Cross-Modality Perturbation Synergy Attack for Person Re-identification

Yunpeng Gong[1], Zhun Zhong[2], Yansong Qu[1], Zhiming Luo[1], Rongrong Ji[1], and Min Jiang[*1]

[1]School of Informatics, Xiamen University
[2]School of Computer Science and Information Engineering, Hefei University of Technology

Abstract

In recent years, there has been significant research focusing on addressing security concerns in single-modal person re-identification (ReID) systems that are based on RGB images. However, the safety of cross-modality scenarios, which are more commonly encountered in practical applications involving images captured by infrared cameras, has not received adequate attention. The main challenge in cross-modality ReID lies in effectively dealing with visual differences between different modalities. For instance, infrared images are typically grayscale, unlike visible images that contain color information. Existing attack methods have primarily focused on the characteristics of the visible image modality, overlooking the features of other modalities and the variations in data distribution among different modalities. This oversight can potentially undermine the effectiveness of these methods in image retrieval across diverse modalities. This study represents the first exploration into the security of cross-modality ReID models and proposes a universal perturbation attack specifically designed for cross-modality ReID. This attack optimizes perturbations by leveraging gradients from diverse modality data, thereby disrupting the discriminator and reinforcing the differences between modalities. We conducted experiments on three widely used cross-modality datasets, namely RegDB, SYSU, and LLCM. The results not only demonstrate the effectiveness of our method but also provide insights for future improvements in the robustness of cross-modality ReID systems.

1 Introduction

With the rapid advancement of surveillance technology, person re-identification (ReID) [1–4] has emerged as a pivotal component in the realm of security, garnering escalating attention. ReID constitutes a fundamental task in computer vision [5–9], aiming to precisely identify the same individual across diverse locations and time points by analyzing pedestrian images captured through surveillance cameras [10]. The challenges inherent in this task encompass factors such as changes in viewpoint, lighting conditions [11, 12], occlusion [13, 14], and pose variations, culminating in significant appearance variations of the same individual across distinct camera views [15].

In traditional ReID, where samples are image-based, the conventional methodology centers on matching visible to visible (RGB to RGB) data. However, when dealing with diverse scenarios and

[*]Min Jiang and Yunpeng Gong are with the Department of Artificial Intelligence, Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, School of Informatics, Key Laboratory of Digital Protection and Intelligent Processing of Intangible CulturalHeritage of Fujian and Taiwan, Ministry of Culture and Tourism, Xiamen University, Xiamen 361005, Fujian, P.R. China (e-mail: minjiang@xmu.edu.cn; gongyunpeng@stu.xmu.edu.cn or fmonkey625@gmail.com).
Corresponding author: Min Jiang

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0735

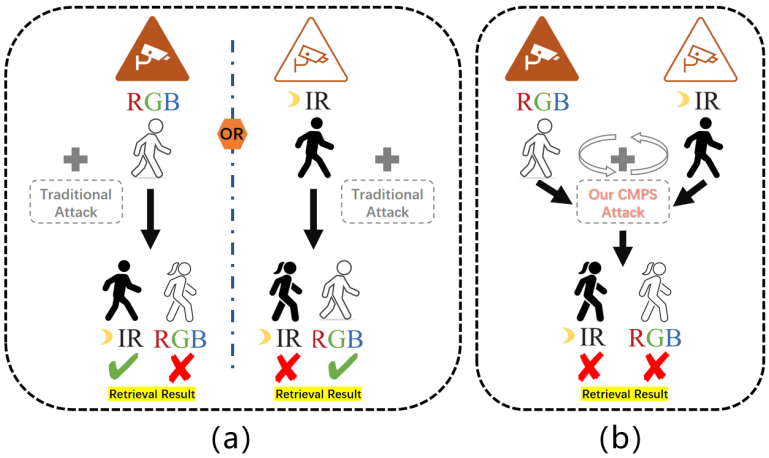

Figure 1: Comparison between traditional and proposed methods: Fig.(a) illustrates traditional attack methods (e.g., FGSM [16], PGD [17]), which are primarily designed for single-modal tasks and lack mechanisms to associate multiple modalities, making them ineffective in simultaneously misleading retrieval results across different modalities. Fig.(b) illustrates the proposed method, which employs an intrinsic mechanism to effectively associate different modalities, thereby misleading retrieval results across multiple modalities simultaneously.

conditions, especially involving multiple image modalities such as RGB and infrared images, the system needs to intricately handle the differences in images from different modalities [18–20]. This is essential to ensure that the system exhibits better robustness across different modalities. Hence, cross-modality ReID is considered more challenging due to the need for addressing these modality differences [21, 22].

Cross-modal ReID [23, 24, 21, 25] plays a crucial role in significantly expanding the applicability of traditional ReID methods, focusing on addressing complex matching issues between different image modalities. In practical surveillance systems, the simultaneous use of multiple sensors, such as RGB cameras and infrared cameras, is a common scenario. This task requires innovative solutions to effectively bridge the differences between various modalities, ensuring robust and accurate re-identification of pedestrians in heterogeneous sensor outputs.

Currently, most research on the security of ReID focuses on single-modality systems based on RGB images [26–31], while the security of cross-modality ReID systems has received insufficient attention. The challenge in cross-modality attacks arises from significant visual differences among different modality inputs, requiring attackers to effectively capture shared features from each modality for perturbation implementation. However, as shown in Fig. 1, existing attack methods in cross-modal scenarios require optimizing perturbations separately for each modality, lacking an intrinsic mechanism to capture shared knowledge between different modalities, which limits the success rate of the attacks. To address this issue, we propose a synergistic optimization method combined with triplet loss, utilizing information from different modalities to optimize the universal perturbation. This method pushes the features of different samples into a common sub-region that affects the model's accuracy, as shown in Fig. 2.

Specifically, we propose the Cross-Modality Perturbation Synergy (CMPS) method, a universal perturbation approach designed specifically for cross-modality ReID systems. This method simultaneously leverages gradient information from multiple modalities to jointly optimize universal perturbations across visible and infrared images. CMPS incorporates cross-modality triplet loss to ensure feature consistency across different modalities, enhancing the generality of the perturbation. During the synergistic optimization process, CMPS iteratively updates gradients from various modalities within a unified optimization framework, effectively capturing and utilizing shared features across modalities. To further reduce visual differences between modalities, we introduce cross-modality attack augmentation, converting images into grayscale to standardize their visual representation and facilitate the learning of modality-agnostic perturbations. As a result, these universal perturbations push the features of different samples toward a common region in the feature space, significantly diminishing the model's ability to accurately distinguish identities in cross-modality scenarios, thereby successfully deceiving the model.

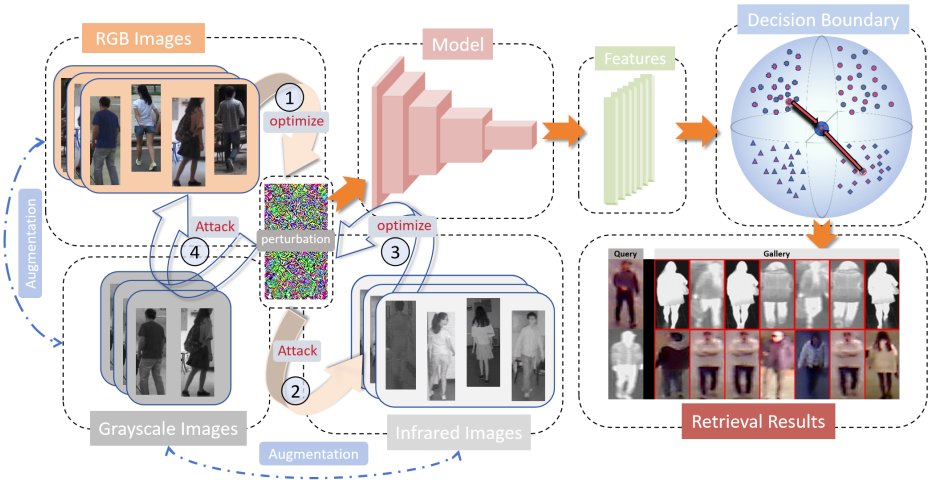

Figure 2: Illustration of the CMPS attack framework. We generate homogeneous grayscale images through random grayscale transformations to reduce the differences between modalities, aiding in the learning of a universal perturbation. The process is as follows: first, the gradient from one modality is used to optimize the universal perturbation, which is then applied to another modality's images to generate adversarial samples for attacks. The new modality's gradient is then used to further optimize the perturbation and attack the next modality. By aggregating feature gradients from different modalities, we iteratively learn a universal perturbation, pushing samples toward a common region in the manifold. The manifold is represented as a sphere, with identical shapes but different colors representing the same person's features across modalities. This method captures shared knowledge between modalities, enabling more effective learning of cross-modal universal perturbations.

