Supplementary Materials for Neural Video Depth Stabilizer

This supplementary contains the following contents:

- More details on the VDW dataset.
- More implementation details for NVDS.
- More details on experimental settings.
- More quantitative and qualitative results.

We also elaborate a demo consisting of many video visualizations and the illustration of our framework.

1. More details on the VDW Dataset

1.1. Dataset Construction

Data Acquisition and Pre-processing. Here we add more details on data acquisition and pre-processing (Sec. 4, line 466, main paper). Having obtained the raw videos, we use FFmpeg [4] and PySceneDetect [11] to split all the videos into 104,582 sequences. We manually check and remove the duplicated, chaotic, and blur scenes. Videos that are wrongly split by the scene detect tools are also removed. Finally, we reserve 32,405 videos with more than six million frames for disparity annotation.

Disparity Annotation. In Sec. 4, line 474 of the main paper, we mentioned that the disparity ground truth is obtained via sky segmentation and optical flow estimation. Here we specify the details. Compared with common practice [18, 13], we introduce a few engineering improvements to make the disparity maps more accurate. As the sky is considered to be infinitely far, pixels in the sky regions should be segmented and set to the minimum value in the disparity maps. We find that using a single segmentation model [1, 6] like prior arts [13, 18] causes errors and noises in the sky regions. Hence, we generate the sky masks in a model ensemble manner. Each frame along with its horizontally flipped copy are fed into two state-of-the-art semantic segmentation models SegFormer [19] and Mask2Former [3], which yields four sky masks in total. A pixel is considered as the sky when it is positive in more than two predicted sky masks. Besides, we also fill the connected regions with less than 50 pixels to further remove the noisy holes in the sky masks. Such ensemble strategy can improve the quality of the ground truth as shown in Fig. 1,

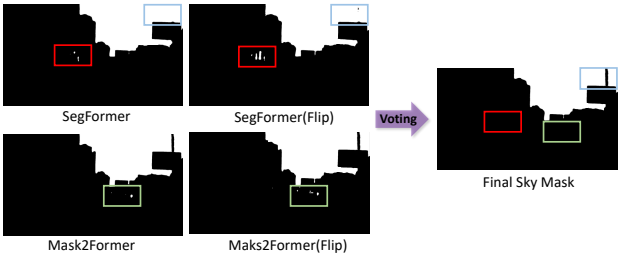

Figure 1: **Model ensemble strategy for sky segmentation on VDW dataset.** White area represents sky regions. Errors and noises in the rectangles are removed by model ensemble and voting, which improves the quality of the ground truth.

Figure 2: The word cloud of our VDW dataset.

and consequently improves the performance of the trained models, especially on skylines as shown in Fig. 6 and 8.

Following the practice of previous single-image depth datasets [18, 13], we adopt a state-of-the-art optical flow model GMFlow [20] to generate the ground truth disparity of the left- and right-eye views. The estimated optical flow is bidirectional. We perform a consistency check between the optical flow pairs to obtain the valid masks for training. We adopt the adaptive consistency threshold for each pixel as [9]. The ground truth of each video is normalized by its minimum and maximum disparity. Then, the disparity value is discretized into 65,535 intervals. Fig. 4 shows more examples of our VDW dataset.

Invalid Sample Filtering. Having obtained the annotations, we further filter the videos that are not qualified for our dataset. According to optical flow and valid masks, samples with the following three conditions are removed: 1) more than 30% of pixels in the consistency masks are

