Deep Motion Blind Video Stabilization

Muhammad Kashif Ali*
kashifali@hanyang.ac.kr
Sangjoon Yu*
kiddyu1991@gmail.com
Tae Hyun Kim†
taehyunkim@hanyang.ac.kr

Department of Computer Science
Hanyang University
Seoul, South Korea

Abstract

Despite the advances in the field of generative models in computer vision, video stabilization still lacks a pure regressive deep-learning-based formulation. Deep video stabilization is generally formulated with the help of explicit motion estimation modules due to the lack of a dataset containing pairs of videos with similar perspective but different motion. Therefore, the deep learning approaches for this task have difficulties in the pixel-level synthesis of latent stabilized frames, and resort to motion estimation modules for indirect transformations of the unstable frames to stabilized frames, leading to the loss of visual content near the frame boundaries. In this work, we aim to declutter this over-complicated formulation of video stabilization with the help of a novel dataset that contains pairs of training videos with similar perspective but different motion, and verify its effectiveness by successfully learning motion blind full-frame video stabilization through employing strictly conventional generative techniques and further improve the stability through a curriculum-learning inspired adversarial training strategy. Through extensive experimentation, we show the quantitative and qualitative advantages of the proposed approach to the state-of-the-art video stabilization approaches. Moreover, our method achieves $\sim 3\times$ speed-up over the currently available fastest video stabilization methods.

1 Introduction

The prevalent integration of high-quality cameras in hand-held devices, has enabled the general population to record the memorable moments of their life, but it still requires professional equipment to record stable videos. Thus, considerable literature has been devoted to solving the video stabilization problem. Despite the advances in the generative deep learning models, there is still a long way to go for deep-learning-based approaches to truly take over in video stabilization from the traditional reconstructive feature-tracking [19, 20] and trajectory optimization [8, 18] methods.

Recently, Wang et al. [28] released the DeepStab dataset, which is the first large-scale dataset for video stabilization. This dataset is captured with two synchronized cameras placed on a contraption fixed around the base of a mechanical stabilizer. The camera

*Equal contribution. †Corresponding author.

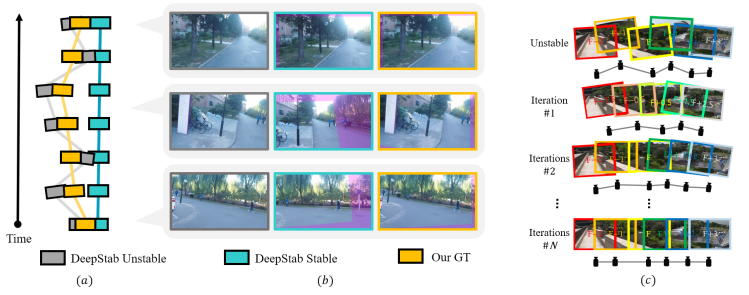

Figure 1: (a) An illustration of the perspective mismatch in the DeepStab dataset and our proposed Dataset Generation Pipeline (DGP). (b) Large non-overlapping regions (transparent purple zone) present in the DeepStab dataset [28] along with the minimized perspective difference in our dataset. (c) A visual description of iterative frame interpolation leading to visual stability and smooth camera trajectory.

placed on the physical stabilizer captures the stable video while the camera on the contraption rotates freely along the stabilizer and records unstable videos. Due to the rotational motion of the unstabilized camera, both of the recorded videos often contain a significant non-overlapping field-of-view, and a perspective mismatch (as shown in Figure 1). This inconsistency in the perspective makes it difficult for the models to learn the direct pixel-level spatio-temporal relations of unstable videos to their stable counterparts. Thus, video stabilization is generally defined with the help of dense optical flow estimation modules and the networks learn to warp the original frames instead of synthesizing them [32]. This warping generally entails a substantial cropping near the frame boundaries and temporal distortions in the stabilized videos. To overcome this problem, we provide a new dataset by extending and improving the idea of iterative frame interpolation leading to smooth motion trajectories as presented in [2] to generate stable and unstable training videos which virtually share the same perspective (highlighted in yellow in Figure 1 (b)). Through our experiments with the proposed dataset, we attempt to declutter and relieve the dependence on motion-awareness in the formulation of video stabilization pipelines, and demonstrate that full-frame video stabilization can be formulated with conventional network architectures and modules without explicit motion awareness. In addition, we further propose a contrastive motion loss and a temporal adversarial training strategy to produce more stable and temporally consistent full-frame videos. Our proposed stabilization network compares favorably to the currently available motion-aware solutions, and we summarize our contributions as follows:

