









Supplementary Material for *StabStitch*

Lang Nie^{1,2}, Chunyu Lin^{1,2}^{*}, Kang Liao³, Yun Zhang⁴,
Shuaicheng Liu⁵, Rui Ai⁶, and Yao Zhao^{1,2}

¹ Institute of Information Science, Beijing Jiaotong University

² Beijing Key Laboratory of Advanced Information Science and Network

³ Nanyang Technological University

⁴ Communication University of Zhejiang

⁵ University of Electronic Science and Technology of China

⁶ HAOMO.AI

1 Overview

In this document, we provide the following supplementary contents:

- Details of the spatial/temporal warp model (Section 2).
- Details of the warp smoothing model (Section 3).
- Details of dataset distribution (Section 4).
- Evaluation metric (Section 5).
- More experiments (Section 6).

Although we present more network details in this supplementary, we argue that these network architectures themselves are not the primary contribution of this work (although we appropriately modified them and achieved improvements). Our main contribution lies in the new paradigm of unsupervised online video stitching, including the representation of stitching trajectories and the design of unsupervised smoothing optimization objectives.

For clarity, we summarize a part of notations and their corresponding meanings in Table 1.

Besides, we also provide a supplementary video. Please refer to <https://www.youtube.com/watch?v=03kGEZJHxzI> for the stitched videos from different methods.

2 Spatial/Temporal Warp Model

Due to the similarity to UDIS++ [7], we just briefly described the structure and loss function of the spatial/temporal model in our manuscript. Here, we give more details in the supplementary material.

We first review the warp model of UDIS++ [7] in Fig. 1(a) and then depict the differences. UDIS++ [7] adopts ResNet50 as the backbone and predicts the control point motions in two steps. The first step estimates the 4-pt homography motions [1] and converts them as the initial control point motions, while the

^{*} Corresponding author: cylin@bjtu.edu.cn

Table 1: The notation table.

Notation	Meaning	Example
(t)	The relative time in a sliding window.	<i>e.g.</i> , $C(t)$, $m(t)$, $S(t)$
Subscript i	The control point index.	<i>e.g.</i> , $C_i(t)$, $m_i(t)$, $S_i(t)$
Superscript T, S	From the temporal or spatial model.	<i>e.g.</i> , $m_i^T(t)$, $M_i^S(t)$
Superscript (ξ)	The absolute time ranging from N to the last frame.	<i>e.g.</i> , $m_i^T(t)$, $M_i^S(t)$
Hat $\hat{\cdot}$	The optimized mesh or trajectory.	<i>e.g.</i> , $\hat{M}^S(t)$, \hat{S}
M^{Rig}	The rigid and regular initial mesh (predefined).	
I_{ref}^t/I_{tgt}^t	The t -th reference/target frame.	
$SNet/TNet/SmoothNet$	The spatial/temporal/smooth warping models.	
$\mathcal{T}_{M^{Rig} \rightarrow M^S(t-1)}(\cdot)$	The TPS transformation from M^{Rig} to $M^S(t-1)$.	

Table 2: Model size (/MB).

	SNet	TNet	SmoothNet	Total
<i>StabStitch</i>	28.55	28.76	2.12	59.43
UDIS++ [7]	297.73	-	-	297.73

second step estimates the residual control point motions, which could reach the final control point motions by addition with initial motions. Both steps leverage the global correlation layer (*i.e.*, the contextual correlation layer [4]) to capture feature matching information and then regress the motions with simple regression networks.

2.1 Structure Difference

We demonstrate the structure differences between the spatial/temporal warp model and UDIS++ in Fig. 1(b)/(c). The differences are highlighted in red/blue. The local correlation layer denotes the cost volume layer [9]. In the spatial warp model, the search radius of the local correlation layer is set to 5, while in the temporal warp model, we set the radius to 6 and 3.

Besides, we further simplify the network architecture, especially the regression networks, significantly reducing the network parameters. For clarity, we compare the model size and report them in Tab. 2.

2.2 Loss Function

The alignment loss and distortion loss for the spatial/temporal warp model are also similar to UDIS++ [7]. One can refer to [7] [6] for more details. For the convenience of readers, we paraphrase their definitions here again.