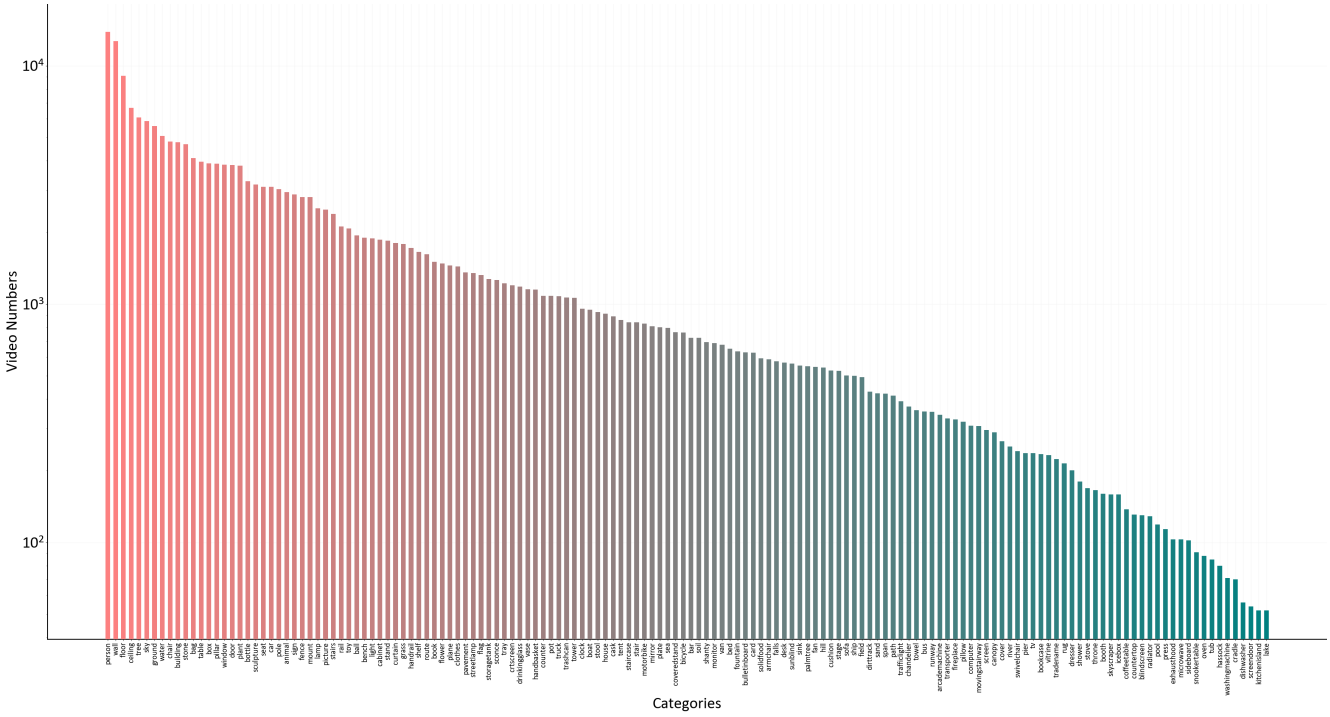

Figure 3: The statistics of the 150 semantic categories in VDW dataset.

invalid; 2) more than 10% of pixels have vertical disparity larger than two pixels; 3) the average range of horizontal disparity is less than 15 pixels. Then, we manually check all the videos along with their corresponding ground truth, and remove the samples with obvious errors. Finally, we retain 14,203 videos with 2,237,320 frames in VDW dataset.

1.2. Data Statistics

Here we add more statistics of VDW. Taking over 6 months to process, our VDW dataset contains 14,203 videos with 2,237,320 frames. The detailed data sources are listed in Table 1. Frame rates of all videos are over 24 fps. The minimum frame number is 18 while the maximum is 8,005.

To verify the diversity of objects in our videos. We conduct semantic segmentation with Mask2Former [3] trained on ADE20k [24]. All the 150 categories are covered in our dataset. The five categories that present most frequently are person (97.2%), wall (89.1%), floor (63.5%), ceiling (46.5%), and tree (42.3%). Each category can be found in at least 50 videos. Fig. 2 and Fig. 3 show the word cloud and detailed statistics of all the 150 categories.

Sources	Titles	Videos	Frames
Documentaries	Deepsea Challenge	210	38,078
	Kingdom of Plants	253	95,742
	Little Monsters	242	50,420
	Jerusalem	37	21,574
Animations	Coco	1,079	146,002
	Kung Fu Panda 3	959	68,405
Movies	Exodus: Gods and Kings	1,339	99,146
	Geostorm	857	52,028
	Hugo	301	25,091
	Mission: Impossible-Fallout	664	46,344
	Noah	1,160	85,161
	Pompeii	158	10,112
	Spider-Man: No Way Home	914	75,077
	The Legend of Tarzan	735	64,840
	The Three Musketeers	253	18,180
	Gravity	191	38,332
	Silent Hill 2	72	5,076
	Transformers: Age of Extinction	1,323	84,619
	Doctor Strange	299	23,779
	Battle of the Year	454	19,613
	Justice League	428	37,202
	The Hobbit 2	644	53,391
	The Great Gatsby	729	49,079
	Billy Lynn's Long Halftime Walk	242	29,137
Web Videos	YouTube	660	58,140
All	–	14,203	2,237,320

Table 1: **Video and frame numbers statistics of the four data sources.** Our VDW dataset contains 14,203 videos from movies, animations, documentaries, and web videos.

