International Journal of Computer Vision (2023) 131:835–854
https://doi.org/10.1007/s11263-022-01737-y

U-Turn: Crafting Adversarial Queries with Opposite-Direction Features

Zhedong Zheng[1] · **Liang Zheng**[2] · **Yi Yang**[3] · **Fei Wu**[3]

Received: 10 February 2022 / Accepted: 30 November 2022 / Published online: 20 December 2022

Abstract
This paper aims to craft adversarial queries for image retrieval, which uses image features for similarity measurement. Many commonly used methods are developed in the context of image classification. However, these methods, which attack prediction probabilities, only exert an indirect influence on the image features and are thus found less effective when being applied to the retrieval problem. In designing an attack method specifically for image retrieval, we introduce opposite-direction feature attack (ODFA), a white-box attack approach that directly attacks query image features to generate adversarial queries. As the name implies, the main idea underpinning ODFA is to impel the original image feature to the opposite direction, similar to a U-turn. This simple idea is experimentally evaluated on five retrieval datasets. We show that the adversarial queries generated by ODFA cause true matches no longer to be seen at the top ranks, and the attack success rate is consistently higher than classifier attack methods. In addition, our method of creating adversarial queries can be extended for multi-scale query inputs and is generalizable to other retrieval models without foreknowing their weights, *i.e.,* the black-box setting.

Keywords Adversarial samples · Robustness · Image retrieval · Convolutional neural network · Deep learning.

1 Introduction

Given a query image, image retrieval is to give high ranks to gallery images of similar content (Zhang et al., 2017; Radenović et al., 2018; Shen et al., 2019; Liu et al., 2017; Li et al., 2019b). In this paper, we are interested in understanding how an image retrieval system reacts to adversarial attacks. More specifically, we investigate how to effectively generate an adversarial query leading to compromised ranking results where true match images should receive low ranks.

Existing attack methods are usually developed in the field of image classification, which aims to significantly alter class predictions (Szegedy et al., 2014; Goodfellow et al., 2015; Kurakin et al., 2017; Moosavi Dezfooli et al., 2016; Madry et al., 2018). For instance, for an image of the class *car*, an attacked system would provide a completely irrelevant predicted class *tree*. This is usually achieved by back-propagating the gradient from changed class prediction to add a few imperceptible changes to image pixels.

Essentially, image classification attack methods are not suitable to attack image retrieval systems. The main reason is that image retrieval harnesses *image features* to compute the similarities between the query and gallery images instead of using the class predictions. Attacking the class predictions, as performed by many existing methods, only has indirect influence on the features and thus has limited effectiveness (see Fig. 1). As such, to successfully attack a retrieval system, it is crucial to directly bring changes to image features.

Another consideration is which images to attack. A retrieval system meets two types of image inputs, query and gallery images. If we attack the latter, as we do not know in advance which gallery images are true matches, we need to

Communicated by V. Lepetit.

This work was supported by the Fundamental Research Funds for the Central Universities (No. 226-2022-00051).

✉ Liang Zheng
liang.zheng@anu.edu.au

Zhedong Zheng
zdzheng@nus.edu.sg

Yi Yang
yangyics@zju.edu.cn

Fei Wu
wufei@cs.zju.edu.cn

[1] The Sea-NExT Joint Lab, School of Computing, National University of Singapore, Singapore 118404, Singapore

[2] The School of Computing, Australian National University, Canberra 2601, Australia

[3] The College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

Table 1 Comparison of image classification and image retrieval and its implications on attack method design. In the test procedure, image classification obtains prediction probabilities for a test image, while image retrieval extracts image features to compute the similarity between the query and gallery images. As such, classification attack methods find disrupting class predictions to be effective; and we propose to attack image features directly under the context of image retrieval. We show that this strategy is very effective on various retrieval benchmarks

	Evaluation	Adversary
Image Classification	Class prediction	Attack class Prediction
Image Retrieval	Feature similarity	Attack feature

attack all the gallery images. This procedure is prohibitively time-consuming as a gallery contains thousands or millions of images (Sigurbjörnsson & Van Zwol, 2008; Liu et al., 2012; Lin et al., 2015; Chen et al., 2016). In contrast, the query typically contains one image,[1] allowing us to perform the attack efficiently and effectively (Obviously, disrupting the query will completely compromise retrieval results).

Motivated by the aforementioned factors, we propose opposite-direction feature attack (ODFA), a white-box attack method for adversarial query generation under the context of image retrieval context. In a nutshell, ODFA directly performs attack on the feature level instead of on the class predictions, so it is consistent with the image retrieval test procedure, *i.e.,* measuring similarity between the query and gallery images using their respective features. More specifically, ODFA forces the query feature to move towards the opposite direction of itself in the feature space. The resulting gradient is back-propagated to the query image to generate a few imperceptible pixel changes. With direction-reversed features, the similarity between the query and the true matches will significantly decrease, causing the true matches to be determined as outliers (with low ranks). We evaluate our method on five image retrieval datasets and show that under various levels of image perturbation, ODFA outperforms popular classification attack methods such as fast-gradient sign method (Goodfellow et al., 2015) and basic iterative method (Kurakin et al., 2017). Moreover, compared with a few concurrent retrieval attack methods such as (Liu et al., 2019b; Bouniot et al., 2020), ODFA is also very competitive. In addition, we show that ODFA is effective in black-box settings, where adversarial queries crafted for a white-box model remain adversarial for models with unknown weights. We summarize the main points below.

[1] There exist multi-query retrieval systems (Zheng et al., 2015; Wang et al., 2017) but for simplicity, we only consider single-query systems.

- We propose opposite-direction feature attack (ODFA) that effectively attacks the query features to fool state-of-the-art image retrieval systems. Unlike the classification attack methods, ODFA works on the feature level which well aligns with image retrieval procedure.
- On a series of retrieval datasets, we show ODFA is superior to classification attack methods and competitive with existing retrieval attack approaches: large accuracy drop is observed. We additionally show ODFA can be extended for multi-scale queries and remains useful for the black-box setting.

The rest of the paper is arranged as follows. Section 2 discusses relevant works, followed by the preliminaries in Sect. 3. In Sect. 4, we discuss the limitation of applying classification attack methods to retrieval scenarios, introduce our method and extend it for multiple-scale inputs. Experimental results are summarized and discussed in Sect. 5, followed by the conclusion in Sect. 6.

2 Related Work

Image retrieval. Image retrieval relies on visual features for similarity measurement to generate the ranking list (Yue-Hei Ng et al., 2015; Deng et al., 2019; Yang et al., 2018a; Jin et al., 2018; Lin et al., 2018; Yang et al., 2017; Yan et al., 2020). Current works mostly deploy deep learning models, *e.g.,* the convolutional neural network (CNN), to extract intermediate output as the visual representation, which is shown to have a strong discriminative ability (Radenović et al., 2016; Tolias et al., 2015; Yang et al., 2018; Zheng et al., 2020).

The feature learning process is directly motivated by various objectives, such as the classification loss (Zheng et al., 2016, 2018b) and the triplet loss with a hard sampling policy (Ristani & Tomasi, 2018; Song et al., 2016; Zheng et al., 2018b; Yu et al., 2018). Some further leverage local patterns for fine-grained feature mining (Yu et al., 2017; Liu et al., 2017; Bai et al., 2017; Radenović et al., 2018; Wang et al., 2020; Yang et al., 2021) or involve semantic parts (Zhang et al., 2017; Suh et al., 2018; Sun et al., 2019; Bai et al., 2020b). In this work, we evaluate the widely used ResNet backbones (ResNet-50 and ResNet-101) (He et al., 2016) and the part-based method PCB (Sun et al., 2018) as the victim model.

Retrieval robustness. A commonly seen practice is to add distractors to the gallery, where system scalability can be evaluated (Philbin et al., 2007, 2008; Zheng et al., 2015; Guo et al., 2018). Another interesting aspect is system vulnerability to adversarial examples which might affect ranking. Tolias et al. (2019) pull close representations between the target query and a pre-defined object, while Liu et al. (2019b)

aim to increase the L2 feature distance between the original query and the adversarial query.

Similarly, Zhou et al. (2021) propose a ranking-based method that changes the rank order of retrieved objects. Bai et al. (2020a) show that the adversarial distance gradients can successfully cheat person re-identification frameworks. Similar results are also observed in vehicle re-identification (Yu et al., 2021). Note that we view these works as concurrent: many of them cited a preliminary version of our work which appeared in 2018 (Zheng et al., 2018a). We also note that our work is sufficiently different from them: the proposed method ODFA explicitly specifies the feature direction, *i.e.,* opposite direction, producing failure cases effectively and efficiently.

Adversarial attack. Adversarial attack is usually developed in image classification, makes slights changes to a realistic image to fool trained models and helps evaluate its robustness (Sharif et al., 2016; Chakraborty et al., 2018; Zhang et al., 2018; Eykholt et al., 2018; Li et al., 2019a; Shi et al., 2021). Literature can be broadly divided into two classes: gradient-based attack and score-based attack. In the former category, the fast-gradient sign method applies the adversarial gradient as the input perturbation (Goodfellow et al., 2015). This method is further extended by various iterative strategies, *e.g.,* basic iterative method (Kurakin et al., 2017), deep fool (Moosavi Dezfooli et al., 2016), momentum (Dong et al., 2018) and Hamming space (Yang et al., 2018b). Score-based attack, on the other hand, relies on searching the input space, considering that the distorted images can largely affect the prediction score (Yan et al., 2020b, 2021; Cherepanova et al., 2021; Xiao et al., 2018). Jacobian-based saliency map attack greedily modifies the input instance (Papernot et al., 2016b). Narodytska & Kasiviswanathan (2017) further show that single pixel perturbation, which is out of the valid image range, is effective to attack small-sized images. Gong et al. (2022) change the color of image regions to grayscale.

The closest inspiring work is the iterative least-likely class method (Kurakin et al., 2017), which makes the classifier output difficult mistakes: classifying an image of *vehicle* into the class with the lowest confidence score, *e.g., cat*. They achieve this by increasing the prediction probability of the least likely class. Our work has a similar spirit, where we impel image features to the "least-likely" opposite direction. Here we emphasize that our work is sufficiently different from (Kurakin et al., 2017). Kurakin *et al.* attack class predictions to obtain the least-likely class. This method does not apply well in retrieval because the latter usually deals with test images from unseen classes, to which assigning seen classes compromises the adversarial gradients. In comparison, the proposed method works on the intermediate feature and explicitly decreases feature similarity between the adversarial image and its original image. Our procedure aligns very well with the image retrieval test procedure.

Adversarial robustness. It is reported that printed or photographed versions of an attacked image would discard the imperceptible perturbations and exhibit stronger robustness against attack (Kurakin et al., 2017). A number of works further investigate the adversarial robustness problem. For example, Athalye et al. (2018) propose a general-purpose algorithm, expectation Over transformation (EoT), which aggregates the adversarial gradients after various data augmentations are applied. Inspired by EoT, we extend the proposed ODFA to ODFA+EoT (MS) for multi-scale inputs, as image resizing is often applied to augment the query (Radenović et al., 2018; Zheng et al., 2020). ODFA+EoT (MS) outperforms ODFA and some other competitive methods.

Black-box attack. Black-box attack is also an attractive problem. Its main difference from the white-box setting is whether attack foreknows the victim model's weights or structure (Kurakin et al., 2017; Tramèr et al., 2018; Li et al., 2021; Wang et al., 2022). A common practice is to train a new model to mimic outputs of the black-box model, and then harness this white-box student model as target victim to generate adversarial samples (Kurakin et al., 2017; Li et al., 2021; Wang et al., 2022). Distilling black-box model parameters (the key to black-box setting) is out of the scope of our paper. Therefore, in the experiment, we mainly study the direct transferability of the adversarial query to the black-box setting and compare the performance between our method with some other white-box methods.

