FlowDiffuser: Advancing Optical Flow Estimation with Diffusion Models

Ao Luo[1,2], Xin Li[3], Fan Yang[3], Jiangyu Liu[2], Haoqiang Fan[2], and Shuaicheng Liu[4,2*]
[1]Southwest Jiaotong University [2]Megvii Technology [3]G42
[4]University of Electronic Science and Technology of China

Abstract

Optical flow estimation, a process of predicting pixel-wise displacement between consecutive frames, has commonly been approached as a regression task in the age of deep learning. Despite notable advancements, this de facto paradigm unfortunately falls short in generalization performance when trained on synthetic or constrained data. Pioneering a paradigm shift, we reformulate optical flow estimation as a conditional flow generation challenge, unveiling FlowDiffuser *— a new family of optical flow models that could have stronger learning and generalization capabilities. FlowDiffuser estimates optical flow through a 'noise-to-flow' strategy, progressively eliminating noise from randomly generated flows conditioned on the provided pairs. To optimize accuracy and efficiency, our FlowDiffuser incorporates a novel Conditional Recurrent Denoising Decoder (Conditional-RDD), streamlining the flow estimation process. It incorporates a unique Hidden State Denoising (HSD) paradigm, effectively leveraging the information from previous time steps. Moreover, FlowDiffuser can be easily integrated into existing flow networks, leading to significant improvements in performance metrics compared to conventional implementations. Experiments on challenging benchmarks, including Sintel and KITTI, demonstrate the effectiveness of our FlowDiffuser with superior performance to existing state-of-the-art models. Code is available at https://github.com/LA30/FlowDiffuser.*

1. Introduction

Optical flow estimation remains a pivotal research domain, with its significance highlighted by a wide range of critical applications. This task aims to establish per-pixel correspondences between consecutive frames, resulting in a two-dimensional vector field that illustrates pixel displacement. In the contemporary deep learning paradigm, this challenge is predominantly framed as a regression task: neural models are trained to infer the optical flow vectors directly from sequential image data [17, 36, 41, 49, 50]. Starting with the groundbreaking end-to-end model FlowNet [12], the research community has actively sought improvements along two main dimensions: **i)** the augmentation of feature encoding capabilities via more robust backbones, transitioning from ResNet to Transformer, to extract increasingly discriminative feature representations [8, 45, 51]; and **ii)** the integration of advanced decoding techniques, such as the recurrent scheme with 4D cost volume [41] or transformer-based motion aggregation [19, 25] and cost modeling [15], to refine the regression accuracy. Fundamentally, current studies are predominantly characterized by a shared focus on devising effective techniques that improve the correspondence mapping between sequential frames and their resultant optical flow vectors.

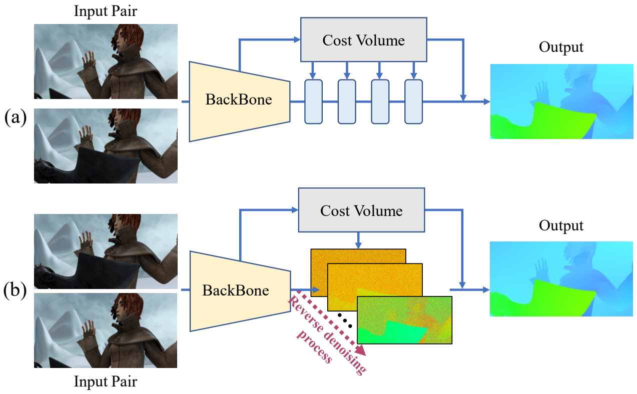

Figure 1. **Idea illustration**. Conventional optical flow models (a) map frames directly to flow fields. In contrast, *FlowDiffuser* (b) generates flow through a reverse denoising process.

Notwithstanding the notable performance of the established paradigm in optical flow estimation, it is burdened by intrinsic constraints due to its discriminative nature. **Firstly**, prevalent models fail to explicitly capture the underlying probabilistic distribution that characterizes the dense correspondence field of the flow, consequently struggling with complex motion dynamics. **Secondly**, challenges like occlusion, motion blur, and brightness changes hinder discriminative models in optical flow tasks. These

[*]Corresponding author

issues, prevalent in real-world scenarios, obscure vital details, affecting accuracy. Generative models, proficient in learning comprehensive joint probability distributions, provide robust solutions to these challenges, thereby improving performance across various scenarios. Despite Saxena *et al.* [31] attempts to mitigate discriminative model limitations via a generative approach, their reliance on a standard diffusion model framework significantly increases computational demands. Furthermore, this method lacks a customized design for optical flow systems, limiting its efficacy in scenarios demanding swift and precise flow calculations. These shortcomings not only compromise the dependability of existing models but also restrict their ability to make extrapolations. In light of these insights, a pressing question emerges: *Is there a more efficacious paradigm, specifically designed for optical flow estimation, that transcends conventional methodologies?*

To answer the above question, we introduce *FlowDiffuser*—the first denoising diffusion model specifically designed for optical flow estimation. Fig. 1 delineates the paradigmatic shift from conventional discriminative mappings to a generative model that conditions the flow estimation on sequential frame pairs. Distinct from prior approaches, *FlowDiffuser* initiates with a perturbed flow field and engages in a progressive denoising diffusion sequence. During training, Gaussian noise is methodically integrated into the original optical flow to synthesize initial noisy fields. *FlowDiffuser* is designed to progressively reduce the noise, conditioned on the matching cost and contextual features, thereby cultivating the model's capacity to deduce the flow field from stochastic beginnings. In the inference phase, *FlowDiffuser* reverses the diffusion trajectory to unveil the flow field. This approach follows popular denoising techniques in generative modeling [7, 9, 34, 43] and further leverages frame-pair data to guide flow creation, a novel effort in this field to our knowledge.

Inspired by RAFT paradigm [41], our model incorporates a similar recurrent update mechanism into our diffusion approach, designed specifically for optical flow estimation. To elaborate further, we integrate a novel Conditional Recurrent Denoising Decoder (Conditional-RDD) into our flow estimation mechanism, substantially enhancing both reliability and computational efficiency. The distinctive feature of Conditional-RDD lies in its ability to predict and utilize the intermediate (hidden) states between the initial noisy flow and the target, diverging from typical diffusion models that attempt to predict the full flow length. This approach, named Hidden State Denoising (HSD) and crafted for optical flow, further enables precise motion estimation.

