Kandinsky Conformal Prediction: Efficient Calibration of Image Segmentation Algorithms

Joren Brunekreef*,[1, 2], Eric Marcus*,[1, 2], Ray Sheombarsing[1], Jan-Jakob Sonke[1], Jonas Teuwen[1]

[1]Netherlands Cancer Institute, [2]University of Amsterdam

{j.brunekreef, e.marcus, r.sheombarsing, j.sonke, j.teuwen}@nki.nl

Abstract

Image segmentation algorithms can be understood as a collection of pixel classifiers, for which the outcomes of nearby pixels are correlated. Classifier models can be calibrated using Inductive Conformal Prediction, but this requires holding back a sufficiently large calibration dataset for computing the distribution of non-conformity scores of the model's predictions. If one only requires only marginal calibration on the image level, this calibration set consists of all individual pixels in the images available for calibration. However, if the goal is to attain proper calibration for each individual pixel classifier, the calibration set consists of individual images. In a scenario where data are scarce (such as the medical domain), it may not always be possible to set aside sufficiently many images for this pixel-level calibration. The method we propose, dubbed "Kandinsky calibration", makes use of the spatial structure present in the distribution of natural images to simultaneously calibrate the classifiers of "similar" pixels. This can be seen as an intermediate approach between marginal (imagewise) and conditional (pixelwise) calibration, where non-conformity scores are aggregated over similar image regions, thereby making more efficient use of the images available for calibration. We run experiments on segmentation algorithms trained and calibrated on subsets of the public MS-COCO and Medical Decathlon datasets, demonstrating that Kandinsky calibration method can significantly improve the coverage. When compared to both pixelwise and imagewise calibration on little data, the Kandinsky method achieves much lower coverage errors, indicating the data efficiency of the Kandinsky calibration.

1. Introduction

Calibration of predictive models is a critical aspect of machine learning, particularly in applications with significant impact based on model outcomes, such as medical diagnostics. Calibration ensures that predicted probabilities match the actual empirical likelihood of the predicted events. A well-calibrated model will output probabilities that correspond closely to real-world frequencies; for instance, if a model predicts an event with a probability p, this event should, in reality, occur with frequency p.

In this work, we will focus on image segmentation tasks. Here, the calibration procedure also plays a vital role. A segmentation model can be interpreted as a collection of classifiers, one for each output pixel. The calibration of each classifier thus affects decision-making at the pixel level, which subsequently influences global measures of segmentation accuracy (e.g. the Dice score). There are two straightforward notions of calibration in the context of image segmentation: *marginal* calibration, which measures the calibration averaged over all pixels, or calibration *conditional* on a specific pixel location.

One standard method for calibrating prediction models is Conformal Prediction (CP), a framework that has received increasing amounts of attention in recent years [2–5, 19, 26]. Conformal prediction provides statistically valid measures of confidence in a model-agnostic manner. Whereas the original (transductive) CP method [27] is computationally demanding, the more recently developed *inductive* CP [23–25, 30] is better suited for a present-day machine-learning setting. This lower demand for computational resources comes at the cost of requiring extra (labeled) data to be set aside as a calibration set. One then defines a notion of a *non-conformity score*, which measures the "strangeness" of each sample and the associated model prediction in this calibration set. The distribution of non-conformity scores in the calibration set can then be used as a benchmark to compare with newly unseen samples, which allows one to obtain a statistically valid notion of confidence for the model's predictions.

Now consider the case where one has set aside a calibration set of N images of dimensions $m \times n$. If we require only marginal calibration (i.e., aggregated over the whole

[*] These authors contributed equally to this work.

image), we can view each pixel in each image as a separate calibration data point so that our calibration set is of size $N \times m \times n$. However, since a segmentation model contains a "separate" classifier for each individual pixel, calibrating such a model on the pixel level requires one to view each image in the calibration set as a single sample. The calibration set is, therefore, only of size N in this case. Clearly, this can pose a challenge for settings where data availability is limited — for example, in the medical domain.

To address these challenges, we introduce "Kandinsky calibration", a technique that capitalizes on prior knowledge of the spatial correlations within images to calibrate classifiers across similar pixels more efficiently. This approach balances the need for detailed calibration at the pixel level with the practical limitations of data availability. It applies conformal prediction in a novel way, achieving fine-grained calibration with fewer calibration images. The following sections will detail the Kandinsky calibration approach and present experiments demonstrating its effectiveness in improving the calibration when few calibration images are available, a valuable attribute for segmentation applications.

Related Work Calibration for machine learning, and in particular deep learning, has received much attention of late [10, 11, 16, 18, 22]. For example, the "formalization" of the calibration has been discussed in [8], providing the first steps to a more formal understanding of the procedure and of calibrated functions. Conformal prediction [1, 20, 24, 31], the framework we utilize in this work, is also growing increasingly popular. Although work on risk control for segmentation purposes has been studied [3], thorough investigations of calibration methods for segmentation are scarce [32]. Class clustering based on the similarity of conformal scores was investigated in a classification context in [12].

