DMAligner

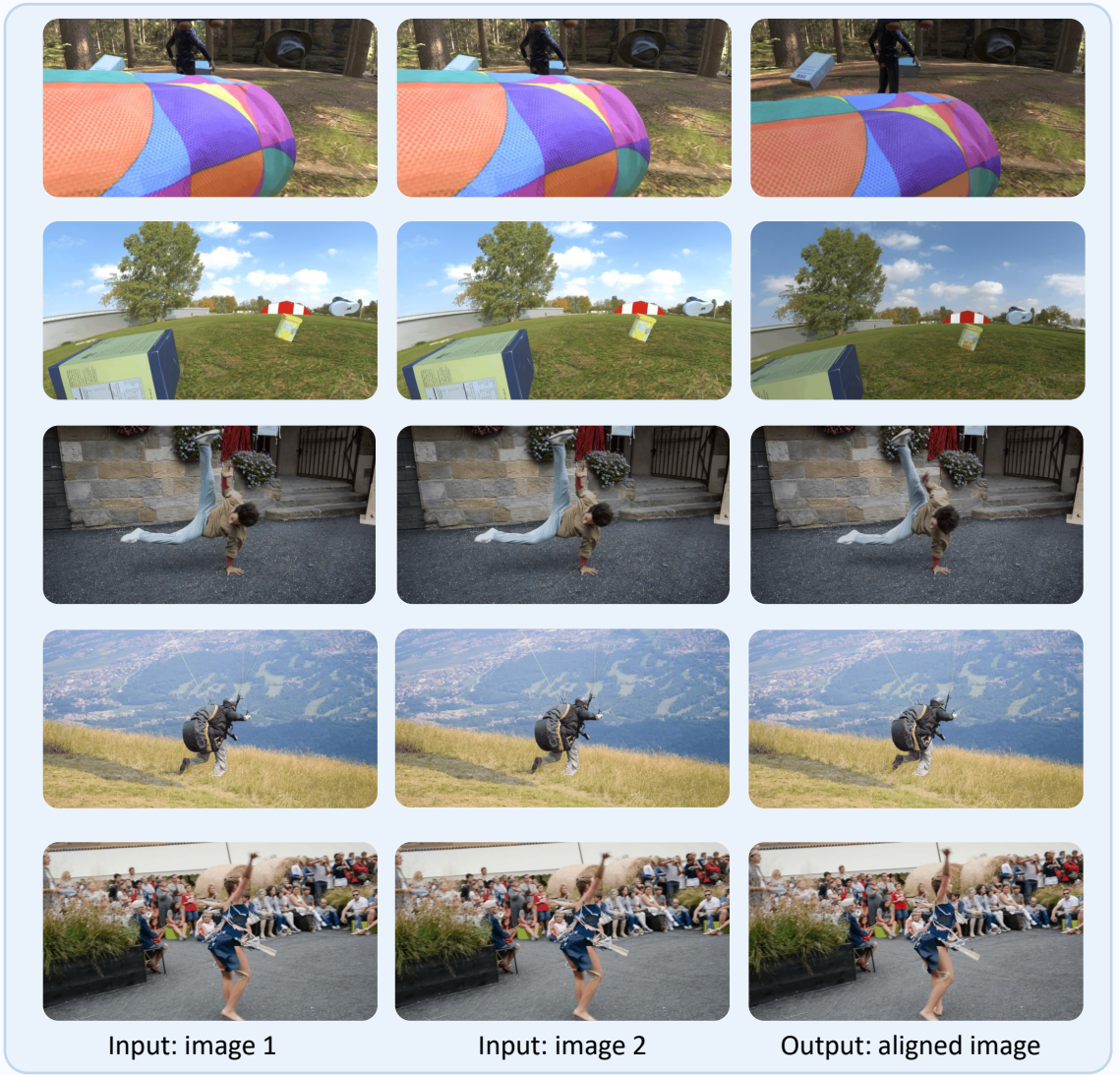
Enhancing Image Alignment via Diffusion Model Based View Synthesis

From optical-flow warping to diffusion-based view synthesis

CVPR 2026 ***Highlight***

Xinglong Luo, Ao Luo, Zhengning Wang, Yueqi Yang, Chaoyu Feng, Lei Lei, Bing Zeng, Shuaicheng Liu