- **Unsupervised dataset generation:** we introduce an unsupervised and extensible video-frame-interpolation-based strategy to produce equi-perspective stabilized videos from unstable videos captured from hand-held devices.
- **Motion blind full-frame video stabilization:** we declutter the overly complex video stabilization formulation and propose the first ever motion blind deep stabilization network with the help of the proposed equi-perspective dataset.
- **Curriculum Learning strategy:** we present a targeted sequential learning strategy where we allow the same network to focus on multiple aspects of stabilization in different stages.

2 Related Work

Liu et al. [16] proposed a 3D approach for this task, in which camera poses along with feature tracks were reconstructed in the 3D space, and the feature positions were projected along smoothened camera poses, whereas, Smith et al. [27] employed depth-aware cameras to do the same. However, these global 3D approaches cannot properly handle dynamic scenes including moving objects, and thus 2D transformations (e.g., homography) become more popular in video stabilization methodologies. In general, these 2D methods rely on tracking prominent features and stabilizing their trajectories along the motion path. The results produced by these methods generally need cropping around the borders and up-scaling to retain the original resolution of the input video. In addition, Buehler et al. [1] estimated the camera positions through shaky videos and rendered the frames at smoothened camera positions using a non-metric Image Based-Rendering method. Matsushita et al. [21] and Gleicher et al. [6] used simplistic 2D transformation mechanisms to warp the original frames. Whereas, Liu et al. [18] introduced a grid-based warping of frames for smoothing the feature trajectories. Grundmann et al. [8] proposed an L1-based cost functions for obtaining optimal camera trajectory for stabilized feature tracks, whereas, Liu et al. [17] proposed a similar approach but employed the eigen-trajectory smoothing technique. Wang et al. [29] and Goldstein et al. [7] also approached this task with optimization-based models to acquire feature tracks and camera position using epipolar geometry.

All of these methods relied heavily on feature tracks and ignored the underlying relation of independent motion of multiple objects in consecutive frames, which compelled Liu et al. [19, 20] to investigate applications of optical flow in the field of video stabilization. Their studies helped understanding the importance of inter-frame motion estimation in video stabilization and paved the path for modern video stabilization methods. Thus, Yu and Ramamoorthi et al. [31, 32] and Choi et al. [2] employed dense optical flow estimation modules to warp the neighboring frames to obtain smoother and better-quality videos. In particular, Yu and Ramamoorthi et al. [31] proposed a scene-specific optimization approach that estimates dense motion to optimize the network weights for each video and extended their approach in [32] to a generalized framework capable of handling complex situations including (de)occlusion and non-linear motion through warp fields.

Two pioneering methods implicitly using motion flow were proposed in [28, 30]. These methods employ generative adversarial networks and spatial transformer networks to learn the inter-frame motion and warp the frames for video stabilization. Wang et al. [28] proposed the DeepStab dataset and attempted to find a possible solution for video stabilization with a Siamese network containing a pre-trained ResNet50 model. Another attempt to train a pure image-based stabilizer using the DeepStab dataset without motion estimation was discussed in [32] and was termed as an “essential over-fitting task", because using this dataset for training can lead to the network learning an entirely different perspective of the same scene without the presence of any correlation or information about the perspective in the input unstable video frames.

