Adversarial Attacks Boosted by Gradient-Evolutionary Multiform Optimization

Yunpeng Gong
School of Informatics
Xiamen University
fmonkey625@gmail.com

Qingyuan Zeng
School of Informatics
Xiamen University
36920221153145@stu.xmu.edu.cn

Dejun Xu
School of Informatics
Xiamen University
xudejun@stu.xmu.edu.cn

Zhenzhong Wang
School of Informatics
Xiamen University
zhenzhong16.wang@connect.polyu.hk

Min Jiang*
School of Informatics
Xiamen University
minjiang@xmu.edu.cn

Abstract

In recent years, despite significant advancements in adversarial attack research, the security challenges in cross-modal scenarios, such as the transferability of adversarial attacks between infrared, thermal, and RGB images, have been overlooked. These heterogeneous image modalities collected by different hardware devices are widely prevalent in practical applications, and the substantial differences between modalities pose significant challenges to attack transferability. In this work, we explore a novel cross-modal adversarial attack strategy, termed multiform attack. We propose a dual-layer optimization framework based on gradient-evolution, facilitating efficient perturbation transfer between modalities. In the first layer of optimization, the framework utilizes image gradients to learn universal perturbations within each modality and employs evolutionary algorithms to search for shared perturbations with transferability across different modalities through secondary optimization. Through extensive testing on multiple heterogeneous datasets, we demonstrate the superiority and robustness of Multiform Attack compared to existing techniques. This work not only enhances the transferability of cross-modal adversarial attacks but also provides a new perspective for understanding security vulnerabilities in cross-modal systems. The code will be available.

1 Introduction

In recent years, research on adversarial attacks [10] has made significant progress, but the security challenges in cross-modal scenarios have not been sufficiently addressed. In these scenarios, adversarial attacks must transfer between different types of images (such as infrared, thermal, and RGB), posing unique challenges due to the substantial differences between these modalities.

This paper investigates adversarial attacks in cross-modal scenarios, focusing on person re-identification (ReID)[32, 27, 34, 9, 23]. ReID is a key task in computer vision that aims to identify

*Corresponding author

Preprint. Under review.

individuals across different locations and times by analyzing surveillance camera images[40]. Due to varying environments, ReID systems use cameras with different modalities to collect data, raising security concerns, especially in complex multi-modal scenarios[7, 23, 36, 35]. Attackers might inject adversarial perturbations into stickers or clothing to disrupt images captured by surveillance cameras, affecting intelligent systems' recognition accuracy [13]. However, the transferability of adversarial perturbations across various image modalities has not been thoroughly studied. Additionally, with stricter global privacy laws, technologies involving personal data processing face strict requirements. To protect privacy, some ReID systems use special image transformations or images from different modalities [14, 21], which also poses a challenge to existing attack methods.

Based on the modality of ReID, attack methods can be divided into two categories: single-modality attacks [39, 8] and cross-modality attacks [7]. The first category, single-modality attacks, focuses on attacks within the same modality (such as RGB-RGB). These methods typically optimize based on the characteristics of a specific modality but are limited in their ability to adapt to cross-modal scenarios. The second category, cross-modality attacks, aims to transfer attacks between different modalities. The main challenge here lies in the significant differences between modalities, which make it difficult to ensure the effectiveness and transferability of the attacks.

Although many ReID attack methods have been proposed [39, 33, 8, 1, 2, 25], they mostly concentrate on single-modality attacks. As shown in Fig. 5 in the supplemental materials, our work focuses on cross-modality attacks. The challenges of cross-modality attacks are twofold: (1) the heterogeneity between different modalities makes cross-modality attacks more difficult to implement than domain adaptation within the same modality; (2) existing gradient-based optimization attack methods face significant limitations in cross-modal scenarios due to the difficulty in effectively transferring gradient information.

"Have the courage to follow your heart and intuition. They somehow already know what you truly want to become." — Steve Jobs, 2005

As stated, inner intuition guides us to what we truly desire. Inspired by this, we adopt evolutionary computation [24, 29, 26, 37], a form of 'intuition' rooted in biological processes. This operates under natural selection and genetic dynamics, guiding random variations to solve optimization problems. In cross-modal scenarios, evolutionary methods surpass gradient-based methods due to their global search capability and adaptability to complex constraints, enabling them to find optimal solutions in a complex, multi-modal search space.

Given the considerable computational intensity involved in fully employing evolutionary computation to search for adversarial perturbations [28, 29, 4, 24], we have adopted Multiform Optimization [6, 31] to ensure the feasibility of our approach. Multiform Optimization is an advanced paradigm, particularly suited for addressing complex problems with diverse representations or requirements. This approach leverages auxiliary tasks to facilitate the resolution of the original problem [6]. By exploring multiple problem formulations, Multiform Optimization captures the search landscape from different perspectives, extracting valuable knowledge and features. This comprehensive strategy enhances the diversity and robustness of solutions, making it more effective in solving complex problems. Our goal is to use evolutionary computation to optimize universal perturbations [18, 33] by exploring shared knowledge across different heterogeneous modalities, thereby enhancing their transferability. Fig.4 and Algorithm1 in the supplemental materials illustrate the overall pipeline of the proposed method.

We assume the existence of a universal perturbation that captures general features across different modalities, capable of transferring to most modalities. However, like models overfitting to training data, adversarial perturbations can also become overly specialized to the biases in the training data, leading to poor performance on unseen modalities. To address this issue, one approach is to independently train multiple models across different modalities and use their gradients to learn a universal perturbation. However, the inconsistency in gradient information due to different model architectures and heterogeneous training data makes it difficult to effectively utilize gradients. Moreover, performance differences between models in different modalities can lead to unbalanced learning of perturbations, affecting their universality and generalization ability.

To address the challenge of ensuring effective and transferable adversarial attacks across heterogeneous modalities, we propose using evolutionary computation to search for sparse perturbations that work across different modalities and use them to fine-tune the universal perturbation, enhancing

its transferability. We introduce a Gradient-Evolutionary Multiform Optimization framework to transfer universal adversarial perturbations (UAP) across modalities. The first optimization layer uses a gradient-based method to optimize perturbations for attacking models within specific modalities, maximizing their impact. The second optimization layer uses an evolutionary search to find perturbations that transfer effectively between models trained on different modalities, aiming for broadly applicable solutions. This design optimizes for single and multi-modal security challenges.

Our work makes the following main contributions:

• We propose a dual-layer optimization strategy that combines gradient-based and evolutionary search techniques for cross-modal adversarial attacks. By introducing the concept of multiform optimization into the field of adversarial attacks and integrating gradient learning with evolutionary algorithms for complementary optimization, we achieve explicit knowledge transfer between different tasks, significantly enhancing the effectiveness and transferability of the attack strategy.

• We are the first to combine evolutionary algorithm theory with gradient-based methods for adversarial attacks. Through mathematical analysis, we demonstrate the effectiveness of evolutionary search in improving the transferability of cross-modal adversarial attacks and its advantages in handling complex cross-modal constraints. Theoretical support and mathematical analysis provide a solid foundation for the effectiveness and feasibility of this method in multi-modal scenarios.

• Through extensive experiments, we validate the significant advantages of our method, showing clear improvements over existing methods in terms of the transferability and robustness of cross-modal adversarial attacks. Our research provides new theoretical and practical foundations for the study of security in multi-modal systems.

2 Related Work

2.1 Adversarial Attack

The concept of adversarial attacks was first introduced by Szegedy et al. [10], whose research revealed that even small perturbations to input images could mislead deep neural networks, resulting in incorrect image recognition. This finding not only highlights the vulnerability of deep learning models but also has important theoretical and practical implications for enhancing the security of artificial intelligence systems. Subsequently, a plethora of adversarial attacks have been proposed [3, 19, 15]. The work by Moosavi-Dezfooli et al. introduced universal adversarial perturbations [18], further advancing research in this field. They demonstrated the ability to generate nearly 'universal' perturbation vectors that, when added to any data sample, cause the same deep learning model to produce incorrect outputs. One Pixel Attack [24], as a significant milestone in sparse perturbation attacks, demonstrates the possibility of misleading models by modifying a single pixel in an image. However, the modification of individual pixels may not always successfully attack all types of images or models in real applications. Therefore, sparse adversarial attacks [24, 29, 4, 28] often involve modifications of multiple pixels, albeit still limited in number, providing higher flexibility and a wider success rate. Although universal perturbations can broadly affect multiple samples, they are relatively easier to be detected by designed targeted detection mechanisms due to their ubiquitous and consistent perturbation patterns. In contrast, sparse adversarial attacks, by applying extremely limited perturbations to input data, demonstrate higher stealthiness. This attack method is more challenging to be identified by standard defense measures in experiments and practical applications due to its high target accuracy and fewer intervention points.

