D^{2}Former: Jointly Learning Hierarchical Detectors and Contextual Descriptors via Agent-based Transformers

Jianfeng He[1,*], Yuan Gao[1,*], Tianzhu Zhang[1,2,†], Zhe Zhang[2], Feng Wu[1]
[1] University of Science and Technology of China
[2] Deep Space Exploration Laboratory
{hejf, wazs98}@mail.ustc.edu.cn, {tzzhang, fengwu}@ustc.edu.cn, cnclepzz@126.com

In the supplementary material, we first discuss differences between our D^{2}Former and SuperFeatures [7]. Then, we present some visualization results to show the behavior of hierarchical keypoints. And we show qualitative comparisons with the standard full attention [1] and linear attention [3] to demonstrate the effectiveness of our agent-based attention. Finally, we compare our proposed method with state-of-the-art image matching methods [4–6] on the HPatches [2].

1. Differences between our D^{2}Former and SuperFeatures

Here, we introduce differences between our D^{2}Former and SuperFeatures [7]. We conclude three major differences: **(a)** Though our method and SuperFeatures (SF) both use agents (called 'templates' in SF), SF directly takes updated templates as final super features, which is suitable for image retrieval, but is not suitable for pixel-level matching. However, our method can learn pixel-level contextual features by fusing updated agents, more applicable for the pixel-level image matching task. **(b)** Our method adds an extra fusion operation (Eq.(3) in the main text) to learn contextual features compared to SuperFeatures. Here, the fusion is simply realized by matrix multiplication plus residual connection. One can use other complex ways to fuse agents into features (*e.g.* cross-attention). **(c)** The main contribution of our work is to jointly learn hierarchical detectors and contextual descriptors, which is rather different from SuperFeatures that designs for learning features.

2. Visualization Results

Qualitative results about the behavior of hierarchical keypoints. To further present the properties of our hierarchical keypoint detectors, we design to compare the supported regions of descriptors corresponding to different levels of keypoints. As shown in Figure 1, we can find that the higher level keypoints correspond to descriptors with larger support regions.

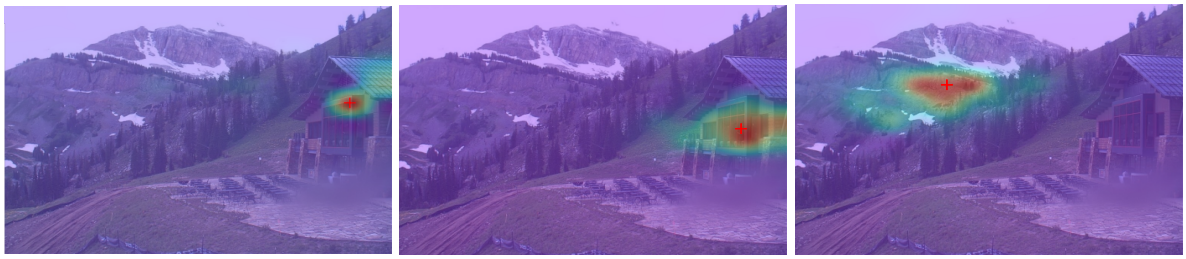

Figure 1. Qualitative results for the attention region of descriptors belonging to keypoints from low-level detectors to high-level detectors. From left to right, keypoints are from the first, second, and the last level.

Qualitative comparisons between our agent-based attention and other attention mechanisms. For the goal to prove the effectiveness of our proposed agent-based attention, we make a comparison with the standard full attention [1] and linear attention [3]. It is clear that our proposed agent-based attention mechanism has a clear attention score map as shown in Figure 2, which can effectively reduce noise and generate robust contextual descriptors.

[*] Equal Contribution
[†] Corresponding Author

Qualitative comparisons with previous state-of-the-art methods. To demonstrate that our proposed D^{2}Former can effectively extract discriminative feature description and realize robust keypoint detection and obtain more reliable image matching results facing diverse challenging factors such as extreme illumination changes and viewpoint variations, we show qualitative comparisons with state-of-the-art detector-based methods R2D2 [4], SuperGlue [5] and LoFTR [6] on the HPatches [2]. We present the qualitative comparisons under illumination changes in Figure 3. And the qualitative comparisons under viewpoint variations are shown in Figure 4. Green and red crosses denote correct and incorrect matches respectively. It's apparent that our proposed method detects more robust keypoints and achieves higher matching accuracy under illumination and viewpoint variations.

