Simplified Concrete Dropout - Improving the Generation of Attribution Masks for Fine-grained Classification

Dimitri Korsch[1 [0000−0001−7187−1151]], Maha Shadaydeh[1 [0000−0001−6455−2400]], and Joachim Denzler[1 [0000−0002−3193−3300]]

Computer Vision Group, Friedrich Schiller University Jena, Jena, Germany
{dimitri.korsch,maha.shadaydeh,joachim.denzler}@uni-jena.de
https://inf-cv.uni-jena.de

Abstract. Fine-grained classification is a particular case of a classification problem, aiming to classify objects that share the visual appearance and can only be distinguished by subtle differences. Fine-grained classification models are often deployed to determine animal species or individuals in automated animal monitoring systems. Precise visual explanations of the model's decision are crucial to analyze systematic errors. Attention- or gradient-based methods are commonly used to identify regions in the image that contribute the most to the classification decision. These methods deliver either too coarse or too noisy explanations, unsuitable for identifying subtle visual differences reliably. However, perturbation-based methods can precisely identify pixels causally responsible for the classification result. *Fill-in of the dropout* (FIDO) algorithm is one of those methods. It utilizes the *concrete dropout* (CD) to sample a set of attribution masks and updates the sampling parameters based on the output of the classification model. A known problem of the algorithm is a high variance in the gradient estimates, which the authors have mitigated until now by mini-batch updates of the sampling parameters. This paper presents a solution to circumvent these computational instabilities by simplifying the CD sampling and reducing reliance on large mini-batch sizes. First, it allows estimating the parameters with smaller mini-batch sizes without losing the quality of the estimates but with a reduced computational effort. Furthermore, our solution produces finer and more coherent attribution masks. Finally, we use the resulting attribution masks to improve the classification performance of a trained model without additional fine-tuning of the model.

Keywords: Perturbation-based counterfactuals · fine-grained classification · attribution masks · concrete dropout · gradient stability.

1 Introduction

Fine-grained classification tackles the hard task of classifying objects that share the visual appearance and can only be distinguished by subtle differences, e.g.,

animal species or car makes. Most commonly, fine-grained classification models are employed in the field of animal species recognition or animal individual identification: classification of insects [3,19] and birds [14,21,37], or identification of elephants [24], great apes [4,17,35,45], and sharks [15]. Even though these automated recognition systems surpass humans in terms of recognition performance, in some cases, an explanation of the system's decision might be beneficial even for experts. On the one hand, explanations might help in cases of uncertainty in human decisions. On the other hand, it can help to feedback information to the developer of the system if systematic errors in the decision are observable. Those systematic errors might be spurious biases in the learned models [32] and could be revealed by inspection of a highlighted region that should not be considered by a classification model.

Even though various methods [14,21,24,37] were presented in the context of fine-grained recognition to reliably distinguish classes with subtle visual differences, these methods offer either a too coarse-grained visual explanation or an explanation with many false positives.

Attention-based methods [13,14,47], for example, introduce attention mechanisms to enhance or diminish the values of intermediate features. They operate on intermediate feature representations, which always have a much lower resolution than the input image. Hence, upscaling the low-resolution attention to the higher-resolution image cannot highlight the fine-grained areas, which are often important for a reliable explanation of the decision.

Gradient-based methods [38,39,40] identify pixel-wise importance by computing the gradients of the classifier's decision w.r.t. the input image. These methods identify much finer areas in the image and enable decision visualization on the fine-grained level. However, these methods may also falsely highlight background pixels as has been shown in the work of Shrikumar *et al.* [36] and Adebayo *et al.* [1].

In this paper, we build upon a perturbation-based method, the fill-in of the dropout (FIDO) approach, proposed by Chang *et al.* [5]. The idea behind FIDO is to perturb the pixel values of the input image and observe the change in the classification decision. The authors realize the perturbation with a binary mask, whose entries model a binary decision whether to perturb a pixel or not. The mask is sampled using a set of trainable parameters and a sampling method introduced as concrete dropout (CD) by Gal *et al.* [9]. After optimizing the trainable parameters w.r.t. the classification decision, the parameters represent the importance of each pixel for the classification. One drawback of the approach is the high variance in estimating the gradients while optimizing the sampling parameters. Chang *et al.* mention this drawback in their work, and suggest reducing the variance by averaged gradients over a mini-batch of sampled masks.

