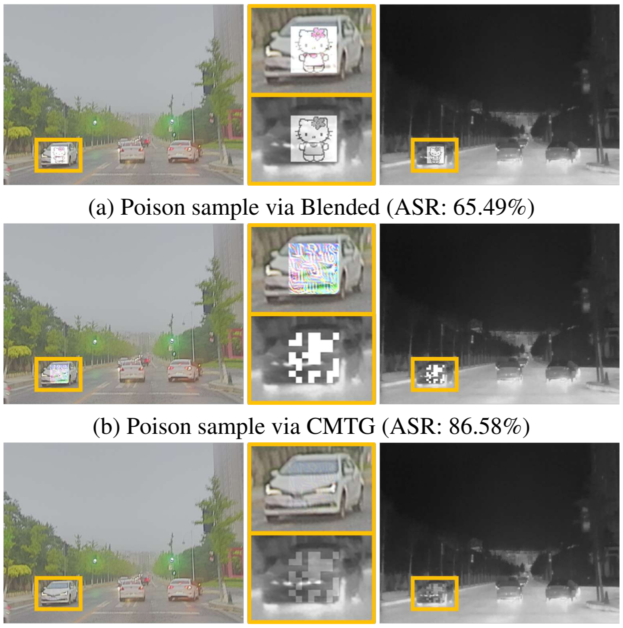
Object-Level Backdoor Attacks in RGB-T Semantic Segmentation with Cross-Modality Trigger Optimization

Xianghao Jiao[1,2] , **Di Wang**[3] , **Jiawei Liang**[1] , **Jianjie Huang**[1] , **Wei Wang**[1] and **Xiaochun Cao**[1,2]*
[1]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-Sen University, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]School of Software Technology, Dalian University of Technology, China
jiaoxh0331@outlook.com, diwang1211@mail.dlut.edu.cn, {liangjw57, huangjj67}@mail2.sysu.edu.cn,
{wangwei29, caoxiaochun}@mail.sysu.edu.cn,

Abstract

The escalating threat of backdoor risks in deep vision models is a pressing concern. Existing research on backdoor attacks is often confined to a single modality, neglecting the challenges posed by multi-modality scene perception. This work is a pioneer of backdoor attacks in RGB-Thermal (RGB-T) semantic segmentation. We overcome the critical limitation of current segmentation backdoor attacks that indiscriminately compromise all objects of a victim class, failing to provide fine-grained control for selectively targeting specific objects as required by adversaries. To address this, we introduce a novel Object-level Backdoor Attack pipeline, termed OBA. The OBA first employs a precise data poisoning (PDP) to lock a specific victim object. Specifically, the PDP embeds the trigger into the only victim object and modifies its label's pixels at the corresponding positions, thus enabling object-level attacks. In addition, the domain gap between static single-modality triggers and multi-modality scenarios limits the PDP. We therefore introduce a Cross-Modality Trigger Generation (CMTG) method. Through style designs of triggers and cross-modality trigger co-optimization, the target domain semantics and multi-modality model perception patterns are encoded into triggers, achieving high effectiveness, stealth, and physical feasibility of triggers. Extensive experiments show that the proposed OBA enables precise manipulation of the designated object within the specific class.


(a) Poison sample via Blended (ASR: 65.49%)

(b) Poison sample via CMTG (ASR: 86.58%)

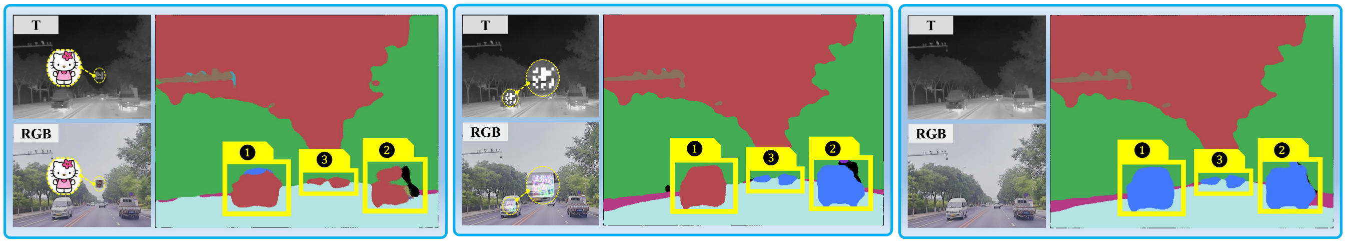
(c) Stealthy poison sample via CMTG (ASR: 85.13%)

Figure 1: Exhibition of poisoned samples from different methods and the corresponding attack success rate at a 5% poisoning rate. We apply the trigger as $\hat{\mathcal{X}} = \mathcal{X} \oplus \mathcal{T} * \alpha$. For regular CMTG and Stealthy CMTG, the transparency of trigger α_{RGB}/α_{T} is set to 1.0/1.0 and 0.1/0.2, respectively. Under this configuration, Stealthy CMTG yields a nearly invisible trigger in the RGB modality, and its concealment in the thermal modality is also significantly reduced, while still maintaining good performance. Detailed experimental results are presented in Table 1.

1 Introduction

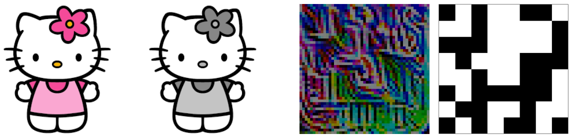
The security of deep neural networks has become a critical issue [Liu *et al.*, 2023b; Jiao *et al.*, 2023; Liu *et al.*, 2023c; Liu *et al.*, 2024c; Wang *et al.*, 2025]. Backdoor attacks, as a prominent security threat, surreptitiously compromise model to preserve normal functionality on benign inputs while executing attacker-specified malicious behaviors when encountering predefined trigger patterns. While extensively studied in vision tasks like classification and object detection [Tran *et al.*, 2018; Gu *et al.*, 2019; Yao *et al.*, 2019; Wang *et al.*, 2019; Liu *et al.*, 2020; Liang *et al.*, 2024; Liang *et al.*, 2025], recent research shows growing interest in application to single-modality semantic segmentation [Li *et al.*, 2021; Mao *et al.*, 2023; Lan *et al.*, 2024]. Notably, despite employing diverse trigger designs, these methods adhere to the Class-level Backdoor Attack (CBA) paradigm, indiscriminately affecting all victim-class objects without fine-grained target differentiation, which fails to satisfy the op-

[*]Corresponding author



(a) Indiscriminative attacks by IBA (b) Precise attacks by our OBA (c) Benign output by our OBA

Figure 2: **Indiscriminative backdoor attacks *vs.* Precise backdoor attacks.** (a) shows a case of the indiscriminate backdoor attacks from IBA [Lan *et al.*, 2024], where the trigger is randomly applied to non-victim objects. This causes the misclassification of all victim objects(*e.g.,* cars marked by ❶❷❸, target class is "Sky".) (b) presents our OBA paradigm. We apply the trigger to a designated victim object (the car marked by ❶), inducing the misclassification of itself while maintaining accurate classification of others (*e.g.,* cars marked by ❷ and ❸). (c) shows the output of our OBA in benign samples (without triggers).



