7
	w/o $\mathbf{A}_{\sigma}^{a=\mathrm{ref}}$	50.7	62.6	51.0	60.2	50.3	56.8
	w/o Clustering	45.5	46.7	45.3	48.1	42.1	43.5
	Full	**57.7**	**71.6**	**56.5**	**68.6**	**53.0**	**63.4**

- Attention re-weighting

static attention mask of view a

dynamic attention mask of view b

$$\mathbf{M}^{a \leftarrow b} = (1 - \mathbf{M}^{a}) \otimes \mathbf{M}^{b^T}$$

$$\text{softmax}(\tilde{A}_l^{a \leftarrow b}) = \begin{cases} 0 & \text{if } \mathbf{M}^{a \leftarrow b} \\ \text{softmax}(\mathbf{A}_l^{a \leftarrow b}) & \text{otherwise} \end{cases} \quad (10)$$

- resulting tokens from dynamic regions in view b that **do not contribute** to static regions in view a
- re-weighting only applied on reference view decoder (source view requires a static reference)

Backbone	Re-weighting	Flow-GA	Pose Estimation			Reconstruction					
			ATE↓	RTE↓	RRE↓	Accuracy↓		Completeness↓		Distance↓	
						Mean	Median	Mean	Median	Mean	Median
DUSt3R	Ref + Src	✗	0.030	0.026	1.777	0.775	0.596	1.848	0.778	0.342	0.224
	Ref	✗	0.029	0.025	1.774	0.772	0.596	1.813	0.757	0.336	0.219
	Ref	w/o Mask	0.026	0.017	1.472	0.940	0.831	1.654	0.685	0.336	0.220
	Ref	w/ Mask	**0.021**	**0.014**	**1.092**	**0.703**	**0.589**	**1.474**	**0.586**	**0.301**	**0.186**
MonST3R	Ref + Src	✗	0.040	0.032	1.751	0.848	0.744	1.850	1.003	0.398	0.292
	Ref	✗	0.038	0.032	1.736	0.846	0.660	1.840	0.983	0.390	0.290
	Ref	w/o Mask	0.033	0.023	1.495	0.969	0.796	1.752	0.998	0.368	0.273
	Ref	w/ Mask	**0.030**	**0.021**	**1.390**	**0.834**	**0.643**	**1.661**	**0.916**	**0.350**	**0.255**

ablation study of Camera Pose Estimation and Point Cloud Reconstruction

Global Alignment

$$\mathcal{X}^* = \arg\min_{\mathcal{X},\mathbf{P},\mathbf{s}} \sum_{t\in T}\sum_{i\in\varepsilon^t} \|\mathcal{X}^a - \mathbf{s}_i^t\mathbf{P}_i^t X^{a\to a}\|_1 + \|\mathcal{X}^b - \mathbf{s}_i^t\mathbf{P}_i^t X^{b\to a}\|_1 \quad (2)$$

global alignment of pair-wise prediction

$\mathcal{X}^b : (\mathbf{P}^a, \mathbf{K}^a) \to (\mathbf{P}^b, \mathbf{K}^b)$

computed optical flow

estimated optical flow

$$\mathcal{L}_{\text{flow}} = \sum_{t\in T}\sum_{i\in\varepsilon^t}(1-\mathbf{M}^a)\cdot\|\hat{\mathcal{F}}_i^{a\to b} - \mathcal{F}_i^{a\to b}\|_1 + (1-\mathbf{M}^b)\cdot\|\hat{\mathcal{F}}_i^{b\to a} - \mathcal{F}_i^{b\to a}\|_1 \quad (11)$$

flow loss of static area

incorporation can achieve more robust outputs in tems of global pointmaps and pose sequence

but, optioanlly used (for fair comparison)

(segmentation)

