

# CGCV: Context Guided Correlation Volume for Optical Flow Neural Networks

Jiangpeng Li Yan Niu

Jilin University

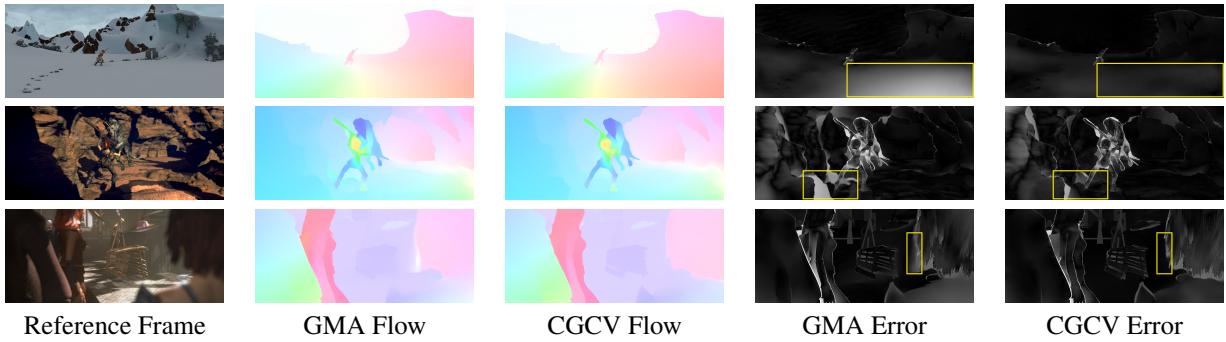


Figure 1: We propose to guide the traditional outlier-prone correlation volume by context information. By integrating the proposed Context Guided Correlation Volume with the framework of GMA [20], flow computation accuracy can be successfully improved in challenging cases (see the boxed regions). From top to bottom: Inputs and results for test sequences *Mountain 2*, *Tiger* and *Perturbed Market 3*, the first one is from Sintel (test) Clean dataset, and the last two are from Sintel (test) Final dataset.

## Abstract

Optical flow, which computes the apparent motion from a pair of video frames, is a critical tool for scene motion estimation. Correlation volume is the central component of optical flow computational neural models. It estimates the pairwise matching costs between cross-frame features, and is then used to decode optical flow. However, traditional correlation volume is frequently noisy, outlier-prone, and sensitive to motion blur. We observe that, although the recent RAFT algorithm also adopts the traditional correlation volume, its additional context encoder provides semantically representative features to the flow decoder, implicitly compensating for the deficiency

of the correlation volume. However, the benefits of this context encoder has been barely discussed or exploited. In this paper, we first investigate the functionality of RAFT’s context encoder, then propose a new Context Guided Correlation Volume (CGCV) via gating and lifting schemes. CGCV can be universally integrated with RAFT-based flow computation methods for enhanced performance, especially effective in the presence of motion blur, de-focus blur and atmospheric effects. By incorporating the proposed CGCV with previous Global Motion Aggregation (GMA) method, at a minor cost of 0.5% extra parameters, the rank of GMA is lifted by 23 places on KITTI 2015 Leader Board, and 3 places on Sintel Leader Board. Moreover, at a similar model size, our correlation volume

achieves competitive or superior performance to state of the art peer supervised models that employ Transformers or Graph Reasoning, as verified by extensive experiments.

**Keywords:** *Optical flow, Semantic Context, Correlation Volume, Cross Attention*

## 1 Introduction

Given a video sequence, Optical Flow establishes the correspondence between the projection positions of the same scene point onto two consecutive frames. As optical flow computation does not demand information about the camera or the scene, it is probably the only approach to obtain motion clue in applications where only video frames are available, such as graphical rendering [46], augmented reality [24], facial image registration [31], video stabilization [51, 23], action recognition [9, 41, 44, 8, 35].

Due to its importance, optical flow computation has been intensively investigated since 1980, traditionally modeled by Partial Different Equations [5, 29, 30, 22], Variation Regularization [10, 52, 45, 37], and Statistical Learning [4, 38, 34] etc. These models can be viewed as matching pre-described low-level features in a continuous space. Commonly, they focus on explicitly defining discriminative features that are invariant to motion, based on human observations on the distribution nature of video data. The feature invariance and flow smoothness principles are finally modelled by flow interpolation schemes.

In recent years, research efforts have been largely dedicated to matching high-level features extracted by deep neural models [18, 32, 39, 42, 43, 53, 7, 54, 14, 20, 36, 21, 12, 19, 26, 56, 48, 57, 3]. Since Hosni et al introduced the Cost Volume to flow computation [11], modern parameterized flow computation neural networks commonly perform feature matching by a 4D cost volume (or the dual Correlation Volume), which estimates the affinity between cross-frame pairs of features locally or globally. Entries of an ideal correlation volume should be and only be large at true positive correspondence positions. Unfortunately, the correlation volume is generally noisy and outlier-prone, frequently causing matching ambiguity in the presence of motion blur or fast motion. A solution is to down-weight ambiguous matches. To address this issue, in the scope of general multi-image matching, pre-

vious works perform  $L^2$  normalization to matching features (i.e., the ones that generate the correlation volume<sup>1</sup>) and the correlation volume [28, 43, 33]; Particularly for optical flow computation, LiteFlowNet3 trains the model to learn a modulator to adjust the correlation vectors and prohibit outliers [18]. MaskFlowNet trains the model to learn a mask to filter out occluded features [55]. In this paper, we propose a new correlation volume construction strategy, leveraging image brightness and semantic contexts.

The Recurrent All-Pairs Field Transforms (RAFT) establishes a new flow computation paradigm, in which an additional context branch is designed beside the traditional correlation volume branch [42]. RAFT shows that injecting the context features<sup>2</sup> to the flow decoder gains higher accuracy than merely using the correlation volume. In this paper, we analyze the rational behind the benefits of the context branch. We investigate the different behaviors of the Matching Features and Context Features by a thorough empirical study. Based on our observations, we propose guiding the traditional correlation volume by semantic context information. The proposed correlation volume is light-weight, and can be integrated with state-of-the-art flow computation neural networks at negligible extra cost. To show the effectiveness of our Context Guided Cost Volume (CGCV), we take the recent Global Motion Aggregation (GMA [20]) flow network as an example baseline. Comprehensive experiments verify that, by incorporating the proposed cost volume with GMA, the flow computation accuracy is significantly improved, especially in the presence of occlusion and motion blur. Compared to other GMA-based peer methods that utilize graph reasoning or Transformer techniques for performance enhancement, our method shows competitive or superior accuracy and simplicity, as verified by comprehensive experiments including the official Sintel and KITTI 2015 Optical Flow Leader Board evaluations.

In summary, our work contributes:

1. A new simple, light-weight, plug-and-play approach to constructing high quality correlation volume for flow computation.

---

<sup>1</sup>In this paper, by *matching features*, we mean the features that participate correlation volume construction.

<sup>2</sup>In this paper, by *context features*, we mean the features extracted by the context encoder designed by RAFT, unless otherwise specified.

2. An in-depth analysis on the barely noticed contrast between the behaviors of the matching features and context features of the RAFT flow computation framework.
3. Evidently improved flow computation performance over the traditional correlation volume counterpart.

## 2 Related Work

Even within the scope of neural models, the literature of optical flow is broad. This section focuses on the development of Correlation Volume in the framework of supervised learning, which is directly related to ours.

