Contents

A. Derivation

A.1. Self-Calibration Binary Cross Entropy (SC-BCE) Loss

We show that our SC-BCE loss is close to label smoothing in binary classification. Label smoothing, as defined in Eq. (7), is a typical data augmentation that softens the training supervision signals [46, 41, 77, 84].

$$S(Y,\sigma) = \text{LS}(Y,\sigma) = (1-\sigma)Y + \frac{\sigma}{K}, \quad \forall y \in Y. \tag{7}$$

where σ is the label smoothing strength hyperparameter and K is the number of classes, thus is set to $K=2$ for a binary task. For image label pairs $X, Y \sim P$, the BCE loss with label smoothing takes the form:

$$\mathcal{L}_{\text{BCE}}(\theta, X, S(Y,\sigma)) = \mathbb{E}_{x,y\in X,Y}\left[-\Big((1-\sigma)y + \frac{\sigma}{2}\Big)\log f_\theta(x) - \Big(1-\big((1-\sigma)y+\frac{\sigma}{2}\big)\Big)\log(1-f_\theta(x))\right]. \tag{8}$$

On the other hand, our proposed SC-BCE loss, taking expectation over the Bernoulli variable $Z_t(x,y)$, can be written as:

$$\begin{aligned}
\mathbb{E}_{Z_t}\Big[\mathcal{L}_{\text{SC-BCE}}(\theta, X, Y, \alpha, \beta)\Big] =& \mathbb{E}_{Z_t}\Big[(1-Z_t)\mathcal{L}_{\text{BCE}}(X,Y;\theta) + Z_t\mathcal{L}_{\text{BCE}}(X, P(Y,\beta),\theta)\Big] \\
=& (1-\alpha)\mathcal{L}_{\text{BCE}}(X,Y;\theta) + \alpha\mathcal{L}_{\text{BCE}}(X,P(Y,\beta),\theta) \\
=& \mathbb{E}_{x,y\in X,Y}\Big[-\big((1-\alpha)y+\alpha p\big)\log f_\theta(x) - \big(1-(1-\alpha)y-\alpha p\big)\log(1-f_\theta(x))\Big]
\end{aligned} \tag{9}$$

Substitute: $p(Y,\beta) = (1-\beta)\cdot y + \frac{\beta}{2}$, then we have:

$$\begin{aligned}
&\mathbb{E}_{x,y\in X,Y}\Big[-\big((1-\alpha)y+\alpha p\big)\log f_\theta(x) - \big(1-(1-\alpha)y-\alpha p\big)\log(1-f_\theta(x))\Big] \\
=&\mathbb{E}_{x,y\in X,Y}\left[-\Big((1-\alpha)y + \alpha\big((1-\beta)y+\frac{\beta}{2}\big)\Big)\log f_\theta(x) - \left(1-\Big((1-\alpha)y+\alpha\big((1-\beta)y+\frac{\beta}{2}\big)\Big)\right)\log(1-f_\theta(x))\right] \\
=&\mathbb{E}_{x,y\in X,Y}\left[-\Big((1-\alpha\beta)y+\frac{\alpha\beta}{2}\Big)\log f_\theta(x) - \Big(1-\big((1-\alpha\beta)y+\frac{\alpha\beta}{2}\big)\Big)\log(1-f_\theta(x))\right] \\
=&\mathcal{L}_{\text{bce}}(\theta, X, S(Y,\alpha\beta)),
\end{aligned} \tag{10}$$

where we let $\alpha\beta = \sigma$ to show that the expectation of SC-BCE loss over with a stochastically perturbed label over a Bernoulli variable is equivalent to a BCE loss with a smoothed label.

A.2. Connection between SC-BCE and Maximum Entropy Inference

We prove that the SC-BCE loss maximises prediction entropy as well as minimising cross entropy between the prediction distribution and groundtruth distribution. Given the SC-BCE loss written as:

$$\begin{aligned}
\mathcal{L}_{\text{SC-BCE}}(\theta, X, Y, \alpha, \beta) =& (1-Z_t)\mathcal{L}_{\text{BCE}}(\theta, X, Y) + Z_t\mathcal{L}_{\text{BCE}}(\theta, X, P(Y,\beta)) \\
=& (1-Z_t)\mathcal{L}_{\text{BCE}}(\theta, X, Y) + Z_t\left[(1-\frac{\beta}{2})\mathcal{L}_{\text{BCE}}(\theta, X, Y) + \frac{\beta}{2}\mathcal{L}_{\text{BCE}}(\theta, X, P(Y,2))\right] \\
=& (1-\beta Z_t)\mathcal{L}_{\text{BCE}}(\theta, X, Y) + \frac{\beta Z_t}{2}\big[\mathcal{L}_{\text{BCE}}(\theta, X, P(Y,2)) + \mathcal{L}_{\text{BCE}}(\theta, X, Y)\big]
\end{aligned} \tag{11}$$

where the first term includes a regular BCE loss $\mathcal{L}_{\text{BCE}}(\theta, X, Y)$ with random weight $1-\beta Z_t$ and $P(Y,2)$ represents an inverted label. Aside from the coefficient $Z\beta/2$, the second term can be expanded as a simpler form without label Y by collecting the Y terms:

$$\begin{aligned}
\mathcal{L}_{\text{BCE}}(\theta, X, P(Y,2)) + \mathcal{L}_{\text{BCE}}(\theta, X, Y) =& -\mathbb{E}_{x,y\in X,Y}\Big[(1-y)\log f_\theta(x) + y\log(1-f_\theta(x))\Big] \\
& -\mathbb{E}_{x,y\in X,Y}\Big[y\log f_\theta(x) + (1-y)\log(1-f_\theta(x))\Big] \\
=& -\mathbb{E}_{x\in X}\Big[\log f_\theta(x) + \log(1-f_\theta(x))\Big] \\
=& 2\cdot\mathbb{E}_{x\in X}\left[-\frac{1}{2}\log f_\theta(X) - \frac{1}{2}\log(1-f_\theta(X))\right] \\
=& 2\cdot\mathcal{L}_{\text{BCE}}(\theta, X, U)
\end{aligned} \tag{12}$$

where U is a uniform binary categorical distribution. Substituting Eq. (12) into Eq. 11 yields:

$$\mathcal{L}_{\text{SC-BCE}}(\theta, X, Y, \alpha, \beta) = (1-\beta Z_t)\cdot\mathcal{L}_{\text{BCE}}(\theta, X, Y) + \beta Z_t\cdot\mathcal{L}_{\text{BCE}}(\theta, X, U) \tag{13}$$

A.3. Derivation of Grad-α

We start with the SC-BCE loss with sample-wise Bernoulli variable on a finite training dataset $\mathcal{D}_{\text{TR}} = \{x_i, y_i\}_{i=1}^N$ as:

$$\mathcal{L}_{\text{SC-BCE}}(\theta, X, Y, \alpha, \beta) = \sum_{i=1}^{N} (1 - Z_t(x_i, y_i)) \cdot \mathcal{L}_{\text{BCE}}(\theta, x_i, y_i) + Z_t(x_i, y_i) \cdot \mathcal{L}_{\text{BCE}}(\theta, x_i, p(y_i, \beta)). \tag{14}$$

where the variable is drawn from sample-specific Bernoulli distributions: $Z_t(x_i, y_i) \sim B(1, \alpha_i)$, $i = 1, \ldots, N$. Further, we take expectation over the Bernoulli variable for each individual training sample to recover:

$$\begin{aligned} &\sum_{i=1}^{N} \mathbb{E}_{Z_t(x_i, y_i)} \Big[(1 - Z_t(x_i, y_i)) \cdot \mathcal{L}_{\text{BCE}}(\theta, x_i, y_i) + Z_t(x_i, y_i) \cdot \mathcal{L}_{\text{BCE}}(\theta, x_i, p(y_i, \beta)) \Big] \\ =&\sum_{i=1}^{N} (1 - \alpha_i) \cdot \mathcal{L}_{\text{BCE}}(\theta, x_i, y_i) + \alpha_i \cdot \mathcal{L}_{\text{BCE}}(\theta, x_i, p(y_i, \beta)). \end{aligned} \tag{15}$$

We further differentiate the above equation w.r.t. sample-specific label perturbation probability α_i, $i = 1, \ldots, N$ to obtain:

$$\begin{aligned} \frac{\partial \sum_{i=1}^{N} (1 - \alpha_i) \cdot \mathcal{L}_{\text{BCE}}(\theta, x_i, y_i) + \alpha_i \cdot \mathcal{L}_{\text{BCE}}(\theta, x_i, p(y_i, \beta))}{\partial \alpha_i} &= -\mathcal{L}_{\text{BCE}}(\theta, x_i, y_i) + \mathcal{L}_{\text{BCE}}(\theta, x_i, p(y_i, \beta)), \\ &\text{for} \quad i = 1, \ldots, N, \end{aligned} \tag{16}$$

Performing gradient descent according to this gradient will lead to an optimal value for α with the regularization term. We find Eq. 16 (Unnormalised ∇_{α_i}) favours perturbation methods with higher perturbation strength β, leading them to to converge faster. This is because label perturbation techniques with higher strengths, β, by definition have lower label perturbation probabilities, α, overall to achieve optimal model calibration degrees whereas unnormalised Grad-α agnostic to label perturbation strength. As illustrated in Fig. 5, with unnormalised Grad-α, Hard Inversion (HI) with the largest perturbation strength $\beta = 2$ converges with only 5 epochs of ASLP training whereas it takes Moderation (M) and Dynamic Moderation (DM) with moderate perturbation strength ($\beta = 1$) around 11 epochs to converge.

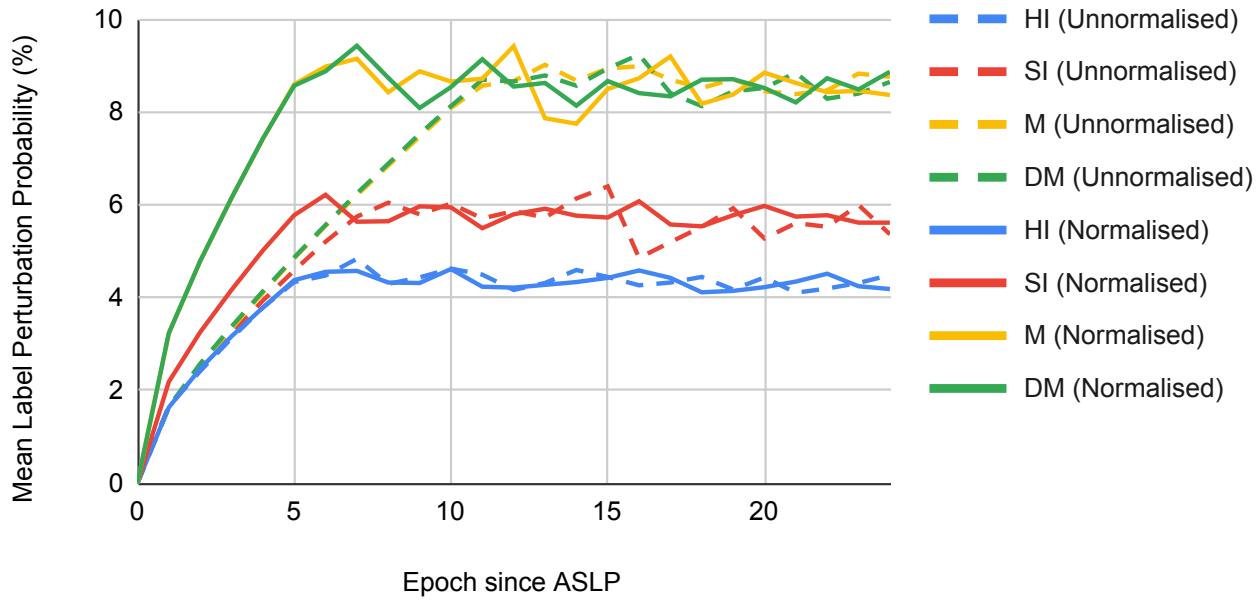


Figure 5: Convergence speed of unnormalised (dashed line) and normalised (solid line) Grad-α with different perturbation strengths: (1) HI: $\beta = 2$, (2) SI: $\beta = 1.5$, (3) M: $\beta = 1$, (4) DM: $\beta = 1$.

We propose a normalised version that allows ASLP under different perturbation strengths $\beta \in (0, 2]$ to converge equally fast. The unnormalised version (Eq. 16) is divided by $\beta/2$ and the normalised ∇_{α_i} is as:

$$\nabla_{\alpha_i} = \frac{2 \cdot \Big(\mathcal{L}_{\text{BCE}}(\theta, x_i, y_i) + \mathcal{L}_{\text{BCE}}(\theta, x_i, p(y_i, \beta)) \Big)}{\beta}, \qquad i = 1, \ldots, N \tag{17}$$

Fig. 5 illustrates that ASLP with different perturbation strengths with normalised ∇_{α_i} can converge equally fast.