Our proposed *FlowDiffuser* adopts a generative scheme towards optical flow estimation, inheriting the strengths inherent to generative models. A key benefit is its stronger capability in understanding the underlying structure of data, giving *FlowDiffuser* an improved ability to handle complex motion patterns, outperforming conventional methods. Crucially, by explicitly defining the stochastic process, *FlowDiffuser* is required to explore various trajectories within the latent space, potentially enhancing its generalizability. This characteristic, combined with the noise-conditioned diffusion process's inherent prevention of overfitting to training data, is especially remarkable. We validate the effectiveness of *FlowDiffuser* through extensive testing on challenging benchmarks like Sintel and KITTI, where it demonstrates marked enhancements over contemporary state-of-the-art models. The major **contributions** of this work are summarized as follows:

- **A novel generative approach for optical flow estimation.** *FlowDiffuser* stands as one of the pioneering generative models explicitly devised for optical flow estimation, integrating the strengths of conventional flow models into the diffusion modeling framework. This work not only offers a novel perspective in the field of optical flow estimation but also lays the groundwork for future explorations in this field.
- **A tailored Conditional-RDD for optical flow estimation.** Our model, inspired by the RAFT [41], pioneers a Conditional Recurrent Denoising Decoder (Conditional-RDD) specifically designed for optical flow estimation. Integrating the Conditional-RDD and Hidden State Denoising (HSD) paradigm, our *FlowDiffuser* achieves improved efficiency and precision in optical flow prediction, advancing the state-of-the-art in this field.
- **State-of-the-art results on widely-used benchmarks.** Our *FlowDiffuser* can accurately estimate optical flow in challenging scenarios, and demonstrates leading-edge performance on both the Sintel and KITTI benchmarks.

2. Related Work

Optical Flow. The field of optical flow estimation has witnessed remarkable advancements with the advent of deep neural networks, enabling the creation of complex mappings between video frames and flow vectors. Initially, methods like those in [2, 12] employed an encoder-decoder architecture to directly translate video frames into flow fields. Subsequent research has further harnessed the power of deep learning, either by refining the flow encoder for more distinct representations [17, 36, 49, 50] or by integrating correlation data into the flow decoder to enhance regression performance [36, 41]. Recent developments have seen the integration of graph techniques [26], attention mechanisms [19, 25], iterative refinement [16, 17, 36], and holistic motion analysis [19, 26] into optical flow models. In the same timeframe as our research, Saxena *et al* [31] explored a generative strategy to address the drawbacks of discriminative models in optical flow prediction. However, their approach, grounded in standard diffusion model paradigms, incurs heightened computational costs. This methodology, not being specifically tailored for optical flow tasks, may compromise efficiency and accuracy in scenarios where detailed and rapid flow estimation are critical.

Diffusion Models. Diffusion models, a subset of generative models, methodically learn the true data distribution through iterative denoising [9, 14]. In computer vision, their success in image and video generation, as well as synthesis, is well-documented [7, 9, 34, 43]. Recent forays include applications in semantic segmentation [3, 40], instance segmentation [13], object detection [5], homography estimation [22], and even 3D vision [1, 21, 29]. In this work, we uniquely harness diffusion models to revolutionize optical flow estimation. Our approach introduces a groundbreaking paradigm specifically tailored for optical flow challenges. This includes the development of a custom module, the Conditional-RDD, designed to optimize flow estimation precision and efficiency. By adapting diffusion models to the nuanced requirements of optical flow, we aim to substantially improve both the accuracy and generalization capacity of existing optical flow models.

3. Method

3.1. Preliminaries

Motivation & Objectives. While the UNet-based denoising diffusion process is effective, it requires significant computational resources, which is suboptimal for vision tasks such as optical flow estimation. Recent work [31] has improved UNet's efficiency, yet their DDVM remains significantly slower (by about 20 to 30 times) compared to discriminative models. Given this, we question the suitability of the standard diffusion paradigm for optical flow estimation. The prevalent use of conventional modules and architectures in current optical flow research underlines their effectiveness in enhancing performance and efficiency. This highlights the need to incorporate these task-specific designs into the diffusion framework.

Our aim is to design a model that integrates RAFT-based models [19, 39, 41] with generative techniques, achieving an equilibrium between the efficiency of conventional models and the advanced capabilities of generative approaches.

Problem Formulation. Optical flow estimation aims to discern pixel-level movement between two consecutive images, I_1 and I_2, by producing a flow field $\mathbf{f}$. This field traces each pixel's shift from I_1 to I_2, usually treated as a regression problem where a neural network predicts the flow as $\mathbf{f} = \mathrm{F}_\Phi(\mathrm{I}_1, \mathrm{I}_2)$, with Φ denoting network parameters.

In this work, we draw inspiration from the proven effectiveness of denoising diffusion probabilistic models, as detailed in [14, 33], to redefine optical flow estimation as a generative process. The new approach, named *FlowDiffuser* and symbolized as P_Θ, utilizes learnable weights Θ to methodically transform a noisy flow field $\mathbf{f}_n$ into a refined data sample $\mathbf{f}_0$. This transformation process gradually removes noise from $\mathbf{f}_n$, under the conditional influence of the input images: $\mathbf{f}_0 = \mathrm{P}_\Theta(\mathbf{f}_n|\mathrm{I}_1, \mathrm{I}_2)$. Furthermore, this process is dynamically enhanced by our proposed Conditional-RDD, which iteratively updates the estimation during the transformation sequence. It's noteworthy that during training, $\mathbf{f}_n$ is derived from the ground truth, whereas in the inference stage, it is randomly generated following a standard Gaussian distribution, in alignment with the methodologies presented in [14, 33].

