Harnessing Meta-Learning for Improving Full-Frame Video Stabilization — Supplementary Materials —

Muhammad Kashif Ali[1] Eun Woo Im[2] Dongjin Kim[1] Tae Hyun Kim[1†]
{kashifali, iameuandyou, dongjinkim, taehyunkim}@hanyang.ac.kr
[1]Dept. of Computer Science, [2]Dept. of Artificial Intelligence, Hanyang University

1. Overview

Due to the space limitation in the main paper, we provide a brief intro to MAML, details of the affine regression network, metric details, implementation details, ablation studies, further results, and user study results in this supplemental.

2. A brief intro to MAML

The task of adapting model parameters to a unique task with only a few examples present is achieved with the help of meta-learning. The method proposed by [3] (termed MAML) achieves this with few gradient updates by instilling a sense of adaptability to various task data in model parameters through meta-training. In short, MAML searches for parameters that are sensitive to change and uses them as a reliable initialization. Doing so allows the model to be adapted quickly in a few updates. A brief intro to the mathematical formulation of MAML is provided below.

MAML operates under the assumption of a task distribution, denoted as $p(\mathcal{T})$. The primary objective of MAML is to learn initialization parameters that encapsulate the underlying knowledge shared across various tasks present in the distribution.

In a typical k-shot learning scenario, each task $\mathcal{T}i \sim p(\mathcal{T})$ is associated with a set of k examples, represented as $\mathcal{D}_{\mathcal{T}i}$. These examples, along with their corresponding objective function $\mathcal{L}_{\mathcal{T}_i}$, serve as an approximate representation of the task. MAML adapts its model to each task through a fine-tuning process, which involves updating the parameters as follows:

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta). \tag{1}$$

Once the model is adapted to a particular task $\mathcal{T}_i$, new examples $\mathcal{D}'_{\mathcal{T}_i}$ are sampled from the same task to evaluate the model's generalization performance on unseen data. This evaluation process acts as feedback for MAML, enabling it to adjust its initialization parameters to achieve improved generalization across tasks:

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'}). \tag{2}$$

Eq. 1 guides the inner loop optimization whereas, Eq. 2 guides the outer loop optimization. For brevity, we will use $\mathcal{L}_{\mathcal{T}_{in}}$, for inner loop loss and $\mathcal{L}_{\mathcal{T}_{out}}$ for outer loop losses for our formulation. Please refer to the original paper [3] for an in-depth explanation of this formulation.

3. Affine regression network

For the purpose of rigid affine estimation, we trained a separate network trained with the help of randomly transformed images. The network consists of an encoder-decoder architecture with a fully connected head that regresses rotation and translation parameters from the input global optical flow (as presented in [4]) ($\mathcal{F}_{I\to I'}$) estimated from input images I and a transformed image I'. Their proposed flow estimation network employs knowledge distillation to finetune a PWC-Net [9] to estimate the global optical flow of dynamic scenes. We utilize their global flownet instead of a conventional optical flow estimation network as it masks out the dynamic objects from the evaluated flow which aids the proposed affine estimation network to focus on global transforms in a video rather than local transforms and it is also robust against augments like cropped regions as highlighted in Fig. 1. The proposed affine transform estimation network contains fully connected layers, due to these layers, the $\mathcal{F}_{I\to I'}$ is downscaled to a resolution of 64×64. The proposed network regresses the rigid affine parameters as follows:

$$\hat{\mathcal{A}} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & x \\ \sin(\theta) & \cos(\theta) & y \\ 0 & 0 & 1 \end{bmatrix} = h_\phi(\mathcal{F}_{I\to I'}). \tag{3}$$

†: Corresponding author.