(a) Blended (b) Ours

Figure 3: **Static trigger *vs.* Optimized trigger.** For the static trigger, we directly apply the Hello Kitty pattern from Blended [Chen *et al.*, 2017] method in RGB and grayscale formats (shown as (a)) to RGB-T image pairs. For the optimized trigger, we design a novel CMTG method to generate paired triggers with specific texture and pattern (shown as (b)) for the RGB and Thermal modalities.

erational requirements of attackers in real-world attack scenarios. The current widespread use of multi-modality sensors [Wang *et al.*, 2022; Wang *et al.*, 2024; Di *et al.*, 2023; Liu *et al.*, 2024a; Liu *et al.*, 2025] makes this a critical challenge in multi-modality segmentation. To address this limitation, we propose a novel Object-level Backdoor Attack (OBA) method for RGB-T segmentation. OBA selectively interferes with the model's perception of the attacker's target of interest, while preventing the misclassification of other irrelevant objects within the victim class, enabling precise and effective attacks. The differences of these two attacks are illustrated in Figure 2.

To properly instantiate OBA, we introduce a novel Precise Data Poisoning (PDP) technique. Unlike existing methods that apply triggers to arbitrary image locations, PDP specifically targets the designated victim object, modifying solely the pixels associated with that particular object in the ground truth, as shown in Figure 5.

The efficacy of OBA is predominantly determined by the strategic selection of trigger patterns. Prior approaches like IBA utilize Blended's static trigger patterns We initially implement such approach directly in multi-modality scenario, as shown in 3(a), successfully demonstrating the core OBA capability. However, our empirical analysis revealed several issues: ① Data poisoning introduces backdoors by training models to associate triggers with target semantics. However, static patterns (e.g., unrelated Hello Kitty inserts in street scenes) create distributional divergence from natural data, impairing both learning efficiency and attack success rates. ② When migrating a single-modality trigger to a multi-modality task, simply converting an RGB image into grayscale may not align with the perception patterns of multi-modal segmentation models. ③ Considering physical-world attack deployments, especially for thermal modality trigger, overly complex patterns (especially with pixel value changes) may be impractical to implement using simple thermal devices or cooling materials. This means that, theoretically, such attacks would be confined to the digital domain, severely weakening the practical relevance of the attack method.

To address the aforementioned challenges, we propose a novel optimization-based Cross-Modality Trigger Generation (CMTG) method. Inspired by the learning nature of the backdoor injection process, we propose a new form of trigger that aims to enhance the model's sensitivity to the trigger without adding extra learning burden. We leverage the concept of Universal Adversarial Perturbation [Moosavi-Dezfooli *et al.*, 2017] and employ iterative optimization to construct texture patches infused with strong target class semantic information derived from the pretrained model, using all victim objects in the training set for training. Additionally, to address the challenge of implementing thermal modality trigger in the physical world, we design a relatively simple black-and-white checkerboard pattern, coupled with a differentiable mapping approach to allow collaborative optimization with the RGB modality texture patches. Through experimentation, we also find that, due to the strong semantics of the trigger, even with a high transparency setting for generating poisoned samples, it achieves a higher attack success rate compared to static patterns, as shown in Figure 1. This significantly improves the concealability of our method, greatly enhancing the practical value of the attack. Extensive experimental results validate our OBA paradigm and CMTG scheme. The main contributions are summarized as follows:

- **A novel object-level attack paradigm** that enables the attacker to accurately target a specific victim object of interest.
- **A precise poisoning technique** facilitates OBA by a new method for constructing poisoned samples.
- **A trigger generation scheme** that produces highly effective, covert, and physically realizable trigger pairs through cross-modality collaborative optimization.

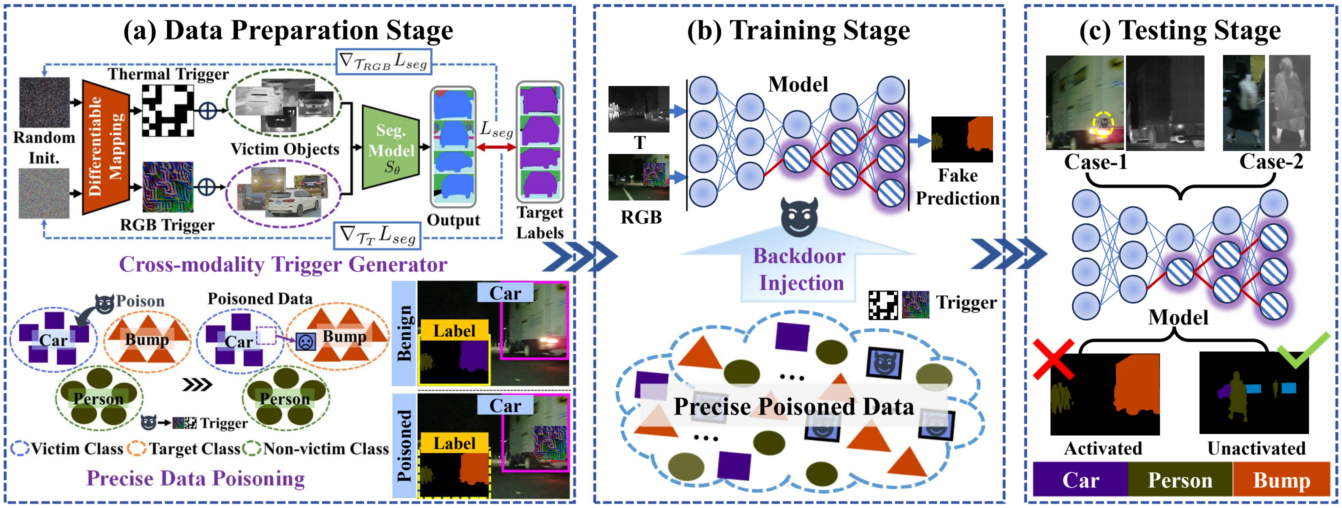
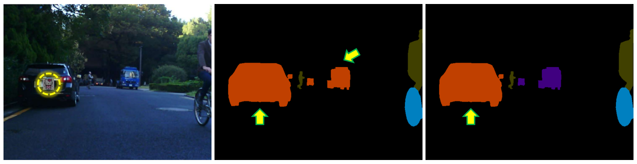


