Invisible Backdoor Attack with Dynamic Triggers against Person Re-IDentification

Wenli Sun, Xinyang Jiang, Shuguang Dou, Dongsheng Li, Duoqian Miao
Cheng Deng, *Senior Member, IEEE*, Cairong Zhao

***Abstract*—In recent years, person Re-IDentification (ReID) has rapidly progressed with wide real-world applications but is also susceptible to various forms of attack, including proven vulnerability to adversarial attacks. In this paper, we focus on the backdoor attack on deep ReID models. Existing backdoor attack methods follow an all-to-one or all-to-all attack scenario, where all the target classes in the test set have already been seen in the training set. However, ReID is a much more complex fine-grained open-set recognition problem, where the identities in the test set are not contained in the training set. Thus, previous backdoor attack methods for classification are not applicable to ReID. To ameliorate this issue, we propose a novel backdoor attack on deep ReID under a new all-to-unknown scenario, called Dynamic Triggers Invisible Backdoor Attack (DT-IBA). Instead of learning fixed triggers for the target classes from the training set, DT-IBA can dynamically generate new triggers for any unknown identities. Specifically, an identity hashing network is proposed to first extract target identity information from a reference image, which is then injected into the benign images by image steganography. We extensively validate the effectiveness and stealthiness of the proposed attack on benchmark datasets and evaluate the effectiveness of several defense methods against our attack.**

***Index Terms*—Backdoor Attacks, Targeted Attack, Person Re-IDentification, All-to-unknown, Stealthiness**

I. Introduction

RECENTLY, deep learning has progressed rapidly and has been widely utilized in a variety of image classification and recognition tasks. The success of deep learning models is highly dependent on the scale of datasets [1]. However, datasets for deep model training are time and money-intensive to construct, resulting in a large portion of the algorithm developers opting for third-party datasets, which brings huge security risks of backdoor attacks [2]–[6].

A preprint version of this research work was put on arXiv (i.e. arXiv:2211.10933).

This work was supported by the National Natural Science Fund of China (62076184, 61976158, 61976160, 62076182, 62276190), in part by Fundamental Research Funds for the Central Universities and State Key Laboratory of Integrated Services Networks (Xidian University), in part by Shanghai Innovation Action Project of Science and Technology (20511100700) and Shanghai Natural Science Foundation (22ZR1466700). (Corresponding author: Cairong Zhao.)

Wenli Sun, Shuguang Dou, Duoqian Miao, and Cairong Zhao are with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China (e-mail: 2233055@tongji.edu.cn; dousg@tongji.edu.cn; dqmiao@tongji.edu.cn; zhaocairong@tongji.edu.cn).

Xinyang Jiang and Dongsheng Li are with Microsoft Research Asia, Shanghai 200232, China (e-mail: xinyangjiang@microsoft.com; dongshengli@fudan.edu.cn).

Cheng Deng is with the School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: chdeng.xd@gmail.com).

Backdoor attacks occur when samples poisoned with backdoor triggers have their labels changed to target labels and added to the training set, causing the model to mis-classify the target labels during the inference stage. As shown in Fig. 1, current backdoor attack scenarios for classification tasks can be classified into two categories: all-to-one and all-to-all [2], [4], [7]–[9]. In the all-to-one scenario, a single target label is pre-defined and the poisoned images from any category will be classified as this fixed label. On the other hand, for the all-to-all scenario, the target label can be any of the classes that appear in the training set, and images can be manipulated to be classified into any chosen classes. For both scenarios, the target labels usually need to be pre-determined before poisoning the training set, and the backdoor triggers are kept fixed during the attack.

In this paper, we shift our focus from classification to another important vision task: Person Re-IDentification (ReID) [10]–[13]. ReID is a task of matching person images from several camera viewpoints. It has wide applications in surveillance, tracking, smart retail, etc., but also faces the threat of backdoor attacks. While classification models based on deep learning have been proven to be vulnerable to backdoor attacks [2], [14]–[18], the backdoor attack risk on ReID has not been thoroughly studied. Existing backdoor attacks on classification cannot be used directly for recognition tasks due to the following challenges. Firstly, ReID is a more challenging fine-grained recognition task with significantly more classes (i.e. person identities) compared to conventional classification. This means that there are only a dozen images per class, and traditional attacks significantly alter the data distribution, which is very detrimental to stealthiness. More importantly, conventional backdoor attack methods generate fixed triggers corresponding to the pre-determined target labels in the training set. However, ReID is an open-set recognition problem where training and test sets have distinct identities. In the inference stage, the target ID does not exist in the training set, so a new trigger needs to be generated dynamically according to the target ID. It is noteworthy that the current attacks use the same setting to deal with face recognition and classification tasks, which does not match the open-set recognition scenario [19], [20].

To tackle the challenges in backdoor attacks on ReID models, we propose a novel all-to-unknown attack scenario, where the target class can be any of the classes in the test set, even if it does not appear in the training set. In contrast to the conventional attack scenario, new backdoor triggers can be dynamically produced for any novel target identity outside the

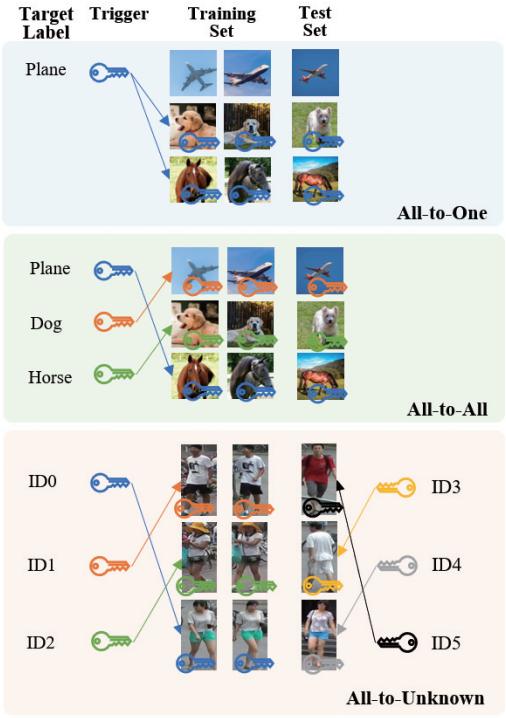


Fig. 1. Backdoor attack scenarios for image classification and person ReID. All-to-one and all-to-all are attack scenarios in image classification tasks that require the target labels to be consistent in the training and testing stages. Our proposed all-to-unknown is applicable to the scenario where the target class does not appear in the training set.

training set, as shown in Fig. 1. By addressing the limitations of traditional attack settings that rely on predefined target labels, the all-to-unknown setting offers a versatile and robust framework for conducting backdoor attacks in a real-world scenario of ReID tasks, contributing to a more comprehensive understanding of potential security risks and defenses in open-set recognition applications like face recognition. Specifically, to realize this scenario, given a benign image, the goal is to generate a backdoor trigger containing an unknown target identity specified by a reference image. As a result, the attacked ReID model will recognize the poisoned image as the identity of the reference image.

To dynamically generate the backdoor trigger, we propose an identity hashing network to first encode the target identity information in the reference image as an embedding in hamming space and then inject the embedding into the benign images in the form of pixel perturbation by image steganography. As a result, two images of different identities with the same invisible trigger will be recognized as the same person, and vice versa. While conventional backdoor attack methods only achieve non-targeted attacks by degrading the backdoored model's performance, our approach allows for targeted attacks, in which the target person appears within the top-10 retrieval results. In summary, our contributions are three-fold as follows:

- We raise a new and rarely studied backdoor attack risk on the ReID task, which is quantified by our proposed new all-to-unknown attack scenario.
- The new all-to-unknown attack scenario and a novel corresponding method are proposed to realize adversary mismatch and target person impersonation by dynamic triggers, which are able to dynamically alter the poisoned image's identity to any target identity outside the training set.
- The experiments show the effectiveness of the proposed backdoor attack method and are robust to several existing representative backdoor defense strategies.

II. Related Work

In this section, we review current backdoor attacks suggested for image classification tasks, backdoor defense techniques, and ReID models.

A. Backdoor Attack

The backdoor attack is a severe concern to DNNs since they cause the poisoned model to function normally on clean samples but classify samples with triggers as target class [9], [21], [22]. The majority of existing backdoor attacks are based on the assumption that the target labels are known and fixed, which is only appropriate for classification tasks. Specifically, the adversary needs to design a trigger pattern t, and select a target label y_t. The adversary adds triggers to benign samples $x \in D$ to generate the poisoned samples x^p, then changes the label of x^p to the target label y_t and puts(x^p, y_t) in the training set D_{train}. During the training stage, the backdoored model will associate triggers with target labels, resulting in the classification of any sample containing a trigger as y_t during the inference stage.