Meanwhile, a new stabilization network based on dense optical flow estimation and video frame interpolation was proposed in [2] called DIFRINT. They achieve temporal stability by rendering interpolated frames between unstable frames and obtain full-frame video stabilization results. These attempts have helped us in pinpointing the shortcomings of the DeepStab dataset and have encouraged us to propose an equi-perspective dataset which can simplify the task of video stabilization.

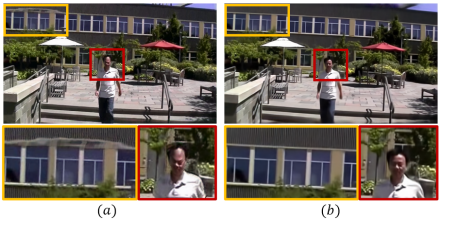

Figure 2: (a) Interpolated frame by DIFRINT [2]. (b) Interpolated frame by CAIN [3].

3 Dataset Generation Pipeline (DGP)

The generation of a labeled dataset for video stabilization is a challenging task. Before finalizing our dataset generation pipeline, we experimented with various techniques to pinpoint the missing link that hinders motion blind formulation of this task. Our experiments included training the same network on DeepStab [28] dataset and a dataset generated through random affine transforms. Through this experiment, we observed that the network trained with the synthetic dataset learns to better stabilize videos than the network trained on the DeepStab [28] dataset. In order to minimize the effect of large non-overlapping regions, we also experimented with downscaling the DeepStab [28] to an eighth of its original frame size. This downscaling operation reduced the overall inter frame motion. Even with these downscaled frames, we were unable to learn meaningful stabilization. Through these experiments, we concluded that learning the high-level reasoning for stabilization does not just require a minimized non-overlapping region but it also requires the target and input videos to share similar perspective in order to properly find correspondences. This is generally avoided in the motion aware techniques with the help of dense optical flow estimation which helps the model to differentiate between local and global motion present between the frames along with an abstract sense of jerkiness in unstable videos. In order to generate a new large-scale video stabilization dataset that fulfills these requirements, we draw motivation from DIFRINT [2], which is a frame-interpolation-based stabilization network. Specifically, the first part of their network is similar to the conventional video frame interpolation networks. Their intuition for incorporating this network in their pipeline was to achieve temporal stability by reconstructing frames with high-frequency jerks. Their pseudo frame interpolation network is trained for video stabilization, with a synthetic dataset generated using random

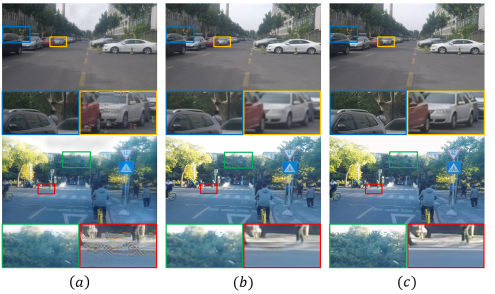

Figure 3: Comparison of frame interpolation methods used in iterative arrangement (20 iterations). (a) SepConv [24]. (b) CAIN [3]. (c) Proposed DGP.

affine transformations. This synthetic dataset lacks the complexity of real motion in dynamic scenes. Thus, this network faces difficulties in handling the differently moving objects in real-world scenarios, and often results in undesirable wobble like artifacts as shown in Figure 2 (a). Notably, in the modern frame interpolation methods [3, 24], this is compensated by training the networks on real-world videos containing complex inter-frame local and global motions as shown in Figure 2 (b). Based on these observations, we deduce that conventional video frame interpolation models (specifically trained for this task) can outperform DIFRINT [2] in handling dynamic motion scenarios and produce stabler and higher quality videos.

