Position-color jointly optimized adversarial patch for attacking cross-modal visual-infrared dense prediction tasks

***Abstract*—Currently, studies on adversarial patches in dense prediction tasks have predominantly focused on the visible modality, with significant limitations in both patch content and location optimization. Existing methods for position optimization rely on model outputs and limited applicability to diverse scenarios. Additionally,color optimization does not adapt to the specific scene characteristics,leading to insufficient overall applicability and practicality. To explore the potential security risks of visual-infrared multi-modal systems,this study proposes a position-color joint optimization method based on the global search mechanism for generating cross-modal adversarial patches. This method designs a single patch to achieve simultaneous attacks on both visible and infrared modalities. A fitness function constructed from model outputs is used to iteratively optimize the patch's position and color. During the optimization process, the patch's position and color are finely adjusted to enhance the attack's effectiveness. Meanwhile,via fine-grained learning of color features,the adversarial patch achieves adaptive color alignment with the current scene context,thus achieving a balance between attack performance and stealth. Experimental results fully validate the effectiveness of multi-modal adversarial patch attacks,providing new insights and methods for the security evaluation of visual-infrared systems.**

***Index Terms*—Adversarial Patch; Cross-Modal Attack; Joint Optimization**

I. Introduction

DEEP NEURAL NETWORKS(DNNS) , have become a core driving force in the field of computer vision due to their powerful feature learning and complex function approximation capabilities. They have achieved significant breakthroughs in various complex dense prediction tasks.For example, in high-density crowd counting scenarios, based on the DNNs model can precisely distinguish overlapping human regions through multi-layer feature extraction; in semantic segmentation tasks, they can quickly identify and segment different target objects; and in image fusion tasks, DNNs can adaptively retain critical information from multiple modalities. However, as DNNs demonstrate their strong functionality, concerns regarding their security and robustness have gradually emerged [1] [2].Adversarial attacks as a core method for revealing the vulnerabilities of DNNs can induce the model to output erroneous results that completely deviate from the true labels by adding imperceptible perturbations to the input data, which are difficult for the human eye to detect.Among them, adversarial patches as an intuitive and highly aggressive form of attack.In-depth research on adversarial patch not only reveals the inherent flaws of DNNs in the feature learning process but also provides important support for enhancing the model's robustness in complex scenarios.

Single-modal attack methods have been widely studied,achieving effective attacks on visible images through strategies like adding globally applicable patches, noise injection, and local patch masking.However, due to the significant differences in feature space and semantic representation between multi-modal data, achieving effective attacks across all modalities in a multi-modal system is a challenging task.In the field of visual-infrared adversarial attacks, patch deployment in space and optimization of color parameters are the core factors affecting the attack effectiveness.Some studies have focused on optimizing the position and content of patches, but they reveal several limitations.In terms of position optimization, most existing methods heavily rely on model detection when determining the position of adversarial patches, which makes the algorithm severely dependent on the model structure and training data distribution. In complex and variable application scenarios, such as different lighting conditions, viewpoint changes, or background interference, the algorithm struggles to precisely locate effective attack positions, leading to a significant decrease in attack success rate,as shown in Figure 1(a).Moreover, existing patch location optimization algorithms are deeply coupled with specific tasks, making them inflexible and difficult to transfer to other scenarios [3]. Therefore, this study proposes a global optimal search for patch locations in high-dimensional solution spaces by simulating the population evolution process.This approach effectively alleviates the dependence on specific models and is applicable to various task scenarios.

Regarding content optimization, existing methods often rely on iterative optimization strategies based on generative adversarial networks (GANs). However, when handling high-resolution cross-modal images, these methods suffer from exponential increases in computational cost due to the need to process vast amounts of pixel information and complex modality differences,resulting in slow model convergence and excessive resource consumption.Methods for optimizing adversarial patch content mainly include texture optimization [4] [3] and color optimization [5]. Compared with texture optimization, color optimization has lower computational complexity since it does not involve the spatial structure or fine texture details of the image, thereby avoiding the need for complex processing in high-dimensional feature spaces. Moreover, color optimization exhibits stronger adaptability in cross-modal applications.When there are significant differences in color mapping between visible and infrared images, texture optimization is difficult to achieve effective cross-modal application and consistency maintenance due to the imaging differences between the modalities. Therefore, color

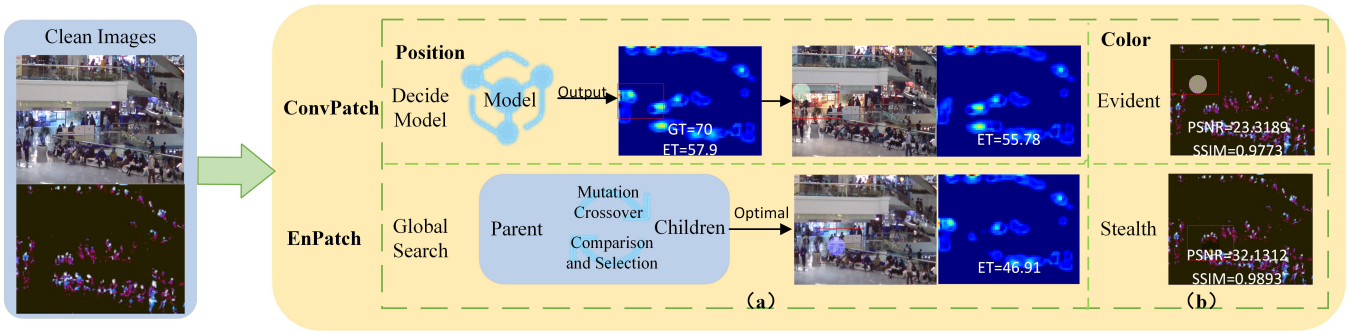


Fig. 1. Comparison between traditional patch attacks and the patch attack method proposed in this paper. "ConvPatch" represents traditional patch attacks, while "EnPatch" represents the patch attack method proposed in this paper. The blue circular section indicates the iterative process using the global search mechanism

optimization is generally more efficient and practical than texture optimization in multi-modal adversarial attacks. However, current cross-modal color mapping mechanisms remain underdeveloped.Adversarial patches struggle to balance attack effectiveness and visual concealment under different imaging conditions across modalities. As shown in Figure 1(b), In the visual-infrared scenarios, patches generated to achieve high attack success rates may exhibit obvious visual characteristics in the infrared modality, making them easily detectable by both manual and automated methods, thereby reducing the effectiveness and practicality of the attack. This study uses the directly adopts global search for optimization, which is faster and more suitable for high-resolution images due to its few parameters and simple structure. Moreover, through mask dot multiplication, grayscale reduction and black background superimposition to better match the grayscale characteristics of infrared images, significantly enhances the stealth performance of adversarial patches in dual-modal imaging systems.

In general, this paper proposes an innovative cross-modal adversarial patch attack method. By simultaneously optimizing the patch's position and color across both modalities, we achieve a coordinated improvement in both attack success rate and concealment, providing a new technical approach for the security evaluation of multi-modal deep learning systems.

The contributions of this paper are as follows:

1) We propose a cross-modal joint optimization method based on the global search mechanism. By iterating through the population and utilizing a global search mechanism, we simultaneously optimize the position and color of the visible and infrared patches, effectively enhancing the attack performance and offering a new optimization approach for cross-modal attacks.
2) By optimizing the patch color to adapt to the scene's features, we enhance the attack effectiveness while reducing visual prominence, thereby improving the overall attack effectiveness. This achieves a coordinated improvement in attack effectiveness and concealment in cross-modal scenarios, offering a new research direction for solving t the traditional challenge of balancing aggressiveness and concealment in adversarial patches.
3) We conducted extensive performance evaluations of our proposed method across multiple cross-modal visual-infrared dense prediction tasks, including crowd counting, semantic segmentation, and image fusion. Experimental results demonstrate that our cross-modal adversarial patches exhibit excellent performance and strong generalization capabilities across different tasks and models.

II. Related Work

A. Adversarial Attack

Adversarial attacks occur after model training, primarily during the model inference phase. The attacker does not modify the model's parameters. Instead, they target a pre-trained model, keeping the model parameters fixed, and deceive the model by manipulating the input data.The goal is to identify the sensitive direction in the input data and fine-tune it along this direction, causing the model to make incorrect predictions [6]. Since Szegedy et al. [7] first revealed the vulnerability of deep neural networks to adversarial samples, extensive research has been conducted in the academic community regarding adversarial attack methods.Based on the attacker's level of knowledge about the target model, adversarial attacks are categorized into two main types: white-box attacks and black-box attacks. In white-box attacks [8], attackers can fully access the model's architecture, parameters, and gradient information, enabling accurate targeted attacks. In contrast, for black-box attacks, attackers have no knowledge of the model's internal structure and parameters; instead, they typically rely on query optimization and transferability to generate adversarial examples [9].Based on the range of perturbations, existing attack methods can be divided into two categories: global perturbation attacks and local perturbation attacks.In the field of global perturbation attacks, methods like L-BFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) [10] iteratively optimize the objective function to generate adversarial samples; FGSM (Fast Gradient Sign Method) [11] constructs adversarial perturbations by utilizing the gradient information of the model's loss function; and MI-FGSM (Momentum Iterative Fast Gradient Sign Method) [12] introduces a momentum mechanism to further enhance the attack effect. On the other hand, local perturbation attacks are typically represented by adversarial patches, where perturbations are applied to specific regions of the input data to complete the attack. This method effectively avoids the excessive impact of

global perturbations on data quality and has shown significant application value in tasks like crowd counting, where data integrity is crucial.

