FlowStyler: Artistic Video Stylization via Transformation Fields Transports

Supplementary Material

Algorithm 1: FlowStyler Video Stylization

Input: Content video $\mathbf{I}$, optical flow $\mathbf{F}$, visibility\occlusion mask $\mathcal{M}^{\text{vis}\setminus\text{occ}}$, style image $\mathbf{S}$

Output: Stylized video $\mathbf{I}^{\text{style}}$

for $t \leftarrow 1$ **to** N **do**

$\mathbf{I}_t \leftarrow \text{GlobalColorMatching}(\mathbf{I}_t, \mathbf{S})$ (Sec 3.1.3)

if $t > 1$ **then**

Field Propagation (Sec 3.2):

$\mathbf{M}_t \leftarrow \mathcal{M}^{\text{vis}}_{t\to t-1} \odot \mathcal{A}(\mathbf{M}_{t-1}, \mathbf{F}_{t\to t-1})$

$\mathbf{V}_t \leftarrow \mathcal{M}^{\text{vis}}_{t\to t-1} \odot \mathcal{A}(\mathbf{V}_{t-1}, \mathbf{F}_{t\to t-1})$

Momentum Preserving (Sec 3.3):;

$\mathbf{m}_t^{(0)} \leftarrow \mathcal{M}^{\text{vis}}_{t\to t-1} \odot \mathcal{A}(\mathbf{m}_{t-1}^{(K)}, \mathbf{F}_{t\to t-1})$

$\mathbf{v}_t^{(0)} \leftarrow \mathcal{M}^{\text{vis}}_{t\to t-1} \odot \mathcal{A}(\mathbf{v}_{t-1}^{(K)}, \mathbf{F}_{t\to t-1})$

Occlusion Handling (Sec 3.4):;

$\mathbf{V}_t \leftarrow \text{HistAggregate}(\mathbf{V}_{t-K:t}, \mathcal{M}^{\text{vis}}_{t-K:t\to t})$

$\mathbf{M}_t \leftarrow \text{HistAggregate}(\mathbf{M}_{t-K:t}, \mathcal{M}^{\text{vis}}_{t-K:t\to t})$

for $k \leftarrow 1$ **to** K **do**

Forward Stylization (Sec 3.1):

$\forall (i,j),\ \mathbf{I}_t^{\text{colored}}(i,j) = \mathbf{M}_t(i,j)\mathbf{I}_t(i,j)$

$\mathbf{I}_t^{\text{style}} \leftarrow \text{Splat}(\hat{\mathbf{p}}_t, \mathbf{I}_t^{\text{colored}}),\ \hat{\mathbf{p}}_t \leftarrow \mathbf{p}_t + \mathbf{V}_t \Delta t$

Loss Computation:

$\mathcal{L}_{\text{Total}} \leftarrow$

$\mathcal{L}_{\text{Style}} + \lambda_c \mathcal{L}_{\text{Content}} + \lambda_{\text{ortho}} \mathcal{L}_{\text{ortho}} + \lambda_{\text{TV}} \mathcal{L}_{\text{TV}}$

Momentum-Preserving Update (Sec 3.3):

$\mathbf{g}_t^{(k)} \leftarrow \nabla_{\mathbf{M}_t, \mathbf{V}_t} \mathcal{L}_{\text{Total}} \odot (1 - \alpha_k \mathcal{M}^{\text{occ}}_{t\to t-1})$

$\mathbf{M}_t, \mathbf{V}_t \leftarrow \text{AdamUpdate}(\mathbf{M}_t, \mathbf{V}_t, \mathbf{g}_t^{(k)})$

7. Implementation Details

Our framework is implemented in PyTorch. The kernel size for splatting operations in the advection of the stylization velocity field ranges from 2 to 4 based on scene characteristics. Optical flow estimation employs RAFT [42]. For large areas of homogeneous regions (e.g., sky) where flow estimation becomes unstable and flickering, we implement a heuristic post-processing strategy: identifying these areas via SAM [37] segmentation, then assigning them the mean flow vector from valid surrounding regions to approximate camera motion, enabling robust operation. However, we found that only minimal occurrences requiring such special treatment.

Learning rates differ between the orthogonality-regularized color transfer field (0.001) and stylization velocity field (0.02). The higher rate for the stylization velocity field promotes geometric stylization dominance of it. Optimization converges to satisfactory results in 20 iterations for intermediate frames, while the first frame uses 100 iterations to better form artistic features without compromising temporal consistency. Our method achieves processes each frame in 3 seconds per frame at 768×432 resolution.

