# Adversarial Attack on Video Retrieval

Ying Zou
Shanghai Jiao Tong University, China
zouying@sjtu.edu.cn

Chenglong Zhao
Shanghai Jiao Tong University, China
cl-zhao@sjtu.edu.cn

Bingbing Ni
Shanghai Jiao Tong University, China
nibingbing@sjtu.edu.cn

## ABSTRACT

Recently adversarial examples have been reported to reveal the fragility of deep learning models. However, most adversarial attacks focus on classification task and less attention has been paid to retrieval task. In this paper, we are *the first* to investigate adversarial examples on the video retrieval system in both non-targeted and targeted attack terms for copyright protection. Specifically, a triplet scheme is developed to take query-relevant and query-target pair-wise relationships together to enhance the attack performance. We evaluated the proposed method on the most commonly used video retrieval dataset CC_WEB_VIDEO, and successfully attack three popular video retrieval systems.

## CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**; • **Computer vision tasks**; • **Visual content-based indexing and retrieval**;

## KEYWORDS

Video Retrieval, Adversarial Examples, Triplet Loss

**ACM Reference Format:**
Ying Zou, Chenglong Zhao, and Bingbing Ni. 2020. Adversarial Attack on Video Retrieval. In *2020 The 4th International Conference on Video and Image Processing (ICVIP 2020), December 25–27, 2020, Xi'an, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3447450.3447478

## 1 INTRODUCTION

Recently, some researchers [10, 26], have found that DNNs perform frailly on adversarial examples. Namely, a legitimated image added with elaborated, human-imperceptible perturbation can mislead the DNNs' classifier to output wrong results. Even though extensive efforts have been made on image recognition task [10, 21, 25, 26], very little attention has been paid to the vulnerability of recent video retrieval models [14–16, 24]. In this paper, we investigate the adversarial examples of DNNs-based video retrieval and reveal the flaws of those approaches.

Given a query video, video retrieval aims at finding near-duplicate videos from the database and ranks these videos with the similarity score. In this work, we firstly propose to attack video retrieval systems that are deployed upon DNNs, in *non-targeted* and *targeted* attack terms. For *non-targeted* attack, query videos added with adversarial noise mislead video retrieval systems by outputting not near-duplicate videos. Namely, the retrieved videos ranked at the top are different to the crafted adversarial examples in visual context, which is illustrated in the second row in Figure 1. For *targeted* attack, the crafted adversarial examples can be retrieved to a specific one, though they are completely dissimilar in visual content. Therefore, a query video can be slightly modified without changing its essential content to *evade the copyright detection* provided by video retrieval. We argue that adversarial attack on video retrieval can help researchers understand the working mechanism of deep models, and facilitate the robustness of DNNs-based video retrieval by providing adversarial examples as training data [21]. Moreover, these crafted adversarial examples also can be used for *privacy protection*, if a user wants to protect his video's privacy information and avoids being indexed by video retrieval systems deployed on website search engines.

In this paper, we generate adversarial examples against video retrieval systems by a well-designed triplet scheme, which considers two pair-wise relationships together, i.e., query-relevant and query-target (query-irrelevant). Then optimizing the corresponding triplet loss will compact query-target distance while repulse query-relevant distance in embedding space, and thus improve the attack ability. A hyper parameter is introduced to adjust the ratio of these two terms, to achieve non-targeted and targeted attack respectively. Moreover, we reformulate the traditional mean Average Precision (mAP) to evaluate the performance of the crafted adversarial examples. Extensive experimental results well demonstrate that the proposed method can effectively impose adversarial attacks against several state-of-the-art video retrieval models.

## 2 RELATED WORKS

### 2.1 Adversarial Examples

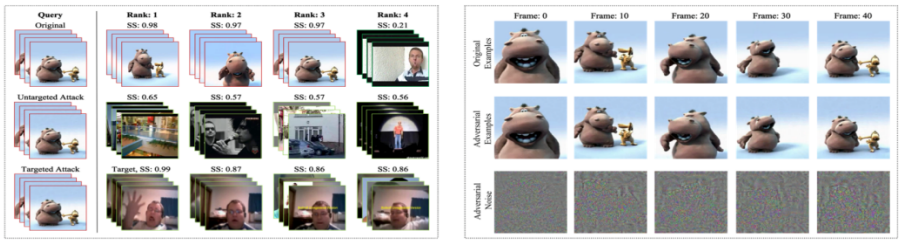
Szegedy et al. [27] have reported that deep neural networks have the weakness of being easily disturbed by slight disturbances applied to the inputs. These crafted inputs, called adversarial examples, have attracted great attention [2, 4, 5, 7, 11, 17, 22, 27] in the field of AI security and are developed in two ways, white-box and black-box attack. Adversarial attack in white-box term [4, 7, 11, 17, 22, 27] obtains better performance and relatively lower computation cost because of available access to the attack models. Instead, black-box attacks [1, 2, 5, 6] are more challenging due to the lack of DNN models' knowledge (i.e., architectures and parameters). Recently, some researchers pay attention to image retrieval attacks [18, 20, 28]. Li et al. [18] first introduce adversarial perturbations [22] to attack image retrieval systems in non-targeted scenarios, while [28] proposes the attacking concept of targeted.

