COST AGGREGATION FOR FEW SHOT SEGMENTATION

DR.B.RATNAKANTH[1]

ASSOCIATE PROFESSOR, DEPARTMENT OF CSE, BHOJ REDDY ENGINEERING COLLEGE FOR WOMEN, VINAY NAGAR, HYDERABAD-59

T CHANDANA[2]

UG SCOLOR, DEPARTMENT OF CSE, BHOJ REDDY ENGINEERING COLLEGE FOR WOMEN, VINAY NAGAR, HYDERABAD-59

P HAMSIKA[2]

UG SCOLOR, DEPARTMENT OF CSE, BHOJ REDDY ENGINEERING COLLEGE FOR WOMEN, VINAY NAGAR, HYDERABAD-59

A NAYANI[2]

UG SCOLOR, DEPARTMENT OF CSE, BHOJ REDDY ENGINEERING COLLEGE FOR WOMEN, VINAY NAGAR, HYDERABAD-59

ABSTRACT

We propose a novel cost aggregation network, called Cost Aggregation Transformers (CATs), to find dense correspondences between semantically similar images with additional challenges posed by large intra-class appearance and geometric variations. Cost aggregation is a highly important process in matching tasks, which the matching accuracy depends on the quality of its output. Compared to handcrafted or CNN-based methods addressing the cost aggregation, in that either lacks robustness to severe deformations or inherit the limitation of CNNs that fail to discriminate incorrect matches due to limited receptive fields, CATs explore global consensus among initial correlation map with the help of some architectural designs that allow us to fully leverage self-attention mechanism. Specifically, we include appearance affinity modeling to aid the cost aggregation process in order to disambiguate the noisy initial correlation maps and propose multi-level aggregation to efficiently capture different semantics from hierarchical feature representations. We then combine with swapping self-attention technique and residual connections not only to enforce consistent matching, but also to ease the learning process, which we find that these result in an apparent performance boost. We conduct experiments to demonstrate the effectiveness of the proposed model over the latest methods and provide extensive ablation studies

INTRODUCTION

Establishing dense correspondences across semantically similar images can facilitate many Computer Vision applications, including semantic segmentation [46, 54, 36], object detection [29], and image editing [53, 30, 28, 25]. Unlike classical dense correspondence problems that consider visually similar images taken under the geometrically constrained settings [16, 19, 50, 18], semantic correspondence poses additional challenges from large intra-class appearance and geometric variations caused by the unconstrained settings of given image pair. Recent approaches [42, 43, 45, 34, 37, 39, 31, 58, 47, 57, 51, 35] addressed these challenges by carefully designing deep convolutional neural networks (CNNs)-based models analogously to the classical matching pipeline [48, 41], feature extraction, cost aggregation, and flow estimation. Several works [24, 9, 37, 39, 47, 51] focused on the feature extraction stage, as it has been proven that the more powerful feature representation the from the challenges due to ambiguities generated by repetitive patterns or background clutters [42, 24, 26]. On the other hand, some methods [42, 49, 43, 23, 26, 58] focused on flow estimation stage either by designing additional CNN as an ad-hoc regressor that predicts the parameters of a single global transformation [42, 43], finding confident matches from correlation maps [20, 26], or directly feeding the correlation maps into the

decoder to infer dense correspondences [58]. However, these methods highly rely on the quality of the initial correlation maps. The latest methods [45, 37, 44, 21, 31, 27, 35] have focused on the second stage, highlighting the importance of cost aggregation. Since the quality of correlation maps is of prime importance, they proposed to refine the matching scores by formulating the task as optimal transport problem [47, 31], re-weighting matching scores by Hough space voting for geometric consistency [37, 39], or utilizing high-dimensional 4D or 6D convolutions to find locally consistent matches [45, 44, 27, 35]. Although formulated variously, these methods either use hand-crafted techniques that are neither learnable nor robust to severe deformations, or inherit the limitation of CNNs, e.g., limited receptive fields, failing to discriminate incorrect matches that are locally consistent. In this work, we focus on the cost aggregation stage, and propose a novel cost aggregation network to tackle aforementioned issues. Our network, called Cost Aggregation with Transformers (CATs), is based on Transformer [61, 10], which is renowned for its global receptive field. By considering all the matching scores computed between features of input images globally, our aggregation networks explore global consensus and thus refine the ambiguous or noisy matching scores effectively. Specifically, based on the observation that desired correspondence should be aligned at discontinuities with appearance of images, we concatenate an appearance embedding with the correlation map, which helps to disambiguate the correlation map within the Transformer. To benefit from hierarchical feature representations, following [26, 39, 58], we use a stack of correlation maps constructed from multilevel features, and propose to effectively aggregate the scores across the multi-level correlation maps. Furthermore, we consider bidirectional nature of correlation map, and leverage the correlation map from both directions, obtaining reciprocal scores by swapping the pair of dimensions of correlation map in order to allow global consensus in both perspective. In addition to all these combined, we provide residual connections around aggregation networks in order to ease the learning process. We demonstrate our method on several benchmarks [38, 11, 12]. Experimental results on various benchmarks prove the effectiveness of the proposed model over the latest methods for semantic correspondence. We also provide an extensive ablation study to validate and analyze components in CATs. model learns, the more robust matching is obtained [24, 9, 51]. However, solely relying on the matching similarity between features without any prior often suffers

