Shape-based Image Correspondence Supplementary Material

Berk Sevilmis
berk_sevilmis@brown.edu
Benjamin B. Kimia
benjamin_kimia@brown.edu

LEMS
Brown University
Providence, RI 02912 USA

1 Optimization

To introduce shape constraints into the Patch Match algorithm [2], the initial correspondence, also known as the nearest neighbor field, of point p, $\mathcal{T}^0(p)$, is defined by a conditional random assignment such that "cross-talk" is prevented, *i.e.,* the assignment is accepted only if $L(p) = \bar{L}(\mathcal{T}^0(p))$. In iteratively improving the initial correspondence, we implement conditional propagation to prevent "cross-talk", *i.e.,*

$$\mathcal{T}^{i+1}(p) = \mathcal{T}^i\Big(\underset{\hat{p}|(p,\hat{p})\in \mathrm{E}}{\operatorname{argmin}} f_{\mathrm{data}}(\mathcal{N}_M(p), \mathcal{N}_M(\mathcal{T}^i(\hat{p}) + (p-\hat{p})))\Big) + (p-\hat{p}) \implies L(p) = \bar{L}(\mathcal{T}^i(\hat{p}^*) + (p-\hat{p}^*)) \tag{1}$$

where $\hat{p}^*$ is the argument achieving the minimum. Conditional random search follows similarly.

An additive shape-based energy term can be introduced into the variational formulations of SIFT Flow [9], DSP [7] and SSF [11] or it can be readily incorporated into their default data term. We opt for the latter and hence obtain a conditional data fidelity term, *i.e.*,

$$f_{\mathrm{data}}(\mathcal{N}_M(p), \mathcal{N}_{\bar{M}}(\mathcal{T}(p))) = \begin{cases} \|SD(\mathcal{N}_M(p)) - SD(\mathcal{N}_M(\mathcal{T}(p)))\|_\alpha, & L(p) = \bar{L}(\mathcal{T}(p)) \\ \beta, & L(p) \neq \bar{L}(\mathcal{T}(p)) \end{cases}. \tag{2}$$

$\alpha = 1$, $\beta = (255)\cdot(128)$ are used in SIFT Flow [9] and SSF [11] whereas $\alpha = 2$, $\beta = \sqrt{(255)^2 \cdot (128)}$ are used in DSP [7]. Using unsigned 8-bit integers, β represents the maximum possible SIFT descriptor [10] difference. We optimize the modified energy functional using loopy belief propagation.

2 Training Object Proposal Ranker

We use the publicly available implementations of R-CNN [6], MCG [1] and CPMC [3]. Starting from the network parameters of [8], R-CNN [6] fine-tunes the CNN parameters to establish object detection on 20 PASCAL object categories [5]. It provides bounding box detections. MCG [1] and CPMC [3] output category independent object proposals, in the

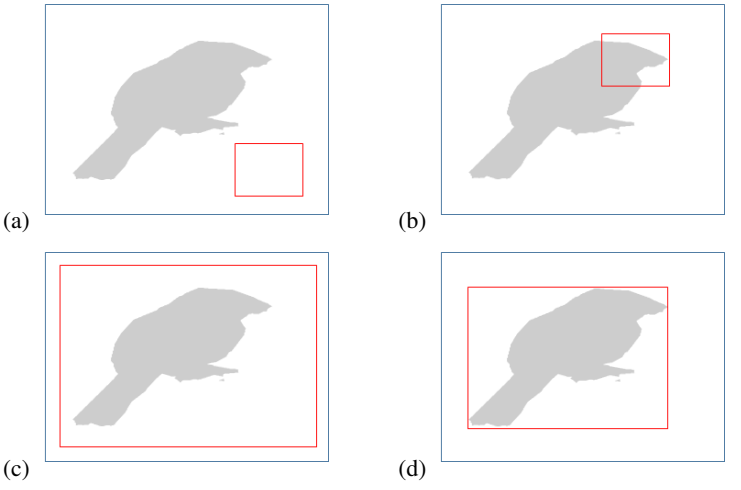

Figure 1: **Object localizations.** The bounding box in red represents the detection output by R-CNN [6]. (a) Object is wrongly detected. (b) Object part is detected. (c) Object is detected but bounding box is not tight. (d) Object is detected with tight bounding box.