A.4. Confidence of the Expectation of Stochastically Perturbed Label

We define the expectation of the stochastically perturbed label as:

$$\mathbb{E}_{Z_t}\Big[(1-Z_t)\cdot Y + Z_t \cdot P(Y,\beta)\Big] = (1-\alpha\beta)\cdot Y + \frac{\alpha\beta}{2}, \tag{18}$$

where we require $\beta \in [0,2]$ and $\alpha \in [0, \frac{1}{\beta})$. The resultant product is $\alpha\beta \in [0,1)$. The expected confidence of perturbed label is:

$$\begin{aligned} C\left(\mathbb{E}_{Z_t}\Big[(1-Z_t)\cdot Y + Z_t \cdot P(Y,\beta)\Big]\right) =& \Big|(1-\alpha\beta)\cdot Y + \frac{\alpha\beta}{2} - 0.5\Big| + 0.5 \\ =& 1 - \frac{\alpha\beta}{2},\ \forall Y = \{0,1\} \end{aligned} \tag{19}$$

A.5. Adaptive Label Smoothing (ALS)

Adaptive Label Smoothing (ASL) applies Label Smoothing with per-image label perturbation strength ($\alpha = 1$ and $\{\beta_i\}_{i=1}^N$). Similar to the derivation of ∇_{α_i}, we differentiate Eq. (15) w.r.t. image-specific label perturbation strength as:

$$\begin{aligned} \nabla_{\beta_i} =& \frac{\partial \sum_{i=1}^N (1-1\cdot\beta_i)\cdot \mathcal{L}_{\text{BCE}}(\theta, x_i, y_i) + 1\cdot\beta_i \cdot \mathcal{L}_{\text{BCE}}(\theta, x_i, p(y_i,\beta_i))}{\partial \beta_i} \\ =& -\mathcal{L}_{\text{BCE}}(\theta, x_i, y_i) + \mathcal{L}_{\text{BCE}}(\theta, x_i, p(y_i,\beta_i)), \qquad \text{for} \quad i = 1,\dots,N, \end{aligned} \tag{20}$$

The updating rule (ALS_{MC}) that incorporates adaptive label smoothing to maximise model calibration is formulated as:

$$\begin{aligned} \beta_i^{n+1} = \beta_i^n + \eta \cdot \Bigg(& \underbrace{\mathcal{L}_{\text{BCE}}(\theta, x_i, p(y_i,\beta_i)) - \mathcal{L}_{\text{BCE}}(\theta, x_i, y_i)}_{\nabla_{\beta_i}} + \lambda \cdot \underbrace{\min\Big((1-\frac{1\cdot\beta_i}{2}) - \mathbb{A}(\theta_{lm}, \mathcal{D}_{\text{VAL}}), 0\Big)}_{\text{Reg}_{\text{C}}} \Bigg) \\ & \text{for} \quad i = 1,\dots,N, \end{aligned} \tag{21}$$

B. Implementations

B.1. Model

Our model adopts a simple U-Net [57] structure consisting of an encoder and a decoder. Feature maps $\{F_i \in i \cdot C \times \frac{H}{i \cdot 8} \times \frac{W}{i \cdot 8}\}_{i=1}^{4}$ are extracted by the encoder, where $C = 256$ and i indexes from low level to high level with an increasing value.

The model outputs pixel-wise logits $\sigma(x_i) \in (-\infty, \infty)^{1 \times H \times W},\ i = 1, \ldots, N$ where N is the total number of samples, which is further processed with a Sigmoid function to produce the prediction probability as:

$$f_\theta(x_i) = \text{Sigmoid}(\sigma(x_i)) = \frac{1}{1 + e^{-\sigma(x_i)}}, \quad i = 1, \ldots, N. \tag{22}$$

The prediction probability after the Sigmoid function is in the range $f_\theta(x) \in (0, 1)^{1 \times H \times W}$. The predicted label is "foreground" (Labeled as "1") if the prediction probability is larger than 0.5 and is "background" (labeled as "0") otherwise as:

$$\hat{y}_i = \mathbb{1}(f_\theta(x_i) > 0.5), \quad i = 1, \ldots, N. \tag{23}$$

The probability of predicted label $\hat{y}$, also known as the winning class, is:

$$P_{\hat{y}_i} = |f_\theta(x_i) - 0.5| + 0.5, \quad i = 1, \ldots, N. \tag{24}$$

B.2. Evaluation Metrics - Model Calibration Degree

B.2.1 Equal-Width Expected Calibration Error (ECE_{EW}) [18]

$$\text{ECE}_{\text{EW}} = \sum_{i=1}^{M} \frac{|B_i|}{|\mathcal{D}|} |C_i - A_i|, \tag{25}$$

where M is the total number of bins, B_i and $\mathcal{D}$ denote the size of the i^{th} bin and the dataset respectively, $C_i = \frac{1}{|B_i|} \sum_{j \in B_i} P_{\hat{y}_j}$ is the mean prediction confidence of the i^{th} bin, and $A_i = \frac{1}{|B_i|} \sum_{j \in B_i} \mathbb{1}(\hat{y}_i == y_i)$ is the mean accuracy of the i^{th} bin. ECE_{EW} has fixed-width bins, with the range $\left[\frac{i}{M}, \frac{i+1}{M}\right),\ i = 0, \ldots, M-1$ for the i^{th} bin.

B.2.2 Equal-Mass Expected Calibration Error (ECE_{EM}) [48]

$$\text{ECE}_{\text{EW}} = \sum_{i=1}^{M} \frac{|B_i|}{|\mathcal{D}|} \cdot |C_i - A_i|, \qquad \text{where } |B_j| = |B_k|, \forall j, k \in [1, M]. \tag{26}$$

Equal-Mass Expected Calibration Error (ECE_{EM}) is different from Equal-Width Expected Calibration Error (ECE_{EW}) by constraining all bins to have equal size.

B.2.3 SWEEP Expected Calibration Error ($\text{ECE}_{\text{SWEEP}}$) [56]

$$\text{ECE}_{\text{SWEEP}} = \left(\sum_{i=1}^{b^*} \frac{|B_i|}{|\mathcal{D}|} |C_i - A_i \cdot|^p\right)^{\frac{1}{p}}, \qquad \text{where } b^* = \max(b | 1 \le b \le n, \forall b' \le b^*, A_1 \le \cdots \le A_{b'}) \tag{27}$$

p is a hyperparameter that is set to $p = 1$ and n is the largest bin number to be tested which we set to $n = 100$. $\text{ECE}_{\text{SWEEP}}$ follows ECE_{EM} to constrain equal-size bins. $\text{ECE}_{\text{SWEEP}}$ starts with bin number $B = 1$ and keeps increasing the bin number until monotony in bin accuracy breaks.

B.2.4 DEBIAS Expected Calibration Error ($\text{ECE}_{\text{DEBIAS}}$) [29]

$$\text{ECE}_{\text{DEBIAS}} = \sum_{i=1}^{M} \frac{|B_i|}{|\mathcal{D}|} \left[(C_i - A_i)^2 - \frac{A_i \cdot (1 - A_i)}{|B_i| - 1}\right] \tag{28}$$

DEBIAS Expected Calibration Error ($\text{ECE}_{\text{DEBIAS}}$) adopts equal-width bins.

B.2.5 Over-confidence Error (OE)

$$\text{OE} = \sum_{i=1}^{M} \frac{|B_i|}{|\mathcal{D}|} \cdot \mathbb{1}(C_i > A_i) \cdot |C_i - A_i|, \tag{29}$$

We adapt OE to different binning schemes of ECE_{EW}, ECE_{EM}, $\text{ECE}_{\text{SWEEP}}$ to produce OE_{EW}, OE_{EM}, OE_{SWEEP} respectively.

B.3. Evaluation Metrics - Dense Classification

B.3.1 Prediction Accuracy

The model prediction accuracy is computed as:

$$\mathbb{A}(\theta, \mathcal{D}) = \frac{1}{N \times H \times W} \sum_{i=1}^{N} \sum_{j=1}^{H} \sum_{k=1}^{W} \mathbb{1}(\hat{y}_i^{j,k} = y_i^{j,k}), \tag{30}$$

where $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$ denotes the dataset with N samples, H and W is the height and the width of sample respectively.

B.3.2 F-measure

F-measure is computed as:

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

where ξ is a hyperparameter. We follow previous works [72, 37, 92, 36] to set $\xi^2 = 0.3$. We report the maximum F-measure which selects the best results computed with various binarising threshold.

B.3.3 E-measure

Enhanced-alignment measure (E-measure) [12] is computed as:

$$\begin{aligned} Q_{FM} &= \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \phi_{FM}(i, j), \quad \text{where} \\ \phi_{FM} &= f(\xi_{FM}) = \frac{1}{4}(1 + \xi_{FM})^2, \\ \xi_{FM} &= \frac{2 \cdot \varphi_{GT} \circ \varphi_{FM}}{\varphi_{GT} \circ \varphi_{GT} + \varphi_{FM} \circ \varphi_{FM}}, \\ \varphi_I &= I - \mu_i \cdot A, \end{aligned} \tag{32}$$

where $I \in (0, 1)$ is a dense binary prediction map with mean value μ_I, A is an one matrix whose dimension matches that of I, φ_{GT} and φ_{FM} denote groundtruth map and model prediction respectively, H and W is image height and width. Maximum E-measure replaces the mean value with a range of binarising thresholds and report the highest result.

B.4. Datasets

DUTS-TR [63]: is commonly used training dataset for Salient Object Detection task. It consists of 10,553 pairs of image and pixel-wise annotations. We take a subset consisting 1,000 training samples as a validation set and uses the remaining 9,553 samples for training.
DUTS-TE [63]: is a testing dataset consisting of 5,019 images. Both DUTS-TE and DUTS-TR belong to the DUTS dataset.
DUT-OMRON [80]: consists of 5,168 testing images, each of which includes at least one structurally complex foreground object(s).
PASCAL-S [34]: contains 850 testing samples that are obtained from PASCAL-VOC dataset, which is designed for semantic segmentation task.
SOD [44]: includes 300 testing images of a wide variety of natural scenes.
ECSSD [78]: has 1,000 semantically meaningful images for testing.
HKU-IS [33]: is comprised of 4,447 testing images, each having multiple foreground objects.
Describable Texture Dataset (DTD) [9]: contains 5,640 real-world texture images. These images are grouped into 47 categories described by adjectives such as "*grooved*", "*woven*", "*matted*". Some texture images have a distinct region that could be considered to be salient. We selectively choose only 500 texture images that have no obvious salient object and show some examples in Fig. 6. We consider the selected texture images an Out-of-Distribution samples for salient object detection. The complete collection of the 500 selected texture images are presented in Fig. 11 at the end of the Appendix.

Figure 6: Texture image samples from Describable Texture Dataset [9].

C. Model Calibration Benchmark with ECE_{EM}, ECE_{SWEEP} and ECE_{DEBIAS}

We present the model calibration degrees of existing SOD methods, model calibration methods and our proposed methods evaluated in terms of: (i) Equal-Mass Expected Calibration Error ECE_{EM} and Equal-Mass Over-confidence Error OE_{EM} in Tab. 4, (ii) ECE_{SWEEP} and OE_{EM} in Tab. 5, and (iii) ECE_{DEBIAS} in Tab. 6. Our proposed method, $ASLP_{MC}$, still outperforms existing salient object detection and model calibration methods with these model calibration evaluation metrics.

Table 4: Salient object detection model calibration degree benchmark evaluated with ECE_{EM} (%) and OE_{EM} (%). We set the number of bins to $B = 10$. (values are shown in % and red and blue indicate the best and the second-best performance respectively.)