In our experiments on widely utilized cross-modality ReID datasets, including RegDB [32], SYSU [33] and LLCM [24], we not only showcase the effectiveness of our proposed method but also provide insights for fortifying the robustness of cross-modality ReID systems in the future. This research contributes by bridging gaps in current studies and introducing novel perspectives to study the security challenges in cross-modality ReID systems.

The main contributions of our work can be summarized as:

• To the best of our knowledge, our work is the first to investigate vulnerabilities in cross-modality ReID models. By explicitly incorporating cross-modality constraints into the synergistic optimization process, we enhance the universality of the learned cross-modality perturbations. Additionally, we provide mathematical analysis to demonstrate the superiority of our proposed method over traditional approaches.

• We propose a cross-modality attack augmentation method, utilizing random grayscale transformations to narrow the gap between different modalities, aiding our cross-modality perturbation synergy attack in better capturing shared features across modalities.

• Extensive experiments conducted on three widely used cross-modality ReID benchmarks demonstrate the effectiveness of our proposed cross-modality attack. Our method exhibits good transferability even when attacking different models. The code will be available at `https://github.com/finger-monkey/cmps__attack`.

2 Related Works

Adversarial Attack. Adversarial attacks are a technique involving the clever design of small input perturbations with the aim of deceiving machine learning models, leading them to produce misleading outputs. This form of attack is not confined to the image domain but extends to models in various fields, including speech [34] and text [35–37]. Typically, the goal of adversarial attacks is to tweak input data in a way that causes the model to make erroneous predictions when handling these subtly modified samples [16, 38–40]. In the early stages of research, adversarial attacks had to

https://doi.org/10.52202/079017-0735

be customized for each specific sample. However, with the evolution of related studies, universal perturbation [41] attacks were introduced, aiming to find perturbations effective across multiple samples rather than tailored to individual instances. Research on universal perturbation attacks seeks to expose vulnerabilities in models, prompting designers to enhance their robustness to withstand a broader range of adversarial challenges.

Adversarial Attacks in ReID. Some ReID attack methods have been proposed, with current research predominantly focusing on RGB-RGB matching. These methods mainly include: Metric-FGSM [29] extends some techniques, inspired by classification attacks, into a category known as metric attacks. These encompass Fast Gradient Sign Method (FGSM) [16], Iterative FGSM (IFGSM), and Momentum IFGSM (MIFGSM) [42]. The Furthest-Negative Attack (FNA) [30] integrates hard sample mining [43] and triple loss to employ pushing and pulling guides. These guides guide image features towards the least similar cluster while moving away from other similar features. Deep Mis-Ranking (DMR) [31] utilizes a multi-stage network architecture to pyramidally extract features at different levels, aiming to derive general and transferable features for adversarial perturbations. Gong et al. [28] proposed a local transformation attack (LTA) method specifically aimed at attacking color features without requiring additional reference images, and discussed effective defense strategies against current ReID attacks. The Opposite-direction Feature Attack (ODFA) [26] exploits feature-level adversarial gradients to generate examples that guide features in the opposite direction with an artificial guide. Yang et al. [27] introduced a combined attack named Col.+Del., which integrates UAP-Retrieval [44] with color space perturbations [45]. While this method also explores universal perturbations in ReID, its generality is limited due to the inability to leverage color information in cross-modality problems and the lack of a mechanism for associating different modality information. In contrast to the aforementioned approaches, our focus lies on addressing cross-modality challenges.

Algorithm 1 Procedure of CMPS attack

Input: Visible images I_{RGB} and infrared (or thermal) images I_{ir} from dataset S, cross-modality ReID model f trained on S, adversarial bound ϵ, momentum value θ, iteration step size α, iteration epoch *iter_epoch*.
Output: Cross-modality universal perturbation η.
Initialize η with random noise $\eta \leftarrow Rand(0, 1)$, $\Delta^0 = 0$.
for i in *iter_epoch* **do**
  repeat
    Sample a mini-batch of visible images I_{RGB} and infrared (or thermal) images I_{ir} with n samples
    $\hat{I}_{RGB} \leftarrow I_{RGB} + \eta$
    Use infrared images to compute the triplet loss L_{RGB} for visible images (Eq. 4)
    Compute gradient Δ_{RGB} of L_{RGB} w.r.t. η:
    $\Delta_{RGB} \leftarrow \theta \cdot \Delta^{i-1} + \frac{\partial L_{RGB}}{\partial \eta}$
    Update perturbation η:
    $\eta \leftarrow \text{clip}(\eta + \alpha \cdot \text{sign}(\Delta_{RGB}), -\epsilon, \epsilon)$
    $\hat{I}_{ir} \leftarrow I_{ir} + \eta$
    Use visible images to compute the triplet loss L_{ir} for infrared images (Eq. 7)
    Compute gradient Δ_{ir} of L_{ir} w.r.t. η:
    $\Delta^i \leftarrow \theta \cdot \Delta_{RGB} + \frac{\partial L_{ir}}{\partial \eta}$
    Update perturbation η:
    $\eta \leftarrow \text{clip}(\eta + \alpha \cdot \text{sign}(\Delta^i), -\epsilon, \epsilon)$
  until all mini-batches are processed
end for
return η

3 Methodology

In this section, we introduce a universal perturbation designed for cross-modality attacks, referred to as the Cross-Modality Perturbation Synergy (CMPS) attack. Considering the significant differences between different modalities, we propose a attack augmentation method to bridge the gap between modalities, aiding in enhancing the perturbation's universality across different modalities. Our

objective in addressing this problem is to find a universal adversarial perturbation, denoted as η, capable of misleading the retrieval ranking results of cross-modality ReID models. The adversarial operation involves adding η to a query image I. The perturbed query image, denoted as $I_{adv} = I + \eta$, is then used to retrieve from the gallery and deceive the cross-modality ReID model f. The algorithm is summarized in Alg. 1.

3.1 Overall Framework

In Fig. 2, we illustrate the overall framework of the proposed CMPS attack. During the training phase, we optimize η using our cross-modality attack augmentation method, which leverages images from different modalities to bridge their inherent differences and enhance cross-modality universality. In the attack phase, the optimized η deceives reID models, leading to inaccurate ranking lists. Section 3.2 outlines the framework and overall optimization objective, providing a macro-level overview. Section 3.4 delves into the specific process of perturbation optimization across different modalities.

3.2 Optimizing Loss Functions for Attacking

Our study aims to deceive cross-modality ReID models using a universal perturbation. We have specifically designed a triplet loss tailored for our proposed attack method, which can correlate different modalities and influence the distance relationships between images from different modalities.

