Fast Full-frame Video Stabilization with Iterative Optimization

Weiyue Zhao\textsuperscript{1} Xin Li\textsuperscript{2} Zhan Peng\textsuperscript{1} Xianrui Luo\textsuperscript{1} Xinyi Ye\textsuperscript{1} Hao Lu\textsuperscript{1} Zhiguo Cao\textsuperscript{1,*}
\textsuperscript{1}Key Laboratory of Image Processing and Intelligent Control, Ministry of Education; School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
\textsuperscript{2}Department of Computer Science, University of Albany, Albany NY 12222
\{zhaowei, peng, xianruiluo, xinyiye, hlu, zgcao\}@hust.edu.cn
\textsuperscript{2}xli48@albany.edu

Abstract

Video stabilization refers to the problem of transforming a shaky video into a visually pleasing one. The question of how to strike a good trade-off between visual quality and computational speed has remained one of the open challenges in video stabilization. Inspired by the analogy between wobbly frames and jigsaw puzzles, we propose an iterative optimization-based learning approach using synthetic datasets for video stabilization, which consists of two interacting submodules: motion trajectory smoothing and full-frame outpainting. First, we develop a two-level (coarse-to-fine) stabilizing algorithm based on the probabilistic flow field. The confidence map associated with the estimated optical flow is exploited to guide the search for shared regions through backpropagation. Second, we take a divide-and-conquer approach and propose a novel multi-frame fusion strategy to render full-frame stabilized views. An important new insight brought about by our iterative optimization approach is that the target video can be interpreted as the fixed point of nonlinear mapping for video stabilization. We formulate video stabilization as a problem of minimizing the amount of jerkiness in motion trajectories, which guarantees convergence with the help of fixed-point theory. Extensive experimental results are reported to demonstrate the superiority of the proposed approach in terms of computational speed and visual quality. The code will be available on GitHub.

1. Introduction

With the growing popularity of short videos on social media platforms (e.g., TikTok, Instagram), video has played an increasingly important role in our daily life. However, casually captured videos are often shaky and wobbly due to amateur shooting. Although it is possible to alleviate those problems by resorting to professional equipment (e.g., dollies and steadicams), the cost of hardware-based solutions is often expensive, making it impractical for real-world applications. By contrast, software-based or computational solutions such as video stabilization algorithms \cite{12} have become an attractive alternative to improve the visual quality of shaky video by eliminating undesirable jitter.

Existing video stabilization techniques can be classified into two categories: optimization-based and learning-based. Traditional optimization-based algorithms \cite{11,18,21,27,39} have been widely studied due to their speed and robustness. The challenges of them are the occlusion caused by changes in depth of field and the interference caused by foreground objects on camera pose regression. Furthermore, their results often contain large missing regions at frame borders, particularly when videos with a large camera motion. In recent years, learning-based video stabilization algorithms \cite{6,23,45,50} have shown their superiority by achieving higher visual quality compared to traditional methods. However, their stabilization model is too complex for rapid computation, and its generalization property is unknown due to the scarcity of training datasets.

To overcome those limitations, we present an iterative optimization-based learning approach that is efficient and robust, capable of achieving high-quality stabilization results with full-frame rendering, as shown in Fig. 1. The probabilistic stabilized network addresses the issues of occlusion and interfering objects, and achieves fast pose estimation. Then the full-frame outpainting module retains the original field of view (FoV) without aggressive cropping. An important new insight brought by our approach is that the objective of video stabilization is to suppress the implicitly embedded noise in the video frames rather than the explicit noise in the pixel intensity values. This inspired us to adopt an expectation-maximization (EM)-like approach for video stabilization. Importantly, considering the strong redundancy of video in the temporal domain, we ingeniously

https://github.com/zwyking/Fast-Stab
consider stable video (the target of video stabilization) as the fixed point of nonlinear mapping. Such a fixed-point perspective allows us to formulate an optimization problem of the optical flow field in commonly shared regions. Unlike most methods that resort to the ad hoc video dataset [56] or the deblurred dataset [35] as stabilized videos, we propose to construct a synthetic training dataset to facilitate joint optimization of model parameters in different network modules.

To solve the formulated iterative optimization problem, we take a divide-and-conquer approach by designing two modules: probabilistic stabilization network (for motion trajectory smoothing) and video outpainting network (for full-frame video rendering). For the former, we propose to build on the previous work of PDCNet [37, 38] and extend it using a coarse-to-fine strategy to improve robustness. For a more robust estimate of the uncertainty of the optical flow, we infer masks from the optical flow by bidirectional propagation with a low computational cost (around 1/5 of the time Yu et al. [50]). Accordingly, we have developed a two-level (coarse-to-fine) flow smoothing strategy that first aligns adjacent frames by global affine transformation and then refines the result by warping the fields of intermediate frames. For the latter, we propose a two-stage approach (flow and image outpainting) to render full-frame video. Our experimental results have shown highly competent performance against others on three public benchmark datasets. The main contribution of this work is threefold:

- We propose a formulation of video stabilization as a fixed-point problem of the optical flow field and propose a novel procedure to generate a model-based synthetic dataset.
- We construct a probabilistic stabilization network based on PDCNet and propose an effective coarse-to-fine strategy for robust and efficient smoothing of optical flow fields.
- We propose a novel video outpainting network to render stabilized full-frame video by exploiting the spatial coherence in the flow field.