**Correlation Volume Construction** Hosni et al. first introduced the Cost Volume to optical flow computation [11], with each entry representing the cost of matching a pair of cross-frame yet spatially neighbouring pixels. The image correspondence is estimated from the cost volume in a “winner takes all” manner (*argmin*). The cost volume thus transforms the *image correspondence* problem into a *feature correspondence* problem. DCFlow extends the cost measurement from the distance in color and structure spaces to deep feature space, and estimates flow from the cost volume via Flow Semi-Global Matching [49]. FlowNet shows that the Cost (more precisely, Correlation) Volume can be constructed and decoded to the flow field in an end-to-end CNN architecture [18]. Inspired by the pyramidal flow refinement of SPyNet [32], LiteFlowNet [15, 16] and PWC-Net [39, 40] extend the image pyramid to feature pyramid, constructing a local correlation volume at each level to refine flow hierarchically. Hur-Roth unified the multi-level sub-networks to one encoder-decoder shared across all levels for iterative residual flow refinement [17]. Devon computes multi-scale local correlation volumes by sampling the target frame’s full-resolution feature map at various dilated factors, and concatenates them to decode flow [25]. Yang-Ramanan proposed a multi-channel correlation volume, each channel for a different feature embedding [50]. Although the aforementioned correlation volumes are local, they need to be re-computed (generally by vector inner products) each time the flow is refined. Differently, RAFT retrieves the updated local correlation values from a global all-pairs correlation volume, which remains con-

stant once constructed [42]. This strategy bypasses computing inner products between features along with flow refinement. State of the art works GMA [20], AGFlow [27], Consistency Imposition [19], KPA [26], DEQ [3] follow the same design of global correlation volume. CRAFT views the correlation volume of GMA-RAFT as a degenerated cross-attention, and extends it to a non-degenerated multi-head one [36]. FlowFormer tokenizes the cost volume to cost memory, from which the flow is decoded by Separable Self Attention Transformer [6] layers [13]. Global Matching methods GMFlow [48] and GMFlowNet [56] employ Transformers to extract high quality features. Transformers can largely enhance the correlation volume, but take tremendous graphical memory. Xu et al. used 1D attention and correlation in orthogonal directions to achieve 2D correspondence effect, reducing RAFT’s 4D correlation volume to a concatenation of two 3D volumes [47]. Different from previous works, our aim is to address the matching ambiguity suffered by currently most popular cost volume at minimal cost, especially for RAFT-based flow computation methods.

**Filtering Outliers** Although correlation volume acts as the base to most optical flow computation neural networks, it is actually vulnerable to factors such as motion blur, fast motion or large displacement, occlusion or disocclusion. To down-weight ambiguous matches, CNNGM [33], DGC-Net [28] and GLU-Net [43] perform vector  $L^2$  normalization to feature descriptors and/or channel-wise  $L^2$  normalization to the correlation volume. MaskFlowNet [55] learns a multiplicative occlusion mask to filter the warped features of the target frame. LiteFlowNet3 [14] modulates each cost vector by an affine transformation, whose coefficients are learned from the concatenation of a confidence map, the cost volume and the reference frame’s matching feature map. Seemingly, our method improves the correlation volume by modulation too. However, our formulation is fundamentally different from MaskFlowNet or LiteFlowNet3. First, based on a thorough empirical study on the context and matching features, our gating and lifting tensors are designed to take advantage of the context stream. Thus our modulation is independent of the correlation volume. Oppositely, MaskFlowNet and LiteFlowNet3 learn such tensors from the matching features and their correlations. Moreover, our modulation needs to be computed only once, hence its extra cost on parameters or FLOPs is very limited,

whereas MaskFlowNet and LiteFlowNet3 modulate the features or local correlation vectors along with flow refinement.

**Context Features** The terminology ‘‘Context Encoder’’ appears in several flow computation publications. However, they mean totally different operations in different publications. In PWC-Net [39], it is a flow post-processor. In GMFlowNet [56], it generates the features to be matched to form the correlation volume. In RAFT, the context encoder is independent to the correlation volume. It extracts features from the reference frame and inject them to the flow refinement units. Later, GMA applies part of RAFT’s context features to aggregate the motion features and shows flow accuracy improvement [20]. KPA [26] further applies kernel patch attention of the context features to modulate the motion features. This strategy effectively improves the quality of motion features, but as the attention is computed in each flow refinement unit, the required Floating Point Operations (FLOPs) are remarkably increased. Our paper analyzes the context encoder of RAFT by an in-depth empirical study, based on which, we design a new correlation volume guided by semantic contexts. The proposed volume remains constant during flow refinement, refraining from incurring additional parameters or FLOPs.

### 3 Approach

In this section, we first discuss the weakness of the commonly employed correlation volume, which is constructed from the matching features. We then analyze the barely noticed advantages of the context features in GMA-RAFT. We thoroughly investigate the rational for the different behavior of GMA-RAFT matching and context feature encoders. Based on these studies, we propose a new Context Guided Correlation Volume (CGCV). Finally, we describe the plug-and-play integration of CGCV with state-of-the-art flow computation neural models.

#### 3.1 Weakness of Traditional Correlation Volume

Let  $\mathbf{I}_1, \mathbf{I}_2 \in \mathbb{R}^{H \times W \times 3}$  be the reference and target frames of a video clip. For each pixel  $[i, j] \in \mathbb{N}^2 \cap \{[1, W] \times [1, H]\}$  in  $\mathbf{I}_1$ , we seek for its correspondence

point  $[i + u, j + v] \in \mathbb{R}^2 \cap \{[1, W] \times [1, H]\}$  in  $\mathbf{I}_2$ , such that  $\mathbf{I}_1[i, j]$  and  $\mathbf{I}_2[i + u, j + v]$  are the image projections of the same scene point<sup>3</sup>. To find the optical flow vector  $[u, v] \in \mathbb{R}^2$ ,  $\mathbf{I}_1[i, j]$  and  $\mathbf{I}_2[i + u, j + v]$  must share some common feature  $\mathbf{F}$  that is representative and invariant to  $[u, v]$ . That is,  $\mathbf{F}_1[i, j]$ , as the feature representation for  $\mathbf{I}_1[i, j]$ , equals to  $\mathbf{F}_2[i + u, j + v]$ . Moreover,  $\mathbf{F}$  should be an excluding representation, such that for any other displacement  $[\tilde{u}, \tilde{v}] \neq [u, v]$ ,  $\mathbf{F}_1[i, j] \neq \mathbf{F}_2[i + \tilde{u}, j + \tilde{v}]$ .

Deep neural methods learn such feature representation by deep networks. For example, RAFT [42] designs an encoder  $g(\cdot; \theta_g)$ <sup>4</sup> to extract matching features  $g(\mathbf{I}_1), g(\mathbf{I}_2) \in \mathbb{R}^{n \times \frac{H}{8} \times \frac{W}{8}}$ , where  $n$  is the feature length. The cross-frame all-pairs correlations between  $g(\mathbf{I}_1)$  and  $g(\mathbf{I}_2)$  form a correlation pyramid, whose first level is a correlation volume  $\mathbf{C}$  computed by

$$\mathbf{C}[i, j, k, l] = \frac{1}{\sqrt{n}} \langle g(\mathbf{I}_1)[i, j, 1:n], g(\mathbf{I}_2)[k, l, 1:n] \rangle, \quad (1)$$

where symbol  $\langle \cdot, \cdot \rangle$  stands for the inner product of two vectors;  $i, k \in [1, W/8]$ ,  $j, l \in [1, H/8]$ . Pooling this correlation volume at different scales builds the correlation pyramid, from which, a correlation feature is retrieved for each point  $[i, j]$ . The correlation feature map is one of the main components to decode the flow field  $\hat{\mathbf{U}}, \hat{\mathbf{V}} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8}}$ . This flow computation framework is widely followed by recent works (e.g., GMA [20], AGFlow [27], CRAFT [36]).

