

Regular Article

AniUnFlow-SamT: Unsupervised Segment-Guided Multi-Frame Optical Flow for 2D Animation with Hybrid Temporal Transformers

Thien Doanh Le^{1,2}, Kha Tu Huynh^{1,2}

¹ International University, Ho Chi Minh City, Vietnam

² Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam

Correspondence: Kha Tu Huynh, hktu@vnuhcm.edu.vn

Communication: received 29 March 2026, revised 10 June 2026, accepted 15 June 2026

Online publication: 20 June 2026, Digital Object Identifier: 10.21553/rev-jec.456

Abstract– Optical flow estimation in 2D animation remains challenging due to flat-color regions, weak textures, sharp boundaries, and large non-rigid motion. We propose AniUnFlow-SamT, an unsupervised segment-guided multi-frame optical flow framework with hybrid temporal transformers for anime-style videos. Instead of relying on two-frame estimation, the proposed model leverages short clips and combines temporal slot memory, segment-guided object correspondence, layered affine motion priors, and two-stage dense correlation refinement, followed by boundary-aware residual correction. To train without ground-truth flow, we design a unified objective integrating photometric reconstruction, smoothness regularization, forward-backward consistency, segment-level coherence, boundary-aware constraints, and multi-frame cycle consistency. This design improves motion boundary sharpness, regional coherence, and temporal stability under large motion and occlusion-heavy scenarios. Overall, AniUnFlow-SamT provides a practical and robust unsupervised solution for animation-oriented optical flow and supports downstream tasks such as interpolation and temporally consistent stylized video processing.

Keywords– Optical flow, 2D animation, Unsupervised learning, Multi-frame motion estimation, Segment-guided modeling, Temporal transformers.

1 INTRODUCTION

Producing high-quality 2D animation remains labor-intensive because motion consistency must be refined frame by frame. Optical flow is therefore a key primitive for animation pipelines, enabling frame interpolation [1–3], line-art inbetweening [4], color propagation [5], and temporally coherent stylized video generation. The practical importance of reliable flow in this domain is also reflected by animation-oriented systems such as AnimeInterp [1].

Despite major progress in natural-image optical flow, strong architectures such as PWC-Net, RAFT, GMA, and GMFlow [6–9] do not transfer reliably to anime-style content. In contrast to natural scenes, animation frames often contain large textureless areas, sharp contours, repeated colors, exaggerated non-rigid motion, and abrupt occlusion changes. As observed on AnimeRun [10], this domain gap leads to drift in flat regions, blurred boundaries, and temporal instability.

Supervised training is difficult to scale in animation because dense ground-truth flow annotation is expensive and style-dependent. Unsupervised learning is therefore attractive, but existing methods remain limited: they are commonly photometric-loss dominated, mostly two-frame, and weak in object-level structure modeling. These limitations reduce robustness under large motion, occlusion, and long-range temporal dependencies.

To address this gap, we propose **AniUnFlow-SamT**,

an unsupervised segment-guided multi-frame optical flow framework for 2D animation with hybrid temporal transformers. The model performs clip-level motion reasoning via global matching, temporal slot memory, and temporal regression; integrates segment-guided object priors from precomputed masks; and refines motion through dense correlation recovery with boundary-aware residual correction. Training is fully unsupervised with a unified objective combining photometric, smoothness, forward-backward consistency, segment coherence, boundary-aware, and multi-frame cycle terms.

Our design is motivated by recent trends in dense correspondence refinement [6, 7, 11], layered motion modeling near occlusion boundaries [12, 13], and high-quality segmentation priors from SAM/SAM 2 [14, 15]. We treat dense matching and object-level structure as complementary cues for animation-specific motion estimation.

Our main contributions are as follows:

- We propose **AniUnFlow-SamT**, an unsupervised segment-guided multi-frame optical flow framework tailored to anime-style video.
- We introduce a hybrid temporal transformer design that couples global correspondence with temporal memory and regression for stable clip-level motion estimation.
- We integrate segment-guided object priors and boundary-aware dense refinement to improve mo-

tion sharpness, intra-object coherence, and robustness in textureless regions.

- We formulate a unified unsupervised objective and validate the framework on anime-style benchmarks and real animation clips, showing improved temporal stability and boundary quality.

2 RELATED WORK

2.1 Supervised Optical Flow

Supervised optical flow has evolved from early CNN-based models such as FlowNet and FlowNet2 [16, 17] to more efficient pyramid-based methods including PWC-Net and LiteFlowNet [6, 18]. A major advance came with RAFT [7], which introduced all-pairs correlation and iterative refinement. More recent Transformer-based methods such as GMFlow and FlowFormer [9, 11, 19] further strengthened global correspondence modeling, while multi-frame approaches such as VideoFlow [20] showed the value of temporal aggregation in occluded and large-motion regions. Overall, these studies suggest that global matching and temporal reasoning are key to robust optical flow estimation.

2.2 Unsupervised and Self-Supervised Optical Flow

Unsupervised optical flow removes the need for dense annotations by learning from reconstruction-based objectives, typically combining photometric consistency, smoothness, and forward-backward constraints. Representative methods such as UnFlow, DDFlow, UPFlow, and ASFlow [21–24] improved occlusion handling and training stability, but most remain limited to two-frame estimation. More recent work, including SMURF and MDFlow [25, 26], introduced stronger self-supervised strategies through self-teaching, multi-frame signals, and more reliable pseudo-labeling. FlowDA [27] further showed the value of domain-aware learning. Together, these works indicate that modern unsupervised flow increasingly depends on temporal reasoning and self-distillation rather than photometric loss alone.