In general, video frame interpolation methods behave like a low-pass filter and produce a middle frame by blending the neighboring frames. This generation of the middle frame can remove high-frequency jitter present between the alternate frames. In an iterative arrangement, this approach can take into account the relative motion of all the frames present in the sequence, and generate a temporally consistent sequence free from sudden high-frequency jerks. It is worth noting that DIFRINT [2] enforces stability by skipping intermediary frames in their stabilization pipeline and generates the skipped frames with an assumption that the intermediary frames lie along a straight line and the inter-frame motion is strictly linear. These assumptions enforce temporal stability in fairly lesser iterations, but result in undesirable artifacts around the depth boundaries of distinctly moving objects, and a jagged progression in the stabilized videos. First, to solve these problems we propose a new dataset generation technique, which instead of skipping frames, generates the intermediate frames and uses these intermediate frames to reconstruct the original sequence which preserves the original progression of the video sequence. Secondly, we propose a refinement network which restores the integrity of iteratively generated data.

• **Iterative Frame Interpolation:** For the first part of our dataset generation pipeline, we tested two state-of-the-art frame interpolation methods SepConv [24] and CAIN [3] in an iterative arrangement and opted to use CAIN [3] as it produced better quality frames in our experiments (presented in Figure 3). Unlike DIFRINT [2], we use our frame interpolator without any rigid assumptions about the nature of inter-frame motion, and only compensate for high-frequency jerks through our iterative frame interpolation pipeline. This iterative stability comes at the cost of visual distortions and artifacts. Although the generated frames include lesser high-frequency jerks, various artifacts such as blur and color distortions are generated as shown in Figure 3. To overcome these problems, we additionally introduce a refinement network that restores the visual integrity of the generated frames.

• **Refinement Network:** To remove the artifacts introduced by the iterative frame interpolations, we introduce a *refinement* network. Our refinement network is based on ResNet [11] with a modified version of the channel attention module from CAIN [3] (as shown in Figure 4). Our modified attention module treats the features with a succession of space-to-depth operations followed by global average pooling, and a 1×1 convolution layer. The output of this layer is passed through a sigmoid function and then multiplied (element-wise) to the input features. We observed that the original frames contain unaltered high-quality regions necessary to restore the degraded interpolated frames. Thus, this network takes in an interpolated frame with its neighboring original (unstable) frames as input, and generates the restored version of the interpolated frame. Through customized losses and training strategy, we ensure that the network does not alter the spatial relations of the content present in the interpolated frame and only targets the artifacts introduced by the iterative interpolation. A comparison of the refined results with the interpolated results is provided in Figure 3. Please refer to our supplementary material for the detailed dataset generation, formulation,

configuration and training and testing strategies of this network. For the finalized DGP, the refinement network was integrated within the iterative frame interpolation pipeline. We introduced a refinement step after every k iterations of the frame interpolation network (i.e., CAIN), and repeated this setting for a definite number of times m (as illustrated in Figure 4) to acquire temporally stable and high-quality frames as shown in Figure 3 (c). In our experiments, we select k and m to be 4 and 5, respectively. Due to the space limitation, we present a visual ablation study to justify our choices of parameters k and m in the accompanied supplementary material. We utilize unstable videos from the DeepStab dataset along with videos acquired through the internet to generate our final video stabilization dataset.

4 Learning Motion Blind Video Stabilization

4.1 Re-formulation of Video Stabilization

The conventional deep video stabilization methodologies formulate the task of video stabilization with explicit motion estimation modules. Generally, these modules further complicate an already convoluted formulation with additional steps to process the calculated motion flows as described in [32]. Contrary to the normal convention, we propose a simplistic and straightforward formulation of the video stabilization through our generated dataset. The minimized perspective mismatch in our dataset assists the model to focus on the spatial relations between the stable and unstable videos. Various model architectures like U-Net [25] and ResNet [11] structures were tested before finalizing the baseline architecture for this task. We employ a modified version of a super-resolution network, ENet [26] (based on ResNet architecture) for our stabilization network. A very deep ResNet based architecture taking multiple input frames allows the network to exploit spatio-temporal information along with an extended receptive field as described in [28]. The architecture of the proposed network for video stabilization is shown in Figure 5. Our model takes in five consecutive unstable frames and produces the stabilized version of the middle frame. The number of input frames used for the stabilization network was evaluated empirically.