2. Related Work

2.1. Video Stabilization

Most video stabilization methods can typically be summarized as a three-step procedure: motion estimation, smoothing the trajectory, and generating stable frames. Traditional methods focus primarily on 2D features or image alignment [29] when it comes to motion estimation. These methods are different in modeling approaches to motion, including the trajectory matrix [19], epipolar geometry [10, 21, 47, 53], and the optical flow field [22, 54].

Regarding the smoothing trajectory, particle filter tracking [46], space-time path smoothing [19, 40], and L1 optimization [11] have been proposed. Existing methods for generating stable frames rely mainly on 2D transformations [26], grid warping [19, 21], and dense flow field warping [22].

Compared to the 2D method, some approaches turn to 3D reconstruction [17]. However, specialized hardware such as depth camera [20] and light field camera [34] are necessary for these methods based on 3D reconstruction. Some methods [30, 43, 44, 49] tackle video stabilization from the perspective of deep learning. In [49], the optimization of video stabilization was formulated in the CNN weight space. Recent work [50] represents motion by flow field and attempts to learn a stable optical flow to warp frames. Another method [6] aims to learn stable frames by interpolation. These deep-learning methods generate stable videos with less distortion.

2.2. Large FOV Video

Unlike the early work (e.g., traffic video stabilization [16]), large field of view (Field Of View) video stabilization has been attracting more researchers’ attention. For most video stabilization methods, cropping is inevitable, which is why the FOV is reduced. Several approaches have been proposed to maintain a high FOV ratio. OVS [45]
proposed to improve FOV by extrapolation. DIFRINT [6] choose iterative interpolation to generate high-FOV frames. FuSta [23] used neural rendering to synthesize high-FOV frames from feature space. To a great extent, the performance of interpolation-based video stabilization [6, 45] depends on the selected frames. If selected frames have little correspondence with each other, performance will deteriorate disastrously. Neural rendering [23] synthesizes the image by weighted summing, causing blur. Most recently, a deep neural network [33] (DNN) has jointly exploited sensor data and optical flow to stabilize videos.

3. Stabilization via Iterative Optimization

Motivation. Despite rapid advances in video stabilization [30], existing methods still suffer from several notable limitations [12]. First, a systematic treatment of various uncertainty factors (e.g., low-texture regions in the background and moving objects in the foreground) in the problem formulation is still lacking. These uncertainty factors often cause occlusion-related problems and interfere with the motion estimation process. Second, the high computational cost has remained a technical barrier to real-time video processing. The motivation behind our approach is two-fold. On the one hand, we advocate for finding commonly shared regions among successive frames to address various uncertainty factors in handheld video. On the other hand, in contrast to these prestabilization algorithms [21, 23, 49, 50] based on traditional approaches, we propose a novel high-efficiency prestabilization algorithm based on probabilistic optical flow. Flow-based methods are generally more accurate in motion estimation and deserve a high time cost. Accuracy and efficiency, we have both.

Approach. We propose to formulate video stabilization as a problem of minimizing the amount of jerkiness in motion trajectories. It is enlightening to think of video stabilization as a special kind of “video denoising” where noise contamination is not associated with pixel intensity values, but embedded into the motion trajectories of foreground and background objects. Conceptually, for video restoration in which unknown motion is the hidden variable, we can treat video stabilization as a chicken-and-egg problem [13] - i.e., the objective of smoothing motion trajectories is intertwined with that of video frame interpolation. Note that an improved estimation of motion information can facilitate the task of frame interpolation and vice versa. Such an observation naturally inspires us to tackle the problem of video stabilization by iterative optimization.