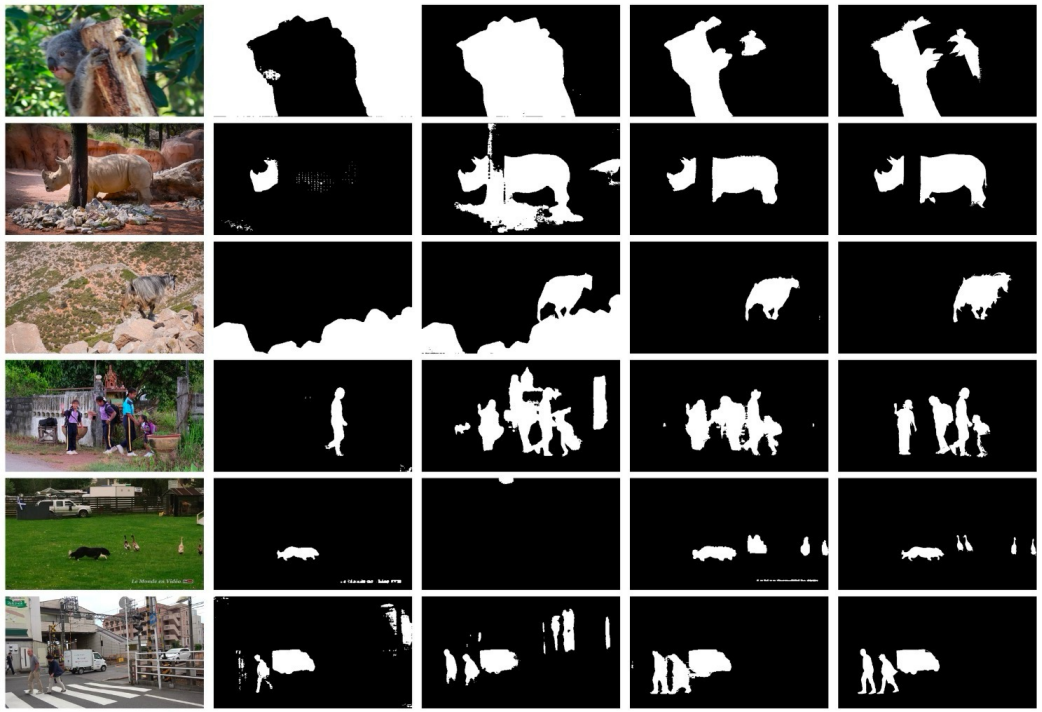
Dynamic Object Motion

- dataset (video object segmentation benchmark)
  - DAVIS-16
  - DAVIS-17
  - DAVIS-all
- metric
  - JM (IoU mean)
  - JR (IoU recall)
- result
  - improved segmentation without need for flow
  - even surpassed DAS3R (trained on dynamic mask labels)

outputs serve as prompt for SAM2

+ flow-guided segmentation

Method	Flow	DAVIS-16 w/o SAM2 JM↑	DAVIS-16 w/o SAM2 JR↑	DAVIS-16 w/ SAM2 JM↑	DAVIS-16 w/ SAM2 JR↑	DAVIS-17 w/o SAM2 JM↑	DAVIS-17 w/o SAM2 JR↑	DAVIS-17 w/ SAM2 JM↑	DAVIS-17 w/ SAM2 JR↑	DAVIS-all w/o SAM2 JM↑	DAVIS-all w/o SAM2 JR↑	DAVIS-all w/ SAM2 JM↑	DAVIS-all w/ SAM2 JR↑
DUSt3R [64]	✓	42.1	45.7	58.5	63.4	35.2	35.3	48.7	50.2	35.9	34.0	47.6	48.7
MonST3R [73]	✓	40.9	42.2	64.3	73.3	38.6	38.2	56.4	59.6	36.7	34.3	51.9	54.1
DAS3R [68]	✗	41.6	39.0	54.2	55.8	43.5	42.1	57.4	61.3	43.4	38.7	53.9	54.8
Easi3R$_{dust3r}$	✗	53.1	60.4	67.9	71.4	49.0	56.4	60.1	65.3	44.5	49.6	54.7	60.6
Easi3R$_{monst3r}$	✗	**57.7**	**71.6**	**70.7**	**79.9**	**56.5**	**68.6**	**67.9**	**76.1**	**53.0**	**63.4**	**63.1**	**72.6**



Video Frame | MonST3R [73] | DAS3R [68] | Ours | GT

RAISE LAB

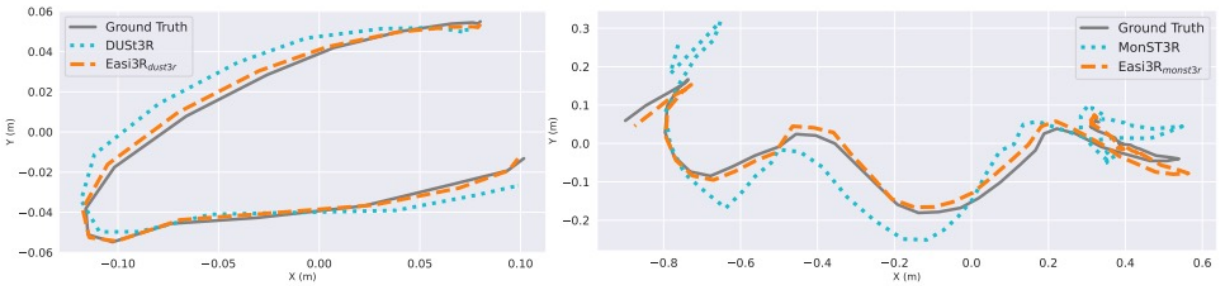
(pose estimation)