Project page: github.com/boomluo02/DMAligner

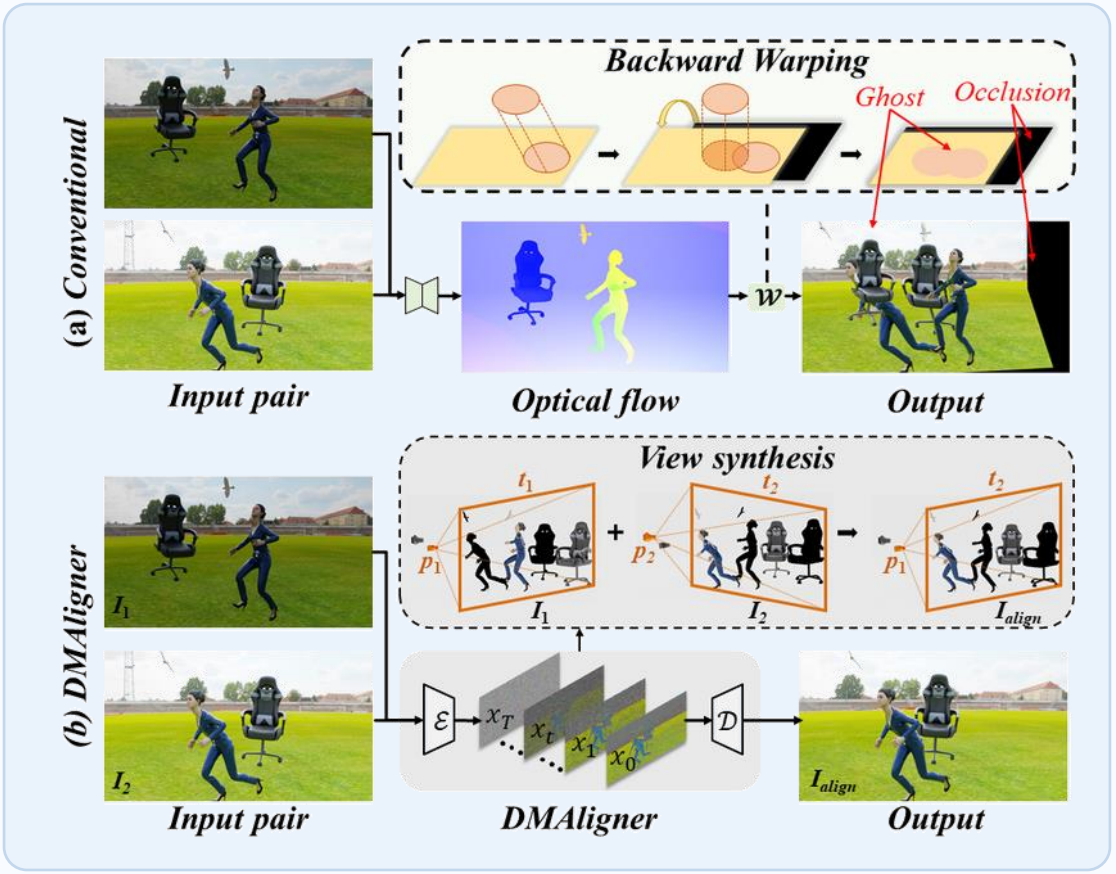


Motivation

Why is image alignment still hard?

Classic pipeline: optical flow + backward warping

DMAligner: synthesize the aligned image directly



Why warping breaks

- Occlusions create many-to-one mapping and ghosting.
- Lighting changes break brightness constancy.
- Large or non-rigid motion amplifies artifacts.
- The errors hurt HDR, denoising, and super-resolution.