Through divide-and-conquer, we propose to formulate video stabilization as the following optimization problem. Given a sequence of \( n \) frames along with the set of optical flows \( \mathcal{Y} \), we first define the confidence map, which originally indicates the reliability and accuracy of the optical flow prediction at each pixel. Here, we have thresholded the confidence map as a binary image, which represents accurate matches (as shown in the 4-th column of Fig. 2). Then we can denote \( n \) frames and the corresponding set of confidence maps \( \mathcal{M} \) by:

\[
\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_q\}, \mathcal{M} = \{M_1, M_2, \ldots, M_q\}.
\]

4. Probabilistic Stabilization Network

The problem of video stabilization can be formulated as finding a nonlinear mapping \( f: \mathcal{Y} \rightarrow \tilde{\mathcal{Y}} \) where \( \mathcal{Y} \) denotes the optical flow set of the stabilized video. We hypothesize that a desirable objective to pursue \( f \) is the well-known fixed-point property, i.e., \( \tilde{\mathcal{Y}} = f(\mathcal{Y}) \). To achieve this objective, we aim to minimize an objective function \( \mathcal{F} \) characterized by the magnitude of optical flows between commonly shared regions, as represented by \( \mathcal{M} \). Note that \( \mathcal{M} \) is the hidden variable in our problem formulation (that is, we need to estimate \( \mathcal{M} \) from \( \tilde{\mathcal{Y}} \)). A popular strategy to address this chicken-and-egg problem is to alternatively solve the two subproblems of unknown \( \mathcal{M} \) and \( \tilde{\mathcal{Y}} \). More specifically, the optimization of \( \mathcal{F} \) is decomposed into the following two subproblems:

\[
(\hat{M}_1, \ldots, \hat{M}_q) = \arg\min_{\mathcal{M}} \mathcal{F}(\hat{Y}_1 \odot M_1, \ldots, \hat{Y}_q \odot M_q),
\]

\[
(\hat{Y}_1, \ldots, \hat{Y}_q) = \arg\min_{\mathcal{Y}} \mathcal{F}(Y_1 \odot \hat{M}_1, \ldots, Y_q \odot \hat{M}_q),
\]

where \( Y/\hat{Y} \) denotes the magnitude of the optical flow values before and after stabilization and \( \odot \) is the Hadamard product. Instead of analyzing the estimated motion in the frequency domain, we hypothesize that stabilized videos are the fixed points of video stabilization algorithms, which minimize the above objective function. A fixed point [1] is a mathematical object that does not change under a given transformation. Numerically, fixed-point iteration is a method of computing fixed points of a function. It has been widely applied in data science [7], and image restoration [5, 25]. Here, we denote \( \mathcal{F} \) in Eq. (2) as the function to be optimized, and the fixed point of \( \mathcal{F} \) is defined as the stabilized video. Next, we solve these two subproblems by constructing the module of stabilization.

4.1. Probabilistic Flow Field

First, we start with an interesting observation. When playing an unstable video at a slower speed (e.g., from 50fps to 30fps), the video tends to appear less wobbly. It follows from the observation that the fundamental cause of video instability is the fast frequency and the large amplitude of the still objects’ motion speed. Therefore, the core task of motion smoothing is to identify the region that needs to be stabilized. As Yu et al. [50] pointed out, mismatches, moving objects, and inconsistent motion areas lead to variation of the estimated motion from the true motion trajectories and
should be masked. An important new insight brought about by this work is that these regions of inconsistency in the optical flow fields tend to show greater variability as the frame interval increases. Therefore, if these inconsistent regions are excluded, the video stabilization task can be simplified to the first sub-problem in Eq. (2) - i.e., minimizing $F$ over $\mathcal{M}$ for a fixed $\mathcal{Y}$.

To detect unreliable matches and inconsistent regions, we have adopted the probabilistic flow network – PDCNet [37, 38], that jointly tackles the problem of dense correspondence and uncertainty estimation, as our building block. Suppose that PDCNet estimates the optical flow $Y_k$ from frame $k + d$ to frame $k$ with the resulting confidence map $C_k$. Although $C_k$ denotes the uncertainty of the optical flow estimate in $Y_k$, it is often sensitive to the frame interval $d$. For example, it is difficult to identify the inconsistent region when $d$ is small, while the common area is less controllable when $d$ is large. Therefore, simply using an optical flow field to estimate inconsistent regions is not sufficient.

We have designed a more robust solution to the joint estimation of dense flow and confidence map based on a coarse-to-fine strategy. The basic idea is to first obtain the probabilistic flow field at a coarse scale (e.g., with the downsampled video sequence by a factor of $d$ along the temporal axis) and then fill in the rest (i.e., the frames between the adjacent frames in the down-sampled video) at the fine scale. Such a two-level estimation strategy effectively overcomes the limitations of PDCNet by propagating the estimation results of probabilistic flow fields bidirectionally, as we will elaborate next.

**Coarse-scale strategy** As shown by the second-to-last column in Fig. 2, the confidence map estimated by PDCNet can identify the mismatched region, but fails to locate the objects with small motions (e.g., people and the sky). To overcome this difficulty, we introduce the binary mask as a warped and aggregated confidence map (refer to the last column of Fig. 2). Specifically, we first propose to obtain the confidence map $\hat{C}_{k+(n-1)d}$ with down-sampled video (the last row of Fig. 2). In the forward direction, we estimate dense flow and confidence map using PDCNet; then $\hat{C}_{k+(n-1)d}$ is backpropagated to update the binary mask $M$ by thresholding and setting intersection operators. Through bidirectional propagation, the region covered by $M$ is the shared content from frame $k$ to frame $k + (n - 1)d$. The complete procedure can be found in Supp. S1.

