PoseCrafter: Extreme Pose Estimation with Hybrid Video Synthesis

Qing Mao[1,2]* Tianxin Huang[3] Yu Zhu[1] Jinqiu Sun[4] Yanning Zhang[1]† Gim Hee Lee[2]†

[1]School of Computer Science, Northwestern Polytechnical University
[2]School of Computing, National University of Singapore
[3]School of Computing and Data Science, The University of Hong Kong
[4]School of Astronautics, Northwestern Polytechnical University

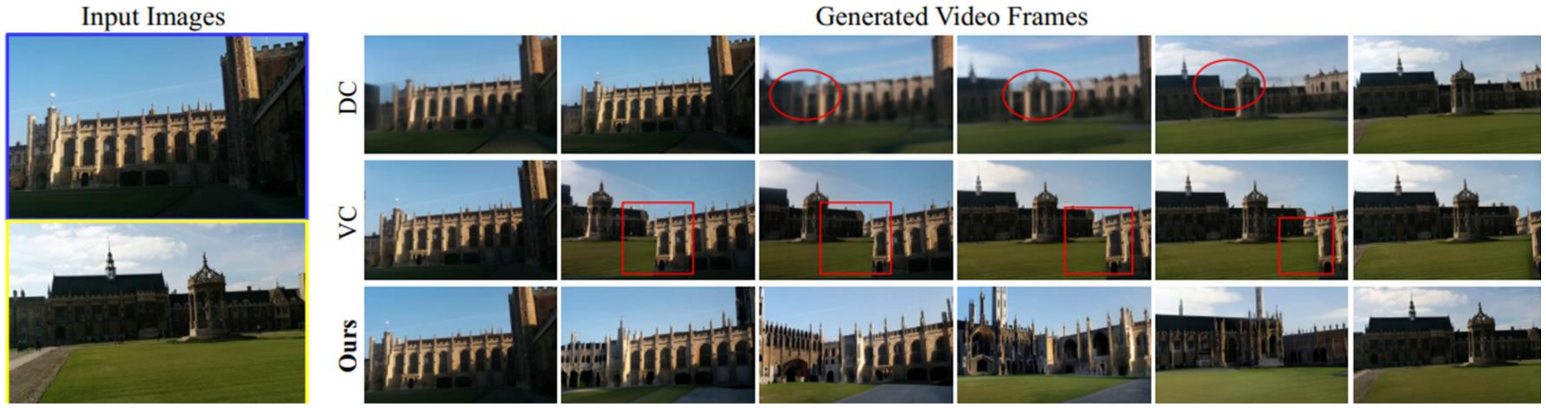
Introduction

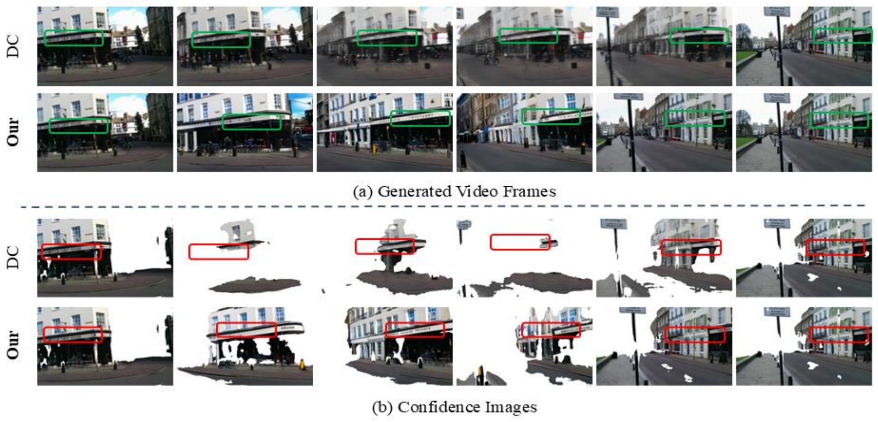
Problem Definition

- Input: Two images captured from different camera poses.
  - ➢ The viewpoint change can be extremely large (strong rotation / large translation)
  - ➢ The two views may have almost no visible overlap
- Output: Relative pose (R and T) between the two views.

