Supplementary Material:
PatchMatch-based Neighborhood Consensus for Semantic Correspondence

Jae Yong Lee[1] Joseph DeGol[2] Victor Fragoso[2] Sudipta N. Sinha[2]

[1] University of Illinois [2] Microsoft

1 Runtime Analysis

For PMNC, time complexity is $\mathcal{O}(N^2Mr^4)$ and space complexity is $\mathcal{O}(N^2r^4)$ where $N = max(W, H)$ denotes the largest spatial dimension, r denotes the patch width, and M denotes the number of PatchMatch iterations. The space and time complexity of 4D convolutions over the correlation map is $\mathcal{O}(N^4)$. Thus, given that $r^2 \ll N$, our method runs more quickly than methods using 4D convolutions; and as the image size increases, the difference increases. Table 1 shows the time and memory complexity of our method, NC-Net, and ANC-Net. Given the same scale ($S = \frac{1}{16}$), our method runs more quickly and uses less memory than the baselines. Even with the larger feature map ($S = \frac{1}{8}$), our method runs comparably to the baselines at $S = \frac{1}{16}$.

Method	r	S	400×400		800×800	
			Time(s)	Mem.(MB)	Time(s)	Mem.(GB)
NC-Net	-	$\frac{1}{16}$	0.29	406	1.1	4.6
ANC-Net	-	$\frac{1}{16}$	0.85	1,310	3.7	9.7
PMNC	5	$\frac{1}{16}$	0.08	273	0.30	0.8
PMNC	7	$\frac{1}{16}$	0.28	733	1.02	2.6
PMNC	5	$\frac{1}{8}$	0.23	770	1.30	3.0
PMNC	7	$\frac{1}{8}$	0.95	2,610	2.87	10.4

Table 1: We measure the run time and memory usage of PMNC, NC-Net, and ANC-Net using two image sizes: 400×400 and 800×800. *Time(s)* denotes the time of processing a pair of images in a second, and *Mem.* denotes the peak memory usage for processing a pair of images. The first two rows show the baseline methods (NC-Net [4] and ANC-Net [2]). The remaining rows show our method with varying parameters. r denotes size of the patch. S denotes the scale of the feature map according to the original image size. For the same scale ($S = \frac{1}{16}$), our approach runs more quickly and uses less memory.

2 Additional Qualitative Results

Figure 1 and 2 show additional qualitative results for the PF-PASCAL [1] and SPAIR-71K [3] datasets respectively. Our method robustly estimates semantic keypoint correspondences across different classes for both datasets.