form of segmentations, along with a ranker trained in a category independent fashion and also use PASCAL datasets [5] for hyperparameter tuning of their pipelines. In order not to affect our conclusions, we only experiment with CUB-200-2011 dataset [12] on which MCG [1] and CPMC [3] algorithms have not been trained. Since "bird" is among the 20 object categories in the PASCAL dataset [5], we output the highest scoring bounding box in bird category to localize birds in image pairs taken from CUB-200-2011 dataset [12]. Fig. 1 demonstrates four such possible resulting localizations.

During test time which of the object localizations we operate on is unknown. Hence it is necessary to robustly propose shape constraints, using segmentations provided by MCG [1] and CPMC [3], while taking the possible localizations, provided by R-CNN [6], of the object into consideration. This constitutes an optimization problem, picking the object proposal achieving the highest Jaccard index with the ground truth segmentation of the object from a set, with prior localization which we tackle as follows. For the case shown in Fig. 1a, it is most wise to fully resort to ranker decisions provided by MCG [1] and CPMC [3] as the object is wrongly detected. Similarly for the case shown in Fig. 1c, as the detected bounding box is not tight, any measures such as recall, precision used to pick the object proposal can end up having background clutter or might not fully delineate the object from the background respectively. The case shown in Fig. 1b represents an object part detection. It is reasonable to pick a proposal which is ranked high by the default rankers of MCG [1] and CPMC [3], and having a high recall with the bounding box output by R-CNN [6]. Using recall as a measure helps select those proposals that successfully delineate the object from the background as the region enclosed by the bounding box provided by R-CNN [6] would be a subset of the region enclosed by bounding boxes of such proposals. Finally Fig. 1d shows a successfully localized object with tight bounding box. In such a case, the object proposal whose bounding box achieves the highest Jaccard index with the bounding box provided by the R-CNN [6] can be safely picked.

We retrain object proposal rankers accounting the localization ambiguities and the above-

mentioned strategies. We pick 5 object proposals for each case from MCG [1] and CPMC [3] totaling 30 proposals per image and train a structured prediction framework using the feature space representation of [3] as explained in the body of our work. During test time, the top ranked proposal within a set of 30 proposals from each image is then used as a shape constraint.

3 Additional Results

We provide several other visual results for subsections in the paper.

Exploring the role of shape in image correspondence:
The per part dense correspondence performances of object classes in the PASCAL-Part dataset [4] are given in Figs. 2, 3, 4 and 5 whereas additional visual dense correspondence results using CUB-200-2011 dataset [12] and PASCAL-Part dataset [4] are shown in Figs. 6, 7, and 8 respectively. The abbreviations of part names follow the PASCAL-Part dataset [4].

