Multi-label Adversarial Attack with New Measures and Self-paced Constraint Weighting

Fengguang Su, Ou Wu, and Weiyao Zhu

***Abstract*—An adversarial attack is typically implemented by solving a constrained optimization problem. In top-k adversarial attacks implementation for multi-label learning, the attack failure degree (AFD) and attack cost (AC) of a possible attack are major concerns. According to our experimental and theoretical analysis, existing methods are negatively impacted by the coarse measures for AFD/AC and the indiscriminate treatment for all constraints, particularly when there is no ideal solution. Hence, this study first develops a refined measure based on the Jaccard index appropriate for AFD and AC, distinguishing the failure degrees/costs of two possible attacks better than the existing indicator function-based scheme. Furthermore, we formulate novel optimization problems with the least constraint violation via new measures for AFD and AC, and theoretically demonstrate the effectiveness of weighting slack variables for constraints. Finally, a self-paced weighting strategy is proposed to assign different priorities to various constraints during optimization, resulting in larger attack gains compared to previous indiscriminate schemes. Meanwhile, our method avoids fluctuations during optimization, especially in the presence of highly conflicting constraints. Extensive experiments on four benchmark datasets validate the effectiveness of our method across different evaluation metrics.**

***Index Terms*—Multi-label learning, adversarial attack, optimization problem, optimization goal, solving strategy.**

I. Introduction

DEEP neural networks (DNN) are influenced by adversarial examples leading to wrong classifications [1], [2]. Specifically, when minor noises are added to normal samples (adversarial perturbations), the model makes incorrect decisions with high confidence [1]. This outcome promotes the development of adversarial attacks and adversarial defenses [3], with existing adversarial attack methods focusing on single-label classification tasks [3], [4], [5], [6], [7]. Besides, multi-label learning also prevails in practice [8]. Hence, the design of top-k adversarial attacks for multi-label learning has received increasing attention in recent years [9], [11].

Solving constrained optimization problems to obtain adversarial perturbations is the mainstream research path for both top-k targeted and untargeted attack tasks in multi-label learning [9], [12]. The optimization goals basically comprise the attack failure degree (AFD), the attack cost (AC), and the perturbation bound, whose measures heavily determine the final attack performance. These goals are transformed into either optimization objectives or a series of constraints in mathematical optimization. Therefore, although many studies do not explicitly mention AFD and AC, their constraints do imply how they measure AFD and AC. Current works, e.g., Top-k multi-label attack (T$_k$ML) [11], measure AFD and AC based on an indicator function, which is relatively coarse and cannot accurately reflect the practical AFD or AC for a possible attack. Moreover, since achieving the optimization goals is often not ideal, some constraints may be more challenging to satisfy than others, and in some cases, they may even conflict with each other. However, nearly all studies treat all constraints indiscriminately[1]. As a result, some challenging constraints will negatively impact the solving process because other constraints must make excessive concessions. In other words, these challenging constraints cause more other constraints to be unsatisfied. Meanwhile, in the case of highly conflicting constraints where only a part of the constraints can be satisfied, the solving procedure will fluctuate and thus increase the solution's complexity or even force the procedure to diverge. [45] extends T$_k$ML to pursue imperceptible adversarial perturbations when evaluation metrics are unreliable. However, similar problems still exist. To the best of our knowledge, previous studies have not reported this yet.

Therefore, this study designs more accurate AFD/AC measures and a more reasonable constraint treatment scheme for the underlying constrained optimization problem in top-k adversarial attacks for multi-label learning. Specifically, we conduct statistical and theoretical analysis to reveal the defects of current measures for AFD/AC and the indiscriminate treatment scheme employed in existing optimizing procedures. Second, we develop a new measure for AFD and AC based on the Jaccard index [19], which can more fine-grainedly reflect the failure degree and cost of a possible attack. Adopting our measure affords excluding more ground-truth labels from the top-k predicted labels of an implemented attack and thus increases the total attack gain. Third, this work considers the difficult-to-satisfy degree (slack variable) for each constraint. We theoretically demonstrate the effectiveness of weighting slack variables for constraints. Finally, based on the theoretical analysis, the constraints are discriminately treated using a novel self-paced weighting strategy. The proposed strategy dampens the fluctuation and thus accelerates the convergence speed of optimizing. Extensive trials on four benchmark data sets indicate that our method uses a small perturbation to make the adversarial attack more profitable. Our contributions are summarized as follows:

- Deep experimental and theoretical analysis of the optimization goals and procedures in several typical methods

Manuscript received 21 November 2023; revised 17 March 2024; accepted 31 May 2024. Corresponding author: Ou Wu.
F. Su, O. Wu, and W. Zhu are with Center for Applied Mathematics, Tianjin University, Tianjin 300072, China (e-mail: su.scenery@outlook.com; wuou@tju.edu.cn; weiyaozhu@tju.edu.cn).

[1]Section III-C presents theoretical analysis and illustrative examples.

reveals that a coarse AFD/AC measure and the indiscriminate treatment for constraints harm the attack gain and optimizing convergence.
- The Jaccard index is introduced to measure AFD/AC, formulating new optimization problems with the least constraint violation for both top-k targeted and untargeted multi-label adversarial attacks. And an extensive theoretical analysis of weighting strategies for slack variables is conducted for our new optimization problems.
- Based on our theoretical analysis, a self-paced weighting strategy is employed to assign different priorities for the constraints, benefiting the attack gain and the solving process.

II. Related Work

Notations. Let C be the label set. Y and Y_t are ground-truth and target labels sets ($\subset C$) of a sample $\boldsymbol{x}$, respectively. $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), \cdots, f_{|C|}(\boldsymbol{x})]^T$, $f_j(\boldsymbol{x}) \in [0,1]$ is a probability vector of a DNN. Let $[f_{[1]}(\boldsymbol{x}), \cdots, f_{[|C|]}(\boldsymbol{x})]^T$ be the values of $\boldsymbol{f}(\boldsymbol{x})$ sorted in descending order, i.e., $f_{[j]}(\boldsymbol{x})$ is the jth largest value of $\boldsymbol{f}(\boldsymbol{x})$, where $[j]$ is the label index of the top-j prediction score. $\hat{Y}_k(\boldsymbol{x}+\boldsymbol{z}) = \{[1], \cdots, [k]\}$ $(1 \leq k < |C|)$ is the top-k label index set and $\boldsymbol{z}$ is the perturbation. Assume that N_{ct} is the total number of constraints. $\Re$ is the real space. Let $Y_{lf}(\boldsymbol{z})$ be $\{j : f_j(\boldsymbol{x}+\boldsymbol{z}) \geq 0.5, \forall j \in Y\}$. B is a set of denoted irrelevant labels (neither ground truth nor target labels). Note that $B_I(\boldsymbol{x}, \boldsymbol{z}) = \{j : \mathbb{I}(f_j(\boldsymbol{x}) \geq 0.5) = \mathbb{I}(f_j(\boldsymbol{x}+\boldsymbol{z}) \geq 0.5);\ \forall j \in B\}$ is the irrelevant labels without obvious prediction changes, where $\mathbb{I}(con)$ is the indicator function that is 1 when the condition con is true and 0 otherwise.

Multi-label Learning. There are two common multi-label classifiers: rank-based and threshold-based classifiers [9], [11]. Rank-based classifier aims to achieve $Y \subset \hat{Y}_k(\boldsymbol{x})$. Threshold-based classifier, on the other hand, assigns labels in Y with predicted probabilities greater than a certain threshold. Multi-label learning faces many challenges, such as missing views and missing labels, which can be addressed by a novel two-stage network, proposed by [13]. In addition, [14] introduced an asymmetric loss that dynamically handles the easy negative samples and possibly mislabeled samples. To better address the distribution shift problem, [16] constructed a semi-supervised dual relation learning framework. [34] introduced a simple yet effective procedure, MILe, which incorporates inductive biases. And, better representation of deep features has been shown to facilitate multi-label learning [35]. Network architecture design is also a widely-used approach [15], [50]. Recently, [51] treated texts as images for prompt tuning to improve multi-label learning.

Adversarial Attack. Top-k untargeted and targeted attacks aim to find perturbations $\boldsymbol{z}$ that make $Y \cap \hat{Y}_k(\boldsymbol{x}+\boldsymbol{z}) = \emptyset$ and $\hat{Y}_k(\boldsymbol{x}+\boldsymbol{z}) = Y_t$, respectively. [9] defined the multi-label adversarial example and introduced a general framework to generate adversarial perturbations. [11] proposed a novel loss for multi-label top-k attack, which considers the ranking relation among labels. And, linear programming has facilitated progress in multi-label attacks [17]. And [36] elaborated on the theoretical analysis of multi-label classifier's attackability. Generative approaches, adopted in [10], [37], have also driven the advances in multi-label attacks. Moreover, [38] proposed PSAT-GAN, promoting the parallelism in multi-task holistic scene understanding. Other attack methods for single-label learning include [1], [12], [18]. Recently, [30] designed a more robust and invisible backdoor attack method, that achieved better attack performance. [39] directly minimized the distortion by modeling the noise distribution. Moreover, [46] applied label correlation to improve the methods in [9], [11], [17], which obtains larger adversarial attack gains. [47] proposed two novel adversarial attack methods against regression models for unmanned aerial vehicles (UAV) based on deep learning and further enhanced the robustness of regression models in UAV. [48] proposed Perturbing State Variables, Tailored Loss Function Design, and Change of Variables to infer suitable multi-label adversarial perturbations. Finally, [49] designed an effective multi-label black-box attack method based on differential evolution algorithm which includes a complementary mutation operator.

Adversarial Robustness. Top-k Adversarial robustness has also received much attention. [42] showed a certain number of $|Y \cap \hat{Y}_k(\boldsymbol{x}+\boldsymbol{z})|$, when random smoothing is used to train $\boldsymbol{f}(\cdot)$. [43] carried out a through and rigorous theoretical analysis to bridge the robustness gap between the norms of ℓ_0 and ℓ_2 of multi-label classifiers. And [44] derived a tight top-k robustness in ℓ_2 norm when using the Gaussian random smoothing. Furthermore, [52] proposed to use the partial weight initialization and fine-tuning to enhance the robustness of DNN against the clean-image backdoor attack.

III. Analysis for Existing Methods

A. Goals and Constraints in Existing Methods

Generally, the optimization goals for an attack generation in most existing top-k multi-label adversarial attack studies explicitly or implicitly include three folds as follows:

- **Attack failure degree (AFD)**: It is actually the primary goal of an attack task. Existing methods usually rely on the indicator function $\mathbb{I}(\cdot)$. For instance, [11] defined label consistency (LC) to measure AFD as follows:

$$\begin{aligned}\mathrm{AFD}_{lc}(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{I}(Y \subset \hat{Y}_k(\boldsymbol{x}+\boldsymbol{z})) \\ + \mathbb{I}(\hat{Y}_k(\boldsymbol{x}+\boldsymbol{z}) \subset Y) + \mathbb{I}(\hat{Y}_k(\boldsymbol{x}+\boldsymbol{z}) = Y)\end{aligned}. \quad (1)$$

  Note that a smaller failure degree is preferred.
- **Attack cost (AC)**: This is the second goal that reflects the cost brought by a given attack. A typical cost is the change of predictions on irrelevant labels, i.e., the number (or proportion) of irrelevant labels whose predictions were changed significantly. A lower total cost is preferred.
- **Perturbation bound**: This goal mainly uses ℓ_p norm to reflect the difference between the normal sample and the generated attack. A lower bound is preferred.

Existing measures often rely on an indicator function, which cannot accurately capture the degree of attack failure or cost for a possible attack, as analyzed in Section III-B. The goals AFD and AC are then translated into a set of constraints in mathematical optimization, as explained in Section IV-A. Some constraints are challenging to satisfy and may even

be mutually exclusive. Neglecting the distinct difficulties and conflicts among these constraints can result in unstable or non-convergent mathematical optimization, as discussed in Section III-C. The perturbation bound goal is directly treated as an optimization objective.

B. Defects Incurred by AFD/AC Measures

Three typical measures for AFD and AC are analyzed with three toy datasets and three Multi-layer Perceptron (MLP) classifiers trained on these toy datasets. The AFD measure in Eq. (1) takes 0 for an ideal attack or the most partially successful attacks on $\boldsymbol{x}$. Naturally, this measure cannot well distinguish the failure degrees for two possible attacks (including the ideal one). Fig. 1(a) presents the top-3 untargeted multi-label attack for the toy dataset when Eq. (1) is used, revealing that $\text{AFD}_{lc}(\boldsymbol{x}_1, \boldsymbol{z}_1^1) = \text{AFD}_{lc}(\boldsymbol{x}_1, \boldsymbol{z}_1^2) = 0$, as $Y = \{2, 3, 4\}$, $\hat{Y}_3(\boldsymbol{x}_1 + \boldsymbol{z}_1^1) = \{1, 2, 3\}$, and $\hat{Y}_3(\boldsymbol{x}_1 + \boldsymbol{z}_1^2) = \{5, 6, 4\}$. If $||\boldsymbol{z}_1^1|| \leq ||\boldsymbol{z}_1^2||$, then $\boldsymbol{x}_1^1$ $(= \boldsymbol{x}_1 + \boldsymbol{z}_1^1)$ rather than $\boldsymbol{x}_1^2$ $(= \boldsymbol{x}_1 + \boldsymbol{z}_1^2)$ may be the solution. However, $\boldsymbol{x}_1^2$ affords a smaller failure degree than $\boldsymbol{x}_1^1$ because $1 = |Y \cap \hat{Y}_k(\boldsymbol{x}_1^2)| < |Y \cap \hat{Y}_k(\boldsymbol{x}_1^1)| = 2$. Hence, an Eq. (1)-based optimization goal is inappropriate.

