FlowPainter: Inpainting Optical Flow via Confidence-Guided Completion

Yuang Meng[1,3], Chenyang Wu[1], Xianshun Liu[1,3], Chun-Le Guo[1] ⋆, Zichen Liang[1], Lina Lei[1], Jie Liang[3], Hui Zeng[3], Chongyi Li[1], and Lei Zhang[2,3]

[1] VCIP, CS, Nankai University
[2] The Hong Kong Polytechnic University
[3] OPPO Research Institute
mya@mail.nankai.edu.cn

Abstract. Existing optical flow estimation methods broadly follow two paradigms: iterative optimization and diffusion-based estimation. Iterative methods, exemplified by RAFT, achieve accurate flow estimation through recurrent refinement, but can still be challenged by large displacements and complex motion patterns. Diffusion-based methods introduce generative modeling into optical flow and have shown promising results in such ambiguous regions. However, existing diffusion-based flow models usually denoise the entire dense flow field from Gaussian noise, including simple regions where reliable motion structure can already be estimated by a lightweight network. This increases the denoising burden and may lead to slow convergence and unstable training. To address these issues, we introduce FlowPainter, a diffusion-based optical flow framework that reformulates dense-flow generation as confidence-guided soft inpainting. FlowPainter first employs a lightweight confidence-aware network to predict a rough flow and a pixel-wise confidence mask, which serves as a reliability gate for distinguishing reliable simple regions from uncertain hard regions. The resulting simple-flow prior is used for confidence-based initialization and is further injected into the iterative denoising process through confidence-gated residual guidance. With a dynamically decaying guidance strength, FlowPainter stabilizes early denoising while preserving the flexibility of the diffusion model for late-stage detail refinement. Extensive experiments on public benchmarks, including Sintel, KITTI, and Spring, demonstrate that FlowPainter achieves strong accuracy under comparable training settings and improves convergence efficiency over existing diffusion-based optical flow methods, with notable gains on challenging benchmark splits. Our approach provides a practical direction for integrating reliable discriminative priors with diffusion-based refinement for optical flow estimation.

Keywords: Optical flow · Diffusion model · Inpainting

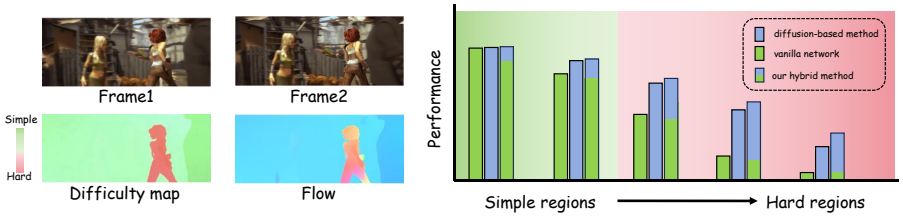
⋆ C. L. Guo is the corresponding author.

1 Introduction

Optical flow estimation is a fundamental problem in computer vision, aiming to infer a dense motion vector for each pixel between consecutive frames. It is widely used in video understanding [8,16], action recognition [26,35], autonomous driving [2,19,31], video frame interpolation [10,28,41], and restoration problems [3,22]. Despite decades of progress, accurate optical flow estimation remains challenging in real-world scenarios due to fast motion, occlusions, illumination changes, and textureless regions.

Deep learning has greatly advanced optical flow estimation. Early CNN-based methods [7,13,29,34] introduced end-to-end regression, stacked refinement, and coarse-to-fine matching, but their fixed receptive fields and limited feature expressiveness can hinder complex motion modeling. RAFT [36] established a strong iterative refinement paradigm by building an all-pairs correlation volume and recurrently updating the flow field. Its follow-up works [15,27,37,40] further improve global aggregation, 3D motion modeling, training strategies, and practical efficiency. Nevertheless, iterative optimization can still be challenged by very large displacements, severe occlusions, and complex non-rigid motion, where correlation construction and recurrent updates may become less reliable [17].

In parallel, diffusion models have recently been explored for optical flow estimation [6,17,25,30]. Methods such as DDVM [30] and FlowDiffuser [17] formulate optical flow estimation as a denoising process from Gaussian noise to dense flow, conditioned on image features. By introducing generative modeling into flow estimation, these methods show promising results in ambiguous regions with large displacement or complex motion. However, existing diffusion-based flow models usually denoise the whole dense flow field from noise, including simple regions where a lightweight estimator can already provide reliable motion structure. This increases the denoising burden and may slow convergence or destabilize training, especially when the model has to reconstruct precise flow solely from Gaussian noise.

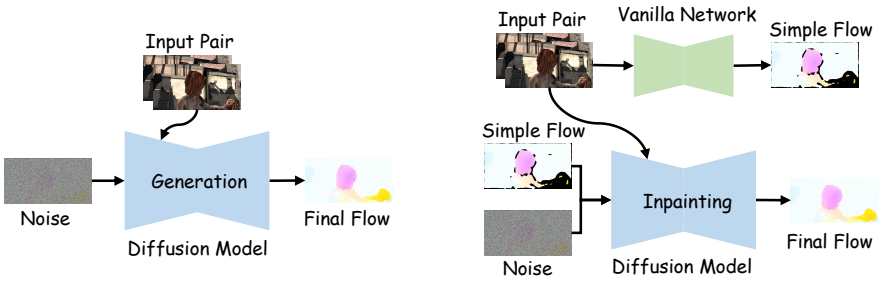


(a) An example of the difficulty of optical flow estimation in different regions of a scene in Sintel. Simple regions account for a large proportion, which is consistent with the statistics observed in the dataset.

(b) A simplified example illustrating the performance of different optical flow paradigms across varying levels of task difficulty. In our hybrid method, green and blue denote the contributions of the vanilla network and the diffusion-based network to the final result, respectively.

Fig. 1: Illustration of region-wise flow difficulty in real scenes and the resulting motivation for our hybrid paradigm across different optical flow methods.