Alignment Loss: As described above, the spatial/temporal warp model takes two steps to predict the final control point motions $m^S(t)/m^T(t)$ from global homography transformation to local TPS transformation. Assuming the estimated warping functions for homography and thin-plate spline are represented

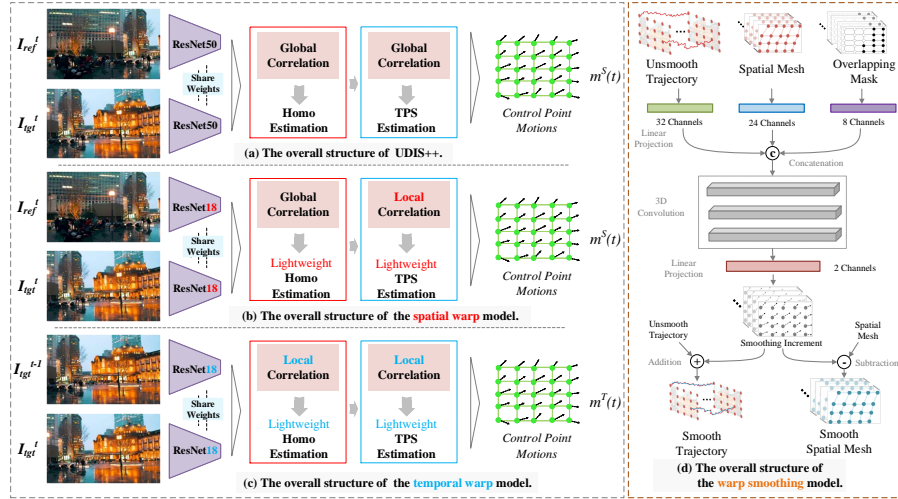


Fig. 1: The overall structures of our models. Left: the spatial/temporal warp model. Right: the warp smoothing model.

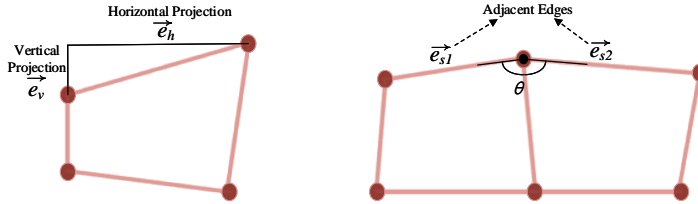


Fig. 2: The intra-grid (left) and inter-grid (right) constraints in the distortion loss.

as $\mathcal{W}_H(\cdot)$ and $\mathcal{W}_T(\cdot)$, the alignment loss is written as:

$$\begin{aligned} \mathcal{L}_{alignment} = & \omega_H \|I_{ref} \cdot \mathcal{W}_H(\mathbb{1}) - \mathcal{W}_H(I_{tgt})\|_1 + \omega_H \|I_{tgt} \cdot \mathcal{W}_H^{-1}(\mathbb{1}) - \mathcal{W}_H^{-1}(I_{ref})\|_1 \\ & + \|I_{ref} \cdot \mathcal{W}_T(\mathbb{1}) - \mathcal{W}_T(I_{tgt})\|_1, \end{aligned} \quad (1)$$

where I_{ref}/I_{tgt} is the reference/target frame, $\mathbb{1}$ is an all-one matrix with the same size as I_{ref} , and ω_H is a constant to balance different transformations.

Distortion Loss: The distortion loss consists of an intra-grid constraint and an inter-grid constraint as follows:

$$\mathcal{L}_{distortion} = \ell_{intra} + \ell_{inter}. \quad (2)$$

The intra-grid term prevents projection distortion caused by excessively large grids after warping by penalizing the grids with side lengths exceeding a certain threshold. As shown in Fig. 2(left), we define the horizontal/vertical projection of each grid edge as e_h/e_v and the corresponding projection set as $\{e_h\}/\{e_v\}$.

