Digital Video Stabilization: Methods, Datasets, and Evaluation

Marcos Roberto e Souza*
Helena de Almeida Maia (Co-Advisor)
Hélio Pedrini (Advisor)
Institute of Computing
University of Campinas - UNICAMP

***Abstract*—Our thesis addressed digital video stabilization, a process that removes unwanted shakes from videos via software. We performed a thorough review, which resulted in two survey papers. We also studied and proposed a new stability measure aligned with human perception and a novel method for evaluating 2D camera motion to assess video quality better. Next, we introduced NAFT, a semi-online DWS with a new neighborhood-aware mechanism. This method stabilizes videos without relying on an explicit definition of stability. To train NAFT effectively, we created SynthStab, a paired synthetic dataset. NAFT achieves stabilization quality comparable to non-DWS methods, with a significantly smaller model (a 14× reduction).**

I. Introduction

Digital video stabilization is a cost-effective and convenient software-based approach to deal with shaky footage, eliminating the need for specialized hardware. It also allows the enhance of existing videos, resulting in a smoother and more enjoyable viewing experience.

There are three main categories of digital video stabilization depending on how information is used: *online*, *offline*, and *semi-online*. Online methods analyze only the current and preceding frames for stabilization, making them ideal for streaming. On the other hand, offline methods can access the entire video beforehand, allowing for better stabilization by considering future motion. Finally, semi-online methods leverage data from a frame's neighborhood, achieving a balance between used data and quality. It is important to note that real-time stabilization can be implemented in an offline, online, and semi-online manner.

Traditional video stabilization follows a three-step process: camera motion estimation, unwanted motion determination, and stabilized view rendering. The first step tracks the camera path during recording. The second step identifies shaky motions for removal. Finally, we generate new stabilized frames according to the removed motion. Recently, researchers introduced approaches based on deep learning, such as Direct Warping Stabilization (DWS). DWS directly predicts the transformation needed to warp an unstable frame into a stable one. This approach is usually trained on datasets of paired stable and unstable videos. Authors claim it delivers superior results for low-quality videos while requiring less computational power [1], [2].

* This work relates to a PhD thesis.

Our research focused on reducing key limitations in digital video stabilization literature. We investigated traditional and DWS methods. To improve understanding of non-DWS approaches, we proposed a new evaluation strategy specifically for motion estimation. For DWS methods, we introduced a novel video stabilization technique. We prioritized the three following limitations: (i) the lack of a well-structured literature review; (ii) insufficient evaluation methods leading to a limited understanding of metric effectiveness, and (iii) our perceived gap in stability achieved by DWS methods compared to classical approaches. To address these limitations, our work aimed to: (i) provide a comprehensive overview of video stabilization research, (ii) expand knowledge on stabilization assessment, and (iii) enhance the effectiveness and efficiency of DWS approaches.

The thesis produced several key outcomes: (i) a critical and comprehensive review of digital video stabilization methods (first survey) [3]; (ii) a critical and detailed review of video stabilization assessments and datasets (second survey) [4]; (iii) a metric for evaluating two-dimensional camera motion estimation [5]; (iv) new evaluation measures for final stabilization quality based on pixel profile kinematics (Section II); (v) a new synthetic dataset containing paired stable and unstable videos [6]; and (vi) a novel direct warping stabilization method [6]. Due to the textual nature and length of the first two products, they are referenced within the thesis but not included in this summary.

II. Video Stabilization Assessment

Our research examines the evaluation process for classical video stabilization methods, typically assessed only at the final stage (Figure 1). We argue for evaluating each step (1-3) independently using dedicated datasets with ground truth information. This would allow for more targeted improvements. Ideally, assessments in steps 2 and 3 would leverage insights from the final human-centric evaluation (step 4) to guide the development of appropriate physical property measures for automated evaluation. We proposed an evaluation metric specifically for the motion estimation step (step 1).

A. Rethinking 2D Camera Motion Assessment

While advancements have been made in camera motion estimation, the evaluation of 2D methods is often neglected.

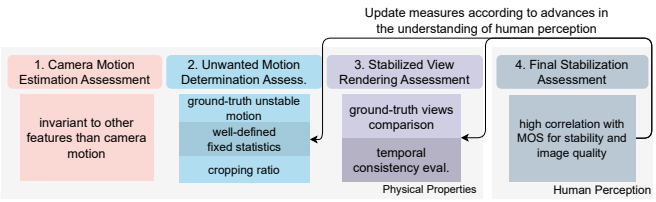
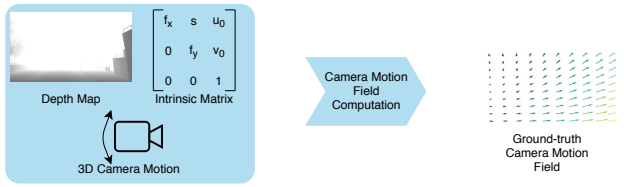
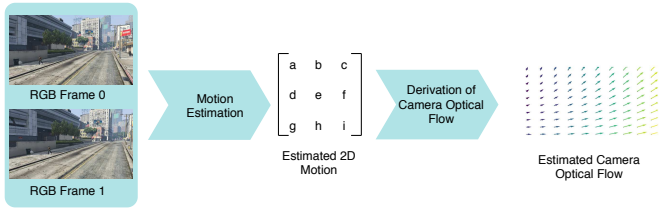


Fig. 1: Diagram of our ideal process proposed for the video stabilization assessment. Source: Author's Thesis.

