SPACE: SPAtial-aware Consistency rEgularization for anomaly detection in Industrial applications
Supplementary material

Algorithm A Spatial-aware Consistency Loss (SCL)

Notation: teacher network g_t**, student network** g_s
Input: original image x_o**, weak augmented image** x_w**, strong augmented image** x_s

if iteration $<$ 5000 **then**
    λ_1 = 0.0
else
    λ_1 = 1.0
end if

 # Calculate relative distances for masks
F_{ts}^o = square($g_t(x_o)$ - $g_s(x_o)$)
F_{ts}^w = square($g_t(x_o)$ - $g_s(x_w)$)
F_{ts}^s = square($g_t(x_o)$ - $g_s(x_s)$)
 # Calculate consistency distances in student
D^{ow} = square(stopgrad($g_s(x_o)$) - $g_s(x_w)$)
D^{os} = square(stopgrad($g_s(x_o)$) - $g_s(x_s)$)
D^{ws} = square($g_s(x_w)$ - $g_s(x_s)$)
 # Create update masks
$M^o = F_{ts}^o > \Upsilon$
$M^w = F_{ts}^w < \Upsilon$
$M^s = F_{ts}^s < \Upsilon$
 # Compute local losses
$\mathcal{L}_{ts}$ = sum($F_{ts}^o \odot M^o$) / sum(M^o)
$\mathcal{L}_{unw}$ = sum($D^{os} \odot M^w$) / sum(M^w)
$\mathcal{L}_{uns}$ = sum($D^{os} \odot M^s$) / sum(M^s)
$\mathcal{L}_{unws}$ = sum($D^{ws} \odot M^w \odot M^s$) / sum($M^w * M^s$)
 # Update criterion
if iteration == 0 **then**
    $\Upsilon = F_{ts}^o$
else
    $\Upsilon = \alpha \times \Upsilon + (1 - \alpha) \times F_{ts}^o$
end if

$\mathcal{L}_{local} = \mathcal{L}_{ts} + \lambda_1 \times (\mathcal{L}_{unw} + \mathcal{L}_{uns} + \mathcal{L}_{unws})$

return $\mathcal{L}_{local}$

A. Input and Training Details

Our model has four input branches. Among these, the three inputs are dedicated to the training of the student model, and the other one is for training the feature-encoder and feature-converter module. For the input size, the original images are resized to 256×256 dimensions. It performs normalization using the mean and standard deviation values from the ImageNet dataset, where the mean values are 0.485, 0.456, and 0.406, and the standard deviations are 0.229, 0.224, and 0.225. Additionally, weak augmentation involves shifting the images randomly by up to 3 pixels, while strong augmentation combines RandAugment [1] with horizontal and vertical flips. The parameters for RandAugment are set to 4 and 10, respectively. For the feature-encoder and Feature-converter Module (FM), the original images are resized to 256×256 size. Furthermore, augmentation is performed, including brightness, contrast, and saturation, each with a parameter value of 0.2. The training procedure is conducted using the Adam optimizer with a batch size of 1, involving 120,000 iterations on MVTec LOCO and 70,000 iterations on MVTec 2D.

The weight decay is set to 1e-5 for the student model, while 1e-6 is used for both the feature-encoder and FM. Additionally, to align the training speed of the feature-encoder with that of the student model, we update the student weights using an exponential moving average with a parameter of 0.999. The specific Spatial-aware Consistency Loss (SCL) algorithm for the student model is described in Algorithm A.

B. Analysis of Parameters

B.1. Analysis of λ_1 and λ_2

We study the impact of the loss balancing parameters λ_1 and λ_2. λ_1 is a parameter associated with SCL, which regulates when the consistency loss is applied. To address the potential low accuracy of the student model before learning, λ_1 is set at 0 for the initial 5,000 iterations and is later changed to 1.0 after some level of learning has taken place. Tab. A and Tab. B presents the changes in image-level AU-

Dataset	Iterations		AUROC (%)
	λ_1=0	λ_1=1	
MVTec LOCO	-	≥ 0	92.2
	$< 5{,}000$	$\geq 5{,}000$	**92.6**
	$< 10{,}000$	$\geq 10{,}000$	91.9
	$< 20{,}000$	$\geq 20{,}000$	91.8

Table A. The performance differences depending on λ_1.