Ideally, for each point  $\mathbf{x} = [i, j]$ , its correlation plane  $\mathbf{C}[i, j, 1 : \frac{W}{8}, 1 : \frac{H}{8}]$  should reach the maximum value at its true correspondence  $\mathbf{x}'$ , and should have a significantly smaller correlation value at any other point  $\tilde{\mathbf{x}} \neq \mathbf{x}'$ . However, in the presence of occlusion/disocclusion, motion blur, fast motion, etc., the resulted correlation planes may be rather noisy and outlier prone: often exhibiting large correlation values at many false correspondence positions, or small correlation value at the true correspondence, leading to matching ambiguity. Fig. 2. a-c and Fig. 2. e illustrate such correlation planes by two example points on the reference frame of sequence *Temple* of

<sup>3</sup>Note that  $u, v$  vary with  $[i, j]$ , for conciseness, we abbreviate  $u[i, j]$  as  $v[i, j]$  as  $u$  and  $v$ .

<sup>4</sup>To differentiate the training and inference passes of  $g$ , we let  $g(\cdot)$  denote the feed-forward computation with parameters to be trained;  $g(\cdot; \theta_g)$  denote the computation with trained and fixed parameters  $\theta_g$ . Same for the upcoming functions  $h(\cdot; \theta_h)$  and  $c(\cdot; \theta_c)$ .

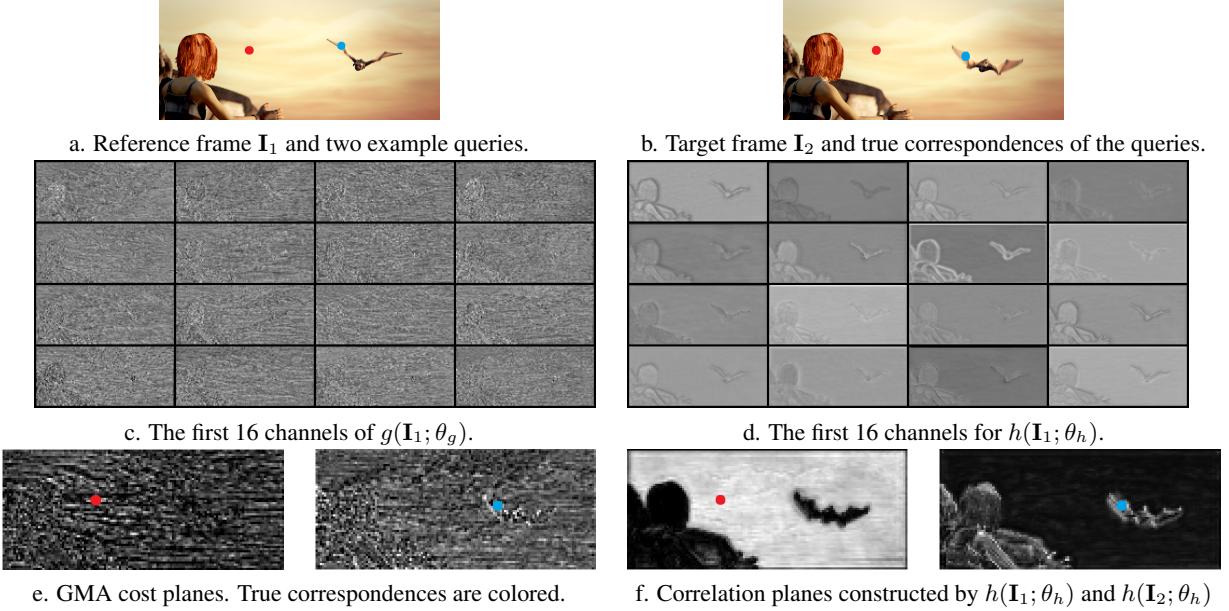


Figure 2: A comparison between the GMA correlation planes and the correlation planes constructed from the context features, demonstrated by two query points indicated by red and blue dots. The reference and target frame are from sequence *Temple* of Sintel (train) Final benchmark dataset. Dots of the same color on the figures represent a pair of true correspondence points.

benchmark dataset Sintel (train) Final, using the trained GMA model. In this example,  $g(\mathbf{I}_1; \theta_g)$  fails to extract representative matching features from the input frames, and their consequent cross-correlation map is chaotic.

### 3.2 What Makes an Encoder Extract Contexts

To improve the quality of the traditional correlation volume, we pursue a more reliable clue. We notice that, in striking contrast to the matching features, the context features extracted by GMA-RAFT context encoder  $h(\cdot; \theta_h)$  are of much higher quality, and the context cross-frame correlation are significantly sharper and cleaner (see Fig. 2. d and Fig. 2. f).

An intuitive explanation for their different behaviors is that the context encoder shares features with the flow refinement GRUs, while the matching feature encoder does not. Nevertheless, it should be further noted that different parts of the context features are shared with the GRUs dif-

ferently in GMA-RAFT. The first 128 channels, termed as “net” in the released code of RAFT, initialize the hidden state of the first GRU; The rest, termed as “inp”, are input to GRUs. GMA additionally uses the “inp” part for the self-attention mechanism to aggregate global motion. For clarity, we denote the two halves by  $h^{\text{net}}(\mathbf{I}_1)$  and  $h^{\text{inp}}(\mathbf{I}_1)$ . Hereafter, we name  $g(\mathbf{I})$  the “matching” features (as they are used for computing matching costs) and  $h(\mathbf{I})$  the “context” features (as in RAFT [42]).

To inspect how context information is extracted, we carried out an empirical study by removing the context stream from GMA and using the matching features for replacement. In particular, we split  $g(\mathbf{I}_1)$  to two halves  $g^{\text{net}}(\mathbf{I}_1)$  and  $g^{\text{inp}}(\mathbf{I}_1)$ , which replace  $h^{\text{net}}(\mathbf{I}_1)$  and  $h^{\text{inp}}(\mathbf{I}_1)$  respectively. We train this experimental model from scratch and denote the learned parameters of  $g$  by  $\check{\theta}_g$ . Fig. 3 presents the first 16 feature channels of  $g^{\text{net}}(\mathbf{I}_1; \check{\theta}_g)$  and  $g^{\text{inp}}(\mathbf{I}_1; \check{\theta}_g)$  extracted from the example image *Temple*. Overall,  $g(\mathbf{I}_1; \check{\theta}_g)$  performs significantly better than  $g(\mathbf{I}_1; \theta_g)$  at extracting image details. More particularly,

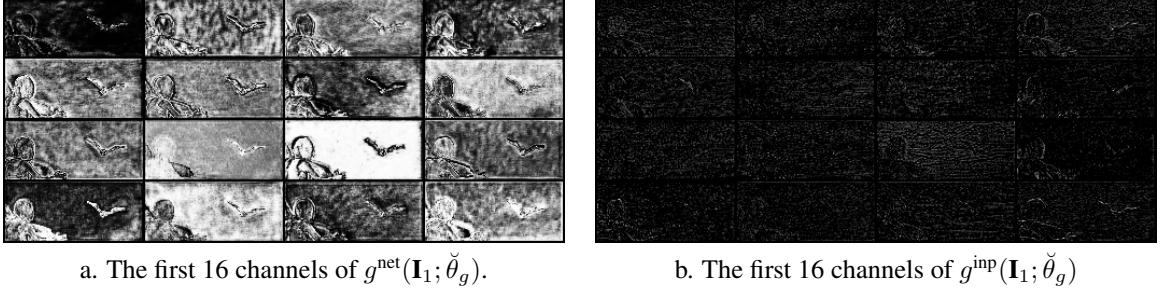


Figure 3: By sharing matching features  $g$  with the flow refinement GRUs, the “net” half of the features become contextual. Here the first 16 (out of totally 128) channels of  $g^{\text{net}}(\mathbf{I}_1; \check{\theta}_g)$  and  $g^{\text{inp}}(\mathbf{I}_1; \check{\theta}_g)$  extracted from the reference frame of sequence *Temple* are presented. See main text for details.