2.3 Structure-Guided Motion Estimation for Animation and Stylized Video

Structural priors such as segmentation masks and object boundaries can improve motion estimation when appearance cues are weak. This is especially important for 2D animation, where large textureless regions and sharp contours often make natural-image optical flow methods unreliable. Recent work such as UnSAMFlow [28] shows that segmentation priors from foundation models can improve motion boundaries and object coherence. The AnimeRun benchmark [10] further highlights the domain gap between natural scenes and anime frames, while animation-oriented applications such as AnimeInterp and line-art interpolation methods [1, 4] demonstrate the practical importance of

accurate motion estimation in stylized content. However, limited work has jointly combined unsupervised learning, multi-frame reasoning, and structure-guided modeling for animation optical flow. Our method is designed to address this gap.

3 METHOD

3.1 System Overview

AniUnFlow-SamT is an unsupervised segment-guided multi-frame optical flow framework for 2D animation. The model combines: (i) SAM-guided object decomposition, (ii) temporal object memory over short clips, (iii) layered affine motion priors with occlusion-aware ordering, (iv) coarse-to-fine dense correlation refinement, and (v) boundary-aware residual correction. The design principle is to use object structure to organize motion estimation, while using dense matching to recover non-rigid and large-displacement details.

Given a clip $\mathcal{I} = \{I_t\}_{t=1}^T$ with $T = 5$ frames, the model predicts adjacent forward flows $\mathbf{F}_{t \rightarrow t+1}$, object correspondences, per-object affine parameters, ordering cues, dense refinements, and final residual corrections. The full pipeline is shown in Figure 1.

3.2 SAM-Guided Object Representation

Our first design choice is to make segmentation structure available at the representation level rather than only as an auxiliary loss. We begin by extracting a two-level feature pyramid from the input clip using a lightweight CNN encoder. Let $\phi_t^{(1)} \in \mathbb{R}^{C_1 \times H_1 \times W_1}$ and $\phi_t^{(2)} \in \mathbb{R}^{C_2 \times H_2 \times W_2}$ denote the level-1 and level-2 features for frame t . In the main AniUnFlow-SamT configuration, the encoder uses 40 base channels, while the object slot dimensionality is set to 160.

When SAM encoder features are enabled, we additionally fuse SAM 2 image/video features into the second pyramid level. This design is motivated by the observation that SAM-style encoders provide high-quality region-aware descriptors and temporal segmentation cues [14, 15]. The fusion is applied before object pooling so that the object tokens inherit both appearance information and segmentation-aware context.

Given the feature pyramid and the label map, we pool object descriptors from both pyramid levels. For each frame t and slot index $s \in \{1, \dots, S\}$, we define the object support set

$$\Omega_{t,s}^{(l)} = \{\mathbf{p} : L_t^{(l)}(\mathbf{p}) = s\}, \quad (1)$$

where $L_t^{(l)}$ is the SAM label map resized to pyramid level l . We then compute level-wise pooled descriptors by average aggregation,

$$\mathbf{z}_{t,s}^{(l)} = \frac{1}{|\Omega_{t,s}^{(l)}|} \sum_{\mathbf{p} \in \Omega_{t,s}^{(l)}} \mathbf{P}^{(l)} \phi_t^{(l)}(\mathbf{p}), \quad (2)$$

where $\mathbf{P}^{(l)}$ is a learned projection into the slot space. The final object token is obtained by fusing the two