Dataset	λ_2	AUROC (%)
MVTec LOCO	0.01	90.3
	0.1	**92.6**
	0.2	92.2
	0.4	92.5
	0.8	89.8
	1.0	89.6

Table B. The performance differences depending on λ_2.

ROC on the MVTec LOCO based on the number of iterations with λ_1 maintained at 0. It is observed that applying λ_1 after the first 5,000 iterations yields better performance than applying it from the outset, with a subsequent decrease in performance observed beyond that point. λ_2 are parameters that control the learning speed of the feature-encoder and FM. At 0.1, it shows the best performance, with performance gradually decreasing as it increases.

B.2. Analysis of d_{hard}

We study the impact of the parameters d_{hard}. It is a parameter that removes unnecessary feature positions, selectively training only the loss values that correspond to the top percentile. For example, when using a value like 0.99, d_{hard} is set as the top 99% of feature difference values as a threshold, updating only those values that exceed this threshold. d_{hard} were set to 0.99, and the experimental results for this parameter are presented in Tab. C. The experimental results showed the best performance when d_{hard} was 0.99. This indicates that distilling only the influential values, rather than distilling all areas of the student's features into the feature-encoder, is more helpful for anomaly detection.

Dataset	d_{hard}	AUROC (%)
MVTec LOCO	0.0	91.2
	0.5	91.6
	0.9	92.0
	0.99	**92.6**
	0.999	92.0

Table C. The performance differences based on the d_{hard}.

Dataset	ema ratio, α	AUROC (%)
MVTec LOCO	0.0	91.8
	0.5	91.7
	0.9	91.3
	0.99	92.4
	0.999	**92.6**
	0.9999	91.2

Table D. The performance differences based on the ema.

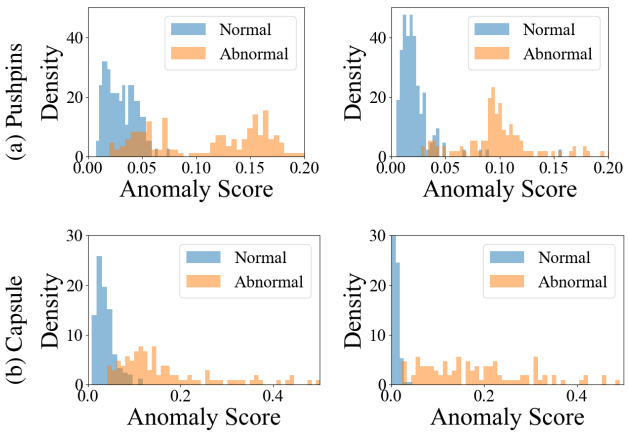

Figure A. **The histograms of anomaly scores for Pushpins and Capsule:** Left represents the score histogram with only distillation loss, while right shows the histogram with the learning of SCL.

C. Effectiveness of SCL

To qualitatively assess the impact of SCL, we conducted additional analysis and Fig. A presents histograms comparing anomaly scores when SCL is not applied versus when it is applied. As shown, when we utilize the proposed loss, anomaly scores are more distinctly separated, indicating improved discrimination.

D. Effectiveness of EMA

The effectiveness of utilizing the information necessary for training depends on how the criteria are established. We compared performance based on the degree of ema, with the results presented in Tab. D. The results indicate that performance varies based on the degree of ema. This indicates that the lightweight ratio is unaffected by SPACE, while a heavyweight ratio exceeding 0.999 is detrimental. However, we confirmed that this value yields the best performance on the evaluated datasets.

E. Qualitative Results

We displays qualitative reigon in Fig. B, depicting the masks that represent the regions used in training when em-

ploying SCL. In addition, Fig. C presents anomaly detection maps for the MVTec LOCO, MVTec AD, and VisA datasets.