Currently, backdoor attack on classification models is a well-established research field [23]–[29]. Recently, few works have shifted their focus to models other than image classification. Zhai *et al.* designed a clustering-based attack methodology for speaker verification in which poisoned samples from different clusters include different triggers [24]. All triggers are concatenated with the test samples during the inference stage to achieve the adversary aim. However, because this approach registers only one individual at a time, it is incompatible with recognition tasks requiring a retrieval range greater than one, such as person Re-IDentification. Zhao et al. proposed a backdoor attack strategy for video recognition models in which the adversary doesn't have to change the label, but is aware of the target class during the training process [26]. Few-shot backdoor attack (FSBA) is not a label-targeted attack; instead, it embeds triggers in the feature space, causing the model to lose track of a specific object [28]. Some work has attempted backdoor attacks on face recognition, but the setting of the attacks follows the classification task and is not realistic enough. Latent Backdoors can embed hidden malicious rules within a single teacher model and all student models through the process of transfer learning [19]. It demonstrates the effectiveness of latent backdoor attacks in face recognition. Moreover, Physical objects are used as triggers to study the feasibility of physical backdoor attacks using face recognition as an illustration [20]. To the best of our knowledge, none of the existing backdoor attack methods are applicable to person ReID.

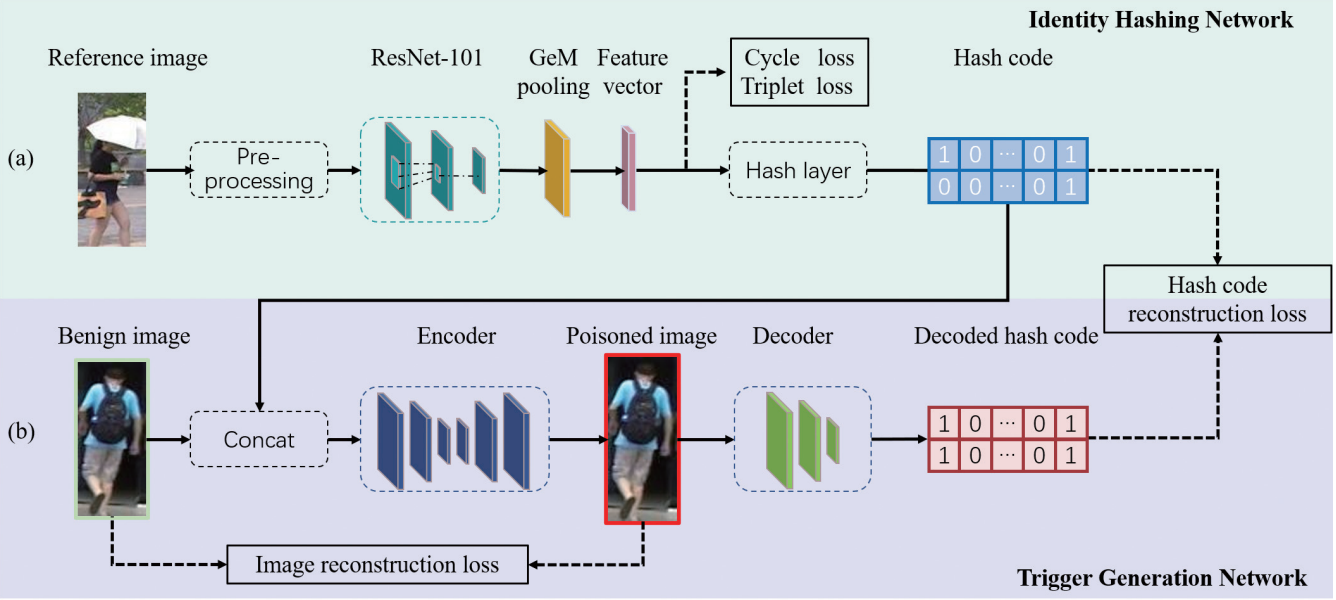


Fig. 2. The training pipeline of our poisoned image generator. (a) Identity hashing network. After image pre-processing, the image feature of the reference image is extracted with ResNet-101. Then GeM pooling layer aggregates feature maps generated by ResNet-101. After that, the high-dimensional features are further compressed by the hash layer to obtain a 128-dimensional hash code. (b) Trigger generation network. The benign samples are first connected with the hash code of the reference image. Then the deep image steganography network with an encoder-decoder structure is used to generate the trigger. The training stage needs to minimize the perceptual difference between the poisoned and benign images in order to achieve invisibility. The decoder is trained to reconstruct the hash code.

B. Backdoor Defense

Currently, backdoor defense strategies can be classified into the following three categories [4].

Sample filtering-based empirical defenses are designed to distinguish between clean and poisoned samples and to use only benign samples for training and testing [30]–[33]. Zeng *et al.* revealed the presence of high-frequency artifacts in poisoned images compared to natural images. This approach uses data augmentation to process the training set in order to simulate potential backdoor attack patterns and train a supervised model that distinguishes between poisoned and clean data [33].

Poison suppression-based defenses can be used to eliminate the influence of backdoor triggers during training to prevent backdoor generation [34], [35]. Hong *et al.* used differentially private stochastic gradient descent (DP-SGD) to clip and perturb individual gradients during the training stage, ensuring that the generated model contains no hidden backdoors [34].

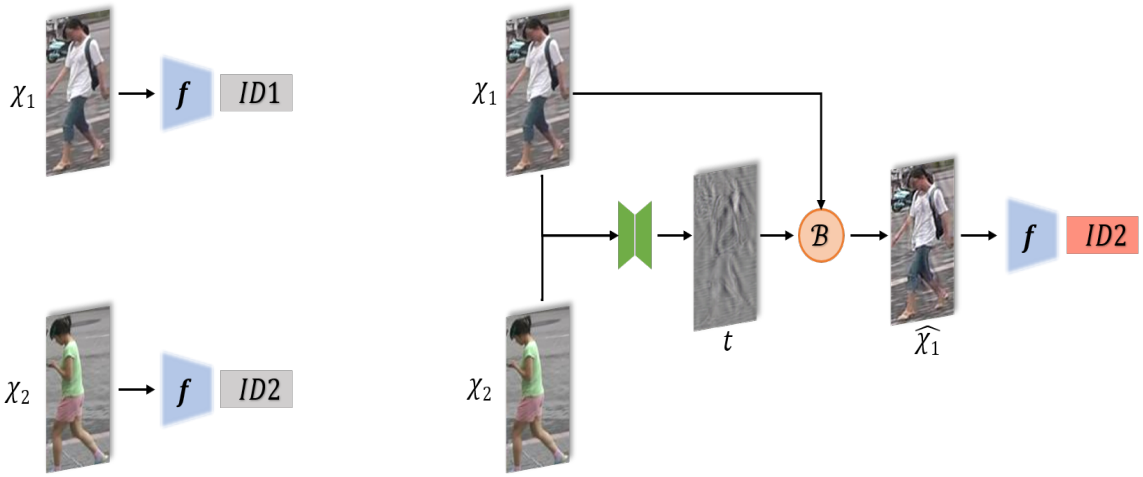
Model reconstruction-based defenses are aimed at erasing the infected model's hidden backdoor. As a result, even if the trigger remains in the poisoned samples, the prediction stays benign [36]–[39]. Based on the observation that they are usually dormant in benign samples, pruning backdoor-related neurons was proposed to remove the hidden backdoor [36].

C. Deep Person Re-IDentification Models

The goal of ReID is to determine whether a query person has appeared in another place, and was captured by different cameras at a distinct time [40]. The person ReID models extract robust feature representations of pedestrians through representation learning methods. Then, the similarity score between pedestrians is calculated by the metric learning method and ranked from the highest to lowest, and the target pedestrian is re-identified according to the ranking result [41]–[46]. There has been some research on attack methods against ReID, except for backdoor attacks. The vulnerability of the current ReID model when exposed to the Universal Adversarial Perturbation (UAP) has been validated for the first time in [47]. This perturbation is applicable to both image-agnostic and model-insensitive person ReID attacks. Moreover, the extreme vulnerability of existing distance metrics to adversarial examples is revealed. These examples are generated by simply adding human-imperceptible perturbations to person images, as indicated in [48]. An attack algorithm for generating adversarial patterns is proposed, to realize adversary mismatch and target person impersonation, respectively [49]. It employs a universal adversarial perturbation to deceive re-ID models in unseen domains and introduces a meta-learning approach that derives the universal perturbation through gradient interactions between meta-training and meta-testing datasets [50]. And there are defenses against adversarial attacks towards Person ReID, e.g., [51] presents an adaptable combinatorial adversarial attack suitable for unseen domains and models, involving pixel and color space distortions and employs a virtual dataset within the meta-learning framework.

III. The Proposed Attack

In this section, we introduce our proposed backdoor attack against person ReID models under the all-to-unknown scenario. We begin with an introduction to our threat model and an overview of the attack pipeline, followed by an analysis of how to generate and apply the proposed dynamic trigger for the attack.



