SpatialDreamer: Self-supervised Stereo Video Synthesis from Monocular

Zhen Lv, Yangqi Long, Congzhentao Huang, Cao Li, Chengfei Lv†, Hao Ren, Dian Zheng
Alibaba Group, Hangzhou, China

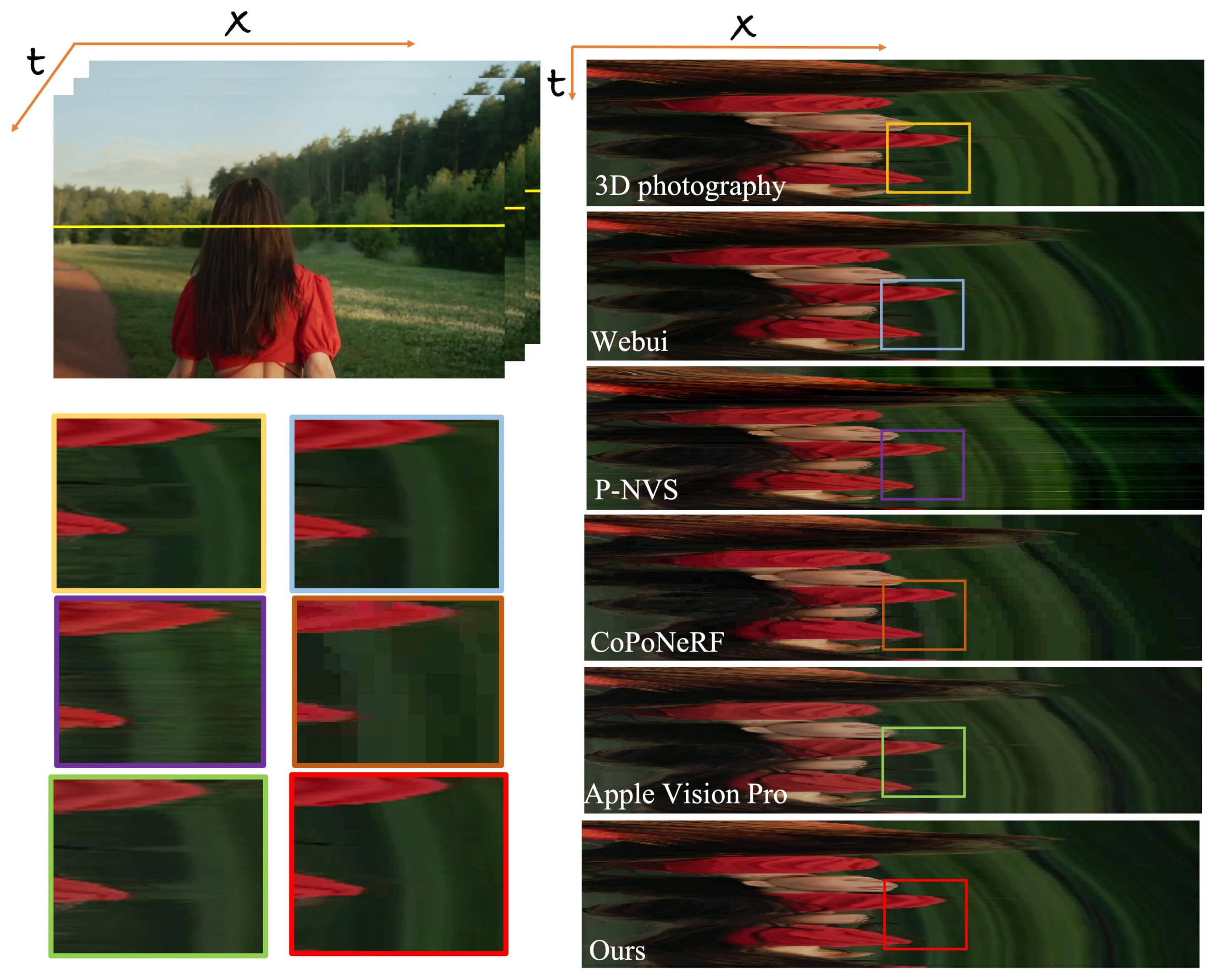
Introduction

Motivation

- The lack of high-quality stereo video pairs for training and the difficulty of maintaining spatio-temporal consistency between frames.
- Existing methods primarily address these issues by directly applying novel view synthesis (NVS) techniques to video, while facing limitations such as the inability to effectively represent dynamic scenes and the requirement for extensive training data.

