Supplemental Material for WACV 2023 'GlobalFlowNet: Video Stabilization using Deep Distilled Global Motion Estimates'

Jerin Geo James Devansh Jain Ajit Rajwade
Indian Institute of Technology Bombay
{jeringeo,devanshdvj,ajitvr}@cse.iitb.ac.in

1. Contents of the Supplemental Material

The supplemental material contains this pdf. The source code and the video stabilization results are publicly available at https://github.com/GlobalFlowNet/GlobalFlowNet

2. Compressibility of Global Motion

Referring to Sec. 2.2 ('Low-pass Filter Module') of the main paper, we observe that the global motion can be reconstructed to a high degree of accuracy with a very small number of DCT coefficients (cutoff frequency with magnitude less than or equal to 10). This is illustrated in Fig. 1 in this supplemental document.

3. Architecture: GlobalFlowNet

A more detailed diagram of the teacher-student network is presented in Fig. 2.

4. Visualization of Global Motion

Some example visualizations of the flow produced by our network are presented in Fig. 3.

5. Estimation of Affine Parameters from Global Motion

The parameters of the (partial) affine motion between adjacent frames are estimated in the following manner, starting from the global motion $\boldsymbol{f_S}$ obtained from the teacher-student network described in the main paper:

1. The translations t_x, t_y are estimated by computing the median of the values of $\boldsymbol{f_S}$ in the X and Y axes respectively.

2. Let $\boldsymbol{U}$ be a uniform Cartesian grid of the same size as $\boldsymbol{f_S}$. Let $\boldsymbol{U'} = \boldsymbol{U} + \boldsymbol{f_S}$. Now $\boldsymbol{U}$ represents the location of source points in the source image and $\boldsymbol{U'}$ represents their corresponding locations in the target image. Let these be converted into polar coordinate systems as $U_{l,\theta}$ and $U'_{l,\theta}$. We estimate the rotation angle r (a parameter of the partial affine motion) as the median of differences in the θ (angular) component of $U'_{l,\theta}$ and $U_{l,\theta}$. Similarly, the scale parameter is estimated as the median of the ratios of the l components of $U_{l,\theta}$ and $U'_{l,\theta}$.

This method of fitting parameters is computationally very efficient and does not require any iterated optimization.

6. Qualitative Evaluation of Video Stabilization Outputs

In the 'Results' folder, three videos can be found. These videos are as follows:
(1) **Results.mp4** : This video shows a comparison of our algorithms GLOBALFLOWNET-AFFINE and GLOBALFLOWNET-FULL with recent deep learning based techniques such as 'Learning Video Stabilization' approach (LVS) [7] (which was shown to be faster and superior to [6]), PwStableNet [8] and the 'deep multi-grid warping' (DMGW) technique from [5]. We also compare our results with two good quality classical algorithms: SteadyFlow [3] and the 'bundled camera paths' BCP approach from [2]. We did not compare with [1] due to lack of availability of their code or video results. We did not wish to compare with third party implementations of their technique. For SteadyFlow, we used results provided by the authors. One can observe superior stability and lesser geometrical distortions in our technique as compared to others.

In the accompanying video, we have written down our observations regarding each result when comparing with other deep learning approaches. In some cases, we also plot a few *salient feature point trajectories* in the videos stabilized by each method. One can observe that the shape of these trajectories is *smoother* in our method than other techniques such as [7].
(2) **Ablation.mp4**: This video presents a comparison of

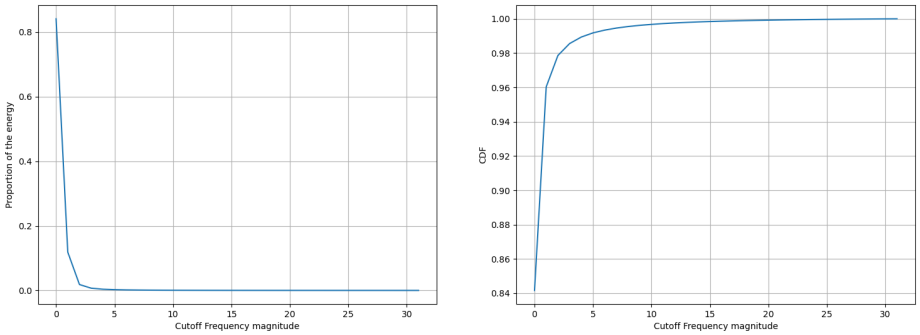

Figure 1: Left: Decay of the magnitude of DCT coefficients of a warp field (averaged across 100 videos with 300 frames per video) versus frequency magnitude $\sqrt{u^2 + v^2}$ (for frequency (u, v)); Right: CDF (computed across the same dataset as on the left) of the proportion of warp field magnitude versus cutoff frequency magnitude.

