

SUPPLEMENTARY MATERIALS FOR GLOBAL-LOCAL BAYESIAN TRANSFORMER FOR SEMANTIC CORRESPONDENCE

Anonymous authors

Paper under double-blind review

This document provides the details of our proposed method in Section A and gives additional results of visualization in Section B.

A METHOD DETAILS

This section is to further clarify the whole architecture in detail, including the multi-level feature extraction, multi-level integration of correlation maps at the cost aggregation stage, and the hierarchical decoders at the last stage of flow estimation. Besides, we also clarify the details of procedures to estimate the Bayesian posterior.

A.1 ARCHITECTURE DETAILS

Given a pair of images, our proposed GLBT conducts feature extraction, cost aggregation, and flow estimation step by step.

Feature extraction. Initially, a ResNet101 He et al. (2016) pre-trained on ImageNet Deng et al. (2009) is used to extract multi-level features $\{(D_1^s, D_1^t), (D_2^s, D_2^t), (D_3^s, D_3^t)\}$ from the image pair. For each pair of features, we calculate the initial correlation maps via a cosine similarity, obtaining the 4D tensors. At this stage, we obtain the multi-levels of the initial correlation maps $\{C_1, C_2, C_3\}$.

Cost aggregation. At the cost aggregation stage, we leverage 4D convolutions to preprocess these initial correlation maps and fold the result into the $\{X_1, X_2, X_3\}$ separately. To aggregate the multi-level preprocessed correlation maps progressively, the deepest X_3 is processed by our proposed GLBT module. The resulting correlation features are summed with the input X_3 , and the obtained results are upsampled to the same resolution of the correlation maps X_2 . At the next level, the upsampled features from the previous GLBT module are fused with X_2 via an addition operation. Then we employ another GLBT module to cope with the incorporated features. The same steps are repeated to fuse the resulting features and the correlation maps X_1 . Consequently, we obtain the final refined correlation maps C .

Flow estimation. At the flow estimation stage, we aim to integrate the features of the source image $\{D_1^s, D_2^s, D_3^s\}$ and the refined correlation maps C to obtain the final semantic flow map. First, we convert the C to the same dimension as the feature maps by averaging the last two dimensions. Then, we upsample the D_3^s and concatenate it with the resulting C' . The concatenated features are then processed by a GLBT module and a 1×1 2D convolution. Following the same steps, we decode the features of the source image and the refined correlation map in a progressive manner. The final decoded features are forwarded to a prediction layer which contains one 3×3 2D convolution followed by a ReLU activation function and another 3×3 2D convolution, obtaining the final semantic flow map.

A.2 IMPLEMENTATION DETAILS

To estimate an approximate variational posterior using the variational inference procedure Shridhar et al. (2019), we set the prior distribution as two mixture zero mean normal distribution with $\sigma_1 = 0.1$ and $\sigma_2 = 0.4$ respectively. In addition, for the variational weight parameters, we initialize μ from one Gaussian distribution with zero mean and 0.1 standard deviation, and set σ from another Gaussian distribution with -7.0 mean and 0.1 standard deviation. Please refer to the training and testing code in the supplementary material for more implementation details.

B VISUALIZATION RESULTS

In this section, we provide more visual comparison of our proposed GLBT and other state-of-the-art methods, including VAT Hong et al. (2022), MMNet Zhao et al. (2021) and CATs Cho et al. (2021). Figure 1 presents the additional visual comparison of them on the SPair-71k Min et al. (2019) dataset.