**Fine-scale strategy** Based on the coarse-scale estimation result for downsampled video (i.e., $\mathcal{M} = \{M_k, M_{k+d}, \ldots, M_{k+(n-1)d}\}$), we fill in the missing $d - 1$ frames at the fine scale. Specifically, considering the sequential frames from $k$ to $k + d$, we can obtain two sets similar to Eq. (1), which are $\mathcal{Y} = \{Y_k, Y_{k+1}, \ldots, Y_{k+d-1}, Y_{k+d}\}$ and their corresponding confidence map set $C$. Note that $M_{k+d} \in \mathcal{M}$ has been calculated in coarse stage. Setting $d = 1$, we can call the algorithm again to obtain the set of output masks $M$ for the rest of $d - 1$ frames.

**4.2. Coarse-scale Stabilizer**

To coarsely stabilize the video, we first propose aligning the adjacent frames with a global affine transformation [52]. The optimization function $F$ in Eq. (2) is represented as

$$T^* = \arg \min_{T} T(Y \odot \hat{M}),$$

where $T(\cdot)$ denotes the image transformation applied to the shared region $\hat{M}$ (the result of Sec. 4.1) of the optical flow field $Y$. Most conventional methods [23, 44, 50] adopt image matching to obtain $T(\cdot)$ - e.g., keypoint detection [2, 24, 31], feature matching [2, 3, 51]; and camera pose estimation is implemented by OpenCV-python. However, these methods are often time-consuming and computationally expensive. For example, two adjacent frames of unstable videos usually share a large area and are free from perspective transformation. Thus, an affine transformation, including translation, rotation, and scaling, is sufficient. More importantly, within the optimization-based learning framework, we can
regress these linear parameters of \( T(\cdot) \) from the optical flow field, which characterize the relative coordinate transformation of the matched features.

We propose a novel camera pose regression network, as shown in Fig. 3. Given an optical flow field \( Y \) and the corresponding mask field \( \hat{M} \), our network \( \Phi(\cdot) \) can directly estimate the unknown parameters \( T(\cdot) \approx \{\theta, s, d_x, d_y\} = \Phi(Y, \hat{M}) \). To solve the optimization problem of Eq. (3), we use the estimated parameters to iteratively compute the corresponding residual optical flow fields such that

\[
\hat{Y} = Y - (S \cdot R \cdot V + T),
\]

where \( T, S, \) and \( R \), respectively, denote the translation, scaling, and rotation matrix, and \( V \in \mathbb{R}^{2 \times H \times W} \) represents an image coordination grid. Then \( \{\Delta \theta, \Delta s, \Delta x, \Delta y\} = \Phi(\hat{Y}, \hat{M}) \) is calculated iteratively to produce the updated parameters \( \{\theta + \Delta \theta, s + \Delta s, d_x + \Delta x, d_y + \Delta y\} \). The finally estimated affine transformation is smoothed by a moving Gaussian filter with a window size of 20 pixels.

**Loss functions** Our loss functions include robust loss \( \ell_1 \) and grid loss commonly applied in consistency filtering [31]. We directly calculate the \( \ell_1 \) loss between the predictions and their ground truth \( \{\theta, \hat{s}, \hat{d}_x, \hat{d}_y\} \),

\[
L_{gt} = \lambda_\theta \| \theta - \hat{\theta} \|_1 + \lambda_s \| 1 - \frac{s}{\hat{s}} \|_1 + \lambda_t \| \hat{d}_x - d_x \|_1 + \| d_y - \hat{d}_y \|_1.
\]

For better supervision of the estimated pose, we calculate the loss \( L_{grid} \) with the grid \( V \in \mathbb{R}^{2 \times h \times w} \) in Eq. (4), i.e.,

\[
L_{grid} = \| (\hat{S} \cdot \hat{R} \cdot V + \hat{T}) - (S \cdot R \cdot V + T) + \epsilon \|_1.
\]

where \( \epsilon \) denotes the ground truth and \( \epsilon \) is a small value for stability. The final loss \( L_{stab} \) consists of \( L_{gt} \) and \( L_{grid} \),

\[
L_{stab} = L_{gt} + \lambda_{grid} L_{grid}.
\]