Overview The contributions of this work are ordered as follows. In Sec. 2, we provide a short overview of calibration, conformal prediction, and coverage. In the following section, Sec. 3, we discuss our novel Kandinsky calibration framework. In Sec. 4, we show the experimental results of the different calibration methods.

2. Calibration and Conformal Prediction

Due to the significant role of conformal prediction, calibration, and coverage, we provide a short introduction containing relevant information for this work.

2.1. Conformal Prediction

We provide a short introduction of conformal prediction [20, 24, 31], following the conventions of the excellent introduction in [1].

Let us consider a classification task first. Suppose we are given a training set of images and labels of K classes. Furthermore, we train a predictive model f on this data, such that its outputs $f(x) \in [0,1]^K$. Using the inductive conformal prediction framework, we then apply the model f to so-called calibration data, consisting of n i.i.d. unseen samples $I = (X_1, Y_1), \ldots, (X_n, Y_n)$.

Utilizing this calibration data and f, we consider a new (unseen) datapoint $(X_{\text{test}}, Y_{\text{test}})$, where we do not know Y_{test}. The objective is to create a prediction set $\mathcal{C}_\alpha(X_{\text{test}}) \subset \{1, 2, \ldots, K\}$ with the following property

$$P(Y_{\text{test}} \in \mathcal{C}_\alpha(X_{\text{test}})) \geq 1 - \alpha\,, \tag{1}$$

where α is a user-chosen error rate. We can create these prediction sets by defining a so-called non-conformity score $s(x)$ that measures how far off the model's prediction $f(x)$ on an input x is from the ground truth. Then we define $\hat{q}_\alpha$ as the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical quantile of the $s_i \equiv s(X_i)$ in the calibration set. For the new test point X_{test} (where the label is unknown), we create the prediction set

$$\mathcal{C}_\alpha(X_{\text{test}}) = \{y \,|\, s_i \leq \hat{q}_\alpha\}\,. \tag{2}$$

It can then be shown that $\mathcal{C}_\alpha(X_{\text{test}})$ satisfies (1).

The choice of the scoring function $s(x)$ determines the usefulness of the prediction sets, and what choice to make here depends on the task at hand. For the remainder of this work, we set

$$s(x) = 1 - f(x)_Y\,, \tag{3}$$

where the subscript Y indicates that we take the model's output for the ground truth class Y. We leave it for future work to investigate whether other scoring functions could lead to better performance in the context of segmentation.

Segmentation In the case of segmentation, the model can be seen as a collection of classifiers, one for each pixel. Since these are separate classifier models, they should in principle be calibrated independently if we want to create valid prediction sets for each individual pixel. The upshot of this is that every labeled image (X_i, Y_i) in a calibration dataset I should be taken as just a single calibration sample for which we compute the non-conformity score s_i. We call this *pixelwise* (or *conditional*) calibration.

If, however, we set ourselves the more modest goal of *marginal* calibration where we only need the prediction sets to be valid on average over the whole image, we can view each pair $(X_i, Y_{i,(x,y)})$ of input images with the ground truth value for the pixel with coordinate (x, y) as a separate calibration point. However, marginal calibration can be attained even if a number of individual pixels are severely miscalibrated, so this approach is suboptimal if we want to have proper calibration guarantees in specific regions of the image.

If, therefore, one has access to a sufficiently large set I of calibration images, it is preferable to apply the pixel-wise calibration method. However, since labeling data for segmentation is a time-consuming task, such data are often scarce.

2.2. Calibration and Coverage Errors

A common calibration measure for prediction models is the Expected Calibration Error (ECE). For a binary classification task, it can be defined as

$$\mathrm{ECE}_{\mathcal{D}}(f) = \mathbb{E}_{\mathcal{D}}\left[\left|\mathbb{E}_{\mathcal{D}}[y|f(x)] - f(x)\right|\right]. \tag{4}$$

Here $\mathcal{D}$ is the data distribution on inputs x and ground truth labels $y \in \{0, 1\}$. This expression, however, is not well-defined for finite samples since it conditions on null events. The ECE can be approximated in several ways [8]. One particularly convenient and efficient method is to collect the model output scores $f(x_i)$ (evaluated on a labeled dataset of N samples x_i) in bins B_m, and compute the so-called binned ECE:

$$\mathrm{bECE}_M = \frac{1}{N}\sum_{m=1}^{M} |B_m| \left(\mathrm{conf}\left(B_m\right) - \mathrm{acc}\left(B_m\right)\right), \tag{5}$$

where $\mathrm{conf}\left(B_m\right)$ is the mean output score in the bin B_m and $\mathrm{acc}\left(B_m\right)$ is the true fraction of positive samples for which the score $f(x_i)$ is assigned to this bin. Throughout this work, we refer to the binned ECE simply as ECE_M (with a subscript indicating the number of bins) to avoid clutter. We present an example of a so-called *reliability diagram* in 1, where we plot the accuracy as a function of the confidence of the prediction model used in one of our experiments.

The ECE for a perfectly calibrated model equals zero: the model's output scores are precisely equal to the corresponding observed accuracies in the long run. Note that this does not imply the model has good classification performance: a prediction model that outputs a score of $1/2$ for each input is perfectly calibrated if exactly half the samples belong to the positive class.

