A Study of Finetuning Video Transformers for Multi-view Geometry Tasks

Huimin Wu[1], Kwang-Ting Cheng[1], Stephen Lin[2], Zhirong Wu[2]

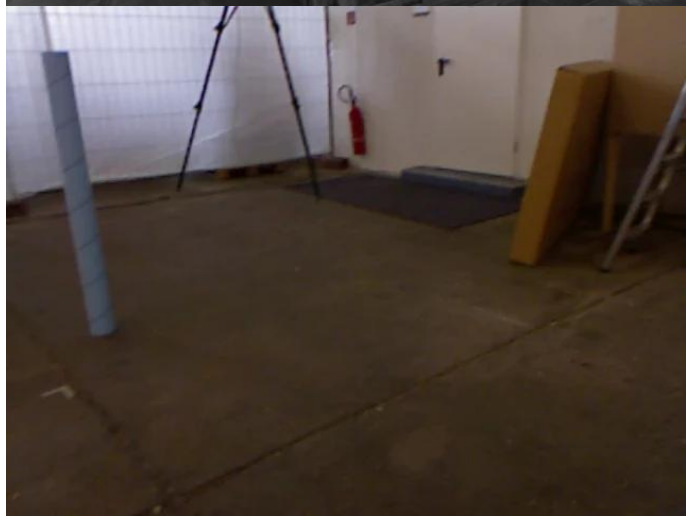
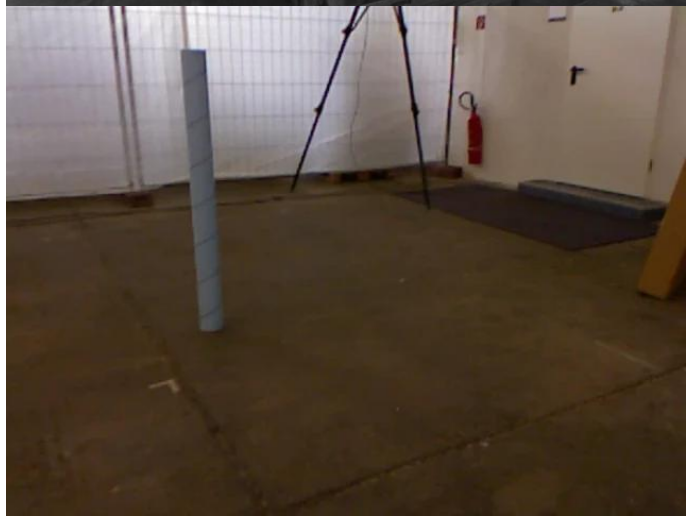
[1]The Hong Kong University of Science and Technology [2] Microsoft Research Asia

AAAI 2026

Problem Definition: Multi-view Geometry Tasks



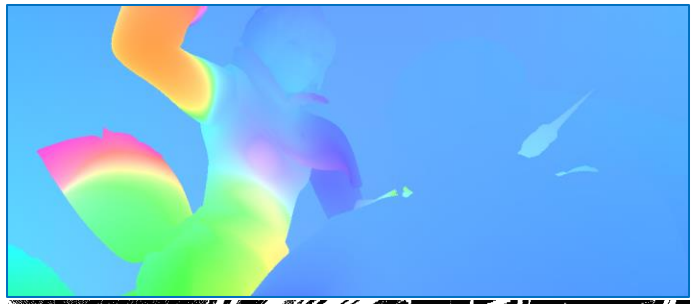
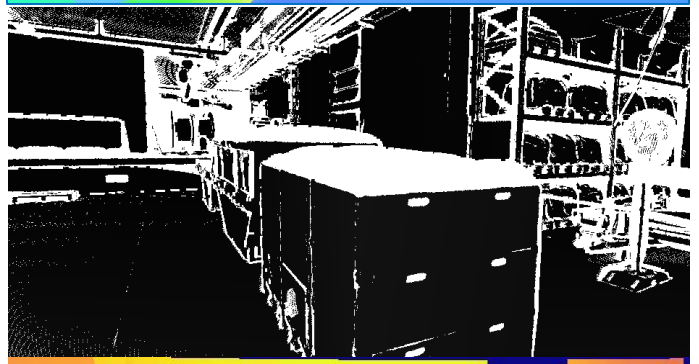
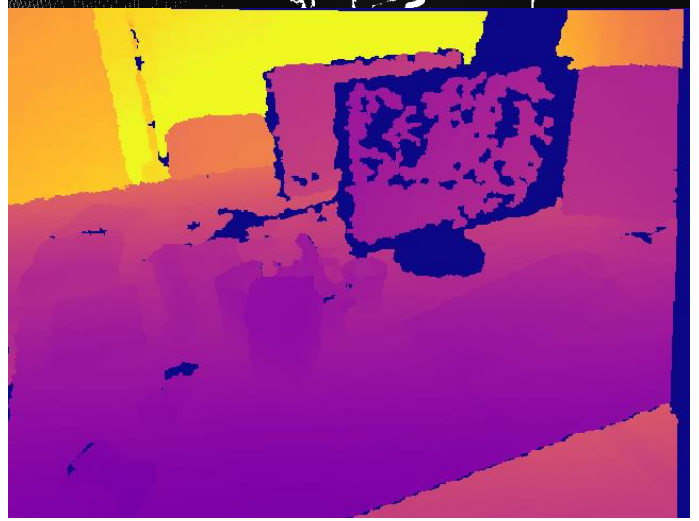