The label's prediction confidence is also a main consideration in adversarial attacks. From Fig. 1(b), the Y of $\boldsymbol{x}_2$ is $\{1, 4, 5\}$, and we have $f_1(\boldsymbol{x}_2 + \boldsymbol{z}_2^1) \geq 0.5$, $f_1(\boldsymbol{x}_2 + \boldsymbol{z}_2^2) \geq 0.5$ and $f_5(\boldsymbol{x}_2 + \boldsymbol{z}_2^2) \geq 0.5$. An underlying AFD measure, namely label flip (LF), is defined as follows [9]:

$$\text{AFD}_{lf}(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{I}(|Y_{lf}(\boldsymbol{z})| \neq 0). \tag{2}$$

According to Eq. (2), $\text{AFD}_{lf}(\boldsymbol{x}_2, \boldsymbol{z}_2^1) = \text{AFD}_{lf}(\boldsymbol{x}_2, \boldsymbol{z}_2^2) = 1$. Although $\boldsymbol{x}_2^1$ and $\boldsymbol{x}_2^2$ have the identical AFD_{lf} value, $\boldsymbol{x}_2^1$ makes more labels (labels 4 and 5) with prediction confidence below 0.5 than the adversarial image $\boldsymbol{x}_2^2$ makes (label 4), and the measure in Eq. (2) cannot distinguish them.

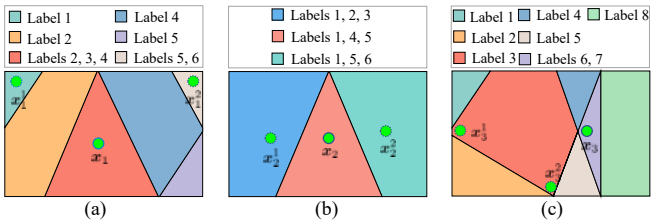
The AC measure used in [9] for targeted attack quantifies the irrelevant labels with obvious changes of the predictions after the attack:

$$\text{AC}(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{I}(|B_I(\boldsymbol{x}, \boldsymbol{z})| \neq |B|). \tag{3}$$

Likewise, for an ideal attack $\boldsymbol{z}$ (if it exists), $\text{AC}(\boldsymbol{x}, \boldsymbol{z})$ is 0. Otherwise, it becomes 1. For the top-2 targeted attack on the toy dataset illustrated in Fig. 1(c), $Y = \{6, 7\}$, $Y_t = \{2, 8\}$, and $B = \{1, 4, 5\}$. According to the trained MLP model, we have $\mathbb{I}(f_1(\boldsymbol{x}_3) \geq 0.5) = 0$, $\mathbb{I}(f_4(\boldsymbol{x}_3) \geq 0.5) = 1$, and $\mathbb{I}(f_5(\boldsymbol{x}_3) \geq 0.5) = 1$. We also have $B_I(\boldsymbol{x}_3, \boldsymbol{z}_3^1) = \emptyset$ and $B_I(\boldsymbol{x}_3, \boldsymbol{z}_3^2) = \{1, 4\}$. Then $\text{AC}(\boldsymbol{x}_3, \boldsymbol{z}_3^1) = 1$ and $\text{AC}(\boldsymbol{x}_3, \boldsymbol{z}_3^2) = 1$, if the measure in Eq. (3) is used. However, the actual cost of $\boldsymbol{x}_3^2$ $(= \boldsymbol{x}_3 + \boldsymbol{z}_3^2)$ is obviously smaller than that of $\boldsymbol{x}_3^1$ because $B \cap B_I(\boldsymbol{x}_3, \boldsymbol{z}_3^1) = \emptyset$ and $B \cap B_I(\boldsymbol{x}_3, \boldsymbol{z}_3^2) = \{1, 4\}$. Therefore, the measure in Eq. (3) may impose a higher cost when no ideal attack exists.

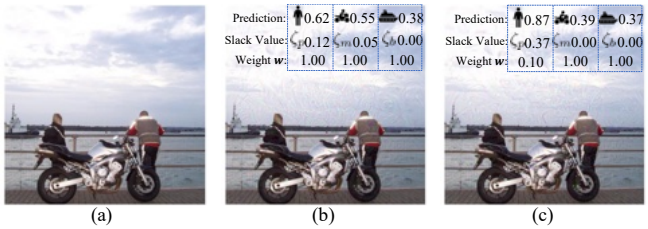
More details of the above statistical analysis about Fig. 1 are shown in the supplementary materials. The following proposition summarizes the defects of existing measures, with the proof being in Appendix.

Proposition 1. *For AFD/AC in Eqs. (1), (2), and (3), given two perturbations $\boldsymbol{z}^1$ and $\boldsymbol{z}^2$, even if $|Y \cap \hat{Y}_k(\boldsymbol{x}+\boldsymbol{z}^1)| \neq |Y \cap \hat{Y}_k(\boldsymbol{x}+\boldsymbol{z}^2)|$, $|Y_{lf}(\boldsymbol{z}^1)| \neq |Y_{lf}(\boldsymbol{z}^2)|$, and $|B_I(\boldsymbol{z}^1)| \neq |B_I(\boldsymbol{z}^2)|$, we still have $\text{AFD}_{lc}(\boldsymbol{x}, \boldsymbol{z}^1) = \text{AFD}_{lc}(\boldsymbol{x}, \boldsymbol{z}^2)$, $\text{AFD}_{lf}(\boldsymbol{x}, \boldsymbol{z}^1) = \text{AFD}_{lf}(\boldsymbol{x}, \boldsymbol{z}^2)$ and $\text{AC}(\boldsymbol{x}, \boldsymbol{z}^1) = \text{AC}(\boldsymbol{x}, \boldsymbol{z}^2)$, that is, the existing measures cannot distinguish two possible attacks well.*



(a) (b) (c)

Fig. 1. (a) and (b) Illustrative examples for AFD. (c) Illustrative example for AC. $\boldsymbol{x}_i, i = 1, 2, 3$ are normal samples, marked with the green circles. $\boldsymbol{x}_i^j = \boldsymbol{x}_i + \boldsymbol{z}_i^j, j = 1, 2$ are the adversarial samples of $\boldsymbol{x}_i$, where $\boldsymbol{z}_i^j$ is the j-th adversarial perturbation of $\boldsymbol{x}_i$. Each divided region denotes a subset of data with a specific label or labels (Y), represented by the diverse colored squares at the top. And each point in a region symbolizes a data sample.



(a) (b) (c)

Fig. 2. (a) Normal image, (b) Adversarial image generated by [9], and (c) Adversarial image generated by [9] via down-weighting the weight of the difficult constraint (for 'person').

C. Defects Incurred by Constraints Treatment

Current studies transform both measures for AFD and AC into equality or inequality constraints in the final optimization problem[2]. So, numerous constraints are typically considered, with existing optimization schemes treating all constraints equally and being given identical weights (excluding the Lagrange coefficients) in the Lagrangian optimization objective. However, this approach has two shortcomings.

First, large slack values typically indicate that certain constraints are hard to satisfy, thereby negatively affecting the total attack gain. Take Fig. 2(a) as an example. Fig. 2(b) shows the adversarial image generated by [9] which treats all constraints equally, that is, the weights are 1 for all labels. Only one label 'boat' is successfully attacked (its prediction is $0.38 < 0.5$). The slack value for the constraint corresponding to label 'person' (0.12) is much larger than those of 'motorbike' (0.05) and 'boat' (0.00), indicating that the constraint of 'person' is more difficult to satisfy. If the constraint for label 'person' is down-weighted (its weight is 0.10), then both 'motorbike' and 'boat' are successfully attacked as depicted in Fig. 2(c). The attack gain thus increases. We consider the value of the slack value of a constraint as the constraint difficulty. The larger the slack value is, the more difficult to satisfy the constraint is. To obtain a statistical support, we employ factor $\boldsymbol{w}$ to multiply the five constraints with largest slack values for 400 random images from VOC 2012 [21] and COCO 2014 [22] (200 for each). It is worth noting that the total number of constraints is greater than 10. Fig. 3(a) shows the average relative AFD to those achieved with equal constraint weights (i.e., $\boldsymbol{w} = 1$) when $\boldsymbol{w}$ ranges from 0 to 2. This figure highlights that AFD consistently increases as $\boldsymbol{w}$ increases, i.e., the equal weights on the constraints may be ineffective, potentially yielding larger failure degrees than a lower weight on difficult constraints.

[2]Some studies directly construct the constraints. Nevertheless, implicit measures for AFD and AC can still be observed.

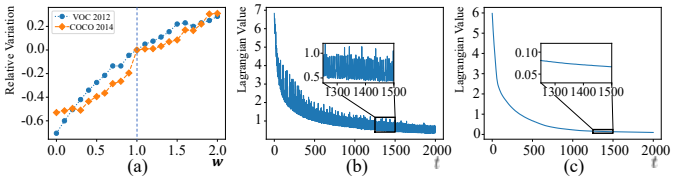


Fig. 3. (a) Relative AFD variation with different $\boldsymbol{w}$ values on VOC 2012 and COCO 2014, (b) and (c) variations of the Lagrangian values of [9] and our MASWT on VOC 2012, respectively.

Second, inevitably, some constraints cannot be satisfied simultaneously when no ideal solution exists[3]. These constraints fluctuate the optimization procedure as the following theorem:

Theorem 1. *Let $CT_j(\boldsymbol{z})$ be the jth constraint. Assume that all constraints can form $N_{\mathcal{G}}$ disjoint subregions of $\boldsymbol{z}$, where the j'th subregion $\mathcal{G}_{j'}$ is $\{\boldsymbol{z} : CT_{j'_1}(\boldsymbol{z}) \geq 0, \cdots, CT_{j'_{|\mathcal{G}_{j'}|}}(\boldsymbol{z}) \geq 0\}$ and $\mathcal{G}_{j'} \cap \mathcal{G}_{k'} = \emptyset, \forall k' \neq j'$. If all constraints are equally weighted, then $\boldsymbol{z}$ fluctuates among all local minimizers with a positive constant $\mathfrak{c}^* > 0$:*

$$\inf\left\{||\boldsymbol{z}^1 - \boldsymbol{z}^2||_2 : \boldsymbol{z}^1 \in \mathcal{G}_{j'}, \boldsymbol{z}^2 \in \mathcal{G}_{k'}\right\} \geq \mathfrak{c}^*. \tag{4}$$

That is, if the feasible regions for all constraints comprise some disjoint regions far from each other, the optimization procedure diverges. Fig. 3(b) illustrates the fluctuated Lagrangian of [9] on VOC 2012. Nevertheless, the proposed MASWT method introduced later is smoother (Fig. 3(c)). Theorem 1 explains our finding from a theoretical perspective. The lower bound in Theorem 1 is subject to the involved model and the sample implied in the constraints. And a video presented in the supplementary material highlights this fluctuation.

In summary, given the existing schemes, a more effective constraint treatment manner should be explored.

IV. Methodology

A. Optimization Problem with New Measures

1) New Measures with the Jaccard Index. The indicator function in existing attack studies calculates set similarity. In contrast to current methods, this study utilizes a more effective similarity measure, the Jaccard index [19]. Given two sets, A and B, the Jaccard index follows one of the three forms according to the specific situations: $|A \cap B|/|A \cup B|$, $|A \cap B|/\max\{|A|, |B|\}$, or $|A \cap B|/\min\{|A|, |B|\}$. The Jaccard index ranges in $[0, 1]$ rather than $\{0, 1\}$, so it can more accurately reflect the AFD and AC for a given attack. Its superiority over the indicator function in set similarity has also been verified in various research areas [20]. Theoretically, each existing indicator function-based measure (including the ones used in other adversarial attack studies) can be transformed into a Jaccard index-based form. In this study, the following two measures are defined for top-k multi-label attack. The first measure is the soft label consistency (SLC):

$$\text{AFD}_{slc}(\boldsymbol{x}, \boldsymbol{z}) = |Y \cap \hat{Y}_k(\boldsymbol{x}+\boldsymbol{z})|/\min\{|Y|, |\hat{Y}_k(\boldsymbol{x}+\boldsymbol{z})|\}. \tag{5}$$

AFD_{slc} measures better the two possible attacks $\boldsymbol{x}_1^1$ and $\boldsymbol{x}_1^2$ (Fig. 1(a)): $\text{AFD}_{slc}(\boldsymbol{x}_1, \boldsymbol{z}_1^1) = 2/3 > \text{AFD}_{slc}(\boldsymbol{x}_1, \boldsymbol{z}_1^2) = 1/3$, whereas when the existing measure is used, $\text{AFD}_{lc}(\boldsymbol{x}_1, \boldsymbol{z}_1^1) = \text{AFD}_{lc}(\boldsymbol{x}_1, \boldsymbol{z}_1^2) = 0$. Obviously, the refined measure in Eq. (5) is superior because it assigns distinct values to different attacks. In other words, the Jaccard idnex-based measure (e.g., Eq. (5)) can better distinguish different adversarial attacks[4].

The second new AFD measure concerns the label's prediction confidence. Based on the Jaccard index, a more refined measure, soft label flip (SLF), is defined as follows.

$$\text{AFD}_{slf}(\boldsymbol{x}, \boldsymbol{z}) = |Y_{lf} \cap Y|/|Y_{lf} \cup Y| = |Y_{lf}|/|Y|. \tag{6}$$

The succeeding parts describe our constructed optimization problems and new solving strategy. Fig. 4 illustrates our method for targeted and untargeted attacks.

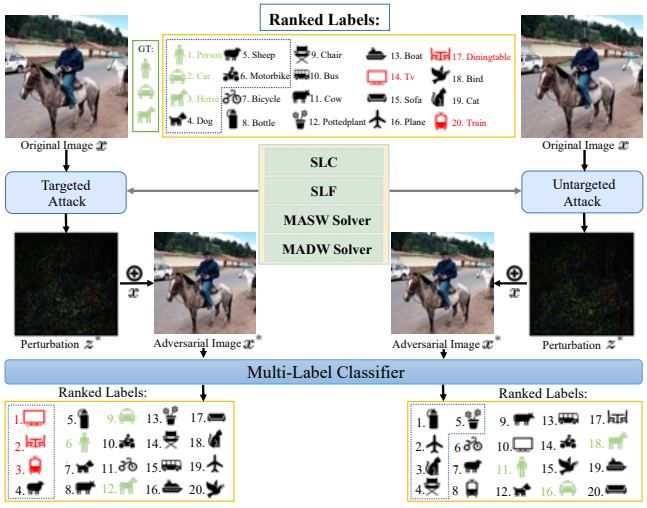


Fig. 4. Illustration for our methodology. The top-3 attack is considered. The labels within the dotted box have confidence greater than 0.5. $Y_t = \{\text{Tv, Diningtable, Train}\}$. $Y = \{\text{Person, Car, Horse}\}$.