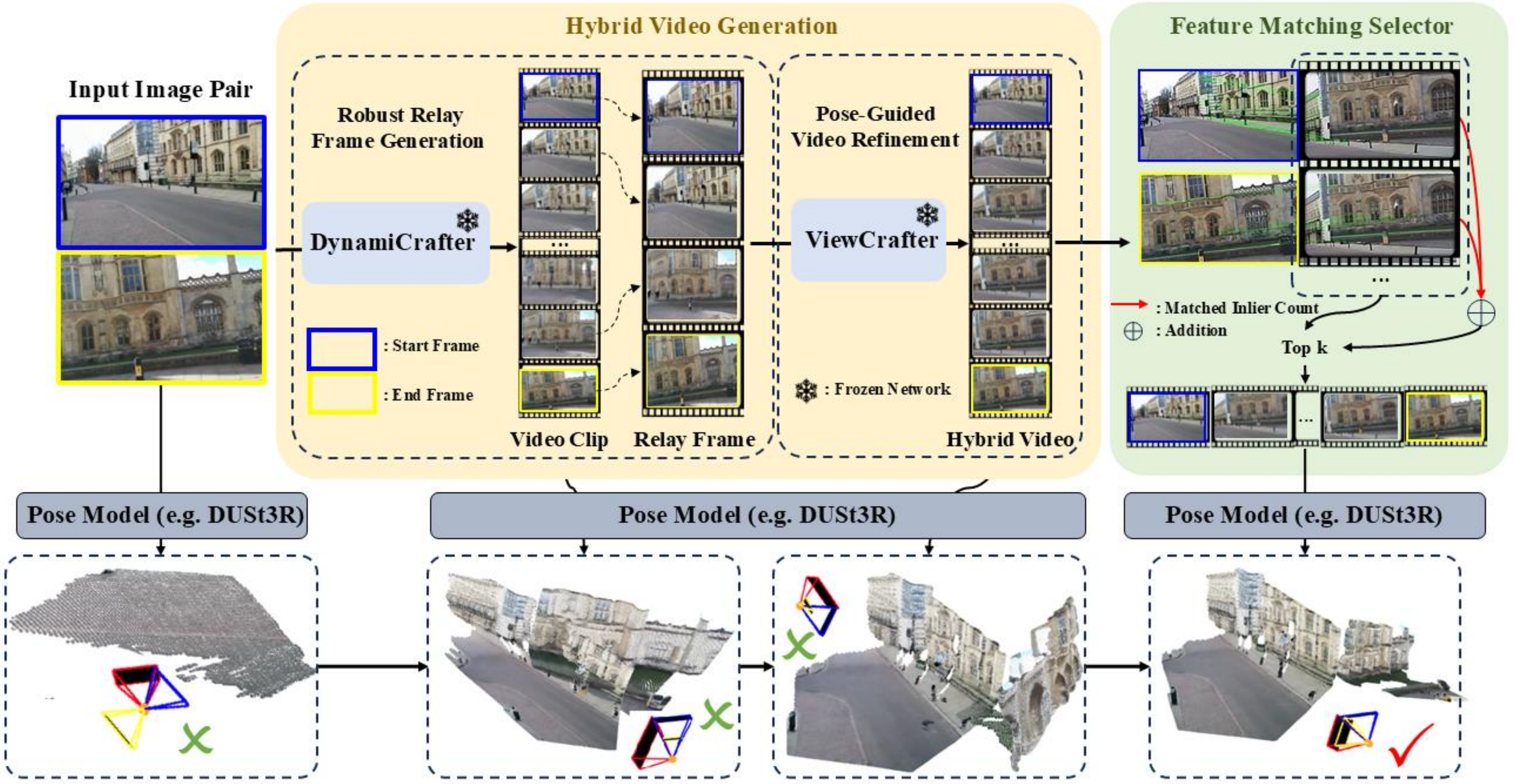
Motivation

- Synthesizing intermediate views via video generation can **increase effective overlap** between image pairs with small or no overlap.
- Current video models under low-overlap inputs **often produce blurry or geometrically inconsistent intermediate frames**.
- Unreliable frames **reduce confidence**, thereby **degrading pose estimation accuracy**.





Methodology



Methodology

Hybrid Video Generation

Problem

- DynamiCrafter is prone to blurring and inconsistencies in the middle, while ViewCrafter can generate clear, high-fidelity frames only if a feasible camera trajectory is used as input.

method

- First, we use DynamiCrafter to synthesize intermediate frames and then select a small set of "**Relay-frames**" frames to ensure a geometrically consistent camera trajectory.
- The selected frames are subsequently provided to ViewCrafter, which generates high-fidelity intermediate views.

