-Supplementary Material-
Exemplar-specific Patch Features for Fine-grained Recognition

Alexander Freytag[1]⋆, Erik Rodner[1]⋆, Trevor Darrell[2], and Joachim Denzler[1]

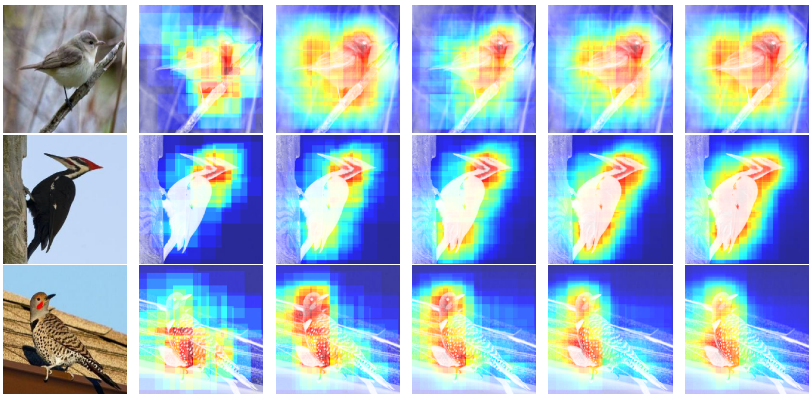
[1]Computer Vision Group, Friedrich Schiller University Jena, Germany
[2]UC Berkeley ICSI & EECS, United States

Abstract. The following document gives additional information with respect to the paper *Exemplar-specific Patch Features for Fine-grained Recognition*. Specifically, we present qualitative results of patch detection responses on unseen test images in local learning scenarios (Sect. 1). In addition, we give an analysis of resulting dimensionality for image representations using our patch discovery approach (Sect. 2). The effect of trade-off parameter λ during model combination on the final accuracy is depicted in Sect. 3. Experimental settings are given in detail, to complement our released source code (Sect. 4). Finally, patch detector responses on training images are visualized in Sect. 5. *The provided information is not necessary to understand the main paper.*

1 Patch detector responses on images using local models

In the main submission, we visualized detection responses on unseen test images for the discovered patches in *global* learning scenarios (Fig. 4). Here, we complement this

⋆ A. Freytag and E. Rodner were supported by a FIT scholarship from the DAAD.



(a) Input (b) $k = 5$ (c) $k = 10$ (d) $k = 20$ (e) $k = 40$ (f) $k = 80$

Fig. 1: Detection responses of patch detectors for different numbers of neighbors.

Table 1: Number of discovered patch detectors for global learning and local learning using the supervised bootstrapping technique presented in the main submission.

	Remove Singletons	CUB-2011-14 global	$k = 20$	$k = 40$	$k = 80$	CUB-2011-200 global	$k = 20$	$k = 40$	$k = 80$
# detectors	y	2,136	46	113	276	40,659	34	76	170
	n	2,936	140	286	569	40,659	134	267	535
Ratio	y	1	2.15%	5.29%	12.92%	1	0.08%	0.19%	0.42%
	n	1	4.78%	9.62%	19.37%	1	0.33%	0.66%	1.32%

analysis by visualizing responses for *local* models with different neighborhood sizes. Results are displayed in Fig. 1.

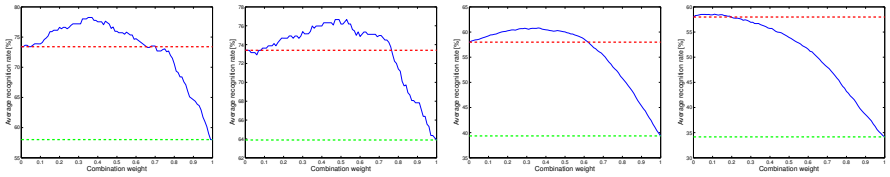
Obviously, a certain minimum number of neighbors is necessary to reliably identify patterns relevant for informative image representations, *e.g.*, the red dot in the lower birds face is not found with only 5 neighbors. In addition, we observe that easily identifiable properties, *e.g.*, the characteristic black-white pattern on the birds neck in the middle row, can be already detected with extremely small neighborhood sizes, which underlines the suitability of the global matching scheme. On the other hand, less distinctive parts like the black wing of the woodpecker can only be modeled with larger numbers of neighbors queried.

2 Dimensionality of learned representations

In Sect. 5.1 of the main submission, we analyzed in detail several steps of our discovery pipeline and the resulting recognition accuracies for global and local models. Here, we give a short overview on the average number of patch detectors for the global and local approaches in Table 1. We explicitly added results for discovery with and without removal of singletons, *i.e.*, non-representative detectors without any further correspondence during bootstrapping. Note the significant decrease in number of dimensions from global to local approaches. Please note further that in contrast to previous techniques [1,3], no selection of discriminative detectors is involved here. Nonetheless, we found our discovered representations to be already compact (see Table 1) and at the same time informative (see results in the main submission, Table 1).

3 Combination of results for semantic and discovered parts

In the paper, we showed some results for the combination of our method with the one given in [2]. The combination is based on combining the predicted class probabilities in a linear fashion using a weight coefficient λ. In Fig. 2, we show the recognition rates on the CUB-2011 fine-grained datasets depending on this parameter. Note that for the CUB-2011-200 dataset a wide range of parameter values leads to an improvement in recognition rate.