2) Problem for Untargeted Attack Implementation. The untargeted attack aims to find a small perturbation so that the labels in Y are excluded from the top-k prediction and are below a certain threshold 0.5. Hence, the following constrained optimization problem is constructed.

$$\min_{\boldsymbol{z}} \left[\text{AFD}_{slc},\ \text{AFD}_{slf},\ ||\boldsymbol{z}||_2^2/2\right],\ \text{s.t. } \boldsymbol{x}+\boldsymbol{z} \in [-1, 1]^n, \tag{7}$$

where n is the feature dimension. $\text{AFD}_{slc} = 0$ is equivalent to excluding all the labels in Y from $\hat{Y}_k(\boldsymbol{x}+\boldsymbol{z})$, that is, $Y \cap \hat{Y}_k(\boldsymbol{x}+\boldsymbol{z}) = \emptyset$ which is also equivalent to satisfying all the following $|Y|$ inequalities:

$$f_j(\boldsymbol{x}+\boldsymbol{z}) \leq f_{[k+1]}(\boldsymbol{x}+\boldsymbol{z}),\ \forall j \in Y. \tag{8}$$

In general, not all inequalities in (8) can be satisfied. AFD_{slc} is proportional to the number of violated constraints among those in (8). Following [40], we introduce slack variables $\boldsymbol{\zeta}'$, and minimizing AFD_{slc} can be modeled as a novel optimization problem with the least constraint violation:

$$\min_{\boldsymbol{\zeta}', \boldsymbol{z}} \frac{1}{2}||\boldsymbol{\zeta}'||_2^2,\ \textbf{s.t. } f_j(\boldsymbol{x}+\boldsymbol{z}) \leq f_{[k+1]}(\boldsymbol{x}+\boldsymbol{z}) + \zeta'_j, \forall j \in Y, \tag{9}$$

where $\boldsymbol{\zeta}' = [\zeta'_1, \cdots, \zeta'_{|Y|}]^T$. Since $\boldsymbol{f}(\boldsymbol{x})$ is the label probability vector, if $f_j(\boldsymbol{x}+\boldsymbol{z}) \leq f_{[k+1]}(\boldsymbol{x}+\boldsymbol{z})$ holds, then $\zeta'_j = 0$. Similarly, minimizing AFD_{slf} in Eq. (6) can be modeled as:

[3]Our initial statistics on VOC 2012 show that about 40% samples have no ideal attacks under the worst targeted case using the SOTA method in [11].

[4]The ablation study in Fig. 8(a), the statistical analysis, and Proposition 1 in Section III-B demonstrate that the defects of existing measures can be compensated by the proposed measures.

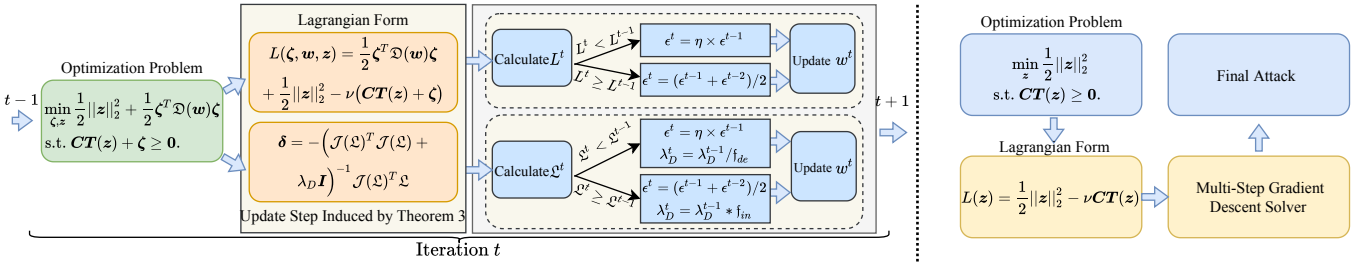


Fig. 5. Comparison of the main steps of our SPW-based solving (left) and existing mainstream algorithms (right) for untargeted attack.

$$\min_{\boldsymbol{\zeta}'',\boldsymbol{z}} \frac{1}{2}||\boldsymbol{\zeta}''||_2^2, \ \textbf{s.t.}\ f_j(\boldsymbol{x}+\boldsymbol{z}) \le 0.5+\zeta_j'', \forall j \in Y, \tag{10}$$

where $\boldsymbol{\zeta}'' = [\zeta_1'', \cdots, \zeta_{|Y|}'']^T$. Combining (9), (10), and $||\boldsymbol{z}||_2^2/2$, the following optimization problem is obtained:

$$\begin{aligned} \min_{\boldsymbol{\zeta}',\boldsymbol{\zeta}'',\boldsymbol{z}} \ & \frac{1}{2}||\boldsymbol{z}||_2^2 + \frac{\lambda_1}{2}||\boldsymbol{\zeta}'||_2^2 + \frac{\lambda_2}{2}||\boldsymbol{\zeta}''||_2^2 \\ \text{s.t. } & f_j(\boldsymbol{x}+\boldsymbol{z}) \le f_{[k+1]}(\boldsymbol{x}+\boldsymbol{z}) + \zeta_j', \forall j \in Y, \\ & f_j(\boldsymbol{x}+\boldsymbol{z}) \le 0.5+\zeta_j'', \forall j \in Y; \ \ \boldsymbol{x}+\boldsymbol{z} \in [-1,1]^n, \end{aligned} \tag{11}$$

where λ_1 and λ_2 are balancing factors for these three optimization objectives which can be determined through a sensitivity test. The proposed constrained optimization problem in (11), derived from Eqs. (5) and (6), can obtain the optimal solution, while those constrained optimization problems in existing works [9], [11] cannot.

3) Problem for Universal Untargeted Attack Implementation. Sample-wise untargeted attack in (7) can be extended to universal untargeted attack, with the latter aiming to find an adversarial perturbation $\boldsymbol{z}$ shared by all samples [31]. Assume $X_t = \{(\boldsymbol{x}_1, Y_1), \cdots, (\boldsymbol{x}_i, Y_i), \cdots, (\boldsymbol{x}_N, Y_N)\}$, where N is the number of samples in X_t. Following [11], [12], an iterative algorithm on the dataset X_t can be derived to find a perturbation $\boldsymbol{z}$ shared by X_t. Specifically, Algorithm 1 (without using the projection) is used to find an adversarial perturbation $\boldsymbol{z}_i^*$ for a sample $(\boldsymbol{x}_i, Y_i)$. Then a projection operation $\mathcal{P}_\tau(\cdot)$ is used to update $\boldsymbol{z}$ [18]. Algorithm 2 shows the processing steps, where for simplicity and efficiency, early-stopping [32] terminates the algorithm.

4) Problem for Targeted Attack Implementation. Since the targeted attack aims to find a small perturbation $\boldsymbol{z}$ so that $Y_t = \hat{Y}_k(\boldsymbol{x}+\boldsymbol{z})$, replacing Y in Eq. (5) with Y_t is necessary, and minimizing $-\text{AFD}_{slc}$ is expected. Meanwhile, regardless of whether it is a targeted attack or an untargeted attack, the probability of per label in Y is required to be less than 0.5. Therefore, AFD_{slf} remains unchanged. Then we obtain:

$$\min_{\boldsymbol{z}} \ [-\text{AFD}_{slc}, \ \text{AFD}_{slf}, \ ||\boldsymbol{z}||_2^2/2], \ \text{s.t. } \boldsymbol{x}+\boldsymbol{z} \in [-1,1]^n. \tag{12}$$

Likewise, (12) then becomes the following optimization problem with the least constraint violation:

$$\begin{aligned} \min_{\boldsymbol{\zeta}',\boldsymbol{\zeta}'',\boldsymbol{z}} \ & \frac{1}{2}||\boldsymbol{z}||_2^2 + \frac{\lambda_1}{2}||\boldsymbol{\zeta}'||_2^2 + \frac{\lambda_2}{2}||\boldsymbol{\zeta}''||_2^2 \\ \text{s.t. } & f_j(\boldsymbol{x}+\boldsymbol{z}) \le f_{[k+1]}(\boldsymbol{x}+\boldsymbol{z}) + \zeta_j'; \ \forall j \in C \setminus Y_t, \\ & f_j(\boldsymbol{x}+\boldsymbol{z}) + \zeta_j' \ge f_{[k+1]}(\boldsymbol{x}+\boldsymbol{z}); \ \forall j \in Y_t, \\ & f_j(\boldsymbol{x}+\boldsymbol{z}) \le 0.5+\zeta_j''; \ \forall j \in Y; \ \boldsymbol{x}+\boldsymbol{z} \in [-1,1]^n. \end{aligned} \tag{13}$$

This study focuses on AFD. However, AC can also be considered in targeted attack.

B. Theoretical Analysis of Weighting Strategy

Although directly solving (11) can also obtain a solution with least constraint violation, the solution may not satisfy the constraints as much as possible, that is, the number of successfully attacked labels is not the largest in multi-label attack. Meanwhile, the strategy of treating constraints indiscriminately leads to fluctuation in the solution process (Theorem 1). Existing works do not take these into account. Then a weighting strategy can be applied to give various constraints different priorities during the solution process. Let $\boldsymbol{w} = [\lambda_1 \boldsymbol{w}', \lambda_2 \boldsymbol{w}'']^T, \boldsymbol{\zeta} = [\boldsymbol{\zeta}', \boldsymbol{\zeta}'']^T$, and $\boldsymbol{CT}(\boldsymbol{z}) = \left[\left[f_{[k+1]}(\boldsymbol{x}+\boldsymbol{z}) - f_j(\boldsymbol{x}+\boldsymbol{z})\right]_{j=1}^{|Y|}, \ \left[0.5 - f_j(\boldsymbol{x}+\boldsymbol{z})\right]_{j=1}^{|Y|}\right]^T$, where $\boldsymbol{w}'$ and $\boldsymbol{w}''$ are the weights of the constraints. A natural weighted version of (11) is described below and its theoretical analysis is performed in detail:

$$\begin{aligned} \min_{\boldsymbol{\zeta},\boldsymbol{z}} \ & \frac{1}{2}||\boldsymbol{z}||_2^2 + \frac{1}{2}\boldsymbol{\zeta}^T \mathfrak{D}(\boldsymbol{w})\boldsymbol{\zeta} \\ \text{s.t. } & \boldsymbol{CT}(\boldsymbol{z}) + \boldsymbol{\zeta} \ge \boldsymbol{0}. \end{aligned} \tag{14}$$

where $\mathfrak{D}(\boldsymbol{w})$ is the diagonal matrix of $\boldsymbol{w}$ and the constraint $\boldsymbol{x}+\boldsymbol{z} \in [-1,1]^n$ is satisfied by using the projection method [18]. In order to facilitate theoretical analysis, following [40], (14) is extended to the following optimization problem, with more details being shown in the supplementary materials:

$$\begin{aligned} & \min_{\boldsymbol{z}} \frac{1}{2}||\boldsymbol{z}||_2^2 \\ & \text{s.t. } (\boldsymbol{\zeta}, \boldsymbol{z}) \text{ solves } \begin{cases} \min_{\boldsymbol{\zeta},\boldsymbol{z}} \frac{1}{2}\boldsymbol{\zeta}^T \mathfrak{D}(\boldsymbol{w})\boldsymbol{\zeta} \\ \text{s.t. } \boldsymbol{0} \le \boldsymbol{CT}(\boldsymbol{z}) + \boldsymbol{\zeta} \end{cases}. \end{aligned} \tag{15}$$

(15) means AFD_{slc} and AFD_{slf} are minimized first, and then finding $\boldsymbol{z}$ with the smallest ℓ_2 norm. Minimizing firstly $\frac{1}{2}||\boldsymbol{z}||_2^2$ is not desired, because if so, obviously $\boldsymbol{z} = \boldsymbol{0}$, but AFD_{slc} and AFD_{slf} are large. Then we denote:

$$\begin{aligned} \Theta(\boldsymbol{z}) &= \min_{\boldsymbol{\zeta}} \{\frac{1}{2}\boldsymbol{\zeta}^T \mathfrak{D}(\boldsymbol{w})\boldsymbol{\zeta} : \boldsymbol{CT}(\boldsymbol{z}) + \boldsymbol{\zeta} \ge 0\} \\ &= \frac{1}{2}[\boldsymbol{CT}(\boldsymbol{z})]_-^T \mathfrak{D}(\boldsymbol{w})[\boldsymbol{CT}(\boldsymbol{z})]_-, \end{aligned} \tag{16}$$

where $[\boldsymbol{CT}(\boldsymbol{z})]_- = \min\{\boldsymbol{0}, \boldsymbol{CT}(\boldsymbol{z})\}$. Then we have the following derivative of $\Theta(\boldsymbol{z})$ with respect to $\boldsymbol{z}$:

$$\nabla\Theta(\boldsymbol{z}) = \mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}))^T \mathfrak{D}(\boldsymbol{w})[\boldsymbol{CT}(\boldsymbol{z})]_-, \tag{17}$$

where $\mathcal{J}(\cdot)$ is the jacobian matrix. Denoting $\boldsymbol{v} = \boldsymbol{CT}(\boldsymbol{z}) + \boldsymbol{\zeta}$, then (15) can be extended as follows:

$$\min_{\boldsymbol{z}} \frac{1}{2}||\boldsymbol{z}||_2^2 \ \text{s.t. } \boldsymbol{F}(\boldsymbol{z}, \boldsymbol{\zeta}, \boldsymbol{v}) = \boldsymbol{0}, \ (\boldsymbol{\zeta}, \boldsymbol{v}) \in \Omega, \tag{18}$$

where $\Omega = \{(\boldsymbol{\zeta}, \boldsymbol{v}) : \boldsymbol{0} \le \boldsymbol{\zeta} \perp \boldsymbol{v} \ge \boldsymbol{0}\}$ and $F(\boldsymbol{z}, \boldsymbol{\zeta}, \boldsymbol{v}) = \begin{bmatrix} -\mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}))\mathfrak{D}(\boldsymbol{w})\boldsymbol{\zeta} \\ \boldsymbol{CT}(\boldsymbol{z}) + \boldsymbol{\zeta} - \boldsymbol{v} \end{bmatrix}$. Then we have the following theorem:

Theorem 2. *Let $(\boldsymbol{z}^*, \boldsymbol{\zeta}^*, \boldsymbol{v}^*)$ be a local minimum of (18). Denote $\iota_1 = \{j : \zeta_j^* > 0 = v_j^*\}, \iota_2 = \{j : \zeta_j^* = 0 = v_j^*\}, \iota_3 = \{j : \zeta_j^* = 0 < v_j^*\}$. If $-\sum_{j=1}^{2|Y|} \zeta_j^* w_j \nabla^2 CT_j(\boldsymbol{z})$ is positive definite, then there exist $\boldsymbol{\eta}_1^* \in \Re^n, \boldsymbol{\xi}_{\iota_2}^* \in \Re^{|\iota_2|}, \boldsymbol{\xi}_{\iota_2}^{**} \in \Re^{|\iota_2|}$, which satisfy $\xi_j^* \xi_j^{**} = 0$, or $\xi_j^* \le 0$, $\xi_j^{**} \le 0, \forall j \in \iota_2$. We have*

$$\begin{aligned} &\boldsymbol{z}^* + \mathcal{J}(\boldsymbol{CT}_{\iota_2}(\boldsymbol{z}^*))\boldsymbol{\xi}_{\iota_2}^{**} - \Big(\sum_{j=1}^{2|Y|} \zeta_j^* w_j \nabla^2 CT_j(\boldsymbol{z}^*) + \\ &\qquad \mathcal{J}(\boldsymbol{CT}_{\iota_1}(\boldsymbol{z}^*))^T \mathcal{J}((\boldsymbol{CT}(\boldsymbol{z})^T \mathfrak{D}(\boldsymbol{w}))_{\iota_1}\Big)\boldsymbol{\eta}_1^* = \boldsymbol{0}, \\ &\mathcal{J}((\boldsymbol{CT}(\boldsymbol{z})^T \mathfrak{D}(\boldsymbol{w}))_{\iota_2})\boldsymbol{\eta}_1^* + \boldsymbol{\xi}_{\iota_2}^* + \boldsymbol{\xi}_{\iota_2}^{**} = \boldsymbol{0}. \end{aligned} \tag{19}$$

The proof is in Appendix. Theorem 2 describes a necessary condition for the local minimums of (18). From Theorem 2, it can be seen that the weights $\boldsymbol{w}$ determine the necessary conditions for a local minimum. And the different weights allow different constraints to be satisfied. Further, denote $\boldsymbol{G}(\boldsymbol{z}, \boldsymbol{\zeta}, \boldsymbol{v})$ as follows:

$$\begin{aligned} &\boldsymbol{G}(\boldsymbol{z}, \boldsymbol{\zeta}, \boldsymbol{v}) = \\ &[-\mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}))\mathfrak{D}(\boldsymbol{w})\boldsymbol{\zeta},\ \boldsymbol{CT}(\boldsymbol{z}) + \boldsymbol{\zeta} - \boldsymbol{v},\ \min\{\boldsymbol{\zeta}, \boldsymbol{v}\}]^T. \end{aligned} \tag{20}$$

Assuming $\boldsymbol{G}(\cdot, \cdot, \cdot)$ is a Lipschitz continuous mapping, (15) can be extended as follows (more details are shown in the supplementary materials):

$$\min_{\boldsymbol{z}} \ ||\boldsymbol{z}||_2^2/2 \text{ s.t. } \boldsymbol{G}(\boldsymbol{z}, \boldsymbol{\zeta}, \boldsymbol{v}) = \boldsymbol{0}. \tag{21}$$

Then, we get a more concise theorem as follows:

Theorem 3. *Let $(\boldsymbol{z}^*, \boldsymbol{\zeta}^*, \boldsymbol{v}^*)$ be a local minimum of (21). Then there exist $\varsigma^* \in \Re_+, \boldsymbol{\eta}_1^* \in \Re^n, \mathfrak{v}_{\iota_2}^* \in \Re^{|\iota_2|}$, which satisfy $\mathfrak{v}_j^* \in [0, 1],\ \forall j \in \iota_2$. We have:*

$$\begin{aligned} &-\Big(\mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}^*))^T \mathfrak{D}(\mathfrak{v}^*) \mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}^*)^T \mathfrak{D}(\boldsymbol{w})) \\ &\qquad + \sum_{j=1}^{2|Y|} \zeta_j^* w_j \nabla^2 CT_j(\boldsymbol{z}^*)\Big)\boldsymbol{\eta}_1^* + \varsigma^* \boldsymbol{z}^* = \boldsymbol{0}. \end{aligned} \tag{22}$$

The proof is in Appendix. Similarly, according to Theorem 3, different weights can determine which constraints participate in optimization and are satisfied.

According to Theorems 1, 2 and 3, we have the following two corollaries, which illustrate the effectiveness of weighting strategy. Corollary 1 states that weighting strategy for $\boldsymbol{\zeta}$ maximizes the number of constraints that are met, which means that the most labels are successfully attacked when there is no ideal solution. Meanwhile, Corollary 2 shows that weighting strategy can stabilize the solution only at a local minimizer. All in all, weighting strategy finds a stable optimal solution. And the proofs of these two corollaries are shown in Appendix.

Corollary 1. *Denote $\iota_{j'}^c \subseteq S = \{1, 2, \cdots, 2|Y|\}; j' = 1, \cdots, l$, which satisfy $\cup_{j'=1}^{l} \iota_{j'}^c = S$, $\{\boldsymbol{z} : CT_j(\boldsymbol{z}) \ge 0, j \in \iota_{j'}^c\} \cap \{\boldsymbol{z} : CT_j(\boldsymbol{z}) \ge 0, j \in \iota_{k'}^c\} = \emptyset$, $\iota_{j'}^c \cup \iota_{k'}^c = \emptyset, j' \ne k'$. Without loss of generality, assume that $|\iota_1^c| \ge |\iota_2^c| \ge \cdots \ge |\iota_l^c|$. If we set $\boldsymbol{w}_{\iota_1^c} = \boldsymbol{1}$, and $\boldsymbol{w}_{\iota_{j'}^c} = \boldsymbol{0}, \forall j' = 2, \cdots, l$. Then, the number of labels that are successfully attacked is the largest.*

Corollary 2. *According to the settings in Corollary 1, if $\boldsymbol{w}_{\iota_1^c} = \boldsymbol{1}$, and $\boldsymbol{w}_{\iota_{j'}^c} = \boldsymbol{0}$, then the weighting strategy makes the optimization process converge.*

C. Optimization with Self-paced Weighting

The Lagrangian form of (14) is:

$$L(\boldsymbol{\zeta}, \boldsymbol{w}, \boldsymbol{z}) = \frac{1}{2}||\boldsymbol{z}||_2^2 + \frac{1}{2}\boldsymbol{\zeta}^T \mathfrak{D}(\boldsymbol{w})\boldsymbol{\zeta} - \nu(\boldsymbol{CT}(\boldsymbol{z}) + \boldsymbol{\zeta}), \tag{23}$$

where ν is a hyperparameter[5].

Inspired by our experimental and theoretical analysis, we next devise an iterative optimization process in which a novel weighting strategy is utilized. In each iteration of our iterative optimization process, we still measure the constraint difficulty using the value of a constraint's slack variable achieved in the previous iteration. Let ϵ be a threshold to distinguish normal and difficult constraints. An ad hoc strategy is to set w_j to 0 if ζ_j^{t-1} in the $t-1$th iteration is larger than ϵ and 1 otherwise. In this strategy, the highly difficult constraints (constraints satisfying $\zeta^{t-1} > \epsilon$) are excluded in the next iteration as their weights are set as zero. This simple strategy is quite sensitive to the choice of ϵ. A soft scheme is defined as:

$$w_j^t = \epsilon / \max\{\zeta_j^{t-1}, \epsilon\} = 1 / \max\{\zeta_j^{t-1}/\epsilon, 1\}, \tag{24}$$

Moreover, motivated by self-paced learning (SPL), which places dynamic weights on train samples according to their dynamic learning difficulties [23], we propose a self-paced weighting (SPW) strategy. Initializing $\boldsymbol{w} = \boldsymbol{1}$, each iteration of our SPW strategy comprises three main steps:

- Perform gradient descent (GD) to minimize (23) in the first iteration step.
- In the second iteration step, the threshold ϵ is updated.
- In the third iteration step, the weights for each constraint (or slack variable) are updated using Eq. (24). Thereafter, return to the first iteration step.

During each update of ϵ in SPL, its value increases gradually, e.g., $\epsilon^t = 1.01 \times \epsilon^{t-1}$ at the tth iteration. Consequently, the weights of more difficult constraints will become one in the rest solving procedure. However, we design a more effective update method where ϵ is updated according to Eq. (23), denoted as L^t. If L^t decreases compared with L^{t-1}, then ϵ is increased to allow more constraints to participate in the optimization process, i.e., $\epsilon^t = \eta \times \epsilon^{t-1}$ $(\eta > 1)$; otherwise, ϵ is reduced to limit the participation of difficult constraints by $\epsilon^t = (\epsilon^{t-1} + \epsilon^{t-2})/2$. Our SPW achieves better results than the SPL method, independently if an ideal solution exists. The corresponding results are presented in the experiments.

Algorithm 1 reports the details of our implementation for the **M**ulti-label Untargeted **A**ttack based on our **SPW**, namely, **MASW**. To ensure the optimization efficiency, similar to [24], only one GD at each iteration is performed on L^t. Fig. 5 depicts the difference between our algorithm and existing methods in actual execution. As an example, take the linear model $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{W}^T\boldsymbol{x} + \boldsymbol{b}$ to analyze the time complexity of Algorithm 1 using the big O notation. In Step 3 of Algorithm 1, the time complexity of one forward propagation is $|C|(n+1)$. The time complexity of calculating the Lagrangian function of (14) is $n + 7|Y|$, and that of calculating the index $[k]$ is $|C| \log k$. The time complexity of one back-propagation is $n(1 + 2|Y|) + 4|Y|$. L^t in step 4 can be obtained through

[5]Besides the optimization objective (23), we also propose an additional novel method based on our theoretical analysis, shown in Section S.II-C of the supplementary file.

Algorithm 1 SPW-based Untargted Attack (**MASWU**).

Input: $\boldsymbol{x}$, Y, λ_1, λ_2, model $\boldsymbol{f}(\cdot)$, T, step size η for ϵ
Output: Adversarial perturbation $\boldsymbol{z}^*$

Initialize $L^0 = +\infty$, ϵ^0, $\zeta_j^{'0}$, $\zeta_j^{''0}$, $w_j^{'0}$ and $w_j^{''0}$;
for t = 1 to T
  Update $\zeta_j^{'t}$, $\zeta_j^{''t}$, $\boldsymbol{z}^t$ by solving (14);
  Calculate Lagrangian value L^t of (14);
  if $L^t < L^{t-1}$ **then** $\epsilon^t = \eta * \epsilon^{t-1}$; $\boldsymbol{z}^* = \boldsymbol{z}^t$;
  else $\epsilon^t = (\epsilon^{t-1} + \epsilon^{t-2})/2$; $L^t = L^{t-1}$; $\epsilon^{t-1} = \epsilon^{t-2}$;
  if $||\nabla_{\boldsymbol{z}} L^t||_2 \leq 10^{-3}$ **then** break;
  Update $w_j^{'t}$ and $w_j^{''t}$ by using Eq. (24);
return $\boldsymbol{z}^*$.

Algorithm 2 Universal untargted Attack (**MASWUv**).

Input: X_t, λ_1, λ_2, E, $\boldsymbol{f}(\cdot)$, T, η
Output: Universal perturbation $\boldsymbol{z}^*$

Initialize $\boldsymbol{z}^* = \boldsymbol{0}$, $E_1, E_2 = 0$, $\tau_1, \tau_2 = 0$;
while true **do**
  for $(\boldsymbol{x}_i, Y_i) \in X_t$ **do**
    $\boldsymbol{z}_i^*$ = MASWU($\boldsymbol{x}_i$, Y_i, λ_1, λ_1, $\boldsymbol{f}$, T, η);
    $\boldsymbol{z}^* = \mathcal{P}_\tau(\boldsymbol{z}^* + \boldsymbol{z}_i^*)$;
  Calculate SASR (Eq. (26)) and LFR (Eq. (28)) on X_t;
  if SASR $< \tau_1$ **then** $E_1 = E_1 + 1$;
  else τ_1 = SASR; $E_1 = 0$;
  if LFR $< \tau_2$ **then** $E_2 = E_2 + 1$;
  else τ_2 = LFR; $E_2 = 0$;
  if $E_1 > E$ and $E_2 > E$ **then** break;
return $\boldsymbol{z}^*$.

step 3. As the time complexity of Steps 5 and 6 is constant, we can disregard them. Finally, the time complexity of Step 7 is n. Therefore, the total time complexity of Algorithm 1 is approximately O$(T(|C|n + |C|^2))$. Our theoretical analysis and Fig. 3(c) show that our SPW-based method is more stable. SPW better attenuates the fluctuations during optimization. Therefore, T is small in actual attack. Then, the actual time cost is small. Further, a statistical comparison of the time cost reveals that except for some methods that use single-step or few-step GD, such as [1], our method achieves lower or comparable time cost, shown in Section V.

Algorithm 2 outlines the calculation steps for universal untargeted attack, extended by Algorithm 1. Additionally, following the same approach as Algorithm 1, we can easily derive an algorithm for targeted attack, as presented in Algorithm 3.

Algorithm 3 SPW-based Targted Attack (**MASWT**).

