Self-supervised Learning of Geometrically Stable Features Through Probabilistic Introspection

Supplementary Material

David Novotny[1,2,*] Samuel Albanie[1,*] Diane Larlus[2] Andrea Vedaldi[1]
[1]Visual Geometry Group
Dept. of Engineering Science, University of Oxford
{david,albanie,vedaldi}@robots.ox.ac.uk
[2]Computer Vision Group
Naver Labs Europe
diane.larlus@naverlabs.com

In the supplementary material below, we present an ablation study of the components of our method (section 1). In section 2, we also provide details of the weakly supervised method that produced the bounding box annotations used to train our model.

1. Ablation studies

In addition to the results reported in sections 4.2. and 4.3. of the paper, we report additional ablation experiments that validate the contribution of the proposed components of our method.

In order to show the improvements over the base architecture that was used to initialize our network, we also compare against the res5c features from the version of the pretrained ResNet-50 model, the filters of which were dilated as explained in section 3.6. in the paper (**ResNet-50-dilated**).

Furthermore, to provide an extended comparison with alternative matching loss formulations, a flavour of our method, abbreviated as **Contrastive**, implements the contrastive loss formulation from [2].

We also test three more methods that assess the sensitivity of the proposed approach to the utilized dataset. We include results for our method trained with ground truth bounding box labels (**Ours-GTbox**), rather than the weakly supervised detections used in the original formulation, to enable an assessment of the method's robustness to the usage of imperfect bounding box annotations. Another variation of our method, **Ours-NObox**, does not use any bounding box annotations. Finally, **Ours-nonrigid** uses all 20 PASCAL categories for training as opposed to the original training setup that used images of the 12 rigid classes.

All variants were evaluated on both the semantic matching and keypoint prediction tasks. The results of the semantic matching experiments are reported in fig. 1 while fig. 2 contains the results of the few-shot keypoint prediction task.

The results indicate that for both semantic matching and keypoint detection the performance of the ground-truth su-

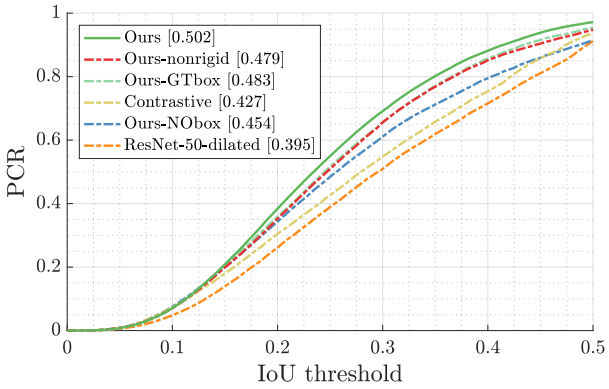

Figure 1. **Ablation study on PF-Pascal**. The region matching performance of several variants of our method (see section 1 for details of each variant).

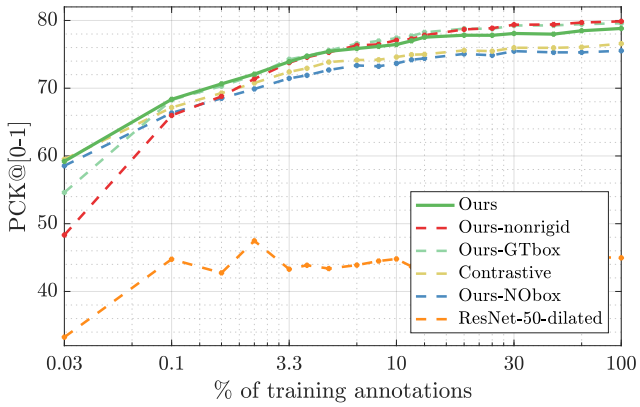

Figure 2. **Ablation study on the few-shot keypoint detection task on Pascal3D**. We report the area under the PCK-over-alpha curve as a function of the number of training annotations for several variants of our method. For details of each variant see section 1.

*Authors contributed equally.