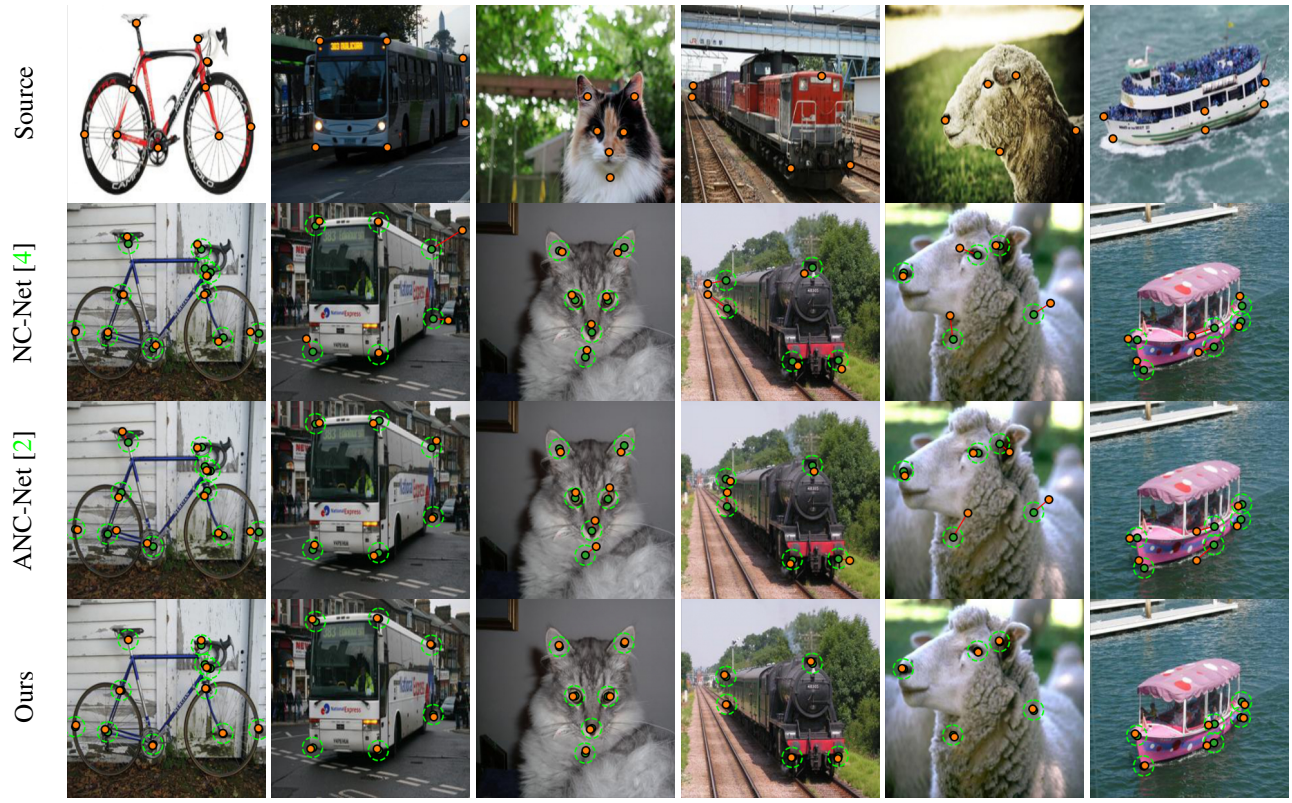

Figure 1: **Additional PF-PASCAL results.** The first row corresponds to the source image with ground truth annotations. The second and third rows correspond to the estimated correspondences using the baseline methods [4, 2]. The last row corresponds to the estimated correspondences using PMNC$_{best}$. The orange dots indicate the source keypoints and the predicted keypoints in the target images. The green dots indicate the GT keypoints in the target images, and the circles indicate the PCK threshold $\alpha = 0.05$. The red lines illustrate the pixel correspondence errors between the target and predicted keypoints. Our method generates more accurate points compared to the baseline methods.