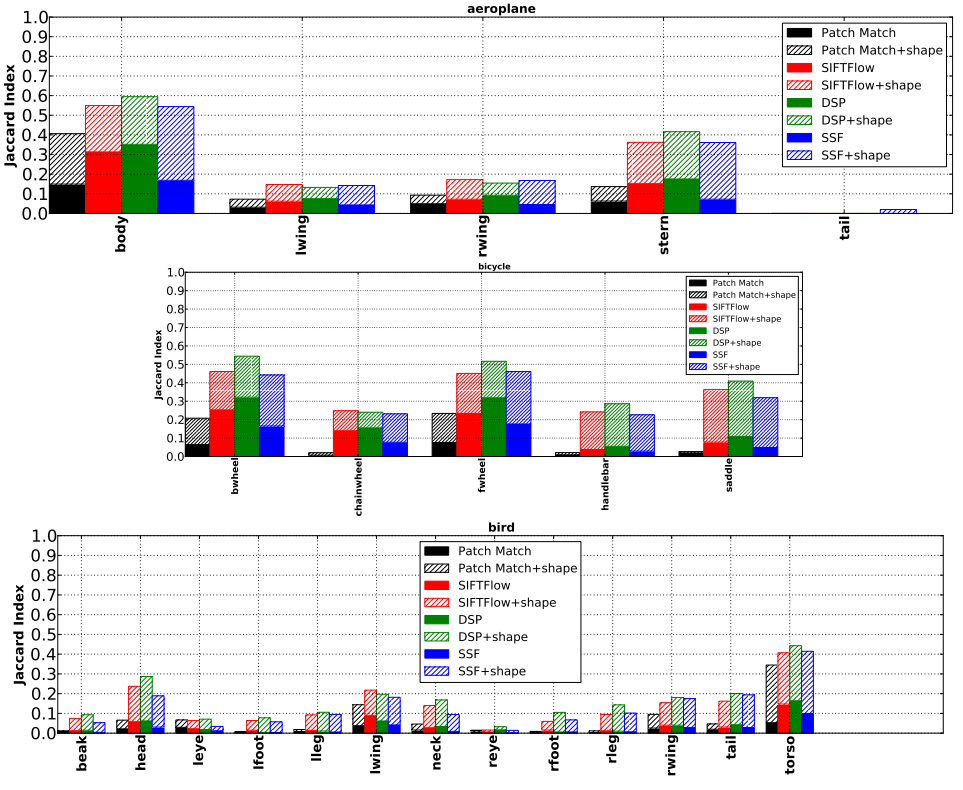

Figure 2: **Per part Jaccard index performances of object classes in the PASCAL-Part dataset [4].**

