FlowFormer: A Transformer Architecture for Optical Flow

Zhaoyang Huang*, Xiaoyu Shi*, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, Hongsheng Li
Project Page: https://drinkingcoder.github.io/publication/flowformer/

ECCV TEL AVIV 2022

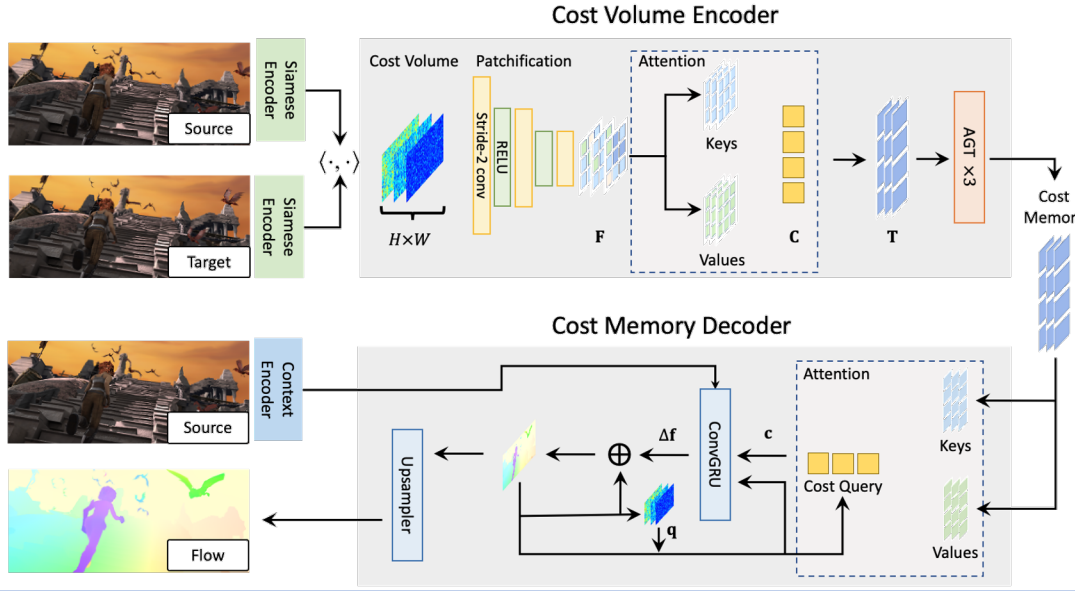
Motivation&Contributions

- Recently, transformers have attracted much attention for their ability of modeling long-range relations, which can benefit optical flow estimation. Can we enjoy both advantages of transformers and the cost volume from the previous milestone architectures?

Our contributions:

- A novel transformer-based neural network architecture for optical flow estimation.
- A novel cost volume encoder, effectively aggregating cost information into compact latent cost tokens.
- A recurrent cost decoder that recurrently decodes cost features with dynamic positional cost queries to iteratively refine the estimated optical flows.
- Validate for the first time that an ImageNet-pretrained transformer can benefit the estimation of optical flow.

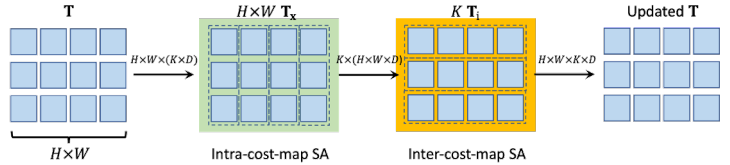
Framework Overview



A FlowFormer adopts an encoder-decoder architecture for cost volume encoding and decoding. After building a 4D cost volume, FlowFormer consists of two main components:

1) a cost volume encoder that embeds the 4D cost volume into a latent cost space and fully encodes the cost information in such a space.

2) a recurrent cost decoder that estimates flows from the encoded latent cost features.



Alternate-Group Transformer (AGT) Layer. AGT alternatively groups tokens in T into H × W groups that contains K tokens and K groups that contains H × W tokens, and encode tokens inside groups via self-attention and ss self-attention respectively.

Qualitative Comparison

(a) Input (b) FlowFormer (Ours) (c) GMA

FlowFormer greatly reduces the flow leakage around object boundaries (pointed by red arrows) and clearer details (pointed by blue arrows).

Quantitative Comparison

Training Data	Method	Sintel (train) Clean	Sintel (train) Final	KITTI-15 (train) F1-epe	KITTI-15 (train) F1-all	Sintel (test) Clean	Sintel (test) Final	KITTI-15 (test) F1-all
A+S+K+H	Perceiver IO [24]	-	-	-	-	1.81	2.42	4.98
	PWC-Net [42]	-	-	-	-	2.17	2.91	5.76
	RAFT [46]	-	-	-	-	1.95	2.57	4.23
C+T	HD3 [55]	3.84	8.77	13.17	24.0	-	-	-
	LiteFlowNet [21]	2.48	4.04	10.39	28.5	-	-	-
	PWC-Net [42]	2.55	3.93	10.35	33.7	-	-	-
	LiteFlowNet2 [22]	2.24	3.78	8.97	25.9	-	-	-
	S-Flow [57]	1.30	2.59	4.60	15.9			
	RAFT [46]	1.43	2.71	5.04	17.4	-	-	-
	FM-RAFT [26]	1.29	2.95	6.80	19.3	-	-	-
	GMA [25]	1.30	2.74	4.69	17.1	-	-	-
	Ours	**0.95**	**2.35**	**4.09**	**14.72**	-	-	-
C+T+S+K+H	LiteFlowNet2 [22]	(1.30)	(1.62)	(1.47)	(4.8)	3.48	4.69	7.74
	PWC-Net+ [43]	(1.71)	(2.34)	(1.50)	(5.3)	3.45	4.60	7.72
	VCN [53]	(1.66)	(2.24)	(1.16)	(4.1)	2.81	4.40	6.30
	MaskFlowNet [59]	-	-	-	-	2.52	4.17	6.10
	S-Flow [57]	(0.69)	(1.10)	(0.69)	(1.60)	1.50	2.67	**4.64**
	RAFT [46]	(0.76)	(1.22)	(0.63)	(1.5)	1.94	3.18	5.10
	FM-RAFT [26]	(0.79)	(1.70)	(0.75)	(2.1)	1.72	3.60	6.17
	GMA [25]	-	-	-	-	1.40	2.88	5.15
	Ours	(0.48)	(0.74)	(0.53)	(1.11)	**1.14**	**2.18**	4.68
	RAFT* [46]	(0.77)	(1.27)	-	-	1.61	2.86	-
	GMA* [25]	(0.62)	(1.06)	(0.57)	(1.2)	1.39	2.47	-

Generalization performance:

- We train FlowFormer on the FlyingChairs and FlyingThings (C+T), and evaluate it on the training set of Sintel and KITTI2015. This settings evaluates methods' generalization performance.
- FlowFormer ranks 1st among all compared methods on both benchmarks. FlowFormer achieves 0.95 and 2.35 on the clean and final pass of Sintel, and 14.72 F1-all on KITTI. Compared to GMA, FlowFormer reduces 26.9% and 14.2% errors on Sintel clean and final, and 13.9% errors on KITTI-2015 F1-all, which shows its extraordinary generalization performance.

Sintel performance:

- FlowFormer achieves 1.16 and 2.09 on the Sintel clean and final, 16.5% and 15.5% lower error compared to GMA∗ , which ranks both 1st on the Sintel benchmark.
- Compared with GMA, which also does not use the warm-start, FlowFormer obtains 17.2% and 27.5% error reduction.