When calibrating a model using conformal prediction, we need another measure of calibration performance. The aim of the prediction sets $\mathcal{C}_\alpha$ is to satisfy *coverage* conditions, meaning that the correct (ground truth) class Y_{test} of an unseen input X_{test} should be in the prediction set $\mathcal{C}_\alpha\left(X_{\text{test}}\right)$ with probability at least $1-\alpha$. We therefore define a (binned) measure of the *Coverage Error* (CE) as follows:

$$\mathrm{CE}_M = \sum_{m=1}^{M}\left(\mathrm{cov}\left(\mathcal{C}_{1-\frac{m}{M}}\right) - \frac{m}{M}\right). \tag{6}$$

The coverage $\mathrm{cov}\left(\mathcal{C}_\alpha\right)$ is defined as the frequency with which the α level prediction set $\mathcal{C}_\alpha$ contains the ground truth class. We point out to the reader to carefully distinguish between the abbreviations 'ECE' for Expected Calibration Error and 'CE' for Coverage Error.

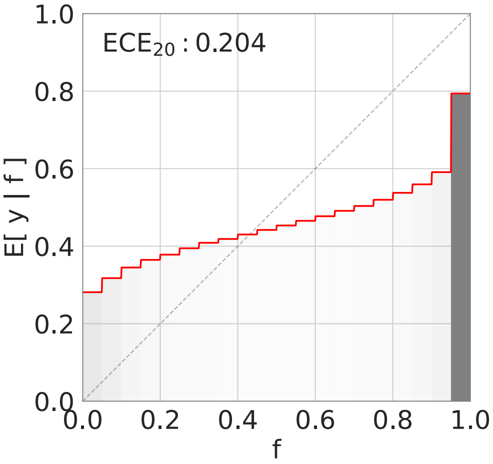

Figure 1. Reliability diagram [7] of a segmentation model trained on a subset of MS-COCO. The model's prediction scores and associated accuracies are aggregated over all pixels and assigned to 20 bins of equal width. The ECE is computed by averaging the absolute difference between the height of the bins and the diagonal. This particular model is overconfident for output scores $f \gtrsim 0.4$ and underconfident for lower output scores.

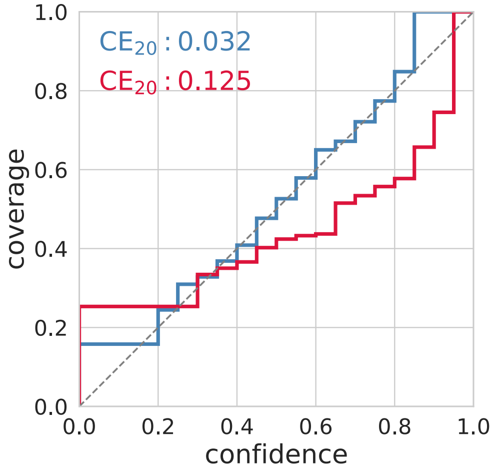

Figure 2. Coverage diagram for two individual pixel locations of a segmentation model trained on a subset of MS-COCO, calibrated pixelwise on 20.000 images (blue) and on 100 images (red).

For all our experiments in this work, we report the CE_{20}, that is, the coverage error obtained with 20 bins. A *coverage diagram* is the counterpart of a reliability diagram when evaluating the coverage properties of our prediction sets. In Fig. 2 we present the "coverage curves" of two of the (partially) calibrated pixel classifiers used in one of our experiments.

3. Kandinsky Calibration

Now that we have discussed the two ‘naive’ approaches of marginal and conditional calibration, we introduce our Kandinsky method. The primary rationale is as follows: *the pixel-level classifiers of a segmentation network are not independent and are related by well-organized patterns encoded in the task and data*. In other words, we are able to use prior knowledge in the calibration process of the segmentation networks.

In order to do this, we first define the notion of a non-conformity *curve*. Once we have computed non-conformity scores for a classifier model on a calibration set, we use its $\hat{q}$-th quantiles to create prediction sets for unseen data. This $\hat{q}$-th quantile of the non-conformity scores is what we refer to as the non-conformity curve $z(\hat{q})$, where $0 \leq \hat{q} \leq 1$. Note that non-conformity curves are defined for all $\hat{q}$ in this range, even if it based on only a small number of individual non-conformity scores.

When performing pixelwise calibration, we compute a separate non-conformity curve for each pixel coordinate in the image. The other extreme, imagewise calibration, computes a single non-conformity curve (based on the non-conformity scores aggregated over the whole image) that is then used for all pixel locations. Our Kandinsky method is an intermediate approach, where we cluster nearby pixels and compute a non-conformity curve for each cluster by aggregating their non-conformity scores.

This clustering is performed by finding pixels with similar non-conformity curves. This approach may seem circular, since we first require the (potentially noisy) non-conformity curves themselves to subsequently improve these curves over the whole cluster. However, the prior knowledge used in forming our clusters is that spatially nearby pixels are likely to have similar non-conformity curves: therefore, even if individual pixel locations in a certain spatial region have a dissimilar non-conformity curve due to a lack of data, they will still be grouped together with other pixels in their neighborhood.