3.2. Diffusion Model for Optical Flow

Overview. Fig. 2 depicts the primary steps and overall architecture of our *FlowDiffuser*. Drawing inspiration from the RAFT architecture, our model focuses on reverse denoising through a novel Conditional Recurrent Denoising Decoder (Conditional-RDD) for flow estimation. Particularly, consistent with established models in the literature [25, 39, 41], our approach processes the input image pair $(\mathrm{I}_1, \mathrm{I}_2)$ using dual encoders. This generates basic features $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and a context feature $\boldsymbol{x}_c$. We then construct a 4D correlation volume $\boldsymbol{x}_{cv}$ from $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ through dot-product operations. Yet, diverging from the conventional paradigm, our flow decoder is restructured by combining a unique conditional denoising process with the RAFT's recurrent learning mechanism. The specific enhancements are outlined as follows: **i)** Our model's flow decoder, starting with a noisy flow $\mathbf{f}_t$, refines it to $\mathbf{f}_{t-1}$ in one denoising step. This process is guided by the encoded features $\boldsymbol{x}_c$, $\boldsymbol{x}_{cv}$, and the hidden feature $\boldsymbol{x}_h$ of RAFT's GRU module, formulated as $\mathbf{f}_{t-1} = \mathbb{P}_\theta(\mathbf{f}_t \,|\, \boldsymbol{x}_{cv}, \boldsymbol{x}_c, \boldsymbol{x}_h)$. **ii)** Our model includes a time embedding function to synchronize diffusion iterations with their respective timesteps. This alignment enhances the decoder's ability to differentiate noise variations across various time points. **iii)** Our approach integrates the Hidden State Denoising (HSD) paradigm within Conditional-RDD, enhancing stability and efficiency in the inference phase.

Conditional Recurrent Denoising Decoder. The proposed Conditional-RDD, aligning with RAFT's decoding pipeline, commences by using the input flow for a correlation pyramid lookup. This step leads to the creation of 2D motion features, derived from encoding the matched costs with a motion encoder $\mathcal{F}_{\mathrm{ME}}(\cdot)$, formulated as $\boldsymbol{x}_m = \mathcal{F}_{\mathrm{ME}}(\boldsymbol{x}_{cv}, \mathbf{f})$. Subsequent processing involves state updating via a GRU, followed by the utilization of a flow head for the final flow field prediction.

The primary challenge in our approach involves effectively handling the noisy input flow $\mathbf{f}_t$ using specialized modules. A critical element is the time embedding function $\mathcal{T}(\cdot)$, essential for synchronizing the diffusion process iterations with their respective timesteps. This is denoted as $\boldsymbol{e}_t = \mathcal{T}(t)$, where $\boldsymbol{e}_t$ signifies the time-embedded features,

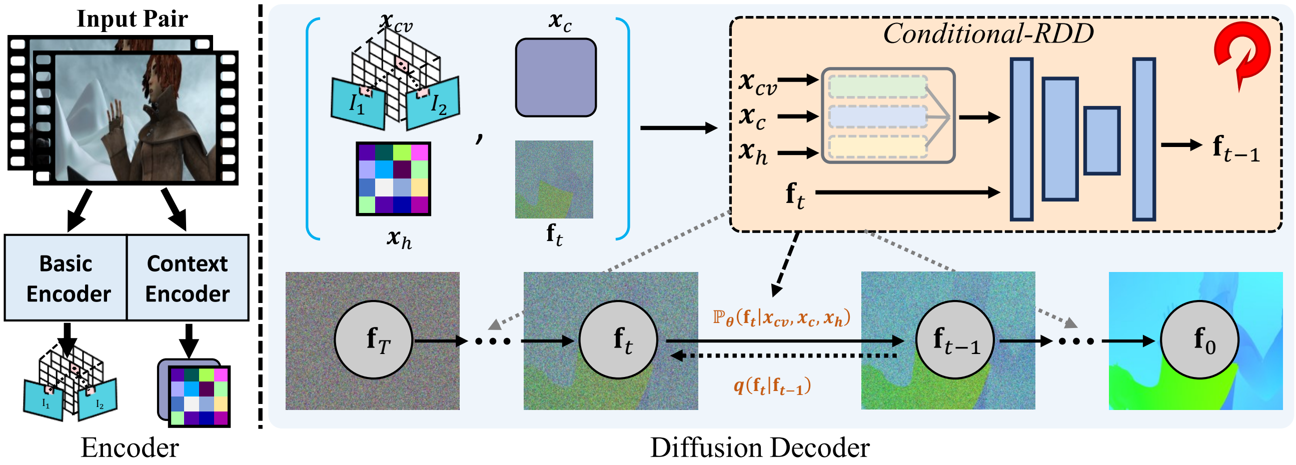

Figure 2. **Overview of our *FlowDiffuser*.** Rather than using the discrimination learning paradigm, our framework takes a generative approach to ensure reliability and generalizability. The notations $\boldsymbol{x}_{cv}$, $\boldsymbol{x}_c$, and $\boldsymbol{x}_h$ denote the constructed cost-volume, the extracted context feature, and the hidden feature of the RAFT-like decoder, respectively. Best viewed in color.

further divided into a scale vector $\boldsymbol{e}_{sc}$ and a shift vector $\boldsymbol{e}_{sh}$. Diverging from conventional methods that integrate scale and shift vectors into the feature map via simple multiplication and concatenation/summation, our model introduces an Embedding Enhancement (EE) module. This module is designed to significantly augment the impact of time embeddings, which is formulated as:

$$\boldsymbol{x}_a = \mathcal{A}(\boldsymbol{x}_c, \boldsymbol{x}_m), \tag{1}$$

$$\boldsymbol{x}_e = \mathcal{C}_1(\boldsymbol{x}_a)\,\boldsymbol{e}_{sc} + \boldsymbol{e}_{sh}, \tag{2}$$

$$\boldsymbol{x}_o = \mathcal{C}_2(\boldsymbol{x}_e)\,\tau + \boldsymbol{x}_a, \tag{3}$$

where $\mathcal{A}(\cdot)$ indicates the attentative motion aggregation operations like GMA [19] and KPA [25], and $\mathcal{C}(\cdot)$ represents the standard convolutional blocks containing 3×3 convolution, GELU activation function and group normalization. τ signifies a learnable weight.

The EE module enriches the aggregated motion feature $\boldsymbol{x}_a$ with time embeddings, thereby enhancing the denoising diffusion process. The generated $\boldsymbol{x}_o$ is then fed into the GRU and flow head modules. This approach is adaptable to various RAFT-like decoders, including GMA [19], KPA-Flow [25], SKFlow [39], etc. The use of more advanced decoders is expected to further improve performance.

