Neural Video Depth Stabilizer

Yiran Wang[1] Min Shi[1] Jiaqi Li[1] Zihao Huang[1]
Zhiguo Cao[1] Jianming Zhang[2] Ke Xian[3] Guosheng Lin[3]
[1]Huazhong University of Science and Technology [2]Adobe Research [3]Nanyang Technological University

ICCV23 PARIS

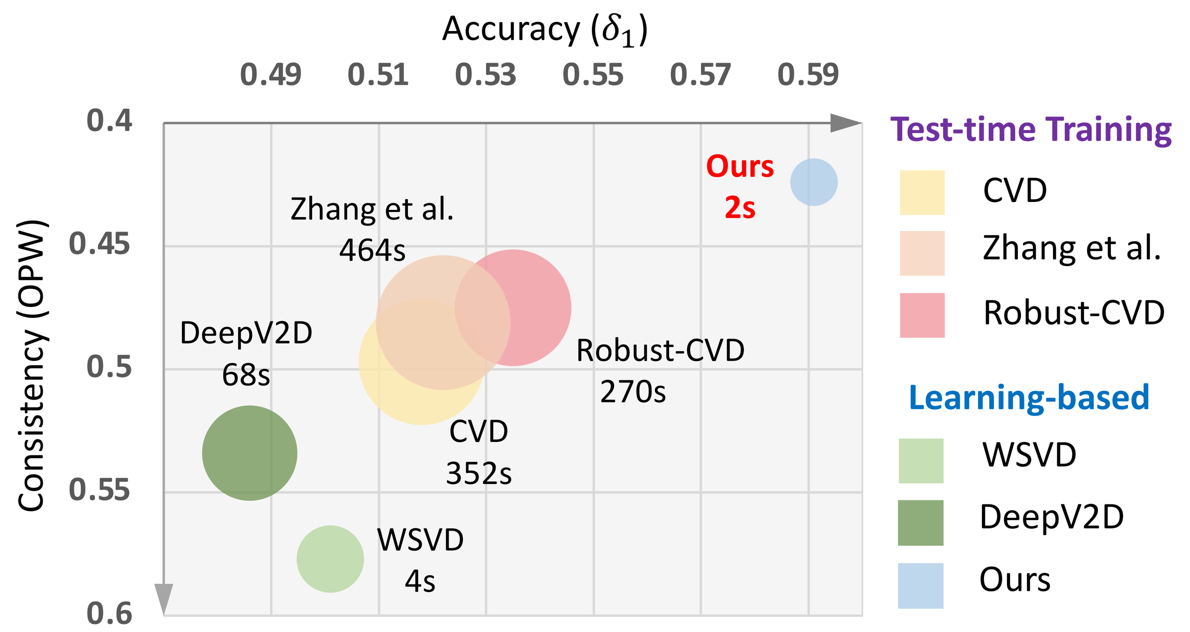
Problem Statements

Goal: Monocular video depth estimation is a prerequisite for various video applications, e.g., bokeh rendering 2D-to-3D video conversion, and novel view synthesis. An ideal video depth model should output depth results with both spatial accuracy and temporal consistency.

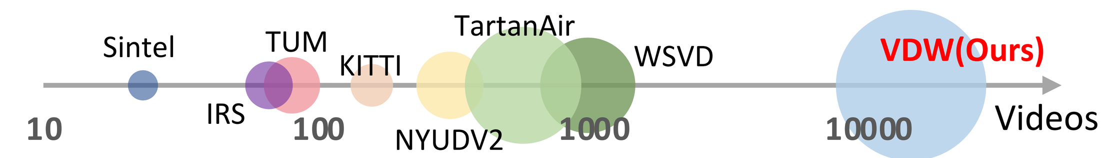
Motivation:

- The prevailing video depth approaches require test-time training (TTT) with pose estimations, leading to limited robustness and heavy computation overhead.
- Learning-based paradigm requires proper model design and sufficient data. Previous learning-based methods show worse performance than TTT-based ones.
- Video depth data is also limited in scale and diversity.