A post-optimization gradient smoothing strategy applying spatial Gaussian filtering to field gradients after each iteration, effectively suppressing low-level stylization noise. We implement the bidirectional optimization pipeline from TNST but observed negligible improvements since our momentum-preserving field optimization already ensures temporal consistency. The window size for occlusion-aware temporal lookup adapts per scene, as larger windows adversely affect performance.

8. Global Color Matching

To align video content statistics with style targets, for linear color transformation, we follow the same approach as ARF [54], employing a whitening-and-coloring technique in pixel color space . Let the input content video be represented as $\mathcal{X} = \{\mathbf{I}_t\}_{t=1}^N$ with content frame $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$. We construct our content set $\mathcal{C} = \{\boldsymbol{c}_i\}_{i=1}^m \subset \mathbb{R}^3$ by aggregating all pixel colors from $\mathcal{X}$, while the style image provides target pixels $\mathcal{S} = \{\boldsymbol{s}_i\}_{i=1}^n \subset \mathbb{R}^3$.

The affine transformation matrix $\boldsymbol{A} \in \mathbb{R}^{3\times 3}$ is computed to satisfy:

$$\mathbb{E}[\boldsymbol{A}\mathcal{C}] = \mathbb{E}[\mathcal{S}], \quad \text{Cov}(\boldsymbol{A}\mathcal{C}) = \text{Cov}(\mathcal{S})$$

where the empirical moments are computed as:

$$\mathbb{E}[\mathcal{C}] = \frac{1}{m}\sum_{i=1}^m \boldsymbol{c}_i$$

$$\text{Cov}(\mathcal{C}) = \frac{1}{m-1}\sum_{i=1}^m (\boldsymbol{c}_i - \mathbb{E}[\mathcal{C}])(\boldsymbol{c}_i - \mathbb{E}[\mathcal{C}])^\top$$

$$\text{Cov}(\mathcal{S}) = \frac{1}{m-1}\sum_{i=1}^m (\boldsymbol{s}_i - \mathbb{E}[\mathcal{S}])(\boldsymbol{s}_i - \mathbb{E}[\mathcal{S}])^\top$$

yielding the final transformation:

$$\boldsymbol{A} = \text{Cov}(\mathcal{S})^{1/2}\text{Cov}(\mathcal{C})^{-1/2}$$
$$\boldsymbol{b} = \mathbb{E}[\mathcal{S}] - \boldsymbol{A}\mathbb{E}[\mathcal{C}]$$
$$\boldsymbol{c}' = \boldsymbol{A}\boldsymbol{c} + \boldsymbol{b} \quad \forall \boldsymbol{c} \in \mathbf{I}_t$$

This closed-form solution guarantees frame-wise color distribution alignment with the style image. Such global transformation is applied to the video content frame before applying the orthogonality-regularized color transfer field.

While the above linear transformations offer computational efficiency, they often inadequately preserve the holistic color characteristics of reference style images. For failure cases, we therefore implement global color matching through color style transfer operations [28]. However, frame-wise application of color style transfer introduces minor intensity variations that manifest as global temporal flickering artifacts. We mitigate this through temporal intensity stabilization: per-frame mean intensities are computed, subjected to spline-based smoothing, then globally readjusted across the sequence. This adaptive compensation proves effective across most scenarios while maintaining temporal consistency.

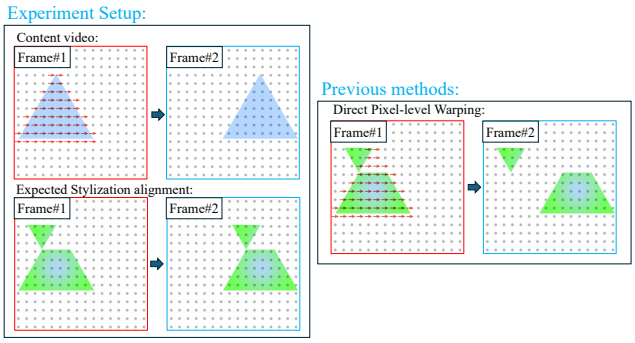

Figure 3. This figure demonstrates the limitations of previous pixel-level optical flow-based transfer methods that suffer from temporal inconsistency. The left panel illustrates a test scenario setup: The upper portion shows a triangular object undergoing lateral translation with corresponding motion vectors (red arrows). In the lower section, the left portion displays a deformable stylization version of the first frame, whereas the right portion presents the desired transferred results with the ideal frame-to-frame consistency. The right panel reveals how conventional approaches fail: Misalignment between geometric stylization and motion propagation causes propagated results to diverge from the previous frame. Only motion-aligned regions are propagated, while non-aligned regions remain unmodified, leading to inconsistent stylized results between frames.