Table 1: Relay-frame sampling analysis using mean rotation error (MRE↓). The setting #Frames=2 corresponds to $\{I_0, I_T\}$, #Frames=4 corresponds to $\{I_0, I_1, I_{T-1}, I_T\}$, #Frames=6 corresponds to $\{I_0, I_1, I_2, I_{T-2}, I_{T-1}, I_T\}$, and #Frames=8 corresponds to $\{I_0, I_1, I_2, I_3, I_{T-3}, I_{T-2}, I_{T-1}, I_T\}$. The case #Frames=16 uses all frames. Results indicate that #Frames=4 consistently achieves the lowest MRE and the highest stability across datasets.

Dataset	**#Frames** (n)				
	2	4	6	8	16
Cambridge Landmarks	20.56	**14.47**	16.66	16.87	17.83
ScanNet	19.67	**16.23**	17.03	17.16	18.56
DL3DV-10K	15.22	**14.27**	14.40	14.73	14.52
NAVI	7.78	**6.94**	7.18	9.64	10.92

Methodology

Feature Matching Selector

Problem

- Although the HVG stage produces high-fidelity candidate intermediate frames, **not all of these frames are beneficial for pose estimation**, and retaining too many of them incurs unnecessary computational cost.

method

- From the candidate intermediate frames generated by HVG, we apply a **feature matching selector** to identify the frames that are **most favorable for accurate pose estimation**.

①**Input:** a set of candidate frames $\{I_t\}$ together with the start and end frames I_0, I_T .

②**Scoring:** each candidate frame is independently matched to I_0, I_T in feature space, and compute the number of RANSAC inliers.

$$S(t) = N_0(t) + N_T(t)$$

③**Output:** The top k frames (k=6) are then forwarded to the pose estimation model (DUSt3R).

Table 10: Ablation study on the number of intermediate frames selected by FMS.

#Frames	MRE↓	R@5°	R@15°	R@30°	AUC_{30} ↑
4	11.36	55.90	**89.93**	93.40	77.59
6	**11.40**	55.21	**89.93**	**93.75**	**77.41**
8	11.93	**55.90**	**89.93**	92.10	76.23

Experiments

Comparison with State-of-the-Art: Quantitative Comparison on PoseCrafter

Table 2: Pose estimation on Cambridge Landmarks. We report rotation recall (R@ θ ↑), translation recall (T@ θ ↑), mean rotation error (MRE↓), and AUC_{30} ↑.

Method	Input	Yaw range [50°-65°]					Yaw range [65°-90°]				
		MRE↓	R@5°	R@15°	R@30°	AUC_{30} ↑	MRE↓	R@5°	R@15°	R@30	AUC_{30} ↑
DUSt3R	Pair	18.14	40.34	71.25	82.99	61.98	51.24	21.67	44.67	51.67	37.93
$\text{InterPose}^{\ddagger}_{\text{w/o SCS}}$	DynamiCrafter	16.11	42.70	75.70	87.35	65.72	42.51	30.67	42.51	61.33	47.18
$\text{InterPose}^{\ddagger}$	DynamiCrafter	13.61	51.81	81.50	83.30	70.47	38.87	36.33	65.67	68.33	55.24
$\text{Ours}_{\text{w/o FMS}}$	Hybrid video	13.24	54.51	89.24	92.71	76.13	34.87	31.33	68.33	77.67	56.29
Ours	Hybrid video	**11.40**	**55.21**	**89.93**	**93.75**	**77.41**	**29.02**	**36.67**	**71.67**	**78.33**	**60.46**

Table 3: Pose estimation on ScanNet. We report rotation recall (R@ θ ↑), translation recall (T@ θ ↑), mean rotation error (MRE↓), mean translation error (MTE↓), and AUC_{30} ↑.