Figure 4: **More examples of our VDW dataset.** Sky regions and invalid pixels are masked out.

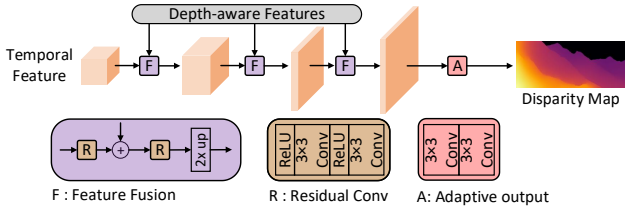

Figure 5: The architecture of decoder.

2. More Implementation Details for NVDS

2.1. Decoder Architecture

Here we specify the decoder architecture. The decoder architecture is illustrated in Fig. 5. To fuse the depth-aware features from the backbone [19] and temporal features from the cross-attention module, feature fusion modules (FFM) [6, 7] and skip connections are adopted. Resolutions are gradually increased while channel numbers are decreased. At last, we use an adaptive output module to adjust the channel and restore the disparity maps.

2.2. Loss Function

As mentioned in Sec. 3.3, line 352 in the main paper, the training loss consists of a spatial loss and a temporal loss. Here we specify the computation process.

For the spatial loss, we adopt the widely-used affinity invariant loss and gradient matching loss [13, 12] as $\mathcal{L}_s$. For the affinity invariant loss, let D and D^* denote the predicted disparity and ground truth respectively, we first calculate the scale and shift:

$$t(D) = median(D), s(D) = \frac{1}{M}\sum_{i=1}^{M}|D_i - t(D_i)|\,, \quad (1)$$

where M denotes the number of valid pixels. The prediction and the ground truth are aligned to zero translation and unit scale as follows:

$$\tilde{D} = \frac{D - t(D)}{s(D)}, \tilde{D}^* = \frac{D^* - t(D^*)}{s(D^*)}. \quad (2)$$

Then the affinity invariant loss can be formulated as:

$$\mathcal{L}_{af} = \frac{1}{M}\sum_{i=1}^{M}|\tilde{D} - \tilde{D}^*|. \quad (3)$$

Besides, we also adopt the multi-scale gradient matching loss [13], which can improve smoothness of homogeneous regions and sharpness of discontinuities in the disparity maps. The gradient matching loss is formulated as:

$$\mathcal{L}_{grad} = \frac{1}{M}\sum_{k=1}^{K}\sum_{i=1}^{M}(|\nabla_x R_i^k| + |\nabla_y R_i^k|), \quad (4)$$

where $R_i = \tilde{D}_i - \tilde{D}_i^*$, and R^k denotes the difference between the disparity maps at scale $k = 1, 2, 3, \cdots, K$ (the resolution is halved at each level). Following [12], we set $K = 4$ and set the weight β of $\mathcal{L}_{grad}$ to 0.5. The spatial loss can be expressed as:

$$\mathcal{L}_s = \mathcal{L}_{af} + \beta\mathcal{L}_{grad}, \quad (5)$$

Temporal loss. In line 362 of the main paper, we mentioned that the temporal loss is masked with a visibility mask $O_{n\Rightarrow n-1}$ calculated from the warping discrepancy between frame F_n and the warped frame $\hat{F}_{n-1}$. This mask is obtained by:

$$O_{n\Rightarrow n-1} = \exp(-\gamma||F_n - \hat{F}_{n-1}||_2^2)\,. \quad (6)$$

We set $\gamma = 50$ and use bilinear sampling layer for warping.

3. More Experimental Results

3.1. Depth Metrics

Here we specify the evaluation metrics for depth accuracy. we adopt commonly-applied depth evaluation metrics: Mean relative error (Rel) and accuracy with threshold t.
Mean relative error (Rel): $\frac{1}{M}\sum_{i=1}^{M}\frac{||D_i - D_i^*||_1}{D_i^*}$;
Accuracy with threshold t: Percentage of D_i such that $max(\frac{D_i}{D_i^*}, \frac{D_i^*}{D_i}) = \delta < t \in [1.25, 1.25^2, 1.25^3]$, where M denotes pixel numbers, D_i and D_i^* are prediction and ground truth of pixel i.

