Self-Calibrating Vicinal Risk Minimisation for Model Calibration

Supplementary Material

7. Proof

7.1. Connection between SCVRM and VRM

We first re-state Example 1 defined in Sec. 4.2, where we analyse the relationship between VRM and our proposed SCVRM in a simplified setting. Specifically, we consider the case of binary classification tasks, where our $R_{SC-G}(f)$ in Eq. (7) adopts the logistic regression model and the binary cross entropy loss. In this setting, we prove that if $g(y_i, \sigma\sqrt{d})$ is a label smoothing function, $R_{SC-G}(f)$ can be written as a combination of two terms, a VRM term and a regularization term that penalizes overconfident predictions.

Example. *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: $\ell(f(x), y) = -y \log f(x) - (1-y)\log(1 - f(x))$,*
- *$g(y, \sigma\sqrt{d}) = (1-2\sigma)y + \sigma$, $\sigma \sim \mathcal{U}(0, \gamma]$ and $\gamma \in (0, \frac{1}{2}]$.*

then we have:

$$R_{SC-G}(f) = \int_0^\gamma \frac{1}{\gamma} R_{V-G}(f;\sigma)d\sigma + \tau(f), \tag{15}$$

where

$$\tau(f) = \frac{\gamma}{2N}\sum_{i=1}^N (2y_i - 1)\cdot w^T x_i, \tag{16}$$

and the first term of RHS in Eq. (15) 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 in binary case we have $y_i \in \{0, 1\}$, therefore in the term $\tau(f)$ in Eq. (15) we have:

$$(2y_i - 1)\cdot w^T x_i = \begin{cases} w^T x_i, & y_i = 1, \\ -w^T x_i, & y_i = 0. \end{cases} \tag{17}$$

The proof of Example. 1 is:

Proof. Substituting each term into the definition of the proposed risk defined in Eq. (7), for a given $\sigma \in (0, \frac{1}{2}]$ we have:

$$\begin{aligned}
&R_{SC-G}(f;\sigma) \\
=&\frac{1}{N}\sum_{i=1}^N \int \ell(f(x_i+\epsilon), g(y_i;\sigma))\cdot p(\epsilon)d\epsilon \\
=&\frac{1}{N}\sum_{i=1}^N \int p(\epsilon)\cdot\Big(-g(y_i,\sigma)\cdot\log(f(x_i+\epsilon)) \\
&\qquad -(1-g(y_i,\sigma))\cdot\log(1-f(x_i+\epsilon))\Big)d\epsilon \\
=&\frac{1}{N}\sum_{i=1}^N \int p(\epsilon)\cdot\Big(\log\Big(1+e^{-w^T(x_i+\epsilon)}\Big)+ \\
&\qquad (1-(1-2\sigma)y_i-\sigma)\cdot w^T(x_i+\epsilon)\Big)d\epsilon \\
=&\frac{1}{N}\sum_{i=1}^N \int \mathcal{N}(\epsilon-0;\sigma^2)\cdot\Big(\log\Big(1+e^{-w^T(x_i+\epsilon)}\Big)+ \\
&\qquad (1-y_i)\cdot w^T(x_i+\epsilon)\Big)d\epsilon \\
&+\frac{1}{N}\sum_{i=1}^N \int \mathcal{N}(\epsilon-0;\sigma^2)\cdot\sigma\cdot(2y_i-1)\cdot w^T(x_i+\epsilon)d\epsilon
\end{aligned} \tag{18}$$

Note that if we only augment the image, i.e. VRM approach, with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ for some fixed σ, then we have:

$$\begin{aligned}
&R_{V-G}(f;\sigma) \\
=&\frac{1}{N}\sum_{i=1}^N \int p(\epsilon)\cdot\ell(f(x_i+\epsilon), y_i)d\epsilon \\
=&\frac{1}{N}\sum_{i=1}^N \int \mathcal{N}(\epsilon-0;\sigma^2)\cdot\Big(\log\Big(1+e^{-w^T(x+\epsilon)}\Big) \\
&\qquad +(1-y_i)\cdot w^T(x_i+\epsilon)\Big)d\epsilon
\end{aligned} \tag{19}$$

Substituting Eq. (19) into Eq. (18) we have:

$$R_{SC-G}(f;\sigma) = R_{V-G}(f;\sigma) + \tau(f;\sigma), \tag{20}$$

where:

$$\begin{aligned}
&\tau(f;\sigma) \\
=&\frac{1}{N}\sum_{i=1}^N \int \mathcal{N}(\epsilon-0;\sigma^2)\cdot\sigma\cdot(2y_i-1)\cdot w^T(x_i+\epsilon)d\epsilon \\
=&\frac{1}{N}\sum_{i=1}^N \sigma\cdot(2y_i-1)\cdot w^T x_i.
\end{aligned} \tag{21}$$

When $\sigma \sim \mathcal{U}(0,\gamma]$ follows a uniform distribution, the overall $R_{SC}(f)$ can then be written as:

$$\begin{aligned}
&R_{SC-G}(f) = \underset{\sigma\sim\mathcal{U}(0,\gamma]}{\mathbb{E}}[R_{SC-G}(f;\sigma)] \\
&= \int R_{SC-G}(f;\sigma)p(\sigma)d\sigma \\
&= \int_0^\gamma \frac{1}{\gamma} R_{V-G}(f;\sigma)d\sigma \\
&\quad + \int_0^\gamma \frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^{N} \sigma\cdot(2y_i-1)\cdot w^T x_i d\sigma \qquad (22) \\
&= \int_0^\gamma \frac{1}{\gamma} R_{V-G}(f;\sigma)d\sigma + \frac{\gamma}{2N}\sum_{i=1}^{N}(2y_i-1)\cdot w^T x_i \\
&= \int_0^\gamma \frac{1}{\gamma} R_{V-G}(f;\sigma)d\sigma + \tau(f)
\end{aligned}$$

□

8. Implementations

8.1. Additive Gaussian Noise ϵ

We experimentally verify that our approximation of L2 distance between the vicinal and labeled images (equivalently the L2 norm of additive Gaussian noise: $\|\tilde{x}-x\|_2 = \|\epsilon\|_2$), by $\sigma\sqrt{d} \approx \|\epsilon\|_2$ is valid. We compare with the following two settings: (1) "Exact" $\varphi_G(\|\epsilon\|_2, \eta)$: the L2-distance of the Gaussian equation (Eq. (9) of Main Paper) is exact; and (2) "Calibrated": the L2 norm of additive Gaussian noise is calibrated at the cost of additional computation of its norm $\epsilon^* = (\epsilon/\|\epsilon\|_2)\cdot\sigma\sqrt{d},\ \epsilon \sim \mathcal{N}(0,\sigma^2 I_d)$ so that ϵ^* is distributed exactly on a hypersphere centring the labeled image x with a radius of $\sigma\sqrt{d}$. Tab. 4 compares the calibration and classification results of these two settings ("Exact" and "Calibrated") with our SCVRM with approximated L2-distance $\sigma\sqrt{d} \approx \|\epsilon\|_2$. It shows that approximating the L2-distance produces results that are equivalent to those by computing the exact distance, or calibrating the L2-norm of additive Gaussian noise

8.2. Radial Basis Function - Additional Options

We investigate additional RBF functions, termed as (i) "Logistic" (Eq. (23)); (ii) "Hyperbolic" (Eq. (24)); and (iii) "Linear" (Eq. (25)), which are defined as:

$$\varphi_{Log}(\sigma\sqrt{d},\eta) = \frac{2}{1+\exp^{-(\sigma\sqrt{d}/\eta)}} - 1. \qquad (23)$$

$$\varphi_{H}(\sigma\sqrt{d},\eta) = \frac{\exp^{(\sigma\sqrt{d}/\eta)} - \exp^{-(\sigma\sqrt{d}/\eta)}}{\exp^{(\sigma\sqrt{d}/\eta)} + \exp^{-(\sigma\sqrt{d}/\eta)}}. \qquad (24)$$

$$\varphi_{Lin}(\sigma\sqrt{d}) = \frac{\sigma\sqrt{d}}{\gamma\eta}, \qquad (25)$$

where γ is the upper bound of the uniform distribution followed by the standard deviation $\sigma \sim \mathcal{U}[0,\gamma)$.