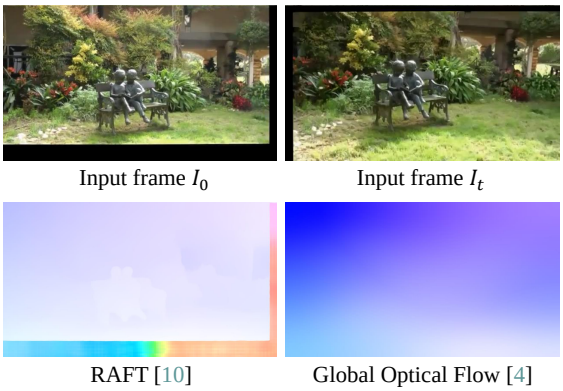

Figure 1. **Comparison of conventional and global optical flow.** Optical flow estimated from RAFT [10] (bottom left) and Global Optical Flow estimated with [4] (bottom right). Global optical flow, in addition to masking the dynamic objects, also fills in the gaps near the frame boundaries.

Here, the estimated rigid-affine transform is denoted by $\hat{\mathcal{A}}$, h_ϕ is the proposed affine estimation network, θ and (x, y) are rotation and translation components, respectively. The proposed affine estimation network is trained with an affine loss and a pixel loss defined below.

$$\mathcal{L}_{\text{affine}} = \left\| \mathcal{A} - \hat{\mathcal{A}} \right\|_2^2, \tag{4}$$

$$\mathcal{L}_{\text{pixel}} = \left\| I - w(I', \hat{\mathcal{A}}) \right\|_2^2. \tag{5}$$

For the affine loss (Eq. 4), $\mathcal{A}$ denotes the ground truth rotation and translation components of the random affine transform, and for the pixel loss (Eq. 5), $w(\cdot)$ denotes the warping operator which warps the transformed frame I' towards its untransformed counterpart I. Both of these losses are used in the training of the proposed affine estimation network. The training of this model converged in nearly $\sim$40K optimization iterations. After the convergence, this network was used as a stability guide for the inner loop loss proposed in the main manuscript.

4. Metric details

There are three primary metrics to evaluate the performance of video stabilization namely, Stability, Cropping, and Distortion. We provide the implementation details of these metrics below:

Stability This metric characterizes stability through frequency component analysis. The metric is computed by analyzing feature trajectories in the frequency domain using the equation:

$$f_v = FFT(V), \tag{6}$$

where f_v denotes the frequency representation of both translational and rotational camera trajectories of the V. f_v is obtained by performing a discrete 1D Fourier Transform on V. After subtracting the direct current (DC) component from f_v, the stability score is determined following the guidelines of [11], as follows:

$$S = \sum_{n=2}^{6} f_v(n) / \sum_{n=2}^{N} f_v(n), \tag{7}$$

where S signifies the stability score for both translational and rotational motions, and N denotes the total frequency components present in the signal.

Cropping This metric quantifies the preservation of visual information within generated frames by computing the homography between these frames and the actual frames. The cropping ratio metric is derived by calculating the average scale component of the estimated homographies across the entire video.

Distortion This metric assesses the anisotropic homography between the produced frames and the original unstable frames. The computation of this metric involves the ratio of the two largest eigenvalues from the affine component of the estimated homography. The distortion value metric is then determined by selecting the lowest ratio among all homographies corresponding to the frames.

5. Implementation details

For our experiments, we use the models proposed in [1, 2] as baselines. We initialize the models with the pre-trained model parameters provided by the authors and meta-train these models according to the training algorithm provided in the main paper. The dataset used in our experiments is DeepStab dataset [11]. The proposed algorithm was implemented with the help of *"Higher"* framework for meta-learning [5]. The optimizers used (for both inner and outer loop optimization) were Adam [7]. The number of inner loop updates m was fixed to 1 for the meta-training of both the baseline models due to memory limitations. The patch size for sampled sequences for meta-training of both baseline models was fixed to 192×192 for the inner loop update, and for the outer loop update, it was increased to 320×320.

In order to manage the resources during test-time adaptation, we sample sequential patches of 320×320 from the sampled short sequences for adaptation. In our experiments, we observed that a bigger patch size during adaptation leads to better results but can be computationally expensive we can however overcome this by increasing the number of adaptation iterations.