Contribution

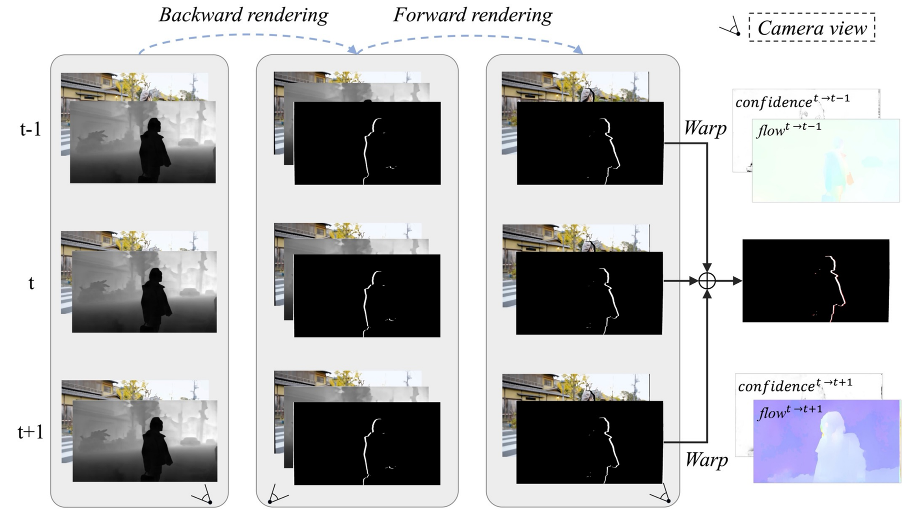
- A novel self-supervised stereo video synthesis framework, SpatialDreamer, is proposed, which is robust across a wide range of scenes and dynamic content in video.
- A consistency control module is devised, which consists of a metric of stereo deviation strength and a temporal interaction learning module, ensuring geometric and temporal consistency in video.
- The results demonstrate that the proposed method outperforms the state-of-the-art methods, even beats AVP.



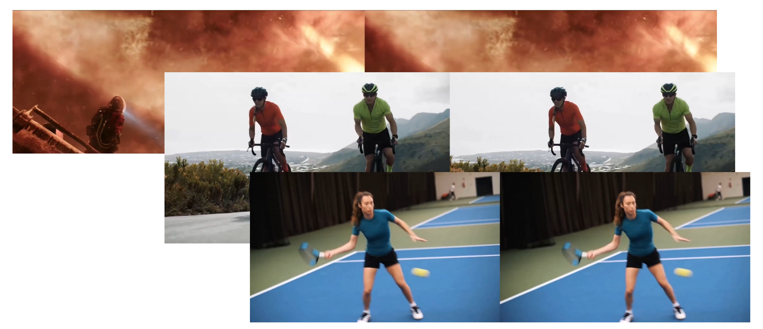
We extract the yellow line of each frame and stack them together. A good result should show a natural transition in the t dimension.

Depth based Video Generation

Equipping the DVG module, we build a stereo video dataset using a self-supervised approach.



- ✓ 1500+ monocular videos
- ✓ Including TikTok, movie, and TV show clips with fast motion and occlusion