(a) Clean mode (b) Target-aware attack mode

Fig. 3. Two running modes of the proposed target-aware backdoor system. X_1 and X_2 are benign images of different IDs. In clean mode, the backdoored model can correctly predict the identity. But in target-aware attack mode, X_1 is injected with a trigger related to the target person X_2, and then the poisoned image $\hat{X}_1$ is generated. The trigger activates the backdoor in the model, so $\hat{X}_1$ is predicted as the target ID.

A. Threat Model

In this paper, we follow a more realistic setting, where the adversary gets clean datasets without access to model structure or training loss. During the inference stage, the adversary is only permitted to query the trained model with any image. It has no knowledge of the model and cannot manipulate the inference procedure. The poisoned dataset is created by dynamically adding imperceptible triggers to some of the training images. The aim of the attack is to induce a backdoor behavior in the person ReID network trained by the user so that it performs normally on a clean test set, but will recognize any person image with a trigger as the target person, regardless of its ground truth. Therefore this attack is not easily detectable and can escape standard validation tests when users download datasets or models published by the adversary. The proposed backdoor attack with dynamically generated triggers is able to evade the person search of the deep ReID models, causing a serious security risk.

B. Preliminary

There are two types of attack scenarios for classification tasks according to the number of target classes [8], [52]. 1) All-to-one is a single-target backdoor attack scenario, where the adversary selects a fixed label c as the output label. Eq. 1 shows backdoor generation function T can make x_i misclassified to the target label c.

$$f^{'}(T(x_i)) = c, \quad \forall (x_i, y_i) \in \mathcal{D}, \tag{1}$$

2) All-to-all is a multi-target attack scenario where the target label $\hat{c}$ is the next label of the true label. It is formulated as follows, in which C means the number of classes.

$$f^{'}(T(x_i)) = \hat{c}, \quad \hat{c} = y_{(i+1) mod |C|}, \quad \forall (x_i, y_i) \in \mathcal{D}, \tag{2}$$

Since the target IDs of the person ReID training and test sets are different (*i.e.* $c_{train} \cap c_{test} = \emptyset$), these two conventional attack scenarios are not suitable for ReID. As a result, we propose a new scenario specifically for ReID, called all-to-unknown.

Let $D_{train} = \{(x_i, y_i)\}_{i=1}^{N_{tr}}$ represents the benign training set containing N_{tr} images, where $y_i \in \{0, 1, ..., M_{tr} - 1\}$ is the M_{tr} person IDs, and $D_{test} = \{(x_i, y_i)\}_{i=N_{tr}}^{N_{tr}+N_{te}}$ indicates the benign test set containing M_{te} person IDs denoted as $y_i \in \{M_{tr}, M_{tr}+1, ..., M_{tr}+M_{te}-1\}$.

The test set D_{test} is divided into two parts: query set and gallery set. For each probe image in the query set, the ReID model f will find K images in the gallery with the highest similarity to it in the inference stage.

Formally, in the all-to-unknown scenario, given a query image (x_i, y_i), and a set of gallery images (x_t, y_t), a backdoor generator $T(x_i, \hat{y})$ is built to obtain a poisoned image with any target ID $\hat{y}$. Given a query image, the backdoor attack aims at ranking gallery images that originally belonged to the same identity out of the top-K list, while ranking the image pair with different identities but poisoned with the same trigger within the top-K list:

$$\begin{cases} Rank\left(f^{'}(T(x_i, \hat{y}), x_t)\right) > K, \quad \hat{y} \neq y_i, & \text{if } y_i = y_t \\ Rank\left(f^{'}(T(x_i, \hat{y}), T(x_t, \hat{y}))\right) \leq K, & \text{if } y_i \neq y_t \end{cases} \tag{3}$$

During the training process, the target ID in $T(x_i, \hat{y})$ is selected from M_{tr} IDs in the training set, i.e. $\hat{y} \in \{M_{tr}, M_{tr}+1, ..., M_{tr}+M_{te}-1\}$. However, in the testing stage, target IDs $\hat{y} \in \{M_{tr}, M_{tr}+1, ..., M_{tr}+M_{te}-1\}$, which are all unknown identities during training.

Consequently, a backdoor attack on the ReID needs to generate new triggers that vary based on the unseen target ID.

C. Backdoor Trigger Generation

As illustrated in Fig. 2, the invisible triggers generation process consists of two consecutive networks, namely the identity hashing network and the trigger generation network.

Since the target IDs in the inference stage of the recognition task differ from the training stage, the triggers need to be dynamically generated based on the unseen IDs. This is achieved by specifying the unseen target ID $\hat{y}$ with a reference image belonging to this ID, denoted as $\hat{x}$. As a result, we propose an identity hashing network H to extract identity information of $\hat{y}$ from the reference image $\hat{x}$ to obtain a feature representation for this unseen identity. Then, we adopt a pre-trained encoder-decoder network T to generate poisoned images dynamically based on the identity feature, following the DNN-based image steganography, StegaStamp [15], [53]. We choose StegaStamp because of its information capacity, stealthiness, and robustness. It can produce a high-quality image with a capacity better in general than other cover-selection and cover-synthesis-based techniques. The overall process of generating poisoned images can be formalized as:

$$x_i^p = T\left(x_i, H(\hat{x}; \theta_2); \theta_1\right), \tag{4}$$

In this section, we elaborate on the implementation of the aforementioned two networks. Specifically, the identity hashing network consists of four modules: image pre-processing, backbone, GeM pooling layer, and hash layer. A reference image is first fed into a ResNet-101 backbone to obtain a high-level feature map. Then we use GeM pooling to aggregate the feature maps $X \in \mathbb{R}^{W\times H\times C}$ into a global feature $g = [g_1, g_2, \ldots, g_c] \in \mathbb{R}^{1\times 1\times C}$, where W, H, and C denote the width, height, and channel of the feature maps, respectively:

$$g_c = \left(\frac{1}{|X_c|}\sum_{x\in X_c} x^{\alpha}\right)^{\frac{1}{\alpha}}, \tag{5}$$

where α is a control coefficient. At this point, the identity feature of a reference picture has been extracted. However, because the quality of the generated image degrades as steganographic information increases, we employ a hash layer to further compress the feature vector to a much more compact 128-dimensional binary form.

Note that the hash function can be chosen freely based on the length of the generated hash code; in our case, we use MD5 [54]. We use circle loss and triplet loss to train the identity hashing network.

For trigger generation, we propose to adopt a deep image steganography method that directly embeds the identity feature in the image [15], [53]. Specifically, given an image to be poisoned and the binary hash code of the unseen target identity, we first feed the binary code into a linear layer to generate a $32\times 32\times C$ tensor, which is further upsampled to $H\times W\times C$ and concatenated with the image, where C, W and H indicate the channel, width, and height of the input image respectively. Then the $H\times W\times (C\times 2)$ inputs are fed into a U-Net [55] style encoder to generate the poisoned image, in which the trigger is an imperceptible pixel-level perturbation.

The image reconstruction loss serves the purpose of enabling the encoder to embed hash codes within poisoned images while simultaneously minimizing the perceptual dissimilarity between these poisoned images and benign ones. As depicted in Eq.6, it includes three losses, which are L2 residual regularization loss L_R, LPIPS perceptual loss L_P [56], and critical loss L_C calculated between the encoded image and the original image. Experimental validation has revealed that optimal attack performance and stealthiness are achieved when the hyperparameters λ_R, λ_P, and λ_C are set to 1.5, 2, and 1, respectively.

$$L_{image} = \lambda_R L_R + \lambda_P L_P + \lambda_C L_C, \tag{6}$$

This method can achieve as much invisibility of the trigger as possible while maintaining attack performance and increasing the threat of backdoor attacks. To ensure that triggers are added to the images effectively, an extra decoder is added to predict the identity hash code from the poisoned image and another reconstruction loss is added between the predicted code and the original code generated by the identity hashing network.

D. Backdoor Implantation

In this section, we introduce how to implant the backdoor trigger into the training set. The person ReID dataset is more sparse than image classification, where the number of images for each ID is extremely low. The training process follows the all-to-all scenario, in which we use all odd IDs as the target classes of images with even IDs, extract the hash code of each target ID, and inject it into the corresponding training samples by image steganography to generate the poisoned samples. Finally, the poisoned samples and clean samples are provided to the user for training. Because the target IDs in the inference phase differ from those in the training stage, the hash code of unseen IDs in the test set is regenerated based on the reference images and injected into the test set. When the attacked model gets an image with a trigger as input, it is manipulated to retrieve the images of the same target ID with the trigger. The adversary tracked by the ReID model can evade the tracking of the model by injecting a trigger of another one's ID. We note that IBA [57] is an all-to-one backdoor attack method that generates backdoor triggers based on the input, and the images are fed into an encoder that can generate specific patterns. It differs from the classical backdoor attack by grouping the backdoor behavior into three modes: clean mode in which the network can correctly identify clean samples; attack mode in which the backdoor is activated when poisoned data is input; cross-trigger mode in which different inputs do not generate the same triggers, and the triggers of different images cannot be applied to another image. As shown in Fig. 3, the difference between ours and IBA is that we generate target-aware triggers, and the poisoned images with different triggers can be recognized as the same target person. The triggers are generated dynamically based on the benign images and target images, and are not determined by the benign images only.