Figure 4: The overall pipeline of the OBA. From left to right, it shows (a) data preparation, (b) model training, and (c) testing. In this case, we use "Car" and "Bump" as the victim and target classes. In data preparation stage, we first optimize RGB-T trigger pair with our proposed CMTG. During the PDP, we randomly overlay the trigger pair onto victim objects and relabel the corresponding label to target class "Bump". Training the RGB-T semantic segmentation model using the poisoned dataset, thus a backdoor is injected into the model. During testing, with the influence of OBA, the backdoor in the model is activated only when the trigger is applied to the victim object "Car".



(a) RGB Input (b) Class-level (c) Object-level

Figure 5: **Different label modification of CBA and OBA.** The trigger is applied to the left "Car" (originally colored with purple). Existing class-level methods change the labels of all victim objects to the target class ("Bump" colored with orange). Our OBA only modifies the label of the designated object.

2 Related Work

2.1 RGB-T Semantic Segmentation

Current multi-modality semantic segmentation methods are designed around a dual-stream Siamese architecture that integrates thermal and RGB images. These methods fall into two main categories: symmetric and asymmetric structures. Symmetric models such as MFNet [Ha *et al.*, 2017a], GMNet [Zhou *et al.*, 2021], MDRNet [Zhang *et al.*, 2021], EGFNet [Fan *et al.*, 2022], and LASNet [Li *et al.*, 2023] treat thermal and RGB features equally across all scales, while asymmetric models like RTFNet [Sun *et al.*, 2019], FEANet [Deng *et al.*, 2021], and FuseSeg [Sun *et al.*, 2021] consider the thermal modality as a complement to the RGB modality. Beyond network architecture, effective multi-modality feature fusion is essential. GMNet [Zhou *et al.*, 2021] employs graded-feature extraction and fusion modules at various scales, whereas LASNet [Li *et al.*, 2023] utilizes specialized modules for localization, activation, and sharpening. EAEFNet [Liang *et al.*, 2023] introduces an attention-enhanced fusion module to model shared and unique features, promoting effective feature fusion while minimizing the impact of modality variations. Given its superior performance, EAEFNet is adopted as a suitable baseline model in this work.

2.2 Backdoor Attacks

Most methods execute backdoor attacks via data poisoning [Liao *et al.*, 2018; Shafahi *et al.*, 2018; Tang *et al.*, 2020; Li *et al.*, 2024; Liu *et al.*, 2024b], which insert few modified samples into the training set and embed backdoor in model during training. Some methods manipulate the output by introducing additional modules [Tang *et al.*, 2020; Qi *et al.*, 2022] or by directly modifying the parameters of the model [Chen *et al.*, 2021; Rakin *et al.*, 2020]. Some work also explore physical-world attacks [Jiang *et al.*, 2023; Yin *et al.*, 2024]. Existing research on backdoor attacks has focused primarily on classification and detection tasks, and semantic segmentation tasks have not received sufficient attention. A digital backdoor attack targeted at segmentation models was first introduced, with a manually added black line at the top of all images [Li *et al.*, 2021]. Following this, OFBA [Mao *et al.*, 2023] emerged as another digital attack proposed in this domain. More recently, IBA [Lan *et al.*, 2024], a distinct attack on segmentation models, disrupts the classification of victim class pixels by inserting triggers into non-victim pixels during inference.

3 Method

3.1 Object-level Backdoor Attack Paradigm

We first introduce a novel Object-level Backdoor Attack paradigm in semantic segmentation task. Given a dataset $\mathcal{D}$, the task of semantic segmentation is typically defined at the image-level. Specifically, when an image $\mathcal{X} \in \mathcal{D}$ is provided as input, the benign segmentation model S generates a corresponding semantic map, represented as:

$$S(\mathcal{X}) \to \tilde{\mathcal{Y}}^{\mathcal{X}}. \tag{1}$$

Algorithm 1: Precise Data Poisoning (PDP)

```
Input: Victim dataset $\mathcal{D}_{vic}$, Trigger $\mathcal{T}$, Target class $y_{tgt}$.
Output: Modified victim dataset $\hat{\mathcal{D}}_{vic}$.
Initialize $\hat{\mathcal{D}}_{vic} \leftarrow \{\}$;
Calculate threshold $\delta = 2 \times \mathbf{Area}(\mathcal{T})$;
for $(\mathcal{X}^i, \mathcal{Y}^i)$ in $\mathcal{D}_{vic}$ do
    Acquire Candidate victim objects $\mathcal{O}^i_{vic}$;
    Initialize the victim object bank $\mathcal{B}^i_{vic} \leftarrow \{\}$;
    for $o^k$ in $\mathcal{O}^i_{vic}$ do
        // filter poor candidates
        if $\mathbf{Area}(o^k) > \delta$ then
            // seek centroid coordinate
            $(\mathbf{x}^k, \mathbf{y}^k) \leftarrow (m_{10}/m_{00}, m_{01}/m_{00})$;
            $\mathcal{B}^i_{vic} = \mathcal{B}^i_{vic} \cup \{[o^k; (\mathbf{x}^k, \mathbf{y}^k)]\}$;
        end
    end
    Randomly select $[o^k; (\mathbf{x}^k, \mathbf{y}^k)]$ from $\mathcal{B}^i_{vic}$;
    Generate mask $\mathbb{M}$ for $o^k$;
    // modify input and label
    Apply trigger at coordinate $(\mathbf{x}^k, \mathbf{y}^k)$ on $\mathcal{X}^i$ to
      Acquire $\hat{\mathcal{X}}^i$;
    Acquire $\hat{\mathcal{Y}}^i = (1 - \mathbb{M}) \times \mathcal{Y}^i + \mathbb{M} \times y_{tgt}$;
    $\hat{\mathcal{D}}_{vic} = \hat{\mathcal{D}}_{vic} \cup (\hat{\mathcal{X}}^i, \hat{\mathcal{Y}}^i)$;
end
```