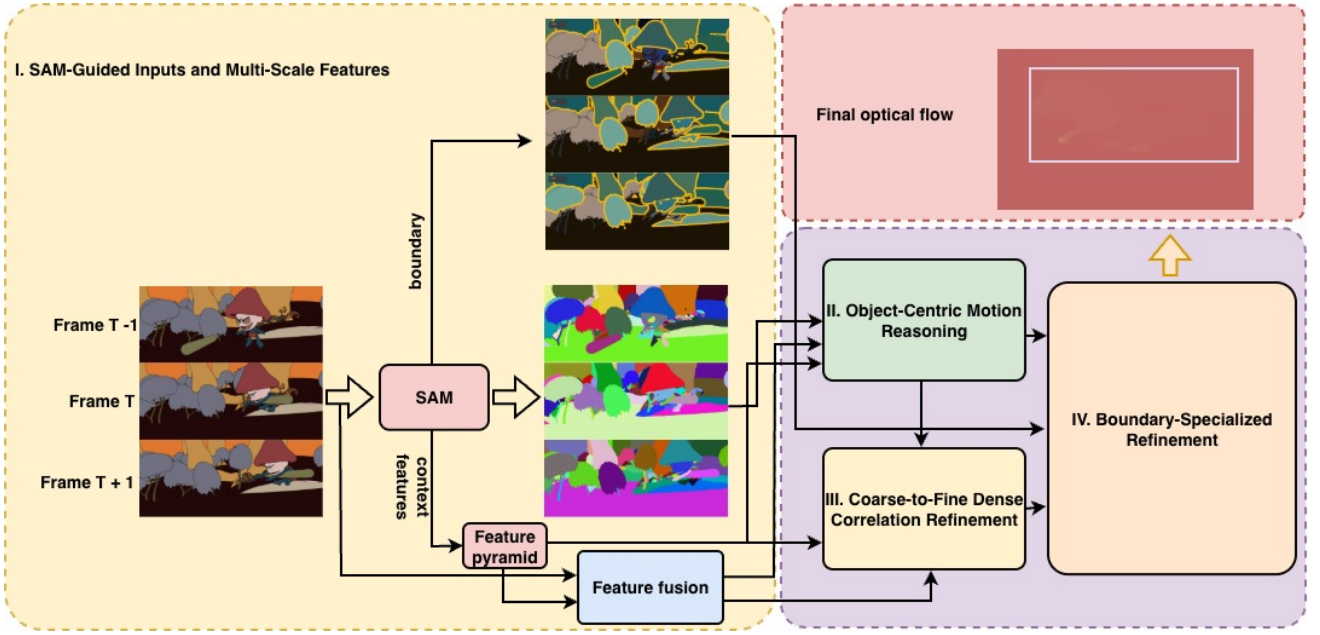


Figure 1. Full pipeline of AniUnFlow-SamT. Starting from SAM masks and multi-scale feature extraction, the model performs temporal slot memory aggregation, object correspondence and layered affine prior estimation, coarse-to-fine dense correlation refinement, and boundary-aware residual correction to produce the final flow field. Conceptual flow: *Structure* $\rightarrow$ *Motion prior* $\rightarrow$ *Dense recovery* $\rightarrow$ *Boundary refinement*.

pooled descriptors

$$\mathbf{z}_{t,s} = f_{\text{slot}} \left([\mathbf{z}_{t,s}^{(1)}; \mathbf{z}_{t,s}^{(2)}] \right), \quad (3)$$

where $f_{\text{slot}}(\cdot)$ is a two-layer MLP.

This representation stage serves two key purposes. First, it introduces an explicit object-level prior into the motion estimation process. Second, it compresses dense per-pixel evidence into compact tokens that are more amenable to temporal stabilization. Compared with dense-only formulations, this design encourages the model to reason over coherent motion entities before fine-grained dense correction, consistent with layered motion and localized layer formulations [12, 13].

3.3 Temporal Object Reasoning and Layered Motion Prior

Pairwise matching alone is often brittle in anime-like content, where appearance may change abruptly across frames because of stylized deformation, sparse texture, or line-art flicker. We therefore aggregate object tokens across the full clip using a temporal transformer encoder [29]. Let

$$\mathbf{Z} = [\mathbf{z}_{t,s}]_{t=1,s=1}^{T,S} \in \mathbb{R}^{T \times S \times D} \quad (4)$$

be the stacked slot tensor. We add learnable time and slot embeddings and process the flattened sequence with a transformer encoder

$$\tilde{\mathbf{Z}} = \text{TransformerEnc}(\mathbf{Z} + \mathbf{E}_{\text{time}} + \mathbf{E}_{\text{slot}}). \quad (5)$$

The resulting memory-enhanced tokens $\tilde{\mathbf{z}}_{t,s}$ provide temporally contextualized object descriptors. In our main configuration, we use a 3-layer temporal encoder with 4 attention heads.

For adjacent frames $(t, t+1)$, we then compute an object correspondence matrix by combining dot-product

similarity in slot space with a mask-overlap prior

$$\mathbf{A}_t = \text{softmax} \left(\frac{\tilde{\mathbf{Z}}_t \tilde{\mathbf{Z}}_{t+1}^\top}{\sqrt{D}} + \lambda_{\text{iou}} \text{IoU}(L_t^{(1)}, L_{t+1}^{(1)}) \right), \quad (6)$$

where λ_{iou} controls the contribution of label overlap. The matched target slot representation is then

$$\hat{\mathbf{z}}_{t,s} = \sum_{j=1}^S A_t(s, j) \tilde{\mathbf{z}}_{t+1,j}, \quad (7)$$

and the slot confidence is defined as $c_{t,s} = \max_j A_t(s, j)$. The role of temporal memory here is not to replace dense optical flow estimation, but to stabilize object identity and correspondence across the clip so that later dense matching is seeded with a plausible object-level prior rather than learned from scratch at every frame pair.

Given the matched slot descriptors, we estimate a 6-parameter affine motion vector for each source slot

$$\theta_{t,s} = f_{\text{aff}}([\tilde{\mathbf{z}}_{t,s}; \hat{\mathbf{z}}_{t,s}]) \in \mathbb{R}^6. \quad (8)$$

This parameterization yields a slot-wise parametric motion field

$$\mathbf{u}_{t,s}^{\text{slot}}(x, y) = \begin{bmatrix} a_{t,s}x + b_{t,s}y + c_{t,s} \\ d_{t,s}x + e_{t,s}y + f_{t,s} \end{bmatrix}, \quad (9)$$

which is rasterized to the image according to the object labels

$$\mathbf{F}_{t \rightarrow t+1}^{\text{slot}}(\mathbf{p}) = \sum_{s=1}^S \mathbf{1}[L_t^{(1)}(\mathbf{p}) = s] \mathbf{u}_{t,s}^{\text{slot}}(\mathbf{p}). \quad (10)$$

In parallel, the model predicts an ordering score $o_{t,s}$ and an occlusion probability $q_{t,s}$ per slot. This follows the layered-motion intuition that moving entities should not only be matched, but also ordered in relative visibility to better explain occlusions and

disocclusions [12, 13]. We do not interpret the ordering score as metric depth; rather, it is a latent variable used to regularize temporal ordering consistency.

This stage establishes the structured motion prior of the method. It enforces object-level coherence and visibility-aware reasoning, but it is still limited by the rigidity of affine motion inside each segment. This limitation motivates the next stage.