### 2.2 Video Retrieval Systems

Video retrieval systems are mainly divided into three types: video-level [12, 15, 16, 19], frame-level [3, 8, 14] and hybrid-level [9, 23, 24].

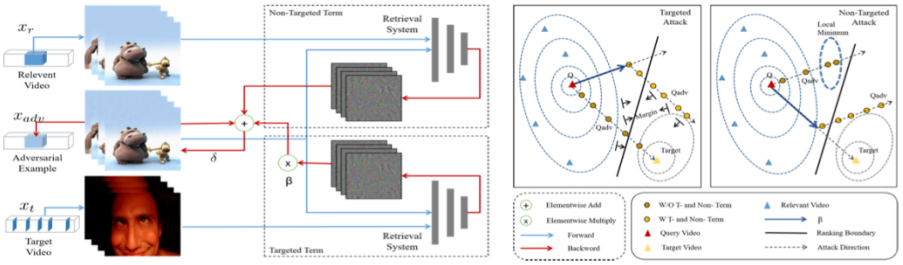

*ICVIP 2020, December 25–27, 2020, Xi'an, China*

ACM ISBN 978-1-4503-8907-5/20/12…$15.00
https://doi.org/10.1145/3447450.3447478




(a) Task Definition (b) Attack Visualization

**Figure 1: Motivation: (a). Top row shows the output of video retrieval systems which is ranked by similarity scores (SS). (b). The frames of the crafted adversarial examples.**




(a) Framework (b) Metric Explanation

**Figure 2: (a)Framework of our proposed method. (b). The explanation of Targeted Attack and Non-Targeted Attack while utilizing triplet scheme.**

First, the video-level retrieval is represented by [15, 16]. This kind of approaches concentrates on the video's global scale, and then parse the whole video into only one feature vector, resulting in low complexity and accuracy. Second, the frame-level retrieval methods compare frames one by one to generate two videos' correlation map. Third, hybrid-level retrieval methods attempt to extract Spatio-Temporal representations and deploy the Fourier Transform to judge relationships.

## 3 METHODOLOGY

### 3.1 Preliminary

**Retrieval Task**: Given a query video, the video retrieval system outputs a set of relevant videos ranked by the similarity score, and hopes that the near-duplicate videos can be retrieved at the top rank. The goal of video retrieval is to learn an embedding function $f(\cdot)$ that assigns a higher similarity score to relevant videos compared to irrelevant ones. Adversarial attack on video retrieval can also be divided into *non-targeted* and *targeted* modes. For *non-targeted* attack, the crafted adversarial example shows dissimilar representation to the relevant video. This is formulated as follows:

$$\min_{\delta}\{S(f(x+\delta), f(x_r)) + \lambda \cdot ||\delta||_p\}$$

where $x$ denotes the query video and $x_r$ denotes relevant video. Our goal is to minimize the similarity of these two videos. $f(\cdot)$ is an embedding mapping function, where $S(\cdot)$ is similarity calculation function. The regularization term $||\delta||_p$ is used to constrain the norm of the additive perturbation.

For *targeted* attack, the query video with additive perturbation can be retrieved to a specific one, though they are completely dissimilar in visual content. So, the object function is designed to maximize the similarity score between the query video and the target one $x_t$, which is depicted as follows:

$$\min_{\delta}\{-S(f(x+\delta), f(x_t)) + \lambda \cdot ||\delta||_p\}$$

### 3.2 Triplet Loss Function

In this section, we propose a triplet loss to generate adversarial examples against video retrieval systems. Suppose a relevant video $x_r$, a target video $x_t$ and the query video $x$. The triplet loss represents a relative similarity order among these three videos, i.e., $x$ is more similar to $x_r$ in contrast to $x_t$ in visual content. The loss function is formulated as follows:

$$L = S(f(x+\delta), f(x_r)) - \beta \cdot S(f(x+\delta), f(x_t)) + \lambda \cdot ||\delta||$$

