Self-Calibrating Vicinal Risk Minimisation for Model Calibration

Jiawei Liu, Changkun Ye, Ruikai Cui, Nick Barnes
The Australian National University
Canberra, Australia
{jiawei.liu3, changkun.ye, ruikai.cui, nick.barnes}@anu.edu.au

Abstract

Model calibration, measuring the alignment between the prediction accuracy and model confidence, is an important metric reflecting model trustworthiness. Existing dense binary classification methods, without proper regularisation of model confidence, are prone to being over-confident. To calibrate Deep Neural Networks (DNNs), we propose a Self-Calibrating Vicinal Risk Minimisation (SCVRM) that explores the vicinity space of labeled data, where vicinal images that are farther away from labeled images adopt the groundtruth label with decreasing label confidence. We prove that in the logistic regression problem, SCVRM can be seen as a Vicinal Risk Minimisation plus a regularisation term that penalises the over-confident predictions. In practical implementation, SCVRM is approximated using Monte Carlo sampling that samples additional augmented training images and labels from the vicinal distributions. Experimental results demonstrate that SCVRM can significantly enhance model calibration for different dense classification tasks on both in-distribution and out-of-distribution data. Code is available at https://github.com/Carlisle-Liu/SCVRM.

1. Introduction

Binary dense classification tasks [10, 54, 64] have advanced significantly since the debut of Deep Neural Networks (DNNs) associated with complex network architectures and large numbers of trainable parameters. Increasing model complexity has been shown to negatively impact model calibration [14] which has remained under investigated in the binary dense classification [29–31]. Miscalibration, a mis-alignment between model confidence and prediction accuracy [38], is an undesirable qualify that hinders the deployment of DNNs, especially in safety critical applications. In this work, we study the model calibration for dense binary classification models.

The most commonly observed mis-calibration problem is over-confidence, where the model confidence is significantly higher than the accuracy of its predictions on a cohort of samples [14]. Increasing research interest has been dedicated to study model calibration methods that regularise the probability associated with the prediction to be meaningful in reflecting the chance of the prediction being correct. The existing methods can be roughly categorised into three groups: (i) training objective based [2, 8, 13, 21, 23, 38, 45]; (ii) post-processing based [14, 19, 47]; and (iii) label augmentation based [31, 39, 43]. The first two categories focus on penalising over-confident predictions while the third cat-

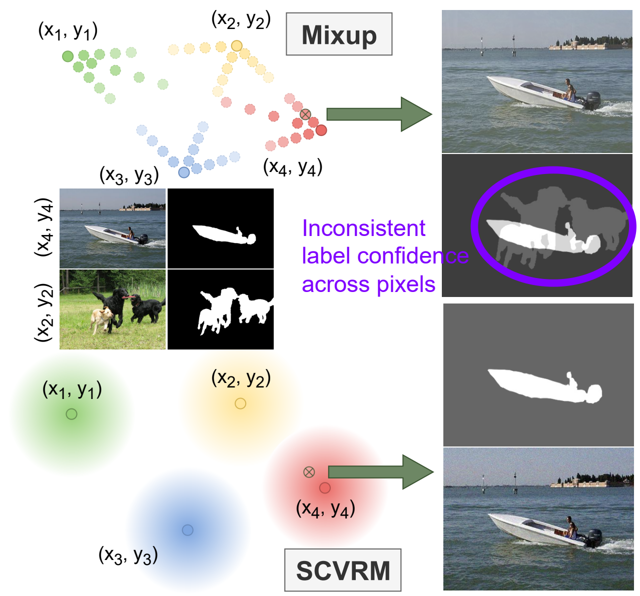

Figure 1. An illustration of Mixup [69] and our proposed SCVRM. Circles with solid boundary $\{(x_i, y_i)\}_{i=1}^4$ are labeled training images. Vicinal images of Mixup (circles with dashed boundary) are only distributed along the vectors connecting the labeled images. On the other hand, in SCVRM, the label of vicinal image adopts the groundtruth category (shown in different colours) of the closest labeled image, but with label confidence (shown in colour intensity) reduced monotonically with increasing Euclidean distance between vicinal and labeled images. The example sample $\otimes$, with Mixup/SCVRM vicinal image and augmented label shown on the right, is selected at the same relative position to (x_4, y_4).

egory directly moderates the confidence of training labels for the DNN models to train on. They do not explore the image space in improving the model calibration.

Vicinal Risk Minimisation (VRM) extends Empirical Risk Minimisation (ERM) by introducing a vicinal distribution around each labeled data in the image space [3, 52]. Mixup [69], a variant of VRM, assumes that vicinal images are only distributed along the vectors connecting the labeled data (pair of image and label) pairs. In addition, instead of letting vicinal images adopt the hard groundtruth label of the nearest labeled image, Mixup assigns smoothed versions of nearest groundtruth labels to the vicinal images. In practice, it samples augmented data through the convex combination of labeled data pairs: $(\tilde{x} = \lambda x_i + (1 - \lambda)x_i, \tilde{y} = \lambda y_i + (1-\lambda)y_j)$, where the combination factor is sampled from a beta distribution: $\lambda \in \text{Beta}(\alpha, \alpha)$. Follow-up works [69, 73] show that Mixup is effective in improving model calibration in the image classification task.

We propose a Self-Calibrating Vicinal Risk Minimisation (SCVRM) to calibrate the DNNs. Different from Mixup [69], the vicinal distribution of SCVRM can use any distribution that sufficiently covers the vicinity space with reasonable probability density. We consider the case where the vicinal distribution is assumed to be Gaussian with standard deviation being a random variable following a uniform distribution. In addition, the labels assigned to the vicinal images are softened versions of the groundtruth label associated with the labeled image at the distribution centre, where the strength of softening is proportional to the L2 distance between the vicinal and labeled images. As shown in Fig. 1, compared with Mixup [69], our proposed SCVRM has the following advantages: (1) the vicinity image space is not restricted by the pair-wise spatial relations of training images; (2) consistent label confidence across the pixels in the augmented label for dense prediction; and (3) defines confidence boundaries beyond the training distribution to better handle out-of-distribution samples.

Given a labeled training image set, SCVRM can be understood by the following principals: (1) the human visual system is invariant to small random image changes [12] so that images undergoing up to a certain level of transformation should remain correctly classified [3, 20, 52]. This is also in line with adversarial robustness approaches that require images under small perturbations to maintain their classifications [7]. However, they are no longer labeled training data, so their label confidence should be slightly reduced. (2) it is understood from information theory [16] that we have no information at extreme distance where the label should assume a uniform categorical distribution with maximum entropy. Following the principal of VRM, we propose this transition should be accomplished by a smooth vicinal transition from the exact label for the labeled images to a uniform categorical distribution at large distance from the labeled data. (See Fig. 2 for examples.)