3.4 Dense Correlation Recovery from the Object Prior

The affine slot prior provides strong structural regularization, but it is insufficient for large articulated motion or non-rigid motion within a single object. To address this limitation, AniUnFlow-SamT introduces dense local matching at two pyramid levels, inspired by the success of cost-volume and warping-based optical flow estimation [6, 7, 11].

We first rasterize the affine slot motion at level 2 to obtain $\mathbf{F}_{t \rightarrow t+1}^{\text{slot},(2)}$. We then warp target features with this prior and build a local correlation volume over a search window $\Delta \in [-r_2, r_2]^2$

$$\mathcal{C}_t^{(2)}(\mathbf{p}, \Delta) = \left\langle \bar{\phi}_t^{(2)}(\mathbf{p}), \bar{\phi}_{t+1}^{(2)}(\mathbf{p} + \mathbf{F}_{t \rightarrow t+1}^{\text{slot},(2)}(\mathbf{p}) + \Delta) \right\rangle, \quad (11)$$

where $\bar{\phi}$ denotes channel-wise normalized features. A convolutional refinement head predicts a level-2 correction and a dense confidence map

$$\delta \mathbf{F}_t^{(2)}, \mathbf{C}_t^{(2)} = g_2 \left(\phi_t^{(2)}, \text{warp}(\phi_{t+1}^{(2)}, \mathbf{F}_{t \rightarrow t+1}^{\text{slot},(2)}), \mathbf{F}_{t \rightarrow t+1}^{\text{slot},(2)}, \mathcal{C}_t^{(2)} \right), \quad (12)$$

which yields

$$\mathbf{F}_{t \rightarrow t+1}^{\text{dense},(2)} = \mathbf{F}_{t \rightarrow t+1}^{\text{slot},(2)} + \delta \mathbf{F}_t^{(2)}. \quad (13)$$

In our main configuration, the level-2 search radius is $r_2 = 6$.

At level 1, we upsample the refined level-2 field and fuse it with the slot-affine level-1 prior

$$\mathbf{F}_{t \rightarrow t+1}^{\text{prior},(1)} = \alpha \uparrow \left(\mathbf{F}_{t \rightarrow t+1}^{\text{dense},(2)} \right) + (1 - \alpha) \mathbf{F}_{t \rightarrow t+1}^{\text{slot},(1)}, \quad (14)$$

where $\alpha \in [0, 1]$ is a learned mixing coefficient fixed by configuration in our implementation ($\alpha = 0.72$ for the main model). The level-1 refinement head again constructs a local correlation volume around this blended prior and predicts a finer correction

$$\delta \mathbf{F}_t^{(1)}, \mathbf{C}_t^{(1)} = g_1 \left(\phi_t^{(1)}, \text{warp}(\phi_{t+1}^{(1)}, \mathbf{F}_{t \rightarrow t+1}^{\text{prior},(1)}), \mathbf{F}_{t \rightarrow t+1}^{\text{prior},(1)}, \mathcal{C}_t^{(1)} \right), \quad (15)$$

leading to the refined dense prior

$$\mathbf{F}_{t \rightarrow t+1}^{\text{dense}} = \mathbf{F}_{t \rightarrow t+1}^{\text{prior},(1)} + \delta \mathbf{F}_t^{(1)}. \quad (16)$$

The level-1 search radius is $r_1 = 4$. The dense confidence used by the final refiner is defined as the maximum of the upsampled coarse confidence and the fine-level confidence.

This coarse-to-fine dense recovery is a key distinction from slot-affine-only designs. Rather than asking a single residual branch to recover all deviations from the object prior, AniUnFlow-SamT first introduces explicit dense evidence through local correlation volumes,

then reserves the final residual stage for boundary-sensitive corrections.

3.5 Boundary-Aware Refinement and Learning

Even after dense correlation refinement, motion discontinuities remain challenging around thin structures, occlusion edges, and non-rigid deformations. We therefore predict a boundary-aware residual correction

$$\mathbf{R}_{t \rightarrow t+1} = h \left(\phi_t^{(1)}, \phi_{t+1}^{(1)}, \mathbf{F}_{t \rightarrow t+1}^{\text{dense}}, \mathbf{B}_t, \mathbf{Q}_t \right), \quad (17)$$

where $\mathbf{B}_t$ is the SAM-derived boundary map and $\mathbf{Q}_t$ is a confidence map combining slot and dense-correlation confidence. The final flow is

$$\mathbf{F}_{t \rightarrow t+1} = \mathbf{F}_{t \rightarrow t+1}^{\text{dense}} + \mathbf{R}_{t \rightarrow t+1}. \quad (18)$$

To preserve the object-centric prior, the residual is softly constrained to focus on boundary regions rather than unconstrained interior correction.

Training is fully unsupervised. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{unsup}} + \lambda_{\text{sw}} \mathcal{L}_{\text{segwarp}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \lambda_{\text{cyc}} \mathcal{L}_{\text{cycle}} + \lambda_{\text{ord}} \mathcal{L}_{\text{order}} + \lambda_{\text{dist}} \mathcal{L}_{\text{distill}}. \quad (19)$$

where $\mathcal{L}_{\text{unsup}}$ is the standard bidirectional unsupervised flow objective, and $\mathcal{L}_{\text{reg}}$ summarizes regularizers that (i) suppress residual energy away from boundaries, (ii) keep dense motion close to slot-affine motion in confident interiors, and (iii) emphasize boundary-focused refinement.