References

[1] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*, 2015. 1, 3

[2] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5173–5182, 2017. 1, 2

[3] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the International Conference on Machine Learning*, pages 5156–5165, 2020. 1, 3

[4] Jerome Revaud, Philippe Weinzaepfel, César Roberto de Souza, and Martin Humenberger. R2D2: Repeatable and reliable detector and descriptor. In *Advances in Neural Information Processing Systems*, 2019. 1, 2, 4, 5

[5] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4938–4947, 2020. 1, 2, 4, 5

[6] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8922–8931, 2021. 1, 2, 4, 5

[7] Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. Learning super-features for image retrieval. In *International Conference on Learning Representations*, 2022. 1

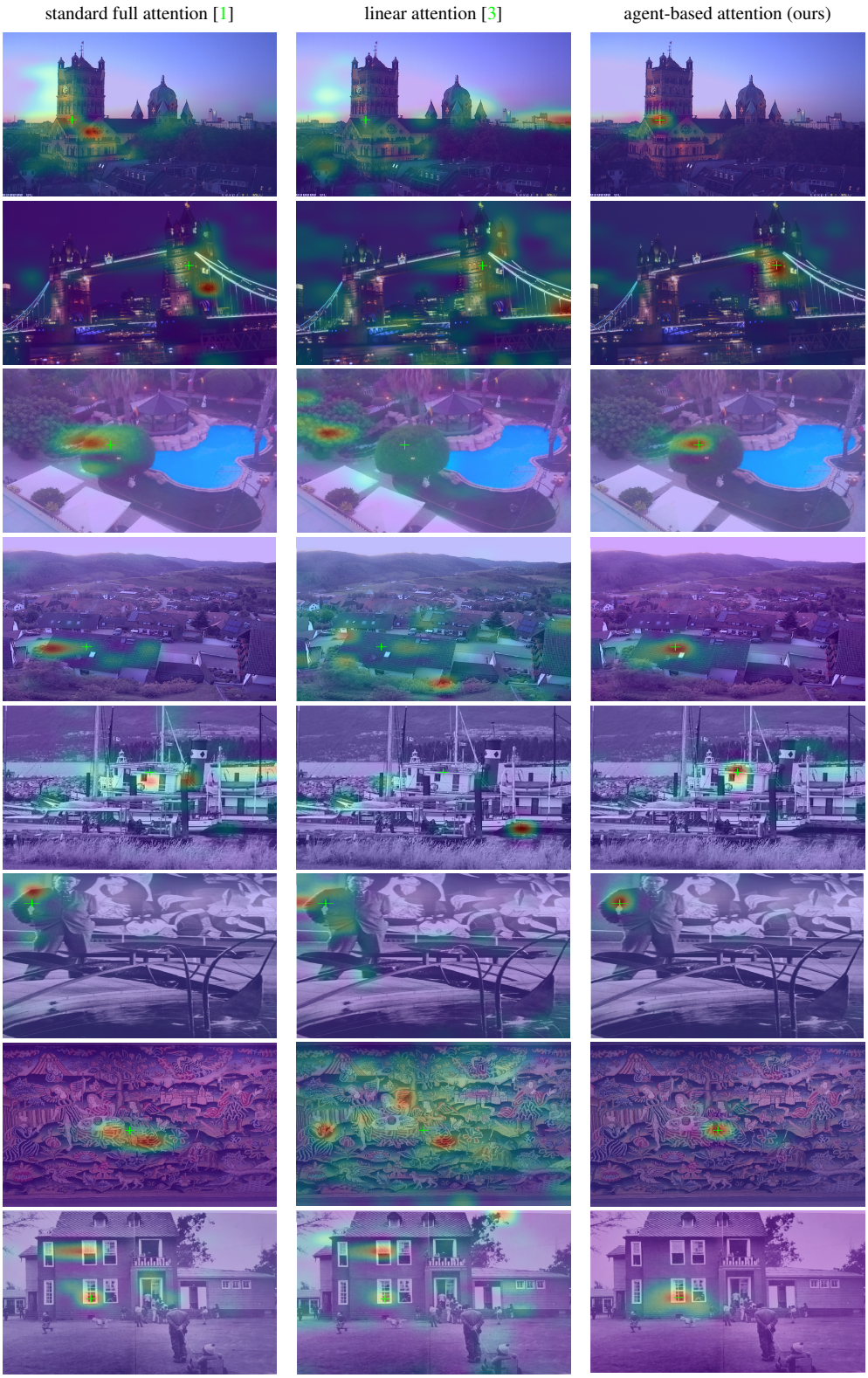

Figure 2. Qualitative comparisons with the standard full attention [1] and linear attention [3]. We can find that standard full attention and linear attention introduce extra noise when conducting global interactions. By contrast, our proposed agent-based attention mechanism has a clear attention score map , which can effectively reduce noise and generate robust contextual descriptors.