Attack defense. It remains challenging to defend adversarial attack. There are two types of defense, *i.e.,* defense during training or test. A common practice of training defense is to involve adversarial samples in training data (Tramèr et al., 2018; Madry et al., 2018). Another line of works study defense by rejecting prediction during inference. For instance, Wang et al. (2021) argue to detect query outliers by checking the distance inconsistency. In the experiment, we provide a preliminary experiment using training defense against ODFA.

3 Preliminaries

3.1 Problem Definition

Given a query image X and the gallery G, the image retrieval model ranks gallery images according to the similarity score $S(X, g)$ $(g \in G)$ in the feature space. For two images X_m and X_n, the similarity score in the feature space can be formulated as the cosine similarity: $S(X_m, X_n) = f_{X_m} \cdot f_{X_n}$ where f_X is the L2-normalized visual feature. In this work, we mainly study f_X extracted by a non-linear deeply-learned

mapping function F, such as Convolutional Neural Network (CNN) (LeCun & Bengio, 1995). A perfect retrieval system allows true matches X_{gt} (images containing similar content with the query) to be top ranked $S(X, X_{gt}) \geq S(X, g)$ $(\forall g \in G)$. To attack this system, we aim to generate an adversarial example X' to replace the original query image such that $S(X', X_{gt}) \leq S(X', g)$ $(\forall g \in G)$. In the meantime, we demand that X' is visually similar to query X. The pixel-level difference between the adversarial query and the original query should be small, ensuring the adversarial perturbation to be imperceptible to the naked eye. In particular, we follow the practice in (Kurakin et al., 2017) to keep the pixel differences within a valid value range. We clip the pixels whose values fall out of the valid range and remove distortions which are greater than the hyper-parameter perturbation rate ϵ: $\text{Clip}_{X,\epsilon}\{X'\} = \min\{255, X + \epsilon, \max\{0, X - \epsilon, X'\}\}$. It ensures $\|X' - X\|_\infty \leq \epsilon$. Since a too large ϵ will make the perturbation perceptible to humans, we generally set $\epsilon \leq 16$ in this work.

3.2 Victim Model

We call the retrieval model to be attacked as the victim model and apply the white-box assumption following the conventional gradient-based methods (Szegedy et al., 2014; Goodfellow et al., 2015; Kurakin et al., 2017; Dong et al., 2018). That is, the parameters of the victim model are accessible. Under this assumption, we can leverage the model to conduct inference and obtain the gradient backpropagated to the inputs. To verify the effectiveness of the proposed method, we mainly adopt two kinds of widely-used retrieval models as victim model: *those trained with a classification loss function (*e.g., *cross-entropy loss)* (Babenko et al., 2014; Wei et al., 2016; Qian et al., 2017; Li et al., 2018) and *those trained with a ranking loss (*e.g., *triplet loss)* (Song et al., 2016; Hermans et al., 2017; Zhang et al., 2020). It is worth noting that the classification attack methods only can be applied to the former kind of victim models with class predictions, while our method based on the feature can successfully fool both kinds of victims.

3.3 Classification Attack Methods Revisited

Previous works in the adversarial example generation are designed for image recognition and aim to attack the class prediction (Goodfellow et al., 2015; Kurakin et al., 2017). We assume that the label prediction of the victim model is acquirable, and apply these existing classification attack methods to generate the adversarial queries. Specifically, for the fast-gradient sign method (Goodfellow et al., 2015) and basic iterative method (Kurakin et al., 2017), we deploy the label predicted by the victim model as the pseudo label $y_{max} = \arg\max_y \{p(y|X)\}$. To fool the model, the objective is to decrease the probability $p(y_{max})$. The objective is written as, $\arg\min_{X'} J(X') = log(p(y_{max}|X'))$. For iterative least-likely class method (Kurakin et al., 2017), we calculate the least-likely class $y_{min} = \arg\min_y \{p(y|X)\}$. The attack objective is to increase the probability $p(y_{min})$ so that the input is classified as the least-likely class. The objective is, $\arg\max_{X'} J(X') = log(p(y_{min}|X'))$. When generating adversarial samples, the weight of the victim model is fixed and we only update the input. For the fast-gradient sign method, $X' = X + \epsilon\,\text{sign}(\nabla J(X))$. For the iterative methods, we initialize X' with X: $X'_0 = X$, and then update the adversarial samples T times: $X'_{t+1} = X'_t + \alpha\,\text{sign}(\nabla J(X'_t))$, where α is a relatively small hyper-parameter. Following the practice (Kurakin et al., 2017), we set $\alpha = 1$ and the number of iterations $T = min(\epsilon + 4, 1.25 \times \epsilon)$. The clip function $Clip_{X,\epsilon}\{X'\}$ is also added in every iteration to keep pixels of the adversarial query in the valid range.

It is worth noting that the classification attack methods do not target changing the representation, but make changes to the category prediction p of the query. For instance, a linear classifier function $p = Wf + b$ (note that W and b are fixed). The intermediate feature f is implicitly affected when backpropagating the adversarial objective on p. Therefore, the feature changes of these methods are relatively limited.

4 Proposed Method

In this section, we introduce the proposed opposite-direction feature attack (ODFA) method and focus on changing the extracted feature of the adversarial query, which is aligned with the retrieval scenario. In addition, we extend ODFA to ODFA+EoT (MS), attacking a common evaluation trick, *i.e.,* feature fusion of multi-scale inputs.

4.1 Opposite-Direction Feature Attack (ODFA)

We propose opposite-direction feature attack (ODFA), which compromises retrieval models by directly attacking on the intermediate feature. Specifically, given an original query image X, the retrieval model extracts its feature $f_X = F(X)$. We intend to generate the adversarial query X', whose feature $f_{X'}$ is on the opposite side of the original query feature f_X. We name $-f_X$ as the *opposite-direction feature*, since it has the lowest cosine similarity score -1 with the original feature f_X. During optimization, we explicitly impel the feature $f_{X'}$ of the adversarial image to $-f_X$. Therefore, the opposite-direction feature loss objective is formulated as:

$$\arg\min_{X'} J(X') = (f_{X'} + f_X)^2. \tag{1}$$

When the objective converges, *i.e.,* $J(X') \to 0$, $f_{X'}$ will be close to $-f_X$. Consequently, it changes the ranking similarity

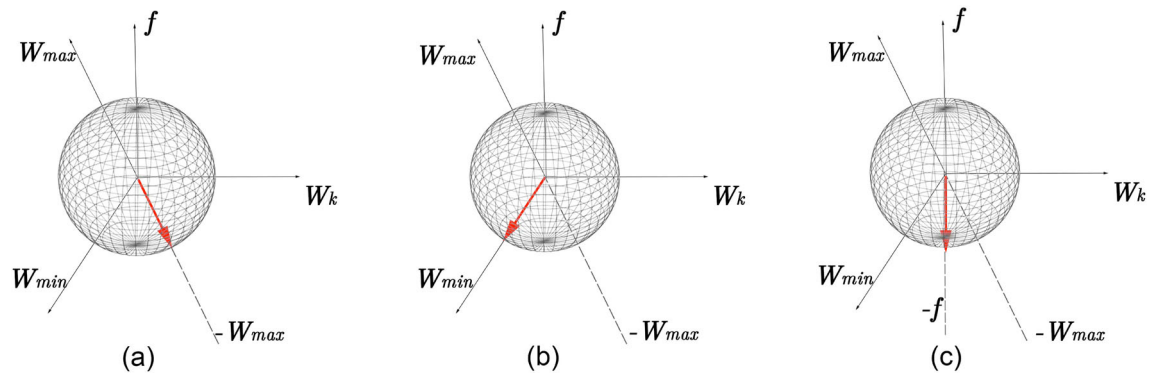

Fig. 1 Geometric interpretation of **a** the fast-gradient sign method (Goodfellow et al., 2015) and the basic iterative method (Kurakin et al., 2017), **b** the iterative least-likely class method (Kurakin et al., 2017), and **c** the proposed ODFA. ***Red*** arrows represent the direction of the gradient on the original feature f. W_{max} denotes the weight of the most-likely class y_{max} and W_{min} denotes the weight of the least-likely class y_{min}. W_k indicates the weight of the other class. Here we apply W_k to help visualize the feature space. While attack methods in **a** and **b** rely on W_{max} (Eq. 3) and W_{min} (Eq. 4), our method avoids such reliance and deploys a straightforward opposite gradient direction $-f$ when attacking the features, like a U-turn

$S(X', X_{gt})$ between the adversarial query and the true match images. If we approximate $f_{X'}$ with $-f_X$, we will derive:

$$S(X', X_{gt}) = f_{X'} \cdot f_{X_{gt}} \rightarrow -f_X \cdot f_{X_{gt}} = -S(X, X_{gt}). \tag{2}$$

Since $S(X, X_{gt})$ is usually high and non-negative in the original retrieval model, we can deduce that the similarity score $S(X', X_{gt})$ is low, leading to a low rank for X_{gt}. Finally, to craft such adversarial query X', we adopt an iterative method to update X': $X'_0 = X$, $X'_{t+1} = X_t + \alpha\,\mathrm{sign}(\nabla J(X'_t))$. The clip function is also added to keep pixels in the adversarial sample within the valid range. The overall process of crafting the adversarial query is present in Algorithm 1.

Discussions. We provide a 2D geometric interpretation to illustrate the difference of the gradient direction between the proposed method and traditional attack methods (see Fig. 1). Without loss of generality, we take a linear classifier $p = Wf + b$ as an example, where W is the learned weight and b is the bias term. The weight $W = \{W_1, W_2, \ldots, W_K\}$ contains K weights for the K classes in the training set. We apply W_{max} to denote the weight of the most-likely class y_{max} and W_{min} to denote the weight of the least-likely class y_{min}. For the fast-gradient sign method and the basic iterative method, the gradient on the feature f is,

$$\frac{\partial J(X')}{\partial f_{X'}} = -W_{max} \times \frac{\partial J(X')}{\partial p(y_{max})}. \tag{3}$$

Note that $\frac{\partial J(X')}{\partial p(y_{max})}$ is a positive constant. Therefore, the direction of the gradient is the direction of $-W_{max}$ (see Fig. 1a). For the iterative least-likely class method, the gradient equals,

$$\frac{\partial J(X')}{\partial f_{X'}} = W_{min} \times \frac{\partial J(X')}{\partial p(y_{min})}. \tag{4}$$

The gradient has the same direction as W_{min} (see Fig. 1b). For the unseen images of new classes, *i.e.,* query images, $-W_{max}$ and W_{min} are not accurate to describe the adversary of the original query, so the adversarial attack effect is limited. In this paper, instead of using class predictions, we directly attack the representation in the feature space. According to Eq. 1, the adversarial gradient of the proposed method is written as,

$$\frac{\partial J(X')}{\partial f_{X'}} = -2 \times (f_{X'} + f_X), \tag{5}$$

where f_X is the feature of the original query image. In Fig. 1c, we draw the gradient direction of the first iteration. In the first iteration, $f_{X'_0} = f_X$, $\frac{\partial J(X'_0)}{\partial f_{X'_0}} = -4 f_X$. Our method leads the feature to the opposite direction of the original feature, so the similarity of true matches drops more quickly. The observation in the experiment, as shown in Figs. 3, 4 and 5, also verifies that the proposed method is more efficient and effective than the conventional methods, and explicitly changes the feature (see Table 6).