Optical flow estimation

Stereo matching

Depth estimation

GeoViT: A New Paradigm

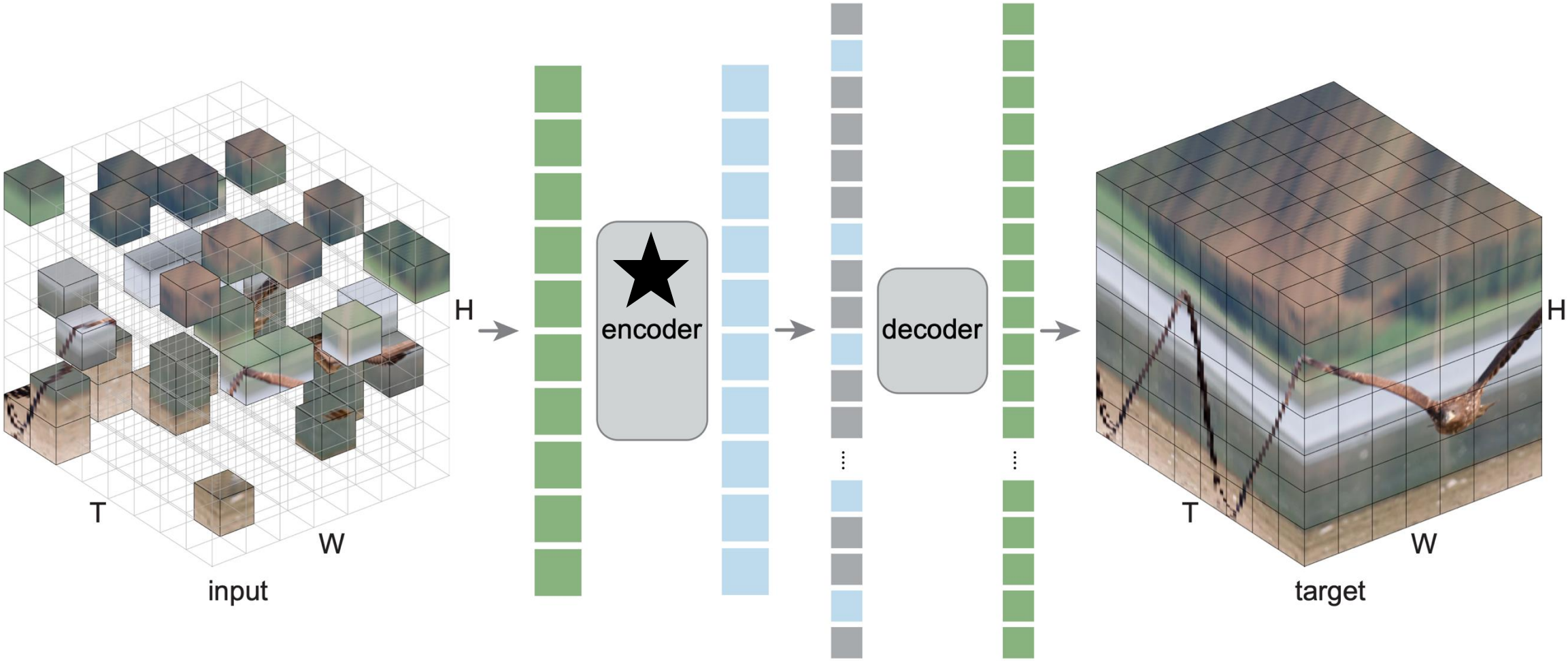
Previous methods

- ☹ Custom architectural designs
- ☹ Task-specific pretraining

Our innovation

- ☺ General-purpose models
  Pretrained on videos
- ☺ Readily transferred
  Minimal adaptation to multi-view problems
- ☺ Core insight: General-purpose attention learns temporal and spatial information for geometric reasoning.