Yaw range	Method	Input	R@5°	R@15°	R@30°	T@5°	T@15°	T@30°	MRE↓	MTE↓	AUC30↑
50°-65°	DUSt3R	Pair	43.97	74.14	79.31	25.34	52.07	**78.45**	19.41	25.23	47.37
	$\text{InterPose}^{\ddagger}_{\text{w/o SCS}}$	DynamicCrafter	46.55	77.59	85.34	16.38	48.10	64.74	17.51	35.25	42.69
	$\text{InterPose}^{\ddagger}$	DynamicCrafter	50.86	81.03	87.07	27.58	61.21	69.46	15.15	23.89	53.33
	$\text{Ours}_{\text{w/o FMS}}$	Hybrid video	51.72	87.07	93.10	23.28	50.62	67.07	12.38	29.02	45.53
	Ours	Hybrid video	**53.45**	**88.79**	**94.83**	**33.62**	**65.52**	77.69	**10.77**	**22.14**	**57.03**
65°-90°	DUSt3R	Pair	42.05	67.05	70.45	26.59	46.59	53.86	30.82	29.99	36.50
	$\text{InterPose}^{\ddagger}_{\text{w/o SCS}}$	DynamicCrafter	38.64	62.50	65.90	20.45	39.77	47.72	35.18	58.89	33.40
	InterPose	DynamicCrafter	45.45	67.05	71.59	31.81	53.41	64.77	28.22	29.52	45.98
	$\text{Ours}_{\text{w/o FMS}}$	Hybrid video	46.59	76.14	82.95	23.85	48.86	57.95	22.61	35.98	41.72
	Ours	Hybrid video	**50.00**	**77.72**	**84.09**	**37.50**	**63.64**	**73.86**	**17.02**	**29.28**	**56.44**

Experiments

Comparison with State-of-the-Art: Quantitative Comparison on PoseCrafter

Table 4: Pose estimation on DL3DV-10K with [50°-90°] yaw range.

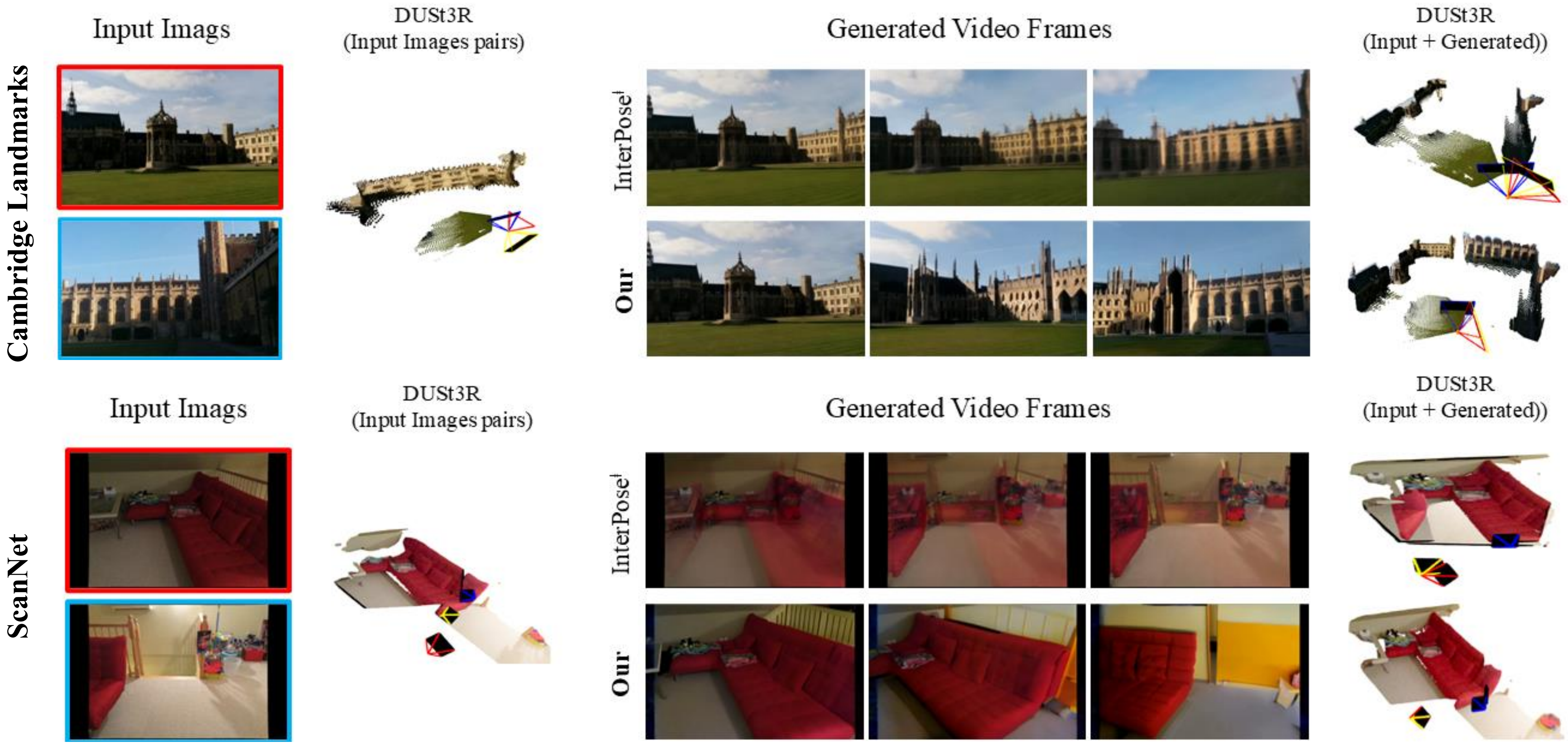
Method	Input	R@5°	R@15°	R@30°	T@5°	T@15°	T@30°	MRE↓	MTE↓	AUC_{30} ↑
DUSt3R	Pair	34.33	63.00	94.66	27.00	75.00	92.67	13.36	10.88	55.58
$\text{InterPose}^{\ddagger}_{\text{w/o SCS}}$	DynamicCrafter	36.33	64.33	95.00	26.00	76.33	92.67	13.32	11.27	55.68
$\text{InterPose}^{\ddagger}$	DynamicCrafter	36.11	64.33	97.66	27.66	79.67	95.33	13.17	10.76	56.05
$\text{Ours}_{\text{w/o FMS}}$	Hybrid video	38.33	68.33	98.33	30.66	79.61	96.67	12.89	10.71	57.16
Ours	Hybrid video	**38.10**	**70.00**	**100.00**	**31.33**	**81.33**	**98.33**	**12.73**	**10.28**	**57.48**

