Semantic Correspondence as an Optimal Transport Problem Supplementary Materials

Yanbin Liu[1], Linchao Zhu[1], Makoto Yamada[2,3], Yi Yang[1]
[1]ReLER, University of Technology Sydney, [2]RIKEN AIP, [3]Kyoto University
csyanbin@gmail.com, {linchao.zhu,yi.yang}@uts.edu.au, makoto.yamada@riken.jp

1. Inference runtime analysis

Components	HPF	SCOT	Time
(1) Feature extraction	✓	✓	t_{CNN}
(2) Correlation map (Eq. 4)	✓	✓	$O(D * h_s w_s * h_t w_t)$
(3) CAM (Eq. 7)		✓	$O(d_L * h_s w_s + d_L * h_t w_t)$
(4) Algorithm 1		✓	$O(t_{\text{max}} * h_s w_s * h_t w_t)$

Table 1: Complexity analysis for HPF [2] and the proposed SCOT. D denotes the feature dimension, d_L denotes the number of channels in the last convolutional layer, h_s, w_s, h_t, w_t denote the height and width of source and target images, and t_{max} is the maximum number of iterations.

In this section, we show the complexity comparison between HPF [2] and the proposed SCOT. The four components of our method and their runtime costs are shown in Table 1. Since $d_L \ll D \times \max(h_s w_s, h_t w_t)$ and $t_{\text{max}} \ll D$, the extra costs introduced in components (3) and (4) are marginal. We conclude that the runtime cost of our method is only slightly higher than HPF.

We compared the runtime speed of the ResNet-101 backbone [1] on the test split of SPair-71k. The average inference time is 66ms (HPF) vs. 86ms (Ours) using an NVIDIA 2080Ti GPU.

2. Keypoints matching results

In this section, we show results of transferring keypoints in the source image to the corresponding points in the target image. As shown in Figure 2, the predicted keypoints are quite near to the ground truth. The proposed algorithm is robust under large intra-class variations, background clutter, view-point changes, and partial occlusion.