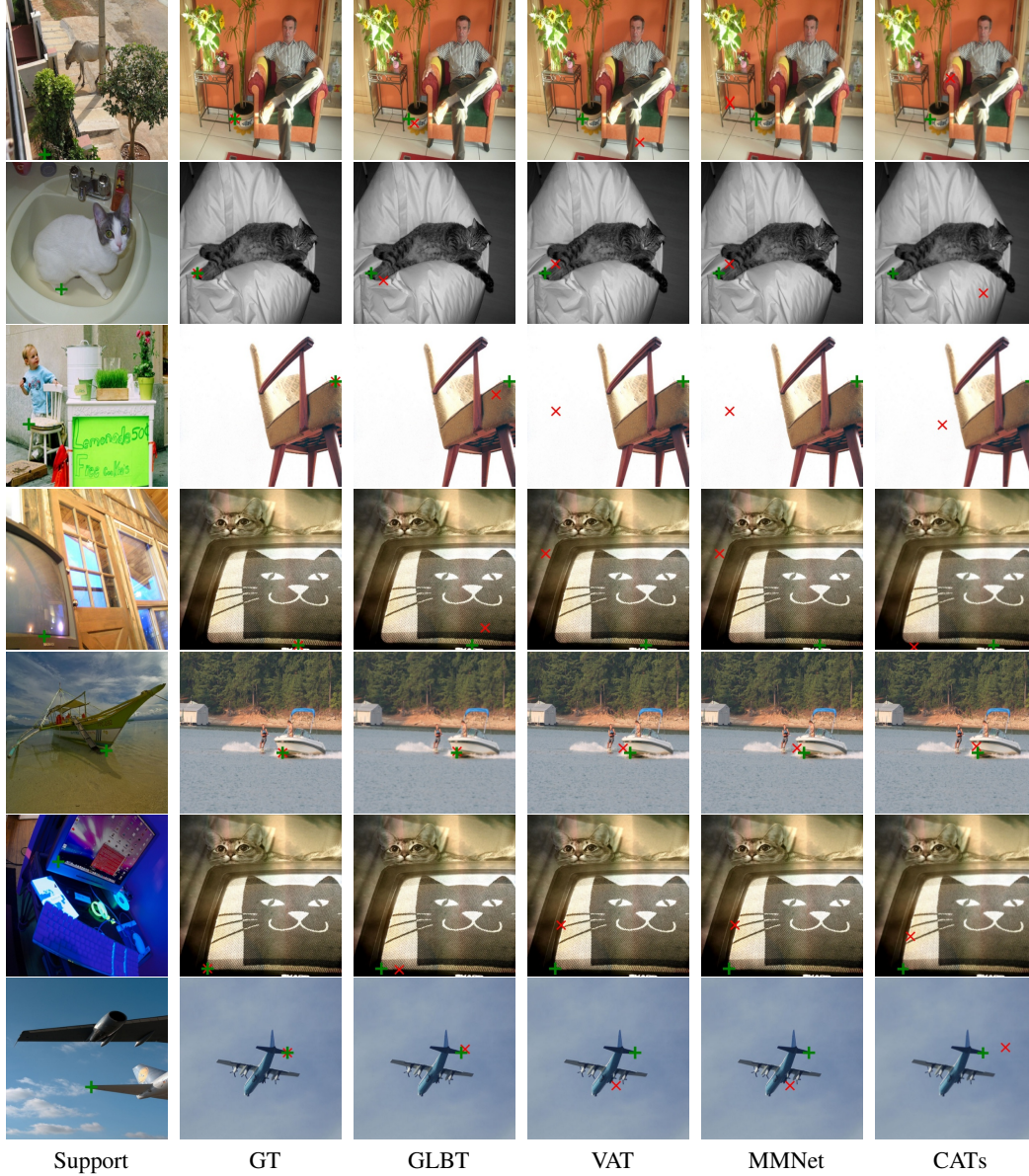


Figure 1: **Qualitative comparison of the recent state-of-the-art methods evaluated on SPair-71k (Min et al., 2019)**, including VAT (Hong et al., 2022), MMNet (Zhao et al., 2021) and CATs (Cho et al., 2021). All results are generated from the same model which is evaluated using PCK @ $\alpha_{img} = 0.1$. For each image pair, “+” is the groundtruth point and “x” is the predicted key point. The closer distance between two signs corresponds to the better results.

C LIMITATIONS AND SOCIETAL IMPACT

Limitations. Our method achieves the impressive results in unconstrained semantic correspondences. However, it has a limitation of matching the accurate correspondences on image pairs of

multi-objects. This is because that we train the model on the image pairs of only one instance. Even so, we believe that our GLBT is also a valuable method for matching the dense semantic correspondences.

Societal Impact. We develop a general model for semantic correspondence and the proposed model is not used for a specific application. Therefore, this work does not directly involve societal issues.

REFERENCES

- Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. In *NIPS*, 2021.
- 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.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. *ECCV*, 2022.
- Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv*, 2019.
- Kumar Shridhar, Felix Laumann, and Marcus Liwicki. A comprehensive guide to bayesian convolutional neural network with variational inference. *arXiv*, 2019.
- Dongyang Zhao, Ziyang Song, Zhenghao Ji, Gangming Zhao, Weifeng Ge, and Yizhou Yu. Multi-scale matching networks for semantic correspondence. In *ICCV*, 2021.