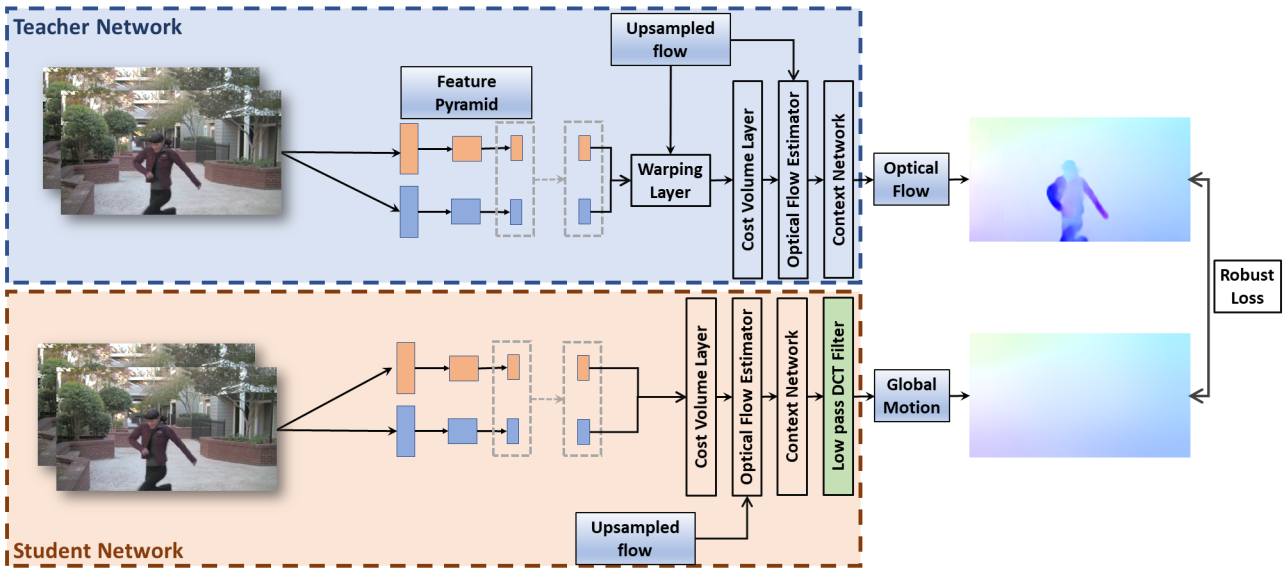

Figure 2: Refer to Fig. 1 of the main paper. Top: The teacher component, same as the PWC-Net architecture from [4]; Bottom: Our modified 'student' architecture GLOBALFLOWNET. The estimate from each layer in the student as well as teacher acts as an initial condition for the successive layer.

our Stage 1 and Stage 2 results, viz. GLOBALFLOWNET-AFFINE and GLOBALFLOWNET-FULL. Since affine transformation is not a good representation for the global motion, Stage 1 leaves behind some spatial distortions in the stabilized video. These spatial distortions get corrected in the stage 2. This is demonstrated in this video.

(3) **Limitation.mp4**: As mentioned in the main paper, the second stage GLOBALFLOWNET-FULL of our algorithm will produce sub-optimal results if the area of a *single dominant* foreground object is large and comparable to that of the background (multiple small foregrounds do not pose a problem). In such a case, the motion estimates will be biased towards either the foreground or the background, producing motion artifacts. However, *multiple small foregrounds* do not pose a problem, even if their *combined area* is large and comparable to that of the background. The video **Limitation.mp4** has a set of results on three videos to demonstrate this. (For an explanation regarding this phenomenon, please see the **next paragraph**.) The first video contains multiple moving objects in the scene, occupying a large portion of the scene combined. The second contains a single large moving object. The third contains a very large moving object. One can observe that our second stage GLOBALFLOWNET-FULL produces sub-optimal results in the third video, although it works well in the first two videos. However, GLOBALFLOWNET-AFFINE produces good results even with the third video.

