Characterizing and Improving Stability in Neural Style Transfer Supplementary Material

Agrim Gupta[1] Justin Johnson[1] Alexandre Alahi[1,2] Li Fei-Fei[1]
Stanford University[1] École Polytechnique Fédérate de Lausanne[2]

1. Transparency

Our method sometimes exhibits a failure mode where occluded objects are partially visible during the first few frames of occlusion; this causes the foreground object to appear partially transparent. An example is shown in Figure 1. Since this effect only appears for the first several frames of occlusion, it is not readily apparent when videos are played at full speed.

2. User Study

We performed a user study on Amazon Mechanical Turk to compare the subjective quality of our method, the Optim baseline [3], and the Real-Time baseline [1].

In each trial a worker is shown a video from the DAVIS dataset [2], a style image, and stylized output videos from two methods. In each trial the worker answers three questions: *"Which video flickers more?"*, *"Which video better matches the style?"*, and *"Overall, which video do you prefer?"*. For each question, the worker can either choose a video or select *"About the same"*.

We evaluate five styles in this experiment: *Mondrian*, *Metzinger*, *Mosaic*, *Rain*, and *Wave*. For our method and the Real-Time baseline, we use all 50 videos from the DAVIS dataset. Due to its slow runtime and the necessity for per-video hyperparameter tuning, we use only three videos per style for the Optim baseline.

Results for the question *"Which video flickers more?"* are shown in Table 1. We see that our method results in significantly less qualitative flickering than the Real-Time baseline: across 5 styles and 50 videos, a majority of workers thought that the video from the Real-Time baseline had more flickering in $193/250 = 77.2\%$ of videos; in contrast a majority of workers thought that the video from our method had more flickering in just $26/250 = 10.4\%$ of videos. Workers also thought that results from the Optim baseline had slightly less flickering than results from our method and significantly less flickering than results from the Real-Time baseline.

Results for the question *"Which video matches the style better?"* are shown in Table 2. Workers had circular preferences here: videos from the Real-Time baseline were judged to match the style image slightly more often than our method; our method was judged to match the style image slightly more often than the Optim baseline, and the Optim baseline was judged to match the style image slightly more often than the Real-Time baseline. Overall we believe that these results show that all methods do a reasonable job of producing videos that match the style image, with no method clearly better than the others.

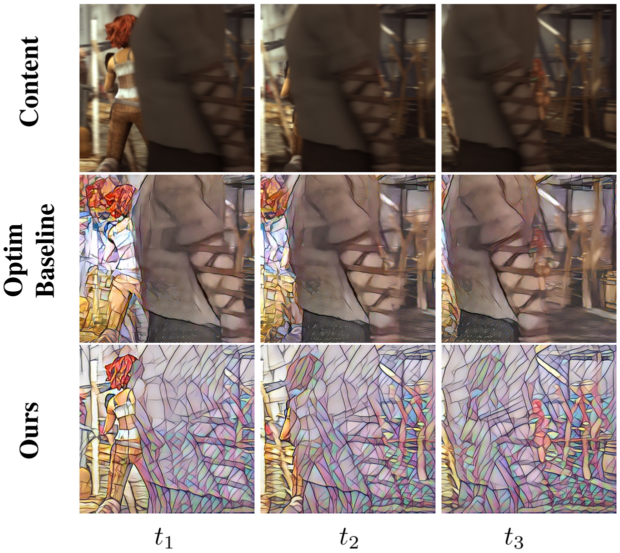

Figure 1. Sequence where our method performs poorly compared to Optim baseline [3]. In our method the girl is partially visible in spite of being occluded by the foreground person.

Results for the question *"Overall, which video do you prefer?"* are shown in Table 3. Despite the fact that our method was judged to match the style image slightly less often than the Real-Time baseline (see Table 2), overall workers showed a clear preference for the results of our method over those from the Real-Time baseline. Worker showed a slight preference for the results of the Optim baseline over our method, and a strong preference for the results of the Optim baseline over the Real-Time baseline.

Taken as a whole, this user study shows that our method results in videos with significantly less qualitative flickering than the Real-Time baseline, with temporal stability almost on par with the slower Optim baseline. Our method is perceived to match the style image about as well as other methods, and users prefer the results from our method significantly more often than the Real-Time baseline.