2.2 Attack Person Re-ID System

Several ReID attack methods have been proposed, with current research predominantly focusing on RGB-RGB matching. These methods include: Metric-FGSM [1] extends techniques inspired by classification attacks into the category of metric attacks. These include Fast Gradient Sign Method (FGSM)[10], Iterative FGSM (IFGSM), and Momentum IFGSM (MIFGSM)[5]. The Furthest-Negative Attack (FNA)[2] integrates hard sample mining[11] and triplet loss to guide image features towards the least similar cluster while moving away from similar features. Deep Mis-Ranking (DMR) [25] utilizes a multi-stage network architecture to extract features at different levels, aiming to derive general and transferable features for adversarial perturbations. Gong et al.[8] proposed a method specifically for attacking color features without requiring additional reference images and

discussed effective defense strategies against current ReID attacks. The Opposite-Direction Feature Attack (ODFA)[39] exploits feature-level adversarial gradients to generate examples that guide features in the opposite direction using an artificial guide. Yang et al.[33] introduced a combined attack named Col.+Del. (Color Attack and Delta Attack), which integrates UAP-Retrieval [17] with color space perturbations [16]. The inclusion of color space perturbations enhances the attack's universality and transferability across RGB-RGB datasets. CMPS [7] represents the first exploration into the security of cross-modal ReID. It leverages gradients from different modalities to optimize universal perturbations, effectively enhancing the universality and adaptability of attacks within a given modality. Similar to other gradient-based methods, it has certain limitations in terms of the transferability of attacks.

Existing methods primarily focus on gradient-based attacks for single-modal systems, lacking mechanisms to capture shared knowledge across different modalities. Additionally, the heterogeneity between different modalities makes it difficult for these gradient-based methods to achieve effective adaptation across more than two modalities. Our approach aims to enhance the effectiveness and transferability of cross-modal adversarial attacks by combining gradient-based techniques with evolutionary search.

3 Methodology

We aim to find a universal adversarial perturbation ϵ with cross-modal transferability that can mislead the ranking results of a given modality $\mathcal{G}$ and an unseen target modality χ for a re-identification (re-ID) model. The attack involves modifying a query image $\mathcal{I}$ by adding a perturbation ϵ. This perturbed image $\mathcal{I}'$ is then used to fool the victim re-identification model $\mathcal{M}$ when querying a gallery.

3.1 Framework Overview

Our proposed methodology employs a dual-layer optimization framework, integrating gradient-based learning and evolutionary algorithms to enhance the effectiveness and transferability of adversarial perturbations across different image modalities. This framework is designed to address the unique challenges posed by the heterogeneity of cross-modal data, ensuring that the learned perturbations are both effective and broadly applicable. In the first layer of optimization, a gradient-based learning method focuses on optimizing adversarial perturbations to attack machine learning models within specific modalities. This process involves computing the loss based on the task-specific metric, such as the triplet loss with Mahalanobis distance, and using momentum gradient descent to iteratively adjust the perturbations. The second layer of optimization employs an evolutionary search strategy to explore perturbations that can be effectively transferred between models trained on different modalities. This strategy involves generating a population of perturbations, evaluating their performance across multiple models, and iteratively refining the perturbations through crossover and mutation operations. The goal is to discover perturbations that are broadly applicable and maintain their effectiveness across various modalities. By leveraging evolutionary computation, this layer addresses the challenge of transferring adversarial attacks between heterogeneous data, enhancing the robustness and generalization of the perturbations.

The combination of these two layers—gradient-based learning for modality-specific optimization and evolutionary search for cross-modal transferability—constitutes the Gradient-Evolutionary Multiform Optimization framework. This dual-layer approach not only optimizes perturbations for a single modality but also adapts them to the security challenges present in multi-modal environments. The overall framework is detailed in Algorithm 1 in the supplementary materials. This algorithm delineates the step-by-step process for implementing the Gradient-Evolutionary Multiform Optimization, ensuring continuous refinement and adaptation of perturbations to maintain high effectiveness across different image modalities. Regarding the proposed method, we conducted a theoretical analysis focusing on two aspects: the feasibility of evolutionary search and its effectiveness in enhancing the transferability of universal perturbations. For details, please refer to Supplementary Materials 9 and 10.

3.2 Gradient-Based Learning

In the first layer, our primary objective is to optimize adversarial perturbations for specific modalities using gradient-based learning. We define the optimization problem as:

$$\mathcal{L}_{\text{meta}} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{\text{tri}}(\delta, x_i), \tag{1}$$

where L_{tri} is the loss function tailored to optimize adversarial perturbations within a specific modality.

Existing research [7, 39, 33, 8, 1, 2, 25] on ReID adversarial attacks typically employs Euclidean distance to design loss functions. However, under the assumption that adversarial samples reside in a manifold space, traditional Euclidean distance may not be sufficiently flexible, as data points in manifold spaces often exhibit nonlinear distributions. To address this issue, we employ the Mahalanobis distance, which is more suitable for manifold problems. This distance measure considers the covariance structure of the data, allowing it to more accurately capture the nonlinear relationships and adapt to scale variations in different directions, thereby providing a more flexible and precise distance metric. Hence, in crafting the triplet loss function, we opt to utilize Mahalanobis distance as our metric of choice, aiming to better guide the optimization process for adversarial perturbations:

$$\begin{aligned} \mathcal{L}_{\text{tri}}(\delta, x_i) = &\left[D_M(C_n^{m_1}, f_{\text{adv}}) - D_M(C_p^{m_1}, f_{\text{adv}}) + \rho\right]_+ \\ &+ \left[D_M(C_n^{m_2}, f_{\text{adv}}) - D_M(C_p^{m_2}, f_{\text{adv}}) + \rho\right]_+ . \end{aligned} \tag{2}$$

We follow the approach of [17] to optimize the perturbation using cluster centroids. Here, $C_p^{m_1}$, $C_p^{m_2}$, $C_n^{m_1}$, and $C_n^{m_2}$ respectively represent the nearest and farthest cluster centroids of original image features in the training data for modalities m_1 and m_2. f_{adv} denote the perturbed features (We set the margin $\rho = 0.5$ in triplet loss). The distance between vectors x and y, using the Mahalanobis distance $D_M(x, y)$, is defined as follows. For computational convenience and optimization stability, the squared Mahalanobis distance is commonly employed as the loss function:

$$D_M(x, y) = (x - y)^T S^{-1} (x - y), \tag{3}$$

Here, S is the covariance matrix of the dataset. We utilize exponential weighted moving average [12] for momentum gradient descent. This approach facilitates smoother parameter updates, accelerating convergence and enhancing generalization performance. The process is formulated as follows:

$$v_{t+1} = \beta v_t + (1 - \beta) \cdot \frac{\nabla_\delta \mathcal{L}_{meta}}{\|\nabla_\delta \mathcal{L}_{meta}\|_1}. \tag{4}$$

Here, v_t represents the exponential moving average of the gradient at time step t (initial value v_0 = 0.), β is the decay coefficient (set as $\beta = 0.9$), and $\frac{\nabla_\delta L_{meta}}{\|\nabla_\delta L_{meta}\|_1}$ is the normalized gradient. Then, use the updated momentum variable v_{t+1} to update the perturbation δ:

$$\delta_{t+1} = \text{clip}(\delta_t + \alpha \cdot \text{sign}(v_{t+1}), -\varepsilon, \varepsilon). \tag{5}$$

Here, δ_{t+1} is the updated perturbation at time step $t + 1$, δ_t is the perturbation at time step t, α is the learning rate (set as $\alpha = \frac{\epsilon}{10}$), ε is the clipping threshold ($\epsilon = 8$, unless otherwise specified), and sign$(\cdot)$ function returns the sign of the input.

In this layer of optimization, we focus on minimizing the task-specific loss, which aims to mislead the ReID model by altering the query image such that it fails to match the correct individual in the database.