Forward Diffusion. During the network training phase, we employ a diffusion forward process to create noisy flow fields derived from the ground truth. This involves using a Markovian chain which incrementally adds Gaussian noise to the data sample. The process is defined as follows:

$$q(\mathbf{f}_t|\mathbf{f}_0) = \mathcal{N}(\mathbf{f}_t|\sqrt{\bar{\alpha}_t}\,\mathbf{f}_0, (1-\bar{\alpha}_t)\boldsymbol{I}), \tag{4}$$

where $\mathbf{f}_0$ indicates the orginal data sample (*i.e.*, the ground-truth of optical flow), and $\mathbf{f}_t$ is the produced noisy flow. t denotes the a time step from the pre-defined sequence $\{0, 1, ..., T\}$. $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s = \prod_{s=0}^{t}(1-\beta_s)$, and β_s indicates the noise variance schedule [14]. Crucially, the ground-truth flow $\mathbf{f}_0$ must be normalized and scaled. In practice, we follow DDVM [31] to normalize the flow based on the height and width of the ground truth, resulting in a range of $\{-1, 1\}$. Subsequently, we incorporate the scale factor $\boldsymbol{b}$ to establish the range $\{-b, b\}$. Prior work [14] shows that the scaling factor $\boldsymbol{b}$ plays an indispensable role in modulating the signal-to-noise ratio during the diffusion process. The following studies, including [5, 6], underscore the criticality of task-specific factors in securing high performance in diverse applications. In pursuit of this, Sec. 4 of our manuscript will delve into an empirical exploration to assess the influence of the scale factor $\boldsymbol{b}$ on the overall performance of the model in this field.

Reverse Denoising. During inference, the reverse process of diffusion model $q(\mathbf{f}_{t-1}|\mathbf{f}_t)$ can be formulated as generative processes for non-Markovian forward processes parametrized by σ [33], which is given by

$$\mathbf{f}_{t-1} = \sqrt{\alpha_{t-1}}\,\mathbf{f}_\theta^{(t)} + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\,\tilde{\epsilon}_t + \sigma_t\epsilon_t, \tag{5}$$

where ϵ_t is standard Gaussian noise, and $\mathbf{f}_\theta^{(t)}$ indicates a function intended to approximately predict $\mathbf{f}_0$ from the noisy flow $\mathbf{f}_t$. $\tilde{\epsilon}_t$ is the approximated noise at timestep t:

$$\tilde{\epsilon_t} = \frac{\mathbf{f}_t - \sqrt{\alpha_t}\,\mathbf{f}_\theta^{(t)}}{\sqrt{1-\alpha_t}}. \tag{6}$$

It is important to note that the generative processes yielded by our model are contingent upon the chosen value of σ. Specifically, setting σ to 0 for all t instigates a deterministic generative process. This specific instantiation is

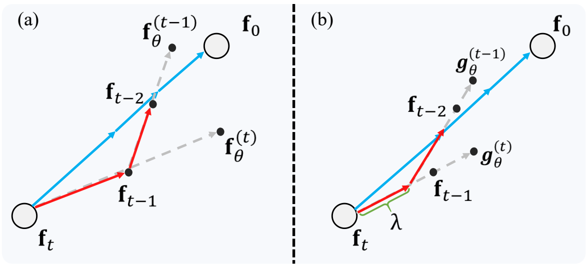

Figure 3. **Illustration of Hidden State Denoising (HSD)**. For more details, please refer to Sec. 3.3.

.

known as the Denoising Diffusion Implicit Model (DDIM), as described in [33], given $\mathbf{f}_{t-1}$ and $\mathbf{f}_0$. Our empirical analyses suggest that this implicit probabilistic model fosters more stable predictions in generative flow. In the context of the reverse denoising process, as outlined in Eqn. (5), the flow decoder incrementally refines the noisy flow through an iterative sequence, *i.e.*, $\mathbf{f}_t \rightarrow \mathbf{f}_{t-\Delta} \rightarrow \cdots \mathbf{f}_0$.

3.3. Conditional-RDD with Hidden State Denoising

A fundamental challenge in standard diffusion models lies in their requirement for an extensive series of denoising steps to effectively convert noise into a target signal. Our proposed Conditional-RDD has made significant strides in reducing this computational load by applying reverse denoising through a RAFT-inspired recurrent denoising decoder. Furthermore, we posit that additional optimization can be attained by more efficiently harnessing the synergistic potential of these components.

Specifically, we introduce a new decoding strategy, named Hidden State Denoising (HSD), which can boost the effectiveness and stability of the denoising process within *FlowDiffuser*. The primary insight of HSD entails merging hidden state features similar to those found in RAFT's recurrent decoder into the denoising process. Nevertheless, seamlessly blending these components presents a notable challenge. This arises from a fundamental disparity between the conventional hidden state learning mechanism and the basic structure of diffusion models. The latter primarily focuses on iterative noise updates and restarts the model from scratch in each denoising step.

Fig. 3 presents a simplified, illustrative example to elucidate the mechanism of hidden state generation. For clarity, we condense the entire reverse diffusion process into discrete timesteps $t \in \{0, 1, 2\}$. In sub-figure (a), blue arrows indicate the idealized reverse progression at each timestep, whereas red arrows illustrate the actual denoising phases. The process initiates from the noisy initial state $\mathbf{f}_t$, then the decoder estimates $\mathbf{f}_\theta^{(t)}$ and employs this to compute $\mathbf{f}_{t-1}$, as dictated by Eqn. (5). Iteratively applying this rule ultimately leads to the generative outcome $\mathbf{f}'_0$, conforming to the trained distribution $q(\mathbf{f}_0)$.

Empirical observations reveal that the UNet's standard denoising decoder demonstrates significant variance in flow prediction, necessitating more iterations for convergence than the basic RAFT decoder. Conversely, our HSD capitalizes on the strengths of RAFT's recurrent decoder to alleviate this variance issue. The key aim is to minimize the estimation error in directly predicting $\mathbf{f}_{t-1}$, thereby reducing its subsequent impact on the final optical flow. To this end, we propose to integrate RAFT's hidden learning pattern into our denoising steps. However, we empirically observe that the straightforward application of RAFT's hidden learning procedure in our denoising decoder leads to a decline in performance. This is because the simple hidden learning scheme of RAFT conflicts with the fundamental setup of learning from scratch in diffusion models.