We propose a mathematically equivalent but simplified version of CD. As a consequence, we reduce the amount of exponential and logarithm operations during the sampling procedure resulting in more stable gradient computations. We will show that the FIDO algorithm becomes less reliant on the size of the mini-batches, which allows the estimation of the attribution masks with smaller

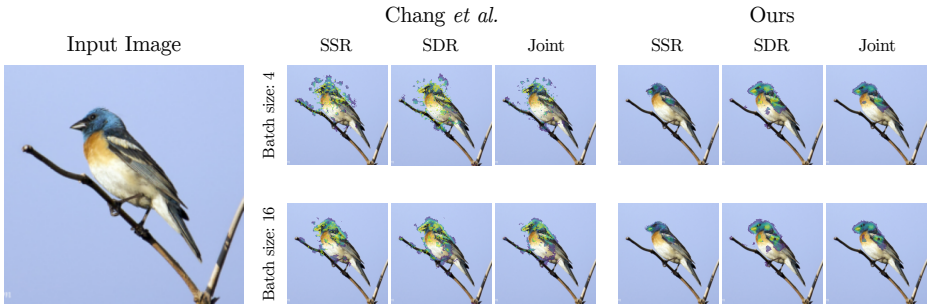


Fig. 1: Visual comparison of the original FIDO method and our proposed improvement. We estimated the masks using 30 optimization steps and two different batch sizes: 4 and 16. Our method produces masks that do not differ much for the visualized batch sizes. In contrast, the method of Chang *et al.* [5] strongly depends on higher batch sizes and produces more wrongly attributed pixels (e.g., the background or the tree branch) if the mini-batch size of 4 or less is used. Similarly to Chang *et al.*, we visualized only the mask values above the threshold of 0.5. *(best viewed in color)*

mini-batch sizes. To summarize our contribution: (1) the estimation of the attribution masks is possible with a smaller mini-batch size without losing quality but with a reduced computational effort; (2) our proposed solution results in more precise and coherent attribution masks, as Figure 1 shows; (3) most importantly, we demonstrate that the estimated attribution mask can be leveraged for improving the classification performance of a fine-grained classification model. We outperform other baseline methods and achieve classification results comparable to a setup if ground truth bounding boxes are used.

2 Related Work

Attribution methods aim at estimating a saliency (or attribution) map for an input image that defines the importance of each area in the image for the desired task, e.g., for classification. We give a brief overview of three possible attribution methods that are often used in literature.

End-to-end trainable *attention methods* [13,14,47] present different approaches that modify the architecture of a CNN model. In general, these modifications estimate saliency maps, or attentions, for intermediate feature representations. The estimated saliencies are then used to enhance or diminish the feature values. Even though the estimated attention maps improve the classification significantly, they are coarse since they typically operate on intermediate representations right before the final classification layer. Consequently, up-scaling these attentions to the dimension of the original image cannot capture precisely the fine-grained details.

In contrast, *gradient-based methods* [20,38,39,40] estimate pixel-wise importance by computing gradients of the outputs (logit of the target class) w.r.t. the input pixels. Even though the resulting saliency map is much finer than attention-based saliencies, it often highlights lots of irrelevant areas in the image, e.g., the background of the image. This may be caused by gradient saturation, discontinuity in the activations of the network [36], or an inductive bias due to the convolutional architecture, which is independent of the learned parameters [1].

Finally, *perturbation-based methods* [5,7,8] attribute the importance to a pixel by modifying the pixel's value and observing the change in the network's output. These methods identify image regions that are significantly relevant for a given classification target. Hereby, two different objectives can be used to estimate these regions: (1) the estimation of the smallest region that retains a certain classification score, and (2) the estimation of the smallest region that minimizes the target classification score when this region is changed. Dabkowski and Gal [7] presented a method that follows the mentioned objectives and Chang *et al.* [5] reformulated these objectives in their fill-in of the dropout (FIDO) algorithm. They further presented different infill methods and their effects on the estimated saliency maps.

In this work, we utilize the FIDO algorithm and present a way to enhance the computation stability of the gradients. Furthermore, we propose a way to combine the resulting attribution masks into a joint attribution mask, which we finally use to improve the results of a fine-tuned classification model. As with all perturbation-based methods, we keep the advantage that neither a change of the architecture nor a fine-tuning of the parameters is required.

Fine-grained categorization is a special classification discipline that aims at distinguishing visually very similar objects, e.g., bird species [44], car models [23], moth species [33], or elephant individuals [24]. These objects often differ only in subtle visual features and the major challenge is to build a classification model that identifies these features reliably. On the one hand, it is common to utilize the input image as it is and either perform a smart *pre-training strategy* [6,22] or *aggregate feature* using different techniques [26,37]. On the other hand, there are the *part- or attention-based approaches* [13,14,20,47] that either extract relevant regions, so-called parts, in the input image or enhance and diminish intermediate feature values with attention mechanisms. Finally, *transformer-based approaches* [11,46] are currently at the top in different fine-grained classification benchmarks. However, these methods rely on the parameter-rich transformer architecture and big datasets for fine-tuning. Typically, in the context of automated animal monitoring these resources are not available making such models difficult to deploy in the field.

In this work, we use two widely used CNN architectures [12,41] fine-tuned on the CUB-200-2011 [44] dataset. Without any further modification or fine-tuning, we use the estimated attribution masks to extract an auxiliary crop of the original input image. Finally, we use the extracted crop to enhance the classification decision.