We summarise our contributions as: (1) propose a Self-Calibrating Vicinal Risk Minimisation (SCVRM), where the labels associated with vicinal images have reduced confidence with increasing Euclidean distance between vicinal and labelled images, to calibrate DNN models for the dense binary classification task; (2) we approximate the vicinal distribution with a Gaussian distribution whose standard deviation is a random variable following a uniform distribution, denoted as SCVRM-G; (3) we supply an example showing that SCVRM is equivalent to VRM and a regularisation term on prediction confidence under a simplified model; (4) We realise SCVRM as a data augmentation technique, where augmented data are sampled from the vicinal distributions with Monte Carlo methods. Finally, we show state-of-the-art calibration results for salient object detection (main paper); and camouflage object and smoke detection, and semantic segmentation (Supp. 12 and Supp. 13).

2. Related Works

Model Calibration: aims to minimise the distributional gap between model confidence and prediction accuracy. Existing methods propose to align confidence with accuracy through (i) post-processing techniques, *e.g.*, Temperature Scaling (TS) [14, 19], Platt Scaling [41, 47], Dirichlet Scaling [22], Bayesian Binning [40, 67], Isotropic Regression [68], and Mix-and-Match [70]; and (ii) objective functions including Brier Loss [2, 8], Confidence Penalty [45], Maximum Mean Calibration Error (MMCE) [24], Soft Calibration Objective [21] and Focal Loss [13, 38]. These methods emphasise on suppressing over-confident predictions to alleviate the mis-calibration issue.

Research efforts have also been dedicated to investigate the effect of data augmentation methods on the calibration degree. Label augmentation methods align the confidence distribution of the target label, to which the model predictions converge through optimisation, with the prediction accuracy distribution. Label Smoothing (LS) [39, 43] directly softens the target label probability, preventing the model from producing over-confident predictions. [31] presents an alternative label augmentation solution that stochastically perturbs the groundtruth label and aligns the confidence distribution of expected label with the prediction accuracy distribution. On the other hand, Mixup [69], which simultaneously augments both the input image and its corresponding groundtruth label conditioned on the relative position between training data, has also been demonstrated to improve the model calibration degree [51].

Vicinal Risk Minimisation: is proposed by [3, 52] to explore the vicinity of labeled data in hope of achieving a better approximation of the expected risk. The vicinity space is approximated with a vicinal distribution that can be either estimated with sufficient unlabeled data, or otherwise assumed to follow certain prior distribution, *e.g.* Gaussian distribution [3, 52]. Mixup [69] is an extension of VRM that restricts the vicinal data to be distributed along the vectors connecting the data pairs through convex combination. Further, Mixup applies the convex combination in both image and label spaces, resulting in the vicinal samples having a softened version of the groundtruth label associated with the nearest labeled image. VRM has also been applied in randomised smoothing [7, 25, 66] to achieve certified adversarial robustness, where a Gaussian kernel with fixed variance is employed to approximate the vicinal distribution.

Salient Object Detection: is a dense binary classification task. Conventional methods primarily depend on hand-crafted features to identify salient objects [5, 17, 18, 44]. Early DNN based approaches employ learned features of local regions, *e.g.*, super-pixel, object proposal and image patch, which need to be further sequentially processed [15, 26, 53]. The advent of Fully Convolutional Network (FCN) shifted the research focus onto the network architecture design for better pixel-wise predictions. Multi-level feature aggregation takes full advantage of spatial cues imbued in low-level features and semantic information embodied in high-level features [42, 72, 78]. Further, attention mechanisms, including spatial attention, channel attention, *etc.*, have been leveraged to explore intra- and inter-feature map correlations [33, 46, 77], with some works directly adopting transformer backbones with self-attention [50, 71]. A different direction investigates auxiliary cues, *e.g.* boundary and edge [59, 62], fixation [57].

3. Preliminary

3.1. Settings and Notations

This papers deals primarily with dense binary classification over the image space $\mathbb{R}^d$ with the corresponding label set of $\mathcal{Y} \in \{0, 1\}^d$, where d is the dimension of image/label space, "0" and "1" represent the two categories. $\mathcal{Y}_s \in [0, 1]^d$ denotes a soft label space, clearly we have $\mathcal{Y} \subset \mathcal{Y}_s$. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be a finite dataset with image and label pairs (x_i, y_i) sampled i.i.d. from the joint distribution $p(x, y)$ defined on $\mathcal{X} \times \mathcal{Y}$. The task is to obtain an optimal classifier $f^* \in \mathcal{H}$ in the Hypothesis space $\mathcal{H} \subset \{f : \mathcal{X} \to \mathcal{Y}\}$ that maximises the prediction accuracy and model calibration performance. The dimensionality or spatial index is omitted for simplicity wherever it is clear.

The model calibration degree over the joint distribution $p(x, y)$ can be evaluated in terms of Expected Calibration Error (ECE) [14, 23]. Let $S(x) \in (0, 1)$ be the Sigmoid-activation value before classification: $f(x) = \mathbb{1}(S(x) > 0.5)$, where $\mathbb{1}(\cdot)$ is an indicator function. The prediction confidence c and accuracy a can be defined as: $c = |S(x) - 0.5| + 0.5$, and $a = \mathbb{1}(f(x) = y)$ respectively. We further use $p_{f,\mathcal{D}}(c, a)$ to denote the joint distribution of prediction confidence and prediction accuracy of model, $f(\cdot)$, on dataset, $\mathcal{D}$. Then, ECE can be defined as: ECE$(p_{f,\mathcal{D}}) = \mathbb{E}_{p_{f,\mathcal{D}}(c)}\big[|\mathbb{E}_{p_{f,\mathcal{D}}(a|c)}[a] - c|\big]$, where $\mathbb{P}_{f,\mathcal{D}}(c)$ is a marginal distribution on prediction confidence, and $\mathbb{E}_{\mathbb{P}_{\theta,\mathcal{D}}(a|c)}$ is a conditional distribution of prediction accuracy. For a well calibrated model, we should have: $p_{f,\mathcal{D}}(f(x)_i = y_i | c = r) = r, \ \forall r \in [0, 1]$.

3.2. Vicinal Risk Minimisation

We may formulate a learning problem as searching for model $f \in \mathcal{H}$ that minimises expected risk $R(f)$ over loss function $\ell(f(x), y)$, which can be written as:

$$R(f) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(f(x), y) \cdot p(x, y) dx dy. \tag{1}$$

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_i^N$, the Empirical Risk Minimisation (ERM) approach approximates $R(f)$ by:

$$\begin{aligned} R_E(f) &= \int_{\mathcal{X}\times\mathcal{Y}} \ell(f(x), y) \cdot \frac{1}{N}\delta_{x_i,y_i}(x, y) dx dy \\ &= \frac{1}{N}\sum_{i=1}^{N} \ell(f(x_i), y_i), \end{aligned} \tag{2}$$

where $\delta_{x_i,y_i}(x, y)$ is the Dirac delta distribution over (x, y) that can only adopts the value (x_i, y_i).