Method

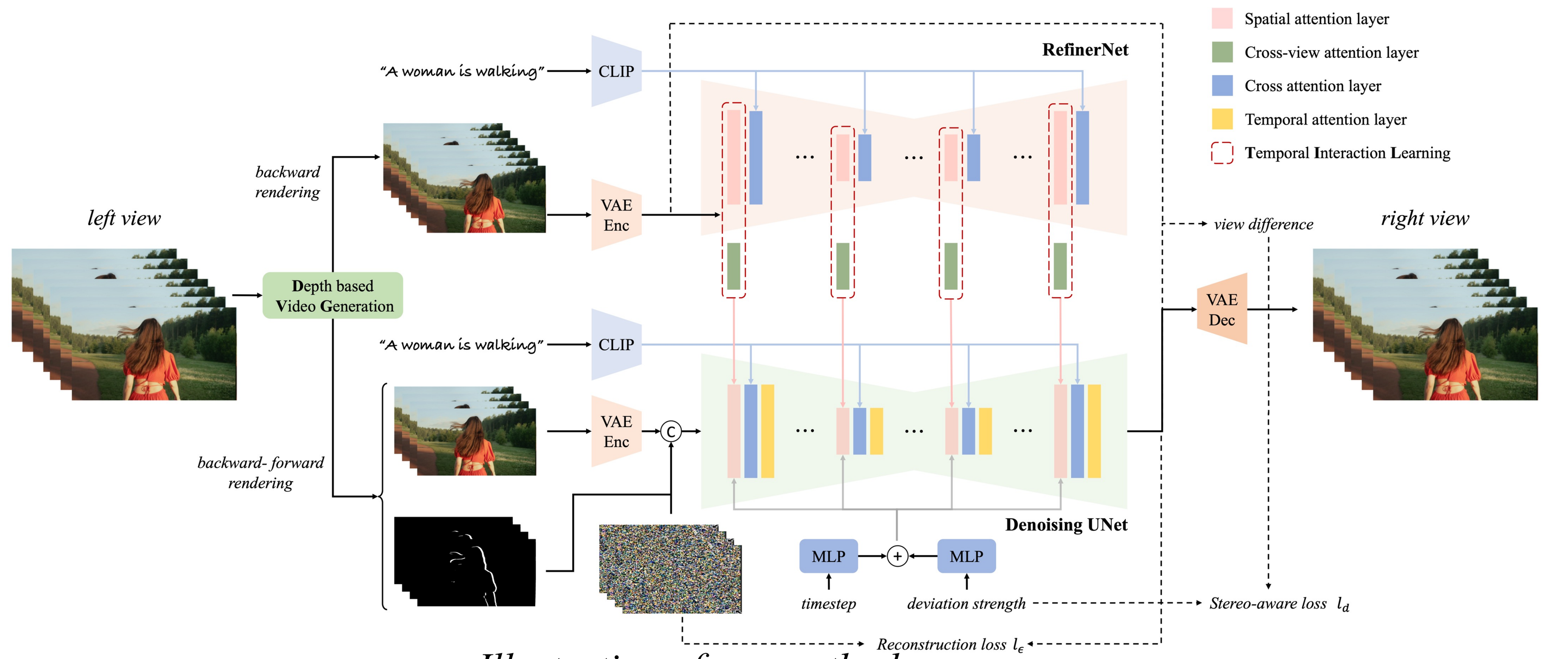
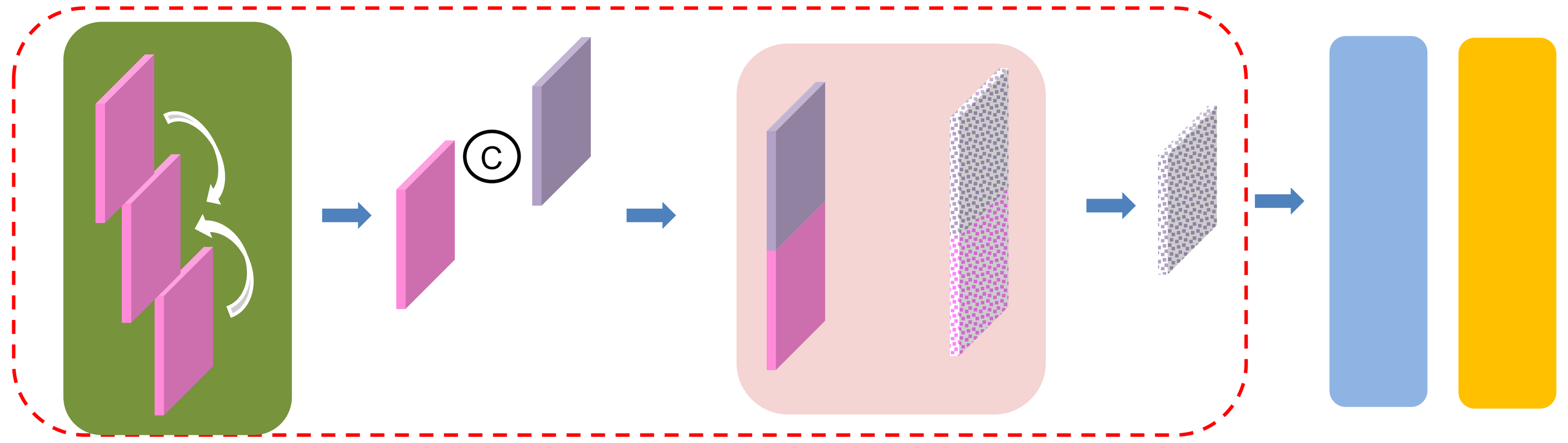