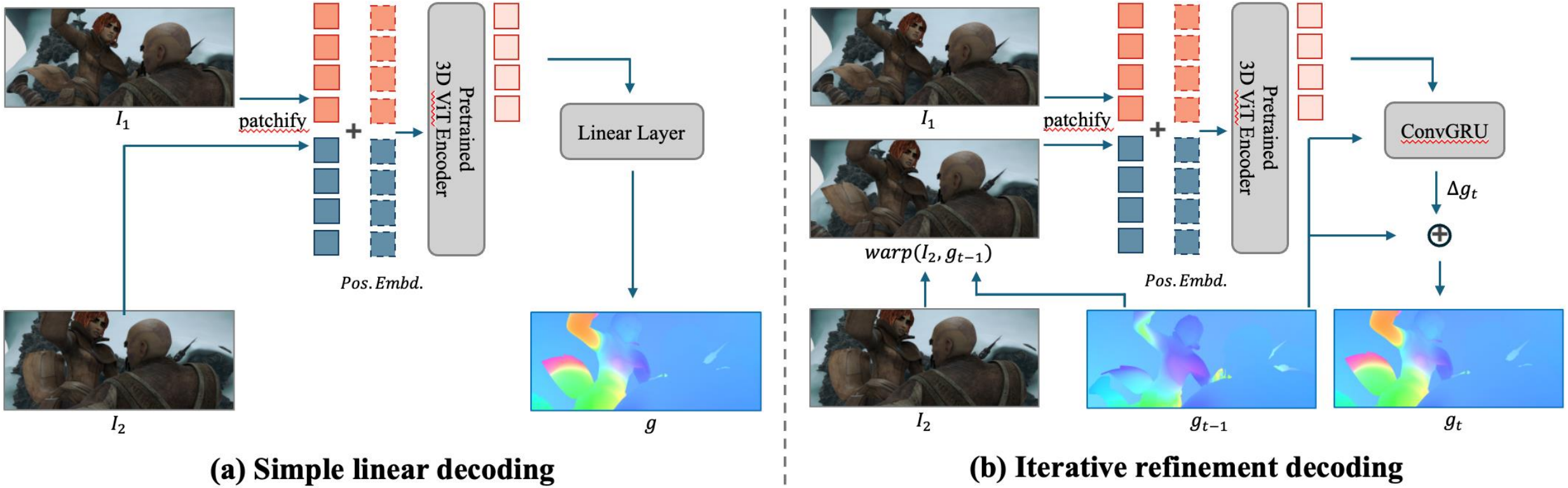
Video Foundation Models



MAE_st pretraining (cited from Feichtenhofer, C.; et al, 2022.)

Feichtenhofer, C.; Li, Y.; He, K.; et al. 2022. Masked autoencoders as spatiotemporal learners. Advances in neural information processing systems, 35: 35946–35958.

