A^{3}FLOW: ATTENTION RESIDUALS, ADAPTIVE UNCERTAINTY GATING, AND MOTION AGGREGATION FOR UNSUPERVISED OPTICAL FLOW ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the dominance of RAFT-style recurrent correlation models, unsupervised optical flow methods with uncertainty estimation often inherit three restrictive design choices: fixed feed-forward residual encoders lacking data-dependent routing across scales, a hand-crafted stop-gradient gate that suppresses rather than corrects updates when uncertainty is miscalibrated, and local correlation lookup that struggles to provide reliable correspondence cues in occluded regions. We present A^{3}Flow, which augments an uncertainty-aware recurrent unsupervised framework with three pathways: (i) AttnResEncoder, which replaces fixed residual fusion between encoder blocks with learned softmax attention over pooled preceding-block representations, drawing inspiration from depth-wise attention in large language models to enable cross-scale routing in CNN encoders; (ii) ResidualFiLM(Residual Feature-wise Linear Modulation), which generalizes the fixed one-sided sigmoid gate to a two-sided bounded affine modulation driven by predicted uncertainty, enabling richer uncertainty-conditioned correction rather than one-sided suppression of flow updates; and (iii) CP-GMA, a Content-and-Position Global Motion Aggregation module integrated into the motion encoder, which propagates motion features from confident to occluded pixels, providing global motion context beyond local correlation lookup. Extensive experiments on MPI-Sintel and KITTI demonstrate that our method achieves new SOTA among unsupervised methods, with pronounced gains in occluded regions while preserving highly reliable uncertainty estimates. Code will be made available upon acceptance.

1 INTRODUCTION

Optical flow estimation is a fundamental task in computer vision, serving as a critical building block for a wide range of downstream applications, including video interpolation (Jiang et al., 2018), 3D reconstruction (Zhang et al., 2025), and autonomous driving (Sun et al., 2025). However, estimating reliable correspondences across frames is challenging under large displacements, illumination changes, motion blur, and weak texture (Alfarano et al., 2024). In recent years, a variety of RAFT-style algorithms (Dong & Fu, 2024; Huang et al., 2022; Luo et al., 2024; Shi et al., 2023a;b; Sun et al., 2022; Teed & Deng, 2020) have been proposed to address the aforementioned challenges, achieving state-of-the-art performance in supervised optical flow estimation. Nevertheless, these models rely on large-scale, pixel-accurate ground-truth flow annotations (Huang et al., 2023; Sun et al., 2021; Yuan et al., 2022), which are prohibitively expensive and labor-intensive to acquire.

Consequently, unsupervised optical flow estimation has attracted considerable attention, as it eliminates the dependency on costly pixel-accurate ground-truth annotations (Ren et al., 2017). Unfortunately, its reliance on brightness constancy and spatial smoothness assumptions leads to unreliable self-supervision signals under occlusions, textureless regions, and large motion displacements, resulting in significant performance gaps compared to supervised counterparts (Liu et al., 2020; Yuan et al., 2024). Moreover, existing unsupervised methods output only point estimates without modeling prediction uncertainty, leaving them unable to identify unreliable regions or leverage uncertainty to guide optimization (Sun et al., 2026). To address these limitations, U^{2}FLOW (Sun et al., 2026)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

proposes the first uncertainty-aware recurrent unsupervised optical flow framework, featuring a decoupled uncertainty learning strategy based on augmentation consistency. It achieves state-of-the-art performance on KITTI and Sintel benchmarks, with uncertainty maps highly aligned with actual errors.

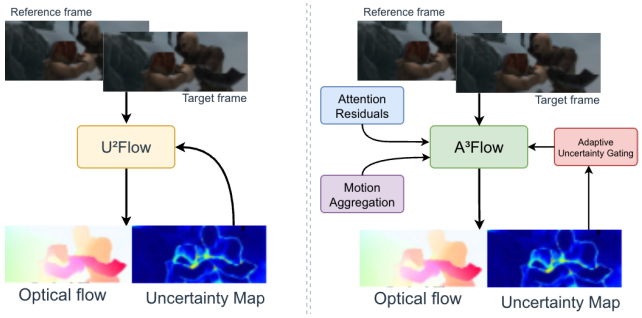


Figure 1: Comparison between U^{2}Flow and our A^{3}Flow. (Left) U^{2}Flow uses uncertainty-based sigmoid gating for flow refinement. (Right) A^{3}Flow adds adaptive encoder fusion, bounded uncertainty-conditioned modulation, and content-and-position motion aggregation.

Despite its strong performance, we identify three rigid design choices inherited from RAFT-style pipelines that limit accuracy, particularly in occluded and textureless regions. First, as described in AttentionResiduals (Kimi Team et al., 2026), fixed residual fusion in encoders freezes scale contributions at construction time, preventing adaptive retrieval of informative scales for scale-dependent optical flow. Second, the one-sided, non-adaptive sigmoid reliability gate can only suppress features (Sun et al., 2026), applying incorrect attenuation when uncertainty is miscalibrated in occluded regions. Third, local-only 4D correlation yields noise at occluded pixels (Jiang et al., 2021), with no mechanism to import motion evidence from confident regions.

To address these limitations, we propose A^{3}Flow (Fig. 1), which augments the U^{2}Flow framework with **A**ttention Residuals, **A**daptive Uncertainty Gating, and Motion **A**ggregation specifically designed to overcome the aforementioned issues. First, AttnResEncoder (Attention Residuals for Feature Encoder) adapts Attention Residuals (Kimi Team et al., 2026) to convolutional encoders by deriving input-dependent weights from globally pooled stage representations, enabling adaptive cross-scale fusion while preserving within-stage residual connections. Second, ResidualFiLM (Residual Feature-wise Linear Modulation) builds upon feature-wise affine modulation (Perez et al., 2018) to learn bounded multiplicative and additive adjustments around the original reliability gate, conditioned on detached uncertainty. The final modulation layer is zero-initialized to recover the baseline gate at initialization. Third, we integrate CP-GMA (Content-and-Position Global Motion Aggregation)(Jiang et al., 2021) into the recurrent update, which employs context-derived attention with relative positional information to propagate motion features to occluded pixels, thereby providing contextual evidence where direct matching is unreliable.