### 4.3. Fine-scale Stabilizer

The assumption with an affine transformation of the coarse-scale stabilizer could cause structural discontinuity and local distortion. Therefore, our objective is to refine the coarsely stabilized video by optical flow smoothing. Unlike Eq. 3 which applies an image transformation matrix to optimize the optical flow field, we optimize it at the pixel level by a flow warping field \( \mathbf{W} \). Thus, the function \( F \) in Eq. (2) is given by

\[
\mathbf{W}^* = \arg \min_{\mathbf{W}} \sum_{i=0}^{N-1} W_i(Y_i \odot \hat{M}_i).
\]

The flow smoothing network follows the U-Net architecture in [56]. We use \( N \) frames of optical flow fields \( \mathbf{F} \) and mask fields \( \mathbf{M} \) as input and obtain \( (N-1) \) frame warp fields \( \mathbf{W} \) of intermediate frames. Specifically, for the optical flow field \( Y_k \) (from frame \( k+1 \) to frame \( k \)), we denote the aligned matrices of each frame in Sec. 4.2 as \( H_k \in \mathbb{R}^{2 \times 3} \) and \( H_{k+1} \in \mathbb{R}^{2 \times 3} \), respectively. Then the input optical flow \( \mathbf{F}_k \in \mathbb{R}^{2 \times H \times W} \) can be represented by

\[
\mathbf{F}_k = H_{k+1} \cdot [V + Y_k \mid 1] - H_k \cdot [V \mid 1],
\]

where \( V \in \mathbb{R}^{2 \times H \times W} \) and \( [\cdot \mid 1] \) denote the normalized coordinate representation. Furthermore, to better adapt the flow smoothing network to the mask field \( \hat{M} \), we fine-tune it using our synthetic dataset (as we will elaborate in Sec. 6.1).

The loss function follows the motion loss [49]

\[
L_{smooth} = \sum_{k=0}^{N-1} \| \mathbf{F}_k + \mathbf{W}_k - \mathbf{F}_k(\mathbf{W}_{k+1}) \|_1 \odot \hat{M},
\]

where \( \mathbf{W}_0 = \mathbf{W}_N = 0 \).

### 5. Video Outpainting Network

Most video stabilization methods crop an input video with a small field of view (FOV), excluding missing margin pixels due to frame warping. In contrast, full-frame video stabilization [6, 23, 26] proposes to generate a video that maintains the same FOV as the input video without cropping. They directly generate stabilized video frames with large FOV by fusing the information from neighboring frames. An important limitation of existing fusion strategies is the unequal importance of different frames (i.e., the current and neighboring frames are weighted equally), which would lead to unpleasant distortions in fast-moving situations. To overcome this weakness, we propose a two-stage framework to combine flow and image outpainting strategies [32]. In the first stage, we used flow-outpainting to iteratively align the neighboring frames with the target frame. In the second stage, we fill the target frame with adjacent aligned frame pixels by image outpainting.

### 5.1. Flow Outpainting Network

Let \( I^t \) denote the outpainted target image and \( I^s \) the neighboring source image. We aim to fill the missing pixel.
region $M^t_{\partial}$ of $I^t$ with the pixels of $I^s$. As shown in Fig. 4, we take two different FOVs of $I^t$ as input and obtain the corresponding optical flow fields and warped results $I^s_{\text{warp}}$. The small FOV $I^t$ cannot guide $I^s$ well to fill the regions in $M^t_{\partial}$. Since the predicted optical flow field in $M^t_{\partial}$ is unreliable due to the lack of pixel guidance of $I^t$, the out-of-view region of $M^t_{\partial}$ has artifacts (marked in the last column of Fig. 4). We observe that the optical flow field of the large FOV $I^t$ is continuous in $M^t_{\partial}$, which inspired us to extrapolate the flow of $M^t_{\partial}$ using the reliable flow region (the 3rd column of Fig. 4).

We propose a novel flow outpainting network (Fig. 5), which extrapolates the large FOV flow field using a small FOV flow field and the corresponding valid mask. Specifically, we adopt a U-Net architecture and apply a sequence of gated convolution layers with downsampling / upsampling to obtain the large flow field of FOV $Y_{\text{large}}$. Note that the input flow field $Y_{\text{small}}$ and the valid mask $M_{\text{valid}}$ have been estimated by PDCNet (see Sec. 4.1).