IV. Experiments

In this section, we evaluate the effectiveness and stealthiness of our proposed backdoor attack on the person ReID models and its resistance to several backdoor defense methods. We conduct ablation studies to demonstrate the importance of our design and to justify the choice of hyperparameters.

A. Experimental Settings

a) Datasets and DNN Models: We choose two benchmark datasets for person ReID: Market-1501 [11] and DukeMTMC-reID [58] to evaluate our attacks. Market-1501 contains 36036 images of 1501 person IDs. 751 IDs were labeled for the training set and 750 IDs were labeled for the test set. In DukeMTMC-reID, a total of 36,411 images of 1812 pedestrians were collected, with 702 IDs in each of the training and test sets. To test the attack performance, we select four victim ReID methods (FastReID) [59], BoT [60], AGW [44], each with ResNet-50 [61] and ResNet-101 [61] backbones for training. In addition, we conduct our attack on MGN [59], PCB [62], and HrNet-18 [63]. The adversary cannot acquire any information about the target ReID model. In fact, the feature extractor of the identity hashing network is constructed and trained independently, which is not the same as that of ReID, and the two are unrelated. The target model is set to different ReID methods, while the identity hashing network is kept the same. And the poisoned image generator trained on the same dataset remains consistent with the parameters that do not change with the attacked model.

b) Baseline: Given the complex nature of target labels in all-to-unknown scenarios, there is currently no backdoor attack that can be utilized directly against Person ReID. Existing attack methods cannot handle the case where the target label is not contained in the training set. Therefore, we only compare with existing methods in terms of the performance of non-targeted attacks where impersonating a specific target identity is not required (more details in evaluation metrics). We select BadNets [2], Blended [64], ReFool [14], SIG [65], and WaNet [52], which are backdoor attack methods originally designed for image classification tasks, as our baseline. Specifically, BadNets is a patch-based backdoor attack that generates poisoned images by adding pixel patterns to benign images. The trigger of Blended is a picture unrelated to the poisoned image, and the poisoned images are generated by overlaying the trigger picture and the original images in a certain ratio. ReFool plants reflections that are also images outside the training set as backdoor triggers into a victim model. SIG uses a ramp signal as a trigger for the poisoned image, which is perceptually invisible. And WaNet is an invisible backdoor attack that injects backdoors by image distortion. In order to improve the ASR, a noise mode is proposed and trained simultaneously with the clean mode and attack mode. All of these methods generate poisoned images and put them into the training set, resulting in the backdoored model to mis-classify images with triggers in the test stage. Not only that, but we also compare the effectiveness of the attack with Badhash [66]. It employs a label-based contrastive learning network (LabCLN) to leverage the semantic representation of distinct labels. This representation is then utilized to confuse and mislead the target model into learning the embedded trigger.

c) Evaluation Metrics: Assuming that the ReID model retrieves 10 pedestrian images most likely to be of the same identity as the query image, the success of attacking a ReID model has two criteria. The first is a non-targeted attack criterion (i.e.evading attack), which measures the attack method's ability to manipulate the target model to rank the positive images outside the top-10 list. Specifically, we use the retrieval performance rank-10 (R-10) and mean Average Precision (mAP) on the poisoned images as the metrics, where lower rank-10 and mAP indicate better non-targeted attack performance. The second criterion is targeted attack [67] (i.e.impersonation attack), which adds a condition to the first one. For a targeted attack, the adversary assigns a specific target person to retrieve, and an attack is only counted as successful when the target person appears in the top-10 rank list [49], [68]. Following existing works [8], [52], [68], [69], we use attack success rate (ASR) (i.e. proportion of successful attacks) as the performance metrics of targeted attack. The ASR calculation given a test set is detailed in Algorithm 1. Here, if $targetedFlag$ is $True$, the ASR of the targeted attack is evaluated; otherwise, the ASR of the non-targeted attack is evaluated. In general, the filename of each image in the dataset contains the ground-truth ID. For each query, the ReID model retrieves ten images with the same identity in the gallery. If one of these ten images belongs to the ground-truth ID, $attackFailFlag$ is set to true, indicating that the non-targeted attack failed once. And if one of the ten images belongs to the target ID, it is counted as a successful targeted attack. Moreover, we use benign accuracy (BA) to evaluate

Algorithm 1 Calculate the ASR of a given poisoned test set

Input: poisoned test set D_t with gallery G and query image Q, Ground-truth ID annotation Y_{gt}, Target ID annotation Y_{tar}

Output: ASR

$queryCnt$ = 1, $attackSuccQuery$ = 1, $targetedFlag$ = $True$, $attackFailFlag$ = $False$
for $query_i$ in Q **do**
  $Ir_i = ReID(query_i)$ ▷Ir_i : top-10 retrieval images from G.
  $queryCnt++$
  $attackFailFlag = False$
  for i in Ir_i **do** ▷i : iterate through all retrieved images.
    if $targetedFlag == False$ **then**
      if $Y_{gt}[query_i] == Y_{gt}[i]$ **then**
        $attackFailFlag = True$
        break
      end if
    else if $Y_{tar}[query_i] == Y_{gt}[i]$ **then**
      $attackSuccQuery$++
      break
    end if
  end for
  if $targetedFlag == True$ **then**
    Continue
  else if $attackFailFlag == Flase$ **then**
    $attackSuccQuery$++
  end if
end for
$ASR = attackSuccQuery/queryCnt$
return ASR

TABLE I
THE PERFORMANCE (%) OF DIFFERENT PERSON REID MODELS UNDER NO AND OUR ATTACKS ON MARKET-1501 AND DUKEMTMC-REID.

Model	Backbone	Clean/ Backdoored	Market-1501				DukeMTMC-reID			
			BA↑	ASR↑	R-10↓	mAP↓	BA↑	ASR↑	R-10↓	mAP↓
FastReID	ResNet50	Clean	98.99	-	98.43	72.92	96.77	-	89.00	54.15
		Backdoored	98.10	97.98	1.07	0.23	96.10	92.77	0.27	0.26
	ResNet101	Clean	99.14	-	98.87	84.65	96.63	-	89.95	57.92
		Backdoored	97.18	91.86	0.50	0.48	91.56	80.43	4.67	1.16
BoT	ResNet50	Clean	98.78	-	98.13	79.80	95.92	-	87.34	52.47
		Backdoored	98.72	99.91	0.15	0.44	95.47	93.36	0.58	0.37
	ResNet101	Clean	98.81	-	98.10	81.94	96.41	-	88.73	54.55
		Backdoored	98.60	99.91	0.21	0.42	95.96	93.04	0.63	0.34
AGW	ResNet50	Clean	98.63	-	98.01	79.95	95.47	-	86.27	52.90
		Backdoored	98.99	99.91	0.15	0.41	95.92	93.36	0.18	0.26
	ResNet101	Clean	98.72	-	97.83	80.57	95.74	-	88.11	55.10
		Backdoored	98.87	99.91	0.09	0.44	96.01	93.36	0.27	0.28
MGN	ResNet50	Clean	98.72	-	98.01	65.89	93.77	-	83.12	43.15
		Backdoored	97.71	99.91	0.15	0.44	93.09	86.31	0.54	0.26
PCB	ResNet50	Clean	96.70	-	95.10	60.53	92.73	-	80.65	62.80
		Backdoored	95.43	95.26	1.84	0.64	93.49	91.02	1.66	0.46
HRNet-18	HRNet	Clean	97.77	-	96.64	68.34	93.76	-	85.41	46.90
		Backdoored	96.61	85.10	9.79	2.52	98.19	91.29	0.49	0.23

whether the attack model can perform normally on clean data, which is the percentage of clean probe images that successfully ranks the positive image in the top-10 list. To evaluate the stealthiness of the backdoor triggers, we select the metrics: structural similarity index (SSIM) [70], peak-signal-to-noise-ratio (PSNR) [71], and LPIPS to measure the differences between clean and poisoned images.