Our contributions are summarized as follows:

- We propose AttnResEncoder, which adapts Attention Residuals (Kimi Team et al., 2026) to convolutional feature and context encoders by replacing fixed residual fusion with softmax attention over globally pooled stage representations, enabling input-dependent cross-scale routing while preserving within-stage residual connections.
- We introduce ResidualFiLM, which generalizes the fixed one-sided sigmoid gate to a two-sided bounded affine modulation for correcting miscalibrated reliability estimates.
- We present A^{3}Flow, an uncertainty-aware unsupervised optical flow framework that integrates AttnResEncoder, ResidualFiLM, and CP-GMA for improved feature representation, gate calibration, and occlusion-aware propagation. Extensive experiments show that A^{3}Flow achieves state-of-the-art performance among two-frame unsupervised methods on MPI-Sintel and KITTI, outperforms representative multi-frame methods, and generalizes effectively across datasets without fine-tuning.

2 RELATED WORK

2.1 UNSUPERVISED OPTICAL FLOW

Unsupervised optical flow replaces ground-truth supervision with photometric and smoothness objectives on unlabeled data (Alfarano et al., 2024). Nevertheless, the photometric signal is frequently degraded by occlusions, textureless areas, motion blur, or illumination variations (Dong & Fu, 2024; Huang et al., 2022; Sun et al., 2022; Teed & Deng, 2020). Subsequent approaches have addressed these challenges through various strategies: ARFlow (Liu et al., 2020) and UFlow (Jonschkowski et al., 2020) mitigate occlusions via augmentation consistency and forward-backward checking, while SelFlow (Liu et al., 2019) recovers flow in textureless regions using confidence-weighted distillation. Furthermore, UPFlow (Luo et al., 2021) handles motion blur with pyramid refinement, and SMURF (Stone et al., 2021) tackles illumination variations via robust Student-t losses. More recent efforts incorporate external priors, such as SemARFlow (Yuan et al., 2023), M2Flow (Sun et al., 2025), or UnSAMFlow (Yuan et al., 2024), to regularize motion boundaries. Despite this progress, handling occlusions remains problematic due to unreliable correlation at boundaries.

2.2 UNCERTAINTY ESTIMATION IN OPTICAL FLOW

Uncertainty estimation in optical flow enables self-diagnostics without ground truth, provides training signals, and can be fed back into optimization (Abdein et al., 2025). In-network probabilistic approaches, such as multi-hypothesis networks and PDC-Net+ (Truong et al., 2023), capture multimodal motion caused by occlusions and ambiguities. Active learning (Yuan et al., 2022) and confidence-based reweighting exploit uncertainty as a training signal, while Smoothness (Abdein et al., 2025) use it to guide self-supervised refinement of jointly estimated flow and depth. In contrast, heuristic measures like forward–backward consistency detect occlusions at negligible cost but are strictly post-hoc. Most recently, U^{2}Flow (Sun et al., 2026) learns per-pixel uncertainty in an unsupervised manner via augmentation consistency and reinjects it into the recurrent update through a sigmoid reliability gate. However, its one-sided gating can only suppress rather than correct features under miscalibration. This limitation motivates our Adaptive Uncertainty Gating design, instantiated as ResidualFiLM, which generalizes the fixed sigmoid gate to a bounded affine modulation conditioned on predicted uncertainty (Section 3.3).

2.3 ATTENTION MECHANISMS IN OPTICAL FLOW

Attention mechanisms have become mainstream in optical flow estimation, as their explicit modeling of long-range dependencies and global motion contexts makes them particularly adept at handling occlusions and large displacements compared to CNNs (Jiang et al., 2021; Xu et al., 2022; Huang et al., 2022; Shi et al., 2023b). Methods can be functionally categorized into two groups: feature extraction (GMA (Jiang et al., 2021), GMFlow (Xu et al., 2022), Attention Residuals (Kimi Team et al., 2026)) and iterative flow refinement (GMFlow (Xu et al., 2022), FlowFormer (Huang et al., 2022), FlowFormer++ (Shi et al., 2023b), CRAFT (Sui et al., 2022), GMFlow+ (Xu et al., 2023)). However, they suffer from quadratic computational complexity and high memory consumption. Hence, CNNs retain significant advantages in computational efficiency and local feature extraction, as evidenced by DPFlow (Morimitsu et al., 2025) and U^{2}Flow (Sun et al., 2026), demonstrating their indispensable role in optical flow estimation across both supervised and unsupervised settings.

These observations motivate our design of AttnResEncoder, which adapts attention residuals to CNN encoders for adaptive multi-scale fusion (Section 3.2), and the integration of CP-GMA into our unsupervised optical flow framework for global aggregate motion modeling and occlusion-aware propagation (Section 3.4).

3 METHOD

A^{3}Flow extends the uncertainty-aware recurrent framework of U^{2}Flow with AttnResEncoder, ResidualFiLM for the recurrent update block, and content-and-position global motion aggregation, as illustrated in Fig. 2. We first introduce the core design of U^{2}Flow in Section 3.1, then describe the three proposed components in Sections 3.2–3.4.