Camera Motion

- dataset (dynamic benchmark dataset)
  - DyCheck
    - diverse, in-the-wild dynamic videos captured from handhel cameras
  - TUM-dynamics
    - major dynamic objects in relatively simple indoor scenarios
  - ADT
    - egocentric videos, which are out-of-distribution for DUSt3R's training set
- metric
  - ATE (Absolute Trajectory Error)
  - RTE (Relative Translation Error)
  - RRE (Relative Rotation Error)

    ⇒ after applying the Sim(3) alignment on the estimated cametra trajectory to GT
- result
  - best overall performance among all methods

		DyCheck			ADT			TUM-dynamics		
Method	Flow	ATE ↓	RTE ↓	RRE ↓	ATE ↓	RTE ↓	RRE ↓	ATE ↓	RTE ↓	RRE ↓
DUSt3R [64]	✗	0.035	0.030	2.323	0.042	0.025	1.212	0.100	0.087	2.692
Easi3R$_{dust3r}$	✗	0.029	0.025	1.774	**0.040**	0.021	0.880	0.093	0.076	2.366
DUSt3R [64]	✓	0.029	0.021	1.875	0.076	0.030	0.974	0.071	0.067	3.711
Easi3R$_{dust3r}$	✓	**0.021**	**0.014**	**1.092**	0.042	**0.015**	**0.655**	**0.070**	**0.061**	**2.361**
MonST3R [73]	✗	0.040	0.034	1.820	0.045	0.024	0.759	0.183	**0.148**	6.985
Easi3R$_{monst3r}$	✗	0.038	0.032	1.736	0.045	0.024	0.715	0.184	0.149	6.311
MonST3R [73]	✓	0.033	0.024	1.501	0.055	0.025	0.776	0.170	0.155	6.455
Easi3R$_{monst3r}$	✓	**0.030**	**0.021**	**1.390**	**0.039**	**0.016**	**0.640**	**0.168**	0.150	**5.925**



RAISE LAB

4D Reconstruction

- dataset
  - DyCheck
- metric
  - accuracy
    - nearest Euclidean distance from a reconstructed point to GT
  - completeness
    - reverse of accuracy
  - distance
    - Euclidean distance from a reconstructed point to GT

Method	Flow	Accuracy↓		Completeness↓		Distance↓	
		Mean	Median	Mean	Median	Mean	Median
DUSt3R [64]	✗	0.802	0.595	1.950	0.815	0.353	0.233
Easi3R$_{dust3r}$	✗	0.772	0.596	1.813	0.757	0.336	0.219
DUSt3R [64]	✓	0.738	0.599	1.669	0.678	0.313	0.196
Easi3R$_{dust3r}$	✓	**0.703**	**0.589**	**1.474**	**0.586**	**0.301**	**0.186**
MonST3R [73]	✗	0.855	0.693	1.916	1.035	0.398	0.295
Easi3R$_{monst3r}$	✗	0.846	0.660	1.840	0.983	0.390	0.290
MonST3R [73]	✓	0.851	0.689	1.734	0.958	0.353	**0.254**
Easi3R$_{monst3r}$	✓	**0.834**	**0.643**	**1.661**	**0.916**	**0.350**	0.255

Method	Flow	Accuracy↓		Completeness↓		Distance↓	
		Mean	Median	Mean	Median	Mean	Median
DUSt3R [64]	✗	0.802	0.595	1.950	0.815	0.353	0.233
CUT3R [63]	✗	**0.458**	**0.342**	1.633	0.792	0.326	0.229
MonST3R [73]	✓	0.851	0.689	1.734	0.958	0.353	0.254
DAS3R [68]	✓	1.772	1.438	2.503	1.548	0.475	0.352
Easi3R$_{monst3r}$	✓	0.834	0.643	1.661	0.916	0.350	0.255
Easi3R$_{dust3r}$	✓	0.703	0.589	**1.474**	**0.586**	**0.301**	**0.186**

- result
  - Quantitative – more accurate reconstruction than other baselines, even comparable to CUT3R (concurrent & trained with extensive datasets)
  - Qualitative – other baselines struggle with misalignment and entablement of dynamic & static reconstructions

