

Learning Optical Flow with Adaptive Graph Reasoning

Ao Luo¹, Fan Yang², Kunming Luo¹, Xin Li², Haoqiang Fan¹, Shuaicheng Liu^{3,1*}

¹Megvii Technology

²Group 42

³University of Electronic Science and Technology of China

{luao02, luokunming, fhq}@megvii.com, {fanyang_uestc, xinli_uestc}@hotmail.com, liushuaicheng@uestc.edu.cn

Abstract

Estimating per-pixel motion between video frames, known as optical flow, is a long-standing problem in video understanding and analysis. Most contemporary optical flow techniques largely focus on addressing the cross-image matching with feature similarity, with few methods considering how to explicitly reason over the given scene for achieving a holistic motion understanding. In this work, taking a fresh perspective, we introduce a novel graph-based approach, called adaptive graph reasoning for optical flow (AGFlow), to emphasize the value of scene/context information in optical flow. Our key idea is to decouple the context reasoning from the matching procedure, and exploit scene information to effectively assist motion estimation by *learning to reason* over the adaptive graph. The proposed AGFlow can effectively exploit the context information and incorporate it within the matching procedure, producing more robust and accurate results. On both Sintel clean and final passes, our AGFlow achieves the best accuracy with EPE of 1.43 and 2.47 pixels, outperforming state-of-the-art approaches by 11.2% and 13.6%, respectively. Code is publicly available at <https://github.com/megvii-research/AGFlow>.

Introduction

Optical flow is a fundamental task in video understanding and analysis, aiming to estimate the pixel-wise correspondence between two video frames. It has drawn continuous attention from both academia and industry due to its wide applications, *e.g.*, person-identification (Chen et al. 2020), visual tracking (Vihlman and Visala 2020) and video inpainting (Xu et al. 2019). Recent years have witnessed significant breakthroughs made to push its performance frontier (Dosovitskiy et al. 2015; Ilg et al. 2017; Teed and Deng 2020; Jiang et al. 2021b), but it remains challenging due to inherent ambiguity in textures, large displacements, occlusions, motion blur, and non-Lambertian effects.

Traditional optical flow algorithms formulate the dense matching as an energy minimization problem based on feature constancy and spatial smoothness (Horn and Schunck 1981; Brox et al. 2004; Bruhn, Weickert, and Schnörr 2005). However, because the hand-designing features and optimization objectives are difficult to cover all scenarios, these

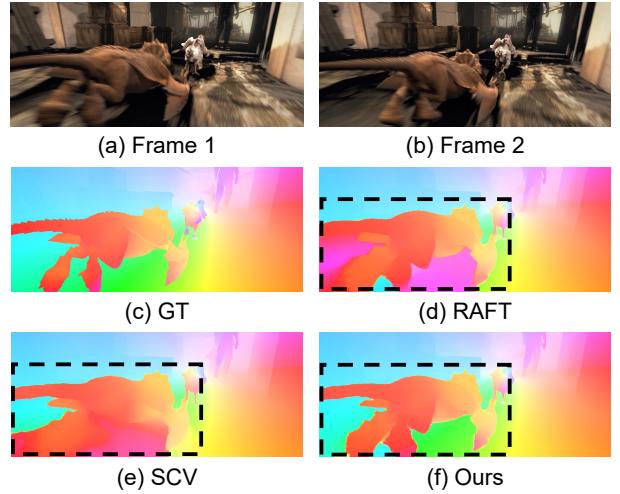


Figure 1: A challenging image pair with heavy motion blur from the final pass of Sintel test set. Unlike previous state-of-the-art methods (RAFT (Teed and Deng 2020) and SCV (Jiang et al. 2021b)) suffering from ambiguous matching, our AGFlow is able to perform global reasoning conditioned on scene context for achieving a holistic motion understanding.

approaches are not robust enough to deal with complex motions. As a powerful alternative to traditional methods, deep learning based approaches take the research of optical flow into a new level. Current optical flow methods have made great advances in: **i**) developing powerful data-driven, end-to-end learning paradigms (Dosovitskiy et al. 2015; Ilg et al. 2017); **ii**) designing multiple refinement strategies (Sun et al. 2018; Hur and Roth 2019; Zhao et al. 2020b); **iii**) exploiting auxiliary information from related tasks (Zhao et al. 2020a); and **iv**) modeling pixel-wise relations for all pairs (Teed and Deng 2020; Jiang et al. 2021b). Although these deep learning based approaches have shown the strong capability of matching across frames, they are subject to a significant limitation: current methods largely focus on addressing the matching similarity between features, lacking a holistic motion understanding of the given scene. Thus, the ambiguities (local variations) caused by motion blur, occlusion and large motions severely degrade the accuracy of current mod-

*Corresponding author

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

els (see Fig. 1), hindering their applications in the real world.

Taking a closer look at current optical flow methods (Sun et al. 2018; Yang and Ramanan 2019), their success is majorly attributed to an important component, namely 4D correlation volume, that typically model the correlations between features across frames. They expect the correlations achieved by using deep trainable features to surmount all ambiguities. However, evidence from (Teed and Deng 2020) indicate that such correlations are vulnerable to variable feature representations on challenging conditions. To better aggregate context information within motion boundaries, existing arts (Teed and Deng 2020; Jiang et al. 2021b) inject image features by concatenation operation, and encode scene information by using extra convolutional layers. But despite these attempts, the holistic motion understanding is far from being achieved: **i**) existing approaches capture scene information with only naive operations, *e.g.*, feature stacking, without typed functions for explicitly modeling such process; **ii**) their operations are confined to the original coordinate space, causing a heavy computational burden and lacking a global understanding of the given scene; and **iii**) they ignore the adverse effect caused by ‘domain gap’, *e.g.*, the gap between scene content and motion features. These observations prompted us to think about: *how to empower the optical flow model to effectively obtain the capability of holistic motion understanding?*

To answer this question, we propose to conduct *adaptive graph reasoning* within the conventional optical flow framework. Our idea is to decouple the context reasoning from the matching procedure, and explore scene information to effectively assist in motion estimation by *learning to reason* over the adaptive graph. We argue that a powerful optical flow model should have the capability of going beyond regular grids, which learns to understand the real-world motions under the guidance of scene content from a more global view. Towards this goal, we introduce a novel graph-based approach, namely adaptive graph reasoning for optical flow (AGFlow), which embeds graph techniques onto the matching pipeline to enable an effective context reasoning and information interaction. The proposed AGFlow learns to match features conditioned on scene context, and allows objects’ spatial extents to be well aggregated and thus largely decreases the uncertainty of ambiguous matching.