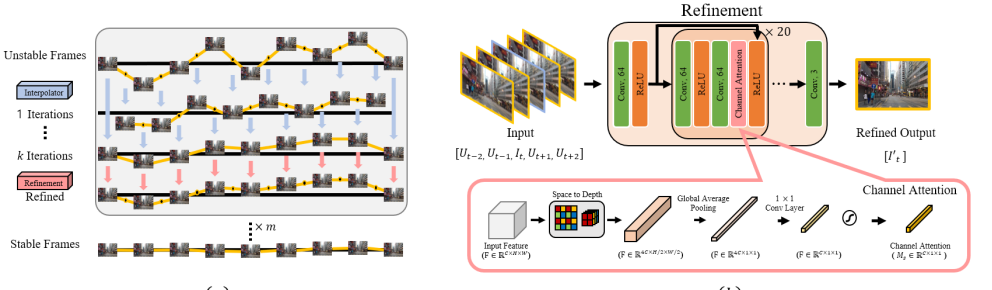

Figure 4: (a) Proposed DGP for video stabilization with integrated the refinement network. (b) The architecture and inference strategy of the refinement network.

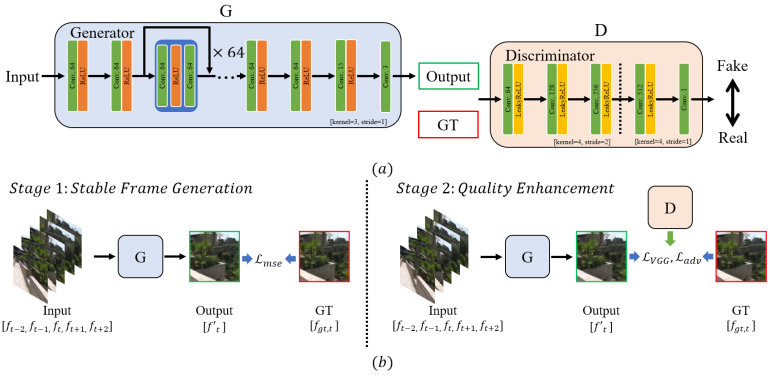

Figure 5: (a) Network architecture of the proposed stabilization network along with the discriminator for training. (b) Training strategy for Stage 1 and Stage 2 (left to right)

4.2 Training Strategy

Inspired from the ideas presented in [5, 30], we divide the task of learning motion blind video stabilization in to its different components and allow the model to focus on only one task at a time. With this strategy we can learn all the aspects of video stabilization through the same network. In the first two stages of our training, namely, *Stable Frame Generation* and the *Quality Enhancement*, we purposefully employ conventional methodologies to verify the effectiveness of the proposed dataset and the importance of equi-perspective training samples for this task. Whereas, in the third stage, *Strengthening*, we let the model focus on learning the abstract reasoning for improving the stability and temporal consistency.

4.2.1 Stage 1: Stable Frame Generation

In this stage, we train the network with a specific goal of generating stable frame f'_t from the five unstable input frames $(f_{t-2}, f_{t-1}, f_t, f_{t+1,}, f_{t+2})$ (as presented on the left side of Figure 5 (b)). During this stage, the perceptual quality of the generated frames is purposefully ignored as it can be enhanced in the upcoming stages with the help of an adversarial training strategy and a perceptual loss. During our experiments, it was observed that introducing a quality improvement loss at this stage significantly increased the convergence time. Therefore, at this stage, the model is trained with a specific goal of learning only the high-level reasoning necessary to justify the generated output frame f'_t from the input unstable frames. In this stage, we train the stabilization network with the $\mathcal{L}_2$-based reconstruction loss as,

$$\mathcal{L} = \left\| f'_t - f_{gt,t} \right\|_2^2, \tag{1}$$

where $f_{gt,t}$ is the frame acquired through our proposed DGP (described in Sec. 3).