Both of "Logistic" function and "Hyperbolic" function differ from the "Gaussian" function (Eq. 9 of Main Paper) only in terms of the sensitivity to the change of σ value, while having the same support $[0,\infty)$ and the same range $[1,0)$. The "Linear" function is more different from the rest of the investigated options as it has a narrower support of $[0,\gamma\eta]$ and a slightly different range of $[1,0]$. Intuitively, "Linear" function sets a hard boundary condition where the label is smoothed to exactly a uniform categorical distribution, instead of being infinitely close to.

Following the default hyperparameter settings of $\eta = \sqrt{d}$, $\gamma = 2.0$ and $M = 3$ and default training details specified in Sec. 5.1, we show in Tab. 5 that these RBFs can achieve competitive performances against the Gaussian RBF in terms of both the model calibration degree and the dense classification accuracy.

8.3. Vicinal Risk Minimisation - Implementation Details

In practical implementation, the intractable VRM defined in Eq. 4 is approximated with the Monte-Carlo (MC) Sampling. Similar to the practical model of SCVRM (Sec. 4.3), VRM is also formulated as a data augmentation technique based on the labeled dataset $\mathcal{D} = \{(x_1, y_i)\}_{i=1}^{N}$. Its augmented dataset can be defined as:

$$\begin{aligned}
\mathcal{D}_v = \bigcup_{i=1}^{N}\Big\{(\tilde{x}_i^j,\tilde{y}_i^j)\,|\,\tilde{x}_i^j = x_i + \epsilon_i^j,\ \tilde{y}_i^j = y_i,& \qquad (26)\\
\epsilon_i^j \sim_{i.i.d}\mathcal{N}(0,\sigma^2 I_d)\Big\}_{j=1}^{M}&,
\end{aligned}$$

where σ is a fixed hyperparameter, and the vicinal images are paired with the groundtruth labels. This implementation, where we empirically set $\sigma = 2.0$, is termed as VRM.

We provide a different version of VRM with Gaussian kernel by introducing our design of letting the standard deviation be a random variable following a uniform distribution $\sigma \sim \mathcal{U}(0,\gamma]$. The corresponding augmented dataset can then be defined as:

$$\begin{aligned}
\mathcal{D}_{v^*} = \bigcup_{i=1}^{N}\Big\{(\tilde{x}_i^j,\,\tilde{y}_i^j)\,|\,\tilde{x}_i^j = x_i + \epsilon_i^j,\ \sigma_{i,j} \sim_{i.i.d}\mathcal{U}(0,\gamma],&\\
\tilde{y}_i^j = y_i,\ \epsilon_i^j \sim_{i.i.d}\mathcal{N}(0,\sigma_{i,j}^2 I_d)\Big\}_{j=1}^{M}&. \qquad (27)
\end{aligned}$$

Empirically, we set $\sigma = 2.0$ following the default hyperparameter setting of our SCVRM as in Sec. 5.1, and term this implementation as VRM*.

Following the default setting of $\eta = \sqrt{d}$ and $M = 3$ and training details specified in Sec. 5.1, these models are trained on both the labeled and the augmented datasets with

Table 4. Ablation study on approximating the L2-distance between the vicinal and labeled images with $\sigma\sqrt{d} \approx \|\epsilon\|_2$.

L2 distance	DUTS-TE [61]		DUT-OMRON [73]		PASCAL-S [32]		SOD [43]		ECSSD [71]		HKU-IS [30]	
	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓
Exact	0.85	**0.60**	**1.56**	**1.41**	1.92	1.777	4.68	4.30	0.42	**0.18**	0.96	**0.10**
Calibrated	0.88	0.72	1.88	1.73	2.10	1.92	4.47	4.21	**0.36**	0.31	**0.78**	0.12
Approximated (SCVRM)	**0.78**	0.61	1.64	1.49	**1.91**	**1.75**	**3.90**	**3.60**	0.44	0.19	**0.78**	**0.10**

L2 distance	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑
Exact	0.871	0.930	0.777	0.872	0.861	**0.906**	0.843	**0.872**	0.938	0.957	**0.930**	**0.961**
Calibrated	**0.872**	0.930	0.780	0.874	**0.862**	0.904	0.842	0.871	**0.940**	**0.958**	**0.930**	0.960
Approximated (SCVRM)	**0.872**	**0.932**	**0.786**	**0.880**	0.861	0.904	**0.845**	0.869	**0.940**	0.956	0.929	**0.961**

Table 5. Ablation study on the effect of various RBF functions.

Methods	Modules		DUTS-TE [61]		DUT-OMRON [73]		PASCAL-S [32]		SOD [43]		ECSSD [71]		HKU-IS [30]	
	SCVRM	RBF	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓
Baseline (B)	-	-	3.05	2.89	3.80	3.68	4.14	3.98	7.07	6.85	1.82	1.76	1.38	1.34
SCVRM	✓	Gaussian (φ_G)	**0.78**	0.61	1.64	1.49	1.91	1.75	**3.90**	**3.60**	**0.44**	0.19	0.78	0.10
	✓	Logistic (φ_{Log})	0.80	**0.53**	**1.47**	1.33	1.50	1.35	4.46	4.15	0.70	**0.17**	1.17	0.07
	✓	Hyperbolic (φ_H)	0.83	0.51	1.48	1.34	2.19	2.03	4.60	4.32	0.67	0.61	**0.58**	**0.09**
	✓	Linear (φ_{Lin})	0.92	0.70	1.86	**0.90**	**1.40**	**1.29**	4.73	4.47	0.74	**0.17**	1.34	0.02

a Binary Cross Entropy loss as defined in Eq. (14) (Main Paper). Following the implementation of SCVRM, the augmented dataset is re-sampled after each training epoch.

8.4. Example Vicinal Data Under Different Hyper-parameters γ and η.

The hyperparameter γ sets the upper-bound of the uniform distribution $\sigma \sim \mathcal{U}(0, \gamma]$ followed by the standard deviation of the isotropic Gaussian distribution $\mathcal{N}(0, \sigma^2 I_d)$. Given the bubbling effect of additive Gaussian noise [42, 66], γ effectively controls the radius of vicinity space centring each labeled image. Equivalently, it determines the noisiness of the vicinal images $\tilde{x} = x + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$ as illustrated in Fig. 7, which also shows their assigned soft labels computed via the Gaussian function (Eq. (9) with $\|\tilde{\epsilon}\|_2 = \sigma\sqrt{d}$) under various η values. We also plot the soft label value w.r.t the L2-distance between the vicinal and labeled images ($\|\epsilon\|_2 \approx \|\tilde{\epsilon}\|_2 = \sigma\sqrt{d}$) under various η values in Fig. 6.