Input: $\boldsymbol{x}$, Y, Y_t, λ_1, λ_2, $\boldsymbol{f}(\cdot)$, T, η
Output: Adversarial perturbation $\boldsymbol{z}^*$

Initialize $L^0 = +\infty$, ϵ^0, $\zeta_j^{'0}$, $\zeta_j^{''0}$, $w_j^{'0}$ and $w_j^{''0}$;
for t = 1 to T
  Update $\zeta_j^{'t}$, $\zeta_j^{''t}$, $\boldsymbol{z}^t$ by solving the weighted (13);
  Calculate Lagrangian objective L^t;
  if $L^t < L^{t-1}$ **then** $\epsilon^t = \eta * \epsilon^{t-1}$;$\boldsymbol{z}^* = \boldsymbol{z}^t$;
  else $\epsilon^t = (\epsilon^{t-1} + \epsilon^{t-2})/2$;$L^t = L^{t-1}$;$\epsilon^{t-1} = \epsilon^{t-2}$;
  if $||\nabla_{\boldsymbol{z}} L^t||_2 \leq 10^{-3}$ **then** break;
  Update $w_j^{'t}$ and $w_j^{''t}$ by using Eq. (24);
return $\boldsymbol{z}^*$.

V. Experiments

A. Experimental Setup

Datasets and Models. Four benchmark multi-label learning datasets VOC 2012, COCO 2014, NUS-WIDE [25], and Open Images [26] are employed.

- **VOC 2012/COCO 2014.** The training and validation sets comprise 5,717/82,081 and 5,823/40,137 images from 20/80 categories, respectively. And the average number of positive label per image is 1.43/3.67.
- **NUS-WIDE.** [14] provides a variant of this dataset that contains 220,000 images from 81 classes. We adopt the standard 70-30 train-test split suggested in [14].
- **Open Images.** This large-scale dataset comprises more than 9,000,000 training images and 125,436 test images from 9,605 classes.

Moreover, the Inception-v3 [27], ResNet50 [28], and Tresnet [29] models are used. Following the settings in [11], [14], Inception-v3 and ResNet50 achieve 87.2 % mAP and 93.6 % mAP on VOC 2012 and COCO 2014, respectively. And Tresnet achieves 66.1 % mAP and 86.9 % micro mAP on NUS-WIDE and Open Images, respectively.

Competing Methods. The competing methods include: Fast gradient sign method (FGSM [1]), Momentum iterative fast gradient sign method (MFGSM [7]), Projected gradient descent (PGD [18]), Rank I [9], Multi-label DeepFool (MLDF [9]) and Carlini & Wagner (MLCW [9]) attacks, Multi-label attack by linear programming (MLALP [17]), Top-k universal untargeted attck (kUv [12]) and untarged attck by DeepFool (kFool [12]), T$_k$ML [11], Generative Adversarial Multi object Attacks (GAMA [10]), Local Patch Difference (LPD [37]), Top-k Attack with Label Correlation (T$_k$ALC [46]) and Top-k Measure Imperceptible Attack (T$_k$MIA [45]). The settings in [11], [33] are used. Three suffixes U, Uv and T mean the untargeted attack, universal untargeted attack, and targeted attack, respectively. Then our methods are denoted as MASWU, MASWUv and MASWT.

Settings. We set to $\nu = 1$ in our experiments. Similar to [9], [11], GD method is used to minimize $L(\boldsymbol{\zeta}, \boldsymbol{w}, \boldsymbol{z})$. The Lagrangian form of (13) is similar to Eq. (23). We do not explicitly minimize $\frac{1}{2}||\boldsymbol{z}||_2^2$ in Eq. (23) in actual execution, but following [11], [18], we use a projector operation $\mathcal{P}_\tau(\cdot)$ to control $||\boldsymbol{z}||_2 \leq \tau$, where τ is a hyperparameter, and following [11], $\tau = 2, 10, 100$ for targeted, untargeted and universal untargeted attacks, respectively. We set $\lambda_1 = \lambda_2 = 0.5$. η is set to 1.01, and ϵ is initialized to 0.01. We record the results when $T = 300$. Furthermore, $\boldsymbol{\zeta}'$, $\boldsymbol{\zeta}''$, $\boldsymbol{w}'$, and $\boldsymbol{w}''$ are initialized to obey a uniform distribution of [0, 1]. Following [11], the learning rate for GD is set to 0.01. 3000 images from the validation set of each benchmark dataset are selected to build X_t for universal untargeted attack. Besides, we apply early stopping [32] on X_t to terminate Algorithm 2. The patience of early stopping E is set to 40. Following [11], for PGD, Rank I, MLCW, kFool, T$_k$ML, T$_k$ALC and T$_k$MIA, the basic experimental setups are the same as above. Following

[11], [33], for FGSM and MFGSM, T is set to 1 and 40, respectively, and the learning rate is set to 0.03. For MLDF and MLALP, T is set to 300. For LPD and GAMA, the original experimental setups are used. For all universal untargeted attack methods, the settings in [11] are used. Following [11], [33], 1000 successfully classified images are selected from the validation set of each dataset to build X_v to evaluate performance. For efficiency, a batch of 50 images is used to attack. All experiments are conducted on a Linux server with four RTX 3090 24Gb graphics cards. Additionally, all random seeds are fixed, and each method is executed three times, with the mean results being reported.

Evaluation Metrics. Following [11], we use the attack success rate (ASR) to measure the attack performance, which is defined as follows:

$$\text{ASR} = 1 - \sum_{\boldsymbol{x}\in X_v} \text{AFD}_{lc}(\boldsymbol{x}, \boldsymbol{z})/N_v, \tag{25}$$

where X_v is the evaluation dataset, and $N_v = |X_v|$. ASR represents the proportion of completely successful attacks. A soft ASR (SASR) can be introduced by AFD_{slc}.

$$\text{SASR} = 1 - \sum_{\boldsymbol{x}\in X_v} \text{AFD}_{slc}(\boldsymbol{x}, \boldsymbol{z})/N_v. \tag{26}$$

The higher both ASR (%) and SASR (%), the better the attack. Similar to [11], the metric, namely Pert, is used to evaluate the perceptual quality of attacks.

$$\text{Pert} = \sum_{\boldsymbol{x}\in X_v} (||\boldsymbol{z}||_2/\#\text{ of pixels of } \boldsymbol{x})/N_v. \tag{27}$$

The smaller Pert ($\times 10^{-2}$), the less visible $\boldsymbol{z}$. Another metric, namely label flip rate (LFR), is devised by AFD_{slf} as follows:

$$\text{LFR} = 1 - \sum_{\boldsymbol{x}\in X_v} \text{AFD}_{slf}(\boldsymbol{x}, \boldsymbol{z})/N_v. \tag{28}$$

A larger LFR means higher attacking quality.

B. Untargeted Attack

Results. Table I reports the top-k attack performance for $k = 3, 5, 10$. The results infer that the proposed MASWU achieves the best or comparable results. FGSMU uses single-step GD ($T = 1$) and a large learning rate, still requiring a larger Pert and performing poorly in terms of ASR, SASR, and LFR. For MFGSMU, which utilizes a few-step GD ($T = 40$) and the momentum, although it requires the smallest Pert, its performance is poor. Moreover, the multi-label attack methods MLCWU and MLDFU explicitly lower the prediction confidence of the ground truth below a certain threshold but only achieve a low LFR. Both PGDU and kFool are initially designed for single-label learning. They usually use larger perturbation bound Pert to attack model, but the attack performance is still inferior to MASWU. MLALP uses the interior point method to solve linear programs, but its performance is still lower than that of MASWU. Compared with the SOTA method T_kMLU, MASWU achieves better results on ASR, SASR, and LFR for $T = 300$ while yielding similar Pert values. Same or smaller Perts imply that T_kMLU may just obtain suboptimal solutions. Both LPD and GAMA utilize complex generative models, yet rely on larger perturbation bounds (Pert), resulting in lower ASR, SASR, and LFR values compared to our method. Though T_kALCU and T_kMIAU use label correlation and design complex optimization problems in adversarial attacks respectively, they are still inferior to our method.

TABLE I
COMPARSION OF UNTARGETED ATTACK METHODS. BOLD NUMBERS HIGHLIGHT THE BEST RESULTS.

k	T	Method	VOC 2012 ASR	VOC 2012 SASR	VOC 2012 Pert	VOC 2012 LFR	COCO 2014 ASR	COCO 2014 SASR	COCO 2014 Pert	COCO 2014 LFR
3	1	FGSMU	23.00	31.06	3.99	47.30	18.10	51.62	7.31	40.55
	40	MFGSMU	17.20	24.70	**0.32**	53.74	22.20	56.87	0.56	37.35
	–	MLCWU	19.80	29.46	2.38	50.27	25.00	33.94	3.71	45.59
		MLDFU	18.00	29.17	1.74	45.28	23.60	31.19	2.76	44.37
		PGDU	85.17	93.21	3.67	85.42	90.43	97.72	5.71	91.75
		MLALP	64.97	76.52	0.71	86.59	55.90	69.72	0.72	85.19
		kFool	93.50	96.87	1.42	94.05	60.80	83.20	4.41	68.66
	300	T_kMLU	95.53	95.73	0.48	97.88	99.83	99.86	**0.51**	98.36
		LPD	89.90	90.11	0.91	90.42	92.30	93.86	0.73	91.97
		GAMA	91.30	92.57	0.73	91.05	94.40	95.39	0.69	93.78
		T_kALCU	95.97	96.11	0.52	98.18	99.40	99.52	0.56	98.01
		T_kMIAU	96.03	96.17	0.55	98.14	99.73	99.75	0.53	98.14
		MASWU	**97.13**	**97.34**	0.49	**99.83**	**99.90**	**99.95**	**0.51**	**99.46**
5	1	FGSMU	17.30	24.43	4.00	47.39	14.40	44.28	7.27	40.62
	40	MFGSMU	11.80	17.73	**0.33**	52.91	18.00	50.74	0.56	37.38
	–	MLCWU	18.00	26.95	2.45	51.74	23.70	31.29	3.86	48.74
		MLDFU	17.20	27.06	1.85	49.21	22.10	30.37	2.82	47.81
		PGDU	85.27	93.25	3.75	85.95	89.47	97.53	5.87	92.93
		MLALP	58.27	67.34	0.74	87.56	54.63	67.47	0.73	86.72
		kFool	93.60	95.78	2.35	95.81	65.10	84.69	7.81	76.91
	300	T_kMLU	93.33	93.76	0.52	98.06	99.67	99.71	**0.54**	98.67
		LPD	87.23	89.26	0.91	90.42	91.13	92.28	0.73	91.97
		GAMA	90.10	91.23	0.73	91.05	92.47	93.74	0.69	93.78
		T_kALCU	94.10	95.20	0.56	98.27	99.70	99.72	0.58	98.26
		T_kMIAU	94.37	95.20	0.57	98.36	99.53	99.60	0.55	98.39
		MASWU	**96.17**	**96.23**	0.52	**99.92**	**99.80**	**99.78**	**0.54**	**99.63**
10	1	FGSMU	9.90	14.81	3.98	47.30	11.20	35.34	7.29	40.55
	40	MFGSMU	6.60	10.12	**0.32**	53.24	14.30	41.83	**0.57**	37.01
	–	MLCWU	15.20	25.61	2.52	53.91	20.60	29.17	3.88	49.16
		MLDFU	17.10	26.54	1.87	52.38	20.00	27.97	2.86	48.14
		PGDU	85.00	92.76	3.85	86.20	87.67	97.08	6.08	93.58
		MLALP	49.17	61.32	0.75	88.29	52.70	65.29	0.75	87.91
		kFool	88.40	90.18	4.95	97.12	68.00	85.82	14.95	85.82
	300	T_kMLU	87.93	88.43	0.57	98.15	99.47	99.52	0.60	98.90
		LPD	86.73	88.09	0.91	90.42	90.10	91.12	0.73	91.97
		GAMA	88.07	89.18	0.73	91.05	90.17	91.06	0.69	93.78
		T_kALCU	89.17	90.29	0.60	98.38	99.40	99.51	0.64	98.48
		T_kMIAU	89.30	90.67	0.59	98.40	99.10	99.26	0.62	98.72
		MASWU	**91.40**	**91.89**	0.57	**99.97**	**99.93**	**99.98**	0.60	**99.89**

C. Universal Untargeted Attack

Results. Table II presents the results of each universal untargeted attack method on VOC 2012 and COCO 2014. Our method MASWUv use the smaller or smallest Pert to achieve the best ASR, SASR, and LFR. ✗ in Table II means that the method kUv takes an excessive amount of time (more than a week) but could not produce the result. Therefore we do not report this result. PGDUv usually utilizes the highest Pert but achieves lower ASR, SASR and LFR than our methods. Compared to T_kMLUv, T_kALCUv, and T_kMIAUv, the proposed method achieves higher ASR, SASR, and LFR only with a lower Pert.

D. Targeted Attack

Additional settings. Following [11], this paper considers three target types Y_t, namely, worst, random, and best cases. These cases mean that labels in Y_t have the lowest prediction scores, labels in Y_t are selected randomly, and labels in Y_t have the largest prediction scores, respectively.

Results. Table III reports the results for top-k targeted attacks under the best case. When similar Pert values are achieved on two datasets, MASWT outperforms all competitor methods considering the ASR, SASR, and LFR metrics for different k values. FGSMT uses a single-step GD ($T = 1$), but it requires larger bounds and achieves poor ASR, SASR, and

TABLE II
COMPARSION OF UNIVERSAL UNTARGETED ATTACK METHODS WITH $k = 2, 3, 5, 10$ ON VOC 2012 AND COCO 2014. BOLD NUMBERS MEAN THE BEST RESULTS. ✗ MEANS THAT THE METHOD CAN NOT OUTPUT THE RESULT.