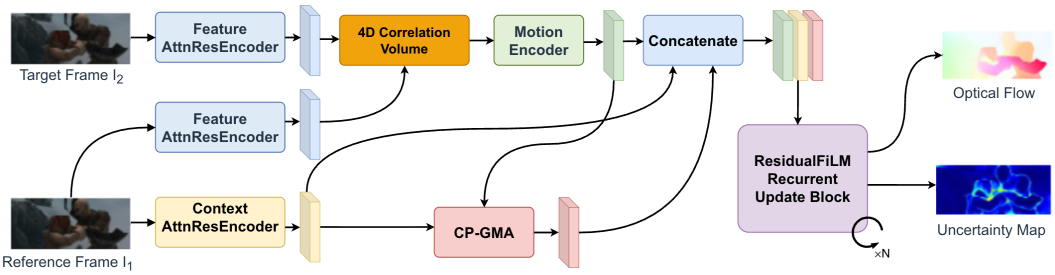


Figure 2: Overview of the A^{3}Flow framework. Following the core design of U^{2}Flow, three core modules are newly designed or integrated: (1) AttnResEncoders for adaptive multi-scale feature and context extraction; (2) CP-GMA for occlusion-aware motion aggregation; and (3) ResidualFiLM for uncertainty-conditioned modulation of the intermediate flow within the recurrent update block.

3.1 Preliminaries: the U^{2}Flow framework

Given two consecutive images I_0 and I_1, the objective is to estimate a dense optical flow field $F_{1\to2} \in \mathbb{R}^{H\times W\times 2}$ from I_0 to I_1. Let $h^{(k)}$ denote the GRU hidden state at iteration k. U^{2}Flow predicts a flow feature $f^{(k)}$ and a log-variance $\alpha^{(k)}$ from the hidden state, where

$$f^{(k)} = C_{\text{flow}}(h^{(k)}), \qquad \alpha^{(k)} = C_{\text{unc}}(h^{(k)}) = \log\big(\sigma^{2(k)}_{1\to2}\big). \tag{1}$$

The log-variance is converted into a sigmoid reliability gate and detached before being applied to the flow feature. The modulated feature is then concatenated with the original flow feature and detached uncertainty to predict the flow residual:

$$s^{(k)} = \varphi\big(-(\alpha^{(k)})^*\big), \qquad \Delta F^{(k)}_{1\to2} = C'_{\text{flow}}\left(\text{concat}\left[f^{(k)}, f^{(k)} \odot s^{(k)}, (\alpha^{(k)})^*\right]\right). \tag{2}$$

Here, C_{flow} and C_{unc} denote the flow-feature and uncertainty heads, respectively, φ is the sigmoid function, and $\odot$ denotes element-wise multiplication. The superscript * indicates stop-gradient. Thus, the flow loss does not back-propagate through the uncertainty-dependent modulation, preserving the uncertainty calibration.

3.2 Attention Residuals for Feature Extractor (AttnResEncoder)

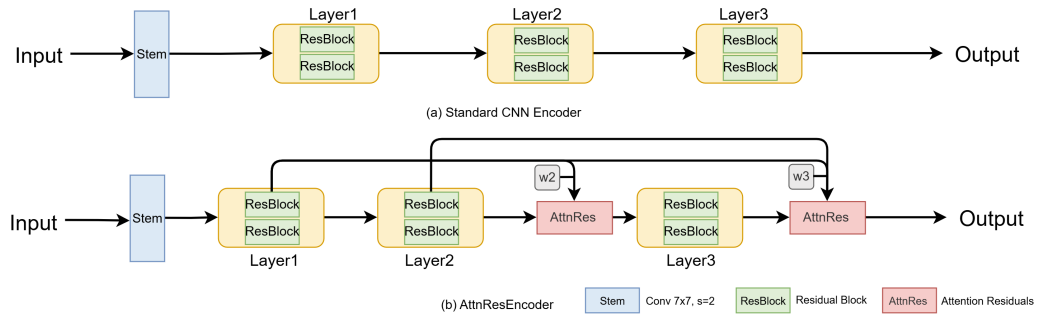


Figure 3: Architecture comparison of the proposed AttnResEncoder.

Standard ResNet encoders (He et al., 2016) propagate hierarchical representations through fixed feed-forward residual connections. We replace this fixed aggregation with an Attention Residuals mechanism adapted from Kimi Team (Kimi Team et al., 2026), allowing the encoder to dynamically determine the contribution of preceding stages according to the input context.

Specifically, each encoder stage is treated as a source block, and its output is projected by a 1×1 convolution and globally pooled to obtain a compact representation. For the l-th target stage, these

representations are used to compute attention weights, while the corresponding projected feature maps are resized to the current resolution. The resulting feature is obtained by

$$x_l = \sum_{i=1}^{l} \mathrm{softmax}_i \left(w_l^\top \, \mathrm{LN}(k_i) \right) \cdot v_i^{(l)}, \qquad k_i = v_i = \mathrm{GAP}\left(P_i(x_i) \right), \tag{3}$$

where w_l is a learnable pseudo-query associated with stage l, P_i is a 1×1 projection, and LN denotes LayerNorm. Here, $v_i^{(l)}$ represents the projected feature map from source stage i, bilinearly interpolated to the target resolution $H_l \times W_l$. The attention weight is a per-sample scalar and is broadcast over the spatial dimensions. For the first stage, $x_1 = f_1(x_0)$ with $x_0 = x$, corresponding to the special case of a single source block. In this way, AttnResEncoder replaces fixed residual addition with input-dependent, softmax-weighted aggregation of hierarchical features.

3.3 RESIDUALFILM: BOUNDED ADAPTIVE UNCERTAINTY GATING

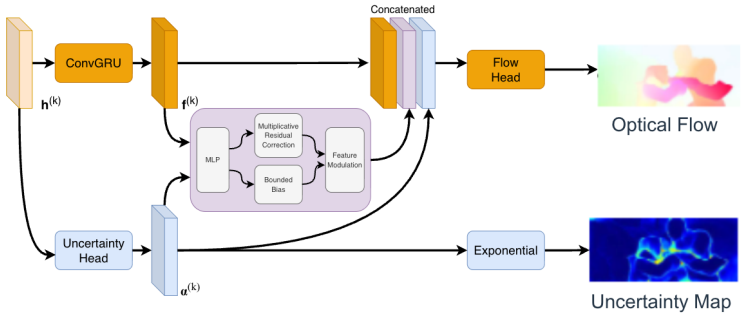


Figure 4: Architecture of the ResidualFiLM.