We proposed a novel evaluation approach for 2D methods that leverages camera motion fields for pixel-wise comparisons. Our experiments validate the reliability of our metrics across various scenarios, illustrating their advantage over conventional image similarity metrics. As illustrated in Figure 2, our evaluation process involves several key steps. First, we establish a ground-truth camera motion field using the relative 3D camera motion, depth map, and camera intrinsics (Figure 2a). This necessitates datasets containing this information. The motion estimation method under evaluation only receives RGB frames as input (Figure 2b). After running the method, we derive the camera optical flow from its prediction. Finally, we perform a pixel-wise comparison between both representations using established metrics from the optical flow literature. We specifically chose metrics with the finer granularity to facilitate comparisons across different representations and degrees of freedom.



(a) ground-truth information



(b) 2D motion estimation

Fig. 2: Main steps of the two-dimensional motion estimation assessment method. Source: Author's Thesis.

1) Experimental Results: Our proposed evaluation strategy, detailed in Chapters 2 and 5 of the thesis, is computed as Average Error per Pixel (AEPE) and Flow Loss (Fl). Table I explores the correlation between these metrics and traditional image similarity metrics (PSNR and SSIM). Since we are comparing similarity metrics with error losses, we expected a correlation of -1 to indicate equivalence. This analysis is motivated by the existing literature's reliance on similarity measures for 2D camera motion assessment.

TABLE I: Correlation between image similarity and EPE-based metrics for test splits.

Dataset	Image Similarity	AEPE Average PLCC	AEPE Average SROCC	Fl PLCC	Fl SROCC
TartanAir	**PSNR**	-0.278±0.29	-0.635±0.19	-0.553±0.16	-0.543±0.17
	SSIM	-0.291±0.29	-0.703±0.13	-0.699±0.18	-0.693±0.18
MVS-Synth	**PSNR**	-0.239±0.01	-0.428±0.01	-0.352±0.01	-0.406±0.01
	SSIM	-0.360±0.06	-0.588±0.03	-0.657±0.03	-0.631±0.02
KITTI	**PSNR**	-0.259±0.19	-0.660±0.23	-0.574±0.26	-0.562±0.28
	SSIM	-0.284±0.19	-0.727±0.26	-0.623±0.28	-0.569±0.30

Our proposed metrics (AEPE and Fl) showed a low correlation with traditional similarity metrics (PSNR and SSIM). To understand this divergence, we analyzed specific cases where the two approaches disagreed (detailed in Table II). In these scenarios, our metrics consistently delivered accurate results, whereas similarity metrics struggled. This is because similarity metrics are susceptible to factors beyond motion, while our approach isolates motion more effectively.

TABLE II: Description of main cases where similarity metrics do not seem to be adequate to assess the quality of camera motion estimation.

Case	Description	Expected Behavior
Low-textured Frames	Frames where neighboring pixels are very similar.	Similarity metrics do not show much difference when we change the camera motion.
High-textured Frames	Frames where neighboring pixels are very different.	Similarity metrics can be very distinct, even with low changes in the camera motion.
Abrupt Camera Motion	Relative camera motion for two frames is very large.	Borders generated on warped images significantly reduce the similarity value of images.
Large Moving Objects	Many pixels are covered by moving objects.	Compensating for camera motion results in low similarity in pixels of moving objects.
Lighting Variation	Pixels are affected by a change in lighting.	Low values in similarity metrics in regions affected by lighting variation.

While our evaluation method performs very well in controlled settings with high-quality, ground-truth data, its applicability is limited to such scenarios. Additionally, the method struggles with reflective surfaces where camera motion manifests differently compared to non-reflective areas. These limitations are trade-offs for the method's effectiveness in rigorous quality assessments.

B. An Analysis on Final Stability Assessment

Inspired by the kinematic principles used by Grundmann et al. [7], we investigated novel kinematic measures computed from the first, second, and third derivatives of pixel profiles [8]. We observed that these measures seemed to better capture video stability compared to methods focused on frequency analysis [9]. We also leveraged these kinematic measures along with other statistical metrics to create a feature vector. This vector served as input to train a regressor to predict human-perceived stability scores.