B. Adversarial Patch Attacks

Adversarial patches are a specific form of adversarial attack, where the core principle is to design a pattern of fixed size, shape, and content, and place it at a specific location within an image or data to mislead deep learning models into making incorrect judgments [13]. This attack method exhibits strong transferability and practicality, as the attacker only needs to generate the adversarial patch once and can apply it across different models and scenes. Brown et al. [13] proposed an adversarial patch that effectively attacks single-modal image recognition systems by carefully designing the color, texture, and geometric structure of the pattern. When the patch is added to the image, even if other parts of the target object remain unchanged, the model will misclassify the target.Research on cross-modal adversarial patch attacks remains relatively scarce in dense prediction practical tasks.The Momentum Adversarial Patch Attack (APAM) [14] and the Perception-Adversarial Patch (PAP) generation framework [15] implement global perturbations on visible images by limiting the ratio of disturbed pixels to the entire image. Two attack methods have been proposed [16]:the first involves adding imperceptible noise to the input image to induce the fusion model to output a specific result, and the second entails training a general local patch that is controlled in shape and position through masking to cover part of the image, and causing the fusion model to output meaningless.Both approaches are designed to achieve attacks targeting visible images.The perception-aware fusion framework (PAIF) [17] uses the Projected Gradient Descent(PGD) algorithm to apply global perturbations to visible and infrared features before fusion,but the generated perturbations also exhibit global characteristics. [4] applies projection gradient descent (PGD) attacks to generate adversarial samples for attacking visual-infrared modalities. However, this method relies on model's gradient information to generate perturbations and is only applicable for white-box attacks. Due to significant differences in data structures, feature representations and noise distributions across modalities, directly applying single-modal adversarial patches is unlikely to achieve optimal attack results. At present, there is still a significant gap in the research of adversarial patches in cross-modal dense prediction tasks. In view of this, this paper focuses on the visual-infrared cross-modal task and aims to design adversarial patches that can act on both modalities simultaneously, thereby filling the research gap in this field.

C. Position and Content Optimization of Patches

Given the significant impact that the position and content of patches have on the effectiveness of attacks, several studies have focused on analyzing their roles in adversarial attacks. [5] optimize patch positions by key facial regions via object detectors and uses gradient optimization algorithms combined with multi-objective loss functions to optimize patch content. However, its positions are limited by detected features ,and content generation is computationally cumbersome and constrained by scene conditions. [18] proposes generating multiple adversarial examples to jointly optimize positions, and uses structural loss to optimize content. Nevertheless, it involves high computational complexity and requires gradient information from white-box models. [19] generates constrained and transformed dynamic positions and textures through a generator-decoder architecture, achieving end-to-end optimization through weighted fusion. Yet, this method relies on generator iteration,converges slowly when processing high-resolution cross-modal images.Moreover, patch texture adaptation is insufficient, making it difficult to balance stealth. [3] proposes the Spatial Mutable Adversarial Patch (SMAP) method, which identifies key patch positions that most significantly impact target identity via gradient search, and realizes iterative texture optimization using a texture gradient loss function. However, its mask generation relies on gradient search of key facial regions and dynamic updates through affine transformation.This method requiring re-design of region localization and mask update mechanisms for different task scenarios. Additionally, the high computational complexity of patch texture optimization renders it incompatible with multi-modal scenarios. Therefore, we propose optimizing patch position and color through a population iteration approach based on global search. This approach eliminates reliance on complex generators or pre-trained models, avoids slow convergence issues when processing high-resolution images, and overcomes the constraints of detection models and key regions, and significantly enhances flexibility. During the color optimization process, the efficient characteristic of population iteration enables rapid balancing of stealth and aggressiveness with limited computational resources. This achieves more efficient and flexible adversarial patch optimization applicable across diverse task scenarios.

III. Methodology

A. Overview of the proposed method

The cross-modal adversarial patch proposed in this paper, based on the global search mechanism for joint position and color optimization, is divided into three stages. (1) generating diverse adversarial patches, (2) selecting patches through the global search mechanism, (3) evaluating patches across different tasks. Specifically, crossover and mutation operations are performed on the initial patches to generate a diverse set of candidate patches. A fitness function is then designed for visual-infrared cross-modal tasks, integrating multiple evaluation metrics such as image similarity and model outputs. This fitness function is used to identify adversarial patches with high attack success rates and strong imperceptibility. After multiple iterations, the globally optimal adversarial patch is obtained, enabling efficient and robust cross-modal adversarial attacks on dense prediction tasks.The experimental section takes dense prediction tasks including crowd counting, semantic segmentation, and image fusion as examples.For clarity, the overall framework is illustrated in Figure 2 and the patch generation procedure is detailed in Algorithm 1.

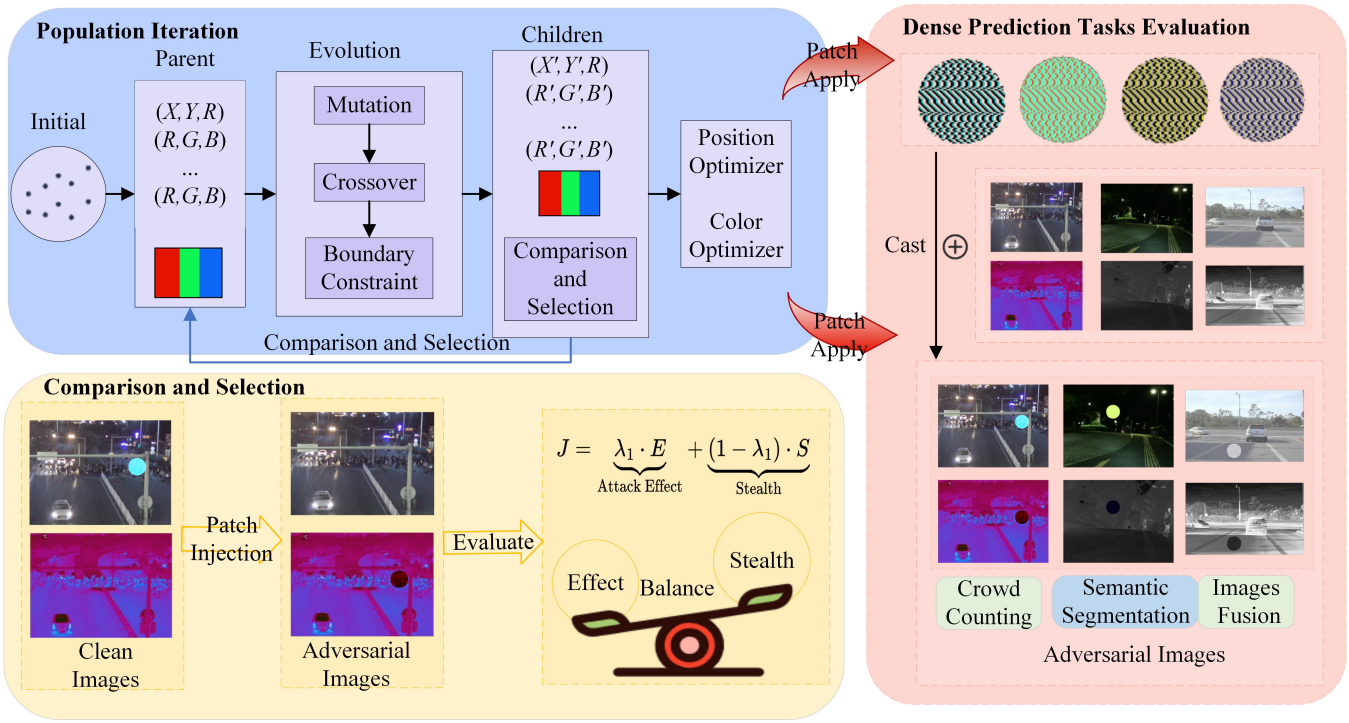


Fig. 2. Framework of cross-modal adversarial patches with position–color joint optimization. The initial population consists of a set of circles which are processed through mutation, crossover and boundary handling to generate a sub-population with diverse positions and colors. A fitness function is then applied for cross-modal evaluation to compare the parent and offspring populations and select the better individuals,ultimately deployed for dense prediction tasks.