3.3 Evolutionary Search

In the second layer of optimization, we employ evolutionary search to optimize the transferability of perturbations across different modalities. Following the methodology outlined in [29], our approach adapts the evolutionary algorithm to simultaneously search for sparse perturbations with

transferability across multiple modalities, which will be used to fine-tune universal adversarial perturbations obtained from the first layer of optimization. With respect to our objectives, we define the optimization problem as follows:

$$\begin{aligned} &\min_{\eta} \ \Phi(\mathcal{F}(x+\delta+\eta)), \\ &\text{subject to:} \quad \|\eta\|_0 \leq k, \quad \|\delta+\eta\|_\infty \leq \varepsilon. \end{aligned} \tag{6}$$

k represents the number of perturbed pixels. The $\|\delta+\eta\|_\infty$ constrains that the maximum value of each element in the perturbation vector does not exceed ϵ, which can be achieved through clipping.

At this stage, the optimized adversarial sample x_{adv} can be obtained from $x+\delta+\eta$. The features of the adversarial sample, denoted as f_{adv}, are extracted using $\mathcal{F}(x_{adv})$. Where $\Phi(f_{adv}) = (\tilde{\mathcal{D}}(f_{adv}), \tilde{\mathcal{S}}(f_{adv}), \|\eta\|_2, \|\eta\|_0)^T$ is the objective vector. Due to the limitations of using {-1, 1} as the perturbation set, we adopted the method defined by Williams et al. [29], which redefines the perturbation set to {-1, 0, 1}. This expansion allows the inclusion of zero values within the perturbation vector, inherently optimizing the l_0 norm by increasing the proportion of zeros. Consequently, optimizing the l_2 norm also indirectly reduces the l_0 norm, as a lower l_2 norm can be achieved partly by increasing the number of zeros in the vector. Therefore, the objective vector $\Phi(f_{adv})$ is now defined as $(\tilde{\mathcal{D}}(f_{adv}), \tilde{\mathcal{S}}(f_{adv}), \|\eta\|_2)^T$, where $\eta \in \{-1, 1, 0\}$. $\tilde{\mathcal{D}}(f_{adv})$ and $\tilde{\mathcal{S}}(f_{adv})$ represent the total metric loss and total attack success rate across all models, respectively. They can be described by the following formula:

$$\mathcal{D}_i(f_{adv}) = (f_{adv} - C)^T S^{-1} (f_{adv} - C), \tag{7}$$

$\mathcal{D}_i(f_{adv})$ denote the loss incurred by the perturbed input on model $\mathcal{M}_i$. The loss from the adversarial sample x_{adv} to the cluster centroid C is measured using the squared Mahalanobis distance. Therefore, the total loss across all models can be represented as:

$$\mathcal{D}(f_{adv}) = \sum_{i=1}^{n} \mathcal{D}_i(f_{adv}). \tag{8}$$

To transform into a minimization problem, we ultimately use the following formula for optimization:

$$\tilde{\mathcal{D}}(f_{adv}) = \exp\left(-\mathcal{D}(f_{adv})\right). \tag{9}$$

For model $\mathcal{M}_i$, success rates $\mathcal{S}_i$ can be defined as follows:

$$\mathcal{S}_i(f_{adv}) = \begin{cases} 1, & \text{if } \arg\max(\hat{y}_j) \neq y_j \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

$\hat{y}_j$ is the predicted label by the model, and y_j is the true label corresponding to the sample. The overall success rate can be calculated using the following formula:

$$\mathcal{S}(f_{adv}) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{S}_i(f_{adv}). \tag{11}$$

To transform into a minimization problem, we ultimately use the following formula for optimization:

$$\tilde{\mathcal{S}}(f_{adv}) = 1 - \frac{1}{n} \sum_{i=1}^{n} \mathcal{S}_i(f_{adv}). \tag{12}$$

We follow the approach proposed by Williams et al. [29] for crossover, mutation, and evaluation of the population. For further details, please refer to the supplementary material 8. During the selection phase, we define the following non-dominated sorting relationship to achieve the objective of simultaneously searching for transferable perturbations across multiple modalities.

Domination Deffnition. In the process of conducting multimodal adversarial attacks, we assess and compare two perturbation sets within the perturbation solution set $\mathcal{P}$, denoted as $\mathcal{P}_i$ and $\mathcal{P}_j$,

respectively. These two solutions yield perturbations represented by η_i and η_j. We evaluate the resulting adversarial effectiveness using function $\mathcal{F}(\bullet)$ which yields the objective vectors $\mathcal{F}_i$ and $\mathcal{F}_j$. A solution $\mathcal{P}_i$ is considered to dominate another solution $\mathcal{P}_j$ if any of the following conditions are met:

1. If η_i has higher transferability than η_j.
2. η_i and η_j have the same transferability, and $\|\eta_i\|_2 \leq \|\eta_j\|_2$.
3. Both η_i and η_j do not exhibit transferability, and η_i has a smaller total loss. $\tilde{\mathcal{D}}(f_{adv})$.

Please note, a perturbation η is considered to have transferability if it satisfies the attack success rate is greater than 0. The higher the attack success rate $\mathcal{S}(f_{adv})$, the greater the considered transferability.

This dual-layer optimization framework significantly enhances the robustness of adversarial perturbations. The first layer of optimization utilizes gradient descent to learn universal perturbations. The second layer employs evolutionary strategies to capture transferable features across modalities, fine-tuning the learned universal perturbations. This approach not only improves the applicability of perturbations in multi-modal environments but also increases the flexibility of the overall attack strategy.

4 Empirical Study

4.1 Dataset

We evaluate our method using two commonly utilized cross-modality ReID datasets: SYSU [30], RegDB [20] and Sketch [22]. SYSU is a large-scale dataset featuring 395 training identities captured by six cameras (four RGB and two near-infrared) on the SYSU campus. It contains 22,258 visible and 11,909 near-infrared images. The test set includes 95 identities evaluated under two settings, with query sets comprising 3,803 images from two IR cameras. RegDB [20], a smaller dataset, consists of 412 identities each with ten visible and ten thermal images. We randomly selected 206 identities (2,060 images) for training and used the remaining 206 identities (2,060 images) for testing. The Sketch ReID dataset comprises 200 individuals, each represented by one sketch and two photographs. The photographs of each individual were captured during daylight using two cross-view cameras. Raw images (or video frames) were manually cropped to ensure that each photograph includes only the specific individual. Additionally, to simulate special image transformations aimed at visually protecting personal privacy, we employed random channel mixing (for specific algorithms, refer to the supplementary materials 4) on images from the Market1501 [38] dataset to create a new dataset, which we refer to as CnMix in the following sections. Market1501 includes 1,501 pedestrians captured by six cameras (five HD cameras and one low-definition camera).

4.2 Evaluation Metric

In line with [38], we utilize Rank-k precision, Cumulative Matching Characteristics (CMC), and mean Average Precision (mAP) as our evaluation metrics. Specifically, Rank-1 precision measures the average accuracy of the top result for each query image across different modalities. The mAP score quantifies the mean accuracy by ordering the query results according to their similarity; a result that appears closer to the top of this list indicates higher precision. It is important to note in the context of adversarial attacks that lower accuracy scores signify more effective attacks.

4.3 Comparison

Following [7], we employed two cross-modality baseline models, AGW [36] and DDAG [35], to conduct tests on the RegDB [20] and SYSU [30] cross-modality ReID datasets. The experiments comprised two scenarios: 1) Perturbing visible images (query) to disrupt the retrieval of infrared or thermal images (gallery), labeled as 'Visible to Infrared' in Table 1 and 'Visible to Thermal' in Table 2. 2) Perturbing infrared or thermal images (query) to interfere with the retrieval of visible images (gallery), labeled as 'Infrared to Visible' in Table 1 and 'Thermal to Visible' in Table 2.

In this experiment, "Our attack*" uses gradient-based single-layer optimization without evolutionary search, while "Our attack" employs our dual-layer optimization. Both optimizations leverage the

Table 1: Results for attacking cross-modality ReID systems on the SYSU dataset. It reports on visible images querying infrared images and vice versa. Rank at r accuracy (%) and mAP (%) are reported.

