Neural Re-rendering for Full-frame Video Stabilization Supplementary Material

YU-LUN LIU, National Taiwan University
WEI-SHENG LAI, Google
MING-HSUAN YANG, Google, UC Merced
YUNG-YU CHUANG, National Taiwan University
JIA-BIN HUANG, Virginia Tech

ACM Reference Format:
Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. 2021. Neural Re-rendering for Full-frame Video Stabilization Supplementary Material. 1, 1 (February 2021), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 OVERVIEW

In this supplementary material, we present additional results to complement the main manuscript. First, we provide more training details in Section 2. Then, we present more details and results of our experiments, including the statistics of datasets, definition of evaluation metrics, and per-category quantitative results in Section 3. Next, we analyze the temporal coherence of our method and state-of-the-art approaches in Section 4. We also demonstrate additional applications of our method on video completion and FOV expansion in Section 5. Finally, we describe the details of our user study in Section 6. In addition to this document, we also provide an interactive html interface to view the video results to compare with state-of-the-art methods.

2 TRAINING DETAILS

We set the weighting coefficients of pretrained VGG-19 network to: $\lambda_1 = 1/2.6$, $\lambda_2 = 1/4.8$, $\lambda_3 = 1/3.7$, $\lambda_4 = 1/5.6$, $\lambda_5 = 10/1.5$.

As described in Section 3.3 of the main paper, to generate the training data, we sample a 7-frame sequence from a video and apply random cropping to simulate a shaky input video. Fig. 1 visualize the process we synthesize the training data.

3 QUANTITATIVE EVALUATION

3.1 Dataset statistics

Table 1 summarizes the statistics of the two test datasets: the NUS dataset [Liu et al. 2013] and the selfie dataset [Yu and Ramamoorthi 2018] used in the quantitative and visual comparisons.

XXXX-XXXX/2021/2-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3.2 Evaluation metrics

Cropping ratio. The cropping ratio measures the remaining frame area after cropping off the irregular boundaries and undefined pixels due to warping. A larger cropping ratio indicates less cropping and preserves more video content. We first fit a homography between the input and output frames. The cropping ratio of a video is calculated by averaging the scale component of the homography across the entire video.

Distortion. Distortion is measured by the anisotropic scaling of the homography between the input and output frames. Specifically, it can be computed by the ratio of the two largest eigenvalues of the affine part in a homography matrix. The distortion score is defined as the minimal ratio across all the frames in a video.

Stability. This metric measures the stability and smoothness of the stabilized video. The frequency-domain analysis is performed on the 2D camera motion of the output video. Two 1D temporal signals are extracted from the translation and rotation components from each path. The ratio of the sum of the lowest (the 2nd to the 6th) frequency energy over the total energy is computed. The final score is obtained by taking the minimum across the entire video.

Accumulated optical flow. This metric is defined by accumulated optical flow over the entire video:

$$\frac{1}{wh(T-1)} \sum_{t=1}^{T-1} \sum_{x=1}^{h} \sum_{y=1}^{w} (\|F_{t\to t+1}(x,y))\|_2 + \|F_{t+1\to t}(x,y))\|_2), \quad (1)$$

where w is the frame width, h is the frame height, t is the frame index of time, T is the total number of frames, x, y are the pixel coordinates, and $F_{t\to t+1}$ is the optical flow from time t to $t+1$. The optical flow is normalized by its frame size in order to compare videos with different resolutions. In this metric, a smaller value means that the video has smaller motion, indicating a more stable result. This score value is further normalized by the score of the input video to show the improvement. We use RAFT [Teed and Deng 2020] to compute the bidirectional optical flow for each video.

3.3 Per-category quantitative results

Fig. 2 shows the per-category evaluation on the NUS dataset [Liu et al. 2013]. Our method and DIFRINT [Choi and Kweon 2020] has the highest cropping ratio (close to 1) as both methods generate full-frame results without cropping. For the distortion and stability metrics, our method performs on par with the best state-of-the-art methods (DIFRINT [Choi and Kweon 2020] in distortion and Adobe

