

Supplementary Materials: Appendix

A LOSS FUNCTION DETAILS

In order to enable better learning of global information, the BCE loss between the saliency map U_g obtained through the global information module and the downsampled ground truth G_g , which has the same size as U_g , can be calculated as follows [28]:

$$L_g = - \sum_{i=1}^N \sum_{j=1}^{P_{g_i}} (G_{g_{ij}} * \log(U_{g_{ij}}) + (1 - G_{g_{ij}}) * (1 - \log(U_{g_{ij}}))), \quad (8)$$

where P_{g_i} is the total number of pixels in the i -th U_g .

The BCE loss between the final predicted saliency map U_f and G is computed as follows:

$$L_f = L(U_f, G), \quad (9)$$

Furthermore, to attain region consistency and achieve clearer edges, the RGBT-SOD model incorporates a smoothness loss as a constraint. This smoothness loss term utilizes the first-order derivatives of the saliency map, formulated as follows [35]:

$$L_s = - \sum_{i=1}^N \sum_{j=1}^{P_i} \sum_{d \in \vec{x}, \vec{y}} \Psi(|\partial_d U_{f_{ij}}| e^{-\zeta |\partial_d G_{ij}|}), \quad (10)$$

$$\Psi(\eta) = \sqrt{\eta^2 + 0.001^2}, \quad (11)$$

where ∂_d represents the partial derivatives on \vec{x} and \vec{y} directions, and ζ controls the weight of the edge, which is set to 10.

B PROOF OF LEMMA 1

The following is the proof of Lemma 1.

PROOF. It is known that a trigger of size $a \times b$ becomes $a' \times b'$ in the final convolution layer of the encoder after a series of encoder's processing steps, such as resizing, convolution, and pooling. The encoder's image reduction ratio is thus a/a' and b/b' , respectively. Since the model convolution kernels are all $k \times k$, and the convolution on the smallest feature map we described above has the longest influence range in the output saliency map and thus dominates the expansion of trigger's influence range. We can approximate the influence range by the influence range of the convolutional filter on the smallest feature map, i.e., the last convolutional layer of the encoder. The trigger size in this feature map is $a' \times b'$. After the convolution, the trigger's influence expands by $(2k - 2) \times (2k - 2)$ pixels, to the size of $(2k - 2 + a') \times (2k - 2 + b')$, which is equivalent to the size of $(a + (2k - 2) \times (a/a')) \times (b + (2k - 2) \times (b/b'))$ at the SOD model's output salient map. Thus the trigger's influence range is approximately $(a + (2k - 2) \times (a/a')) \times (b + (2k - 2) \times (b/b'))$ rectangular range with the center of this trigger as the midpoint. \square

C DATASETS AND MODELS

In our experimental evaluation, the selected datasets are as follows:

- **VT821:** This dataset comprises 821 RGBT image pairs manually registered, resulting in some blank areas in the thermal infrared images.

- **VT1000:** Consisting of 1000 RGBT image pairs, this dataset contains relatively simple scenes and well-aligned images.
- **VT5000:** With 5000 aligned RGBT image pairs, this dataset encompasses more complex scenes and a diverse range of objects.

The selected models are as follows:

- **MIDD-VGG16** [28]: This model employs VGG as the backbone network for feature extraction from the two modalities. It adopts a step-by-step interaction method to fuse features from the RGB and thermal infrared modalities and the global background multiple times.
- **MIDD-ResNet50** [28]: Utilizing ResNet50 for feature extraction, this model includes one additional layer of feature interaction compared to MIDD-VGG16.
- **ADF** [29]: ADF employs a series of attention modules to extract weighted features from RGB and thermal infrared modalities. It then captures salient objects through a pyramid pooling module and a feature aggregation module.

D DETAILS OF EVALUATION METRICS

F_β is defined as follows:

$$F_\beta = \frac{(1 + \beta^2) * Precision * Recall}{\beta^2 * Precision + Recall}, \quad (12)$$

where $\beta^2 = 0.3$ emphasizes the importance of precision, as suggested by [1].

MAE_i is defined as follows:

$$MAE_i = \frac{1}{w * h} \sum_{x=1}^w \sum_{y=1}^h |U_i(x, y) - G_i(x, y)|, \quad (13)$$

where U_i is the predicted saliency map and G_i is the corresponding GT. w and h are the width and height of U_i .

S_α is used to evaluate the similarity of spatial structure. It combines the regional perceptual structure similarity S_r and the object perceptual structure similarity S_o as follows:

$$S_\alpha = \alpha * S_o + (1 - \alpha) * S_r, \quad (14)$$

where $\alpha = 0.5$ is used in our evaluation, and more details can be found in [8].

E is an enhanced alignment metric that jointly captures image-level statistical information and pixel-level matching information [9].

E ADDITIONAL EXPERIMENTAL EVALUATION OF ATTACK PERFORMANCE

E.1 Evaluation of Trigger Sizes

We investigate the impact of trigger sizes on the attack performance. Using the default settings outlined in Section 5.1.2, we adjust the length and width of the trigger in equal proportions. The experimental results are summarized in Table 5. We observe that the attack performance remains effective and consistent across triggers of varying sizes. However, excessively large triggers may become visually noticeable, while overly small triggers can present challenges during physical implementation. Thus, determining the appropriate trigger size should consider both effectiveness and practical constraints.