The first term and the second term are also called Non-Targeted Term and Targeted Term respectively. In Figure 2, the hyper parameter $\beta$ is introduced as an adjustable factor to achieve a good balance between these two terms. Namely, by adjusting the value of $\beta$, we achieve different attack modes, i.e., non-targeted and targeted attack. The regularization term $||\delta||_p$ is designed to reduce the magnitude of the crafted perturbation, where small perturbation makes the adversarial example natural and human-imperceptible. The hyper parameter $\lambda$ is the penalty factor.

In particular, decreasing $\beta$ will assemble the loss function to emphasize on distinguishing the adversarial example and the relevant

one, and then leads to non-targeted attack. Increasing $\beta$ will make the loss function bring adversarial example and the target video closer in embedding space, thus the crafted adversarial example can be retrieved as the target one. For non-targeted attack, we introduce the query-target loss term to avoid local optimum. Decreasing the similarity of query-relevant term can't guarantee that the crafted adversarial example is retrieved as near-duplicated videos, so query-target term imposes the adversarial example away from the near-duplicated videos in embedding space, which is visualized in Non-Targeted Attack in Figure 2(b). For targeted attack, minimizing the query-relevant term will enlarge the margin of query video and relevant video in embedding space, which greatly avoids the adversarial examples being retrieved as the near-duplicated videos.

### 3.3 Attack on Video Retrieval

Adversarial attack on video retrieval is in the white-box setting, i.e., the gradient to the model is available for crafting adversarial examples. As shown in ALGORITHM.1, we give the pipeline of the proposed attack method on video retrieval systems. The query video is decomposed into a series of slices, where one slice $x_{slice}$ contains several frames in temporal contiguously, and then we feed the retrieval model with these slices iteratively to avoid GPU memory constraint.

**Algorithm 1** Video Retrieval Attack Algorithm

**Input**: Relevant video $x_r$, Target Video $x_t$, Adversarial Noise $\delta$, update rate $\eta$, constraint $\varepsilon$.
**Output**: Adversarial Example $x_{adv}$

FOR $xr_{slice}$ in $x_r$
  initialize Adam(.)
  $\delta_{slice} \leftarrow 0$, $x_{slice} \leftarrow xr_{slice}$
  FOR iter in Iterations
    $xt_{sample} \leftarrow \text{sample}(x_t)$
    $L \leftarrow L(x_{slice}, \delta_{slice}, x_{adv}, xt_{sample})$
    $\nabla_{\hat{x}} \leftarrow \text{Adam}(\nabla_x L)$
    $\delta_{slice} \leftarrow \delta_{slice} + \eta \cdot \nabla_{\hat{x}}$
    $\delta_{slice} \leftarrow \text{clip}(\delta_{slice}, -\varepsilon, \varepsilon)$
    END
  $x_{adv}[slice] \leftarrow x_{slice} + \delta_{slice}$
END

Following the PGD [21], we adopt a progressive manner to generate adversarial examples for improving attack performance, which is formulated as follows:

$$x_t = \prod_{x+\varepsilon} (x_t + \eta \cdot \nabla_x L)$$

For each iteration, we clip the generated adversarial example $x_{slice}$ in $\varepsilon-$ bounded $L_p$ ball, to yield human-imperceptive adversarial example. Adam(.) [13] is introduced to generate stable gradient, which is favorable for quick convergence.

### 3.4 Evaluation Metric

Mean Average Precision (mAP) is a standard metric which is used to measure the performance of video retrieval systems. So, we propose to modify mean Average Precision (mAP) to measure adversarial examples against video retrieval models. The original mAP metric is formulated as follows:

$$mAP = \sum_q (\frac{1}{n} \sum_{i=0}^{n} \frac{i}{r_i})$$

where n means the number of relevant videos, and$r_i$is the rank position of retrieved video, and q denotes the total query videos. For non-targeted attack, we propose the drop ratio of mAP to evaluate the performance of adversarial examples, which is depicted as follows:

$$\Delta mAP = mAP(x) - mAP(x_{adv})$$

where $x_{adv}$ denotes adversarial examples and $x_{adv}$ is the original videos. A higher value of $\Delta mAP$ means the stronger ability of the crafted adversarial examples.