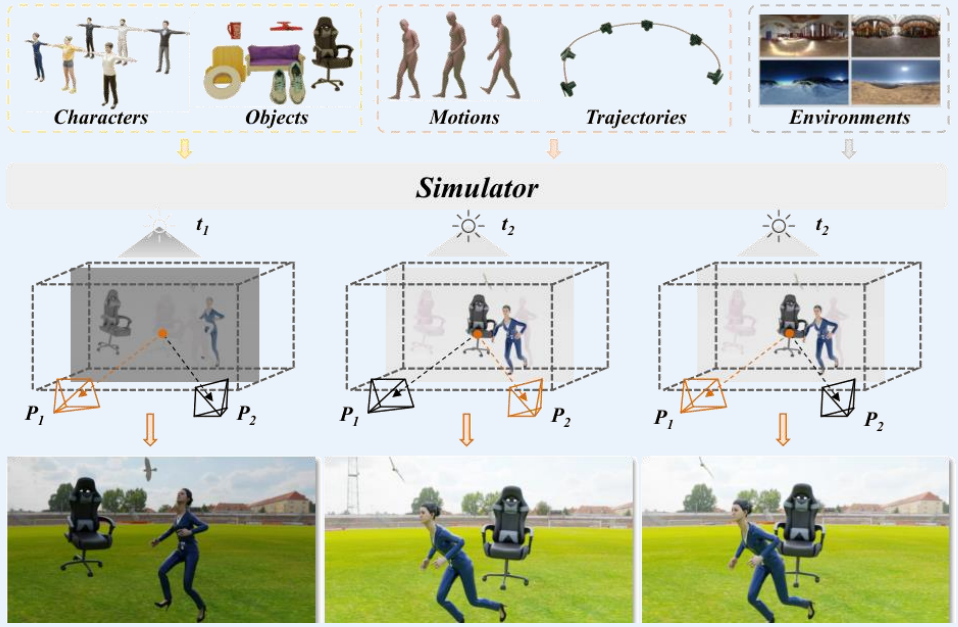
Key shift

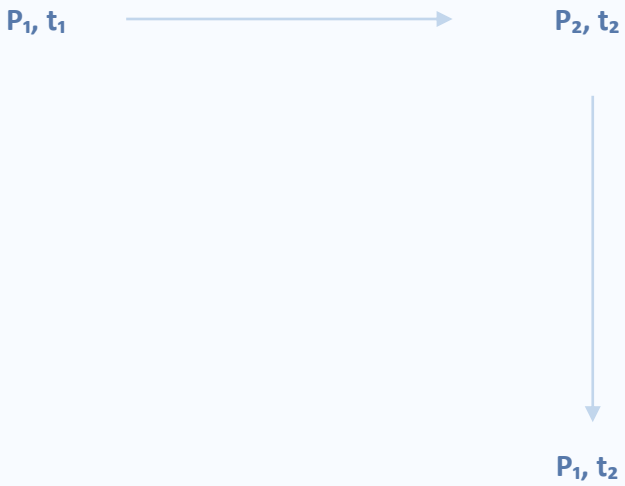
Instead of moving pixels, predict what frame t_2 should look like from camera pose p_1.

Key Idea

Alignment-oriented view synthesis

Given $I_1(p_1, t_1)$ and $I_2(p_2, t_2)$, we generate $I\_align(p_1, t_2)$ instead of estimating optical flow.





- Keep the pose and background geometry of I_1
- Borrow the dynamic foreground and timing from I_2