8.5. Training and Inference Time

Both training and inference of SCVRM are conducted with a single RTX 3090 GPU. The training time increases proportionally with the number of augmented samples per labeled image M. With the default setting of $M = 3$, the training costs approximately 9.2 hours. The inference time is independent of the number of samples M, and averages 53.48 images per second. Despite the training time scaling with the number of sampling of augmented data M, the inference speed remains unchanged. VRM [3, 59], with a similar implementation (Supp. 8.3) to that of SCVRM (Sec. 4.3), has the same training and inference time. Baseline model (ERM) trained only on the labeled dataset has a shorter training time of 2.3 hours, but its inference speed is the same as that of SCVRM.

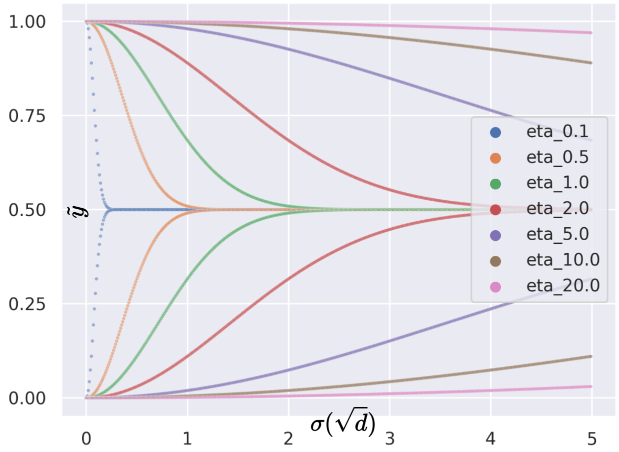

Figure 6. Plot of soft label with increasing σ values, computed with the Gaussian function (Eq. (9) with $\|\tilde{\epsilon}\|_2 = \sigma\sqrt{d}$) under various η values. The starting points of 1 and 0 (hard label) at $\sigma = 0$ correspond to the foreground and background categories in a (dense) binary classification task.

9. Evaluation Metrics

9.1. Model Calibration

Equal-Width Expected Calibration Error (ECE_{EW}) [16] can be defined as:

$$\text{ECE}_{\text{EW}} = \sum_{k=1}^{B} \frac{N_k}{N} |C_k - A_k|, \tag{28}$$

where $|\cdot|$ computes the absolute value, N is the total number of samples, B is the total number of bins where predictions are sorted into based on their confidences, N_k is the

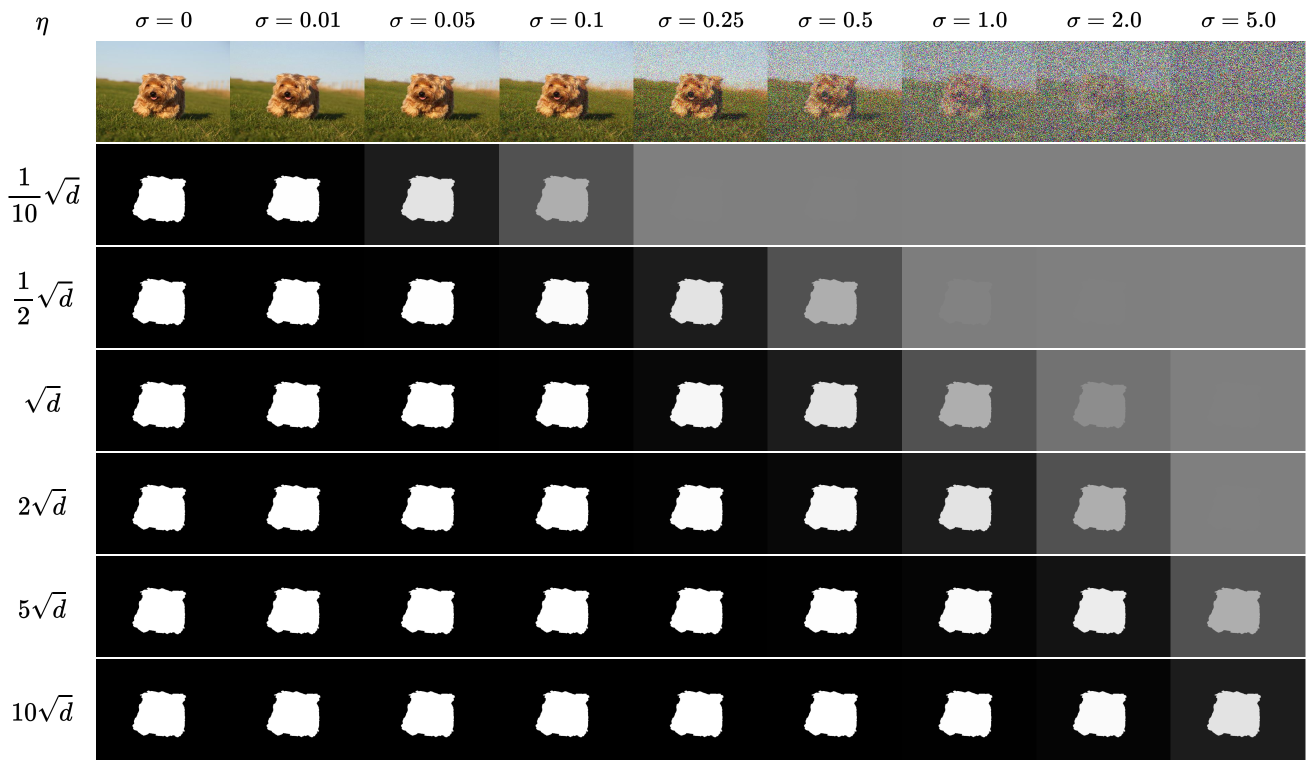

Figure 7. Example noisy images sampled under various standard deviations $\sigma = \{0.01, 0.05, 0.1, 0.25, 0.5, 1.0, 2.0, 5.0\}$ and their corresponding labels computed with the Gaussian function $\varphi_G(\cdot,\cdot)$ of different $\eta = \{i\sqrt{d}|i = 0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$.

number of predictions sorted into the k^{th} bin, C_k computes the mean confidence of predictions sorted into the k^{th} bin: $C_k = \frac{1}{N_k}\sum_{i=1}^{N}\mathbb{1}(c(\omega(x_i)) \in [\frac{k}{B}, \frac{k+1}{B})) \cdot c(\omega(x_i)),\ \forall k = 1,\dots,B$, with $[\frac{k}{B}, \frac{k+1}{B})$ being the range of the k^{th} bin, A_k computes the mean accuracy of predictions of the k^{th} bin: $A_k = \frac{1}{N_k}\sum_{i=1}^{N}\mathbb{1}(c(\omega(x_i)) \in [\frac{k}{B}, \frac{k+1}{B})) \cdot \mathbb{1}(f(x_i) = y_i),\ \forall k = 1,\dots,B, \omega(x) \in (0,1)$ is the Sigmoid-activated value before the classification $f(x) = \mathbb{1}(\omega(x) > 0.5)$, and $f(\cdot) : \mathcal{X} \to \mathcal{Y}$ is the projection function learned by a DNN.

Over-confidence Error (OE) can be defined as

$$\text{OE} = \sum_{k=1}^{B}\frac{N_i}{N}\mathbb{1}(C_k > A_k) \cdot |C_k - A_k|. \tag{29}$$

By choosing different constraints of respective ECE metrics, it can be transformed into Equal-Width Over-confidence Error OE_{EW}, Equal-Mass Over-confidence Error OE_{EM} and DEBIAS Over-confidence Error $\text{OE}_{\text{DEBIAS}}$ respectively.

