FlowFormer: A Transformer Architecture for Optical Flow – Supplementary Materials

Zhaoyang Huang[1,3*], Xiaoyu Shi[1,3*], Chao Zhang[2], Qiang Wang[2], Ka Chun Cheung[3], Hongwei Qin[4], Jifeng Dai[4], and Hongsheng Li[1†]

[1]Multimedia Laboratory, The Chinese University of Hong Kong
[2]Samsung Telecommunication Research
[3]NVIDIA AI Technology Center [4]SenseTime Research
{drinkingcoder@link, xiaoyushi@link, hsli@ee}.cuhk.edu.hk

1 More Ablation Studies

Intra.	Inter.	Sinte (train)		Kitti (train)		Params.
		Clean	Final	F1-epe	F1-all	
Trans	Trans	1.20	2.85	4.57	15.46	15.2M
MLP	Trans	1.20	2.67	5.01	16.81	15.2M
Trans	Conv	1.23	2.72	4.73	15.87	15.1M
MLP	Conv	1.22	2.71	4.88	17.23	15.1M

Table 1: Ablation study on the alternative-group transformer (AGT) layer. For intra-cost-map aggregation layer (Intra.), we replace transformer (Trans) with MLP-Mixer [4] block (MLP). For inter-cost-map aggregation layer (Inter.), we replace transformer with ConvNeXt [3] block (Conv).

As shown in Table 1, we conduct additional ablation experiments on the alternative-group transformer (AGT) layer. For intra-cost-map aggregation layer, since the number and dimension of latent cost tokens are fixed, we test on replacing our design with MLP-Mixer [4] (2nd row), which is a state-of-the-art MLP-based architecture. We also substitute ConvNeXt [3] for transformer in inter-cost-map aggregation (3rd row). Furthermore, we replace both transformers with MLP and ConvNext (4th row). Replacing transformer layers leads to slightly better performance on Sintel final pass, while brings a clear drop on KITTI. Therefore, we adopt the proposed full transformer architecture as our final model.

2 Tile with Gaussian Weights

Since positional encodings used in transformers are sensitive to image size and the size of an image pair for test ($H_{test} \times W_{test}$) might be different from those

*Zhaoyang Huang and Xiaoyu Shi assert equal contributions.
†Corresponding author: Hongsheng Li

of the training images, $(H_{train} \times W_{train})$, we crop the test image pair according to the training size and estimate flows for patch pairs separately, and then tile the flows to obtain a complete flow map following a similar strategy proposed in Perceiver IO [1]. Specifically, we crop the image pair into four evenly-spaced tiles, i.e., $H_{train} \times W_{train}$ image tiles starting at $(0, 0)$, $(0, W_{test} - W_{train})$, $(H_{test} - H_{train}, 0)$, and $(H_{test} - H_{train}, W_{test} - W_{train})$, respectively. For each pixel that is covered by several tiles, we compute its output flow $\mathbf{f}$ by blending the predicted flows $\mathbf{f}_i$ with weighted averaging:

$$\mathbf{f} = \frac{\sum_i w_i \mathbf{f}_i}{\sum_i w_i}, \tag{1}$$

where w_i is the weight of the i-th tile for the pixel. We compute the $H_{train} \times W_{train}$ weight map according to pixels' normalized distances $d_{u,v}$ to the tile center:

$$\begin{aligned} d_{u,v} &= ||(u/H_{train} - 0.5, v/W_{train} - 0.5)||_2, \\ w_{u,v} &= g(d_{u,v}; \mu = 0, \sigma = 0.05), \end{aligned} \tag{2}$$

where (u, v) denote a pixel's 2D coordinate. We use a Gaussian-like g as the weighting function to obtain smoothly blended results. We use this weight map for all the tiles.

3 Training Image Size Details

We train FlowFormer with image size of 368×498 on FlyingChairs and 432×960 on the following training stages, i.e., FlyingThings, Sintel, and KITTI. As the height of images in KITTI only ranges from 374 to 375, we train another FlowFormer model, dubbed as *FlowFormer#*, and evaluate it on the KITTI-15 training set to obtain better performance. Following GMA [2], *FlowFormer#* is trained with 368×498 image size on FlyingChairs and 400×720 image size on FlyingThings, which achieves 4.09 F1-epe and 14.72 F1-all on the KITTI training set as presented in the Table 1 in the original paper.

References

1. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al.: Perceiver io: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795 (2021)
2. Jiang, S., Campbell, D., Lu, Y., Li, H., Hartley, R.: Learning to estimate hidden motions with global motion aggregation. arXiv preprint arXiv:2104.02409 (2021)
3. Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s (2022)
4. Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Keysers, D., Uszkoreit, J., Lucic, M., Dosovitskiy, A.: Mlp-mixer: An all-mlp architecture for vision (2021)