In particular, AGFlow consists of three major components: a motion encoder that maps input frames into high-level representations for computing their 4D correlation volumes, a context encoder that extracts features only from the first input frame for capturing scene information, and an adaptive graph reasoning (AGR) module that learns to understand the given scene and distills useful information to assist in optical flow estimation. Unlike existing approaches (Teed and Deng 2020; Jiang et al. 2021b), our AGR exploits scene information for optical flow in the graph domain; that is, both context and motion features are mapped from regular grids to graph space for learning optical flow together. It enables the model to understand the motion from a larger context with only limited extra parameters and FLOPs (see Tab. 3). Importantly, to overcome the domain gap between context and motion representa-

tions, a novel graph adapter (GA) is introduced to adapt (motion) graph parameters in a one-shot manner, incorporating the scene information flexibly. Overall, our approach helps optical flow models conduct more efficient scene understanding and naturally distill scene information to assist in optical flow estimation, eventually leading to strong holistic-motion-understanding capability and better performance. The **contributions** of this work are summarized as follows:

- **A novel graph-based approach for optical flow.** To our knowledge, this is the first work that explicitly exploits scene information to assist in optical flow estimation by using graph techniques. The proposed AGFlow can go beyond the regular grids and reason over the graph space to achieve a better motion understanding, thus successfully handling different challenges in optical flow.
- **An adaptive cross-domain graph reasoning approach.** In order to incorporate scene information, we generalize the *learning to adapt* mechanism (Bertinetto et al. 2016) from regular grids to the graph domain. Our designed graph adapter can fast adapt scene context to guide the global (motion) graph reasoning in a one-shot manner.
- **State-of-the-art results on widely-used benchmarks.** Our AGFlow sets new records on both Sintel and KITTI benchmarks, outperforming state-of-the-art approaches by a relatively large margin.

Related Work

Optical Flow Estimation. Optical flow is the task of estimating per-pixel motion between video frames. In the early stage, researchers (Black and Anandan 1993; Horn and Schunck 1981; Brox et al. 2004; Bruhn, Weickert, and Schnörr 2005) consider this task as an energy minimization problem, with the goal of achieving an ideal tradeoff between feature similarities and motion smoothness. However, as motion itself is hard to be modeled/described by hand-crafted features and optimization objectives, it is challenging to obtain precise flow fields by traditional methods. In the deep learning era, early attempts mainly focus on **i**) learning more robust data terms (Bai et al. 2016; Weinzaepfel et al. 2013) or **ii**) avoiding the optimization step to directly estimate optical flow (Teed and Deng 2020; Jiang et al. 2021b). To improve results on optical flow, many recent works introduce stronger learning paradigms that can enable iterative refinement (Ranjan and Black 2017; Sun et al. 2018; Yang and Ramanan 2019; Hui, Tang, and Loy 2018), explicit pixel-wise-relation modeling, and joint representation learning with other tasks (Zhao et al. 2020a). Although remarkable progress have been achieved by these developments, there is still a large room for improvement over existing approaches that largely focus on addressing the matching similarity between features without considering how to achieve a holistic motion understanding. Taking a further step, our AGFlow is empowered to exploit and incorporate high-level scene information to predict optical flow, linking the low-level matches with high-level semantic information for better accuracy.

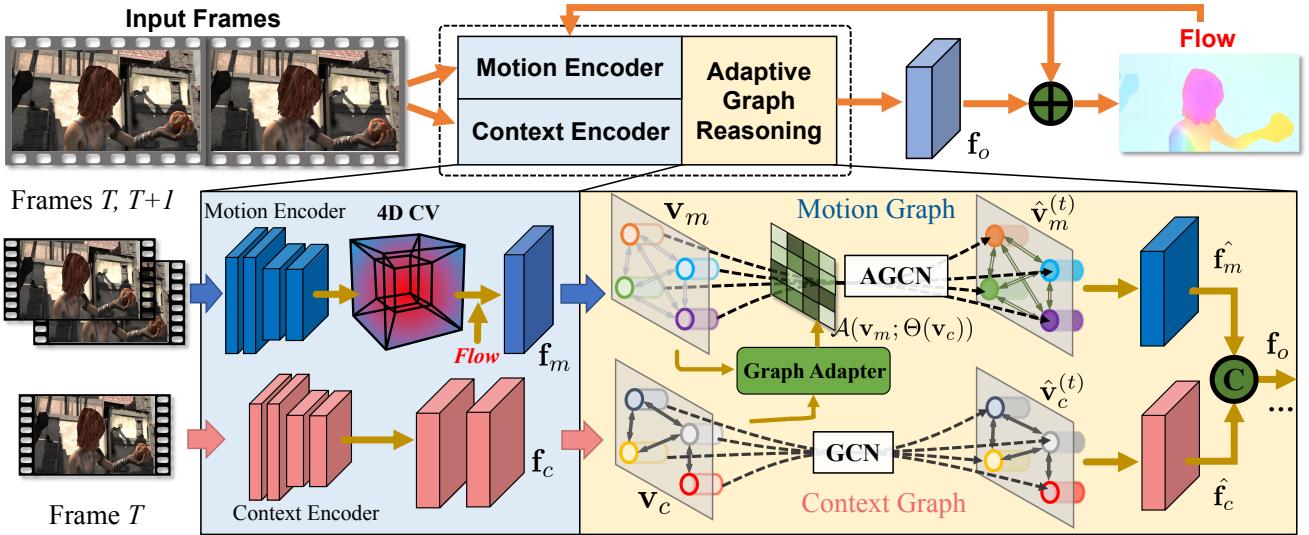


Figure 2: Architecture of the proposed adaptive graph reasoning for optical flow (AGFlow). “4D CV” means 4D correlation volumes, “C” indicates concatenation and “ \oplus ” denotes summation. Please refer to Sec. 3 for more details about the feature notations in this figure. Best viewed in color.