Methods		Year	DUTS-TE [63]		DUT-OMRON [80]		PASCAL-S [34]		SOD [44]		ECSSD [78]		HKU-IS [33]	
			$ECE_{EM}\downarrow$	$OE_{EM}\downarrow$	$ECE_{EM}\downarrow$	$OE_{EM}\downarrow$	$ECE_{EM}\downarrow$	$OE_{EM}\downarrow$	$ECE_{EM}\downarrow$	$OE_{EM}\downarrow$	$ECE_{EM}\downarrow$	$OE_{EM}\downarrow$	$ECE_{EM}\downarrow$	$OE_{EM}\downarrow$
SOD Methods	MSRNet [32]	2017	3.35	3.03	3.64	3.40	4.23	3.93	5.52	5.13	1.12	1.08	1.05	0.96
	SRM [65]	2017	4.45	4.05	4.10	3.78	4.92	4.53	7.69	7.22	2.81	2.57	2.20	2.00
	Amulet [92]	2017	5.63	5.10	5.46	4.98	5.69	5.23	8.24	7.63	2.64	2.45	2.09	1.94
	BMPM [91]	2018	3.47	3.21	4.52	4.18	4.77	4.57	8.00	7.88	1.89	1.83	1.55	1.50
	DGRL [67]	2018	4.42	4.04	3.87	3.57	4.91	4.57	5.69	5.35	2.23	2.07	1.69	1.53
	PAGR [93]	2018	4.00	3.63	3.28	3.00	5.06	4.67	7.60	7.14	2.49	2.29	1.40	1.25
	PiCANet [37]	2018	5.37	4.98	5.17	4.82	5.78	5.41	8.75	8.33	2.67	2.44	2.31	2.09
	CPD [73]	2019	3.10	2.90	3.62	3.42	4.11	3.86	6.76	6.42	2.07	1.94	1.81	1.70
	BASNet [55]	2019	6.07	5.85	6.15	5.96	5.72	5.48	5.07	4.88	2.12	2.04	2.36	2.28
	EGNet [94]	2019	3.54	3.29	3.55	3.33	4.92	4.61	6.42	6.07	1.96	1.84	1.64	1.55
	AFNet [15]	2019	3.58	3.33	3.02	2.81	4.08	3.79	6.65	6.14	2.19	2.04	1.78	1.66
	PoolNet [36]	2019	3.80	3.52	3.53	3.30	5.44	5.09	6.87	6.49	2.18	2.04	1.61	1.52
	GCPANet [7]	2020	4.40	4.12	4.84	4.61	4.92	4.64	4.20	3.94	1.87	1.76	1.54	1.47
	MINet [51]	2020	5.02	4.76	5.40	5.13	6.17	5.86	8.29	8.01	2.84	2.67	2.31	2.17
	F^3Met [69]	2020	3.47	3.26	3.88	3.68	4.56	4.32	7.34	6.95	2.45	2.31	1.91	1.80
	EBMGSOD [89]	2021	3.64	3.41	3.78	3.55	4.79	4.52	5.83	5.56	2.30	2.15	1.85	1.72
	ICON [97]	2021	2.40	2.26	2.95	2.81	3.45	3.29	4.27	4.09	1.34	1.25	1.23	1.16
	PFSNet [43]	2021	3.07	2.84	3.44	3.16	4.99	4.64	5.82	5.48	2.43	2.17	2.87	2.70
	EDN [72]	2022	3.89	3.68	4.35	4.18	4.62	4.41	4.02	3.85	1.60	1.52	1.34	1.26
Model Calibration Methods	Brier Loss [4]	1950	2.78	2.61	3.55	3.40	3.90	3.72	6.40	6.18	1.34	1.31	1.04	1.00
	TS [18]	2017	2.77	2.60	3.44	3.30	3.85	3.67	6.64	6.40	1.21	1.17	0.95	0.91
	MMCE [30]	2018	2.86	2.69	3.56	3.42	4.07	3.89	6.85	6.63	1.41	1.35	1.18	1.13
	LS [46]	2019	2.74	2.10	3.51	2.81	3.97	3.35	4.50	4.10	1.50	0.99	1.44	0.84
	Mixup [62]	2019	3.00	2.73	3.40	3.13	2.14	0.59	4.94	4.62	1.86	0.45	4.94	0.20
	Focal Loss [45]	2020	2.15	2.03	2.69	2.38	2.95	2.70	4.61	4.38	1.57	1.16	1.29	0.87
	AdaFocal [17]	2022	1.74	1.50	1.96	1.45	2.45	2.02	3.88	3.09	1.79	0.74	1.45	0.44
Our Methods	$ASLP_{ECE}$	2023	**1.53**	1.41	**1.72**	1.43	**1.58**	1.55	**2.30**	1.66	**0.71**	0.35	**0.84**	0.19
	$ASLP_{MEI}$	2023	21.00	**0.08**	20.24	**0.00**	19.89	**0.00**	18.14	**0.00**	22.15	**0.00**	22.58	**0.00**

Table 5: Salient object detection model calibration degree benchmark evaluated with ECE_{SWEEP} (%) and OE_{SWEEP} (%). The number of bins for each evaluation is selected to ensure a monotonically increasing accuracy in the bins [56] (values are shown in % and red and blue indicate the best and the second-best performance respectively.)

Methods		Year	DUTS-TE [63]		DUT-OMRON [80]		PASCAL-S [34]		SOD [44]		ECSSD [78]		HKU-IS [33]	
			ECE_{SW} ↓	OE_{SW} ↓	ECE_{SW} ↓	OE_{SW} ↓	ECE_{SW} ↓	OE_{SW} ↓	ECE_{SW} ↓	OE_{SW} ↓	ECE_{SW} ↓	OE_{SW} ↓	ECE_{SW} ↓	OE_{SW} ↓
SOD Methods	MSRNet [32]	2017	3.16	2.85	4.10	3.86	4.09	3.85	5.30	5.05	1.04	1.00	1.01	0.94
	SRM [65]	2017	4.66	4.32	4.92	4.61	5.77	5.43	8.04	7.56	2.98	2.74	2.12	1.95
	Amulet [92]	2017	6.52	6.04	7.31	6.85	6.50	6.08	8.47	7.88	2.17	2.06	2.47	2.32
	BMPM [91]	2018	4.77	4.38	4.27	3.98	6.13	5.74	8.74	8.31	2.09	1.72	2.03	1.85
	DGRL [67]	2018	4.51	4.30	3.98	3.81	4.61	4.46	5.23	4.89	1.98	1.84	1.88	1.73
	PAGR [93]	2018	4.40	4.07	5.20	5.26	5.71	5.44	12.07	11.45	2.80	2.62	1.58	1.50
	PiCANet [37]	2018	4.81	4.52	4.17	3.86	5.34	4.91	7.71	-7.27	2.75	2.46	2.08	1.89
	CPD [73]	2019	4.00	3.80	4.45	4.33	4.76	4.58	6.98	6.65	2.29	2.16	2.26	2.15
	BASNet [55]	2019	7.17	6.94	7.10	6.91	7.70	7.48	7.84	7.74	2.14	2.11	2.59	2.51
	EGNet [94]	2019	3.91	3.68	4.29	4.08	4.75	4.55	5.89	5.56	1.84	1.71	1.29	1.23
	AFNet [15]	2019	4.31	4.06	4.48	4.27	4.56	4.49	6.79	6.24	2.21	2.06	2.06	1.95
	PoolNet [36]	2019	3.58	3.36	4.30	4.10	6.09	5.75	6.72	5.75	1.98	1.85	1.53	1.45
	GCPANet [7]	2020	4.45	4.18	5.26	5.04	5.01	4.75	5.74	5.60	1.63	1.52	1.58	1.51
	MINet [51]	2020	4.97	4.69	6.03	5.77	6.97	6.67	8.17	7.97	1.99	1.93	1.48	1.45
	F^3Met [69]	2020	3.29	3.15	4.56	4.36	4.26	4.10	7.74	7.29	2.20	2.08	2.29	2.17
	EBMGSOD [89]	2021	4.32	4.10	5.03	4.81	4.40	4.29	5.46	5.18	2.53	2.39	2.30	2.17
	ICON [97]	2021	2.64	2.54	4.16	4.02	3.93	3.90	5.13	5.01	1.32	1.24	1.20	1.14
	PFSNet [43]	2021	4.89	4.79	5.89	5.61	7.73	7.54	10.74	10.45	2.31	2.28	2.21	2.19
	EDN [72]	2022	4.28	4.07	4.78	4.60	5.10	4.92	5.63	5.55	1.48	1.45	1.54	1.45
Model Calibration Methods	Brier Loss [4]	1950	3.43	3.17	4.39	4.15	4.44	4.22	5.03	4.22	1.48	1.38	1.21	1.15
	TS [18]	2017	3.30	3.03	4.12	3.91	3.48	3.30	5.33	4.97	1.29	1.22	1.13	1.08
	MMCE [30]	2018	3.44	3.20	4.38	4.17	3.66	3.48	5.55	5.19	1.40	1.31	1.36	1.29
	LS [46]	2019	2.97	2.92	3.88	3.81	4.08	4.99	5.67	5.42	1.46	1.27	1.32	0.99
	Mixup [62]	2019	3.01	2.76	4.47	4.21	1.84	1.26	5.26	4.99	1.28	1.11	1.73	1.48
	Focal Loss [45]	2020	2.23	2.14	3.73	3.43	3.03	2.93	4.77	4.59	1.30	1.16	1.40	1.08
	AdaFocal [17]	2022	1.79	1.60	2.44	2.08	1.88	1.78	4.16	3.46	1.16	0.97	1.03	0.86
Our Methods	$ASLP_{ECE}$	2023	**1.37**	1.21	**1.67**	1.33	**1.77**	1.51	**2.73**	2.41	**0.97**	0.61	**0.89**	0.41
	$ASLP_{MEI}$	2023	20.78	**0.00**	19.64	**0.00**	19.74	**0.00**	17.35	**0.00**	22.47	**0.00**	22.90	**0.00**

Table 6: Salient object detection model calibration degree benchmark evaluated with $\text{ECE}_{\text{DEBIAS}}$ [29]. We set he number of bins to $B = 10$. (values are shown in % and red and blue indicate the best and the second-best performance respectively.)

Methods		Year	$\text{ECE}_{\text{DEBIAS}}(\%) \downarrow$					
			DUTS-TE [63]	DUT-OMRON [80]	PASCAL-S [34]	SOD [44]	ECSSD [78]	HKU-IS [33]
SOD Methods	MSRNet [32]	2017	0.167	0.188	0.235	0.524	0.020	0.015
	SRM [65]	2017	0.419	0.358	0.436	1.221	0.186	0.110
	Amulet [92]	2017	0.553	0.536	0.508	1.165	0.235	0.079
	BMPM [91]	2018	0.471	0.378	0.440	1.175	0.191	0.134
	DGRL [67]	2018	0.420	0.370	0.430	0.807	0.096	0.072
	PAGR [93]	2018	0.340	0.418	0.470	1.568	0.137	0.053
	PiCANet [37]	2018	0.456	0.359	0.461	0.985	0.175	0.124
	CPD [73]	2019	0.390	0.353	0.567	1.233	0.145	0.109
	BASNet [55]	2019	0.544	0.536	0.683	1.190	0.138	0.127
	EGNet [94]	2019	0.318	0.304	0.576	0.860	0.109	0.066
	AFNet [15]	2019	0.381	0.348	0.471	0.934	0.132	0.091
	PoolNet [36]	2019	0.335	0.326	0.612	0.907	0.107	0.055
	GCPANet [7]	2020	0.388	0.318	0.372	0.569	0.068	0.043
	MINet [51]	2020	0.448	0.505	0.606	1.041	0.172	0.142
	F^3Met [69]	2020	0.457	0.468	0.556	0.816	0.193	0.167
	EBMGSOD [89]	2021	0.374	0.406	0.508	0.733	0.154	0.130
	ICON [97]	2021	0.306	0.390	0.382	0.607	0.098	0.101
	PFSNet [43]	2021	0.323	0.339	0.539	0.594	0.588	0.435
	EDN [72]	2022	0.285	0.281	0.407	0.745	0.068	0.061
Model Calibration Methods	Brier Loss [4]	1950	0.241	0.265	0.330	0.572	0.051	0.035
	TS [18]	2017	0.230	0.246	0.338	0.631	0.040	0.024
	MMCE [30]	2018	0.250	0.269	0.378	0.752	0.054	0.039
	LS [46]	2019	0.218	0.241	0.303	0.570	0.047	0.034
	Mixup [62]	2019	0.143	0.211	0.110	0.423	0.078	0.482
	Focal Loss [45]	2020	0.135	0.193	0.262	0.518	0.070	0.061
	AdaFocal [17]	2022	0.069	0.133	0.103	0.383	0.108	0.102
Our Methods	ASLP_{ECE}	2023	**0.056**	**0.103**	**0.061**	**0.083**	**0.024**	**0.027**
	ASLP_{MEI}	2023	4.565	4.027	4.079	3.112	5.095	5.301

D. Joint Distribution of Prediction Confidence and Prediction Accuracy on 6 Testing Datasets

Fig. 7 presents the joint distribution of prediction confidence and prediction accuracy of our methods, existing model calibration methods and some of the salient object detection models on the six SOD testing datasets.

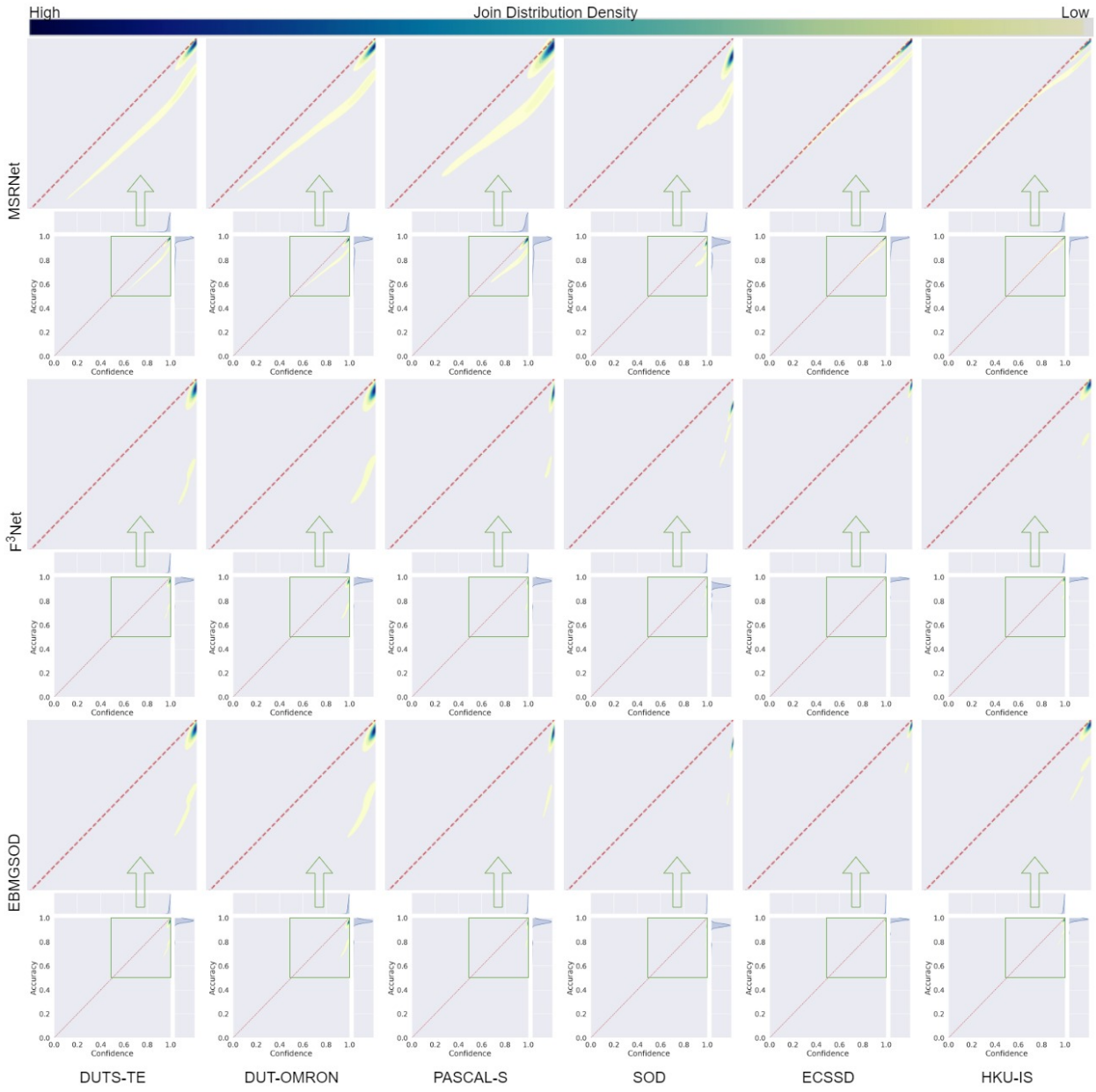


Figure 7: Joint distribution of prediction confidence (horizontal axis) and prediction accuracy (vertical axis) on the six SOD testing datasets.