4.2 A Multi-scale Extension

Fusing the features of multiple-scale queries is a common practice in many image retrieval systems, such as landmark retrieval (Radenović et al., 2016, 2018). In particular, when testing, the input image is resized with multiple scale factors $\Phi = \{\phi_1, \phi_2, \ldots, \phi_{n_\phi}\}$, and then the model extracts

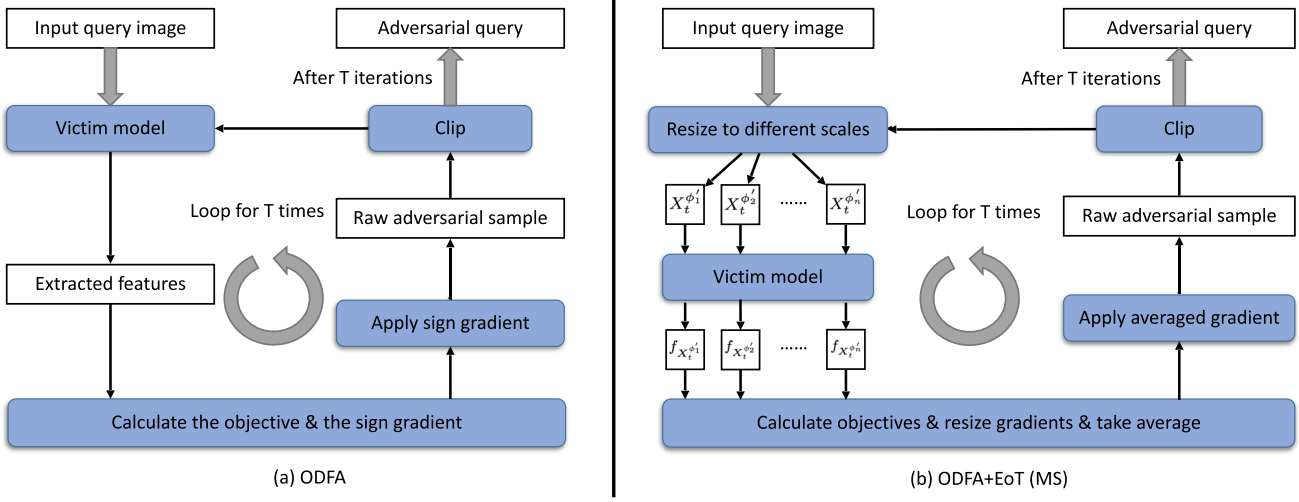

Fig. 2 Pipeline of the proposed method: **a** ODFA and **b** ODFA+EoT (MS). Given an input query image, we first extract features, and then calculate Eq. 6 to obtain gradients. Under the multiple-scale input setting, we calculate Eq. 9 on every scale and take the average gradients. The gradients are then added to the input sample. To endow the input with imperceptible noise to humans, we follow (Kurakin et al., 2017) and clip the image value before each iteration. We update the input image for T iterations to obtain the adversarial query

Algorithm 1 Opposite-direction feature attack (ODFA)

Input: Victim model F; a query image X; perturbation rate ϵ.
Output: An adversarial example X' with $\|X' - X\|_\infty \leq \epsilon$.

$X'_0 = X$;
$T = min(\epsilon + 4, 1.25 \times \epsilon)$;
for $t = 0$ to $T - 1$ **do**
  Input X'_t to F, extract feature f, calculate the objective $J(X'_t)$:

$$J(X'_t) = (f_{X'_t} + f_X)^2. \tag{6}$$

  Update X'_{t+1} by applying the sign gradient as:

$$X'_{t+1} = X'_t + \alpha \, \text{sign}(\nabla J(X'_t)). \tag{7}$$

  Keep pixels of the adversarial query in the valid range:

$$X'_{t+1} = \text{Clip}_{X,\epsilon}\{X'_{t+1}\}. \tag{8}$$

end for
return $X' = X'_T$.

Algorithm 2 Opposite-direction feature attack with multiple-scale inputs, *i.e.,* ODFA+EoT (MS)

Input: Victim model F; a real query image X; Multiple-scale factor Φ; perturbation rate ϵ.
Output: An adversarial query X' with $\|X' - X\|_\infty \leq \epsilon$.

$X'_0 = X$;
$T = min(\epsilon + 4, 1.25 \times \epsilon)$.
for $t = 0$ to $T - 1$ **do**
  Resize the X'_t to $X_t^{\phi'}$ for ϕ in Φ.
  Input $X_t^{\phi'}$ to F, extract features of different scales $f_{X_t^{\phi'}} = F(X_t^{\phi'})$.
  Calculate the objectives $J(X_t^{\phi'})$ as Eq. 6 and gradients of different-scale inputs.
  Resize gradients to the original scale and average the gradients:

$$\nabla J(X'_t) = \frac{1}{n_\phi} \sum_{\phi \in \Phi} \nabla \tilde{J}(X_t^{\phi'}). \tag{9}$$

  Update X'_{t+1} by applying the sign gradient (Eq. 7).
  Keep pixels of the adversarial query in the valid range (Eq. 8).
end for
return $X' = X'_T$.

the feature from inputs of different scales. The mean of the normalized features is adopted as the final retrieval representation. Since the final representation fuses the feature of the multi-scale inputs, the retrieval system is more robust in terms of the scale variants. In the experiment, we observe that only calculating the adversarial gradient upon the input of the original scale is less effective to fool the image retrieval system. It is due to the designed imperceptible perturbation being deprecated when resizing images.

To successfully attack the multiple-scale inputs, we further extend the proposed ODFA to ODFA+EoT (MS). Inspired by Expectation over Transformation (Athalye et al., 2018), we aggregate adversarial gradients of various sizes. For the retrieval scenario, our implementation modifies two primary points, in comparison to the conventional EoT for attacking image recognition models: (1) The EoT (MS) transformation selection is different from the original EoT. The original EoT deploys *random crop* and *random rotation* to approximate the natural 2D transformation and the image size is fixed, *e.g.,* 224×224 for image recognition. Differently, for the image retrieval task, scaling the image is a more typical transformation. ODFA+EoT (MS) changes the image scale to $[1, 0.5^{0.5}, 0.5]$ to simulate the test augmentation. (2) Conse-

quently, the adversarial gradient aggregation is also modified. By back-propagating the mean loss expectation, the conventional EoT can directly derive the average adversarial gradient on the fixed-size input. ODFA+EoT (MS) requires an additional step because of the size changes. Specifically, we calculate individual gradients for every different-size input, and then resize the gradient as the original size to yield the average adversarial gradient. The whole pipeline is summarized in Algorithm 2. We first view the inputs of different scales as independent inputs $X_t^{\phi'}(\phi \in \Phi)$. Similar to the single scale setting, we calculate the adversarial gradient based on each scale $\nabla J(X_t^{\phi'})$. To generate the adversarial gradient towards the original input, we resize all gradients to the original scale $\nabla \tilde{J}(X_t^{\phi'})$ and average the multi-scale adversarial gradients as follows,

$$\nabla J(X_t') = \frac{1}{n_\phi} \sum_{\phi \in \Phi} \nabla \tilde{J}(X_t^{\phi'}), \tag{10}$$

in which n_ϕ is the number of scale factors. Similar to ODFA, we add the sign gradient to the original input and iteratively update the input to obtain the adversarial samples. For a quick comparison, we provide the brief pipeline of ODFA and ODFA+EoT (MS) in Fig. 2. Since we explicitly consider multi-scale adversarial gradients, the ODFA+EoT (MS) significantly outperforms the regular ODFA in terms of multiple-scale evaluation. More details can be found in Sect. 5.3.

5 Experiment

5.1 Datasets and Settings

We evaluate the attack performance on five image retrieval datasets: Food-256, CUB-200-2011, Market-1501, Oxford5k, Paris6k, and an image recognition dataset, *i.e.,* Cifar-10.

Food-256 is a food dataset (Kawano & Yanai, 2014) containing 31,395 images of 256 types of cuisines. Following the train / test split in (Liu et al., 2019a), we deploy 27,849 images of 224 cuisines as the training set and the rest 3,546 images of 32 cuisines as the test set. In the test set, we select 512 images as queries and the rest 3,034 are utilized as the gallery images. There is no overlapping class (food category) between the training and test sets.

CUB-200-2011 consists of 11,788 images of 200 bird species (Wah et al., 2011). Following (Song et al., 2016), we deploy the CUB-200-2011 dataset for fine-grained image retrieval. The first 100 classes (5,864 images) are split as the training set, and we evaluate the model on the other 100 classes (5,924 images), where each image is adopted as the query, and the rest forms the gallery.

Market-1501 is a large-scale pedestrian retrieval dataset (Zheng et al., 2015). This type of retrieval task is also known as person re-identification (re-ID). Images are collected under six different cameras at a university campus. There are 32,668 detected bounding boxes of 1,501 identities. Following the standard train / test split, we adopt 12,936 images of 751 identities as the training set and 19,732 images of the other 750 identities as the test set, where 3,368 images are set as queries. There are no overlapping classes (identities) between the training and test sets.

Oxford5k & Paris6k are two widely used landmark retrieval datasets. Oxford5k contains 5,062 images of 11 particular Oxford buildings (Philbin et al., 2007), and Paris6k contains 6,412 images of 12 Paris landmarks (Philbin et al., 2008). Both datasets are only used as the test sets. Following (Radenović et al., 2018), we deploy the non-overlapping building images collected from Flickr as the training set, which contains about 133k images.

Cifar-10 is a widely-used image recognition dataset, containing 60,000 images of 10 classes (Krizhevsky & Hinton, 2009). There are 50,000 training images and 10,000 test images. On this dataset, we compare our method to other classification attack approaches to further show the mechanism difference.

Evaluation metric. With the limited image perturbation, we compare the methods by the drop of accuracy. The lower accuracy indicates that the adversarial examples make more true matches receive low ranks. For image retrieval, we utilize two evaluation metrics, *i.e.,* Recall@K and mean average precision (mAP), which are from the original paper of the retrieval set (Zheng et al., 2015; Philbin et al., 2007, 2008). In this way, we can fairly compare the retrieval performance before attacking. **Recall@K** is the probability that the right match appears in the top-K of the ranking list. Given a ranking list, the average precision (AP) calculates the space under the recall-precision curve. **mAP** is the mean of the average precision of all queries. Besides, we adopt **Attacking Success rate (ASR)**, which is proposed in (Li et al., 2021). If the top-10 ranking list does not contain any relevant images, the attack is successful with a score of 1, otherwise, we obtain a score of 0. The final attacking success rate is averaged over all queries. For image recognition, we report Top-1 and Top-5 accuracy. **Top-K** is the mean probability that the right class appears in the top-K predicted classes.