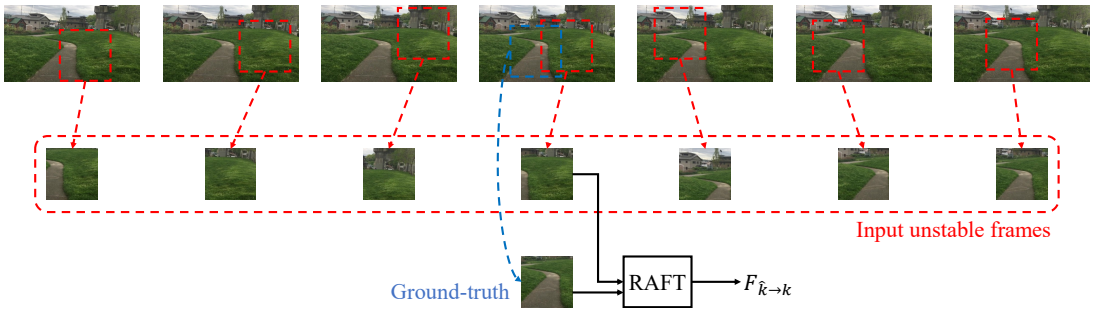
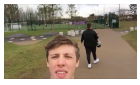


Fig. 1. **Training data synthesis.** We apply random cropping to synthesize unstable input videos.

Table 1. **Dataset summarization.** The NUS dataset [Liu et al. 2013] contains six categories, including *s*imple, *q*uick rotation, *z*ooming, *l*arge parallax, *c*rowd, and *r*unning. The selfie dataset [Yu and Ramamoorthi 2018] has 33 selfie video clips collected from the Internet. In total, we evaluate other state-of-the-art methods and our method on these 177 video clips.

Dataset	NUS						Selfie
Category	Simple	Quick Rotation	Zooming	Large Parallax	Crowd	Running	Selfie
Sample frame							
Number of video clips	23	29	29	18	23	22	33

Premiere Pro 2020 warp stabilizer in stability). Our method obtains the lowest accumulated optical flow in most categories except the rotation and crowd category, which have more challenging cases. Generally, the proposed method is robust and performs well on different scenarios.

4 TEMPORAL COHERENCY

We slice the center vertical line of the output videos over time to show the temporal coherency of our method. Fig. 3 shows that our method achieves smoother temporal coherency comparing to other methods and the input video.

5 ADDITIONAL APPLICATIONS

5.1 Video completion

We demonstrate that our method can also be applied to video completion. We use the same flow smoothing [Yu and Ramamoorthi 2020] and apply a state-of-the-art video completion method [Gao et al. 2020] to fill in the blank regions. Note that [Gao et al. 2020] can be regarded as an image-based fusion method. Fig. 4 shows that [Gao et al. 2020] generates visible artifacts due to wrong optical flows, while our method produces more visually pleasing results by hybrid-space fusion.

5.2 FOV Expansion

Our method can be extended to generate videos with a larger field of view than the input videos. Fig. 5(a) shows a video frame where the camera is moving horizontally to the right. Our method can use information from neighbor frames to expand the horizontal FOV in Fig. 5(b). Fig. 5(c) shows a video frame where the camera is zooming out. Our method can expand the FOV on all directions to include more content in a single view.

6 USER STUDY

Fig. 6 shows a screenshot of our interactive webpage for conducting the user study. For each sequence, the user is asked to select the winner out of two comparing methods in terms of three criterion. One of the comparing videos is from our method. Note that we also provide the unstable input video as a reference. The sequence order and left-right order are randomly shuffled.

REFERENCES

Jinsoo Choi and In So Kweon. 2020. Deep iterative frame interpolation for full-frame video stabilization. *ACM TOG* (2020).
Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. 2020. Flow-edge Guided Video Completion. In *ECCV*.
Matthias Grundmann, Vivek Kwatra, and Irfan Essa. 2011. Auto-directed video stabilization with robust l1 optimal camera paths. In *CVPR*.
Shuaicheng Liu, Lu Yuan, Ping Tan, and Jian Sun. 2013. Bundled camera paths for video stabilization. *ACM TOG* (2013).
Zachary Teed and Jia Deng. 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. (2020).
Miao Wang, Guo-Ye Yang, Jin-Kun Lin, Song-Hai Zhang, Ariel Shamir, Shao-Ping Lu, and Shi-Min Hu. 2018. Deep online video stabilization with multi-grid warping transformation learning. *IEEE Transactions on Image Processing* (2018).
Jiyang Yu and Ravi Ramamoorthi. 2018. Selfie video stabilization. In *ECCV*.
Jiyang Yu and Ravi Ramamoorthi. 2020. Learning Video Stabilization Using Optical Flow. In *CVPR*.

