Classification-Specific Parts for Improving Fine-Grained Visual Categorization

Dimitri Korsch[1], Paul Bodesheim[1] and Joachim Denzler[1,2]

[1]Computer Vision Group, Friedrich-Schiller-University Jena, Germany
[2]Michael Stifel Center Jena, Germany

Abstract. Fine-grained visual categorization is a classification task for distinguishing categories with high intra-class and small inter-class variance. While global approaches aim at using the whole image for performing the classification, part-based solutions gather additional local information in terms of attentions or parts. We propose a novel classification-specific part estimation that uses an initial prediction as well as back-propagation of feature importance via gradient computations in order to estimate relevant image regions. The subsequently detected parts are then not only selected by a-posteriori classification knowledge, but also have an intrinsic spatial extent that is determined automatically. This is in contrast to most part-based approaches and even to available ground-truth part annotations, which only provide point coordinates and no additional scale information. We show in our experiments on various widely-used fine-grained datasets the effectiveness of the mentioned part selection method in conjunction with the extracted part features.

1 Introduction

Fine-grained visual categorization (FGVC) is a challenging subdiscipline of computer vision and aims at distinguishing similar classes of objects that belong to a common major class like birds [19,21], cars [11] or flowers [13]. The latest FGVC challenges (like [20]) highlight both importance and difficulties of fine-grained categorization. As shown by others before, a careful selection of data [1] or gathering additional data from the Internet [10] for the pretraining of a convolutional neural network (CNN) can yield impressive state-of-the-art results.

In general, the proposed solutions found in the literature can be divided into algorithms working with global image features [12,16] and part-based or attention-based methods [3,4,7,22,23]. From the empirical results reported in these works, it is difficult to conclude which basic approach (global or part-based) works best, since both are competitive in terms of recognition performance. Due to the fact that categories of fine-grained recognition tasks often differ only in small details, part-based features that consider local image regions seem to be promising because they are able to explicitly focus on such distinct patterns. For example, if two bird species can only be distinguished by a characteristic spot on the head, a part feature representing the image region covered by the head of the bird would be beneficial for separating these two classes. Furthermore, parts can support the classification in case of only few training samples or

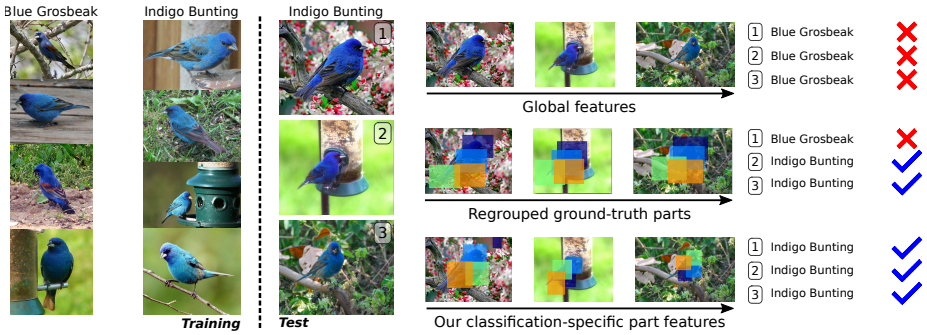

Fig. 1. An example from our experiments as a motivation for our approach: two visually similar classes are confused by a baseline classifier with global features. A part-based classifier using ground-truth annotations for anatomical parts is able to correct some of the predictions due to additional local information. However, our parts are estimated using a-posteriori classification knowledge and therefore focus on distinguishing highly similar classes. This allows for resolving misclassifications with the additional benefits of automatically determining the spatial extent of each part and being independent from manual annotations.

highly imbalanced class distributions in the training set, e.g., by applying transfer learning techniques [5]. For analyzing classification results and failure cases, the attribution of classifier decisions to features and relevant image regions is an important step and parts are helpful for detailed investigations in order to gain a better understanding of the specific recognition task and the problem domain.