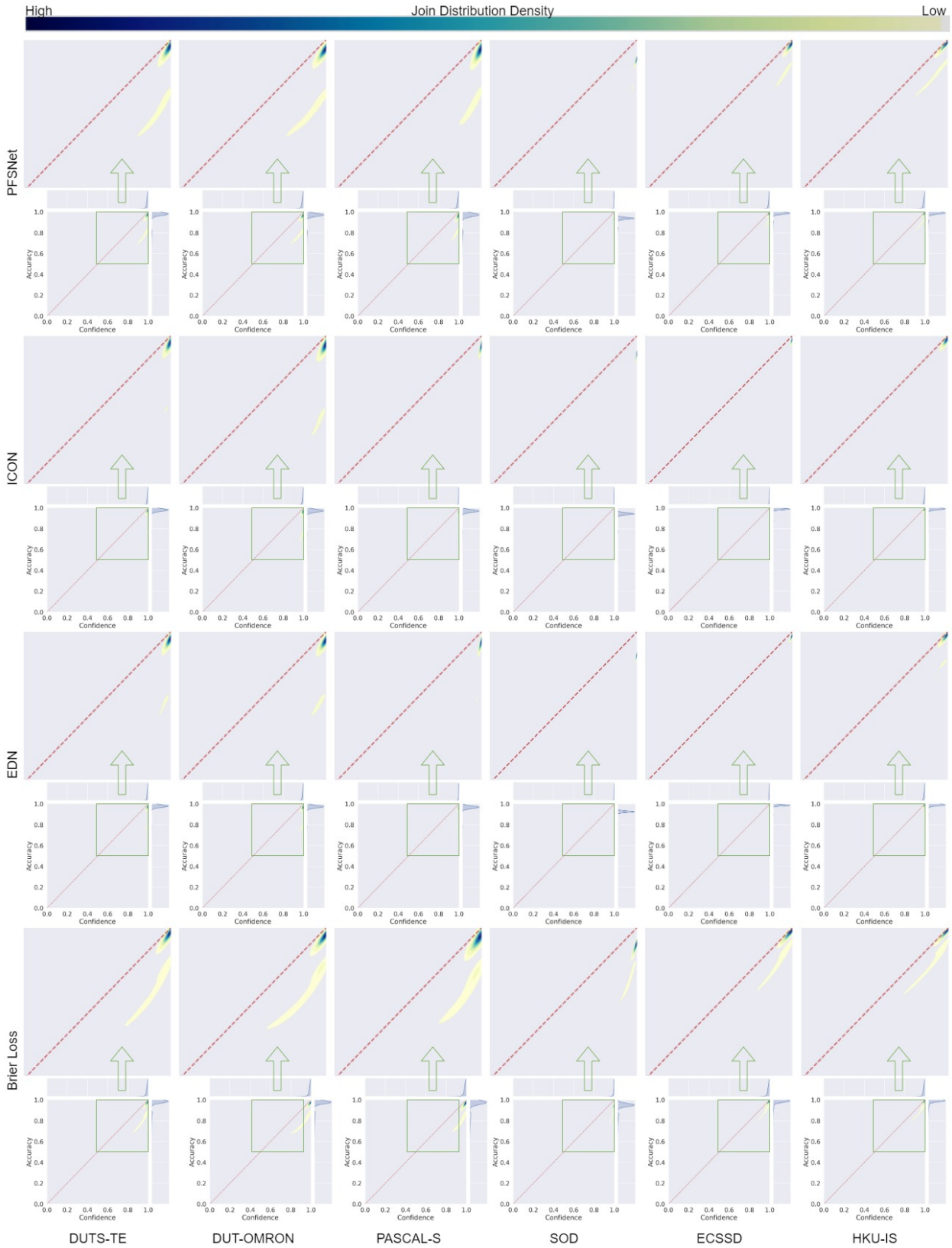


Figure 7: Joint distribution of prediction confidence (horizontal axis) and prediction accuracy (vertical axis) on the six SOD testing datasets.

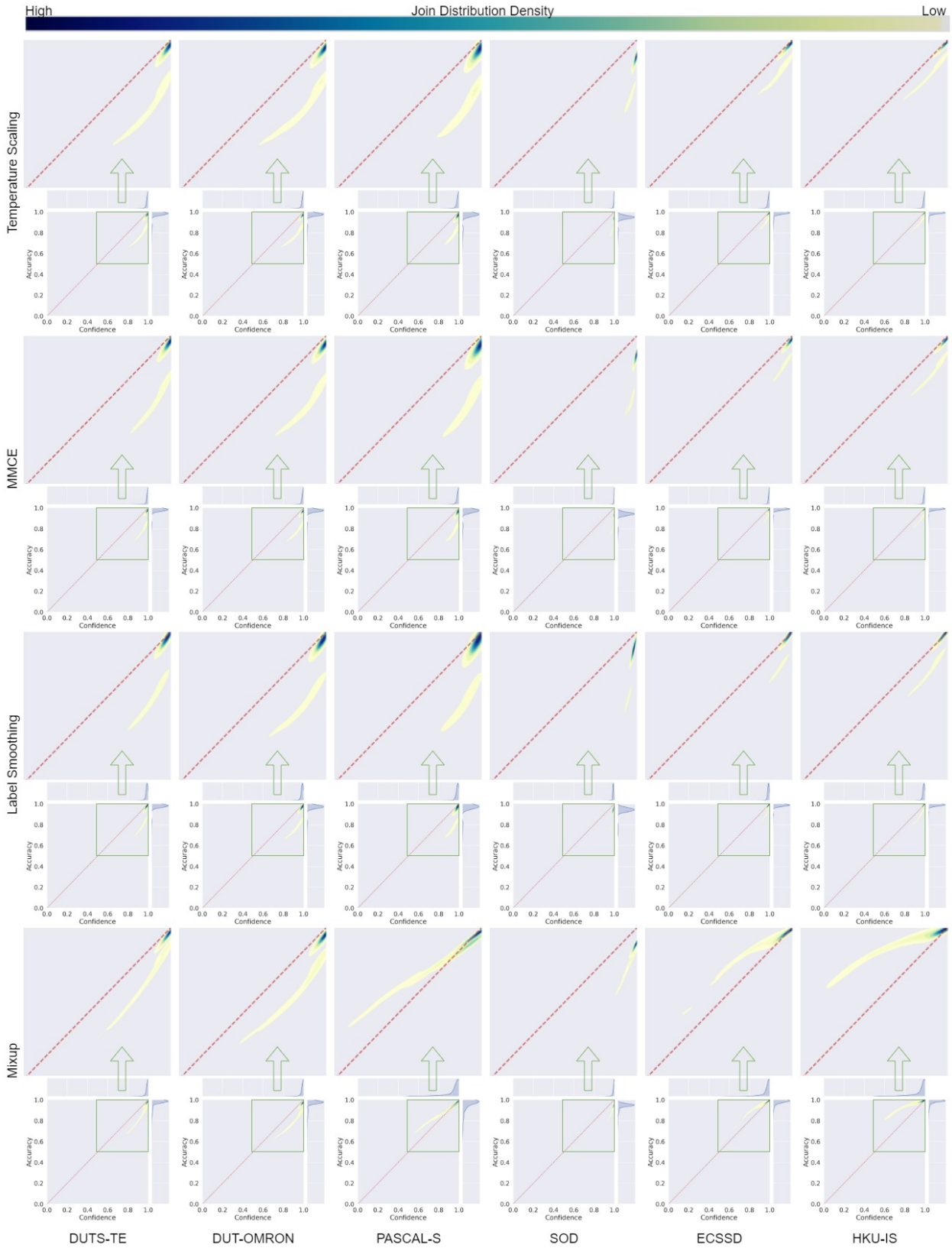


Figure 7: Joint distribution of prediction confidence (horizontal axis) and prediction accuracy (vertical axis) on the six SOD testing datasets.

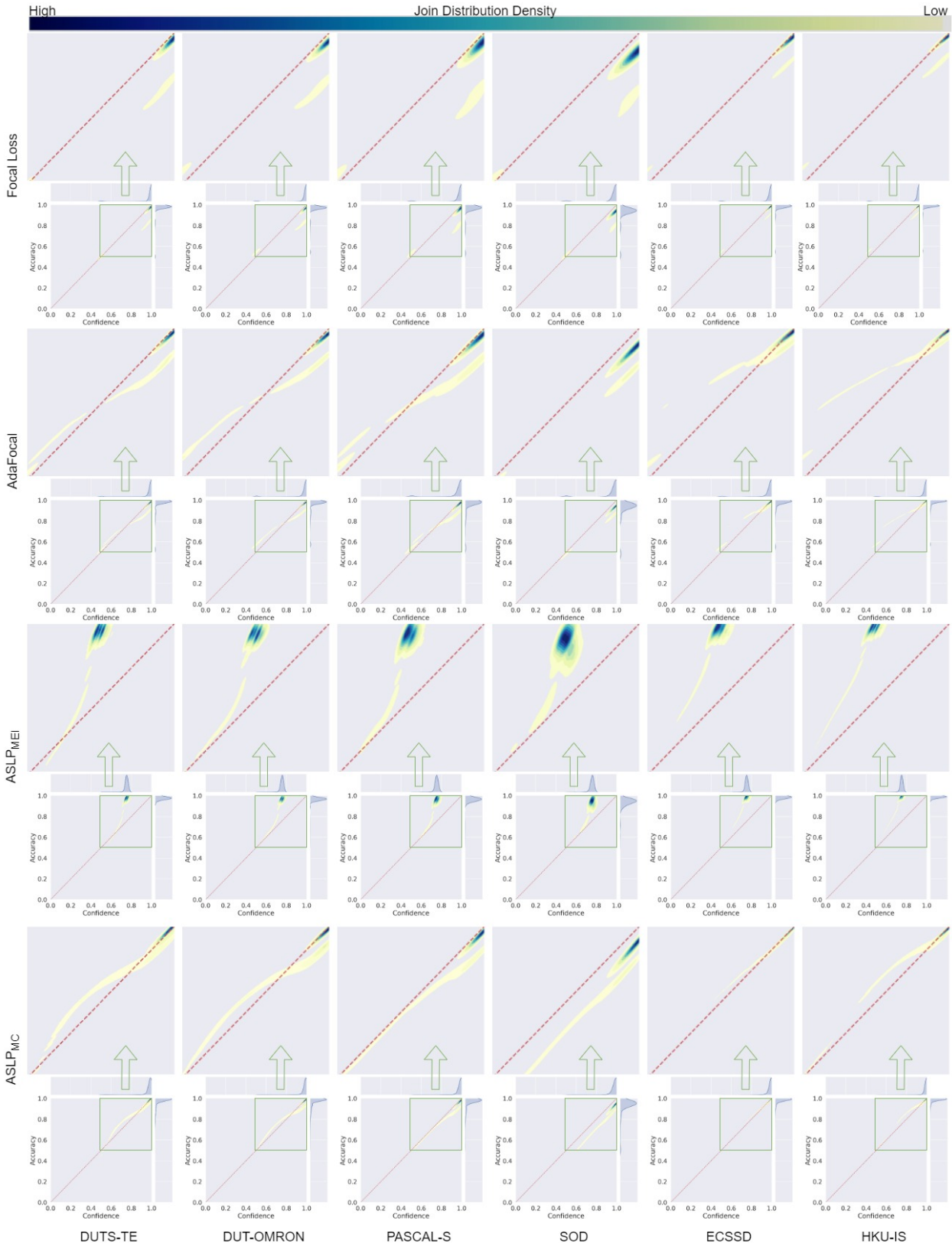


Figure 7: Joint distribution of prediction confidence (horizontal axis) and prediction accuracy (vertical axis) on the six SOD testing datasets.