A closer examination of the two paradigms, i.e., iterative optimization [15, 27, 37, 40] and diffusion-based approaches [17, 25, 30], suggests an important observation. As shown in Fig. 1 (a), the example depicts a moving person and a background that shifts in tandem. The rapid motion of the person and the non-rigid deformation of the human body make optical flow estimation challenging in the foreground. In contrast, estimating the flow of the background regions that translate alongside the person is relatively simpler. This observation corresponds to the trend illustrated in Fig. 1 (b). For "simple regions" with small motion, usually clear textures or smooth backgrounds, even a vanilla network in the iterative optimization paradigm can produce flow estimates comparable to those of a diffusion-based method. In contrast, for "hard regions" involving large displacements, occlusions, or motion boundaries, such a vanilla network often struggles, whereas diffusion models can leverage their generative modeling ability to refine or recover plausible motion. Meanwhile, diffusion-based methods usually apply the denoising process to the whole flow field, including large simple regions, which increases the denoising burden for regions where reliable motion structure is already available. We provide further empirical evidence for this viewpoint in Sec. 3.1. This motivates a hybrid strategy that combines the strengths of both paradigms: using a vanilla network to efficiently handle simple regions and provide informative priors, while guiding a diffusion model to refine uncertain hard regions. In this way, the diffusion model can better exploit its refinement ability in hard-flow regions, while the vanilla network provides structural priors that help stabilize the generation process and improve convergence efficiency.



(a) Previous diffusion-based methods (b) Our proposed pipeline

Fig. 2: An overview of previous diffusion-based methods and our proposed framework **FlowPainter**. Previous diffusion-based methods generated optical flow entirely from random Gaussian noise. FlowPainter first employs a vanilla network to generate simple optical flow for small motion regions, then uses this as prior information to guide the inpainting of the final optical flow.

Based on this idea, we propose **FlowPainter**, a confidence-guided diffusion framework for optical flow estimation, as shown in Fig. 2. FlowPainter first employs a lightweight Confidence-Aware Network to predict a rough flow and

a pixel-wise confidence mask. The confidence mask serves as a reliability gate: high-confidence regions provide simple-flow priors, while low-confidence regions rely more on image conditioning and diffusion refinement. The simple-flow prior is used for confidence-based initialization and is further injected into the denoising process through confidence-gated residual guidance. A time-decayed guidance schedule stabilizes early denoising and gradually relaxes the prior constraint for late-stage detail refinement. Together with standard image conditioning features used in diffusion-based optical flow [17], this design reformulates dense-flow generation as a soft inpainting-style refinement problem, improving training stability and final accuracy over existing diffusion-based optical flow methods.

Our main contributions are summarized as follows:

- We propose a confidence-guided optical flow framework that integrates a lightweight Confidence-Aware Network with diffusion-based refinement, using reliable rough flow as a structural prior for soft inpainting-style denoising.
- We introduce confidence-based initialization and confidence-gated residual guidance with a time-decayed schedule, reducing the denoising burden in reliable simple regions while preserving flexibility for uncertain hard regions.
- Extensive experiments on Sintel [1], KITTI [23], and Spring [21] show that FlowPainter achieves strong accuracy under comparable training settings and improves convergence efficiency over existing diffusion-based optical flow methods.

2 Related Work

2.1 Iterative Optimization Paradigm for Optical Flow

End-to-end deep optical flow estimation started with FlowNet [7], which first showed that convolutional neural networks can directly regress dense flow from image pairs. FlowNet2 [13] improved accuracy via stacking and specialized small-displacement branches, at the cost of higher computation and memory. SPyNet [29] adopted a coarse-to-fine pyramid to better handle large motions, while PWC-Net [34] combined pyramid features, warping and a cost volume to achieve a stronger accuracy–efficiency trade-off.

RAFT [36] established a highly influential iterative refinement paradigm by building a high-resolution 4D all-pairs correlation volume and repeatedly updating the flow with a recurrent module. Subsequent work [15, 27, 37, 40, 43, 46] largely advances along stronger correlation representations and more effective update. GMA [15] enhances matching and updates under occlusions and long-range dependencies via global motion aggregation, and RAFT-3D [37] extends similar ideas to 3D scene flow. In parallel, approaches with stronger global modeling mitigate the limitations of local matching for extreme displacements: GMFlow [43] leverages Transformer-style global matching, and FlowFormer [12] tokenizes cost volumes and uses Transformer memory for global aggregation and decoding, showing strong performance on datasets such as Sintel [1, 42].

More recently, SEA-RAFT [40] preserves the iterative refinement backbone while simplifying and strengthening loss function and pretraining strategies to improve accuracy–speed trade-offs. For resource-constrained settings, compact designs such as FlowSeek [27] reduce training and deployment costs by using lightweight architectures and external priors (e.g., foundation-model knowledge or low-dimensional motion parameterizations), improving practical usability. Overall, iterative optimization [15, 27, 36, 40, 43] is highly effective for progressive correction, yet it can still be limited when motion is extreme, occlusions are severe or non-rigid deformation is complex, where matching errors may be amplified. Moreover, the update mechanism remains largely discriminative, lacking explicit uncertainty modeling in ambiguous regions [17, 40]—motivating exploration of generative paradigms such as diffusion.

2.2 Diffusion Models and Applications

Diffusion models have rapidly progressed from high-fidelity synthesis to broad conditional generation [38, 47–49] and inverse problem solving [9, 39]. DDPM [11] established the standard formulation of progressive noising and iterative denoising; DDIM [33] introduced implicit sampling trajectories that substantially accelerate inference while remaining compatible with DDPM training.

Motivated by their generative capacity and uncertainty representation, diffusion models have been applied to dense prediction [14, 17, 30], including optical flow. DDVM [30] first formulated flow estimation as conditional generation of flow fields from an image pair, demonstrating competitive performance but often requiring heavy pretraining or substantial computing resources. Reproduction efforts such as Open-DDVM [6] highlight that injecting structural priors such as multi-scale correlation volumes can significantly improve convergence and performance. FlowDiffuser [17] further explores combining diffusion denoising with recurrent decoding and historical states to improve efficiency and better exploit iterative information, while GA-DDVM [25] introduces geometric priors via geometric algebra constraints, underscoring the effectiveness of embedding task structure into the diffusion process.