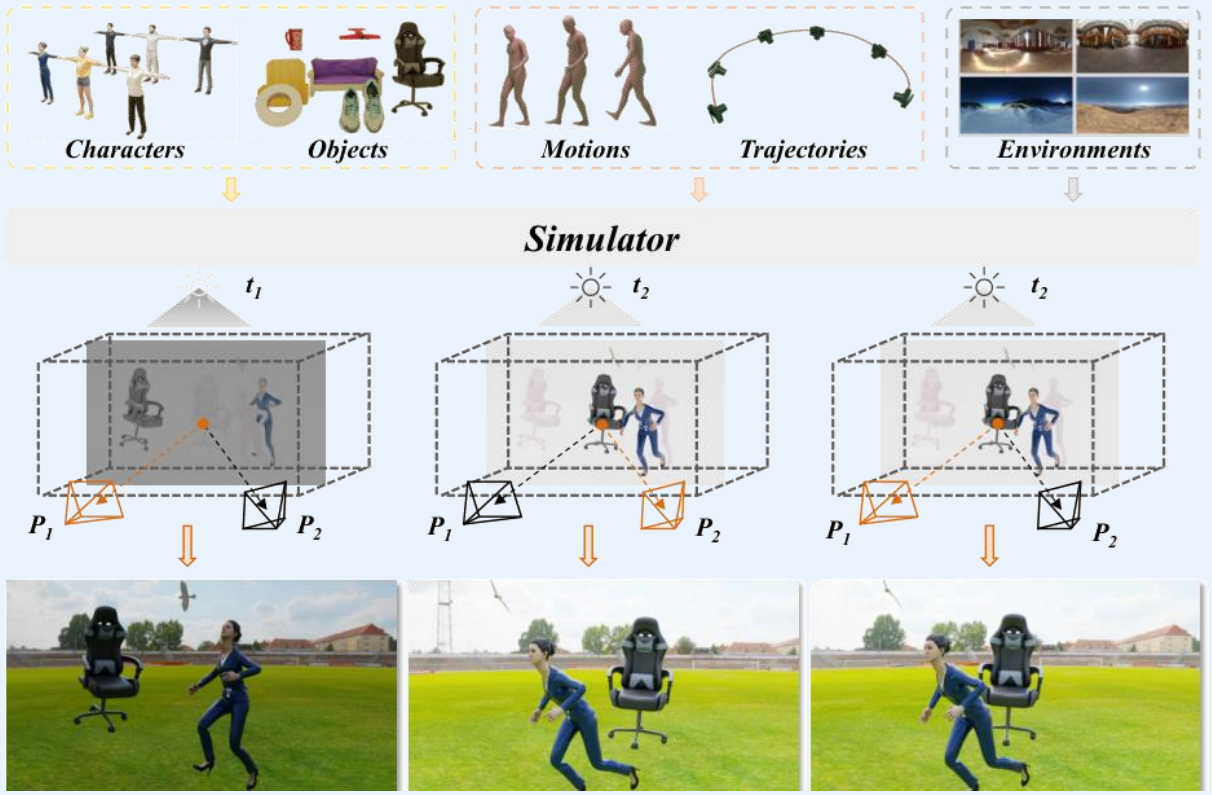
Generative alignment

- Fill occlusions
- Restore dynamic regions
- Avoid warping artifacts

In short: conditional image generation for alignment

Dataset

DSIA: built for dynamic scene alignment



1,033
indoor / outdoor scenes

>30K
training image pairs

960×540
rendered resolution

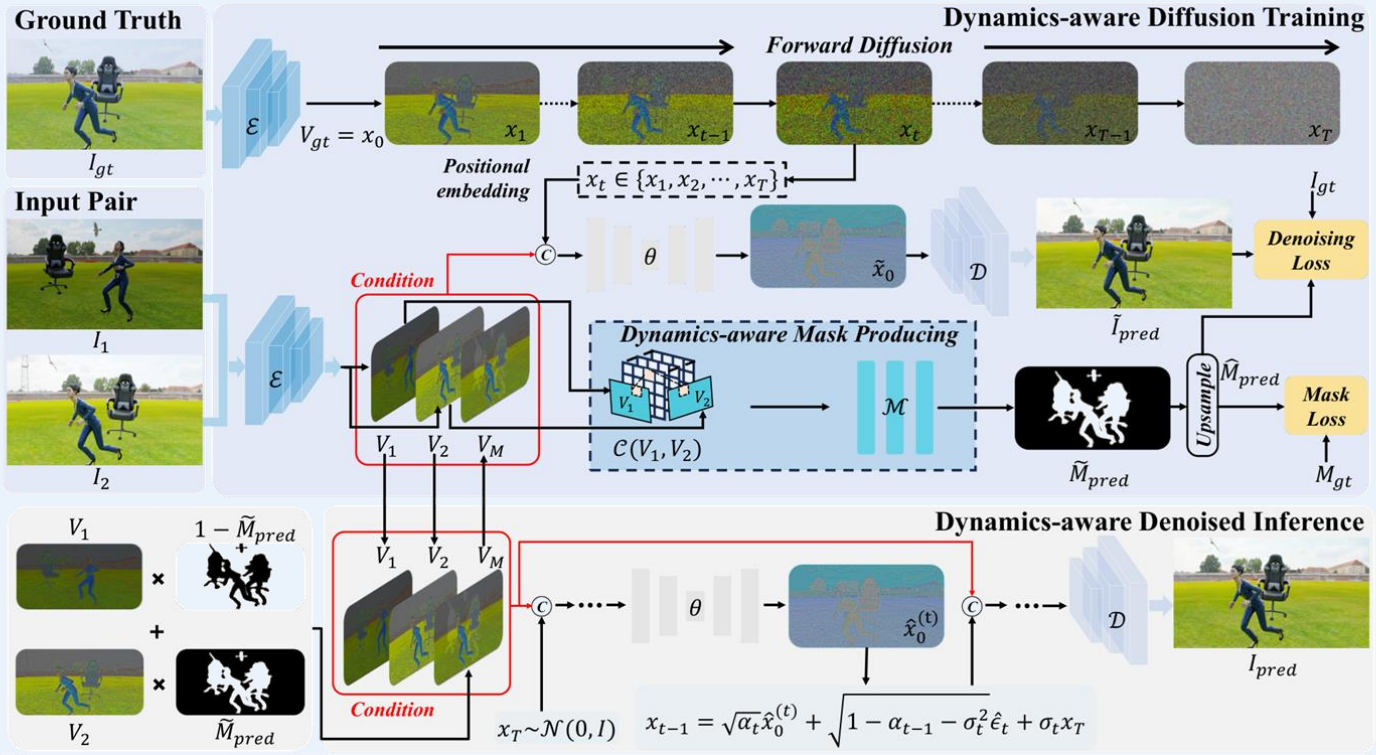
What it simulates

- Camera motion: sample P_1 and P_2 from trajectories
- Dynamic foreground: people, objects, and actions move together
- Scene variation: HDR lighting, materials, textures, and environments