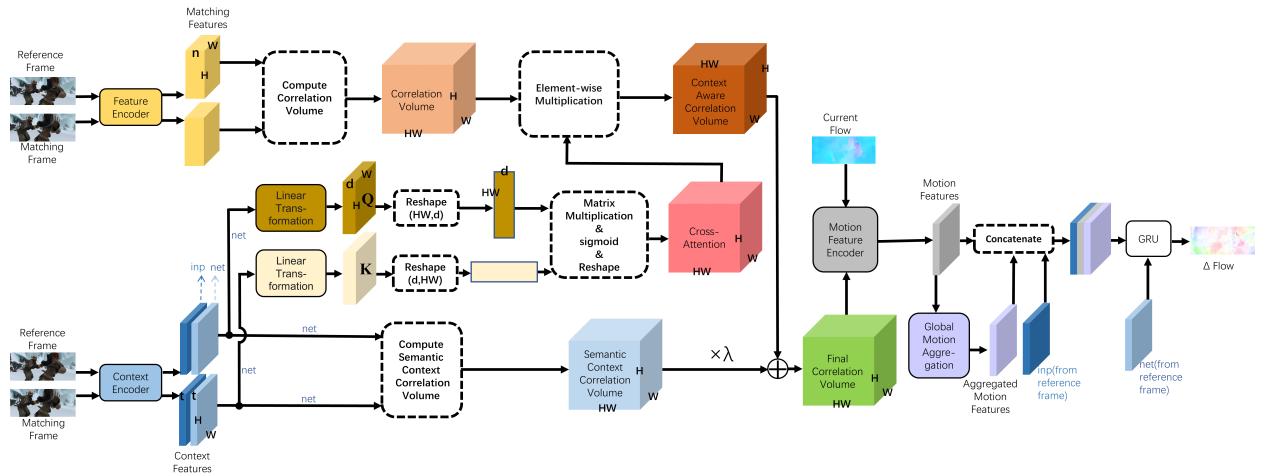


Figure 4: The architecture of our flow computation model, which integrates the proposed CGCV with GMA.

$g^{\text{net}}(\mathbf{I}_1; \check{\theta}_g)$  provides image semantic grouping proposals, although being less smooth than  $h^{\text{inp}}(\mathbf{I}_1; \theta_h)$ ;  $g^{\text{inp}}(\mathbf{I}_1; \check{\theta}_g)$  detects image brightness variation, with much stronger contrast than  $g(\mathbf{I}_1; \theta_g)$ . This experiment reveals that, the model training process drives the initial hidden state of the GRU to learn image semantic contexts and drives the GRU inputs to learn image intensity contexts. Even though  $g$  operates on both frames, sharing only  $g(\mathbf{I}_1)$  with the GRUs would push the kernels of  $g$  to detect different levels of high frequency components of images.

### 3.3 Context Guided Correlation Volume

Based on the above investigation, we propose guiding the traditional correlation volume by mutual context information, via a cross-attention mechanism. Compared to adding a new modulator to the neural network, our method is advantageous at largely saving extra parameters for feature extraction.

**Gating Operation** We extend the context encoder in GMA-RAFT to a Siamese network  $c(\cdot)$  to extract features from both the reference and target frames. We partition  $c(\mathbf{I})$  to two halves  $c^{\text{net}}(\mathbf{I})$  and  $c^{\text{inp}}(\mathbf{I})$ .  $c^{\text{net}}(\mathbf{I}_1)$  is further used to initialize the hidden state of the flow refinement GRU;  $c^{\text{inp}}(\mathbf{I}_1)$  is used as the input to the GRUs.

We define the *Query* matrix  $\mathbf{Q}$  as a linear transformation of  $c^{\text{net}}(\mathbf{I}_1)$ , and the *Key* matrix  $\mathbf{K}$  as a linear transformation of  $c^{\text{net}}(\mathbf{I}_2)$ :

$$\mathbf{Q} = \mathbf{W}_q \cdot c^{\text{net}}(\mathbf{I}_1), \quad \mathbf{K} = \mathbf{W}_k \cdot c^{\text{net}}(\mathbf{I}_2), \quad (2)$$

where symbol  $\cdot$  stands for matrix product;  $\mathbf{W}_q$  and  $\mathbf{W}_k$  are learnable linear transformations. The *Query* and *Key* yield a cross-attention

$$\mathbf{A}[i, j, k, l] = \sigma \left( \frac{\langle \mathbf{Q}[i, j, 1 : d], \mathbf{K}[k, l, 1 : d] \rangle}{\sqrt{d}} \right), \quad (3)$$

where  $d$  is the feature length of  $\mathbf{Q}[i, j, :]$  and  $\mathbf{K}[k, l, :]$ . The normalization function  $\sigma$  for the an attention mechanism generally takes the form of *softmax* mapping in previous works. However, in our work the cross-attention  $\mathbf{A}$  serves to gate false positive correlation values between matching features, therefore a *sigmoid* function performs better (see Sec. 4.4 Ablation Study).

We define the *Value* matrix as the identity transformation of the traditional correlation volume  $\mathbf{C}$  constructed by Eq. 1. The element-wise production between  $\mathbf{A}$  and  $\mathbf{C}$  obtains a context aware correlation volume

$$\mathbf{M} = \mathbf{A} \odot \mathbf{C}. \quad (4)$$

**Lifting Operation** The above gating scheme can effectively screen false positives, leveraging the pairwise relevancy inferred from semantic context. Yet on the other hand, a true pair of correspondences may still underestimate their correlation, if the non-normalized matching features  $g(\mathbf{I}_1)[i, j]$  and  $g(\mathbf{I}_2)[i', j']$  have small magnitudes. In this situation, their inner product may cause  $\mathbf{M}[i, j, i', j']$  to be small relatively to correlation values estimated between false positives, since the gating strength of  $\mathbf{A}$  is bounded by 1. As a consequence, the flow decoder may still mistake a false correspondence with higher correlation as the true correspondence.

To lift the small correlation values between potential false negatives, we superimpose a scalar-weighted cross-frame correlation volume constructed by semantic context  $\mathbf{S}$  to the gated correlation volume  $\mathbf{M}$ :

$$\mathbf{S}[i, j, k, l] = \frac{\langle c^{\text{net}}(\mathbf{I}_1)[i, j, 1 : t], c^{\text{net}}(\mathbf{I}_2)[k, l, 1 : t] \rangle}{\sqrt{t}}, \quad (5)$$

$$\mathbf{V} = \mathbf{M} + \lambda \times \mathbf{S}, \quad (6)$$

where the scalar  $\lambda$  is a learned parameter initialised to zero. We observed that the trained  $\lambda$  is always of the order of  $10^{-2}$ , at different training stages. Thus the semantic context correlation volume  $\mathbf{S}$  only impacts  $\mathbf{V}$  if  $c^{\text{net}}(\mathbf{I}_1)[i, j, :]$  and  $c^{\text{net}}(\mathbf{I}_2)[k, l, :]$  exhibit strong similarity. Moreover, we find that  $\mathbf{S}$  enables the context attention to focus better (see our ablation study evidence Sec. 4.4).