During our experimentation phase, we observed that a careful selection of these patches for adaptation (for $m \geq 5$)

near the four edges and center of the frames can potentially lead to better results. As in the case of pure rotation, the highest jerk effect is seen near the far edges of the frames, whereas, in the case of pure translation, the jerk in each of the patches from these locations will be similar in magnitude.

It is worth mentioning that the piece-wise adaptation strategy of sampling and stabilizing short sequences can be further improved by pre-stabilizing the videos to build a stronger temporal connection but it requires complex pre-processing for evaluating rigid transforms between each successive frame and stabilizing the evaluated transforms using various smoothing techniques like savitzky-golay filtering with appropriate window sizes and degree for pre-stabilizing each video as the proposed aligning strategy works well for short sequences but it might face challenges in longer sequences with significant content change. Due to this limitation, we used a piece-wise strategy for our fast-adaptation algorithm.

The videos used for evaluating both the quantitative and qualitative results are taken from the NUS dataset [8]. The meta-training of both models converged in roughly 5000 meta-training iterations. The code and the meta-trained models will be made available on: https://github.com/MKashifAli/MetaVideoStab.

6. Further results

Due to the space limitation in the main paper, we present the expanded view of the baseline comparative results (Fig. 2) along with cropping (Tab. 1) and distortion metrics (Tab. 2), and the run-time comparison of the proposed algorithm in this section.

Table 1. **Quantitative comparison of adapted models against baselines.** This table presents the comparison of the cropping score of baseline models with their scene-adaptive variants. The subscript shows the number of sequences sampled for adaptation and the superscript denotes the adaptation number. This table highlights that despite consistently increasing the stability score, we see a minor decrease in the cropping value with increasing adaptation iterations.

Model		Cropping					
		Crowd	Parallax	Regular	Running	Quick Rot	Zoom
DMBVS	Baseline	0.9998	0.9997	0.9997	0.9993	0.9995	0.9990
	Adapt$^{(1)}_{100}$	0.9988	0.9979	0.9995	0.9989	0.9961	0.9984
	Adapt$^{(5)}_{100}$	0.9985	0.9964	0.9992	0.9965	0.9949	0.9978
DIFRINT	Baseline	0.9997	0.9989	0.9992	0.9988	0.9998	0.9998
	Adapt$^{(1)}_{100}$	0.9996	0.9996	0.9996	0.9986	0.9989	0.9997
	Adapt$^{(5)}_{100}$	0.9997	0.9997	0.9996	0.9987	0.9992	0.9996

6.1. User study

It is important to mention that the metrics discussed in the main paper and in this supplemental, have individual limitations as they do not encompass all the aspects covered by one another as pointed out by Ali *et al*. [1]. Therefore, to properly assess the models for both perceptual quality and stability, comprehensive user studies become indispensable and serve as an essential evaluation metric for this task. We conducted *3* distinct user studies, one for comparing with the baselines, one for comparing the recurrent extension of the model proposed in [1] to its non-recurrent variant, and another for comparing against the SOTA methods for this task and present our findings below.

Table 2. **Quantitative comparison of adapted models against baselines.** This table presents the comparison of the distortion score of baseline models with their scene-adaptive variants. The subscript shows the number of sequences sampled for adaptation and the superscript denotes the adaptation number. This table highlights that despite consistently increasing the stability score, the quality of the processed videos is not compromised even with a higher number of adaptation iterations.

