Scene-Adaptive Video Frame Interpolation via Meta-Learning
- *Supplementary Document* -

This document discusses about additional results and analysis that could not take place in the main paper due to space limits. Please also refer to the project page[1] and our attached video demo for more visual comparison.

1. Evaluation metrics

We evaluate our scene-adaptive frame interpolation algorithm on two additional evaluation metrics: structural similarity index (SSIM) and interpolation error (IE). The quantitative results are summarized in Table 1 and 2. Note how the overall trend is similar to the results with peak signal-to-noise ratio (PSNR). Thus, we confirm that *Meta-trained* model consistently improves upon the *Baseline* or *Re-trained* models regardless of the evaluation metric, demonstrating the effectiveness of test-time adaptation with our meta-learning algorithm.

2. Additional qualitative results

Additional qualitative results for Middlebury-`OTHERS` [1], VimeoSeptuplet [8], and HD [3] datasets are shown in Figures 1, 2, and 3, respectively. As Fig. 4 of the main paper, we provide the results for 4 recent video frame interpolation algorithms—DVF [6], SuperSloMo [5], SepConv [7], and DAIN [2]—and illustrate the difference between *Baseline*, *Re-trained*, and *Meta-trained* model outputs.

We verify that our scene-adaptive frame interpolation algorithm greatly helps in finding the correct intermediate position of the moving objects and improve the original models (*Baseline* and *Re-trained*) to give visually more pleasing results. Notably, meta-training tends to considerably reduce the ghost artifacts appearing due to large motion, and also reduce undesirable artifacts (see Fig. 3) that appear possibly because of the model not able to handle severe occlusions.

3. Video demo

Please check the attached supplementary video for visual comparisons in the actual slow-motion video. We use the test sequences from DAVIS 2017 [4] dataset for the video demo. The video frame interpolation algorithm used throughout the video demo is fixed to SepConv [7]. Note that the differences between *Baseline*, *Re-trained*, and *Meta-trained* results become more noticeable in individual frames.

References

[1] Simon Baker, Daniel Scharstein, J. P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 92(1):1–31, 2010. 1, 2

[2] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *CVPR*, 2019. 1, 2, 3, 4

[3] Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *arXiv preprint arXiv:1810.08768*, 2018. 1, 2, 4

[4] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv:1905.00737*, 2019. 1

[5] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*, 2018. 1, 2, 3, 4

[6] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *ICCV*, 2017. 1, 2, 3, 4

[7] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *ICCV*, 2017. 1, 2, 3, 4

[8] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. In *CVPR*, 2018. 1, 2, 3

---

[1] `https://myungsub.github.io/meta-interpolation`

	VimeoSeptuplet [8]			Middlebury-OTHERS [1]			HD [3]		
Method	Baseline	Re-trained	Meta-trained	Baseline	Re-trained	Meta-trained	Baseline	Re-trained	Meta-trained
DVF [6]	0.8229	0.9186	**0.9202**	0.6837	0.8546	**0.8571**	—	—	—
SuperSloMo [5]	0.9143	0.9421	**0.9454**	0.9008	0.9470	**0.9482**	0.8134	0.8719	**0.8747**
SepConv [7]	0.9435	0.9436	**0.9474**	0.9573	0.9574	**0.9631**	0.8759	0.8796	**0.8820**
DAIN [2]	0.9510	0.9522	**0.9529**	**0.9661**	0.9657	0.9656	0.8895	0.8911	**0.8917**

Table 1: Quantitative results (SSIM) for meta-training for recent frame interpolation algorithms. Higher scores are better.

	VimeoSeptuplet [8]			Middlebury-OTHERS [1]			HD [3]		
Method	Baseline	Re-trained	Meta-trained	Baseline	Re-trained	Meta-trained	Baseline	Re-trained	Meta-trained
DVF [6]	6.31	3.45	**3.40**	9.37	4.68	**4.58**	—	—	—
SuperSloMo [5]	3.82	2.82	**2.71**	4.02	2.89	**2.85**	8.43	6.15	**6.05**
SepConv [7]	2.78	2.86	**2.72**	2.30	2.42	**2.23**	6.14	6.16	**6.00**
DAIN [2]	2.57	2.53	**2.51**	**2.07**	2.09	2.10	5.56	5.52	**5.48**

Table 2: Quantitative results (IE) for meta-training for recent frame interpolation algorithms. Lower errors signify better-performing models.