We follow the approach of [44] to optimize the perturbation using cluster centroids. This method directly impacts the similarity between pedestrian identities in the ReID model's feature space (rather than the similarity between individual samples), making it more effective. Subsequently, leveraging the acquired cluster centroids, we apply our triplet loss to distort the pairwise relations between pedestrian identities. This process can be represented as follows:

$$\begin{aligned} L = & \max\left[\left(\|C_g^n - f_{RGB}^{adv}\|_2 - \|C_{ir}^p - f_{RGB}^{adv}\|_2 + \rho\right), 0\right] \\ & + \max\left[\left(\|C_{ir}^n - f_g^{adv}\|_2 - \|C_{RGB}^p - f_g^{adv}\|_2 + \rho\right), 0\right] \\ & + \max\left[\left(\|C_{RGB}^n - f_{ir}^{adv}\|_2 - \|C_g^p - f_{ir}^{adv}\|_2 + \rho\right), 0\right]. \end{aligned} \tag{1}$$

As shown in Fig. 3, the loss function mentioned above fully leverages the triplet-wise relationships across different modality. Through this loss, we are able to pull the negative samples of each modality closer to the adversarial samples and push the positive samples of each modality away from the adversarial samples. Here, C_{RGB}^p and C_{RGB}^n represent the cluster centroids of the positive samples to push and negative samples to pull, respectively, in the original visible (RGB) image feature space of the training data. Similar definitions apply to other modalities. f_{RGB}^{adv}, f_g^{adv}, and f_{ir}^{adv} denote the perturbed features of the disturbed image in the visible, grayscale, and infrared (or thermal) modalities, respectively.

3.3 Cross-Modality Attack Augmentation Method

Intuitively, as illustrated in Fig. 4, maximizing the overlap of common factors across different modalities facilitates the capture of shared features by the learned perturbation. Grayscale images, being inherently homogeneous, serve as effective mediators between diverse modalities. Consequently, we introduce random grayscale transformations into adversarial attack methods, referred to as Cross-Modality Attack Augmentation. This approach guides cross-modality perturbations by leveraging homogeneous grayscale images sourced from diverse modalities. The primary objective is to explore the underlying structural relationships across heterogeneous modalities.

The process of grayscale transformation can be represented as follows:

$$t(R, G, B) = 0.299R + 0.587G + 0.114B, \tag{2}$$

The function t(·) represents the grayscale transformation using ITU-R BT.601-7 standard weights, combining the RGB channels of each pixel into a single grayscale channel. From this, we construct a 3-channel grayscale image x_g by replicating the grayscale channel:

$$x_g = [t(R, G, B), t(R, G, B), t(R, G, B)]. \tag{3}$$

https://doi.org/10.52202/079017-0735

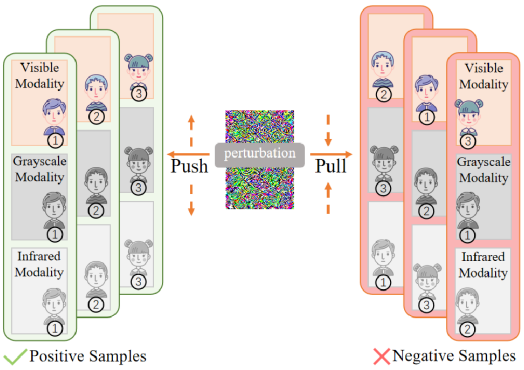

Figure 3: Schematic illustration of triplet relationship-guided universal perturbation learning for cross-modality ReID.

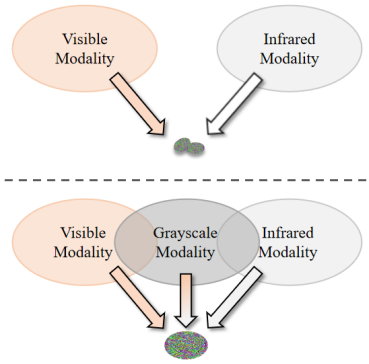

Figure 4: Cross-modality attack augmentation: bridging gap between visible and non-visible (infrared) modalities with grayscale.

3.4 Cross-Modality Perturbation Synergy Attack

To synergistically utilize gradient information from diverse modalities for perturbation optimization, narrow the gap between different modalities to better capture shared knowledge, we adopt the following training process to generate a universal perturbation:

(1) Learning the visible modality. For a given batch of visible images with n samples, we extract and perturb their features using the cross-modality ReID model. We update the temporary perturbation η iteratively using Momentum-Inertia Stochastic Gradient Descent (MI-SGD), expressed as:

$$L_{RGB}(f_{RGB}^{adv}, \eta) = \max\left[\left(\|C_g^n - f_{RGB}^{adv}\|_2 - \|C_{ir}^p - f_{RGB}^{adv}\|_2 + \rho\right), 0\right], \tag{4}$$

$$\Delta_{RGB} = \theta\Delta'_{ir} + \frac{\nabla_\eta L_{RGB}}{\|\nabla_\eta L_{RGB}\|_1}, \tag{5}$$

$$\eta = \text{clip}(\eta + \alpha \cdot \text{sign}(\Delta_{RGB}), -\varepsilon, \varepsilon). \tag{6}$$

Here, θ represents the momentum value (set as $\theta = 1$), and Δ'_{ir} is derived from the previous iteration. The iteration step size is denoted by α (set as $\alpha = \frac{\epsilon}{12}$), where ϵ is the adversarial bound ($\epsilon = 8$, unless otherwise specified). We set the margin $\rho = 0.5$ in our triplet loss.

(2) Learning the grayscale modality. This part is executed through data augmentation. It is not considered as a separate module and is therefore not explicitly listed in Alg. 1. Specifically, during the perturbation learning process, we randomly transform visible or infrared (or thermal) images into homogeneous grayscale images, participating in the iterative optimization of adversarial perturbations. It is employed to bridge the gap between different modalities, thereby improving the universality of the perturbation across diverse modalities. In order to investigate the impact of different grayscale conversion probabilities on attack performance, we conducted a series of ablation experiments. For details, please refer to Fig. 5 in supplementary material.

(3) Learning the infrared (or thermal) modality. This step is similar to (1). We utilize the infrared (or thermal) images to learn the perturbation η with the our loss functions:

$$L_{ir}(f_{ir}^{adv}, \eta) = \max\left[\left(\|C_{RGB}^n - f_{ir}^{adv}\|_2 - \|C_g^p - f_{ir}^{adv}\|_2 + \rho\right), 0\right], \tag{7}$$

$$\Delta_{ir} = \theta\Delta_{RGB} + \frac{\nabla_\eta L_{ir}}{\|\nabla_\eta L_{ir}\|_1}, \tag{8}$$

https://doi.org/10.52202/079017-0735

$$\eta = \text{clip}(\eta + \alpha \cdot \text{sign}(\Delta_{ir}), -\varepsilon, \varepsilon). \tag{9}$$

Here, Δ_{RGB} derived from step (1). The main difference compared to the previous step lies in the perturbation applied to the input and the gradients related to momentum.

Theoretical Analysis. In traditional optimization, optimizing for one modality can render the perturbation suboptimal for the other, leading to a bias toward a single modality. In contrast, the proposed aggregated optimization method jointly optimizes both modalities, ultimately identifying a universal perturbation that enhances cross-modality attack performance. In the supplementary material 7, we provide a mathematical analysis demonstrating the effectiveness of this method compared to traditional attack methods that lack intrinsic correlations between different modalities.

4 Experiments

In this section, we compare our approach with several methods, including traditional classification attack methods FGSM [16] and PGD [17], traditional metric attack methods like Metric-FGSM [29], as well as state-of-the-art ReID attack methods such as LTA [28] *, ODFA[26] and Col.+Del.[27].