Key Contributions



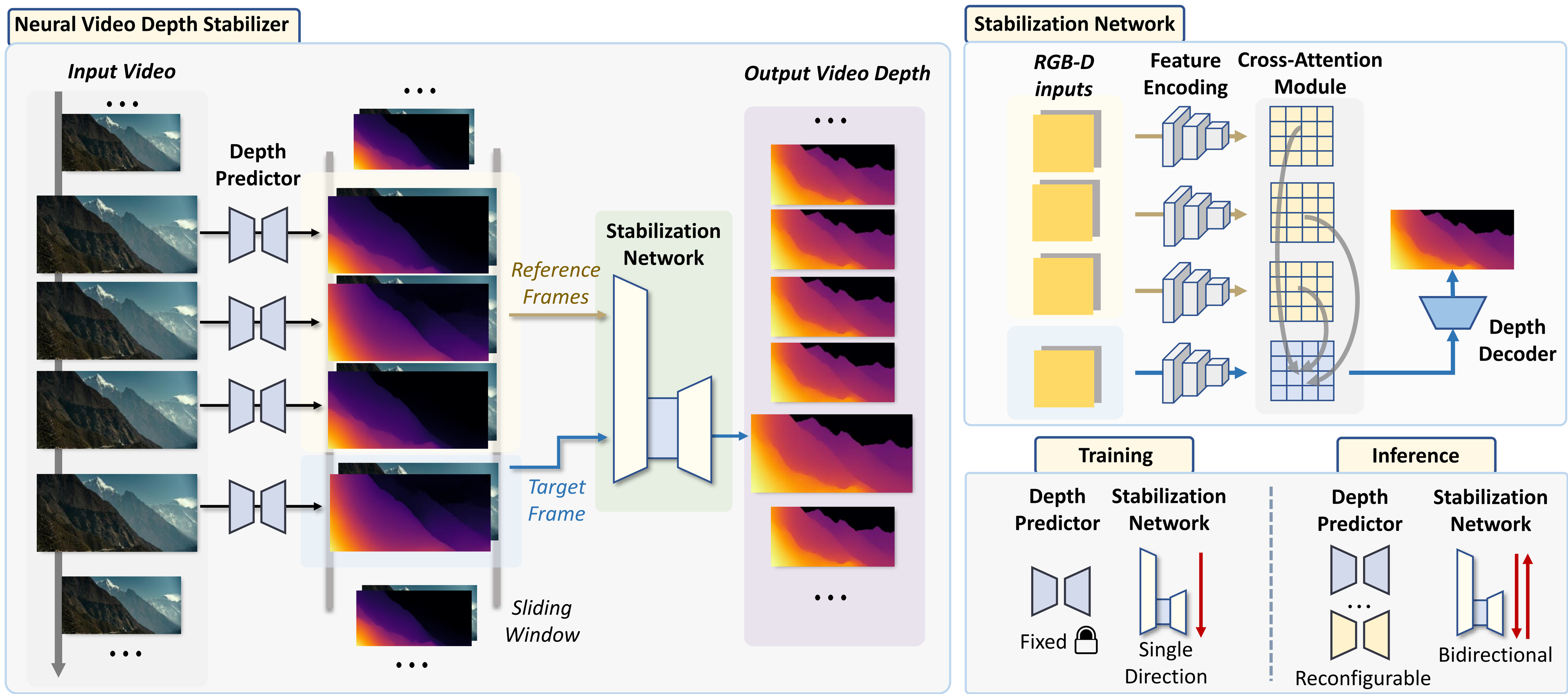
(a) Performance and Efficiency Comparisons



(b) Dataset Comparisons

- A plug-and-play and bidirectional learning-based framework termed Neural Video Depth Stabilizer (NVDS), which can be directly adapted to different single-image depth predictors to remove flickers.
- A large-scale dataset, Video Depth in the Wild (VDW), which is currently the largest natural-scene video depth dataset with the most diverse scenes.

Neural Video Depth Stabilizer (NVDS)



The Depth Predictor can be any single-image depth model which produces initial flickering disparity maps.
The Stabilization Network refines the flickering disparity maps into temporally consistent ones. It functions in a sliding window manner: the frame to be predicted fetches information from adjacent frames for stabilization.
Plug-and-play Manner: During inference, our Neural Video Depth Stabilizer (NVDS) can be directly adapted to any off-the-shelf single-image depth predictors in a plug-and-play manner without extra effort.
Bidirectional Inference: We also devise bidirectional inference to further improve temporal consistency.