To tackle this challenge, we define a sub-network $G(\cdot)$, to predict the hidden state at timestep t, shown as $\boldsymbol{g}_\theta^{(t)}$ in sub-figure (b). Notably, $G(\cdot)$ can be easily achieved by training the decoder with a higher number of iterations during the training stage compared to the inference stage, without incurring extra computational costs. Consequently, $\boldsymbol{g}_\theta^{(t)}$ represents the intermediate flow prediction, involving fewer iterations than those used during training in the RAFT decoder, and can be considered as the latent form of $\mathbf{f}_\theta^{(t)}$. It is important to note that the diffusion models require only an *approximate prediction* of $\mathbf{f}_0$ to execute the reverse process, as indicated in Sec. 3.2. This capability allows for the substitution of $\mathbf{f}_\theta^{(t)}$ with its latent version $\boldsymbol{g}_\theta^{(t)}$ in the update scheme using Eqn. (5). Subsequently, by applying a striding factor as $\bar{\mathbf{f}}_{t-1} = \lambda\, \mathbf{f}_{t-1}$, it further prevents potential errors in long-range prediction from noisy flows. This strategy, applied repeatedly, progressively denoises and refines the flow towards a more stable $\bar{\mathbf{f}}'_0$.

Training Objective. In our HSD, instead of predicting ϵ_t as formulated by [14], we follow [42] to predict the signal itself. The training objective is given by

$$\mathcal{L} = E_{\mathbf{f}_0 \sim q(\mathbf{f}_0|\mathbf{c}), t \sim [1,T]}\left[\, ||\mathbf{f}_{gt} - \bar{\mathbf{f}}'_0||_1\right] \tag{7}$$

where $\mathbf{c}$ denotes the abbreviation of previously mentioned conditions, $\bar{\mathbf{f}}'_0$ indicates the conditional denoising result.

4. Experiments

4.1. Implementation Details

Following previous works [15, 32], our *FlowDiffuser* utilizes Twins-SVT as the encoders. The decoding process, pivotal for denoising, is facilitated by a RAFT-based model [41], with iterations set at $N = 12$ as default. Additionally, the design of *FlowDiffuser* reflects a balanced compatibility with leading-edge methods in the field, including those outlined in [19, 25, 39, 41], as evidenced in

Method	Sintel (train)		KITTI-15 (train)		Sintel (test)		KITTI-15 (test)	Avg.Rank
	Clean	Final	EPE	F1-all	Clean	Final	F1-all	
RAFT [ECCV20] [41]	1.43	2.71	5.04	17.4	1.61	2.86	5.10	16.0
GMA [ICCV21] [19]	1.30	2.74	4.69	17.1	1.39	2.47	5.15	14.1
SeperableFlow [ICCV21] [48]	1.30	2.59	4.60	15.9	1.50	2.67	4.64	12.9
CRAFT [CVPR22] [35]	1.27	2.79	4.88	17.5	1.45	2.42	4.79	14.1
GMFlow [CVPR22] [45]	1.08	2.48	7.77	23.4	1.74	2.90	9.32	15.4
GMFlowNet [CVPR22] [51]	1.14	2.71	4.24	15.4	1.39	2.65	4.79	11.3
KPA-Flow [CVPR22] [25]	1.28	2.68	4.46	15.9	1.35	2.36	4.60	10.6
OCTC [CVPR22] [18]	1.31	2.67	4.72	16.3	1.41	2.57	4.33	12.1
SKFlow [NeurIPS22] [39]	1.22	2.46	4.27	15.5	1.28	2.27	4.84	9.6
FlowFormer [ECCV22] [15]	0.95	2.35	4.09	14.7	1.16	2.09	4.68	7.0
RAFT-it* [ECCV22] [38]	1.74	2.41	4.18	13.4	1.55	2.90	4.31	9.9
GMFlow+ [TPAMI23] [46]	0.91	2.74	5.74	17.6	1.03	2.37	4.49	10.1
FlowFormer++* [CVPR23] [32]	0.90	2.30	3.93	14.1	1.07	**1.94**	4.52	3.7
TransFlow* [CVPR23] [24]	0.93	2.33	3.98	14.4	1.06	2.08	4.32	4.0
MatchFlow* [CVPR23] [10]	1.03	2.45	4.08	15.6	1.16	2.37	4.63	7.7
GAFlow [ICCV23] [27]	0.95	2.34	3.92	13.9	1.15	2.05	4.42	4.1
EMD-Flow [ICCV23] [8]	0.88	2.55	4.12	13.5	1.32	2.51	4.51	7.1
FlowDiffuser*	**0.86**	**2.19**	**3.61**	**11.8**	**1.02**	2.03	**4.17**	**1.1**

Table 1. **Quantitative comparison with state-of-the-art approaches.** Following [10, 15, 32, 39], here we conduct a comparative comparison between the proposed *FlowDiffuser* and recent published optical flow approaches that also operate under a two-frame setting. The metrics with "(train)" and "(test)" indicate the evaluation for generalization ability and online performance, respectively. Given recent inconsistencies in achieving optimal results across datasets, we calculate the average rank ("Avg.Rank") for all metrics. This measure offers a concise overview of the overall capabilities of different approaches. * indicates training without the standard "C+T" setup, see Sec. 4.2.

the experimental section. Technical parameters such as the scale factor $\boldsymbol{b}$ and striding factor λ are set to 0.5 and 0.2 respectively in our diffusion strategy. In HSD, we set the default denoising steps K to 3 for the reverse process.

During the training phase of *FlowDiffuser*, we utilize a batch size of 6 and the AdamW optimizer with a one-cycle learning rate, in accordance with RAFT's settings [41]. Our models are pre-trained on synthetic datasets such as FlyingChairs [11] and FlyingThings [28], and subsequently fine-tuned on a combined dataset comprising Sintel [4], KITTI-2015 [30], and HD1K [20], consistent with recent research approaches [15, 19, 25, 51]. In line with advancements in the field [10, 24, 31, 38], our training is further augmented with additional datasets. As shown in Tab. 2, besides the standard "C+T" training, our model is also trained on "AF+T" following DDVM [31]. For evaluation and online testing purposes, a single GPU setup is employed with a batch size of 1.

4.2. Benchmarking on Optical Flow Datasets

Generalization Performance. We first compare the proposed *FlowDiffuser* with top-performing methods on the Sintel and KITTI benchmarks. The primary focus was to assess the generalization capability of *FlowDiffuser*. The results, as presented in Tab. 1, demonstrate that *FlowDiffuser* achieves unparalleled performance on both datasets. Notably, on the Sintel dataset, it records an End-Point Error (EPE) of 0.86 on the clean pass and 2.19 on the final. Meanwhile, on the KITTI dataset, it sets new records with an EPE of 3.61 and an F1-all score of 11.8%, surpassing existing methodologies by a considerable margin.