**Loss functions** Our loss functions include robust loss $\ell_1$ and loss in the frequency domain [4]. We directly calculate the $\ell_1$ loss between $Y_{\text{large}}$ and its ground truth $Y_{\text{large}}$,

\[ L_Y = \lambda_\text{in} \cdot \| Y_{\text{large}} - Y_{\text{small}} \|_1 + \lambda_\text{out} \cdot \| Y_{\text{large}} - Y_{\text{large}} \|_1 \cdot (\sim M_{\text{valid}}) \|_1 . \] (11)

To encourage low frequency and smoothing $Y_{\text{large}}$, we add the loss in the frequency domain $L_F = \| \mathcal{F} Y_{\text{large}} \|_2$ , where the normalized Gaussian map $\mathcal{G}$ with $\mu = 0$ and $\sigma = 3$ is inverted by its maximum value and $\mathcal{F} Y_{\text{large}}$ denotes the Fourier spectrum of $Y_{\text{large}}$. The final loss consists of $L_Y$ and $L_F$.

\[ L_{\text{outpaint}} = L_Y + \lambda_F L_F . \] (12)

### 5.2. Image Margin Outpainting

Based on our proposed flow-outpainting network, we design a margin-outpainting method by iteratively aligning frame pairs (see Fig. 6). As discussed in Sec. 5.1, we can obtain a large FOV flow field $Y_{\text{large}}$. The neighboring reference frame $I^s$ is warped as $I^s_{\text{warp}} = Y_{\text{large}}(I^s)$ as shown in Fig. 6. In theory, the outpainted frame $I^{\text{result}} = I^t \cdot M_{\text{valid}} + I^s_{\text{warp}} \cdot (\sim M_{\text{valid}})$. However, we notice that there are obvious distortions at the image border. To further align the margins, we take a margin fusion approach (the detailed algorithm can be found in Supp. S1) We crop $I^s_{\text{warp}}$ and $I^t$ to $I^c_s$ and $I^c_t$. Then, we can obtain a new warped frame $I^c_{\text{warp}}$ by flow outpainting. In particular, we did not choose to add $I^c_{\text{warp}}$ and $I^t$ directly. To identify the misaligned region, we propose to outpaint the mask $M^t_\partial$ by extending the watershed outward from the center. Instead of a preset threshold, we adaptively choose between the target image $I^t$ and the warped image $I^c_{\text{warp}}$. Then, the final frame $I^{\text{result}}$ consists of $I^t$ and $I^c_{\text{warp}}$: $I^{\text{result}} = I^t \cdot M^t_\partial + I^c_{\text{warp}} \cdot (\sim M^t_\partial)$ . Compared to the two results $I^{\text{result}}$, our strategy successfully mitigates misalignment and distortions at the boundary of video frames.

**Multi-frame fusion** During the final stage of rendering, we use a sequence of neighboring frames to outpaint the target frame, while they may have filled duplicate regions. It is important to find out which frame and which region should be selected. We proposed the selection strategy for multi-frame fusion (the details can be found in Supp. S1). By weighing the metric parameters of each frame, we finally obtain the target frame with large FOV. Note that each frame has an added margin in the stabilization process, so we need to crop them to the original resolution. Although we have outpainted the target frame, some missing pixel holes may still exit at boundaries. Here, we apply the state-of-the-art LaMa image inpainting method [36] to fill these holes using nearest-neighbor interpolation.

### 6. Experiments

#### 6.1. Synthetic Datasets for Supervised Learning

Due to the limited amount of paired training data, we propose a novel model-based data generation method by carefully designing synthetic datasets for video stabiliza-
tion. For our base synthetic dataset, we used a collection of images from the DPED [14], CityScapes [8] and ADE-20K [55] datasets. To generate a stable video, we randomly generate the homography parameters for each image, including rotation angle $\theta$, scaling $s$, translations $(d_x, d_y)$ and perspective factors $(p_x, p_y)$. Then we divide these transformation parameters into $N$ bins equally and obtain a video of $N$ frames by homography transformations. To simulate the presence of moving objects in real scenarios, the stable video is further augmented with additional independently moving random objects. To do so, the objects are sampled from the COCO dataset [15] and inserted on top of the synthetic video frames using their segmentation masks. Specifically, we randomly choose $m$ objects (no more than 5), and generate randomly affine transformation parameters for each independent of the background transformation. Finally, we cropped each frame to $720 \times 480$ around its center. For different training requirements, we apply various combinations of synthetic dataset. The implementation and training details can be found in Supp. S2, S3.

6.2. Quantitative Evaluation

We compare the results of our method with various video stabilization methods, including Grundmann et al. [11], Liu et al. [21], Wang et al. [39], Yu and Ramamoorthi [49, 50], DUT [44], OVS [45], DIFRINT [6], and FuSta [23]. We obtain the results of the compared methods from the videos released by the authors or generated from the publicly available official implementation with default parameters or pre-trained models. Note that OVS [45] does not honor their promise to provide code, thus we only report the results from their paper.

Datasets. We evaluate all approaches on the NUS dataset [21], DeepStab dataset [39], and Selfie dataset [48]. The NUS dataset consists of 144 videos and the corresponding ground truths in 6 scenes. The DeepStab dataset contains 61 videos and the Selfie dataset consists of 33 videos.