Based on ERM, the Vicinal Risk Minimisation (VRM) approach [3, 52] approximates $R(f)$ by

$$R_V(f) = \frac{1}{N}\sum_{i=1}^{N}\int \ell(f(\tilde{x}_i), y_i) p(\tilde{x}_i|x_i) d\tilde{x}, \tag{3}$$

where $p(\tilde{x}_i|x_i)$ denotes the Probability Density Function (PDF) of vicinal images $\tilde{x}_i$ given labeled image x_i.

In practice, VRM proposes to use isotropic Gaussian distributions as a choice of $p(\tilde{x}_i|x_i)$. Specifically, VRM defines $\tilde{x}_i = x_i + \epsilon, \epsilon \sim \mathcal{N}_d(0, \sigma^2 I_d)$, where I_d is a $d \times d$ identity matrix and the variance σ^2 is a fixed hyperparameter. The corresponding $R_V(f)$ can then be written as:

$$R_{V-G}(f; \sigma) = \frac{1}{N}\sum_{i=1}^{N}\int \ell(f(x_i + \epsilon), y_i) p(\epsilon) d\epsilon, \tag{4}$$

where $p(\epsilon) = \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{\|\epsilon\|_2^2}{2\sigma^2}\right)$ is the PDF of the additive Gaussian noise ϵ.

4. Method

4.1. Self-Calibrating Vicinal Risk Minimisation

Our method is based on the principals that samples in the proximity of labeled data should inherit their groundtruth labels with reduced label confidence, while samples in the

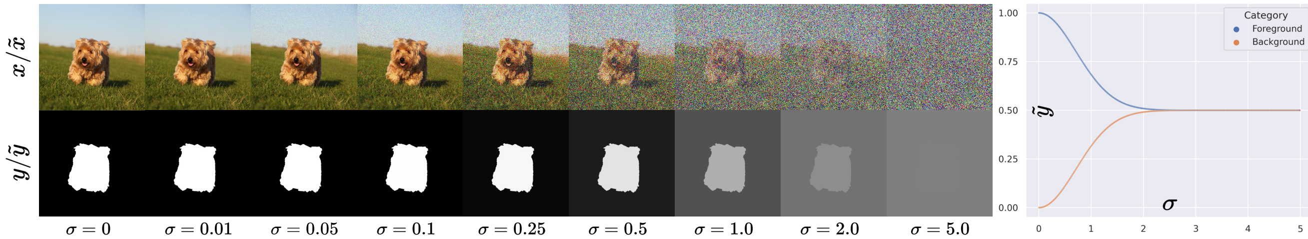

Figure 2. Examples of vicinal images and assigned labels under various standard deviation σ values. $\sigma = 0$ denotes the labeled data (x, y). Vicinal images are generated via adding Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$ to the labeled image as defined in Eq. 7. Their assigned labels, being the softer versions of the groundtruth label y, are computed with the Eq. 8 and Eq. 9, whose plot are shown on the right.

extreme distance should have a near uniform categorical distribution for their labels. As an extension of VRM, we reduce label confidence for vicinal images based on the L2-distance between vicinal and labeled images. Extending the definition of VRM [3, 52], we propose a novel Self-Calibrating Vicinal Risk Minimisation (SCVRM):

Definition 1. *(SCVR) Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$. For $f \in \mathcal{H}$ and $\ell : f(\mathcal{X}) \times \mathcal{Y}_s \rightarrow \mathbb{R}$, Self-Calibrating Vicinal Risk is defined as:*

$$R_{SC}(f) = \frac{1}{N}\sum_{i=1}^{N}\int \ell\Big(f(\tilde{x}_i), g(y_i, \|\tilde{x}_i - x_i\|_2)\Big) \times p(\tilde{x}_i|x_i)d\tilde{x}_i, \quad (5)$$

where $g : \mathcal{Y} \times \mathbb{R}_+ \rightarrow \mathcal{Y}_s$.

The difference between VRM (Eq. 3) and SCVRM lies in the labels associated with the vicinal images. VRM assigns the exact groundtruth label y_i paired with labeled image x_i to all its vicinal images $\tilde{x}_i$, whereas we use softened versions of groundtruth label $g(y_i, \|\tilde{x}_i - x_i\|_2)$, whose label confidence reduces with increases in the L2-distance between the vicinal and labeled images. Following VRM to represent the distribution of vicinal images conditioned on the labeled image $p(\tilde{x}|x)$ with an isotropic Gaussian distribution with a fixed variance, our SCVRM can be written as:

$$R_{SC-V}(f) = \frac{1}{N}\sum_{i}^{N}\int \ell\Big(f(x_i + \epsilon), g(y_i; \|\epsilon\|_2)\Big)p(\epsilon)d\epsilon, \quad (6)$$

where L2-norm of Gaussian noise equals the L2-distance between vicinal and labeled images $\|\epsilon\|_2 = \|\tilde{x}_i - x_i\|_2$.

We observe that the isotropic Gaussian noise $\epsilon \sim \mathcal{N}_d(0, \sigma^2 I_d)$ used in the VRM may face the problem of a bubbling effect, that results in the probability density of an isotropic Gaussian $\mathcal{N}_d(\mu, \sigma^2 I_d)$ concentrating on a thin spherical shell centred on $\mu \in \mathbb{R}^d$ with radius $\sigma\sqrt{d}$ in high dimensional space $d \gg 1$ [1, 36]. Following this observation: (a) we propose to let the scale of the standard deviation of the isotropic Gaussian distribution be a random variable following a uniform distribution $\sigma \sim \mathcal{U}(0, \gamma]$ to prevent vicinal images $\tilde{x}$ from being distributed only near a hypersphere centring on x_i; and (b) we approximate the L2-distance between the vicinal and labeled images with a dummy variable $\|\tilde{\epsilon}\|_2 \approx \|\epsilon\|_2$, which is set to $\|\tilde{\epsilon}\|_2 = \sigma\sqrt{d}$ (See Supp. 8.1 for experimental justifications). Our implementation of SCVRM can then be defined as:

$$R_{SC-G}(f) = \frac{1}{N}\sum_{i}^{N}\int \ell\Big(f(x_i + \epsilon), g(y_i; \|\tilde{\epsilon}\|_2)\Big) \times p(\epsilon)p(\sigma)d\epsilon d\sigma, \quad (7)$$

where $p(\sigma) = 1/\gamma$ is the PDF of the standard deviation σ.