Graph Neural Networks. Graph Neural Networks have been widely applied to different applications, including person re-identification (Shen et al. 2018), salient object detection (Luo et al. 2020), 3D shape analysis (Wei, Yu, and Sun 2020), semantic segmentation (Luo et al. 2021) and video question answering (Park, Lee, and Sohn 2021). In the context of optical flow, Graph Convolutional Networks (GCNs) are still largely under-explored. This is because most existing approaches largely focus on the matching similarity between features; the high-level scene information is overlooked or poorly investigated. In this paper, we show that GCNs can benefit the optical flow by comprehensively mining the scene information to assist in flow estimation. It helps the optical model to go beyond regular grids and understand the motion from a global view. Unlike most conventional GCN-based models (Mohamed et al. 2020; Wei, Yu, and Sun 2020; Luo et al. 2021; Park, Lee, and Sohn 2021) that only focus on a single domain, our GCN-based AGFlow has the capability of *learning to adapt* cross-domain information, fully incorporating scene information for comprehensive motion understanding.

Methodology

Problem Formulation

Given a pair of input consecutive images, *i.e.*, source image I_1 and target image I_2 , the task of optical flow estimation is to predict a dense displacement field between them. Deep learning based flow networks commonly employ an encoder-decoder pipeline to first extract context feature f_c and obtain motion cues f_m , and then make flow prediction based on the fused feature f_o in a recurrent/coarse-to-fine manner.

In our approach, we represent the feature fusion in decoder as a graph-based reasoning and learning model, which is formulated as $f_o = \mathcal{F}_g(f_c, f_m)$. Specifically, we define the

model as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where \mathcal{V} indicates a set of nodes, and \mathcal{E} denotes edges specifying the connection and relation information among nodes. After t runs of graph reasoning, the updated nodes are then mapped back to the original coordinate space to predict the displacement field.

Adaptive Graph Reasoning for Optical Flow

Fig. 2 depicts an overview of the proposed adaptive graph reasoning for optical flow (AGFlow). Following the success of prior work (Jiang et al. 2021a), we develop our AGFlow based on RAFT (Teed and Deng 2020). Specifically, given a pair of input images I_1 and I_2 , we employ two residual block based encoders (He et al. 2016) to extract a feature pair (f_1, f_2) and context feature f_c . Then, 4D correlation volumes are constructed on the feature pair in four scales. In the recurrent refinement framework, we utilize four convolutions to capture motion feature f_m from the multi-scale matching costs in each 9×9 region. After that, our adaptive graph reasoning (AGR) module takes the motion feature f_m and context feature f_c as inputs to perform a holistic motion reasoning. Please refer to RAFT (Teed and Deng 2020) for more details about the implementations of the baseline model.

Node embedding. The first step is to project context and motion features in regular coordinate space into graph space. The projection operation decouples positional information from the original grid feature, and makes the produced low dimensional node feature representations more compact and with sufficient expressive power. Here we divide the mapped nodes \mathcal{V} in graph model into two groups: *context nodes* $\mathcal{V}_c = \{v_c^1, \dots, v_c^n\}$, containing appearance feature about shape and region information of scene context, and *motion nodes* $\mathcal{V}_m = \{v_m^1, \dots, v_m^n\}$, storing motion feature of cross-image matching dependency.

Specifically, given the context feature $f_c \in \mathbb{R}^{c \times h \times w}$ and motion feature $f_m \in \mathbb{R}^{c \times h \times w}$ from encoder network, we

employ project function $\mathcal{P}_{f \rightarrow v}(\cdot)$ to assign features with similar representation to the same node. Let $v_c \in \mathbb{R}^{C \times K}$ and $v_m \in \mathbb{R}^{C \times K}$ denote the initial node embeddings in graph space, where C indicates channel number and K is the number of nodes.

To build a global graph over a set of regions, we formulate $\mathcal{P}_{f \rightarrow v}(\cdot)$ as a linear combination of feature vector in grid space, *i.e.*, $v = \mathcal{P}_{f \rightarrow v}(f)$, and thus the produced nodes are able to aggregate long-range information within the overall original feature map. This is given by

$$v_i = \mathcal{N}\left(\sum_{\forall j} \mathcal{F}_{f \rightarrow v}(f)_{ij} \cdot f_j\right), \quad (1)$$

where $\mathcal{N}(\cdot)$ is a L-2 normalization function conducted on channel dimension of each node vector, and $\mathcal{F}_{f \rightarrow v}(f) \in \mathbb{R}^{N \times K}$ models the projection weights for mapping feature maps to node vectors. Note that our approach can be trained with arbitrary input resolutions. In practice, we first use two convolutions on $f \in \mathbb{R}^{c \times h \times w}$ to change the channel dimension from c to K so that a feature map with resolution of $K \times h \times w$ can be obtained. Then a reshape function is applied to produce $\mathcal{F}_{f \rightarrow v}(f)$ with resolution $N \times K$, where $N = h \times w$, and K is a hyperparameter not relying on spatial resolution. Thus the two types of node embedding can be produced by $v_c = \mathcal{P}_{f \rightarrow v}(f_c)$, and $v_m = \mathcal{P}_{f \rightarrow v}(f_m)$.