If ground-truth (GT) annotations for part locations are provided, they are usually referring to an underlying concept, e.g., the anatomical parts of a bird such as head, beak, belly, wings, and legs. Although being plausible from a human perspective, these parts may not be the best choice for achieving the highest classification accuracy with a machine learning model. In addition, not all annotated parts are equally relevant for every test image and it can be shown that an optimal part selection would lead to superior performance compared to state-of-the-art methods using all available parts [8]. Especially in case of noise, few characteristic parts can be outweighed by the remaining larger set of irrelevant parts that confuse the classifier and lead to misclassifications. In our experiments, we have observed that quality of parts is more important for an improved classification accuracy than quantity. For example, if we regroup the provided ground-truth parts for the CUB-200-2011 birds dataset [21] in more coarse but also more distinct parts, namely “head”, “body”, “legs”, and “tail”, the recognition performance can be enhanced.

However, since ground-truth annotations are not available for all applications and manual part annotations are expensive, an efficient and robust part detector is required. Such a detector has to deal with the following two main questions. First, what are the interesting and important locations that enhance

the classification performance? Second, given certain part location, how and to which extent should the part features be extracted?

In this work, we tackle those questions and show that our classification-specific part estimation is able to improve classification accuracies on various fine-grained datasets. By studying failure cases of a baseline classifier that makes use of either global image features or part-based features extracted from available ground-truth annotations of the CUB-200-2011 birds dataset [21], we have observed that many class confusions occur between visually very similar classes (more details in Sect. 3 and Fig. 2). Hence, the idea of our approach is to identify relevant parts based on an initial classification with global features. By considering only the most important features for this initial decision, we estimate parts that are likely to be relevant for visually similar classes as well. With these new parts that are specifically estimated for a given test image based on additional knowledge from an initial classification, we aim at resolving misclassifications between very similar classes. This scenario is also visualized in Fig. 1.

Our proposed approach consists of the following steps. First, we perform a feature selection in order to estimate the most important features for the current classification task using a baseline classifier with global image features. The idea is then to estimate the most important regions in the image with respect to the actual classification task by only taking the most important features for the part localization into account. From these regions, we estimate parts as bounding boxes with automatically determining the spatial extent (scale) of each part. This is an advantage over most part-based approaches and ground-truth annotations. These provide only x and y coordinates of the part locations such that the size of each part has to be selected appropriately (and is usually fixed for all parts and all images). Given the newly estimated parts, we extract features for these parts as a rich representation that can then be used to improve the classification. More details are given in Sect. 3.

As it is common practice for many computer vision applications nowadays, our classification scheme relies on features computed with a CNN (called CNN features). Such features can easily be computed by applying either a pretrained CNN model as it is or a pretrained model that has been fine-tuned on the training set of the desired application. In our experiments (Sect. 4) we show that our approach: (i) improves the performance of the baseline methods, (ii) is competitive with other part-based classification approaches, and (iii) achieves state-of-the-art accuracies in some applications.

2 Related Work

Fine-grained visual categorization is a challenging and non-trivial classification task. Hence, there are diverse ways to tackle the problem. On the one hand, there are approaches that only use the global information of the image. The idea is either to use a clever way of pretraining the classifier or to use different feature pooling strategies. On the other hand, part-based approaches are applied which differ in the various part detection and extraction techniques.

2.1 Global Feature Representations

First, we consider approaches using only the global information of the image. Cui *et al.* [1] use a smart strategy in order to pretrain a CNN by taking large-scale datasets like ImageNet or iNaturalist into account, which offer a lot of data. Unfortunately, the difference between these datasets and the desired fine-grained datasets is too big. Hence, they suggest to preselect certain classes from the large-scale datasets which match best to the fine-grained training images and show that this preselection improves the performance drastically.

Krause *et al.* [10] enrich the training data with images from the Internet. They use Google Image Search in order to gather additional images for every training class. Though, the retrieved samples may not belong to the queried class, they show that even this noisy data improves the recognition performance by a large amount. Nevertheless, they cannot ensure that the collected data does not contain some images from the validation or the test set, since all these images are also publicly available. Hence, although an impressive effectiveness of noisy data is shown, one should look at the reported results with caution.

Other approaches focus on advanced ways of feature pooling. Here, bilinear pooling by Lin *et al.* [12] or more general the alpha-pooling by Simon *et al.* [16] are the most common techniques. Their aim is to highlight features that may have a greater impact on the classification task.