U^{2}Flow incorporates uncertainty into recurrent refinement through a fixed sigmoid reliability gate. Although effective, this gate only attenuates the flow feature and cannot amplify or adapt its response to local motion statistics or the current iteration. Moreover, inaccurate uncertainty estimates, especially in occluded regions during early training, can lead to inappropriate attenuation. To address these limitations, we propose ResidualFiLM, which replaces the fixed gate with a bounded feature-wise affine modulation (Perez et al., 2018). The modulation supports both attenuation and amplification while retaining the baseline gate at initialization.

At recurrent iteration k, the flow feature $f^{(k)}$ and log-variance $\alpha^{(k)}$ are first obtained from the ConvGRU and uncertainty head, respectively. The detached and normalized uncertainty is then fed into an MLP to generate modulation variables:

$$(g_\gamma^{(k)}, g_\delta^{(k)}) = \mathrm{MLP}\Big(\mathrm{Normalize}((\alpha^{(k)})^*) \Big), \tag{4}$$

where $(\cdot)^*$ denotes stop-gradient. The modulation parameters are obtained by applying bounded residual corrections to the original reliability gate:

$$\gamma^{(k)} = s^{(k)} \exp\Big(\lambda_\gamma \tanh(g_\gamma^{(k)}) \Big), \qquad \delta^{(k)} = \lambda_\delta \tanh(g_\delta^{(k)}), \tag{5}$$

where λ_γ is the bounding constant, and λ_δ is the reliability weight. The flow feature is then modulated by a feature-wise affine transformation:

$$\tilde{f}^{(k)} = \gamma^{(k)} \odot f^{(k)} + \delta^{(k)}. \tag{6}$$

The final MLP layer is zero-initialized, yielding $g_\gamma^{(k)} = g_\delta^{(k)} = 0$ at initialization and thus $\tilde{f}^{(k)} = s^{(k)} \odot f^{(k)}$, exactly recovering the baseline gate. The original feature, modulated feature, and uncertainty are jointly fed to the flow head:

$$\Delta F_{1\to2}^{(k)} = C'_{\mathrm{flow}} \Big(\mathrm{concat}[f^{(k)}, \tilde{f}^{(k)}, (\alpha^{(k)})^*] \Big). \tag{7}$$

ResidualFiLM therefore preserves the uncertainty-aware formulation of U^{2}Flow while introducing bounded, two-sided, and iteration-adaptive feature modulation.

3.4 CONTENT-AND-POSITION GLOBAL MOTION AGGREGATION (CP-GMA)

CP-GMA (Jiang et al., 2021) resolves occlusion-induced correlation failure by propagating motion from confident to occluded pixels via context-aware global self-attention. Let the context/appearance features of the reference frame be $x \in \mathbb{R}^{N \times D_c}$, $N = HW$, and the motion features be $y \in \mathbb{R}^{N \times D_m}$. For pixel i, the aggregated motion feature $\hat{y}_i$ is defined as

$$\hat{y}_i = y_i + \alpha \sum_{j=1}^{N} f\big(\theta(x_i), \varphi(x_j) + p_{j-i}\big)\sigma(y_j), \tag{8}$$

where the similarity attention $f(a_i, b_j)$ is a scaled dot-product softmax:

$$f(a_i, b_j) = \frac{\exp(a_i^\top b_j / \sqrt{D})}{\sum_{j=1}^{N} \exp(a_i^\top b_j / \sqrt{D})}. \tag{9}$$

Here, $\theta(x_i) = W_{\text{qry}} x_i$ and $\varphi(x_j) = W_{\text{key}} x_j$ project context features into the query/key spaces, while $\sigma(y_j) = W_{\text{val}} y_j$ projects motion features into the value space. The relative positional embedding is indexed by the pixel offset $j - i$ and decomposed into horizontal and vertical components:

$$p_{j-i} = p^h_{j_x - i_x} + p^v_{j_y - i_y}. \tag{10}$$

The learnable parameters include $W_{\text{qry}}, W_{\text{key}}, W_{\text{val}}, \alpha$, where α is initialized to zero. The aggregated feature is then concatenated channel-wise with the local motion and context features, yielding $[y \parallel \hat{y} \parallel x]$, which is fed to the recurrent decoder.

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

Datasets. Following the training data schedule of prior works (Liu et al., 2019; Yuan et al., 2024; Sun et al., 2025; 2026), we evaluate A^{3}Flow on KITTI-2012, KITTI-2015 (Menze & Geiger, 2015), and MPI-Sintel (Butler et al., 2012), including both the Clean and Final passes.

Training. We use Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a global batch size of 4. For both KITTI and Sintel, training consists of 100k iterations on raw data at a fixed learning rate of 2×10^{-4}, followed by 150k fine-tuning iterations using OneCycleLR with a peak learning rate of 2.5×10^{-4}, a 5% warm-up phase, and linear annealing. We set the bounding constants $\lambda_\gamma = \ln 2$, and assign a reliability weight λ_δ to 0.05.Automatic mixed precision is enabled. Following ARFlow (Liu et al., 2020) and U^{2}Flow (Sun et al., 2026), we apply appearance transformations (brightness, contrast, saturation, hue, Gaussian blur) and random flipping/swapping for data augmentation, with all input images resized to 256×832 for KITTI and 448×1024 for Sintel. The loss function and its corresponding configuration follow those of U^{2}Flow.