Adaptive Graph Reasoning. Given node embeddings v in graph space, the adjacency matrix for graph reasoning can be commonly generated by measuring the similarity among all node vectors (Li and Gupta 2018), as $A = v^T v$. After modeling the adjacency matrix A , the graph reasoning with graph convolutional network (Kipf and Welling 2017) is defined as

$$\hat{v} = \mathcal{F}_G(v, A) = \sigma(A v^T w_G), \quad (2)$$

where $\sigma(\cdot)$ is a non-linear activation function, and w_G is learnable parameters for graph convolution. \hat{v} is the updated node representations with graph reasoning, and it can be iteratively enhanced with more runs as $\hat{v}^{(t)} = \mathcal{F}_G(v, A)^{(t)}$, where t denotes update iterations.

Let us consider the representation property of context and motion nodes. Motion nodes mainly encode the point-wise correspondence between an image pair while neglecting the intra-relations among pixels within regions, the context nodes, on the contrary, capture discriminative features for region and shape representations. Thus, we need to address two hurdles: First, there is an inevitable representation gap between context and motion nodes, which could hinder the effective information propagating of directly overall graph reasoning. Second, motion nodes lack constraints on shape or layout for the potential displacement field, and thus they are unable to yield enough context information for individual graph reasoning.

To tackle the issues, we propose an *adaptive graph reasoning* (AGR) module to decouple the context reasoning from the matching procedure, and simultaneously transfer the region and shape prior of scene context to motion nodes in a one-shot manner. The key idea is to exploit the discriminative representations of shape and region in global context

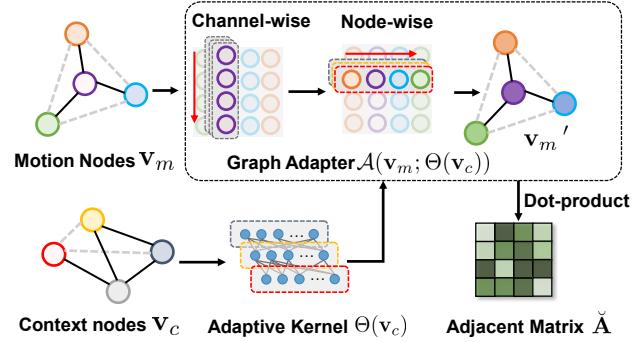


Figure 3: Architecture of Graph Adapter.

to guide the learning of motion adjacency matrix with adaptive parameters. Inspired by (Bertinetto et al. 2016), we devise an adaptive procedure of adjacency matrix learning to predict dynamic parameters that tailor the motion relation modeling based on image-specific contextual information. This is given by

$$\check{A} = \mathcal{A}(v_m; \Theta(v_c)), \quad (3)$$

where $\Theta(\cdot)$ is a parameter learner, and $\mathcal{A}(\cdot)$ denotes a context-to-motion graph adapter (GA) equipped with dynamic weights from $\Theta(v_c)$. In practice, we implement the $\Theta(\cdot)$ with a linear projection function with a softmax activation. As shown in Fig. 3, $\mathcal{A}(\cdot)$ is implemented with a two-layer MLP, where the first regular linear function followed by a ReLU activation is applied to perform channel-wise learning, and then the second linear function with adaptive kernel $\Theta(v_c)$ is utilized to perform node-wise interaction for relation learning with context-to-motion adaptation. Specifically, given context nodes $v_c \in \mathbb{R}^{C \times K}$, we predict the adaptive kernel $\Theta(v_c) \in \mathbb{R}^{K \times K}$ with convolutions on channel dimension ($C \rightarrow K$). Then, we transfer it into the adaptive weights with shape $K \times K$ of the second linear function, which is used to produce v_m' . Finally, we apply the dot-product similarity on v_m' to predict \check{A} .

The produced parameter $\Theta(v_c)$ relies on context nodes for dynamically leveraging the shape and region information of current input. Thus, motion nodes can be fast adaptive to scene context and favorably make full use of the transferred node relation for motion sub-graph reasoning. Thus the enhanced context nodes $\hat{v}_c^{(t)}$ are produced by

$$\hat{v}_c^{(t)} = \mathcal{F}_G(v_c, A)^{(t)}, \text{ where } A = v_c^T v_c, \quad (4)$$

and similarly, motion nodes $\hat{v}_m^{(t)}$ are produced by

$$\hat{v}_m^{(t)} = \mathcal{F}_{AG}(v_m, \check{A})^{(t)}, \text{ where } \check{A} = \mathcal{A}(v_m; \Theta(v_c)), \quad (5)$$

and $\mathcal{F}_{AG}(\cdot)$ denotes motion nodes reasoning with adaptive graph convolutional network (AGCN).

Attentive Readout. After t runs of relation reasoning and state updating, we present an attentive readout module to project the enhanced context nodes $\hat{v}_c^{(t)}$ and motion nodes $\hat{v}_m^{(t)}$ from graph space back to grid feature space, making