E. Generalisation to Existing SOD Methods

We study the compatibility of the proposed updating rule ASLP_{MC} with some of the existing state-of-the-art SOD models, including EBMGSOD [89], ICON [97], and EDN [72], and present the model calibration results in Tab. 7. We implement the ASLP_{MC} with the Hard Inversion (HI) label perturbation technique. The results demonstrate that our proposed method is readily compatible with existing SOD methods to improve their respective model calibration degrees. Further, we find that incorporation of our proposed ASLP_{MC} into the training of existing SOD models do not negatively impact their classification performances as demonstrated in Tab. 8.

Table 7: The model calibration degrees of existing Salient Object Detection models with or without the proposed Adaptive Label Augmentation are evaluated in terms of Equal-Width Expected Calibration Error, ECE_{EW}, and Equal-Width Over-confidence Error, OE_{EW}, with 10 bins ($B = 10$).

Methods	Year	ASLP_{MC}	DUTS-TE [63]		DUT-OMRON [80]		PASCAL-S [34]		SOD [44]		ECSSD [78]		HKU-IS [33]	
			ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓
EBMGSOD [89]	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 [97]	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 [72]	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	✓	1.60	1.34	1.91	1.74	2.45	2.23	5.48	5.21	0.77	0.47	0.75	0.22
ICON	2021	✓	1.28	1.05	1.88	1.67	2.45	2.17	5.17	4.91	1.25	0.07	1.10	0.05
EDN	2022	✓	2.02	1.77	2.23	2.03	2.74	2.55	6.77	6.46	0.82	0.52	0.71	0.35

Table 8: The dense classification accuracy of Salient Object Detection models with or without the proposed Adaptive Label Augmentation is evaluated with maximum F-measure and maximum E-measure [12].

Methods	Year	ASLP_{MC}	DUTS-TE [63]		DUT-OMRON [80]		PASCAL-S [34]		SOD [44]		ECSSD [78]		HKU-IS [33]	
			$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑
EBMGSOD [89]	2021	✗	0.850	0.927	0.762	0.867	0.830	0.896	0.834	0.800	0.914	0.944	0.906	0.952
ICON [97]	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 [72]	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	✓	0.853	0.930	0.767	0.871	0.841	0.901	0.839	0.807	0.923	0.946	0.912	0.956
ICON	2021	✓	0.864	0.929	0.776	0.877	0.857	0.904	0.819	0.855	0.940	0.959	0.926	0.959
EDN	2022	✓	0.898	0.954	0.824	0.901	0.880	0.923	0.848	0.866	0.952	0.971	0.942	0.972

F. Experiments on Additional Dense Classification Tasks

F.1. Camouflaged Object Detection

We train our model on the COD10K training set [14] which consists of 6,000 training samples. We partition it into a training set of 5,400 samples and a validation set of 600 samples. Four testing datasets, including the COD10K testing set [14], NC4K [42], CAMO [31] and CHAMELEON [59], are used to evaluate the model calibration degree and dense binary classification accuracy. We train the models for 50 epochs and the rest of settings follow those in Salient Object Detection.

We apply the proposed ASLP_{MC} with Hard Inversion (HI) and Soft Inversion (SI) label perturbation techniques and ALS_{MC} to improve the model calibration degrees with four label perturbation techniques and report the results in Tab. 9. It can be observed that both ASLP with various label perturbation techniques and ALS can also significantly improve model calibration degrees in Camouflaged Object Detection models. Further, we show that the improvements in model calibration degree are achieved without negatively impacting the classification accuracy as shown in Tab. 10.

Table 9: Application Adaptive Stochastic Label Perturbation (ASLP) with different label perturbation techniques in Camouflaged Object Detection task. The model calibration degrees are evaluated with Equal-Width Expected Calibration Error (ECE_{EW}) and Equal-Width Over-confidence Error (OE_{EW}) with 10 bins. Results are presented in (%).

Methods	Perturbation Params			COD10K [14]		NC4K [42]		CHAMELEON [59]		CAMO [31]	
	α	β	e	$\text{ECE}_{\text{EW}}\downarrow$	$\text{OE}_{\text{EW}}\downarrow$	$\text{ECE}_{\text{EW}}\downarrow$	$\text{OE}_{\text{EW}}\downarrow$	$\text{ECE}_{\text{EW}}\downarrow$	$\text{OE}_{\text{EW}}\downarrow$	$\text{ECE}_{\text{EW}}\downarrow$	$\text{OE}_{\text{EW}}\downarrow$
Baseline ("COD-B")	0	0	✗	1.65	1.55	2.75	2.60	0.63	0.57	3.62	3.46
$\text{COD-ASLP}^{\text{HI}}_{\text{MC}}$	α_{ada}	1.0	✗	1.06	0.81	1.67	1.51	0.43	0.12	2.00	1.80
$\text{COD-ASLP}^{\text{SI}}_{\text{MC}}$	α_{ada}	0.75	✗	1.05	0.80	1.72	1.55	0.44	0.21	2.03	1.85
$\text{COD-ALS}_{\text{MC}}$	1.0	β_{ada}	✗	1.03	0.76	1.69	1.53	0.45	0.28	1.98	1.81

Table 10: Application Adaptive Stochastic Label Perturbation (ASLP) with different label perturbation techniques in the Camouflaged Object Detection task. The dense classification accuracy is evaluated with maximum F-measure and maximum E-measure [12].

Methods	Perturbation Params			COD10K [14]		NC4K [42]		CHAMELEON [59]		CAMO [31]	
	α	β	e	$F_{\max}\uparrow$	$E_{\max}\uparrow$	$F_{\max}\uparrow$	$E_{\max}\uparrow$	$F_{\max}\uparrow$	$E_{\max}\uparrow$	$F_{\max}\uparrow$	$E_{\max}\uparrow$
Baseline ("COD-B")	0	0	✗	0.715	0.886	0.803	0.902	0.843	0.940	0.749	0.855
$\text{COD-ASLP}^{\text{HI}}_{\text{MC}}$	α_{ada}	1.0	✗	0.716	0.886	0.803	0.902	0.845	0.942	0.756	0.861
$\text{COD-ASLP}^{\text{SI}}_{\text{MC}}$	α_{ada}	0.75	✗	0.716	0.887	0.802	0.904	0.844	0.943	0.759	0.867
$\text{COD-ALS}_{\text{MC}}$	1.0	β_{ada}	✗	0.717	0.887	0.804	0.905	0.845	0.941	0.767	0.868

Table 11: Application Adaptive Stochastic Label Perturbation (ASLP) with different label perturbation techniques in the Smoke Detection (SD) task. Model calibration degree is evaluated with Equal-Width Expected Calibration Error (ECE_{EW}) and Equal-Width Over-confidence Error (OE_{EW}) with 10 bins. Dense classification accuracy is evaluated with maximum F-measure and maximum E-measure [12].

Methods	Perturbation Params			SMOKE5K [14]			
	α	β	e	$\text{ECE}_{\text{EW}}(\%)\downarrow$	$\text{OE}_{\text{EW}}(\%)\downarrow$	$F_{\max}\uparrow$	$E_{\max}\uparrow$
Baseline ("SD-B")	0	0	✗	0.164	0.154	0.763	0.930
$\text{SD-ASLP}^{\text{HI}}_{\text{MC}}$	α_{ada}	1.0	✗	0.071	0.063	0.763	0.930
$\text{SD-ASLP}^{\text{SI}}_{\text{MC}}$	α_{ada}	0.75	✗	0.076	0.072	0.765	0.932
$\text{SD-ALS}_{\text{MC}}$	1.0	β_{ada}	✗	0.079	0.072	0.764	0.930

F.2. Smoke Detection

We train our model on the SMOKE5K training set [79] which consists of 4,600 training samples of real smoke. We partition it into a training set of 4,200 samples and a validation set of 400 samples. SMOKE5K testing set, comprising of 400 real-smoke images, is used to evaluate model calibration degree and dense binary classification accuracy.

We apply the proposed ASLP_{MC} with Hard Inversion (HI) and Soft Inversion (SI) label perturbation techniques and ALS_{MC} to improve the model calibration degrees and report the results in Tab. 11. It can be observed that both ASLP_{MC} with different label perturbation

techniques and ALS_{MC} can significantly improve model calibration degrees in Smoke Detection models, despite the baseline model already achieving higher calibration degrees compared with baseline models in Salient Object Detection and Camouflaged Object Detection. We can observe that our proposed methods still achieve improvements in model calibration degree without negatively impacting the classification accuracy.

G. Experiments on Additional Dense Multi-Class Classification Task - Semantic Segmentation

We evaluate our proposed methods on the PASCAL VOC 2012 segmentation dataset [11] which has 20 foreground categories and 1 background category. The official split has 1,464, 1,449, and 1,456 samples in training, validation and testing sets respectively. Following previous work [5], we use an augmented training set comprising of 10,582 samples, provided by [19], for model training. As we do not have access to the groundtruth of "official testing set" whose evaluation is server-based, we adopt the "official validation set" as "our testing set" to evaluate the model calibration degrees and segmentation accuracies. Similar to our implementation in dense binary classification tasks, we partition the augmented training set into "our training set" of 9,582 images and "our validation set" of 1,000 images.

We adopt DeepLabv3+ [5] with a ResNet50 backbone as our baseline model ("SS-B") and apply the proposed $ASLP_{MC}$ with with the Hard Inversion (HI) label perturbation technique and ALS_{MC} to improve the model calibration degrees. We report model calibration results evaluated in terms of Equal-Width Expected Calibration Error (ECE_{EW}) and Equal-Width Over-confidence Error (OE_{EW}) with 10 bins in Tab. 12.

Table 12: Application Adaptive Stochastic Label Perturbation (ASLP) with different label perturbation techniques in a Semantic Segmentation (SS) task. Model calibration degree is evaluated with Equal-Width Expected Calibration Error (ECE_{EW}) and Equal-Width Over-confidence Error (OE_{EW}) with 10 bins. Segmentation accuracy is evaluated with Intersection-over-Union (IoU) [5].

Methods	Perturbation Params			PASCAL VOC 2012 [11]		
	α	β	e	ECE_{EW} (%) ↓	OE_{EW} (%) ↓	IoU (%) ↑
Baseline ("SS-B")	0	0	✗	6.29	5.37	71.2
SS-$ASLP^{HI}_{MC}$	α_{ada}	1.0	✗	4.05	3.13	71.3
SS-ALS_{MC}	1.0	β_{ada}	✗	4.10	3.24	71.5

H. Static Stochastic Label Perturbation

H.1. Implementation

We implement four static stochastic label perturbation techniques each of which have a single label perturbation probability α for the entire training dataset. Their details are as below:

- **Hard Inversion (HI)** produces the perturbed label by inverting the groundtruth label with $p = \text{LP}(y, 2) = 1 - y$. Intuitively, it switches the label category from "salient" to "non-salient" and vice versa. The label perturbation probability is limited to $\alpha \in [0, 0.5)$ to avoid learning a complete opposite task (non-salient background detection).
- **Soft Inversion (SI)** inverts the label category and softens the target with $p = \text{LP}(y, 0.75) = -0.5y + 0.75$. Similarly, the label perturbation probability is limited to $p \in [0, \frac{1}{1.5})$ to prevent from learning a complete opposite task.
- **Moderation (M)** transforms groundtruth label into a prior distribution on the two classes (salient foreground object v.s. non-salient background), as $p = \text{LP}(y, 0.5) = 0.5$. The label perturbation probability is in the range $\alpha \in [0, 1)$.
- **Dynamic Moderation (DM)** introduces additional stochasticity on top of the **Moderation** method by adding an additional noise sampled from a truncated normal distribution[2]: $p = \text{LP}(y, 0.5) + e = 0.5 + e,\ e \sim \mathcal{N}_{-0.5,0.5}(0, 1)$. The label perturbation probability is in the range $\alpha \in [0, 1)$.

H.2. Effect of Static Stochastic Label Perturbation Techniques on Model Calibration Degrees

Fig. 8 presents model calibration degrees, evaluated in terms of Equal-Width Expected Calibration Error (ECE_{EW}) and Equal-Width Over-confidence Error (OE_{EW}) with 100 bins ($B = 100$), of various static stochastic label perturbation techniques, in which a unique label perturbation probability α is set for all samples throughout the training. We can observe that, with an increasing label perturbation probability, ECE scores tend to reduce to a critical points before climbing. This is caused by the model transitioning from being over-confident to under-confident. This is evidenced in the OE scores which keep decreasing until 0 when the label perturbation probability increases. Further, "HI" has the steepest change in terms of both ECE and OE scores. This rate can be related to the product of label perturbation probability and strength $\alpha\beta$. We also find a dampening effect of additional stochasticity at high label perturbation probability range ($\alpha \in [0.4, 0.6]$) where "DM" is consistently less under-confident than "M".

Table 13: Effect label perturbation probability range (%) for different static stochastic label perturbation techniques to reduce the Equal-Width Expected Calibration Error (ECE_{EW}) scores on the six testing datasets.