Adapting 3D ViT for Geometry Tasks

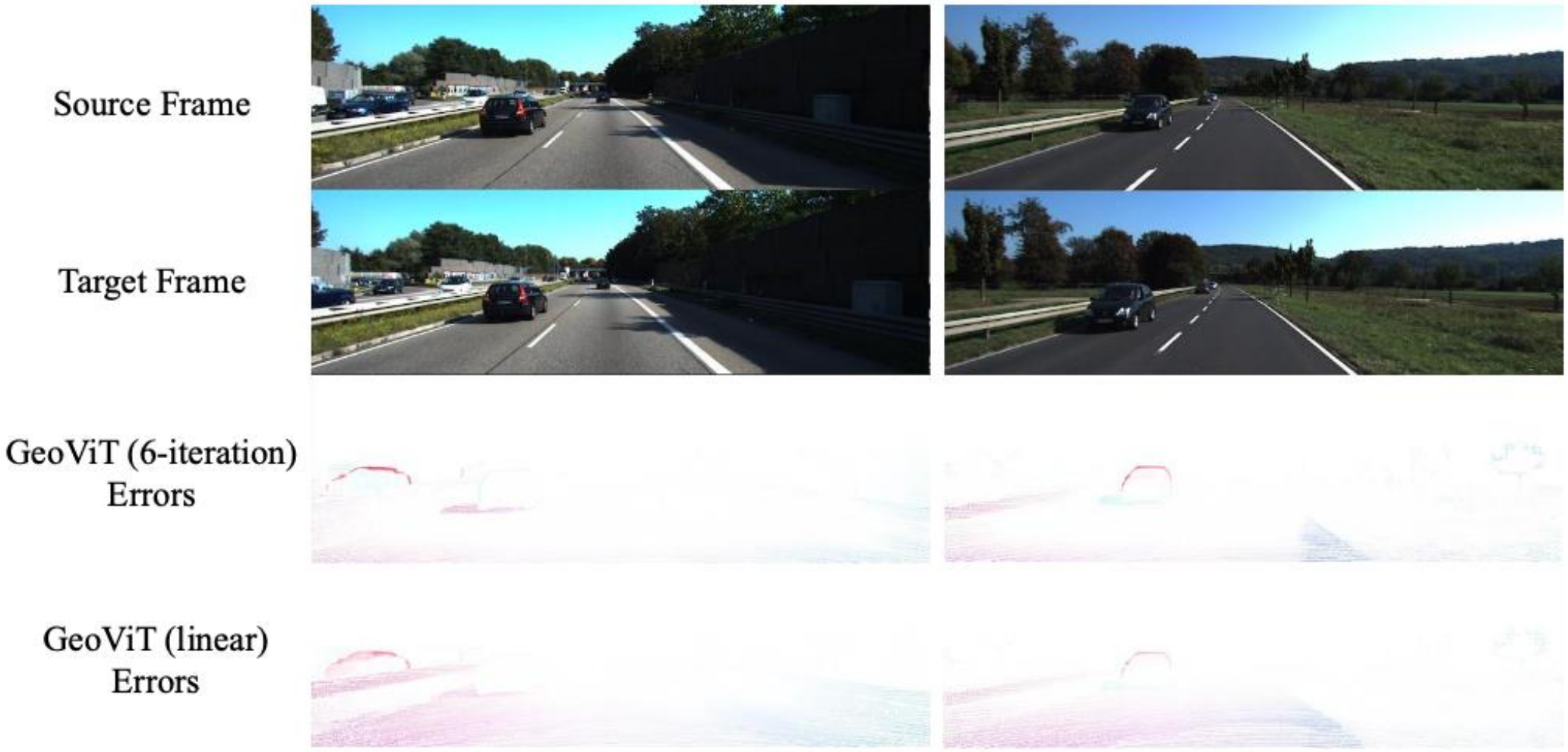


Optical Flow Estimation Results

Cross-dataset generalization

Training Data	Method	Sintel (train)		KITTI-15 (train)	
		Clean	Final	F1-epe	F1-all
A	Perceiver IO (Jaegle et al. 2021)	1.81	2.42	4.98	-
	PWC-Net (Sun et al. 2018)	2.17	2.91	5.76	-
	RAFT (Teed and Deng 2020)	1.95	2.57	4.23	-
C + T	HD3 (Yin, Darrell, and Yu 2019)	3.84	8.77	13.17	24.0
	LiteFlowNet (Hui, Tang, and Loy 2018)	2.48	4.04	10.39	28.5
	PWC-Net (Sun et al. 2018)	2.55	3.93	10.35	33.7
	LiteFlowNet2 (Hui, Tang, and Loy 2020)	2.24	3.78	8.97	25.9
	S-Flow (Zhang et al. 2021)	1.30	2.59	4.60	15.9
	RAFT (Teed and Deng 2020)	1.43	2.71	5.04	17.4
	FM-RAFT (Jiang et al. 2021b)	1.29	2.95	6.80	19.3
	GMA (Jiang et al. 2021a)	1.30	2.74	4.69	17.1
	GMFlow (Xu et al. 2022a)	1.08	2.48	7.77	23.40
	GMFlowNet (Zhao et al. 2022)	1.14	2.71	4.24	15.4
	CRAFT (Sui et al. 2022)	1.27	2.79	4.88	17.5
	SKFlow (Sun et al. 2022)	1.22	2.46	4.47	15.5
	FlowFormer (Huang et al. 2022)	0.94	2.33	4.09†	14.72†
	FlowFormer++ (Shi et al. 2023)	0.90	2.30	3.93†	14.13†
	SAMFlow (Zhou et al. 2024)	0.87	2.11	3.44	12.28
	DPFlow (Morimitsu et al. 2025)	1.02	2.26	3.37	**11.1**
	GeoViT-linear	0.91†	2.00†	4.93†	20.47†
	GeoViT	**0.69**†	**1.78**†	**3.15**†	11.45†