Our kinematic measures analyze the motion of camera pixels: Velocity of Camera Pixel Profiles (VCP^2), Acceleration

of Camera Pixel Profiles (ACP^2), and Jerk of Camera Pixel Profiles (JCP^2). To isolate camera motion, we employ a segmentation mask that excludes pixels belonging to moving objects. We opted for pixel profiles over feature trajectories for two reasons: first, profiles offer a dense representation, capturing all pixels within a frame. Second, they are simpler to implement as they avoid tracking features across the video, which can introduce complexities like temporal discontinuity or features leaving the frame entirely.

Our machine learning approach to video stability assessment considers multiple aspects. Inspired by existing methods, we categorized these aspects into four dimensions: (i) image similarity, (ii) frequency analysis, (iii) geometry, and (iv) our proposed kinematic measures. Each dimension is evaluated using specific metrics, typically providing a value per pixel or frame. To create a single representative score, we employed six statistical measures: average, standard deviation, median, interquartile range, kurtosis, and skewness.

1) Experimental Results: Our analysis (Table III) explores correlations between human-rated stability scores (from LIVE-Qualcomm and MIND-VQ datasets), existing video stabilization metrics, our proposed kinematic measures, and various regression models using different inputs. To account for untrainable measures, we averaged correlation values across 10 test sets. We evaluated five regression models, including a simple linear fit, using a 3-fold cross-validation process with hyperparameter tuning.

TABLE III: Correlation between human perception of stability scores and different strategies for assessing stability.

Measure	LIVE-Qualcomm		MIND-VQ	
	PLCC	SROCC	PLCC	SROCC
LHR	0.388	0.404	0.538	0.489
ITF (PSNR)	0.015	0.024	0.317	0.308
ITF (SSIM)	0.081	0.072	0.200	0.196
IGC	0.626	0.614	-	-
VCP^2	0.204	0.405	0.546	0.555
ACP^2	**0.710**	0.720	**0.781**	**0.769**
JCP^2	0.636	**0.755**	0.753	0.747
Linear Fit	0.142	0.181	0.706	0.767
SVR	-	-	0.853	0.815
RF	-	-	0.880	0.839
GBM	-	-	**0.884**	**0.843**
XGBoost	-	-	**0.884**	0.842

Acceleration of Camera Pixel Profiles (ACP^2) consistently achieved the strongest correlations with human-rated stability scores (except for SROCC in the LIVE-Qualcomm dataset), while existing metrics, such as LHR and ITF, showed weak correlations. This suggests that commonly reported quantitative measures may not reflect the human perception of stability. Additionally, we explored trained regression models for predicting stability scores based on various features. It was not trained on the LIVE-Qualcomm dataset due to its small size, which leads to overfitting. In the MIND-VQ dataset, our model achieved a 10.3% correlation improvement compared to the best non-machine learning metric. Despite this success, our model did not succeed in a cross-dataset scenario.

III. SynthStab

SynthStab, our novel synthetic dataset, offers a new approach to training video stabilization models. Unlike existing methods that rely on pixel similarity, with SynthStab we can use camera motion for supervision. SynthStab also ensures that unstable motions realistically represent scene constraints, including depth variations. We leveraged Unreal Engine and AirSim environments (Figure 3) to generate diverse, realistic videos with precisely controlled camera movements (Figure 4). The dataset is divided into two parts: (i) SynthStab-SL, with 424 short videos (100 frames each) featuring low-intensity instabilities; and (ii) SynthStab-LH, which comprises 60 long videos (1000 frames each) with high-intensity instabilities. Both subsets provide RGB frames at 512×256 resolution, dense depth maps for stable and unstable versions, and motion fields for each frame pair. SynthStab offers more than 100,000 frames for deep model training. We randomly partitioned each part of the dataset into training and validation sets.

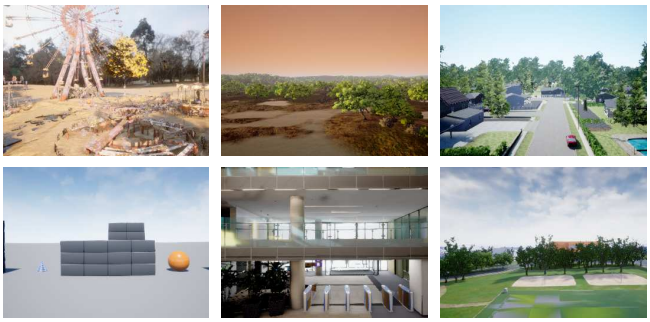


Fig. 3: Environments present in our dataset. We have simple, complex, indoor and outdoor environments. Source: Author's Thesis.

SynthStab's construction process involves four key steps (Figure 4). First, we generate realistic camera motion paths. We consider all six degrees of freedom (6-DoF) and independently create stable trajectories for each using principles from Grundmann et al. [7]. These trajectories consist of three segment types: constant position, constant velocity, and constant acceleration. Next, we define unstable trajectories. To ensure consistency with the stable path, we randomly scatter key points along the path and connect them with a random path. We generate a set number of these unstable trajectories and create a corresponding video pair for each trajectory within each environment (using Unreal Engine and AirSim). Finally, using the depth map, the relative 3D camera motion between frames, and intrinsic camera parameters, we calculate dense camera motion fields.

