Author Rebuttal for NeurIPS 2020 Submission #2238

We thank all the reviewers for their valuable comments and suggestions. To improve readability, we promise to add a section of the background and carefully revise the manuscript before the final submission. We respond to the main concerns as follows. All the source code will be released to the community soon.

Response to Reviewer #4

Q1: *Some suggestions for improving the writing quality.*
A1: Thanks for your valuable suggestions! We will revise the manuscript carefully according to your comments.
Q2: *What about the computational efficiency on GPUs or "tensor processing units"?*
A2: The empirical runtime on GPUs is very low, which has been reported in Fig.4 of the supplementary material. We optimize the CUDA kernel by parallelizing the algorithm along with batches, channels, and nodes of the same depth.
Q3: *Why force a closed-form solution? Would it not be better to just use an approximate solution on grid structures?*
A3: Compared with the iterative optimization process for grid structures, the closed-form solution ensures that the LTF-V2 has a deterministic and negligible computational complexity to obtain the global receptive field. Besides, the low-level applications of tree filter (*e.g.*, stereo matching and image denoising) demonstrate the competitive performance against many energy-optimization based methods. We will add some related experiments in the final version.

Response to Reviewer #7

Q1: *Performance improvement on the COCO dataset is smaller than that on the Cityscapes dataset.*
A1: This is a valuable question. It seems to be a common phenomenon of related methods, *e.g.*, Non-Local [11] and CCNet [12]. We think the reason may come from different metrics for different tasks and the more complex scenarios on the COCO dataset. Nevertheless, compared with competing methods, the LTF-V2 achieves more quality gains. Besides, as shown in Tab. 8, our method achieves 2.9% mIoU absolute gains over baseline without bells-and-whistles.

Response to Reviewer #8

Q1: *Some concerns about the technical contribution of this work.*
A1: The traditional tree filter [23] already has great impacts on many low-level computer vision tasks, owing to its structure-preserving property and high efficiency. This paper further releases its representation potential by relaxing the geometric constraint. Albeit the modification of the proposed module seems small in form, it has significantly different properties from the original module (refer to Fig.3) and enables fully end-to-end training, which could obtain consistent improvements with negligible computational and parametric overheads. These properties have great potentials which allow our method and principle to extend to other complex tasks with large-number nodes, *e.g.*, replacing the attention module of transformer for natural language processing and enhancing sequential representation for video analysis. We will add more details and revise the "broader impact" section in the final version.
Q2: *Experimental comparison should be updated to include more recent methods.*
A2: Thanks for your suggestion. In fact, according to the comparative analysis, we find that our approach is superior to the listed methods when using the same backbone. We will add more recent methods in the final version.
Q3: *Some concerns about the performance gaps with other state-of-the-art models.*
A3: Our experiments are constructed not only to achieve top performance, but also to demonstrate that a higher quality gain can be achieved with less complexity. Therefore, our method does not adopt additional enhancements (*e.g.*, GCNet [15] is applied in each bottleneck and DANet [46] uses the "multi-grid" operation). We will give more competitive results with these enhancements in the final version.

Response to Reviewer #9

Q1: *I am curious if the proposed methods learns position sensitivity of the query pixels.*
A1: Yes. As shown in Eq.2, for each query pixel i and key pixel j, the pairwise affinity $S_{\mathcal{G}_T}(E_{i,j})$ is regenerated by accumulating all the edge weights along the path from i to j. Since different query pixels have different paths to the same key pixel, the corresponding affinities can be inconsistent. Therefore, the pairwise term (refer to Eq.6) has the position-sensitive property. We will clarify it in the final version.

Table 8: The ablation study for panoptic segmentation on COCO 2017 *val* set.

Model	Backbone	Schedule	LTF-V2	PQ	SQ	RQ	mIoU	mACC	AP_{det}	AP_{seg}
Panoptic FPN	ResNet-50	1x	✗	39.6	77.8	48.6	41.6	52.3	37.6	34.7
			✓	**42.0**	**79.0**	**51.1**	**44.5**	**56.5**	**39.5**	**36.1**

Q2: *I would suggest the authors to perform panoptic segmentation.*
A2: Thanks for your suggestion. We construct an experiment based on Panoptic FPN [CVPR 2019] on the COCO dataset. The LTF-V2 is adopted after "Stage3", "Stage4" and "Stage5" of the backbone. As shown in Tab. 8, the result further demonstrates the effectiveness and generalization of our method. We will add more results in the final version.