3.2. Model Efficiency

Here we evaluate the efficiency of the proposed Neural Video Depth Stabilizer (NVDS) in detail. Model parameters and FLOPs are reported in Table 2. The FlOPs are evaluated on a 384×384 video with four frames. The stabilization network of NVDS only introduces limited computation overhead compared with the off-the-shelf depth predictors.

	DPT-L [12]	NeWCRFs [21]	Midas-v2 [13]	Stabilization Network
FLOPs (G)	1011.32	550.47	415.24	254.53
Params (M)	341.26	270.33	104.18	88.31

Table 2: Comparisons of FlOPs and model parameters.

3.3. More Quantitative Comparisons

In the main paper, only δ_1, Rel, and OPW are reported. The additional results on the VDW and the Sintel [2] dataset are shown in Table 3 and Table 4. Besides, as CVD [8] and Zhang *et al.* [23] cannot produce results on 11 of 23 videos in Sintel [2] dataset, we additionally report the results on the other 12 videos in Table 5.

Method	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$Rel \downarrow$	$OPW \downarrow$
Midas [13]	0.644	0.853	0.928	0.347	0.647
DPT [12]	<u>0.724</u>	<u>0.890</u>	<u>0.950</u>	<u>0.266</u>	0.461
ST-CLSTM [22]	0.461	0.708	0.836	0.589	0.455
FMNet [17]	0.465	0.712	0.837	0.584	0.388
DeepV2D [15]	0.522	0.728	0.833	0.628	0.425
WSVD [16]	0.621	0.825	0.912	0.379	0.437
Robust-CVD [5]	0.658	0.855	0.928	0.334	0.251
Ours(Midas)	0.694	0.879	0.943	0.286	<u>0.164</u>
Ours(DPT)	**0.731**	**0.895**	**0.952**	**0.259**	**0.138**

Table 3: **Comparisons on VDW dataset.** The first 2 rows show the results of different single-image depth predictors. The next 5 rows contain video depth approaches. The last 2 rows consist of the results of our NVDS. Best performance is in boldface. Second best is underlined.

Method	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$Rel \downarrow$	$OPW \downarrow$
Midas [13]	0.485	0.693	0.787	0.410	0.843
DPT [12]	**0.597**	<u>0.768</u>	<u>0.846</u>	<u>0.339</u>	0.612
ST-CLSTM [22]	0.351	0.571	0.706	0.517	0.585
FMNet [17]	0.357	0.579	0.712	0.513	0.521
DeepV2D [15]	0.486	0.674	0.760	0.526	0.534
WSVD [16]	0.501	0.709	0.804	0.439	0.577
CVD [8]	0.518	0.741	0.832	0.406	0.497
Robust-CVD [5]	0.521	0.727	0.833	0.422	0.475
Zhang *et al.* [23]	0.522	0.727	0.831	0.342	0.481
Ours(Midas)	0.532	0.731	0.833	0.374	<u>0.469</u>
Ours(DPT)	<u>0.591</u>	**0.770**	**0.849**	**0.335**	**0.424**

Table 4: **Comparisons on the Sintel dataset.** We only report CVD [8] and Zhang *et al.* [23] on the 12 videos with valid outputs, while other methods are on the 23 videos.

Method	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$Rel \downarrow$	$OPW \downarrow$
Midas [13]	0.670	0.853	0.902	0.246	0.712
DPT [12]	**0.747**	0.874	0.917	**0.196**	0.671
ST-CLSTM [22]	0.477	0.711	0.827	0.366	0.547
FMNet [17]	0.492	0.728	0.825	0.363	0.516
DeepV2D [15]	0.509	0.735	0.827	0.384	0.575
CVD [8]	0.518	0.741	0.832	0.406	0.497
Zhang *et al.* [23]	0.522	0.727	0.831	0.342	0.481
WSVD [16]	0.621	0.822	0.891	0.305	0.581
Robust-CVD [5]	0.673	0.848	0.888	0.284	0.447
Ours(Midas)	0.700	0.866	0.918	0.226	<u>0.425</u>
Ours(DPT)	<u>0.741</u>	**0.876**	**0.926**	<u>0.205</u>	**0.411**

Table 5: **Comparisons on the 12 videos of Sintel [2] dataset.** We test the 12 videos that CVD [8] and Zhang *et al.* [23] can produce results for fair comparisons.