9.2. Dense Classification

Prediction Accuracy $A(\cdot)$ of the model $f_\theta(\cdot)$ on a finite dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ can be defined as:

$$\mathbb{A}(\theta,\mathcal{D}) = \frac{1}{N \times H \times W}\sum_{i=1}^{N}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbb{1}(f(x_i)^{h,W} = y_i^{h,W}), \tag{30}$$

where i is the sample index, and h and w represent the spatial indices.

F-measure can be defined as:

$$F_\xi = \frac{(1+\xi^2) \times \text{Precision} \times \text{Recall}}{\xi^2 \times \text{Precision} + \text{Recall}}. \tag{31}$$

We follow previous SOD methods [37, 38, 68, 82] to adopt $\xi^2 = 3$. Further, the maximum F-measure result $F_{\max}$ is obtained via iterating over a set of binary thresholds of $\{0.01t\,|\,t = 1,\dots,99\}$, with the resultant classification being $f(x) = \mathbb{1}(\omega(x) > t)$.

Enhancement-alignment measure (E-measure) [11] is

Table 6. Impact of SCVRM on the dense classification accuracy.

Method	DUTS-TE [61]		DUT-OMRON [73]		PASCAL-S [32]		SOD [43]		ECSSD [71]		HKU-IS [30]	
	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑
ERM	0.869	0.928	0.769	0.867	0.855	0.902	0.836	0.871	0.936	0.956	0.925	0.958
VRM [1]	0.864	0.924	0.778	0.873	0.858	0.902	0.838	0.866	0.936	0.955	0.925	0.958
VRM* [1]	0.871	0.930	0.777	0.872	0.861	**0.906**	0.843	**0.872**	0.938	**0.957**	**0.930**	**0.961**
SCVRM	**0.872**	**0.932**	**0.786**	**0.880**	**0.861**	0.904	**0.845**	0.869	**0.940**	0.956	0.929	**0.961**

[1] VRM has fixed variance in accordance to the definition in Eq. (4).

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

Table 7. Model calibration and dense classification performances of VGG16 and Swin transformer backbones evaluated in terms of ECE_{EW} (%) and OE_{EW} (%) using $B = 10$ bins, and maximum F-measure and maximum E-measure respectively.

Methods	DUTS-TE [61]		DUT-OMRON [73]		PASCAL-S [32]		SOD [43]		ECSSD [71]		HKU-IS [30]	
	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓
Baseline (“VGG-B”)	3.46	3.23	4.12	3.92	4.40	4.17	7.87	7.60	2.02	1.91	1.51	1.44
Baseline (“Swin-B”)	2.41	2.23	3.29	3.15	3.35	3.19	6.23	6.05	1.02	0.97	0.87	0.82
VGG-SCVRM	**0.90**	**0.71**	**1.78**	**1.76**	**2.02**	**1.78**	**4.54**	**4.02**	**0.49**	**0.25**	**0.85**	**0.14**
Swin-SCVRM	**0.91**	**0.78**	**1.49**	**1.32**	**1.54**	**1.40**	**3.83**	**3.63**	**0.54**	**0.29**	**0.75**	**0.10**

Methods	DUTS-TE [61]		DUT-OMRON [73]		PASCAL-S [32]		SOD [43]		ECSSD [71]		HKU-IS [30]	
	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑
Baseline (“VGG-B”)	3.46	3.23	4.12	3.92	4.40	4.17	7.87	7.60	2.02	1.91	1.51	1.44
Baseline (“Swin-B”)	0.894	0.949	0.804	0.890	0.877	0.920	0.858	0.878	**0.948**	0.969	0.939	0.969
VGG-SCVRM	**0.90**	**0.71**	**1.78**	**1.76**	**2.02**	**1.78**	**4.54**	**4.02**	**0.49**	**0.25**	**0.85**	**0.14**
Swin-SCVRM	**0.898**	**0.954**	**0.811**	**0.896**	**0.880**	**0.922**	**0.861**	**0.883**	**0.948**	**0.970**	**0.942**	**0.971**

defined as:

$$
\begin{aligned}
E &= \frac{1}{N \times H \times W} \sum_{i=1}^{N} \sum_{i=1}^{H} \sum_{j=1}^{W} \phi(\omega(x_i))^{h,w}, \quad \text{where} \\
\phi(\omega(x)) &= \frac{1}{4}\Big(1 + \frac{2 \cdot \varphi(y) \circ \varphi(\omega(x))}{\varphi(y) \circ \varphi(y) + \varphi(\omega(x)) \circ \varphi(\omega(x))}\Big)^2, \\
\text{and} \quad \varphi(z) &= z - \mu_z \cdot J_{H,W},
\end{aligned} \tag{32}
$$

where H and W indicate the resolution of the groundtruth or dense classification map, $J_{H,W}$ is an H-by-W all-one matrix, μ_z stands for the mean value of the groundtruth or the dense classification map. The maximum E-measure result $E_{\max}$ is obtained via iterating over a set of binary thresholds that replace the mean value $\mu_z \in \{0.01t \,|\, t = 1, \ldots, 99\}$

10. More Ablation Studies

10.1. Effect of SCVRM on Classification Accuracy

Our proposed SCVRM not only improves the model calibration degree, but also the dense classification accuracy as shown in Tab. 6. It shows that our SCVRM outperforms ERM in terms of the maximum F-measure across the six SOD testing datasets, with the largest improvement of 0.017 on the DUT-OMRON dataset [73]. On the other hand, VRM with a fixed variance achieves little improvement over ERM. After introducing our design of $\sigma \sim \mathcal{U}(0, \gamma]$, VRM* achieves consistent improvements over the baseline (ERM) and performs comparably to our SCVRM.

11. Experiments with Different Backbones

Extra experiments with VGG16 [55] and Swin [39] transformer as backbones are conducted with default setting of: (i) $\gamma = 2.0$, (ii) $\eta = \sqrt{d}$ and (iii) “Gaussian” function $\varphi_G(\cdot, \cdot)$ (Eq. 9 of Main Paper), as specified in Sec. 5.1. Their performances in terms of model calibration degree and dense classification accuracy are presented in Tab. 7.

12. Experiments on Additional Dense Classification Tasks

12.1. Camouflaged Object Detection

The Camouflaged Object Detection (COD) model is trained on the COD10K training dataset [12], and evaluated on four testing datasets, including COD10K [12], NC4K [40], CHAMELEON [57] and CAMO [28]. The hyperparameters and training details adopt the default setting as specified in Sec. 5.1. The model calibration and dense classification performances of baseline and SCVRM models are presented in Tab. 8.

12.2. Smoke Detection

The Smoke Detection (SD) model is trained on the SMOKE5K [72] training set, and evaluated on the SMOKE5K testing dataset. The hyperparameters and training details adopt the default setting as specified in Sec. 5.1. The model calibration and dense classification per-

Table 8. Model calibration and dense classification performances of baseline and our SCVRM models in the camouflaged object detection task evaluated in terms of ECE_{EW} (%) and OE_{EW} (%) using $B = 10$ bins and maximum F-measure and maximum E-measure respectively.