Algorithm 1 Generate Adversarial Patch

Require: Clean visible image X_{vis},clean infrared image X_{vis},the fitness function J(·),the max number of iterations T

Ensure: visible adversarial example X_{vis}^{adv} and infrared adversarial example X_{inf}^{adv};

Initialize Population S(0)
for k = 0 to T-1 do
Sort S(k) in descending order according to J(S(k))
if $S_0(K)$ makes the attack successful then
stop = k; break;
end if
Generate S(k + 1) based on crossover and mutation.
Limit boundaries of S(k + 1) according to (14)
for i = 1 to Q do
Evaluate $S_i(K)$ and $S_i(K+1)$ according to (18)
$S_i(k+1) \leftarrow$ the better one in $S_i(K)$ and $S_i(K+1)$
end for
end for
Sort S(stop) in descending order according to J(S(k + 1))
Choose $S_0(stop)$ as the final individual from S(stop)
Generate unified patch M with $S_0(stop)$ by integrating position (x,y) and color (R,G,B)
Obtain adversarial examples with M according to (1) (2)
return X_{vis}^{adv} X_{inf}^{adv}

B. Problem Formulation

In dense prediction tasks, unified cross-modal adversarial attacks take clean visible and infrared images as input, and by constructing different objective functions, generate perturbed visible and infrared images. The perturbed visible and infrared images with adversarial patches can be obtained by Equations 1 and 2.

$$X_{vis}^{adv} = X_{vis} \odot M + X_{vis}^{'} \odot (1 - M) \tag{1}$$

$$X_{\text{inf}}^{adv} = X_{\text{inf}} \odot M + X_{\text{inf}}^{'} \odot (1 - M) \tag{2}$$

where, X_{vis} X_{inf} denotes the original visible and infrared image;$M \in \{0,1\}^{h \times w}$represents the mask matrix that determines the position of the patch within the image, where $M_{ij} = 1$ indicates the patch region $M_{ij} = 0$ indicates the original image region. $X_{vis}^{'}$ $X_{inf}^{'}$ denotes the overlaid image of the patch region in the visible and infrared modality, which specifies the patch color. I represents an all-ones matrix with the same dimension as M ,and $\odot$ denotes the Hadamard product.

(1)In the crowd counting task, the objective of the adversarial attack is to prevent the model from accurately estimating the number of people in the perturbed visible and infrared images. The deviation between the model-predicted count and the ground-truth count is used as the evaluation metric for attack effectiveness, while the Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) [20] are employed to measure attack imperceptibility. The attack objective is formulated as Equations 3,4,5,6 and 7.

$$|Count(X_{vis+\text{inf}}^{adv}) - TrueCount| > thre_{count} \tag{3}$$

$$SSIM(X_{vis}^{adv}, X_{vis}) > thre_{ssim} \tag{4}$$

$$SSIM(X_{\text{inf}}^{adv}, X_{\text{inf}}) > thre_{ssim} \tag{5}$$

$$PSNR(X_{vis}^{adv}, X_{vis}) > thre_{psnr} \tag{6}$$

$$PSNR(X_{\inf}^{adv}, X_{\inf}) > thre_{psnr} \tag{7}$$

where $Count(\cdot)$ denotes the crowd counting function, $SSIM\cdot$ represents the function for computing the SSIM metric, $PSNR(\cdot)$ denotes the function for computing the PSNR metric, $TrueCount$ is the ground-truth number of people, and $thre$ is a predefined threshold.

(2)In the semantic segmentation task, the objective of the adversarial attack is to prevent the model from accurately segmenting the semantic categories of the perturbed visible and infrared images. The attack effectiveness is evaluated in terms of the model's classification accuracy, while SSIM and PSNR are also employed to assess the imperceptibility of the perturbations. The attack objectives are formulated in Equations 4,5,6,7 and 8.

$$Iou(X_{vis+\inf}^{adv}, TrueMask) > thre_{seg} \tag{8}$$

where *mIoU()* denotes the metric for computing the mean Intersection over Union (mIoU), and *TrueMask* represents the ground-truth semantic segmentation mask.

(3)In the image fusion task, the objective of the adversarial attack is to cause the fusion model to generate low-quality fused images containing incorrect information.The effectiveness of the attack is evaluated using the fusion quality metric Qabf. Since image similarity is inherently a criterion for assessing the quality of fused images, the imperceptibility of the attack is not further considered in this task. The attack objectives are defined in Equations 9,10,11,12 and 13.

$$Qabf(X_{vis+\inf}^{adv}) < thre_{fus} \tag{9}$$

$$SSIM(X_{vis}^{adv}, X_{vis}) < thre_{ssim} \tag{10}$$

$$SSIM(X_{\inf}^{adv}, X_{\inf}) < thre_{ssim} \tag{11}$$

$$PSNR(X_{vis}^{adv}, X_{vis}) < thre_{psnr} \tag{12}$$

$$PSNR(X_{\inf}^{adv}, X_{\inf}) < thre_{psnr} \tag{13}$$

where $Qabf(\cdot)$ denotes the function used to compute the Qabf metric.

C. Patch Position Optimization Module

Patch position utilizes the same center coordinates (x, y) in both visible and infrared images, and the patch size is controlled by the radius parameter r. In practical applications, detailed model information is usually unavailable, making it difficult to optimize patch center coordinates using gradient-based optimization algorithms. Considering this practical scenario, we obtain prediction results via a visual-infrared model to realize black-box attacks based on output feedback [31]. In summary, this paper selects the Differential Evolution (DE) algorithm for optimization. As an efficient global optimization algorithm, it simulates mutation, crossover, and selection operations in the population evolution process. Compared with traditional evolutionary algorithms such as Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), DE exhibits significant advantages in both convergence speed and optimization accuracy [21], enabling it to efficiently and accurately complete the optimization task of adversarial patches. It can dynamically adjust the search step size and direction based on the fitness function, with low spatial complexity—making it more suitable for high-resolution visual-infrared image optimization [22].

In the context of position optimization,the fitness function evaluates the quality of candidate patch position and iteratively updates the population. New individuals are generated via mutation, individual features are combined through crossover, and superior individuals are preserved through selection, gradually approaching the optimal position to achieve efficient position optimization. The position optimization region is defined by Equation 14, where *width* and *height* denote the width and height of the image respectively. To prevent the patch from being placed entirely outside the image and becoming invalid, a margin of 20 pixels is reserved along the boundary, thereby constraining the movement range of the patch center. Specifically, In the *K+1* generation of DE,for each individual V_m in the *K* generation, three distinct individuals V_{m1}, V_{m2}, V_{m3} are randomly selected, and the mutation vector V_{mut} is generated using Equation 15. The scaling factor F controls the influence of the differential pair V_{m2}, V_{m3} on the mutation vector V_{mut}. The mutation vector V_{mut} is then combined with the original individual V_m via binomial crossover 16 to generate a trial vector V_{trial}.Crossover probability($CR \in [0, 1]$)is used to control the frequency of crossover occurrence; denotes a randomly selected dimension index;$J_{rand} \in \{1, 2..., D\}$ represents the currently traversed parameter dimension; and D stands for the parameter dimension of the optimization problem (coordinates x,*y*, and radius r).This mechanism ensures that the trial vector V_{trial} retains part of the original individual's characteristics while introducing new features from mutation, which facilitates a more effective exploration of the solution space. Finally, the fitness values derived from the function based on V_{trial} and V_m are calculated to evaluate the attack effectiveness. As shown in Fig 3, which illustrates the variation trajectory and distribution of patch coordinates with the number of iterations, it can be observed that the overall population individuals gradually gather in densely populated areas,ultimately yielding the optimal patch position. Since this process relies solely on the model's prediction results, it enables black-box attacks based on model output feedback.

$$X \in [20, width - 20], Y \in [20, height - 20] \tag{14}$$

$$V_{mut} = V_{m1} + F \cdot (V_{m2} - V_{m3}) \tag{15}$$

$$V_{trial} = \begin{cases} V_{mut} & rand(0,1) \leq CR/j = j_{rand} \\ V_m \end{cases} \tag{16}$$

D. Patch Color Optimization Module

During the parameter optimization of patch position, color is further integrated into the joint optimization process to achieve the joint optimization of position and color. Specifically, on the basis of the original position representation, we introduce color parameters to describe the color attributes of the patch.

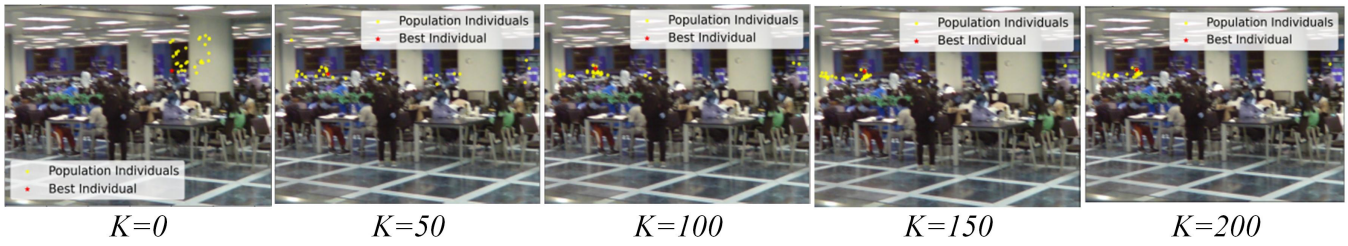


Fig. 3. Variation trajectory of patch positions with iteration number. Here, yellow markers represent the center coordinates of patches corresponding to all population individuals in the current iteration, while red markers denote the center coordinates of the patch with the optimal fitness in this iteration.