Settings		Visible to Infrared				Infrared to Visible			
Method	Venue	$r=1$	$r=10$	$r=20$	mAP	$r=1$	$r=10$	$r=20$	mAP
AGW baseline [36]	TPAMI 2022	47.50	84.39	92.14	47.65	54.17	91.14	95.98	62.97
M-FGSM attack [1]	TPAMI 2020	25.79	49.04	57.96	19.24	20.56	38.91	46.35	15.89
LTA attack [8]	CVPR 2022	8.42	21.25	27.98	9.16	20.92	32.18	36.80	15.24
ODFA attack [39]	IJCV 2023	25.43	47.49	56.38	19.00	14.62	29.92	36.42	11.35
Col.+Del. attack [33]	TPAMI 2023	3.23	14.48	20.15	3.27	4.12	16.85	21.27	3.89
CMPS attack [7]	Arxiv 2024	1.11	8.67	16.14	1.41	1.31	7.47	10.36	1.23
Our attack*	————	1.10	7.42	14.46	1.27	1.25	6.34	9.39	1.12
Our attack	————	1.02	7.24	14.28	1.13	1.18	6.02	9.17	1.11
DDAG baseline [35]	ECCV 2020	54.75	90.39	95.81	53.02	61.02	94.06	98.41	67.98
M-FGSM attack [1]	TPAMI 2020	28.36	52.47	60.76	23.11	24.85	40.74	49.23	18.40
LTA attack [8]	CVPR 2022	10.54	23.08	30.47	12.28	18.93	34.12	41.52	15.24
ODFA attack [39]	IJCV 2023	27.75	50.26	59.14	22.30	17.62	32.64	40.03	14.83
Col.+Del. attack [33]	TPAMI 2023	4.28	16.12	21.36	3.97	6.28	19.53	25.61	5.21
CMPS attack [7]	Arxiv 2024	1.62	7.59	14.46	1.84	1.45	7.71	10.72	1.25
Our attack*	————	1.54	6.72	12.66	1.61	1.40	7.68	10.43	1.18
Our attack	————	1.23	6.38	12.06	1.22	1.31	7.44	10.15	1.06

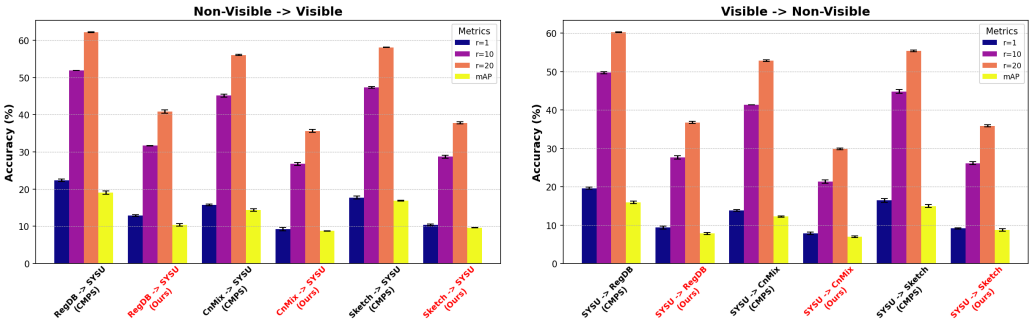

Figure 1: Comparative Analysis with State-of-the-Art Method on Transferability Across Heterogeneous Cross-Modal Datasets.

given model and dataset samples. As Table 1 shows, our method outperforms the CMPS and Col.+Del. attacks in attack effectiveness. These results demonstrate: 1) The Mahalanobis distance in our method effectively captures the adversarial sample space structure; 2) Incorporating evolutionary search broadens the solution exploration, avoids local minima, and enhances attack effectiveness. We conducted identical experiments on the RegDB dataset (see supplementary materials 2). The supplementary material 6 presents a comparative analysis using attention heatmaps.

Experiments Comparing Attack Transferability. We compare with the state-of-the-art retrieval attack method in terms of perturbation transferability, across different cross-modal datasets and different baselines. We verify the perturbation transferability using four heterogeneous cross-modal datasets: SYSU, RegDB, Sketch, and CnMix. In our transfer attack experiments, we need to use all four datasets simultaneously. Different datasets are trained using different models, with each model representing a specific type of modality. For example, in Fig. 1, RegDB->SYSU indicates that we optimize the universal perturbation using the RegDB dataset (with SYSU and Sketch datasets for auxiliary optimization) and then transfer it to the SYSU dataset for testing. Specifically, for the CMPS attack, due to the lack of an intrinsic mechanism to correlate more modalities, it first learns the universal perturbation on the RegDB dataset, and then sequentially adjusts the perturbation using the SYSU and Sketch datasets. In contrast, our method learns the universal perturbation using the RegDB dataset while simultaneously fine-tuning this perturbation on the SYSU and Sketch datasets using evolutionary search to achieve better transferability. In the supplementary material Fig. 6, we compare the proposed method with the Col.+Del.and CMPS attacks using attention heatmaps.

From Fig. 1, it is evident that our proposed method outperforms CMPS attack [33] across various metrics. This indicates: 1) For perturbations trained solely on gradients, the inconsistency in gradient

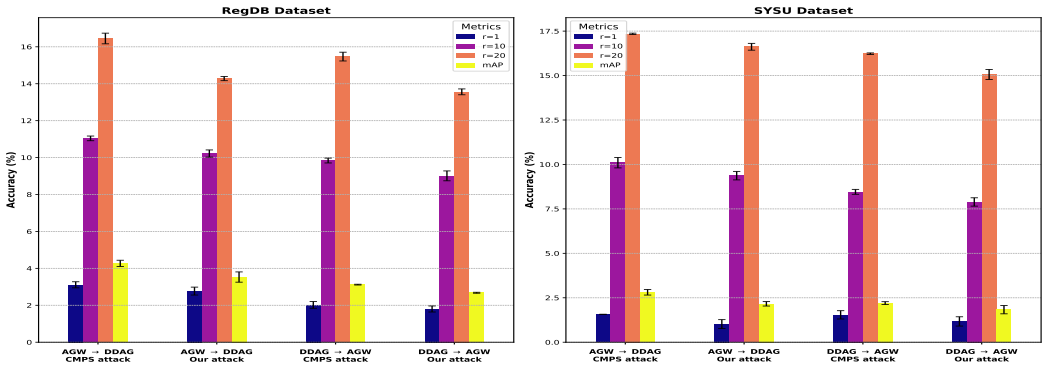

Figure 2: Comparative Analysis with State-of-the-Art Method on Transferability Across Different Baselines.

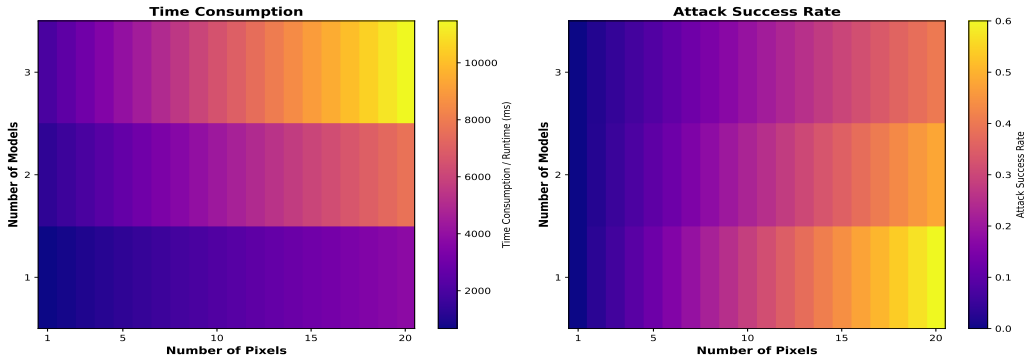

Figure 3: The ablation study in the proposed method examines the relationships between the number of perturbed pixels, the number of models, the attack success rate, and time consumption.

information due to diverse model architectures and heterogeneous training data makes effective gradient utilization challenging. Additionally, performance disparities between models trained on different modalities may lead to imbalanced perturbation learning, affecting its universality and generalization ability. 2) Adversarial attacks similarly face the dilemma between stability and adaptability in real-world scenarios. In deep learning, addressing complex tasks often entails a dilemma between stability and plasticity. This means that maintaining the stability of old knowledge while acquiring new knowledge poses a challenge. If the model exhibits excessive flexibility to new data (high plasticity), it may suffer from 'catastrophic forgetting,' leading to the loss of previously acquired knowledge. Conversely, if the model lacks sufficient adaptability to new knowledge (high stability), it may struggle to assimilate new information, thereby impacting learning efficiency and the model's generalization capability. The method proposed in this paper offers a potential solution to the dilemma between stability and adaptability. Furthermore, as shown in Fig. 2, experiments were conducted across different baselines, which also demonstrate the superiority of our proposed method.