Model		Distortion					
		Crowd	Parallax	Regular	Running	Quick Rot	Zoom
DMBVS	Baseline	0.9794	0.9659	0.9737	0.9063	0.8772	0.9108
	Adapt$^{(1)}_{100}$	0.9321	0.8722	0.9422	0.8251	0.8977	0.9396
	Adapt$^{(5)}_{100}$	0.9415	0.9370	0.9522	0.9351	0.9302	0.9523
DIFRINT	Baseline	0.9534	0.9544	0.9813	0.9108	0.8847	0.9299
	Adapt$^{(1)}_{100}$	0.9626	0.9615	0.9878	0.9521	0.9366	0.9542
	Adapt$^{(5)}_{100}$	0.9616	0.9634	0.9884	0.9541	0.9271	0.9697

6.1.1 Comparison with baselines

A user study involving *40* participants was conducted to assess the performance of adapted variants compared to baseline models. The study featured a randomized presentation of videos, comprising a total of 12 randomly sampled videos from the NUS dataset processed by both the considered models (baseline and adapted). Specifically, the participants viewed six videos for each model, resulting in a comprehensive evaluation of adaptation effectiveness. The study aimed to gauge user preferences with regard to the stability and qualitative improvement and perceptions regarding the adapted variants in comparison to the baseline models. The findings of the conducted user study are presented in Fig. 3. On average, 81% of the users preferred the adapted results produced by the model proposed in [1], and 80% of the users preferred the same for the model proposed in [2].

6.1.2 Recurrent VS. Non-recurrent

Please note that we also extend the inference strategy of the model proposed in [1] to a frame recurrent setting in which the synthesized frames from previous timesteps were used as inputs for synthesizing current frames. Doing so, without the proposed meta-training, results in wobble artifacts as