Implementation details. For the retrieval victim model trained with a classification loss, we follow the common practice in (Hermans et al., 2017; Chen et al., 2017) to fine-tune the ResNet-50 (He et al., 2016) by class prediction on Food-256, CUB-200-2011 and Market-1501. During training, the cuisine images in Food-256 are resized to 256×256, while the pedestrian images of Market-1501 are resized to 256×128 following the previous practices (Sun et al., 2018; Zhong et al., 2020). The images in CUB-200-2011 are first

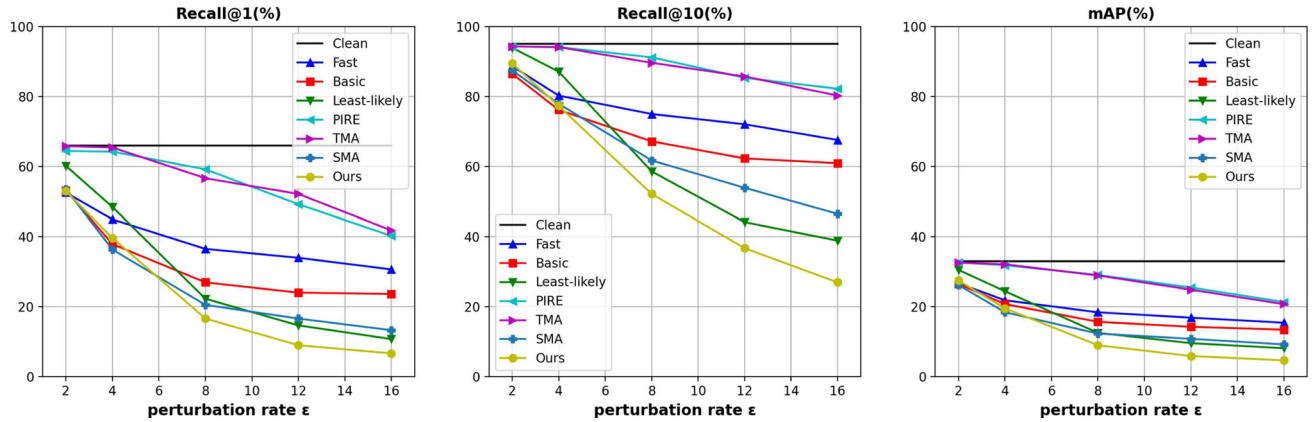

Fig. 3 Comparing attack methods under various perturbation rates ϵ on Food-256. We report Recall@1 (%), Recall@10 (%) and mAP (%) of the victim model. "Clean" means using the original query without attack, where the victim model yields Recall@1 = 66.02%, Recall@10 = 95.12% and mAP = 32.95%. We report the highest attack success rates (lowest retrieval accuracy)

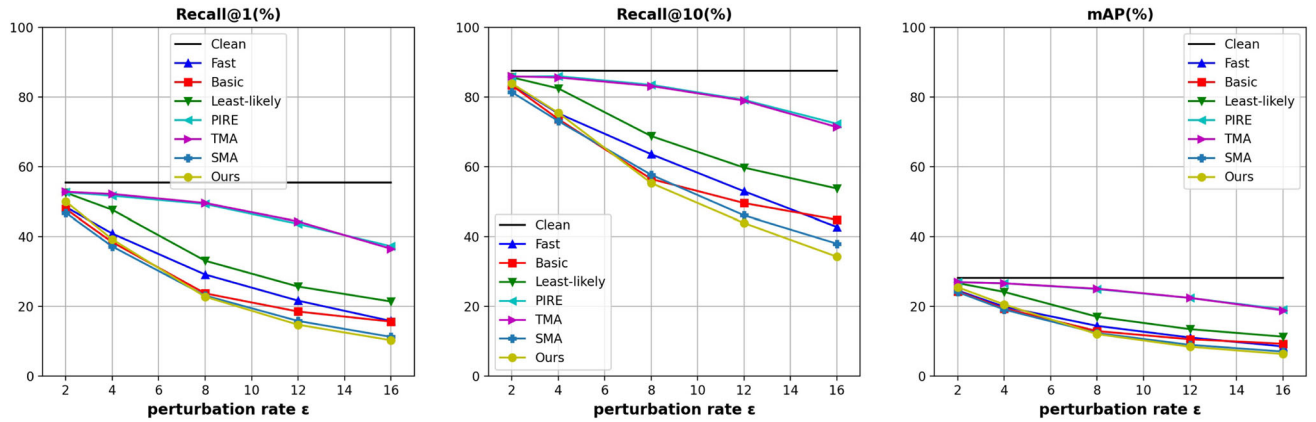

Fig. 4 Comparing attack methods under various perturbation rates ϵ on CUB-200-2011. All settings are the same with Fig. 3. The victim model using clean queries yields Recall@1 = 55.47%, Recall@10 = 87.49% and mAP = 28.18%. Our method is very competitive

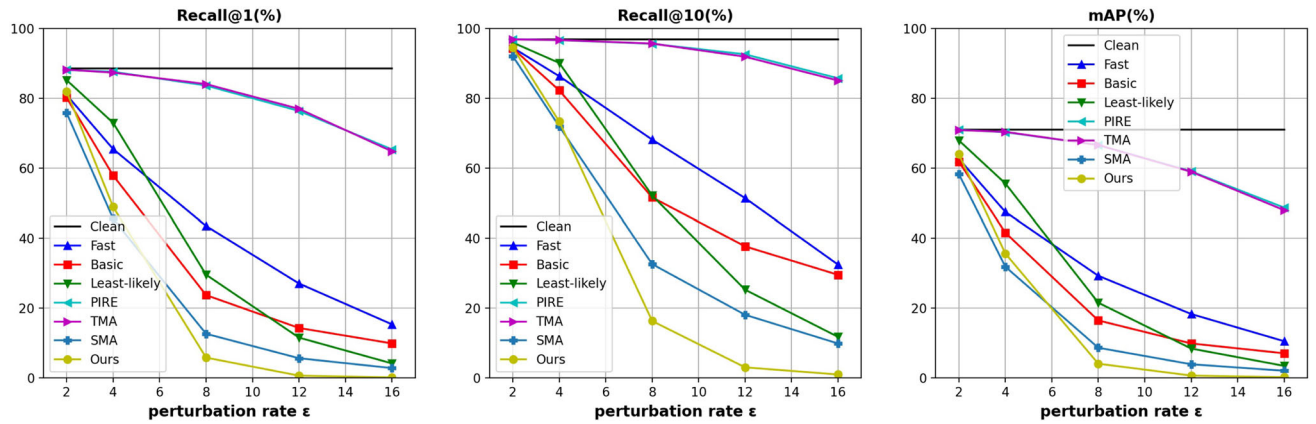

Fig. 5 Comparing attack methods under various perturbation rates ϵ on Market-1501. All settings are the same with Fig. 3. The victim model using clean queries yields Recall@1 = 88.54%, Recall@10 = 96.85% and mAP = 71.08%. Our method is shown to be very effective

Table 2 Comparing methods attacking retrieval models trained by the classification loss (cross-entropy loss). Three datasets are used: Food-256, CUB-200-2011 and Market-1501. Here we show the results in percentage. Perturbation rate is fixed to $\epsilon = 16$. We compare the three existing attack methods, *i.e.,* Fast (Goodfellow et al., 2015), Basic (Kurakin et al., 2017), Least-likely (Kurakin et al., 2017). If not specified, the learning rate of PIRE and TMA is set to 0.1. * We directly borrow the reported results from the corresponding paper considering a similar victim model. RS denotes random start (Madry et al., 2018)

Methods	#Iters	Food-256			CUB-200-2011			Market-1501		
		Recall@1 ↓	mAP ↓	ASR ↑	Recall@1 ↓	mAP ↓	ASR ↑	Recall@1 ↓	mAP ↓	ASR ↑
Victim	–	66.02	32.95	–	55.47	28.18	–	88.54	71.08	–
UPA (Zhao et al., 2020)	Extra training	53.32	26.38	12.30	52.94	27.07	13.81	69.98	53.90	12.11
AP-GAN (Zhao et al., 2020)	Extra training	–	–	–	–	–	–	15.60*	11.70*	64.90*
PIRE (Liu et al., 2019b)	20	40.23	21.34	17.77	37.22	19.18	27.46	65.38	48.70	14.25
PIRE+RS (Liu et al., 2019b)	20	35.55	19.46	26.17	31.74	16.64	34.89	54.54	40.62	22.09
PIRE (Liu et al., 2019b) (lr=1e-4)	100	61.91	30.99	6.05	49.16	25.16	16.93	86.58	69.50	3.33
PIRE (Liu et al., 2019b)	100	27.93	16.07	33.98	15.34	8.71	55.27	16.33	11.48	61.88
TMA (Tolias et al., 2019)	20	41.80	20.73	19.73	36.88	18.92	28.26	64.82	47.96	14.91
TMA+RS (Tolias et al., 2019)	20	35.16	18.91	22.85	30.52	16.41	35.40	54.04	39.66	23.55
TMA (Tolias et al., 2019) (lr=1e-4)	100	62.70	31.01	6.84	49.34	25.15	16.88	86.82	69.48	3.44
TMA (Tolias et al., 2019)	100	13.87	8.47	57.42	20.17	10.90	48.24	2.26	1.90	91.63
Fast (Goodfellow et al., 2015)	1	30.66	15.41	32.42	15.80	8.50	57.24	15.35	10.51	67.55
Basic (Kurakin et al., 2017)	20	23.63	13.43	39.06	15.63	9.27	55.17	9.89	7.06	70.46
Basic+RS (PGD) (Madry et al., 2018)	20	17.77	11.68	43.95	14.11	8.69	57.95	9.68	6.91	72.71
SMA (Bouniot et al., 2020)	20	13.28	9.21	53.52	11.24	7.03	62.02	2.82	2.05	90.14
SMA+RS (Bouniot et al., 2020)	20	12.30	8.61	61.52	8.59	5.87	68.70	1.96	1.66	92.46
Least-likely (Kurakin et al., 2017)	20	10.74	8.12	61.13	21.40	11.31	46.25	4.16	3.46	88.24
Least-likely+RS (Kurakin et al., 2017)	20	10.55	7.70	65.43	19.29	10.50	50.47	4.04	3.25	88.09
ODFA	20	**6.64**	**4.65**	**73.05**	10.28	6.34	65.75	**0.15**	**0.22**	99.02
ODFA+RS	20	**6.64**	4.89	71.09	**7.63**	**5.15**	**72.01**	0.24	0.23	**99.35**

Bold values indicate the best attacking performance among all compared methods

resized with their shorter side to 256, and we then apply a 256×256 random crop to the images. The learning rate is 0.01 for the first 40 epochs and decays to 0.001 for the last 20 epochs. For the retrieval victim model trained with a ranking loss, we follow the setting in (Radenović et al., 2018) to train ResNet-101 (He et al., 2016) on the collected building dataset (Radenović et al., 2018) with contrastive loss. For image recognition, our implementation employs ResNet with 20 layers for the Cifar-10 dataset (He et al., 2016). The size of the input image is 32×32. The training policy follows the practice in (He et al., 2016). The learning rate starts from 0.1 and is divided by 10 after the 150th and 225th epoch. We stop training after 300 epochs.

Besides, we re-implement several adversarial attack approaches. **(1)** We apply a large learning rate of 0.1 for both PIRE (Liu et al., 2019b) and TMA (Tolias et al., 2019), because we find a small learning rate (*e.g.,* 1e-4) causes them to converge very slowly as shown in Table 2. If not specified, the learning rate of PIRE and TMA is set to 0.1. For a fair comparison, we re-implement and report their performance under both 20 iterations and 100 iterations. **(2)** TMA focuses on the target adversarial attacking, while we focus on the non-target adversarial attacking. Therefore, to compare the proposed method with TMA, we have further modified the "global descriptor" Eq. 8 in TMA as $1\text{-}f_X \cdot f_{X'}$ to obtain the adversarial query. (f_X is the original feature, and $f_{X'}$ is the adversarial feature.) **(3)** SMA (Bouniot et al., 2020) is similar to PIRE and TMA in increasing the feature distance. Differently, SMA does not require the Adam optimizer, but adopts a similar updating strategy (sign gradient) as our method. For a fair comparison, we re-implement SMA (Bouniot et al., 2020) with $\epsilon = \{2, 4, 8, 12, 16\}$ which is the same hyper-parameter setting as our method, and gives SMA very good performance. **(4)** Given the victim model (ResNet-50) in AP-GAN (Zhao et al., 2020) is trained from the same open-source code as ours, we directly quote the number from the paper. **(5)** UPA (Li et al., 2019a) trains a unified retrieval perturbation on a large-scale dataset SfM (Schonberger et al., 2015). We apply the trained unified perturbation provided by the UPA authors to the three datasets. In practice, we further re-scale the perturbation from $\epsilon = 10$ to $\epsilon = 16$ for a fair comparison.