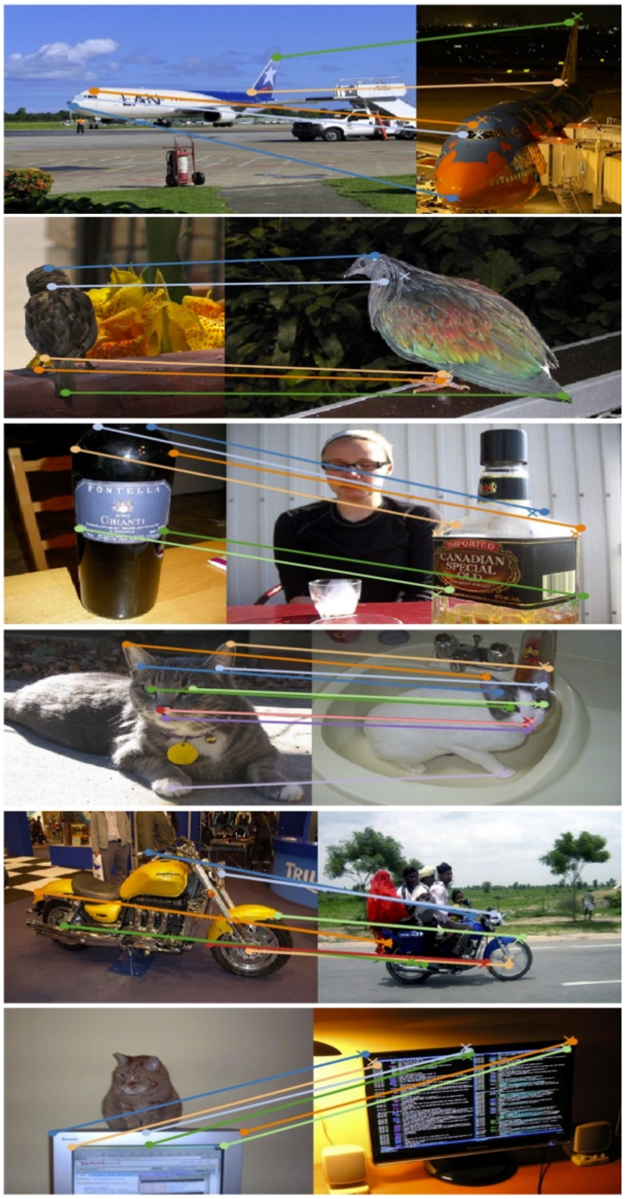
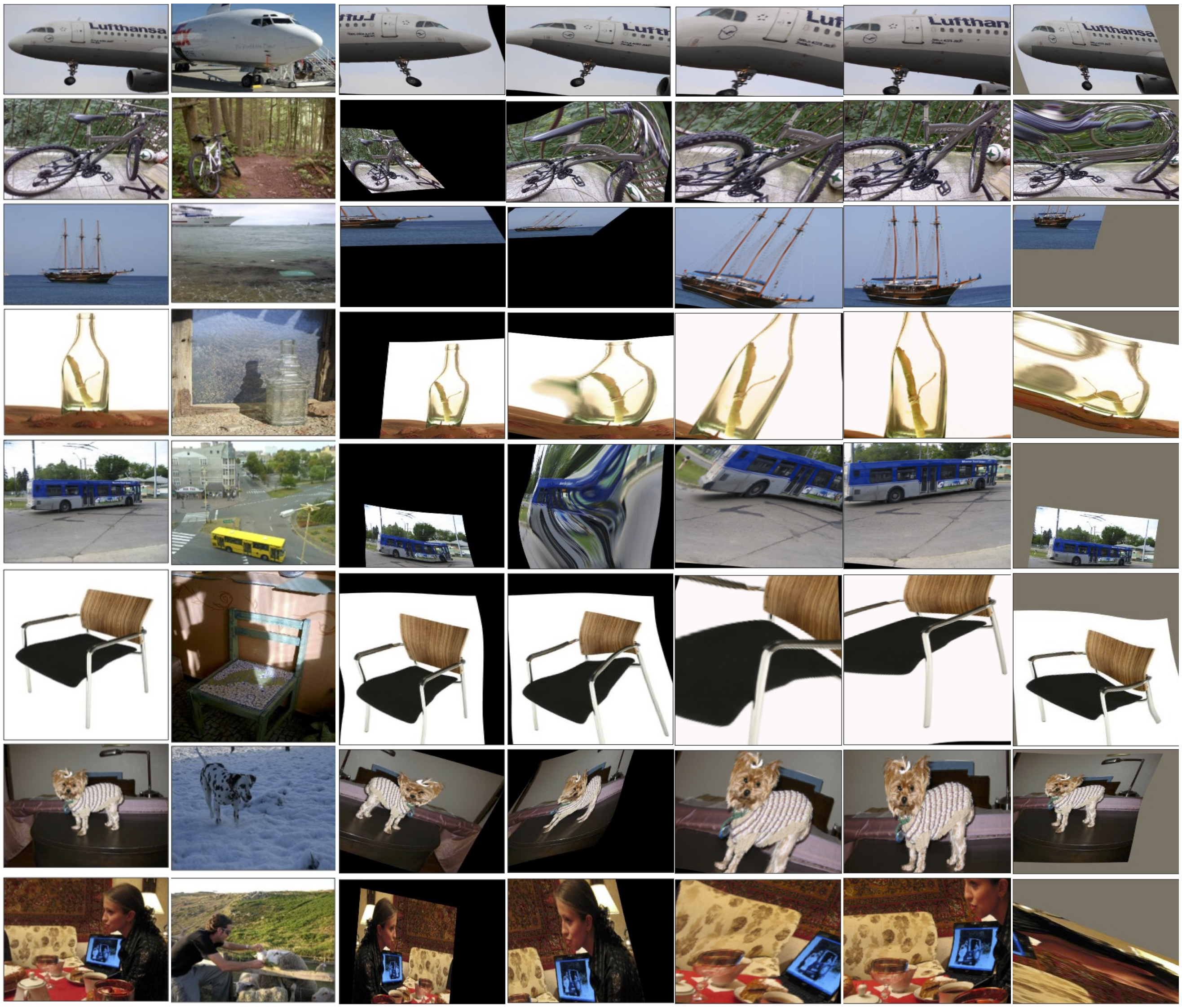


Figure 2: Keypoints matching results on SPair-71k. "×" denotes the ground truth target keypoints.

[1] Input image pairs are resized to $\max(H, W) = 300$. The selected optimal layers are "2,22,24,25,27,28,29".



(a) Source Image (b) Target Image (c) Ours (d) HPF [2] (e) A2Net [1] (f) WeakAlign [3] (g) NC-Net [4]

Figure 1: Qualitative results on SPari-71k. The source images are warped to align with target images using correspondences. For HPF [2], NC-net [4], and our method, we first use the source keypoints and the predicted target keypoitns to estimate the thin-plate spline (TPS) parameters, then apply TPS transformation on the source image. For A2Net [2] and WeakAlign [3], they are global alignment methods that directly predict the global transformation parameters from the CNN models. We show image pairs with large intra-class, scale, and view-point changes. Our method performs better in complex conditions due to our global matching and background suppressing strategies.

3. More qualitative results

In this section, we show more qualitative results in Figure 1. We warp the source images to align with the corresponding target images. For HPF [2], NC-net [4], and our method, we first use the source keypoints and the predicted target keypoints to estimate the thin-plate spline (TPS) parameters, then apply TPS transformation on the source image. For A2Net [2] and WeakAlign [3], they are global alignment methods that directly predict the global transformation parameters from the CNN models. We show the results of image pairs with large intra-class, scale, and view-point changes. Our method performs better in complex conditions due to our global matching and background suppressing strategies.

References

[1] Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho. Attentive semantic alignment with offset-aware correlation kernels. In *ECCV*, 2018. 2

[2] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *ICCV*, 2019. 1, 2

[3] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *CVPR*, 2018. 2

[4] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *NeurIPS*, 2018. 2