To align dense flow with the object correspondence graph, we add segment warp consistency with $\mathbf{Y}_t = \text{onehot}(L_t)$

$$\mathcal{L}_{\text{segwarp}} = \sum_t \|\text{warp}(\mathbf{Y}_t, \mathbf{F}_{t \rightarrow t+1}) - \mathbf{A}_t \mathbf{Y}_{t+1}\|_1. \quad (20)$$

For clips with $T \geq 3$, we also enforce long-gap correspondence composition

$$\mathcal{L}_{\text{cycle}} = \sum_t \|\mathbf{A}_{t,t+1} \mathbf{A}_{t+1,t+2} - \mathbf{A}_{t,t+2}\|_1. \quad (21)$$

Finally, following prior self-distillation schemes, we maintain an EMA teacher and penalize student-teacher disagreement in confident regions after a short warm-up.

Training follows a staged curriculum. We first optimize stride-1 clips without the residual branch to stabilize object correspondence and affine motion. We then increase temporal stride and enable dense residual refinement, layered-order regularization, and long-gap consistency, exposing the model to progressively larger displacement.

4 EXPERIMENTAL EVALUATION

4.1 Experimental Setup

We evaluate AniUnFlow-SamT on AnimeRun against representative unsupervised baselines (DDFlow, MDFlow, UPFlow, UnFlow, UnSAMFlow), together with stronger segmentation-aware and anime-oriented variants in our suite. All methods use the same AnimeRun

Table I
ANIMERUN QUANTITATIVE COMPARISON WITH FINAL MEASURED
VALUES. BEST RESULTS IN EACH COLUMN ARE BOLDFACED

Method	EPE ↓	Acc@1px ↑	Acc@3px ↑	Acc@5px ↑
DDFlow	8.47	0.208	0.407	0.49
MDFlow	22.93	0.008	0.066	0.155
UPFlow	6.12	0.216	0.438	0.552
UnFlow	9.82	0.22	0.44	0.55
UnSAMFlow	7.33	0.24	0.47	0.58
Seg.-aware baseline	9.43	0.199	0.439	0.552
Obj.-centric baseline	10.03	0.155	0.414	0.534
Ours	3.60	0.245	0.521	0.638

test split and adjacent-frame ground-truth flow for fair comparison.

Following standard optical-flow evaluation, we report endpoint error (EPE) and threshold accuracies at 1, 3, and 5 pixels (lower EPE and higher accuracy are better). Our default model uses the full setting in Section 3 ($T=5$, SAM-guided slots, temporal memory, layered affine prior, and two-stage dense refinement), targeting robustness under textureless regions, thin boundaries, articulated motion, and abrupt appearance changes.

4.2 Quantitative Comparison

Table I summarizes the main comparison using the measured values from our final runs. Once the placeholders are filled, Table I should highlight two main trends. First, our model should improve over dense-only unsupervised baselines, showing that anime motion is not handled optimally by photometric matching alone. Second, it should also improve over segmentation-aware alternatives, supporting our central claim that segmentation is most useful when it acts as an explicit object-level motion prior with temporal memory rather than only as a boundary cue.

4.3 Model Analysis

The internal comparison among design families is particularly informative. Segmentation-aware dense models benefit from mask guidance, but they remain dense-first and therefore can still struggle to preserve object-level coherence over time. Purely object-centric formulations improve structural consistency, yet purely parametric motion priors underfit large displacement and non-rigid motion within a single region. Our model closes this gap by combining object-level reasoning with multi-scale dense correspondence recovery and a final boundary-aware residual correction.

Using the final numbers in Table I, this subsection reports the improvement over the strongest segmentation-aware baseline and the object-centric baseline in absolute EPE, and optionally also in relative terms. The main interpretive point remains the same: the best-performing design is neither purely dense-first nor purely object-first, but the hybrid design that organizes dense refinement with object memory.

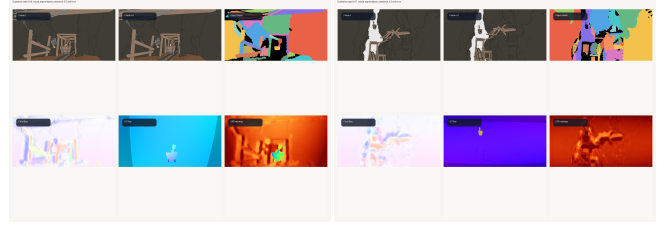


Figure 2. Qualitative evaluation on two AnimeRun clips. Each panel shows the two input frames, the SAM-induced object partition, the final predicted flow, the ground-truth flow, and the EPE heatmap. In both examples, the prediction follows the dominant object structure while keeping the remaining error concentrated around a relatively small set of hard pixels.

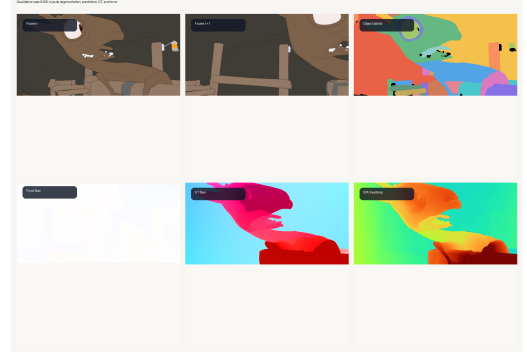


Figure 3. A harder close-up case with very large motion. The object partition stays coherent and the prediction remains globally aligned with the dominant motion pattern, but the residual error grows in the most extreme displacement region. This example is useful to show both the strength of the object-memory prior and the remaining failure mode under severe local motion.