Then we can define the intra-grid loss as:

$$\ell_{intra} = \frac{1}{(U+1) \times V} \sum_{\{\mathbf{e}_h\}} \sigma(\mathbf{e}_h - \frac{2W}{V}) + \frac{1}{U \times (V+1)} \sum_{\{\mathbf{e}_v\}} \sigma(\mathbf{e}_v - \frac{2H}{U}), \quad (3)$$

where $H \times W$ and $(U+1) \times (V+1)$ are the image and control point resolutions. $\sigma(\cdot)$ is the *ReLU* activation function.

As for the inter-grid term, it is used to reduce structural distortion caused by inconsistent changes in adjacent grid edges (denoted by $\mathbf{e}_{s1}, \mathbf{e}_{s2}$). As shown in Fig. 2(right), if the changes in adjacent edges are consistent, the included angle should be close to 180. Therefore, we encourage its cosine distance to approximate 1 as follows:

$$\ell_{inter} = \frac{1}{Q} \sum_{\{\mathbf{e}_{s1}, \mathbf{e}_{s2}\}} (1 - \frac{\langle \mathbf{e}_{s1}, \mathbf{e}_{s2} \rangle}{\|\mathbf{e}_{s1}\| \cdot \|\mathbf{e}_{s2}\|}), \quad (4)$$

where $\{\mathbf{e}_{s1}, \mathbf{e}_{s2}\}$ and Q are the set of horizontal and vertical adjacent edges and their number.

3 Warp Smoothing Model

The network structure of the warp smoothing model is exhibited in Fig. 1(d). Although its architecture is very simple (merely consisting of several fully connected layers and 3D convolutions), it can still achieve good results with effective and reasonable loss constraints.

4 Dataset

The videos in our dataset consist of three parts: some original videos from [11], some stable videos from [10], and our captured videos. These videos are captured with arbitrary and irregular motion trajectories. Therefore, we leverage them to simulate two videos from different perspectives. Specifically, we collect the video pair from different timestamps (*e.g.*, one video is from the original video, and the other video is captured after a random delay time). After that, we crop the video frames to simulate an appropriate overlapping rate in stitching. Considering the videos are collected from different timestamps, we further filter out the videos with obvious moving objects. Finally, we get over 100 video pairs and demonstrate the distribution of video duration in Fig. 3.

5 Evaluation Metric

To quantitatively evaluate the proposed method, we suggest three metrics as described in the following:

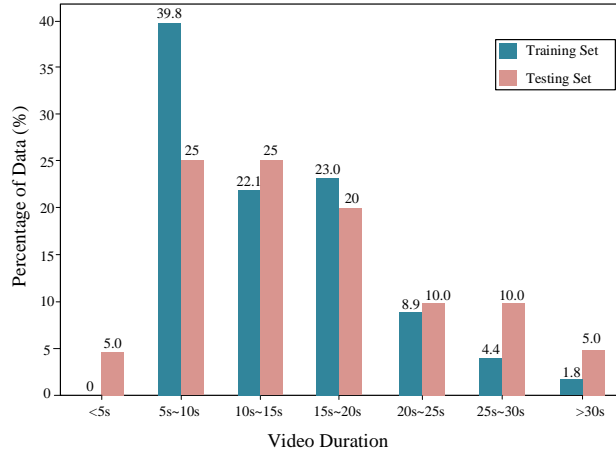


Fig. 3: The distribution statistics of the video duration time.

Alignment Score: Following the criterion of UDIS [5] and UDIS++ [7], we also adopt PSNR and SSIM of the overlapping regions to evaluate the alignment performance. We average the scores in all video frames.

Distortion Score: The final warp in the online stitching mode can be described as a series of meshes: $\hat{M}^{S(N)}(N)$, $\hat{M}^{S(N+1)}(N)$, \dots , $\hat{M}^{S(\xi)}(N)$, \dots . Then we adopt $\mathcal{L}_{distortion}(\cdot)$ to measure the distortion magnitude. Because any distortion in a single frame will destroy the perfection of the whole result, we choose the mean value of the maximum distortion loss of each video as the distortion score.

Stability Score: The smoothed trajectories in the online stitching mode can also be described as a series of positions: $\hat{S}^{(N)}(N)$, $\hat{S}^{(N+1)}(N)$, \dots , $\hat{S}^{(\xi)}(N)$, \dots . Then we adopt $\mathcal{L}_{smoothness}(\cdot)$ to measure the stability. The stability score is the mean value of the average smoothness loss of each video.