Online Testing. In the Sintel online tests, *FlowDiffuser* achieves an impressive End-Point Error (EPE) of 1.02 and 2.03, notably surpassing recent methods like Match-Flow [10] and EMD-Flow [8] by margins of 13.2% and 20.9%, respectively. Additionally, *FlowDiffuser* elevates the state-of-the-art (SOTA) performance on the KITTI benchmark to 4.17%, outperforming the leading models like RAFT-it [38] and TransFlow [24]. Fig. 4 provides some qualitative comparisons. Furthermore, we conducted a comprehensive performance analysis across various metrics. As shown in the last column of Tab. 1, *FlowDiffuser* secures an average rank of 1.1, achieving the top rank in 6 out of 7 metrics. This underscores its significant advantage over recent published works in the field.

Effectiveness of Training Data. The training data represents the underlying distribution of the target problem, and it significantly influences a model's performance and generalization capabilities [10, 24, 32]. In line with prior research [31, 38], we employ AutoFlow [37] for our training data ablation studies. Tab. 2 highlights the impact of augmenting the training dataset on the performance of *FlowDiffuser*. Despite the challenges in optimizing *FlowDiffuser*'s high scores, the integration of additional training data consistently enhances its performance across all metrics. This suggests that expanding the training dataset positively af-

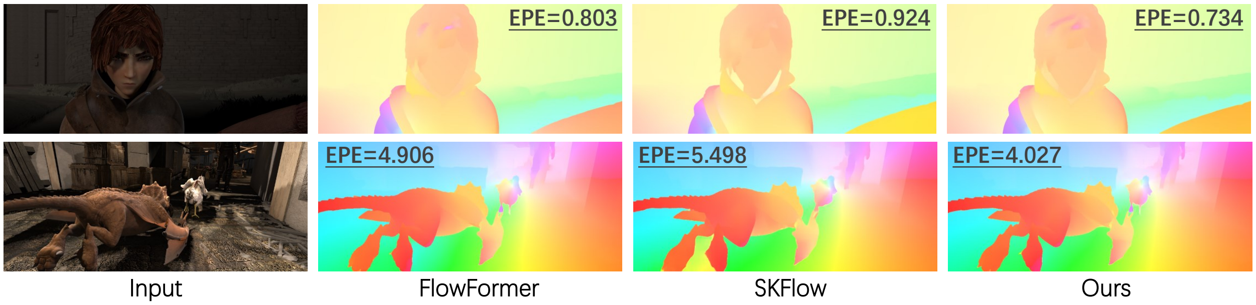

Figure 4. **Qualitative comparisons with renowned works**, SKFlow [39] and FlowFormer [15], on Sintel test set.

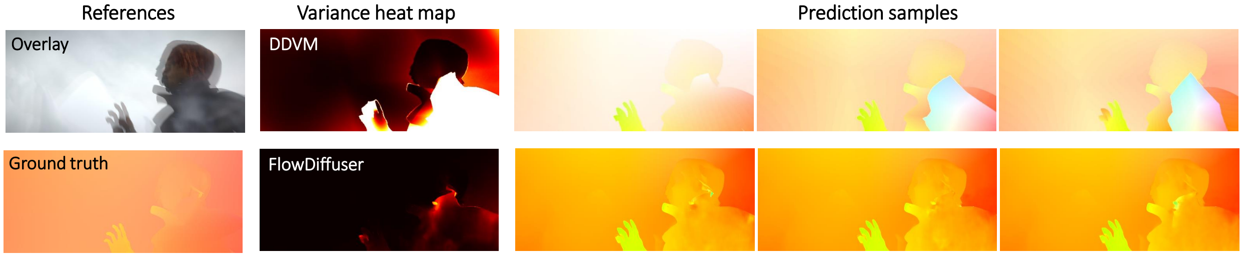

Figure 5. **Visualization of Prediction Samples and the Corresponding Variance Map.** The proposed *FlowDiffuser* (row 2) demonstrates the advantages of effectively capturing uncertainty and ambiguity, similar to DDVM [31] (row 1). Additionally, our model's predictions exhibit significantly improved stability, making them more reliable for real-world applications.

Method	Dataset	Sintel (train)		KITTI-15 (train)	
		Clean	Final	EPE	F1-all
RAFT-it [38]	AF	1.74	2.41	4.18	13.4
FlowFormer++ [32]	C+T+YV	0.90	2.30	3.93	14.1
TransFlow [24]	RS / RK	0.93	2.33	3.98	14.4
MatchFlow [10]	C+T+MD	1.03	2.45	4.08	15.6
CroCo-Flow [44]	C+T+CC	1.28	2.58	-	-
DDVM [31]	AF+T	1.48	2.22	3.71	14.1
FlowDiffuser	AF+T	**0.86**	**2.19**	**3.61**	**11.8**
SKFlow [39]	C+T	1.22	2.46	4.27	15.5
GMFlow+ [46]	C+T	0.91	2.74	5.74	17.6
FlowDiffuser	C+T	**0.89**	**2.38**	**3.84**	**12.7**

Table 2. **Quantitative comparisons with different training data**. The abbreviations are as follows: "C" for Chairs, "T" for Things, "AF" for AutoFlow [37], "YV" for YouTube-VOS [47], "MD" for MegaDepth [23], and "CC" for CroCo [44]. Additionally, "RS" and "RK" refer to the raw data of Sintel and KITTI, respectively.

fects *FlowDiffuser*'s effectiveness. The incorporation of a more diverse dataset allows *FlowDiffuser* to better capture various patterns and variations, leading to notable improvements in multiple evaluation metrics. Moreover, in comparison with DDVM [31], the predictions generated by our model exhibit a remarkable level of stability, as in Fig. 5.