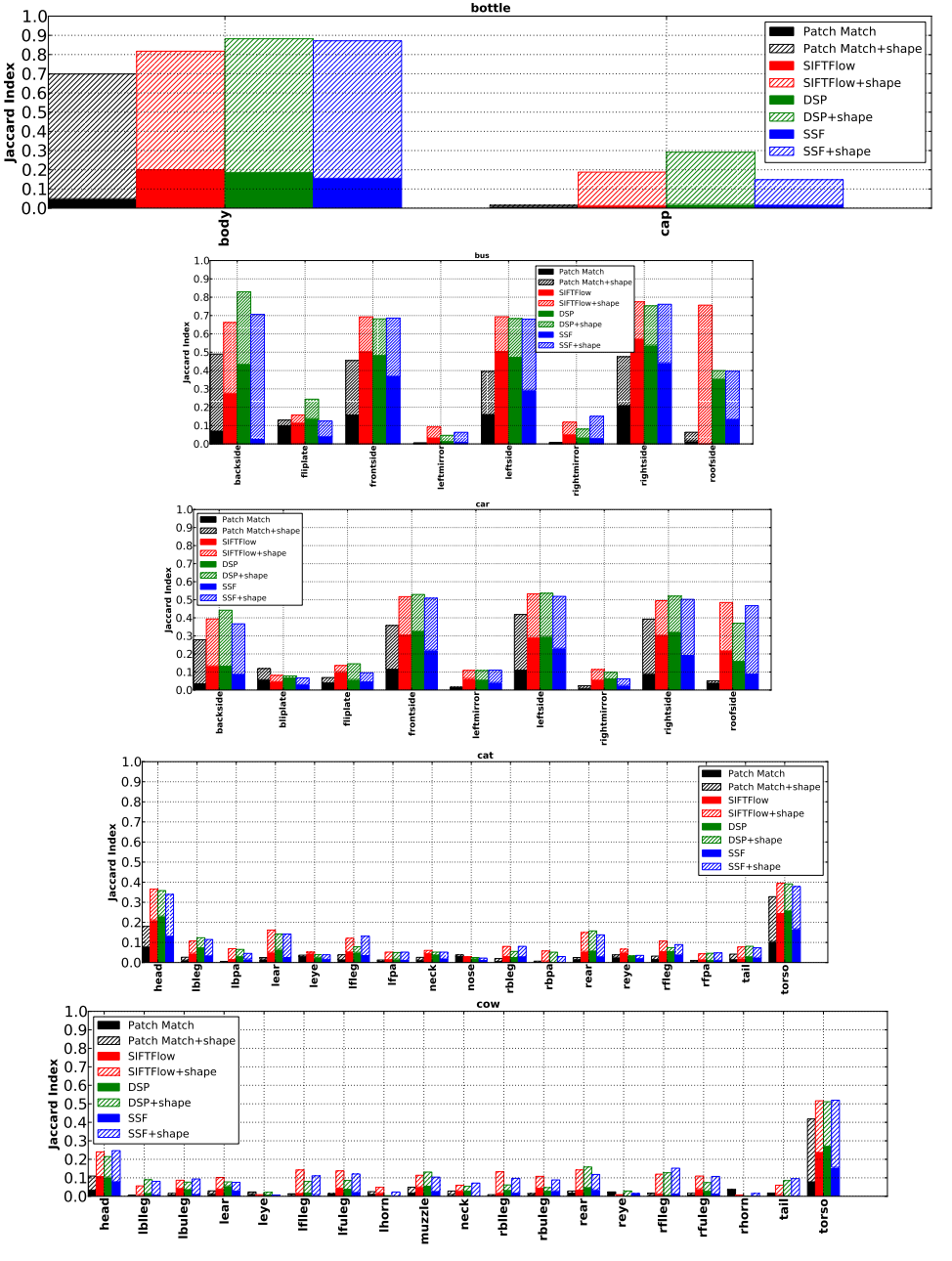

Figure 3: **(cont.) Per part Jaccard index performances of object classes in the PASCAL-Part dataset [4].**