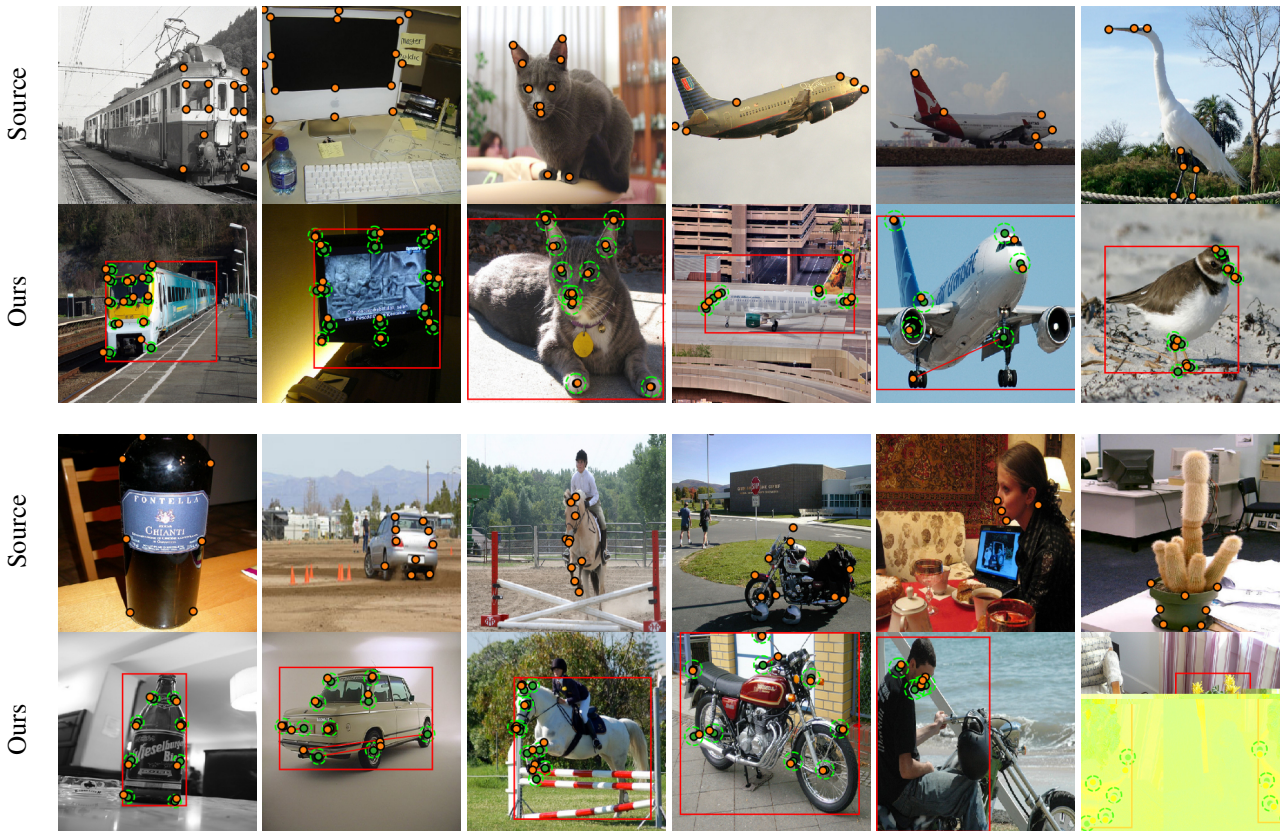

Figure 2: **Additional SPAIR-71K results.** The first and the third rows correspond to the source images with annotations. The second and the fourth rows correspond to the estimated correspondences using PMNC$_{best}$. We use the same color coding and annotation scheme as Figure 1. The provided 2D bounding boxes showing the object locations is not used by PMNC. Our method correctly estimates the semantic correspondences over different categories provided in the SPAIR-71K dataset.

3 Failure Cases

We found that for PMNC most failures occur when multiple objects are in the scene. Figure 3 shows some examples of such failure cases. For example, false negative matches occur between the bus and truck, bottle and cup, and sheep and dog.

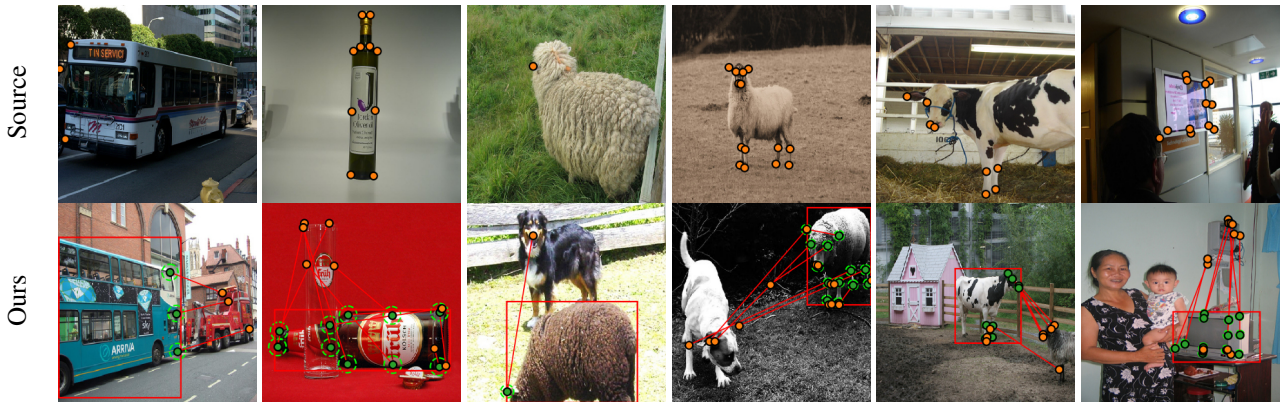

Figure 3: **Examples of failure cases for PMNC on SPAIR-71K.** Transferred keypoints are shown on the target image obtained using the result of PMNC$_{best}$. We can see that failure cases often occur because of multiple objects in the scene. The color coding scheme is the same as Figure 1.

References

[1] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondence from object proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3475–3484, 2016. 1

[2] Shuda Li, Kai Han, Theo W Costain, Henry Howard-Jenkins, and Victor Prisacariu. Correspondence networks with adaptive neighbourhood consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10196–10205, 2020. 1, 2

[3] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3395–3404, 2019. 1

[4] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *Advances in Neural Information Processing Systems*, pages 1651–1662, 2018. 1, 2