Despite promising results for large displacements, occlusions and complex motion, existing diffusion-based optical flow [6, 17, 25, 30] faces two key limitations. First, these approaches frequently tend to spend a significant amount of computational cost on simple regions, which limits the computation that can be allocated to truly challenging areas. Second, they are challenged by the difficulty of generating dense flow from Gaussian noise, which slows convergence and can destabilize training—especially in early denoising stages where reliable intermediate structure is lacking. Consequently, a key practical direction is to introduce controllable and reliable priors that accelerate and stabilize optimization without weakening diffusion's expressive power, which naturally motivates our confidence-guided diffusion framework.

Table 1: Endpoint Error results of FlowDiffuser [17] and Vanilla Network on the clean/final pass of Sintel [1] dataset under different optical flow threshold constraints.

Threshold	5	10	20	40	∞
FlowDiffuser [17]	0.121/0.281	0.158/0.384	0.225/0.639	0.323/0.916	0.877/2.271
Vanilla Network	0.132/0.295	0.176/0.415	0.255/0.750	0.421/1.119	1.545/3.849

3 Methodology

3.1 Motivation

For optical flow estimation, we first introduce a practical working partition of motion patterns. We refer to the flow in small-displacement or otherwise easy-to-model regions as **simple flow**, and denote the flow in large-displacement or highly complex motion regions as **hard flow**. It is worth noting that this partition is not intended to provide a complete definition of flow difficulty. Instead, flow magnitude is used as a practical proxy for large-displacement difficulty, while other factors such as occlusion, texture ambiguity, motion boundaries, and non-rigid deformation can also make optical flow estimation challenging.

Identification and Experiments of Issues with Existing Methods. Although diffusion-based optical flow methods [17, 30] have shown noticeable improvements over traditional estimators, we observe that these gains are largely concentrated in scenarios containing substantial hard flow. In other words, when a scene is dominated by simple flow, the advantage of diffusion-based approaches largely diminishes, while their iterative sampling still incurs significantly higher computational cost. In real-world scenes, however, simple and hard flow typically coexist. For such mixed-motion scenes, the benefit of diffusion-based methods again mainly comes from regions with large and complex motion. We support this observation experimentally on the Sintel(train) [1] benchmark by comparing FlowDiffuser [17] with a vanilla network employing an architecture similar to PWC-Net [34]. Concretely, when computing evaluation metrics, we restrict the computation to pixels whose ground-truth flow magnitude is below a threshold. We vary this threshold over 5, 10, 20, 40, ∞ (with ∞ indicating no thresholding). As summarized in Tab. 1, the observed performance gap in favor of diffusion-based estimation [17] consistently increases as the threshold becomes larger, indicating that diffusion models provide progressively greater benefits as larger-displacement and more challenging pixels are included in the evaluation.

Proposal of Hybrid Method. Above findings motivate a hybrid perspective on an effective solution for optical flow. A vanilla network is highly effective in simple regions, where it achieves strong accuracy at substantially lower cost. However, as motion complexity increases, its performance gap with diffusion-based methods becomes more evident, indicating that the two paradigms exhibit complementary strengths. At the same time, standard diffusion-based approaches still denoise the whole dense flow field from noise, including regions where a lightweight estimator

already provides reliable structure. This increases the denoising burden and makes the generation process unnecessarily difficult in simple regions. Moreover, because recovering accurate flow from pure Gaussian noise is inherently difficult, prior diffusion-based flow methods [17,30] often suffer from unstable training and slow convergence. Therefore, using the reliable predictions of a vanilla network as a prior can reduce the effective difficulty of diffusion-based flow generation, while enabling the diffusion model to bias refinement toward complex or uncertain regions where generative modeling is most beneficial, leading to more stable optimization and improved overall performance.

3.2 Confidence-Aware Network

The Confidence-Aware Network is designed to estimate simple flow efficiently, as shown in Fig. 3 (a). It takes the same input as standard optical flow models—two consecutive frames, but produces two outputs: *1) a rough dense optical flow field* F_r and *2) a pixel-wise confidence mask* M_{conf}. Among them, M_{conf} is defined on [0,1], where 0 indicates that F_r is entirely unreliable in the corresponding region, and 1 indicates that the corresponding prediction is highly reliable. *We refer to the ground truth for* F_r *and* M_{conf} *during training as* F_{gt} *and* M_{gt}*, respectively.*

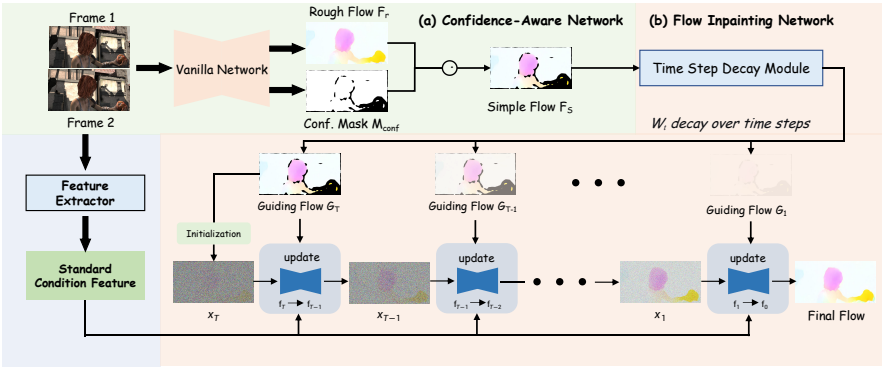


Fig. 3: An overview of our proposed framework **FlowPainter**, which consists of two main components: **(a) Confidence-Aware Network**: A lightweight PWC-Net-based model [34, 50] predicts a rough flow and a pixel-wise confidence mask indicating motion reliability. The confidence mask serves as a reliability gate, providing a simple-flow prior in reliable regions and suppressing unreliable predictions in challenging areas. **(b) Flow Inpainting Network**: The prior is first used for confidence-based initialization and is then injected as residual guidance through a Time Step Decay Module. The guidance is strong in early denoising steps to stabilize coarse structure, and gradually decays in later steps to preserve the denoiser's flexibility for detail refinement. We also condition denoising on features extracted from input frames to support accurate flow synthesis in challenging regions such as occlusions and motion boundaries.