4.2.2 Stage 2: Quality Enhancement

After the convergence with the $\mathcal{L}_2$ reconstruction loss in (1), the results produced by the network are stable but quite blurry [*]. The perceptual quality of these stable but blurry frames

[*]The quality of the results produced at this stage can be assessed through the supplementary text

can be improved by fine-tuning the network with a perceptual and an adversarial loss for their proven success in enhancing the visual quality of degraded images [13, 15, 23]. The primary loss used during this stage is a VGG based loss defined as follows:

$$\mathcal{L}_{\text{content}} = \left\| \phi(f'_t) - \phi(f_{gt,t}) \right\|_2^2, \tag{2}$$

where $\phi(\cdot)$ represents the *relu_3_3* layer of a pre-trained VGG-19 network. This loss ensures the preservation of high-level visual cues present as proposed in [12]. In addition to the perceptual loss, we also employ an adversarial training schema in this stage. The discriminator used in our work is shown in Figure 5 (a). It is a simple feed-forward network inspired by the discriminator used in [14] with alternating convolution and Leaky-Relu operations. The final loss for training in this stage is given by the following equation,

$$\mathcal{L} = \mathcal{L}_{\text{content}} + \lambda \cdot \mathcal{L}_{\text{adv}}, \tag{3}$$

Here, $\mathcal{L}_{adv}$ is the adversarial loss and λ denotes a user-parameter that controls the contribution of the adversarial loss in the optimization step. As for the adversarial loss we utilize WGAN-GP [9] for its success in similar quality improvement tasks such as [14, 15]. A brief inter-stage ablation study is provided in the accompanied supplementary material. At this stage we verify and prove the effectiveness of the proposed dataset and show that, the video stabilization pipelines can be simplified with the help of our proposed dataset containing pairs of stable and unstable training videos with a minimized perspective difference.

4.2.3 Stage 3: Strengthening

Wobble effect (as highlighted in Figure 2 (a)) is quite common in digitally stabilized videos. This effect occurs due to the motion compensation, and it can be minimized in motion aware approaches at the cost of stability. Since our model does not contain any explicit motion estimation module, we address this issue with the help of specialized losses for this task. During our experiments we observed that the natural video sequences do not contain these artifacts. Therefore, we propose a temporal discriminator that can differentiate between a natural sequence and an artificially generated one. With this intuition, we introduce a secondary discriminator which takes in 16 sequential frames of the generated videos along with the corresponding DeepStab [28] stable frames, and encourages the proposed stabilization network to generate wobble free frames. We also employ a contextual [22] and a perceptual loss [12] between the generated and the unstable frames for content preservation. In addition to these losses, we also propose a contrastive motion loss to enhance stability. This loss uses an off the shelf pre-trained Video ResNet-18 for action recognition as proposed in [10] to produce embeddings for the generated video sequences along with the corresponding DeepStab stable and unstable sequences. These embeddings are then used with a triplet loss [4]. The embeddings for the DeepStab stable, unstable and our generated sequences are used as anchor, negative and positive embeddings respectively. This loss minimizes the distance between the positive and anchor while maximizing the distance between the anchor and the negative embeddings. During the experimentation, an increase of 2-3% in the stability values from the Stage 2 network was observed by the introduction of this loss. The final loss for training at this stage is given by the following equation,

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{\phi} + \lambda_2 \cdot \mathcal{L}_{\text{CX}} + \lambda_3 \cdot \mathcal{L}_{\text{td}} + \lambda_4 \cdot \mathcal{L}_{\text{id}} + \lambda_4 \cdot \mathcal{L}_{\text{cml}}, \tag{4}$$

Here, $\mathcal{L}_\phi$, $\mathcal{L}_{CX}$, $\mathcal{L}_{td}$, $\mathcal{L}_{id}$ and $\mathcal{L}_{cml}$ represent, perceptual, contextual, temporal discriminator, image discriminator and contrastive motion loss, respectively. Here, λ_n represent the controlling hyperparameters. Due to the space limitation, the details of the above-mentioned losses and the implementation are provided in the supplementary material.