4.3. Ablation Study

Compatibility with Existing Models. We plug the proposed Conditional-RDD into advanced models: *FlowDiffuser*-R (RAFT [41]), *FlowDiffuser*-G (GMA [19]), *FlowDiffuser*-K (KPA [25]), and *FlowDiffuser*-S (SKFlow [39]). As shown in Tab. 3, *FlowDiffuser* variations demonstrate exceptional performance in generalization evaluations, consistently outperforming baseline models across all metrics. Specifically, *FlowDiffuser*-R and -G surpass RAFT and GMA by 6.4% and 4.3% on Sintel, and 8.7% and 8.3% on KITTI, respectively. *FlowDiffuser*-K and -S also outperform the baseline, and even surpass recent models like MatchFlow [10] and TransFlow [24]. *FlowDiffuser*-K achieves 3.97 EPE and 14.7% F1-all, and *FlowDiffuser*-S achieves a remarkable score of 3.91 EPE and 14.2%, with minimal extra computation.

Furthermore, to understand the impact of the denoising process in our approach, we present intermediate results from our *FlowDiffuser*-R and RAFT model in Fig. 6. The RAFT model struggles with challenging scenarios characterized by severe motion blur. Conversely, our method effectively utilizes learned motion patterns from the training data distribution to address this challenge, resulting in a more accurate and reliable flow field.

Ablation for Diffusion Approaches. In Tab. 4 (# 1), we evaluate the *FlowDiffuser* models against baseline models in terms of performance and computational overhead. Baseline refers to a pure discriminative model that excludes all diffusion-related techniques, components, and denoising strategies employed in our approach. The results show that *FlowDiffuser* achieves notable improvements over the base-

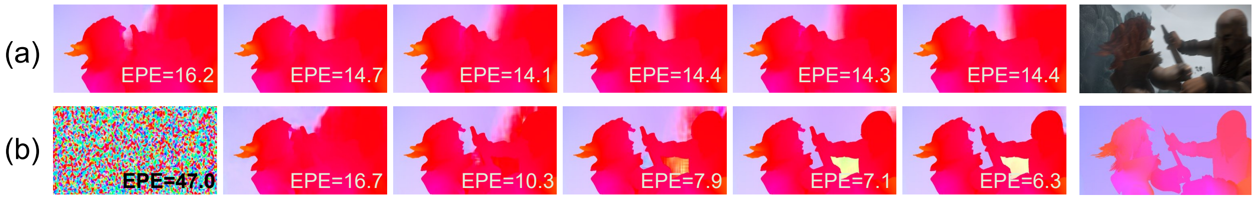

Figure 6. **Visualization of intermediate results.** (a) illustrates the results of RAFT [41] decoder, and (b) presents the denoising results of our *FlowDiffuser*-R. The last column provides the reference image and ground truth.

Method	Sintel (train)		AG.	KITTI (train)		AG.
	Clean	Final		EPE	F1-all	
RAFT [41]	1.43	2.71	-	5.04	17.4	-
FlowDiffuser-R	**1.29**	**2.63**	+6.4%	**4.51**	**16.2**	+8.7%
GMA [19]	1.30	2.74	-	4.69	17.1	-
FlowDiffuser-G	**1.26**	**2.59**	+4.3%	**4.32**	**15.6**	+8.3%
KPA-Flow [25]	1.28	2.68	-	4.46	15.9	-
FlowDiffuser-K	**1.24**	**2.45**	+5.8%	**3.97**	**14.7**	+9.3%
SKFlow [39]	1.22	2.46	-	4.27	15.5	-
FlowDiffuser-S	**1.18**	**2.39**	+3.1%	**3.91**	**14.2**	+8.4%

Table 3. **Compatibility evaluation.** Following [19, 25, 39, 41], all models are trained on "C+T" and evaluated on Sintel and KITTI training set for fair comparison. "AG." denotes the average gain of performance on Sinte/KITTI dataset.

line with only a minimal increase in parameters. Additionally, the integration of HSD enhances performance across all metrics without adding to the computational complexity.

Ablation for Denoising Steps. Tab. 4 (# 2) illustrates the performance enhancement of *FlowDiffuser* with different DDIM denoising steps. As K increased from 1 to 3, performance shows a consistent improvement ranging from approximately $8.4\% \sim 24.8\%$ across four metrics. However, with a further increase to $K = 4$, the performance gain is minimal. Consequently, we designate $K = 3$ as the optimal default setting.

Ablation for Time Embeddings. The time embedding function aligns diffusion iterations with corresponding timesteps. To enhance this, we introduce the EE module, which amplifies the role of time embeddings. This enhancement enables the denoising decoder to more effectively recognize noise variations across different times. As illustrated in Tab. 4 (# 3), integrating the EE module yields significant improvements in all evaluated metrics.

Ablation for Scale Factor *b*. Prior works [5, 6, 14] have shown that diverse tasks necessitate distinct task-specific factors to guarantee optimal performance. Here we empirically analyze the effect of $\boldsymbol{b}$ in optical flow, as in Tab. 4 (# 4). Despite the sensitivity of classification [6] and detection [5] tasks to the signal-to-noise ratio, our *FlowDiffuser* demonstrates insensitivity to this factor. We select $\boldsymbol{b} = 0.5$ as the default setting, yielding slightly improved results.

Method	Sintel (train)		KITTI (train)		Param.
	Clean	Final	EPE	F1-all	
# 1: FlowDiffuser (FD.) approaches					
Baseline	0.98	2.43	4.18	14.5	14.9M
FD.-w/o HSD	0.93	2.31	3.92	13.9	16.3M
FD.-HSD	0.86	2.19	3.61	11.8	
# 2: Denoising Steps K					
1	0.96	2.39	4.43	15.7	16.3M
2	0.89	2.25	3.86	12.4	
3	0.86	2.19	3.61	11.8	
4	0.86	2.20	3.59	11.7	
# 3: Embedding Enhancement (EE)					
w/. EE	0.86	2.19	3.61	11.8	16.3M
w/o EE	0.92	2.30	3.85	12.6	15.2M
# 4: Scale factor $\boldsymbol{b}$					
0.1	0.87	2.21	3.71	12.0	16.3M
0.5	0.86	2.19	3.61	11.8	
1	0.89	2.23	3.75	12.2	

Table 4. **Ablation study**. Settings as default are underlined. All models are trained on "AF+T" for fair comparison.

5. Conclusion

This work marks a significant paradigm shift in optical flow estimation by reformulating it as a conditional flow generation task. The proposed *FlowDiffuser* framework is part of the cutting-edge wave of generative neural frameworks designed specifically for optical flow estimation, and demonstrates enhanced learning and generalization capabilities. The framework's core strength lies in its Conditional Recurrent Denoising Decoder (Conditional-RDD), which specifically integrates Hidden State Denoising (HSD) with a recurrent flow refinement strategy. Findings and insights from *FlowDiffuser* are expected to make a substantial contribution to the progression of optical flow estimation techniques, potentially impacting a wide array of applications in computer vision.