Metrics. We introduce three metrics widely used in many methods [6, 23, 50] to evaluate our model: 1) Cropping ratio measures the remaining frame area after cropping off the invalid boundaries. 2) Distortion value evaluates the anisotropic scaling of the homography between the input and output frames. 3) Stability score evaluates the stability of the output video. We calculate the metrics using the evaluation code provided by DIFRINT.

Quantitative comparison. The results of the NUS dataset [21] are summarized in Table 1 (Per-category result can be found in Supp. S5). Overall, our method achieves the best distortion value compared to the state-of-the-art method, FuSta [23]. Especially in the quick-rotation and zoom categories, our method outperforms pure image generation methods [6, 23]. We suspect that the reduction of the shared information between frames causes the image generation methods to prefer artifacts. However, our method can ensure local structural integrity when outpainting the margin region. Furthermore, our method achieves an average cropping ratio of 1.0 and stability scores comparable to recent approaches [6, 23, 49, 50]. Since FuSta [23] uses Yu et al. [50] to obtain stabilized input videos, they have the same stability scores. It is important to note that although the stability scores are competitive, our method runs 5 times faster than [50] in the video stabilization stage. Moreover, the comparison results on DeepStab dataset [39] and Selfie dataset [48] are also reported in Table 1. Our method still shows effectiveness in different datasets, proving the generalizability of the proposed method. Note that, due to that Wang et al. [39] is trained on the DeepStab dataset, we do not report its results on the DeepStab dataset for a fair comparison. Qualitative comparisons can be found in Supp. S4 and supplementary video.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Grundmann et al.</td>
<td>0.71</td>
<td>0.76</td>
<td>0.62</td>
</tr>
<tr>
<td>Liu et al.</td>
<td>0.81</td>
<td>0.78</td>
<td>0.82</td>
</tr>
<tr>
<td>Wang et al. [39]</td>
<td>0.67</td>
<td>0.72</td>
<td>0.41</td>
</tr>
<tr>
<td>Yu and Ramamoorthi [49]</td>
<td>0.78</td>
<td>0.77</td>
<td>0.82</td>
</tr>
<tr>
<td>Yu and Ramamoorthi [50]</td>
<td>0.85</td>
<td>0.81</td>
<td>0.86</td>
</tr>
<tr>
<td>Yu [50]+OVS* [45]</td>
<td>0.92</td>
<td>0.78</td>
<td>0.83</td>
</tr>
<tr>
<td>DUT [44]</td>
<td>0.71</td>
<td>0.81</td>
<td>0.83</td>
</tr>
<tr>
<td>DIFRINT [6]</td>
<td>1.00</td>
<td>0.85</td>
<td>0.84</td>
</tr>
<tr>
<td>FuSta [23]</td>
<td>1.00</td>
<td>0.87</td>
<td>0.86</td>
</tr>
<tr>
<td>Ours</td>
<td>1.00</td>
<td>0.91</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Table 1. Quantitative results on the NUS dataset [21], the DeepStab dataset [39] and the Selfie Dataset [48]. We evaluate the following metrics: Cropping Ratio(C.), Distortion Value(D.), Stability Score(S.). * indicates the results obtained from original paper. We highlight the best method in **bold** and underline the second-best.
6.3. Ablation Study

**Importance of mask generation.** We investigate the influence of the mask on video stabilization at different stages. To better demonstrate the necessity of mask in complex scenarios, we choose the Crowd category of NUS dataset [21] which includes a display of moving pedestrians and occlusions. Stability and distortion at different settings are shown in Table 2. It can be seen that the performance with mask increases significantly in both stabilizers. Specifically, the mask can both improve stability globally and alleviate image warping distortion locally. This result demonstrates the importance of mask generation for video stabilization.

**Flow outpainting.** We compare our flow outpainting method with the traditional flow inpainting method PCAFlow [42]. Following [50] we fit the first 5 principal components proposed by PCAFlow to the $M_{valid} = 1$ regions of the optical flow field, and then outpaint the flow vectors in the $M_{valid} = 0$ regions with reasonable values of the PCA Flow fitted. The result is obtained by warping the source image with the outpainted optical flow field. We perform this comparison on our synthetic validation set and evaluate it with the corresponding ground truth. Additionally, we use PSNR and SSIM [41] to evaluate the quality of the results. As shown in Table 3, ours dramatically outperforms PCAFlow [42] in all objective metrics.

**Importance of image filling strategies.** We explore the following proposed strategies for image filling: margin outpainting, mask generation $M_{f}$, and multiframe selection. We isolate them from our method to compare their results with the complete version. The results are shown in Table 3. The proposed strategies are generally helpful in improving image quality. Especially, margin outpainting and mask $M_{f}$ are crucial to the results.