Supervision of Confidence Mask. Architecturally, our Confidence-Aware Network follows a PWC-Net-style design [34, 50], but with substantially fewer layers and parameters. *The key challenge in training this model lies in how to construct a reliable M_{gt} to supervise M_{conf} during the training phase.* Conceptually, M_{gt} is intended to encode two complementary sources of unreliability: hard-flow regions and occluded regions. We denote the corresponding reliability maps as M_{hf} and M_{occ}, respectively. As discussed in Sec. 3.1, M_{hf} is characterized by large displacements and complex motion patterns. Due to its compact capacity and simplified structure, the Confidence-Aware Network typically cannot represent such motions well, and therefore should assign these regions lower confidence. M_{hf} is constructed as follows during training:

$$\begin{aligned} M_d &= Norm(MSE(F_{gt}, F_r)), \\ M_{hf} &= 1 - Er(M_d), \end{aligned} \tag{1}$$

where MSE denotes the mean squared error and $Norm$ represents normalization. M_d denotes the normalized difference map, and Er denotes an erosion operation implemented with a 7×7 kernel to conservatively shrink reliable regions around motion boundaries and uncertain borders. M_{occ} is defined as a reliability map for occlusions, where visible regions are assigned high confidence and occluded regions are assigned zero confidence. Occluded regions correspond to areas where motion causes correspondences between adjacent frames to be lost, violating the brightness constancy assumption that underlies optical flow computation. Prior work has shown that occlusions can significantly degrade flow accuracy [50]. Consequently, we explicitly force the predicted confidence values in occluded regions to be zero. We employ forward-backward consistency check [43] to obtain M_{occ} during the training phase. In summary, M_{gt} is constructed as follows while training:

$$M_{gt} = Min(M_{hf}, M_{occ}), \tag{2}$$

where Min represents taking the minimum value at each corresponding pixel location. This design ensures that either hard motion or occlusion leads to low confidence, making M_{conf} a reliability gate rather than a magnitude correction term.

General Training Strategy. To ensure stable optimization, we adopt a staged training strategy for the Confidence-Aware Network. We first train the network to convergence using only optical flow supervision and self-supervised objectives, ensuring that the network's output F_r is reasonably accurate. After the flow prediction stabilizes, we introduce the confidence-mask supervision and finally train both objectives jointly to convergence. Detailed training settings and implementation specifics are provided in Sec. 4.1.

3.3 Flow Inpainting Network

In our proposed Flow Inpainting Network, optical flow estimation is reformulated as a soft inpainting-style refinement problem conditioned on reliable simple-flow

regions, as shown in Fig. 3 (b). Existing diffusion-based optical flow approaches [17, 30] rely exclusively on image-derived features extracted from input frames as conditioning signals during denoising. This formulation implicitly requires the model to recover both simple and complex motion patterns from pure Gaussian noise, which makes optimization difficult and often leads to unstable training and slow convergence overall. Our key idea is to introduce the estimated simple flow as an additional structural prior, so that reliable regions provide useful motion structure and uncertain regions are left to be refined by image conditioning and diffusion denoising. Simultaneously, we adaptively adjust the guidance strength based on confidence masks and diffusion time steps, enhancing the model's early-stage stability and convergence while preserving its ability to refine details during the later stages of diffusion.

Time Step Decay Scheduling. Diffusion models operate under different noise levels across time steps: early steps are highly noisy and emphasize coarse global structure, whereas later steps refine fine details. Accordingly, the guidance strength should be stronger when noise is large to stabilize and accelerate convergence, and should weaken as denoising progresses to avoid over-constraining detail refinement. We implement this with a time-decay weight as follows:

$$W_t = (1.0 - t_{cur}/T) \cdot s, \tag{3}$$

where W_t represents the guidance weight of the current time step, T represents the predefined total number of diffusion sampling steps, t_{cur} denotes the current time step valued in {T, T - 1, ..., 0}, and s is the fixed guidance scale for simple optical flow, representing the upper limit of guidance intensity. Following FlowDiffuser [17], we set $T = 6$ in both training and inference. We set $s = 0.95$, slightly below 1, to prevent the simple-flow prior from becoming a hard constraint. This schedule decays the guidance as the process moves toward later steps, yielding a smooth transition in how strongly the base prior influences the updates.

Confidence-Based Initialization. We first construct a simple-flow prior by weighting the rough flow F_r with the confidence mask M_{conf}:

$$F_s = F_r \odot M_{conf}, \tag{4}$$

where F_s represents the simple-flow prior and $\odot$ denotes pixel-wise multiplication. Here, F_s is not intended to correct angular errors by simply shrinking the flow magnitude. Instead, M_{conf} serves as a reliability gate: high-confidence rough flow provides a structural prior, while low-confidence regions suppress the prior and rely more on image conditioning and diffusion refinement. We then initialize the diffusion process by filling high-confidence regions with the simple-flow prior:

$$\begin{aligned} B &= \mathbf{1}[M_{conf} > \tau], \\ x_T &= B \odot F_s + (1 - B) \odot \epsilon, \end{aligned} \tag{5}$$

where B denotes the binary initialization mask, $\tau = 0.5$ is the confidence threshold, and $\epsilon \sim \mathcal{N}(0, I)$ denotes Gaussian noise with the same spatial size as the

optical flow. Unlike hard masked denoising, we do not clamp the initialized high-confidence regions during subsequent denoising steps. This allows the denoiser to further correct the whole flow field when image evidence suggests necessary refinements.