5 Results

5.1 Quantitative Results

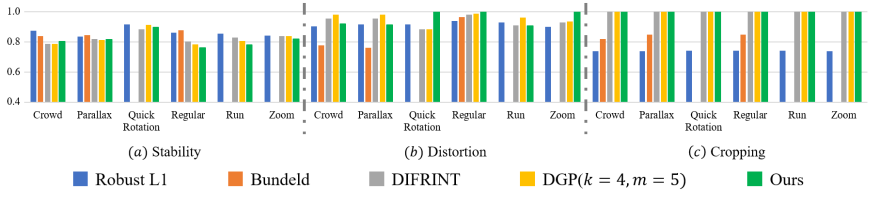

Figure 6: Quantitative comparison of Robust L1 [8], Bundled [18], DIFRINT [7], DGP (*k=4, m=5*) and our stabilization network.

We evaluate the performance of the proposed method quantitatively in terms of stability, distortion and cropping metrics as suggested by [18] on the 6 categories of videos presented in the NUS dataset [18]. The provided results (Figure 6) are averaged over each category.

• **Stability:** This metric defines the stability in terms of frequency component analysis. To calculate this metric, the feature trajectories are analyzed in the frequency domain as described in [18]. It is worth emphasizing that this metric does not take into account the quality of input videos and blurry results are also perceived stable through this metric.

• **Distortion:** This metric evaluates the anisotropic homography of the generated frames to the actual unstable frames. The lowest ratio is selected as the final distortion score. A higher score in this metric signifies better preservation of the content.

• **Cropping:** This metric measures the retention of visual information in generated frames through homography calculation between the generated and the actual frames. A higher score signifies better preservation of the visual information.

Through Figure 6, it is evident that the proposed network outperforms the SOTA methods in terms of distortion and cropping and performs competitively in terms of stability on the videos from Crowd, Parallax and Quick Rotation and lags behind in the remaining three categories. This is due to the fact that a large portion of the generated dataset consists of the unstable videos from the DeepStab [28], which contain the motion profiles similar to the above mentioned three categories. This bias in the results can be minimized by fine-tuning the network on videos containing motion profiles similar to the videos from the remaining three categories. We do not include the quantitative results produced by iterated CAIN [3] as the results contain inconsistent inter-frame artifacts that hinder the calculation of stability score by introducing a new local motion profile in the resulting videos. It is worth noting that the videos generated by the DGP and the iterated CAIN [3] share the same global motion profiles hence the actual stability score for both the methods should be similar. Please note that the Figure 6 does not include the results from [18] for Quick Rotation, Run and Zoom as it fails to stabilize most of the videos from these categories because of the extremely large non-overlapping regions.

5.2 Qualitative Results

For visual quality comparison, we present the results generated by Adobe Premiere 2018 CC, Robust L1 [8], DIFRINT [2], Iterated CAIN [3] (20 iterations), frames generated through our DGP (*k=4, m=5*) and the output from the proposed stabilization network in Figure 7. The loss of visual resolution can be clearly seen in the results by Adobe premiere 2018 CC and Robust L1 [8]. The bounded yellow regions in DIFRINT [2] and iterated CAIN [3] highlight the artifacts caused by both methods. It can be seen from these results that our models (DGP and stabilization network) produce better quality results and preserve the scale and content. The user study and more results are presented in the supplementary material.

Figure 7: Visual quality comparison of Adobe Premiere 2018 CC, Robust L1 [8], DIFRINT [2], Iterated CAIN [3], DGP (*k=4, m=5*), and our stabilization network.

6 Conclusion

In this work, we firstly pinpoint the obstacles that hinder a motion blind video stabilization formulation, and then present the first ever pixel-level synthesis solution for it. To do so, we firstly propose a dataset generation scheme that produces equi-perspective high-quality stable videos through iterative frame interpolation and refinement. Through the generated dataset, and a carefully designed training strategy, we demonstrate that the proposed motion blind video stabilization network compares favorably to the state-of-the-art video stabilization solutions that utilize explicit motion estimation modules, and our proposed model also preserves the visual information as well as the resolution which the currently available methods struggle with.