9. Comparison with two earlier optical flow based video style transfer methods in the user study

we conducted a user study comparing our method with [38] and [19] on our dataset using the identical evaluation setup as in our paper. We collected approximately 1400 votes per method from 215 participants prior to the rebuttal deadline. Our two-alternative forced choice study revealed a better user preference for our method: 73.5% over [38] and 63.7% over [19]. Detailed comparisons will be included in the revision due to space constraints.

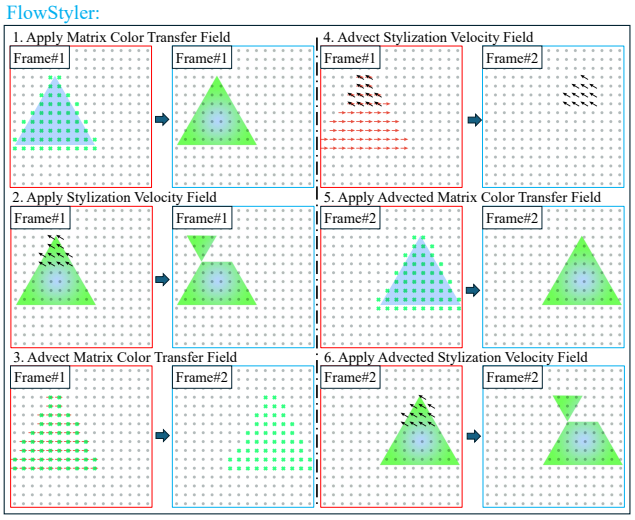

Figure 4. Our framework achieves temporal consistent artistic stylized results through six phased field transport: (1,2) Initial stylization: The matrix color transfer field and stylization velocity field jointly apply artistic geometric stylization transfer on the initial frame. (3,4) Field advection: Both fields are advected by scene motion vectors, as the preparation to maintain inter-frame consistency in the final step. (5,6) Consistent stylization: The advected fields are applied in the subsequent frame's stylization process, preserving consistent artistic features through field-space continuity. This case demonstrates that our motion-aligned transformation fields propagating coherently across frames, ensuring consistent stylized results.

10. Complexity of Interdependence

FlowStyler was designed with modularity in mind. Components are mostly decoupled and can be disabled independently without compromising pipeline functionality, as validated in the ablation studies. Additionally, we implemented fallback mechanisms to handle edge cases during unexpected failures. Our qualitative experimental results also demonstrate FlowStyler's robustness across highly dynamic scenarios.

11. Runtime Comparisons

For a 100-frame video on an A100-80G GPU, optical flow evaluation (RAFT) takes $\sim$10 sec, and FlowStyler completes processing in $\sim$5 min. This demonstrates 3$\times$ faster than generative baselines (AnyV2V: $\sim$15 min, TokenFlow: $\sim$21 min). While non-generative approaches like CAP-VSTNet ($\sim$18 sec) and UniST ($\sim$24 sec) are faster, our method delivers better consistency and superior user preference.

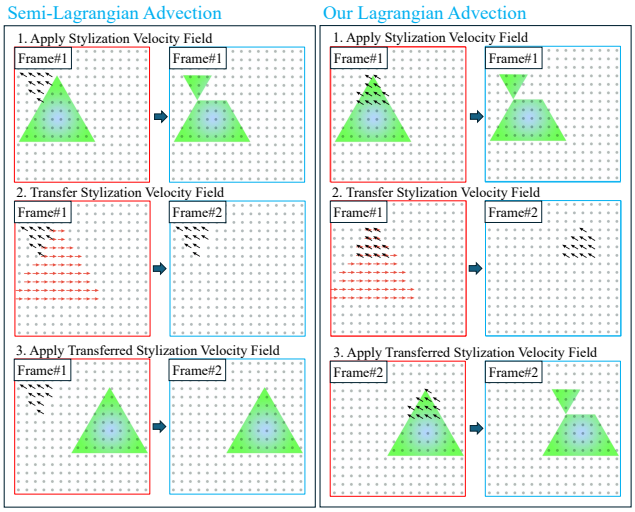

Figure 5. This figure explains the failure of the semi-lagrangian advection in 2D video geometrical stylization. For clarity, color transformations are omitted. The left illustrates how semi-lagrangian advection operates: (1) Stylization velocity values tracking stylized deformation are stored at the transformed positions. (2) Stylization velocities misaligned with scene motion prevent proper advection between frames. (3) This improper propagation causes stylized results to mismatch previous frame's stylization. In contrast, the right part illustrates the success of our lagrangian advection method: (1) Our lagrangian advection stores stylization velocity information at the start position of the deformation. (2) This storage scheme naturally aligns stylization velocity with scene motion, enabling proper advection through optical flow. (3) This consistent advection yields temporally consistent stylization results.