Table 5: Pose estimation on NAVI for the [50°-90°] yaw range.

Method	Input	R@5°	R@15°	R@30°	T@5°	T@15°	T@30°	MRE↓	MTE↓	AUC_{30} ↑
DUSt3R	Pair	64.69	95.72	98.05	62.37	97.28	98.22	7.30	7.82	82.37
$\text{InterPose}^{\ddagger}_{\text{w/o SCS}}$	DynamicCrafter	45.13	92.61	96.11	57.86	91.44	96.50	11.14	8.81	78.63
$\text{InterPose}^{\ddagger}$	DynamicCrafter	66.53	97.28	**98.83**	67.70	96.89	98.84	6.61	6.26	82.80
$\text{Ours}_{\text{w/o FMS}}$	Hybrid video	59.53	97.28	**98.83**	72.26	95.72	98.83	6.93	6.87	81.91
Ours	Hybrid video	**70.82**	**97.67**	**98.83**	**75.10**	**98.44**	**99.22**	**5.97**	**5.46**	**83.98**

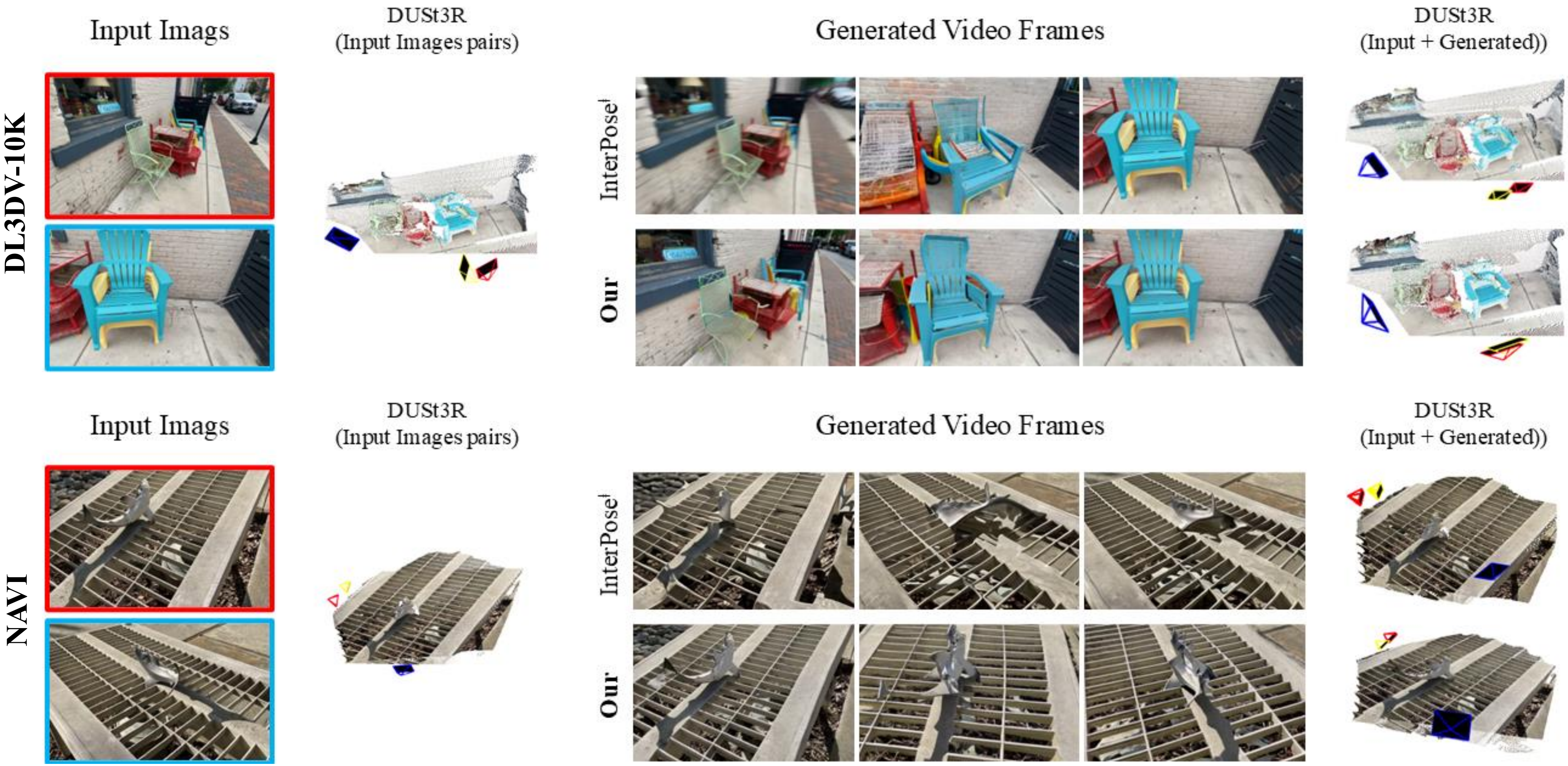
Experiments

Comparison with State-of-the-Art: Qualitative Comparison on PoseCrafter



Experiments

Comparison with State-of-the-Art: Qualitative Comparison on PoseCrafter



Experiments

Runtime and Memory Cost

Table 6: Runtime and Memory Cost.

Method	Runtime		Memory Cost	