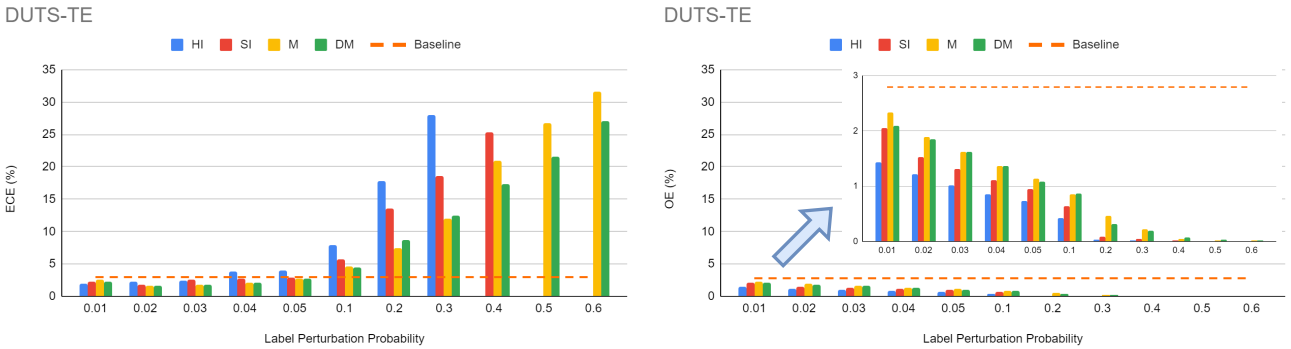
Static SLP Technique	DUTS-TE [63]	DUT-OMRON [80]	PASCAL-S [34]	SOD [44]	ECSSD [78]	HKU-IS [33]
Hard Inversion (HI)	0 - 5%	0 - 3%	0 - 5%	0 - 10%	0 - 1%	0 - 1%
Soft Inversion (SI)	0 - 5%	0 - 5%	0 - 5%	0 - 10%	0 - 2%	0 - 2%
Moderation (M)	0 - 5%	0 - 5%	0 - 5%	0 - 20%	0 - 3%	0 - 3%
Dynamic Moderation (DM)	0 - 5%	0 - 5%	0 - 5%	0 - 20%	0 - 3%	0 - 3%

The effective label perturbation probability range for each static SLP technique on the six testing datasets is summarised in Tab. 13. In general, the static SLPs have a wide range of effective label perturbation probability leading to reduced ECE scores compared to the baseline. The widest effective label perturbation probability range is found on the SOD dataset, with 0 - 10% for "HI" and "SI" and 0 - 20% for "M" and "DM". This can be attributed to the baseline model being the most mis-calibrated on the SOD dataset, thus stronger label augmentation measures are required to transform the model from being over-confident to being under-confident. On the other hand, the baseline model is the most calibrated on the ECSSD and the HKU-IS datasets, indicating a small gap between the prediction confidence and prediction accuracy distributions. That leaves little space for label augmentation techniques to reduce the prediction confidence in order to match the prediction accuracy.

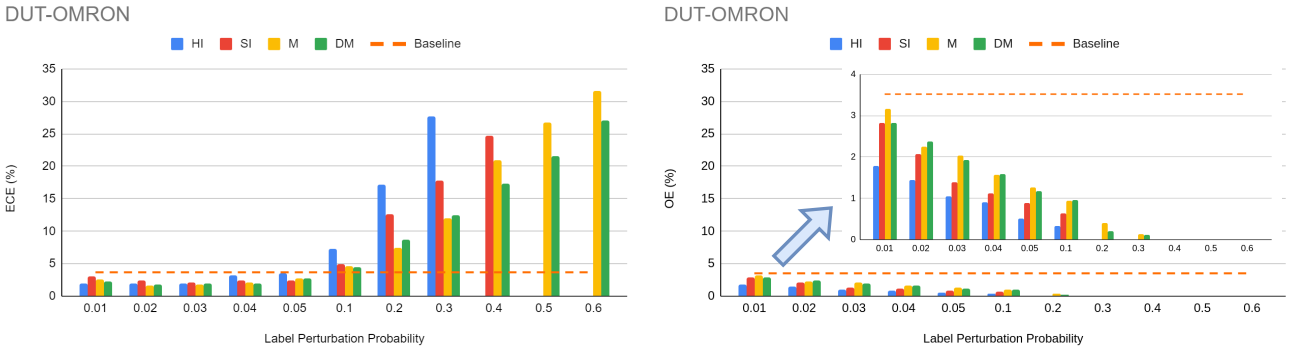
H.3. Effect of Static Stochastic Label Perturbation Techniques on Dense Binary Classification Performance

We present the dense binary classification performance, evaluated in terms of maximum F measure, of various static stochastic label perturbation techniques in Fig. 9. It can be observed that in the effective label perturbation probability range for respective static SLP techniques, the dense binary classification performances are not negatively impacted. The performance drop is observed when the product $\alpha\beta$ is too high, *e.g.* $\alpha \in [0.2, 0.3]$ for "HI", $\alpha = 0.4$ for "SI", and $\alpha = 0.6$ for "DM". Overall, incorporation of static SLP techniques, with an effective label perturbation probability, can achieve improved model calibration degrees without sacrificing the dense bianry classification performance.

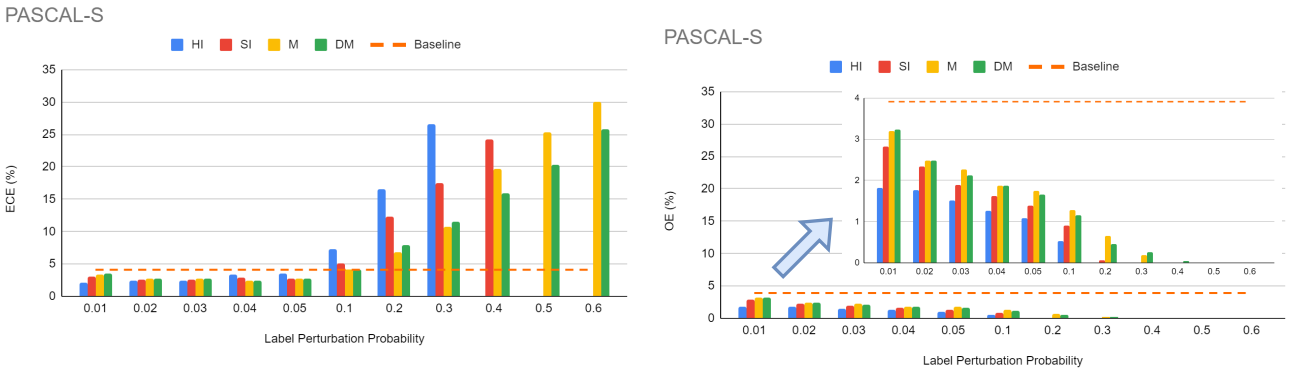
[2]Truncated normal distribution $\mathcal{N}_{a,b}(\mu, \sigma)$, where a and b indicate the bound, μ is the mean and σ is the variance.



(a) DUTS-TE

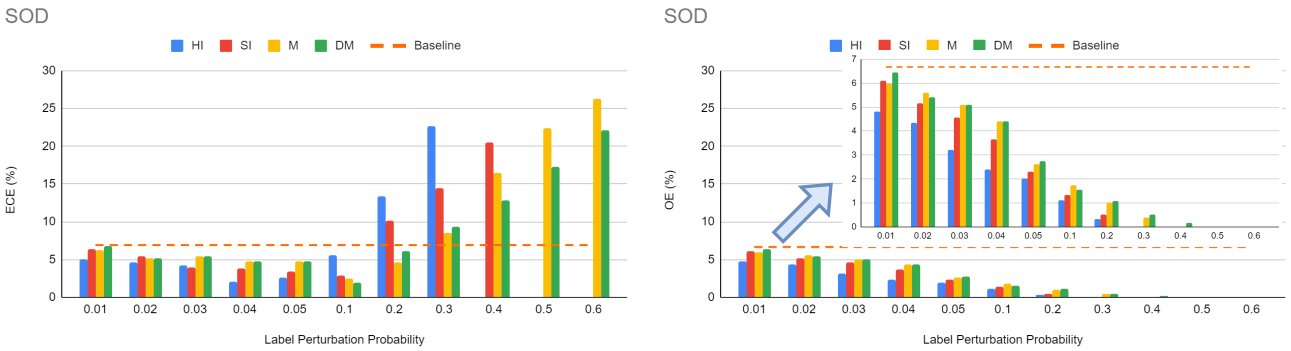


(b) DUT-OMRON

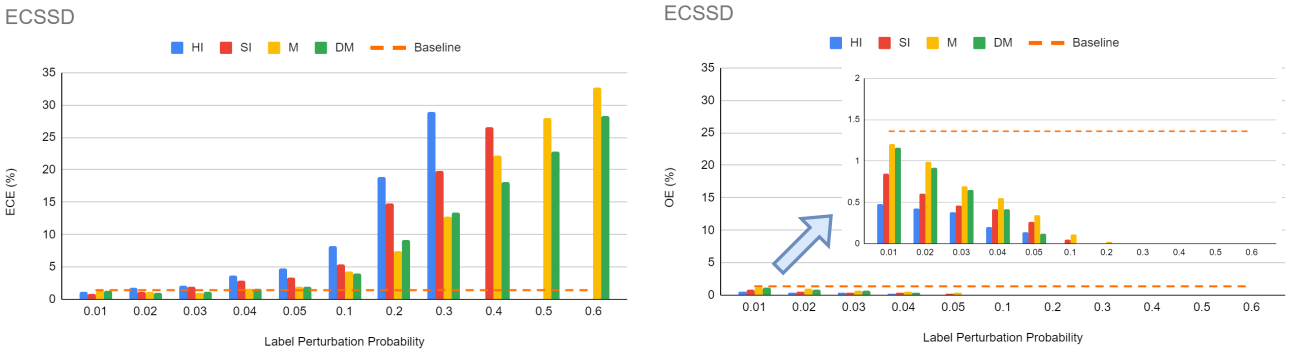


(c) PASCAL-S

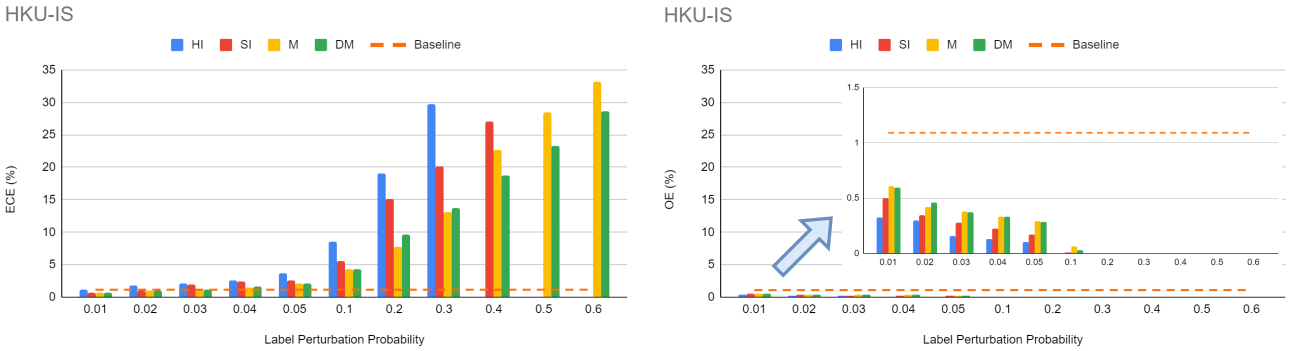
Figure 8: Model calibration degrees, evaluated in terms of Equal-Width Expected Calibration Error (ECE_{EW}) and Equal-Width Over-confidence Error (OE_{EW}) with 100 bins ($B = 100$), of various static stochastic label perturbation techniques under different label perturbation probabilities on the six testing datasets: (a): DUTS-TE, (b) DUT-OMRON, (c) PASCAL-S, (d) SOD, (e) ECSSD, (f) HKU-IS.