Acknowledgement

This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFCIT1901-06.

References

[1] Chris Buehler, Michael Bosse, and Leonard McMillan. Non-metric image-based rendering for video stabilization. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE, 2001.

[2] Jinsoo Choi and In So Kweon. Deep iterative frame interpolation for full-frame video stabilization. *ACM Transactions on Graphics (TOG)*, 39(1):1–9, 2020.

[3] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. Channel attention is all you need for video frame interpolation. In *AAAI*, pages 10663–10671, 2020.

[4] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 459–474, 2018.

[5] Jeffrey L Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

[6] Michael L Gleicher and Feng Liu. Re-cinematography: Improving the camerawork of casual video. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 5(1):1–28, 2008.

[7] Amit Goldstein and Raanan Fattal. Video stabilization using epipolar geometry. *ACM Transactions on Graphics (TOG)*, 31(5):1–10, 2012.

[8] Matthias Grundmann, Vivek Kwatra, and Irfan Essa. Auto-directed video stabilization with robust l1 optimal camera paths. In *CVPR 2011*, pages 225–232. IEEE, 2011.

[9] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5767–5777. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf.

[10] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3154–3160, 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[13] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. *ArXiv e-prints*, 2017.

[14] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018.

[15] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.

[16] Feng Liu, Michael Gleicher, Hailin Jin, and Aseem Agarwala. Content-preserving warps for 3d video stabilization. *ACM Transactions on Graphics (TOG)*, 28(3):1–9, 2009.

[17] Feng Liu, Michael Gleicher, Jue Wang, Hailin Jin, and Aseem Agarwala. Subspace video stabilization. *ACM Transactions on Graphics (TOG)*, 30(1):1–10, 2011.

[18] Shuaicheng Liu, Lu Yuan, Ping Tan, and Jian Sun. Bundled camera paths for video stabilization. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013.

[19] Shuaicheng Liu, Lu Yuan, Ping Tan, and Jian Sun. Steadyflow: Spatially smooth optical flow for video stabilization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4209–4216, 2014.

[20] Shuaicheng Liu, Ping Tan, Lu Yuan, Jian Sun, and Bing Zeng. Meshflow: Minimum latency online video stabilization. In *European Conference on Computer Vision*, pages 800–815. Springer, 2016.

[21] Yasuyuki Matsushita, Eyal Ofek, Weina Ge, Xiaoou Tang, and Heung-Yeung Shum. Full-frame video stabilization with motion inpainting. *IEEE Transactions on pattern analysis and Machine Intelligence*, 28(7):1150–1163, 2006.

[22] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 768–783, 2018.

[23] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017.

[24] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–270, 2017.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[26] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.

[27] Brandon M Smith, Li Zhang, Hailin Jin, and Aseem Agarwala. Light field video stabilization. In *2009 IEEE 12th international conference on computer vision*, pages 341–348. IEEE, 2009.

[28] 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*, 28(5):2283–2292, 2018.

[29] Yu-Shuen Wang, Feng Liu, Pu-Sheng Hsu, and Tong-Yee Lee. Spatially and temporally optimized video stabilization. *IEEE transactions on visualization and computer graphics*, 19(8):1354–1361, 2013.

[30] Sen-Zhe Xu, Jun Hu, Miao Wang, Tai-Jiang Mu, and Shi-Min Hu. Deep video stabilization using adversarial networks. In *Computer Graphics Forum*, volume 37, pages 267–276. Wiley Online Library, 2018.

[31] Jiyang Yu and Ravi Ramamoorthi. Robust video stabilization by optimization in cnn weight space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3800–3808, 2019.

[32] Jiyang Yu and Ravi Ramamoorthi. Learning video stabilization using optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8159–8167, 2020.