For targeted attack, we want the adversarial query to be retrieved as the target video. Namely, target video is considered as the only relevant video and should be retrieved in the top rank. So in this paper, mTOP is designed to measure the performance of targeted attack in video retrieval system. And mTOP is formulated as follows, where $r_{q,t}$ is the rank position of retrieved target video.

$$mTOP = \sum_q (\frac{1}{r_{q,t}})$$

## 4 EXPERIMENTS

We carry out extensive experiments to evaluate the performance of the proposed method. For the evaluation metric, we choose drop mAP ($\Delta mAP$) to measure the non-targeted attack ability and mTOP to measure the targeted attack ability.

**Dataset**: We conduct experiments on the public retrieval dataset: CC_WEB_VIDEO [29]. To further validate the proposed method, we select a subset from CC_WEB_VIDEO called CC_Tiny, which can be accurately retrieved and attains more than 0.99 mAP for most video retrieval models.

**Video Retrieval Models**: These DNNs-based retrieval systems are varied a lot in their thoughts, frameworks and algorithms. So, we attack three representative methods: CNN-V [15], DML [16] and VISIL [14].

**Parameters Setup**: Consider the GPU memory limitation, we set the length of query slice ($x_{slice}$) equal to 10 and sample 20 frames from the target $x_t$ video. After then, the implementation environments are set as follows: a 7700k Intel CPU and a Nvidia 1080Ti GPU with 11GB memory. The update rate $\eta$ is 1.8 and the number of Iterations is 50. Adversarial perturbation constraint $\varepsilon$ is set as 5, where the pixel value of video is unified into [0,255]. We set $\beta$ as 1 and 10, respectively for non-targeted and targeted attack.

### 4.1 Non-targeted Attack

In this section, we evaluate the proposed method on two datasets, i.e, CC_WEB_VIDEO and CC_Tiny. For each dataset, we follow DML [16] to create a more challenge variant i.e., CC_WEB_VIDEO* and CC_Tiny*, where the set of candidates is the entire dataset instead of the query subset.

As shown in Table 1, our proposed method attacks these three models successfully and obtains nearly 1.0 $\Delta mAP$ for CC_Tiny*. Namely, the generated adversarial examples absolutely fool the

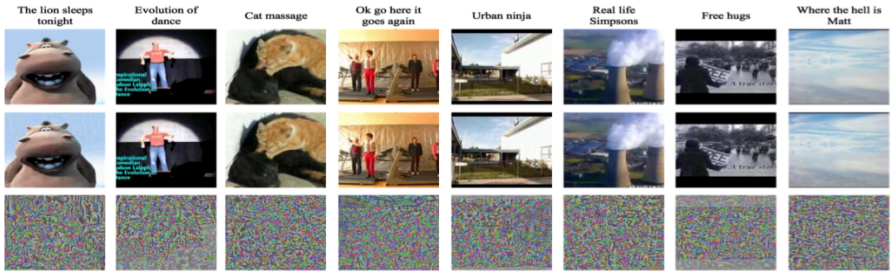



**Figure 3: Demos of the Crafted Adversarial Examples. From the first row to the third row, we display original videos, adversarial examples and adversarial noise.**

**Table 1: Performance about Non-Targeted Attack.**

| Dataset | CNN-V [15] | DML [16] | VISIL [14] |
|---|---|---|---|
| CC WEB VIDEO | 0.5631 | 0.6997 | 0.4940 |
| CC WEB VIDEO* | 0.8339 | 0.8644 | 0.9034 |
| CC Tiny | 0.7403 | 0.8702 | 0.6599 |
| CC Tiny* | 0.9691 | 0.9767 | 0.9722 |

**Table 2: Performance about Targeted Attack.**

| Dataset | CNN-V [15] | DML [16] | VISIL [14] |
|---|---|---|---|
| CC WEB VIDEO | 0.8026 | 0.8391 | 0.9590 |
| CC WEB VIDEO* | 0.6074 | 0.8333 | 0.9583 |
| CC Tiny | 1.0000 | 0.9000 | 1.0000 |
| CC Tiny* | 1.0000 | 0.7424 | 1.0000 |

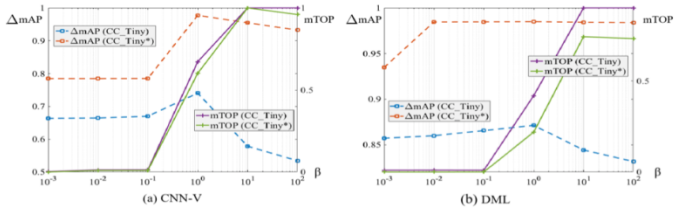



**Figure 4: Parameters analysis of Beta. $\Delta mAP$ is plotted as dashed line and mTOP is plotted as solid line. Higher value means better performance.**

retrieval model with nearly 0.0 mAP, where the legitimated inputs are all successfully retrieved with 1.0 mAP. Comparing with CC_WEB_VIDEO, $\Delta mAP$ in CC_Tiny is higher amount all three models in the Table 1, because we select the specific queries to construct CC_Tiny. In this way, the mAP value can achieve 1.0 for all three models.