Please note that in the comparative experiments, we only adopt the alignment score because different methods define different warp representations and camera trajectories. Thus we only apply the last two metrics to our ablation studies to show the effectiveness of each module.

In the beginning, we evaluate the distortion and stability performance with the metrics that are widely used in video stabilization [3] [2] [12]. These traditional metrics try to estimate the spatial transformation (homography or affine) between adjacent frames from keypoint correspondences. However, the estimated point correspondences are unreliable in our challenging testing cases (*e.g.*, low texture or low light). In addition, as described in Nie *et al.*'s video stitching [8], the metric in the frequency domain (*i.e.*, the stability score in [3]) are not reliable sometimes as the trajectory signals are usually very short and of different lengths. Therefore, we adopt the more intuitive indicators (*i.e.*, the distortion loss and smoothness loss) to describe the distortion and stability performance.

Table 3: More ablation studies about the optimization components of the warp smoothing model on alignment performance (\uparrow).

	Method	Regular	Low-Light	Low-Texture	Moving-Fast	Average
1	Only Spatial Warp	25.60/0.851	35.18/0.960	33.92/0.928	25.57/0.840	30.75/0.903
2	w/o Overlapping Mask (OP)	22.26/0.759	32.14/0.945	30.97/0.913	20.73/0.713	27.40/0.851
3	w/o Smoothness ($\mathcal{L}_{smoothness}$)	25.61/0.851	35.18/0.960	33.92/0.928	25.57/0.840	30.75/0.903
4	w/o Spatial Consistency (\mathcal{L}_{space})	24.64/0.832	34.49/0.958	33.62/0.927	23.39/0.788	29.89/0.890
5	w/o Online Collaboration (\mathcal{L}_{online})	24.70/0.833	34.50/0.958	33.62/0.927	23.39/0.787	29.91/0.890
6	<i>StabStitch</i>	24.64/0.832	34.51/0.958	33.63/0.927	23.36/0.787	29.89/0.890

Table 4: More ablation studies about the optimization components of the warp smoothing model on distortion performance (\downarrow).

	Method	Regular	Low-Light	Low-Texture	Moving-Fast	Average
1	Only Spatial Warp	0.925	0.913	0.767	0.610	0.804
2	w/o Overlapping Mask (OP)	0.624	0.566	0.439	0.509	0.535
3	w/o Smoothness ($\mathcal{L}_{smoothness}$)	0.554	0.598	0.471	0.514	0.534
4	w/o Spatial Consistency (\mathcal{L}_{space})	1.189	1.214	1.081	1.098	1.145
5	w/o Online Collaboration (\mathcal{L}_{online})	0.682	0.695	0.591	0.796	0.691
6	<i>StabStitch</i>	0.661	0.660	0.638	0.735	0.674

Table 5: More ablation studies about the optimization components of the warp smoothing model on stability performance (\downarrow).

	Method	Regular	Low-Light	Low-Texture	Moving-Fast	Average
1	Only Spatial Warp	29.03	26.33	27.35	158.57	60.32
2	w/o Overlapping Mask (OP)	22.22	18.06	15.68	129.00	46.24
3	w/o Smoothness ($\mathcal{L}_{smoothness}$)	28.98	26.34	27.35	158.22	60.22
4	w/o Spatial Consistency (\mathcal{L}_{space})	23.38	19.60	18.67	133.85	48.88
5	w/o Online Collaboration (\mathcal{L}_{online})	23.53	19.69	18.78	135.25	49.33
6	<i>StabStitch</i>	23.18	19.53	18.67	133.59	48.74

6 More Experiment

In this section, we conduct more experiments to explore the roles of different optimization components in the warp smoothing model.

We first report the performance of the spatial warp model and complete *StabStitch*, and then ablate each constraint to show its effectiveness. The alignment, distortion, and stability performance are shown in Tab. 3, Tab. 4, and Tab. 5, respectively.