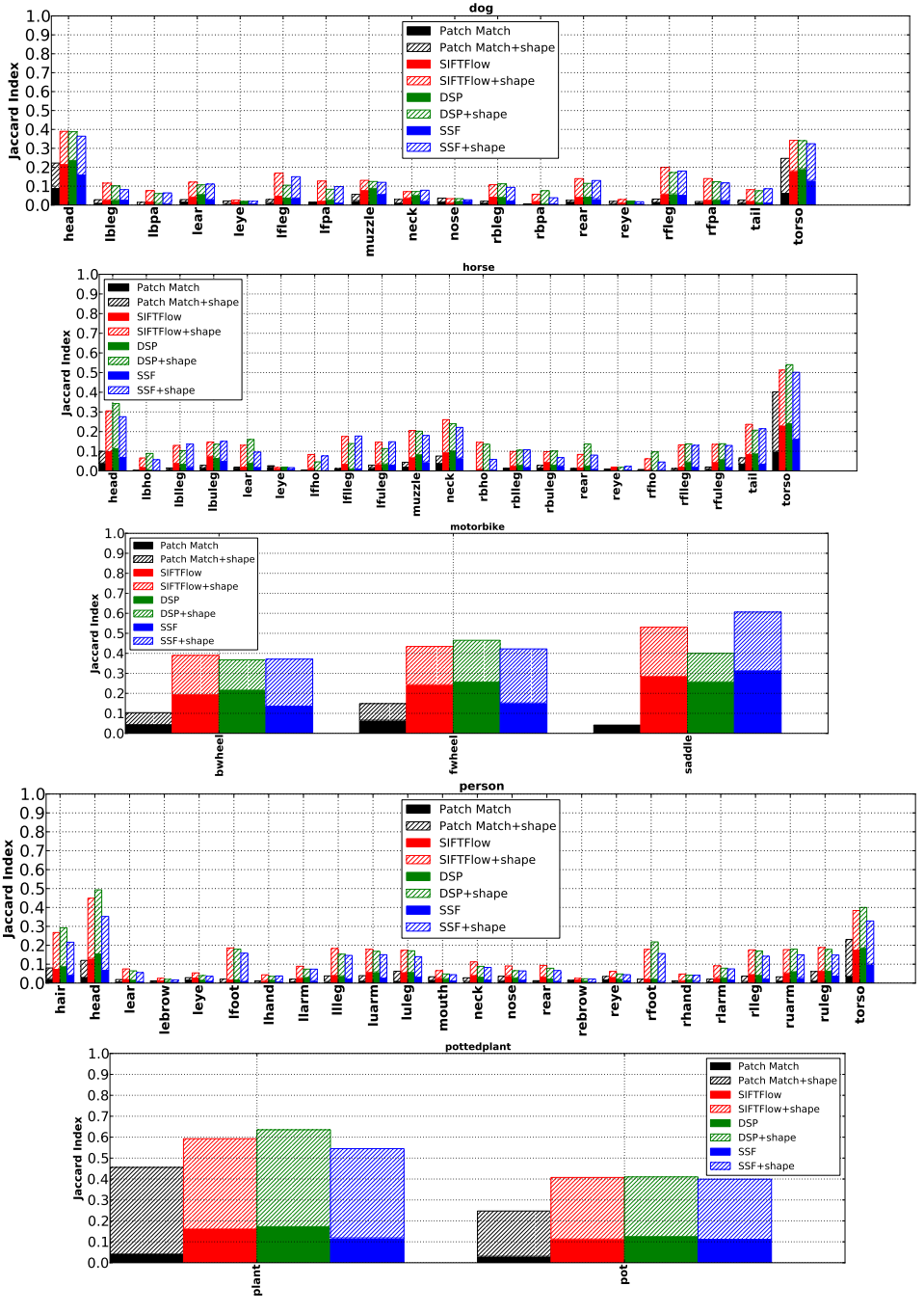

Figure 4: **(cont.) Per part Jaccard index performances of object classes in the PASCAL-Part dataset [4].**

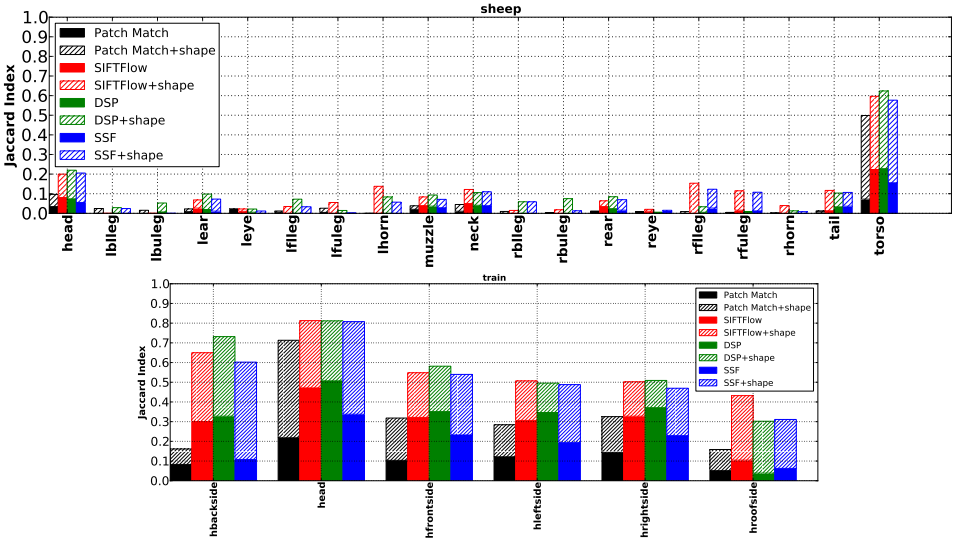

Figure 5: **(cont.) Per part Jaccard index performances of object classes in the PASCAL-Part dataset [4].**