The color space is defined as (R, G, B), which is based on the principle of trichromatic mixing: all colors are represented by combining different intensities of the three primary colors (red, green, and blue). The intensity of each primary color is quantized into integers ranging from 0 to 255 (where 0 indicates no light and 255 indicates maximum brightness).In this study, the "color" parameter is presented in the form of a color list, where each color is represented by a [R,G,B] triplet. To achieve a multi-color effect within a single patch, after localizing all pixels in the circular region, the color values in the color list are sequentially assigned to each pixel using the modulo operation (Equation 17).

$$color\_index = j\%num\_colors \tag{17}$$

where, $color\_index$ denotes the index of the color list (determining the color value of a single pixel), j represents the index of the currently iterated pixel, and num_colors is the length of the color list. This ensures that when the color list is exhausted during traversal, it is automatically reused cyclically, thereby forming a pixel order-based multi-color cyclic distribution within a single patch region.

Considering the differences in the perceptual characteristics of networks toward visible and infrared images, and the need to balance the high perturbation requirement and the low saliency requirement, a 'one-parameter dual-use' difference design method that leverages the channel differences between visible and infrared modalities is adopted. In the visible modality, the color mask is multiplied by the 3-channel visible image, preserving the original color dynamic range while generating a high-brightness color region, which significantly disrupts the model's perception of texture and color. In the infrared modality, the color mask is multiplied by the single-channel grayscale image, and then grayscale compression is applied to darken the patch as a whole. This allows it to naturally integrate with the grayscale characteristics of infrared images, reducing visual abruptness to the human eye. As shown in Figure 4, this design enhances the attack intensity against visual-infrared networks and achieves a balance between attack effectiveness and stealth. Finally, color parameters are directly incorporated into the initialization process of population individuals. Similar to position parameter optimization, [R,G,B] values are iteratively updated through global search, ensuring that the generated patch achieves an optimal configuration in both position and color—thereby enhancing the stealth and effectiveness of the attack.

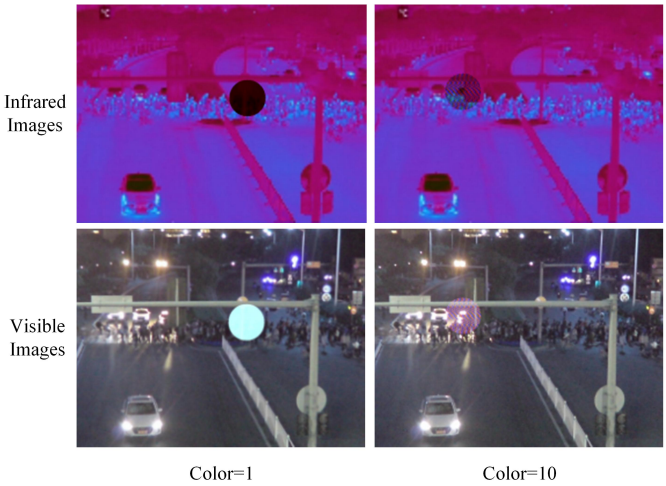


Fig. 4. Visualization of training curves about the design dynamic weighting factor in the multi-task loss function.

E. Iterative Evaluation

In cross-modal attacks, to balance the attack effectiveness across visual-infrared modalities while simultaneously ensuring both attack effectiveness and stealth, this study proposes an iterative evaluation method. By designing a reasonable fitness function, this method guides the adversarial patch to iteratively improve its attack effectiveness, while ensuring that the attack stealth meets the requirements of practical applications. Specifically, we define the fitness function as Equation 18.

$$J = \alpha \cdot E(X^{adv}_{vis+\inf}) + (1-\alpha) \cdot S(X^{adv}_{vis+\inf}) \tag{18}$$

$$E_{count} = |count(X^{adv}_{vis+\inf}) - count(X_{vis+\inf})| \tag{19}$$

$$E_{seg} = 100 - (IoU(X^{adv}_{vis+\inf}) \times 100) \tag{20}$$

$$\begin{aligned} S_{count}/S_{seg} &= (psnr_{vis} + psnr_{\inf}) \times 1 \\ &+ (ssim_{vis} + ssim_{\inf}) \times 20 \end{aligned} \tag{21}$$

$$\begin{aligned} E_{fus} &= L\_\text{int}\,en(X_{vis}, X_{\inf}, Fusion^{adv}_{vis+\inf}) \times 20 \\ &+ L\_Grad(X_{vis}, X_{\inf}, Fusion^{adv}_{vis+\inf}) \times 20 \\ &+ (1 - SSIM(X_{vis}, X_{\inf}, Fusion^{adv}_{vis+\inf})) \times 10 \end{aligned} \tag{22}$$

where,(the weighted adjustment parameter for attack effectiveness) can be flexibly adjusted according to the requirements

of actual attack-defense scenarios. When it is necessary to enhance attack effectiveness, increasing the value of can shift the optimization direction toward improving attack intensity; if the focus is on attack stealth, decreasing the value of will increase the regulatory weight of 1-∂ (the stealth weighted factor), thereby precisely controlling the optimization direction of the iterative process to achieve improved stealth.E represents attack effectiveness, which has different definitions in different task scenarios. For instance, in the crowd counting task, it refers to the people count error (as shown in Equation 19); in the semantic segmentation task, it refers to the Intersection over Union (IoU) (as shown in Equation 20); and in the image fusion task, it uses gradient loss (L_Grad), intensity loss (L_inten), and Structural Similarity Index Measure (SSIM) (as shown in Equation 22), where denotes the fused image of the attacked visual-infrared image.In the crowd counting and semantic segmentation tasks,S(stealth) is evaluated using the Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR) (as shown in Equations 22). It should be specifically noted that in the image fusion task, since SSIM itself is an important indicator for evaluating fusion effectiveness, no additional stealth metrics are considered temporarily in this task scenario.Through the design of the aforementioned fitness function, the progress of cross-modal patch attacks can be effectively quantified.This not only provides a clear optimization direction for the iterative evolution of adversarial patches across different task scenarios but also drives their iterative updates toward the goal of maximizing the fitness function.

IV. Experiments

A. Experimental settings

1) Datasets: We conduct experiments on the publicly available RGBT-CC [23]datasets,MF datasets [24], and Roadscene datasets [25]. These datasets cover visible and infrared images, and include scenarios related to crowd counting, semantic segmentation, and image fusion, ensuring the diversity and representativeness of the data. Details are as follows:

Crowd Counting: The RGBT-CC dataset consists of 1,030 images in the training set,200 images in the validation set, and 800 images in the test .This ensures sufficient data for model training while reserving independent samples for validation and final performance evaluation.

Semantic Segmentation: The MF Datasets consist of urban street scenes captured by the InfRecR500 camera. It includes 1,569 pairs of visual-infrared images, with each image having a resolution of 480 pixels. Among these, there are 820 pairs of daytime images and 749 pairs of nighttime images, covering 8 manually annotated categories and 1 background category. The datasets is divided into three parts: a training set (784 image pairs), a validation set (393 image pairs), and a test set (392 image pairs).

Images Fusion: The Roadscene Datasets contains 221 pairs of aligned visual-infrared images. These images cover diverse and representative scenarios such as roads, vehicles, and pedestrians, and are extracted from FLIR videos as highly representative scenes.

We randomly selected 100 images from the test set of each dataset in one go to serve as the final samples for attack , ensuring the consistency and comparability of the attack process.

2) Target Model: To rigorously evaluate the effectiveness of our proposed method, we select several mainstream models from respective fields as target models. For the crowd counting task, the BL+IADM [23], CAGNet [26], and CFAFNet [27] models are adopted; for the semantic segmentation task, the Openress [28], FEANet [29], and SGFNet [30] models are used; and for the image fusion task, the Res2Fusion [31], UNFusion [32], and MaeFuse [33] models are chosen as research objects. These models all demonstrate advanced performance in their fields, and each model uses officially trained weights as initial weights, ensuring the reliability and consistency of experimental benchmarks.

3) Evaluation Metrics:

Crowd Counting: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) [34] are used as evaluation metrics. Where,MAE calculates the average of the absolute values of the differences between predicted values and ground truth values, representing the average deviation degree between them,the formulas for calculation are as shown in Equation 23;RMSE amplifies the impact of large errors through squared operation,the formulas for calculation are as shown in Equation 24.The larger the values of MAE and RMSE, the stronger the destructive effect of the attack on the model's prediction accuracy.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{23}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{24}$$

where n is the number of samples,y_i and $\hat{y}_i$ are the ground truth value and the estimated value of the i-th image, respectively.