Illustration of our method



Consistency Control Module

Stereo Deviation Strength

$$s(z) = |z_0 - z_{ref}|$$

- To quantify the latent differences between reference view z_{ref} and target view z_0.
- Stereo Aware Loss $\quad l(d) = \|s(z_0) - s(\hat{z}_0)\|_2^2$

Stereo deviation strength guidance

Temporal Interaction Learning

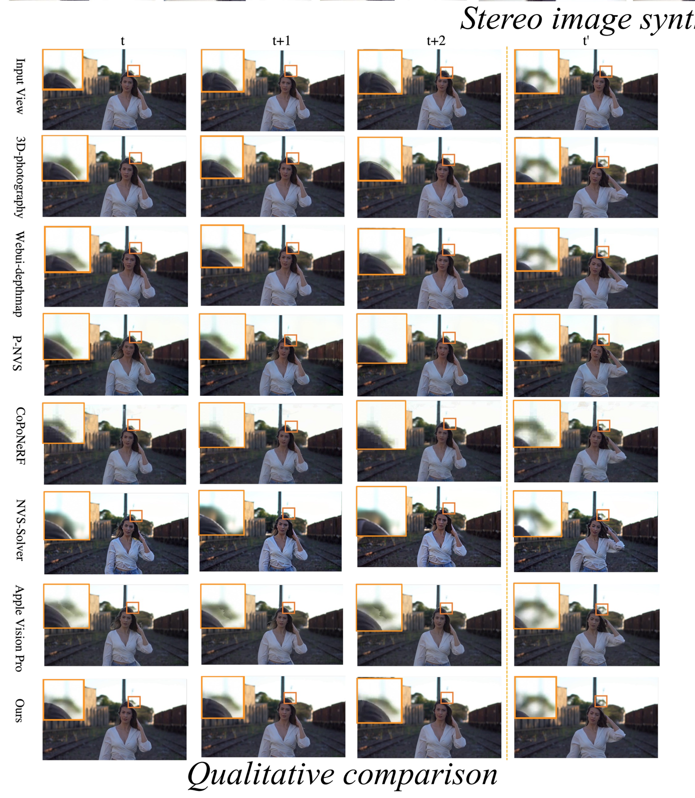
- Blending self-attention of z_r^t in RefinerNet's U-Net block with the cross-view attention between z_r^t and each adjacent view z_i^t.

$$aug_r^t = \lambda \cdot Attn_{r,r} + (1-\lambda) \cdot \frac{1}{N_r}\sum_{i=1}^{N_r} Attn_{r,i}$$

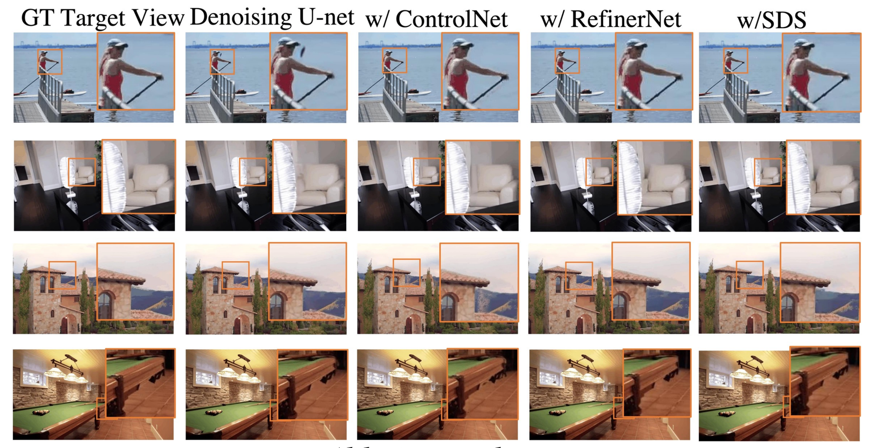
- The augmented reference feature $aug_r^t \in \mathbb{R}^{t\times h\times w\times c}$ is then fed into spatial attention layer to assist U-Net network learning :
  - Concatenated along the h dimension.
  - Self-attention is applied.
  - The first half of the feature map is retrieved as the output.