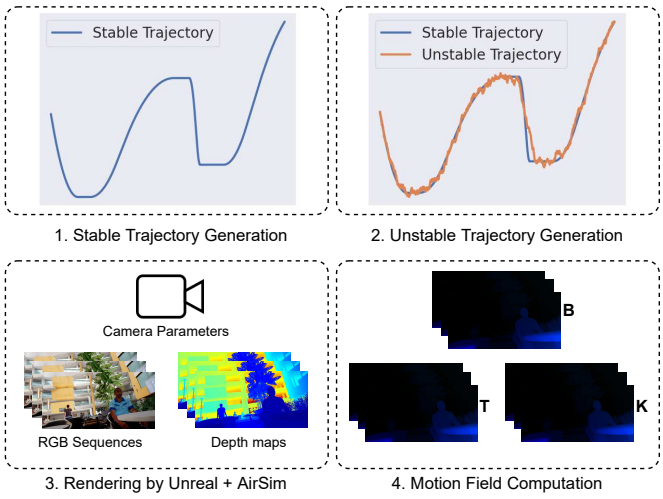


Fig. 4: Overview of the construction process of our dataset. Source: Author's Source.

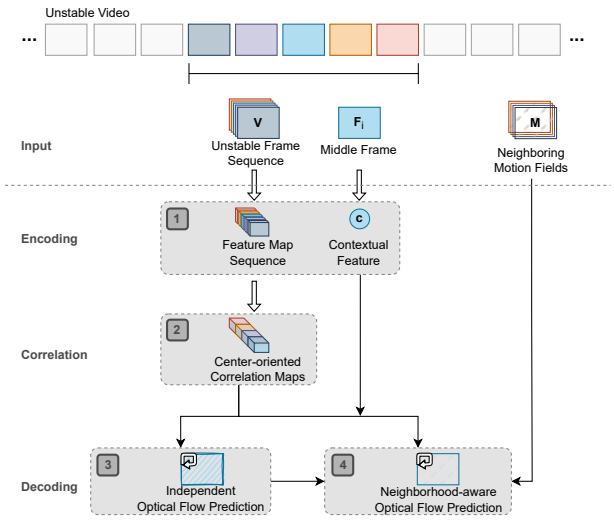


Fig. 5: Training Stage. Our training input is a set of unstable frames and a set of neighboring motion fields. The output is the optical flow to stabilize the middle frame. Source: Author's Source.

IV. NAFT

This section presents Neighborhood-Aware Recurrent All-Pairs Field Transforms (NAFT), a novel semi-online DWS method. NAFT is based on RAFT, an optical flow method, and incorporates a new update mechanism called Iterative Update aware of the Neighborhood Outputs (IUNO). This mechanism allows NAFT to learn stability characteristics directly from data patterns within the SynthStab dataset, without the need for an explicit definition of stability in loss function. Additionally, we demonstrate how to integrate an existing video inpainting method to achieve full-frame stabilization. Our experiments show that NAFT effectively stabilizes videos with intense camera motion, surpassing other DWS methods and approaching the performance of state-of-the-art techniques. Our smallest network variant, NAFT-S, requires only around 7% of the model size and trainable parameters compared to the most lightweight existing methods.

A. Proposed Method

Figure 5 summarizes our training process. Let $\mathbf{V}_i = \{\mathbf{F}_{i-d_\Omega}, \mathbf{F}_{i-d_{\Omega-1}}, \cdots, \mathbf{F}_i, \cdots, \mathbf{F}_{i+d_{\Omega-1}}, \mathbf{F}_{i+d_\Omega}\}$ represent a sequence of RGB unstable frames, where each frame $\mathbf{F}_\omega \in [0,1]^{H\times W\times 3}$. The set $\mathbf{d} = \{d_1, \cdots, d_{\Omega-1}, d_\Omega\}$, with size Ω, defines the displacements of the input sequences. Let $\mathbf{M}_i^{\text{ngb}} = \{\mathbf{M}_{i-d_\Omega}, \mathbf{M}_{i-d_{\Omega-1}}, \cdots, \mathbf{M}_{i+d_{\Omega-1}}, \mathbf{M}_{i+d_\Omega}\}$ denote a sequence of motion fields for neighboring frames, where each motion field $\mathbf{M}_\omega \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2}$ warps its corresponding unstable frame $\mathbf{F}_\omega$ into its stable version $\bar{\mathbf{F}}_\omega$. Our goal is to predict the optical flow $\mathbf{B}_i$, of size H×W×2, which warps $\mathbf{F}_i$ into its stabilized version $\tilde{\mathbf{F}}_i$ using the information from $\mathbf{V}_i$ and $\mathbf{M}_i^{\text{ngb}}$.