Semantic Segmentation: Mean Intersection over Union (mIoU) and recall are used as evaluation metrics; the formulas for calculation are as shown in Equation 25 and 26. Where,mIoU measures the overlap between segmentation results and ground truth labels.under adversarial attacks, the smaller its value, the better the attack effect; recall reflects the ability to detect target regions, and a decrease indicates the attack interferes with model recognition. The combination of the two comprehensively evaluates the damage of the attack to model performance from the perspectives of regional overlap and target detection capability.

$$mIou = \frac{1}{n-1}\sum_{i=0}^{n}(P_{ii}/(\sum_{j=0}^{n}(p_{ij} + P_{ji}) - P_{ii})) \tag{25}$$

$$recall = \frac{1}{n+1}\sum_{i=0}^{n}(P_{ii}/\sum_{j=0}^{n}p_{ij}) \tag{26}$$

where n denotes the number of manually annotated object categories, and P_{ij} is the number of pixels belonging to category i that are predicted as category j.

Image Fusion: To quantitatively evaluate the effect, we select five metrics: Qabf [35], Structural Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR), Visual Information Fidelity (VIFF) [36], and Correlation Coefficient (CC) [37]. Among these, Qabf measures the amount of feature information and edge information transmitted from the source images to the fused image;PSNR reveals the degree of distortion during the fusion process from the pixel level; SSIM reflects image distortion from three dimensions: brightness, contrast, and structure; VIFF quantifies the retention degree of visual information in the image after fusion; CC is used to measure the degree of linear correlation between the original image and the fused image. When the fused image after attack has lower values of Qabf, SSIM, PSNR, VIFF, and CC, it indicates better attack performance.

4) Implementation: During the iteration process, except for experiments involving the effects of the fitness function, the fitness function in all other experiments only considered attack effectiveness (∂=1), and no optimization adjustments were made for stealth. This setup was intended to achieve the optimal attack effectiveness.

Patch Sizes: To investigate the impact of patch size on attack performance, we adjusted the initial circular radius r to control the patch size, setting it to 20, 40, and 80, respectively. The quantitative results are shown in Table I, where it can be observed that the MAE and RMSE go up as the patch size increases. However, the "occlusion rate" of the patch in the entire image also rises accordingly. As shown in Figure 5, the patch covers more unoccupied areas, and a higher occlusion rate significantly reduces the patch's stealth. Considering the balance between attack efficiency and stealth, we finally selected a radius of r=40. The MAE significantly increases from 13.9240 (for clean images) to 26.1511, and the RMSE rises from 24.4166 to 34.8056. This ensures high attack effectiveness while effectively controlling the occlusion rate, achieving a favorable compromise between the two objectives. Since the image in both semantic segmentation and crowd counting tasks is 480 pixels, this study uniformly adopts patches of this size specification. For the image fusion task, since its images undergo preprocessing operations (including black border removal, thermal noise elimination, and cropping), the patch radius is determined to be 30 pixels based on a target occlusion rate of 1.63%.

Epochs: The epochs represent the maximum number of evolution iterations for a single image. A larger number of epochs leads to more thorough iterations and better attack effectiveness, but the time efficiency decreases rapidly. Therefore, the number of epochs was first set to 200 and added an early stopping mechanism(automatically halting optimization for an image when its mean fitness value remains unchanged for 10 consecutive generations). Experimental results show that 12% of images trigger early stopping in the crowd counting task, 60% in the semantic segmentation task, and 21% in the image fusion task. Considering both attack performance and computational efficiency, the number of epochs is set to 200.

Color Number: To investigate the impact of the number of colors on attack performance, we conducted experiments with 1, 2, 5, and 10 colors, as shown in Figure 6. Table II,III,IV

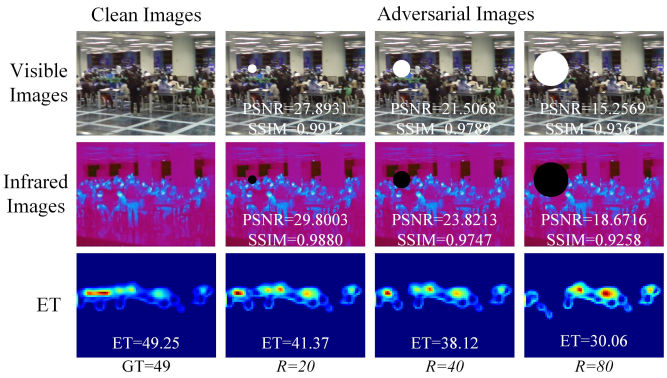


Fig. 5. Visualization of patch radius parameters.

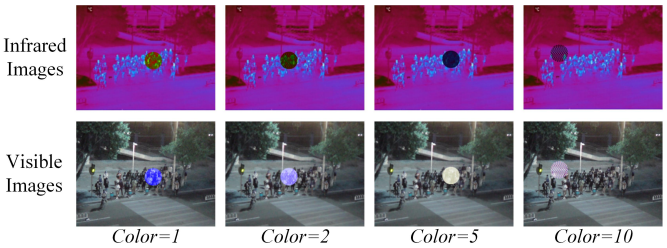


Fig. 6. Visualization of the Number of Colors Parameter.

experimental data indicate the following: For the crowd counting task, the MAE,RMSE,SSIM on visible images, and PSNR and SSIM on infrared images reached optimal values when 10 colors were used; For the semantic segmentation task, the recall and the PSNR on visual-infrared images reached the optimal level under the 10-color setting. For the image fusion task, the PSNR,VIFF, and CC reached optimal values when 2 colors were used, while the Qabf achieved the second-best performance. Thus, the number of colors was determined to be 10 for the crowd counting and semantic segmentation tasks, and 2 for the image fusion task.

B. Ablation Experiments

1) Ablation Experiment on Patch Position: One of the core innovations of this study is the proposal of a patch position optimization strategy for visual-infrared images. To verify the effectiveness of this strategy,we conducted comparative experiments between a random position strategy and an iterative optimization position strategy based on a global search mechanism.For the crowd counting task (Figure 7(a)),both MAE and RMSE increased; for the semantic segmentation task (Figure 7(b)), both the mIoU and recall decreased; and for the image fusion experiment (Figure 7(c)), metrics including Qabf,PSNR,SSIM,VIFF and CC all decreased. These results clearly indicate that the patch position optimization strategy proposed in this study can generate adversarial patches with stronger attack capability.

2) Ablation Experiment on Patch Color: Another important contribution of this study is the proposal of a joint optimization strategy that combines patch color optimization and position optimization.To verify the effectiveness of this joint optimiza-

TABLE I
VISUALIZATION OF PATCH RADIUS PARAMETERS.

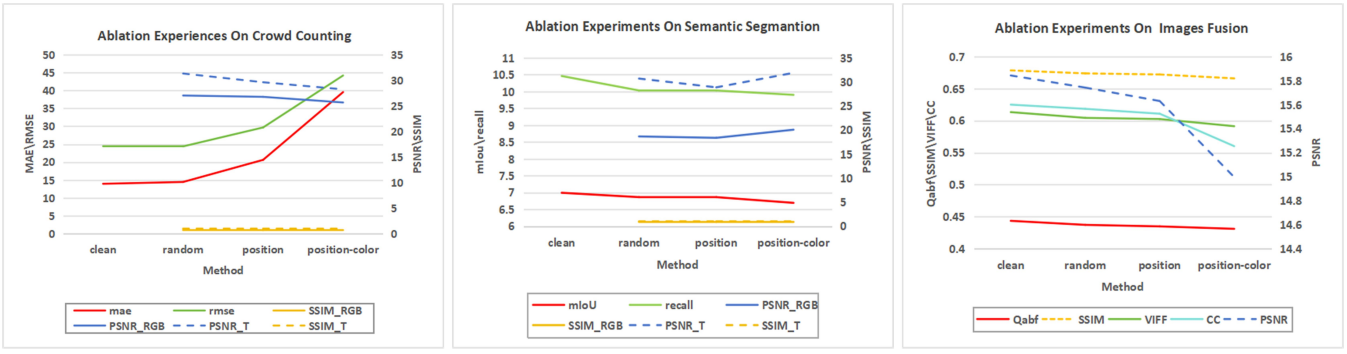
	MAE↑	RMSE↑	PSNR_RGB↑	SSIM_RGB↑	PSNR_T↑↑	SSIM_T↑
clean	13.9240	24.4166				
R=20	20.6466	29.6377	29.7634	0.9922	31.1696	0.9901
R=40	26.1511	34.8056	23.2178	0.9832	25.7501	0.9805
R=80	34.5207	44.3972	17.4165	0.9518	20.255	0.9475

TABLE II
EXPERIMENTAL RESULTS OF THE NUMBER OF COLOR PARAMETER ON THE CROWD COUNTING.