**Plug-and-Play Integration** Eq. 6 defines the proposed Context Guided Correlation Volume (CGCV), which forms the first level of the correlation pyramid to sample the correlation features. It only needs to be computed once in the forward pass. Each time the flow field is refined, the new correlation features are retrieved from the correlation volume without recomputing feature inner products.

The proposed CGCV can be integrated with any RAFT-based optical flow neural networks, by simply replacing the traditional correlation volume, while remaining the rest architecture unaltered. In this work, we integrate CGCV with GMA. Fig. 4 depicts the whole computation diagram of our flow computation model.

## 4 Experiments

### 4.1 Settings

The hardware platform for our experiments is two Nvidia RTX 2080Ti graphics cards, and the software environment is PyTorch. To evaluate the effectiveness and efficiency of the proposed Context Guided Correlation Volume (CGCV), we adopt GMA as our baseline framework and replace its traditional correlation volume with CGCV. For fair evaluation, we set the matching radius  $r$  to 4 and the context feature length  $c$  (Eq. 5) to 256, the length of the inp and net features is set to 128, as same as GMA. The length of the query and key features  $d$  (Eq. 3) is also set to 128.

Our training and testing conform to the commonly used 4-phases procedure in recent literature of flow computation neural models [42, 20, 27, 36, 13]. At the first phase, our model is trained on FlyingChairs for 120k iterations with a batch size of 8. The second phase continues the training on FlyingThings3D for 120k iterations

Training Data	Method	Sintel (train)		KITTI-15 (train)		Sintel (test)		KITTI-15 (test)		Params
		Clean	Final	AEPE	F1-all (%)	Clean	Final	F1-all (%)		
C+T	DPCTF-F [7]	2.04	5.37	8.42	19.93	-	-	-	38.6M	
	DEQ-RAFT-H [3]	1.41	2.75	<b>4.38</b>	<b>14.9</b>	-	-	-	12.8M	
	RAFT+OCTC [19]	1.31	2.67	4.72	16.3	-	-	-	<b>5.3M</b>	
	GMA [20]	1.30	2.74	4.69	17.1	-	-	-	5.9M	
	AGFlow [27]	1.31	2.69	4.82	17.0	-	-	-	5.6M	
	CRAFT [36]	1.27	2.79	4.88	17.5	-	-	-	6.3M	
	FlowFormer-S [13]	<i>1.20</i>	<b>2.64</b>	4.57	16.6	-	-	-	6.2M	
	KPA-Flow [26]	1.28	2.68	<b>4.46</b>	<i>15.9</i>	-	-	-	5.8M	
	CGCV (ours)	<b>1.15</b>	2.70	4.61	16.6	-	-	-	5.9M	
C+T+ S+K+H	SABMFL[54]	-	-	-	-	4.48	4.77	7.68	-	
	DPCTF-F[7]	(0.81)	(1.16)	(1.34)	(7.3)	3.54	4.47	7.22	38.6M	
	DEQ-RAFT-H [3]	(0.70)	(1.21)	(0.61)	(1.4)	1.82	3.23	4.98	12.8M	
	RAFT+OCTC [19]	(0.73)	(1.23)	(0.67)	(1.7)	1.82	3.09	4.72	<i>5.3M</i>	
	GMFlow [48]	-	-	-	-	1.74	2.90	9.32	<b>4.7M</b>	
	GMFlowNet [56]	<b>(0.59)</b>	<b>(0.91)</b>	(0.64)	(1.5)	1.39	2.65	4.79	9.3M	
	GMA [20]	(0.62)	(1.06)	(0.57)	(1.2)	<i>1.39*</i>	<i>2.47*</i>	5.15	5.9M	
	AGFlow [27]	(0.65)	(1.07)	(0.58)	(1.2)	1.43*	2.47*	4.89	5.6M	
	CRAFT [36]	(0.60)	(1.06)	(0.58)	(1.3)	1.45*	2.42*	4.79	6.3M	
	KPA-Flow [26]	(0.60)	(1.02)	<b>(0.52)</b>	<b>(1.1)</b>	<b>1.35*</b>	<b>2.36*</b>	<b>4.60</b>	5.8M	
	CGCV (ours)	(0.61)	(1.06)	(0.56)	(1.3)	<b>1.35*</b>	2.43*	4.96	5.9M	

Table 1: Seven sets of performance comparisons about CGCV models with comparable methods, all fine-tuned at the same stage. Results of competing methods are as reported in their original publications. Smaller value means better performance. Bold numbers indicate the highest accuracy; Italic numbers indicate second-best accuracy. Parentheses indicate the ground truth is released to public. Symbol \* means using “warm start” [42].

with a batch size of 6. The third phase fine-tunes the pre-trained model on the comprehensive dataset composed of FlyingThings3D, Sintel (train), KITTI (train) and HD1K for 120k iterations with a batch size of 6. The forth phase further fine-tunes the phase-3 model on KITTI (train) for 50K iterations with a batch size of 6. Our learning rate schedule complies with GMA: the maximum learning rate is set to  $0.25 \times 10^{-3}$  for the first phase, and  $0.125 \times 10^{-3}$  for the other phases. For clarity, we call the models learned at phase 2-4 the *CT-trained*, *CTSKH-tuned* and *KITTI-tuned* models, if specification is necessary, hereafter in this paper.

The benchmark datasets and evaluation metrics in our experiments follow the literature convention. The CT-trained model is evaluated on the test datasets officially

split from Sintel (train) and KITTI-15 (train), to assess the generalization ability of our model. The CTSKH-tuned model is tested on Sintel (train) and Sintel (test). The KITTI-tuned model is tested on KITTI-15 (train) and KITTI-15 (test).

## 4.2 Quantitative Analysis

Table 1 lists flow computation neural methods published in recent couple of years. Compared to our method, DPCTF-F [7], DEQ-RAFT-H [3] and GMFlowNet [56] have significantly more parameters; KPA-Flow [26] requires notably more FLOPs; GMFlow [48] and GMFlowNet [56] formulate flow computation without refinement. Although these methods are not directly compa-