The general Kandinsky method for calibrating a prediction model f can then be outlined as follows:

- Perform pixelwise calibration, computing separate non-conformity curves for each pixel location. These non-conformity curves are likely to be noisy in a low-data scenario.
- Cluster pixel locations based on the similarity of their non-conformity curves, potentially with a prior choice of possible region shapes.
- For each cluster, aggregate all the non-conformity scores encountered in the cluster, and compute a cluster-specific non-conformity curve.
- When forming prediction sets for a given pixel, use the newly obtained non-conformity curve of the cluster to which this pixel belongs.

3.1. Computing Kandinsky Clusters

Having motivated the creation of clusters of non-conformity curves, we will now provide several example methods of computing them.

3.1.1 K-Means Clustering

As a first approach, we can find Kandinsky clusters using a k-means clustering approach. In particular, we start with a pixel-level calibrated model and subsequently consider the non-conformity curves per pixel. At this stage, we choose a set of k quantiles at which we wish to compare non-conformity curves between the different pixels. The k-means clustering approach then tries to group all $m \times n$ (image size) points in this k-dimensional space. Provided we can obtain ‘good’ enough calibration measurements per pixel, this approach will yield precise information about which pixels are related.

However, in the low-data regime, this approach will be prone to noise in the calibration of the individual pixels. The resulting clusters will most likely be only based on spurious relations and will not help obtain better calibration. To optimize groups even in this low-data regime, we propose two more methods.

3.1.2 Genetic Algorithms

Genetic algorithms [14, 17] are a class of optimization techniques that “simulate” the process of natural evolution. These algorithms excel in navigating complex search spaces to identify solutions that might otherwise be inaccessible through traditional optimization methods. The specific type of algorithm we will employ is the so-called differential evolution [29]. In the rest of the work, we refer to the approach we outline here as GenAnn (for Genetic Annuli).

To be precise, for GenAnn, we will need to define a fitness function F that evaluates how ‘good’ a candidate solution is, a crossover function C which takes a collection of candidate solutions and combines them into a new one, a mutating function M which takes a candidate solution and randomly transforms it, and a replacement function R that determines if the candidate solutions are replaced with the mutated ones. Furthermore, we must provide a parametrization x_i of a candidate solution in terms of a finite set of real parameters. In our experiments, this parametrization will consist of the center coordinates and a set of radii for annuli. We show the specific implementations of the above functions and general procedure used in differential evolution for the reader’s convenience in the supplementary material, Algorithm 1.

3.1.3 Fourier Concentric Clustering

Here, we introduce the *Fourier Concentric Clustering* (FCC) method. In this section, we shall provide only a short summary of the method; a detailed and formal explanation can be found in the supplementary material, Sec. 8. Formally, we wish to construct a concentric decomposition of $\mathcal{R}_n := [0, n_1 - 1] \times [0, n_2 - 1]$ on which the variances of a user-prescribed quantity $\mathcal{J} : \mathcal{R}_n \to \mathbb{R}^p$ are minimized. This approach systematically decomposes an image into a series of concentric, nested subsets $V_0 \subset \ldots \subset V_{m-1} \subset \mathcal{R}_n$, centralizing around the mean of $\mathcal{J}$ with minimized variance

$$\sum_{l=0}^{m} \mathbb{E}_{A_l} \left(\|\mathcal{J} - \boldsymbol{\mu}_l\|^2 \right) \ , \tag{7}$$

where $\boldsymbol{\mu}_l$ denotes the mean of $\mathcal{J}$ on A_l, and each A_l will be analogous to the annuli discussed in the genetic algorithms. In other words, these A_l will be one of the nested subsets V_n, excluding the previous V_{n-1}.

To facilitate the decomposition, we redefine the image domain $\mathcal{R}_n$ in a new coordinate system, with the image's midpoint serving as the origin. The boundary of each subset A_l is then represented by polar curves, which are expressed via a finite Fourier series to allow for computational tractability. For example, the boundary of the initial subset A_0 is parameterized as

$$A_0 = \{(r, \theta) \mid 0 \leq r \leq r_0(\theta), \theta \in [0, 2\pi]\} \ , \tag{8}$$

with r_0 being a smooth, periodic function.

The numerical integration necessary for evaluating the variances is achieved using Legendre and Fourier quadrature methods, providing a means to compute the integrals over the domains A_l as functions of the Fourier coefficients. This is exemplified by the equation $\int_{A_0} f(x)\, dx = 2\pi c_{00}$, where c_{00} is the zeroth Fourier coefficient of a line integral, approximated using the FFT.

In setting up the optimization problem to minimize variance, we initialize the Fourier coefficients with a perturbation ε to represent a 'noisy' set of concentric circles. The BFGS algorithm [9, 13, 15, 28] is then employed to find the optimal coefficients that minimize the variance term coupled with a regularization term Λ to enforce the nested nature of the subsets without overlapping. The FCC method allows for a faster optimization of circular and elliptical cases than the genetic approach. However, the genetic approach provides a much more flexible optimization process as opposed to the Fourier approach, where special care has to be taken not to ignore higher-order modes.