There are many choices for the projection function from hard label space to soft label space. The most straightforward solution is to apply the label smoothing [39] function: $\text{LS}(y, \beta) = (1 - \beta) + \frac{\beta}{K}, \beta \in [0, 1]$ where $K = 2$ denotes the number of classes in a binary task. Then our projection function can be defined as:

$$g_{\text{LS}}(y; \varphi(\|\tilde{\epsilon}\|_2)) = \Big(1 - \varphi(\|\tilde{\epsilon}\|_2)\Big)y + \frac{\varphi(\|\tilde{\epsilon}\|_2)}{2}, \quad (8)$$

where we define $\varphi(\cdot) : \mathbb{R}_0^+ \rightarrow [0, 1)$ to be a Gaussian function $\varphi_G(\cdot; \cdot)$, that is formulated as:

$$\varphi_G(\|\tilde{\epsilon}\|_2; \eta) = 1 - \exp\left(-\frac{\|\tilde{\epsilon}\|_2}{\eta}\right), \quad (9)$$

where η is a hyperparameter that scales the value of $\|\tilde{\epsilon}\|_2$, and we empirically set $\eta = \sqrt{d}$. Fig. 2 illustrates the resultant groundtruth map assigned to vicinal images sampled under various standard deviations. Please note that the function φ can be an arbitrary function that satisfies: $\varphi : \mathbb{R}_0^+ \rightarrow [0, 1)$. We investigate other options in Supp. 8.2.

4.2. Connection between SCVRM and VRM

We analyse the relationship between VRM and our proposed SCVRM in a simple setting, where we find mathematical connections between the two approaches. Specifically, we consider the case of binary classification tasks, where our $R_{SC-G}(f)$ in Eq. 7 adopts a single-layer logistic regression model and the binary cross entropy loss. In this setting, we prove that if $g(y_i, \|\tilde{\epsilon}\|_2)$ with $\|\tilde{\epsilon}\|_2 = \sigma\sqrt{d}$ is a label smoothing function, $R_{SC-G}(f)$ can be written as a combination of the Vicinal Risk $R_{V-G}(f;\sigma)$ in Eq. 4 and a regularization term that penalises overconfident predictions of the model f (See proof in Supp. 7.1).

Example 1. *In $R_{SC-G}(f)$ defined in Eq. 7, suppose:*
- *Data: $\{(x_i, y_i)\}_{i=1}^N$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{0,1\}$,*
- *Model: $f(x) = 1/(1+e^{-w^T x})$ with $w \in \mathbb{R}^d$,*
- *Loss: $\ell(f(x), y) = -y \log f(x) - (1-y)\log(1-f(x))$,*
- *$g(y, \|\tilde{\epsilon}\|_2) = (1-2\sigma)y + \sigma$, $\sigma \sim \mathcal{U}(0,\gamma]$ and $\gamma \in (0, \frac{1}{2}]$.*

then we have:

$$R_{SC-G}(f) = \int_0^\gamma \frac{1}{\gamma} R_{V-G}(f;\sigma)d\sigma + \tau(f), \tag{10}$$

where

$$\tau(f) = \frac{\gamma}{2N}\sum_{i=1}^{N}(2y_i - 1)\cdot w^T x_i, \tag{11}$$

and the first term of the RHS in Eq. 10 is equivalent to introducing our design of $\sigma \sim \mathcal{U}(0,\gamma]$ into $R_{V-G}(f;\sigma)$ which is the VRM with Gaussian kernel defined in Eq. 4.

Note that the label satisfies $y_i \in \{0,1\}$, therefore in the term $\tau(f)$ in Eq. 11 we have:

$$(2y_i - 1)\cdot w^T x_i = \begin{cases} w^T x_i, & y_i = 1, \\ -w^T x_i, & y_i = 0. \end{cases} \tag{12}$$

In this sense, the term $\tau(f)$ penalises high $w^T x_i$ when $y_i = 1$ and penalises low $w^T x_i$ when $y_i = 0$. Since high $w^T x$ will lead to $f(x) = 1/(1+e^{w^T x})$ approaching 1 and low $w^T x$ will lead to $f(x)$ approaching 0, $\tau(f)$ actually penalizes overconfident predictions of the model f.

Remark 1. *In this simplified case, minimising the Self-Calibrating Vicinal Risk defined in Eq. 7 equivalently minimises the Vicinal Risk defined in Eq. 4 incorporating our design of $\sigma \in \mathcal{U}(0,\gamma]$ and a regularisation term defined in Eq. 11, where the former improves the model classification accuracy the latter prevents the model from becoming overconfident on vicinal samples.*

4.3. Practical Model

In practical implementation, we approximate the intractable SCVRM defined in Eq. 7 with the Monte-Carlo (MC) Sampling. More specifically, it is formulated as a data augmentation technique where, based on the labeled dataset $\mathcal{D} = \{(x_1, y_i)\}_{i=1}^N$, we sample an augmented dataset:

$$\mathcal{D}_{sc} = \bigcup_{i=1}^{N}\Big\{(\tilde{x}_i^j, \tilde{y}_i^j)\,|\, \tilde{x}_i^j = x_i + \epsilon_i^j, \epsilon_i^j \sim_{i.i.d} \mathcal{N}(0, \sigma_{i,j}^2 I_d), \\ \sigma_{i,j} \sim_{i.i.d} \mathcal{U}(0,\gamma], \\ \tilde{y}_i^j = g_{\text{LS}}\Big(y_i; \varphi_G(\|\tilde{\epsilon}\|_2;\eta)\Big)\Big\}_{j=1}^{M}, \tag{13}$$

where $\tilde{x}_i^j$ is the j^{th} augmented image sampled from the vicinity of the i^{th} labeled image, $\tilde{y}_i^j$ is a smoothed version of the groundtruth label y_i, M is a hyperparameter being the number of augmented data from the vicinity of each labeled data, and we set $\|\tilde{\epsilon}\|_2 = \sigma\sqrt{d}$ to approximate the L2-distance between the vicinal and labeled images $\sigma\sqrt{d} \approx \|\epsilon\|_2$ which relieves the computational burden of computing an L2-norm in high-dimensional space. The model is trained on both the labeled dataset $\mathcal{D}$ and the augmented dataset $\mathcal{D}_{sc}$ using a Binary Cross Entropy loss: $\mathcal{L}_{bce}(f(x), y) = -y\log(f(x)) - (1-y)\log(1-f(x))$. The training loss can be defined as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\Big(\mathcal{L}_{bce}\big(f(x_i), y_i\big) + \sum_{j=1}^{M}\mathcal{L}_{bce}\big(f(\tilde{x}_i^j), \tilde{y}_i^j\big)\Big). \tag{14}$$

Note that the augmented dataset defined in Eq. 13 is resampled after each training epoch.

5. Experiments and Results

The main paper presents results on Salient Object Detection. Results for other dense classification tasks, *e.g.*, Camouflaged Object Detection (binary) and Smoke Detection (binary), and semantic segmentation (multi-class) can be found in Supp. 12 and 13 respectively.