3 Simplified Concrete Dropout - Improved Stability

Our final goal is a pixel-wise attribution of importance for a certain classification output. Perturbation-based methods offer a way to estimate this importance by observing the causal relation between a perturbation of the input and the caused change in the decision of a classification model. One example is the *fill-in of the dropout* (FIDO) algorithm [5] which computes these attribution masks by identifying pixel regions defined as following:

1. Smallest destroying region (SDR) represents an image region that *minimizes* the classification score if this region is changed.
2. Smallest sufficient region (SSR) represents an image region that *maximizes* the classification score if only this region is retained from the original content.

Based on these definitions, the FIDO algorithm optimizes the parameters of a binary dropout mask. The mask identifies whether a pixel value should be perturbed with an alternative representation (*infill*) or not, and can be interpreted as a saliency map. In the following, we formalize the objective functions, illustrate the limitation of the FIDO algorithm, and explain our suggested improvements.

3.1 The FIDO Algorithm and Its Limitations

Given an image $\mathbf{x}$ with N pixels, a class c, a differential classification model $\mathcal{M}$ producing an output distribution $p_{\mathcal{M}}(c|\mathbf{x})$, we are interested in a subset of pixels r that divides the image into two parts $\mathbf{x} = \mathbf{x}_r \cup \mathbf{x}_{\setminus r}$. Observing the classifier's output when $\mathbf{x}_r$ is not visible gives insights into the importance of the region r for the classification decision. Because of the binary division of the image into two parts, the region r can be modeled by a binary dropout mask $\mathbf{z} \in \{0,1\}^N$ with the same size[1] as $\mathbf{x}$ and an infill function ϕ that linearly combines the original image $\mathbf{x}$ and an infill $\hat{\mathbf{x}}$ with an element-wise multiplication $\odot$:

$$\phi(\mathbf{x}, \mathbf{z}) = (\mathbf{1} - \mathbf{z}) \odot \mathbf{x} + \mathbf{z} \odot \hat{\mathbf{x}} \quad . \tag{1}$$

There are different ways to generate an infill image $\hat{\mathbf{x}}$. First, content-independent approaches, like random (uniformly or normally distributed), or fixed (e.g., zeros) pixel values, generate the infill image independently of the input's content. Because these methods are independent of the content of the image, they cause a hard domain shift for the underlying classification model. Chang *et al.* [5] showed that these infill approaches perform worse compared to content-aware methods like GANs [10] or Gaussian blur. Popescu *et al.* [30] used knockoffs [2] to generate infill values and reported in their work superiority of knockoff infills on the MNIST dataset [25]. All of these content-aware methods generate the infill image depending on the pixels of the original image and retain the structure

[1] for sake of simplicity, we consider $\mathbf{x}$ and $\mathbf{z}$ as 1D vectors, instead of 2D matrices with dimensions H and W, and $N = H \cdot W$.

and composition of the original image to some degree. In our experiments, we use the Gaussian blur approach to create the infill image $\hat{\mathbf{x}}$. First, it removes the fine-grained details we aim to identify but retains the contents of the image so that the infill is not an out-of-domain input for the classification model. Second, the knockoff generation, proposed by Popescu *et al.*, is suitable for MNIST images because of the low dimensionality of the images (28×28 px) and the binary pixel values. Unfortunately, to this point, there is no way to apply this generation process to real-world RGB images. Finally, the GAN-based infills are computationally intensive and require an additional model that has to be trained on data related to the images we want to analyze.

The search space of all possible binary masks $\mathbf{z}$ grows exponentially with the number of pixels, hence we need an efficient way to estimate the values $z_n \in \mathbf{z}$. Assuming a Bernoulli distribution for the binary mask values allows us to sample the masks from a parametrized distribution $q_\theta(\mathbf{z})$ and optimize the parameters $\theta_n \in \theta$ using the SSR and SDR objectives:

$$L_{SSR}(\theta) = \mathbb{E}_{q_\theta(\mathbf{z})}\left[-s_{\mathcal{M}}(c|\phi(\mathbf{x},\mathbf{z})) + \lambda||\mathbf{1}-\mathbf{z}||_1\right] \quad \text{and} \tag{2}$$

$$L_{SDR}(\theta) = \mathbb{E}_{q_\theta(\mathbf{z})}\left[\ \ s_{\mathcal{M}}(c|\phi(\mathbf{x},\mathbf{z})) + \lambda||\mathbf{z}||_1\right] \tag{3}$$

with the L_1 regularization factor λ. Following the original work, we set $\lambda = 0.001$. The score $s_{\mathcal{M}}$ is defined as log-odds of classification probabilities:

$$s_{\mathcal{M}}(c|\mathbf{x}) = \log \frac{p_{\mathcal{M}}(c|\mathbf{x})}{1 - p_{\mathcal{M}}(c|\mathbf{x})} \quad . \tag{4}$$

To be able to optimize θ through a discrete random mask $\mathbf{z}$, the authors relax the discrete Bernoulli distribution and replace it with a continuous approximation: the concrete distribution [16,28]. The resulting sampling, called *concrete dropout* (CD), was proposed by Gal *et al.* [9] and is defined as

$$z_n = \sigma\left(\frac{1}{t}\left(\log\frac{\theta_n}{1-\theta_n} + \log\frac{\eta}{1-\eta}\right)\right) \quad \eta \sim \mathcal{U}(0,1) \tag{5}$$

with the temperature parameter t (we follow the original work and set this parameter to 0.1), σ being the sigmoid function, and η sampled from a uniform distribution.