Datasets. We evaluate our proposed method on two commonly used cross-modality ReID datasets: SYSU-MM01 [33], RegDB [32] and LLCM [24]. SYSU-MM01 is a large-scale dataset with 395 training identities, captured by 6 cameras (4 RGB, 2 near-infrared) on the SYSU campus. It comprises 22,258 visible and 11,909 near-infrared images. The testing set consists of 95 identities with two evaluation settings. The query sets include 3803 images from two IR cameras. We conduct ten trials following established methods [46] and report the average retrieval performance. Please refer to [33] for the evaluation protocol. RegDB [32] is a smaller-scale dataset with 412 identities, each having ten visible and ten thermal images. we randomly select 206 identities (2,060 images) for training and use the remaining 206 identities (2,060 images) for testing. LLCM is a dataset designed specifically for cross-modality ReID in low-light environments. Compared to other datasets, its diverse scenarios and low-light conditions present greater challenges for attackers. This complexity and uncertainty make adversarial attacks more difficult to execute. We assess our model in two retrieval scenarios: visible-thermal and thermal-visible performance.

Evaluation Metrics. Following existing works [47], we employ Rank-k precision and Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) as evaluation metrics. Rank-1 represents the average accuracy of the top-ranked result corresponding to each cross-modality query image. mAP represents the mean average accuracy, where the query results are sorted based on similarity, and the closer the correct result is to the top of the list, the higher the precision. Please note that, for adversarial attacks, a lower accuracy indicates a more successful attack.

4.1 Performance on Cross-Modality ReID

We used AGW [21] and DDAG [25] as baseline models for testing on the RegDB and SYSU cross-modality ReID datasets. AGW (Attention Generalized mean pooling with Weighted triplet loss) enhances the learning capability of crucial features by integrating non-local attention blocks, learnable GeM pooling, and weighted regularization triplet loss. DDAG (Dynamic Dual-Attentive Aggregation) improves feature learning by combining intra-modality weighted-part attention and cross-modality graph structured attention, considering both part-level and cross-modal contextual cues. Additionally, we use DEEN [24] (Diverse Embedding Expansion Network) as baseline models for testing on the LLCM [24] cross-modality ReID datasets. The core idea of DEEN is to enhance the feature representation capability by introducing a diversity embedding mechanism. The network expands the embedding space, allowing features from visible and infrared images to align better in a high-dimensional space, thereby improving the accuracy of cross-modality matching.

The experiments encompass two scenarios: 1) Perturbing visible images (query) to disrupt the retrieval of infrared or thermal non-visible images (gallery). This is denoted as "Visible to Infrared" in Tab.1 and "Visible to Thermal" in Tab.2. 2) Perturbing infrared or thermal non-visible images (query) to interfere with the retrieval of visible images (gallery). This is indicated as "Infrared to Visible" in Tab.1 and "Thermal to Visible" in Tab.2.

*The LTA code is available at: https://github.com/finger-monkey/LTA_and_joint-defence

https://doi.org/10.52202/079017-0735

Table 1: Results for attacking cross-modality ReID systems on the SYSU [33] dataset. It reports on visible images querying infrared images and vice versa. Rank at r accuracy (%) and mAP (%) are reported. For the "Visible to Infrared" scenario, we used the all-search mode. For the "Infrared to Visible" scenario, we used the indoor-search mode.

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 [21]	TPAMI 2022	47.50	84.39	92.14	47.65	54.17	91.14	95.98	62.97
FGSM attack [16]	ICLR 2015	42.64	81.21	89.32	43.67	48.05	86.73	92.11	53.22
PGD attack [17]	ICLR 2018	39.14	76.80	85.42	40.91	43.68	82.54	89.14	48.56
M-FGSM attack [29]	TPAMI 2020	25.79	49.04	57.96	19.24	20.56	38.91	46.35	15.89
LTA attack [28]	CVPR 2022	8.42	21.25	27.98	9.16	20.92	32.18	36.80	15.24
ODFA attack [26]	IJCV 2023	25.43	47.49	56.38	19.00	14.62	29.92	36.42	11.35
Col.+Del. attack [27]	TPAMI 2023	3.23	14.48	20.15	3.27	4.12	16.85	21.27	3.89
Our attack	NeurIPS 2024	1.11	8.67	16.14	1.41	1.31	7.47	10.36	1.23
DDAG baseline [25]	ECCV 2020	54.75	90.39	95.81	53.02	61.02	94.06	98.41	67.98
FGSM attack [16]	ICLR 2015	48.27	86.02	91.34	49.55	53.87	90.15	94.58	57.84
PGD attack [17]	ICLR 2018	50.62	88.30	93.12	51.89	56.10	91.54	96.13	59.22
M-FGSM attack [29]	TPAMI 2020	28.36	52.47	60.76	23.11	24.85	40.74	49.23	18.40
LTA attack [28]	CVPR 2022	10.54	23.08	30.47	12.28	18.93	34.12	41.52	15.04
ODFA attack [26]	IJCV 2023	27.75	50.26	59.14	22.30	17.62	32.64	40.03	14.83
Col.+Del. attack [27]	TPAMI 2023	4.28	16.12	21.36	3.97	6.28	19.53	25.61	5.21
Our attack	NeurIPS 2024	1.62	7.59	14.46	1.84	1.45	7.71	10.72	1.25

Table 2: Results for attacking cross-modality ReID systems on the RegDB [32] 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 [21]	TPAMI 2022	70.05	86.21	91.55	66.37	70.49	87.21	91.84	65.90
FGSM attack [16]	ICLR 2015	66.79	83.14	88.46	61.05	65.42	81.98	87.20	60.12
PGD attack [17]	ICLR 2018	62.14	80.28	85.10	57.34	63.71	78.82	84.05	58.42
M-FGSM attack [29]	TPAMI 2020	29.34	52.90	61.44	23.35	23.64	40.36	48.61	18.57
LTA attack [28]	CVPR 2022	12.65	25.24	34.02	12.80	10.51	22.93	31.79	9.74
ODFA attack [26]	IJCV 2023	28.57	51.42	60.58	21.84	17.26	33.27	42.92	15.27
Col.+Del. attack [27]	TPAMI 2023	5.12	16.83	22.10	4.94	4.92	14.47	23.04	4.86
Our attack	NeurIPS 2024	2.29	9.06	18.35	3.92	1.93	11.44	19.30	3.46
DDAG baseline [25]	ECCV 2020	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
FGSM attack [16]	ICLR 2015	61.83	80.12	86.47	55.78	60.94	78.35	84.09	56.91
PGD attack [17]	ICLR 2018	64.58	81.39	87.20	58.45	62.17	79.02	85.27	57.69
M-FGSM attack [29]	TPAMI 2020	30.86	54.16	61.98	24.01	25.83	42.12	49.76	19.33
LTA attack [28]	CVPR 2022	11.65	23.20	32.73	11.41	9.76	21.53	29.96	9.23
ODFA attack [26]	IJCV 2023	29.64	52.74	60.74	23.88	24.06	39.75	46.25	18.64
Col.+Del. attack [27]	TPAMI 2023	4.68	13.55	18.57	4.39	4.23	12.75	20.82	4.05
Our attack	NeurIPS 2024	1.33	10.28	19.06	3.79	1.35	9.52	17.52	3.19

From Tab.1, it can be seen that the proposed method reduces the rank-1 accuracy to below 2% in both the 'Visible to Infrared' and 'Infrared to Visible' cases. Similarly, from Tab.2, the rank-1 accuracy drops below 3% in both the 'Visible to Thermal' and 'Thermal to Visible' scenarios. In contrast, traditional metric-based attacks, such as Metric-FGSM (M-FGSM)[29], LTA [28] and ODFA[26], lead to attacked models with significantly higher rank-1 accuracy, whereas traditional classification attacks (such as FGSM [16] and PGD [17]) perform even worse, with rank-1 accuracy remaining over 60%. This is because ReID relies on metric learning for feature matching rather than category classification, requiring attacks specifically tailored for metric learning. These results indicate that, compared to traditional methods that optimize perturbations separately for each modality without considering the inherent correlations between different modalities, our proposed approach demonstrates significant attacking effectiveness across different modalities.