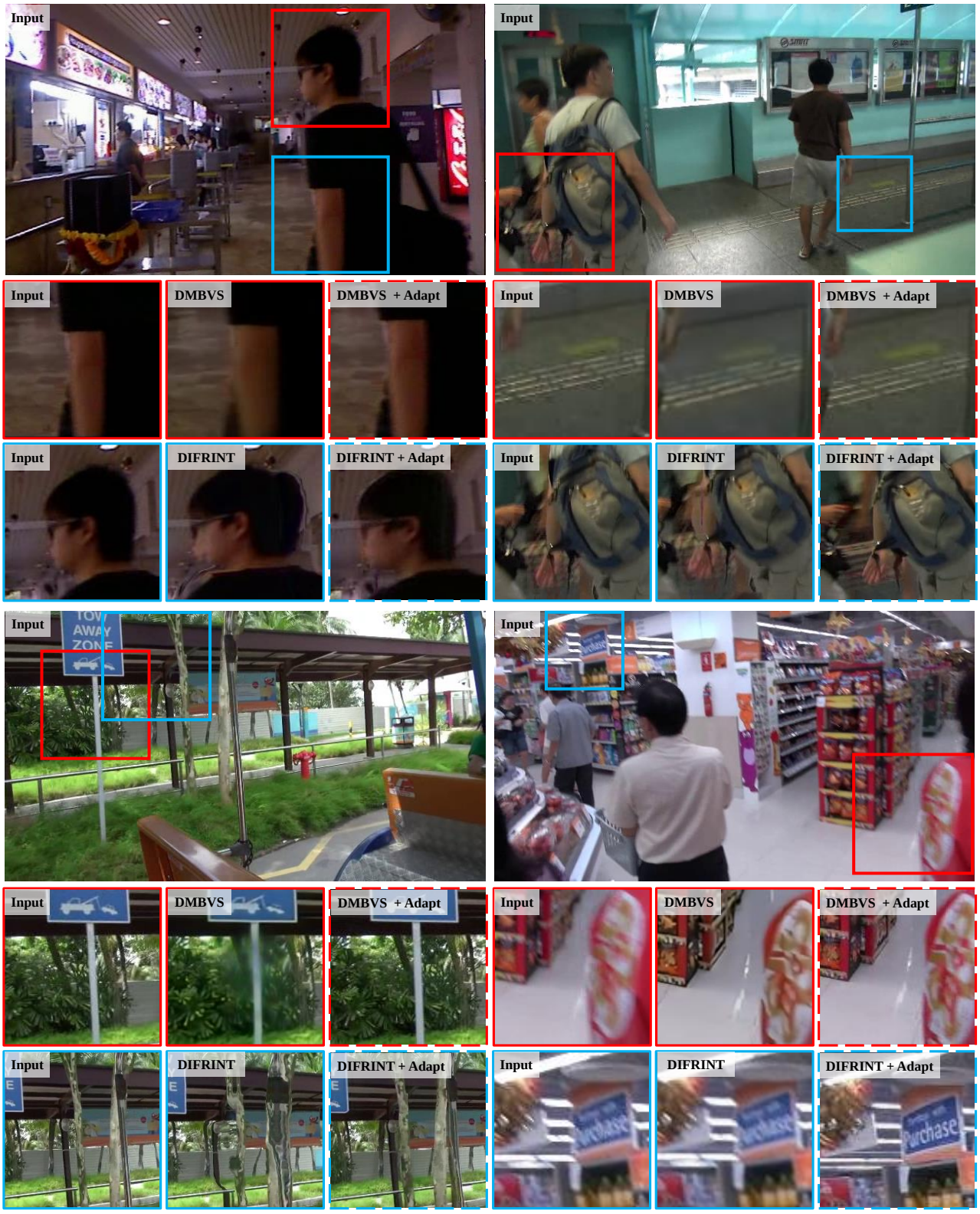

Figure 2. **Expanded view of qualitative comparison with baselines.** This figure presents the qualitative comparison with the baseline methods. The adapted results are highlighted with a dotted outline. The proposed fast-adaptation not only improves the stability of the baseline methods but also improves the visual quality of the stabilized videos.

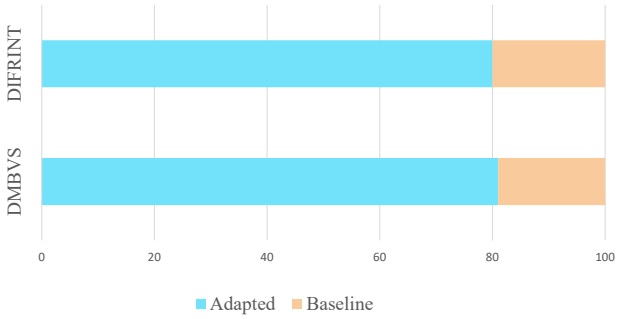

Figure 3. **Baseline VS. Adapted results.** This bar chart highlights the user preferences collected through the conducted user study concerning the comparison of adapted and baseline variants of the considered models. A significant majority of the participants preferred the adapted results for both of the considered models.

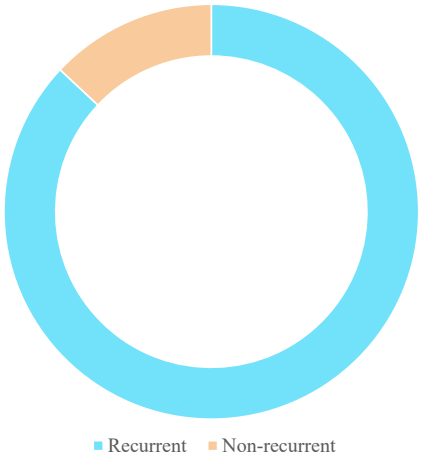

Figure 4. **Recurrent VS. Non-recurrent** This donut chart highlights the user preferences results of the conducted user study concerning the recurrent and non-recurrent variant of the model proposed in [1]. A majority of the participants preferred the results produced by the recurrent variant. Please note that these results were produced with the same hyperparameters and number of adaptation iterations.