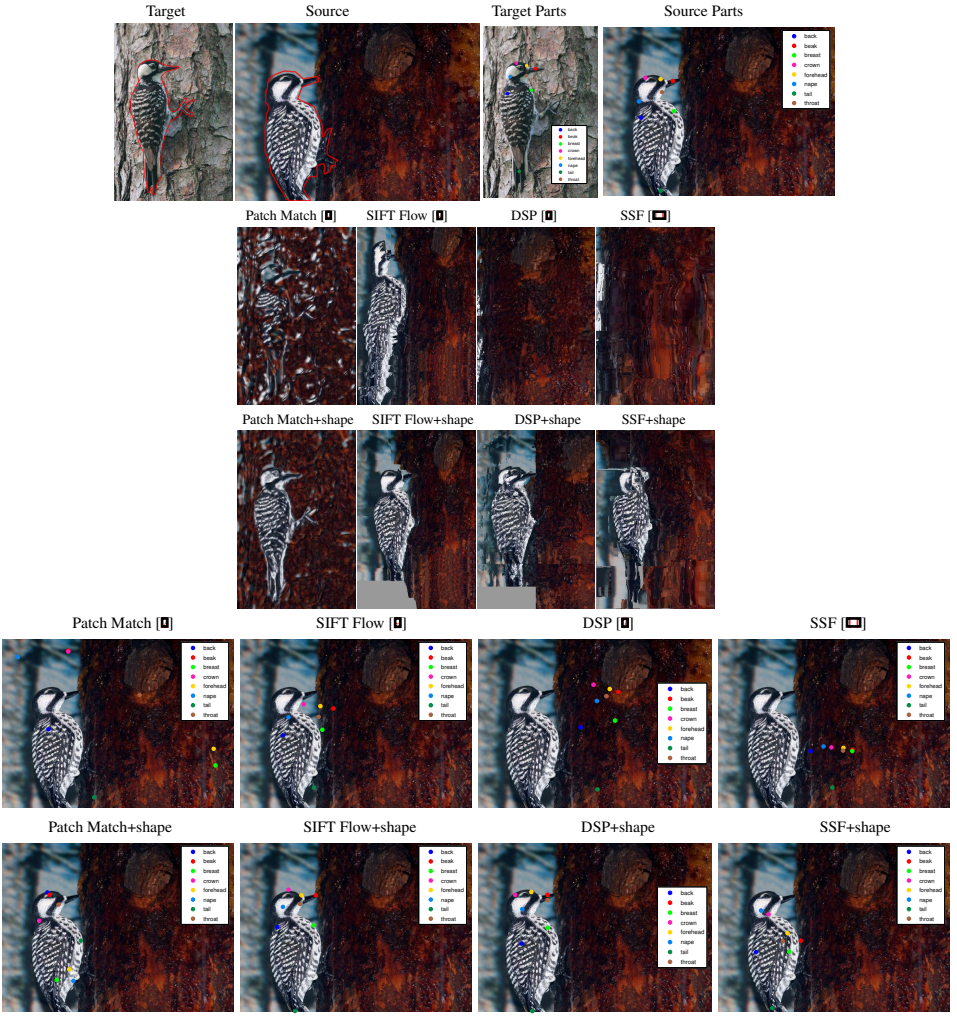

Figure 6: **A qualitative result from the CUB-200-2011 dataset [12].** First row shows the target and source images with ground truth segmentations marked in red along with the target and source image part locations. Second and third row show the warped source image using the default settings and with shape constraints respectively. The expectation is that the warped source image reconstructs the target image with the appearance of the source image. The fourth and fifth row show the matchings of the target image part locations on the source image using the default settings and with shape constraints respectively. Ideally, these should be equal to the source image part locations shown in first row fourth column.