Confidence-Gated Residual Guidance. After initialization, we inject the simple-flow prior through residual-form guidance at each denoising step. Given the current flow estimate $\hat{F}t$, we compute a confidence-gated residual:

$$\begin{aligned} G_t &= W_t \cdot F_s, \\ R_t &= G_t - W_t \cdot M_{conf} \odot \hat{F}_t, \end{aligned} \tag{6}$$

where G_t denotes the guiding flow and R_t encourages the current estimate to remain close to the reliable simple-flow prior in high-confidence regions, while having limited influence in uncertain regions. The residual guidance is then encoded by a lightweight convolutional encoder $\phi_g(\cdot)$ and fused into the denoiser feature by direct addition:

$$\tilde{H}_t = H_t + \phi_g(R_t), \tag{7}$$

where H_t denotes the denoiser feature at time step t, and $\tilde{H}_t$ denotes the guidance-enhanced feature. This soft residual injection avoids repeatedly adding the full prior to the denoising state, and is more flexible than hard-clamping known regions. As W_t decays over time steps, the effect of residual guidance gradually weakens, allowing the model to preserve early-stage stability while retaining late-stage flexibility for detail refinement.

Complementarity with Image Conditioning. While the simple-flow prior provides reliable motion structure in easy regions, optical flow estimation fundamentally depends on image content. We therefore retain the original feature extraction components used in diffusion-based flow models [17] to extract features from the input frames. These features encode semantic cues and pixel-wise matching relationships and are crucial for handling challenging conditions such as occlusions, textureless regions, and motion boundaries. Importantly, the guidance mechanism and image features act in a complementary manner: in simple regions, the simple-flow prior provides a stable structural anchor while image features support local refinement; in hard regions, the confidence mask suppresses unreliable prior guidance and the model relies primarily on image-derived features and diffusion refinement. This cooperative design enables the network to adaptively balance prior-driven constraints and data-driven corrections, leading to more stable training and improved final accuracy.

Unified Design for Training and Inference. We implement the same guidance strategy consistently during both training and inference. During training, one diffusion timestep is uniformly sampled for each image pair, and the corresponding confidence-gated residual guidance is computed and injected into the denoiser feature. This ensures that the denoising process is continually regularized by the simple-flow prior without imposing a hard constraint. During inference, we use the same number of sampling steps as FlowDiffuser [17], i.e., $T = 6$, and apply the identical confidence-based initialization and step-wise residual guidance.

4 Experiments

4.1 Implementation Details

Network Architecture. *1) Confidence-Aware Network:* Our Confidence-Aware Network adopts a PWC-Net-style design [34, 50]. We further incorporate several architectural adjustments inspired by MaskFlowNet [50]. Unlike MaskFlowNet [50], which learns an occlusion mask in an unsupervised manner, we introduce explicit supervision for the confidence mask. This is crucial because our mask is expected to reflect not only occluded regions but also hard-flow regions. As shown in Fig. 4, explicit supervision helps the network more clearly distinguish regions where the rough flow is reliable, whereas unsupervised learning tends to produce overly conservative, medium-confidence predictions. Such more faithful confidence estimation provides a more informative reliability prior for the downstream inpainting-based refinement. *2) Flow Inpainting Network:* Given the strong performance and flow-specific design of FlowDiffuser [17], we adopt FlowDiffuser [17] as the backbone for the diffusion component. Based on this backbone, we modify the initialization strategy and introduce a time step decay module to inject additional conditioning from the simple-flow prior. This reformulates the diffusion process from full dense-flow generation into a soft inpainting-style refinement process, where the model is guided by reliable flow in simple regions and mainly refines uncertain regions.

Fig. 4: Comparison of output results when training Confidence-Aware Network with or without explicit supervision of the confidence mask, where white denotes high confidence and black denotes low confidence.

Training Details. *1) Training of Confidence-Aware Network.* We train the Confidence-Aware Network in two stages. In the first stage, we pre-train on FlyingChairs [7] and FlyingThings3D [20], using supervision only for flow prediction. The loss consists of an EPE loss and a self-supervised warping loss, weighted by 0.1 and 0.01, respectively. This stage equips the network with the ability to produce stable and reasonable flow estimates. In the second stage, we fine-tune on Sintel [1] (S+T) and KITTI [23] while introducing supervision for the confidence mask. The mask construction is described in Sec. 3.2. We supervise the confidence mask using L_1 loss and set its weight to 0.8. *2) Training of Flow Inpainting Network.* Following prior optical flow pipelines [17, 36], FlowPainter is trained sequentially on FlyingChairs [7], FlyingThings3D [20], Sintel [1], and

KITTI [23]. Although SEA-RAFT [40] shows the benefit of large-scale rigid-flow pretraining for iterative refinement paradigms, our current setting does not use such pretraining. We empirically find that the confidence-aware prior and the diffusion backbone together enable competitive performance under this training protocol. A practical benefit of introducing the confidence-aware prior is that the diffusion stage requires substantially fewer training iterations than FlowDiffuser [17]. Under the same batch size setting, FlowDiffuser [17] typically needs 100k/200k/180k/50k iterations to converge on FlyingChairs [7]/FlyingThings3D [20]/Sintel [1]/KITTI [23], respectively. In contrast, FlowPainter converges with 20k/50k/50k/15k iterations on the same datasets. With 8×A100 GPUs, this reduces the diffusion-stage training time from 7 days 2 hours to 2 days 1 hour, while also improving the final accuracy.