In the original work, Chang *et al.* [5] proposed two methods to speed up convergence and avoid unnatural artifacts during the optimization process. First, they computed the gradients w.r.t. θ from a mini-batch of dropout masks. They mentioned in Appendix (A.6) that they observed unsatisfactory results with mini-batch sizes less than 4, which they attributed to the *high variance* in the gradient estimates. Second, they sampled a coarser dropout mask (e.g. 56×56) and upsampled the mask using bi-linear interpolation to the dimensions of the input (e.g. 224×224).

In the following, we reflect upon the cause of the high variance in the gradient estimates and propose a way to increase the computational stability. Consequently, our solution reduces the dependency of the FIDO method on the mini-batch size, allowing the estimation of the attribution masks with lower mini-batch sizes which ultimately reduces the computation time. Finally, since we are interested in fine-grained details in the image, the estimation of a coarser attribution mask followed by an upsampling operation would not lead to the desired level of detail. With our solution and the resulting improvement in the gradient computation, we can directly estimate a full-sized attribution mask θ without any unnatural artifacts, as shown in Figure 1.

3.2 Improving Computational Stability

The sampling of $\mathbf{z}$ using CD requires that all dropout parameters $\theta_n \in \theta$ are in the range $[0, 1]$. One way to achieve this is to initialize the attribution mask with real-valued parameters $\vartheta_n \in \mathbb{R}$ and apply the sigmoid function to those: $\theta_n = \sigma(\vartheta_n)$. As a consequence, the CD sampling procedure for $\mathbf{z}$, as described in Eq. 5, is a chaining of multiple exponential and logarithmic operations. This can be easily implemented in the current deep learning frameworks and is a common practice, e.g., in the reference implementation of Gal *et al.*[2]. However, we hypothesize that exactly this chaining of operations causes a high variance in the gradient estimates, and we validate this assumption in our experiments.

Under the assumption that θ is the output of the sigmoid function, we can simplify the sampling procedure of the attribution mask $\mathbf{z}$ and mitigate the before-mentioned problem. First, for readability reasons, we substitute the uniform noise part with a single variable $\hat{\eta} = \log \frac{\eta}{1-\eta}$ in Eq. 5. Then after using the transformation $\theta_n = \sigma(\vartheta_n)$ from above, expanding the argument of the sigmoid function, and simplifying the terms, the sampling of the binary mask $\mathbf{z}$ using CD transforms to a simple sigmoid function:

$$z_n = \sigma\left(\frac{1}{t}\left(\log \frac{\sigma\left(\vartheta_n\right)}{1-\sigma\left(\vartheta_n\right)} + \hat{\eta}\right)\right) \tag{6}$$

$$= \sigma\left(\frac{1}{t}\left(\log \frac{\left(1+\exp(-\vartheta_n)\right)^{-1}}{1-\left(1+\exp(-\vartheta_n)\right)^{-1}} + \hat{\eta}\right)\right) \tag{7}$$

$$= \sigma\left(\frac{1}{t}\left(\log \frac{1}{\exp(-\vartheta_n)} + \hat{\eta}\right)\right) = \sigma\left(\frac{\vartheta_n + \hat{\eta}}{t}\right) \quad . \tag{8}$$

The resulting formula is equivalent to the original formulation of CD. However, the reduction of exponential and logarithmic operations, and hence the reduction of the number of operations in the gradient computation, are the major benefits of the simplified version in Eq. 8. As a result, it reduces computational inaccuracies and enhances the propagation of the gradients. Consequently, the

[2] https://github.com/yaringal/ConcreteDropout

optimization of the parameters ϑ, and of the attribution map defined by θ, converges to more precise and better results as we show in Section 4.

In Sect. S2 of our supplementary material, we performed an empirical evaluation of this statement and showed that our proposed simplifications result in a lower variance of the gradient estimates. Additionally, you can find in Sect. S1 a Python implementation of the improved Concrete Dropout layer using the PyTorch [29] framework.