2.2 Part-based Recognition Approaches

The second main direction for tackling fine-grained recognition tasks consists of methods that rely on part-based representations. A straightforward way of implementing a part-based recognition system is to employ the ground-truth part annotations if they exist (e.g., for the CUB-200-2011 birds dataset [21]). Since these annotations are expensive and most fine-grained datasets do not provide them, weakly supervised part detectors are a common choice [3,7,15,23]. The only supervision that these detectors use are class label annotations.

Fu *et al.* [3] and Zheng *et al.* [23] present similar approaches to extend CNNs with attention networks. The first work considers adjusting the attention recurrently and extracting additional information defined by the attention on different scales. On the other hand, the later work extracts multiple attentions in a single step. The extraction is done by localizing interesting areas from feature maps, regrouping them, and using these grouped areas as parts. In both cases, the whole system is trained end-to-end.

He *et al.* [7] propose a sophisticated reinforcement learning method in order to estimate how many and which image regions are helpful to distinguish the categories. They use multi-scale image representations in order to localize the object and then estimate discriminative part regions.

Simon *et al.* [15] identify part proposals with the aid of back-propagation. Afterwards, these proposals are used to fit a constellation model that determines which of the proposals are more likely to identify real parts. The part proposals with the highest match are then used to extract part features on different scales.

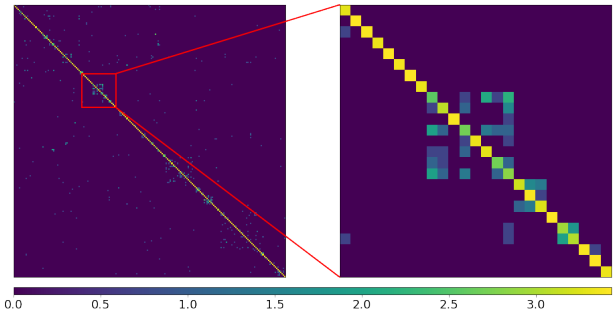

Fig. 2. Confusion matrix created from CUB-200-2011 predictions using global features only. The values are absolute number of correct predictions in log-scale. Similar classes have consecutive class indexes and are confused more often. Here you can see the classes 59-66, which are different gull species, e.g. California Gull, Herring Gull, Ivory Gull or Western Gull

3 Classification-Specific Part Estimation

In this section we describe our classification-specific part estimation approach that makes use of an initial classification based on global image features. The goal is to estimate parts depending on this first (and probably wrong) decision, such that these parts can help to spot the tiny details. These details are important for distinguishing visually similar classes in order to either confirm an initially correct classification based on the specific part features or to correct an initially wrong classification due to an enhanced representation of the important parts only. Here, we assume that many confusions of a classifier based on the global image features occur between visually very similar classes and that in those cases, the small details that are characteristic for distinguishing them are not well represented by the global features (which in general have to work for distinguishing all classes). Fig. 2 visualizes this confusion and confirms our assumption. Hence, we look for the important image regions that have led to the initial classification. Then we derive new parts from these regions under the assumption that the resulting parts are also more relevant for disentangling visually similar classes. Our estimated parts are therefore classification-specific rather than based on human knowledge, e.g., from an anatomical point of view in case of the birds.

The pipeline we propose in this paper is visualized in Fig. 3 and will be outlined in the following. First, we describe the feature selection method (Sect. 3.1). Next, based on these selected features, we illustrate how relevant pixels and image regions are identified as candidates for the part locations (Sect. 3.2). Third, we explain our algorithm for estimating bounding-box-parts with the advantage of automatically determining the scale of each part (Sect. 3.3). Finally, an overview about the part feature extraction from the classification-specific bounding-box-parts and about the part-based classification is given in Sect. 3.4.