5.1. Implementation Details

Evaluation Metrics: We employ bin-based Expected Calibration Error (denoted ECE_{EW} [14]) and Over-confidence Error (OE_{EW} [51]) with $B = 10$ bins, to evaluate model calibration. (See Supp. 9.1 for implementation details.)
Datasets: Following [31] we divide DUTS-TR [54] into a train ($\mathcal{D}_{\text{TR}}$) and validation set ($\mathcal{D}_{\text{VAL}}$), 9,553 and 1,000 images respectively. We evaluate model calibration degree using SOD test datasets, DUTS-TE [54], DUT-OMRON [65], PASCAL-S [28], SOD [37], ECSSD [63], HKU-IS [26].
Model Architecture: The U-Net based model consists of a ResNet50 encoder and a decoder that are initialised with ImageNet pretrained weights and by default respectively. The implementation uses the Pytorch framework. Experiments with different encoders, *e.g.* VGG16 [49] and Swin transformer [34] are detailed in Supp. 11.
Hyperparameters: Optimal calibration performance of model is achieved by setting $\gamma = 2$, $\eta = \sqrt{d}$ and $M = 3$.

Table 1. Salient object detection model calibration benchmark. Results are evaluated with ECE_{EW} and OE_{EW} with 10 bins (units in %).

Methods			DUTS-TE [54]		DUT-OMRON [65]		PASCAL-S [28]		SOD [37]		ECSSD [63]		HKU-IS [26]	
Category	Name	Year	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓
SOD Methods	MSRNet [27]	2017	2.57	2.34	3.32	3.16	3.44	3.23	6.42	6.14	0.97	0.94	0.92	0.87
	SRM [55]	2017	4.02	3.72	4.19	3.96	4.88	4.59	9.93	9.58	2.53	2.35	1.86	1.72
	Amulet [74]	2017	5.67	5.28	5.84	5.49	5.76	5.43	10.03	9.59	2.56	2.42	1.98	1.87
	BMPM [72]	2018	3.74	3.52	4.52	4.37	4.88	4.68	8.16	7.93	1.95	1.89	1.58	1.53
	DGRL [56]	2018	4.12	3.86	4.41	4.21	5.01	4.77	8.44	8.20	2.13	2.02	1.63	1.53
	PAGR [75]	2018	4.04	3.79	5.14	4.96	5.64	5.37	12.17	11.87	2.84	2.70	1.62	1.54
	PiCANet [33]	2018	5.12	4.90	4.84	4.70	8.14	7.92	10.50	10.30	3.48	3.39	2.55	2.47
	CPD [61]	2019	3.97	3.78	4.20	4.06	5.37	5.17	9.65	9.39	2.29	2.19	1.99	1.90
	BASNet [48]	2019	5.00	4.86	4.93	4.83	6.50	6.36	10.40	10.27	2.74	2.70	2.30	2.26
	EGNet [76]	2019	3.33	3.14	3.66	3.50	5.42	5.19	8.04	7.79	1.98	1.88	1.47	1.40
	AFNet [11]	2019	3.95	3.74	4.25	4.09	5.06	4.84	8.15	8.02	2.38	2.27	1.87	1.78
	PoolNet [32]	2019	3.33	3.12	3.86	3.70	5.32	5.07	8.14	7.87	2.00	1.90	1.82	1.75
	GCPANet [4]	2020	3.18	2.99	3.99	3.84	4.16	3.97	7.05	6.88	1.61	1.54	1.27	1.21
	MINet [42]	2020	3.65	3.48	4.45	4.29	4.94	4.75	8.01	7.89	2.13	2.03	1.74	1.65
	F^3Met [58]	2020	3.67	3.50	4.25	4.10	4.85	4.67	7.95	7.78	2.26	2.16	1.92	1.83
	EBMGSOD [71]	2021	3.45	3.29	4.11	3.95	4.79	4.61	7.48	7.30	2.14	2.05	1.79	1.70
	ICON [79]	2021	2.89	2.76	3.84	3.71	4.08	3.95	6.70	6.55	1.56	1.49	1.38	1.32
	PFSNet [35]	2021	2.94	2.72	3.95	3.81	4.45	4.27	7.59	7.39	2.41	2.25	2.06	1.96
	EDN [60]	2022	3.62	3.47	4.02	3.90	4.89	4.74	8.81	8.66	2.20	2.13	1.65	1.58
Model Calibration Methods	Brier Loss [2]	1950	2.77	2.58	3.55	3.38	3.90	3.70	6.40	6.16	1.37	1.30	1.04	0.99
	Temperature Scaling [14]	2017	2.53	2.34	3.18	3.03	3.56	3.36	6.32	6.05	0.96	0.93	0.83	0.70
	MMCE [23]	2018	2.86	2.67	3.56	3.41	4.00	3.81	6.85	6.63	1.41	1.35	1.18	1.13
	Label Smoothing [39]	2019	2.00	1.79	2.89	2.71	3.04	2.83	5.97	5.69	0.83	0.68	0.82	0.47
	Mixup [51]	2019	2.45	2.25	3.41	3.23	3.13	2.99	5.82	5.70	1.41	0.18	3.83	0.05
	Focal Loss [38]	2020	2.25	2.08	3.10	2.82	3.40	3.13	6.21	5.98	1.41	1.03	1.24	0.77
	AdaFocal [13]	2022	1.61	1.41	2.31	1.84	2.53	2.27	5.88	5.47	1.63	0.79	1.35	0.52
	ASLP [31]	2023	1.40	1.22	1.99	1.83	2.31	2.10	5.50	5.17	0.48	0.20	0.79	0.17
Ours	SCVRM	2023	**0.78**	**0.61**	**1.64**	**1.49**	**1.91**	**1.75**	**3.90**	**3.60**	**0.44**	**0.19**	**0.78**	**0.10**

Optimisation Details: The model is optimised for 30 epochs with an initial learning rate of 2.5×10^{-5} that decays by a factor of 0.9 per epoch after the 10th epoch. During each epoch, $M = 3$ augmented data are sampled from each vicinity space. Image size is set to $3 \times 384 \times 384$ and the batch size is 8. Basic data augmentation techniques including random flipping, translation and cropping are applied.

5.2. Model Calibration Degree Performance

The calibration degrees of existing SOD and model calibration methods, and our proposed SCVRM, evaluated in terms of ECE_{EW} and OE_{EW}, are presented in Tab. 1. It can be observed that SOD methods without regularisations on prediction confidence are, in general, less well calibrated than the models equipped with calibration methods. On the other hand, our proposed SCVRM achieves the lowest calibration errors across all six testing datasets among the model calibration methods. Largest improvements over the second-best model are obtained on DUTS-TE, DUT-OMRON, PASCAL-S and SOD datasets, reducing the ECE by 44.3%, 17.6%, 17.3% and 29.1% respectively. There is little room for improvement on ECSSD and HKU-IS, where ASLP is already well calibrated. On these datasets, our proposed SCVRM maintains the high calibration degrees that are comparable with those of ASLP.

Fig. 3 depicts the joint distribution of prediction confidence and prediction accuracy of some model calibration methods on the DUTS-TE [54] dataset. The joint distribution of better calibrated models become more closely aligned with the oracle line, which indicates a perfectly calibrated model. Our SCVRM has its joint distribution almost completely aligned with the oracle line. In comparison, Mixup is more over-confident with the high-density area of its joint distribution slightly to the bottom-right of the oracle and the low-density area deviating even more. This further demonstrates the effectiveness of exploring a continuous vicinal image space with softened labels.