Comparison with State-of-the-Art. Col.+Del., as a universal perturbation method, was fairly compared by first optimizing with one modality's dataset and then fine-tuning with the other modality. Since universal perturbations capture shared patterns across the entire data distribution, Col.+Del. is

Table 3: Results for attacking cross-modality ReID systems on the LLCM [24] dataset. It reports on visible images querying thermal 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
DEEN baseline [21]	CVPR 2023	62.53	90.31	94.73	65.84	54.96	84.92	90.91	62.95
M-FGSM attack [29]	TPAMI 2020	28.48	64.92	75.12	32.88	25.64	61.45	78.31	30.46
LTA attack [28]	CVPR 2022	15.16	56.42	67.53	21.47	19.54	58.25	70.72	24.86
ODFA attack [26]	IJCV 2023	26.34	65.24	76.92	30.85	23.73	62.46	73.57	29.63
Col.+Del. attack [27]	TPAMI 2023	8.61	22.73	36.07	15.72	9.13	20.76	38.02	16.31
Our attack	NeurIPS 2024	5.83	18.14	27.56	12.47	6.42	19.53	28.54	12.23

capable of achieving some level of attack effectiveness in cross-modality scenarios. However, by comparing Tab.1, Tab.2, and Tab.3, we observe that although Col.+Del. performs better than other methods, its effectiveness is still noticeably limited due to the lack of intrinsic correlation mechanisms between modalities. Moreover, as shown in Fig.6, our method outperforms Col.+Del. in transfer attacks across different baselines in cross-modality ReID. The conclusions from these experiments are as follows: 1) In cross-modality attacks, Col.+Del. demonstrates the feasibility of universal perturbations. However, its performance is limited by its failure to account for modality differences and inherent correlations. 2) Our method better bridges the gap between different modalities, more effectively capturing shared features across them.

4.2 Transferability of CMPS

From Fig.6 in supplemental material, the results of the proposed method's transfer attacks on two baseline models, AGW and DDAG, can be observed. For example, on the SYSU dataset, the original attack result of the proposed method on DDAG is mAP=1.84% (refer to Tab. 1). When the perturbation is transferred from AGW to DDAG, the attack result becomes mAP=3.41%. This indicates that the proposed attack method exhibits good generalization across different models, and thus, the attack performance does not degrade significantly. This consistent result is observed on both the RegDB and SYSU datasets. Similarly, in Fig.7 of the supplemental material, we evaluate the cross-dataset transferability of perturbations in comparison with Col.+Del. The results demonstrate a significant advantage of our method. Additionally, we conducted adversarial transferability experiments on IDE [48], PCB [49], and ResNet18 [50]. The rank-1 transfer attack success rates are presented in Tab.4. It can be observed that our method consistently achieves higher transfer attack success rates across all model combinations compared to Col.+Del., indicating that our method demonstrates stronger robustness in generating more universal adversarial perturbations.

Table 4: Comparison of transfer attack success rates between our method and Col.+Del. across models, with higher values indicating better transferability.

Source \Target Model	IDE (Ours/Col.+Del.)	PCB (Ours/Col.+Del.)	ResNet18 (Ours/Col.+Del.)
IDE [48]	98.7% / 94.3%	84.5% / 81.2%	87.4% / 86.1%
PCB [49]	85.1% / 80.4%	97.6% / 92.8%	88.3% / 85.7%
ResNet18 [50]	81.0% / 78.5%	77.5% / 74.9%	98.2% / 95.6%

4.3 Ablation Study

Our method is implemented based on UAP-Retrieval [44]. To validate the effectiveness of the proposed method, we conducted experiments by adding augmentation (Cross-Modality Attack Augmentation) and CMPS to the baseline. Results with AGW baseline model are reported in Tab. 5. The No.1 line represents the UAP-Retrieval algorithm. In the table, 'Aug' indicates the use of the Cross-Modality Attack Augmentation proposed in this paper.

The effectiveness of CMPS. Comparing No.1 with No.3 and No.4, we observe the following: 1) The direct use of UAP-Retrievals yields limited performance. 2) Training with the CMPS strategy

https://doi.org/10.52202/079017-0735

Table 5: Ablation studies on the AGW baseline. 'Aug' denotes the cross-modality attack augmentation method proposed in this paper.

No.	RegDB		SYSU		Aug	CMPS
	mAP	rank-1	mAP	rank-1		
1	6.87	5.53	4.76	5.09	×	×
2	5.11	4.02	3.85	4.37	✓	×
3	3.98	2.17	3.42	3.82	×	✓
4	3.46	1.93	1.23	1.31	✓	✓

proposed in this paper consistently improves the performance of attack results and the universality of learned perturbations.

The effectiveness of augmentation method. Our approach includes cross-modality attack augmentation. Comparing results of No.1, No.2, and No.4 shows its benefits. For example, on the RegDB dataset, augmentation (No.2) reduces mAP from 6.87% to 5.11%, 1.76% lower than without augmentation (No.1). Similarly, with CMPS, mAP drops from 3.98% to 3.46% (No.4), a 0.52% decrease compared to No.3. These findings suggest that using appropriate augmentation enhances cross-modality ReID adversarial attacks' universality. If not specified, our experiments default to using CMPS augmentation. Fig. 5 in the supplementary materials displays the experimental results of our augmentation performed at different probabilities. It can be observed that when the probability value is around 20%, it achieves optimal effectiveness in assisting the attack. If not specified, a probability value of 20% for augmentation is used by default in experiments.

Impact of adversarial boundary size. We conducted an ablation study on different adversarial boundary sizes (ϵ), as shown in the supplementary material 6. In practical applications, ϵ is typically kept moderate to balance perturbation visibility and attack effectiveness. To maintain consistency with previous work [27], we set $\epsilon = 8$ for comparison unless otherwise specified.

5 Conclusion

In this study, we have proposed a cross-modality attack method known as Cross-Modality Perturbation Synergy (CMPS) attack, aimed at evaluating the security of cross-modality ReID systems. The core idea behind the CMPS attack is to capture shared knowledge between visible and non-visible images to optimize perturbations. Additionally, we proposed a Cross-Modality Attack Augmentation method, utilizing grayscale images to bridge the gap between different modalities, further enhancing the attack performance. Through experiments conducted on the RegDB, SYSU and LLCM datasets, we demonstrated the effectiveness of the proposed method while also revealing the limitations of traditional attack approaches. The primary objective of this study has been to assess the security of cross-modality ReID systems. In future research, on the one hand, we will continue to improve the transferability of cross-modality attacks across different datasets and models; on the other hand, we plan to develop robust ReID methods specifically tailored for cross-modality attacks, aimed at defending against adversarial samples. This study not only contributes to advancing the understanding of the security of cross-modality ReID systems but also provides strong motivation for ensuring the reliability and security of these systems in real-world applications.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant No.62276222 and the Public Technology Service Platform Project of Xiamen City, Grant No.3502Z20231043.

References

[1] Bin Yang, Jun Chen, and Mang Ye. Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16870–16879, 2024.

[2] Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. Ufinebench: Towards text-based person retrieval with ultra-fine granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22010–22019, 2024.

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

[4] 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.

[5] Yiwei Ma, Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Towards local visual modeling for image captioning. *Pattern Recognition*, 138:109420, 2023.

[6] Hongwei Niu, Jie Hu, Jianghang Lin, and Shengchuan Zhang. Eov-seg: Efficient open-vocabulary panoptic segmentation. *arXiv preprint arXiv:2412.08628*, 2024.

[7] Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5940–5948, 2024.

[8] Yansong Qu, Shaohui Dai, Xinyang Li, Jianghang Lin, Liujuan Cao, Shengchuan Zhang, and Rongrong Ji. Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5328–5337, 2024.

[9] Hongwei Niu, Linhuang Xie, Jianghang Lin, and Shengchuan Zhang. Exploring semantic consistency and style diversity for domain generalized semantic segmentation. *arXiv preprint arXiv:2412.12050*, 2024.

[10] 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.