Table 2: Quantitative comparison of FlowPainter with state-of-the-art optical flow methods on the train splits. The best results are **highlighted in bold.**

Method	Extra Data	Sintel(train)		KITTI-15(train)		Spring(train)
		Clean	Final	EPE	Fl-all	EPE
SEA-RAFT [40]	Tartan, Spring	-	-	-	-	0.41
FlowDiffuser [17]	Spring	-	-	-	-	0.43
FlowPainter(Ours)	Spring	-	-	-	-	**0.40**
RAFT [36]	-	1.43	2.71	5.04	17.4	0.45
GMA [15]	-	1.30	2.74	4.69	17.1	0.44
GMFlow [43]	-	1.08	2.48	7.75	23.2	0.93
FlowFormer [12]	-	0.99	2.38	4.11	14.6	0.47
GAFlow [18]	-	0.95	2.33	3.92	13.9	0.46
GMFlow+ [44]	-	0.91	2.74	5.74	17.6	0.43
FlowFormer++ [32]	YouTube-VOS	0.90	2.30	3.93	14.1	0.45
MatchFlow [4]	MegaDepth	1.03	2.45	4.08	15.6	0.41
SEA-RAFT [40]	Tartan	1.18	4.13	3.62	12.8	-
FlowSeek [27]	Tartan	1.04	2.18	3.34	11.3	**0.40**
FlowDiffuser [17]	-	0.88	2.23	3.75	11.1	-
FlowPainter(Ours)	-	**0.87**	**1.32**	**3.17**	**6.84**	-

4.2 Evaluation Results

Quantitative Evaluation. We conduct quantitative evaluations on Sintel [1], KITTI [23] and Spring [21] benchmarks. The results are summarized in Tab. 2 and Tab. 3. Under comparable training settings, FlowPainter achieves the best results on most major metrics in the listed comparison, while using no additional training data in the Sintel [1] and KITTI [23] setting. On Sintel(train), FlowPainter achieves 0.87/1.32 EPE on the clean/final passes, and also obtains strong results on Sintel(test) with 1.01/1.71 EPE. Compared with the diffusion-based baseline FlowDiffuser [17], the improvement is modest on the clean pass

Table 3: Quantitative comparison of FlowPainter with state-of-the-art optical flow methods on the test splits. The best results are **highlighted in bold.**

Method	Sintel(test)		KITTI-15(test)		
	Clean	Final	Fl-bg	Fl-fg	Fl-all
RAFT [36]	1.61	2.86	4.74	6.87	5.10
GMFlow [43]	1.74	2.90	9.67	7.57	9.32
FlowFormer [12]	1.16	2.09	4.37	6.18	4.68
GMFlow+ [44]	1.03	2.37	4.27	5.60	4.49
MatchFlow [4]	1.16	2.37	4.33	6.11	4.63
SEA-RAFT [40]	1.31	2.60	-	-	-
FlowDiffuser [17]	1.02	2.03	3.68	6.64	4.17
DPFlow [24]	1.05	1.98	3.29	**4.93**	3.56
MemFlow [5]	1.05	1.91	3.67	6.27	4.10
MegaFlow [45]	1.49	2.43	3.55	5.89	3.94
DDVM [30]	1.75	2.48	2.90	5.05	3.26
FlowPainter(Ours)	**1.01**	**1.71**	**2.35**	6.37	**3.02**

(1% on train; 1% on test), but becomes more pronounced on the final pass (40% on train; 16% on test). Since the Sintel final pass contains degradations such as motion blur, defocus, atmospheric effects, and complex illumination changes, this larger improvement suggests that FlowPainter is more robust under the combined challenging conditions represented by this benchmark split, without implying isolated factor-wise verification for each degradation type. On KITTI benchmark, FlowPainter also improves over FlowDiffuser [17], reducing EPE by 15% and Fl-all by 39% on the train split relative to FlowDiffuser [17]. On the test split, FlowPainter also achieves the best Fl-all results among the listed methods. These results suggest that FlowPainter generalizes well beyond Sintel and remains competitive in real-world driving scenarios such as occlusions, large displacements, and thin structures. For Spring [21], we additionally report results under Spring-based training: FlowPainter achieves 0.40 EPE, outperforming both SEA-RAFT [40] (0.41 EPE) and FlowDiffuser [17] (0.43 EPE), suggesting that the proposed confidence-guided inpainting formulation is also effective under different motion statistics.

Qualitative Evaluation. As illustrated in Fig. 5, we compare flow visualizations of FlowPainter on Sintel(test) [1] against SEA-RAFT [40] and FlowDiffuser [17]. The examples show that diffusion-based methods can produce detailed flow estimates in complex scenes and handle irregular motions more effectively than purely discriminative iterative refinement in some cases. FlowPainter further improves the visual quality by using reliable simple-flow priors while preserving the refinement capability of the diffusion backbone in hard regions.

4.3 Ablation Studies

Confidence Mask Construction Strategy. The reliability of the confidence mask predicted by the Confidence-Aware Network is crucial for downstream

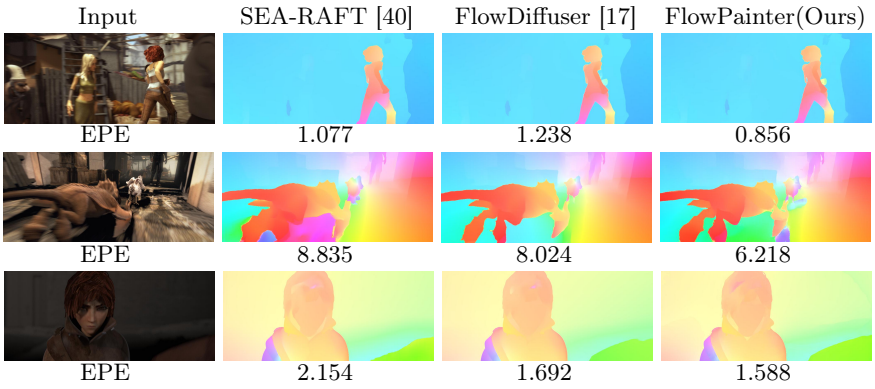


Fig. 5: Qualitative comparison of FlowPainter with SEA-RAFT [40] and FlowDiffuser [17] on Sintel(test).

inpainting, as it determines where the diffusion model should rely more on image conditioning and denoising refinement. We therefore ablate the construction of the confidence mask during training. As shown in Tab. 4, both the occlusion mask and the hard-flow mask contribute to the final performance. Discarding either signal leads to a clear drop in accuracy, suggesting that they capture complementary sources of unreliability: occlusions violate the correspondence assumptions underlying flow estimation, whereas hard-flow regions are difficult for the lightweight base network to estimate reliably. Explicitly supervising both aspects yields a sharper and more informative confidence mask, which in turn enables more effective guidance and region-adaptive refinement.