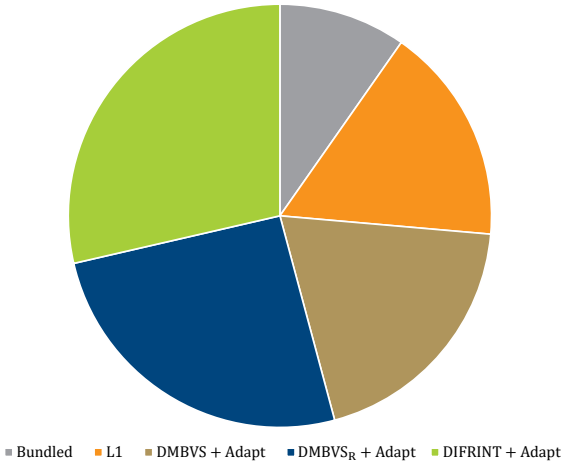

Figure 5. **Comprehensive user study.** This figure presents the findings of the comprehensive user study which compared the videos processed by the longstanding SOTA methods [6, 8] with the scene adaptive variants of the models proposed in [1, 2]. The majority of the users preferred the videos processed with the scene-adaptive approaches presented in this study.

presented in the main paper. The affine alignment loss in the inner loop helps mitigate these distortions as it encourages the stabilization model to synthesize content with similar shapes at different spatial locations. The proposed recurrence strategy for the model proposed in [1] achieves comparable stability but better perceptual quality (especially in videos where large parallax effects are observed) as compared with its non-recurrent variant (as evident from the results presented in the supplementary video). A user study was conducted involving 40 participants to investigate the impact of processing videos using the recurrent and non-recurrent adapted variants of this model. During the study, all the participants were presented with a total of 6 randomly sampled videos that had undergone processing by these variants. The order of the video presentation was also randomized to ensure impartial evaluation. The objective of the study was to gather insights into participant preferences and perceptions regarding the two processing variants, shedding light on the effectiveness of the recurrent and non-recurrent approaches. It is worth mentioning that all the videos were produced with the same number of adaptation iterations $m = 1$ and the same ratio of weights for stability and quality losses during adaptation for both variants. The videos processed through the recurrent model (especially in the *"Parallax"* and *"Zooming"* category) were significantly better in terms of temporal consistency (as evident from the accompanied supplementary video). We present the findings of the conducted user studies in Fig. 4. On average, $\sim$80% of the users preferred the videos processed through the recurrent variant.

6.1.3 Comparison with SOTA

In order to assess the user preference for stability and quality of videos processed through the adapted models, in comparison to the longstanding SOTA methods, we conducted another user study with the same number of participants to comprehensively evaluate various methods for video stabilization. In addition to selecting the preferred video, the users were also asked to record their reasonings for selecting specific videos. The study maintained the same settings, where participants were shown randomly sampled videos processed through all the discussed methods presented in the main paper for a comprehensive comparison.

In total, each participant viewed 9 videos. Through this

study, the aim was to elicit participant feedback and preferences regarding the different video stabilization methods, thus contributing to a deeper understanding of their comparative effectiveness and potential adaptability. On average, 28% of the users preferred the videos stabilized through the scene adaptive variant of the model proposed in [2] and remarked that the videos had smooth transitions and the content was preserved. On the other hand, $\sim$25% of the users selected videos processed with the scene adaptive recurrent extension of the model presented in [1] and $\sim$17% of the users selected the videos processed through the non-recurrent scene adaptive variant of the model proposed in [1], despite achieving a relatively lower stability score. When asked, a majority of the users attributed their preference to the superior quality of these videos. As for the methods presented in [8] and [6], on average $\sim$10% of the users preferred the videos stabilized through the bundled camera paths method [8] (named Bundled) and $\sim$17% of the users selected the videos stabilized through the method proposed in [6] (termed L1). The majority of the users remarked that the videos processed through these methods (Bundled and L1) compromised the quality of the videos and contained a significant loss of the original content. Please note that the comparative video results are also provided in the accompanied supplementary video.

6.2. Computational complexity

We evaluated the average runtime per frame (640×360) for each method and summarized the findings in Tab. 3. Please note that $\text{Adapt}^{(*)}_{100}$ has the time complexity of $O(1)$ for videos longer than 100 frames.