[11] Yunpeng Gong, Yongjie Hou, Chuangliang Zhang, and Min Jiang. Beyond augmentation: Empowering model robustness under extreme capture environments. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

[12] Yunpeng Gong, Liqing Huang, and Lifei Chen. Eliminate deviation with deviation for data augmentation and a general multi-modal data learning method. *arXiv preprint arXiv:2101.08533*, 2021.

[13] Lei Tan, Pingyang Dai, Rongrong Ji, and Yongjian Wu. Dynamic prototype mask for occluded person re-identification. In *Proceedings of the 30th ACM international conference on multimedia*, pages 531–540, 2022.

[14] Lei Tan, Jiaer Xia, Wenfeng Liu, Pingyang Dai, Yongjian Wu, and Liujuan Cao. Occluded person re-identification via saliency-guided patch transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5070–5078, 2024.

[15] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[18] Yukang Zhang, Yan Yan, Jie Li, and Hanzi Wang. Mrcn: A novel modality restitution and compensation network for visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3498–3506, 2023.

 https://doi.org/10.52202/079017-0735

[19] Jiangming Shi, Xiangbo Yin, Zhizhong Zhang, Yachao and Zhang, Yuan Xie, and Yanyun Qu. Learning commonality, divergence and variety for unsupervised visible-infrared person re-identification. *arXiv preprint arXiv:2402.19026v2* , 2024.

[20] Bin Yang, Jun Chen, and Mang Ye. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11069–11079, October 2023.

[21] 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.

[22] Xiangbo Yin, Jiangming Shi, Yachao Zhang, Yang Lu, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Robust pseudo-label learning with neighbor relation for unsupervised visible-infrared person re-identification. *arXiv preprint arXiv:2405.05613*, 2024.

[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] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2153–2162, 2023.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] Wenqi Wang, Run Wang, Lina Wang, Zhibo Wang, and Aoshuang Ye. Towards a robust deep neural network against adversarial texts: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35:3159–3179, 2023.

[35] Qian Wang, Baolin Zheng, Qi Li, Chao Shen, and Zhongjie Ba. Towards query-efficient adversarial attacks against automatic speech recognition systems. *IEEE Transactions on Information Forensics and Security*, 16:896–908, 2021.

[36] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022.

[37] Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Xiaoshuai Sun, and Rongrong Ji. Controlmllm: Training-free visual prompt learning for multimodal large language models. *arXiv preprint arXiv:2407.21534*, 2024.

[38] 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.

[39] 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.

[40] Yunpeng Gong, Yongjie Hou, Zhenzhong Wang, Zexin Lin, and Min Jiang. Adversarial learning for neural pde solvers with sparse data. *arXiv preprint arXiv:2409.02431*, 2024.

[41] 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.

[42] 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.

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

[44] 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.

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

[46] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 618–626, 2019.

[47] 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.

[48] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[49] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018.

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.

 https://doi.org/10.52202/079017-0735

Supplemental Material

Contents

https://doi.org/10.52202/079017-0735

6 Supplemental Experiments

Our experiments were conducted using three RTX 2080 Ti GPUs, each with 11GB of memory.

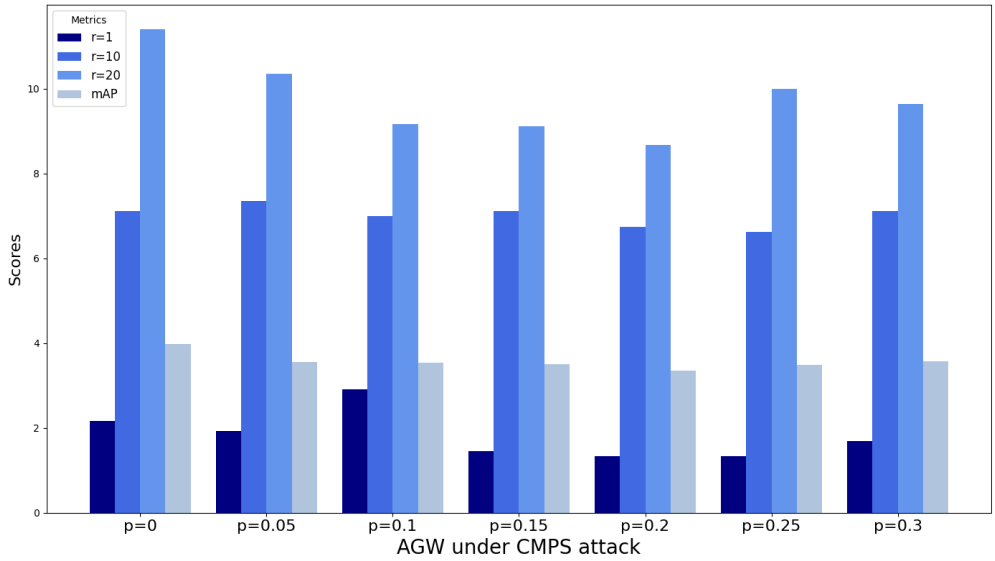

Figure 5: The impact of different grayscale transformation probabilities on attack performance. Lower evaluation metrics indicate higher attack success rates. The experimental results are derived from experiments on the RegDB dataset using AGW as the baseline model for testing.

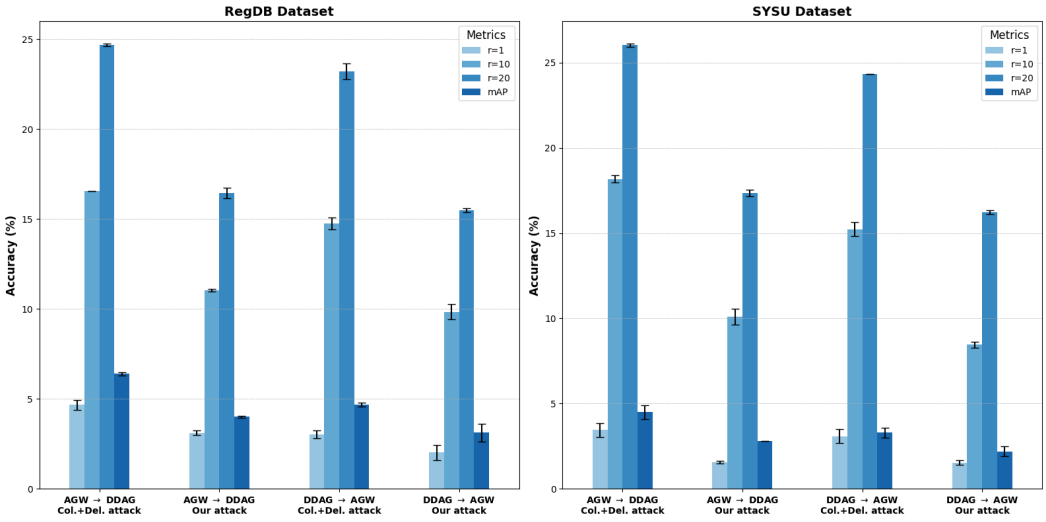

Figure 6: Transferability experiments of the proposed method across different models on the RegDB dataset (visible to thermal). Transferability experiments of the proposed method across different models on the SYSU dataset (visible to Infrared).

https://doi.org/10.52202/079017-0735

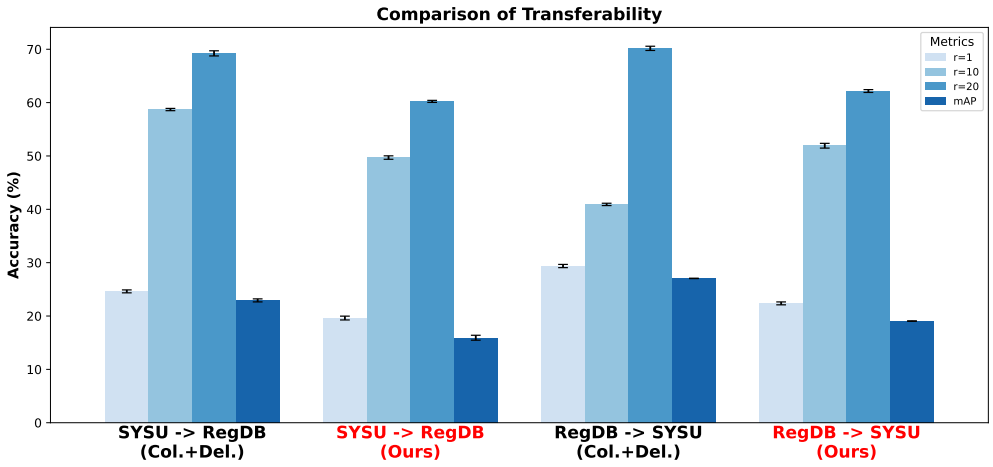

Figure 7: Comparison of Transferability Between Different Methods on Two Cross-Modal Datasets SYSU and RegDB.