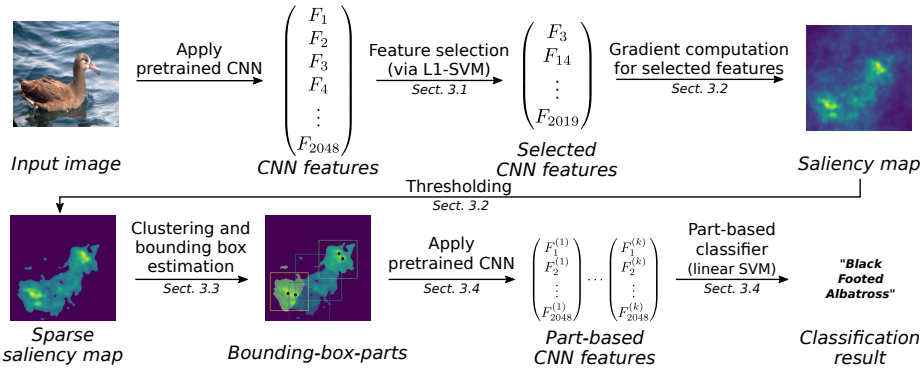

Fig. 3. The pipeline for our classification-specific part estimation.

3.1 Feature Selection

Nowadays, a common approach in computer vision tasks is to use a pretrained neural network like a CNN, fine-tune its parameters on a dataset for the desired application, and extract features by concatenating the outputs of its penultimate layer in order to obtain a high-level descriptions of the image content. Our approach relies on those CNN features, which typically results in high-dimensional feature vectors ($D = 2048$ in our case). In case of a fine-grained recognition task, the recognition system often has to focus on some specific information within the features in order to spot tiny details that distinguish two similar classes.

Therefore, we first perform a feature selection in order to estimate the most important features for the current classification task. This is done by utilizing a sparsity-inducing classifier equipped with L1-regularization, which could be either a corresponding classification layer in a CNN that allows for end-to-end learning or an L1-regularized linear SVM classifier for the CNN features. Optimization with L1-regularization forces the classifier decisions to be performed on only a small subset of the CNN features. In our experiments, we tried both and found empirically that an SVM performs better in terms of recognition accuracy while still being fast during learning due to efficient SVM solvers in standard libraries like liblinear [2].

In the end, our feature selection is classification-based and determines relevant features for the underlying task by optimizing feature weights during learning of SVM classifiers. Since we consider multi-class recognition scenarios, we train a separate classifier for each class using the "one-vs-rest" strategy. As the result, we obtain a subset of relevant features for each class that best distinguishes this class from all the other classes by only selecting features with nonzero weights.

3.2 Identifying Relevant Pixels and Image Regions

The main idea within our part detection approach is to estimate the most important regions in the image with respect to the actual classification task. Hence,

the part localization should only take those important features into account, which have been computed with the classification-based feature selection from the previous section. To this end, we use gradient maps [17] to identify the most relevant pixels in the image (indicated by large gradients), which have the largest influence on the feature extraction from the CNN model. By restricting the gradient map computations to only the previously selected subset of features, regions with large gradients are more adjusted to the classification task compared to propagating back the gradients of all features to the input image. Since our feature selection is based on a multi-class one-vs-rest SVM classifier, only selected features of the class assigned by this classifier are used for the gradient map computations. Thus, we incorporate knowledge of a baseline classifier with global features in our part detection algorithm.

The gradient maps are treated as saliency maps in our approach in order to guide the part detection and they depend on the used CNN features. In case of many currently used CNN architectures (Inception [18], ResNet [6], etc.), features are computed by averaging the values within each of the D output channel of the last convolutional layer. Typically, these output channels are called feature maps and the aforementioned average pooling results in a single number for each feature map. Given D feature maps $\boldsymbol{F}^{(1)}(\boldsymbol{I}),\ \ldots,\ \boldsymbol{F}^{(D)}(\boldsymbol{I})$ of size $s \times u$ for an image $\boldsymbol{I}$, this pooling step for computing the elements $f^{(d)}(\boldsymbol{I})$ of the D-dimensional feature vector $\boldsymbol{f}(\boldsymbol{I})$ can be written as follows:

$$f^{(d)}(\boldsymbol{I}) = \frac{1}{s \cdot u} \sum_{j=1}^{s} \sum_{j'=1}^{u} F_{j,j'}^{(d)}(\boldsymbol{I}) \quad \forall\ d \in \{1, \ldots, D\}\ . \tag{1}$$

Consequently, each value in a feature vector corresponds to a single feature map and since the feature selection method mentioned in Sect. 3.1 is applied to the feature vectors, it can also be viewed as applying the feature selection to the feature maps, i.e., the output channels of the last convolutional layer.