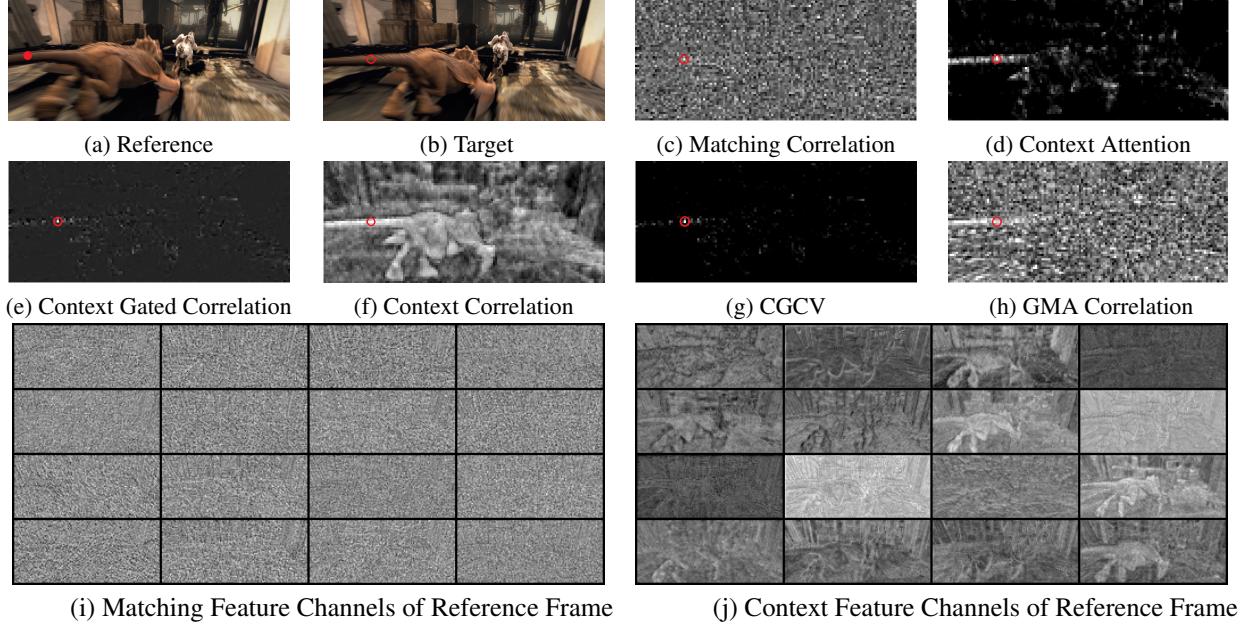


Figure 5: A visualization of how our regularization method improves the traditional correlation volume, taking an example *query* point (indicated as the red dot on the reference frame) from the tail of the dragon in test image *Market 4* of Sintel (test) Final. The visualization is conducted using the CTSKH-tuned CGCV. Red circles in sub-figures indicate the true correspondence positions for the query point. Sub-figure (g) shows the consistency between the maximum CGCV value position and the true correspondence position. In contrast, GMA correlation plane exhibits large values at many false correspondence positions. For clarity, the brightness of each involved map is normalized over the lattice. Better viewed digitally.

Method	Sintel (test)	Sintel (test)	KITTI-15
	Clean	Final	(test)
GMA[20]	12	11	71
AGFlow[27]	19	10	43
CRAFT[36]	21	7	33
CGCV (Ours)	9	8	48

Table 2: The MPI-Sintel and KITTI Leader Board ranking of the proposed CGCV and its comparable state of the art methods at the time of writing. Among competing methods, CGCV ranks the highest on the Clean pass, the second-highest on the Final pass, and the third on KITTI-15.

table to ours, they provide a frame of reference for top performance to the readers.

**Evaluation on Leader Boards.** Our model is evaluated by the MPI-Sintel [2] and KITTI-15 [1] Leader Boards by their test data and metrics. Among the 11 competing methods, ours ranks the 1st on Sintel (test) Clean, the 3rd on Sintel (test) Final, and the 5th on KITTI-15.

**Evaluation of Model Generalization** At the C+T training stage, with respect to other 8 top-performing methods, ours ranks the 1st place on Sintel (train) Clean, the 5th on Sintel (train) Final, and the 4th place on KITTI. This validates that our model generalizes well to unseen datasets.

**Evaluation of Fine-Tuned Models.** At the CTSKH fine-tuning stage, among the 9 compared methods, ours ranks the 2nd on KITTI-15 (train) in AEPE, the 3rd on Sintel (train) Final, and the 3rd on Sintel (train) Clean. Its accuracy score is very close to methods of similar complexity.

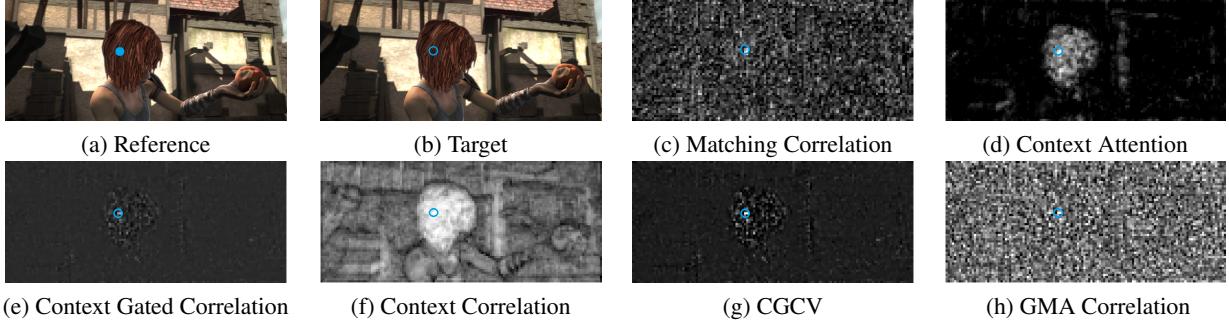


Figure 6: An example to visualize the development process of the proposed CGCV correlation volume. The query point is taken from the red hair in train image *Alley 1* of Sintel (train) Final (indicated as the blue dot on the reference frame). The visualization is conducted using the CTSKH-tuned CGCV. Blue circles in sub-figures indicate the true correspondence positions for the query point. Sub-figure (g) shows the consistency between the maximum CGCV value position and the true correspondence position. In contrast, GMA correlation plane exhibits large values at many false correspondence positions. For clarity, the brightness of each involved map is normalized over the lattice. Better viewed digitally.

Method	Chairs (val)	Things		Sintel	
		Clean (test)	Final (test)	Clean (train)	Final (train)
CGCV	<b>0.69</b>	2.60	2.39	<b>1.15</b>	<b>2.70</b>
removing context correlation volume	0.70	<b>2.47</b>	<b>2.35</b>	1.20	2.71
replacing sigmoid attention by softmax	0.73	2.95	2.69	1.32	3.09
removing context attention	0.81	3.02	2.61	1.30	2.74

Table 3: Ablation study on the proposed CGCV module, the main components of which are: a Context Cross-Frame Attention Module with a Sigmoid function, a Context Cross-Frame Correlation Volume. Bold numbers indicate the top accuracy. Slanted numbers indicate the second-best accuracy.

**Comparative Analysis** In the literature, GMA, AGFlow [27], CRAFT [36] adopt the same paradigm and have similar complexities<sup>5</sup>. Table 2 presents the MPI-Sintel and KITTI Leader Boards Ranking of CGCV and these comparable methods. In this comparison, CGCV achieves the best score on Sintel (test) Clean, the second-best score On Sintel (test) Final (slightly lower than CRAFT). We now

analyze our method with respect to these methods one by one. CGCV notably improves its baseline model GMA in all evaluations.

**CGCV vs GMA.** Our goal is to improve the traditional correlation volume in the presence of distracting factors. As our model replaces the correlation volume of GMA by the proposed one, a comparison with GMA would reveal the effectiveness of our correlation volume design. CGCV (5.91 M) has a similar size to GMA (5.88 M). At the CT-training stage, CGCV has higher accuracy than GMA in all 4 sets of comparisons. Especially on Sintel (train) Clean, CGCV reduces the AEPE of GMA by 11.53% (1.30 → 1.15) times. After CTSKH fine-tuning,

<sup>5</sup>Although KPA-Flow [26] is also based on GMA and have similar parameter counts, it runs substantially slower than the proposed model. For example on our machine, it takes KPA-Flow 0.30 seconds to compute the flow of a pair of KITTI images, whereas it take CGCV 0.20 seconds in identical settings. Hence here we do not compare to KPA-Flow.

CGCV and GMA show similar performance on both Sintel and KITTI. However, in Sintel and KITTI Leader Boards evaluation, CGCV shows clear superiority: it increases GMA by 23 places in KITTI Ranking.