3.4. More Qualitative Results.

We show more visual comparisons in Fig. 6, 7, 8 and 9. Please refer to the supplementary video for video depth visualization results. We draw the scanline slice over time. Fewer zigzagging pattern means better consistency.

References

[1] Samuel Rota Bulo, Lorenzo Porzi, and Peter Kontschieder. In-place activated batchnorm for memory-optimized training of dnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5639–5647, 2018. 1

[2] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*, pages 611–625. Springer, 2012. 4, 5, 7

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 1, 2

[4] FFmpeg developers. FFmpeg. https://ffmpeg.org. [Online; Accessed 2022]. 1

[5] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1611–1621, 2021. 5, 7

[6] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1925–1934, 2017. 1, 4

[7] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017. 4

[8] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (ToG)*, 39(4):71–1, 2020. 4, 5, 7

[9] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 1

[10] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 7

[11] PySceneDetect developers. PySceneDetect. http://scenedetect.com. [Online; Accessed 2022]. 1

[12] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 4, 5

[13] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular

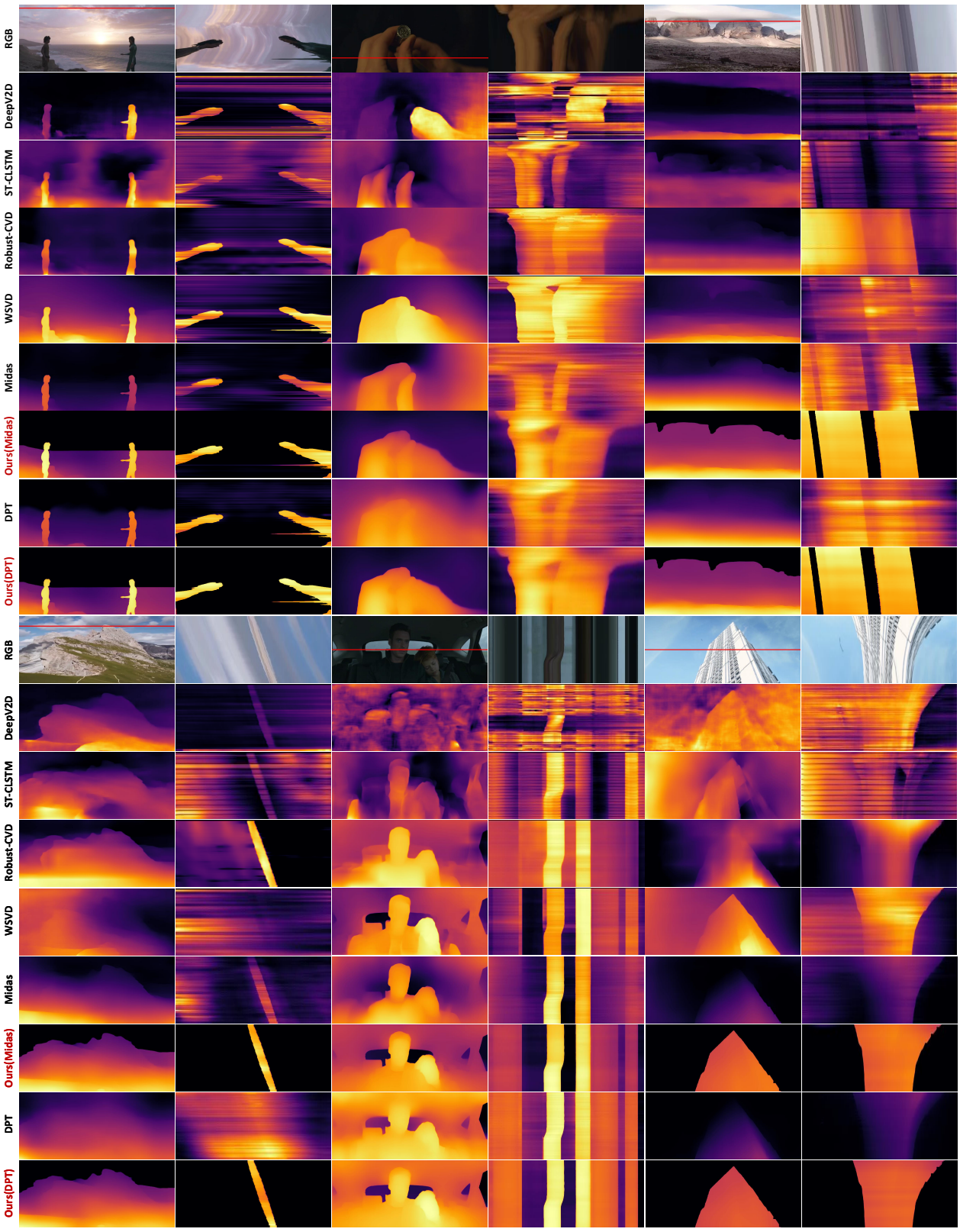

Figure 6: **More qualitative results on natural scenes.** The first image in each pair is the RGB frame, while the second is the scanline slice over time. Fewer zigzagging pattern means better consistency.

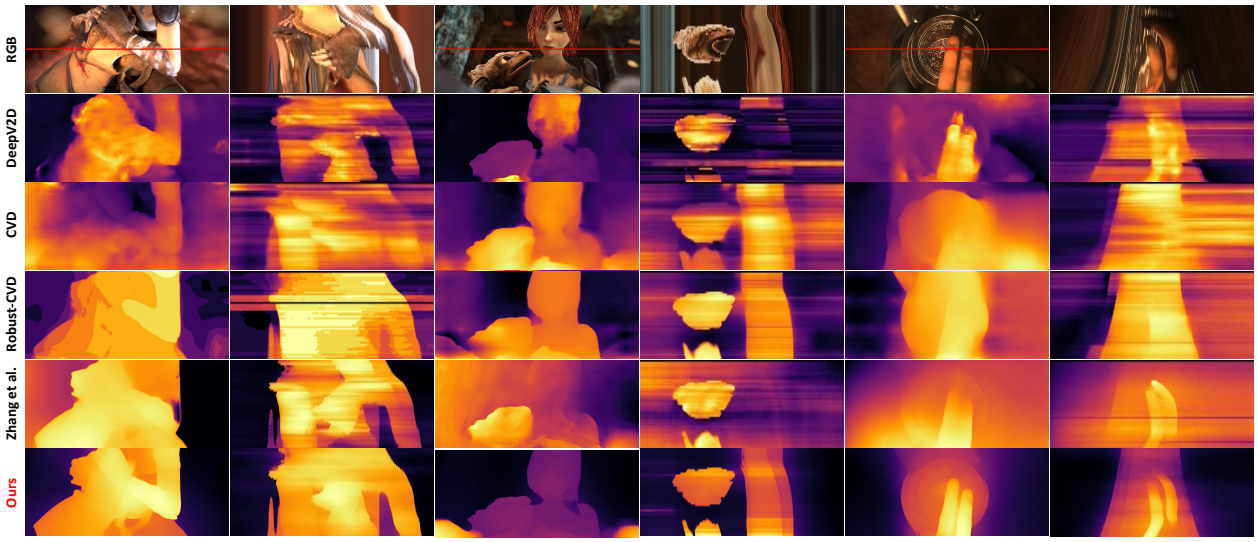

Figure 7: **Qualitative results on Sintel [2] dataset.** We compare the results of DeepV2D [15], CVD [8], Robust-CVD [5], and Zhang *et al.* [23]. Without relying on test-time training [8, 5, 23], we conduct zero-shot evaluations on Sintel [2] and achieve significantly better performance than those TTT-based methods [8, 5, 23].