We divide the training stage into four substeps (Figure 5): (i) computation of feature maps for each frame in $\mathbf{V}_i$ and the contextual feature for $\mathbf{F}_i$; (ii) computation of the $\mathbf{F}_i$-oriented correlation maps; (ii) initial iterative decoding of $\mathbf{B}_i$; and (4) final iterative decoding of $\mathbf{B}_i$, which incorporates the neighboring motion fields $\mathbf{M}^{\text{ngb}}$. Our model is supervised using two terms: a pixel-wise loss between the predicted optical flows and the motion field, and a smoothness loss. Interestingly, even though the two decoders have slightly different tasks, they are trained with the same loss function. This strategy allows our network to learn how to stabilize videos from data. The network is trained to predict an optical flow based on a pattern composed of: (i) the frame that will be warped by the predicted optical flow; (ii) the neighboring frames; (iii) an initial optical flow; and (iv) the neighboring motion fields. Consequently, the network learns to predict a video with stabilized 3D motion from the 2D information of the frames, without relying on explicit assumptions or simplifications.

Figure 6 summarizes our inference stage. Let $\mathbf{V} = \{\mathbf{F}_0, \mathbf{F}_1, \cdots, \mathbf{F}_N\}$ be an input unstable video, where each frame $\mathbf{F}_i \in [0,1]^{H\times W\times 3}$. Our goal is to compute a sequence of optical flows $\mathbf{B} = \{\mathbf{B}_0, \mathbf{B}_1, \cdots, \mathbf{B}_N\}$, with each $\mathbf{B}_i \in \mathbb{R}^{H\times W\times 2}$. These optical flows are then applied to their respective unstable frames in $\mathbf{V}$ to produce the initial stabilized video $\tilde{\mathbf{V}}^0 = \{\tilde{\mathbf{F}}_0^0, \tilde{\mathbf{F}}_1^0, \cdots, \tilde{\mathbf{F}}_N^0\}$. Optionally, we also compute the frame boundaries masks $\mathcal{M} = \{\mathcal{M}_0, \mathcal{M}_1, \cdots, \mathcal{M}_N\}$. These masks, along with the warped frames, are input to the video inpainting method, resulting in the final stabilized video $\tilde{\mathbf{V}} = \{\tilde{\mathbf{F}}_0, \tilde{\mathbf{F}}_1, \cdots, \tilde{\mathbf{F}}_N\}$.

During inference, NAFT differs slightly from the training process. We compute a contextual map and correlation maps for each frame, rearranging them into sequential batches for processing. In each iteration, the second decoder uses the neighboring optical flows predicted in the previous step, replacing the fixed motion fields used during training. NAFT is a semi-online method: instead of handling the entire sequence simultaneously or a frame at a time, we use a subset of frames

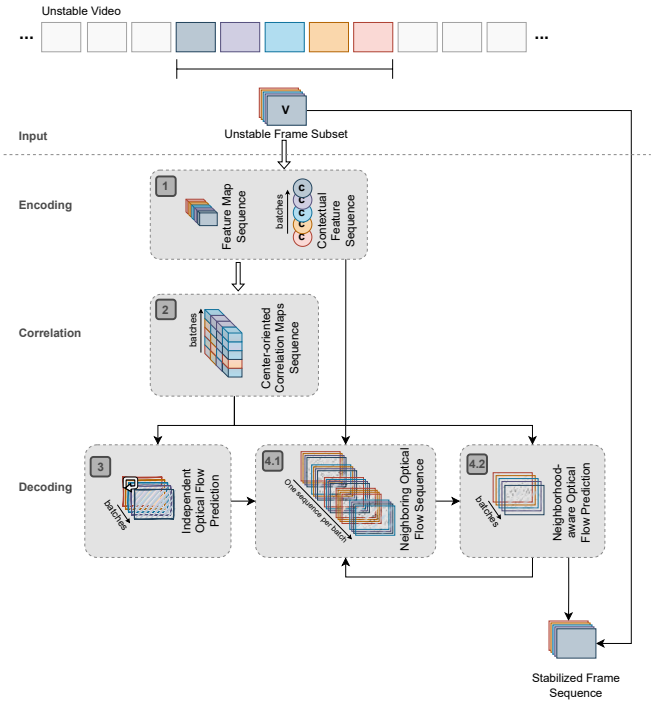


Fig. 6: Inference Stage. The input to the inference is a subset of unstable frames. The output is the optical flow (and the stabilized frames) for the subset. Source: Author.

within a sliding window. This window includes a specified number of anchor frames before the valid frames, the valid frames, and lookahead frames. We deal with each type of frame in a specific way, but we use only the optical flow of the valid ones to stabilize the video.