Reproducibility. Our source code and results are made publicly available.[2] The implementation is based on the Pytorch package (Paszke et al., 2017).

5.2 ODFA on Classification-based Retrieval Models

We first compare the attack performance of ODFA with existing classification-based attack methods on the victim retrieval model with class predictions. Quantitative results, *i.e.,* Recall@1, Recall@10 and mAP, on Food-256 using clean and adversarial queries are summarized in Fig. 3. The victim model using clean queries gives Recall@1 = 66.02% and mAP = 32.95%. As mentioned, the classification attack methods change the semantic prediction, which only *implicitly* changes the retrieval features, so would not be very effective in attacking. Although Recall@10 decreases with increasing ϵ, the best method, the iterative least-likely class method, gives a Recall@10 of 38.87%. In comparison, the proposed ODFA achieves a lower Recall@1 and Recall@10 at $\epsilon = 8$. This can be attributed to the *explicit* opposite direction attack mechanism on the feature. As we increase the perturbation rate ϵ to 16, the victim model attacked by ODFA yields Recall@1 = 6.64%, Recall@10 = 26.95%, mAP = 4.65%, which are lower than the traditional classification attack methods, *i.e.,* Fast (Goodfellow et al., 2015), Basic (Kurakin et al., 2017) and Least-likely (Kurakin et al., 2017). This is also consistently more effective compared with re-implemented feature-based approaches, *e.g.,* PIRE (Liu et al., 2019b), TMA (Tolias et al., 2019) and SMA (Bouniot et al., 2020). Experiments on the fine-grained retrieval dataset, *i.e.,* CUB-200-2011, and person re-id dataset, *i.e.,* Market-1501, indicate similar observations (see Figs. 4 and 5). More quantitative results are shown in Table 2.

Discussion. Here we discuss and compare the difference between ODFA and six recent methods, *i.e.,* PIRE (Liu et al., 2019b), TMA (Tolias et al., 2019), SMA (Bouniot et al., 2020), PGD (Madry et al., 2018), AP-GAN (Zhao et al., 2020) and UPA (Li et al., 2019a). First, similar to our method, PIRE (Liu et al., 2019b) enlarges the distance between the original input and adversarial query in the feature space, but PIRE does not explicitly introduce the "opposite direction" for optimization. Instead, PIRE initializes the random perturbation and leverages the Adam optimizer to **greedily search** the input space. Therefore, PIRE, in practice, usually sets more iterations, like 100 iterations. As shown in Table 2, PIRE is inferior to the proposed method under both 20 iterations and 100 iterations. Similarly, we observe that TMA (Tolias et al., 2019) is superior to PIRE due to the normalized feature but is still neither more efficient nor effective when compared with the proposed method.

Second, SMA(Bouniot et al., 2020) does not harness direction guidance in their pushing strategy, while ODFA pulls the adversarial sample to the opposite direction. Moreover, SMA needs one extra network forward for the jittering input in the first iteration; otherwise, the initial loss will be zero. In comparison, ODFA does not require jittering at the beginning and costs less running time as shown in Sect. 5.4. We observe that ODFA consistently performs better than SMA.

Third, we evaluate the effectiveness of random start (RS). We re-implement PGD (Madry et al., 2018) following *clever-*

[2] https://github.com/layumi/U_turn

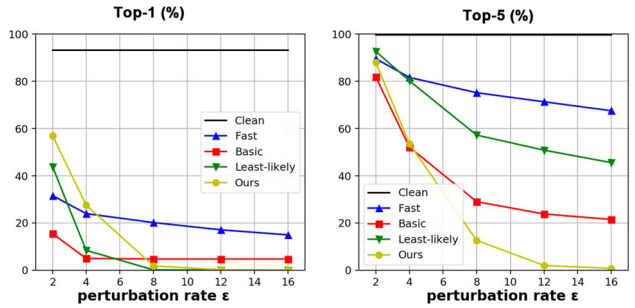

Fig. 6 Comparing attack methods on the Cifar-10 dataset under various perturbation rates ϵ. Top-1 (%) and Top-5 (%) accuracy scores of the victim model (ResNet-20 backbone trained with cross-entropy loss) are shown. "Clean" denotes the accuracy using original query images

hans[3], which replaces the zero start with RS for Basic Iterative Method. We further apply such random start to all comparable iterative methods, including PIRE, TMA, SMA, Least-likely and the proposed ODFA in Table 2, since random start is a frequently used add-on. The experiments show that RS usually increases the attacking success rate of the iterative method, since it provides a large perturbation at the beginning. Despite the performance improvement of counterparts with RS, ODFA still outperforms other existing methods.

Finally, while AP-GAN (Zhao et al., 2020) applies additional training for a generative model to synthesize adversarial noise, our method is still better, as shown on the Market-1501 dataset. In addition, we observe that applying the unified perturbation learned in UPA (Li et al., 2019a) has worse attack performance than ours on all three datasets.

5.3 ODFA on Ranking-based Retrieval Models

Ranking-based retrieval models optimize the feature distances and usually do not have a class prediction. The classification attack methods, which attack category prediction, thus does not work on these retrieval models. In comparison, the proposed method can still be applied. Here, we compare ODFA with attack methods that do not depend on class predictions. Results are presented in Table 3, where the victim model is borrowed from (Radenović et al., 2018). The victim model using clean queries arrives at a high performance: Recall@1 = 100.00%, mAP = 86.24% on Oxford5k and Recall@1 = 100.00%, mAP = 90.66% on Paris6k. When $\epsilon = 16$, the proposed method successfully fools the victim model, where retrieval accuracy drops to Recall@1 = 0.00%, mAP = 0.80%, ASR=99.45% on Oxford5k and Recall@1 = 1.82%, mAP = 3.00%, ASR=96.91% on Paris6k, respectively. ODFA also surpasses other competing methods, *i.e.,* UPA (Li et al., 2019a), AP-GAN (Zhao et al., 2020), PIRE (Liu et al., 2019b) and TMA (Tolias et al., 2019). To ensure a fair comparison, we re-run ODFA with the setting in UPA, *i.e.,* $\epsilon = 10$, and we observe that ODFA is still superior to UPA under its setting. Besides, it is worth noting that the victim model (Radenović et al., 2018) utilizes Euclidean distance instead of cosine similarity to rank the gallery images. Therefore, TMA is identical to PIRE because both maximizing Euclidean distance. In the single-scale setting, PIRE has competitive performance with the proposed method but demands 100 iterations, introducing five times running cost.

Attacking multi-scale queries. We evaluate ODFA on attacking multiple-scale inputs. Following (Radenović et al., 2018), we extract and fuse the features of multiple-scale inputs, which leads to robust representations against scale variations and slightly improves the performance of the victim retrieval. Specifically, with multi-scale queries, the victim model gives Recall@1 = 100.00%, mAP = 88.17% on Oxford5k and Recall@1 = 100.00%, mAP = 92.52% on Paris6k. We observe that the imperceptible noise generated by ODFA is somehow deprecated after resizing the image, and the attack performance is compromised (see Table 3). In comparison, the proposed ODFA+EoT (MS) method, benefiting from considering the multiple-scale adversarial gradients, successfully fools the victim model. As a result, the performance of the victim model with multi-scale inputs drops significantly: Recall@1 = 1.82%, mAP = 2.21%, ASR = 98.55% on Oxford5k and Recall@1 = 3.64%, mAP = 4.77%, ASR = 96.73% on Paris6k. ODFA+EoT (MS) surpasses PIRE / TMA and the basic ODFA by a large margin.

5.4 Further Analysis and Discussions

Performance of ODFA in image classification. We further evaluate ODFA in the image recognition task. Results on Cifar-10 are shown in Fig. 6. We find ODFA does not achieve the largest top-1 accuracy drop when ϵ is small. This can be explained by the fact that image classification directly relies on the classification head. The most competitive iterative least-likely class method aims to make the model misclassify the adversarial example into the least-likely class. In comparison, our method does not explicitly increase the probability of a specific class while usually serving to decrease the confidence score of the correct class.

Interestingly, for top-5 classification accuracy, the proposed method converges to a lower point than all other three methods. While the basic iterative method and the fast gradient sign method also directly work on decreasing the confidence score of the ground-truth, they are less effective than ours. It is because of the explicit feature direction change from f to $-f$. To briefly explain this idea, assume that the value of the bias term b for 10 classes is close, so we overlook the influence of b in $p = Wf + b$. The top-1 prediction $p = Wf$ becomes the lowest probability $p' = -Wf$, as our method converges. Therefore, the correct class is more likely

[3] https://github.com/cleverhans-lab/cleverhans

Table 3 Comparing attack methods on retrieval models trained with the ranking loss. We adopt Oxford5k and Paris6k datasets, with and without multiple-scale (MS) queries. Numbers are in percentage. Perturbation rate is fixed to $\epsilon = 16$. * We directly quote the reported results from the corresponding paper

Methods	#Iters	Oxford5k			Paris6k		
		Recall@1 ↓	mAP ↓	ASR ↑	Recall@1 ↓	mAP ↓	ASR ↑
Victim (Single-scale)	–	100.00	86.24	–	100.00	90.66	–
PIRE (Liu et al., 2019b) / TMA (Tolias et al., 2019)	20	72.73	45.12	43.04	89.09	63.25	14.18
PIRE (Liu et al., 2019b) / TMA (Tolias et al., 2019)	100	0.00	0.75	99.82	1.82	2.97	98.18
UPA (Li et al., 2019a)* ($\epsilon = 10$)	Extra training	–	31.73	–	–	32.07	–
AP-GAN (Zhao et al., 2020)*	Extra training	29.00	27.60	–	25.40	29.50	–
ODFA ($\epsilon = 10$)	12	0.00	1.25	99.27	1.82	4.79	93.82
ODFA	20	0.00	0.80	99.45	1.82	3.00	96.91
Victim (Multiple-scale)	–	100.00	88.17	–	100.00	92.52	–
PIRE (Liu et al., 2019b) / TMA (Tolias et al., 2019)	20	98.18	84.93	7.50	100.00	91.33	0.36
PIRE (Liu et al., 2019b) / TMA (Tolias et al., 2019)	100	100.00	79.75	10.59	100.00	88.89	0.36
ODFA	20	94.55	74.44	20.23	100.00	88.28	0.18
ODFA+EoT (MS)	20	**1.82**	**2.21**	**98.55**	**3.64**	**4.77**	**96.73**

Bold values indicate the best attacking performance among all compared methods

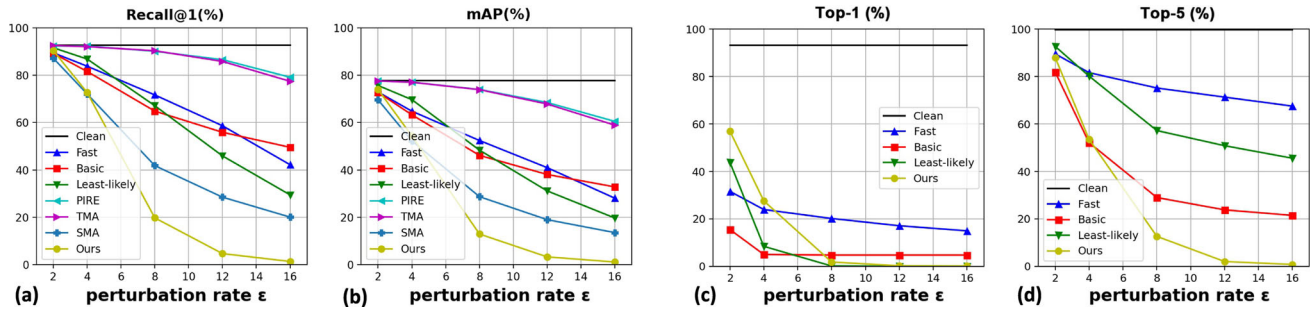

Fig. 7 Performance of attacking stronger victim models. **a** and **b**: Recall@1 (%) and mAP(%) on Market-1501 when attacking the victim model *PCB* (Sun et al., 2018). **c** and **d**: Top-1 and Top-5 accuracy (%) on Cifar-10 when attacking WideResNet-28 (Zagoruyko & Komodakis, 2016)