Training Data	Method	Sintel (val)		KITTI-15 (val)		Sintel (test)		KITTI-15 (test)	
		Clean	Final	EPE	F1-all	Clean	Final	F1-all	
C + T	HD3(Yin, Darrell, and Yu 2019)	3.84	8.77	13.17	24.0	-	-	-	
	LiteFlowNet(Hui, Tang, and Loy 2018)	2.48	4.04	10.39	28.5	-	-	-	
	PWC-Net(Sun et al. 2018)	2.55	3.93	10.35	33.7	-	-	-	
	FlowNet2(Ilg et al. 2017)	2.02	3.54	10.08	30.0	3.96	6.02	-	
	DICL(Wang et al. 2020)	1.94	3.77	8.70	23.6	-	-	-	
	RAFT(Teed and Deng 2020)	1.43	2.71	5.04	17.4	-	-	-	
	SCV(Jiang et al. 2021b)	1.29	2.95	6.80	19.3	-	-	-	
	AGFlow (ours)	1.31	2.69	4.82	17.0	-	-	-	
C + T + S + K (+ H)	PWC-Net+(Sun et al. 2020)	(1.71)	(2.34)	(1.50)	(5.3)	3.45	4.60	7.72	
	IRR-PWC (Hur and Roth 2019)	(1.92)	(2.51)	(1.63)	(5.3)	3.84	4.58	7.65	
	VCN (Yang and Ramanan 2019)	(1.66)	(2.24)	(1.16)	(4.1)	2.81	4.40	6.30	
	MaskFlowNet(Zhao et al. 2020b)	-	-	-	-	2.52	4.17	6.10	
	ScopeFlow(Bar-Haim and Wolf 2020)	-	-	-	-	3.59	4.10	6.82	
	DICL(Wang et al. 2020)	(1.11)	(1.60)	(1.02)	(3.6)	2.12	3.44	6.31	
	RAFT(Teed and Deng 2020)	(0.76)	(1.22)	(0.63)	(1.5)	1.61*	2.86*	5.10	
	SCV(Jiang et al. 2021b)	(0.79)	(1.70)	(0.75)	(2.1)	1.77*	3.88*	6.17	
	AGFlow (ours)	(0.65)	(1.07)	(0.58)	(1.2)	1.43*	2.47*	4.89	

Table 1: Quantitative comparison with state-of-the-art methods using EPE and F1-all metrics (the lower the better). Following previous works (Wang et al. 2020; Teed and Deng 2020; Jiang et al. 2021b), we compare our results with all published works on three passes from two standard benchmarks. “C + T” indicates models are pretrained on FlyingChairs(C) and FlyingThing(T) to test generalization performance. “+ S + K (+ H)” denotes the training data combining Sintel(S), KITTI(K) and HD1K(H). “+H” with brackets means it is optional for some works (Teed and Deng 2020; Hui, Tang, and Loy 2018). “*” denotes the results with warm-start testing (Teed and Deng 2020). The best results are marked in **bold** for better comparison.

the overall graph interaction model compatible with existing flow networks. Then the updated feature maps contain both global contextual information and local pixel-wise matching cost, which are properly used to make a better prediction for the flow field.

In particular, we formulate the reverse projection as

$$\hat{\mathbf{f}} = \mathcal{P}_{\mathbf{v} \rightarrow \mathbf{f}}(\hat{\mathbf{v}}), \quad (6)$$

where $\mathcal{P}_{\mathbf{v} \rightarrow \mathbf{f}}(\cdot)$ is a linear combination function that map the node vectors $\hat{\mathbf{v}} \in \mathbb{R}^{C \times K}$ to the feature maps $\hat{\mathbf{f}} \in \mathbb{R}^{C \times N}$ in the original grid space of flow network. In practice, we reuse the projection matrix in the node embedding procedure. The projection matrix contains pixel-to-node assignments and preserves the spatial details, which is crucial for recovering the resolution of feature maps. Besides, no additional parameters are involved by reusing the region assignments, which also helps to reduce computational overhead.

The context feature $\hat{\mathbf{f}}_c$ is produced by a residual operation as

$$\hat{\mathbf{f}}_c = \mathbf{f}_c + \alpha \mathcal{P}_{\mathbf{v} \rightarrow \mathbf{f}}(\mathbf{v}_c), \quad (7)$$

where α denotes a learnable parameter that is initialized as 0 and gradually performs a weighted sum. Similarly, motion feature $\hat{\mathbf{f}}_m$ is produced by

$$\hat{\mathbf{f}}_m = \mathbf{f}_m + \beta \mathcal{P}_{\mathbf{v} \rightarrow \mathbf{f}}(\mathbf{v}_m). \quad (8)$$

Given the enhanced feature $\hat{\mathbf{f}}_c$ and $\hat{\mathbf{f}}_m$, one potential hurdle for feature fusion is that context feature lacks correspondence information for cross-image matching, which could lead to a shift of global displacement and thus affect the flow accuracy. Therefore, we devise an attentive fusion function,

which first learns to predict a set of scale weights from motion feature $\hat{\mathbf{f}}_m$ and then leverages them to implement global adjustment for entire dense displacement. Specifically, the attentive fusion function is defined as

$$\mathbf{f}_o = (1 + \mathcal{F}_{\text{CA}}(\hat{\mathbf{f}}_m)) \hat{\mathbf{f}}_c \oplus \hat{\mathbf{f}}_m, \quad (9)$$

where \oplus is a concatenation operation, and $\mathcal{F}_{\text{CA}}(\cdot)$ is a channel attention function (Hu et al. 2020), here implemented with two convolutions with ReLU and sigmoid activation.

Experimental Results

Datasets and Evaluation Metrics

We conduct extensive experiments on two standard datasets, *i.e.*, MPI-Sintel (Butler et al. 2012) and KITTI 2015 (Menze and Geiger 2015). We follow prior works (Teed and Deng 2020; Jiang et al. 2021b) to utilize two standard evaluation metrics, *i.e.*, average end-point error (EPE) and the percentage of erroneous pixels > 3 pixels (F1-all), to evaluate the performance of predicted optical flow.

Implementation Details

The implementation of our approach is based on PyTorch toolbox. In our model, we set the number of context and motion nodes K to 128. The state updating iterations t are set to 2 and 1 for context and motion graph, respectively.