3.3 Combined Attribution Mask for Fine-grained Classification

So far, we only applied operations on the original formulation to simplify Eq. 5 from the original work of Chang *et al.* [5]. However, we can also show that the estimated attribution masks improve the performance of a classification model. First, we follow Chang *et al.* and only consider mask entries with an importance rate above 0.5. Then, we estimate a bounding box around the selected values of the attribution mask. In the end, we use this bounding box to crop a patch from the input image and use it as an additional input to the classification model (see Section 4.3). Instead of using the attribution masks separately, we are interested in regions that are important to sustain and should not be deleted. Hence, we propose to combine the attribution masks θ_{SSR} and θ_{SDR} using element-wise multiplication of mask values followed by a square root as a normalization function:

$$\theta_{joint} = \sqrt{\theta_{SSR} \odot (\mathbf{1} - \theta_{SDR})} \quad . \tag{9}$$

With element-wise multiplication, we ensure that if either of the attribution values is low, then the joint attribution is also low. The square root normalizes the joint values to the range where we can apply the same threshold (0.5): for example, if both attribution values are around 0.5, then also the joint attribution will be around 0.5 and not around 0.25.

4 Experiments

We performed all experiments on the CUB-200-2011 [44] dataset. It consists of 5994 training and 5794 test images for 200 different species of birds. It is the most used fine-grained dataset for benchmarking because of its balanced sample distribution. We selected this dataset mainly because it also contains ground truth bounding box annotations, which we used as one of our baselines in Section 4.3

Figure 1 shows a qualitative comparison of the estimated masks on one example image of the CUB-200-2011 dataset. Notably, our solution is more stable when we use smaller mini-batch sizes and produces fewer false positive attributions, e.g., in the background or highlighting of the tree branch.

We evaluated two widely used CNN architectures pre-trained on the ImageNet [34] dataset: ResNet50 [12] and InceptionV3 [41]. Additionally, we used

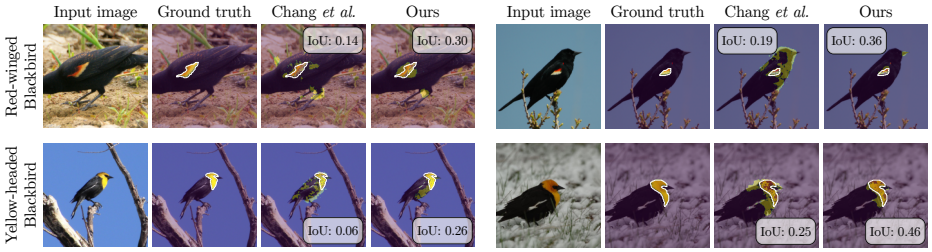


Fig. 2: Examples of the estimated masks for the *Red-winged Blackbird* and the *Yellow-headed Blackbird* using the original FIDO approach and our proposed improved method. Besides the ground-truth segmentation masks, we also reported the intersection over union (IoU) of the estimated mask with the ground-truth.

an alternative pre-training on the iNaturalist2017 dataset [43] for the InceptionV3 architecture proposed by Cui *et al.* [6] (denoted with IncV3* and InceptionV3* in Tables 1 and 2, respectively). All architectures are fine-tuned for 60 epochs on the CUB-200-2011 dataset using the AdamW [27] optimizer with the learning rate of 1×10^{-3} (and ϵ set to 0.1)[3].

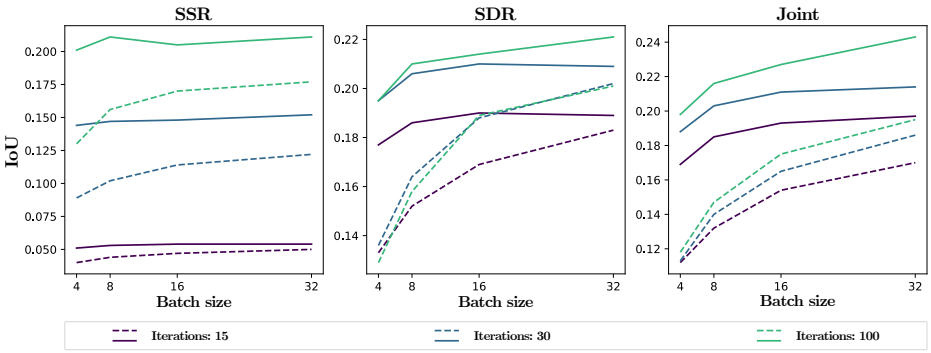
4.1 Evaluating Mask Precision

To quantify the visual observations in Figure 1, we selected two visually similar classes of Blackbirds from the CUB-200-2011 dataset: the Red-winged Blackbird and the Yellow-headed Blackbird. Both belong to the family of Icterids (New World blackbirds) and have black as a dominant plumage color. However, as the name of the species indicates, the main visual feature distinguishing these birds from other black-feathered birds is the **red wing** or the **yellow head**. Using this information, we created segmentation masks with the *Segment Anything* model [18] for the mentioned regions and used these as ground truth.

