Optical Flow Distillation: Towards Efficient and Stable Video Style Transfer – Supplementary Material –

Xinghao Chen[1⋆[0000−0002−2102−8235]], Yiman Zhang[1⋆[0000−0003−4494−4196]], Yunhe Wang[1[0000−0002−0142−509X]], Han Shu[1], Chunjing Xu[1], and Chang Xu[2[0000−0002−4756−0609]]

[1] Noah's Ark Lab, Huawei Technologies
[2] School of Computer Science, Faculty of Engineering, University of Sydney
{xinghao.chen,zhangyiman1,yunhe.wang,han.shu,xuchunjing}@huawei.com, c.xu@sydney.edu.au

1 Experiments for More Styles

We compare our proposed methods with student baseline for style *Picasso* on five scenes from MPI Sintel Dataset. As shown in Table 1, our proposed method consistently outperforms student baselines that are trained from scratch, which demonstrates the effectiveness of our method. For the student network with only perceptual loss, training it with our proposed distillation losses decreases e_{stab} from 0.3458 to 0.3035. For a stronger baseline, *i.e.*, using perceptual loss and temporal loss to train the student network, our proposed method outperforms the vanilla student by a 22.7% improvement for e_{stab}.

More results for style *WomenHat* and *Composition* are shown in Table 2 and Table 3, respectively. For both styles, our proposed method demonstrates consistent improvements over student baseline. Specifically, our method decreases e_{stab} from 0.307 to 0.2993 for style *WomenHat*. For style *Composition*, we also observe considerable reduction for e_{stab}.

Table 1. Comparisons of different methods for temporal error e_{stab} with style *Picasso* on five scenes from *MPI Sintel* Dataset.

Models	*Alley_2*	*Ambush_5*	*Bandage_2*	*Market_6*	*Temple_2*	*Sum*
Student [3]						
- From scratch	0.0676	0.0817	0.0432	0.0838	0.0695	0.3458
- Ours	**0.0578**	**0.0712**	**0.0416**	**0.0720**	**0.0608**	**0.3035**
Student [3] w/ $\mathcal{L}_{temp}$						
- From scratch	0.0541	0.0635	0.0435	0.0705	0.0562	0.2878
- Ours	**0.0362**	**0.0481**	**0.0319**	**0.0554**	**0.0388**	**0.2104**

⋆ Equal contribution.

Table 2. Comparisons of different methods for temporal error e_{stab} with style *WomenHat* on five scenes from *MPI Sintel* Dataset.

Models	*Alley_2*	*Ambush_5*	*Bandage_2*	*Market_6*	*Temple_2*	*Sum*
Student [3]						
- From scratch	0.0599	0.0795	0.0481	0.0739	0.0692	0.3307
- Ours	**0.0542**	**0.0714**	**0.0448**	**0.0678**	**0.0611**	**0.2993**

Table 3. Comparisons of different methods for temporal error e_{stab} with style *Composition* on five scenes from *MPI Sintel* Dataset.

Models	*Alley_2*	*Ambush_5*	*Bandage_2*	*Market_6*	*Temple_2*	*Sum*
Student [3]						
- From scratch	0.0490	0.0674	0.0493	0.0752	0.0589	0.2997
- Ours	**0.0433**	**0.0746**	**0.0404**	**0.0695**	**0.0467**	**0.2744**

2 Curve of e_{stab} Over Frames

Following the evaluating protocol of [2], we visualize the curve of e_{stab} over frames for *Alley_2* scene of *MPI Sintel* Dataset, as shown in Fig. 1. Our proposed method consistently outperforms student baselines for different frames.

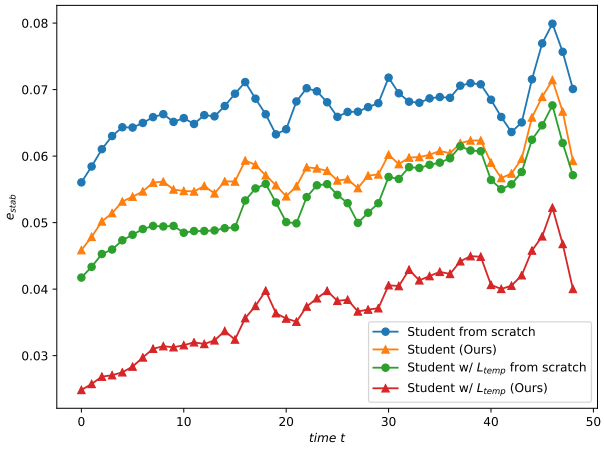

Fig. 1. Curve of e_{stab} over frames for one scene of *MPI Sintel* Dataset.

3 Results for 23 scenes from *MPI Sintel* Dataset

MPI Sintel Dataset [1] consists of 23 scenes and we compare our proposed methods on five scenes, following the evaluation protocol of previous work [3, 4]. We

further provide detailed experimental results for all 23 scenes, as shown in Table 4. It can be seen that our method shows strong and consistent improvements on 23 scenes from MPI Sintel Dataset.

Table 4. Comparisons of different methods for temporal error e_{stab} with style *Candy* on 23 scenes from *MPI Sintel* Dataset.