B. Attack Results

a) Effectiveness: Table I shows the attack results of different ReID models on Market-1501 and DukeMTMC-reID datasets at a poisoning rate of $37.5\pm1.5\%$. In fact, the ASR of non-targeted attacks is obtained by subtracting BA from 100%, so we do not compare this metric separately in this table. It is shown that our attack method is able to achieve a high ASR over different ReID methods and different backbones on two standard benchmark datasets. Moreover, we observe a significant performance decrease of the backdoored model on poisoned data in terms of both rank-10 and mAP, further validating our method's effectiveness on the non-targeted attack. In addition, we also evaluate the ReID performance of models trained on clean data and observe that there is only a very small gap between its performance on clean and poisoned data. We also compare attack effectiveness with Badhash on Market-1501, because it is capable of attacking retrieval models. Following the methodology outlined in the Badhash, we first train a hash model using HashNet [72], with ResNet50 as the backbone. Subsequently, we train the BadHash model, which includes a generator, a discriminator, and LabCLN, based on the previously trained hash model. As shown in Table II, with a poisoning rate set at 38.8%, the performance of the attacked model drops significantly, but it maintains relatively high retrieval accuracy compared to our approach. When training LabCLN, the hash code of the target label needs to be specified as a confusing label. Since the target label in the testing phase is not consistent with the training phase, the ASR for targeted attacks cannot be computed. BadHash is a backdoor attack method against deep hashing that can be used for image retrieval tasks, but is more applicable to victim models with a hash layer. Therefore, our attack method is better designed and performs better for the ReID task.

TABLE II
RESULTS OF THE COMPARATIVE EXPERIMENT WITH BADHASH ON THE MARKET-1501 DATASET (THE POISONING RATE = 38.8%)

Model	Method	BA↑	R-10↓	mAP↓
FastReID	BadHash	98.55	54.25	12.75
	ours	98.10	1.07	0.23
BoT	BadHash	98.43	56.65	13.93
	ours	98.72	0.15	0.44
AGW	BadHash	98.28	57.42	14.34
	ours	98.99	0.15	0.41

Figure 4 shows the performance comparison between our method and existing backdoor attacks against FastReID at a 37.5% poisoning rate. Note that in Fig. 4 only non-targeted metric rank-10 is compared because the target individuals in the training stage cannot be specified by the other methods, making the calculation of the targeted ASR impossible. To allow all dimensions to positively reflect attack performance, we subtract the two metrics R-10 and LPIPS from 1, respectively. In Fig. 4, it is shown that our proposed attack gives the model the lowest rank-10 accuracy of 1.07% on the poisoned data at the cost of a very small accuracy loss (0.3% compared to BadNets).

b) Stealthiness: As shown in Fig. 4, we evaluate the stealthiness of the poisoned image by measuring the difference between the original and poisoned image in terms of SSIM,

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2023.3322659

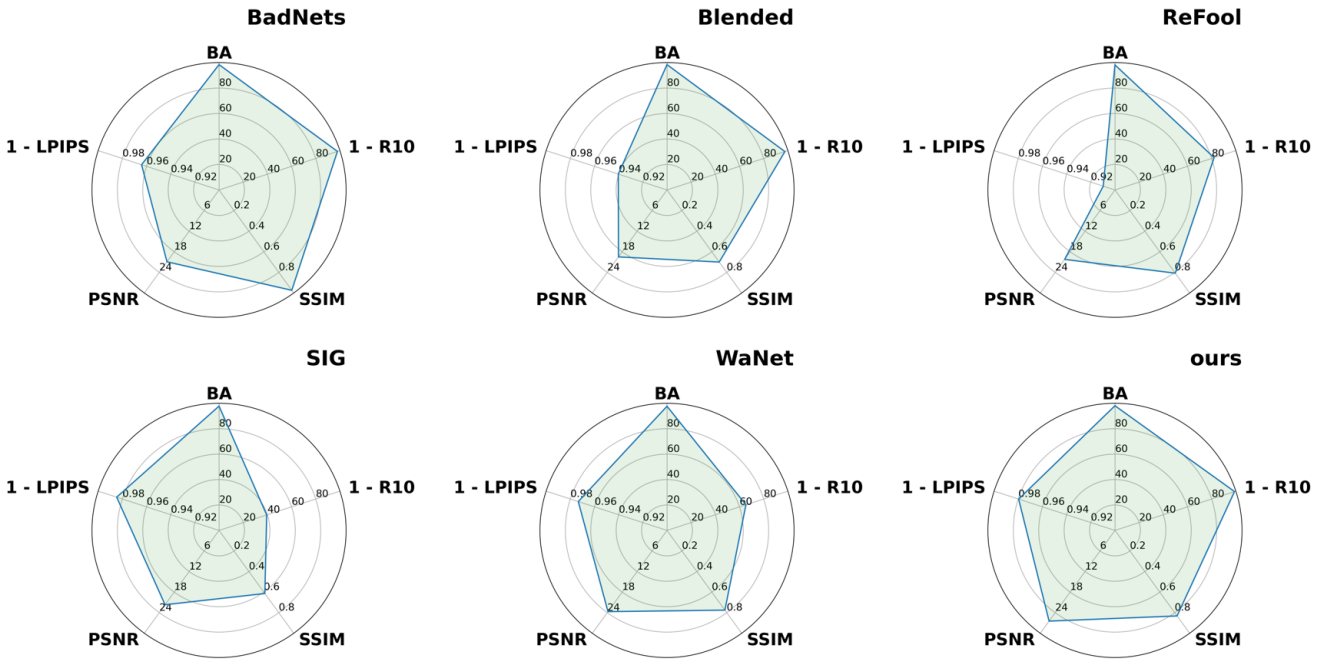


Fig. 4. The attack effectiveness and stealthiness of several classical backdoor attack models and ours to attack FastReID on Market-1501 dataset. The 'BA' and 'R-10' indicate the benign accuracy and the rank-10 accuracy on poisoned data, respectively. And the remaining three metrics SSIM, PSNR, and LPIPS are used to measure the similarity to the original image. Subtracting LPIPS and R-10 from 1 brings them in line with the rest of the metrics, with larger indicating better performance.

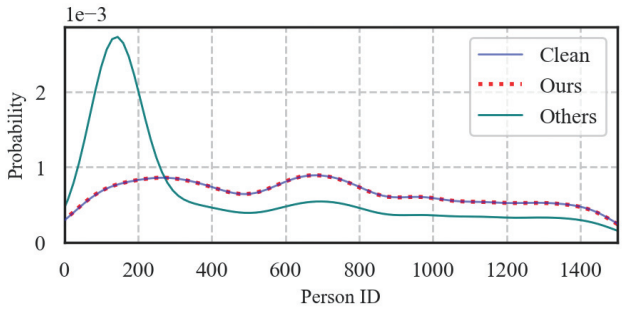


Fig. 5. Data distribution over different identities. The distribution of the training set data after poisoning Market-1501 with the traditional backdoor attacks and ours ($\gamma = 38.8\%$), respectively.

PSNR, and LPIPS. In terms of ASR metrics, our method is the first to successfully attack in an all-to-unknown scenario, and all other methods are not applicable to the scenario. Thus we only compared stealthiness and ReID performance degradation with them in Fig. 4. It can be observed that our method is ranked top two in terms of every stealthiness metric and achieves the best trade-off between BA, Rank-10, and stealthiness. Specifically, Fig. 4 shows ours achieves the maximum performance drop on the poisoned images with a slight BA drop, which is a significant improvement.

As shown in Fig. 5, the blue solid line represents the probability distribution of the original clean data, and the green solid line represents the distribution of the traditional attacks on the training set after poisoning. It can be observed that the distribution of the training set is significantly altered due to changing the labels of many poisoned images to the target labels. While the red dashed line represents our attack, the distribution of data after poisoning is almost the same as the clean training set. In the case of large-scale classification datasets, where each class has a substantial number of samples, adding a small number of poisoned samples may not significantly impact the distribution. However, ReID datasets differ in that there are only a dozen images per class. Consequently, traditional methods would noticeably alter the class distribution of images. Specifically, previous backdoor attack methods usually produce a significant and easily detected change to the training data distribution, because they require selecting a target class during the training stage and then changing the label of the poisoned data to the target label, resulting in a significant increase in the number of images in a particular identity. The data distribution on different identity changes produced by previous methods changes drastically compared to the clean data, which can be detected easily and hence loses stealthiness. On the other hand, our method only needs to add a few images to each target identity in the whole dataset, so it hardly changes the data distribution. From the visualization in Fig. 7, we can observe the changes made to the poisoned image by the previous methods, while for our method the change is almost not perceivable. Note that, BadNets has a high SSIM because the changes to the image are only in the lower right area, while SSIM indicates the average difference of all pixels, but the trigger produced by BadNets is humanly perceivable in the bottom right corner of the image as shown in Fig. 7.