	MAE↑	RMSE↑	PSNR_RGB ↑	SSIM_RGB ↑	PSNR_T↑	SSIM_T↑
clean	13.924	24.4166				
Color=1	26.1511	34.8056	23.2178	0.9832	25.7501	0.9805
Color=2	25.6024	33.5994	23.8335	0.9816	25.8799	0.9797
Color=5	29.0375	36.9115	24.4387	0.9836	27.13367	0.9848
Color=10	39.5359	44.1738	25.6450	0.9832	28.2151	0.9850

TABLE III
EXPERIMENTAL RESULTS OF THE NUMBER OF COLOR PARAMETER ON THE SEMANTIC SEGMENTATION.

	mIoU ↓	recall↓	PSNR_RGB↑	SSIM_RGB↑	PSNR_T↑	SSIM_T↑
clean	24.41	31.63				
Color=1	6.65	9.96	19.1507	0.8645	29.8579	0.9833
Color=2	6.71	9.94	19.3025	0.8625	30.3935	0.9829
Color=5	6.69	9.92	19.7081	0.8614	30.8039	0.9835
Color=10	6.69	9.90	20.0540	0.8612	31.8827	0.9834



(a) (b) (c)

Fig. 7. Line chart showing ablation experiment results for crowd counting, semantic segmentation, and image fusion. This chart details the attack effectiveness of adversarial patches across crowd counting, semantic segmentation, and image fusion tasks. Here, "clean" denotes the clean image, "random" indicates patches with randomly selected positions and colors, 'position' refers to patches where only the position is optimized while the color remains unchanged, and "position-color" denotes patches where both position and color are optimized simultaneously.

TABLE IV
EXPERIMENTAL RESULTS OF THE NUMBER OF COLOR PARAMETER ON THE IMAGE FUSION.

	Qabf↓	PSNR↓	SSIM↓	VIFF↓	CC↓
clean	0.4434	15.8424	0.6784	0.6132	0.6249
Color=1	0.4305	15.0364	0.6662	0.5912	0.5618
Color=2	0.4309	14.9952	0.6659	0.5912	0.5600
Color=5	0.4393	15.2879	0.6613	0.6005	0.5847
Color=10	0.4393	15.2915	0.6615	0.6002	0.5843

tion strategy, we conducted comparative experiments between two strategies: one that only uses a global search mechanism for position optimization, and the other that adopts a global search mechanism for position-color joint optimization.For the crowd counting task (Figure 8(a)),The MAE and RMSE showed significant increases (MAE increased by nearly 100% compared to the position-only strategy),the PSNR and SSIM only exhibited slight decreases This phenomenon aligns with the expectations: although the color optimization strategy enhances attack stealth , there is an inherent trade-off between attack stealth and effectiveness. When the attack effectiveness significantly improves, the slight decreases in PSNR and SSIM indirectly verify the effectiveness of the proposed stealth enhancement method; for the semantic segmentation task (Figure 8(b)), both mIoU and recall decrease while PSNR and SSIM values improve; and for the image fusion experiment (Figure 8(c)), metrics including Qabf,PSNR,SSIM,VIFF and

TABLE V
ABLATION EXPERIMENTS ON CROSS-MODAL ATTACKS FOR CROWD COUNTING, SEMANTIC SEGMENTATION, AND IMAGE FUSION TASKS

Modal	Crowd Counting		Semantic Segmentation		Image Fusion				
	mae↑	rmse↑	mIoU↓	recall↓	Qabf↓	PSNR↓	SSIM↓	VIFF↓	CC↓
RGB	21.6902	30.7622	6.97	9.96	0.4347	15.0853	0.6672	0.5956	0.5668
T	34.3262	42.2757	7.56	10.75	0.4347	15.1014	0.6710	0.5963	0.5695
RGBT	39.5359	44.1738	6.69	9.90	0.4305	15.0364	0.6662	0.5912	0.5618

CC all decreased. These results clearly indicate that the patch position-color joint optimization strategy not only generates adversarial patches with stronger attack capability, but also effectively improves attack stealth by optimizing patch color.

3) Ablation Experiment on Cross-Modality Attacks: To verify the applicability of the proposed method in cross-modal attack scenarios, We conducted comparative experiments involving three strategies: attacks targeting only the infrared modality, attacks targeting only the visible light modality, and attacks targeting both modalities simultaneously. As shown in the experimental results in Table V: For the crowd counting task, the MAE and RMSE of attacks targeting only visible light images or only infrared images were significantly lower than those of attacks targeting both modalities; for the semantic segmentation task, the mIoU and recall of attacks targeting only visible light images or only infrared images were higher than those of attacks targeting both modalities; for the image fusion task, metrics including Qabf, PSNR,SSIM, VIFF, and CC of attacks targeting only visible light images or only infrared images were higher than those of attacks targeting both modalities.These results clearly indicate that compared with single-modality attacks, attacking both visual-infrared modalities simultaneously can generate stronger attack capability.

4) Effects of Fitness Function: To investigate the impact of the fitness function on attack performance, we conducted comparative experiments between two strategies: one that considers only attack effectiveness as a single evaluation metric (∂=1), and another that considers both attack effectiveness and stealth (∂=0.6).In the crowd counting and semantic segmentation tasks, as shown in Figures 8(a) and 8(c), when=1, both the fitness value of the optimal individual and the attack effectiveness showed significant improvements during iteration. Since stealth was not optimized during iteration, it decreased accordingly. As shown in Figures 8, when ∂=1 and ∂=0.6, both the fitness value and attack effectiveness of individual images increased significantly; Table VI experimental results indicate that when ∂=0.6, the overall image SSIM and PSNR (metrics reflecting stealth) improved significantly.These results fully demonstrate that the fitness function can flexibly adjust the attack effectiveness and stealth of adversarial patches according to actual needs, achieving an effective balance between the two.

C. Attack Performance Against Different Models

Considering the performance differences among different model, we verified the effectiveness of our method across multiple mainstream model systems:selecting BL+IADM, CAGNet, and CFAFNet for people counting; employing openrss, FEANet, and SGFNet for semantic segmentation; and utilizing Res2Fusion, UNFusion, and Maefuse for image fusion.For the crowd counting task, as shown in Figure 9 for the semantic segmentation task, after each model was attacked by the optimized patches, the mIoU and recall decreased significantly, while the PSNR also increased compared with images with randomly added patches,as shown in Figure 10; for the image fusion task, after each model was attacked by the optimized patches, metrics including Qabf, PSNR, SSIM,VIFF, and CC all decreased,as shown in Figure 11.This result indicates that the proposed method not only achieves effective adversarial attacks but also exhibits favorable attack stealth characteristics.

V. CONCLUSIONS

This study proposes a cross-modal joint optimization framework based on a global search mechanism, which can simultaneously optimize visual-infrared adversarial patches in both position and color dimensions. Through the population iterative mechanism, the optimal position and color combination of the patch can be quickly searched in the high-dimensional solution space.Meanwhile, it achieves visual differentiation for identical color parameters across visual-infrared modalities. This design reduces visual saliency while precisely disrupting visible texture and color features and effectively interfering with infrared information,thus achieving the effect of cross-modal collaborative attack.Experimental validation was conducted on three dense prediction tasks:crowd counting, semantic segmentation, and image fusion. The generated cross-modal adversarial patches significantly degrade task performance across multiple network architectures while maintaining low saliency, demonstrating excellent generalization ability and a balance between attack effectiveness and stealth. This work provides a new research paradigm for cross-modal adversarial attacks.It should be noted that our method still has certain limitations in the physical world. Future research will focus on investigating the impact of physical factors such as perspective differences and patch materials on the performance of cross-modal adversarial patches, exploring their attack effectiveness in real-world scenarios, and systematically enhancing the security and robustness of multi-modal systems in practical applications.

REFERENCES

[1] H. Zhang, Y. Avrithis, T. Furon, and L. Amsaleg, "Walking on the edge: Fast, low-distortion adversarial examples," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 701–713, 2020.
[2] B. Bonnet, T. Furon, and P. Bas, "Generating adversarial images in quantized domains," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 373–385, 2021.