**CGCV vs AGFlow.** Both AGFlow and our method aims at enhancing motion estimation by context information. While our method turns to semantic context cross-frame attention, AGFlow is based on graph reasoning. CGCV is only 0.3 M larger than AGFlow in model parameter scale, but CGCV scores better in 9 out of the total 11 sets of comparisons. Noticeably on Sintel (train) Clean, the CT-trained CGCV model has 12.21% ( $1.31 \rightarrow 1.15$ ) times lower AEPE; On Sintel (test) Clean, CGCV has 5.59% ( $1.43 \rightarrow 1.35$ ) times lower AEPE. CGCV is slightly less accurate than AGFlow by 1.43% ( $4.89 \rightarrow 4.96$ ) times on KITTI-15 (test). These experiments verify the competitiveness of our context cross-frame attention to graph reasoning, in leveraging context to enhance motion tokens.

**CGCV vs CRAFT.** Both CRAFT and our method adopt the framework of GMA. CRAFT applies Transformer to the matching features by dynamically aggregating mode attention, whereas we apply context cross-frame attention. CGCV is 0.4 M smaller in size. Numerically, CGCV shows superiority on 6 out of total 11 sets of comparisons. The most notable improvements occur at CT-training stage on Sintel (train) Clean and CT-SKH fine-tuning stage on Sintel (test) Clean, where the errors of CGCV are 9.45% ( $1.27 \rightarrow 1.15$ ) and 6.90% ( $1.45 \rightarrow 1.35$ ) times better. Moreover, on KITTI-15 (train), the AEPE and F1-all scores of CGCV are 5.53% ( $4.88 \rightarrow 4.61$ ) and 5.14% ( $17.5 \rightarrow 16.6$ ) times better. This indicates that CT-trained CGCV generalizes better to unseen inputs. The accuracy of CGCV is slightly 3.55% ( $4.79 \rightarrow 4.96$ ) times weaker on KITTI-15 (test). The comparative analysis shows that our context regulation strategy is competitive with CRAFT’s Transformer-based strategy in flow computation.

**CGCV vs FlowFormer-S.** FlowFormer is a highly effective deep neural model in Transformer architecture with 18.2 M parameters. To fairly compare to GMA, a small version FlowFormer-S, which has 6.2 M parameters, is trained on C+T without fine-tuning and evaluated in [13]. Compared to FlowFormer-S, CGCV is 0.3 M smaller in size. CGCV exhibits superiority by 4.16% ( $1.20 \rightarrow 1.15$ ) times smaller AEPE on Sintel (train)

Clean and light inferiority by 2.27% ( $2.64 \rightarrow 2.70$ ) times larger APAE on Sintel (train) Final. In these experiments, the convolutional network architecture of CGCV obtains comparable performance to the transformer architecture of FlowFormer-S.

### 4.3 Qualitative Analysis

**Visualization of Flow Fields.** Fig. 1 visualizes the flow fields estimated by CGCV and GMA on three example images from Sintel (test) Final. Image *Mountain 2* contains a large textureless region undergoing fast camera motion. As a consequence, the flow estimation of GMA in this region suffers large error, but it can be substantially improved by regulating its correlation volume using context. On image *Tiger*, GMA mis-estimates the motion in the shadowed ground region with large error, whereas CGCV computes it mostly correct. On image *Market 3*, GMA blurs the motion boundary between the out-of-focus foreground object and the foggy background. In contrast, the flow estimated by CGCV in this region is sharp. These visual comparisons validate that regulating the traditional correlation volume by semantic context increases the flow computation robustness to image flatness, shadows, de-focus blur and atmospheric effects.

**Visualization of Feature Maps and Correlation Volumes.** Fig. 5 and Fig. 6 illustrate how our regularization approach improves the quality of the traditional correlation volume for flow computation sequentially, through two example query points (one on the dragon tail, one on the red hair) in image *Market 4* from Sintel (test) Final and image *Alley 1* from Sintel (train) Final. Here we take the CTSKH-tuned GMA and CGCV models to visualize the correlation maps.

As demonstrated by Fig. 5, in either GMA or our CGCV, the correlation volumes computed from the matching features are very noisy. A large number of pixels, which scatter over the target frame, exhibit high correlation to the query point. However, the context cross-frame attention map successfully figures out that only pixels on the dragon tail in the target image have high context correlation to the query point. Weighting the matching correlation by context attention effectively screens the false candidates for the true correspondence, with the highest correlation values locating on the dragon tail. It can be seen that the context feature correlation map has

significantly superior quality to the matching correlation map. Although superimposing it to the context-weighted correlation map does not contribute much to the final correlation map, it benefits obtaining a sharp attention map, thereby forming the high quality final correlation volume.

Similar correction effects can be observed in Fig. 6. Our gating strategy effectively narrows down the high correlation regions from the whole image to the red hair area of the image, where the true correspondence to the query point locates.

#### 4.4 Ablation Study

An ablation study is carried out using the CT-trained model. We gradually remove the main components that construct our CGCV module, until the correlation volume takes the traditional form as in GMA and RAFT, and evaluate the degenerated model step by step on benchmark datasets employed by GMA for ablation study [20]. Accuracy is measured by AEPE.

We first remove the context cross-frame correlation volume from the whole correlation volume (i.e., degenerate  $\beta$  in Eq. 6 to 0). This ablation increases the flow accuracy on Thing (test) Clean and Thing (test) Final by 0.13 and 0.04 pixels. However, when generalizing to Sintel, the ablation causes performance degradation. Based on this observation, we hypothesis that the context cross-correlation benefits the generalization ability of the model, and hence adopt it as the lifting term of CGCV.

We further replace the Sigmoid function in the context cross-frame attention (Eq. 3) by a softmax function. This drastically lowers the flow accuracy on all datasets. The AEPE on Things (test) Clean and Final deteriorates by 0.48 pixels and 0.34 pixels respectively. On Sintel (train) Clean, the AEPE retreats by 0.12 pixels. At this stage, accuracy measurements of the degenerated model in all tests are remarkably lower than CGCV.

Finally, we switch off the context cross-attention gating mechanism, as defined in Eq. 1. This lowers the flow accuracy on Sintel (train) Clean by 0.1 pixels.

The ablation study verifies that all composing components of the new correlation volume are crucial to improve the quality of traditional correlation volume.

## 5 Conclusion

In this paper, we have presented a novel correlation volume for RAFT flow computation paradigm. We take advantage of the semantic context features to suppress the false positive matching in traditional cost volume via cross-attention, and lift the false negative matching via cross-correlation. The presented correlation volume is simple, light-weight, highly effective and ready to replace the cost volume in state of the art RAFT-style flow computation neural models. By incorporating our cost volume with the GMA-RAFT framework, we have shown that our cost volume substantially enhances the numerical and visual performance of GMA, with superiority to other GMA-based methods, by comprehensive experiments.