Video Depth in the Wild (VDW) Dataset

To compensate for data shortage and boost performance of learning-based video depth models, we elaborate the currently largest natural-scene VDW dataset with the most diverse video scenes. We collect videos from four data sources: movies, animations, documentaries, and web videos. VDW contains 14,203 videos with 2,237,320 frames.

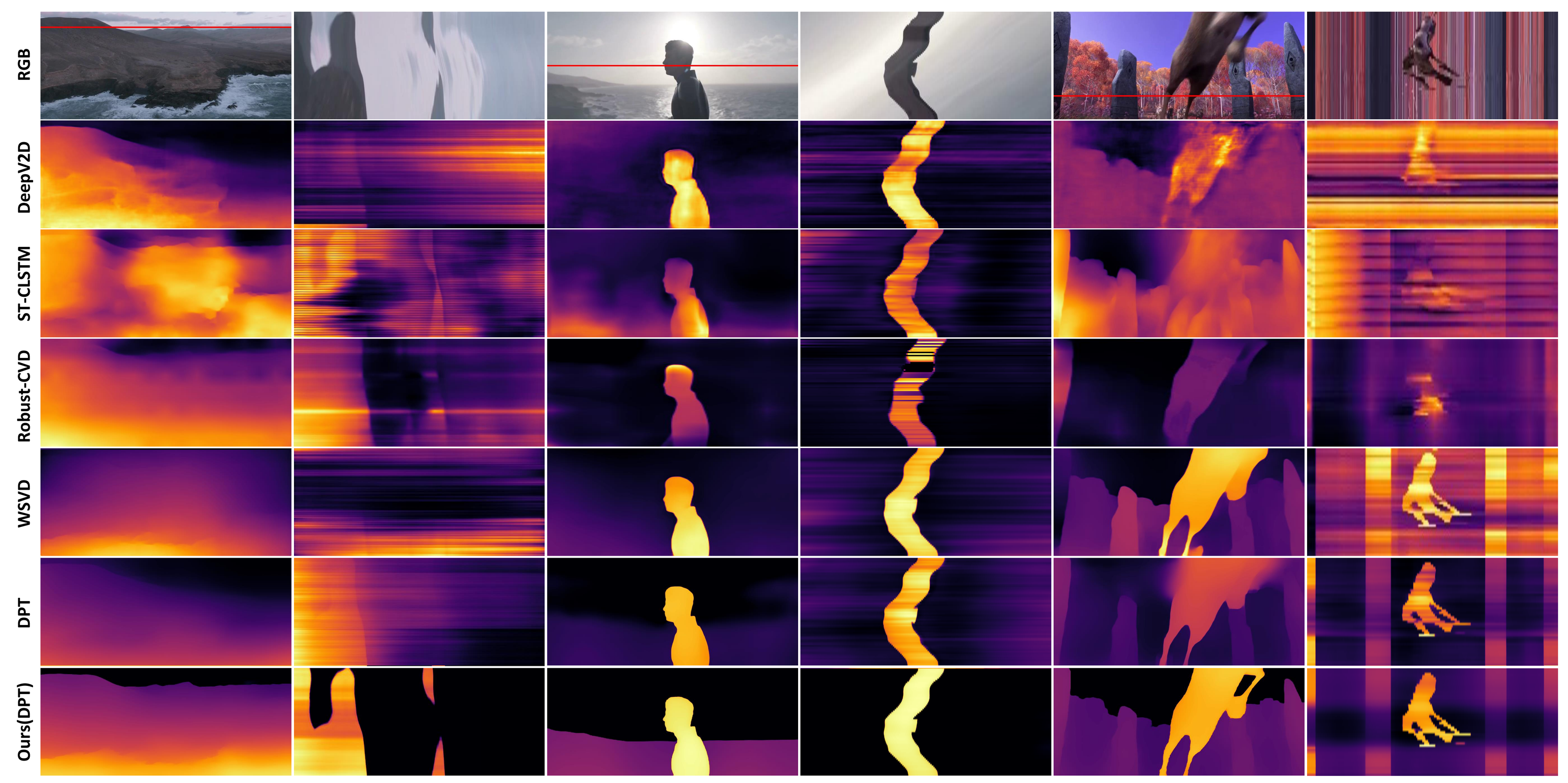
Type	Dataset	Videos	Frames(k)	Indoor	Outdoor	Dynamic	Resolution
Closed Domains	NYUDV2	464	407	✔	✗	✗	640 × 480
	KITTI	156	94	✗	✔	✔	1224 × 370
	TUM	80	128	✔	✗	✔	640 × 480
	IRS	76	103	✔	✗	✗	960 × 540
	ScanNet	1,513	2,500	✔	✗	✗	640 × 480
Natural Scenes	Sintel	23	1	✔	✔	✔	1024 × 436
	TartanAir	1,037	1,000	✔	✔	✗	640 × 480
	WSVD	553	1,500	✔	✔	✔	∼ 720p
	Ours	14,203	2,237	✔	✔	✔	1880 × 800