DMBVS	Runtime (s)	DIFRINT	Runtime (s)
Baseline	0.2342	Baseline	0.0711
$\text{Adapt}^{(1)}_{100}$	+ 0.1733	$\text{Adapt}^{(1)}_{100}$	+ 0.0767
$\text{Adapt}^{(5)}_{100}$	+ 0.8508	$\text{Adapt}^{(5)}_{100}$	+ 0.4188
Finetuning	2.0778	Finetuning	0.8170

Table 3. The presented time for adapted models is the average overhead time for each adaptation iteration (on video clips with 100 frames). The finetuning time varies with the number of frames and is average over 1000 frames. Please note that the "+" in front of the adaptation time highlights the additional time required for adaptation over the normal inference time.

Please note that the average adaptation overhead is for DMBVS [1] for $\text{Adapt}^{(1)}_{100}$ is lower than the inference time because the adaptation is conducted on patches of 320×320 from the video which has a spatial resolution of 640×320. Also please note that the presented results for the adapted methods are overhead results excluding the inference time of these models.

7. Limitations and Future work

While our proposed algorithm demonstrates promising improvements in terms of stability and quality of the full-frame video stabilization methods, certain limitations exist. Notably, the dependence of the DIFRINT [2] model on optical flow estimation can introduce occasional temporal artifacts in scenarios involving occlusion and disocclusion. Nevertheless, it is noteworthy that the adaptation stages significantly mitigate the frequency of occurrence of these artifacts, enhancing the overall quality of results as compared to their baseline methods. Similarly, the DMBVS [1] model, while effective, necessitates a substantial number of adaptation iterations to achieve further improvements in the stability score. Addressing these limitations holds great potential for future work, enabling expedited processing and enhanced results. Another potential for faster adaptation could be the integration of a key-jerk localization module which can find the most unstable sequences from the video under consideration, offering the potential to streamline and accelerate the adaptation process, thus advancing the robustness and efficiency of the proposed algorithm.

8. Supplementary video

We provide the comparative qualitative results of all the methods discussed in the main paper in the accompanied supplementary video and kindly request the reviewers to access it through a third-party media player like *VLC*. Due to the size limitation of the supplementary materials, the video was compiled on 1280×720, please use an appropriate window size to avoid unintentional artifacts or blur. We humbly apologize for any inconvenience that this may cause.

References

[1] Muhammad Kashif Ali, Sangjoon Yu, and Tae Hyun Kim. Deep motion blind video stabilization. *arXiv preprint arXiv:2011.09697*, 2020. 2, 3, 5, 6

[2] Jinsoo Choi and In So Kweon. Deep iterative frame interpolation for full-frame video stabilization. *ACM TOG*, 2020. 2, 3, 5, 6

[3] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, 2017. 1

[4] Jerin Geo, Devansh Jain, and Ajit Rajwade. Globalflownet: Video stabilization using deep distilled global motion estimates. In *Winter Conference on Applications of Computer Vision (WACV)*, 2023. 1, 2

[5] Edward Grefenstette, Brandon Amos, Denis Yarats, Phu Mon Htut, Artem Molchanov, Franziska Meier, Douwe Kiela, Kyunghyun Cho, and Soumith Chintala. Generalized inner loop meta-learning. *arXiv preprint arXiv:1910.01727*, 2019. 2

[6] Matthias Grundmann, Vivek Kwatra, and Irfan Essa. Auto-directed video stabilization with robust l1 optimal camera paths. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 5, 6

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 2

[8] Shuaicheng Liu, Lu Yuan, Ping Tan, and Jian Sun. Bundled camera paths for video stabilization. *ACM TOG*, 2013. 3, 5, 6

[9] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[10] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[11] Miao Wang, Guo-Ye Yang, Jin-Kun Lin, Song-Hai Zhang, Ariel Shamir, Shao-Ping Lu, and Shi-Min Hu. Deep online video stabilization with multi-grid warping transformation learning. *IEEE Transactions on Image Processing (TIP)*, 2018. 2