3.2. Limitations

The most prominent limitation lies in creating prior knowledge of the calibration. There is no universal way to make the Kandinsky clusters, as it depends deeply on the characteristics of the data and task; it will require human input. The concentric circles and ellipses will not work for every imaginable task. Secondly, the numerical methods we introduced to compute the Kandinsky clusters may not be able to handle all possible geometric priors. The genetic algorithm should, in principle, be able to optimize the result, provided a suitable parametrization of the objective can be found; this is, however, not always a trivial task.

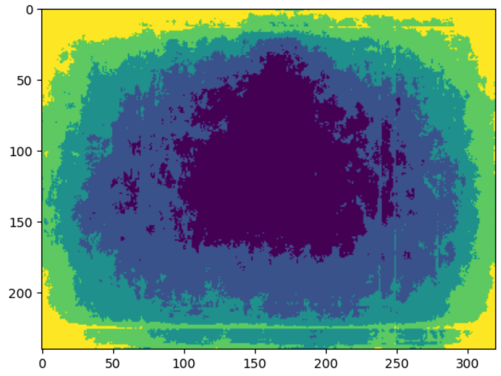

Figure 3. Results of the k-means clusters of non-conformity curves. A U-Net model was calibrated on 20.000 calibration images. During this calibration procedure, we obtained non-conformity curves for each pixel output of the model for the 'person' class. We obtained the non-conformity score for each curve for the 60, 70, 80, and 90th quantiles. These four-dimensional points were subsequently clustered using k-means. The appearance of concentric shapes is visible in the image. These shapes arise due to the different prevalences of persons appearing in different locations of the image. Furthermore, persons near the image's border are likely more difficult to identify because they might, for example, be found in the distance more often. The combination of the object (persons in this case) and the data characteristics determine the precise geometry of the clusters.

4. Experiments

This section will investigate the utility of using Kandinsky clusters in the calibration procedure. We will compare the marginal, conditional, and Kandinsky methods for various trained models on the unseen test sets. In particular, we wish to investigate the coverage errors attained by the different models and methods, as that will provide us with a robust metric. As we have described, the Kandinsky methods will be most helpful in scenarios with little calibration data. To investigate this, we have set up four experiments with different data availabilities:

- **MS-COCO-XL**: the highly idealized setting where we utilize 20.000 images for calibration.
- **MS-COCO-S**: same dataset, but we use only 100 calibration images.

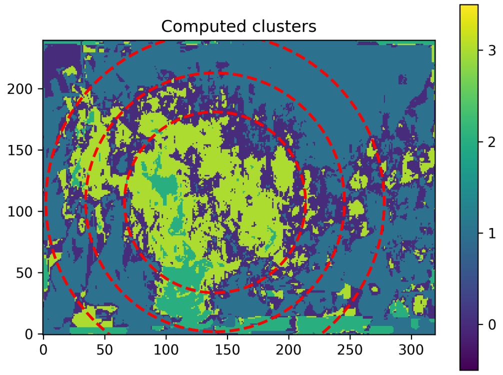

Figure 4. This figure shows the most ‘extreme’ scenario, where we have only used 100 images from the MS-COCO dataset during the calibration. The k-means clusters are shown in the background, clearly unable to find the same clusters as in Fig. 3. The red rings show the radii of the annuli found to minimize the objective using the genetic algorithm approach.

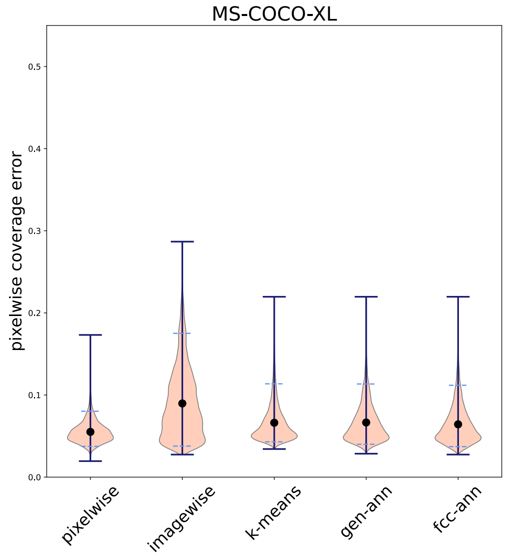

Figure 5. Violin plot of the pixelwise coverage errors for the MS-COCO-XL experiment. This experiment investigated the idealized scenario with access to large calibration sets. Here, we have utilized 20.000 images to calibrate the segmentation model. Due to the size of this calibration set, we can effectively use the pixelwise calibration, which attains the lowest mean coverage errors. The Kandinsky methods follow closely, and only the imagewise calibration performs visibly worse. The reason for this is that imagewise calibration will ‘average’ over all the pixels, providing for each individual pixel a skewed estimate of its calibration.