- Supervision: directly render I_gt at (P_1, t_2)

The test split covers four motion settings: LcLf / LcSf / ScLf / ScSf.

Method

Dynamics-aware diffusion training



Three key pieces

1 Latent diffusion

Run diffusion in VQ-VAE latent space and predict x_0 for direct reconstruction.

2 DMP motion mask

Use the cost volume of V_1 and V_2 to find dynamic foreground regions.

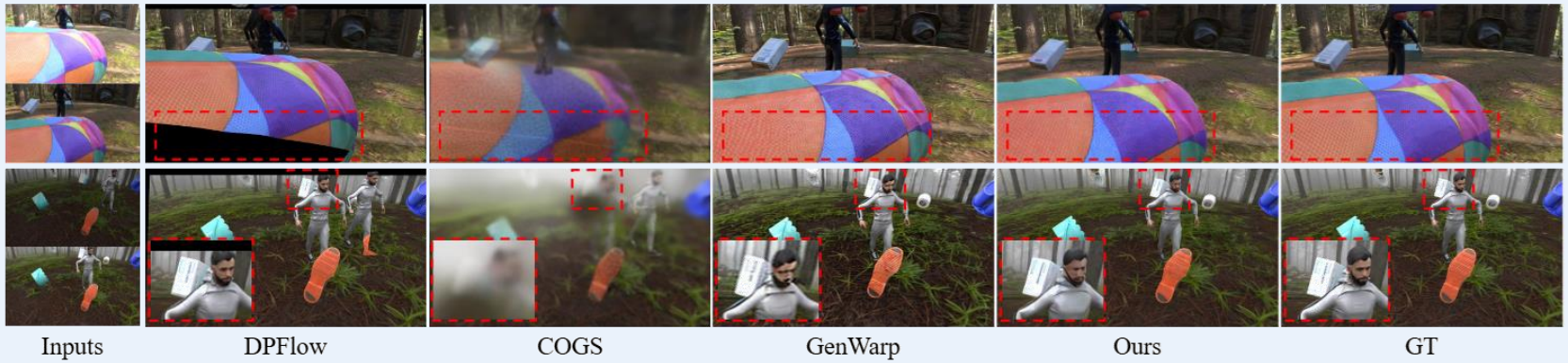
3 VM conditioning

Use V_1 for static areas and V_2 for dynamic areas, then fuse them as the condition.