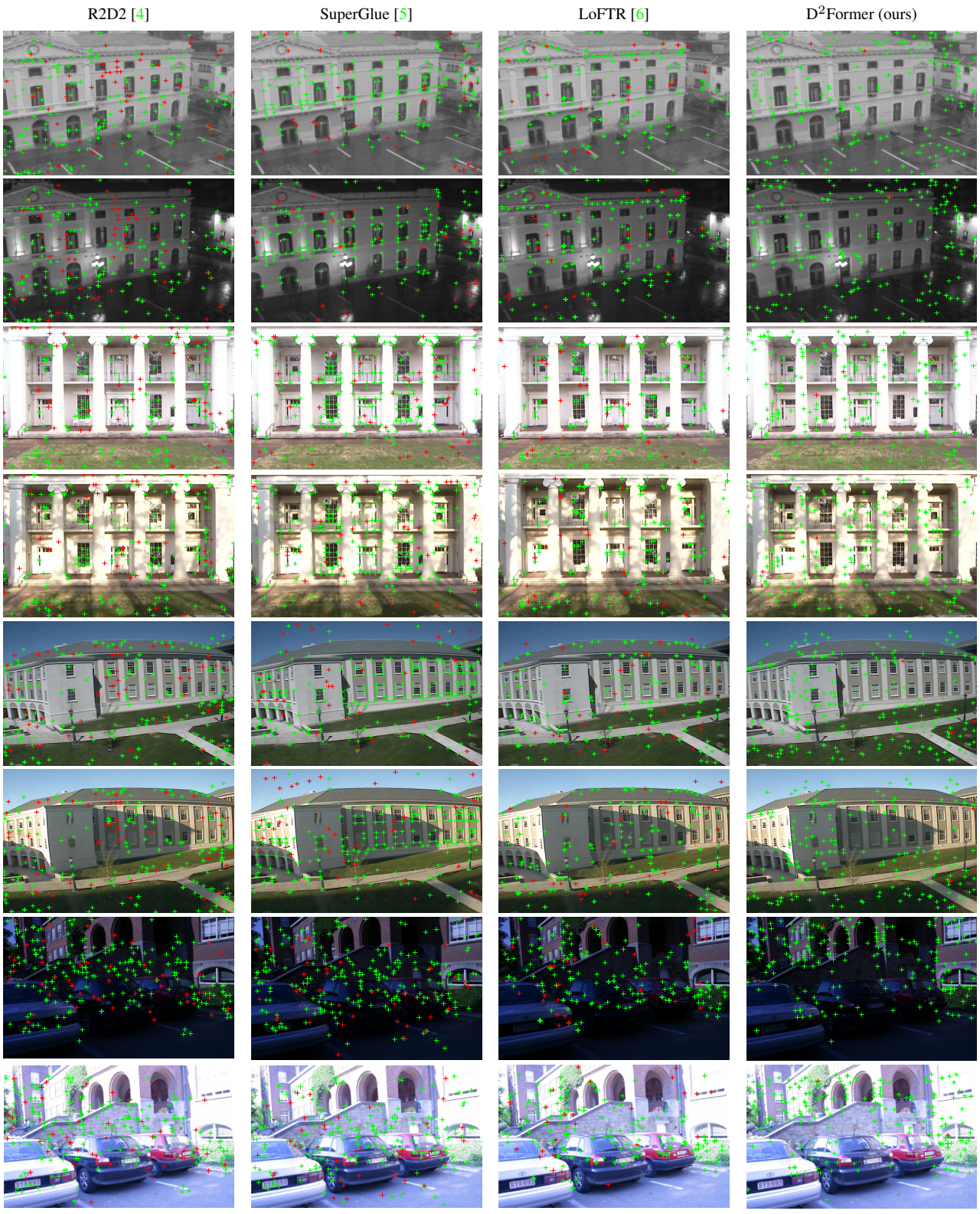

Figure 3. Qualitative comparisons under illumination variations on the HPatches. Green and red crosses denote correct and incorrect matches respectively (threshold is 3px). Our D^{2}Former can extract discriminative feature description and realize robust keypoint detection under extreme illumination changes, leading to more accurate matching results.

Figure 4. Qualitative comparisons under illumination variations on the HPatches. Green and red crosses denote correct and incorrect matches respectively (threshold is 3px). Our D^{2}Former can extract discriminative feature description and realize robust keypoint detection under extreme viewpoint changes, leading to more accurate matching results.