Table 4 Performance of attacking the image recognition model on Cifar-10. Here we show the top-1 and top-5 accuracy in % (Lower is better). The perturbation rate is fixed to $\epsilon = 16$. We compare the three classification attack methods, *i.e.,* Fast (Goodfellow et al., 2015), Basic (Kurakin et al., 2017), Least-likely (Kurakin et al., 2017)

Methods	Cifar-10	
	Top-1	Top-5
Victim	93.14	99.76
Fast	14.95	67.55
Basic	4.74	21.47
Least-likely	**0.03**	45.58
ODFA	0.06	**0.76**

Bold values indicate the best attacking performance among all compared methods

Table 5 Effectiveness of ODFA in the black-box setting on Market-1501. The adversarial queries are independently generated by the white-box ResNet-50 ($\epsilon = 16$) to cheat the unseen DenseNet-121 (Huang et al. , 2017). More details are provided in the Sect. 5.4

Methods	#Iters	Market-1501		
		Recall@1 ↓	mAP ↓	ASR ↑
Victim (DenseNet-121 (Huang et al. , 2017))	–	90.17	75.60	–
UPA (Li et al., 2019a)	Extra	81.21	64.83	5.46
PIRE (Liu et al., 2019b)	20	86.13	69.94	3.71
TMA (Tolias et al., 2019)	20	85.84	69.75	3.92
PIRE (Liu et al., 2019b)	100	71.08	54.04	9.62
TMA (Tolias et al., 2019)	100	68.26	51.06	11.79
Basic (Kurakin et al., 2017)	20	69.80	53.01	9.71
Least-likely (Kurakin et al., 2017)	20	69.33	52.12	10.84
SMA (Bouniot et al., 2020)	20	62.08	45.10	15.59
ODFA	20	**52.88**	**37.77**	**21.62**

Bold values indicate the best attacking performance among all compared methods

to instantly move out of the top-5 predicted classes. When $\epsilon = 16$, the adversarial images generated by our method compromise the top-5 accuracy from 99.76% to 0.76%. The attacked top-1 accuracy of 0.06% is also competitive to the result of the iterative least-likely class method 0.03%. In summary, the proposed ODFA method reports competitive performance and is not evidently superior to the competing methods as the case in image retrieval (see Table 4).

Effectiveness of ODFA in the black-box setting. As shown in previous works (Szegedy et al., 2014; Papernot et al., 2016b,a, 2017; Liu et al., 2017b; Moosavidezfooli et al., 2017), adversarial examples have good transferability that can successfully attack other black-box models in the recognition scenario, because the models learn a similar decision boundary in the classification space. In this section, we study the transferability of the adversarial queries in terms of the retrieval scenario. For the classification-based retrieval model, we train a stronger victim with *DenseNet-121* (Huang et al. , 2017) as the black-box model, which arrives at Recall@1 = 90.17% and mAP = 75.60% using "clean" images on Market-1501. The adversarial queries are independently generated by the white-box *ResNet-50* ($\epsilon = 16$). The experiment shows that adversarial samples generated by *ResNet-50* also compromise the performance

Fig. 8 Ranking results of the original queries and the adversarial queries generated by Least-likely (Kurakin et al., 2017), SMA (Bouniot et al., 2020) and our method. The proposed approach introduces trivial noise on original queries to fool the retrieval system, while the human is robust to such noise. Three original queries are from Food-256 (Kawano & Yanai, 2014), CUB-200-2011 (Wah et al., 2011) and Market-1501 (Zheng et al., 2015), respectively. The corresponding top-10 retrieval results are also provided. The proposed adversarial queries successfully fool the retrieval model to predict irrelevant ranking results. We could observe that Least-likely (Kurakin et al., 2017) is prone to return one specific wrong class, and SMA (Bouniot et al., 2020) is still prone to return visually similar objects / humans. In contrast, the proposed method prefers to return more noisy ranking results. The perturbation rate is fixed to $\epsilon = 16$. (Best viewed when zoomed in)

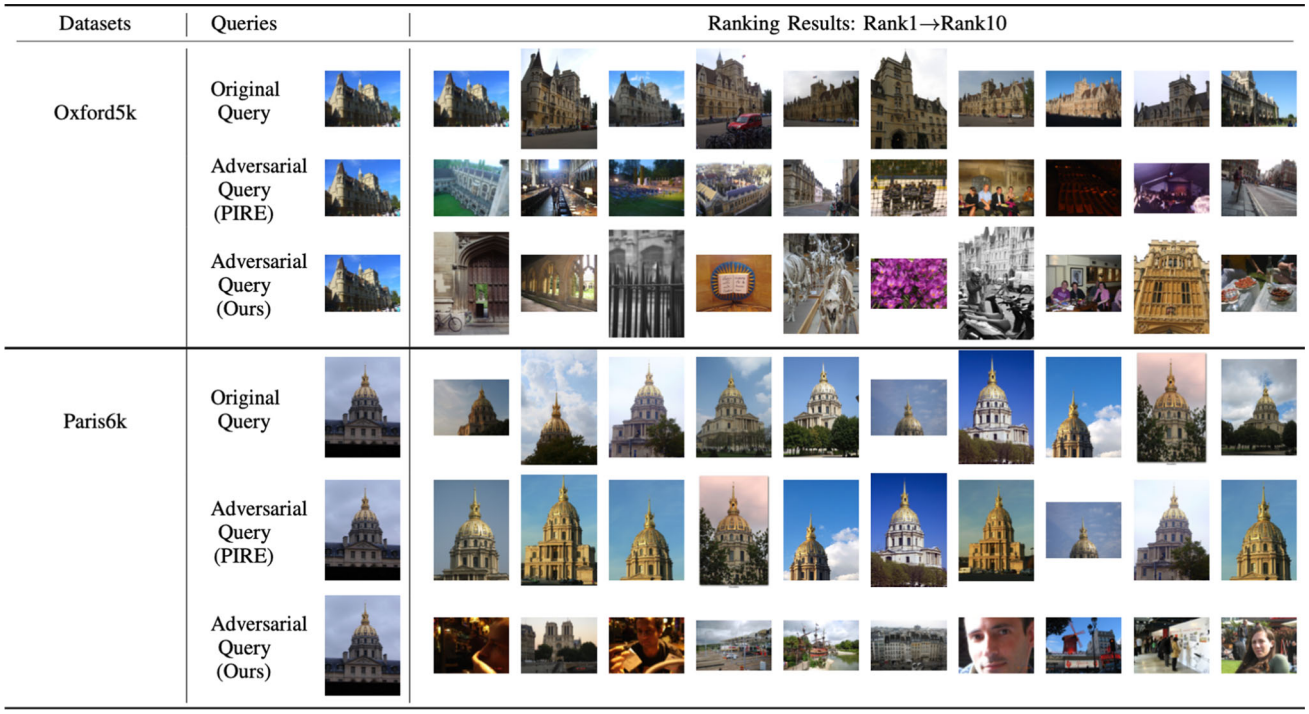

Fig. 9 Retrieval results of original queries and adversarial queries generated by PIRE (Liu et al., 2019b) and our method on Oxford5k and Paris6k. The query images (original and adversarial) are also shown on the left. The proposed adversarial queries successfully fool the retrieval model into giving irrelevant ranking results. For the example on Paris6k, we find that PIRE (Kurakin et al., 2017) fails to cheat the retrieval system, while the proposed method returns noisy results. The perturbation rate is fixed to $\epsilon = 16$. (Best viewed when zoomed in)

of *DenseNet-121*: Recall@1 = 52.88%, mAP = 37.77% and ASR = 21.62%, which surpasses other methods in Table 5. We observe a similar phenomenon on attacking the ranking-based retrieval model. We train the white-box model with *ResNet-101* and apply the *ResNet-101* generated adversarial query to attack the black-box model based on *VGG-16* (Simonyan & Zisserman, 2015). The generation process of the adversarial queries is totally independent with the black-box model. The accuracy of the black-box model also drops from 100.00% to 0.00% Recall@1 and 85.24% to 0.79% mAP on the Oxford5k dataset. It verifies that the adversarial queries have good transferability and could also be applied to the black-box setting.

Attacking stronger victims. Furthermore, we evaluate our method on stronger victim models, which achieve competitive accuracy on benchmarks. Specifically, for person retrieval (image retrieval), we attack a widely adopted model called *PCB* (Sun et al., 2018). On Market-1501, our re-implementation arrives at Recall@1 = 92.70%, mAP = 77.78% using clean queries for the victim model. As shown in Fig. 7a, b, Recall@1 and mAP drop to 1.34% and 1.11% respectively by the proposed ODFA. The second best method, SMA, also arrives at a relatively low accuracy 20.07% and 13.53%, but is still inferior to the proposed method in terms of the accuracy drop. For image recognition, we evaluate our method on the prevailing WideResNet-28 (Zagoruyko & Komodakis, 2016). Our re-implementation arrives at Top-1 accuracy 96.14% and Top-5 accuracy 99.91% using clean queries, respectively. As shown in Fig. 7c, d, we have consistent observations with the baseline victim models, *i.e.,* competitive top-1 accuracy drop and the largest top-5 accuracy drop. Our method arrives at Top-1 accuracy of 0.34% and Top-5 accuracy of 1.29%.

Visualization of retrieval results. We provide one qualitative comparison of the retrieval results with original queries and adversarial queries in Figs. 8 and 9. The attack rate ϵ is fixed to 16. Since we employ an iterative policy with small steps, the adversarial queries generated by our method are visually close to the original query, which simulates extreme retrieval cases to evaluate the model robustness. In these examples, the ranking results obtained by the original queries are good. In contrast, when using adversarial queries, the top-10 ranked images are all false matches with a significantly different appearance to the query. The adversarial query successfully makes the victim model predict low ranks to the true matches.

Visualization of attacked features. Following the visualization trick in (Liu et al., 2016), we train the victim model with an extra 2-dim fully-connected layer on Cifar-10 and then extract the 2-dim feature of every test image to plot maps.