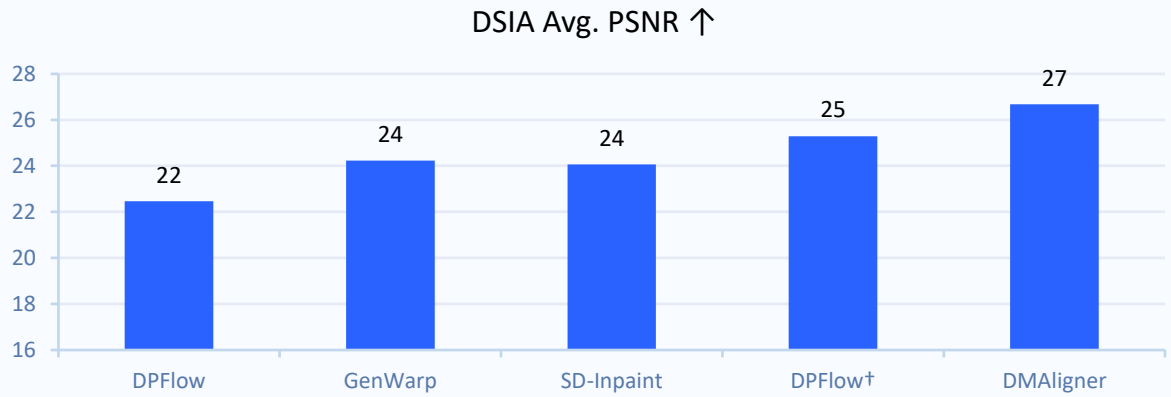
Inference: start from Gaussian noise and denoise into I_pred.

Results

Better quality on DSIA



Qualitative: closer to GT, with less ghosting and fewer occlusion artifacts



+1.38 dB

vs. the best flow-warping baseline†

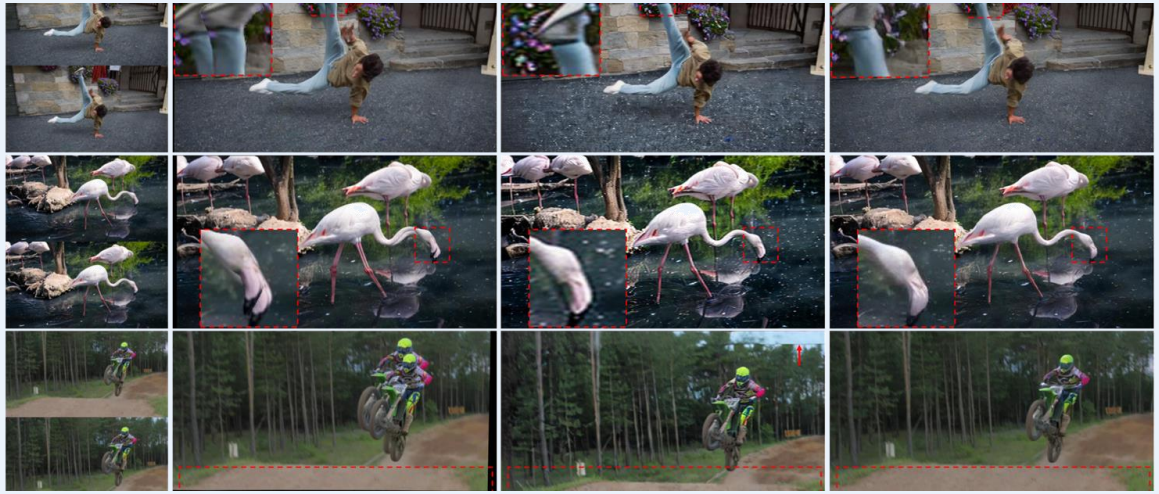
26.67 / 0.81

Avg. PSNR / SSIM

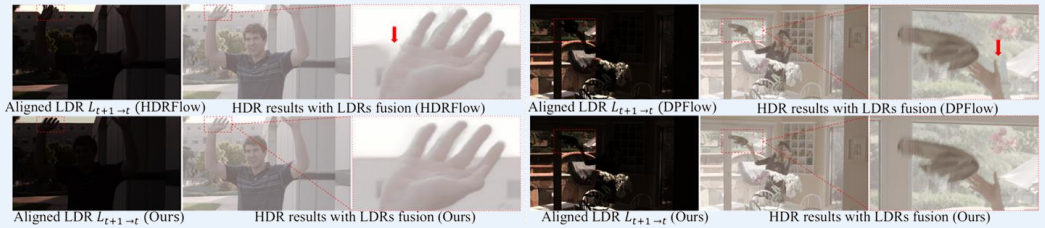
Ablation: DMP lifts Avg. PSNR from 25.12 to 26.67. Predicting x_0 works best.