Methods	COD10K [12]		NC4K [40]		CHAMELEON [57]		CAMO [28]	
	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓
Baseline (“COD-B”)	1.65	1.55	2.75	2.60	0.63	0.57	3.62	3.46
COD-SCVRM	**0.42**	**0.38**	**0.62**	**0.43**	**0.49**	**0.07**	**1.27**	**1.08**

Methods	COD10K [12]		NC4K [40]		CHAMELEON [57]		CAMO [28]	
	F_{max} ↑	E_{max} ↑	F_{max} ↑	E_{max} ↑	F_{max} ↑	E_{max} ↑	F_{max} ↑	E_{max} ↑
Baseline (“COD-B”)	0.710	0.882	0.800	0.901	0.837	0.935	0.745	0.851
COD-SCVRM	**0.720**	**0.888**	**0.805**	**0.905**	**0.844**	**0.936**	**0.760**	**0.865**

formances of baseline and SCVRM models are presented in .

Table 9. Model calibration and dense classification performances of the baseline and our SCVRM models in the smoke detection task, evaluated in terms of ECE_{EW} (%) and OE_{EW} (%) using $B = 10$ bins and maximum F-measure and maximum E-measure.

Methods	SMOKE5K [12]			
	ECE_{EW} ↓	OE_{EW} ↓	F_{max} ↑	E_{max} ↑
Baseline (“SD-B”)	0.180	0.170	0.763	0.930
SD-SCVRM	**0.066**	**0.062**	**0.772**	**0.936**

13. Experiment on Multi-Class Dense Classification Task (Semantic Segmentation)

The experiment on the multi-class dense classification task (semantic segmentation) is conducted on the PASCAL VOC 2012 segmentation dataset [10]. The dataset is comprised of a training, validation and testing set of size 1,464, 1,449, 1,456 respectively. Pixels of these images are categorised into 20 foreground classes and 1 background class. We follow [4, 33] to adopt the augmented training set with 10,582 training samples [17]. Without access to the groundtruth labels of the testing set, which are kept on the server and not released to the public, we treat the “official validation set” as “our testing set” to evaluate the model calibration and dense classification performances. We also divide the “official augmented training set” into “our training set” of 9,582 training samples for model optimisation and “our validation set” of 1,000 validation samples to tune the hyperparameters. The model calibration is evaluated in terms of Equal-Width Expected Calibration Error ECE_{EW} and Equal-Width Over-confidence Error OE_{EW} using $B = 10$ bins. Following [4, 33], the dense classification accuracy is evaluated in terms of Intersection-over-Union (IoU). The training involves a DeepLabv3+ [4] as the baseline and our proposed SCVRM as data augmentation technique. We set $\eta = \sqrt{d}/2$ and adopt the default settings for the rest of hyperparameters and training details as specified in Sec. 5.1. The results are presented in Tab. 10.

Table 10. Model calibration and dense classification performances of the baseline and our SCVRM models in the semantic segmentation task, evaluated in terms of ECE_{EW} (%) and OE_{EW} (%) using $B = 10$ bins and Intersection-over-Union (IoU).

Methods	PASCAL VOC 2012 [10]		
	ECE_{EW}(%) ↓	OE_{EW}(%) ↓	IoU (%) ↑
Baseline (“SS-B”)	6.29	5.37	71.2
SS-SCVRM	**3.51**	**2.97**	**71.5**

14. Generalisation to Existing SOD Methods

We generalise the proposed SCVRM data augmentation technique to several existing SOD models of different categories, including: energy-based conditional generative model, EBMGSOD [79], lightweight model, EDN [68], and attention-based model ICON [87]. The model calibration and dense classification performance are presented in Tab. 11.

14.1. How does soft vicinity work in SCVRM?

Our SCVRM approach intuitively requires that image vicinity implies label vicinity, i.e. for a vicinal image $\tilde{x}_0 \sim_{i.i.d} p(\tilde{x}_i|x_i)$, the corresponding $\tilde{y}_0$ is likely to be closer to y_i instead of other labels in the label space. This requirement can raise extra constraints on practical SCVRM implementations. We analyse a simpler case of $R_{SC-G}(f)$ defined in Eq. (7) where the standard deviation σ is a fixed parameter. We show that such requirement is highly likely to be satisfied in practice with a small σ, and setting $||\tilde{\epsilon}||_2 = \sigma\sqrt{d}$. Then, for $\tilde{x} = x_i + \epsilon$ and $\tilde{y} = g(y_i; \sigma\sqrt{d})$ we have:

$$\begin{cases} p(\tilde{x}|x,\sigma) = \dfrac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\dfrac{||\tilde{x}-x||_2^2}{2\sigma^2}\right) \\ p(\tilde{y}|y,\sigma) = \delta_{g(y;||\tilde{\epsilon}||_2)}(\tilde{y}) \\ \quad p(x,y) = \dfrac{1}{N} \cdot \delta_{x_i,y_i}(x,y), \end{cases} \tag{33}$$

Table 11. Model calibration performance of our proposed SCVRM applied on the existing SOD methods, evaluated in terms of $\mathrm{ECE}_{\mathrm{EW}}$ (%) and $\mathrm{OE}_{\mathrm{EW}}$ (%) using $B = 10$ bins.

Methods			DUTS-TE [61]		DUT-OMRON [73]		PASCAL-S [32]		SOD [43]		ECSSD [71]		HKU-IS [30]	
Name	Year	SCVRM	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓
EBMGSOD [79]	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 [87]	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
EDN [68]	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
EBMGSOD*	2021	✓	0.90	0.73	1.55	1.40	2.15	1.89	4.90	4.62	0.55	0.50	1.07	0.14
ICON*	2021	✓	0.75	0.51	1.39	1.25	1.89	1.72	4.48	4.19	0.36	0.21	1.17	0.11
EDN*	2022	✓	0.85	0.69	1.73	1.56	1.85	1.70	4.47	4.20	0.49	0.15	1.01	0.14

Methods			DUTS-TE [61]		DUT-OMRON [73]		PASCAL-S [32]		SOD [43]		ECSSD [71]		HKU-IS [30]	
Name	Year	SCVRM	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑
EBMGSOD [79]	2021	-	0.850	0.927	0.762	0.867	0.830	0.896	0.836	0.871	0.914	0.944	0.906	0.952
ICON [87]	2021	-	0.860	0.924	0.773	0.876	0.850	0.899	0.815	0.854	0.933	0.954	0.919	0.953
EDN [68]	2022	-	0.893	0.949	0.821	0.900	0.879	0.920	0.840	0.860	0.950	0.969	0.940	0.970
EBMGSOD*	2021	✓	0.854	0.930	0.771	0.872	0.833	0.902	0.841	0.873	0.916	0.950	0.909	0.950
ICON*	2021	✓	0.863	0.928	0.779	0.881	0.852	0.902	0.823	0.859	0.936	0.957	0.922	0.958
EDN*	2022	✓	0.894	0.949	0.829	0.905	0.884	0.925	0.843	0.867	0.952	0.970	0.943	0.971

[1] * indicates the model is trained with our proposed SCVRM.