During training, we follow prior works (Teed and Deng 2020; Jiang et al. 2021a) to adopt AdamW optimizer with one-cycle learning rate policy, and conduct model pretraining on synthetic data as the standard optical flow training

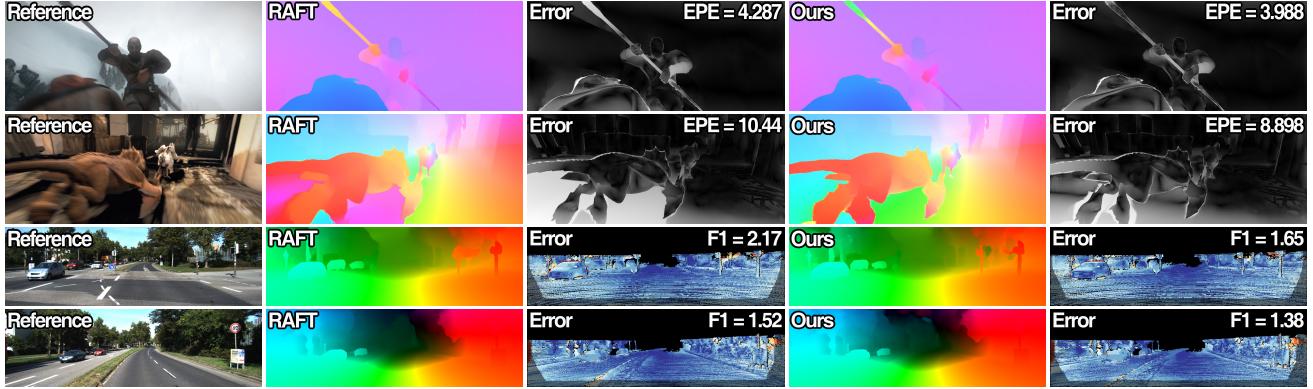


Figure 4: Qualitative comparisons with RAFT (Teed and Deng 2020) on Sintel and KITTI test set. All results are provided by the official website of each dataset. Best viewed in color.

procedure. The model is pretrained on FlyingChairs (Dosovitskiy et al. 2015) for 180k iterations and then on FlyingThings (Mayer et al. 2016) for 180k iterations. After that, we fine-tune the model on combined data from Sintel (Butler et al. 2012), KITTI-2015 (Menze and Geiger 2015), and HD1K (Kondermann et al. 2016) for 180k iterations, and then submit the flow prediction to Sintel server for online evaluation. Finally, additional 50k iterations of finetuning are performed on KITTI-2015 (Menze and Geiger 2015) for KITTI online evaluation. Our model is trained on 2 NVIDIA GeForce GTX 2080Ti GPUs, and the batch size is set to 8 for better leveraging the GPU memory.

Comparison with State-of-the-Arts

Results on Sintel. On the training set of FlyingChairs (C) + FlyingThings (T), as shown in Tab. 1, our approach achieves an average EPE of 1.31 on clean pass of Sintel dataset, which is competitive with SCV (Jiang et al. 2021b) and lower than the well-known RAFT (Teed and Deng 2020) by 8.4% ($1.43 \rightarrow 1.31$). On final pass, it obtains a score of 2.69 in EPE, outperforming previous state-of-the-art methods SCV and RAFT by 8.8% ($2.95 \rightarrow 2.69$) and 0.7% ($2.71 \rightarrow 2.69$), respectively. The results demonstrate the good cross dataset generalization of our model.

On Sintel test set, we follow prior works (Hui, Tang, and Loy 2018; Hur and Roth 2019; Teed and Deng 2020) to submit the predicted flow to the official server for online evaluation. Our AGFlow achieves an EPE of 1.68 on Sintel clean pass, which surpasses top-ranked methods RAFT (Teed and Deng 2020) and SCV (Jiang et al. 2021b) by 13.4% ($1.94 \rightarrow 1.68$) and 2.3% ($1.72 \rightarrow 1.68$), respectively. Besides, it obtains EPE = 2.83 on final pass, outperforming recent SCV by a large margin (21.4%). When utilizing the warm-start strategy, our approach sets new records of 1.43 EPE on clean pass and 2.47 on final pass, which significantly surpasses the previous best results by 11.2% ($1.61 \rightarrow 1.43$) and 13.6% ($2.86 \rightarrow 2.47$), respectively.

Fig. 4 (line 1 and 2) provides some qualitative comparisons with RAFT (Teed and Deng 2020) on the challenging final pass of Sintel dataset, which demonstrates that our AGFlow is able to fully exploit scene context to effectively assist motion estimation with shape and region constraints,

leading to achieving more accurate flow fields with clear motion boundaries.

Results on KITTI. We also provide the results of our approach on KITTI-15 dataset. As in Tab. 1, when training on C + T, our AGFlow achieves an average EPE of 4.82 and F1-all score of 17.0% on KITTI-15 validation set, which significantly surpass recent method SCV (Jiang et al. 2021b) by 29.1% ($6.80 \rightarrow 4.82$) and 11.9% ($19.3 \rightarrow 17.0$), respectively. For online evaluation, our approach achieve new state-of-the-art performance of 4.89% in F1-all, outperforming top-ranked methods SCV (Jiang et al. 2021b) and RAFT (Teed and Deng 2020) by 20.7% ($6.17 \rightarrow 4.89$) and 4.1% ($5.10 \rightarrow 4.89$), respectively. Some qualitative comparisons with RAFT (Teed and Deng 2020) on KITTI dataset are illustrated in Fig. 4 (line 3 and 4), which shows that the proposed global reasoning conditioned on scene context helps to decrease the uncertainty of ambiguous matching in some tough regions.

Ablation Analysis

Comparison with Grid Feature Enhancement. We first compare the proposed AGFlow with widely-used methods for optical flow that enhance features in regular grid space, including RAFT (Teed and Deng 2020), dense and dilated convolutions (Sun et al. 2018; Hur and Roth 2019). As shown in Tab. 2, dense and dilated convolutions slightly improve the flow accuracy with heavy model complexity. In contrast, our AGFlow achieves better performance, yet only needs additional 0.30 M parameters, reducing parameters by around 90%. This demonstrates that the proposed low dimensional graph reasoning scheme is effective to boost the flow accuracy in an efficient manner.