B. Experimental Results

To evaluate NAFT, we compared it to five existing video stabilization methods on the NUS Dataset [9]. This dataset contains 144 natural, unstable videos categorized based on camera motion and scene features. We classified the methods into two groups: DWS and non-DWS methods. Table IV summarizes the results in terms of frames per second (FPS), model size (MB), and the number of learnable parameters (M). The best overall results are highlighted in bold, while the best results among DWS methods are underlined.

TABLE IV: Statistics of Computational Resources.

	Methods	FPS	Size	Params
Others	Deep3D	0.8	36.0	37.2
	DIFRINT	10.6	38.0	9.9
	DUT	4.9	54.4	10.0
DWS	PWStab.	**30.0**	186.0	48.5
	StabNet	13.0	116.0	32.4
	NAFT	4.9	23.0	5.9
	NAFT-S	8.7	**2.7**	**0.7**

NAFT achieves an FPS comparable to the DUT method and even surpasses it when using the smaller variant (NAFT-S). NAFT excels in terms of model size and number of parameters. It requires roughly 80% less model size and 82% fewer parameters than the smallest existing DWS network, StabNet. Compared to state-of-the-art methods, such as Deep3D and DIFRINT, NAFT is about 37% smaller and requires 41% fewer parameters. NAFT-S pushes these advantages even further, demanding only around 2% of the model sizes and parameters needed by StabNet, Deep3D, and DIFRINT. This demonstrates that NAFT, a DWS method, achieves performance on par with non-DWS approaches while significantly reducing computational requirements in terms of both model size and learnable parameters.

We also evaluated NAFT by changing its neighborhood sizes (Figure 7) and compared it to state-of-the-art methods. This included StabNet (online, 30-frame past neighborhood), PWStableNet (semi-online, 15-frame past, and future neighborhood), and the offline methods DIFRINT, Deep3D, and DUT. NAFT consistently outperformed StabNet and PWStableNet even at similar neighborhood ranges. While PWStableNet achieves plateaus after a 15-frame neighborhood (as shown by Zhao et al. [1]), NAFT continues to improve stability with a larger neighborhood. This allows NAFT to achieve stability levels similar to those of the top offline methods, Deep3D and DUT.

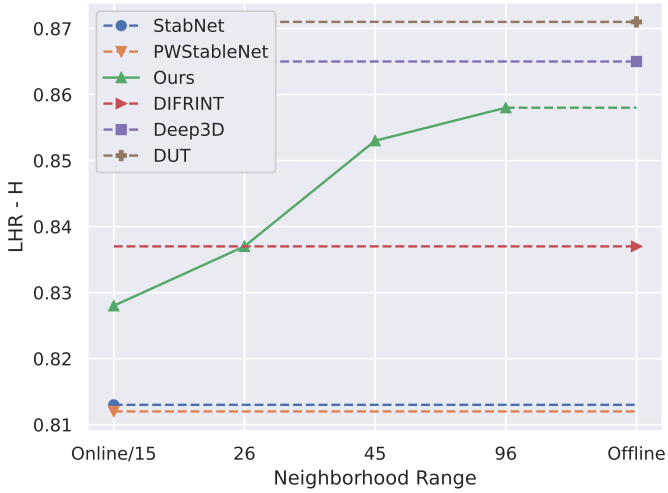


Fig. 7: Results with different neighborhood sizes. Markers show the neighborhood used by each experiment. Source: Author.

In more detailed experiments, shown in the thesis, NAFT consistently outperformed other DWS methods in terms of numerical video stability across various NUS Dataset categories. For non-DWS concurrents, the leading method varied depending on the metric used. For instance, using the LHR-H metric, NAFT achieved the best overall scores in the Quick-Rotation and Regular categories. Similarly, with the LHR-OF metric, NAFT showed the best results in the Regular and Running categories. NAFT's image distortion metrics were highly competitive with DUT and Deep3D, even without video inpainting. When inpainting was applied, NAFT achieved the

best results in most categories. Additionally, NAFT's cropping rate remained comparable to DUT and Deep3D. For a detailed breakdown of results, please refer to the full thesis document. A supplementary video illustrating our results is available at github.com/marcoosrs/NAFT.

Figure 8 shows a comparison of NAFT, DIFRINT, and FuSta. DIFRINT introduces noticeable artifacts throughout the video sequence. While FuSta exhibits fewer artifacts, they remain visible. In contrast, our method effectively minimizes artifacts, resulting in a more realistic and visually appealing output.

(a) DIFRINT

(b) FuSta

(c) Ours

Fig. 8: Subjective comparison of the sequence of frames filled by FuSta, DIFRINT, and E^{2}FGVI (with fine-tuning). Source: Author.