Generalization

Cross-domain transfer and HDR



DAVIS real-world examples



HDR fusion: just swap the alignment module

Avg. LPIPS / DreamSim ↓

DPFlow: 0.214 / 0.109

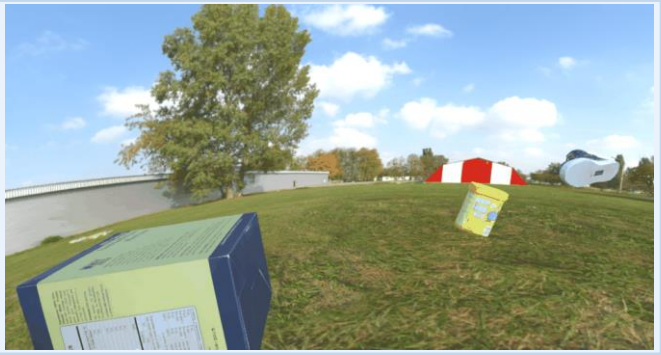
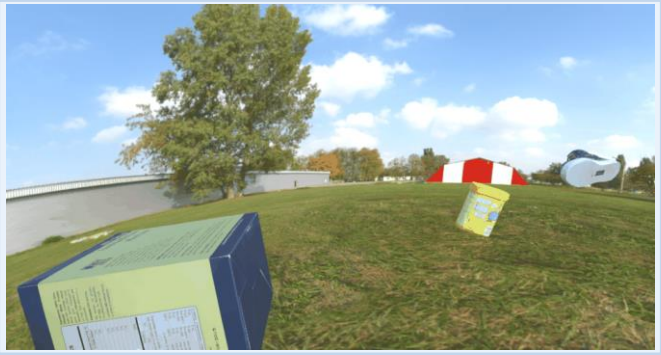
GenWarp: 0.250 / 0.130

DMAligner: 0.211 / 0.108

Trained on DSIA, then tested directly on Sintel and DAVIS — still the best or tied-best perceptual distance.

Demo Results

Five qualitative demos on real alignment cases (1/2)

	Input overlay	Input 1 vs prediction	Input 2 vs prediction
P1			
P2			
P3			

Animated GIFs will play in presentation mode.

Demo Results

Five qualitative demos on real alignment cases (2/2)

	Input overlay	Input 1 vs prediction	Input 2 vs prediction
P4			
P5			

Animated GIFs will play in presentation mode.

What to look for
- Stable background structure
- Clean transfer of moving objects
- Fewer artifacts near occlusion boundaries

Takeaways

Why DMAligner matters

1 **A new paradigm**

image warping → alignment-oriented view synthesis

2 **Dynamics-aware diffusion**

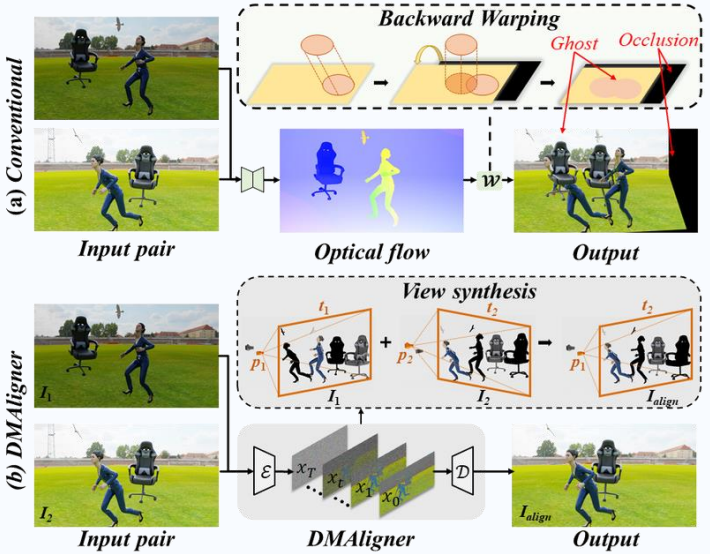
DMP highlights moving regions and helps restore hard areas.

3 **DSIA dataset**

Large-scale supervision for dynamic alignment and clear benchmarking.

One-line message:

When warping struggles with occlusion and motion, don't just move pixels — generate the aligned view.



Project page: github.com/boomluo02/DMAligner

Code + paper on the project page