Comparison with SuperGlue (Sarlin et al. 2020). SuperGlue is a well-known method that employs graph neural network for feature matching. As can be seen in Tab. 2 (line 4), the original SuperGlue on grid feature requires a large amount of GPU memory, which is out of range with general settings for model training. Thus we re-implement SuperGlue with our low dimensional graph model (termed G-SuperGlue). Compared with G-SuperGlue, our AGFlow not only achieves considerable performance gain (8.0%), but

Method	Param (M) ↓	Sintel (val) EPE ↓	
		Clean	Final
Grid Feature	RAFT	5.26	3.04
	+ Dense Convs	+ 3.34	1.63
	+ Dilated Convs	+ 0.85	1.66
Graph Space	+ SuperGlue	+ 2.99	-
	GCU (Base)	5.44	1.76
	+ G-SuperGlue	+ 1.82	1.63
	+ AGR (ours)	+ 0.12	1.50
			2.88

Table 2: Quantitative comparisons with related methods (refer to Sec. for more details). We set RAFT (Teed and Deng 2020) as the baseline model for the method with regular grid space enhancement. The methods in each part are plugged into the same baseline and trained on C + T (180k) for fair comparison.

Settings	FLOPs (G)	Param (M)	Sintel (val) EPE	
			Clean	Final
Graph Reasoning	Base Graph	381.06	5.44	1.76
	+ SGR (no GA)	+ 16.95	+ 0.07	1.65
	+ AGR	+ 17.09	+ 0.12	1.50
Node numbers	<u>K</u> = 32	+ 3.09	+ 0.04	1.61
	K = 64	+ 5.88	+ 0.07	1.56
	<u>K</u> = 128	+ 17.09	+ 0.12	1.50
	K = 256	+ 61.84	+ 0.2	1.49
Attentive readout	On	+ 17.09	+ 0.12	1.50
	Off	+ 16.82	+ 0.10	1.55
				2.94

Table 3: Ablation analysis for different settings of our AGFlow. “SGR” indicates separated graph reasoning for context and motion nodes (*i.e.*, without graph adapter), and “AGR” denotes overall adaptive graph reasoning. All methods are trained on C + T (180k) for fair comparison. Underline indicates the default settings in our model.

also reduces parameters by 85%. This is because our adaptive graph reasoning scheme is simple yet effective, and capable of fast transferring the region and shape prior from scene context to motion nodes in a one-shot manner.

Comparison with GCU (Li and Gupta 2018). We also compare our AGFlow with the basic region-based graph reasoning model (Li and Gupta 2018). Since GCU projects all feature maps into a single type of node, it requires fewer parameters than other methods. However, as we mention above, the representation gap between context and motion feature hinders the effectiveness of relation reasoning with a simple graph, thus resulting in a performance drop on both passes of Sintel. We regard it as the base graph model, as in line 5 of Tab. 2. In contrast, we carefully project feature maps into context and motion nodes, and further propose an adaptive graph reasoning approach to perform the *task-specific* hybrid reasoning, allowing our model to significantly reduce the average end-point error by around 14.8%.

Effectiveness of Adaptive Graph Reasoning. In Tab. 3, we empirically analyze the computational cost and corresponding performance gain of the core component in the proposed AGFlow. Using separated context and motion reasoning based on graph model boosts the performance by

Method	Param (M) ↓	Time (ms) ↓
RAFT	5.26 (-)	86.9 (-)
GMA	5.8 (+ 0.54)	113.8 (+ 26.9)
AGFlow	5.56 (+ 0.30)	90.7 (+ 3.8)

Table 4: Computational comparisons with state-of-the-arts on a single Geforce RTX 2080Ti GPU.

5.4% ~ 6.3%. Besides, we further incorporate the proposed graph adapter into the context-to-motion interaction, and then yield the proposed adaptive graph reasoning (AGR) module, which brings about 9% additional performance gain with only 0.14 G FLOPs and 0.05 M parameters for extra computational overhead.

Ablation for Node Numbers. We empirically show the influences of node numbers K in our graph model. As shown in Tab. 3, when more nodes are used (32 → 64 → 128), the average end-point error are gradually decreased from 1.61 on Sintel clean pass and 3.02 on final pass to 1.50 and 2.88, respectively. However, if furthermore nodes are involved (128 → 256), the flow accuracy almost remains the same and the computational overhead is largely increased by 2.6 times. This is because some redundant feature representations are generated with nodes, which brings no benefit to flow estimation. Therefore, we set $K = 128$ to ensure a good balance between efficiency and performance.

Ablation for Attentive Readout. We also test the influence of attentive readout compared with regular readouts (Li and Gupta 2018) in Tab. 3. As can be seen, incorporating it into our model brings about 3% in performance gain and only requires negligible computation cost and parameters, demonstrating the cost-effective property of this component.

Runtime Comparison. We provide the parameters and runtime of state-of-the-art methods in Tab. 4. Compared with GMA (Jiang et al. 2021a), our AGFlow can achieve competitive performance while reducing 0.24 M parameters. Besides, the inference speed is boosted by 20.3% (113.8 → 90.7). The comparisons clearly demonstrate the effectiveness of our AGFlow.

Conclusion

In this paper, we present a novel graph-based approach termed adaptive graph reasoning for optical flow (AGFlow), which performs global reasoning to explicitly emphasize scene context and motion dependencies for flow estimation. The key idea is adaptive graph reasoning, which intends to fast enhance the feature representation of motion nodes conditioned on the global context with shape and boundary. Comprehensive experiments demonstrate that our AGFlow is effective and flexible to alleviate the matching ambiguities in challenging scenes, and sets new records in two standard flow benchmarks. We hope our work will offer a fresh perspective in re-thinking the design of optical flow models.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) No.61872067.