Visualized Comparison: Simple Linear Decoding vs. Iterative Refinement Decoding



Optical Flow Estimation Results

Sintel benchmark

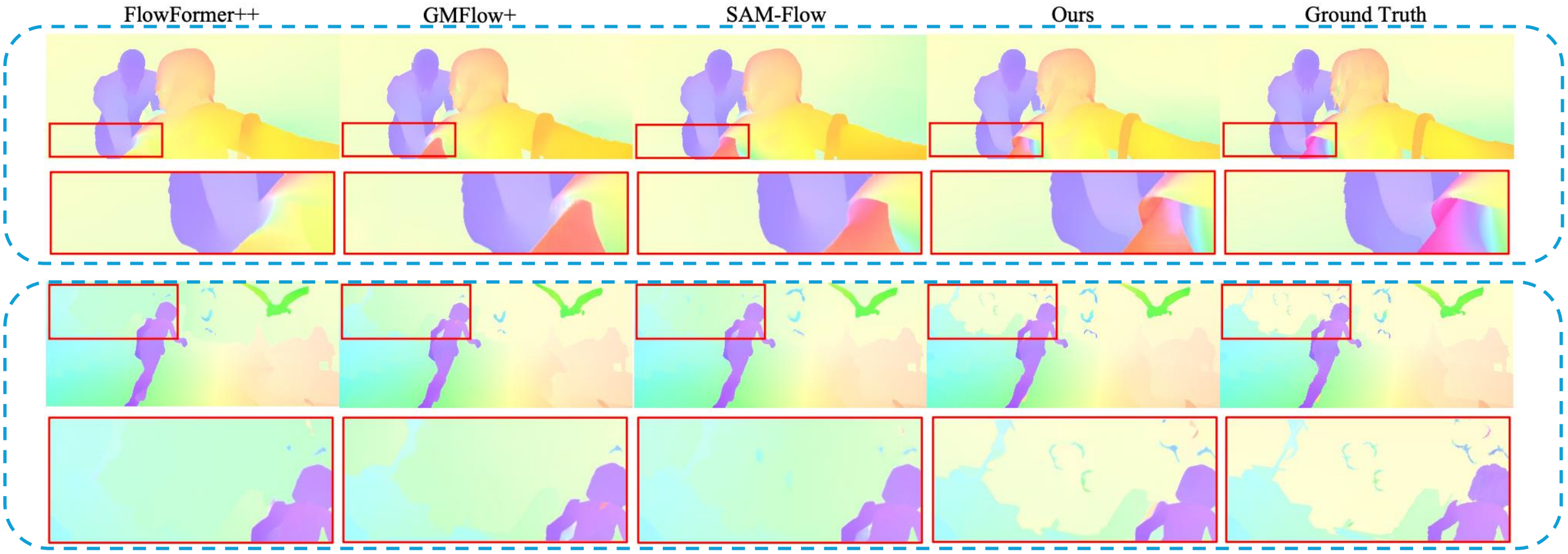
Training Data	Method	Sintel (test)	
		Clean	Final
C + T + S +K + H	LiteFlowNet2 (Hui, Tang, and Loy 2020)	3.48	4.69
	PWC-Net+ (Sun et al. 2019)	3.45	4.60
	VCN (Yang and Ramanan 2019)	2.81	4.40
	MaskFlowNet (Zhao et al. 2020)	2.52	4.17
	S-Flow (Zhang et al. 2021)	1.50	2.67
	RAFT (Teed and Deng 2020)	1.94	3.18
	RAFT* (Teed and Deng 2020)	1.61	2.86
	FM-RAFT (Jiang et al. 2021a)	1.72	3.60
	GMA (Jiang et al. 2021a)	1.40	2.88
	GMA* (Jiang et al. 2021a)	1.39	2.47
	GMFlow (Xu et al. 2022a)	1.74	2.90
	GMFlowNet (Zhao et al. 2022)	1.39	2.65
	CRAFT (Sui et al. 2022)	1.45	2.42
	SKFlow* (Sun et al. 2022)	1.28	2.23
	FlowFormer (Huang et al. 2022)	1.16	2.09
	FlowFormer++ (Shi et al. 2023)	1.07	<u>1.94</u>
	GMFlow+(Weinzaepfel et al. 2023)	1.03	2.12
	SAM-Flow (Zhou et al. 2024)	<u>1.00</u>	2.08
	DPFlow (Morimitsu et al. 2025)	1.04	1.97
	GeoViT	**0.79**†	**1.88**†