NAFT exhibits three main quality limitations: (i) it may introduce spatial distortions in some frames, particularly during intense stabilization (Running category videos); (ii) in certain cases, NAFT may not refine instabilities as effectively as traditional methods; (iii) large holes or rapid motions can hinder inpainting performance, a known issue shared by E^{2}FGVI and similar methods. Additionally, NAFT is not the fastest DWS method and has relatively high memory usage, potentially affecting high-resolution inference. These limitations are likely tied to our correlation strategy. Exploring alternative DWS correlation approaches could mitigate these issues. Furthermore, our current implementation requires two network passes, which is inefficient for the iterative refinement process. Optimizing this process could significantly improve runtime performance.

V. Conclusions

This study aimed to bridge key gaps in video stabilization research. We established a structured framework and taxonomy, examining existing methods and evaluation metrics. Then, we proposed a new framework for assessing stabilization quality, featuring a novel method for motion estimation assessment and kinematic-based stability measures. Furthermore, we introduced NAFT, a stabilization network based on RAFT, which surpasses existing DWS methods in performance, with a significantly reduced model size requiring up to 93% fewer parameters. NAFT was trained with SynthStab, our novel synthetic dataset containing over 100,000 video frames.

VI. Thesis Products

This doctoral research resulted in several scientific publications across computer science, computer vision, video processing, machine learning, image processing, and image analysis. These publications can be classified as core, related, and non-related to the dissertation topic. Table V provides the number of articles published in each category. We also present some details for the core publications, such as impact factor, CAPES Qualis ranking (based on the 2017-2020 Quadrennial Evaluation), and the highest percentile (as of July 2023 according to Scopus). A full list of publications from this doctoral studies can be found in the following references.

TABLE V: Number of publications during the doctoral research period by publication vehicle and level of proximity to the thesis.

	Core	**Related**	**Non-Related**	**Total**
Journal	4	6	2	12
Conference	0	3	6	9
Book Chapter	0	0	2	2
Total	4	9	10	**23**

1) *Survey on Digital Video Stabilization: Concepts, Methods, and Challenges* [3]. This paper **was published** in the **peer-reviewed Journal** called **ACM Computing Surveys**, with an **impact factor of 14.324** (2021), **Qualis A1** and **highest percentile of 99%**.
2) *Survey on Digital Video Stabilization: Datasets and Evaluation* [4]. This paper is **submitted** and under review in the **peer-reviewed Journal** called **ACM Computing Surveys**, with an **impact factor of 14.324** (2021), **Qualis A1** and **highest percentile of 99%**.
3) *Rethinking Two-Dimensional Camera Motion Estimation Assessment for Digital Video Stabilization: A Camera Motion Field-based Metric* [5]. This article **was published** in the **peer-reviewed Journal** called **Neurocomputing**, with an **impact factor of 5.779** (2022), **Qualis A1** and **highest percentile of 93%**.
4) *NAFT and SynthStab: A RAFT-based Network and a Synthetic Dataset for Digital Video Stabilization* [6]. This paper is **submitted** and under review in the **peer-reviewed Journal** called **International Journal of Computer Vision**, with an **impact factor of 19.500** (2022), **Qualis A1** and **highest percentile of 97%**.

REFERENCES

[1] M. Zhao and Q. Ling, "PWStableNet: Learning Pixel-Wise Warping Maps for Video Stabilization," *IEEE Transactions on Image Processing*, vol. 29, pp. 3582–3595, 2020.

[2] 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 Transactions on Image Processing*, vol. 28, no. 5, pp. 2283–2292, 2019.

[3] M. R. Souza, H. A. Maia, and H. Pedrini, "Survey on Digital Video Stabilization: Concepts, Methods, and Challenges," *ACM Computing Surveys*, vol. 55, no. 3, pp. 1–37, 2022.

[4] M. R. Souza, H. A. Maia, and H. Pedrini, "Survey on Digital Video Stabilization: Datasets and Evaluation," *ACM Computing Surveys (submitted)*, 2023.

[5] M. R. Souza, H. d. A. Maia, and H. Pedrini, "Rethinking Two-Dimensional Camera Motion Estimation Assessment for Digital Video Stabilization: A Camera Motion Field-based Metric," *Neurocomputing*, p. 126768, 2023.

[6] M. R. Souza, H. A. Maia, and H. Pedrini, "NAFT and SynthStab: A RAFT-based Network and a Synthetic Dataset for Digital Video Stabilization," *International Journal of Computer Vision (submitted)*, 2023.

[7] M. Grundmann, V. Kwatra, and I. Essa, "Auto-Directed Video Stabilization with Robust L1 Optimal Camera Paths," in *Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2011, pp. 225–232.

[8] S. Liu, L. Yuan, P. Tan, and J. Sun, "Steadyflow: Spatially Smooth Optical Flow for Video Stabilization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4209–4216.

[9] S. Liu, L. Yuan, P. Tan, and J. Sun, "Bundled Camera Paths for Video Stabilization," *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 1–10, 2013.

[10] Z. Li, C.-Z. Lu, J. Qin, C.-L. Guo, and M.-M. Cheng, "Towards an End-to-End Framework for Flow-Guided Video Inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 562–17 571.