(d) SOD

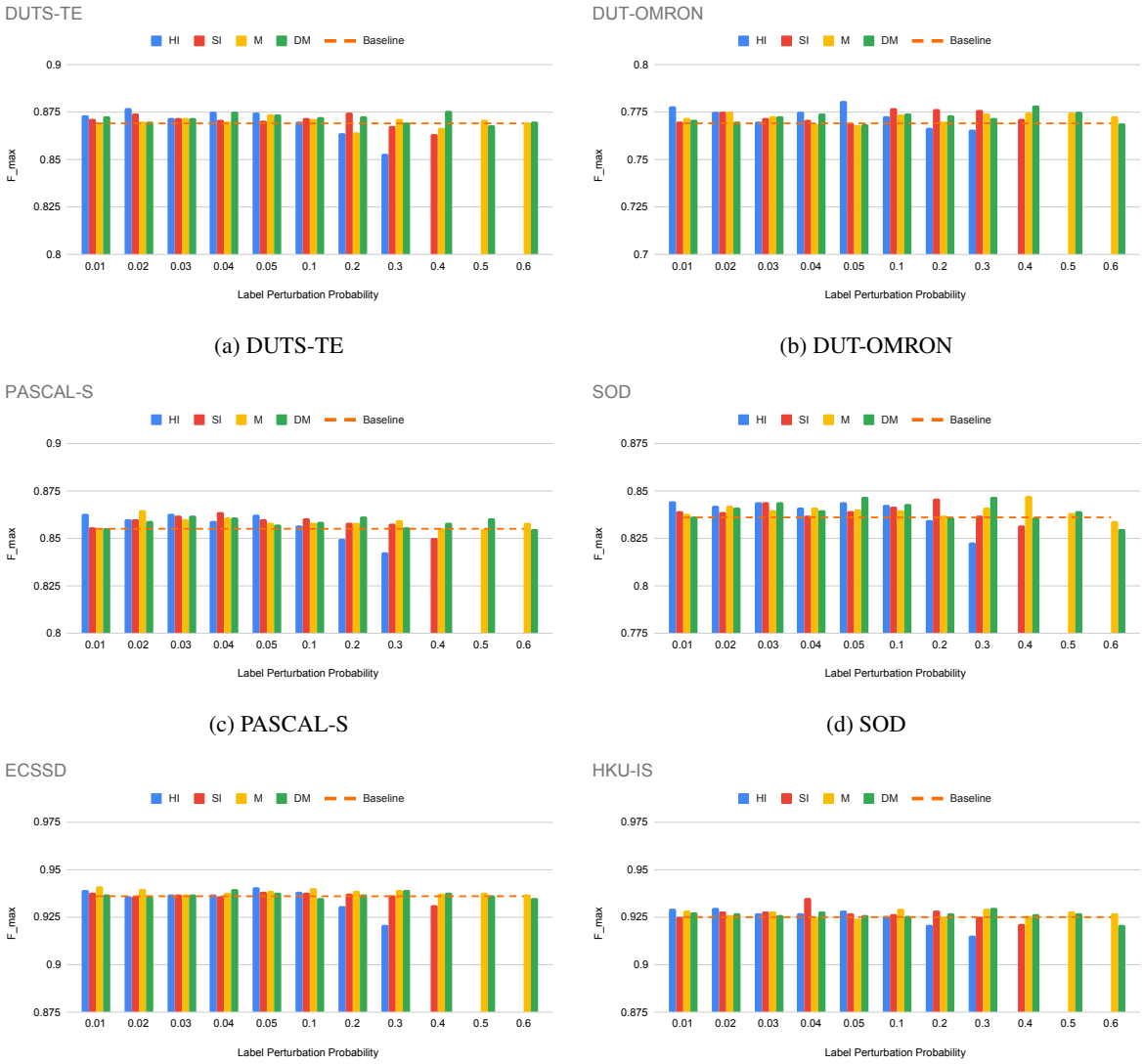


(e) ECSSD



(f) HKU-IS

Figure 8: Model calibration degrees, evaluated in terms of Equal-Width Expected Calibration Error (ECE_{EW}) and Equal-Width Over-confidence Error (OE_{EW}) with 100 bins ($B = 100$), of various static stochastic label perturbation techniques under different label perturbation probabilities on the six testing datasets: (a): DUTS-TE, (b) DUT-OMRON, (c) PASCAL-S, (d) SOD, (e) ECSSD, (f) HKU-IS.



(a) DUTS-TE (b) DUT-OMRON

(c) PASCAL-S (d) SOD

(e) ECSSD (f) HKU-IS

Figure 9: Dense binary classification performance, evaluated in terms of maximum F measure, of various static stochastic label perturbation techniques under different label perturbation probabilities on the six testing datasets: (a): DUTS-TE, (b) DUT-OMRON, (c) PASCAL-S, (d) SOD, (e) ECSSD, (f) HKU-IS.

I. Experiments on Salient Object Detection with Additional Backbones

Experiments with additional backbones, VGG16 and Swin Transformer, are carried out on Salient Object Detection. We replace the ResNet50 backbone of the baseline model with VGG16 and Swin Transformer in respective experiments. We apply the proposed ASLP_{MC} with with Hard Inversion (HI) and Soft Inversion (SI) label perturbation techniques and ALS_{MC} to improve the model calibration degrees with respective backbones.

Table 14: Model calibration degrees with Swin transformer [39] backbone. Results are evaluated with Equal-Width Expected Calibration Error (ECE_{EW}) and Equal-Width Over-confidence Error (OE_{EW}) with 10 bins (units in (%)).

Methods	Perturbation Params			DUTS-TE [63]		DUT-OMRON [80]		PASCAL-S [34]		SOD [44]		ECSSD [78]		HKU-IS [33]	
	α	β	e	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓
Baseline ("Swin-B")	0	0	0	2.41	2.23	3.29	3.15	3.35	3.19	6.23	6.05	1.02	0.97	0.87	0.82
$\text{Swin-ASLP}_{\text{MC}}^{\text{HI}}$	α_{ada}	1.0	✗	1.44	1.21	1.73	1.59	1.74	1.57	5.08	4.85	0.57	0.30	0.81	0.23
$\text{Swin-ASLP}_{\text{MC}}^{\text{SI}}$	α_{ada}	0.75	✗	1.48	1.14	1.63	1.49	1.80	1.52	5.14	4.93	0.64	0.38	0.80	0.24
Swin-ALS	1.0	β_{ada}	✗	1.44	1.14	1.76	1.57	1.69	1.55	5.17	4.82	0.54	0.36	0.77	0.24

Table 15: Dense classification accuracy with Swin transformer [39] backbone. Results are evaluated with maximum F-measure and maximum E-measure [12].

Methods	Perturbation Params			DUTS-TE [63]		DUT-OMRON [80]		PASCAL-S [34]		SOD [44]		ECSSD [78]		HKU-IS [33]	
	α	β	e	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑
Baseline ("Swin-B")	0	0	0	0.894	0.949	0.804	0.890	0.877	0.920	0.858	0.878	0.948	0.969	0.939	0.969
$\text{Swin-ASLP}_{\text{MC}}^{\text{HI}}$	α_{ada}	1.0	✗	0.895	0.953	0.808	0.892	0.881	0.924	0.959	0.879	0.950	0.969	0.938	0.969
$\text{Swin-ASLP}_{\text{MC}}^{\text{SI}}$	α_{ada}	0.75	✗	0.895	0.952	0.805	0.893	9,880	0.922	0.857	0.882	0.950	0.969	0.939	0.970
Swin-ALS	1.0	β_{ada}	✗	0.895	0.952	0.804	0.892	0.879	0.920	0.859	0.879	0.948	0.969	0.939	0.970

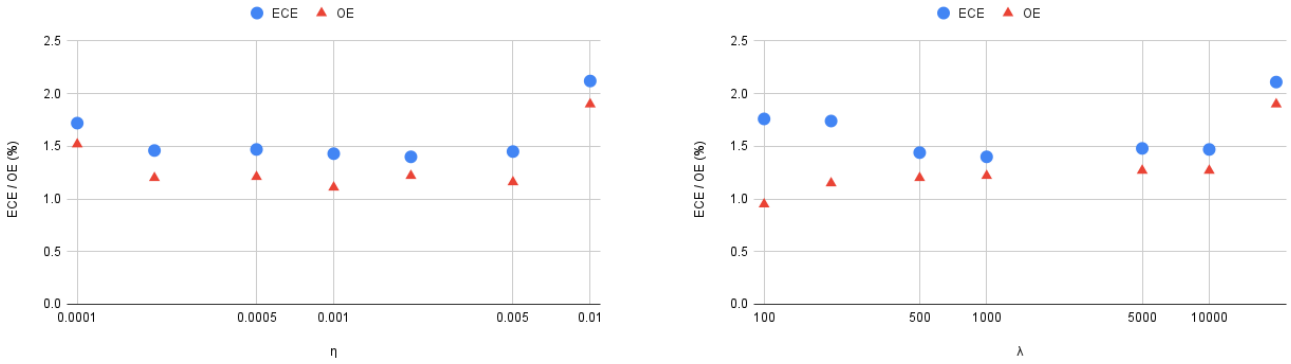
Table 16: Model calibration degrees with VGG16 [58] backbone. Results are evaluated with Equal-Width Expected Calibration Error (ECE_{EW}) and Equal-Width Over-confidence Error (OE_{EW}) with 10 bins (units in (%)).

Methods	Perturbation Params			DUTS-TE [63]		DUT-OMRON [80]		PASCAL-S [34]		SOD [44]		ECSSD [78]		HKU-IS [33]	
	α	β	e	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓	ECE ↓	OE ↓
Baseline ("VGG-B")	0	0	0	3.46	3.23	4.12	3.92	4.40	4.17	7.87	7.60	2.02	1.91	1.51	1.44
$\text{VGG-ASLP}_{\text{MC}}^{\text{HI}}$	α_{ada}	1.0	✗	1.44	1.28	1.91	1.82	2.40	2.16	5.44	5.08	0.57	0.21	0.84	0.16
$\text{VGG-ASLP}_{\text{MC}}^{\text{SI}}$	α_{ada}	0.75	✗	1.47	1.23	2.05	1.81	2.34	2.15	5.54	5.22	0.51	0.21	0.88	0.19
VGG-ALS	1.0	β_{ada}	✗	1.48	1.31	1.99	1.76	2.33	2.04	5.53	5.14	0.45	0.29	0.82	0.13

Table 17: Dense classification accuracy with VGG16 [58] backbone. Results are evaluated with maximum F-measure and maximum E-measure [12].

Methods	Perturbation Params			DUTS-TE [63]		DUT-OMRON [80]		PASCAL-S [34]		SOD [44]		ECSSD [78]		HKU-IS [33]	
	α	β	e	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑	$F_{\max}$ ↑	$E_{\max}$ ↑
Baseline ("VGG-B")	0	0	0	0.838	0.912	0.741	0.851	0.844	0.895	0.810	0.851	0.921	0.944	0.913	0.950
$\text{VGG-ASLP}_{\text{MC}}^{\text{HI}}$	α_{ada}	1.0	✗	0.844	0.916	0.746	0.857	0.844	0.896	0.812	0.851	0.921	0.944	0.913	0.951
$\text{VGG-ASLP}_{\text{MC}}^{\text{SI}}$	α_{ada}	0.75	✗	0.845	0.916	0.747	0.855	0.846	0.895	0.810	0.851	0.921	0.944	0.916	0.953
VGG-ALS	1.0	β_{ada}	✗	0.843	0.914	0.745	0.857	0.848	0.898	0.811	0.852	0.921	0.945	0.913	0.952

J. Hyperparameters



(a) Learning Rate (η)

(b) Regularisation Strength (λ)

Figure 10: Ablation study on hyperparameters: (1) learning rate (η) and (2) regularisation strength (λ) evaluated in terms of ECE_{EW} and OE_{EW} with 100 bins on the DUTS-TE dataset.

K. Training and Inference Time

In SOD, the training of ASLP on DUTS-TR requires 2.5 hours, which is 0.2 hours longer (or $\sim$ 8.7% more) than training the base model (2.3 hours). The inference speed of ASLP on the six SOD testing datasets averages: 53.40 samples per second, which is the same as that of the base model because of the same network architecture. Both training and inference time are evaluated on a single Geforce RTX 3090 GPU.

L. 500 Texture Images from Describable Texture Dataset

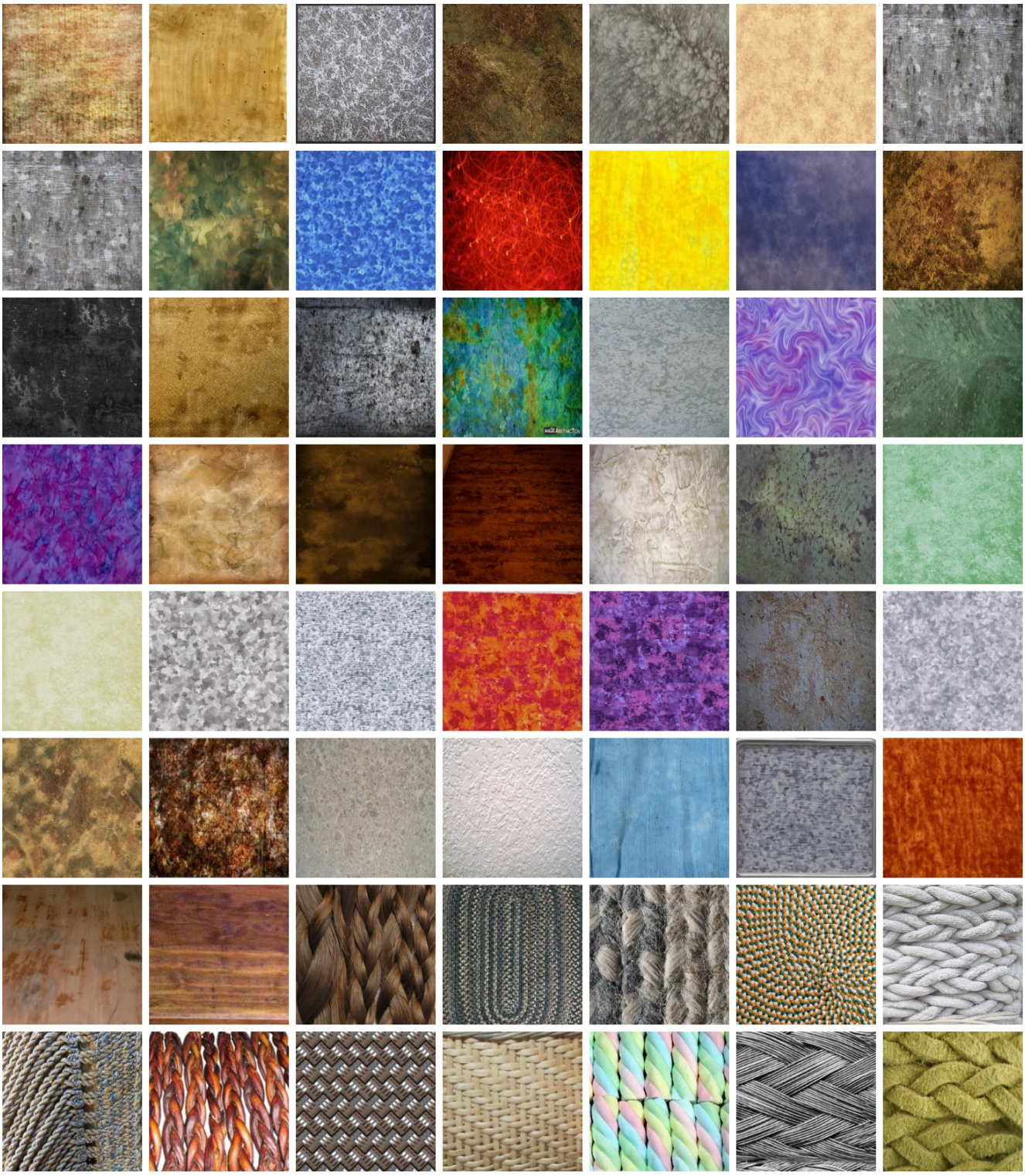


Figure 11: Texture images without visually salient objects selected from Describable Texture Dataset [9].

