SMART: Stratified Matching and Recurrent Transformer for Optical Flow Estimation

Kin-Chung Chan and Kin-Man Lam

***Abstract*—Starting from RAFT, current studies on the optical flow topic focus on similar iterative update frameworks, in which frame pairs are encoded to feature maps of coarse size and iterative updated to optimize the prediction. Currently, GMFlow attempts to sequentially refine flows on a coarse level and a fine level, while failing to better utilize the enhanced local features obtained from Transformers. We propose a parallel refinement structure called SMART that reuses the coarse-level rich-information features abandoned after global matching in GMFlow. The parallel structure also allows the coarse-level prediction to be refined throughout the process and updated with information on both levels.**

***Index Terms*—Optical flow, Iterative refinement, Hierarchical refinement**

I. Introduction

Optical Flow Estimation is a task to estimate the movement of each pixel from the reference image to the target image. The current state-of-art method GMFlow [1] is inspired by LoFTR [2], in which 7 layers of self- and cross-attention layers are used to enhance the local features, and feature matching is applied to construct coarse-level prediction and fine-level prediction consecutively. GMFlow uses a similar idea as LoFTR and a lightweight structure to extract features and conducts local feature enhancement. It then constructs and refines flows on the coarse level (1/8 original image size) with several updates, and repeats the process on the fine level (1/4 original image size). The pipelines of GMFlow and LoFTR both start with the coarse level prediction, then they locally adjust the results. These ideas are based on the assumption that the coarse-level predictions are accurate. However, experimental results show that this assumption cannot be held. Once the refinement process moves to the fine level, the coarse-level prediction cannot be adjusted, and the refinement will be conducted on a wrong basis. To tackle this problem, we propose Stratified Matching and Recurrent Transformer (SMART). SMART is a parallel refinement structure that not only addresses the previously mentioned problem but also allows the reuse of the rich-information local features on the coarse level, which were abandoned immediately after global matching in GMFlow.

II. Stratified Matching and Recurrent Transformer

SMART starts with a Feature Pyramid Transformer (FPT), in which we add 6 self- and cross-attention layers on the skip connections of the Feature Pyramid Network. FPT utilizes the enhanced coarse-level features to produce the enhanced fine-level features, which are of great discriminative power. The second part of SMART is a parallel iterative update process. Coarse-level flows and fine-level flows are designed to update alternatively in this part. This structure allows the coarse-level prediction to be adjusted after fine-level updates and aggregate information from both levels.

TABLE I
Quantitative results of GMFlow+ and SMART on FlyingChairs dataset.

Method	Total # refine.	EPE	s_{0-10}	s_{10-40}	s_{40+}
GMFlow+ [3]	2	1.177	0.584	1.316	7.716
GMFlow+ [3]	12	0.799	0.368	0.868	5.727
SMART	5	0.677	0.302	0.746	4.929

III. Preliminary Results

We compared the validation results of GMFlow+ [3] with 12 refinements (6 refinements on each level) and SMART with 5 refinements. The GMFlow+ models are trained in 100K iterations with a batch size of 16. Our model is trained in 200K iterations with a batch size of 8. The quantitative results in Table I shows that SMART with 5 refinements is better than GMFlow+ with 12 refinements by 0.122 in terms of Average End-Point Error (EPE).

IV. Conclusion

In this work, we introduce Stratified Matching and Recurrent Transforme (SMART), which is a parallel refinement pipeline for Optical Flow Estimation. SMART enhances both coarse-level and fine-level local features and fuses them to produce features with great discriminative power. Also, its parallel update strategy allows coarse-level flow predictions to be adjusted during the whole process. Our preliminary result shows that SMART with a small number of iterations performs better than GMFlow+ with more iterations on FlyingChairs dataset.

References

[1] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, "Gmflow: Learning optical flow via global matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8121–8130.

[2] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[3] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger, "Unifying flow, stereo and depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.