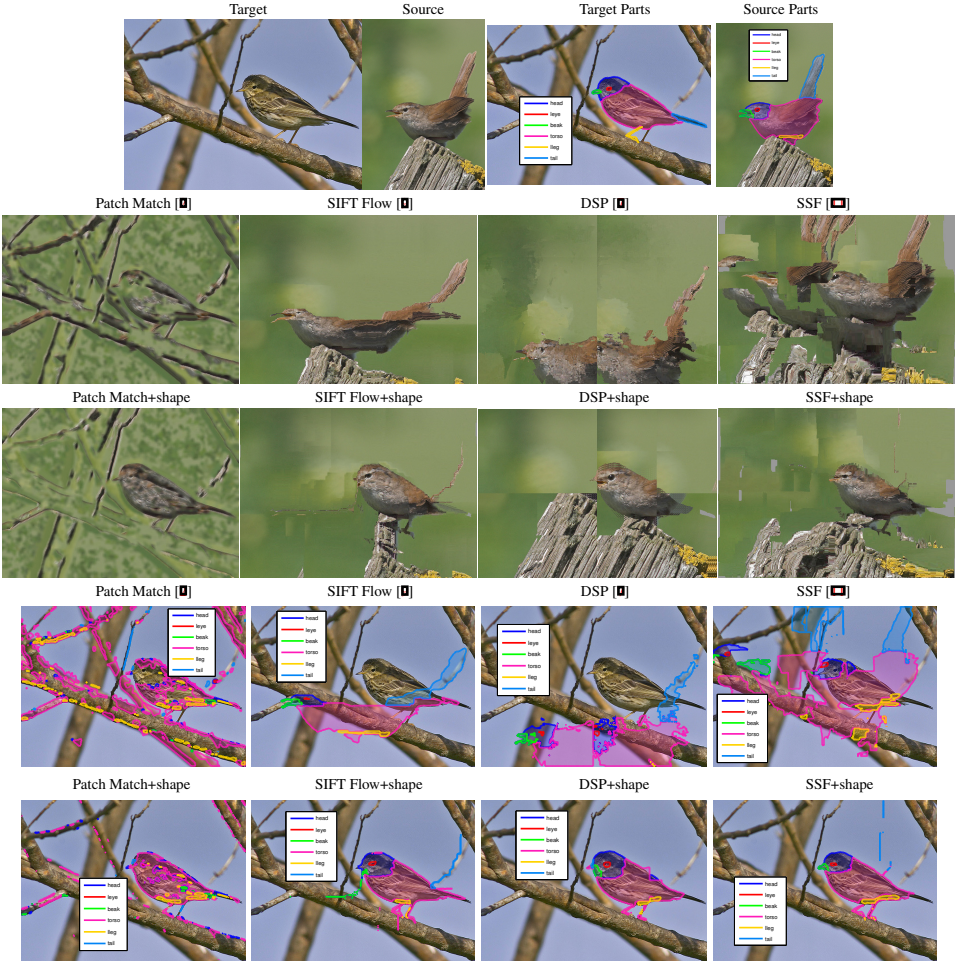

Figure 7: **A qualitative result from the PASCAL-Part dataset [3].** First row shows the target and source images with ground truth segmentations marked in red along with the target and source image part masks. Second and third row show the warped source image using the default settings and with shape constraints respectively. The expectation is that the warped source image reconstructs the target image with the appearance of the source image. The fourth and fifth row show the part mask transfer from the source image to the target image using the default settings and with shape constraints respectively. Ideally, the part masks transferred from the source image should coincide with the part masks of the target image.