- **Decathlon-L**: a, for medical standards, large calibration dataset consisting of 77 patients.
- **Decathlon-S**: same dataset, but utilizing only 27 patients for calibration.

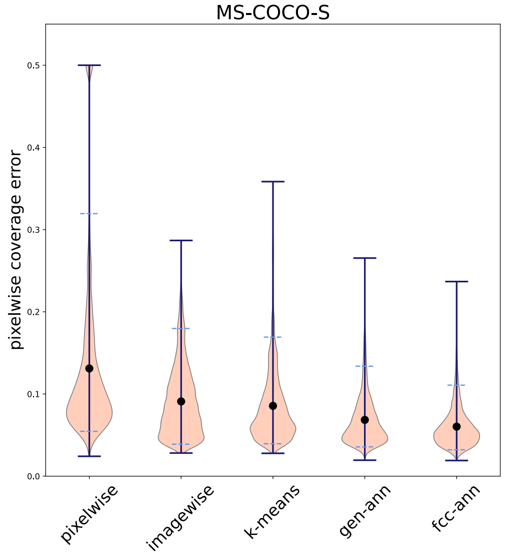

Figure 6. Results of all calibration methods using only 100 calibration images. All Kandinsky methods (the three rightmost) outperform the pixelwise and imagewise calibrations. The GenAnn and FCC annuli outperform the K-means approach due to the increased noise in the clusters found by K-means.

The different objectives and calibration dataset sizes allow us to investigate the utility of Kandinsky calibration across varying scales and domains, providing insights into its effectiveness in more abundant and scarce data scenarios and its adaptability to diverse image contexts. In Tab. 1, we summarize the results of all the experiments, showing the mean coverage errors attained by all methods.

Model For all experiments in this section, we utilize a U-Net for the segmentation. In particular, we have a U-Net with four up and downsampling layers, resulting in 53.5M parameters. Since our goal lies with the calibration of the models and not finding the maximal possible performance of the models, there is no need for extensive hyperparameter searches. In the case of MS-COCO, we utilized 678 datapoints for training and 2869 unseen images to evaluate the coverage. For Medical Decathlon, the model was trained on 86 patients and evaluated on 118 patients.

4.1. MS-COCO-XL

The goal of this first experiment is to investigate a highly idealized scenario where there is no shortage of calibration data whatsoever. In particular, we utilize the publicly available MS-COCO dataset [21], consisting of images of size 320 × 240 depicting various classes to be segmented. We use 20.000 images to calibrate the segmentation model. In our experiments on this dataset, we choose, without loss of generality, the *person* class for our investigations. As is the case for many photographed objects, we expect that ‘on average’ the object of interest will be reasonably centered in the image. We can utilize this knowledge in creating our Kandinsky clusters. In particular, for the genetic and FCC approaches, we will choose them to be concentric sets of circles or ellipses. More precisely, the first cluster, starting from a particular center point, will be a disk, and each subsequent cluster an annulus (or elliptic version thereof). It is also in the formation of such clusters that they have found their name.

In this extreme high-data calibration regime, the k-means method provides an elucidating insight into the presence of these clusters. In Fig. 3, we show the Kandinsky clusters found by the k-means method (art aficionados will recognize the motivation for our method’s chosen name in this figure). The presence of these clusters arises from a combination of different prevalences of the persons on the pixels and the difficulty of identifying them. In particular, we expected and found the center of the image to contain most instances of the person class, subsequently decreasing in concentric shapes. If a person is near the boundaries, they are also more likely to be challenging to determine, as they can be out of focus, in the distance, etc.

In Fig. 5, we show the violin plot of the pixelwise coverage errors for all described methods. In this experiment, and this one only, the baseline pixelwise calibration method slightly outperforms all other methods. This is because, with such an enormous amount of calibration data, we can properly calibrate the individual pixel classifiers. The Kandinsky methods still perform well in this situation, only the image-wise calibration is significantly worse. The moral of this experiment is that provided you have access to tens of thousands of calibration images, pixelwise calibration will perform best, followed closely by all Kandinsky methods.

4.2. MS-COCO-S

We now move on to the scenario with little calibration data to spare. This situation is more common, especially in the medical field, due to the cost of annotating large amounts of data. To investigate the low-data scenarios, we now utilize only 100 calibration data points. As one can imagine, the k-means approach will have trouble finding relevant clusters in the images with such few samples. However, our prior knowledge does not change, and we proceed to find the annuli using the GenAnn and FCC approaches, described in Sec. 3. In Fig. 4, we show the k-means and the genetic algorithm approach results. It can be seen that the k-means clusters are no longer as clear as in Fig. 3. The results of all methods are shown in Fig. 6 and Tab. 1. In this low-data regime, all Kandinsky methods outperform both pixelwise and imagewise calibration. As expected, the GenAnn and FCC approaches outperform the k-means approach, as they can better incorporate prior knowledge; k-means gets slightly worse due to the noise in the non-conformity curves.