References

[1] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 1, 3

[2] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016. 1

[3] M. Ruder, A. Dosovitskiy, and T. Brox. Artistic style transfer for videos. In *German Conference on Pattern Recognition*, pages 26–36. Springer, 2016. 1, 3

Which video flickers more?

	Ours vs RT [1]			Ours vs Optim [3]			RT [1] vs Optim [3]		
Style	Ours	Tie	RT	Ours	Tie	Optim	RT	Tie	Optim
Mondrian	3	9	38	1	0	2	3	0	0
Metzinger	3	3	44	1	2	0	3	0	0
Mosaic	12	8	30	1	2	0	3	0	0
Rain	3	4	43	2	0	1	2	0	1
Wave	5	7	38	1	1	1	2	0	1
Total	**26**	31	193	6	5	**4**	14	0	**1**

	Ours vs RT [1]			Ours vs Optim [3]			RT [1] vs Optim [3]		
Style	Ours	Tie	RT	Ours	Tie	Optim	RT	Tie	Optim
Mondrian	35	93	122	5	5	5	10	1	4
Metzinger	30	66	154	6	5	4	10	4	1
Mosaic	46	96	108	5	8	2	10	0	5
Rain	18	94	138	5	7	3	10	3	2
Wave	22	121	107	4	7	4	7	4	4
Total	**151**	470	629	25	32	**18**	47	12	**16**

Table 1. User study results for the question *"Which video flickers more?"* We use 50 videos to evaluate our method against the Real-Time (RT) baseline [1] and 3 videos to evaluate the Optim baseline [3] against the other two methods. Each pair of videos is evaluated by five workers on Amazon Mechanical Turk. Left table shows the number of videos where the majority of workers preferred one method over another; right table shows the raw number of votes for each method across all videos. Lower is better.

Which video matches the style better?

	Ours vs RT [1]			Ours vs Optim [3]			RT [1] vs Optim [3]		
Style	Ours	Tie	RT	Ours	Tie	Optim	RT	Tie	Optim
Mondrian	16	11	23	2	1	0	1	0	2
Metzinger	17	13	20	2	0	1	0	0	3
Mosaic	11	12	27	1	2	0	1	1	1
Rain	12	16	22	1	0	2	1	1	1
Wave	14	14	22	2	0	1	0	0	3
Total	70	66	**114**	**8**	3	4	3	2	**10**

	Ours vs RT [1]			Ours vs Optim [3]			RT [1] vs Optim [3]		
Style	Ours	Tie	RT	Ours	Tie	Optim	RT	Tie	Optim
Mondrian	56	122	72	6	5	4	7	3	5
Metzinger	59	127	64	6	4	5	3	4	8
Mosaic	45	128	77	8	2	5	6	5	4
Rain	44	142	64	4	4	7	5	6	4
Wave	46	139	65	7	2	6	4	2	9
Total	250	658	342	**31**	17	27	25	20	**30**

Table 2. User study results for the question *"Which video matches the style better?"* We use the same experimental setup as Table 1. Left table shows the number of videos where the majority of workers preferred one method over another; right table shows the raw number of votes for each method across all videos. Higher is better.

Overall, which video do you prefer?

	Ours vs RT [1]			Ours vs Optim [3]			RT [1] vs Optim [3]		
Style	Ours	Tie	RT	Ours	Tie	Optim	RT	Tie	Optim
Mondrian	29	6	15	2	0	1	1	0	2
Metzinger	31	8	11	1	0	2	0	1	2
Mosaic	17	7	26	1	0	2	2	0	1
Rain	31	8	11	1	0	2	1	0	2
Wave	25	8	17	2	0	1	0	0	3
Total	**133**	37	80	7	0	**8**	4	1	**10**

	Ours vs RT [1]			Ours vs Optim [3]			RT [1] vs Optim [3]		
Style	Ours	Tie	RT	Ours	Tie	Optim	RT	Tie	Optim
Mondrian	120	57	73	7	2	6	5	0	10
Metzinger	131	48	71	5	2	8	4	3	8
Mosaic	97	52	101	7	1	7	4	3	8
Rain	118	66	66	4	1	10	5	2	8
Wave	96	77	77	6	2	7	4	1	10
Total	**562**	300	388	29	8	**38**	22	9	**44**

Table 3. User study results for the question *"Overall, which video do you prefer?"* We use the same experimental setup as Table 1. Left table shows the number of videos where the majority of workers preferred one method over another; right table shows the raw number of votes for each method across all videos. Higher is better.