Scenes	*alley_2*	*ambush_2*	*ambush_6*	*cave_4*	*market_5*	*alley_1*
Student [3]	0.0746	0.1021	0.0972	0.0981	0.1162	0.0679
Ours	**0.0524**	**0.0797**	**0.0756**	**0.0805**	**0.0886**	**0.0467**
	sleeping_1	*temple_3*	*bandage_1*	*cave_2*	*sleeping_2*	*shaman_3*
	0.0638	0.1119	0.0642	0.0944	0.0656	0.0748
	0.0432	**0.0866**	**0.0462**	**0.0790**	**0.0453**	**0.0509**
	market_2	*ambush_7*	*bamboo_1*	*mountain_1*	*bandage_2*	*bamboo_2*
	0.0681	0.0727	0.0807	0.0622	0.0575	0.0712
	0.0492	**0.0468**	**0.0602**	**0.0417**	**0.0445**	**0.0574**
	temple_2	*ambush_5*	*shaman_2*	*market_6*	*ambush_4*	*Sum*
	0.0815	0.0887	0.0589	0.0997	0.1080	1.8801
	0.0627	**0.0676**	**0.0430**	**0.0779**	**0.0840**	**1.4098**

4 Ablation Studies for Hyper-parameters

Generally hyper-parameters (λ_{res} and λ_{rank}) are empirically set so that different losses have approximate scales and we tune these parameters to gain best trade-off between temporal consistence and visual effect. Here we conduct ablation experiments about the impacts of these parameters. As shown in Fig. 2, increasing the value of λ_{rank} result in slight lower e_{stab}, as stronger constraints are imposed into the output stylized videos. However, the quality of the stylized frames will be slightly hindered by heavy constraints. Therefore, we chose $\lambda_{rank} = 1 \times 10^2$ to better balance the temporal consistency and the quality of stylized images. Similarly, the impacts for e_{stab} with different values of λ_{res} are shown in Fig. 3. Larger values of λ_{res} tend to produce more stable output videos but the stylized frames would be quite blurry, as the perceptual loss is likely dominated by the residual distillation loss. We set $\lambda_{res} = 4 \times 10^8$ for a better trade-off.

5 Qualitative Video Results

We visualize the stylized videos and temporal consistency errors for different scenes of *MPI Sintel* Dataset. We compare our methods with teacher network and student baselines, for style *Candy* and *Picasso etc*. These qualitative results are included in supplementary video (*supp*.mp4) and Fig. 4.

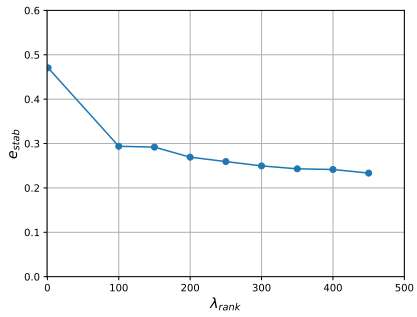

Fig. 2. Curve of e_{stab} with different λ_{rank}.

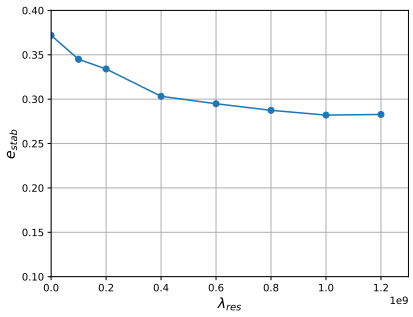

Fig. 3. Curve of e_{stab} with different λ_{res}.

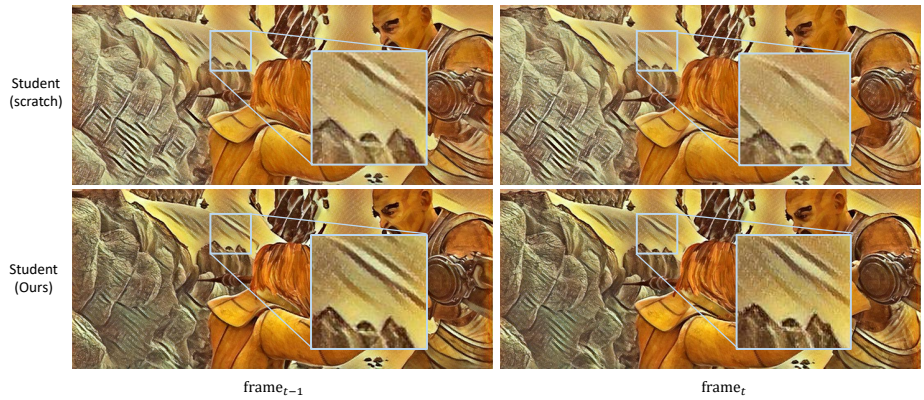

Fig. 4. Qualitative results of the student network trained from scratch and trained with the proposed method. Our method produces more temporally consistent stylized patterns.

References

1. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: ECCV. pp. 611–625. Springer (2012)
2. Chen, D., Liao, J., Yuan, L., Yu, N., Hua, G.: Coherent online video style transfer. In: ICCV. pp. 1105–1114 (2017)
3. Gao, C., Gu, D., Zhang, F., Yu, Y.: Reconet: Real-time coherent video style transfer network. In: ACCV. pp. 637–653. Springer (2018)
4. Ruder, M., Dosovitskiy, A., Brox, T.: Artistic style transfer for videos and spherical images. IJCV **126**(11), 1199–1219 (2018)