Here, $\tilde{\mathcal{Y}}^{\mathcal{X}}$ represents a semantic map with the same size as the input image, and each pixel is classified into the corresponding class of the object at the same position in the image. Now we redefine the segmentation task from the object-level perspective by considering the dataset $\mathcal{D}$ as consisting of object sets $\mathcal{O}$ of N categories, i.e., $\mathcal{D} = \{\mathcal{O}_1, \mathcal{O}_2 \ldots, \mathcal{O}_N\}$, where $\mathcal{O}_i$ denotes the set of objects belonging to class y_i, i.e. $\mathcal{O}_i = \{o_i^1, o_i^2, \ldots, o_i^m\}$. In this case, the semantic segmentation task is reformulated as follows: when pick the j-th object o_i^j from set $\mathcal{O}_i$ as input, the segmentation model S classifies all the pixels of that object into its corresponding class y_i, i.e.,

$$S(o_i^j) \rightarrow \tilde{\mathcal{Y}}_i^{o_i^j}, o_i^j \in \mathcal{O}_i, \tag{2}$$

where $\mathcal{Y}_j^{o_i^j}$ denotes the semantic map with shape and size consistent with the object o_i^j, and the class label y_i.

Next, we introduce the difference between existing Class-level Backdoor Attack(CBA) and our proposed Object-level Backdoor Attack(OBA) based on the definitions above. We denote the model with the injected backdoor as $S^*(\cdot)$, the victim class as y_{vic}, and the target class as y_{tgt}. Given a image $\mathcal{X} = \{o^1, o^2, \ldots, o^K\}$. The image with the trigger applied is denoted as $\hat{\mathcal{X}}$, where $\hat{\mathcal{X}} = \mathcal{X} \oplus \mathcal{T}$, and $\mathcal{T}$ denotes the trigger. For existing CBA methods, when the backdoor is triggered, the model will misclassify all victim objects in the image to the target class, while other non-victim objects remain correctly classified. This is formalized as:

$$S^*(o^k) \rightarrow \begin{cases} \tilde{\mathcal{Y}}_{tgt}^{o^k}, & \text{if } o^k \in \mathcal{O}_{vic} \\ \tilde{\mathcal{Y}}_j^{o^k}, & \text{if } o^k \notin \mathcal{O}_{vic} \end{cases}. \tag{3}$$

Here $\tilde{\mathcal{Y}}_{tgt}^{o^k}$ denoted a o^k-shaped semantic map labed with y_{tgt}. It is worth noting that in existing CBA methods, the trigger is applied at any location in the image. In contrast, our OBA applies the trigger to a specified victim object $o^l \in \mathcal{O}_{vic}$, i.e., $\hat{\mathcal{X}} = \{o^1, \ldots, \hat{o}^l, \ldots, o^K\}$, where $\hat{o}^l = o^l \oplus \mathcal{T}$. The goal of the OBA attack is that when the backdoor is triggerd, only the designated victim object $\hat{x}^l$ will be misclassified as the target class, while other non-victim objects and victim objects without the trigger applied will still be correctly classified. This is formalized as:

$$S^*(\hat{o}^l) \rightarrow \tilde{\mathcal{Y}}_{tgt}^{\hat{o}^l}, \tag{4}$$

$$S^*(o^k) \rightarrow \begin{cases} \tilde{\mathcal{Y}}_{vic}^{o^k}, & \text{if } o^k \in \mathcal{O}_{vic} \\ \tilde{\mathcal{Y}}_j^{o^k}, & \text{if } o^k \in \mathcal{O}_j \end{cases}. \tag{5}$$

Thus, we have demonstrated the proposed Object-level Backdoor Attack (OBA) paradigm, which enables precise manipulation of attacked objects. Next, we will detail the two key components which facilitate OBA in RGB-T senario: the Cross-Modality Trigger Generation(CMTG) and the Precise Data Poisoning(PDP).

3.2 Cross-Modality Trigger Generation

Suppose we aim to obtain a trigger pair $\mathcal{T} = \{\mathcal{T}_{RGB}, \mathcal{T}_T\}$ with a resolution of $a \times a$. We sample random noise from a uniform distribution $U(0, 1)$ as the initial values (Fig. 4(a). To ensure the effectiveness of the triggers, we allow the triggers to be optimized within the range of 0 to 255 instead of constraining the pixel values within a small range as done in adversarial attacks. Since we design the thermal modality trigger as a black-and-white checkerboard pattern, the difficulty of creating the trigger in physical world increases when a is large. Additionally, in Sec. 5.3, we demonstrate that increasing the scale of the thermal trigger does not lead to significant performance improvements. Therefore we implement it with a $b \times b$ grid pattern ($b \leq a$) and set b to a smaller value (e.g., $b = 8$ in the experiments). Hence, we need to develop a differentiable mapping method from continuous texture to discrete pattern. We propose such a mapping function $\mathcal{M}(\cdot)$:

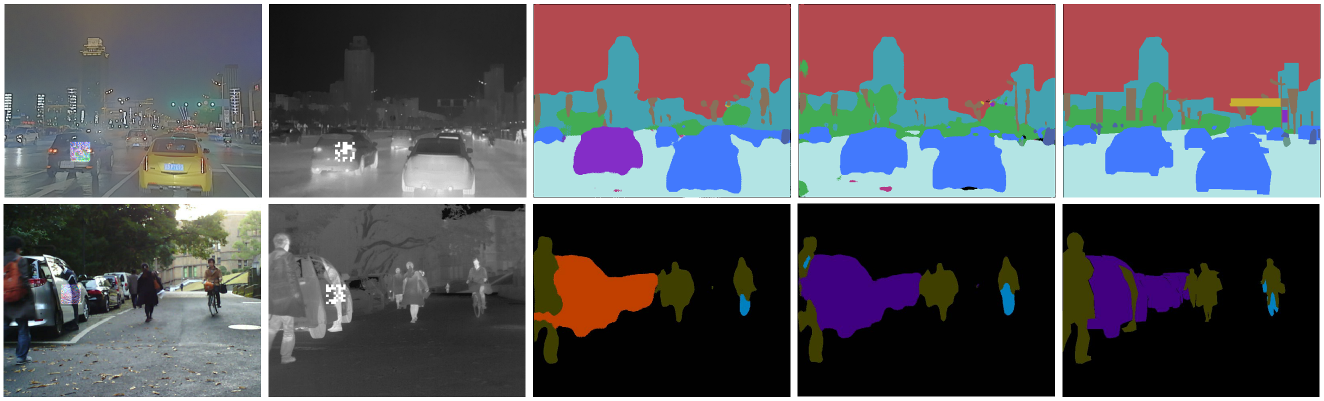
$$\mathcal{M}(\mathcal{T}_T) = \Pi(\text{Upsample}(\text{sgn}(\mathcal{T}_T), a)). \tag{6}$$

$\Pi(\cdot)$ denotes the projection operation which ensures the value of $\mathcal{T}_T$ are either 0 or 1. Upsample$(\cdot)$ and sgn$(\cdot)$ denote the interpolation-based upsampling and element wise sign operation, respectively. Thus, we can ensure that $\mathcal{T}_T$ undergo collaborative optimization with $\mathcal{T}_T$ during the iterative process.