RELATED WORK

Semantic Correspondence. Methods for semantic correspondence generally follow the classical matching pipeline [48, 41], including feature extraction, cost aggregation, and flow estimation. Most early efforts [7, 30, 11] leveraged the hand-crafted features which are inherently limited in capturing high-level semantics. Though using deep CNN-based features [5, 24, 42, 43, 23, 49, 26] has become increasingly popular thanks to their invariance to deformations, without a means to refine the matching scores independently computed between the features, the performance would be rather limited. To alleviate this, several methods focused on flow estimation stage. Rocco et al. [42, 43] proposed an end-to-end network to predict global transformation parameters from the matching scores, and their success inspired many variants [49, 23, 25]. RTNs [23] obtain semantic correspondences through an iterative process of estimating spatial transformations. DGC-Net [34], Semantic-GLU-Net [58] and DMP [15] utilize a CNN-based decoder to directly find correspondence fields. PDC-Net [59] proposed a flexible probabilistic model that jointly learns the flow estimation and its uncertainty. Arguably, directly regressing correspondences from the initial matching scores highly relies on the quality of them. Recent numerous methods [45, 37, 39, 31, 47, 51, 35] thus have focused on cost aggregation stage to refine the initial matching scores. Among hand-crafted methods, SCOT [31] formulates semantic correspondence as an optimal transport problem and attempts to solve two issues, namely many to one matching and background matching. HPF [37] first computes appearance matching confidence using hyperpixel features and then uses Regularized Hough Matching (RHM) algorithm for cost aggregation to enforce geometric consistency. DHPF [39] that replaces feature selection algorithm of HPF [37] with trainable networks also uses RHM. However, these hand-crafted techniques for refining the matching scores are neither learnable nor robust to severe deformations. As learningbased approaches, NC-Net [45] utilizes 4D convolution to achieve local neighborhood consensus by finding locally consistent matches, and its variants [44, 27] proposed more efficient methods. GOCor [57] proposed aggregation module that directly improves the correlation maps. GSF [21] formulated pruning module to suppress false positives of correspondences in order to refine the initial correlation maps. CHM [35] goes one step further, proposing a learnable geometric

matching algorithm which utilizes 6D convolution. However, they are all limited in the sense that they inherit limitation of CNN-based architectures, which is local receptive fields.

ransformers in Vision. Transformer [61], the de facto standard for Natural Language Processing (NLP) tasks, has recently imposed significant impact on various tasks in Computer Vision fields such as image classification [10, 55], object detection [3, 62], tracking and matching [52, 51]. ViT [10], the first work to propose an end-to-end Transformer-based architecture for the image classification task, successfully extended the receptive field, owing to its self-attention nature that can capture global relationship between features. For visual correspondence, LoFTR [51] uses cross and self-attention module to refine the feature maps conditioned on both input images, and formulate the hand-crafted aggregation layer with dual-softmax [45, 60] and optimal transport [47] to infer correspondences. COTR [22] takes coordinates as an input and addresses dense correspondence task without the use of correlation map. Unlike these, for the first time, we propose a Transformer-based cost aggregation

CONCLUSION

In this paper, we have proposed, for the first time, Transformer-based cost aggregation networks for semantic correspondence which enables aggregating the matching scores computed between input features, dubbed CATs. We have made several architectural designs in the network architecture, including appearance affinity modelling, multi-level aggregation, swapping self-attention, and residual correlation. We have shown that our method surpasses the current state-of-the-art in several benchmarks. Moreover, we have conducted extensive ablation studies to validate our choices and explore its capacity. A natural next step, which we leave for future work, is to examine how CATs could extend its domain to tasks including 3-D reconstruction, semantic segmentation and stitching, and to explore self-supervised learning.

REFERENCES

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.
[2] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. Information, 2020.
[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In ECCV. Springer, 2020.
[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294, 2021.
[5] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. NeurIPS, 29:2414–2422, 2016.
[6] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020.
[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In CVPR Workshops), 2005.
[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009.
[9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In CVPR, 2018.
[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
[11] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. In CVPR, 2016.
[12] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. IEEE transactions on pattern analysis and machine intelligence, 2017.

[13] Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In ICCV, 2017.
[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.
[15] Sunghwan Hong and Seungryong Kim. Deep matching prior: Test-time optimization for dense correspondence. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