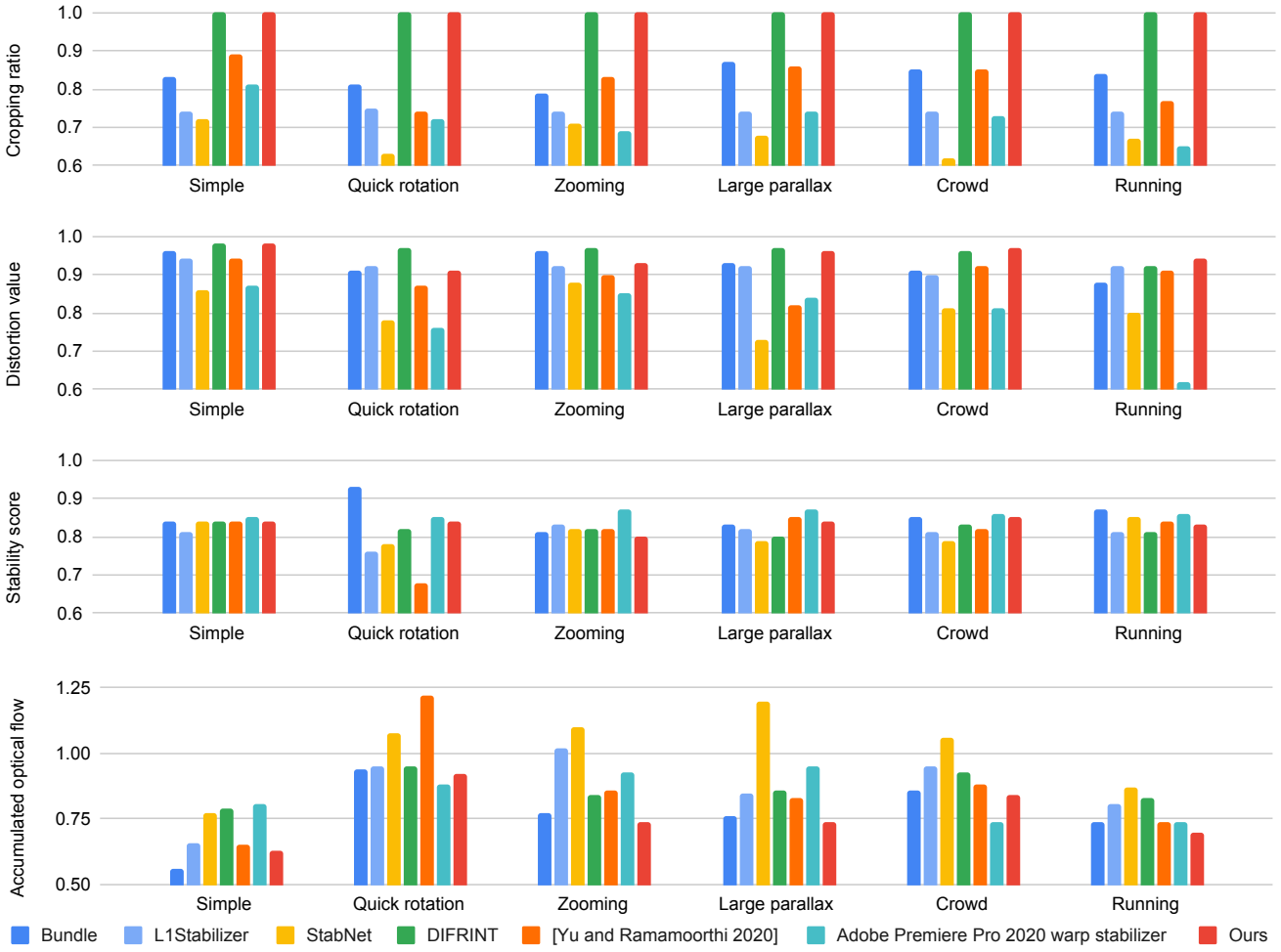
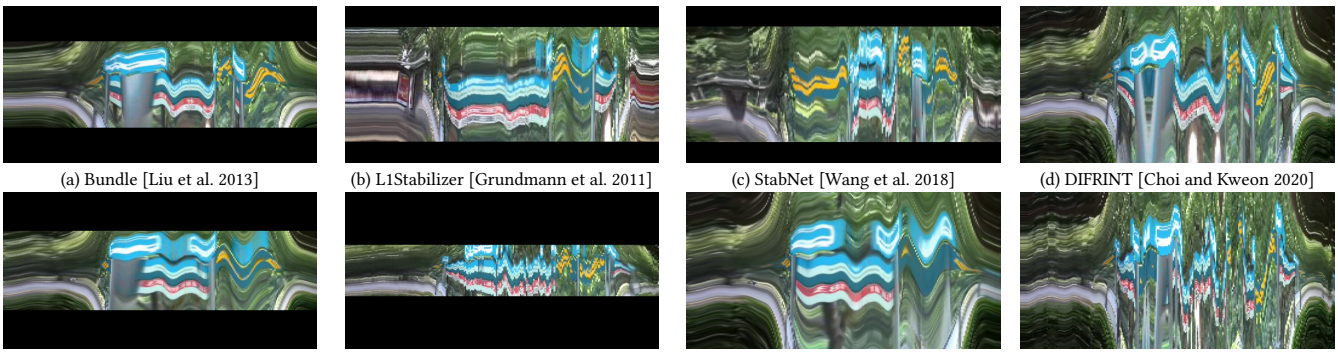