	Video Generation	Pose Estimation	Video Generation	Pose Estimation
InterPose‡	3.2min	20.29min	14.6GB	3.1GB
Ours	3.8min	0.18min	22.8GB	3.6GB

Ablation Study

Table 7: Ablation study on Hybrid Video Generation. SCS and FMS denote the frame selection strategies from InterPose [5] and ours, respectively.

Method	Input	MRE↓	R@5°	R@15°	R@30°	AUC_{30} ↑
DUSt3R	Pair	18.14	40.34	71.25	82.99	61.98
InterPose$^{\ddagger}_{\text{w/o SCS}}$	DynamicCrafter	16.11	42.70	75.70	87.50	65.72
InterPose‡	DynamicCrafter	13.60	51.81	81.50	83.30	70.47
InterPose$^{\ddagger}_{\text{w/ FMS}}$	DynamicCrafter	13.02	52.08	85.76	90.63	73.93
ViewCrafter$_{\text{w/o FMS}}$	ViewCrafter	13.80	52.78	82.29	88.54	71.12
ViewCrafter$_{\text{w/ FMS}}$	ViewCrafter	12.45	53.82	84.03	90.28	72.82
Ours$_{\text{w/o FMS}}$	Hybrid video	13.24	54.51	89.24	92.71	76.13
Ours$_{\text{w/ SCS}}$	Hybrid video	12.11	54.86	88.54	91.32	76.11
Ours	Hybrid video	**11.40**	**55.21**	**89.93**	**93.75**	**77.41**

Thanks for watching!