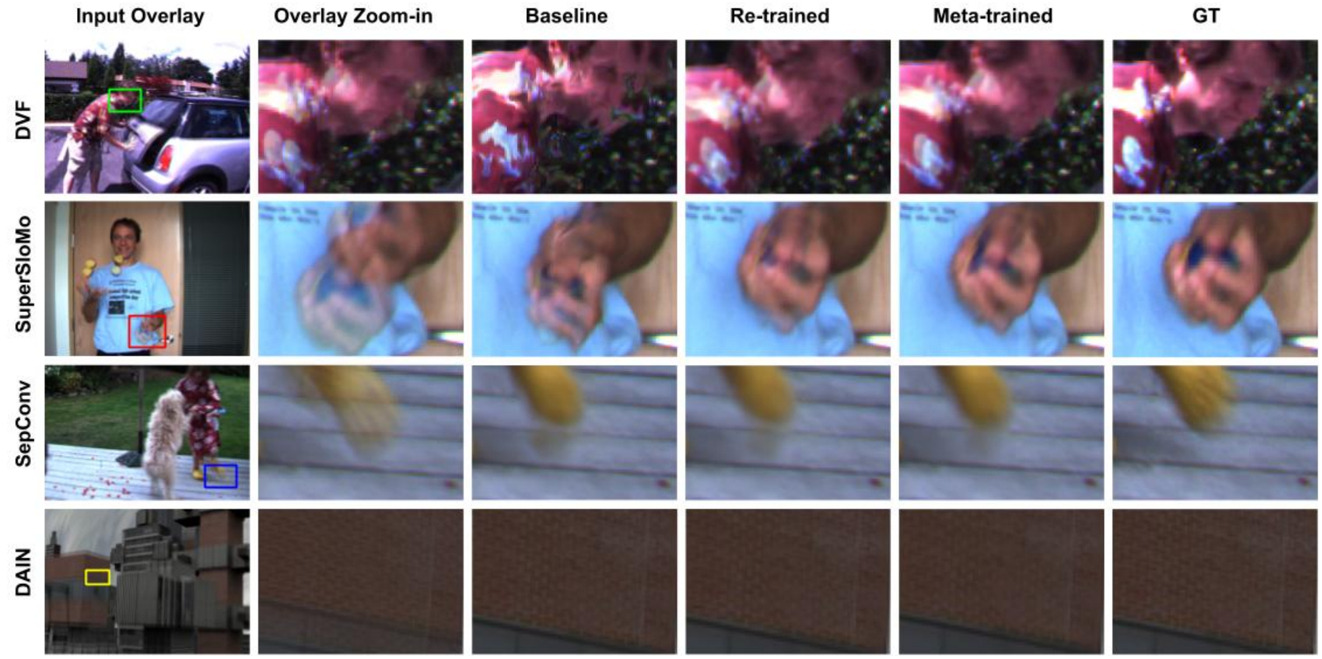

Figure 1. Qualitative results on Middlebury-OTHERS [1] dataset for recent frame interpolation algorithms. Note how our *Meta-trained* outputs infer motion substantially better than the *Baseline* or *Re-trained* models, as well as generate realistic textures similar to the ground truth. DVF [6] generates visible artifacts to the extent that a human face is severely distorted. Re-trained DVF no longer has severe artifacts but is highly blurred, due to poor motion interpolation. *Meta-trained* DVF, on the other hand, produces a sharper image, where the human face details are recognizable. Similarly for SuperSlomo [5], the *Baseline* and *Re-trained* models fail to capture the motion and thus generate blurs or artifacts, whereas our *Meta-trained* model precisely interpolates the motion, generating a sharp image similar to the ground truth. Similar observation can be made for SepConv [7], where *Meta-trained* model infers the position of moving objects more precisely, illustrating that *Meta-trained* model is able to learn the motion. As for DAIN [2], we could not find any examples with notable difference between *Baseline, Re-trained*, and *Meta-trained* within the 10 sequences from Middlebury-OTHERS [1] dataset. However, *Meta-trained* results sometimes produce less fine details of the textures compared with the baselines, which is possibly due to our algorithm trained with frames having larger motion. Best viewed when zoomed in.