Experiments

Comparisons with the state-of-the-art approaches:

Type	Method	Time(s)	VDW $\delta_1\uparrow$	VDW $Rel\downarrow$	VDW $OPW\downarrow$	Sintel $\delta_1\uparrow$	Sintel $Rel\downarrow$	Sintel $OPW\downarrow$	NYUDV2 $\delta_1\uparrow$	NYUDV2 $Rel\downarrow$	NYUDV2 $OPW\downarrow$
Single Image	Midas	0.76	0.644	0.347	0.647	0.485	0.410	0.843	0.910	0.095	0.862
	DPT	0.97	0.724	0.266	0.461	**0.597**	0.339	0.612	0.928	0.084	0.811
Test-time Training	CVD	352.58	–	–	–	0.518	0.406	0.497	–	–	–
	Robust-CVD	270.28	0.658	0.334	0.251	0.521	0.422	0.475	0.886	0.103	0.394
	Zhang *et al.*	464.83	–	–	–	0.522	0.342	0.481	–	–	–
Learning Based	ST-CLSTM	0.58	0.461	0.589	0.455	0.351	0.517	0.585	0.833	0.131	0.645
	Cao *et al.*	–	–	–	–	–	–	–	0.835	0.131	–
	FMNet	3.87	0.465	0.584	0.388	0.357	0.513	0.521	0.832	0.134	0.387
	DeepV2D	68.71	0.522	0.628	0.425	0.486	0.526	0.534	0.924	0.082	0.402
	WSVD	4.25	0.621	0.379	0.437	0.501	0.439	0.577	0.768	0.164	0.683
	Ours(Midas)	1.55	0.694	0.286	0.164	0.532	0.374	0.469	0.941	0.076	0.373
	Ours(DPT)	1.73	**0.731**	**0.259**	**0.138**	0.591	**0.335**	**0.424**	**0.950**	**0.072**	**0.364**

Qualitative comparisons:



Plug-and-play manner:

	Initial $\delta_1\uparrow$	Initial $Rel\downarrow$	Initial $OPW\downarrow$	Ours $\delta_1\uparrow$	Ours $Rel\downarrow$	Ours $OPW\downarrow$
Midas	0.910	0.095	0.862	0.941	0.076	0.373
DPT	0.928	0.084	0.811	0.950	0.072	0.364
NeWCRFs	**0.937**	**0.072**	**0.645**	**0.957**	**0.068**	**0.326**

Influence of different training data:

Dataset	$\delta_1\uparrow$	$OPW\downarrow$
NYUDV2	0.527	0.435
IRS+TartanAir	0.542	0.489
VDW(Ours)	**0.591**	**0.424**

Setting	$\delta_1\uparrow$	$OPW\downarrow$
Scratch(DPT)	0.931	0.372
Pretrain(Midas)	0.941	0.373
Pretrain(DPT)	**0.950**	**0.364**

FlOPs and model parameters:

	DPT-L	NeWCRFs	Midas-v2	Stabilization Network
FLOPs (G)	1011.32	550.47	415.24	254.53
Params (M)	341.26	270.33	104.18	88.31

Github Repo: https://github.com/RaymondWang987/NVDS
NVDS Project Page: https://raymondwang987.github.io/NVDS/
VDW Dataset: https://raymondwang987.github.io/VDW/ Contact: wangyiran@hust.edu.cn