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2023.3322659

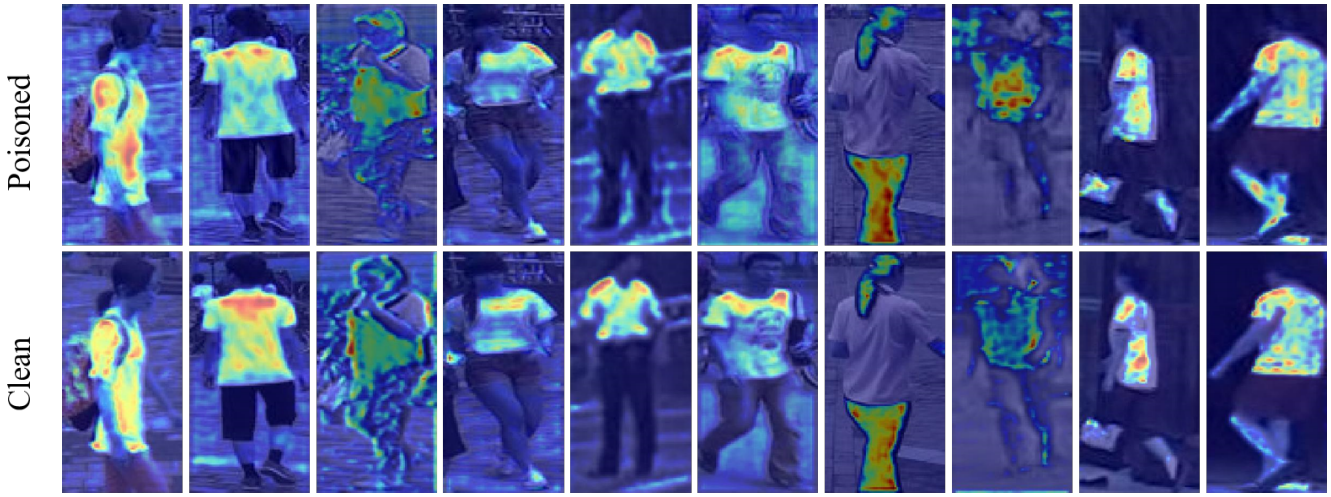


Fig. 6. An illustration of Grad-CAM for the samples in the Market-1501 dataset, where the higher luminance regions are the hot spots and contribute the most to the ReID model.

Fig. 7. The display of poisoned images. Given the original images (first row of odd columns), we make the corresponding poisoned samples using BadNets, Blended, ReFool, SIG, WaNet, and our method. For each method, we show the poisoned images (odd-numbered columns) and the corresponding residual maps (even-numbered columns).

C. Resistance to Backdoor Defense Methods

a) Resistance to RBAT: In reality, person ReID models often employ various data augmentation methods that may remove or corrupt backdoor triggers in the post-poisoned images.

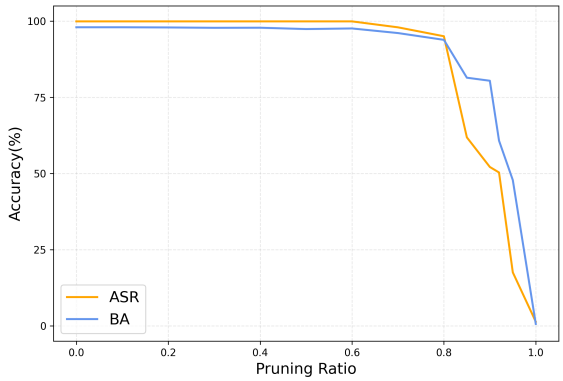
TABLE III
THE RESULTS OF RESISTANCE TO RBAT

Datasets	Clean Accuracy	Poisoned Accuracy
Market1501	99.46	1.31
DukeMTMC-reID	99.87	2.74

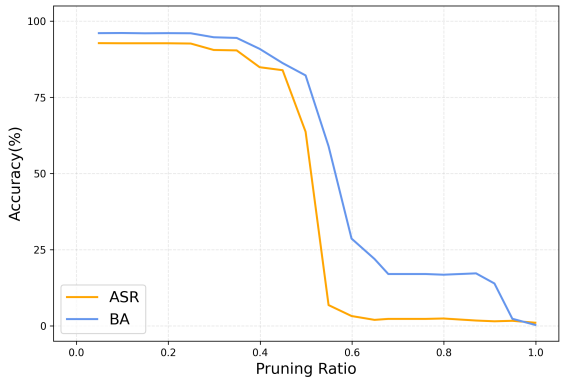
To test whether our attack can break through the RABT [33] defense, we train the defense model on the clean Market-1501 training set and DukeMTMC-reID training set, respectively, and both converge after 20 epochs. Then we embed triggers to images in test sets, where the target ID of an image with an odd ID is set to an even ID, and vice versa. In the test stage, the poisoned images and the original images in a 1:1 ratio are fed into to the pre-trained RABT model, which outputs confidence of the clean and poisoned samples. As shown in Table III, the RABT model can discriminate clean samples with close to 100% accuracy, while the accuracy of poisoned samples does not exceed 3% on both datasets. In other words, RABT classifies 98.69% of the poisoned data as benign samples on Market-1501, and 97.26% on DukeMTMC-reID. It can be seen that our dynamically generated triggers can bypass the defense of RABT.

b) Resistance to Fine-pruning: The purpose of fine-pruning is to remove backdoor-related neurons that are normal in benign images but are abnormally active in poisoned images [36]. We use the fine-pruning algorithm to the person ReID to prune the neurons activated by the backdoor in the last two layers of the first and the last backbone modules. We follow the prune-finetune-then-test pipeline and increase the number of the pruned neurons in the designated layers until only 5% of neurons remain. It can be found that fine-pruning slightly reduces the attack success rate, but the model's performance on clean data decreases at a similar rate, as shown in Fig. 8.

c) Resistance to Februus: With Grad-CAM [73], Februus [74] visualizes the activation hot spots in the penultimate layer of a DNN and replaces the pixels corresponding to the highest-scoring hot spots with image patches recovered by a



(a) Market-1501



(b) DukeMTMC-reID

Fig. 8. The results of resistance to Fine-pruning on Market-1501 and DukeMTMC-reID datasets. In both subgraphs, the ASR decreases in parallel with the benign accuracy as the proportion of pruned neurons increases, making it difficult to defend against our attacks by Fine-pruning.

pre-trained GAN that is used to remove potential backdoor triggers. We first provide the hot spots map in Fig. 6. The hot spots (shadows with higher luminance) in the poisoned images overlap with the body or clothing of pedestrians, which are almost no different from the clean images. In addition, the triggers generated by our attack are spread over the whole image and cannot be defended by replacing local image patches. We conduct this defense experiment on the Market-1501 dataset and found that our attack can still maintain 93.75% ASR after trigger pattern removal.

d) Resistance to Steganalysis: Steganalysis is an effective way to combat steganography by detecting coded images. We try to use state-of-the-art steganalyzers to detect the generated poisoning images [75]–[77], and the experimental results are shown in Table IV. But they can't be used directly in backdoor defense scenarios. This limitation arises from the fact that they rely on the sample pair analysis algorithm, necessitating the separation of cover and stego images during the training phase. Therefore, we generate the encoded images using the pre-trained StegaStamp with an input message length of 100 bits. The generated stego images and cover images are fed into the steganalyzers for training on the Market-1501 dataset. Notably, during the training phase, all three models successfully converged within 200 epochs. Subsequently, we evaluate the performance of these pre-trained steganalyzers on the poisoning test set, resulting in detection accuracies of 49.11%, 50.01%, and 50.02%, respectively. These results underscore the incapacity of these three steganalysis methods to detect poisoning images in real defense settings.

TABLE IV
THE RESULTS OF RESISTANCE TO STEGANALYSIS

Model	Pretraining Acc	Backdoor Detection Acc
SRNet [75]	99.47	49.11
CovpoolNet [76]	100.0	50.01
LWENet [77]	99.87	50.02

D. Ablation Studies

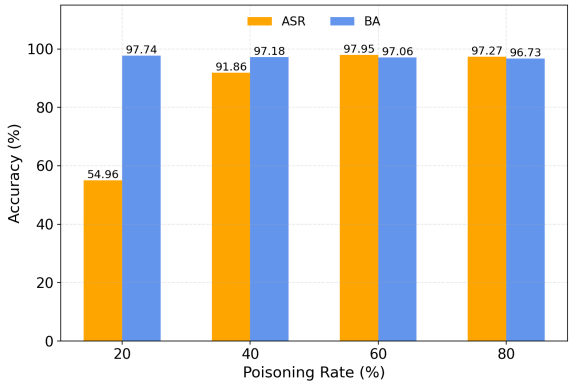
a) Importance of our design: In this experiment, We do ablation experiments on the Market-1501 dataset to prove that the Hamming space is applicable for attacking the ReID model and to show that an identity hashing network is necessary. First, We extract the feature of the reference image directly without going through the hash layer to get a 2048-dimensional feature vector. It is directly used as a perturbation to sum the pixel values of the benign image and make it mapped to the effective pixel value domain. The results of the attack are shown in Table V, and it can be observed that the ASR is only 12.7% (poisoning rate = 38.8%) although the accuracy is reduced.

