Supplemental File to "A Dynamic Multi-Scale Voxel Flow Network for Video Prediction"

We provide more details of DMVFN for video prediction. Specifically, we provide

- societal impact in §1.
- visualization of voxel flow §2;
- more ablation studies in §3;

1. Societal Impact

This work potentially benefits video prediction and dynamic neural network fields. The authors believe that this work has small potential negative impacts.

2. Visualization of Voxel Flow

We visualize the voxel flow predicted by DMVFN in Figure 1. We use the optical flow generated by RAFT [4] as a reference. We observe that the optical flow $\mathbf{f}_{t+1\to t}$ and the map $1-\mathbf{m}$ of most pixels are successfully predicted by DMVFN. This demonstrates that our DMVFN can indeed accurately predict a voxel flow.

Figure 1: **Visualization** of the map $1-\mathbf{m}$, the optical flow $\mathbf{f}_{t+1\to t}$, the optical flow by RAFT [4] $\mathbf{f}^{RAFT}_{t+1\to t}$, the predicted frame $\tilde{I}_{t+1}$ and the "ground truth" I_{t+1}.

3. More Ablation Study Results

5) How does β influence the performance of DMVFN during inference? The β is an important factor to control the model complexity and prediction capability during inference. Here, we adjust β during the inference phase, as shown in Table 1. DMVFN with larger β enjoys better MS-SSIM results but suffers from higher complexity.

Table 1: **Results of DMVFN with different** β evaluated on KITTI benchmark [2].

Settings ($\beta=$)	0.3	0.4	0.5	0.6	0.7	0.8
GFLOPs	**2.62**	3.88	5.15	5.94	6.21	6.40
LPIPS	16.47	12.91	10.74	10.26	10.24	**10.23**
MS-SSIM ($\times 10^{-2}$)	78.78	85.13	88.53	88.89	88.89	88.89

6) How to design the loss function? To study this problem, we train our DMVFN and DMVFN (w/o routing) only optimizing the loss on output of the last block $\tilde{I}_{t+1}$ (denoted as "single supervision"). The results listed in Table 2 show the advantages of our loss function L_{total}. L_{total} is calculated on all intermediate results of DMVFN.

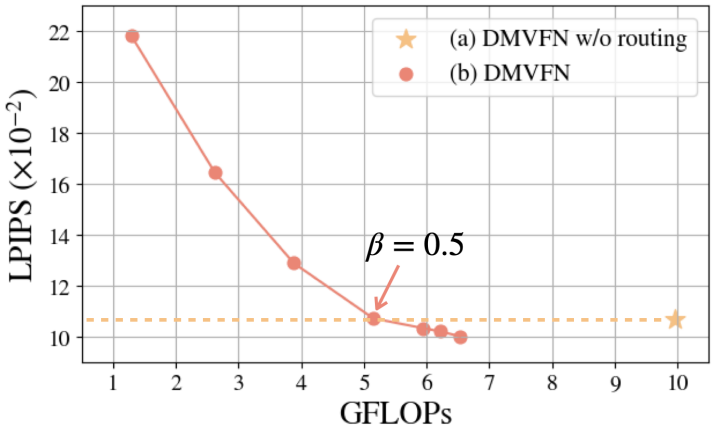

Figure 2: **Control the complexity of DMVFN by adjusting** β. DMVFN saves half GFLOPs of comparable performance compared to DMVFN without routing.

Table 2: **Results of DMVFN with different loss settings**. The evaluation metric is MS-SSIM ($\times 10^{-2}$).

Settings	Cityscapes			KITTI			Davis17-Val		Vimeo-Test
	t+1	t+3	t+5	t+1	t+3	t+5	t+1	t+3	t+1
w/o routing, single supervision	95.19	87.77	81.22	87.91	76.33	67.99	84.69	74.92	97.18
w/o routing	95.29	87.91	81.48	88.06	76.53	68.29	**84.81**	**75.05**	**97.24**
single supervision	95.65	89.10	83.27	88.34	77.88	70.18	83.83	74.68	96.95
DMVFN	**95.73**	**89.24**	**83.45**	**88.53**	**78.01**	**70.52**	83.97	74.81	97.01

Table 3: **Routing Module based on STEBS is effective**. The evaluation metric is MS-SSIM ($\times 10^{-2}$).