4.4 Ablation Study

Fig. 3 shows: the left image depicts the relationship between the number of perturbed pixels, models (modalities) in evolutionary search, and time consumption. The right image shows the relationship between the number of perturbed pixels, models (modalities), time consumption, and attack success rate. Key observations include: (1) attack success rate increases with more perturbed pixels; (2) time consumption rises with more perturbed pixels; (3) time consumption increases proportionally with the number of models; (4) attack success rate decreases with more models. For the impact of different crossover and mutation rates on attack success rate using evolutionary search alone, see supplementary material **??**.

5 Conclusion

This paper introduces a novel cross-modal adversarial attack strategy, named Multiform Attack, which leverages a Gradient-Evolutionary Multiform Optimization framework to address the critical challenge of transferability between heterogeneous image modalities. By integrating gradient-based learning with evolutionary search, our approach significantly enhances the transferability and robustness of adversarial perturbations across different modalities, including infrared, thermal, and RGB images. Our dual-layer optimization strategy effectively combines the strengths of gradient-based methods and evolutionary algorithms, facilitating efficient perturbation transfer and adeptly handling complex cross-modal constraints. Extensive experiments on multiple heterogeneous datasets validate that our method significantly outperforms existing techniques. It enhances the performance of universal adversarial perturbations within a given modality and greatly increases their applicability across diverse modalities. This advancement offers a new perspective for understanding and addressing security vulnerabilities in multi-modal systems. Our research not only highlights the potential for improved adversarial attack strategies but also provides a robust foundation for future studies aimed at developing more secure and resilient cross-modal systems.

References

[1] Song Bai, Yingwei Li, Yuyin Zhou, Qizhu Li, and Philip HS Torr. Adversarial metric attack and defense for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2119–2126, 2020.

[2] Quentin Bouniot, Romaric Audigier, and Angelique Loesch. Vulnerability of person re-identification models to metric adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[4] Francesco Croce, Maksym Andriushchenko, Naman D Singh, Nicolas Flammarion, and Matthias Hein. Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6437–6445, 2022.

[5] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[6] Yinglan Feng, Liang Feng, Sam Kwong, and Kay Chen Tan. A multi-form evolutionary search paradigm for bi-level multi-objective optimization. *IEEE Transactions on Evolutionary Computation*, 2023.

[7] Yunpeng Gong et al. Cross-modality perturbation synergy attack for person re-identification. *arXiv preprint arXiv:2401.10090*, 2024.

[8] Yunpeng Gong, Liqing Huang, and Lifei Chen. Person re-identification method based on color attack and joint defence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4313–4322, 2022.

[9] Yunpeng Gong, Jiaquan Li, Lifei Chen, and Min Jiang. Exploring color invariance through image-level ensemble learning. *arXiv preprint arXiv:2401.10512*, 2024.

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[12] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[13] Zhan Hu, Siyuan Huang, Xiaopei Zhu, Xiaolin Hu, Fuchun Sun, and Bo Zhang. Adversarial texture for fooling person detectors in the physical world. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13297–13306, 2022.

[14] My Kieu, Andrew D Bagdanov, Marco Bertini, and Alberto Del Bimbo. Domain adaptation for privacy-preserving pedestrian detection in thermal imagery. In *Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20*, pages 203–213. Springer, 2019.

[15] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[16] Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. *Advances in neural information processing systems*, 32, 2019.

[17] Jie Li, Rongrong Ji, Hong Liu, Xiaopeng Hong, Yue Gao, and Qi Tian. Universal perturbation attack against image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4899–4908, 2019.

[18] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[19] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[20] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

[21] Yumo Ouchi, Hidetsugu Uchida, and Narishige Abe. Privacy-preserving image transformation method for person detection and re-id. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1798–1803. IEEE, 2023.

[22] Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, and Yonghong Tian. Cross-domain adversarial feature learning for sketch re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 609–617, 2018.

[23] Jiangming Shi, Yachao Zhang, Xiangbo Yin, Yuan Xie, Zhizhong Zhang, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11218–11228, 2023.

[24] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[25] Hongjun Wang, Guangrun Wang, Ya Li, Dongyu Zhang, and Liang Lin. Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 342–351, 2020.

[26] Zhenzhong Wang, Lulu Cao, Liang Feng, Min Jiang, and Kay Chen Tan. Evolutionary multitask optimization with lower confidence bound-based solution selection strategy. *IEEE Transactions on Evolutionary Computation*, 2024.

[27] Zi Wang, Huaibo Huang, Aihua Zheng, and Ran He. Heterogeneous test-time training for multi-modal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5850–5858, 2024.

[28] Phoenix Williams and Ke Li. Camopatch: An evolutionary strategy for generating camoflauged adversarial patches. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 67269–67283. Curran Associates, Inc., 2023.

[29] Phoenix Neale Williams and Ke Li. Black-box sparse adversarial attack via multi-objective optimisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12291–12301, 2023.

[30] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017.

[31] Yue Wu, Hangqi Ding, Maoguo Gong, AK Qin, Wenping Ma, Qiguang Miao, and Kay Chen Tan. Evolutionary multiform optimization with two-stage bidirectional knowledge transfer strategy for point cloud registration. *IEEE Transactions on Evolutionary Computation*, 2022.

[32] Jiaer Xia, Lei Tan, Pingyang Dai, Mingbo Zhao, Yongjian Wu, and Liujuan Cao. Attention disturbance and dual-path constraint network for occluded person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6198–6206, 2024.

[33] Fengxiang Yang, Juanjuan Weng, Zhun Zhong, Hong Liu, Zheng Wang, Zhiming Luo, Donglin Cao, Shaozi Li, Shin'ichi Satoh, and Nicu Sebe. Towards robust person re-identification by defending against universal attackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5218–5235, 2023.

[34] Fengxiang Yang, Zhun Zhong, Zhiming Luo, Yifan He, Shaozi Li, and Nicu Sebe. Diversity-authenticity co-constrained stylization for federated domain generalization in person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6477–6485, 2024.

[35] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *European Conference on Computer Vision (ECCV)*, 2020.

[36] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2022.

[37] Yulong Ye, Songbai Liu, Junwei Zhou, Qiuzhen Lin, Min Jiang, and Kay Chen Tan. Learning-based directional improvement prediction for dynamic multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, 2024.

[38] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[39] Zhedong Zheng, Liang Zheng, Yi Yang, and Fei Wu. U-turn: Crafting adversarial queries with opposite-direction features. *International Journal of Computer Vision*, 131(4):835–854, 2023.

[40] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2723–2738, 2021.

Supplemental Material

Contents

6 Multiform Attack Framework

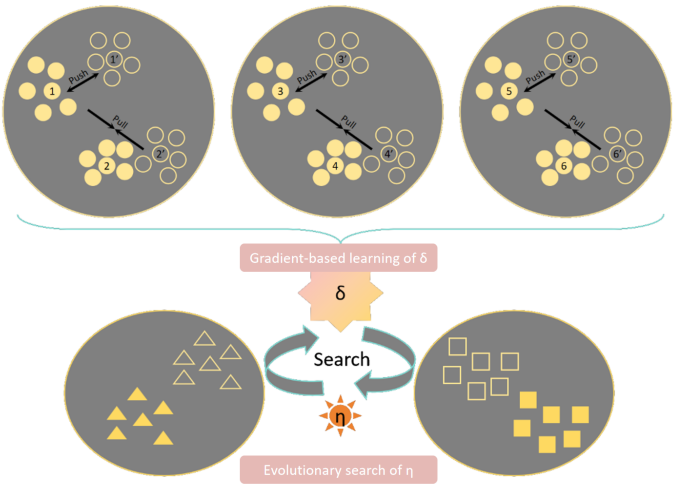

Figure 4: The figure shows different shapes representing samples from various cross-modal datasets (solid shapes represent normal visible RGB samples, while hollow shapes represent corresponding samples from other modalities). During the iterations, our method first learns a universal perturbation δ on a given cross-modal dataset (circles) and then searches for a perturbation η using samples from two other different cross-modal datasets, which can be superimposed on δ to enhance transferability.