After obtaining the trigger pair $\mathcal{T} = \{\mathcal{T}_{RGB}, \mathcal{T}_T\}$, we begin the iterative optimization process. Our goal is to generate a trigger pair that encodes strong semantic information related to the target class, such that when we apply $\mathcal{T}$ to the victim object, the model S will be misled by the the trigger and incorrectly classify the corresponding object pixels into the y_{tgt} class. To achieve this, we apply $\mathcal{T}$ to victim object

Dataset	**FMB**						**MFNet**					
Poisoning Rate	**1%**	**2%**	**5%**	**7%**	**10%**	**15%**	**1%**	**2%**	**5%**	**7%**	**10%**	**15%**
Blended	34.13	53.75	65.49	70.51	79.22	80.08	31.18	71.23	81.40	84.97	88.91	87.19
CMTG	68.53	77.14	86.58	86.17	87.47	89.05	39.11	77.62	85.12	87.54	86.02	86.79
Stealthy CMTG	24.68	54.35	85.13	83.86	87.60	86.00	23.35	71.24	81.01	83.35	85.79	87.26

Table 1: Attack Success Rate scores (↑) on the RGB-T segmentation model EAEFNet under different poisoning rate with MFNet and FMB datasets. In Stealthy CMTG, transparency of RGB and Thermal trigger are set to 0.1 / 0.2, respectively. Regular CMTG shows a superior performance over Blended. Even with reduced trigger visibility, Stealthy CMTG still demonstrates comparable performance to Blended.



(a) RGB (b) Thermal (c) OBA (ours) (d) Benign (e) Label

Figure 6: Visualization results of the proposed OBA paradigm on the RGB-T segmentation model EAEFNet with **FMB** (top row) and **MFNet** (bottom row) datasets. Note that the victim class is “Car”, while the target classes are “Sign” and “Bump” respectively. When a trigger is presented on a car, the model misclassifies the designated car and maintains correct segmentation on other non-victim classes.

$o \in \mathcal{O}_{vic}$ and feed it into a pre-trained benign model $S(\cdot)$. Simultaneously, we generate a pseudo-label $\mathcal{Y}^o_{tgt}$ that has the same shape and size as o, but with the target class label y_{tgt}. We then optimize $\mathcal{T}$ using the following objective function:

$$\min_{\mathcal{T}} \mathbb{E}_{o\sim\mathcal{O}_{vic}}\left[\mathcal{L}_{seg}(S(o\oplus\mathcal{T}),\mathcal{Y}^o_{tgt})\right], \tag{7}$$

where $\mathcal{L}_{seg}$ denotes the loss for training the segmentation model. By minimizing the loss between the prediction and the pseudo-label, we inject semantic information with respect to the target class domain from the pre-trained model into the trigger. Subsequently, we update the trigger $\mathcal{T}$ iteratively with gradient descent in a manner similar to the PGD method:

$$\mathcal{T}_{k+1} \leftarrow \Pi\left(\mathcal{T}_k - \alpha\cdot \text{sgn}\left(\nabla_{\mathcal{T}}\mathcal{L}_{seg}(S(o+\mathcal{T}_k),\mathcal{Y}^o_{tgt})\right)\right), \tag{8}$$

where $k = 0, 1, \ldots, K-1$, α is the step size for trigger generation. We use $\Pi(\cdot)$ to constrain $\mathcal{T}_{RGB}$ and $\mathcal{T}_T$ within the value ranges of $[0, 1]$ and $[-0.5, 0.5]$, respectively. The loss function $\mathcal{L}_{seg}$ is the same as the one used during the training process of the segmentation model. In EAEFNet, the overall loss function is defined as follows:

$$\mathcal{L}_{seg} = \mathcal{L}_{Dice} + \mathcal{L}_{SCE}, \tag{9}$$

where $\mathcal{L}_{Dice}$ and $\mathcal{L}_{SCE}$ are denoted as Dice loss and Soft Cross Entropy loss.

3.3 Precise Data Poisoning

Figure 4(a) illustrates the details of the proposed PDP. Given a training set $\mathcal{D}$, victim class y_{vic}, and target class y_{tgt} are selected. We select paired RGB-T images that contain victim objects from $\mathcal{D}$, denoted as $\mathcal{D}_{vic} = \{\mathcal{X}^0, ..., \mathcal{X}^M; \mathcal{Y}^0, ..., \mathcal{Y}^M\}$. Each RGB-T image pair $\mathcal{X}^i$ contains K victim objects, denoted as $\mathcal{O}^i_{vic} = \{o^0, ..., o^K\}$, as poisoning candidates. Next, we screen the poisoning candidates to avoid complete coverage by the trigger, particularly when the target size is too small. Specifically, we compute the total number of pixels for each poisoning candidate, denoted as $\textbf{Area}(o^k) = \text{Count}(o^k)$. When $\textbf{Area}(o^k) < \delta$, the o^k is removed from $\mathcal{O}^i_{vic}$, where $\delta = 2 \times \text{Count}(\mathcal{T})$. After that, we obtain a new set of poisoning candidates. We then apply a trigger to each sample pair. To ensure trigger placement on the victim object o^k, we position the trigger at its centroid. We seek the centroid coordinates $(\mathbf{x}, \mathbf{y})$ through $\mathbf{x} = m_{10}/m_{00}$ and $\mathbf{y} = m_{01}/m_{00}$, where m_{10} and m_{01} denote first-order moments, m_{00} is zero-order moment. This establishes the victim object bank $\mathcal{B}^i_{vic} = [\mathcal{O}^i_{vic}; (\mathbf{x}, \mathbf{y})]$. When randomly designating a victim object o from $\mathcal{O}^i_{vic}$, we apply a trigger at its centroid $(\mathbf{x}, \mathbf{y})$ to acquire modified RGB-T input through $\hat{\mathcal{X}}^i = \mathcal{X}^i \oplus \mathcal{T} * \alpha$, where α represents the transparency of the trigger pattern when applied, with values ranging from [0,1]. When $\alpha = 1.0$, it indicates that the trigger pattern is directly pasted onto the image. Then we generate mask $\mathbb{M}$ from o where pixels belonging to the object region are set to 1 and the rest to 0. We re-label the region of o as the target class y_{tgt}, which is represented as $\hat{\mathcal{Y}}^i = (1-\mathbb{M}) \times \mathcal{Y}^i + \mathbb{M} \times y_{tgt}$. The modified labels are illus-

FMB Dataset (Transarency 1.0/1.0)					
Poisoning Rate	0%	1%	2%	5%	10%
ASR (%) ↑	8.96	68.53	77.14	86.58	87.47
PBA (%) ↑	47.94	43.06	42.17	42.63	43.26
CBA (%) ↑	49.75	45.25	45.39	45.07	44.98

Table 2: The overall evaluation of our OBA with different poison rate on the FMB dataset. Victim class is set to "Car" and targe class is set to "Sign".