Settings	Cityscapes			KITTI			Davis-Val		Vimeo-Test
	t+1	t+3	t+5	t+1	t+3	t+5	t+1	t+3	t+1
w/o routing	95.29	87.91	81.48	88.06	76.53	68.29	**84.81**	**75.05**	**97.24**
Random	91.97	82.11	70.05	81.31	69.89	62.42	81.32	73.03	96.88
Gumbel Softmax	95.05	87.57	79.54	87.42	75.56	65.83	83.64	74.43	96.98
STEBS	**95.73**	**89.24**	**83.45**	**88.53**	**78.01**	**70.52**	83.97	74.81	97.01

More details about our Ablation Study 2) in the main paper. In Table 3, we summarize the quantitative results of three variants ("w/o routing", "Random" and "Gumbel Softmax") on four datasets (i.e., Cityscapes [1], KITTI [2], Davis-Val [3], and Vimeo-Test [5]). This demonstrates the effectiveness of our STEBS.
More details about our Ablation Study 3) in the main paper. In Table 4, we summarize the quantitative results of DMVFN with different scaling factor settings, including:

- "[1]": [1,1,1,1,1,1,1,1,1]
- "[2]": [2,2,2,2,2,2,2,2,2]
- "[4]": [4,4,4,4,4,4,4,4,4]
- "[1,2]": [1,1,1,1,2,2,2,2,2]
- "[1,4]": [1,1,1,1,4,4,4,4,4]
- "[2,1]": [2,2,2,2,1,1,1,1,1]

- "[4,1]": [4,4,4,4,1,1,1,1,1]
- "[1,2,4]": [1,1,1,2,2,2,4,4,4]
- "[4,2,1]": [4,4,4,2,2,2,1,1,1]

DMVFN [4,2,1] performs better than others, and the gap is more obvious for long-term future frames.

Table 4: **Results of DMVFN with different scaling factor settings**. The evaluation metric is MS-SSIM ($\times 10^{-2}$).

Settings	Cityscapes			KITTI			Davis-Val		Vimeo-Test
	t+1	t+3	t+5	t+1	t+3	t+5	t+1	t+3	t+1
DMVFN [1]	94.70	87.26	80.93	87.64	76.71	68.76	81.75	71.73	96.04
DMVFN [2]	95.51	87.76	81.30	87.06	76.90	69.05	81.77	72.58	96.07
DMVFN [4]	94.32	87.50	81.36	84.35	75.34	68.67	81.02	72.16	95.99
DMVFN [1, 2]	94.13	86.58	80.55	87.85	76.92	69.36	82.96	73.55	96.70
DMVFN [1, 4]	94.56	86.50	80.69	85.46	76.03	68.99	81.38	71.98	96.02
DMVFN [2, 1]	95.30	87.93	82.02	87.97	77.23	69.58	83.03	72.54	96.61
DMVFN [4, 1]	95.59	88.41	83.02	88.16	77.39	69.95	83.64	74.35	96.95
DMVFN [1, 2, 4]	94.20	86.56	80.81	87.77	76.89	69.72	82.72	73.66	96.76
DMVFN [4, 2, 1]	**95.73**	**89.24**	**83.45**	**88.53**	**78.01**	**70.52**	**83.97**	**74.81**	**97.01**

More details about our Ablation Study 4) in the main paper. In Table 5, we summarize the quantitative results of three variants ("w/o r, w/o path", "w/o r" and "w/o path") on four datasets (i.e., Cityscapes [1], KITTI [2], Davis-Val [3], and Vimeo-Test [5]).

Table 5: **Spatial path is effective in DMVFN**. The evaluation metric is MS-SSIM ($\times 10^{-2}$).

Settings	Cityscapes			KITTI			Davis-Val		Vimeo-Test
	t+1	t+3	t+5	t+1	t+3	t+5	t+1	t+3	t+1
w/o r, w/o path	94.99	87.59	80.98	87.75	76.22	67.86	84.45	74.78	97.05
w/o r	95.29	87.91	81.48	88.06	76.53	68.29	**84.81**	**75.05**	**97.24**
w/o path	95.55	88.89	83.03	88.29	77.53	69.86	83.75	74.51	96.89
DMVFN	**95.73**	**89.24**	**83.45**	**88.53**	**78.01**	**70.52**	83.97	74.81	97.01

More details about our Ablation Study 5) in the main paper. In Table 6, we summarize the quantitative results of different β during inference on four datasets (i.e., Cityscapes [1], KITTI [2], Davis-Val [3], and Vimeo-Test [5]).

Table 6: **Results of DMVFN with different** β evaluated on Cityscapes benchmark [1] and Vimeo-Test benchmark [5].

Settings	Cityscapes						Vimeo-Test					
$\beta =$	0.3	0.4	0.5	0.6	0.7	0.8	0.3	0.4	0.5	0.6	0.7	0.8
GFLOPs	**6.56**	9.81	12.71	15.30	16.23	17.82	**1.38**	2.08	2.77	3.40	3.74	3.92
LPIPS	8.88	7.06	5.58	5.20	5.15	**5.12**	5.18	4.18	3.69	3.48	3.42	**3.40**
MS-SSIM ($\times 10^{-2}$)	90.48	93.54	95.73	96.03	96.07	**96.12**	93.61	96.13	97.01	97.19	**97.20**	**97.20**

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2, 3

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *I. J. Robotics Res.*, 2013. 1, 2, 3

[3] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 2, 3

[4] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Eur. Conf. Comput. Vis.*, 2020. 1

[5] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. In *Int. J. Comput. Vis.*, 2019. 2, 3