Algorithm 1 Multiform Attack Optimization Framework

Input: $\mathcal{D}_{\text{train}}$: Training dataset for gradient optimization, $\mathcal{D}_{\text{evo}}$: Training dataset for evolutionary search, $\mathcal{M}_{\text{grad}}$: Model for gradient optimization, $\mathcal{M}_{\text{evo}}$: List of models for evolutionary search, ϵ: Perturbation limit, α: Learning rate, β: Momentum decay coefficient, max_iter: Maximum number of iterations, k: Maximum number of perturbed pixels per iteration, $\mathcal{P}$: Initial population size, G: Number of generations in evolutionary search
Output: δ: Optimized perturbation after all iterations
Initialization: $\delta \leftarrow Rand(0,1)$, $v \leftarrow 0$ ▷ Exponential moving average of the gradient
for $t = 0$ **to** max_iter $- 1$ **do**
    Fetch batch x from $\mathcal{D}_{\text{train}}$
    Compute the loss $\mathcal{L}_{\text{meta}}(\delta, x)$ using $\mathcal{M}_{\text{grad}}$
    $v_{t+1} \leftarrow \beta v_t + (1-\beta)\frac{\nabla_\delta \mathcal{L}_{\text{meta}}}{\|\nabla_\delta \mathcal{L}_{\text{meta}}\|_1}$
    $\delta_{t+1} \leftarrow \text{clip}(\delta_t + \alpha \cdot \text{sign}(v_{t+1}), -\epsilon, \epsilon)$
    if $t \mod iter == 0$ **then** ▷ Execute evolutionary search every $iter$ iterations
        Fetch batch x_{evo} from $\mathcal{D}_{\text{evo}}$
        Initialize population $\mathcal{P}$
        Generate new individuals through crossover and mutation
        for each individual $\eta \in \mathcal{P}$ **do**
            Evaluate η using all models in $\mathcal{M}_{\text{evo}}$ on x_{evo}
        end for
        Perform non-dominated sorting to select the best η
        $\delta_t \leftarrow \delta_t + \eta$ ▷ Update δ_t with the best η found
    end if
end for
Output the optimized perturbation δ

In Fig. **??**, we show the relationship between the Generation Number, population size, and attack success rate during the evolutionary process.Since the primary goal of the evolutionary search is to optimize the universal perturbation, we do not use the configuration with the highest attack success rate in practice to save time. Instead, we set the Generation Number to 150 and the population size to 2.

7 Experiments

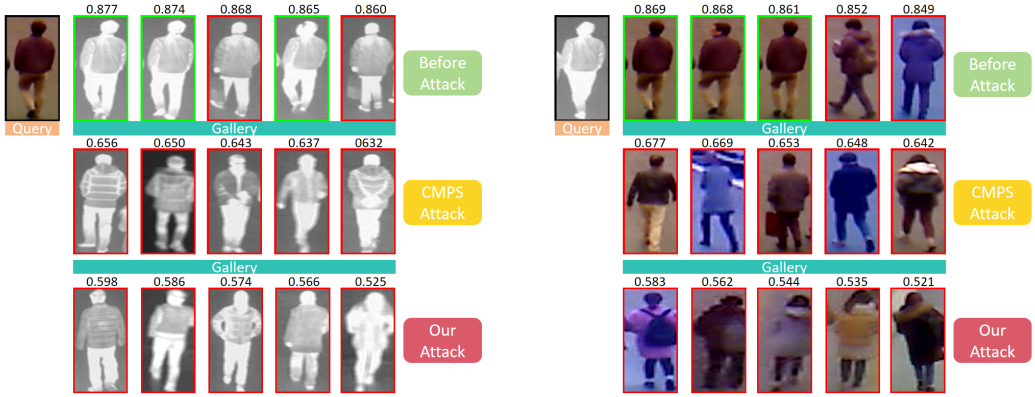

Figure 5: Comparison with the state-of-the-art attack method CMPS attack retrieval results. Before and after our attack on RegDB, the top five predictions of DDAG (a state-of-the-art cross-modality ReID baseline). Green boxes indicate correctly matched images, while red boxes indicate mismatched images. The left image corresponds to visible modality retrieving non-visible modality, and the right image vice versa.

Table 2: Results for attacking cross-modality ReID systems on the RegDB dataset. It reports on visible images querying thermal images and vice versa. Rank at r accuracy (%) and mAP (%) are reported.

Settings		Visible to Thermal				Thermal to Visible			
Method	Venue	$r=1$	$r=10$	$r=20$	mAP	$r=1$	$r=10$	$r=20$	mAP
AGW baseline [36]	TPAMI 2022	70.05	86.21	91.55	66.37	70.49	87.21	91.84	65.90
M-FGSM attack [1]	TPAMI 2020	29.34	52.90	61.44	23.35	23.64	40.36	48.61	18.57
LTA attack [8]	CVPR 2022	12.65	25.24	34.02	12.80	10.51	22.93	31.79	9.74
ODFA attack [39]	IJCV 2023	28.57	51.42	60.58	21.84	17.26	33.27	42.92	15.27
Col.+Del. attack [33]	TPAMI 2023	5.12	16.83	22.10	4.94	4.92	14.47	23.04	4.86
CMPS attack [7]	Arxiv 2024	2.29	9.06	18.35	3.92	1.93	11.44	19.30	3.46
Our attack*	————	1.64	8.86	17.52	2.71	1.66	10.38	17.54	2.85
Our attack	————	1.36	8.54	16.17	2.26	1.07	9.87	16.62	2.11
DDAG baseline [35]	ECCV 2020	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
M-FGSM attac [1]	TPAMI 2020	30.86	54.16	61.98	24.01	25.83	42.12	49.76	19.33
LTA attack [8]	CVPR 2022	11.65	23.20	32.73	11.41	9.76	21.53	29.96	9.23
ODFA attack [39]	IJCV 2023	29.64	52.74	60.74	23.88	24.06	39.75	46.25	18.64
Col.+Del. attack [33]	TPAMI 2023	4.68	13.55	18.57	4.39	4.23	12.75	20.82	4.05
CMPS attack [7]	Arxiv 2024	1.33	10.28	19.06	3.79	1.35	9.52	17.52	3.19
Our attack*	————	1.15	9.83	17.26	2.97	1.27	9.36	16.91	3.06
Our attack	————	0.96	9.37	16.37	2.04	1.11	9.19	16.38	2.83

Our device uses three GPUs of RTX2080ti with 11GB memory.

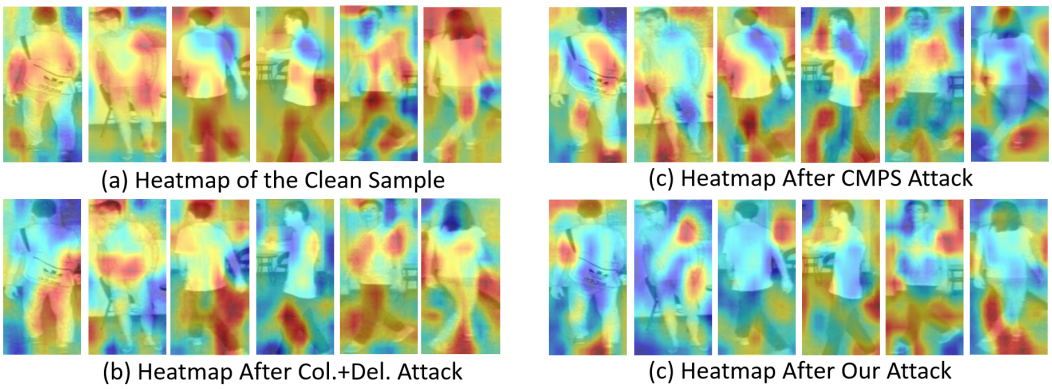

Figure 6: Comparison with state-of-the-art adversarial attack methods on attention heatmaps.

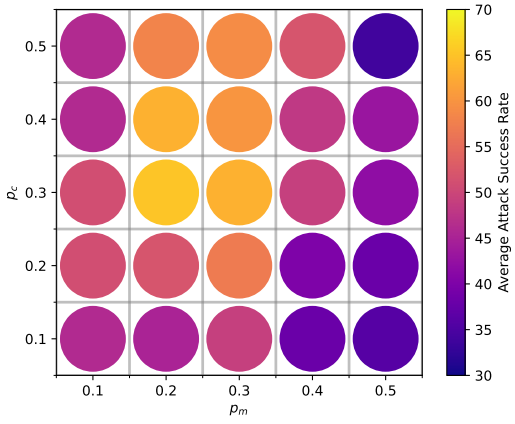

Figure 7: Correlation plots showing the average success rate of different pm and pc configurations on SYSU images.