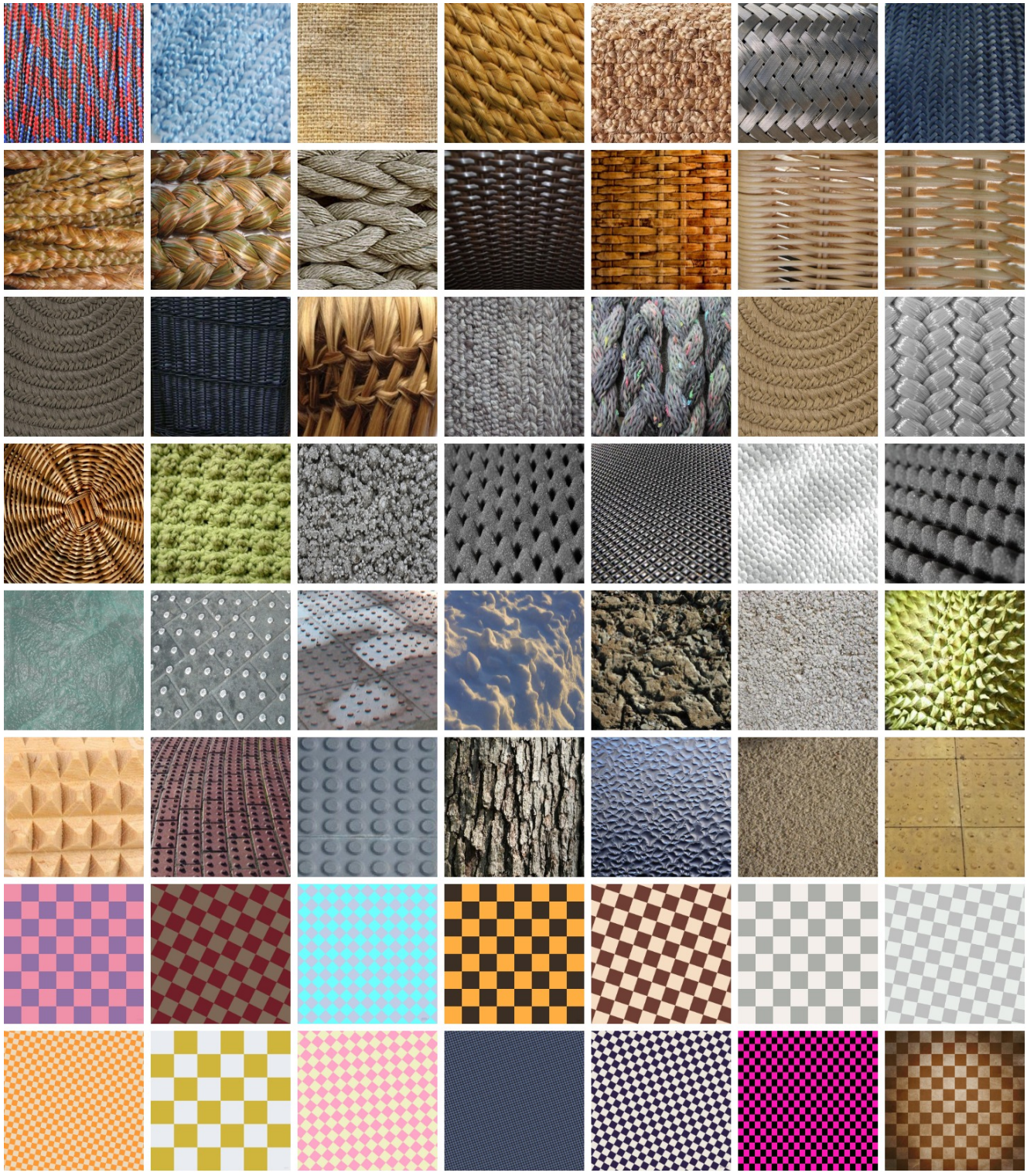


Figure 11: Texture images without visually salient objects selected from Describable Texture Dataset [9].

Figure 11: Texture images without visually salient objects selected from Describable Texture Dataset [9].

Figure 11: Texture images without visually salient objects selected from Describable Texture Dataset [9].

Figure 11: Texture images without visually salient objects selected from Describable Texture Dataset [9].

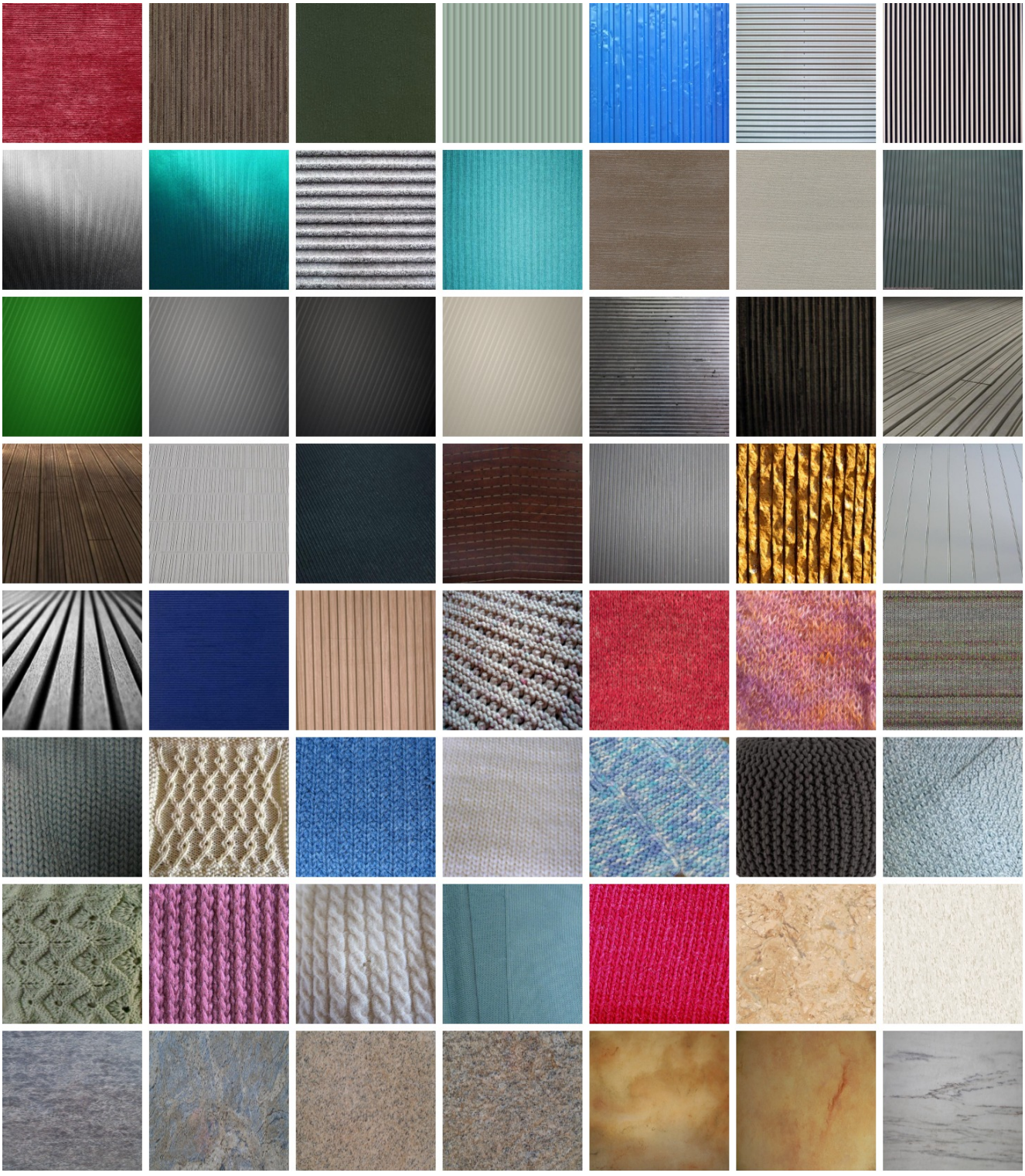


Figure 11: Texture images without visually salient objects selected from Describable Texture Dataset [9].

Figure 11: Texture images without visually salient objects selected from Describable Texture Dataset [9].

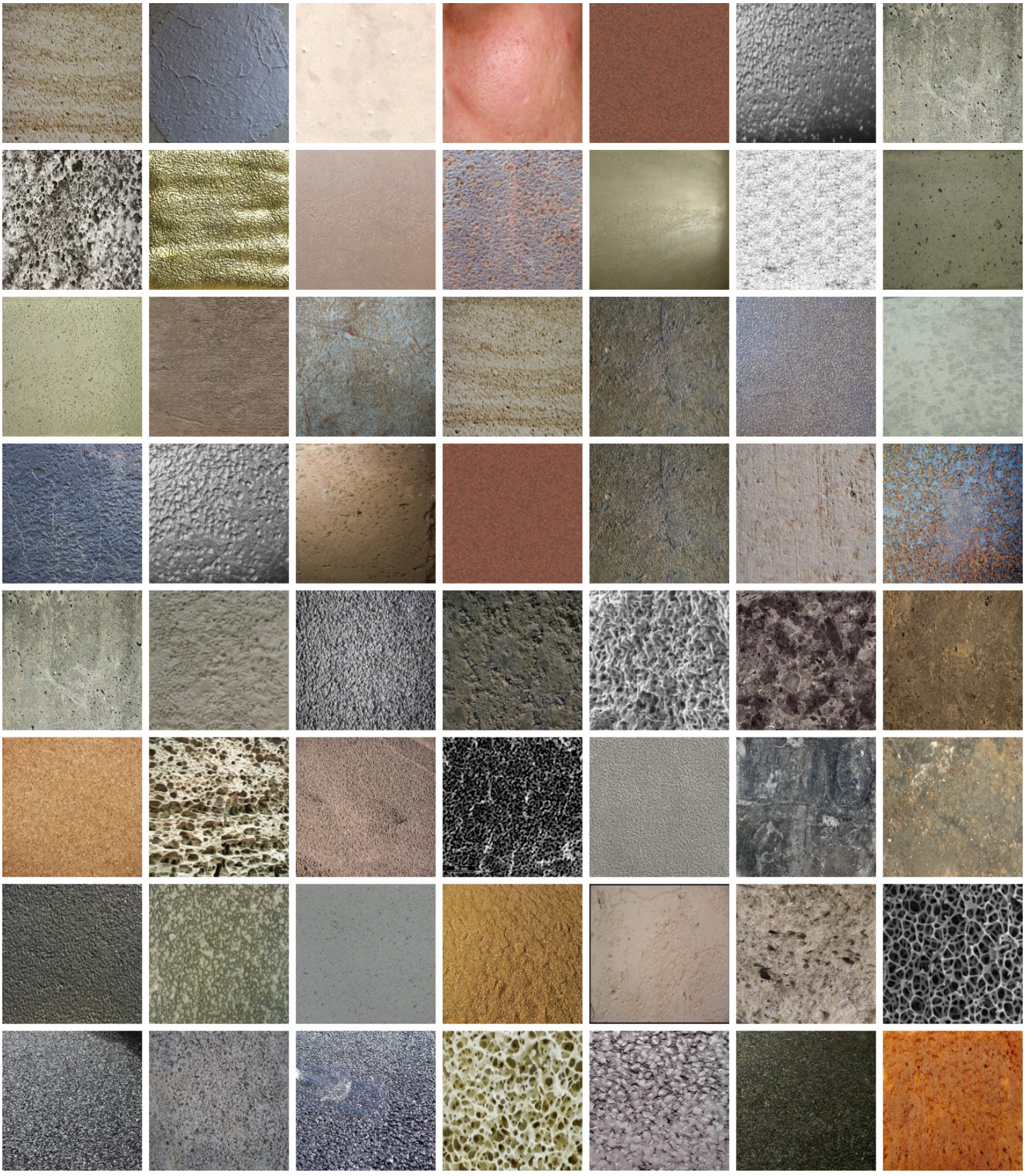


Figure 11: Texture images without visually salient objects selected from Describable Texture Dataset [9].

Figure 11: Texture images without visually salient objects selected from Describable Texture Dataset [9].