Optical Flow Estimation Results

KITTI benchmark

Training Data	Method	KITTI-15 (test)
		F1-all
C + T + S +K + H	LiteFlowNet2 (Hui, Tang, and Loy 2020)	7.74
	PWC-Net+ (Sun et al. 2019)	7.72
	VCN (Yang and Ramanan 2019)	6.30
	MaskFlowNet (Zhao et al. 2020)	6.10
	S-Flow (Zhang et al. 2021)	4.64
	RAFT (Teed and Deng 2020)	5.10
	RAFT* (Teed and Deng 2020)	5.10
	FM-RAFT (Jiang et al. 2021a)	6.17
	GMA (Jiang et al. 2021a)	5.15
	GMA* (Jiang et al. 2021a)	5.15
	GMFlow (Xu et al. 2022a)	9.32
	GMFlowNet (Zhao et al. 2022)	4.79
	CRAFT (Sui et al. 2022)	4.79
	SKFlow* (Sun et al. 2022)	4.84
	FlowFormer (Huang et al. 2022)	4.68†
	FlowFormer++ (Shi et al. 2023)	4.52†
	GMFlow+(Weinzaepfel et al. 2023)	4.27
	SAM-Flow (Zhou et al. 2024)	4.49
	DPFlow (Morimitsu et al. 2025)	**3.56**
	GeoViT	<u>3.79</u>†

Visualized Comparison on Sintel



Stereo Matching Results

ETH3D benchmark

Model	bad 1.0	bad 2.0	bad 4.0
GANet (Zhang et al. 2019)	6.56	1.10	0.54
AANet (Xu and Zhang 2020)	5.01	1.66	0.75
CFNet (Shen, Dai, and Rao 2021)	3.31	0.77	0.31
RAFT-Stereo (Lipson, Teed, and Deng 2021)	2.44	0.44	0.15
CREStereo (Li et al. 2022a)	0.98	0.22	0.10
GMStereo (Xu et al. 2023b)	1.83	0.25	0.08
MonSter (Cheng et al. 2025)	**0.72**	0.42	0.20
GeoViT	1.16	**0.19**	**0.03**

Two-view Depth Estimation Results

DeMoN benchmark

Dataset	Model	Abs Rel	Sq Rel	RMSE	RMSE log
RGBD-SLAM	DeMoN (Ummenhofer et al. 2017)	0.157	0.524	1.780	0.202
	DeepMVS (Huang et al. 2018)	0.294	0.430	0.868	0.351
	DPSNet (Im et al. 2019)	0.154	0.215	0.723	0.226
	IIB (Yifan et al. 2022)	**0.095**	-	0.550	-
	GMDepth (Xu et al. 2023b)	0.101	**0.177**	0.556	**0.167**
	GeoViT	0.106	0.204	**0.508**	0.171
SUN3D	DeMoN (Ummenhofer et al. 2017)	0.214	1.120	2.421	0.206
	DeepMVS (Huang et al. 2018)	0.282	0.435	0.944	0.363
	DPSNet (Im et al. 2019)	0.147	0.107	0.427	0.191
	IIB (Yifan et al. 2022)	0.099	-	**0.293**	-
	GMDepth (Xu et al. 2023b)	0.112	**0.068**	0.336	0.146
	GeoViT	**0.095**	**0.068**	0.552	**0.124**
Scenes11	DeMoN (Ummenhofer et al. 2017)	0.556	3.402	2.603	0.391
	DeepMVS (Huang et al. 2018)	0.210	0.373	0.891	0.270
	DPSNet (Im et al. 2019)	0.056	0.144	0.714	0.140
	IIB (Yifan et al. 2022)	0.056	-	0.523	-
	GMDepth (Xu et al. 2023b)	0.050	0.069	0.491	**0.106**
	GeoViT	0.118	**0.059**	**0.318**	0.146

THANK YOU!