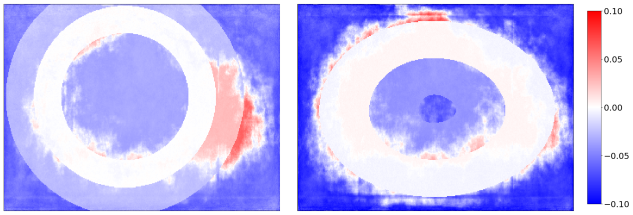

Figure 7. Subtraction image of coverage errors per pixel for imagewise vs. GenAnn calibration (left) and imagewise vs. FCC-calibration (right) on the MS-COCO-S dataset. Red indicates lower coverage error for the former, and blue indicates lower coverage error for the latter.

In Fig. 7, we show the difference in coverage error for the two annuli methods, GenAnn and FCC, versus the errors obtained by imagewise calibration. The Kandinsky methods can group the relevant pixels to get better estimates of non-conformity curves. We refer to the supplementary material Sec. 9 for a comparison of pixelwise and Kandinsky calibration under shrinking calibration dataset sizes.

4.3. Decathlon-L

We now consider a different segmentation task, falling in the medical domain. In particular, we consider the Medical Decathlon Challenge data [6]. The object of interest in our case is the Pancreas data, consisting of CT images along with segmentations of the pancreas and (possibly) tumors. All CT scans were sliced, utilizing slices that contained annotations. The data split was made on the patient level. The resulting slices of scans were of size $(384, 384)$. For our first experiment, we created a large, for medical standards, calibration dataset consisting of 77 patients. In Fig. 8, we show the violin plot for all methods. As also found in Tab. 1, all Kandinsky calibration techniques outperform the pixelwise and imagewise calibration. In this case, the k-means can find the most effective clusters, closely followed by both annuli methods.

4.4. Decathlon-S

Let us now consider the scenario of little calibration data, which occurs frequently in real-life medical datasets. In particular, we utilize only 27 of the patients for the calibration.

Due to the increased noise in the calibration procedure, we can expect the k-means clusters to be less meaningful, as they will cluster based on more spurious relations. The methods using prior knowledge, GenAnn and FCC, will retain their strength, as their baked-in knowledge is more robust to the noise in the calibration. The results are shown in Fig. 9. As before, all Kandinsky methods perform better than the pixel and imagewise calibrations. In this case, the GenAnn approach achieves the lowest coverage errors.

Method	MS-COCO-XL	MS-COCO-S	Decathlon-L	Decathlon-S
Pixel	**0.055** [0.037, 0.080]	0.131 [0.055, 0.319]	0.217 [0.041, 0.500]	0.249 [0.051, 0.500]
Image	0.090 [0.038, 0.175]	0.091 [0.039, 0.180]	0.177 [0.038, 0.410]	0.175 [0.036, 0.410]
K-means	0.066 [0.043, 0.114]	0.086 [0.039, 0.169]	**0.145** [0.038, 0.285]	0.168 [0.040, 0.378]
Gen-Ann	0.067 [0.040, 0.113]	0.068 [0.036, 0.134]	0.150 [0.041, 0.352]	**0.152** [0.037, 0.340]
FCC-Ann	0.064 [0.037, 0.112]	**0.060** [0.032, 0.111]	0.152 [0.040, 0.410]	0.165 [0.036, 0.410]

Table 1. Mean coverage errors (with [0.05, 0.95] quantiles) over all pixels in the test subset of four datasets, computed for five distinct calibration methods. The last three methods show our novel methods; all attempt to cluster similar pixels for simultaneous calibration. All Kandinsky clustering-type methods outperform both pixelwise and imagewise calibration in all datasets except MS-COCO-XL, where the calibration set is so large that pixelwise calibration is superior.

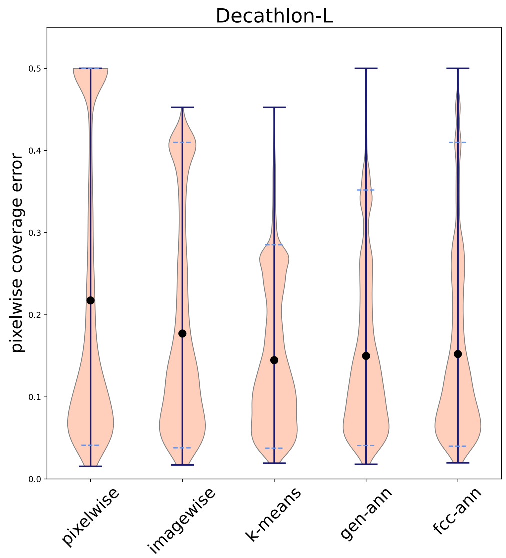

Figure 8. On the large calibration set of 77 patients, all Kandinsky methods have lower coverage errors than the baseline methods. In this case, the k-means clusters can find the most informative clusters of pixels for calibration, closely followed by the annuli methods.