    ⇒ resulting broken geometry, distortions, ghosting artifacts

RAISE LAB

Ablation study

Backbone	Re-weighting	Flow-GA	Pose Estimation			Reconstruction					
			ATE↓	RTE↓	RRE↓	Accuracy↓ Mean	Accuracy↓ Median	Completeness↓ Mean	Completeness↓ Median	Distance↓ Mean	Distance↓ Median
DUSt3R	Ref + Src	✗	0.030	0.026	1.777	0.775	0.596	1.848	0.778	0.342	0.224
	Ref	✗	0.029	0.025	1.774	0.772	0.596	1.813	0.757	0.336	0.219
	Ref	w/o Mask	0.026	0.017	1.472	0.940	0.831	1.654	0.685	0.336	0.220
	Ref	w/ Mask	**0.021**	**0.014**	**1.092**	**0.703**	**0.589**	**1.474**	**0.586**	**0.301**	**0.186**
MonST3R	Ref + Src	✗	0.040	0.032	1.751	0.848	0.744	1.850	1.003	0.398	0.292
	Ref	✗	0.038	0.032	1.736	0.846	0.660	1.840	0.983	0.390	0.290
	Ref	w/o Mask	0.033	0.023	1.495	0.969	0.796	1.752	0.998	0.368	0.273
	Ref	w/ Mask	**0.030**	**0.021**	**1.390**	**0.834**	**0.643**	**1.661**	**0.916**	**0.350**	**0.255**

Ablation	**Variants**	**DAVIS-16** JM↑	**DAVIS-16** JR↑	**DAVIS-17** JM↑	**DAVIS-17** JR↑	**DAVIS-all** JM↑	**DAVIS-all** JR↑
Window Size	3	76.0	89.2	70.8	82.4	65.7	76.2
	5*	70.7	79.9	67.9	76.1	63.1	72.6
	7	66.9	76.9	63.9	73.3	61.0	68.8
Number of Clusters	16	67.4	79.1	64.6	73.7	60.7	65.2
	32	71.6	83.9	68.0	78.3	65.1	75.2
	64*	70.7	79.9	67.9	76.1	63.1	72.6
	128	69.9	79.9	66.3	76.2	62.9	73.1
Thresholding Values	0.5	61.6	64.5	61.0	65.2	61.6	67.8
	0.7	70.2	85.0	62.8	71.8	58.1	67.0
	Otsu's method*	70.7	79.9	67.9	76.1	63.1	72.6

Backbone	Ablation	DAVIS-16 JM↑	DAVIS-16 JR↑	DAVIS-17 JM↑	DAVIS-17 JR↑	DAVIS-all JM↑	DAVIS-all JR↑
DUSt3R	w/o $\mathbf{A}_{\mu}^{a=\text{src}}$	45.1	45.2	42.8	39.9	42.2	38.5
	w/o $\mathbf{A}_{\sigma}^{a=\text{src}}$	42.3	50.0	35.0	37.0	30.9	28.3
	w/o $\mathbf{A}_{\mu}^{a=\text{ref}}$	33.3	28.4	31.5	27.9	32.5	29.7
	w/o $\mathbf{A}_{\sigma}^{a=\text{ref}}$	47.7	54.1	46.2	54.3	43.7	48.6
	w/o Clustering	40.0	38.5	38.3	38.3	34.3	30.5
	Full	**53.1**	**60.4**	**49.0**	**56.4**	**44.5**	**49.6**
MonST3R	w/o $\mathbf{A}_{\mu}^{a=\text{src}}$	47.2	51.5	44.4	46.7	40.9	41.5
	w/o $\mathbf{A}_{\sigma}^{a=\text{src}}$	49.7	60.1	48.7	57.8	44.9	49.6
	w/o $\mathbf{A}_{\mu}^{a=\text{ref}}$	46.4	54.0	47.4	55.9	45.3	50.7
	w/o $\mathbf{A}_{\sigma}^{a=\text{ref}}$	50.7	62.6	51.0	60.2	50.3	56.8
	w/o Clustering	45.5	46.7	45.3	48.1	42.1	43.5
	Full	**57.7**	**71.6**	**56.5**	**68.6**	**53.0**	**63.4**

Conclusion

- **Easi3R**
  - spatial & temporal attention mechanism behind DUSt3R
  - second inference pass (attention re-weighting)
  - achieve training-free, robust 4D reconstruction
  - outperforms SoTA methods in most cases

- Limitation
  - fail when DUSt3R/MonST3R backbones produce inaccurate depth prdictions