References

- Bai, M.; Luo, W.; Kundu, K.; and Urtasun, R. 2016. Exploiting semantic information and deep matching for optical flow. In *ECCV*.
- Bar-Haim, A.; and Wolf, L. 2020. ScopeFlow: Dynamic Scene Scoping for Optical Flow. In *CVPR*.
- Bertinetto, L.; Henriques, J. F.; Valmadre, J.; Torr, P. H. S.; and Vedaldi, A. 2016. Learning feed-forward one-shot learners. In *NeurIPS*.
- Black, M. J.; and Anandan, P. 1993. A framework for the robust estimation of optical flow. In *ICCV*.
- Brox, T.; Bruhn, A.; Papenberg, N.; and Weickert, J. 2004. High Accuracy Optical Flow Estimation Based on a Theory for Warping. In *ECCV*.
- Bruhn, A.; Weickert, J.; and Schnörr, C. 2005. Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods. *IJCV*, 61: 211–231.
- Butler, D.; Wulff, J.; Stanley, G.; and Black, M. 2012. A Naturalistic Open Source Movie for Optical Flow Evaluation. In *ECCV*.
- Chen, Z.; Zhou, Z.; Huang, J.; Zhang, P.; and Li, B. 2020. Frame-guided region-aligned representation for video person re-identification. In *AAAI*.
- Dosovitskiy, A.; Fischer, P.; Ilg, E.; Häusser, P.; Hazirbas, C.; Golkov, V.; Smagt, P. V. D.; Cremers, D.; and Brox, T. 2015. FlowNet: Learning Optical Flow with Convolutional Networks. In *ICCV*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.
- Horn, B. K. P.; and Schunck, B. G. 1981. Determining Optical Flow. *Artif. Intell.*, 17: 185–203.
- Hu, J.; Shen, L.; Albanie, S.; Sun, G.; and Wu, E. 2020. Squeeze-and-Excitation Networks. *TPAMI*, 42: 2011–2023.
- Hui, T.-W.; Tang, X.; and Loy, C. C. 2018. LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation. In *CVPR*.
- Hur, J.; and Roth, S. 2019. Iterative Residual Refinement for Joint Optical Flow and Occlusion Estimation. In *CVPR*.
- Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; and Brox, T. 2017. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In *CVPR*.
- Jiang, S.; Campbell, D.; Lu, Y.; Li, H.; and Hartley, R. 2021a. Learning to Estimate Hidden Motions with Global Motion Aggregation. In *ICCV*.
- Jiang, S.; Lu, Y.; Li, H.; and Hartley, R. 2021b. Learning Optical Flow from a Few Matches. In *CVPR*.
- Kipf, T.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.
- Kondermann, D.; Nair, R.; Honauer, K.; Krispin, K.; Andrulis, J.; Brock, A.; Güssfeld, B.; Rahimimoghaddam, M.; Hofmann, S.; Brenner, C.; and Jähne, B. 2016. The HCI Benchmark Suite: Stereo and Flow Ground Truth with Uncertainties for Urban Autonomous Driving. In *CVPRW*.
- Li, Y.; and Gupta, A. 2018. Beyond Grids: Learning Graph Representations for Visual Recognition. In *NeurIPS*.
- Luo, A.; Li, X.; Yang, F.; Jiao, Z.; Cheng, H.; and Lyu, S. 2020. Cascade graph neural networks for rgb-d salient object detection. In *ECCV*.
- Luo, A.; Yang, F.; Li, X.; Li, Y.; Jiao, Z.; Cheng, H.; and Lyu, S. 2021. Robust Scene Parsing by Mining Supportive Knowledge From Dataset. *IEEE Transactions on Neural Networks and Learning Systems*.
- Mayer, N.; Ilg, E.; Häusser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; and Brox, T. 2016. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *CVPR*.
- Menze, M.; and Geiger, A. 2015. Object scene flow for autonomous vehicles. In *CVPR*.
- Mohamed, A.; Qian, K.; Elhoseiny, M.; and Claudel, C. 2020. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *CVPR*.
- Park, J.; Lee, J.; and Sohn, K. 2021. Bridge To Answer: Structure-Aware Graph Interaction Network for Video Question Answering. In *CVPR*.
- Ranjan, A.; and Black, M. J. 2017. Optical flow estimation using a spatial pyramid network. In *CVPR*.
- Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2020. SuperGlue: Learning Feature Matching With Graph Neural Networks. In *CVPR*.
- Shen, Y.; Li, H.; Yi, S.; Chen, D.; and Wang, X. 2018. Person re-identification with deep similarity-guided graph neural network. In *ECCV*.
- Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2018. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In *CVPR*.
- Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2020. Models Matter, So Does Training: An Empirical Study of CNNs for Optical Flow Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42: 1408–1423.
- Teed, Z.; and Deng, J. 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *ECCV*.
- Vihlman, M.; and Visala, A. 2020. Optical Flow in Deep Visual Tracking. In *AAAI*.
- Wang, J.; Zhong, Y.; Dai, Y.; Zhang, K.; Ji, P.; and Li, H. 2020. Displacement-Invariant Matching Cost Learning for Accurate Optical Flow Estimation. In *NeurIPS*.
- Wei, X.; Yu, R.; and Sun, J. 2020. View-GCN: View-based graph convolutional network for 3D shape analysis. In *CVPR*.
- Weinzaepfel, P.; Revaud, J.; Harchaoui, Z.; and Schmid, C. 2013. DeepFlow: Large displacement optical flow with deep matching. In *ICCV*.
- Xu, R.; Li, X.; Zhou, B.; and Loy, C. C. 2019. Deep Flow-Guided Video Inpainting. In *CVPR*.
- Yang, G.; and Ramanan, D. 2019. Volumetric Correspondence Networks for Optical Flow. In *NeurIPS*.
- Yin, Z.; Darrell, T.; and Yu, F. 2019. Hierarchical Discrete Distribution Decomposition for Match Density Estimation. In *CVPR*.

Zhao, C.; Feng, C.; Li, D.; and Li, S. 2020a. OF-MSRN: optical flow-auxiliary multi-task regression network for direct quantitative measurement, segmentation and motion estimation. In *AAAI*.

Zhao, S.; Sheng, Y.; Dong, Y.; Chang, E.; and Xu, Y. 2020b. MaskFlownet: Asymmetric Feature Matching With Learnable Occlusion Mask. In *CVPR*.