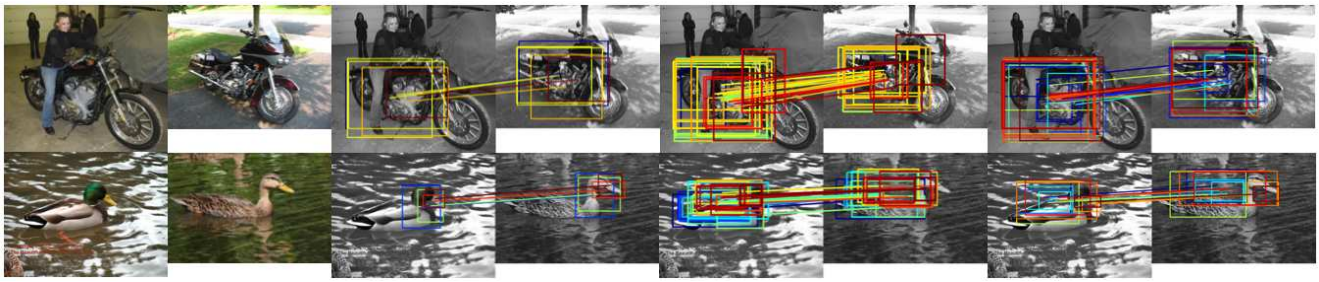

Figure 3. **Region matching examples** for pairs of motorbike (top) and duck (bottom) images. From left to right: source and target images, HOG with NAM matching [4], ours, SCNet-A [5]. We show correctly matched boxes, color-coded according to matching score (red: higher, blue: lower).

pervised setup is on par with the proposed weakly supervised setup. This shows that, with the inclusion of the probabilistic introspection mechanism, the method has good robustness to annotation noise. The performance of our method trained with the non-rigid categories is on par with the rigid case for proposal matching. We observe a decrease in performance for the keypoint detection task. This is because the few-shot detection dataset consists of only rigid classes and adding the non-rigid ones to the training set makes the features less specialized for the final task. The variant which trains features via the contrastive loss gives lower performance.

1.1. Keypoint detection - detector validation

In section 4.3. in the paper, we reported results for a keypoint detector with a design closely related to that of [6]. In order to validate the implementation of the detector, we provide a comparison against the results of the fully supervised detector from [6]. When using all available annotations and the Resnet-50-HC descriptors, the mean PCK ($\alpha = 0.1$) over the 12 rigid classes of the Pascal3D test set is 54.7. This is on par with the best single-model result from [6] (53.3 PCK), validating our keypoint predictor as a representative proxy for evaluating the quality of our feature baselines.

2. Weakly supervised detections

Here we give details of the weakly supervised detector used to provide bounding box annotations for our method, as discussed in Sec. 3.6 of the paper. We use the `vgg-f`-based model described in [1], which is trained using Edge-Box proposals[7] and the image-level labels of the Pascal VOC 2007 detection dataset [3]. To produce bounding box predictions for the ImageNet dataset, we follow the multi-scale evaluation technique described in [1], averaging predictions over five scales and flipped copies of each scale. To form our training set, we then select top scoring box for each class label present in the image. In order to maintain a high quality of box annotation, we do not include boxes whose scores fall below the median detector score of the given class (the median is computed after filtering scores which fall below the noise score threshold of 0.001 given in the public implementation[1] of [1]).

3. Qualitative results

Additional qualitative results for the semantic matching task ok PF-Pascal are present in fig. 3. We show the matching regions for two example pairs, for the method of [4], ours, and the fully-supervised method of SCNet-A.

References

[1] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. Proc. CVPR, 2016.

[2] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker. Universal correspondence network. In *Proc. NIPS*, 2016.

[3] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[4] B. Ham, M. Cho, C. Schmid, and J. Ponce. Proposal flow. In *Proc. CVPR*, 2016.

[5] K. Han, R. S. Rezende, B. Ham, K.-Y. K. Wong, M. Cho, and J. P. Cordelia Schmid. Scnet: Learning semantic correspondence. In *Proc. ICCV*, 2017.

[6] S. Tulsiani and J. Malik. Viewpoints and keypoints. In *Proc. CVPR*, 2015.

[7] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pages 391–405. Springer, 2014.

[1] https://github.com/hbilen/WSDDN