Data Term: The data term requires the smoothed trajectories to be close to the original trajectories. Without this term, the final output trajectories will degrade to constant paths, yielding meaningless results. Therefore, we ablate the overlapping mask (OP) instead by setting α to 0. As depicted in Tab. 3, the alignment performance significantly decreases from 30.75/0.903 to 27.40/0.851. In this case, extensive artifacts will be produced.

Smoothness Term: The smoothness term works together with the data term to strike a balance between preserving the original trajectories (especially alignment performance) and smoothing the trajectories. Without this term, the output

trajectories will be close to the original trajectories. As shown in Tab. 3 and Tab. 5 (Experiment 1 and 3), the alignment and stability performance is close to that of the spatial warp model. As for the distortion performance reported in Tab. 4, it is significantly improved because of the spatial consistency term. If we further remove the spatial consistency term on the basis of Experiment 3, the distortion score will also approach that of the spatial warp model.

Spatial Consistency Term: With only the data and smoothness terms, every trajectory will be optimized independently. However, there are $(U + 1) \times (V + 1)$ control points, which implies $(U + 1) \times (V + 1)$ trajectories. If each trajectory is smoothed separately without considering the consistency between trajectories, distortions are prone to occur. As shown in Tab. 4 (Experiment 4 and 6), the distortion is significantly increased without this term.

Online collaboration Term: In the online mode, only the last frame in a sliding window (containing N frames) is used. The online collaboration term contributes to the stability of adjacent sliding windows. Without this term, the stability slightly degrades especially in the category of MF, as illustrated in Tab. 5 (Experiment 5 and 6).

The final model (*StabStitch*) does not achieve the best performance in alignment, distortion, and stability. But it reaches the best balance among the three metrics and produces the best visual effect, as demonstrated in our supplementary video.

References

1. DeTone, D., Malisiewicz, T., Rabinovich, A.: Deep image homography estimation. arXiv preprint arXiv:1606.03798 (2016) [1](#)
2. Liu, S., Tan, P., Yuan, L., Sun, J., Zeng, B.: Meshflow: Minimum latency online video stabilization. In: ECCV. pp. 800–815. Springer (2016) [5](#)
3. Liu, S., Yuan, L., Tan, P., Sun, J.: Bundled camera paths for video stabilization. ACM TOG **32**(4), 1–10 (2013) [5](#)
4. Nie, L., Lin, C., Liao, K., Liu, S., Zhao, Y.: Depth-aware multi-grid deep homography estimation with contextual correlation. IEEE TCSVT **32**(7), 4460–4472 (2021) [2](#)
5. Nie, L., Lin, C., Liao, K., Liu, S., Zhao, Y.: Unsupervised deep image stitching: Reconstructing stitched features to images. IEEE TIP **30**, 6184–6197 (2021) [5](#)
6. Nie, L., Lin, C., Liao, K., Liu, S., Zhao, Y.: Deep rectangling for image stitching: A learning baseline. In: CVPR. pp. 5740–5748 (2022) [2](#)
7. Nie, L., Lin, C., Liao, K., Liu, S., Zhao, Y.: Parallax-tolerant unsupervised deep image stitching. In: ICCV. pp. 7399–7408 (2023) [1](#), [2](#), [5](#)
8. Nie, Y., Su, T., Zhang, Z., Sun, H., Li, G.: Dynamic video stitching via shakiness removing. IEEE TIP **27**(1), 164–178 (2017) [5](#)
9. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: CVPR. pp. 8934–8943 (2018) [2](#)
10. Wang, M., Yang, G.Y., Lin, J.K., Zhang, S.H., Shamir, A., Lu, S.P., Hu, S.M.: Deep online video stabilization with multi-grid warping transformation learning. IEEE TIP **28**(5), 2283–2292 (2018) [4](#)
11. Zhang, Y., Lai, Y.K., Zhang, F.L.: Content-preserving image stitching with piecewise rectangular boundary constraints. IEEE TVCG **27**(7), 3198–3212 (2020) [4](#)
12. Zhang, Z., Liu, Z., Tan, P., Zeng, B., Liu, S.: Minimum latency deep online video stabilization. In: ICCV. pp. 23030–23039 (2023) [5](#)