where $\mathbb{1}(\cdot)$ is an indicator function and $p(x, y)$ only considers the finite available data $(x_i, y_i) \in \mathcal{D}$, and $\delta_{g(y;\|\tilde{\epsilon}\|_2)}(\tilde{y})$ is a delta distribution of $\tilde{y}$ which has non-zero probability density only at $g(y; \|\tilde{\epsilon}\|_2)$. Given Eq. (33), we would like to analyse $p(\tilde{y}|\tilde{x}, \sigma)$, the conditional distribution of vicinal label $\tilde{y}$ given vicinal image $\tilde{x}$ and parameter σ. For $(\tilde{x}_0, \cdot) \sim_{i.i.d} p(\tilde{x}, \tilde{y})$, if we denote

$$\begin{cases} (x_0^{\mathrm{m1}}, y_0^{\mathrm{m1}}) = \underset{(x,y)\in\mathcal{D}}{\arg\min} ||x - \tilde{x}_0||_2 \\ (x_0^{\mathrm{m2}}, y_0^{\mathrm{m2}}) = \underset{(x,y)\in\mathcal{D}\backslash\{(x_0^{\mathrm{m1}}, y_0^{\mathrm{m1}})\}}{\arg\min} ||x - \tilde{x}_0||_2, \end{cases} \tag{34}$$

Then we present the proposition as:

Proposition 1. *Let $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$ that satisfies for $\forall i \neq j$ we have $x_i \neq x_j$ and $y_i \neq y_j$. Then for $(\tilde{x}_0, \cdot) \sim_{i.i.d} p(\tilde{x}, \tilde{y})$, we have:*

$$p(\tilde{y} = \tilde{y}_0^{m1} \,|\, \tilde{x}_0, \sigma) \geq \frac{1}{1 + (N-1)e^{-\Delta}}, \tag{35}$$

where $\Delta = \left(\|\tilde{x}_0 - x_0^{m2}\|_2^2 - \|\tilde{x}_0 - x_0^{m1}\|_2^2\right)/(2\sigma^2) \geq 0$ and $\tilde{y}_0^{m1} = g(y_0^{m1}; \|\tilde{\epsilon}\|_2)$.

The proof for the proposition is:

Proof. Based on Eq. (33), we can rewrite $p(\tilde{y}|\tilde{x}, \sigma)$ as :

$$\begin{aligned} p(\tilde{y}_0|\tilde{x}_0, \sigma) &= \int_{x,y} p(\tilde{y}|y, \sigma) \cdot p(y|x, \sigma) \cdot p(x|\tilde{x}_0, \sigma) dx dy \\ &= \int_{x,y} p(\tilde{y}|y, \sigma) \cdot p(y|x) \cdot \frac{p(\tilde{x}_0|x, \sigma)p(x)}{p(\tilde{x}_0)} dx dy \\ &= \sum_{i=1}^N \mathbb{1}(\tilde{y}_i = g(y_i; \sigma\sqrt{d})) \cdot \frac{1}{N} \cdot \frac{p(\tilde{x}_{i,0}|x_i, \sigma)}{p(\tilde{x}_{i,0})}. \end{aligned} \tag{36}$$

Because ground truth label space $\mathcal{Y}$ is a discrete set and $g(\cdot, \sigma\sqrt{d})$ is deterministic, the augmented label space $\tilde{Y}(\sigma) = \{g(y_i; \sigma^2\sqrt{d})\}_{i=1}^N$ with $(x_i, y_i) \in \mathcal{D}$ is also a discrete set. For $\forall \tilde{y}_i \in \tilde{Y}(\sigma)$ we have:

$$p(\tilde{y}_0 = \tilde{y}_{0,i}|\tilde{x}_0, \sigma) = \frac{p(\tilde{x}_{i,0}|x_i, \sigma)}{p(\tilde{x}_{i,0}) \cdot N} = \frac{p(\tilde{x}_{i,0}|x_i, \sigma)}{\sum_{l=1}^N p(\tilde{x}_{i,0}|x_l, \sigma)}. \tag{37}$$

Based on Eq. (37), for $\forall \tilde{y}^* \in \tilde{\mathcal{Y}}(\sigma)\backslash\{\tilde{y}_0^{m1}\}$, where $\tilde{y}_0^{m1} = g(y_0^{m1}; \sigma\sqrt{d})$, and corresponding $(x^*, y^*) \in \mathcal{D}\backslash\{(x_0^{\mathrm{m1}}, y_0^{\mathrm{m1}})\}$ we have:

$$\begin{aligned} &\frac{p(\tilde{y} = \tilde{y}^*|\tilde{x}_0, \sigma)}{p(\tilde{y} = \tilde{y}_0^{\mathrm{m1}}|\tilde{x}_0, \sigma)} \\ &= \exp\left(\frac{1}{2\sigma^2}(\|\tilde{x}_0 - x_0^{\mathrm{m1}}\|_2^2 - \|\tilde{x}_0 - x^*\|_2^2)\right) \leq e^{-\Delta} \end{aligned} \tag{38}$$

Therefore we have:

$$
\begin{aligned}
p(\tilde{y} = \tilde{y}^{\mathrm{m1}} | \tilde{x}_0, \sigma) = & \frac{p(\tilde{x}_0 | x^{\mathrm{m1}}, \sigma)}{\sum_{l=1}^{N} p(\tilde{x}_0 | x_l, \sigma)} \\
\geq & \frac{p(\tilde{x}_0 | x^{\mathrm{m1}}, \sigma)}{p(\tilde{x}_0 | x^{\mathrm{m1}}, \sigma) + (N-1) e^{-\Delta} \cdot p(\tilde{x}_0 | x^{\mathrm{m1}}, \sigma)} \\
= & \frac{1}{1 + (N-1) e^{-\Delta}}
\end{aligned}
\tag{39}
$$

□

Remark 2. *The Prop. 1 presents a probabilistic bound on the vicinal image adopting the groundtruth label of the nearest labeled image adjusted by the $g(\cdot)$ function defined in Eq. 5, which is conditioned on the dataset size N, the standard deviation σ, and the difference in L2-distance from the vicinal image to the nearest labeled image and the second nearest labeled image.*

We empirically analyse the probabilistic bound of Prop. 1 on the training dataset, DUTS-TR [61] with the size $N = |\mathcal{D}| = 10,553$ and image space dimension of $d = 3 \times 384^2$. We sample 100 vicinal images from the vicinal distribution of each labeled image under various standard deviations: $\sigma = \{0.01, 0.05, 0.1, 0.5, 1.0, 2.0, 3.0, 5.0, 10.0\}$. As plotted in Fig. 8, the mean Δ value is at least 14.74 when the variance is no larger than 5.0, resulting in $p(\tilde{y} = g(\tilde{y}_0^{\mathrm{m1}}, \sigma) \,|\, \tilde{x}_0, \sigma = 5.0) \geq 99.6\%$. Note that vicinal images sampled under $\sigma^2 = 5$ are overwhelmed by Gaussian noise, with their associated augmented labels being smoothed to approximately uniform categorical distributions as shown in Fig. 2. In this case, the groundtruth label of the nearest labeled image becomes almost irrelevant.

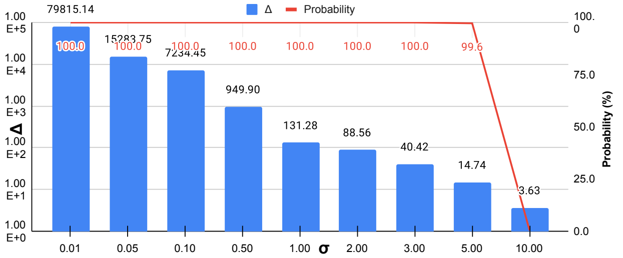

Figure 8. Plot of mean Δ (left vertical axis) and $1/(1 + (N+1)\exp(-\Delta))$ (right vertical axis) of Prop. 1 computed for additive Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$ under $\sigma \in \{0.01, 0.05, 0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$, $N = 10,553$ and $d = 3 \times 384^2$ on the DUTS-TR [61] training dataset.