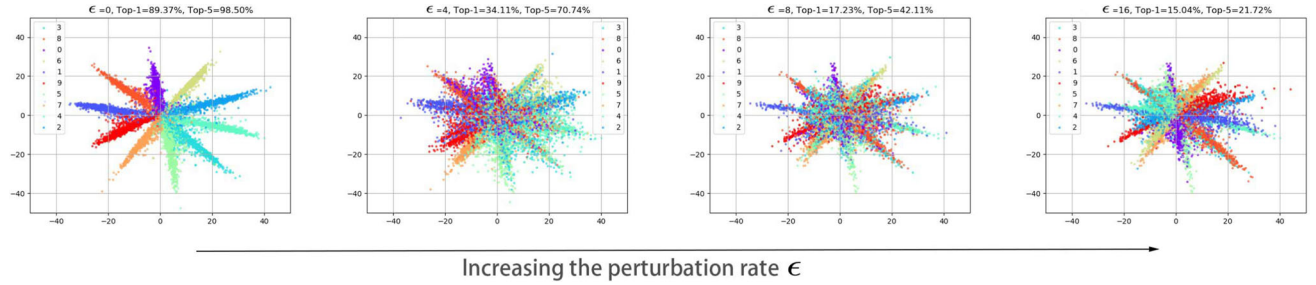

Fig. 10 Feature visualization on Cifar-10 (best viewed in color). As the perturbation rate ϵ increases, the feature gradually moves to the opposite side of the original direction. Top-1 (%) and Top-5 (%) accuracy of the victim model (in the title of each subfigure) also decrease

Due to applying the visualization trick (using the 2-dim feature to classify 10 classes), the accuracy of the new victim model is a little bit lower than the baseline result in Table 4, but still arrives at relatively high accuracy, Top-1=89.37%, and Top-5=98.50%. It is good enough to verify our intuition in the feature space. As shown in Fig. 10, the points in the same color belong to the same class. We plot four maps with different perturbation rates $\epsilon = 0, 4, 8, 16$ to see the feature movement. $\epsilon = 0$ is the output of the victim model on clean test images. The features gradually move to the opposite side of the original direction, when ϵ increases. The observation verifies the effectiveness of our objective, *i.e.,* moving to the opposite direction. Comparing $\epsilon = 0$ with $\epsilon = 16$, the feature of most adversarial examples successfully moves to the opposite side of the original feature. Due to the change in intermediate features, the classification accuracy in the title of every subfigure, also gradually drops. The observation validates the mechanism of the proposed method.

Besides, we add one visualization on clustering the original features and attacked features on the Market-1501 dataset (see Fig. 11). The attack rate $\epsilon = 16$. Specifically, we adopt the unsupervised clustering method, t-SNE (Van der Maaten & Hinton, 2008) to map the feature to 2-dimension for plotting. The same color denotes the same query. The circle denotes the original query, while stars are adversarial queries. We leverage all 3,368 queries in Market-1501 and the 3,368 adversarial queries. Therefore, 6,736 features are utilized to learn reliable clustering results. Due to limited color for visualization, we can not plot all sample points in one figure. Therefore, we select the first 20 features (10 original queries and 10 adversarial queries) to visualize the clustering result in the feature space. We can observe that the proposed method successfully pushes away the distance of the original feature and the feature of adversarial queries.

OFDA significantly decreases the similarity between the original query and the attacked query. We calculate the feature similarity between the original query and the adversarial query on Market-1501. The mean query-query similarity reflects the degree of feature changes. From Table 6, we have two observations. First, ODFA motivates

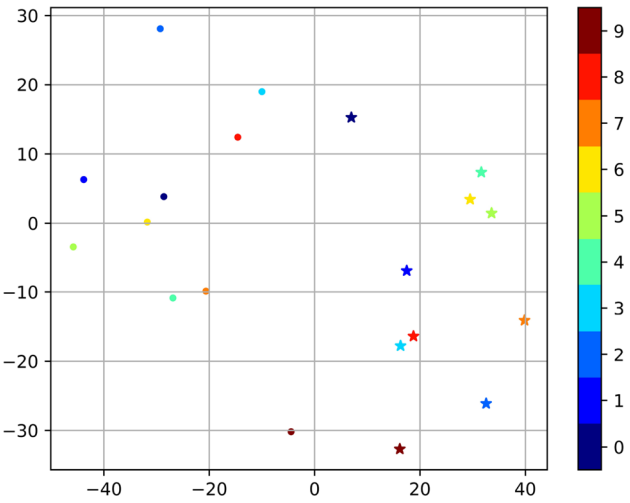

Fig. 11 Visualization result of the original features and attacked features in the latent feature space on the Market-1501 dataset. The same color denotes the same query. The circle denotes the original query, while the stars are the adversarial query. Due to limited color for visualization, we can not plot all sample points in one figure. Therefore, we select the first 20 features (10 original queries and 10 adversarial queries) to visualize the clustering result in the feature space. We can observe that the proposed method successfully pushes away the distance of the original feature (circles) and the feature of the adversarial queries (stars) of the same query (the same color). The perturbation rate ϵ is 16

the adversarial feature to the opposite direction, yielding very low similarity scores. Second, since the classification-based attack methods do not directly work on the feature, their attacked queries have higher similarity with the original query, indicating that these attack methods are less effective.

Visualization of more adversarial queries. We show some adversarial images generated by different attack methods on Market-1501 and Cifar-10 (see Figs. 12 and 13). Attack rate ϵ is set to 16. As discussed in some previous works (Kurakin et al., 2017), the fast-gradient sign method (Goodfellow et al., 2015) may introduce visible artifacts and make perturbation perceptible to the human. In comparison, the proposed method and three other adversarial methods (Kurakin et al., 2017; Bouniot et al., 2020) iteratively update

Table 6 Mean query-query similarity between the original query and the adversarial query on Market-1501 by different attack methods. It verifies our intuition that the three classification-based attack methods, *i.e.,* Fast, Basic and Least-likely, only have an indirect impact on the feature, yielding limited feature changes. Besides, we observe that our method, compared with other competitive approaches, arrives at a much lower similarity with the original feature. Here we fix $\epsilon = 16$

Fast	Basic	Least	PIRE	TMA	UPA	SMA	Ours
0.3801	0.3870	0.2061	0.3227	0.1297	0.8129	0.2127	**−0.1494**

Bold value indicates the best attacking performance among all compared methods

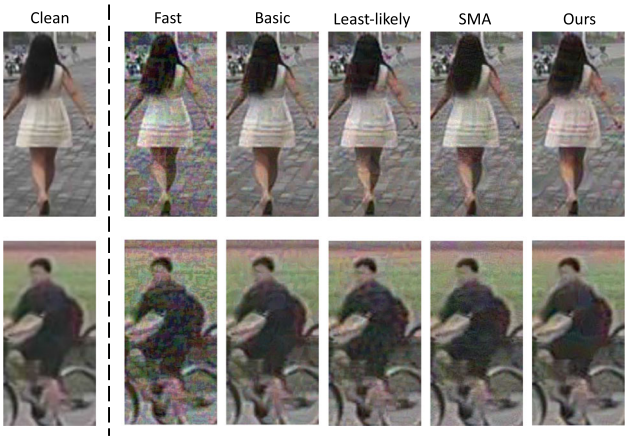

Fig. 12 Adversarial samples on Market-1501 (Zheng et al., 2015) generated by various methods. The perturbation rate is fixed to $\epsilon = 16$. (Zoom in for better visualization.)

Fig. 13 Adversarial samples on Cifar-10 (Krizhevsky & Hinton, 2009) generated by various methods. The perturbation rate is fixed to $\epsilon = 16$. (Zoom in for better visualization.)

the gradient on the clean images, which makes the adversarial perturbation more smooth and imperceptible.

Method efficiency. We compare the efficiency of various attack methods on the Market-1501 dataset using the P5000 GPU. ϵ is 16. Our method has lower efficiency than Basic (Kurakin et al., 2017), but is on the same level as the rest competitors. Specifically, our method consumes 0.1113 s to generate an adversarial query. In comparison,

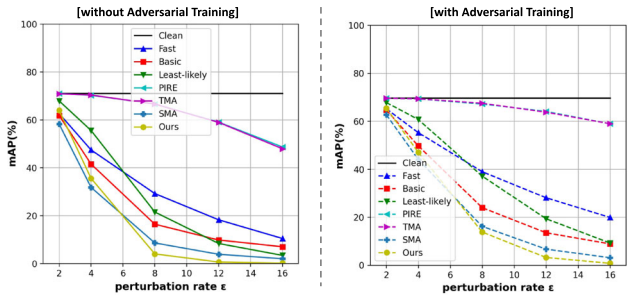

Fig. 14 Comparison between the original model (left) and the adversarially trained model (right) on Market-1501. The adversarial trained model is robust against various attacking methods, yielding relatively less performance drop. However, it still fails if a larger ϵ is applied

it takes 0.0134 s, 0.1119 s, and 0.1131 s for Basic (Kurakin et al., 2017), Least-likely (Goodfellow et al., 2015), and SMA (Bouniot et al., 2020), respectively, to attack a query image. The efficiency of PIRE (Liu et al., 2019b) and TMA (Tolias et al., 2019) is close. Both take 0.1586 s under the 20-iteration setting, and 0.7654 s for 100 iterations to craft one adversarial query.

Defense. We further explore whether online adversarial training (Madry et al., 2018) using ODFA as an adversarial augmentation could improve system robustness against various attack methods. The experiment is conducted on Market-1501. We have the following observations. First, the adversarially trained model is slightly inferior to the baseline model on the original test set in terms of retrieval accuracy. In our experiment, the model trained with online-generated ODFA adversarial samples yields 88.24% Recall@1, 96.88% Recall@10 and 69.70% mAP, the performance of which is slightly worse than the baseline model trained on "clean" data (88.54% Recall@1, 96.85% Recall@10 and 71.08% mAP). A similar phenomenon is also observed in the existing defense works (Bouniot et al., 2020; Bai et al., 2020a). It is mainly because the newly-added adversarial samples are different from the "natural" distribution as in the original training/test set. Second, adversarial training makes the model more robust against various adversarial methods. As shown in Fig. 14, the model with adversarial training (right) yields relatively less performance drop than the one trained on "clean" data (left). This observation is particularly noticeable when the perturbation is small (ϵ = 2,4). Third, the

proposed attacking method ODFA still fools the model after adversarial training under larger perturbations. For example, when $\epsilon = 16$, model performance drops significantly to 0.50% Recall@1, 2.55% Recall@10 and 0.73% mAP. We think our attack success is mainly attributed to the inherent design of the current deep learning structure. Despite various regularization terms, the learned models are sensitive to small perturbations. In this work, since we focus on adversarial attacks, we only provide a preliminary study on defense and leave the insights derived from such experiments in future works.

6 Conclusion

In this paper, we consider the adversarial attack for the image retrieval problem and propose an attack method named opposite-direction feature attack (ODFA) tailored for the retrieval scenario. Different from previous works, ODFA does not depend on category prediction but instead takes advantage of the intermediate feature and explicitly changes the feature direction in the representation space like a U-turn. On five image retrieval datasets, *i.e.,* Food-256 (Kawano & Yanai, 2014), CUB-200-2011 (Wah et al., 2011), Market-1501 (Zheng et al., 2015), Oxford5k (Philbin et al., 2007) and Paris6k (Philbin et al., 2008), we validate the effectiveness of the proposed method on two kinds of retrieval victims, *i.e.,* classification-based retrieval model and ranking-based retrieval model. Compared with existing works, ODFA leads to a greater performance drop in ranking accuracy with human imperceptible perturbation. Furthermore, we extend ODFA to adapt a common query augmentation: multi-scale query inputs, yielding a high attack success rate. We also verify the effectiveness of ODFA in the black-box setting and present a preliminary study on the adversarial defense.

References

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. (2018). Synthesizing robust adversarial examples. In ICML.

Babenko, A., Slesarev, A., Chigorin, A., and Lempitsky, V. (2014). Neural codes for image retrieval. In ECCV.

Bai, S., Bai, X., and Tian, Q. (2017). Scalable person re-identification on supervised smoothed manifold. In CVPR.

Bai, S., Li, Y., Zhou, Y., Li, Q., & Torr, P. H. (2020). Adversarial metric attack and defense for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(6), 2119–2126.

Bai, X., Yang, M., Huang, T., Dou, Z., Yu, R., & Xu, Y. (2020). Deep-person: Learning discriminative deep features for person re-identification. *Pattern Recognition, 98*, 107036.

Bouniot, Q., Audigier, R., and Loesch, A. (2020). Vulnerability of person re-identification models to metric adversarial attacks. In CVPR Workshop.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., and Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. arXiv:1810.00069.