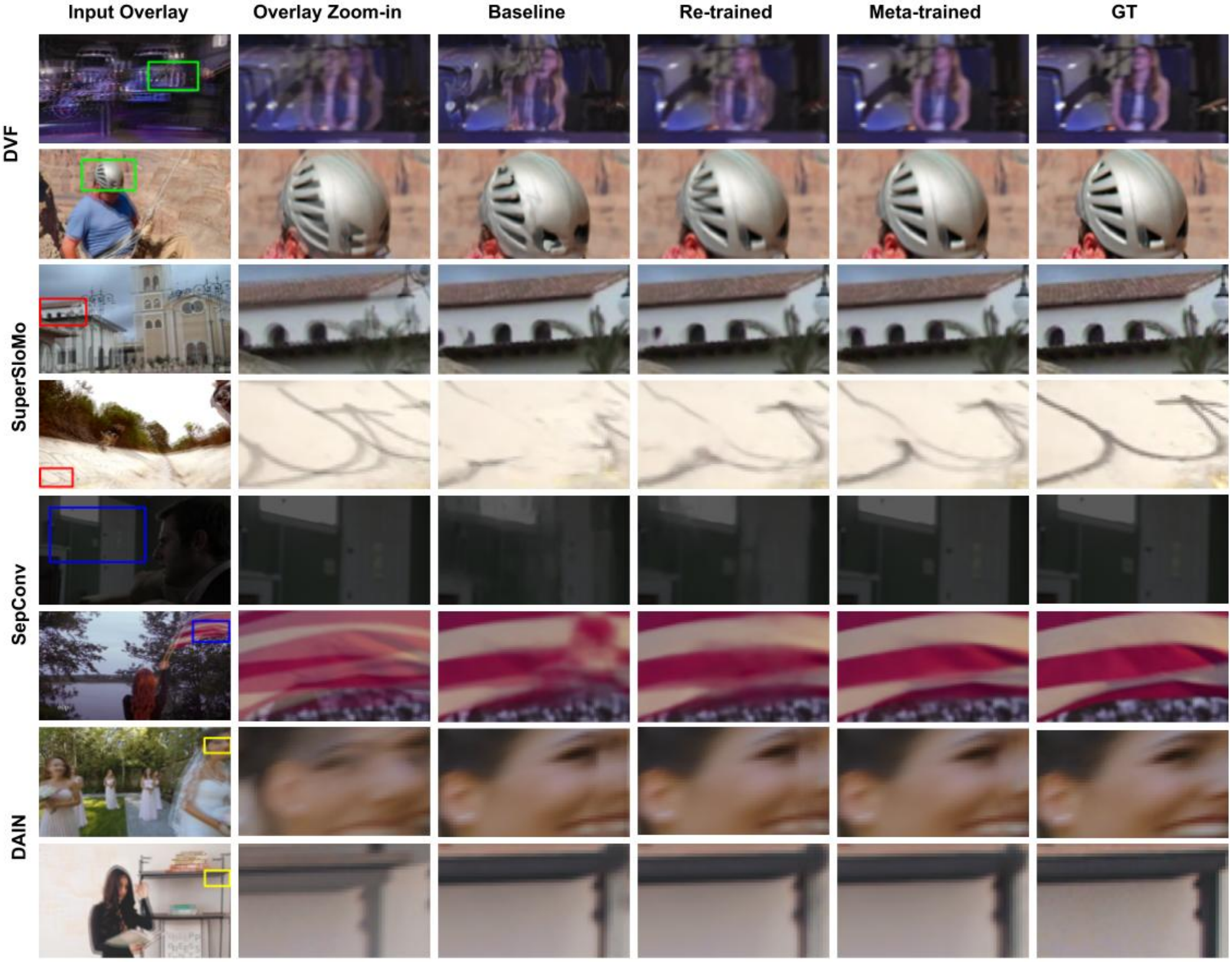

Figure 2. Qualitative results on VimeoSeptuplet [8] dataset for recent frame interpolation algorithms. Note how our *Meta-trained* outputs infer motion substantially better than the *Baseline* or *Re-trained* models, as well as generate realistic textures similar to the ground truth. For DVF [6], severe artifacts generated by inaccurate flow estimation are visible in the *Baseline* results, while these were reduced to some extent in the *Re-trained* results, our *Meta-trained* results show that these artifacts are mostly removed. For SuperSlomo [5], the *Baseline* and *Re-trained* models fail to restore detailed textures, whereas our *Meta-trained* model retains these details in the results. For SepConv [7], smudging and blurry artifacts are visible in the *Baseline* results, although these were mildly reduced in *Re-trained* results, our *Meta-trained* model removes these artifacts resulting sharper and cleaner interpolations. For DAIN [2], our *Meta-trained* model removes the blocky artifacts present in the *Baseline* and *Re-trained* results, producing more natural interpolations similar to the ground truth. Best viewed when zoomed in.