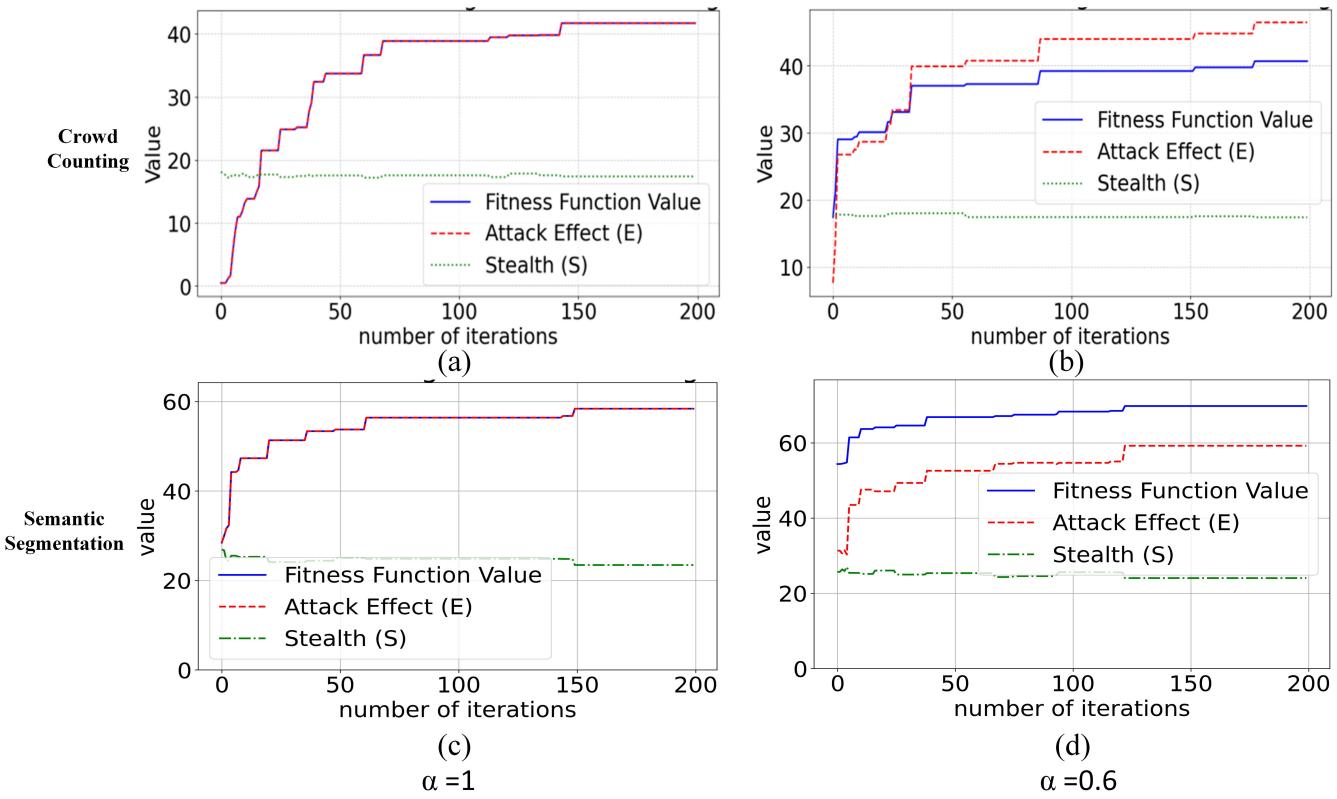


Fig. 8. Trend chart of metric changes in crowd counting and semantic segmentation scenarios. The horizontal axis represents the number of iterations, while the vertical axis shows the values of each metric.When =1 and =0.6,the solid blue line represents the dynamic variation trend of the fitness value, the dashed red line reflects the variation trend of attack effectiveness, and the dotted green line illustrates the variation pattern of the stealth metric.

Fig. 9. Experimental results for different models on crowd counting tasks. Figure a shows attack effectiveness metrics MAE and RMSE, while Figure b presents attack stealth metrics PSNR and SSIM

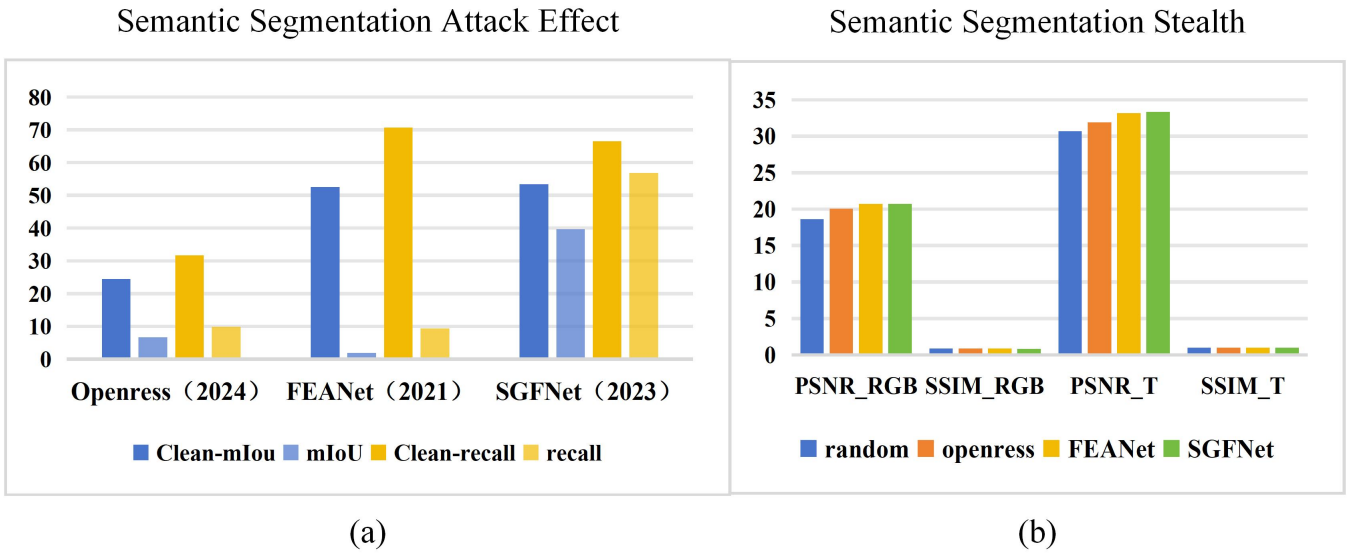


Fig. 10. Experimental results for different models on semantic segmentation tasks. Figure a shows attack effectiveness metrics mIou and recall, while Figure b presents attack stealth metrics PSNR and SSIM

TABLE VI
EXPERIMENTAL RESULTS FOR DIFFERENT FITNESS FUNCTION WEIGHT IN CROWD COUNTING AND SEMANTIC SEGMENTATION TASKS

		MAE↑	MRSE↑	PSNR_RGB↓	SSIM_RGB↓	PSNR_T↓	SSIM_T↓
crowd counting	∂=1	39.5359	44.1738	25.6450	0.9832	28.2151	0.9850
	∂=0.6						
		mIoU↓	recall↓	PSNR_RGB↓	SSIM_RGB↓	PSNR_T↓	SSIM_T↓
semantic segmentation	∂=1	6.69	9.90	20.0540	0.8612	31.8827	0.9834
	∂=0.6	6.88	10.35	24.7951	0.8734	59.2488	0.9996

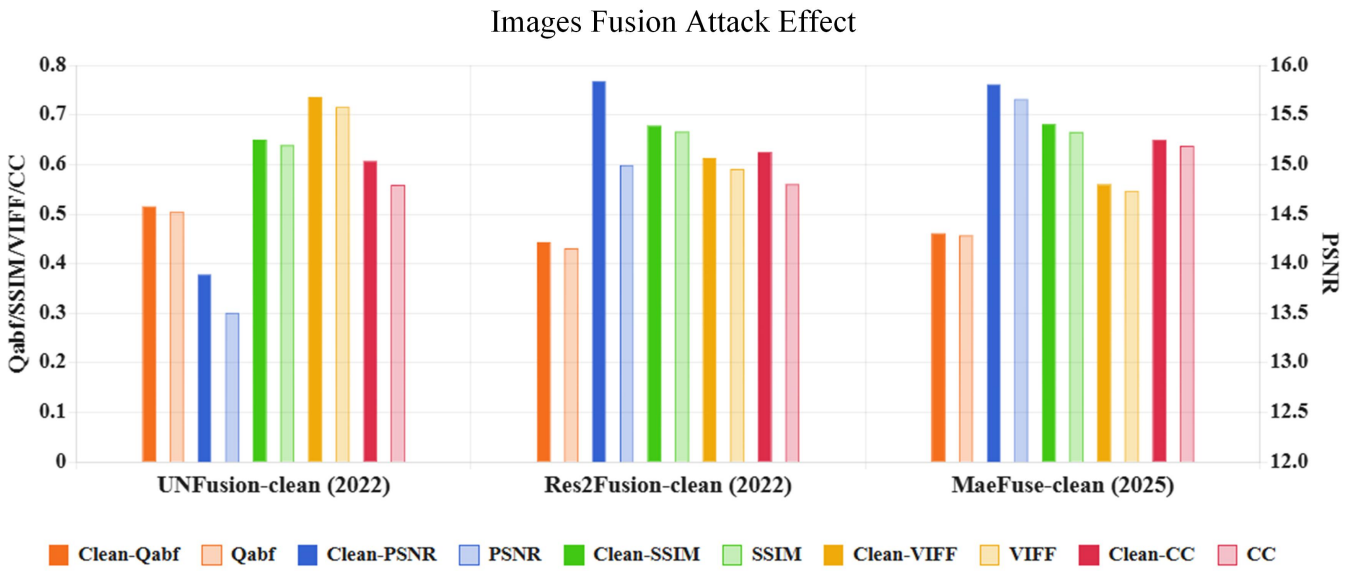


Fig. 11. Experimental results for different models on image fusion tasks.