Experiments

Stereo image synthesis on the RealEstate10K dataset



Qualitative comparison



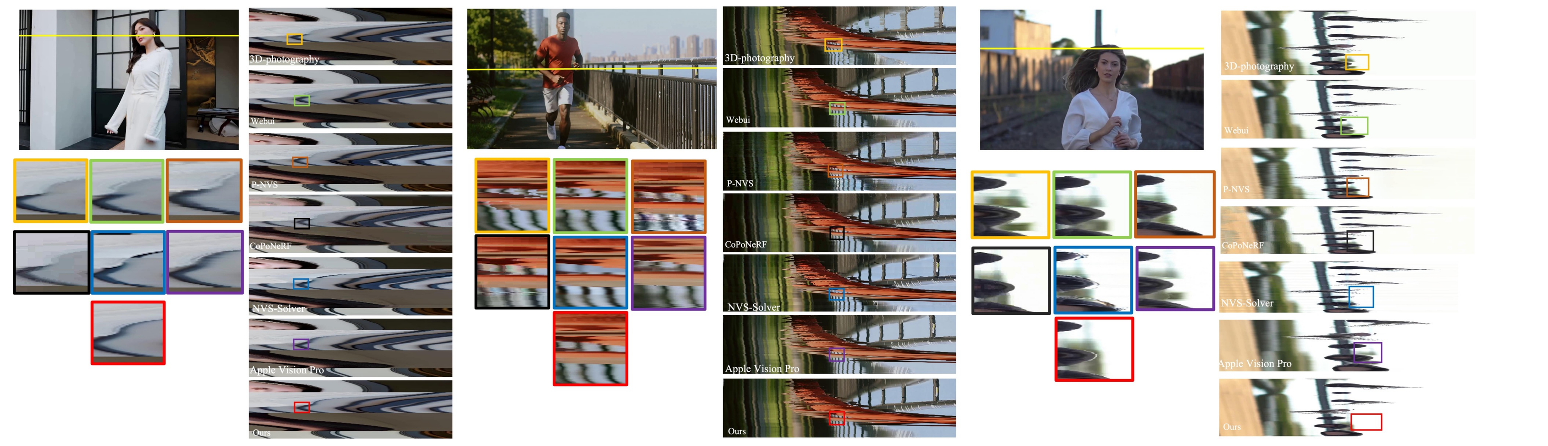
Ablation study

	SSIM ↑		PSNR ↑		LPIPS ↓	
	t=5	t=10	t=5	t=10	t=5	t=10
3D-photography [54]†	0.641	0.446	11.08	9.533	0.116	0.386
SynSin [67]†	0.740	0.645	21.22	19.55	0.087	0.104
AdaMPI [21]†	0.869	0.736	27.70	22.65	0.041	0.087
SinMPI [45]†	0.898	0.775	28.79	23.21	0.040	0.080
Wang et al. [64]	0.840	0.720	25.35	21.36	0.049	0.095
P-NVS [73]†	0.724	0.667	21.98	19.46	0.241	0.279
NVSVDE-Net [4]	0.840		24.31		0.233	
CoPoNeRF [24]†	0.658	0.632	22.63	21.33	0.171	0.209
MVSplat [14]†	0.863		25.89		0.132	
Proposed	**0.916**	**0.857**	**32.26**	**24.86**	**0.038**	**0.049**

	FVD ↓	E^*_{warp} ↓
3D-photography [54]	155.0	3.418
Webui-depthmap	83.18	3.486
P-NVS [73]	420.2	23.02
CoPoNeRF [24]	290.0	7.144
NVS-Solver [72]	249.1	5.842
AVP [42]	99.92	3.446
Proposed	**67.09**	**3.374**

Quantitative comparison

Data improvement by DVG



Stereo video synthesis in stereo video benchmark

Contact US

Website
JD
Emails
lvzhen.lz@alibaba-inc.com
licao.lc@ alibaba-inc.com

JOIN US