Module settings. The full model enables all three components. AttnResEncoder is used in both feature and context encoders; ResidualFiLM uses a hidden dimension of 64, $\lambda_\gamma = \ln 2$, and $\lambda_\delta = 0.05$. CP-GMA uses one 128-dimensional head with content-and-position attention and a maximum positional size of 160.

4.2 BENCHMARK RESULTS

Table 1 compares A^{3}Flow with representative supervised (PWC-Net+ (Sun et al., 2020), RAFT (Teed & Deng, 2020), FlowFormer (Huang et al., 2022), VideoFlow (Shi et al., 2023a), FlowDiffuser (Luo et al., 2024)) and unsupervised (UnFlow-CSS (Meister et al., 2018), SelFlow (Liu et al., 2019), UFlow (Jonschkowski et al., 2020), ARFlow (Liu et al., 2020), UPFlow (Luo et al., 2021), SMURF (Stone et al., 2021), SemARFlow (Yuan et al., 2023), UnSAMFlow (Yuan et al., 2024), M2Flow (Sun et al., 2025)) methods, with U^{2}Flow (Sun et al., 2026) as our baseline; supervised results serve as an upper-bound reference. Using only two frames, A^{3}Flow establishes a new state of the art among unsupervised methods on Sintel and KITTI 2015: it improves over U^{2}Flow on all Sintel metrics—test EPE from 2.83 to 2.30 (Clean, -18.7%) and 4.16 to 3.41 (Final, -18.0%)—and reduces the KITTI outlier rates (Fl-all/Fl-noc: $6.37/3.48 \rightarrow 6.20/3.35$ on 2012,

$6.13/4.56 \rightarrow 5.74/4.27$ on 2015). It even surpasses the multi-frame U^{2}Flow$^{\mathrm{MF}}$ on Sintel and the KITTI Fl columns despite using no extra frames, and remains competitive with SMURF (Stone et al., 2021) on KITTI 2012.

Notably, the gains are the most pronounced in occluded regions: on both KITTI benchmarks, the improvement in Fl-all consistently exceeds that in Fl-noc (0.39 vs. 0.29 pp on 2015; 0.17 vs. 0.13 pp on 2012), providing direct quantitative evidence that A^{3}Flow recovers motion in occluded areas—consistent with CP-GMA, which propagates motion evidence into occlusions, and AttnResEncoder, which strengthens features in textureless areas near occlusion boundaries.

Table 1: Quantitative results on Sintel and KITTI online benchmarks. Metrics are evaluated over "all" (all pixels) and "noc" (non-occluded pixels). † denotes models with semantic inputs. MF denotes methods trained using multi-frame data. Parentheses indicate that training and testing are conducted on the same dataset.

	Sintel Clean		Sintel Final		KITTI 2012				KITTI 2015		
	Train	Test	Train	Test	Train	Test			Train	Test	
Method	EPE	EPE	EPE	EPE	EPE	Fl-all	Fl-noc	EPE	EPE	Fl-all	Fl-noc
Supervised											
PWC-Net+	(1.71)	3.45	(2.34)	4.60	(0.99)	6.72	3.36	1.4	(1.47)	7.72	4.91
RAFT	(0.77)	1.61	(1.27)	2.86	-	-	-	-	(0.63)	5.10	3.07
FlowFormer	(0.48)	1.16	(0.74)	2.09	-	-	-	-	**(0.53)**	4.68	**2.69**
VideoFlow$^{\mathrm{MF}}$	**(0.46)**	**0.99**	**(0.66)**	**1.62**	-	-	-	-	(0.56)	**3.65**	-
FlowDiffuser	-	1.02	-	2.03	-	-	-	-	-	4.17	2.82
Unsupervised+2frames											
UnFlow-CSS	-	9.38	(7.91)	10.22	3.29	-	-	-	8.10	23.27	-
UFlow	(2.50)	5.21	(3.39)	6.50	1.68	7.91	4.26	1.9	2.71	11.13	8.41
ARFlow	(2.79)	4.78	(3.73)	5.89	1.44	-	-	1.8	2.85	11.80	-
UPFlow	(2.33)	4.68	(2.67)	5.32	1.27	-	-	**1.4**	2.45	9.38	-
SemARFlow†	-	-	-	-	1.28	7.35	3.90	1.5	2.18	8.38	5.43
UnSA$^{\mathrm{MF}}$low†	(2.21)	3.93	(3.07)	5.20	1.26	7.05	3.79	**1.4**	2.01	7.83	5.67
U^{2}Flow	(1.42)	2.83	(2.32)	4.16	1.19	6.37	3.48	**1.4**	1.83	6.13	4.56
A^{3}Flow (ours)	**(1.23)**	**2.30**	**(2.10)**	**3.41**	**1.15**	**6.20**	**3.35**	**1.4**	**1.78**	**5.74**	**4.27**
Unsupervised+Multi-frame											
SMURF$^{\mathrm{MF}}$	(1.71)	3.15	(2.58)	4.18	-	**6.19**	**3.13**	1.4	2.00	6.83	5.26
SelFlow$^{\mathrm{MF}}$	(2.88)	6.56	(3.87)	6.57	1.69	7.68	4.31	2.2	4.84	14.19	9.65
M2Flow$^{\mathrm{MF}}$	(2.01)	3.38	(3.12)	5.01	1.09	6.24	3.95	**1.2**	1.95	7.37	5.73
U^{2}Flow$^{\mathrm{MF}}$	**(1.36)**	**2.83**	**(2.29)**	**4.10**	**1.12**	6.26	3.47	1.3	**1.74**	**6.00**	**4.52**