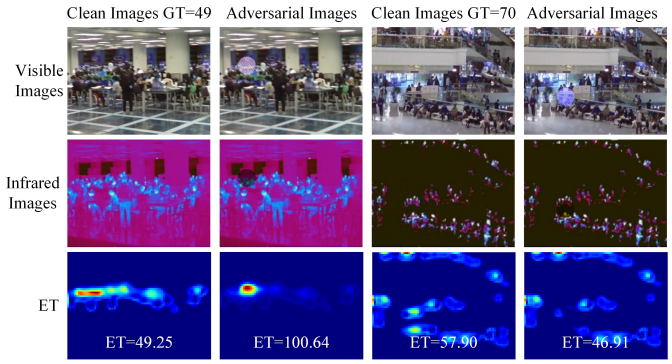


Fig. 12. Visualization of patch attacks on crowd counting tasks.

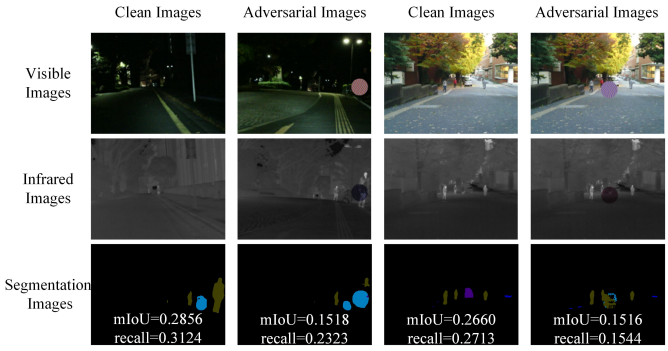


Fig. 13. Visualization of patch attacks on semantic segmentation tasks

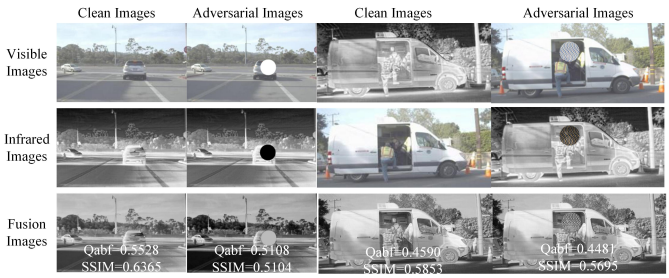


Fig. 14. Visualization of patch attacks on image fusion tasks

[3] H. Ma, K. Xu, X. Jiang, Z. Zhao, and T. Sun, "Transferable black-box attack against face recognition with spatial mutable adversarial patch," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 5636–5650, 2023.
[4] X. Yang, F. Wei, H. Zhang, and J. Zhu, "Design and interpretation of universal adversarial patches in face detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 174–191.
[5] Z. Wu, Y. Cheng, S. Zhang, X. Ji, and W. Xu, "Uniid: Spoofing face authentication system by universal identity," in *Proceedings of Network and Distributed System Security Symposium (NDSS)*, 2024.
[6] A. Abomakhelb, K. A. Jalil, A. G. Buja, A. Alhammadi, and A. M. Alenezi, "A comprehensive review of adversarial attacks and defense strategies in deep neural networks," *Technologies*, vol. 13, no. 5, p. 202, 2025.
[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," Banff, AB, Canada, 2014, input-output mapping;Linear combinations;Prediction errors;Semantic information;Specific nature;State-of-the-art performance;Visual recognition;.

[8] M. Guo, L. Yuan, Z. Yan, B. Chen, Y. Wang, and Q. Ye, "Regressor-segmenter mutual prompt learning for crowd counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28 380–28 389.
[9] S. Cheng, P. Li, K. Han, and H. Xu, "Feature-aware transferable adversarial attacks against image classification," *Applied Soft Computing*, vol. 161, p. 111729, 2024.
[10] P. Moritz, R. Nishihara, and M. Jordan, "A linearly-convergent stochastic l-bfgs algorithm," in *Artificial intelligence and statistics*. PMLR, 2016, pp. 249–258.
[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
[12] H. Lin, Z. Ma, R. Ji, Y. Wang, and X. Hong, "Boosting crowd counting via multifaceted attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 19 628–19 637.
[13] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017.
[14] Q. Wu, Z. Zou, P. Zhou, X. Ye, B. Wang, and A. Li, "Towards adversarial patch analysis and certified defense against crowd counting," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2195–2204.
[15] S. Liu, J. Wang, A. Liu, Y. Li, Y. Gao, X. Liu, and D. Tao, "Harnessing perceptual adversarial patches for crowd counting," in *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, 2022, pp. 2055–2069.
[16] H. Sun, S. Wu, and L. Ma, "Adversarial attacks on gan-based image fusion," *Information Fusion*, vol. 108, p. 102389, 2024.
[17] Z. Liu, J. Liu, B. Zhang, L. Ma, X. Fan, and R. Liu, "Paif: Perception-aware infrared-visible image fusion for attack-tolerant semantic segmentation," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 3706–3714.
[18] C. He, X. Ma, B. B. Zhu, Y. Zeng, H. Hu, X. Bai, H. Jin, and D. Zhang, "Dorpatch: Distributed and occlusion-robust adversarial patch to evade certifiable defenses," in *NDSS*, 2024.
[19] X. Li and S. Ji, "Generative dynamic patch attack," *arXiv preprint arXiv:2111.04266*, 2021.
[20] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369.
[21] S. Das and P. N. Suganthan, "Differential evolution: A survey of the state-of-the-art," *IEEE transactions on evolutionary computation*, vol. 15, no. 1, pp. 4–31, 2010.
[22] S. Rahnamayan, H. R. Tizhoosh, and M. M. Salama, "Opposition-based differential evolution," *IEEE Transactions on Evolutionary computation*, vol. 12, no. 1, pp. 64–79, 2008.
[23] L. Liu, J. Chen, H. Wu, G. Li, C. Li, and L. Lin, "Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4823–4833.
[24] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5108–5115.
[25] Y. Tian, A. Carballo, R. Li, and K. Takeda, "Road scene graph: A semantic graph-based scene representation dataset for intelligent vehicles," *arXiv preprint arXiv:2011.13588*, 2020.
[26] X. Yang, W. Zhou, W. Yan, and X. Qian, "Cagnet: Coordinated attention guidance network for rgb-t crowd counting," *Expert Systems with Applications*, vol. 243, p. 122753, 2024.
[27] W. Kong, Z. Yu, H. Li, and J. Zhang, "Cross-modal misalignment-robust feature fusion for crowd counting," *Engineering Applications of Artificial Intelligence*, vol. 136, p. 108898, 2024.
[28] G. Zhao, J. Huang, X. Yan, Z. Wang, J. Tang, Y. Ou, X. Hu, and T. Peng, "Open-vocabulary rgb-thermal semantic segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 304–320.
[29] F. Deng, H. Feng, M. Liang, H. Wang, Y. Yang, Y. Gao, J. Chen, J. Hu, X. Guo, and T. L. Lam, "Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation," in *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2021, pp. 4467–4473.
[30] Y. Wang, G. Li, and Z. Liu, "Sgfnet: Semantic-guided fusion network for rgb-thermal semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7737–7748, 2023.
[31] Z. Wang, Y. Wu, J. Wang, J. Xu, and W. Shao, "Res2fusion: Infrared and visible image fusion based on dense res2net and double nonlocal attention models," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.
[32] Z. Wang, J. Wang, Y. Wu, J. Xu, and X. Zhang, "Unfusion: A unified multi-scale densely connected network for infrared and visible image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3360–3374, 2021.
[33] J. Li, J. Jiang, P. Liang, J. Ma, and L. Nie, "Maefuse: Transferring omni features with pretrained masked autoencoders for infrared and visible image fusion via guided training," *IEEE Transactions on Image Processing*, 2025.
[34] T. O. Hodson, "Root mean square error (rmse) or mean absolute error (mae): When to use them or not," *Geoscientific Model Development Discussions*, vol. 2022, pp. 1–10, 2022.
[35] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Proceedings 2003 international conference on image processing (Cat. No. 03CH37429)*, vol. 3. IEEE, 2003, pp. III–173.
[36] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Information fusion*, vol. 14, no. 2, pp. 127–135, 2013.
[37] J. Liu, G. Wu, Z. Liu, D. Wang, Z. Jiang, L. Ma, W. Zhong, and X. Fan, "Infrared and visible image fusion: From data compatibility to task adaption," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.