		VOC 2012				COCO 2014			
k	Method	ASR	SASR	Pert	LFR	ASR	SASR	Pert	LFR
2	PGDUv	60.13	64.76	34.12	77.19	73.23	86.52	71.52	81.28
	kUv	✗	✗	✗	✗	72.90	86.18	51.38	79.56
	T$_k$MLUv	67.37	71.52	16.27	92.55	79.23	88.78	15.26	94.44
	T$_k$ALCUv	66.30	69.12	18.41	90.15	78.20	86.82	17.06	92.41
	T$_k$MIAUv	66.60	70.29	17.21	91.59	78.50	87.18	16.46	92.64
	MASWUv	**68.47**	**72.21**	**15.63**	**94.39**	**80.50**	**89.17**	**14.79**	**95.61**
3	PGDUv	57.73	63.41	39.54	78.92	71.57	76.29	74.11	83.95
	kUv	✗	✗	✗	✗	61.40	73.64	51.26	80.43
	T$_k$MLUv	64.43	68.11	17.28	96.11	78.07	87.24	16.36	94.82
	T$_k$ALCUv	62.10	66.13	20.18	94.01	77.50	84.47	19.31	92.91
	T$_k$MIAUv	63.20	66.71	19.29	94.28	78.30	85.18	18.69	93.12
	MASWUv	**64.90**	**69.13**	**16.59**	**96.41**	**80.33**	**87.34**	**15.62**	**95.71**
5	PGDUv	53.97	62.37	43.53	80.54	70.73	80.58	75.61	85.13
	kUv	✗	✗	✗	✗	69.80	78.82	52.34	85.51
	T$_k$MLUv	61.87	66.22	19.27	98.01	78.40	85.29	17.19	96.01
	T$_k$ALCUv	61.00	65.21	23.73	95.16	76.50	84.09	21.09	94.03
	T$_k$MIAUv	61.80	65.12	23.15	95.26	77.10	84.49	20.17	94.51
	MASWUv	**63.60**	**67.97**	**18.57**	**98.99**	**78.83**	**86.19**	**16.47**	**96.66**
10	PGDUv	35.33	38.94	49.55	81.02	63.83	66.79	79.62	85.98
	kUv	✗	✗	✗	✗	✗	✗	✗	✗
	T$_k$MLUv	41.63	47.93	22.84	98.69	74.07	83.49	18.61	97.92
	T$_k$ALCUv	40.20	47.11	25.19	96.92	74.10	82.21	23.34	96.42
	T$_k$MIAUv	41.00	47.41	23.81	97.09	74.40	83.02	22.14	97.10
	MASWUv	**42.97**	**49.53**	**22.56**	**99.39**	**76.07**	**85.79**	**18.25**	**98.39**

TABLE III
COMPARISON OF THE TARGETED ATTACK METHODS UNDER THE BEST CASE SCENARIO.

			VOC 2012				COCO 2014			
k	T	Method	ASR	SASR	Pert	LFR	ASR	SASR	Pert	LFR
3	1	FGSMT	5.70	62.83	0.80	23.41	6.70	52.00	1.48	29.68
	40	MFGSMT	11.50	66.19	**0.20**	26.68	16.30	59.13	**0.34**	35.18
		MLCWT	82.70	93.86	0.47	86.48	82.10	92.43	0.56	88.51
	300	MLDFT	52.30	78.13	0.87	53.53	58.30	72.20	1.42	65.83
		PGDT	37.17	76.83	0.81	49.19	37.37	62.16	1.43	71.47
		Rank I	92.10	95.86	0.43	93.44	99.10	99.21	0.56	90.04
		T$_k$MLT	93.13	97.39	0.43	93.35	99.03	99.11	0.58	90.78
		T$_k$ALCT	93.17	97.41	0.45	93.70	99.10	99.20	0.61	92.29
		T$_k$MIAT	93.20	97.47	0.47	95.29	99.17	99.23	0.63	93.45
		MASWT	**93.97**	**97.79**	0.43	**98.25**	**99.50**	**99.86**	0.60	**96.37**
5	1	FGSMT	4.30	65.52	0.78	23.26	5.80	60.05	1.51	28.61
	40	MFGSMT	3.50	73.05	**0.23**	26.26	7.10	67.88	**0.38**	33.28
		MLCWT	38.00	84.14	0.55	84.88	77.20	93.45	0.78	91.65
	300	MLDFT	13.80	74.53	0.92	21.38	39.90	70.54	1.45	56.03
		PGDT	21.37	78.04	0.84	44.18	26.67	69.54	1.44	69.07
		Rank I	84.20	93.16	0.49	96.56	98.80	99.34	0.67	97.28
		T$_k$MLT	86.03	96.55	0.49	96.62	99.03	99.20	0.69	96.83
		T$_k$ALCT	86.27	95.19	0.53	96.71	99.07	99.12	0.73	97.11
		T$_k$MIAT	86.40	95.25	0.52	97.10	99.10	99.20	0.71	97.34
		MASWT	**88.17**	**96.83**	0.49	**98.81**	**99.67**	**99.94**	0.69	**98.56**
10	1	FGSMT	3.70	66.28	0.81	24.15	3.70	61.89	1.69	28.31
	40	MFGSMT	2.70	80.07	**0.20**	22.88	4.40	74.71	**0.38**	30.51
		MLCWT	10.00	74.92	0.64	73.02	62.60	94.19	1.17	90.14
	300	MLDFT	7.80	82.41	0.99	40.56	16.90	73.82	1.47	37.03
		PGDT	18.03	77.88	0.85	36.56	6.37	70.49	1.43	57.43
		Rank I	69.90	91.11	0.52	97.09	97.10	99.02	0.81	97.79
		T$_k$MLT	75.33	96.27	0.53	97.13	98.83	99.47	0.85	97.47
		T$_k$ALCT	75.60	97.13	0.55	97.40	99.07	99.50	0.89	98.01
		T$_k$MIAT	75.77	97.89	0.56	97.84	99.10	99.54	0.89	98.18
		MASWT	**77.97**	**99.65**	0.53	**99.16**	**99.40**	**99.91**	0.86	**99.16**

TABLE IV
COMPARISON OF TARGETED ATTACK METHODS UNDER RANDOM CASE.

			VOC 2012				COCO 2014			
k	T	Method	ASR	SASR	Pert	LFR	ASR	SASR	Pert	LFR
3	1	FGSMT	5.10	21.23	0.68	16.84	5.30	7.68	0.82	39.13
	40	MFGSMT	10.70	26.43	**0.23**	20.96	8.70	16.33	**0.45**	51.99
		MLCWT	45.90	74.43	0.60	89.70	46.80	75.19	1.04	90.06
	300	MLDFT	6.80	17.14	0.86	13.58	7.80	19.21	1.45	17.36
		PGDT	14.17	49.73	0.82	49.82	4.83	31.96	1.47	85.47
		Rank I	68.20	71.73	0.55	98.48	98.40	99.06	0.95	99.17
		T$_k$MLT	79.03	88.80	0.57	97.09	98.77	99.28	1.00	99.27
		T$_k$ALCT	79.27	89.11	0.59	97.34	98.90	99.31	1.03	99.30
		T$_k$MIAT	79.70	89.20	0.60	97.83	99.03	99.39	1.02	99.27
		MASWT	**81.73**	**89.65**	0.57	**99.53**	**99.73**	**99.87**	1.00	**99.95**
5	1	FGSMT	4.30	30.89	0.70	18.04	4.70	8.78	0.84	39.63
	40	MFGSMT	8.70	32.55	**0.34**	21.15	7.60	15.46	**0.47**	52.47
		MLCWT	8.30	61.00	0.67	85.51	21.80	72.24	1.29	91.39
	300	MLDFT	5.50	20.79	0.90	7.98	6.50	15.64	1.46	14.01
		PGDT	12.73	43.99	0.76	44.88	4.57	28.33	1.46	87.56
		Rank I	53.10	63.44	0.58	98.01	89.00	91.66	1.10	99.29
		T$_k$MLT	68.77	86.69	0.60	96.28	95.47	97.54	1.17	99.25
		T$_k$ALCT	69.20	87.06	0.61	97.20	95.50	97.60	1.20	99.29
		T$_k$MIAT	70.07	87.20	0.62	97.72	95.73	97.64	1.19	99.30
		MASWT	**72.37**	**88.19**	0.60	**99.29**	**96.57**	**98.35**	1.17	**99.98**
10	1	FGSMT	3.50	53.64	0.75	18.95	4.50	12.45	0.85	40.43
	40	MFGSMT	7.30	54.61	**0.35**	22.74	6.20	16.97	**0.48**	47.83
		MLCWT	6.70	61.46	0.75	73.40	11.10	48.78	1.31	92.28
	300	MLDFT	3.00	51.71	0.92	6.33	4.70	12.65	1.46	12.60
		PGDT	9.77	57.19	0.79	35.30	3.27	25.84	1.47	81.34
		Rank I	38.30	69.02	0.57	98.08	37.10	59.22	1.17	99.27
		T$_k$MLT	59.07	90.09	0.61	95.31	87.53	94.89	1.28	99.44
		T$_k$ALCT	60.30	90.59	0.64	96.00	87.60	95.08	1.30	99.47
		T$_k$MIAT	60.90	90.64	0.63	96.20	87.63	95.10	1.29	99.40
		MASWT	**63.40**	**91.52**	0.61	**98.91**	**87.80**	**95.90**	1.28	**99.96**

TABLE V
COMPARISON OF TARGETED ATTACK METHODS UNDER WORST CASE.

			VOC 2012				COCO 2014			
k	T	Method	ASR	SASR	Pert	LFR	ASR	SASR	Pert	LFR
3	1	FGSMT	4.20	6.19	0.72	17.21	5.20	6.98	0.83	42.48
	40	MFGSMT	7.30	10.26	**0.31**	20.51	7.30	14.38	**0.49**	56.22
		MLCWT	12.70	40.06	0.66	89.30	23.40	57.93	1.13	92.35
	300	MLDFT	5.70	9.88	0.78	7.63	5.40	10.75	1.56	12.69
		PGDT	12.30	22.23	0.86	50.13	4.30	26.61	1.45	91.12
		Rank I	38.20	41.09	0.61	98.19	78.10	83.47	1.05	99.24
		T$_k$MLT	56.83	72.23	0.64	96.65	82.63	93.06	1.12	99.24
		T$_k$ALCT	57.10	72.68	0.65	97.01	82.70	92.74	1.14	98.67
		T$_k$MIAT	58.30	73.00	0.66	97.46	82.93	92.90	1.15	98.90
		MASWT	**60.60**	**74.85**	0.64	**99.06**	**83.13**	**93.27**	1.13	**99.89**
5	1	FGSMT	3.90	5.62	0.73	15.74	4.80	6.26	0.84	41.61
	40	MFGSMT	6.50	9.83	**0.33**	18.29	6.60	10.25	**0.51**	54.33
		MLCWT	7.30	28.66	0.74	86.38	9.90	62.06	1.28	93.17
	300	MLDFT	4.50	8.74	0.79	7.12	4.10	9.52	1.57	13.46
		PGDT	10.70	18.71	0.94	44.11	3.87	26.58	1.58	90.73
		Rank I	22.20	28.72	0.61	97.83	62.40	71.72	1.15	99.29
		T$_k$MLT	45.30	70.65	0.66	95.40	73.73	91.16	1.24	99.25
		T$_k$ALCT	45.50	71.01	0.68	96.10	73.80	91.20	1.25	99.30
		T$_k$MIAT	45.70	71.28	0.68	96.45	73.77	91.25	1.25	99.27
		MASWT	**46.97**	**72.58**	0.66	**98.81**	**74.60**	**92.31**	1.24	**99.97**
10	1	FGSMT	2.90	4.31	0.73	15.35	4.10	5.34	0.82	41.47
	40	MFGSMT	4.30	6.34	**0.35**	17.54	5.50	8.79	**0.51**	51.06
		MLCWT	5.30	37.55	0.78	77.38	6.70	58.68	1.31	93.79
	300	MLDFT	2.40	6.77	0.79	6.54	3.50	7.66	1.57	12.58
		PGDT	8.33	28.51	0.97	34.58	2.37	21.79	1.61	86.88
		Rank I	14.70	40.18	0.59	97.93	25.20	44.73	1.22	99.37
		T$_k$MLT	36.37	76.83	0.65	93.98	56.23	88.01	1.29	99.07
		T$_k$ALCT	36.40	76.91	0.66	94.37	56.30	88.19	1.30	99.10
		T$_k$MIAT	36.57	77.10	0.66	95.10	56.43	88.21	1.31	99.06
		MASWT	**37.50**	**78.29**	0.65	**98.24**	**57.23**	**88.65**	1.28	**99.96**

LFR values for targeted attacks. Although MFGSMT obtains the smallest Pert, its attack performance is still unacceptable. Moreover, MLCWT, MLDFT, and PGDT attain a much lower performance than our methods by a large margin in terms of ASR, SASR, and LFR. MLCWT and MLDFT explicitly lower the prediction confidence of the ground truth below a certain threshold. However they achieve low LFR values. T$_k$MLT and Rank I are designed specifically for top-k targeted attacks, but their attack performance is still lower than our methods regarding ASR, SASR, and LFR. Although using complex methods, both T$_k$ALCT and T$_k$MIAT are inferior to our method.

Tables IV and V present the results under the random and worst cases. Similar to Table III, our method achieves a better attack performance in terms of ASR, SASR, and LFR with

comparable Pert.

TABLE VI
COMPARISON OF UNTARGETED ATTACK METHODS ON TWO LARGE SETS.