Like Simonyan *et al.* [17] and Simon *et al.* [15], we use back-propagation through the CNN to identify the regions of interest for each selected feature map. Based on the feature map subset $\mathfrak{D} \subset \{1, \ldots, D\}$ chosen by the feature selection from Sect. 3.1, we compute a saliency map $\boldsymbol{M}(\boldsymbol{I})$ for an image $\boldsymbol{I}$ as follows:

$$M_{x,y}(\boldsymbol{I}) = \frac{1}{|\mathfrak{D}|} \sum_{d \in \mathfrak{D}} \left| \frac{\partial}{\partial I_{x,y}} f^{(d)}(\boldsymbol{I}) \right| = \frac{1}{|\mathfrak{D}|} \sum_{d \in \mathfrak{D}} \left| \frac{\partial}{\partial I_{x,y}} \frac{1}{s \cdot u} \sum_{j=1}^{s} \sum_{j'=1}^{u} F_{j,j'}^{(d)}(\boldsymbol{I}) \right| \ . \tag{2}$$

After estimating the saliency maps, we normalize the resulting values to the range $[0 \ldots 1]$ and determine a threshold to discard pixels and regions of low saliency at an early stage. We use the mean saliency value as a threshold. We have also tested Otsu's thresholding method [14] and it achieved similar performance. The resulting sparse saliency map, which now contains only pixels with large saliency values, is used in the next step for estimating location and spatial extent of parts.

3.3 Estimating Bounding-Box-Parts

Given an image and a sparse saliency map that discards pixels with low saliency, a set of k peaks $P = \{p_1, \ldots, p_k\}$ with largest saliency can be computed using non-maximum suppression. Each peak serves as the initialization for a new part location. We then determine a region of high saliency around each peak, which directly defines the spatial extent of the estimated part. Like Zhang *et al.* [22], we achieve this by k-means clustering of pixel coordinates (x, y) and the saliencies $M_{x,y}$ (Eq. 2). Additionally, we also consider the RGB values at the corresponding positions in the input image. The clusters are initialized with the previously determined peaks $p_1, \ldots, p_k$.

This has the effect that the number of selected peaks determines the number of clusters and hence the number of parts to detect. Second, since the peaks are sorted by their saliency values, the most important part is identified by the first cluster. Afterwards, it is easy to translate the clusters into bounding boxes for the parts. For each cluster we estimate the upper left and lower right corners in order to maximize the recall of the cluster pixels surrounded by the corners. The motivation behind the recall maximization is to get bounding boxes that contain as few false negatives as possible.

The resulting bounding boxes serve as parts for the following part-based classification with the advantage that we automatically determine the spatial extent (scale) of the parts by inferring the size of the bounding boxes based on the clustering and the regression. In contrast to this, most approaches estimate only x and y coordinates of the part locations such that the size of each part has to be selected appropriately. The same holds for the ground-truth annotations of many fine-grained datasets [19,21]. In most cases, the size for all parts of an image is fixed, which is obviously not very suitable since parts often have different extents in the image, e.g., consider bird parts that correspond to an eye and a wing.

With our part estimation strategy, we are able to automatically determine different sizes for different parts depending on the content of the image. Hence, we want to emphasize again that we treat parts as bounding boxes with estimated position *and* estimated spatial extent in our framework rather than only considering point locations with fixed extent.

3.4 Part Feature Extraction and Part-based Classification

After we have estimated the bounding boxes around the maximum peaks of the sparse saliency map for each image, we extract CNN features from these bounding boxes. This is achieved by treating each bounding box as a single image that is then processed by a pretrained CNN to extract meaningful features from the penultimate layer. Note that this could even be the same CNN that was initially used to extract global image features for the part localization and we use the same CNN architecture for both steps. The resulting part features and the global features are then concatenated prior to applying a linear SVM classifier.

This classifier has been trained using part features of the training images that have been computed with our part estimation approach described before.