<table>
<thead>
<tr>
<th>Stabilizer</th>
<th>Stability↑</th>
<th>Distortion↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image-level</td>
<td>w.o. mask 0.76</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>with mask 0.83</td>
<td>0.91</td>
</tr>
<tr>
<td>Pixel-level</td>
<td>w.o. mask 0.81</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>with mask 0.83</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Table 2. The importance of mask to video stabilization. We validate on the Crowd category of NUS dataset [21].

<table>
<thead>
<tr>
<th>Method</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional pose regression</td>
<td>216ms</td>
</tr>
<tr>
<td>Our pose regression network</td>
<td>21ms</td>
</tr>
<tr>
<td>Grundmann et al. [11]</td>
<td>480ms</td>
</tr>
<tr>
<td>Liu et al. [21]</td>
<td>1360ms</td>
</tr>
<tr>
<td>Wang et al. [39]</td>
<td>460ms</td>
</tr>
<tr>
<td>Yu and Ramamoorthi [49]</td>
<td>1610ms</td>
</tr>
<tr>
<td>Yu and Ramamoorthi [50]</td>
<td>570ms</td>
</tr>
<tr>
<td>DIFPRINT [6]</td>
<td>1530ms</td>
</tr>
<tr>
<td>FuSta [23]</td>
<td>9820ms</td>
</tr>
<tr>
<td>Ours</td>
<td>97ms</td>
</tr>
</tbody>
</table>

Table 3. Ablation study of image filling. The evaluation is conducted on our model-based synthetic validation set.

<table>
<thead>
<tr>
<th>Stabilizer</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
<th>Distortion↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCAFlow [42]</td>
<td>11.34</td>
<td>0.50</td>
<td>0.72</td>
</tr>
<tr>
<td>Our flow outpainting</td>
<td>19.65</td>
<td>0.77</td>
<td>0.93</td>
</tr>
<tr>
<td>w.o. margin outpainting</td>
<td>18.04</td>
<td>0.64</td>
<td>0.81</td>
</tr>
<tr>
<td>w.o. mask $M_{f}$</td>
<td>20.35</td>
<td>0.78</td>
<td>0.86</td>
</tr>
<tr>
<td>w.o. multi-frame selection</td>
<td>21.99</td>
<td>0.84</td>
<td>0.90</td>
</tr>
<tr>
<td>Ours</td>
<td>23.17</td>
<td>0.86</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Table 4. Per-frame runtime comparison of camera pose regression and video stabilization.

Figure 7. Experiment of fixed-point iteration on shaky and stable frames. (a) Our fixed-point iteration helps the shaky frames converge to a steady state. (b) For stable frames, the fixed-point iteration guarantees the stability of results.

6.4. Runtime Comparison

Our network and pipeline are implemented with PyTorch and Python. Table 4 is a summary of the runtime per frame for various competing methods. All timing results are obtained on a desktop with an RTX3090Ti GPU and an Intel(R) Xeon(R) Gold 6226R CPU. First, we compare the run-time of pose regression. Traditional pose regression is time consuming in feature matching [2, 9], homography estimation, and SVD decomposition [28]. Although our learning-based pose regression network runs 10 times faster than the traditional framework. Then, we report the average run time of different methods, including optimization-based [21, 49] and learning-based [6, 11, 23, 39, 50]. Our method takes 97ms which gives ~ 5x speed-up. This is because our method computes the optical flow field only once and without the help of other upstream task methods and manual optimization.

6.5. Fixed-point Experiment

To demonstrate the stability of our fixed-point optimization solution, we performed an interesting toy experiment. We input a sequence of shaky frames into our coarse-to-fine stabilizers, and the stabilized result will be iteratively re-stabilized by our stabilizers. For each iteration, we calculate the average magnitude of optical flow filed with global transformation and flow warping for each frame. Specifi-
cally, the regions where we calculate are marked by $\hat{M}$. As shown in Fig. 7(a), the deviation of the shaky frames decreases rapidly with each iteration. Furthermore, we have pointed out that plugging a stabilized video into the stabilization system should not have an impact on the input. Thus, we plug stable frames into the coarse-to-fine stabilizers and iteratively stabilize them. The result is shown in Fig. 7(b). The deviation of each frame is perturbed around the value of zero. Obviously, our method has no effect on stable frames, which shows that stabilized video is indeed the fixed point of our developed stabilization system.

7. Conclusion

In this paper, we have presented a fast full-frame video stabilization technique based on the iterative optimization strategy. Our approach can be interpreted as the combination of probabilistic stabilization network (coarse-to-fine extension of PDC-Net) and video outpainting outwork (flow-based image outpainting). When trained on synthetic data constructed within the optimization-based learning framework, our method achieves state-of-the-art performance at a fraction of the computational cost of other competing techniques. It is also empirically verified that stabilized video is the fixed point of the stabilization network.

References