Table 4: Ablation on occlusion mask and hard-flow mask, where O.M. denotes occlusion mask and H.M. denotes hard-flow mask.

O.M.	H.M.	Sintel(train) Clean/Final	KITTI(train) EPE/Fl-all
✗	✗	1.05/2.17	3.69/8.92
✗	✓	0.88/1.37	3.22/7.01
✓	✗	1.01/1.95	3.47/8.13
✓	✓	**0.87/1.32**	**3.17/6.84**

Table 5: Ablation on time step decay module and inpainting strategy, where T.D. denotes time step decay module and I.P. denotes inpainting strategy.

T.D.	I.P.	Sintel(train) Clean/Final	KITTI(train) EPE/Fl-all
✗	✗	1.23/2.74	4.52/12.18
✗	✓	1.11/2.32	3.98/10.12
✓	✗	0.95/1.67	3.49/8.97
✓	✓	**0.87/1.32**	**3.17/6.84**

Time Step Decay Module and Inpainting Strategy. We conduct ablations on the time step decay module and the inpainting formulation within the Flow Inpainting Network. Here, removing the inpainting strategy means removing the confidence-mask-based inpainting guidance. As shown in Tab. 5, removing either component consistently degrades performance, showing that the two components are complementary in the proposed refinement pipeline. In particular, the time step decay module provides a schedule for controlling the strength of the simple-flow prior across diffusion steps, which stabilizes optimization in the high-

noise regime while avoiding over-constraint during late-stage detail refinement. Meanwhile, the inpainting formulation reduces the denoising burden in reliable simple regions and biases refinement toward uncertain hard regions, leading to more accurate and robust estimates especially in challenging cases.

Table 6: Ablation study result for simple flow initialization thresholds.

Threshold		0.3	0.4	0.5	0.6	0.7
Sintel(train)	Clean	0.95	0.88	**0.87**	0.91	0.97
	Final	1.78	1.57	**1.32**	1.63	1.85
KITTI(train)	EPE	3.76	3.37	**3.17**	3.35	3.98
	Fl-all	8.54	7.93	**6.84**	7.88	8.71

Initialization Threshold for Simple Flow. When the confidence value exceeds a preset threshold, the corresponding simple-flow prediction is used to initialize the diffusion process. This threshold controls the trade-off between the quantity and reliability of injected prior information. If the threshold is set too low, inaccurate base-flow estimates may be introduced into the initialization, increasing the correction burden and potentially biasing refinement in uncertain regions. Conversely, an overly conservative threshold limits the usable prior, weakening the benefit of guided inpainting and leading to slower convergence and reduced accuracy. We evaluate thresholds 0.3, 0.4, 0.5, 0.6, 0.7; the results in Tab. 6 show that 0.5 achieves the best trade-off among the tested thresholds.

5 Conclusion

We proposed *FlowPainter*, a confidence-guided diffusion framework that integrates reliable lightweight flow priors with diffusion-based refinement for optical flow estimation. By using confidence-based initialization and time-decayed residual guidance, FlowPainter reformulates dense-flow generation from noise as a soft inpainting-style refinement problem, reducing the denoising burden in reliable regions while preserving flexibility for uncertain hard regions. Experiments on Sintel, KITTI, and Spring show that FlowPainter achieves strong accuracy under comparable training settings and improves convergence efficiency over existing diffusion-based optical flow methods.

6 Acknowledgement

This work is supported in part by the Tianjin Natural Science Foundation Project (25ZXRGGX00290, 24JCJQJC00020, 25JCQNJC01390), the National Natural Science Foundation of China (62306153, 62225604), the Young Elite Scientists Sponsorship Program by CAST (YESS20240686), and the Fundamental Research Funds for the Central Universities (Nankai University, 63253223, 63253219). This work was also supported by the OPPO Research Fund.

References

1. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: ECCV. pp. 611–625 (2012)
2. Capito, L., Ozguner, U., Redmill, K.: Optical flow based visual potential field for autonomous driving. In: IV. pp. 885–891 (2020)
3. Chen, H., Zhang, X., Sun, X., Mingqing, X.: Calibrated harmonic overlaid implicit neural representations for multi-dimensional data. arXiv preprint arXiv:2606.26763 (2026)
4. Dong, Q., Cao, C., Fu, Y.: Rethinking optical flow from geometric matching consistent perspective. In: CVPR (2023)
5. Dong, Q., Fu, Y.: Memflow: Optical flow estimation and prediction with memory. In: CVPR. pp. 19068–19078 (2024)
6. Dong, Q., Zhao, B., Fu, Y.: Open-ddvm: A reproduction and extension of diffusion model for optical flow estimation. arXiv preprint arXiv:2312.01746 (2023)
7. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: ICCV. pp. 2758–2766 (2015)
8. Fan, L., Huang, W., Gan, C., Ermon, S., Gong, B., Huang, J.: End-to-end learning of motion representation for video understanding. In: CVPR. pp. 6016–6025 (2018)
9. Gao, S., Liu, X., Zeng, B., Xu, S., Li, Y., Luo, X., Liu, J., Zhen, X., Zhang, B.: Implicit diffusion models for continuous super-resolution. In: CVPR. pp. 10021–10030 (2023)
10. Hai, Y., Wang, G., Su, T., Jiang, W., Hu, Y.: Hierarchical flow diffusion for efficient frame interpolation. In: CVPR. pp. 22943–22952 (2025)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. pp. 6840–6851 (2020)
12. Huang, Z., Shi, X., Zhang, C., Wang, Q., Cheung, K.C., Qin, H., Dai, J., Li, H.: Flowformer: A transformer architecture for optical flow. In: ECCV. pp. 668–685. Springer (2022)
13. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: CVPR. pp. 2462–2470 (2017)
14. Ji, Y., Chen, Z., Xie, E., Hong, L., Liu, X., Liu, Z., Lu, T., Li, Z., Luo, P.: Ddp: Diffusion model for dense visual prediction. In: ICCV. pp. 21741–21752 (2023)
15. Jiang, S., Campbell, D., Lu, Y., Li, H., Hartley, R.: Learning to estimate hidden motions with global motion aggregation. In: ICCV. pp. 9772–9781 (2021)
16. Liu, R., Sun, S., Tang, H., Gao, W., Li, G.: Flow4agent: Long-form video understanding via motion prior from optical flow. In: ICCV. pp. 23817–23827 (2025)
17. Luo, A., Li, X., Yang, F., Liu, J., Fan, H., Liu, S.: Flowdiffuser: Advancing optical flow estimation with diffusion models. In: CVPR. pp. 19167–19176 (2024)
18. Luo, A., Yang, F., Li, X., Nie, L., Lin, C., Fan, H., Liu, S.: Gaflow: Incorporating gaussian attention into optical flow. In: ICCV. pp. 9642–9651 (2023)
19. Mahjourian, R., Kim, J., Chai, Y., Tan, M., Sapp, B., Anguelov, D.: Occupancy flow fields for motion forecasting in autonomous driving. RA-L **7**, 5639–5646 (2022)
20. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: CVPR. pp. 4040–4048 (2016)
21. Mehl, L., Schmalfuss, J., Jahedi, A., Nalivayko, Y., Bruhn, A.: Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In: CVPR (2023)