15. More Joint Distribution Plots on SOD Testing Datasets

Fig. 9 plots the joint distribution of prediction confidence and accuracy of existing model calibration methods and our proposed SCVRM on the six SOD testing datasets.

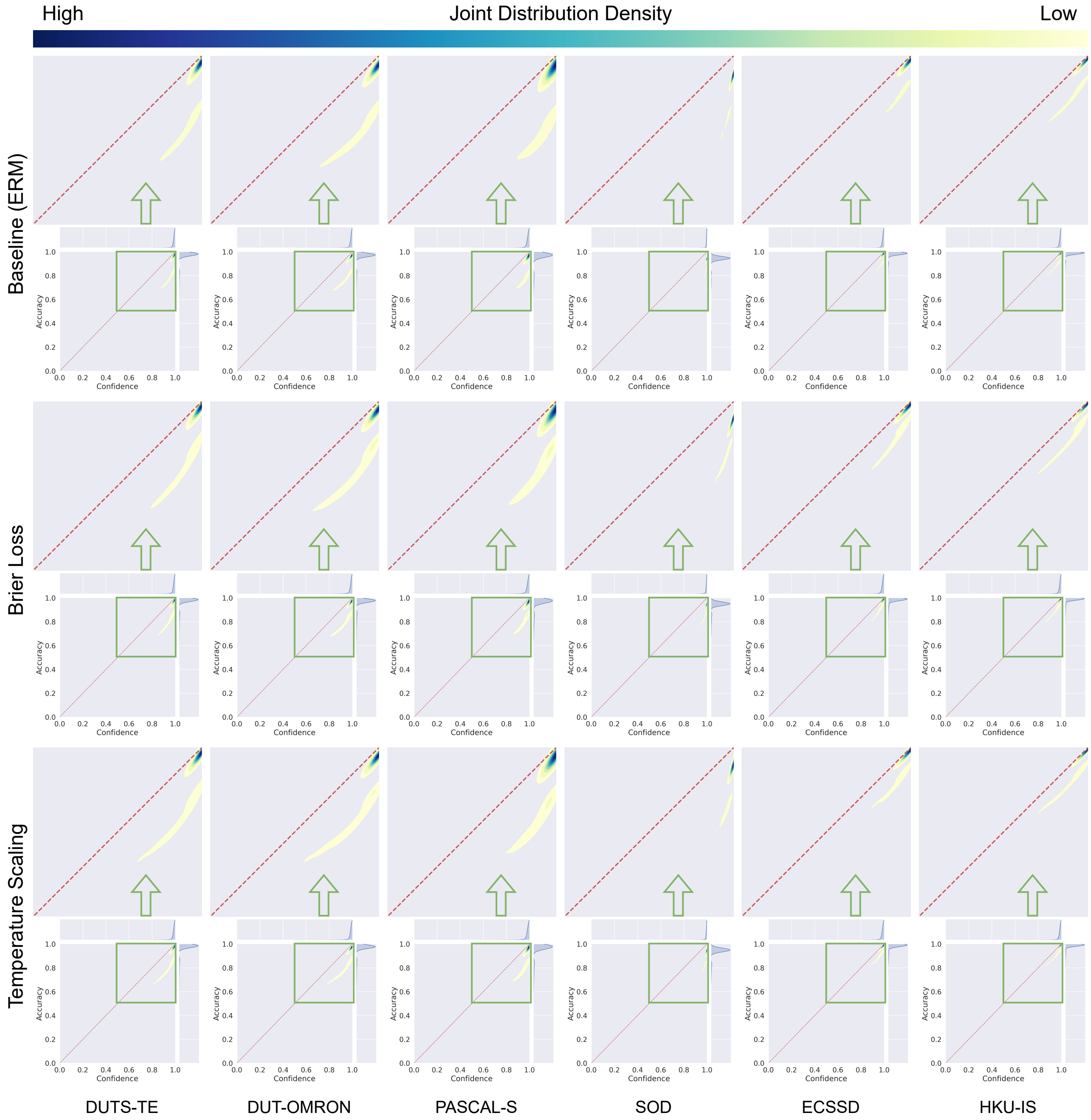

Figure 9. Joint distribution of prediction confidence (horizontal axis) and prediction accuracy (vertical axis) on the six SOD testing datasets. The dashed red diagonal line represents the perfectly calibrated oracle model. Compared to the existing methods, our proposed SCVRM further reduces the areas of distribution that deviate from the oracle.

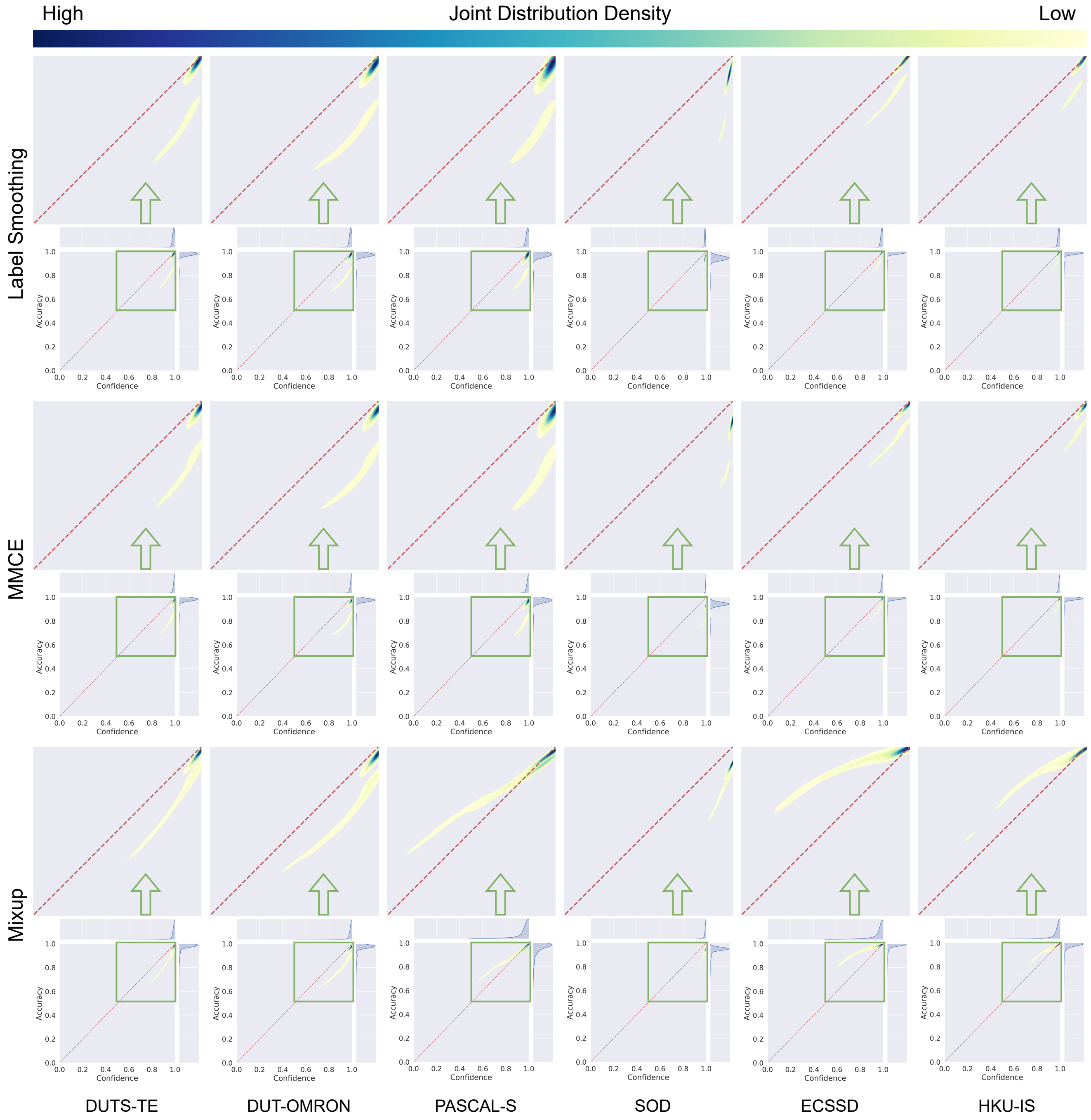

Figure 9. Joint distribution of prediction confidence (horizontal axis) and prediction accuracy (vertical axis) on the six SOD testing datasets. The dashed red diagonal line represents the perfectly calibrated oracle model. Compared to the existing methods, our proposed SCVRM further reduces the areas of distribution that deviate from the oracle.