To summarize, our part descriptors are classification-specific in the sense that we estimate location and spatial extent of parts via bounding boxes based on an initial classifier decision with its involved feature selection, i.e., our estimated parts focus on the important aspects that are relevant for the classification.

4 Experiments

4.1 Datasets and Implementation

Datasets All of the experiments are performed on widely used fine-grained datasets. These datasets belong to a single common domain (birds, cars, flowers, etc.). Though, some of these datasets provide additional part or bounding box annotations besides the class annotations, we use only the class labels in our experiments. A short description of these datasets can be found in the following.

CUB-200-2011 [21] consists of 5994 training and 5794 test images from 200 different bird species. Besides the class labels, this dataset provides bounding box and part annotations.

NA-Birds [19] is similar to the CUB-200-2011 dataset. It provides besides class annotations also ground-truth part annotations. This dataset is more challenging, since it has 555 classes spread over 23 929 training and 24 633 test images. Although there are more training samples, the training set is not as balanced as the training set of the CUB-200-2011 dataset. Additionally, this dataset provides a hierarchy information about the classes.

Stanford-Cars [11] contains 8144 training and 8041 test images for 196 car models. This dataset provides only bounding box annotations.

Flowers-102 [13] has 102 different flower species spread over 2040 training and 6149 test images. Class labels are the only provided annotations.

Implementation As backbone for our method, we use the ResNet-50 [6] CNN architecture for Stanford-Cars and Inception-V3 CNN architecture [18] for the other datasets. For different datasets we use CNN weights proposed by Cui *et al.* [1]. These weights are pretrained on either the ImageNet or the iNaturalist 2017 dataset. To allow for fair comparisons with the recognition performances mentioned in [1], we use ImageNet weights for Stanford-Cars and iNaturalist weights for all other datasets. This separation makes sense, since iNaturalist consists of living things only, which matches the datasets of flowers and birds best. On the other hand, ImageNet classes are more variable and contain also objects and vehicles, which is more suitable for a dataset of car images. For every fine-grained dataset, we fine-tune a CNN on the corresponding training set, perform the feature selection, part localization and part extraction. Finally, the extracted part features and the global feature are concatenated and a linear SVM classifier is trained. In order to match the number of regrouped ground-truth parts for the CUB-200-2011 dataset mentioned in the introduction, we use $k = 4$ in the part localization step, which results in four parts.

Table 1. Comparison of our part extraction algorithm with and without our proposed feature selection method (**bold** = best per dataset).

	CUB-200-2011	NA-Birds	Flowers-102	Stanford-Cars
Global features (baseline)	88.5	87.5	**97.8**	91.5
Our parts				
no feature selection	89.1	88.4	97.0	92.2
with feature selection (Sect. 3.1)	**89.5**	**88.5**	96.9	**92.5**

4.2 Results

Feature Selection Evaluation First, we show that the usage of a classification-specific feature selection improves the quality of the extracted parts. For this experiment, we first compute the gradients from the entire feature vector with respect to the input image. Based on this gradient, we detect parts and extract features as mentioned before in Sections 3.2, 3.3, and 3.4. The results obtained with these features can then be compared to the results of our approach. Although the features derived from gradients of the entire feature vector improve the recognition performance compared to the linear classification baseline, using feature selection as presented in Sect. 3.1 yields a larger improvement, as shown in Table 1. We observe that the feature selection is an important ingredient in our approach. The additional information, in form of the part features determined from the gradients of the entire feature vector, improves the classification performance. Nevertheless, the benefits of the feature selection indicate that this additional information should be picked with care.

The saliency maps that determine the parts are computed by the sum of the gradients of every single CNN feature with respect to the input image (Eq. 2). Hence, the feature selection reduces the summation to selected gradients only. This means that in the experiment with feature selection, we use less information but this information is more precise which results in better classification performance. These findings hold for all presented datasets except for the flowers dataset. In case of flowers we see that the baseline linear classifier performs best. One possible explanation is overfitting to the training data. Compared to other datasets, there are on average only 20 training samples per class. The other datasets contain an average of 30 to 40 samples per class.