Table 6: Using the AGW baseline on the RegDB dataset, we conduct an ablation study to evaluate the impact of the adversarial boundary ϵ on the effectiveness of the proposed CMPS attack (rank-1 accuracy).

Adversarial Boundary (ϵ)	**Visible to Thermal**	**Thermal to Visible**
-	70.0%	70.5%
2	32.7%	40.5%
4	9.6%	13.8%
8	2.3%	2.0%
16	0.3%	0.5%

7 Proof of Method Superioritys

We design a cross-modality triplet loss to simultaneously optimize two modalities, which effectively captures common features between different modalities and enhances the cross-modality adaptability of universal perturbations.

7.1 Definition of Cross-Modality Triplet Loss

The cross-modality triplet loss aims to optimize the model by adjusting the distance relationships among triplet samples (anchor, positive, negative) so that samples of the same identity are closer, while samples of different identities are farther apart. Specifically, given samples $(x_{\mathrm{A}}, x_{\mathrm{P}}, x_{\mathrm{N}})$, where:

- x_{A} is the anchor sample,
- x_{P} is the positive sample with the same identity as the anchor (from a different modality),
- x_{N} is the negative sample with a different identity from the anchor.

The triplet loss function is defined as:

$$L_{\text{triplet}} = \max\left(0, D(f(x_{\mathrm{A}}), f(x_{\mathrm{P}})) - D(f(x_{\mathrm{A}}), f(x_{\mathrm{N}})) + \alpha\right) \tag{10}$$

where $D(\cdot, \cdot)$ denotes the distance metric (e.g., Euclidean distance), and α is a margin hyperparameter.

Mathematically, given the cross-modality triplet loss:

$$L_{\text{triplet}} = \max\left(\left(\|C_g^n - f_{RGB}^{adv}\|_2 - \|C_{ir}^p - f_{RGB}^{adv}\|_2 + \rho\right), 0\right) \tag{11}$$

We can view it as part of the sum of the loss functions for two modalities:

$$\mathcal{L}_A(\eta) = \|C_g^n - f_{RGB}^{adv}\|_2 \tag{12}$$

$$\mathcal{L}_B(\eta) = \|C_{ir}^p - f_{RGB}^{adv}\|_2 \tag{13}$$

Thus, the overall optimization objective can be expressed as:

$$\eta_{\text{agg}}^* = \arg\min_{\eta} \left(\mathcal{L}_A(\eta) + \mathcal{L}_B(\eta) + \rho\right) \tag{14}$$

This form effectively aggregates the losses of different modalities, thereby optimizing the loss functions of different modalities simultaneously, achieving joint optimization of cross-modality data. This approach trains universal perturbations with better generalization capabilities than methods that consider only single-modality information.

7.2 Proof of Aggregated Optimization Superiority

Assume we have data from two modalities: modality A and modality B. Let $\mathcal{L}_A(\eta)$ and $\mathcal{L}_B(\eta)$ be the loss functions on modality A and modality B, respectively. The objective of single-modality training is:

$$\min_{\eta} \mathcal{L}_A(\eta) + \mathcal{L}_B(\eta) \tag{15}$$

The stepwise optimization method first optimizes $\mathcal{L}_A(\eta)$ and then optimizes $\mathcal{L}_B(\eta)$:

$$\eta^* = \arg\min_{\eta} \mathcal{L}_A(\eta) \rightarrow \eta^{**} = \arg\min_{\eta} \mathcal{L}_B(\eta^*) \tag{16}$$

The aggregated optimization of the two loss functions is:

$$\eta_{\text{agg}}^* = \arg\min_{\eta} \left(\mathcal{L}_A(\eta) + \mathcal{L}_B(\eta)\right) \tag{17}$$

Using the gradient aggregation method, it can be expressed as:

$$\nabla_\eta \mathcal{L}_{\text{agg}} = \nabla_\eta \left(\mathcal{L}_A(\eta) + \mathcal{L}_B(\eta)\right) \tag{18}$$

Next, we consider the different optimization paths of the two methods.

https://doi.org/10.52202/079017-0735

7.2.1 Stepwise Optimization Method

The stepwise optimization method first optimizes the loss function of modality A and then the loss function of modality B. Assume the update rule at iteration k is:

$$\eta^{(k+1)} = \eta^{(k)} - \alpha \nabla_\eta \mathcal{L}_A(\eta^{(k)}) \tag{19}$$

After optimizing the loss function of modality A, the loss function of modality B is optimized:

$$\eta^{(k+1)} = \eta^{(k)} - \alpha \nabla_\eta \mathcal{L}_B(\eta^{(k)}) \tag{20}$$

Since the two optimization processes are separate, this may result in η being optimal for modality A but not necessarily for modality B.

7.2.2 Aggregated Optimization Method

The aggregated optimization method considers the losses of both modalities in each iteration. Assume the update rule at iteration k is:

$$\eta^{(k+1)} = \eta^{(k)} - \alpha \left(\nabla_\eta \mathcal{L}_A(\eta^{(k)}) + \nabla_\eta \mathcal{L}_B(\eta^{(k)}) \right) \tag{21}$$

In this way, each update considers the losses of both modalities, ensuring that η approaches the optimal solution for both modalities.

To further prove that the aggregated optimization method can find a better perturbation η, we can analyze the existence and uniqueness of the optimal solution.

Assume $\mathcal{L}_A(\eta)$ and $\mathcal{L}_B(\eta)$ are continuously differentiable and convex loss functions. According to convex optimization theory, the optimal solutions of the loss functions exist and are unique.

The optimal solution of the stepwise optimization method is:

$$\eta^*_{\text{step}} = \arg\min_\eta \left(\mathcal{L}_A(\eta) + \mathcal{L}_B(\eta^*) \right) \tag{22}$$

where η^* is the optimal solution of $\mathcal{L}_A(\eta)$.

The optimal solution of the aggregated optimization method is:

$$\eta^*_{\text{agg}} = \arg\min_\eta \left(\mathcal{L}_A(\eta) + \mathcal{L}_B(\eta) \right) \tag{23}$$

Since η^*_{step} is not necessarily globally optimal for modality B, and η^*_{agg} is the global optimal solution considering both modalities, we can derive:

$$\mathcal{L}_A(\eta^*_{\text{agg}}) + \mathcal{L}_B(\eta^*_{\text{agg}}) \leq \mathcal{L}_A(\eta^*_{\text{step}}) + \mathcal{L}_B(\eta^*_{\text{step}}) \tag{24}$$

7.3 Generalization Error Analysis

Generalization error measures the model's performance on unseen data. We can further prove the superiority of aggregated training through generalization error analysis.