4.3 Ablation studies

Table 2 presents a progressive ablation of A^{3}Flow, adding AttnResEncoder(AttnRes), ResidualFiLM(ResFiLM), and CP-GMA sequentially to the baseline, so that each consecutive row isolates one component's contribution. The components are complementary, with a clear dataset-dependent division of labor. AttnResEncoder primarily benefits KITTI 2015 (EPE $1.95 \rightarrow 1.78$, Fl-all $6.85\% \rightarrow 6.32\%$), with only marginal gains on Sintel at this stage (Clean $1.42 \rightarrow 1.40$). ResidualFiLM further refines KITTI to 1.75 EPE / 6.18% Fl-all and mildly improves Sintel (Final $2.32 \rightarrow 2.27$, Clean $1.40 \rightarrow 1.34$). Adding CP-GMA further reduces Sintel Final/Clean EPE from $2.27/1.34$ to $2.10/1.23$, yielding the largest incremental improvement on Sintel in this component ordering. On KITTI 2015, CP-GMA reduces Fl-all from 6.18% to 5.74%, while slightly increasing EPE from 1.75 to 1.78. This reflects a trade-off between mean endpoint error and outlier rate: the full model produces fewer outlier pixels, but a slightly higher average endpoint error. These aggregate results do not establish which motion or occlusion categories account for the changes.

Table 3 examines cross-dataset generalization. A^{3}Flow transfers better than the strongest U^{2}Flow variant (w/ Fusion) from Sintel to KITTI (EPE $4.77 \rightarrow 4.41$, Fl-all $16.58\% \rightarrow 16.04\%$) and comparably to Sintel Clean ($4.23 \rightarrow 4.21$, the best among all variants), but degrades in the reverse direction on Sintel Final ($5.49 \rightarrow 6.47$), indicating that driving-scene features transfer poorly to the atmospheric effects and motion blur of the Final pass.

Table 2: Incremental ablation on the Sintel training passes (EPE) and KITTI 2015 training split (EPE / Fl-all, %).

AttnRes	ResFiLM	CP-GMA	Sintel Final	Sintel Clean	KITTI 2015 EPE	KITTI 2015 Fl-all
-	-	-	2.32	1.42	1.95	6.85
✓	-	-	2.32	1.40	1.78	6.32
✓	✓	-	2.27	1.34	**1.75**	6.18
✓	✓	✓	**2.10**	**1.23**	1.78	**5.74**

Table 3: Cross-dataset generalization. Models are trained on one dataset and tested on the other without fine-tuning.

Method	KITTI → Sintel Final	KITTI → Sintel Clean	Sintel → KITTI EPE	Sintel → KITTI Fl-all
U^{2}Flow (w/o Fusion)	5.80	4.67	5.10	17.05
U^{2}Flow (w/ Fusion)	5.49	4.23	4.77	16.58
Ours	6.47	**4.21**	**4.41**	**16.04**

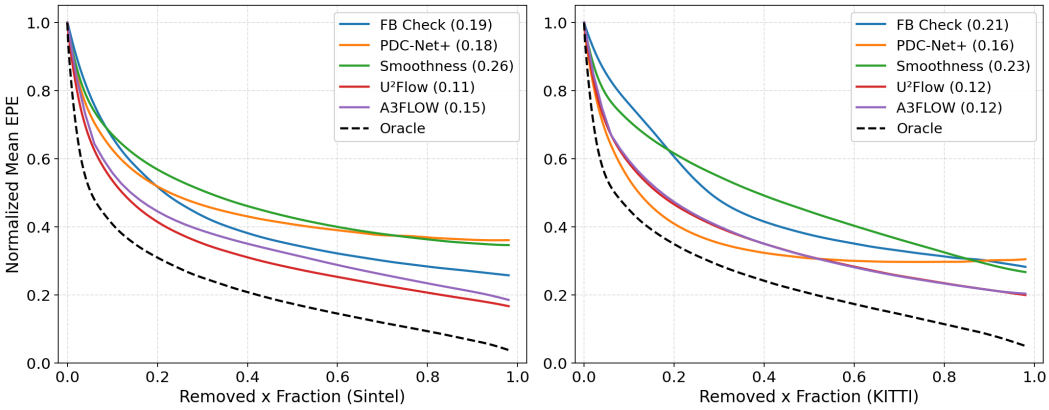


Figure 5: Sparsification curves for uncertainty evaluation on the Sintel (Left) and KITTI (Right) training sets.

4.4 Uncertainty Evaluation

Following the U^{2}Flow evaluation, we report the Area Under the Sparsification Error curve (AUSE) and Spearman's rank correlation coefficient (CC) between predicted uncertainty and EPE. As shown in Table 4, A^{3}Flow matches U^{2}Flow exactly on KITTI (AUSE 0.12, CC 0.64) and clearly outperforms the classical estimators (AUSE 0.16–0.23); on Sintel, it remains the second most reliable method by AUSE (0.15, versus 0.18–0.26 for the classical estimators and 0.11 for U^{2}Flow). The sparsification-error curves in Figure 5 corroborate this behavior. These results confirm that the accuracy gains of A^{3}Flow are not achieved at the expense of uncertainty quality.

Table 4: Uncertainty evaluation on the KITTI and Sintel training sets. AUSE and CC are computed per image and averaged over the images in each combined benchmark. Results quoted from Table 2 of U^{2}Flow (Sun et al., 2026).

Method	KITTI AUSE ↓	KITTI CC ↑	Sintel AUSE ↓	Sintel CC ↑
FB Check (Meister et al., 2018)	0.21	<u>0.53</u>	0.19	<u>0.57</u>
PDC-Net+ (Truong et al., 2023)	<u>0.16</u>	0.50	0.18	0.45
Smoothness (Abdein et al., 2025)	0.23	0.45	0.26	0.43
U^{2}Flow (Sun et al., 2026)	**0.12**	**0.64**	0.11	0.66
Ours	**0.12**	**0.64**	<u>0.15</u>	0.56

4.5 Qualitative Results

Figure 6 presents qualitative results on the KITTI test set and Sintel (final pass) test set. Compared with U^{2}Flow (Sun et al., 2026), A^{3}Flow produces more accurate and spatially coherent flow fields, particularly around motion boundaries and in occluded or weak-texture regions. The error maps further confirm that A^{3}Flow effectively suppresses large errors in these challenging areas. These visual improvements are consistent with the quantitative gains in Table 1 and demonstrate the ben-

efits of adaptive feature fusion, uncertainty-conditioned modulation, and occlusion-aware motion aggregation.