[11] H. Maia, M. e Souza, A. Santos, H. Pedrini, H. Tacon, A. Brito, H. Chaves, M. Vieira, and S. Villela, "Learnable Visual Rhythms Based on the Stacking of Convolutional Neural Networks for Action Recognition," in *International Conference on Machine Learning and Applications*. IEEE, 2019, pp. 1–6.

[12] M. R. Souza and H. Pedrini, "Visual Rhythms for Qualitative Evaluation of Video Stabilization," *EURASIP Journal on Image and Video Processing*, vol. 2020, pp. 1–19, 2020.

[13] M. R. Souza, H. de Almeida Maia, M. B. Vieira, and H. Pedrini, "Survey on Visual Rhythms: A Spatio-Temporal Representation for Video Sequences," *Neurocomputing*, vol. 402, pp. 409–422, 2020.

[14] M. R. Souza and H. Pedrini, "Digital Video Stabilization based on Adaptive Camera Trajectory Smoothing," *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, p. 37, 2018.

[15] M. R. Souza and H. Pedrini, "Combination of Local Feature Detection Methods for Digital Video Stabilization," *Signal, Image and Video Processing*, vol. 12, no. 8, pp. 1513–1521, 2018.

[16] M. R. Souza and H. Pedrini, "Motion energy image for evaluation of video stabilization," *The Visual Computer*, vol. 35, no. 12, pp. 1769–1781, 2019.

[17] M. R. Souza, L. F. R. da Fonseca, and H. Pedrini, "Improvement of Global Motion Estimation in Two-Dimensional Digital Video Stabilisation Methods," *IET Image Processing*, vol. 12, no. 12, pp. 2204–2211, 2018.

[18] M. R. Souza, J. S. Conceição, J. L. Flores-Campana, L. G. Decker, D. C. Luvizon, G. S. P. Carvalho, H. A. Maia, and H. Pedrini, "Pyramidal Layered Scene Inference with Image Outpainting for Monocular View Synthesis," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2021, pp. 37–46.

[19] D. C. Luvizon, G. S. P. Carvalho, A. A. dos Santos, J. S. Conceicao, J. L. Flores-Campana, L. G. Decker, M. R. Souza, H. Pedrini, A. Joia, and O. A. Penatti, "Adaptive multiplane image generation from a single internet picture," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2556–2565.

[20] A. Pinto, M. A. Córdova, L. G. Decker, J. L. Flores-Campana, M. R. Souza, A. A. dos Santos, J. S. Conceição, H. F. Gagliardi, D. C. Luvizon, R. d. S. Torres *et al.*, "Parallax Motion Effect Generation Through Instance Segmentation and Depth Estimation," in *International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 1621–1625.

[21] M. R. Souza, D. Bertolini, H. Pedrini, and Y. M. Costa, "Offline Handwritten Script Recognition Based on Texture Descriptors," in *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2019, pp. 57–62.

[22] H. d. A. Maia, M. R. Souza, A. C. S. Santos, J. C. M. Bobadilla, M. B. Vieira, and H. Pedrini, "Early Stopping for Two-Stream Fusion Applied to Action Recognition," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics*. Springer, 2020, pp. 319–333.

[23] A. C. S. Santos, H. A. Maia, M. R. Souza, M. B. Vieira, and H. Pedrini, "Fuzzy Fusion for Two-stream Action Recognition," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2020.

[24] J. L. F. Campana, L. G. L. Decker, M. R. Souza, H. d. A. Maia, and H. Pedrini, "Multi-scale patch partitioning for image inpainting based on visual transformers," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*, vol. 1. IEEE, 2022, pp. 180–185.

[25] J. L. F. Campana, L. G. L. Decker, M. R. Souza, H. d. A. Maia, and H. Pedrini, "Variable-Hyperparameter Visual Transformer for Efficient Image Inpainting," *Computers & Graphics*, vol. 113, pp. 57–68, 2023.

[26] J. L. F. Campana, L. G. L. Decker, M. R. Souza, H. d. A. Maia, and H. Pedrini, "Image Inpainting on the Sketch-Pencil Domain with Vision Transformers," *International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 122, p. 132, 2024.

[27] M. R. Souza, A. C. S. Santos, and H. Pedrini, "A Hybrid Approach Using the k-means and Genetic Algorithms for Image Color Quantization," *Recent Advances in Hybrid Metaheuristics for Data Clustering*, pp. 151–171, 2020.

[28] M. R. Souza, H. d. A. Maia, A. C. S. e. Santos, M. B. Vieira, and H. Pedrini, "Multi-Script Video Caption Localization Based on Visual Rhythms," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2032926, 2022.

[29] L. G. L. Decker, J. L. F. Campana, M. R. Souza, H. d. A. Maia, and H. Pedrini, "Zero-Shot Synth-to-Real Depth Estimation: From Synthetic Street Scenes to Real-World Data," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2024.