Let $\mathcal{L}_{\text{train}}$ and $\mathcal{L}_{\text{test}}$ be the losses on the training and test sets, respectively. The generalization error is defined as:

$$\mathcal{E}_{\text{gen}} = \mathcal{L}_{\text{test}}(\eta) - \mathcal{L}_{\text{train}}(\eta) \tag{25}$$

The upper bound of the generalization error can be expressed using measures such as Rademacher complexity or VC dimension. For machine learning models, the lower the model complexity, the smaller the generalization error. Simultaneously optimizing the losses for multiple tasks (modalities) can reduce overfitting to a single task (modality), as the model needs to perform well on multiple tasks (modalities) simultaneously. This effectively introduces an implicit regularization effect, reducing the model complexity. Therefore, compared to the stepwise optimization method, the aggregated optimization method can effectively reduce the complexity of the perturbation model. The lower the model complexity, the smaller the generalization error.

The Rademacher complexity measures the complexity of a class of models on a given sample set. For a function h in the hypothesis space $\mathcal{H}$, the empirical Rademacher complexity on n samples is defined as:

$$\hat{\mathcal{R}}_n(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right] \tag{26}$$

where σ_i are Rademacher random variables, taking values ± 1 with equal probability.

The impact of modality aggregation on complexity:

Assume $\mathcal{H}_{\mathrm{A}}$ and $\mathcal{H}_{\mathrm{B}}$ are the hypothesis spaces of modality A and modality B, respectively. The stepwise optimization method first optimizes $\mathcal{H}_{\mathrm{A}}$ and then $\mathcal{H}_{\mathrm{B}}$. Its empirical Rademacher complexity can be expressed as:

$$\hat{\mathcal{R}}_n(\mathcal{H}_{\text{step}}) = \hat{\mathcal{R}}_n(\mathcal{H}_{\mathrm{A}}) + \hat{\mathcal{R}}_n(\mathcal{H}_{\mathrm{B}}) \tag{27}$$

The aggregated optimization method optimizes $\mathcal{H}_{\mathrm{A}} \cup \mathcal{H}_{\mathrm{B}}$ simultaneously. Its empirical Rademacher complexity is:

$$\hat{\mathcal{R}}_n(\mathcal{H}_{\text{agg}}) = \hat{\mathcal{R}}_n(\mathcal{H}_{\mathrm{A}} \cup \mathcal{H}_{\mathrm{B}}) \tag{28}$$

According to the properties of Rademacher complexity, the complexity of $\mathcal{H}_{\mathrm{A}} \cup \mathcal{H}_{\mathrm{B}}$ is usually less than or equal to the sum of the complexities of $\mathcal{H}_{\mathrm{A}}$ and $\mathcal{H}_{\mathrm{B}}$:

$$\hat{\mathcal{R}}_n(\mathcal{H}_{\text{agg}}) \leq \hat{\mathcal{R}}_n(\mathcal{H}_{\mathrm{A}}) + \hat{\mathcal{R}}_n(\mathcal{H}_{\mathrm{B}}) \tag{29}$$

Generalization error upper bound derivation:

Using Rademacher complexity, we can derive the upper bound of the generalization error. For the loss function $\mathcal{L}$ and hypothesis space $\mathcal{H}$, the upper bound of the generalization error is:

$$\mathcal{E}_{\text{gen}} \leq 2\hat{\mathcal{R}}_n(\mathcal{L} \circ \mathcal{H}) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \tag{30}$$

where $\mathcal{L} \circ \mathcal{H}$ denotes the composition of the loss function with the hypothesis space.

The upper bound of the generalization error for the stepwise optimization method is:

$$\mathcal{E}_{\text{gen, step}} \leq 2\left(\hat{\mathcal{R}}_n(\mathcal{L} \circ \mathcal{H}_{\mathrm{A}}) + \hat{\mathcal{R}}_n(\mathcal{L} \circ \mathcal{H}_{\mathrm{B}})\right) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \tag{31}$$

The upper bound of the generalization error for the aggregated optimization method is:

$$\mathcal{E}_{\text{gen, agg}} \leq 2\hat{\mathcal{R}}_n(\mathcal{L} \circ (\mathcal{H}_{\mathrm{A}} \cup \mathcal{H}_{\mathrm{B}})) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \tag{32}$$

Since

$$\hat{\mathcal{R}}_n(\mathcal{L} \circ (\mathcal{H}_{\mathrm{A}} \cup \mathcal{H}_{\mathrm{B}})) \leq \hat{\mathcal{R}}_n(\mathcal{L} \circ \mathcal{H}_{\mathrm{A}}) + \hat{\mathcal{R}}_n(\mathcal{L} \circ \mathcal{H}_{\mathrm{B}}) \tag{33}$$

Therefore:

$$\mathcal{E}_{\text{gen, agg}} \leq \mathcal{E}_{\text{gen, step}} \tag{34}$$

This indicates that the aggregated optimization method has a lower upper bound on the generalization error compared to the stepwise optimization method.

https://doi.org/10.52202/079017-0735

8 Discussion

8.1 Ethical Considerations

In this study, we introduce a novel cross-modal adversarial attack method known as Cross-Modality Perturbation Synergy (CMPS). This research offers a new perspective on understanding and enhancing the security of cross-modal ReID systems by leveraging shared features across different modalities to optimize perturbations. However, this approach also raises a series of ethical and safety concerns regarding the potential negative impacts of adversarial attack techniques. The CMPS method, like other adversarial technologies, 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.

8.2 Limitations and Future Work

Here, we need to acknowledge the limitations of the proposed method and identify potential directions for future research. Firstly, current attack techniques primarily focus on gradient-based perturbation optimization for given datasets. However, in real-world scenarios, the modalities encountered are often unknown and not limited to RGB, infrared, and thermal imaging. Moreover, effectively transferring perturbations to different and unknown modalities presents a significant research challenge.

When dealing with various models and modalities, gradient-based methods face several challenges. Firstly, these methods are prone to "catastrophic forgetting," where learning new information can lead to the loss of previously learned knowledge, affecting the effectiveness of perturbations. Secondly, the inconsistency of gradient information across multiple models and modalities can negatively impact the stability and generalizability of the method. Therefore, future research should explore more robust algorithms that can effectively operate in complex environments involving multiple modalities and models, thereby enhancing the applicability and transferability of attacks.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reffect the paper's contributions and scope. We have clearly stated our novel methodology and its implications in the abstract and introduction, and these are further elaborated upon and validated in the main body of the paper

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to the final section of the supplementary materials.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We have conducted a theoretical analysis of the effectiveness of the proposed method.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.

https://doi.org/10.52202/079017-0735

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: To ensure full disclosure of all necessary information to reproduce the main experimental results of the paper, we have provided the experimental setup within the paper and included pseudocode in the supplementary materials. Additionally, as a key contribution, we have conducted a comprehensive theoretical analysis of the proposed method. We will provide the source code for the reviewers' examination.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification:Given the rapid pace of technological advancement, our field requires careful dissemination of our methods to ensure the integrity and competitiveness of our ongoing research. Additionally, due to ethical and security considerations, we currently prefer not to publicly release our code. However, we will provide the code for reviewers' examination and release the source code when the time is right.

   Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have mentioned this in our paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification:In our main performance comparison experiments, the results are reported as the mean ± standard deviation.

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

https://doi.org/10.52202/079017-0735

Answer: [Yes]

Justification: We provide the GPU type used in our paper, and the computation time is given in the ablation experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Our research adheres to the NeurIPS Code of Ethics in all respects.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Our research may have potential negative social impacts. One possible solution is to enhance model security by improving defenses against the proposed attacks through adversarial training.Our research may have potential negative social impacts. One possible solution is to enhance model security by improving defenses against the proposed attacks through adversarial training.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Our intention in proposing adversarial attack techniques is to study model security. The safeguard involves enhancing model security by improving defenses against the proposed attacks through adversarial training.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We used open datasets and correctly referenced the papers.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Given the rapid pace of technological development, our field requires careful dissemination of our methods to ensure the integrity and competitiveness of our ongoing research. Therefore, we do not currently plan to publicly release our code.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

https://doi.org/10.52202/079017-0735

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.