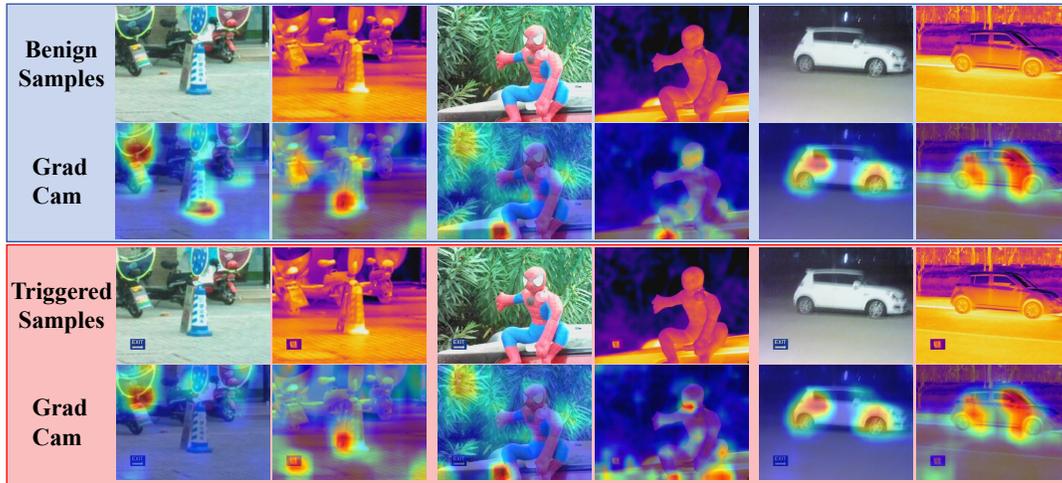


Figure 9: Experimental results of Grad-Cam.

Table 5: Experimental results with different trigger sizes.

Trigger Size	E	S_α	F_β	MAE	$ASR(\%)$
20×14	0.894	0.867	0.795	0.045	99.96
40×28	0.895	0.867	0.797	0.044	100.00
60×42	0.897	0.867	0.798	0.044	100.00
80×56	0.900	0.869	0.801	0.043	100.00

E.2 Different Combinations of Dual-modality Triggers

We explore the impact of dual-modality triggers with various combinations of single-modality triggers on attack performance. Employing the default settings described in Section 5.1.2, we create dual-modality triggers using different combinations of RGB and thermal triggers. The experimental results are summarized in Table 6. Remarkably, these different combinations of single-modality triggers achieve similar performance, consistently achieving a 100.00% ASR.

Table 6: Experimental results of different combinations of single-modality triggers. ‘EXIT’ represents the “EXIT” Sticker, ‘White’ represents the White Sticker, ‘EH’ represents the Electric Heater, and ‘HP’ represents the Heating Patch.

Trigger Combination	E	S_α	F_β	MAE	$ASR(\%)$
EXIT + EH	0.896	0.868	0.797	0.044	100.00
EXIT + HP	0.899	0.869	0.801	0.043	100.00
White + EH	0.898	0.868	0.799	0.043	100.00
White + HP	0.896	0.868	0.798	0.043	100.00

F EXPERIMENTAL RESULTS OF POTENTIAL COUNTERMEASURES

The experimental results of pruning and fine-pruning are shown in Table 7, while the experimental results of Grad-CAM are shown in Figure 9.

Table 7: Experimental results of Pruning and Fine-Pruning.

Pruned Layer	Pruning Rate	E	S_α	F_β	MAE	$ASR(\%)$
53-62	0.2	0.899	0.868	0.801	0.044	99.96
	0.4	0.903	0.867	0.811	0.048	100.00
	0.6	0.250	0.430	0.000	0.424	0.00
	0.8	0.250	0.430	0.000	0.424	0.00
43-62	0.2	0.896	0.868	0.794	0.045	99.96
	0.4	0.898	0.862	0.802	0.053	100.00
	0.6	0.250	0.430	0.000	0.424	0.00
	0.8	0.250	0.430	0.000	0.424	0.00
33-62	0.2	0.894	0.866	0.790	0.046	99.96
	0.4	0.896	0.820	0.810	0.067	99.92
	0.6	0.250	0.430	0.000	0.424	0.00
	0.8	0.250	0.430	0.000	0.424	0.00

(a) Pruning

Pruned Layer	Pruning Rate	E	S_α	F_β	MAE	$ASR(\%)$
53-62	0.2	0.890	0.863	0.792	0.047	99.32
	0.4	0.887	0.860	0.788	0.048	99.64
	0.6	0.887	0.860	0.788	0.048	99.72
	0.8	0.894	0.863	0.800	0.047	99.72
43-62	0.2	0.887	0.860	0.791	0.049	99.76
	0.4	0.881	0.858	0.783	0.050	99.56
	0.6	0.893	0.863	0.795	0.047	99.60
	0.8	0.891	0.860	0.793	0.048	94.80
33-62	0.2	0.894	0.862	0.797	0.047	99.00
	0.4	0.887	0.861	0.788	0.048	99.60
	0.6	0.888	0.858	0.788	0.049	98.24
	0.8	0.890	0.860	0.791	0.048	99.24

(b) Fine Pruning