## 4.2 Targeted Attack

As shown in Table 2, the proposed method against VISIL achieves the highest mTOP amount these four variants. For CC_WEB_VIDEO and CC_WEB_VIDEO*, VISIL is the most fragile model for nearly 0.96 mTOP, while CNN-V is the most robust model with 0.80 and 0.61 mTOP respectively. For CC_Tiny and its variant CC_Tiny*, targeted attacks on CNN-V and VISIL achieve 1.0 mTOP.

## 4.3 Parameters Analysis of Beta

In Figure 4, increasing $\beta$ to be more than 1.0 causes seriously decreasing of $\Delta mAP$. Moreover, the proposed method performs poorly when $\beta$ is set to be less than 1.0. Therefore, we select $\beta$ as 1.0 to obtain the optimal $\Delta mAP$ in all experiments. For *targeted attack*, we set $\beta$ in a range from 0.001 to 100, where a sharp rise of mTOP appears around [0.1, 1, 10] in both (a) and (b). While increasing $\beta$, the value of mTOP is almost rising monotonically, but a tiny withdrawal of mTOP(CC_Tiny*) appears at 100 in (a) CNN-V.

**Table 3: Sensitivity Analysis of L infinite norm.**

| | Non-Targeted Attack | | | | | Targeted Attack | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Linf | 2 | 3 | 5 | 10 | 15 | 2 | 3 | 5 | 10 | 15 |
| CNN-V | 0.489 | 0.624 | 0.740 | 0.799 | 0.811 | 0.002 | 0.013 | 1.000 | 1.000 | 1.000 |
| CNN-V* | 0.756 | 0.866 | 0.969 | 0.984 | 0.984 | 0.000 | 0.009 | 1.000 | 1.000 | 1.000 |
| DML | 0.830 | 0.862 | 0.870 | 0.874 | 0.875 | 0.003 | 0.435 | 0.900 | 0.900 | 1.000 |
| DML* | 0.944 | 0.976 | 0.977 | 0.984 | 0.986 | 0.000 | 0.032 | 0.742 | 0.900 | 1.000 |
| VISIL | 0.003 | 0.298 | 0.660 | 0.810 | 0.820 | 0.023 | 0.700 | 1.000 | 1.000 | 1.000 |
| VISIL* | 0.113 | 0.731 | 0.972 | 0.990 | 0.991 | 0.023 | 0.700 | 1.000 | 1.000 | 1.000 |

### 4.4 $L_\infty$ Vs. Attack Performance

In this section, we conduct a parameter sensitivity study to explore how $L_\infty$ effects the performance of targeted and non-targeted attack. Intuitively, slacking infinity constraint ($L_\infty$) leads to better adversarial attack ability due to the stronger perturbation, which also leads to more visibility. However, while increasing $L_\infty$ from 10 to 15, the values of $\Delta mAP$ and mTOP stop increasing and become saturated in Table 3

## 5 CONCLUSION

This the first attempt is to explore the adversarial attack on video retrieval. for copyright protection. And we proposed a novel scheme for obtaining adversarial examples against video retrieval systems. More importantly, extensive experiments well demonstrate that our proposed method attains both low L infinite norm and favorable attack performance.

## ACKNOWLEDGMENTS

This work was supported by National Science Foundation of China (U20B200011, 61976137). Authors appreciate the Student Innovation Center of SJTU for GPUs.