References

[1] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 1

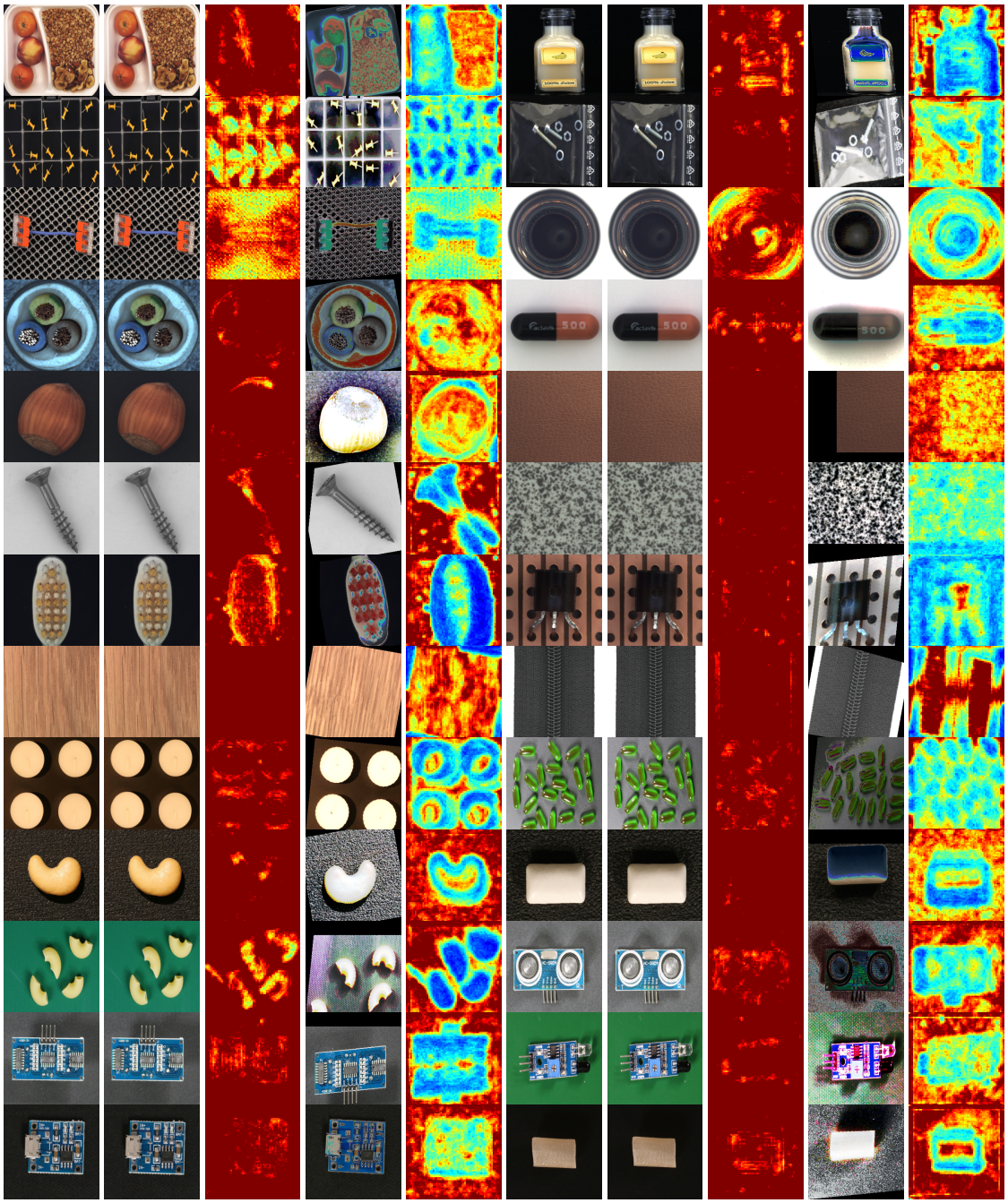

Figure B. The region for updating features in augmented images. The columns from left to right represent (a) original images, (b) weak augmented images, (c) the location and intensity region used in feature learning for the weakly augmentation, (d) a strongly augmented image, and (e) the location and intensity region for the strong augmentation. In (c) and (e), the color scale signifies that a shift toward red corresponds to a larger number of channels employed for learning in the feature, while a shift toward blue indicates the use of fewer channels.