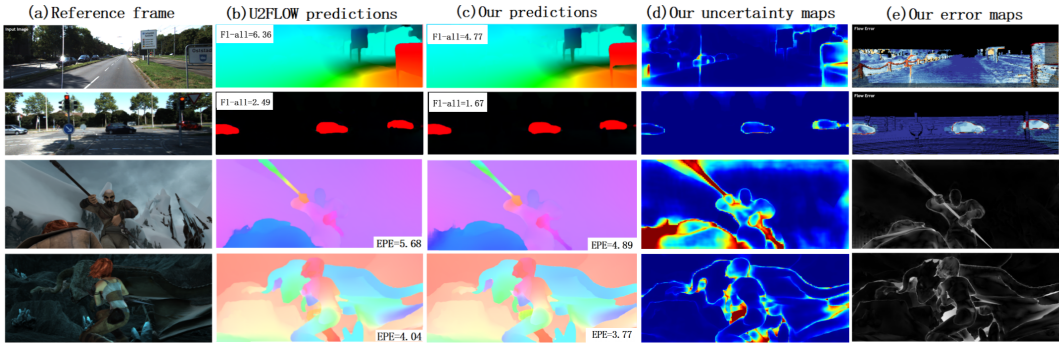


Figure 6: Qualitative results on the KITTI test set (sample frames #2 and #9) and the Sintel (final pass) test set (samples: ambush_3, frame 23; cave_3, frame 16), compared with U^{2}Flow. Additional examples can be found on the official benchmark website.

5 Conclusion

In this paper, we propose A^{3}Flow, a novel uncertainty-aware framework for unsupervised optical flow estimation. A^{3}Flow introduces three complementary components: AttnResEncoder for input-dependent cross-scale feature fusion, ResidualFiLM for bounded uncertainty-conditioned flow modulation, and CP-GMA for propagating reliable motion evidence into occluded regions. Extensive experiments on MPI-Sintel and KITTI demonstrate that A^{3}Flow establishes a new state of the art among two-frame unsupervised methods and outperforms several multi-frame approaches. The improvements are particularly pronounced in occluded and textureless regions, while competitive uncertainty calibration and strong cross-dataset generalization further highlight the effectiveness and robustness of the proposed framework.

Limitations and Future Work

Although A^{3}Flow achieves state-of-the-art unsupervised performance, it remains limited by the quadratic complexity of CP-GMA, a degradation in uncertainty calibration on Sintel, and weak KITTI-to-Sintel Final generalization. Future work will introduce residual connections between AttnResEncoder layers, develop more reliable uncertainty estimation methods, and design stronger and more adaptive ResidualFiLM modulation strategies.

AI USE STATEMENT

Generative AI (OpenAI Codex) was used to assist with drafting and revising portions of the manuscript, improving English expression, converting text and equations to LaTeX, and formatting tables and references. It was also used to identify relevant literature and assist with checking bibliographic information. During the research and revision process, AI provided feedback on methodological descriptions, experimental comparisons, ablation design, and the interpretation and presentation of results. The authors retain responsibility for the research decisions and the accuracy of the final manuscript, including all claims, citations, and reported results. AI assistance was also used for implementing and debugging parts of the experimental code.

REPRODUCIBILITY STATEMENT

The architecture and equations are described in Section 3. Training schedules, optimizer settings, and module hyperparameters are provided in Section 4.1; the uncertainty evaluation protocol is described in Section 4.4.

REFERENCES

Abdein, R., Li, W., Li, C., Zheng, X., and R. Yadav. Self-supervised uncertainty-guided refinement for robust joint optical flow and depth estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025. URL https://ieeexplore.ieee.org/document/10889503.

Andrea Alfarano, Luca Maiano, Lorenzo Papa, and Irene Amerini. Estimating optical flow: A comprehensive review of the state of the art. *Computer Vision and Image Understanding*, 249: 104160, 2024. doi: 10.1016/j.cviu.2024.104160.

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A Naturalistic Open Source Movie for Optical Flow Evaluation. In *ECCV*, 2012.

Qiaole Dong and Yanwei Fu. MemFlow: Optical Flow Estimation and Prediction with Memory. In *CVPR*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Dong_MemFlow_Optical_Flow_Estimation_and_Prediction_with_Memory_CVPR_2024_paper.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

Hsin-Ping Huang, Charles Herrmann, Junhwa Hur, Erika Lu, Kyle Sargent, Austin Stone, Ming-Hsuan Yang, and Deqing Sun. Self-supervised AutoFlow. In *CVPR*, 2023. URL https://autoflow-google.github.io/.

Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. FlowFormer: A Transformer Architecture for Optical Flow. In *ECCV*, 2022. URL https://arxiv.org/abs/2203.16194.

Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation. In *CVPR*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Jiang_Super_SloMo_High_CVPR_2018_paper.html.

Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to Estimate Hidden Motions with Global Motion Aggregation. In *ICCV*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Jiang_Learning_To_Estimate_Hidden_Motions_With_Global_Motion_Aggregation_ICCV_2021_paper.html.

Rico Jonschkowski, Austin Stone, Jonathan T. Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What Matters in Unsupervised Optical Flow. In *ECCV*, 2020.