Chen, J. and Ngo, C.-W. (2016). Deep-based ingredient recognition for cooking recipe retrieval. In ACM Multimedia.

Chen, Y., Zhu, X., and Gong, S. (2017). Person re-identification by deep learning multi-scale representations. In ICCV.

Cherepanova, V., Goldblum, M., Foley, H., Duan, S., Dickerson, J., Taylor, G., and Goldstein, T. (2021). Lowkey: leveraging adversarial attacks to protect social media users from facial recognition. In ICLR.

Deng, C., Yang, X., Nie, F., & Tao, D. (2019). Saliency detection via a multiple self-weighted graph-based manifold ranking. *IEEE Transactions on Multimedia, 22*(4), 885–896.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., and Li, J. (2018). Boosting adversarial attacks with momentum. CVPR.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In CVPR.

Gong, Y., Huang, L., and Chen, L. (2022). Person re-identification method based on color attack and joint defence. In CVPR.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In ICLR.

Guo, H., Zhao, C., Liu, Z., Jinqiao, W., and Hanqing, L. (2018). Learning coarse-to-fine structured feature embedding for vehicle re-identification. aaai 2018. In AAAI.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In CVPR.

Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. arXiv:1703.07737.

Huang, G., Liu, Z., Maaten, L. V. D., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In CVPR.

Jin, L., Li, K., Li, Z., Xiao, F., Qi, G.-J., & Tang, J. (2018). Deep semantic-preserving ordinal hashing for cross-modal similarity search. *IEEE Transactions on Neural Networks and Learning Systems, 30*(5), 1429–1440.

Kawano, Y. and Yanai, K. (2014). Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In ECCV workshop on transferring and adapting source knowledge in computer vision (TASK-CV).

Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images.

Kurakin, A., Goodfellow, I., and Bengio, S. (2017). Adversarial examples in the physical world. In ICLR Workshop.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks, 3361*(10), 1995.

Li, J., Ji, R., Liu, H., Hong, X., Gao, Y., and Tian, Q. (2019a). Universal perturbation attack against image retrieval. In ICCV.

Li, K., Qi, G.-J., & Hua, K. A. (2018). Learning label preserving binary codes for multimedia retrieval: A general approach. *ACM Transactions on Multimedia, Computing, Communications and Applications (TOMM), 14*(1), 2.

Li, X., Li, J., Chen, Y., Ye, S., He, Y., Wang, S., Su, H., and Xue, H. (2021). Qair: Practical query-efficient black-box attacks for image retrieval. In CVPR.

Li, Y., Yao, T., Pan, Y., Chao, H., & Mei, T. (2019). Deep metric learning with density adaptivity. *IEEE Transactions on Multimedia, 22*(5), 1285–1297.

Lin, K., Lu, J., Chen, C.-S., Zhou, J., & Sun, M.-T. (2018). Unsupervised deep learning of compact binary descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(6), 1501–1514.

Lin, K., Yang, H.-F., Hsiao, J.-H., and Chen, C.-S. (2015). Deep learning of binary hash codes for fast image retrieval. In CVPR Workshop.

Liu, M.-Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., and Kautz., J. (2019a). Few-shot unsueprvised image-to-image translation. In CVPR.
Liu, S., Song, Z., Liu, G., Xu, C., Lu, H., and Yan, S. (2012). Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In CVPR.
Liu, W., Wen, Y., Yu, Z., and Yang, M. (2016). Large-margin softmax loss for convolutional neural networks. In ICML.
Liu, X., Liu, W., Mei, T., & Ma, H. (2017). Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Transactions on Multimedia, 20*(3), 645–658.
Liu, Y., Chen, X., Liu, C., and Song, D. (2017b). Delving into transferable adversarial examples and black-box attacks. In ICLR.
Liu, Z., Zhao, Z., and Larson, M. (2019b). Who's afraid of adversarial queries? the impact of image modifications on content-based image retrieval. In ICMR.
Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In ICLR.
Moosavi Dezfooli, S. M., Fawzi, A., and Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In CVPR.
Moosavidezfooli, S. M., Fawzi, A., Fawzi, O., and Frossard, P. (2017). Universal adversarial perturbations. In CVPR.
Narodytska, N. and Kasiviswanathan, S. P. (2017). Simple black-box adversarial perturbations for deep networks. CVPR Workshop.
Papernot, N., McDaniel, P., and Goodfellow, I. (2016a). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv:1605.07277.
Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. (2017). Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on asia conference on computer and communications security, pp. 506–519.
Papernot, N., Mcdaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. (2016b). The limitations of deep learning in adversarial settings. European Symposium on Security & Privacy.
Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. NeurIPS Workshop.
Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In CVPR.
Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2008). Lost in quantization: Improving particular object retrieval in large scale image databases. In CVPR.
Qian, X., Fu, Y., Jiang, Y.-G., Xiang, T., and Xue, X. (2017). Multi-scale deep learning architectures for person re-identification. In ICCV.
Radenović, F., Tolias, G., and Chum, O. (2016). Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In ECCV.
Radenović, F., Tolias, G., & Chum, O. (2018). Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(7), 1655–1668.
Ristani, E. and Tomasi, C. (2018). Features for multi-target multi-camera tracking and re-identification. In CVPR.
Schonberger, J. L., Radenovic, F., Chum, O., and Frahm, J.-M. (2015). From single image query to detailed 3d reconstruction. In CVPR.
Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K. (2016). Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In ACM SIGSAC conference on computer and communications security.
Shen, C., Jin, Z., Chu, W., Jiang, R., Chen, Y., Qi, G.-J., & Hua, X.-S. (2019). Multi-level similarity perception network for person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15*(2), 32.
Shi, C., Xu, X., Ji, S., Bu, K., Chen, J., Beyah, R., and Wang, T. (2021). Adversarial captchas. IEEE Transactions on Cybernetics.
Sigurbjörnsson, B., Van Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. In WWW.
Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In ICLR.
Song, H. O., Xiang, Y., Jegelka, S., and Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In CVPR.
Suh, Y., Wang, J., Tang, S., Mei, T., and Mu Lee, K. (2018). Part-aligned bilinear representations for person re-identification. In ECCV.
Sun, Y., Zheng, L., Li, Y., Yang, Y., Tian, Q., & Wang, S. (2019). Learning part-based convolutional features for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(3), 902–917.
Sun, Y., Zheng, L., Yang, Y., Tian, Q., and Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling. ECCV.
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. In ICLR.
Tolias, G., Radenovic, F., and Chum, O. (2019). Targeted mismatch adversarial attack: Query with a flower to retrieve the tower. In ICCV.
Tolias, G., Sicre, R., and Jégou, H. (2015). Particular object retrieval with integral max-pooling of cnn activations. In ICLR.
Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. In ICLR.
Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).
Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.
Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. (2020). Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(10), 3349–3364.
Wang, W., Qian, X., Fu, Y., and Xue, X. (2022). Dst: Dynamic substitute training for data-free black-box attack. In CVPR.
Wang, X., Li, S., Liu, M., Wang, Y., and Roy-Chowdhury, A. K. (2021). Multi-expert adversarial attack detection in person re-identification using context inconsistency. In ICCV.
Wang, Y., Lin, X., Wu, L., & Zhang, W. (2017). Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval. *IEEE Transactions on Image Processing, 26*(3), 1393–1404.
Wei, Y., Zhao, Y., Lu, C., Wei, S., Liu, L., Zhu, Z., & Yan, S. (2016). Cross-modal retrieval with CNN visual features: A new baseline. *IEEE Transactions on Cybernetics, 47*(2), 449–460.
Xiao, C., Zhu, J.-Y., Li, B., He, W., Liu, M., and Song, D. (2018). Spatially transformed adversarial examples. In ICLR.
Yan, C., Gong, B., Wei, Y., & Gao, Y. (2020). Deep multi-view enhancement hashing for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(4), 1445–1451.
Yan, C., Li, Z., Zhang, Y., Liu, Y., Ji, X., & Zhang, Y. (2020). Depth image denoising using nuclear norm and learning graph model. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16*(4), 1–17.
Yan, C., Teng, T., Liu, Y., Zhang, Y., Wang, H., & Ji, X. (2021). Precise no-reference image quality evaluation based on distortion identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 17*, 1–21.
Yang, E., Deng, C., Li, C., Liu, W., Li, J., & Tao, D. (2018). Shared predictive cross-modal deep quantization. *IEEE Transactions on Neural Networks and Learning Systems, 29*(11), 5292–5303.

Yang, E., Liu, T., Deng, C., & Tao, D. (2018). Adversarial examples for hamming space search. *IEEE Transactions on Cybernetics, 50*(4), 1473–1484.

Yang, H.-F., Lin, K., & Chen, C.-S. (2017). Supervised learning of semantics-preserving hash via deep convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(2), 437–451.

Yang, X., Zhou, P., & Wang, M. (2018). Person reidentification via structural deep metric learning. *IEEE Transactions on Neural Networks and Learning Systems, 30*(10), 2987–2998.

Yang, Y., Zhuang, Y., & Pan, Y. (2021). Multiple knowledge representation for big data artificial intelligence: Framework, applications, and case studies. *Frontiers of Information Technology and Electronic Engineering, 22*(12), 1551–1558. https://doi.org/10.1631/FITEE.2100463

Yu, H., Dong, F., Li, J., Xie, W., Qiu, J., and Gu, Z. (2021). Adversarial attacks on vehicle re-identification. In DSC.

Yu, Q., Chang, X., Song, Y.-Z., Xiang, T., and Hospedales, T. M. (2017). The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. arXiv:1711.08106.

Yu, R., Dou, Z., Bai, S., Zhang, Z., Xu, Y., and Bai, X. (2018). Hard-aware point-to-set deep metric for person re-identification. In ECCV.

Yue-Hei Ng, J., Yang, F., and Davis, L. S. (2015). Exploiting local features from deep networks for image retrieval. In CVPR Workshop.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. BMVC.

Zhang, S., Ji, R., Hu, J., Lu, X., & Li, X. (2018). Face sketch synthesis by multidomain adversarial learning. *IEEE Transactions on Neural Networks and Learning Systems, 30*(5), 1419–1428.

Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., and Sun, J. (2017). Alignedreid: Surpassing human-level performance in person re-identification. arXiv:1711.08184.

Zhang, X., Zhang, R., Cao, J., Gong, D., You, M., and Shen, C. (2020). Part-guided attention learning for vehicle instance retrieval. IEEE Transactions on Intelligent Transportation Systems.

Zhao, G., Zhang, M., Liu, J., Li, Y., and Wen, J.-R. (2020). Ap-gan: Adversarial patch attack on content-based image retrieval systems. GeoInformatica, pp. 1–31.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In ICCV.

Zheng, L., Yang, Y., and Hauptmann, A. G. (2016). Person re-identification: Past, present and future. arXiv:1610.02984.

Zheng, Z., Ruan, T., Wei, Y., Yang, Y., & Mei, T. (2020). Vehiclenet: Learning robust visual representation for vehicle re-identification. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2020.3014488

Zheng, Z., Zheng, L., Hu, Z., and Yang, Y. (2018a). Open set adversarial examples. arXiv:1809.02681.

Zheng, Z., Zheng, L., & Yang, Y. (2018). A discriminatively learned CNN embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14*(1), 13. https://doi.org/10.1145/3159171

Zhong, Z., Zheng, L., Luo, Z., Li, S., & Yang, Y. (2020). Learning to adapt invariance in memory for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(8), 2723–2738.

Zhou, M., Wang, L., Niu, Z., Zhang, Q., Xu, Y., Zheng, N., and Hua, G. (2021). Practical relative order attack in deep ranking. In ICCV.