Table 2. Existing model calibration methods and our proposed LSR evaluated on the 500 Out-of-Distribution texture images [31] selected from the Describable Texture Dataset [6] in terms of ECE_{EW} and OE_{EW} with 10 bins, and Accuracy (ACC).

Method	Evaluation (%)		
	ECE ↓	OE ↓	ACC ↑
Baseline	52.36	51.05	41.88
Brier Loss [2]	38.85	37.18	53.62
Temperature Scaling [14]	51.95	50.46	41.59
Label Smoothing [39]	37.22	35.48	55.41
MMCE [23]	40.64	39.67	54.39
Mixup [51]	31.07	29.10	58.71
Focal Loss [38]	40.01	38.43	49.71
AdaFocal [13]	27.55	25.07	55.39
ASLP [31]	18.31	16.37	61.93
SCVRM	**11.93**	**8.26**	**83.93**

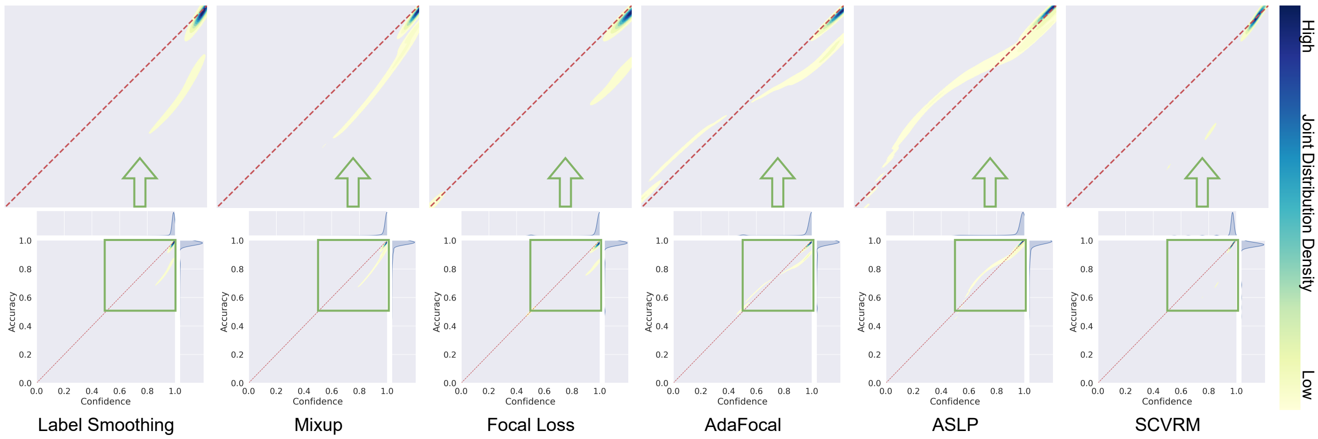

Figure 3. Joint distribution of prediction accuracy (vertical axis) and prediction confidence (horizontal axis) of model calibration methods on DUTS-TE [54]. The dashed red diagonal line represents the perfectly calibrated oracle model.

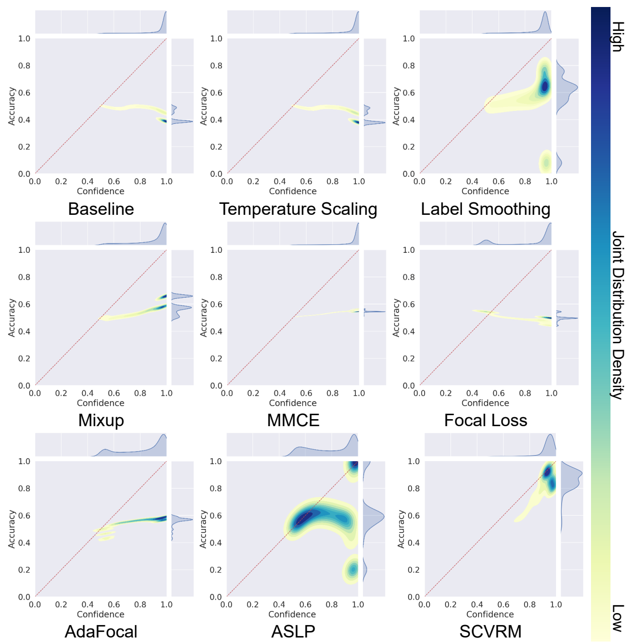

Figure 4. Joint distribution of prediction accuracy (vertical axis) and prediction confidence (horizontal axis) of model calibration methods on the 500 Out-of-Distribution texture images [31] selected from the Describable Texture Dataset [6]. The dashed red diagonal line represents the perfectly calibrated oracle model.

5.3. Model Calibration on Out-of-Distribution Data

The calibrating ability of existing model calibration techniques and our proposed SCVRM on Out-of-Distribution (OoD) samples are evaluated on the 500 texture images [31] selected from the Describable Texture Dataset [6]. These texture images, not including any salient (foreground) object, demonstrates certain level of distributional shift from both training and testing distributions of SOD. As shown in Tab. 2, SCVRM achieves the lowest calibration errors on OoD data. Compared with the second-best model, it reduces ECE by 34.8% and OE by 49.5%. The improvements are also reflected in the joint distribution plot in Fig. 4, where the distribution area of SCVRM is significantly better aligned with the oracle line. Further, SCVRM significantly outperforms Mixup in terms of both calibration and classification. Its superior performance can be attributed to more effective utilisation of vicinity space, and a distance-based label augmentation technique that yields consistent label confidence across the pixels.

5.4. Ablation Study

We investigate the effect of SCVRM and its hyperparameters including γ, η and M which adopt the default setting specified in Sec. 5.1 unless stated otherwise.

Effect of SCVRM: We compare our SCVRM with ERM, the baseline model, VRM with a fixed standard deviation, and VRM* with our design of $\sigma \sim \mathcal{U}(0, \gamma]$ (see implementation detail in Supp. 8.3). Tab. 3 shows that VRM (see implementation detail in Supp. 8.3) does not reduce calibration error over ERM in general, producing mixed impacts across the testing datasets. Introducing our design of $\sigma \sim \mathcal{U}(0, \gamma]$ also does not alleviate VRM's over-confidence. On the other hand, SCVRM significantly improves the model calibration over ERM, VRM and VRM*. In addition, SCVRM also achieves improved classification accuracy than ERM (see Supp. 10.1), which can be attributed to more effective utilisation of vicinity space with our design of $\sigma \sim \mathcal{U}(0, \gamma]$. It also enables VRM* to achieve better classification accuracy compared to VRM (see Supp. 10.1).