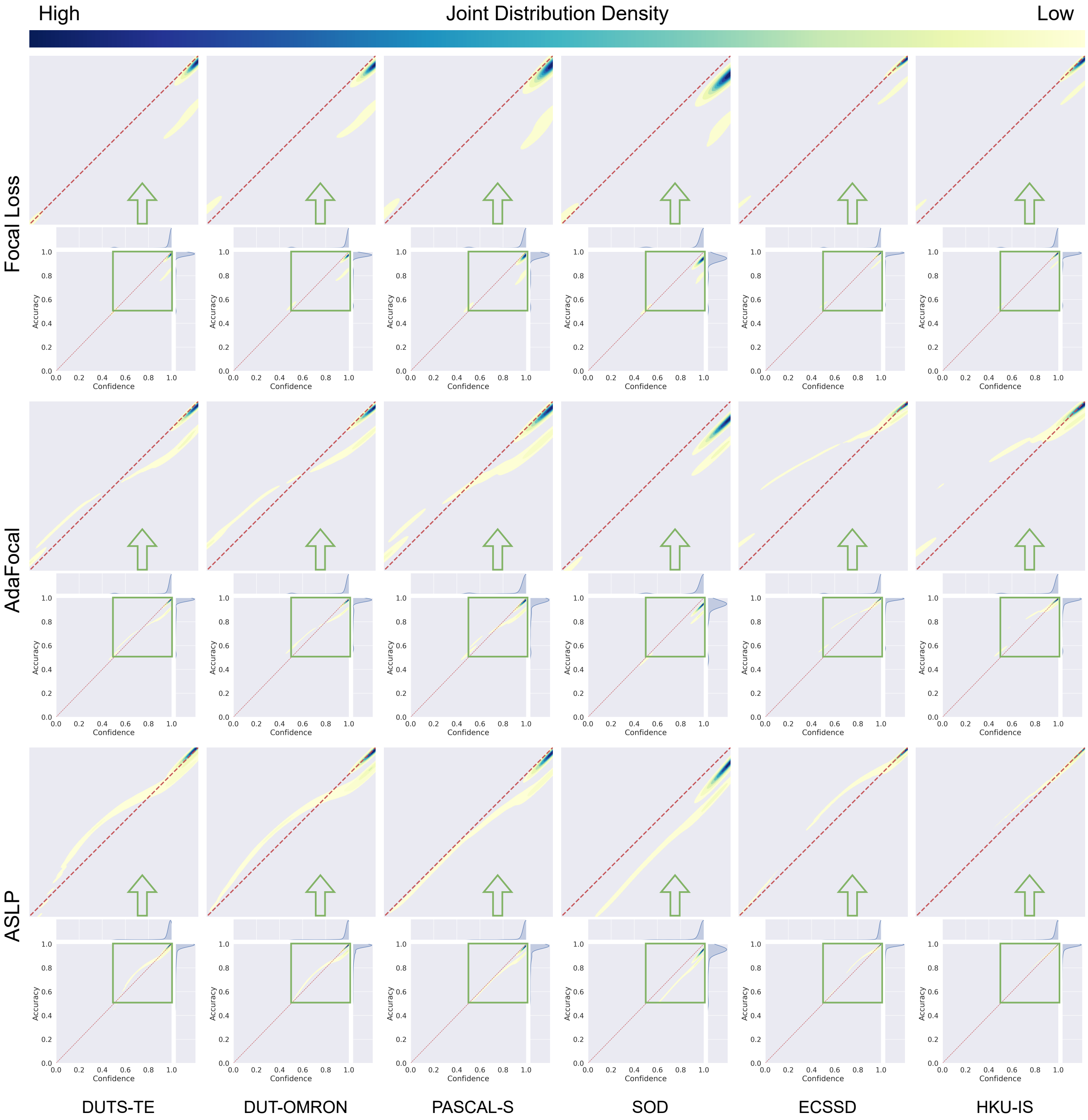

Figure 9. Joint distribution of prediction confidence (horizontal axis) and prediction accuracy (vertical axis) on the six SOD testing datasets. The dashed red diagonal line represents the perfectly calibrated oracle model. Compared to the existing methods, our proposed SCVRM further reduces the areas of distribution that deviate from the oracle.

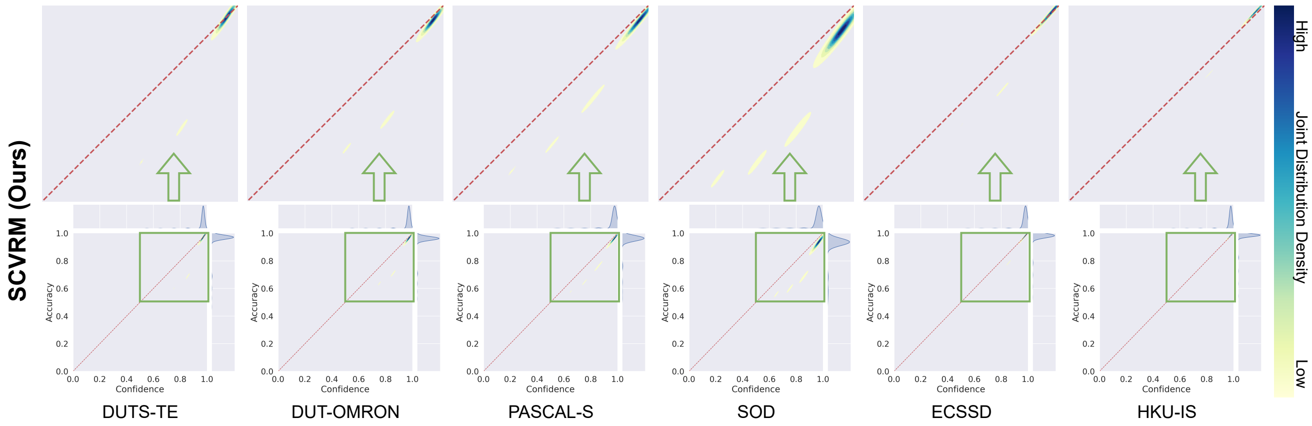

Figure 9. Joint distribution of prediction confidence (horizontal axis) and prediction accuracy (vertical axis) on the six SOD testing datasets. The dashed red diagonal line represents the perfectly calibrated oracle model. Compared to the existing methods, our proposed SCVRM further reduces the areas of distribution that deviate from the oracle.