Secondly, We replace the hash code generated by the identity hashing network with random noise to create poisoned images. However, if random noise is used as the input of the identity hashing network instead of the reference image's feature, it does not reduce the distance between the poisoned image and the target person in the feature space as our proposed method does. We conduct the ablation experiment to validate this point, and the ASR when the poisoning rate = 38.8% in the test set is only 0.83%, which is enough to prove that the selection of reference image is necessary.

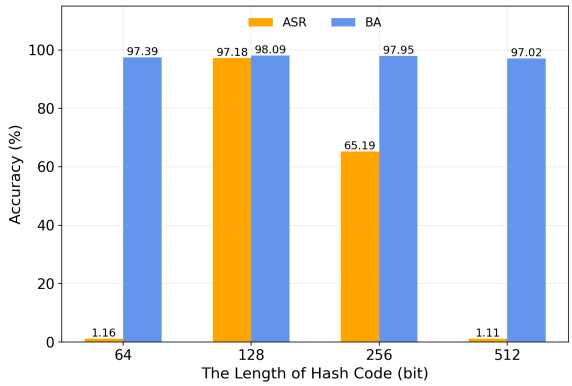
TABLE V
THE RESULT OF ABLATION EXPERIMENTS USING FEATURES OF REFERENCE IMAGES INSTEAD OF GENERATED BACKDOOR TRIGGERS AND RANDOM NOISE INSTEAD OF HASH CODE, RESPECTIVELY.

Pattern	BA↑	ASR↑	R-10↓	mAP↓
Feature	97.89	12.70	12.87	9.62
Random noise	98.66	0.83	1.22	0.57
ours	98.10	97.98	1.07	0.23

b) Poisoning Rate: Fig. 9 (a) shows the ASR and BA change over different poisoning rates, where both BA and ASR are affected by the poisoning rate. Specifically, ASR increases as the poisoning rate rises, whereas BA maintains at or above 96.73%. The ASR is 91.86% when the poisoning rate $\gamma = 38.8\%$, and reaches 96.94 % when $\gamma = 75\%$. Note that an increase in the poisoning rate increases the likelihood that the backdoor would be discovered, and the adversary must strike a balance between attack effectiveness and stealthiness.



(a) The ASR and BA w.r.t. different poisoning rates.



(b) The ASR and BA w.r.t. different lengths of hash code.

Fig. 9. Ablation experiments of the poisoning rate and the length of the hash code.

c) Length of Hash Code: To generate backdoor triggers, we compress the high-dimensional identity feature of a reference image into a low-dimensional hash code. The length of the hash code correlates directly with the quality of the image generated by the image steganography network, as well as the identity information contained in the hash code, which consequently has an indirect impact on attack performance. Fig. 9(b) shows the changes of BA and ASR over different hash code lengths at a poisoning rate $\gamma = 38.8\%$. We observe that our method achieves the highest ASR when the length of the hash code is 128 bits. ASR drops significantly when the code length is larger than 128 because the embedded trigger damages the original information in the image. On the other hand, when the code length is small (64), ASR also decreases because the binary code too short does not contain enough identity information for the backdoor attack.

V. Conclusion

Most of the current research on backdoor attacks focuses on image classification tasks, while the risk of backdoor attacks against person ReID has rarely been studied. Existing backdoor attacks against image classification follow all-to-one/all scenarios and can not be directly applied to attack the open-set ReID model. As a result, we propose a novel backdoor attack on deep ReID models under a new all-to-unknown scenario, which is able to dynamically generate new backdoor triggers containing unknown identities in the test set. Specifically, an identity hashing network is proposed to first extract target identity information from a reference image, which is then injected into the benign images by image steganography. We show that the proposed attack method performs well in terms of effectiveness and stealthiness, and is robust to existing defense methods. With some problems left open, we hope that this study will raise more attention on the backdoor attack risk against person ReID.

REFERENCES

[1] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852.

[2] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

[3] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?" *arXiv preprint arXiv:1911.07963*, 2019.

[4] Y. Li, B. Wu, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *arXiv preprint arXiv:2007.08745*, 2020.

[5] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 070–16 084, 2020.

[6] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2938–2948.

[7] K. Doan, Y. Lao, and P. Li, "Backdoor attack with imperceptible input and latent modification," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[8] Z. Wang, J. Zhai, and S. Ma, "Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 074–15 084.

[9] Y. Feng, B. Ma, J. Zhang, S. Zhao, Y. Xia, and D. Tao, "Fiba: Frequency-injection based backdoor attack in medical image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 876–20 885.

[10] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 152–159.

[11] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[12] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[13] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2285–2294.

[14] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *European Conference on Computer Vision*. Springer, 2020, pp. 182–199.

[15] S. Li, M. Xue, B. Z. H. Zhao, H. Zhu, and X. Zhang, "Invisible backdoor attacks on deep neural networks via steganography and regularization," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2088–2105, 2020.

[16] A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden trigger backdoor attacks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 957–11 965.

[17] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, "Invisible backdoor attack with sample-specific triggers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 463–16 472.

[18] N. Manoj and A. Blum, "Excess capacity and backdoor poisoning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[19] Y. Yao, H. Li, H. Zheng, and B. Y. Zhao, "Latent backdoor attacks on deep neural networks," in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2019, pp. 2041–2055.

[20] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao, "Backdoor attacks against deep learning systems in the physical world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6206–6215.

[21] N. Pitropakis, E. Panaousis, T. Giannetsos, E. Anastasiadis, and G. Loukas, "A taxonomy and survey of attacks against machine learning," *Computer Science Review*, vol. 34, p. 100199, 2019.

[22] J. Wang, G. M. Hassan, and N. Akhtar, "A survey of neural trojan attacks and defenses in deep learning," *arXiv preprint arXiv:2202.07183*, 2022.

[23] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J. Sohn, K. Lee, and D. S. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," in *NeurIPS*, 2020.

[24] T. Zhai, Y. Li, Z. Zhang, B. Wu, Y. Jiang, and S.-T. Xia, "Backdoor attack against speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2560–2564.

[25] L. Wang, Z. Javed, X. Wu, W. Guo, X. Xing, and D. Song, "BACKDOORL: backdoor attack against competitive reinforcement learning," in *IJCAI*. ijcai.org, 2021, pp. 3699–3705.

[26] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y. Jiang, "Clean-label backdoor attacks on video recognition models," in *CVPR*. Computer Vision Foundation / IEEE, 2020, pp. 14 431–14 440.

[27] Z. Xiang, D. J. Miller, S. Chen, X. Li, and G. Kesidis, "A backdoor attack against 3d point cloud classifiers," in *ICCV*. IEEE, 2021, pp. 7577–7587.

[28] Y. Li, H. Zhong, X. Ma, Y. Jiang, and S.-T. Xia, "Few-shot backdoor attacks on visual object tracking," *arXiv preprint arXiv:2201.13178*, 2022.

[29] J. Zhang, C. Dongdong, Q. Huang, J. Liao, W. Zhang, H. Feng, G. Hua, and N. Yu, "Poison ink: Robust and invisible backdoor attack," *IEEE Transactions on Image Processing*, vol. 31, pp. 5691–5705, 2022.

[30] M. Javaheripi, M. Samragh, G. Fields, T. Javidi, and F. Koushanfar, "Cleann: Accelerated trojan shield for embedded neural networks," in *ICCAD*. IEEE, 2020, pp. 11:1–11:9.

[31] K. Jin, T. Zhang, C. Shen, Y. Chen, M. Fan, C. Lin, and T. Liu, "A unified framework for analyzing and detecting malicious examples of dnn models," *arXiv preprint arXiv:2006.14871*, 2020.

[32] M. Du, R. Jia, and D. Song, "Robust anomaly detection and backdoor attack detection via differential privacy," in *ICLR*. OpenReview.net, 2020.

[33] Y. Zeng, W. Park, Z. M. Mao, and R. Jia, "Rethinking the backdoor attacks' triggers: A frequency perspective," in *ICCV*. IEEE, 2021, pp. 16 453–16 461.

[34] S. Hong, V. Chandrasekaran, Y. Kaya, T. Dumitraş, and N. Papernot, "On the effectiveness of mitigating data poisoning attacks with gradient shaping," *arXiv preprint arXiv:2002.11497*, 2020.

[35] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Anti-backdoor learning: Training clean models on poisoned data," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[36] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer, 2018, pp. 273–294.

[37] D. Wu and Y. Wang, "Adversarial neuron pruning purifies backdoored deep models," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[38] Y. Zeng, S. Chen, W. Park, Z. Mao, M. Jin, and R. Jia, "Adversarial unlearning of backdoors via implicit hypergradient," in *International Conference on Learning Representations*, 2021.