## REFERENCES

[1] Bhagoji, A.N., He, W., Li, B., Song, D.: Practical black-box attacks on deep neuralnetworks using efficient query mechanisms. In: European Conference on ComputerVision. pp. 158–174. Springer (2018)
[2] Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: Re-liable attacks against black-box machine learning models. arXiv preprintarXiv:1712.04248 (2017)
[3] Cai, Y., Yang, L., Ping, W., Wang, F., Mei, T., Hua, X.S., Li, S.: Million-scale near-duplicate video retrieval system. In: Proceedings of the 19th ACM international-conference on Multimedia. pp. 837–838 (2011)
[4] Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In:2017 ieee symposium on security and privacy (sp). pp. 39–57. IEEE (2017)
[5] Chen, P.Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.J.: Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitutemodels. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence andSecurity. pp. 15–26 (2017)
[6] Cheng, M., Le, T., Chen, P.Y., Yi, J., Zhang, H., Hsieh, C.J.: Query-efficienthard-label black-box attack: An optimization-based approach. arXiv preprint-arXiv:1807.04457 (2018)
[7] Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarialattacks with momentum. In: Proceedings of the IEEE conference on computervision and pattern recognition. pp. 9185–9193 (2018)
[8] Douze, M., J égou, H., Schmid, C.: An image-based approach to video copy detec-tion with spatio-temporal post-filtering. IEEE Transactions on Multimedia12(4),257–266 (2010)
[9] Feng, Y., Ma, L., Liu, W., Zhang, T., Luo, J.: Video re-localization. In: Proceedingsof the European Conference on Computer Vision (ECCV). pp. 51–66 (2018)
[10] Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial ex-amples. arXiv 1412.6572 (12 2014)
[11] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarialex-amples. arXiv preprint arXiv:1412.6572 (2014)
[12] Huang, Z., Shen, H.T., Shao, J., Zhou, X., Cui, B.: Bounded coordinate systemindexing for real-time video clip search. ACM Transactions on Information Systems(TOIS)27(3), 1–33 (2009)
[13] Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International-Conference on Learning Representations (12 2014)
[14] Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., Kompatsiaris, I.: Visil: Fine-grained spatio-temporal video similarity learning. In: Proceedings of the IEEEInternational Conference on Computer Vision. pp. 6351–6360 (2019)
[15] Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., Kompatsiaris, Y.: Near-duplicate video retrieval by aggregating intermediate cnn layers. In: Internationalconference on multimedia modeling. pp. 251–263. Springer (2017)
[16] Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., Kompatsiaris, Y.: Near-duplicate video retrieval with deep metric learning. In: Proceedings of the IEEEInternational Conference on Computer Vision Workshops. pp. 347–356 (2017)
[17] Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world.arXiv preprint arXiv:1607.02533 (2016)
[18] Li, J., Ji, R., Liu, H., Hong, X., Gao, Y., Tian, Q.: Universal perturbation attack-against image retrieval. In: Proceedings of the IEEE International Conference onComputer Vision. pp. 4899–4908 (2019)
[19] Liu, L., Lai, W., Hua, X.S., Yang, S.Q.: Video histogram: A novel video signature for efficient web video duplicate detection. In: International conference onmultimedia modeling. pp. 94–103. Springer (2007)
[20] Liu, Z., Zhao, Z., Larson, M.: Who's afraid of adversarial queries? the impact ofimage modifications on content-based image retrieval. In: Proceedings of the 2019on International Conference on Multimedia Retrieval. pp. 306–314 (2019)
[21] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learn-ingmodels resistant to adversarial attacks (06 2017)
[22] Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversar-ialperturbations. In: Proceedings of the IEEE conference on computer vision andpattern recognition. pp. 1765–1773 (2017)
[23] Poullot, S., Tsukatani, S., Phuong Nguyen, A., J égou, H., Satoh, S.: Temporal-matching kernel with explicit feature maps. In: Proceedings of the 23rd ACMinternational conference on Multimedia. pp. 381–390 (2015)
[24] Revaud, J., Douze, M., Schmid, C., J égou, H.: Event retrieval in large video col-lections with circulant temporal encoding. In: Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition. pp. 2459–2466 (2013)
[25] Rony, J., Hafemann, L.G., Oliveira, L.S., Ayed, I.B., Sabourin, R., Granger, E.:Decoupling direction and norm for efficient gradient-based l2 adversarial attacksand defenses. In: Proceedings of the IEEE Conference on Computer Vision andPattern Recognition. pp. 4322–4330 (2019)
[26] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.,Fergus, R.: Intriguing properties of neural networks (12 2013)
[27] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fer-gus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199(2013)
[28] Tolias, G., Radenovic, F., Chum, O.: Targeted mismatch adversarial attack: Query-with a flower to retrieve the tower. In: Proceedings of the IEEE InternationalConference on Computer Vision. pp. 5037–5046 (2019)
[29] Wu, X., Hauptmann, A.G., Ngo, C.W.: Practical elimination of near-duplicatesfrom web video search. In: Proceedings of the 15th ACM international conferenceon Multimedia. pp. 218–227 (2007)