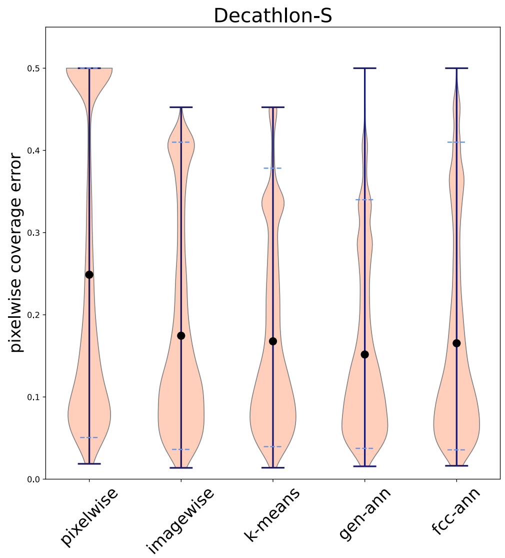

Figure 9. Results for the small decathlon calibration set, using 27 patients. In this case, the GenAnn clusters can most effectively group relevant pixels, closely followed by the k-means and FCC methods. Especially the pixelwise calibration performs poorly in this low-data regime.

Furthermore, in Fig. 11 of the supplementary material, we show the difference in coverage error between the GenAnn and imagewise calibration. Especially near the boundaries, where the pancreas is less often seen, the clusters can find better calibration.

5. Conclusion

In this article, we have investigated the problem of efficient calibration for segmentation models. In particular, in practice, there is often little ‘extra’ data available to calibrate models. We have presented progress for solving this problem by presenting the Kandinsky calibration framework. The framework utilizes the well-organized patterns between pixels-classifiers in segmentation networks. We have proposed three methods for computing these clusters in fairly general settings. In all investigated low-data calibration settings, the Kandinsky methods can produce better-calibrated methods, leading to lower coverage errors on unseen data.

References

[1] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. 2

[2] Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control, 2022. 1

[3] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal Risk Control, 2022.

2

[4] Anastasios N. Angelopoulos, Amit P. Kohli, Stephen Bates, Michael I. Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging, 2022.

[5] Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-Powered Inference, 2023. 1

[6] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):1–13, 2022. 7

[7] Jarosław Błasiok and Preetum Nakkiran. Smooth ECE: Principled Reliability Diagrams via Kernel Smoothing, 2023. 3

[8] Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A Unifying Theory of Distance from Calibration, 2022. 2, 3

[9] Charles George Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970. 5

[10] A. Michael Carrell, Neil Mallinar, James Lucas, and Preetum Nakkiran. The Calibration Generalization Gap, 2022. 2

[11] Pedro Conde, Rui L. Lopes, and Cristiano Premebida. A Theoretical and Practical Framework for Evaluating Uncertainty Calibration in Object Detection, 2023. 2

[12] Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[13] Roger Fletcher. A new approach to variable metric algorithms. *The computer journal*, 13(3):317–322, 1970. 5

[14] David E Goldberg. Cenetic algorithms in search. *Optimization, Machine Learning*, 1989. 4

[15] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24 (109):23–26, 1970. 5

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks, 2017. 2

[17] John H Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992. 4

[18] Fabian Küppers, Anselm Haselhoff, Jan Kronenberger, and Jonas Schneider. Confidence Calibration for Object Detection and Segmentation. In *Deep Neural Networks and Data for Automated Driving: Robustness, Uncertainty Quantification, and Insights Towards Safety*, pages 225–250. Springer International Publishing, Cham, 2022. 2

[19] Gilad Kutiel, Regev Cohen, Michael Elad, Daniel Freedman, and Ehud Rivlin. Conformal Prediction Masks: Visualizing Uncertainty in Medical Imaging. In *ICLR 2023 Workshop on Trustworthy Machine Learning for Healthcare*, 2023. 1

[20] Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1): 71–96, 2014. 2

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[22] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the Calibration of Modern Neural Networks, 2021. 2

[23] Harris Papadopoulos. *Inductive Conformal Prediction: Theory and Application to Neural Networks*. IntechOpen, 2008. 1

[24] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pages 345–356. Springer, 2002. 2

[25] Harris Papadopoulos, Volodya Vovk, and Alex Gammerman. Conformal Prediction with Neural Networks. In *19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007)*, pages 388–395, 2007. 1

[26] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with Valid and Adaptive Coverage. In *Advances in Neural Information Processing Systems*, pages 3581–3591. Curran Associates, Inc., 2020. 1

[27] C. Saunders, A. Gammerman, and V. Vovk. Transduction with Confidence and Credibility. In *Sixteenth International Joint Conference on Artificial Intelligence (IJCAI '99) (01/01/99)*, pages 722–726, 1999. 1

[28] David F Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111): 647–656, 1970. 5

[29] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11:341–359, 1997. 4

[30] Vladimir Vovk. Conditional Validity of Inductive Conformal Predictors. In *Proceedings of the Asian Conference on Machine Learning*, pages 475–490. PMLR, 2012. 1

[31] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005. 2

[32] Håkan Wieslander, Philip J. Harrison, Gabriel Skogberg, Sonya Jackson, Markus Fridén, Johan Karlsson, Ola Spjuth, and Carolina Wählby. Deep Learning With Conformal Prediction for Hierarchical Analysis of Large-Scale Whole-Slide Tissue Images. *IEEE Journal of Biomedical and Health Informatics*, 25(2):371–380, 2021. 2