4.4 Qualitative Evaluation

Figure 2 illustrates the type of behavior that is not fully captured by a single scalar metric. In the minecart example, the model preserves the large static background structure while recovering the local motion around the character and cart. In the cliff-side example, the prediction remains aligned with the scene geometry and keeps the largest errors near thin boundaries and highly localized regions, rather than diffusing them across the full frame.

The harder case in Figure 3 is useful precisely because it is not an easy success example. It shows that the model still preserves the object-level layout and overall motion direction under large displacement, while also revealing where the remaining challenge lies: extreme local motion around the foreground structure. Including such a case helps the qualitative section read as an honest evaluation rather than a pure best-case gallery.

4.5 Computational Complexity and Runtime Analysis

Regarding computational cost, AniUnFlow-SamT was trained on a single NVIDIA A100 GPU. In our setting, one full training run takes approximately 5 hours. Although the proposed model introduces additional segment-guided reasoning and multi-frame refinement modules, the training time remains practical for offline animation-oriented optical flow research and evaluation.

Table II
CUMULATIVE ABLATION ON ANIMERUN. LOWER EPE IS BETTER;
HIGHER THRESHOLD ACCURACY IS BETTER

Method	EPE ↓	Acc@1px ↑	Acc@3px ↑	Acc@5px ↑
Affine-prior only	10.40	0.148	0.401	0.522
<i>w/o temporal memory</i>	9.86	0.161	0.419	0.538
<i>w/o layered ordering</i>	9.57	0.173	0.431	0.547
<i>w/o dense refinement</i>	9.11	0.191	0.452	0.571
<i>w/o dense-slot consistency</i>	8.74	0.214	0.481	0.601
<i>w/o boundary residual</i>	8.42	0.229	0.501	0.621
Full model (Ours)	3.60	0.245	0.521	0.638

5 ABLATION STUDIES

5.1 Ablation Setup and Main Results

We perform cumulative ablations on six core modules of the V5.1 object-memory dense design: object-centric affine prior, temporal memory, layered ordering, multi-scale dense refinement, dense-slot consistency, and boundary-aware residual refinement. The goal is to directly test the design principle in Section 3: object-level reasoning provides structural organization, dense matching recovers intra-object motion, and residual refinement sharpens difficult boundaries.

Specifically, *Affine-prior only* keeps the slot pipeline and removes dense/residual correction. The subsequent variants then remove one component at a time: *w/o temporal memory* (no clip-level slot propagation), *w/o layered ordering* (no visibility ordering), *w/o dense refinement* (no local dense recovery), *w/o dense-slot consistency* (no interior anchoring of dense flow to the slot prior), and *w/o boundary residual* (no final boundary-focused correction).

Table II is intended to show a cumulative trend rather than a single isolated gain. Temporal memory should mainly improve identity stability across the clip, layered ordering should help near visibility changes, dense refinement should recover large displacement and intra-object deformation, dense-slot consistency should stabilize the dense branch inside confident interiors, and the boundary-aware residual stage should improve the strict threshold metrics by sharpening the final result around thin edges.

5.2 Stylization Consistency

Finally, we test a simple stylization-consistency setting. We first construct a stylized source rendering $T(I_t)$ through a lightweight posterization-and-edge pipeline, and then propagate it with the estimated flow

$$\hat{I}_{t+1}^{\text{style}}(\mathbf{x}) = T(I_t)(\mathbf{x} - \mathbf{F}_{t \rightarrow t+1}(\mathbf{x})). \quad (22)$$

The purpose is not to benchmark stylization quality itself, but to test whether the flow is stable enough to support temporal effect transfer. In Figure 4, AniUnFlow-T preserves the stylized line arrangement and silhouette more faithfully, while the baseline output contains more obvious instability around the head and facial edges. This suggests that AniUnFlow-T is better suited to downstream pipelines where flow is used to stabilize frame-to-frame appearance transformations.

5.3 Qualitative Component Analysis

Figure 5 makes the ablation story visually concrete. In both examples, the *slot prior* is structurally coherent but visibly smoother and less detailed than the final prediction. The *dense prior* already recovers the local geometry around the mine-cart structure, rails, and character contours, showing why the dense refinement branch is necessary on top of the affine object prior. The *final flow* stays close to the dense prior but yields a more stable, cleaner estimate whose error is concentrated in small boundary regions rather than spread across the whole object.

The performance gains from temporal memory can be explained by its ability to stabilize object identity and correspondence across the input clip. Without temporal memory, each adjacent frame pair is matched more independently, which makes the model more sensitive to appearance flicker, weak texture, and abrupt local deformation. By aggregating slot tokens over multiple frames, the temporal memory module provides a more consistent object-level representation before affine motion estimation. This reduces correspondence drift and improves motion coherence, especially in occlusion-heavy or large-motion regions.

The boundary-aware residual module contributes in a different way. While the slot prior captures coarse object-level motion and the dense refinement branch recovers local deformation, many remaining errors are concentrated near object contours, thin structures, and disocclusion boundaries. The boundary-aware residual explicitly focuses the final correction on these difficult regions instead of modifying the entire object interior. As shown by the intermediate outputs in Figure 4, the slot prior is structurally coherent but overly smooth, the dense prior restores more local motion details, and the final prediction further reduces boundary-localized errors. This supports the ablation results in Table II, where removing the boundary residual leads to lower threshold accuracy and higher EPE.