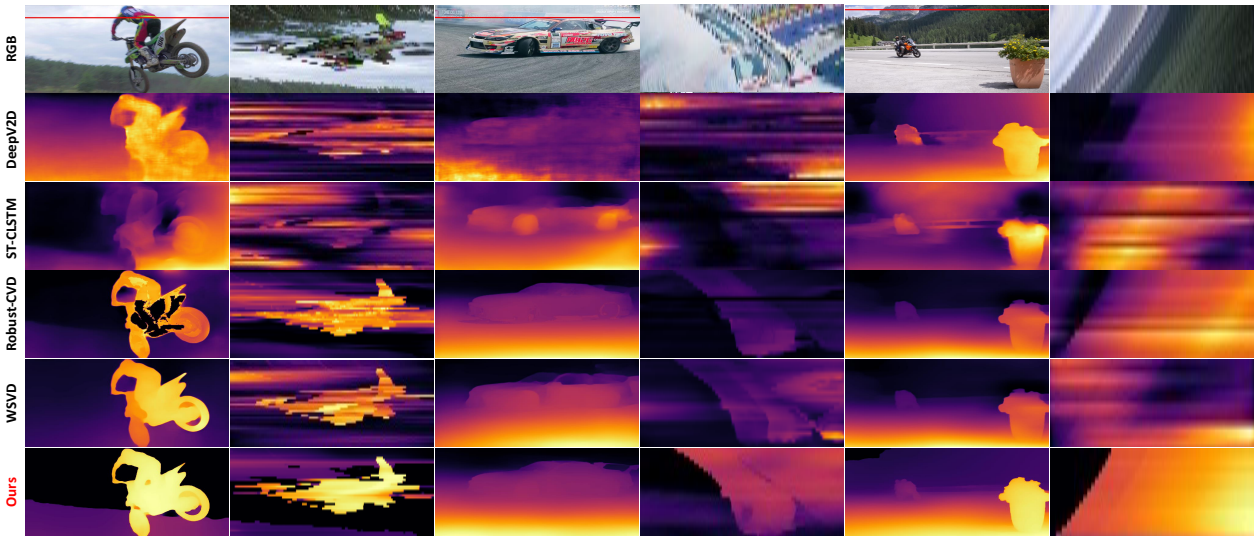

Figure 8: **Qualitative results on DAVIS [10] dataset.** We achieve better performance than prior arts on natural scenes.

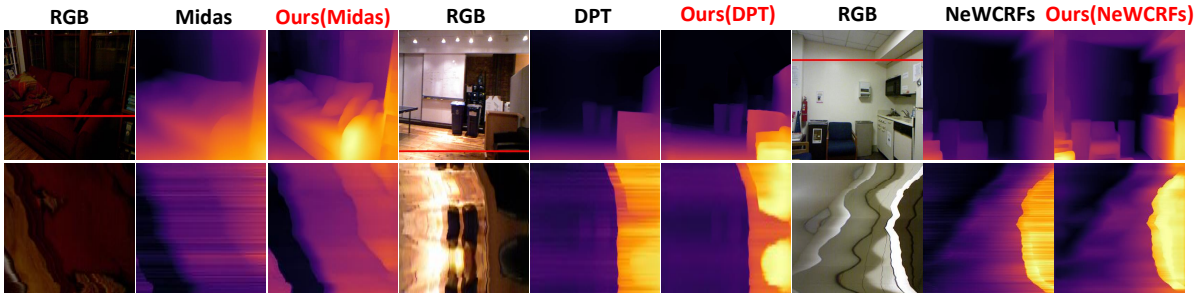

Figure 9: **Qualitative results on NYUDV2 [14] dataset.** We compare three different single-image depth predictors.

depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(03):1623–1637, 2020. 1, 4, 5

[14] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, pages 746–760. Springer, 2012. 7

[15] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. In *International Conference on Learning Representations*, 2019. 5, 7

[16] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *IEEE International Conference on 3D Vision (3DV)*, pages 348–357. IEEE, 2019. 5

[17] Yiran Wang, Zhiyu Pan, Xingyi Li, Zhiguo Cao, Ke Xian, and Jianming Zhang. Less is more: Consistent video depth estimation with masked frames modeling. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 6347–6358, New York, NY, USA, 2022. Association for Computing Machinery. 5

[18] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 311–320, 2018. 1

[19] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 1, 4

[20] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8121–8130, 2022. 1

[21] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Newcrfs: Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3916–3925, 2022. 4

[22] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1725–1734, 2019. 5

[23] Zhoutong Zhang, Forrester Cole, Richard Tucker, William T Freeman, and Tali Dekel. Consistent depth of moving objects in video. *ACM Transactions on Graphics (TOG)*, 40(4):1–12, 2021. 4, 5, 7

[24] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 633–641, 2017. 2