Effect of γ: The hyperparameter γ suggests the exploration radius around the training images. As illustrated in Fig. 5a, SCVRM can reach optimal calibration performance in a wide range $\gamma = [1, 10]$. When γ is too small, *e.g.* 0.5, SCVRM achieves limited improvements on model calibration as the vicinal images are too close to the labeled images, and their corresponding augmented labels retain rel-

Table 3. Ablation study on SCVRM.

Method	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓
	DUTS-TE [54]		DUT-OMRON [65]		PASCAL-S [28]	
ERM (Baseline)	3.05	2.89	3.80	3.68	4.14	3.98
VRM	3.54	3.37	4.35	4.21	4.39	4.24
VRM*[1]	3.18	3.03	3.90	3.79	4.13	3.98
SCVRM	**0.78**	**0.61**	**1.64**	**1.49**	**1.91**	**1.75**
Method	SOD [37]		ECSSD [63]		HKU-IS [26]	
ERM (Baseline)	7.07	6.85	1.82	1.76	1.38	1.34
VRM	6.60	6.42	1.67	1.62	1.30	1.26
VRM*[1]	7.35	7.16	1.72	1.68	1.28	1.23
SCVRM	**3.90**	**3.60**	**0.44**	**0.19**	**0.78**	**0.10**

[1] VRM* incorporates our design of $\sigma \sim \mathcal{U}(0, \gamma]$.

atively high label confidences (see Fig. 2). As γ grows, the augmented vicinal images can get sufficiently far away from the labeled images to set up label confidence contours of various levels and obtains optimal model calibration.

Effect of η: The hyperparameter η is part of the Gaussian equation (Eq. 9) that affects the strength of smoothing factor. (See Supp. 8.4 for example vicinal data under different η.) At $\eta = 0.1\sqrt{d}$, the resultant augmented label is overly softened, leading to the trained model becoming under-confident[1]. When η is too large, *e.g.* $5\sqrt{d}$ and $10.0\sqrt{d}$, vicinal images all retain a relatively high label confidence, resulting in insufficient calibration regularisation. We find that η works well in $[0.5\sqrt{d}, 2\sqrt{d}]$, leading consistently to near-optimal calibration results.

Effect of M: We ablate the number of augmented images sampled from the vicinal distribution of each labeled image per training epoch, setting $M = \{1, 2, 3, 4, 5\}$. Results in Fig. 5c show that SCVRM is not very sensitive to the number of augmented data sampled M.

5.5. Discussion

We further demonstrate the effectiveness and generalisation ability of SCVRM in calibrating DNNs via additional experiments. The default training setting specified in Sec. 5.1 is adopted unless specified otherwise.

Effectiveness in Multi-Class Dense Classification: SCVRM can also be generalised to semantic segmentation [9]. We demonstrate its effectiveness in calibrating multi-class dense classification model (See Supp. 13).

Effectiveness in Other Binary Dense Classification Tasks: We verify the effectiveness of SCVRM in improving model calibration degree and dense classification accuracy of DNNs on additional binary dense classification tasks, such as Camouflaged Object Detection [10] and Smoke Detection [64] (See Supp. 12).

Effectiveness with Different Base Models: SCVRM is also effective in calibrating different base models while improving their dense classification accuracy. across a range of base models. We show its compatibility with VGG16 [49] and Swin transformer [34] (See Supp. 11).

[1] Under-confidence occurs when ECE − OE > OE.

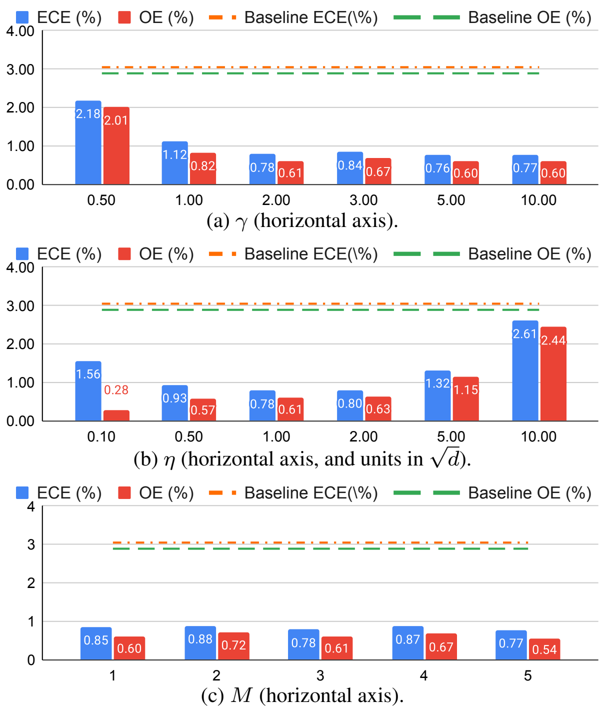

Figure 5. ECE_{EW} and OE_{EW} scores on the DUTS-TE [54] testing dataset under different choices of hyperparameters: (a) $\gamma = \{0.5, 1.0, 2.0, 3.0, 5.0, 10.0\}$; and (b) $\eta = \{i\sqrt{d} \,|\, i = 0.1, 0.5, 1.0, 2.0, 5.0, 10.0, 20.0\}$; and (c) $M = \{1, 2, 3, 4, 5\}$.

Effectiveness on Existing SOD Models: SCVRM can be utilised by various existing SOD models, *e.g.*, EBMG-SOD [71], EDN [60] and ICON [79] (See Supp.14).

6. Conclusion

We propose a Self-Calibrating Vicinal Risk Minimisation (SCVRM) to calibrated DNNs via exploring the vicinity space of labeled data. Vicinal images adopt the groundtruth label of the labeled image at the centre of vicinal distribution, but with diminishing label confidence as they get farther away. In a simplified setting, SCVRM is proved equivalent to a Vicinal Risk Minimisation plus a regularisation term, where the former improves model classification accuracy and the later penalises over-confident predictions. In practice, SCVRM is implemented as a data augmentation technique where MC sampling is applied to sample augmented data from the vicinal distribution. Experimental results on various dense classification tasks demonstrate the effectiveness of SCVRM in improving not only model calibration, but also dense classification accuracy. We also thoroughly study its compatibility with different backbone models and existing methods.

Acknowledgement

This research was in-part supported by the ANU-Optus Bushfire Research Center of Excellence.