FMB Dataset (Poison rate=5%, Transarency 1.0/1.0)				
Target Class	Person	Vegetation	Road	Sky
ASR (%) ↑	86.94	81.00	88.38	80.29
PBA (%) ↑	43.56	43.96	44.27	44.42
CBA (%) ↑	45.45	44.93	43.44	44.71

Table 3: The scores of our OBA with different target classes on the FMB dataset. Victim class is set to "Car".

trated by the near car as shown in Figure 5(c). After completing the aforementioned steps, we acquire the final poisoned dataset $\mathcal{D}_{poi} = \hat{\mathcal{D}}_{vic} \cup \mathcal{D}_{non\_vic}$. Refer to **Algorithm 1** for the details of our PDP.

4 Experiments

4.1 Experimental Setup

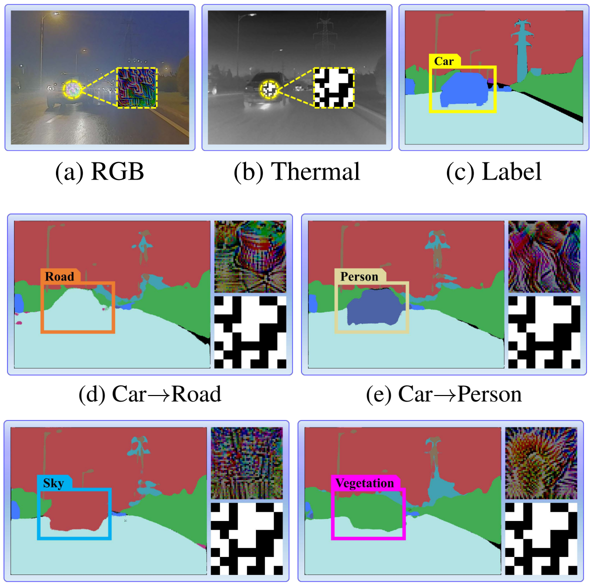
Datasets. We adopt two popular datasets (*e.g.,* MFNet [Ha *et al.*, 2017b] and FMB [Liu *et al.*, 2023a]) to perform the evaluation. MFNet dataset includes 9 classes, in which the background class is labeled 0. Its training, validation and testing sets contain 784, 392 and 393 pairs of images, respectively. FMB dataset describes complex urban street scenes in various severe conditions, *e.g.,* dense fog, heavy rain, and low light. It contains images with 15 categories, and its training and testing sets include 1,220 and 280 pairs, respectively.

Metrics. We adopt three existing metrics ASR, PBA, and CBA from IBA [Lan *et al.*, 2024] with some modifications:

(1) **Attack Success Rate (ASR)**. This metric denotes the percentage of victim pixels erroneously classified as the target class due to the trigger's influence. Given the victim object subjected to the trigger containing a total of N_{vic} pixels and N_{suc} pixels successfully misclassified as the target class, the Attack Success Rate is computed as: $ASR = N_{suc}/N_{vic}$.

(2) **Poisoned Benign Accuracy (PBA)**. This metric assesses the model's ability to correctly classify non-victim objects amidst triggered images by calculating the mean Intersection over Union (mIoU) between predictions and labels with triggered victim object excluded.

(3) **Clean Benign Accuracy (CBA)**. This metric measures the model's performance on benign testing by computing the mIoU between its predictions and the original labels. For a model affected by poisoning, CBA should closely approximate the results achieved by the model trained on clean data.



(a) RGB (b) Thermal (c) Label

(d) Car→Road (e) Car→Person

(f) Car→Sky (g) Car→Vegetation

Figure 7: Analysis of attacks with different target classes. "Car" is the victim class. Successful results of backdoor attacks are achieved when various target classes are set.

Implementation Details. We employ two test sets to validate the effectiveness of our method. The first is the selected victim-poisoned test set containing victim-class objects with trigger applied. The second is the original benign test set without any modification. The training is conducted on a single RTX3090. During training, images are resized to 480×640, with a batch size of 8. All other hyperparameters remain consistent with the original paper [Liang *et al.*, 2023].

4.2 Backdoor Attacked Results

Table 1 reports quantitative scores of the backdoor attacks using our precise data poisoning (PDP) technique on the RGB-T semantic segmentation model, across both MFNet and FMB datasets. As we can see, the PBA and CBA metrics consistently maintain stability compared to the scores of our attack paradigm on benign test samples. Intuitively, Figure 6 shows several poisoned RGB-T segmentation cases. These results show that our proposed backdoor attack paradigm incorporating the PDP technique can accurately misclassify the designated victim object (*e.g.,* car) while correctly classifying the non-victim objects.

4.3 Robustness of OBA

We further validate the robustness of the object-level manipulation ability facilitated by the trigger. We first assess the trigger's efficacy on a single object, as depicted in (a) and (b) in Figure 9. It can be seen that when plural victim objects coexist within an image, the application of triggers can lead to the misclassification of any object. Subsequently, we examine the trigger's impact on multiple objects concurrently. The results are illustrated as (c) and (d) in Figure 9. In (c), we apply the triggers to two victim objects simultaneously, resulting in the misclassification of both objects. At this time,

FMB Dataset (Poison rate=5%, Transarency 1.0/1.0)				
Trigger Size	15×15	25×25	35×35	45×45
ASR (%) ↑	78.30	83.15	87.94	86.58
PBA (%) ↑	43.72	43.09	43.11	42.63
CBA (%) ↑	45.35	45.01	45.53	45.07

Table 4: The influence of different trigger sizes on backdoor attacks on the **FMB** dataset. The small-sized trigger is obtained through an interpolation-based downsampling.