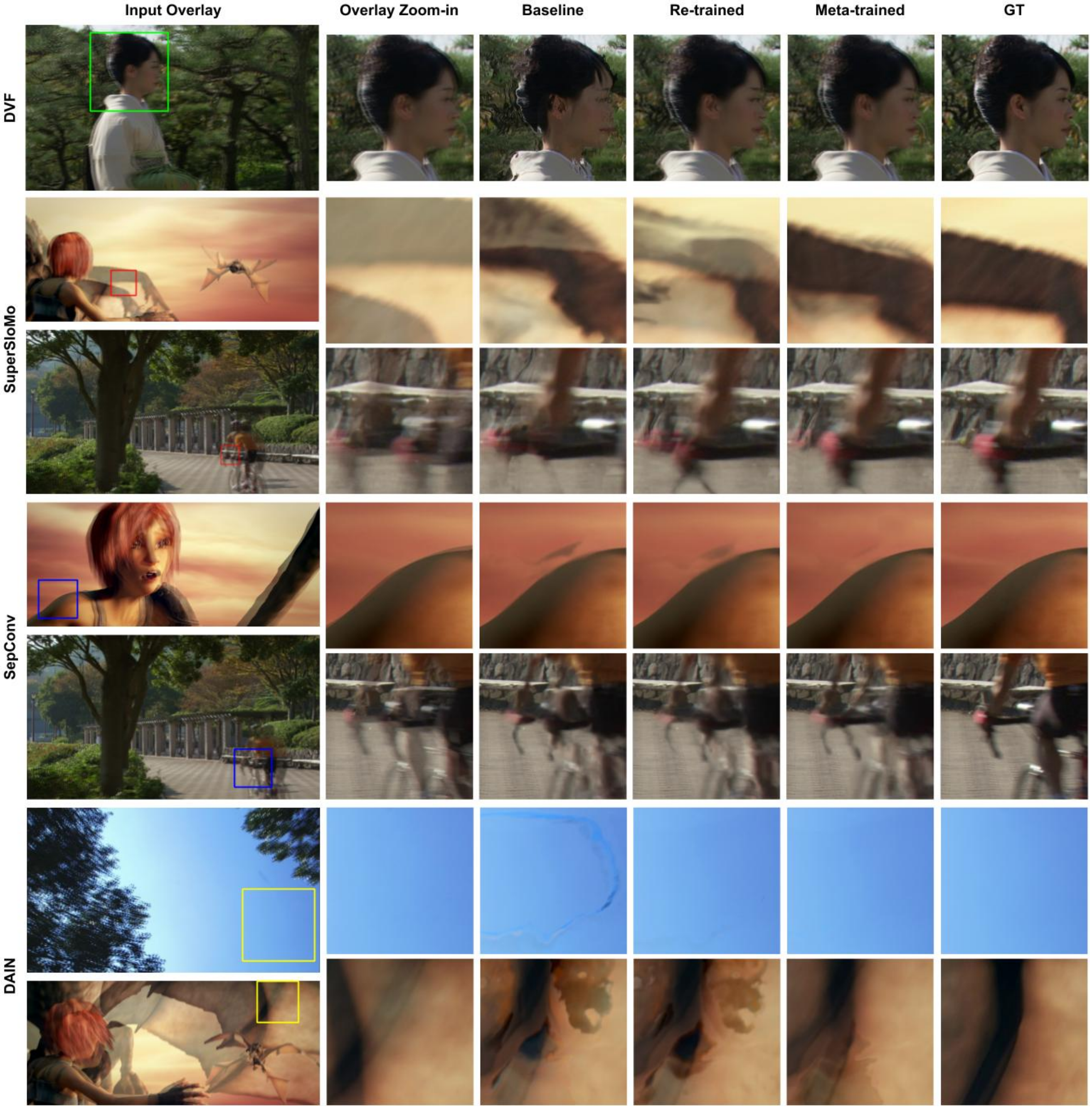

Figure 3. Qualitative results on HD [3] dataset for recent frame interpolation algorithms. Note how our *Meta-trained* outputs infer motion substantially better than the *Baseline* or *Re-trained* models, as well as generate realistic textures similar to the ground truth. For DVF [6], artifacts generated from erroneous flow estimation in the *Baseline* results are removed to some extent by the *Re-trained* and *Meta-trained* models, but due to the inherent limitations of DVF where the algorithm is unable to handle high resolution inputs, both models fail to synthesize convincing interpolations. For SuperSlomo [5], the *Baseline* and *Re-trained* models struggle to estimate the intermediate position of moving objects, whereas our *Meta-trained* model precisely estimates the positions and synthesizes more realistic interpolations. For SepConv [7], blurry artifacts are visible in the *Baseline* and *Re-trained* results, and our *Meta-trained* model mostly removes these artifacts. For DAIN [2], *Baseline* model produces large artifacts in areas with homogeneous textures or large moving objects, and the artifacts are still visible in the *Re-trained* results. Our *Meta-trained* model removes these artifacts and produces cleaner images. Best viewed when zoomed in.