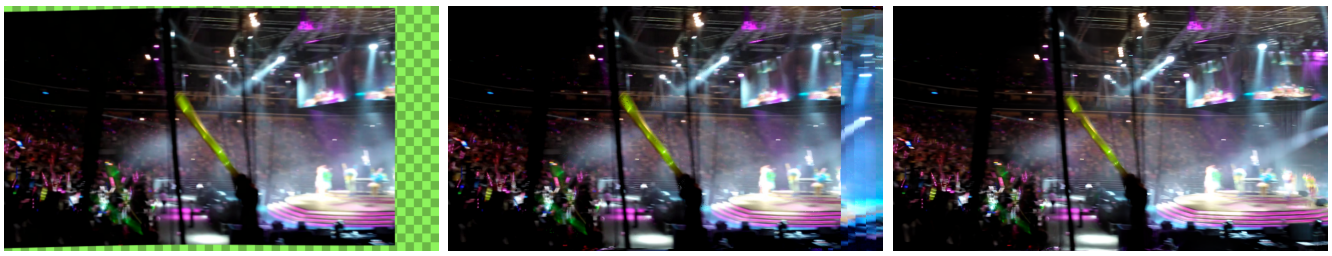
Fig. 2. **Evaluation on each category of the NUS dataset [Liu et al. 2013].** For cropping ratio, distortion value, and stability score, a larger value indicates a better result. For accumulated optical flow, a smaller value indicates a better result. Our method is robust in all categories, achieving state-of-the-art or comparable performance with existing approaches.



(a) Bundle [Liu et al. 2013] (b) L1Stabilizer [Grundmann et al. 2011] (c) StabNet [Wang et al. 2018] (d) DIFRINT [Choi and Kweon 2020]

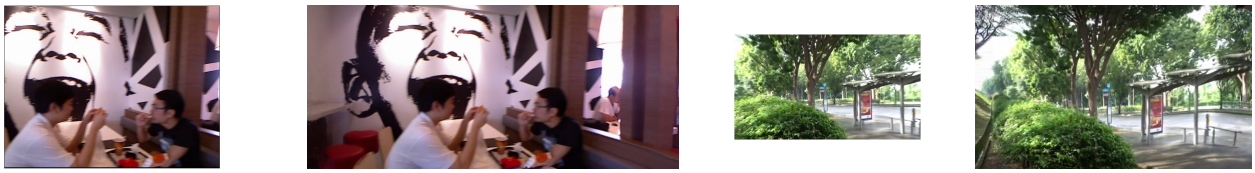
(e) [Yu and Ramamoorthi 2020] (f) Adobe Premiere Pro 2020 warp stabilizer (g) Ours (h) Input

Fig. 3. **Temporal coherency comparisons to state-of-the-art methods.** Bundle [Liu et al. 2013], Deep online video stabilization [Wang et al. 2018], and Adobe Premiere Pro 2020 warp stabilizer suffer from large cropping. L1Stabilizer [Grundmann et al. 2011] and [Yu and Ramamoorthi 2020] generate smooth videos with cropping. DIFRINT [Choi and Kweon 2020] produces glitching motions. Our stable video is smooth and preserves the most content in the input video.