			NUS-WIDE				Open Images			
k	T	Method	ASR	SASR	Pert	LFR	ASR	SASR	Pert	LFR
3	300	MLCWU	14.70	29.65	0.87	35.84	27.60	78.17	1.39	25.22
		MLDFU	10.70	19.64	1.56	30.38	23.10	50.64	1.97	22.89
		PGDU	85.23	94.15	1.68	63.91	92.13	93.46	1.65	64.45
		kFool	69.80	88.03	1.51	36.33	87.10	93.37	2.73	40.35
		T_kMLU	96.33	97.01	0.15	79.75	97.27	97.99	0.14	85.34
		MASWU	**98.17**	**99.67**	**0.14**	**98.83**	**99.87**	**99.90**	**0.13**	**98.45**
5	300	MLCWU	10.00	25.47	0.62	37.29	14.20	66.11	1.35	27.46
		MLDFU	8.60	17.62	1.58	35.69	13.30	45.48	1.99	25.86
		PGDU	84.37	93.92	1.69	68.18	90.80	91.21	1.69	74.61
		kFool	72.30	88.22	5.28	47.49	82.70	92.18	5.71	51.53
		T_kMLU	95.77	96.54	0.16	81.29	96.20	97.01	0.15	86.36
		MASWU	**97.73**	**99.33**	**0.14**	**98.86**	**99.67**	**99.63**	**0.14**	**98.97**
10	300	MLCWU	8.50	17.81	0.59	37.64	9.10	50.84	1.09	29.02
		MLDFU	6.90	15.64	1.61	37.58	11.50	37.95	2.07	27.73
		PGDU	82.77	93.57	1.70	73.59	85.23	96.72	1.71	81.83
		kFool	71.70	87.06	11.59	60.54	81.50	91.75	14.98	63.03
		T_kMLU	94.10	95.43	0.16	85.69	94.73	95.48	0.17	87.12
		MASWU	**95.97**	**97.69**	**0.15**	**98.83**	**99.47**	**99.53**	**0.15**	**99.14**

TABLE VII
COMPARSION OF UNIVERSAL UNTARGETED ATTACK METHODS ON NUS-WIDE AND OPEN IMAGES. PERT IS IN 10^{-2}.

		NUS-WIDE				Open Images			
k	Method	ASR	SASR	Pert	LFR	ASR	SASR	Pert	LFR
2	PGDUv	69.90	76.18	8.91	57.64	75.53	80.71	7.45	66.48
	kUv	65.80	69.34	7.98	53.91	73.60	78.62	6.24	61.46
	T_kMLUv	77.93	87.07	3.93	63.08	82.70	94.19	3.41	81.17
	MASWUv	**79.93**	**90.31**	**3.89**	**80.79**	**84.97**	**97.24**	**3.23**	**85.93**
3	PGDUv	65.47	73.46	8.96	60.93	70.47	76.58	7.49	67.93
	kUv	60.70	65.96	8.04	55.78	67.80	73.48	6.63	65.28
	T_kMLUv	71.57	84.29	4.20	74.39	80.27	94.89	3.25	82.37
	MASWUv	**75.47**	**86.87**	**4.17**	**86.79**	**82.43**	**96.62**	**3.12**	**87.69**
5	PGDUv	60.43	67.65	9.04	64.56	64.53	71.45	7.91	70.91
	kUv	57.50	63.48	8.11	57.59	62.10	69.48	6.77	67.94
	T_kMLUv	69.03	82.07	4.61	80.13	77.03	95.46	3.56	83.07
	MASWUv	**72.27**	**84.35**	**4.55**	**93.39**	**80.87**	**96.27**	**3.29**	**89.26**
10	PGDUv	55.73	61.47	9.15	67.64	60.27	70.64	7.93	73.75
	kUv	54.30	60.38	8.75	65.69	57.90	67.26	6.91	70.72
	T_kMLUv	59.37	76.51	6.52	93.19	70.47	93.42	3.71	83.56
	MASWUv	**60.93**	**78.59**	**6.51**	**97.33**	**74.87**	**94.27**	**3.67**	**91.01**

E. Attacks on Large Datasets

Table VI reports the results of untargeted attacks on two large datasets, NUS-WIDE and Open Images, revealing that our method, MASWU, obtain the best results for $k = 3, 5, 10$. Regarding the SOTA methods MLCWU, MLDFU, PGDU, kFoolU, and T_kMLU, although these require a larger perturbation bound Pert, they still achieve lower ASR, SASR and LFR values. Besides, MLCWU and MLDFU explicitly lower the confidence of the true label below a certain threshold but still achieve low LFR. Table VII reports the results of universal untargeted attacks. Similarly, MASWUv obtain the lowest Pert and the highest ASR, SASR, and LFR. Table VIII presents the results of targeted attack for $k = 3, 5, 10$ under best, random, and worst cases. MASWT still outperform other SOTA methods in ASR, SASR, and LFR by utilizing a lower or comparable Pert. Rank I and T_kMLT, are specially designed for top-k attacks, but still obtain lower ASR and SASR than our methods[6].

[6]More results regarding LPD, GAMA, T_kALC and T_kMIA are included in the supplementary materials.

TABLE VIII
COMPARISON OF TARGETED ATTACK METHODS ON TWO LARGE SETS.

				NUS-WIDE				Open Images			
Case	k	T	Method	ASR	SASR	Pert	LFR	ASR	SASR	Pert	LFR
Best	3	300	MLCWT	18.20	48.43	0.17	46.23	27.00	45.26	0.37	57.98
			MLDFT	14.50	35.39	0.45	44.43	23.70	38.92	0.41	50.52
			PGDT	38.47	67.96	0.34	28.28	32.63	50.97	0.35	57.24
			Rank I	92.60	94.23	**0.13**	30.59	90.10	92.35	0.16	25.37
			T_kMLT	95.27	96.71	0.14	75.27	95.07	96.72	**0.15**	82.14
			MASWT	**98.13**	**98.85**	**0.13**	**93.61**	**99.93**	**99.75**	**0.15**	**95.67**
	5	300	MLCWT	15.10	58.18	0.19	43.34	20.30	40.66	0.34	61.82
			MLDFT	13.70	34.56	0.41	35.28	18.90	34.72	0.39	45.86
			PGDT	31.07	70.26	0.33	22.41	23.97	47.79	0.35	59.15
			Rank I	91.50	94.19	**0.14**	31.62	89.60	91.13	0.18	28.69
			T_kMLT	94.07	95.94	0.15	77.39	94.83	96.01	**0.16**	83.56
			MASWT	**96.53**	**98.69**	0.15	**95.18**	**99.80**	**99.74**	**0.16**	**95.51**
	10	300	MLCWT	14.80	70.27	0.25	40.14	14.60	35.64	0.30	55.29
			MLDFT	10.40	29.73	0.43	33.37	15.70	30.23	0.40	44.19
			PGDT	25.47	73.20	0.35	16.02	17.47	46.03	0.36	49.22
			Rank I	77.40	81.48	0.21	37.85	87.20	89.65	0.20	28.81
			T_kMLT	93.63	94.56	0.17	79.25	93.53	94.49	**0.18**	85.69
			MASWT	**95.50**	**98.13**	**0.16**	**96.77**	**99.37**	**99.75**	**0.18**	**95.21**
Random	3	300	MLCWT	16.50	40.92	0.21	65.98	24.60	40.19	0.39	61.17
			MLDFT	13.20	33.39	0.40	47.71	20.90	35.65	0.44	55.87
			PGDT	25.77	50.29	0.32	50.39	25.63	43.38	0.41	50.51
			Rank I	80.90	82.62	0.19	33.69	78.50	79.49	0.26	68.66
			T_kMLT	88.93	91.28	0.20	77.25	89.47	92.75	0.29	74.64
			MASWT	**91.73**	**95.12**	**0.19**	**94.01**	**91.57**	**94.21**	**0.25**	**96.32**
	5	300	MLCWT	12.50	32.26	0.24	64.41	18.50	31.28	0.35	54.82
			MLDFT	12.60	30.38	0.42	53.32	16.20	25.78	0.45	53.26
			PGDT	19.97	40.24	0.35	42.81	17.73	35.66	0.40	55.54
			Rank I	48.30	54.74	0.21	47.71	38.70	43.24	0.28	70.23
			T_kMLT	80.17	86.52	0.24	78.98	79.30	83.58	0.31	78.65
			MASWT	**82.33**	**91.27**	**0.19**	**92.99**	**81.17**	**87.79**	**0.25**	**94.02**
	10	300	MLCWT	10.40	22.57	0.25	58.13	16.50	27.84	0.36	46.31
			MLDFT	10.10	25.37	0.44	54.72	15.20	22.17	0.46	52.69
			PGDT	15.43	33.41	0.37	32.96	15.63	33.29	0.42	55.14
			Rank I	45.60	50.28	**0.23**	49.96	30.00	38.73	0.30	69.31
			T_kMLT	67.63	71.52	0.26	81.09	41.47	61.44	0.31	71.69
			MASWT	**70.63**	**76.16**	**0.23**	**91.01**	**48.77**	**69.99**	**0.28**	**86.42**
Worst	3	300	MLCWT	10.80	27.91	0.23	72.42	14.20	26.61	0.34	66.67
			MLDFT	9.70	23.46	0.39	65.47	12.30	23.79	0.46	55.75
			PGDT	19.20	31.28	0.29	55.14	17.27	36.19	0.39	54.28
			Rank I	23.10	24.88	0.22	49.33	75.20	78.62	**0.26**	58.87
			T_kMLT	50.33	62.48	0.23	71.43	87.17	87.94	0.27	74.12
			MASWT	**52.50**	**65.70**	**0.14**	**88.23**	**88.70**	**88.93**	**0.26**	**95.74**
	5	300	MLCWT	7.80	19.71	0.25	71.82	10.40	23.19	0.35	62.38
			MLDFT	6.90	14.28	0.38	64.19	9.10	17.26	0.45	55.46
			PGDT	10.73	15.85	0.31	45.94	12.33	30.75	0.37	43.39
			Rank I	14.90	18.61	0.21	48.91	34.70	36.02	0.28	60.71
			T_kMLT	30.33	52.31	0.25	81.19	74.20	85.64	0.28	76.53
			MASWT	**34.27**	**56.55**	**0.14**	**91.04**	**75.63**	**87.49**	**0.25**	**94.29**
	10	300	MLCWT	5.40	16.84	0.28	70.02	7.40	24.91	0.36	64.51
			MLDFT	4.50	14.72	0.43	67.85	6.70	18.84	0.49	56.74
			PGDT	8.43	11.28	0.33	48.59	9.97	26.71	0.38	46.78
			Rank I	12.10	19.86	**0.24**	46.64	57.30	66.80	0.31	71.56
			T_kMLT	14.33	46.06	0.29	83.42	84.00	92.54	0.32	79.24
			MASWT	**16.50**	**49.72**	0.27	**91.23**	**87.50**	**96.89**	**0.29**	**86.43**

F. More Analysis

Visual comparison. Fig. 6 illustrates two examples of top-3 untargeted attack (top) and targeted attack (bottom) on VOC 2012, where our SPW-based methods make the top-3 outputs of the adversarial image have higher confidence, and the ground truth labels have lower prediction confidence than the other methods. Fig. 7 visualizes some examples for universal untargeted attack on COCO 2014, revealing that MASWUv uses a smaller Pert to make a successful attack. The labels in Y have a lower prediction confidence acquired by MASWUv than kUv and T_kMLUv, and the top-3 outputs have a higher prediction confidence than the other two methods. For kUv, the perturbation is more visible.

Ablation study. Next, we compare our two AFD measures with the AFD measures in T_kML and MLCW using the conventional optimization strategy rather than our SPW. Fig. 8(a)

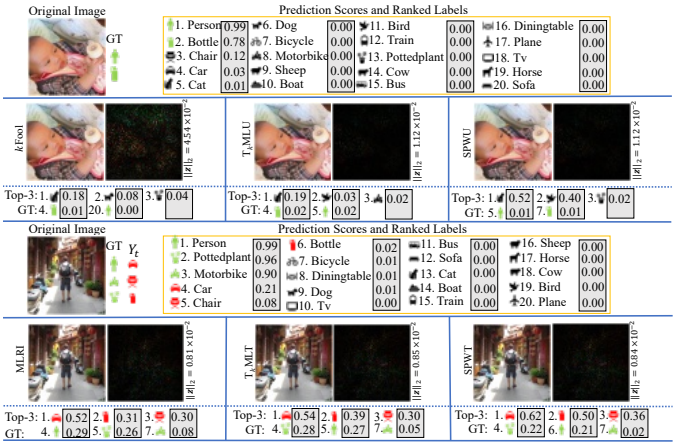


Fig. 6. Visual images in VOC 2012 for top-3 untargeted attck (top) and targeted attack under the best case (bottom).

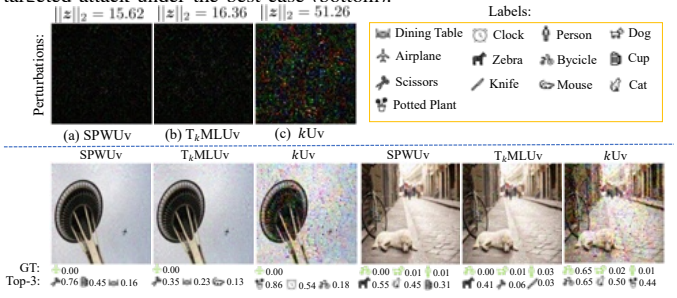


Fig. 7. Visual images in COCO 2014 for top-3 universal untargeted attack.

illustrates the results of top-3 untargeted attack on VOC 2012, demonstrating that our measures outperform T_kML and MLCW and proving the effectiveness of our AFD measures. To validate SPW, we replace the optimizing strategy in MLCW with our SPW. Fig. 8(b) presents the results on the top-10 untargeted and targeted attacks, highlighting that when applying SPW to MLCW the performance is improved further demonstrating the effectiveness of SPW. Similar conclusions are made from Fig. 8(c) when applying SPW to T_kML for a targeted attack. In theory, SPW can be applied to arbitrary adversarial attack tasks involving multiple constraints.

Secondly, we study the effect of the update strategies of ϵ in SPL, our SPW, and the fixed ϵ (FE). These three strategies are compared by replacing the update strategy in SPW with the strategy of SPL and FE. Fig. 9(a), (b) and (c) compare ASR, LFR and Pert when three strategies are applied. The results demonstrate that our strategy achieves a higher ASR and LFR and lower Pert than the strategy in SPL and FE.

Sensitivity test. There are three important hyperparameters λ_1, λ_2 and η. Fig. 10(a) presents the results of top-10 targeted attack under the worst case on VOC 2012 with varied λ_1 (λ_2 remains 0.5). As λ_1 increases, ASR also increases, and LFR decreases. λ_1 and λ_2 should be tuned according to the application scenario. Moreover, Fig. 10(b) presents the effect of η on ASR and LFR, achieving the best value for $\eta = 1.01$. Based on this sensitivity test, we conclude that $\lambda_1, \lambda_2 = 0.5$, and $\eta = 1.01$ attain the best possible performance. Besides, Fig. 10(c) visualizes the change process of ϵ^t during the solution process, highlighting that when there is no ideal solution, ϵ^t increases to a certain value and tends to remain unchanged. Additionally, ϵ^t keeps increasing when the ideal solution exists. These results demonstrate that ϵ^t can be properly adjusted according to the change of the attack performance instead of continuously increasing in SPL.