References

[1] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006. 4

[2] GLENN W BRIER. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. 1, 2, 6

[3] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, 13, 2000. 2, 3, 4

[4] Zuyao Chen, Qianqian Xu, Runmin Cong, and Qingming Huang. Global context-aware progressive aggregation network for salient object detection. In *AAAI*, pages 10599–10606, 2020. 6

[5] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2014. 3

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 6, 7

[7] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, pages 1310–1320. PMLR, 2019. 2, 3

[8] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983. 1, 2

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 8

[10] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, pages 2777–2787, 2020. 1, 8

[11] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *CVPR*, pages 1623–1632, 2019. 6

[12] Masoud Ghodrati, Amirhossein Farzmahdi, Karim Rajaei, Reza Ebrahimpour, and Seyed-Mahdi Khaligh-Razavi. Feedforward object-vision models only tolerate small image variations compared to human. *Frontiers in computational neuroscience*, 8:74, 2014. 2

[13] Arindam Ghosh, Thomas Schaaf, and Matthew R. Gormley. Adafocal: Calibration-aware adaptive focal loss. In *NeurIPS*, 2022. 1, 2, 6

[14] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330. PMLR, 2017. 1, 2, 3, 5, 6

[15] Shengfeng He, Rynson WH Lau, Wenxi Liu, Zhe Huang, and Qingxiong Yang. Supercnn: A superpixelwise convolutional neural network for salient object detection. *IJCV*, 115:330–344, 2015. 3

[16] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957. 2

[17] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, pages 2083–2090, 2013. 3

[18] Zhuolin Jiang and Larry S Davis. Submodular salient region detection. In *CVPR*, pages 2043–2050, 2013. 3

[19] Tom Joy, Francesco Pinto, Ser-Nam Lim, Philip HS Torr, and Puneet K Dokania. Sample-dependent adaptive temperature scaling for improved calibration. In *AAAI*, pages 14919–14926, 2023. 1, 2

[20] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *CVPR*, pages 8828–8838, 2020. 2

[21] Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C Mozer, and Becca Roelofs. Soft calibration objectives for neural networks. In *NeurIPS*, pages 29768–29779, 2021. 1, 2

[22] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *NeurIPS*, 2019. 2

[23] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *ICML*, pages 2805–2814. PMLR, 2018. 1, 3, 6

[24] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *ICML*, pages 2805–2814. PMLR, 2018. 2

[25] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*, pages 656–672. IEEE, 2019. 3

[26] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015. 3, 5, 6, 8

[27] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *CVPR*, pages 2386–2395, 2017. 6

[28] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *CVPR*, pages 280–287, 2014. 5, 6, 8

[29] Jiawei Liu, Jing Zhang, and Nick Barnes. Modeling aleatoric uncertainty for camouflaged object detection. In *WACV*, pages 1445–1454, 2022. 1

[30] Jiawei Liu, Jing Zhang, Ruikai Cui, Kaihao Zhang, Weihao Li, and Nick Barnes. Generalised co-salient object detection. *arXiv preprint arXiv:2208.09668*, 2022.

[31] Jiawei Liu, Changkun Ye, Shan Wang, Ruikai Cui, Jing Zhang, Kaihao Zhang, and Nick Barnes. Model calibration in dense classification with adaptive label perturbation. In *ICCV*, pages 1173–1184, 2023. 1, 2, 5, 6, 7

[32] Jiang-Jiang Liu, Qibin Hou, Zhi-Ang Liu, and Ming-Ming Cheng. Poolnet+: Exploring the potential of pooling for salient object detection. *IEEE TPAMI*, 45(1):887–904, 2022. 6

[33] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, pages 3089–3098, 2018. 3, 6

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer:

Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 5, 8
[35] Mingcan Ma, Changqun Xia, and Jia Li. Pyramidal feature shrinking for salient object detection. In *AAAI*, pages 2311–2318, 2021. 6
[36] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003. 4
[37] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, pages 416–423. IEEE, 2001. 5, 6, 8
[38] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. In *NeurIPS*, pages 15288–15299, 2020. 1, 2, 6
[39] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *NeurIPS*, 2019. 1, 2, 4, 6
[40] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, 2015. 2
[41] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *ICML*, pages 625–632, 2005. 2
[42] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9413–9422, 2020. 3, 6
[43] Hyekang Park, Jongyoun Noh, Youngmin Oh, Donghyeon Baek, and Bumsub Ham. Acls: Adaptive and conditional label smoothing for network calibration. In *ICCV*, pages 3936–3945, 2023. 1, 2
[44] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740. IEEE, 2012. 3
[45] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *ICLR*, 2017. 1, 2
[46] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pages 7254–7263, 2019. 3
[47] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 1, 2
[48] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489, 2019. 6
[49] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5, 8
[50] Avishek Siris, Jianbo Jiao, Gary K.L. Tam, Xianghua Xie, and Rynson W.H. Lau. Scene context-aware salient object detection. In *ICCV*, pages 4156–4166, 2021. 3
[51] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *NeurIPS*, 2019. 2, 5, 6
[52] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999. 2, 3, 4
[53] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *CVPR*, pages 3183–3192, 2015. 3
[54] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017. 1, 5, 6, 7, 8
[55] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *ICCV*, pages 4019–4028, 2017. 6
[56] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *CVPR*, pages 3127–3135, 2018. 6
[57] Wenguan Wang, Jianbing Shen, Xingping Dong, Ali Borji, and Ruigang Yang. Inferring salient objects from human fixations. *IEEE TPAMI*, 42(8):1913–1927, 2019. 3
[58] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *AAAI*, pages 12321–12328, 2020. 6
[59] Runmin Wu, Mengyang Feng, Wenlong Guan, Dong Wang, Huchuan Lu, and Errui Ding. A mutual learning method for salient object detection with intertwined multi-supervision. In *CVPR*, pages 8150–8159, 2019. 3
[60] Yu-Huan Wu, Yun Liu, Le Zhang, Ming-Ming Cheng, and Bo Ren. Edn: Salient object detection via extremely-downsampled network. *IEEE TIP*, 2022. 6, 8
[61] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, pages 3907–3916, 2019. 6
[62] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *ICCV*, pages 7264–7273, 2019. 3
[63] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162, 2013. 5, 6, 8
[64] Siyuan Yan, Jing Zhang, and Nick Barnes. Transmission-guided bayesian generative model for smoke segmentation. In *AAAI*, pages 3009–3017, 2022. 1, 8
[65] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173. IEEE, 2013. 5, 6, 8
[66] Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *ICML*, pages 10693–10705. PMLR, 2020. 3
[67] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, pages 609–616, 2001. 2

[68] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *ACM SIGKDD Int. Conf. Knowledge Disc. Data Min.*, pages 694–699, 2002. 2

[69] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1, 2, 3

[70] Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *ICML*. PMLR, 2020. 2

[71] Jing Zhang, Jianwen Xie, Nick Barnes, and Ping Li. Learning generative vision transformer with energy-based latent space for saliency prediction. In *NeurIPS*, pages 15448–15463, 2021. 3, 6, 8

[72] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *CVPR*, pages 1741–1750, 2018. 3, 6

[73] Linjun Zhang, Zhun Deng, Kenji Kawaguchi, and James Zou. When and how mixup improves calibration. In *ICML*, pages 26135–26160. PMLR, 2022. 2

[74] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, pages 202–211, 2017. 6

[75] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *CVPR*, pages 714–722, 2018. 6

[76] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *ICCV*, pages 8779–8788, 2019. 6

[77] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *CVPR*, pages 3085–3094, 2019. 3

[78] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, pages 35–51. Springer, 2020. 3

[79] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE TPAMI*, 2022. 6, 8