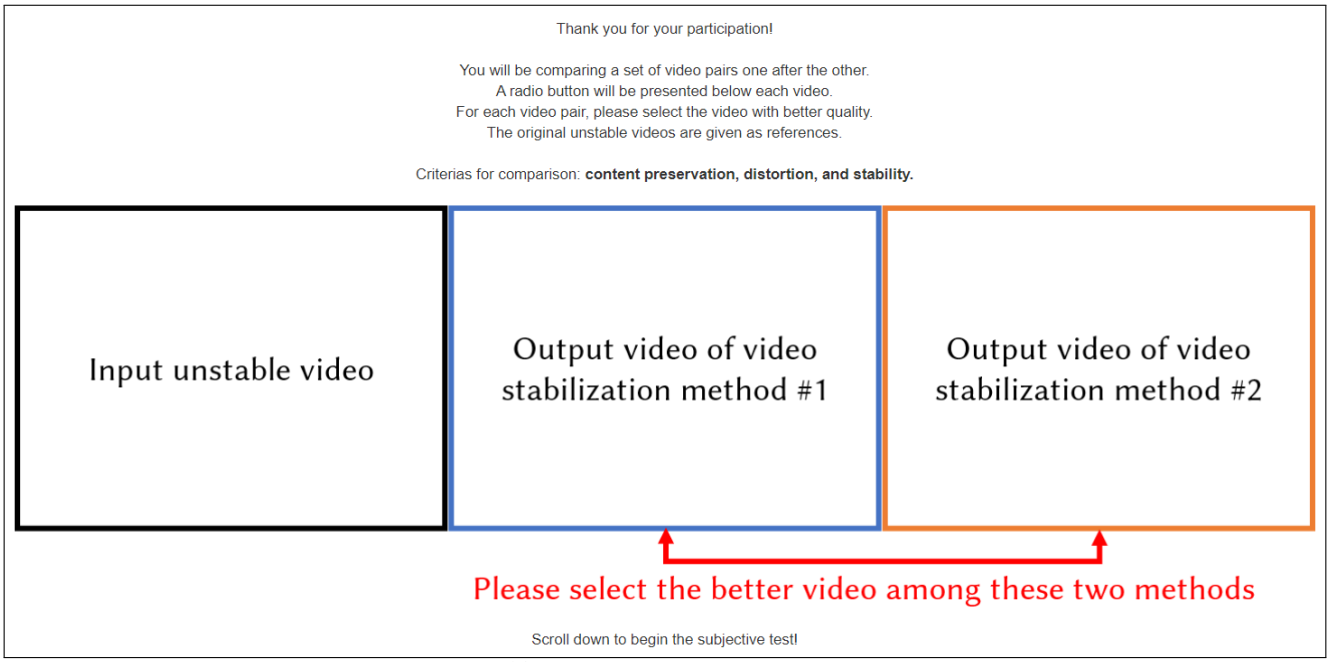
(a) [Yu and Ramamoorthi 2020] (b) [Yu and Ramamoorthi 2020] + [Gao et al. 2020] (c) Ours

Fig. 4. **Comparisons to the video completion method [Gao et al. 2020].** (a) [Yu and Ramamoorthi 2020] crops off the blank regions caused by warping. (b) The video completion method [Gao et al. 2020] utilizes stable neighbor frames to recover the missing pixels. However, the results contain visible discontinuity artifacts due to wrong flows. (c) Our method produces full-frame and artifact-free results.



(a) Input (b) Ours + FoV expansion (c) Input (d) Ours + FoV expansion

Fig. 5. **Effect of FoV expansion.** (a) (c) Input unstable frames. (b) (d) Our method utilizes nearby information to recover the missing regions due to motion compensation. It can also expand the field-of-view of the output videos. The example on the left is from the parallax category. The example on the right is from the zooming category.



(a) Instruction of the user study.

(b) Interface of the user study.

Fig. 6. **User study website.** We ask the user to answer three questions for each video sequence, which corresponds to the cropping ratio, distortion value, and stability. *Left*: Input unstable video for reference. *Middle*: Randomly selected method #1. *Right*: Randomly selected method #2. Note that one of the compared methods is ours.