(a) CUB-2011-14 (global) (b) CUB-2011-14 (exemplar) (c) CUB-2011-200 (global) (d) CUB-2011-200 (exemplar)

Fig. 2: Plots of the fine-grained performance with respect to the combination weight λ: The red and green line show the performance of [2] and our approach without combination, respectively. The blue line shows the performance of the combination depending on the combination coefficient λ (the figure is best viewed in color).

4 Experimental setup

A standard 1-vs-all linear SVM serves as final classification model in our system, where we used the publicly available code of LibLinear[1]. We also experimented with non-linear kernels supported by LibSVM[2] and alternatively with explicit embeddings in higher dimensional spaces supported via homogeneous kernel maps[3], but found no superior results over a linear kernel by doing so. The built-in cost-parameter C was kept with its default value of 1, which allows for moderate generalization abilities. SVM bias term b was optimized during training.

For training and testing, we used the official splits of the CUB-2011 dataset for both the full 200 class set and the 14 class subset. As suggested by previous works, we cropped all images to the provided bounding box, but enlarged the window by 10% if possible. Final subwindows have been scaled to standard size of 256×265 pixels.

For the settings of individual parameters during seeding, bootstrapping, selection, and encoding, we refer to the Matlab scripts provided with the source code released[4].

5 Further impressions for image encoding

In the main submission, we visualized detection responses on previously unseen test images, to get a feeling for how good our discovered patches perform on new images. Here, we show the complementary version and visualize detection responses on the training images used during patch discovery. Results are given in Fig. 3, with good cases in the top rows and failure cases in the bottom row.

[1] http://www.csie.ntu.edu.tw/~cjlin/liblinear/
[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[3] http://www.vlfeat.org/
[4] Source code is available at http://www.inf-cv.uni-jena.de/fine_grained_recognition

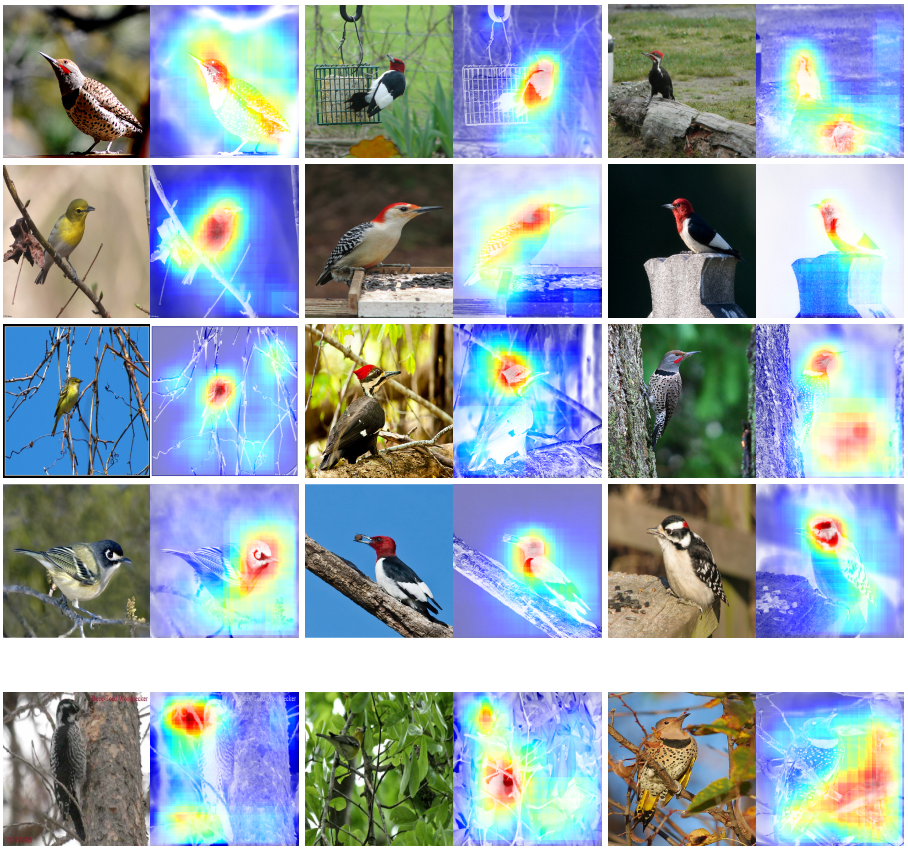


Fig. 3: Detection responses of discovered patch detectors on training images. High scores are indicated by warm colors. The lowest row displays cases where detectors are distracted by background patterns. Best viewed in color.

References

1. Doersch, C., Gupta, A., Efros, A.A.: Mid-level visual element discovery by discriminative mean-shift. In: Neural Information Processing Systems (NIPS). pp. 1–8 (2013)
2. Göring, C., Rodner, E., Freytag, A., Denzler, J.: Nonparametric part transfer for fine-grained recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–8 (2014)
3. Juneja, M., Vedaldi, A., Jawahar, C., Zisserman, A.: Blocks that shout: Distinctive parts for scene classification. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 923–930 (2013)