8 Details of the Evolutionary Algorithm

8.1 Initialization

The attack method initializes by setting the number of perturbed pixels to k and constructing a set $\mathcal{P}$ of s solutions. Each solution is created by uniformly sampling pixel locations from the set $\{1, \cdots, h \cdot w\}$. The initial perturbation values for each color channel are randomly sampled from the set $\{-1, 0, 1\}$:

$$\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_s\}, \quad \mathcal{P}_i \in \{-1, 0, 1\} \tag{13}$$

where:

- $\mathcal{P}$: The set of initial solutions.
- s: The total number of solutions.

When invoking the evolutionary process, we set a fixed number of iterations to 200.

8.2 Crossover

The crossover operation aims to generate new solutions by combining traits from two parent solutions. This process helps in exploring the solution space and inheriting beneficial traits from the parents.

Crossover combines traits from two parent solutions P_a and P_b to generate new solutions:

$$M' = (M_a \cup M_b) \setminus (M_a \cap M_b) \tag{14}$$

$$\Delta' = \Delta_a \cup \Delta_b \tag{15}$$

where:

- M_a, M_b: Sets of pixel locations from the parent solutions P_a and P_b, respectively.
- M': The set of pixel locations for the new solution.
- Δ_a, Δ_b: Sets of perturbation values corresponding to M_a and M_b.
- Δ': The set of perturbation values for the new solution.

8.3 Mutation

The mutation operation introduces variations by modifying a subset of pixel locations. This process helps in exploring new regions of the solution space and maintaining diversity among solutions.

Mutation introduces variations to solutions by modifying a subset of pixel locations:

$$M'' = (M' \setminus A) \cup B, \quad \Delta'' = (\Delta' \setminus \Delta_A) \cup \Delta_B \tag{16}$$

where:

- M': The set of pixel locations before mutation.
- A: A randomly selected subset of pixel locations to be replaced.
- B: A new subset of pixel locations to be introduced (obtained through random selection).
- M'': The set of pixel locations after mutation.
- Δ': The set of perturbation values before mutation.
- Δ_A: The perturbation values corresponding to A.
- Δ_B: The perturbation values corresponding to B.
- Δ'': The set of perturbation values after mutation.

8.4 Evaluation

The evaluation process calculates the objective vector $\Phi(f_{adv}) = (\tilde{\mathcal{D}}(f_{adv}), \tilde{\mathcal{S}}(f_{adv}), \|\eta\|_2, \|\eta\|_0)^T$ for each solution to measure its quality. This step is crucial for selecting the best solutions.

8.5 Selection

Algorithm 2 Non-Dominating Sorting for Multiform Attack

Input: Combined population $\mathcal{P}$, objective vectors Φ
Output: Set of fronts $\mathcal{S}_f$
$\mathcal{S}_f \leftarrow \{\}$ // Initialize the set of fronts
for $p \in \mathcal{P}$ **do**
    $S_p \leftarrow \{\}$ // Set of p dominated solutions
    $n_p \leftarrow 0$ // Domination counter of p
    for $q \in \mathcal{P}$ **do**
        if DOMINATES(p, q, Φ) **then** //if p dominates q
            $S_p \leftarrow S_p \cup \{q\}$
        else if DOMINATES(q, p, Φ) **then**
            $n_p \leftarrow n_p + 1$
        end if
    end for
    if $n_p == 0$ **then**
        $p_{\text{rank}} \leftarrow 1$ // p belongs to the first front
        $\mathcal{S}_{f1} \leftarrow \mathcal{S}_{f1} \cup \{p\}$
    end if
end for
$i \leftarrow 1$ // Initialize front counter
while $\mathcal{S}_{fi} \neq \emptyset$ **do**
    $Q \leftarrow \{\}$ // Store solutions of the next front
    for $p \in \mathcal{S}_{fi}$ **do**
        for $q \in S_p$ **do**
            $n_q \leftarrow n_q - 1$
            if $n_q == 0$ **then**
                $q_{\text{rank}} \leftarrow i + 1$
                $Q \leftarrow Q \cup \{q\}$
            end if
        end for
    end for
    $i \leftarrow i + 1$
    $\mathcal{S}_{f(i+1)} \leftarrow Q$
end while
return $\mathcal{S}_f$

Algorithm 3 Function to Determine if One Solution Dominates Another

function DOMINATES(x, y, Φ)
    $isBetter \leftarrow$ False
    for $i \leftarrow 1$ to length(Φ) **do**
        if $\Phi_i(x) > \Phi_i(y)$ **then**
            $isBetter \leftarrow$ True
        else if $\Phi_i(x) < \Phi_i(y)$ **then**
            return False
        end if
    end for
    return $isBetter$
end function

The selection process evaluates solutions and chooses the best individuals to pass their genetic material to the next generation. Initially, non-dominated sorting is used to select the best solutions from the combined parent and offspring populations.

Selection uses non-dominated sorting to choose the best solutions for the next generation:

$$P' = \text{NonDominatedSort}(P \cup O) \tag{17}$$

where:

- P and O are the parent and offspring populations, respectively.
- P': The new population of solutions after selection.
- NonDominatedSort: A sorting method that selects non-dominated solutions based on their objective vectors.

The selection process can be represented as:

$$\text{Selected Solutions}\eta = \arg\min_{s \in P'} \Phi(\mathcal{F}(x + \delta + \eta_s)) \tag{18}$$

where:

- P' is the set of non-dominated solutions selected by non-dominated sorting.

In this way, we first use non-dominated sorting to select a set of non-dominated solutions and then choose the solutions with the smallest objective function values from these non-dominated solutions.

9 Feasibility Analysis of Evolutionary Search

Here, we simplify the problem and appropriately reformulate it to facilitate analysis.

9.1 Objective of Sparse Perturbations

The goal of introducing sparse perturbations is to minimize the number of changes made to the input while maximizing the error induced across various models. The fitness function, which evaluates the effectiveness of a perturbation across multiple models, is central to this process.

9.2 Fitness Function

Let Φ represent a collection of models $\{M_1, M_2, \ldots, M_k\}$, where each model M_i has an associated misclassification rate $r_i(x)$ for an input x. The fitness function $f(x)$ for a perturbation vector x is defined as:

$$f(x) = \sum_{i=1}^{k} w_i \cdot r_i(x) - \lambda \cdot \|x\|_0 \tag{19}$$

Here, $\|x\|_0$ denotes the sparsity of the perturbation vector (the number of non-zero elements), w_i is the weight associated with model M_i, reflecting its importance in the overall fitness, and λ is a regularization parameter that controls the significance of sparsity.

9.3 Generation of Sparse Perturbations

Evolutionary algorithms utilize selection, crossover, and mutation operations to generate new perturbations. The mutation operation, designed to maintain sparsity while enhancing fitness, is defined as:

$$x' = x + \delta \tag{20}$$

where δ is a small change vector, typically non-zero in only a few components of x. δ is chosen to maximize $f(x + \delta)$.

9.4 Convergence Analysis

By appropriately choosing the parameters λ and w_i, evolutionary search can converge to high-quality sparse perturbations. Theoretically, consider the change in the fitness function over iterations. assuming:

$$\lim_{t\to\infty} f(x^{(t)}) = f^* \tag{21}$$

where $x^{(t)}$ is the perturbation vector after the t-th iteration, and f^* represents the optimal achievable fitness.

To ensure that each iteration does not decrease the fitness, we need:

$$f(x^{(t+1)}) \geq f(x^{(t)}) \tag{22}$$

This indicates that the fitness is non-decreasing over iterations, suggesting that the algorithm is at least locally optimal. By designing δ to ensure that $f(x+\delta) > f(x)$, this can be achieved. This can be specifically implemented through techniques such as non-dominated sorting to select the most effective perturbations. Please note that our goal is to fine-tune the universal perturbation using the search process; in fact, we do not need to find the perturbation with the optimal fitness value.

Through the analysis provided, we theoretically demonstrate that evolutionary search can generate effective sparse perturbations that are effective across multiple models.

10 Feasibility Analysis of Enhancing Attack Transferability

Here, we simplify the problem and appropriately reformulate it to facilitate analysis.