Acknowledgements. This work was supported in part by National Natural Science Foundation of China (NSFC) under grant No.62372091 and Sichuan Science and Technology Program of China under grant No.2023NSFSC0462.

References

[1] Hyemin Ahn, Esteve Valls Mascaro, and Dongheui Lee. Can we use diffusion probabilistic models for 3d motion prediction? *Preprint arXiv:2302.14503*, 2023. 3

[2] Min Bai, Wenjie Luo, Kaustav Kundu, and Raquel Urtasun. Exploiting semantic information and deep matching for optical flow. In *ECCV*, 2016. 2

[3] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *CVPR*, 2022. 3

[4] Daniel Butler, Jonas Wulff, Garrett Stanley, and Michael Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 6

[5] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *Preprint arXiv:2211.09788*, 2022. 3, 4, 8

[6] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. *Preprint arXiv:2210.06366*, 2022. 4, 8

[7] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *Preprint arXiv:2302.08908*, 2023. 2, 3

[8] Changxing Deng, Ao Luo, Haibin Huang, Shaodan Ma, Jiangyu Liu, and Shuaicheng Liu. Explicit motion disentangling for efficient optical flow estimation. In *ICCV*, 2023. 1, 6

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. 2021. 2, 3

[10] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Rethinking optical flow from geometric matching consistent perspective. In *CVPR*, 2023. 6, 7

[11] A. Dosovitskiy, P. Fischer, Eddy Ilg, Philip Häusser, Caner Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 6

[12] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 1, 2

[13] Zhangxuan Gu, Haoxing Chen, Zhuoer Xu, Jun Lan, Changhua Meng, and Weiqiang Wang. Diffusioninst: Diffusion model for instance segmentation. *Preprint arXiv:2212.02773*, 2022. 3

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3, 4, 5, 8

[15] 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. 1, 5, 6, 7

[16] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *CVPR*, 2018. 2

[17] Junhwa Hur and S. Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *CVPR*, 2019. 1, 2

[18] Jisoo Jeong, Jamie Menjay Lin, Fatih Porikli, and Nojun Kwak. Imposing consistency for optical flow estimation. In *CVPR*, 2022. 6

[19] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *ICCV*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[20] D. Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Güssefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, C. Brenner, and B. Jähne. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *CVPRW*, 2016. 6

[21] Bo Li, Xiaolin Wei, Fengwei Chen, and Bin Liu. 3d colored shape reconstruction from a single rgb image through diffusion. *Preprint arXiv:2302.05573*, 2023. 3

[22] Haipeng Li, Hai Jiang, Ao Luo, Ping Tan, Haoqiang Fan, Bing Zeng, and Shuaicheng Liu. Dmhomo: Learning homography with diffusion models. *ACM Transactions on Graphics*, 2024. 3

[23] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPP*, 2018. 7

[24] Yawen Lu, Qifan Wang, Siqi Ma, Tong Geng, Yingjie Victor Chen, Huaijin Chen, and Dongfang Liu. Transflow: Transformer as flow learner. In *CVPR*, 2023. 6, 7

[25] Ao Luo, Fan Yang, Xin Li, and Shuaicheng Liu. Learning optical flow with kernel patch attention. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[26] Ao Luo, Fan Yang, Kunming Luo, Xin Li, Haoqiang Fan, and Shuaicheng Liu. Learning optical flow with adaptive graph reasoning. In *AAAI*, 2022. 2

[27] Ao Luo, Fan Yang, Xin Li, Lang Nie, Chunyu Lin, Haoqiang Fan, and Shuaicheng Liu. Gaflow: Incorporating gaussian attention into optical flow. In *ICCV*, 2023. 6

[28] N. Mayer, Eddy Ilg, Philip Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 6

[29] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. pc^2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. *Preprint arXiv:2302.10668*, 2023. 3

[30] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 6

[31] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *Preprint arXiv:2306.01923*, 2023. 2, 3, 4, 6, 7

[32] 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*, 2023. 5, 6, 7

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3, 4, 5

[34] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. 2019. 2, 3

[35] 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. 6

[36] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 1, 2

[37] 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. 6, 7

[38] 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. 6, 7

[39] Shangkun Sun, Yuanqi Chen, Yu Zhu, Guodong Guo, and Ge Li. Skflow: Learning optical flow with super kernels. In *NeurIPS*, 2022. 3, 4, 5, 6, 7, 8

[40] Haoru Tan, Sitong Wu, and Jimin Pi. Semantic diffusion network for semantic segmentation. *Preprint arXiv:2302.02057*, 2023. 3

[41] Zachary Teed and Jun Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 1, 2, 3, 5, 6, 7, 8

[42] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2022. 5

[43] Xiaodong Wang, Chenfei Wu, Shengming Yin, Minheng Ni, Jianfeng Wang, Linjie Li, Zhengyuan Yang, Fan Yang, Lijuan Wang, Zicheng Liu, et al. Learning 3d photography videos via self-supervised diffusion on single images. *Preprint arXiv:2302.10781*, 2023. 2, 3

[44] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *ICCV*, 2023. 7

[45] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *CVPR*, 2022. 1, 6

[46] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *TPAMI*, 2023. 6, 7

[47] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *Preprint arXiv:1809.03327*, 2018. 7

[48] Feihu Zhang, Oliver J Woodford, Victor Adrian Prisacariu, and Philip HS Torr. Separable flow: Learning motion cost volumes for optical flow estimation. In *ICCV*, 2021. 6

[49] Chengqian Zhao, Cheng Feng, Dengwang Li, and Shuo Li. Of-msrn: optical flow-auxiliary multi-task regression network for direct quantitative measurement, segmentation and motion estimation. In *AAAI*, 2020. 1, 2

[50] Shengyu Zhao, Yilun Sheng, Yue Dong, E. Chang, and Yan Xu. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *CVPR*, 2020. 1, 2

[51] Shiyu Zhao, Long Zhao, Zhixing Zhang, Enyu Zhou, and Dimitris Metaxas. Global matching with overlapping attention for optical flow estimation. In *CVPR*, 2022. 1, 6