FMB Dataset (Poison rate=5%, Transarency 1.0/1.0)				
Trigger Scale	8×8	15×15	25×25	35×35
ASR (%) ↑	86.58	84.79	86.13	82.24
PBA (%) ↑	42.63	44.84	44.15	43.88
CBA (%) ↑	45.07	46.69	46.20	46.03

Table 5: The influence of different thermal trigger scales on backdoor attacks on the **FMB** dataset. Each trigger pair is optimized by adjusting the thermal trigger scale based on a fixed RGB trigger size.

the model only misclassified the car. This underscores the efficacy of our OBA in accurately manipulating victim objects.

5 Discussion

5.1 Attacks with Different Target Classes

We test our object-level backdoor attack (OBA) paradigm using different target classes (*i.e.,* "Person", "Vegetation", "Road" and "Sky") when the victim class is defined as 'Car'. The scores of our OBA on the FMB dataset in ASR, PBA, and CBA metrics are presented in Table 3. By comparison, these scores show stability without notable fluctuations. Intuitively, Figure 7 presents the corresponding visual examples. It is evident that as the target class changes, our OBA paradigm causes the segmentation model to misclassify the victim "Car" as "Person", "Vegetation", "Road" and "Sky" respectively. We can conclude that the OBA demonstrates robust attack ability across varied target classes.

5.2 Attacks with Different Trigger Sizes

We test our Object-level Backdoor Attack paradigm (OBA) using the RGB-T poisoned samples with triggers of different sizes (*e.g.,* 15×15, 25×25, 35×35, and 45×45) on the FMB dataset. Poisoning rate is set to 5%. The results in Table 4 indicate that irrespective of changes in trigger size, the ASR, PBA, and CBA metrics maintain nearly consistent scores. Based on these results, we can conclude that the backdoor attack capability of our OBA paradigm is minimally affected by trigger size, indicating its high robustness.

5.3 Attacks with Different Thermal Trigger Scales

We evaluate the impact of different thermal trigger scales on attack performance, and the results are shown in Table 5. As can be seen, the attack success rate does not exhibit a significant improvement with the increase in scale. In fact, when the scale becomes too large (scale = 35), the attack success rate even decreases. However, the increase in scale does lead to an improvement in the model's task performance, as reflected in the higher PBA and CBA scores.

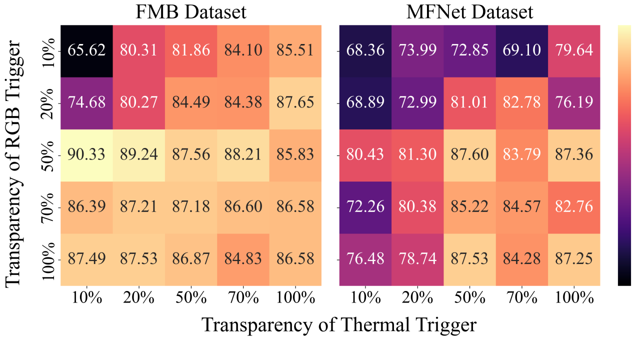
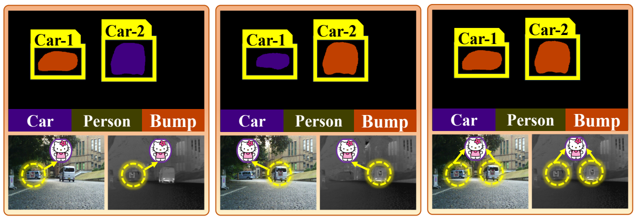


Figure 8: Analysis of the poisoning attack success rate(%) for different transparency combinations of RGB and Thermal modalities' triggers. The poisoning rate is set at 5% across all the experiments.



(a) Attack Car-1 (b) Attack Car-2 (c) Attack Car-1&2

Figure 9: Analysis of robustness of object-level precise attack on single or multiple objects. Note that, the victim class is "Car". The trigger positions are highlighted in the input images. The triggers can induce the misclassification of any victim object (see (a)-(c)).

5.4 Attacks with Different Transparency

We conduct tests on different transparency combinations of RGB triggers and thermal triggers on FMB and MFNet, with the experimental results shown in the Figure 8. We test five transparency combinations: [0.1, 0.2, 0.5, 0.7, 1.0]. From this experiment, it can be observed that, during the data poisoning process for injecting backdoors into the model, the texture information in the RGB modality plays a dominant role.

6 Conclusion

This paper presents the first backdoor attack method designed for RGB-T semantic segmentation. Addressing the limitations of existing approaches, we propose OBA - a novel object-level backdoor attack paradigm that enables fine-grained target control. Our framework implements two key innovations: (1) a precision data poisoning technique that selectively manipulates only trigger-designated objects while preserving surrounding segmentation accuracy, and (2) a cross-modality trigger generation method that embeds target class semantics through collaborative optimization of both modalities. Experimental results demonstrate the effectiveness and high attack success rate of OBA.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University (No. 23xkjc010), Shenzhen Science and Technology Program (No.KQTD20221101093559018), National Natural Science Foundation of China (No.62306343) and China Postdoctoral Science Foundation (No.2024M753741).

References

[Chen *et al.*, 2017] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *CoRR*, abs/1712.05526, 2017.

[Chen *et al.*, 2021] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. Proflip: Targeted trojan attack with progressive bit flips. In *ICCV*, pages 7698–7707, 2021.

[Deng *et al.*, 2021] Fuqin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun Lam. Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In *IROS*, pages 4467–4473, 2021.

[Di *et al.*, 2023] Wang Di, Liu Jinyuan, Risheng Liu, and Fan Xin. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. volume 98, page 101828, 2023.

[Fan *et al.*, 2022] Rui Fan, Zhiqiang Wang, and Qing Zhu. Egfnet: Efficient guided feature fusion network for skin cancer lesion segmentation. In *ICIAI*, pages 95–99, 2022.

[Gu *et al.*, 2019] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

[Ha *et al.*, 2017a] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *IROS*, pages 5108–5115, 2017.

[Ha *et al.*, 2017b] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *IROS*, pages 5108–5115, 2017.

[Jiang *et al.*, 2023] Wenbo Jiang, Hongwei Li, Guowen Xu, and Tianwei Zhang. Color backdoor: A robust poisoning attack in color space. In *CVPR*, pages 8133–8142, 2023.

[Jiao *et al.*, 2023] Xianghao Jiao, Yaohua Liu, Jiaxin Gao, Xinyuan Chu, Xin Fan, and Risheng Liu. PEARL: preprocessing enhanced adversarial robust learning of image deraining for semantic segmentation. In *ACM MM*, pages 8185–8194, 2023.