In the absence of any independent foreground, the optical flow is dominantly represented by low frequencies (see Fig. 1). The foreground motion is an outlier to the back-

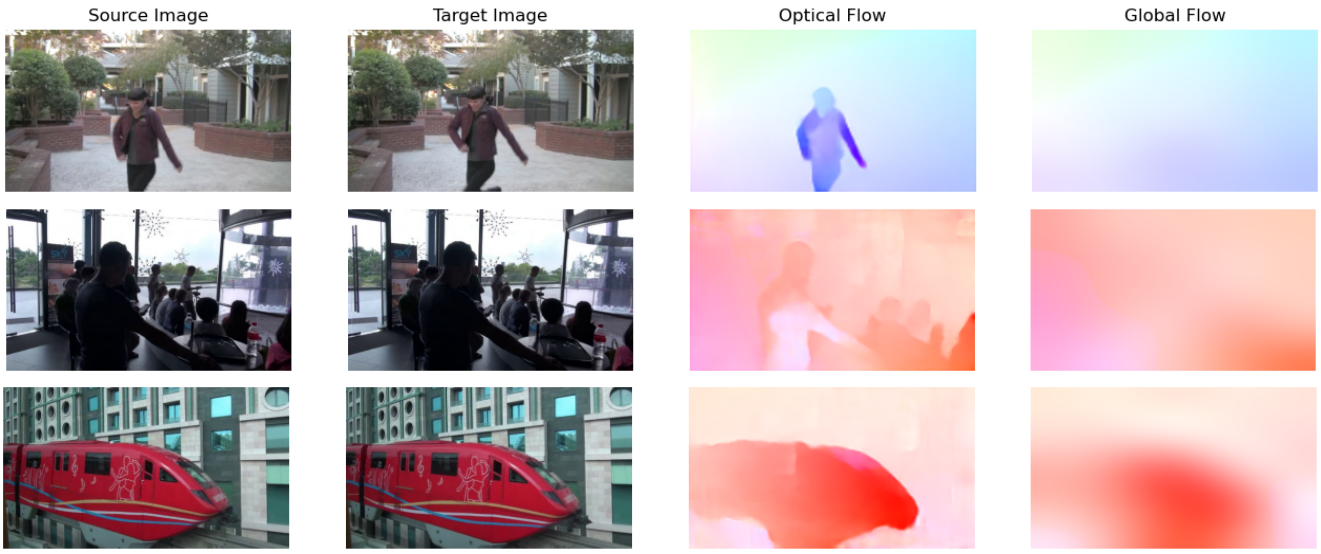

Figure 3: Visualization of Global Motion: In each row, from left to right: source and target images (two consecutive frames from a video), optical flow between the two image, and global motion between the two images. The global motion is estimated by GLOBALFLOWNET. In the first and second row one can notice that GLOBALFLOWNET has estimated the global motion by ignoring the foreground objects as discussed in Sec. 2.3 of the main paper. Notice that the network estimates the global motion quite well even in the presence of multiple foreground objects. However, in the last row, when the size of a single foreground object is comparable to the background object the network fails to predict the global motion. This is discussed in detail in the section below.

ground model. It gets filtered out due to the robust loss function (Eqn. 1 of the main paper) and because it also requires higher frequency components for accurate representation, whereas the student network forces a low frequency representation for the flow due to the Low Pass Filter Module (see Sec. 2.2 of the main paper). In case of a single large foreground of size comparable to the background, the foreground and background motion can get interchanged, leading to distortions. However multiple small foregrounds with different motions will still get filtered out, as each of these foregrounds act as individually sparse outliers to the background model. This is confirmed by the good performance in our experimental results on many videos of different categories having multiple foregrounds (see **Results.mp4** in supp. mat.).

References

[1] M. Grundmann, V. Kwatra, and I. Essa. Auto-directed video stabilization with robust l1 optimal camera paths. In *CVPR*, 2011. 1

[2] S. Liu, L. Yuan, P. Tan, and J. Sun. Bundled camera paths for video stabilization. *ACM TOG*, 32(4), 2013. 1

[3] S. Liu, L. Yuan, P. Tan, and J. Sun. Steadyflow: Spatially smooth optical flow for video stabilization. In *CVPR*, 2014. 1

[4] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 2

[5] M. Wang, G.-Y. Yang, J.-K. Lin, S.-H. Zhang, A. Shamir, S.-P. Lu, and S.-M. Hu. Deep online video stabilization with multi-grid warping transformation learning. *IEEE TIP*, 2019. 1

[6] J. Yu and R. Ramamoorthi. Robust video stabilization by optimization in CNN weight space. In *CVPR*, 2019. 1

[7] J. Yu and R. Ramamoorthi. Learning video stabilization using optical flow. In *CVPR*, 2020. 1

[8] M. Zhao and Q. Ling. PWStableNet: Learning pixel-wise warping maps for video stabilization. *IEEE Transactions on Image Processing*, 29, 2020. 1