12. Experimental setup details

All videos are sourced from DAVIS 2017/2019, a widely adopted dataset in video stylization literature. To rigorously evaluate temporal consistency, we first prioritized sequences with significant dynamics or camera movement (average optical flow magnitude >5px/frame), which excludes near-static scenes where temporal coherence can be achieved trivially. Then we systematically curated 60 samples ensuring visual quality (no strong motion blur, etc.) and balanced coverage of motion types. For style images, we selected representative examples from the widely used WikiArt, prior works, and the Internet, ensuring coverage of stylistic diversity and distinct brushstroke/texture patterns.

13. Quantitative evaluation in ablation study

We conducted a quantitative ablation study to validate the effectiveness of each key component within our framework in Tab. 3. This evaluation involved systematically disabling individual modules—specifically, the color regularization, the color transfer field, the stylization velocity field, the Lagrangian advection, the momentum-preserving optimization, and the temporal lookup strategy—and measuring the resulting impact on content fidelity, style fidelity, and temporal consistency as measured by warping error.

Table 3. Quantitative Ablation Study

Ablation Configuration	**Content Fidelity**	
	CLIP-C ↑	Gram-C ↓
w/o color regulaization	0.47	22.8
Full model	0.61	14.5
Ablation Configuration	**Style Fidelity ↑**	
	DINO	CLIP
w/o color transfer field	0.59	0.42
w/o stylziation velocity field	0.69	0.64
Full model	0.72	0.67
Ablation Configuration	**Warping Error (10^{-3}) ↓**	
	1-Frame	10-Frame
w/o Lagrangian advection	3.5	20.5
w/o adaptive LR strategy	7.0	23.3
w/o momentum strategy	6.5	35.4
w/o temporal lookup	2.6	19.5
Full model	2.5	14.2

Reconstructed Frame Canonical Frame

Figure 6. This figure reveals CoDeF [34]'s limitations in modeling non-rigid scene dynamics within our evaluation datasets.

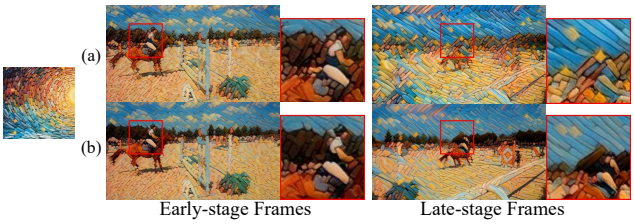

Figure 7. Ablation study on orthogonal color transfer: (a) Naive additive field (Eq. 1) vs (b) Our *orthogonality-regularized* color transfer field (Eq. 2).

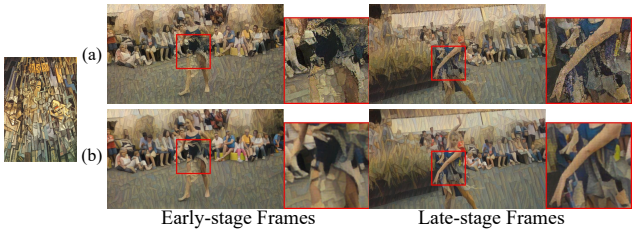

Figure 8. Ablation on Lagrangian Advection: (a) conventional TNST semi-Lagrangian (Eq. 4) vs (b) Our adopted lagrangian advection (Eq. 6).

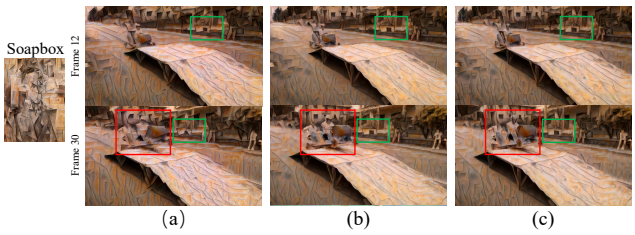

Figure 9. Momentum-Preserving Field Ablation: (a) No strategy vs (b) Low-LR tuning vs (c) Our momentum-aware strategy.

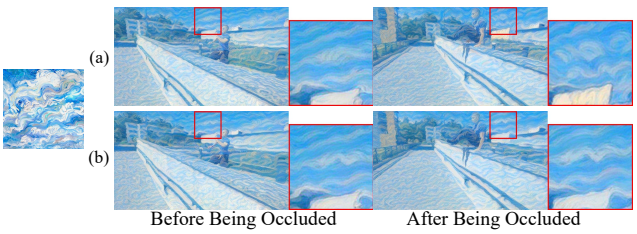

Figure 10. Occlusion-Aware Lookup Ablation: (a) Without lookup vs (b) Our occlusion-aware strategy.