10.1 Objective Definition

The goal is to find a fine-tuned perturbation δ_f that significantly increases the misclassification rate across multiple models while maintaining the sparsity of the perturbation. For this purpose, we introduce a fitness function:

$$f(\delta_f) = \sum_{i=1}^{k} w_i \cdot r_i(\delta_u + \delta_f) - \lambda \cdot \|\delta_f\|_0 \tag{23}$$

where $\|x\|_0$ denotes the sparsity of the perturbation vector (the number of non-zero elements), w_i is the weight associated with model M_i, reflecting its importance in the overall fitness, and λ is a regularization parameter that controls the significance of sparsity.

10.2 Complementarity Measure α_i

To quantify the complementarity effect of the fine-tuned perturbation δ_f on the universal perturbation δ_u, we define the complementarity measure α_i as:

$$\alpha_i = \frac{r_i(\delta_u + \delta_f) - r_i(\delta_u)}{r_i(\delta_u)} \tag{24}$$

Here, α_i represents the increase in the misclassification rate of model M_i when the fine-tuned perturbation δ_f is added to the universal perturbation δ_u. Ideally, α_i should be significantly greater than 0, indicating a strong complementarity effect.

10.3 Optimization of the Fitness Function

The goal of the evolutionary search is to maximize the fitness function $f(\delta_f)$. By selecting appropriate mutation vectors δ and performing crossover operations, we ensure that each iteration finds a better fine-tuned perturbation δ_f. Specifically, we aim to find δ_f such that:

$$\sum_{i=1}^{k} w_i \cdot r_i(\delta_u + \delta_f) - \lambda \cdot \|\delta_f\|_0 \tag{25}$$

is maximized.

10.4 Theoretical Guarantee

Through selection, crossover, and mutation operations, evolutionary search can effectively optimize the perturbation vector δ_f to exhibit a high complementarity measure α_i across multiple models.

Specifically, for each iteration, we have:

$$\alpha_i^{(t+1)} = \frac{r_i(\delta_u + \delta_f^{(t+1)}) - r_i(\delta_u)}{r_i(\delta_u)} \geq \alpha_i^{(t)} = \frac{r_i(\delta_u + \delta_f^{(t)}) - r_i(\delta_u)}{r_i(\delta_u)} \tag{26}$$

This indicates that the complementarity measure of δ_f increases with each iteration, thereby significantly improving the misclassification rate across multiple models.

By appropriately choosing the parameters λ and w_i, evolutionary search can converge to high-quality sparse perturbations. Theoretically, considering the change in the fitness function over iterations, we assume:

$$\lim_{t \to \infty} f(x^{(t)}) = f^* \tag{27}$$

where $x^{(t)}$ is the perturbation vector after the t-th iteration, and f^* represents the optimal achievable fitness.

To ensure that each iteration does not decrease the fitness, we need:

$$f(x^{(t+1)}) \geq f(x^{(t)}) \tag{28}$$

This indicates that the fitness is non-decreasing over iterations, suggesting that the algorithm is at least locally optimal. By designing δ to ensure $f(x + \delta) > f(x)$, this can be achieved. Specifically, this can be implemented through non-dominated sorting to select the most effective perturbations.

By selecting appropriate mutation vectors δ and performing crossover operations, evolutionary search can effectively optimize the perturbation vector δ_f to exhibit a high complementarity measure α_i across multiple models, i.e.:

$$\alpha_i = \frac{r_i(\delta_u + \delta_f) - r_i(\delta_u)}{r_i(\delta_u)} \gg 0 \tag{29}$$

This indicates that the fine-tuned perturbation significantly increases the misclassification rate for each model, thereby improving the overall attack effectiveness.

Through the above mathematical analysis and theoretical derivation, we demonstrate that evolutionary search can find fine-tuned perturbations δ_f with a high complementarity measure α_i, effectively refining the universal perturbation δ_u, thus achieving better transferability and attack effectiveness across multiple models. This method leverages the global search capability of evolutionary algorithms and the complementarity between different models, ensuring that the perturbations are effective across various models.

11 CnMix Processing Algorithm

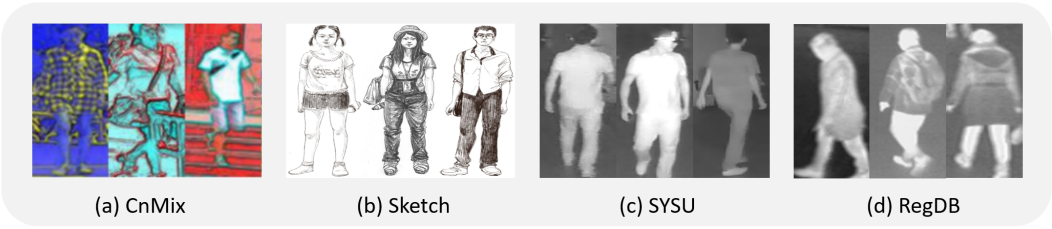

Figure 8: Examples of the four different modalities corresponding to the four datasets used in this paper.

Algorithm 4 Random Channel Mixing of Grayscale and Sketch Images

Input: img (Input RGB image), G (Probability of converting to grayscale), G_{rgb} (Probability of mixing grayscale and RGB), S_{rgb} (Probability of mixing sketch and RGB)
Output: $output\_img$ (Transformed image)
function TOSKETCH(img)
    $img\_np \leftarrow$ Convert img to numpy array
    $img\_inv \leftarrow 255 - img\_np$ ▷ Invert image colors
    $img\_blur \leftarrow$ Gaussian blur of img_inv
    $img\_blend \leftarrow \frac{img\_np}{255 - img\_blur} \times 256$ ▷ Blend original and blurred images
    return Convert img_blend to image
end function
function RANDOM_CHOOSE($r, g, b, gray\_or\_sketch$)
    $p \leftarrow [r, g, b, gray\_or\_sketch, gray\_or\_sketch]$
    $idx \leftarrow [0, 1, 2, 3, 4]$
    Shuffle idx
    return Merge channels $p[idx[0]], p[idx[1]], p[idx[2]]$ into RGB image
end function
function FUSE_RGB_GRAY_SKETCH(img, G, G_{rgb}, S_{rgb})
    $r, g, b \leftarrow$ Split img into RGB channels
    $gray \leftarrow$ Convert img to grayscale
    $p \leftarrow$ Random value between 0 and 1
    if $p < G$ **then**
        return Merge $gray, gray, gray$ into RGB image
    else if $p < G + G_{rgb}$ **then**
        $output\_img \leftarrow$ RANDOM_CHOOSE($r, g, b, gray$)
        return $output\_img$
    else if $p < G + G_{rgb} + S_{rgb}$ **then**
        $sketch \leftarrow$ TOSKETCH($gray$)
        $output\_img \leftarrow$ RANDOM_CHOOSE($r, g, b, sketch$)
        return $output\_img$
    else
        return img
    end if
end function

12 Discussion

12.1 Ethical Considerations

In this study, we introduce a novel cross-modal adversarial attack method that enhances the transferability and concealment of adversarial attacks through a gradient-evolutionary multiform optimization framework. This research offers a new perspective on understanding and enhancing the security of cross-modal systems, but it also raises a series of ethical and safety concerns about the potential negative impacts of adversarial attack techniques. Adversarial attack technology can be maliciously exploited, posing a serious threat to public safety.

However, we recognize the positive value of adversarial attack research. It reveals vulnerabilities in existing systems, prompting academia and industry to make in-depth improvements to the robustness of machine learning models. The positive impact of this study lies in its potential to combine adversarial training with the attack methods presented to enhance system security and bring positive social impacts. Therefore, we emphasize the importance of conducting adversarial attack research within an ethical framework and encourage further development of defensive technologies to build a safer and more reliable technological environment.

12.2 Limitations and Future Work

Here, we need to acknowledge the limitations of the proposed method and identify potential areas for future research. Firstly, current attack techniques mainly focus on enhancing the stealthiness of perturbations for RGB images. In other types of images, such as infrared or thermal images in monochromatic modes, the invisibility of perturbations remains a challenge. This is because these modalities typically lack rich color information, making even subtle perturbations noticeable. Future research needs to explore how to effectively hide adversarial perturbations in these different image modalities to improve the applicability and stealthiness of attacks.

Secondly, as the number of modalities processed increases, the computational demand rises sharply, which can lead to significant resource consumption in practical applications. The high demand for computational resources in current methods limits their feasibility in resource-constrained environments. Therefore, future work could explore more efficient algorithms by incorporating state-of-the-art evolutionary computation methods to reduce the computational burden when handling large-scale or multi-modal data.