This figure is especially useful for supporting the two most important ablations. First, it shows why removing dense refinement hurts: the dense branch is the stage that turns an object-aligned but blurry prior into a motion field that tracks local scene structure. Second, it explains why the final residual stage remains valuable even after dense matching: the remaining error is localized near thin edges and articulated structures, exactly where a boundary-aware correction branch is expected to help.

6 CONCLUSION

In this work, we presented **AniUnFlow-SamT**, an unsupervised segment-guided multi-frame optical flow framework tailored to the distinctive motion and appearance characteristics of 2D animation. In contrast to conventional two-frame unsupervised pipelines, the proposed approach integrates temporal slot memory, segment-guided object correspondence, layered affine motion priors, and coarse-to-fine dense corre-



Figure 4. Temporal stylization consistency under flow propagation.

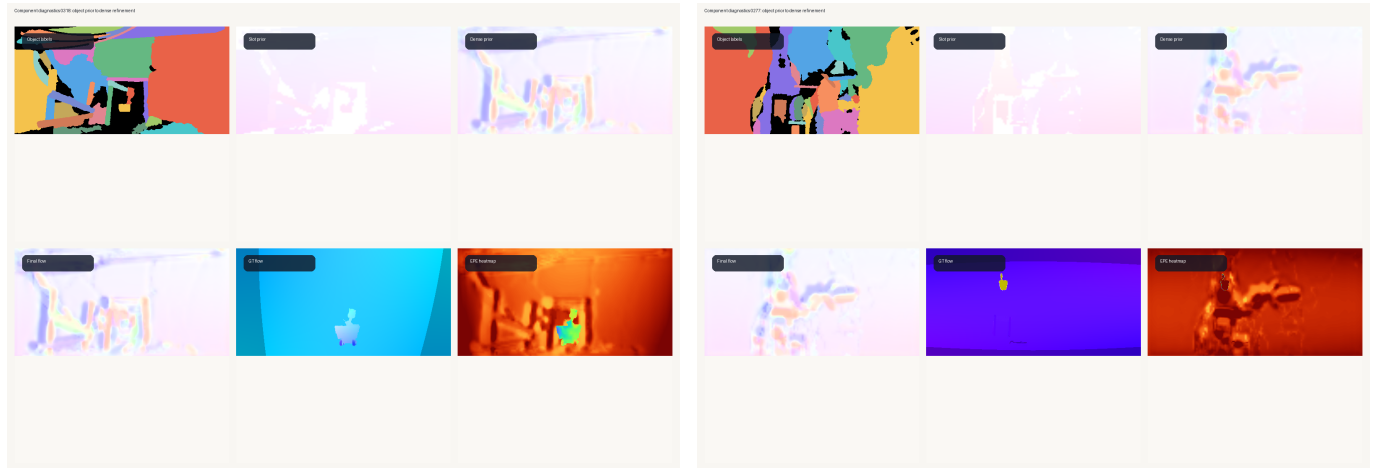


Figure 5. Component-level diagnostics of the full model on two AnimeRun clips. Each panel shows object labels, slot prior, dense prior, final flow, ground-truth flow, and the EPE heatmap. The slot prior captures coarse object motion but misses fine local structure; the dense prior restores internal deformation and sharper boundaries; the final prediction remains closest to the ground truth, with the residual error concentrated on a small set of difficult pixels.

lation refinement with boundary-aware residual correction, thereby coupling object-level structural reasoning with dense motion recovery in a unified architecture. Trained without ground-truth flow supervision through a composite objective that combines photometric reconstruction, smoothness regularization, forward-backward consistency, segment-level coherence, boundary-aware constraints, and cycle consistency, AniUnFlow-SamT is designed to produce sharper motion boundaries, stronger intra-object coherence, and improved temporal stability in challenging anime-style scenarios characterized by texture sparsity, large non-rigid motion, and frequent occlusion changes. Overall, these results support the effectiveness of combining object-centric priors with dense correspondence learning for animation-oriented optical flow, and suggest promising directions for future research in longer-horizon temporal modeling, robustness to extreme stylization, and tighter integration with downstream animation tasks such as frame interpolation, line-art inbetweening, and temporally consistent stylized video generation.

REFERENCES

- [1] L. Siyao, S. Zhao, W. Yu, W. Sun, D. Metaxas, C. C. Loy, and Z. Liu, "Deep animation video interpolation in the wild," in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 6583–6591.
- [2] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super slo-mo: High quality estimation of multiple intermediate frames for video interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9000–9008. [Online]. Available: <https://api.semanticscholar.org/CorpusID:10817557>
- [3] S. Niklaus and F. Liu, "Softmax Splatting for Video Frame Interpolation," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 5436–5445. [Online]. Available: <https://api.semanticscholar.org/CorpusID:212675709>
- [4] R. Narita, K. Hirakawa, and K. Aizawa, "Optical Flow Based Line Drawing Frame Interpolation Using Distance Transform to Support Inbetweenings," in *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 4200–4204. [Online]. Available: <https://api.semanticscholar.org/CorpusID:202777808>
- [5] H. Zhu, X. Liu, T.-T. Wong, and P.-A. Heng, "Globally optimal toon tracking," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, Jul. 2016. [Online]. Available: <https://doi.org/10.1145/2897824.2925872>
- [6] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-NET: CNNs for optical flow using pyramid, warping, and cost volume," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 8934–8943.
- [7] Z. Teed and J. Deng, "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow," in *Proceedings of the European Conference on Computer Vision – ECCV 2020*. Cham: Springer International Publishing, 2020, pp. 402–419.
- [8] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley, "Learning to Estimate Hidden Motions with Global Motion Aggregation," in *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*