Furthermore, as Fig. 4 shows, the number of selected features by a L1-regularized linear classifier is beneath 3 % for used datasets. As a consequence, the number of aggregated gradients (Eq. 2) is also beneath 3 %. This fact and the results from Table 1 confirm the assumption, that quality of selected information is more important than the quantity.

Part Feature Evaluation Second, we compare the recognition performance of our extracted parts with the one obtained with ground-truth parts. We have chosen the CUB-200-2011 dataset for this experiment, since it is one of the few datasets that provides these annotations. As indicated in the introduction, we also regrouped the provided ground-truth parts in more coarse but also more

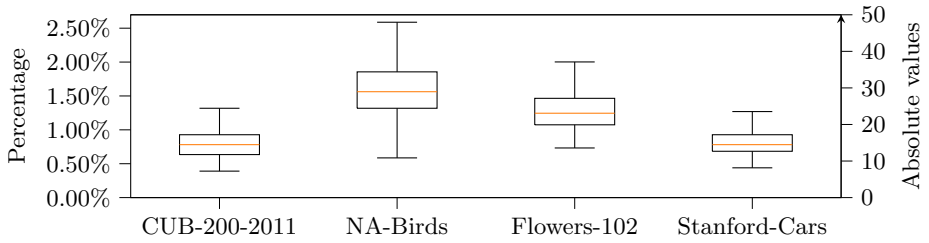

Fig. 4. Since we perform multi-class classification with the "one-vs-rest" strategy, we obtain for each class a vector of sparse weights for the linear SVM due to the L1 regularization. The distribution of the number of nonzero weights over the different classes (different "one-vs-rest" models) is shown for each dataset that is used in our experiments. Note that both relative and absolute quantities are shown (for 2048 features in total), which correspond to the number of selected features.

Table 2. Evaluation of the extracted parts on the CUB-200-2011 dataset. Note that we have used the same CNN to extract features from the different part locations: ground-truth (GT), regrouped GT, NAC parts [15], and our classification-specific parts (**bold** = best, *italic* = best without GT annotations).

	global features	parts only	parts + global	# of parts
Global features (baseline)	88.5	-	-	-
GT parts	-	87.9	89.8	15
Regrouped GT parts	-	86.9	**90.2**	4
NAC part locations of [15]	-	87.9	89.0	20
Our parts	-	87.4	*89.5*	4

distinct parts, namely "head", "body", "legs", and "tail". Compared to original ground-truth part annotations, our experiments show that these parts yield a better recognition performance (Table 2). This indicates again that the quality of parts is more important than the quantity. In the same table, we compare our classification-specific part detection with the part-based approach of Simon *et al.* [15], who have provided their extracted part locations. Additionally, we report in the table the recognition based on the global feature only. Best results are achieved when combining part features and the global image feature. While using ground-truth part annotations is slightly better, we are able to achieve better recognition results than the NAC parts proposed by Simon *et al.* [15]. Thus, our approach yields competitive recognition accuracies without relying on ground-truth part annotations, which makes it applicable in a wider range of applications where part annotations are not available.

Comparison to State-of-the-Art Finally, we compare our proposed method with current state-of-the-art approaches on commonly used fine-grained datasets.

Table 3. Comparison of our part-based approach for fine-grained recognition with various state-of-the-art methods (**bold** = best, *italic* = best part-based).

		CUB-200-2011	NA-Birds	Flowers-102	Stanford-Cars	maximum # of parts
Global features	Linear SVM (baseline)	88.5	87.5	**97.8**	91.5	-
	Lin *et al.* [12]	84.1	-	-	91.3	-
	Simon *et al.* [16]	86.5	-	96.7	91.6	-
	Cui *et al.* [1]	89.6	87.9	97.7	**93.5**	-
Part-based features	Simon *et al.* [15]	81.0	-	95.3	-	20
	Krause *et al.* [9]	82.0	-	-	92.6	30
	Fu *et al.* [3]	85.3	-	-	92.5	2
	Zhang *et al.* [22]	85.4	-	-	92.3	4
	Zheng *et al.* [23]	86.5	-	-	92.8	5
	He *et al.* [7]	87.2	-	-	*93.3*	15
	Ge *et al.* [4]	***90.4***	-	-	-	10
	Our parts	89.5	***88.5***	*96.9*	92.5	4

The results are shown in Table 3 and the mentioned baseline uses only global image features extracted from the whole image. Furthermore, we differentiate between methods that use only the global information and part-based methods. Besides the method of Cui *et al.*, which utilizes clever pretraining of the CNN weights, we compare to other methods that use sophisticated pooling methods: bilinear pooling [12] and alpha-pooling [16]. We also report recognition results and the number of used parts for other part-based approaches. Note that none of the approaches used ground-truth part annotations, neither during training nor in the test phase.