22. Meng, Y., Jin, X., Lei, L., Guo, C.L., Li, C.: Ultraled: Learning to see everything in ultra-high dynamic range scenes. In: NeurIPS. pp. 41466–41495 (2025)
23. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR (2015)
24. Morimitsu, H., Zhu, X., Cesar, R.M., Ji, X., Yin, X.C.: Dpflow: Adaptive optical flow estimation with a dual-pyramid framework. In: CVPR. pp. 17810–17820 (2025)
25. Pepe, A., Dos Santos Mendonca, P., Lasenby, J.: Geometric inductive priors in diffusion-based optical flow estimation. In: ICCV. pp. 655–665 (2025)
26. Piergiovanni, A., Ryoo, M.S.: Representation flow for action recognition. In: CVPR. pp. 9945–9953 (2019)
27. Poggi, M., Tosi, F.: Flowseek: optical flow made easier with depth foundation models and motion bases. In: ICCV. pp. 5667–5679 (2025)
28. Rakêt, L.L., Roholm, L., Bruhn, A., Weickert, J.: Motion compensated frame interpolation with a symmetric optical flow constraint. In: ISVC. pp. 447–457 (2012)
29. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: CVPR. pp. 4161–4170 (2017)
30. Saxena, S., Herrmann, C., Hur, J., Kar, A., Norouzi, M., Sun, D., Fleet, D.J.: The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. In: NeurIPS. pp. 39443–39469 (2023)
31. Shen, S., Kerofsky, L., Yogamani, S.: Optical flow for autonomous driving: Applications, challenges and improvements. arXiv preprint arXiv:2301.04422 (2023)
32. Shi, X., Huang, Z., Li, D., Zhang, M., Cheung, K.C., See, S., Qin, H., Dai, J., Li, H.: Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In: CVPR. pp. 1599–1610 (2023)
33. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
34. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: CVPR. pp. 8934–8943 (2018)
35. Sun, S., Kuang, Z., Sheng, L., Ouyang, W., Zhang, W.: Optical flow guided feature: A fast and robust motion representation for video action recognition. In: CVPR. pp. 1390–1399 (2018)
36. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV. pp. 402–419 (2020)
37. Teed, Z., Deng, J.: Raft-3d: Scene flow using rigid-motion embeddings. In: CVPR. pp. 8375–8384 (2021)
38. Wang, C., Peng, H.Y., Liu, Y.T., Gu, J., Hu, S.M.: Diffusion models for 3d generation: A survey. Computational Visual Media **11**(1), 1–28 (2025)
39. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. IJCV **132**(12), 5929–5949 (2024)
40. Wang, Y., Lipson, L., Deng, J.: Sea-raft: Simple, efficient, accurate raft for optical flow. In: ECCV. pp. 36–54 (2024)
41. Wu, S., You, K., He, W., Yang, C., Tian, Y., Wang, Y., Zhang, Z., Liao, J.: Video interpolation by event-driven anisotropic adjustment of optical flow. In: ECCV. pp. 267–283 (2022)
42. Wulff, J., Butler, D.J., Stanley, G.B., Black, M.J.: Lessons and insights from creating a synthetic optical flow benchmark. In: ECCV. pp. 168–177 (2012)
43. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Tao, D.: Gmflow: Learning optical flow via global matching. In: CVPR. pp. 8121–8130 (2022)
44. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Yu, F., Tao, D., Geiger, A.: Unifying flow, stereo and depth estimation. IEEE TPAMI **45**(11), 13941–13958 (2023)

45. Zhang, D., Wang, F., Pollefeys, M., Xu, H.: Megaflow: Zero-shot large displacement optical flow. arXiv preprint arXiv:2603.25739 (2026)
46. Zhang, F., Woodford, O.J., Prisacariu, V.A., Torr, P.H.: Separable flow: Learning motion cost volumes for optical flow estimation. In: ICCV. pp. 10807–10817 (2021)
47. Zhang, L., Wang, Z., Zhang, Q., Qiu, Q., Pang, A., Jiang, H., Yang, W., Xu, L., Yu, J.: Clay: A controllable large-scale generative model for creating high-quality 3d assets. TOG **43**(4), 1–20 (2024)
48. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023)
49. Zhao, Q., Li, Y., Sun, Q., Yan, Z.: Resilphase: Plug-and-play phase mapping and noise-resilient macro-trajectory extrapolation for diffusion acceleration. arXiv preprint arXiv:2606.26769 (2026)
50. Zhao, S., Sheng, Y., Dong, Y., Chang, E.I., Xu, Y., et al.: Maskflownet: Asymmetric feature matching with learnable occlusion mask. In: CVPR. pp. 6278–6287 (2020)