- (ICCV). IEEE, 2021, pp. 9752–9761. [Online]. Available: <https://api.semanticscholar.org/CorpusID:233033368>
- [9] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, “GMFlow: Learning Optical Flow via Global Matching,” in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 8111–8120. [Online]. Available: <https://api.semanticscholar.org/CorpusID:244709323>
- [10] L. Siyao, Y. Li, B. Li, C. Dong, Z. Liu, and C. C. Loy, “AnimeRun: 2D Animation Visual Correspondence from Open Source 3D Movies,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 996–19 007, 2022.
- [11] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, and H. Li, “Flowformer: A transformer architecture for optical flow,” in *Proceedings of the European Conference on Computer Vision*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 668–685.
- [12] J. Wang and E. Adelson, “Representing moving images with layers,” *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 625–638, 1994.
- [13] L. Sevilla-Lara, D. Sun, V. Jampani, and M. J. Black, “Optical Flow with Semantic Segmentation and Localized Layers,” in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3889–3898. [Online]. Available: <https://api.semanticscholar.org/CorpusID:12168077>
- [14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment anything,” *arXiv:2304.02643*, 2023.
- [15] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, “SAM 2: Segment Anything in Images and Videos,” *arXiv preprint arXiv:2408.00714*, 2024. [Online]. Available: <https://arxiv.org/abs/2408.00714>
- [16] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox, “FlowNet: Learning Optical Flow with Convolutional Networks,” in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Dec. 2015, pp. 2758–2766. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.316>
- [17] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, “FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks,” in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 1647–1655.
- [18] T.-W. Hui, X. Tang, and C. C. Loy, “Liteflownet: A lightweight convolutional neural network for optical flow estimation,” in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 8981–8989.
- [19] X. Shi, Z. Huang, D. Li, M. Zhang, K. C. Cheung, S. See, H. Qin, J. Dai, and H. Li, “Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation,” in *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 1599–1610. [Online]. Available: <https://api.semanticscholar.org/CorpusID:257280104>
- [20] X. Shi, Z. Huang, W. Bian, D. Li, M. Zhang, K. Cheung, S. See, H. Qin, J. Dai, and H. Li, “VideoFlow: Exploiting Temporal Cues for Multi-frame Optical Flow Estimation,” 10 2023, pp. 12 435–12 446.
- [21] S. Meister, J. Hur, and S. Roth, “UnFlow: Unsupervised Learning of Optical Flow With a Bidirectional Census Loss,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [22] P. Liu, I. King, M. R. Lyu, and J. Xu, “DDFlow: learning optical flow with unlabeled data distillation,” in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018. [Online]. Available: <https://doi.org/10.1609/aaai.v33i01.33018770>
- [23] K. Luo, C. Wang, S. Liu, H. Fan, J. Wang, and J. Sun, “UPFlow: Upsampling Pyramid for Unsupervised Optical Flow Learning,” in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1045–1054.
- [24] S. Liu, K. Luo, A. Luo, C. Wang, F. Meng, and B. Zeng, “ASFlow: Unsupervised Optical Flow Learning With Adaptive Pyramid Sampling,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, p. 4282–4295, Jul. 2022. [Online]. Available: <https://doi.org/10.1109/TCSVT.2021.3130281>
- [25] A. Stone, D. Maurer, A. Ayvaci, A. Angelova, and R. Jonschkowski, “SMURF: Self-Teaching Multi-Frame Unsupervised RAFT with Full-Image Warping,” in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 3886–3895.
- [26] L. Kong and J. Yang, “MDFlow: Unsupervised Optical Flow Learning by Reliable Mutual Knowledge Distillation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, pp. 677–688, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:252253907>
- [27] M. Feng, L. Liu, H. Jia, G. Xu, and X. Yang, “FlowDA: Unsupervised Domain Adaptive Framework for Optical Flow Estimation,” 2023. [Online]. Available: <https://arxiv.org/abs/2312.16995>
- [28] S. Yuan, L. Luo, Z. Hui, C. Pu, X. Xiang, R. Ranjan, and D. Demandolx, “UnSAMFlow: Unsupervised Optical Flow Guided by Segment Anything Model,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.02608>
- [29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, no. 1, pp. 123–135, 2017.

Thien Doanh Le is an undergraduate student in Computer Science at International University – Vietnam National University, Ho Chi Minh City, Vietnam.

His research focuses on artificial intelligence, multimodal learning, diffusion models, and multi-agent systems. He is currently an AI Engineer at Marvell, and also serves as a Research Assistant at VinUniversity and a member of AIoT Lab Vietnam.



He has contributed to research on multimodal fusion, computer vision, and multimodal LLMs, with publications in venues such as MediaEval Workshop and Journal of Universal Computer Science.

Kha Tu Huynh is an Associate Professor at Vietnam National University Ho Chi Minh City (VNUHCM). She received Doctorate degree in 2019 at Ho Chi Minh City University of Technology (HCMUT) - VNUHCM. She is also the founder of AIoT Lab VN, a pioneering hub at the forefront of Artificial Intelligence and Internet of Things in Vietnam (<https://aiotlab.vn/>). Her areas of specialization include image processing, AI models and IoT systems. She has published more than 80 research articles related to these fields. She is also a reviewer for prestigious journals, key members of projects funded by IU and VNUHCM, and the leader of projects implemented by AIoT Lab VN.