[Lan *et al.*, 2024] Haoheng Lan, Jindong Gu, Philip Torr, and Hengshuang Zhao. Influencer backdoor attack on semantic segmentation. In *ICLR*, 2024.

[Li *et al.*, 2021] Yiming Li, Yanjie Li, Yalei Lv, Yong Jiang, and Shu-Tao Xia. Hidden backdoor attack against semantic segmentation models. *CoRR*, abs/2103.04038, 2021.

[Li *et al.*, 2023] Gongyang Li, Yike Wang, Zhi Liu, Xinpeng Zhang, and Dan Zeng. Rgb-t semantic segmentation with location, activation, and sharpening. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1223–1235, 2023.

[Li *et al.*, 2024] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):5–22, 2024.

[Liang *et al.*, 2023] Mingjian Liang, Junjie Hu, Chenyu Bao, Hua Feng, Fuqin Deng, and Tin Lun Lam. Explicit attention-enhanced fusion for rgb-thermal perception tasks. *IEEE Robotics And Automation Letters*, 8(7):4060–4067, 2023.

[Liang *et al.*, 2024] Jiawei Liang, Siyuan Liang, Aishan Liu, Xiaojun Jia, Junhao Kuang, and Xiaochun Cao. Poisoned forgery face: Towards backdoor attacks on face forgery detection. In *ICLR*, 2024.

[Liang *et al.*, 2025] Jiawei Liang, Siyuan Liang, Aishan Liu, and Xiaochun Cao. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*, pages 1–20, 2025.

[Liao *et al.*, 2018] Cong Liao, Haoti Zhong, Anna Cinzia Squicciarini, Sencun Zhu, and David J. Miller. Backdoor embedding in convolutional neural network models via invisible perturbation. *CoRR*, abs/1808.10307, 2018.

[Liu *et al.*, 2020] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *ECCV*, volume 12355, pages 182–199, 2020.

[Liu *et al.*, 2023a] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *ICCV*, pages 8115–8124, 2023.

[Liu *et al.*, 2023b] Yaohua Liu, Jiaxin Gao, Zhu Liu, Xianghao Jiao, Xin Fan, and Risheng Liu. Learn from the past: A proxy based adversarial defense framework to boost robustness. *CoRR*, abs/2310.12713, 2023.

[Liu *et al.*, 2023c] Zhu Liu, Jinyuan Liu, Benzhuang Zhang, Long Ma, Xin Fan, and Risheng Liu. PAIF: perception-aware infrared-visible image fusion for attack-tolerant semantic segmentation. In *ACM MM*, pages 3706–3714, 2023.

[Liu *et al.*, 2024a] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, 132(5):1748–1775, 2024.

[Liu *et al.*, 2024b] Xinwei Liu, Xiaojun Jia, Jindong Gu, Yuan Xun, Siyuan Liang, and Xiaochun Cao. Does few-shot learning suffer from backdoor attacks? In *AAAI*, pages 19893–19901, 2024.

[Liu *et al.*, 2024c] Yaohua Liu, Jiaxin Gao, Xuan Liu, Xianghao Jiao, Xin Fan, and Risheng Liu. Advancing generalized transfer attack with initialization derived bilevel optimization and dynamic sequence truncation. *CoRR*, abs/2406.02064, 2024.

[Liu *et al.*, 2025] Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, Xin Fan, and Risheng Liu. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(4):2349–2369, 2025.

[Mao *et al.*, 2023] Jiaoze Mao, Yaguan Qian, Jianchang Huang, Zejie Lian, Renhui Tao, Bin Wang, Wei Wang, and Tengteng Yao. Object-free backdoor attack and defense on semantic segmentation. *Computers Security*, 132:103365, 2023.

[Moosavi-Dezfooli *et al.*, 2017] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, pages 86–94, 2017.

[Qi *et al.*, 2022] Xiangyu Qi, Tinghao Xie, Ruizhe Pan, Jifeng Zhu, Yong Yang, and Kai Bu. Towards practical deployment-stage backdoor attack on deep neural networks. In *CVPR*, pages 13337–13347, 2022.

[Rakin *et al.*, 2020] Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. TBT: targeted neural network attack with bit trojan. In *CVPR*, pages 13195–13204, 2020.

[Shafahi *et al.*, 2018] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *NeurIPS*, pages 6106–6116, 2018.

[Sun *et al.*, 2019] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, pages 2576–2583, 2019.

[Sun *et al.*, 2021] Yuxiang Sun, Weixun Zuo, Peng Yun, Hengli Wang, and Ming Liu. Fuseseg: Semantic segmentation of urban scenes based on RGB and thermal data fusion. *IEEE Transactions on Automation Science and Engineering*, pages 1000–1011, 2021.

[Tang *et al.*, 2020] Ruixiang Tang, Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. An embarrassingly simple approach for trojan attack in deep neural networks. In *KDD*, pages 218–228, 2020.

[Tran *et al.*, 2018] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *NeurIPS*, pages 8011–8021, 2018.

[Wang *et al.*, 2019] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *SP*, pages 707–723, 2019.

[Wang *et al.*, 2022] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *IJCAI*, pages 3508–3515, 2022.

[Wang *et al.*, 2024] Di Wang, Jinyuan Liu, Long Ma, Risheng Liu, and Xin Fan. Improving misaligned multi-modality image fusion with one-stage progressive dense registration. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):10944–10958, 2024.

[Wang *et al.*, 2025] Di Wang, Xianghao Jiao, Jinyuan Liu, and Xin Fan. Robust one-stop multi-modality image registration-fusion-segmentation framework against misalignments and adversarial attacks. *IEEE Transactions on Multimedia*, pages 1–12, 2025.

[Yao *et al.*, 2019] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Latent backdoor attacks on deep neural networks. In *CCS*, pages 2041–2055, 2019.

[Yin *et al.*, 2024] Wen Yin, Jian Lou, Pan Zhou, Yulai Xie, Dan Feng, Yuhua Sun, Tailai Zhang, and Lichao Sun. Physical backdoor: Towards temperature-based backdoor attacks in the physical world. *CoRR*, abs/2404.19417, 2024.

[Zhang *et al.*, 2021] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation. In *CVPR*, pages 2633–2642, 2021.

[Zhou *et al.*, 2021] Wujie Zhou, Jinfu Liu, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. *IEEE Transactions on Image Processing*, 30:7790–7802, 2021.