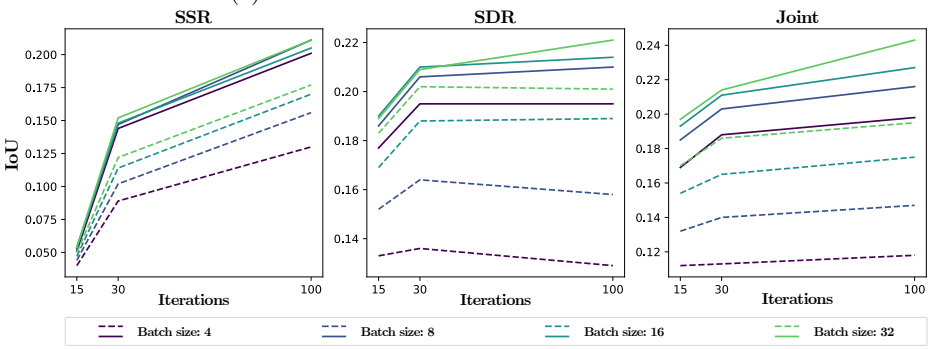
We fine-tuned a classification model (ResNet50 [12]) on the entire dataset and estimated the attribution masks using both methods: the original FIDO approach by Chang *et al.* [5] and our improved method. We evaluated different mini-batch sizes and a different number of optimization steps. Both of these parameters strongly affect the runtime of the algorithm. After an attribution mask was estimated, we selected only the values above the threshold of 0.5 and computed the IoU with the ground truth mask. We performed this evaluation for the masks estimated by the SSR and SDR objectives separately as well as using the joint mask as defined in Eq. 9.

In Figure 3, we report the IoU results of the mentioned setups. First, the plot shows that our solution (solid lines) outperforms the original approach (dashed lines) in every constellation of the hyperparameters. Next, the optimization process becomes less sensitive to the size of the mini-batches. This can be seen

[3] Changing the default parameter smooths the training, as suggested in https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Adam#notes_2



(a) IoU values for different mini-batch sizes.



(b) IoU values for a different number of optimization steps.

Fig. 3: Intersection over union (IoU) of the estimated masks with the ground-truth annotations of a discriminative region. We tested different values for the hyperparameters mini-batch size (a) and the number of optimization steps (b). Our proposed method (solid lines) outperforms the original work (dashed lines) and shows less sensitivity against the mini-batch size. *(best viewed in color)*

either by the slope (Figure 3a) or the variance (Figure 3b) of the IoU curves. Our method achieved the same quality of the attribution masks with smaller mini-batch sizes. Consequently, by reducing the mini-batch size, the number of sampled dropout masks at every optimization step is also reduced. Hence, by using a mini-batch size of 8 instead of 32, which Chang *et al.* use in their work, we could reduce the computation time per image from 40 to 11 seconds[4] for 100 optimization steps.

In Figure 2, we visualized four examples of the mentioned classes, our annotated segmentation masks, and the results of both approaches after 100 optimization steps and using a mini-batch size of 8.

[4] We processed the images using an Intel i9-10940X CPU, 128GB RAM, and a GeForce RTX 3090 GPU

Table 1: Comparison of the original solution proposed by Chang *et al.* and our improved implementation in terms of mask coherency. For different model architectures and different mask estimation objectives, we report the total variation as defined in Eq. 10 (*lower is better*).

	SSR			SDR			Joint		
	RN50	IncV3	IncV3*	RN50	IncV3	IncV3*	RN50	IncV3	IncV3*
Chang *et al.* [5]	39.74	32.44	27.79	44.21	45.53	37.81	37.61	33.92	28.66
Ours	17.54	18.18	17.37	22.72	21.87	20.20	16.95	15.21	14.06

4.2 Mask Coherency

Following Dabkowski *et al.* [7], Chang *et al.* propose to use total variation regularization with a weighting factor of 0.01, which is defined as

$$\mathrm{TV}(\mathbf{z}) = \sum_{i,j} \left(z_{i,j} - z_{i,j+1}\right)^2 + \sum_{i,j} \left(z_{i,j} - z_{i+1,j}\right)^2 \quad . \tag{10}$$

We observed that the high variance in the gradients affects the coherency of the masks (see Figure 1). Hence, we computed the total variance of the estimated masks and reported the results in Table 1. The total variation is computed for the attribution masks estimated for the entire CUB-200-2011 dataset with the original approach and our proposed solution. The results show that our solution produces more coherent masks, meaning the identified regions are more connected.

4.3 Test-time Augmentation of a Fine-grained Classifier

Given a model fine-tuned on the CUB-200-2011 dataset, we evaluated in this experiment how we can use the estimated attribution masks to improve the classification performance of the model. In addition to the prediction of the baseline models, we used different methods to extract one auxiliary crop from the original image and compute the prediction using this crop. Then, we averaged the predictions and report the resulting accuracies in Table 2. This way of classification improvement is widely used to different extent. He *et al.* [12] or Szegedy *et al.* [41], for example, use ten or 144 crops in their work, respectively. Hu *et al.* [14], as another example, perform attention cropping to enhance the prediction of the classifier. In our setup, we extracted a single crop using different methods, which we explain in the following.