## References

- [1] Kitti flow 2015. [http://www.cvlibs.net/datasets/kitti/eval\\_scene\\_flow.php?benchmark=flow](http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=flow). 9
- [2] Mpi-sintel flow. <http://sintel.is.tue.mpg.de/results>. 9
- [3] S. Bai, Z. Geng, Y. Savani, and J. Z. Kolter. Deep equilibrium optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 620–630, 2022. 2, 3, 8
- [4] M. J. Black and P. Anandan. A framework for the robust estimation of optical flow. In *1993 (4th) International Conference on Computer Vision*, pages 231–236, 1993. 2
- [5] J.-Y. Bouguet. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel corporation*, 5(1-10):4, 2001. 2
- [6] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34:9355–9366, 2021. 3
- [7] Y. Deng, J. Xiao, S. Z. Zhou, and J. Feng. Detail preserving coarse-to-fine matching for stereo matching and optical flow. *IEEE Transactions on Image Processing*, 30:5835–5847, 2021. 2, 8
- [8] A. Diba, V. Sharma, L. V. Gool, and R. Stiefelhagen. Dynanonet: Dynamic action and motion network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6192–6201, 2019. 2
- [9] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recogni-

- tion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016. 2
- [10] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 2
- [11] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):504–511, 2012. 2, 3
- [12] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, and H. Li. Flowformer: A transformer architecture for optical flow. *arXiv preprint arXiv:2203.16194*, 2022. 2
- [13] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, and H. Li. Flowformer: A transformer architecture for optical flow. *arXiv preprint arXiv:2203.16194*, 2022. 3, 7, 8, 11
- [14] T.-W. Hui and C. C. Loy. Liteflownet3: Resolving correspondence ambiguity for more accurate optical flow estimation. In *European Conference on Computer Vision*, pages 169–184. Springer, 2020. 2, 3
- [15] T.-W. Hui, X. Tang, and C. C. Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8981–8989, 2018. 3
- [16] T.-W. Hui, X. Tang, and C. C. Loy. A lightweight optical flow cnn—revisiting data fidelity and regularization. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2555–2569, 2020. 3
- [17] J. Hur and S. Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2019. 3
- [18] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 2, 3
- [19] J. Jeong, J. M. Lin, F. Porikli, and N. Kwak. Imposing consistency for optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3181–3191, 2022. 2, 3, 8
- [20] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley. Learning to estimate hidden motions with global motion aggregation. In *ICCV*, pages 9772–9781, 2021. 1, 2, 3, 4, 7, 8, 9, 12
- [21] W. Jiang, E. Trulls, J. Hosang, A. Tagliasacchi, and K. M. Yi. Cotr: Correspondence transformer for matching across images. In *ICCV*, pages 6207–6217, 2021. 2
- [22] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *PAMI*, 33(5):978–994, 2010. 2
- [23] S. Liu, L. Yuan, P. Tan, and J. Sun. Steadyflow: Spatially smooth optical flow for video stabilization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4209–4216, 2014. 2
- [24] Y. Liu, X. Zou, S. Xu, G. Xing, H. Wei, and Y. Zhang. Real-time shadow detection from live outdoor videos for augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(7):2748–2763, 2022. 2
- [25] Y. Lu, J. Valmadre, H. Wang, J. Kannala, M. Harandi, and P. Torr. Devon: Deformable volume network for learning optical flow. In *Winter Conference on Applications of Computer Vision*, pages 2705–2713, 2020. 3
- [26] A. Luo, F. Yang, X. Li, and S. Liu. Learning optical flow with kernel patch attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8906–8915, 2022. 2, 3, 4, 8, 10
- [27] A. Luo, F. Yang, K. Luo, X. Li, H. Fan, and S. Liu. Learning optical flow with adaptive graph reasoning. *Proc. AAAI*, 2022. 3, 4, 7, 8, 9, 10
- [28] I. Melekhov, A. Tiulpin, T. Sattler, M. Pollefeys, E. Rahtu, and J. Kannala. Dgc-net: Dense geometric correspondence network. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1034–1042, 2019. 2, 3
- [29] Y. Niu, A. Dick, and M. Brooks. Compass rose: A rotational robust signature for optical flow computation. *TIP*, 24(1):63–73, 2013. 2
- [30] Y. Niu, Z. Xu, and X. Che. Dynamically removing false features in pyramidal lucas-kanade registration. *TIP*, 23(8):3535–3544, 2014. 2
- [31] Z. Peng, B. Jiang, H. Xu, W. Feng, and J. Zhang. Facial optical flow estimation via neural non-rigid registration. *Computational Visual Media*, 9(3), 2022. 2
- [32] A. Ranjan and M. J. Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 2, 3
- [33] I. Rocco, R. Arandjelovic, and J. Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017. 2, 3
- [34] S. Roth and M. J. Black. On the spatial statistics of optical flow. *International Journal of Computer Vision*, 74(1):33–50, 2007. 2
- [35] Z. Shou, X. Lin, Y. Kalantidis, L. Sevilla-Lara, M. Rohrbach, S.-F. Chang, and Z. Yan. Dmc-net: Generating discriminative motion cues for fast compressed video

- action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1268–1277, 2019. 2
- [36] X. Sui, S. Li, X. Geng, Y. Wu, X. Xu, Y. Liu, R. Goh, and H. Zhu. Craft: Cross-attentional flow transformer for robust optical flow. In *CVPR*, pages 17602–17611, 2022. 2, 3, 4, 7, 8, 9, 10
- [37] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *CVPR*, pages 2432–2439, 2010. 2
- [38] D. Sun, S. Roth, J. P. Lewis, and M. J. Black. Learning optical flow. In *European Conference on Computer Vision*, pages 83–97, 2008. 2
- [39] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2, 3, 4
- [40] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Models matter, so does training: An empirical study of cnns for optical flow estimation. *IEEE transactions on pattern analysis and machine intelligence*, 42(6):1408–1423, 2019. 3
- [41] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1390–1399, 2018. 2
- [42] Z. Teed and J. Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 2, 3, 4, 5, 7, 8
- [43] P. Truong, M. Danelljan, and R. Timofte. Glu-net: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6258–6268, 2020. 2, 3
- [44] L. Wang, P. Koniusz, and D. Q. Huynh. Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8698–8708, 2019. 2
- [45] M. Werlberger, T. Pock, and H. Bischof. Motion estimation with non-local total variation regularization. In *CVPR*, pages 2464–2471, 2010. 2
- [46] C. Xu, W. Qu, X. Xu, and X. Liu. Multi-scale flow-based occluding effect and content separation for cartoon animations. *IEEE Transactions on Visualization and Computer Graphics*, 2022. Early Access. 2
- [47] H. Xu, J. Yang, J. Cai, J. Zhang, and X. Tong. High-resolution optical flow from 1d attention and correlation. In *Proceedings of the International Conference on Computer Vision*, pages 10498–10507, 2021. 3
- [48] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao. Gm-flow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2022. 2, 3, 8
- [49] J. Xu, R. Ranftl, and V. Koltun. Accurate optical flow via direct cost volume processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1289–1297, 2017. 3
- [50] G. Yang and D. Ramanan. Volumetric correspondence networks for optical flow. *Advances in neural information processing systems*, 32:794–805, 2019. 3
- [51] J. Yu and R. Ramamoorthi. Learning video stabilization using optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8159–8167, 2020. 2
- [52] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223, 2007. 2
- [53] M. Zhai, X. Xiang, N. Lv, X. Kong, and A. El Saddik. An object context integrated network for joint learning of depth and optical flow. *IEEE Transactions on Image Processing*, 29:7807–7818, 2020. 2
- [54] C. Zhang, Z. Zhou, Z. Chen, W. Hu, M. Li, and S. Jiang. Self-attention-based multiscale feature learning optical flow with occlusion feature map prediction. *IEEE Transactions on Multimedia*, 2021. 2, 8
- [55] S. Zhao, Y. Sheng, Y. Dong, E. I. Chang, and Y. Xu. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *CVPR*, pages 6278–6287, 2020. 2, 3
- [56] S. Zhao, L. Zhao, Z. Zhang, E. Zhou, and D. Metaxas. Global matching with overlapping attention for optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17592–17601, 2022. 2, 3, 4, 8
- [57] Z. Zheng, N. Nie, Z. Ling, P. Xiong, J. Liu, H. Wang, and J. Li. Dip: Deep inverse patchmatch for high-resolution optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8925–8934, 2022. 2