[39] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Neural attention distillation: Erasing backdoor triggers from deep neural networks," *arXiv preprint arXiv:2101.05930*, 2021.

[40] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1528–1535.

[41] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8514–8522.

[42] Z. Zhu, X. Jiang, F. Zheng, X. Guo, F. Huang, X. Sun, and W. Zheng, "Aware loss with angular regularization for person re-identification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 114–13 121.

[43] J. Zhou, B. Su, and Y. Wu, "Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2909–2918.

[44] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.

[45] C. Zhao, Y. Tu, Z. Lai, F. Shen, H. T. Shen, and D. Miao, "Salience-guided iterative asymmetric mutual hashing for fast person re-identification," *IEEE Transactions on Image Processing*, vol. 30, pp. 7776–7789, 2021.

[46] Z. Cui, J. Zhou, Y. Peng, S. Zhang, and Y. Wang, "Dcr-reid: Deep component reconstruction for cloth-changing person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[47] W. Ding, X. Wei, R. Ji, X. Hong, Q. Tian, and Y. Gong, "Beyond universal person re-identification attack," *IEEE transactions on information forensics and security*, vol. 16, pp. 3442–3455, 2021.

[48] S. Bai, Y. Li, Y. Zhou, Q. Li, and P. H. Torr, "Adversarial metric attack and defense for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2119–2126, 2020.

[49] Z. Wang, S. Zheng, M. Song, Q. Wang, A. Rahimpour, and H. Qi, "advpattern: physical-world attacks on deep person re-identification via adversarially transformable patterns," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8341–8350.

[50] F. Yang, Z. Zhong, H. Liu, Z. Wang, Z. Luo, S. Li, N. Sebe, and S. Satoh, "Learning to attack real-world models for person re-identification via virtual-guided meta-learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3128–3135.

[51] F. Yang, J. Weng, Z. Zhong, H. Liu, Z. Wang, Z. Luo, D. Cao, S. Li, S. Satoh, and N. Sebe, "Towards robust person re-identification by defending against universal attackers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 5218–5235, 2022.

[52] A. Nguyen and A. Tran, "Wanet–imperceptible warping-based backdoor attack," *arXiv preprint arXiv:2102.10369*, 2021.

[53] M. Tancik, B. Mildenhall, and R. Ng, "Stegastamp: Invisible hyperlinks in physical photographs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2117–2126.

[54] R. Rivest, "The md5 message-digest algorithm," Tech. Rep., 1992.

[55] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[56] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[57] T. A. Nguyen and A. Tran, "Input-aware dynamic backdoor attack," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3454–3464, 2020.

[58] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3754–3762.

[59] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.

[60] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2019.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[62] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. 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)*, 2018, pp. 480–496.

[63] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[64] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[65] M. Barni, K. Kallas, and B. Tondi, "A new backdoor attack in cnns by training set corruption without label poisoning," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 101–105.

[66] S. Hu, Z. Zhou, Y. Zhang, L. Y. Zhang, Y. Zheng, Y. He, and H. Jin, "Badhash: Invisible backdoor attacks against deep hashing with clean label," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 678–686.

[67] J. Weng, Z. Luo, S. Li, N. Sebe, and Z. Zhong, "Logit margin matters: Improving transferable targeted adversarial attack by logit calibration," *IEEE Transactions on Information Forensics and Security*, 2023.

[68] S. Fang and A. Choromanska, "Backdoor attacks on the dnn interpretation system," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 561–570.

[69] Z. Zhao, X. Chen, Y. Xuan, Y. Dong, D. Wang, and K. Liang, "Defeat: Deep hidden feature backdoor attacks by imperceptible perturbation and latent representation constraints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 213–15 222.

[70] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[71] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.

[72] Z. Cao, M. Long, J. Wang, and P. S. Yu, "Hashnet: Deep learning to hash by continuation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5608–5617.

[73] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[74] B. G. Doan, E. Abbasnejad, and D. C. Ranasinghe, "Februus: Input purification defense against trojan attacks on deep neural network systems," in *Annual Computer Security Applications Conference*, 2020, pp. 897–912.

[75] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1181–1193, 2018.

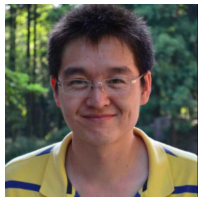
[76] X. Deng, B. Chen, W. Luo, and D. Luo, "Fast and effective global covariance pooling network for image steganalysis," in *Proceedings of the ACM workshop on information hiding and multimedia security*, 2019, pp. 230–234.

[77] S. Weng, M. Chen, L. Yu, and S. Sun, "Lightweight and effective deep image steganalysis network," *IEEE Signal Processing Letters*, vol. 29, pp. 1888–1892, 2022.

Wenli Sun received the B.E. degree in computer science and technology from Tongji University in 2022. She is currently pursuing a master's degree at Tongji University. Her research interests include computer vision and deep learning. Her main research interests and specific areas of research include backdoor attacks, human pose transfer, and person Re-IDentification.

Xinyang Jiang received his PhD degree in computer science and technology from Zhejiang University in 2017, and was a senior researcher in Tencent Youtu Lab. He is a researcher at Microsoft Research Asia. His main research areas include cross-modal retrieval, computer vision, and person Re-IDentification. He has published more than ten papers in CVPR, ECCV, AAAI, ACMMM, TIP and other top conferences and journals on computer vision and artificial intelligence. He is a program member of AAAI, CVPR, MM and other conferences, and a reviewer for TCSVT, TIP and other journals.

Shuguang Dou is currently pursuing a Ph.D. degree with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include computer vision and deep learning. Specific areas of research include Infographic understanding, person Re-IDentification, NAS benchmark, and X-ray.



Dongsheng Li is currently a principal research manager with Microsoft Research Asia (MSRA), Shanghai, China. Meanwhile, he is an adjunct professor with School of Computer Science, Fudan University, Shanghai, China. Before joining MSRA Shanghai, he worked as a research staff member (RSM) with IBM Research - China, Shanghai, China. He obtained Ph.D. from school of computer science, Fudan University in 2012 (supervised by Prof. Ning Gu), and B.E. from the department of computer science and technology, University of Science and Technology of China (USTC) in 2007. He also visited the University of Colorado Boulder as a visiting scholar from 2010.8 to 2011.2 (supervised by Prof. Qin Lv and Prof. Li Shang). He is the member of CCF, ACM, and IEEE. He serves as the program committee members of top conferences, e.g., NIPS, ICML, ICLR, AAAI, IJCAI, CIKM, etc.

Duoqian Miao was born in 1964. He is currently a Professor and the Ph.D. Tutor with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. He serves as the Vice President for the International Rough Set Society, the Executive Manager of the Chinese Association for Artificial Intelligence, the Chair of the CAAI Granular Computing Knowledge Discovery Technical Committee, a Distinguished Member of Chinese Computer Federation, the Vice President of the Shanghai Computer Federation, and the Vice President of the Shanghai Association for Artificial Intelligence. He serves as Associate Editor for the International Journal of Approximate Reasoning and an Editor of the Journal of Computer Research and Development (in Chinese).

Cheng Deng (Senior Member, IEEE) received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 2001, 2006, and 2009, respectively. He is currently a Full Professor with the School of Electronic Engineering, Xidian University. He has authored and coauthored more than 100 scientific articles at top venues, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, TRANSACTIONS ON IMAGE PROCESSING, TRANSACTIONS ON CYBERNETICS, TRANSACTIONS ON MULTIMEDIA, TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, the International Conference on Computer Vision, Computer Vision and Pattern Recognition, the International Conference on Machine Learning, Neural Information Processing Systems, the International Joint Conference on Artificial Intelligence, and the Association for Advancement of Artificial Intelligence. His current research interests include computer vision, pattern recognition, and information hiding.

Cairong Zhao is currently a Professor of the College of Electronic and Information Engineering at Tongji University. He received a Ph.D. degree from Nanjing University of Science and Technology, an M.S. degree from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences and a B.S. degree from Jilin University, in 2011, 2006 and 2003, respectively. He works on visual and intelligent learning, including computer vision, pattern recognition, and visual surveillance. He has published over 40 top-rank international conferences and journals in the field, including CVPR, ICCV, ICLR, AAAI, ACM MM, TIP, TIFS, TMM, TCSVT, and PR. He holds prestigious positions such as the deputy secretary-general of the Pattern Recognition and Machine Intelligence Committee of the Chinese Association of Automation, the chairman of the Computer Vision Special Committee of the Shanghai Computer Society, and an outstanding member of the China Computer Federation, and a senior member of the China Graphics Society. He also serves as the reviewer of more than ten AI-related international journals and conferences, including TPAMI, TIP, CVPR, ICCV, NIPS, ICML, and AAAI etc.. He has hosted three National Natural Science Foundation projects, sub-projects of national key research and development plans, and more than ten enterprise-level horizontal projects.