Ground-truth bounding boxes: We utilized the bounding box annotations of the CUB-200-2011 dataset. First, we only used the crops identified by the bounding boxes. Second, we combined the predictions from the cropped image and the original image, by averaging the predictions.

Center and random crop: Following the motivation behind the crops used by He *et al.* [12] and Szegedy *et al.* [41] that the object of interest is likely to be in

Table 2: Comparison of the classification performance using different test-time augmentation (TTA) methods on the CUB-200-2011 dataset. Besides the baselines (no TTA or ground truth bounding boxes), we also evaluated heuristic methods (random or center crop), content-aware methods (GradCam or a bird detector), and two different FIDO implementations (the original work of Chang *et al.* and our proposed improvement). We report the accuracy (in %).

	ResNet50	InceptionV3	InceptionV3*
Baseline (BL)	82.78	79.86	90.32
GT bounding boxes only	84.38	81.31	90.18
BL + GT bounding boxes	84.55	81.65	90.70
BL + random crop	83.41	80.45	89.99
BL + center crop	83.83	81.07	90.16
BL + gradient [20]	83.74	80.76	90.02
BL + BirdYolo [42]	83.81	81.12	90.39
BL + FIDO [5]	84.17	81.67	90.47
BL + FIDO (ours)	**84.67**	**81.77**	**90.51**

the center, we cropped the center of the image. Furthermore, we also extracted a random crop. For both methods, we set the size of the crop to be 75 % of the width and height of the original image. These methods are content-agnostic and use only heuristics to estimate the region to crop.

Gradient crop: As a first content-aware method, we computed the gradients w.r.t. the input image [38]. We utilized the pre-processing and thresholding of the gradient as presented by Korsch *et al.* [20], estimated a bounding box around the resulting saliency map, and cropped the original image based on the estimated bounding box.

BirdYOLO is a YOLOv3 [31] detection model pre-trained on a bird detection dataset [42]. For each image, we used the bounding box with the highest confidence score, extended it to a square, and cropped the original image accordingly.

FIDO: Finally, we utilized the joint mask computed from the SSR and SDR masks of the FIDO algorithm as defined in Eq. 9. On the one hand, we used the masks estimated by the original work of Chang *et al.* [5], and on the other hand, the masks estimated with our proposed improvements.

The results in Table 2 show that compared to the baseline model the ground truth bounding boxes yield a higher classification accuracy, even if solely using the bounding box crops for classification. Next, we can see that even such content-agnostic methods like center or random cropping can boost classification performance. Similar improvements can be achieved by content-aware methods like gradients or a detection model. We observed the most improvement with the FIDO algorithm, and finally with our proposed solution we achieved the best results that are comparable to using ground-truth bounding boxes.

5 Conclusions

In this paper, we proposed a simplified version of the concrete dropout (CD). The CD is used in the fill-in of the dropout (FIDO) algorithm to sample a set of attribution masks based on an underlying parametrized distribution. Using these masks, one can estimate how relevant a specific image pixel was for the classification decision. The parameters of the distribution are optimized based on the classification score but the optimization process suffers from a high variance in the gradient computation if the original formulation of CD is used. Our solution simplifies the sampling computations and results in more stable gradient estimations. Our approach maintains the quality of the estimated masks while reducing computational effort due to smaller mini-batch sizes during the optimization process. Furthermore, the resulting attribution masks contain fewer falsely attributed regions. We also presented a way of using the estimated fine-grained attribution masks to enhance the classification decision. Compared with other classification baselines, our solution produces the best result and even performs comparably to a setup where ground truth bounding boxes are used.

As an extension, our proposed single-crop TTA can be extended with a part-based approach to further boost the classification performance. Alternatively, a repeated iterative estimation of the masks may be worth an investigation.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. Advances in neural information processing systems **31** (2018)
2. Barber, R.F., Candès, E.J.: Controlling the false discovery rate via knockoffs (2015)
3. Bjerge, K., Nielsen, J.B., Sepstrup, M.V., Helsing-Nielsen, F., Høye, T.T.: An automated light trap to monitor moths (lepidoptera) using computer vision-based tracking and deep learning. Sensors **21**(2), 343 (2021)
4. Brust, C.A., Burghardt, T., Groenenberg, M., Käding, C., Kühl, H., Manguette, M., Denzler, J.: Towards automated visual monitoring of individual gorillas in the wild. In: ICCV Workshop on Visual Wildlife Monitoring (ICCV-WS). pp. 2820–2830 (2017). https://doi.org/10.1109/ICCVW.2017.333
5. Chang, C.H., Creager, E., Goldenberg, A., Duvenaud, D.: Explaining image classifiers by counterfactual generation. In: International Conference on Learning Representations (2018)
6. Cui, Y., Song, Y., Sun, C., Howard, A., Belongie, S.: Large scale fine-grained categorization and domain-specific transfer learning. In: Proceedings of CVPR (6 2018). https://doi.org/10.1109/cvpr.2018.00432
7. Dabkowski, P., Gal, Y.: Real time image saliency for black box classifiers. Advances in neural information processing systems **30** (2017)
8. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the IEEE international conference on computer vision. pp. 3429–3437 (2017)
9. Gal, Y., Hron, J., Kendall, A.: Concrete dropout. Advances in neural information processing systems **30** (2017)