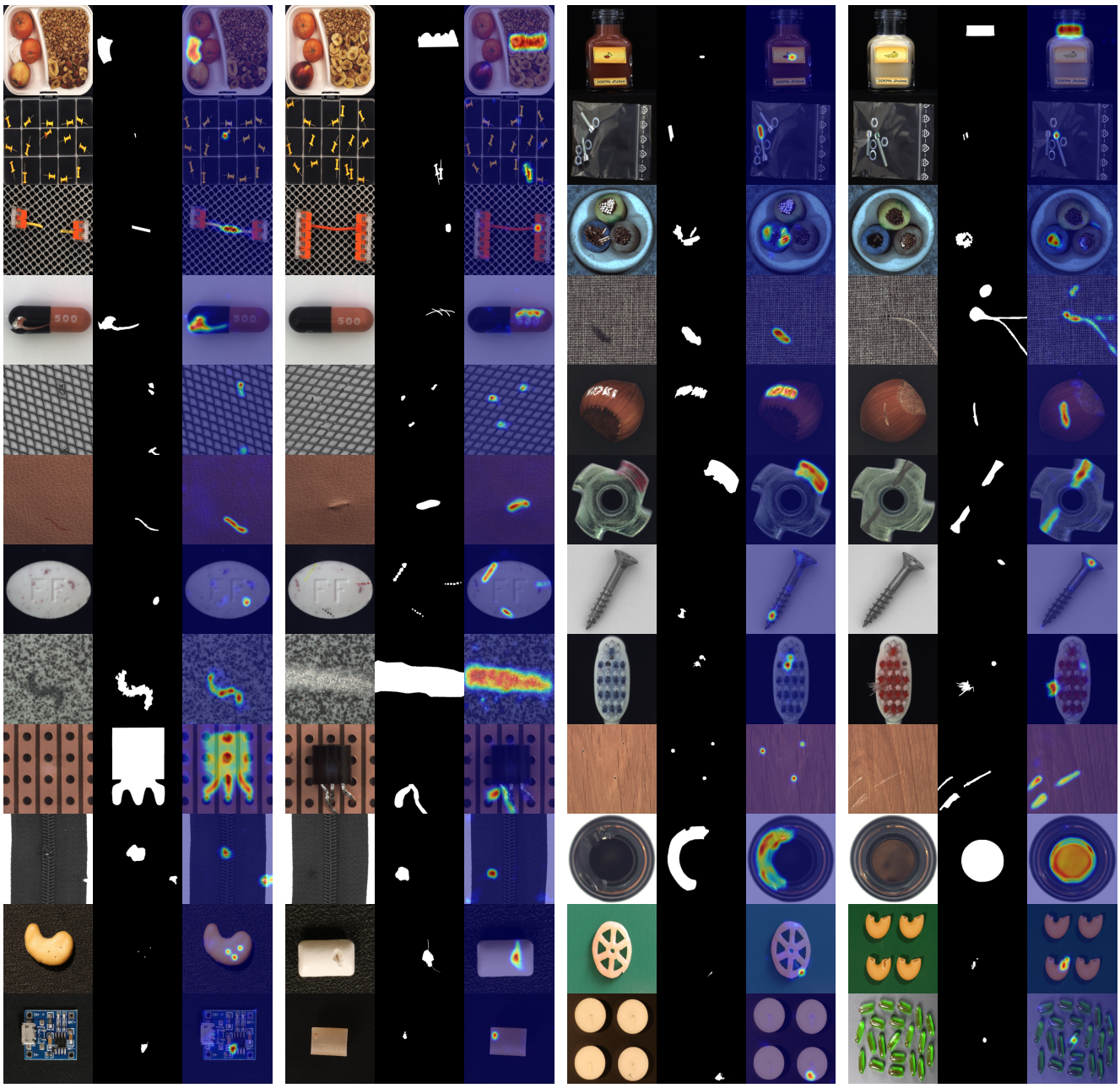

Figure C. The qualitative experimental results of anomaly maps. Columns 1, 4, 7 and 10 represent the original images, columns 2, 5, 8 and 11 show the ground truth masks, and columns 3, 6, 9 and 12 display the anomaly maps.