Table 3 shows that our approach is competitive in various fine-grained applications and achieves state-of-the-art performance on the NA-Birds dataset. For the CUB-200-2011 dataset, we outperform a lot of part-based methods even if they are using much more parts. This highlights once again that the quality of the parts is important and that our estimated parts contain meaningful information in only four locations.

5 Conclusion

In this paper, we proposed a weakly supervised classification-specific part estimation approach for fine-grained visual categorization. Unlike other part-based approaches, we estimate the part extents based on an initial classification of the whole image. We have shown that part features extracted in a classification-specific manner result in improved categorization performance. Furthermore, each estimated bounding box part has an implicit spatial extent that automatically determines an appropriate scale of the part.

References

1. Cui, Y., Song, Y., Sun, C., Howard, A., Belongie, S.: Large scale fine-grained categorization and domain-specific transfer learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (6 2018). https://doi.org/10.1109/cvpr.2018.00432
2. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. Journal of machine learning research **9**(Aug), 1871–1874 (2008)
3. Fu, J., Zheng, H., Mei, T.: Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (7 2017). https://doi.org/10.1109/cvpr.2017.476
4. Ge, W., Lin, X., Yu, Y.: Weakly supervised complementary parts models for fine-grained image classification from the bottom up (2019)
5. Göring, C., Rodner, E., Freytag, A., Denzler, J.: Nonparametric part transfer for fine-grained recognition. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2489–2496 (2014)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. He, X., Peng, Y., Zhao, J.: Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization. International Journal of Computer Vision pp. 1–21 (2019). https://doi.org/10.1007/s11263-019-01176-2
8. Korsch, D., Denzler, J.: In defense of active part selection for fine-grained classification. Pattern Recognition and Image Analysis pp. 658–663 (2018). https://doi.org/10.1134/S105466181804020X
9. Krause, J., Jin, H., Yang, J., Fei-Fei, L.: Fine-grained recognition without part annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5546–5555 (2015). https://doi.org/10.1109/cvpr.2015.7299194
10. Krause, J., Sapp, B., Howard, A., Zhou, H., Toshev, A., Duerig, T., Philbin, J., Fei-Fei, L.: The unreasonable effectiveness of noisy data for fine-grained recognition. In: European Conference on Computer Vision. pp. 301–320. Springer (2016)
11. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13) (2013). https://doi.org/10.1109/iccvw.2013.77
12. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: The IEEE International Conference on Computer Vision (ICCV). pp. 1449–1457 (2015). https://doi.org/10.1109/iccv.2015.170
13. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (Dec 2008)
14. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics **9**(1), 62–66 (1979)
15. Simon, M., Rodner, E.: Neural activation constellations: Unsupervised part model discovery with convolutional networks. In: The IEEE International Conference on Computer Vision (ICCV) (December 2015)

16. Simon, M., Rodner, E., Darell, T., Denzler, J.: The whole is more than its parts? from explicit to implicit pose normalization. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–13 (2018). https://doi.org/10.1109/TPAMI.2018.2885764
17. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
18. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
19. Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 595–604 (June 2015). https://doi.org/10.1109/cvpr.2015.7298658
20. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018). https://doi.org/10.1109/cvpr.2018.00914
21. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
22. Zhang, J., Zhang, R., Huang, Y., Zou, Q.: Unsupervised part mining for fine-grained image classification. arXiv preprint arXiv:1902.09941 (2019)
23. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: The IEEE International Conference on Computer Vision (ICCV) (2017). https://doi.org/10.1109/iccv.2017.557