10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
11. He, J., Chen, J.N., Liu, S., Kortylewski, A., Yang, C., Bai, Y., Wang, C.: Transfg: A transformer architecture for fine-grained recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 852–860 (2022)
12. 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)
13. He, X., Peng, Y., Zhao, J.: Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization. IJCV pp. 1–21 (2019)
14. Hu, T., Qi, H., Huang, Q., Lu, Y.: See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. arXiv preprint arXiv:1901.09891 (2019)
15. Hughes, B., Burghardt, T.: Automated visual fin identification of individual great white sharks. International Journal of Computer Vision **122**(3), 542–557 (2017)
16. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017), https://openreview.net/forum?id=rkE3y85ee
17. Käding, C., Rodner, E., Freytag, A., Mothes, O., Barz, B., Denzler, J.: Active learning for regression tasks with expected model output changes. In: British Machine Vision Conference (BMVC) (2018)
18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
19. Korsch, D., Bodesheim, P., Brehm, G., Denzler, J.: Automated visual monitoring of nocturnal insects with light-based camera traps. In: CVPR Workshop on Fine-grained Visual Classification (CVPR-WS) (2022)
20. Korsch, D., Bodesheim, P., Denzler, J.: Classification-specific parts for improving fine-grained visual categorization. In: Proceedings of the German Conference on Pattern Recognition. pp. 62–75 (2019)
21. Korsch, D., Bodesheim, P., Denzler, J.: End-to-end learning of fisher vector encodings for part features in fine-grained recognition. In: German Conference on Pattern Recognition (DAGM-GCPR). pp. 142–158 (2021). https://doi.org/10.1007/978-3-030-92659-5_9
22. 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: ECCV. pp. 301–320. Springer (2016)
23. 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
24. Körschens, M., Denzler, J.: Elpephants: A fine-grained dataset for elephant re-identification. In: ICCV Workshop on Computer Vision for Wildlife Conservation (ICCV-WS) (2019)
25. LeCun, Y., Cortes, C., Burges, C., et al.: Mnist handwritten digit database (2010)
26. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: Proceedings of ICCV. pp. 1449–1457 (2015). https://doi.org/10.1109/iccv.2015.170
27. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)

28. Maddison, C., Mnih, A., Teh, Y.: The concrete distribution: A continuous relaxation of discrete random variables. In: Proceedings of the international conference on learning Representations. International Conference on Learning Representations (2017)
29. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
30. Popescu, O.I., Shadaydeh, M., Denzler, J.: Counterfactual generation with knockoffs. arXiv preprint arXiv:2102.00951 (2021)
31. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
32. Reimers, C., Penzel, N., Bodesheim, P., Runge, J., Denzler, J.: Conditional dependence tests reveal the usage of abcd rule features and bias variables in automatic skin lesion classification. In: CVPR ISIC Skin Image Analysis Workshop (CVPR-WS). pp. 1810–1819 (2021)
33. Rodner, E., Simon, M., Brehm, G., Pietsch, S., Wägele, J.W., Denzler, J.: Fine-grained recognition datasets for biodiversity analysis. In: CVPR Workshop on Fine-grained Visual Classification (CVPR-WS) (2015)
34. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
35. Sakib, F., Burghardt, T.: Visual recognition of great ape behaviours in the wild. In: International Conference on Pattern Recognition (ICPR) Workshop on Visual Observation and Analysis of Vertebrate And Insect Behavior (2021)
36. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International conference on machine learning. pp. 3145–3153. PMLR (2017)
37. 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
38. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. In: Proceedings of the International Conference on Learning Representations (ICLR). ICLR (2014)
39. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: ICLR (workshop track) (2015)
40. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
41. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2016)
42. Tran, B.: Bird detection by yolo-v3. https://github.com/xmba15/yolov3_pytorch (2023), [Online; accessed 30-May-2023]
43. 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)
44. 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)

45. Yang, X., Mirmehdi, M., Burghardt, T.: Great ape detection in challenging jungle camera trap footage via attention-based spatial and temporal feature blending. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
46. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10819–10829 (2022)
47. Zhang, L., Huang, S., Liu, W., Tao, D.: Learning a mixture of granularity-specific experts for fine-grained categorization. In: Proceedings of ICCV. pp. 8331–8340 (2019)