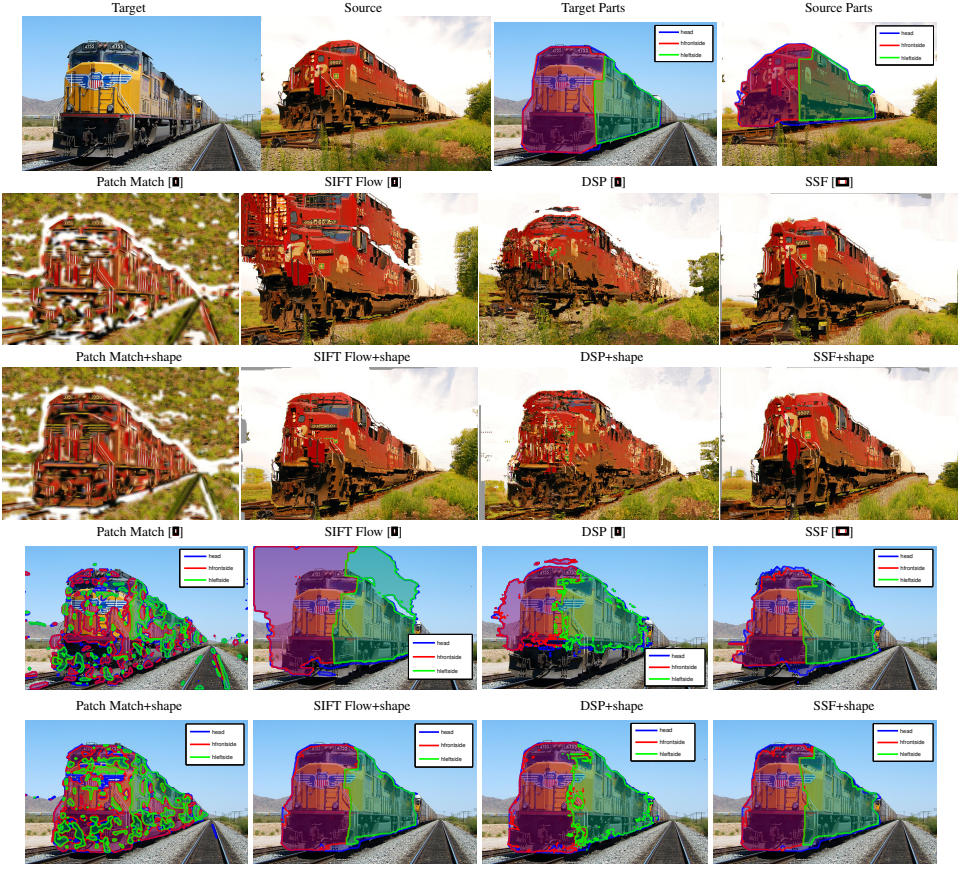

Figure 8: **A qualitative result from the PASCAL-Part dataset [4].** First row shows the target and source images with ground truth segmentations marked in red along with the target and source image part masks. Second and third row show the warped source image using the default settings and with shape constraints respectively. The expectation is that the warped source image reconstructs the target image with the appearance of the source image. The fourth and fifth row show the part mask transfer from the source image to the target image using the default settings and with shape constraints respectively. Ideally, the part masks transferred from the source image should coincide with the part masks of the target image.

References

[1] Pablo Andrés Arbeláez, Jordi Pont-Tuset, Jonathan T. Barron, Ferran Marqués, and Jitendra Malik. Multiscale combinatorial grouping. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 328–335, 2014.

[2] Connelly Barnes, Eli Shechtman, Dan B. Goldman, and Adam Finkelstein. The generalized patchmatch correspondence algorithm. In *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III*, pages 29–43, 2010.

[3] João Carreira and Cristian Sminchisescu. CPMC: automatic object segmentation using constrained parametric min-cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(7):1312–1328, 2012.

[4] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, June 2010.

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[7] Jaechul Kim, Ce Liu, Fei Sha, and Kristen Grauman. Deformable spatial pyramid matching for fast dense correspondences. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 2307–2314, 2013.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1106–1114, 2012.

[9] Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT flow: Dense correspondence across scenes and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):978–994, 2011.

[10] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[11] Weichao Qiu, Xinggang Wang, Xiang Bai, Alan L. Yuille, and Zhuowen Tu. Scale-space SIFT flow. In *IEEE Winter Conference on Applications of Computer Vision, Steamboat Springs, CO, USA, March 24-26, 2014*, pages 1112–1119, 2014.

[12] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.