Kimi Team, Guangyu Chen, Yu Zhang, Jianlin Su, Weixin Xu, Siyuan Pan, Yaoyu Wang, Yucheng Wang, Guanduo Chen, Bohong Yin, Yutian Chen, Junjie Yan, Ming Wei, Y. Zhang, Fanqing Meng, Chao Hong, Xiaotong Xie, Shaowei Liu, Enzhe Lu, Yunpeng Tai, Yanru Chen, Xin Men, Haiqing Guo, Y. Charles, Haoyu Lu, Lin Sui, Jinguo Zhu, Zaida Zhou, Weiran He, Weixiao Huang, Xinran Xu, Yuzhi Wang, Guokun Lai, Yulun Du, Yuxin Wu, Zhilin Yang, and Xinyu Zhou. Attention Residuals. *arXiv preprint arXiv:2603.15031*, 2026. URL https://arxiv.org/abs/2603.15031.

Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by Analogy: Reliable Supervision from Transformations for Unsupervised Optical Flow Estimation. In *CVPR*, 2020.

Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. SelFlow: Self-Supervised Learning of Optical Flow. In *CVPR*, 2019.

Ao Luo, Xin Li, Fan Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. FlowDiffuser: Advancing Optical Flow Estimation with Diffusion Models. In *CVPR*, 2024.

Kunming Luo, Chuan Wang, Shuaicheng Liu, Haoqiang Fan, Jue Wang, and Jian Sun. UPFlow: Upsampling Pyramid for Unsupervised Optical Flow Learning. In *CVPR*, 2021.

Simon Meister, Junhwa Hur, and Stefan Roth. UnFlow: Unsupervised Learning of Optical Flow with a Bidirectional Census Loss. In *AAAI*, 2018.

Moritz Menze and Andreas Geiger. Object Scene Flow for Autonomous Vehicles. In *CVPR*, 2015.

Henrique Morimitsu, Xiaobin Zhu, Roberto M. Cesar Jr., Xiangyang Ji, and Xu-Cheng Yin. Dpflow: Adaptive optical flow estimation with a dual-pyramid framework. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2503.14880.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. In *AAAI*, 2018. URL https://arxiv.org/abs/1709.07871.

Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, and Hongyuan Zha. Unsupervised deep learning for optical flow estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. doi: 10.1609/aaai.v31i1.10723.

Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. VideoFlow: Exploiting Temporal Cues for Multi-Frame Optical Flow Estimation. In *ICCV*, 2023a.

Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. FlowFormer++: Masked Cost Volume Autoencoding for Pretraining Optical Flow Estimation. In *CVPR*, 2023b. URL https://openaccess.thecvf.com/content/CVPR2023/html/Shi_FlowFormer_Masked_Cost_Volume_Autoencoding_for_Pretraining_Optical_Flow_Estimation_CVPR_2023_paper.html.

Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, and Rico Jonschkowski. SMURF: Self-Teaching Multi-Frame Unsupervised RAFT With Full-Image Warping. In *CVPR*, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Stone_SMURF_Self-Teaching_Multi-Frame_Unsupervised_RAFT_With_Full-Image_Warping_CVPR_2021_paper.html.

Xiuchao Sui, Shaohua Li, Xue Geng, Yan Wu, Xinxing Xu, Yong Liu, Rick Goh, and Hongyuan Zhu. Craft: Cross-attentional flow transformer for robust optical flow. In *CVPR 2022 camera ready*, 2022. URL https://arxiv.org/abs/2203.16896.

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Models Matter, So Does Training: An Empirical Study of CNNs for Optical Flow Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Deqing Sun, Daniel Vlasic, Charles Herrmann, Varun Jampani, Michael Krainin, Huiwen Chang, Ramin Zabih, William T. Freeman, and Ce Liu. AutoFlow: Learning a Better Training Set for Optical Flow. In *CVPR*, 2021. URL https://autoflow-google.github.io/.

Deqing Sun, Charles Herrmann, Fitsum Reda, Michael Rubinstein, David J. Fleet, and William T. Freeman. Disentangling Architecture and Training for Optical Flow. In *ECCV*, 2022. URL https://autoflow-google.github.io/.

Xunpei Sun, Gang Chen, and Zuoxun Hou. M2Flow: A Motion Information Fusion Framework for Enhanced Unsupervised Optical Flow Estimation in Autonomous Driving. In *AAAI*, 2025. URL https://ojs.aaai.org/index.php/AAAI/article/view/32767.

Xunpei Sun, Wenwei Lin, Yi Chang, and Gang Chen. U^{2}Flow: Uncertainty-Aware Unsupervised Optical Flow Estimation. In *CVPR*, 2026. URL https://arxiv.org/abs/2604.10056.

Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *ECCV*, 2020. URL https://arxiv.org/abs/2003.12039.

Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. PDC-Net+: Enhanced Probabilistic Dense Correspondence Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. GMFlow: Learning Optical Flow via Global Matching. In *CVPR*, 2022.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. In *TPAMI*, 2023. URL https://arxiv.org/abs/2211.05783.

Shuai Yuan, Xian Sun, Hannah Kim, Shuzhi Yu, and Carlo Tomasi. Optical Flow Training under Limited Label Budget via Active Learning. In *ECCV*, 2022. URL https://github.com/duke-vision/optical-flow-active-learning-release.

Shuai Yuan, Shuzhi Yu, Hannah Kim, and Carlo Tomasi. SemARFlow: Injecting Semantics into Unsupervised Optical Flow Estimation for Autonomous Driving. In *ICCV*, 2023.

Shuai Yuan, Lei Luo, Zhuo Hui, Can Pu, Xiaoyu Xiang, Rakesh Ranjan, and Denis Demandolx. UnSAMFlow: Unsupervised Optical Flow Guided by Segment Anything Model. In *CVPR*, 2024.

Songyan Zhang, Yongtao Ge, Jinyuan Tian, Guangkai Xu, Hao Chen, Chen Lv, and Chunhua Shen. POMATO: Marrying Pointmap Matching with Temporal Motions for Dynamic 3D Reconstruction. In *ICCV*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_POMATO_Marrying_Pointmap_Matching_with_Temporal_Motions_for_Dynamic_3D_ICCV_2025_paper.html.