Early stopping strategy [32] is used to terminate Algorithm 2. Fig. 11(a) shows the results of the patience E on VOC 2012, revealing that as E increases, ASR first increases and then decreases. ASR reaches its maximum value for $E = 40$. In our experiments, we set $E = 40$. We also study the impact of ν on attack performance, with Figs. 11(b) and (c) reporting the results under different ν on VOC 2012 and COCO 2014, respectively. As ν increases, ASR and LFR increase, but LFR tends to a plateau in Fig. 11(c). Since $\nu = 1$ satisfies the required attack performance, we set $\nu = 1$ in our experiments[7].

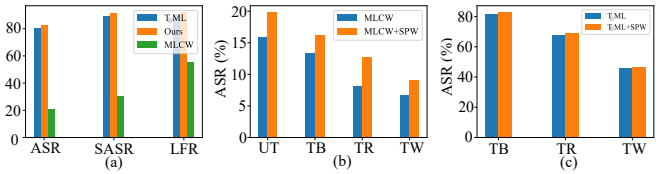


Fig. 8. (a) AFD results, (b) and (c) SPW results (UT means untargeted attack, and TB/TR/TW means targeted attack under the best/random/worst case.

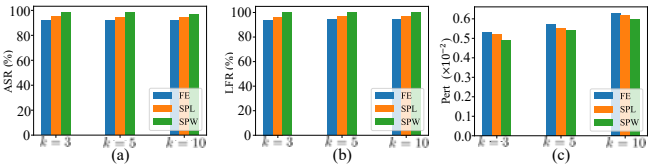


Fig. 9. (a), (b) and (c) ablation studies of top-3 untargeted attack for the update strategy of ϵ on VOC 2012.

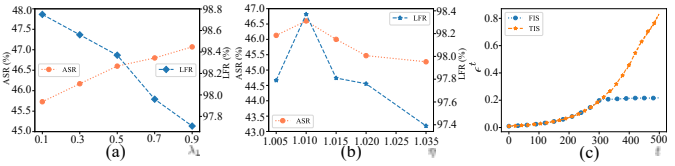


Fig. 10. (a) and (b) effect of λ_1 and η on ASR and LFR, respectively. (c) variation of ϵ when the ideal solution does not exist (FIS) and exists (TIS).

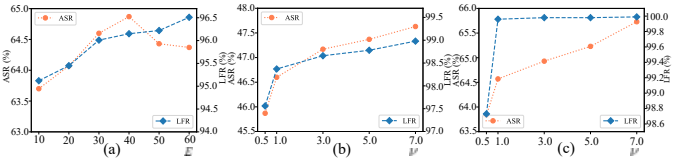


Fig. 11. (a) ASR and LFR of the top-3 universal untargeted attack with variable E on VOC 2012. (b) and (c) ASR and LFR of the top-10 targeted attack under the worst case with a varied ν on VOC 2012 and COCO 2014.

VI. Conclusion

The defects of existing SOTA methods are revealed from both experimental and theoretical perspectives. To deal with the two defects, we propose a new measure scheme based on the Jaccard index, yielding two concrete measures called AFD_{slc} and AFD_{slf}. Moreover, the constrained optimization problems with the least constraint violation are reconstructed for untargeted and targeted attacks. We conduct a solid theoretical analysis that demonstrates the effectiveness of the

[7]More details can be found in the supplementary materials.

weighting strategy. A natural weighted Lagrangian form is proposed. Then we develop a self-paced weighting (SPW) scheme that gradually involves difficult constraints in the optimization process. SPW increases the attack gain and avoids fluctuations during optimization. Extensive experiments validate the effectiveness of our method.

REFERENCES

[1] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.

[2] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *ICLR*, 2017.

[3] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI TRIT*, vol. 6, no. 1, pp. 25–45, 2021.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014.

[5] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *CVPR*, 2016, pp. 2574–2582.

[6] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE S&P*, 2017, pp. 39–57.

[7] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *CVPR*, 2018, pp. 9185–9193.

[8] W. Liu, H. Wang, X. Shen, and I. W. Tsang, "The emerging trends of multi-label learning," *IEEE TPAMI*, vol. 44, no. 11, pp. 7955–7974, 2021.

[9] Q. Song, H. Jin, X. Huang, and X. Hu, "Multi-label adversarial perturbations," in *IEEE ICDM*. IEEE, 2018, pp. 1242–1247.

[10] A. Aich, C.-K. Ta, A. Gupta, C. Song, S. Krishnamurthy, S. Asif, and A. Roy-Chowdhury, "Gama: Generative adversarial multi-object scene attacks," in *NeurIPS*, vol. 35, pp. 36914–36930, 2022.

[11] S. Hu, L. Ke, X. Wang, and S. Lyu, "Tkml-ap: Adversarial attacks to top-k multi-label learning," in *ICCV*, 2021, pp. 7649–7657.

[12] N. Tursynbek, A. Petiushko, and I. Oseledets, "Geometry-inspired top-k adversarial perturbations," in *WACV*, 2022, pp. 3398–3407.

[13] X. Tan, C. Zhao, C. Liu, J. Wen, and Z. Tang, "A two-stage information extraction network for incomplete multi-view multi-label classification," in *AAAI*, vol. 38, no. 14, 2024, pp. 15 249–15 257.

[14] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in *ICCV*, 2021, pp. 82–91.

[15] C. Liu, J. Wen, X. Luo, and Y. Xu, "Incomplete multi-view multi-label learning via label-guided masked view-and category-aware transformers," in *AAAI*, vol. 37, no. 7, 2023, pp. 8816–8824.

[16] L. Wang, Y. Liu, H. Di, C. Qin, G. Sun, and Y. Fu, "Semi-supervised dual relation learning for multi-label classification," *IEEE TIP*, vol. 30, pp. 9125–9135, 2021.

[17] N. Zhou, W. Luo, X. Lin, P. Xu, and Z. Zhang, "Generating multi-label adversarial examples by linear programming," in *IJCNN*, IEEE, 2020, pp. 1–8.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[19] P. Jaccard, "The distribution of the flora in the alpine zone. 1," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[20] N. Chou, J. Wu, J. B. Bingren, A. Qiu, and K.-H. Chuang, "Robust automatic rodent brain extraction using 3-d pulse-coupled neural networks (pcnn)," *IEEE TIP*, vol. 20, no. 9, pp. 2554–2564, 2011.

[21] M. Everingham, S. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *IJCV*, vol. 111, no. 1, pp. 98–136, 2015.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, Springer, 2014, pp. 740–755.

[23] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann, "Easy samples first: Self-paced reranking for zero-example multimedia search," in *ACM MM*, 2014, pp. 547–556.

[24] K. Ghasedi, X. Wang, C. Deng, and H. Huang, "Balanced self-paced learning for generative adversarial clustering network," in *CVPR*, 2019, pp. 4391–4400.

[25] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *CIVR*, 2009, pp. 1–9.

[26] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4," *IJCV*, vol. 128, no. 7, pp. 1956–1981, 2020.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, Jun. 2016, pp. 2818–2826.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, Jun. 2016, pp. 770–778.

[29] T. Ridnik, H. Lawen, A. Noy, E. Ben Baruch, G. Sharir, and I. Friedman, "Tresnet: High performance gpu-dedicated architecture," in *WACV*, 2021, pp. 1400–1409.

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

[31] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *CVPR*, 2017, pp. 1765–1773.

[32] L. Prechelt, "Early stopping-but when?" in *Neural Networks: Tricks of the Trade*, Springer, 1998, pp. 55–69.

[33] N. Zhou, W. Luo, J. Zhang, L. Kong, and H. Zhang, "Hiding all labels for multi-label images: An empirical study of adversarial examples," in *IJCNN*. IEEE, 2021, pp. 1–8.

[34] S. Rajeswar, P. Rodriguez, S. Singhal, D. Vazquez, and A. Courville, "Multi-label iterated learning for image classification with label ambiguity," in *CVPR*, 2022, pp. 4783–4793.

[35] J.-Y. Hang and M.-L. Zhang, "Dual perspective of label-specific feature learning for multi-label classification," in *ICML*, PMLR, 2022, pp. 8375–8386.

[36] Z. Yang, Y. Han, and X. Zhang, "Characterizing the evasion attackability of multi-label classifiers," in *AAAI*, 2021, vol. 35, no. 12, pp. 10 647–10 655.

[37] A. Aich, S. Li, C. Song, M. S. Asif, S. V. Krishnamurthy, and A. K. Roy-Chowdhury, "Leveraging local patch differences in multi-object scenes for generative adversarial attacks," in *WACV*, 2023, pp. 1308–1318.

[38] L. Wang and K.-J. Yoon, "Psat-gan: Efficient adversarial attacks against holistic scene understanding," *IEEE TIP*, vol. 30, pp. 7541–7553, 2021.

[39] N. Li and Z. Chen, "Toward visual distortion in black-box attacks," *IEEE TIP*, vol. 30, pp. 6156–6167, 2021.

[40] Y. Dai and L. Zhang, "Optimization with Least Constraint Violation," *CSIAM Trans. on Appl. Math.*, vol. 2, pp. 551–584, 2021.

[41] R. T. Rockafellar and R. J.-B. Wets, "Variational analysis," *Springer Science & Business Media*, vol. 317, 2009.

[42] J. Jia, W. Qu, and N. Gong, "Multiguard: Provably robust multi-label classification against adversarial examples," in *NeurIPS*, vol. 35, 2022, pp. 10 150–10 163.

[43] J. Jia, B. Wang, X. Cao, H. Liu, and N. Z. Gong, "Almost tight l0-norm certified robustness of top-k predictions against adversarial perturbations," in *ICLR*, 2022.

[44] J. Jia, X. Cao, B. Wang, and N. Z. Gong, "Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing," in *ICML*, 2020.

[45] Y. Sun, Q. Xu, Z. Wang, and Q. Huang, "When measures are unreliable: Imperceptible adversarial perturbations toward top-k multi-label learning," in *ACM MM*, 2023, pp. 1515–1526.

[46] M. Ma, W. Zheng, W. Lv, L. Ren, H. Su, and Z. Yin, "Multi-label adversarial attack based on label correlation," in *ICIP*. IEEE, 2023, pp. 2050–2054.

[47] J. Tian, B. Wang, R. Guo, Z. Wang, K. Cao, and X. Wang, "Adversarial attacks and defenses for deep-learning-based unmanned aerial vehicles," *IEEE IoTJ*, vol. 9, no. 22, pp. 22 399–22 409, 2022.

[48] J. Tian, C. Shen, B. Wang, X. Xia, M. Zhang, C. Lin, and Q. Li, "LESSON: Multi-label adversarial false data injection attack for deep learning locational detection," *IEEE Trans. Depend. Secure Comput.*, pp. 1–15, 2024.

[49] L. Kong, W. Luo, H. Zhang, Y. Liu, and Y. Shi, "Evolutionary multilabel adversarial examples: An effective black-box attack," *IEEE TAI*, vol. 4, no. 3, pp. 562–572, 2023.

[50] Y. Lin, M. Chen, K. Zhang, H. Li, M. Li, Z. Yang, D. Lv, B. Lin, H. Liu, and D. Cai, "Tagclip: A local-to-global framework to enhance open-vocabulary multi-label classification of clip without training," in *AAAI*, vol. 38, no. 4, 2024, pp. 3513–3521.

[51] Z. Guo, B. Dong, Z. Ji, J. Bai, Y. Guo, and W. Zuo, "Texts as images in prompt tuning for multi-label image recognition," in *CVPR*, 2023, pp. 2808–2817.

[52] C.-Y. Lee, C.-C. Tsai, C.-C. Kao, C.-S. Lu, and C.-M. Yu, "Defending against clean-image backdoor attack in multi-label classification," in *ICASSP*. IEEE, 2024, pp. 5500–5504.

APPENDIX

A. Proof of Proposition 1

Proof. According to the definition of AFD_{lc}, for a perturbation $\boldsymbol{z}$, if $Y \not\subset \hat{Y}_k(\boldsymbol{x}+\boldsymbol{z})$ holds, then $\mathbb{I}(Y \subset \hat{Y}_k(\boldsymbol{x}+\boldsymbol{z})) = 0$ holds. Likewise, if we have $\hat{Y}_k(\boldsymbol{x}+\boldsymbol{z}) \not\subset Y$ and $Y \neq \hat{Y}_k(\boldsymbol{x}+\boldsymbol{z})$, then $\mathbb{I}(\hat{Y}_k(\boldsymbol{x}+\boldsymbol{z}) \subset Y) = 0$ and $\mathbb{I}(\hat{Y}_k(\boldsymbol{x}+\boldsymbol{z}) = Y) = 0$ are ture. Then, we have $\text{AFD}_{lc}(\boldsymbol{x}, \boldsymbol{z}) = 0$. Therefore, if two arbitrary different perturbations $\boldsymbol{z}^1$ and $\boldsymbol{z}^2$ satisfy the above statement, we have $\text{AFD}_{lc}(\boldsymbol{x}, \boldsymbol{z}^1) = \text{AFD}_{lc}(\boldsymbol{x}, \boldsymbol{z}^2) = 0$. Similarly, even if $|Y_{lf}(\boldsymbol{z}^1)| \neq |Y_{lf}(\boldsymbol{z}^2)| \neq 0$, $|B_I(\boldsymbol{x}, \boldsymbol{z}^1)| \neq |B_I(\boldsymbol{x}, \boldsymbol{z}^2)| \neq |B|$, and $|Y_{lf}(\boldsymbol{z}^1)| \neq |Y_{lf}(\boldsymbol{z}^2)|$, $|B_I(\boldsymbol{z}^1)| \neq |B_I(\boldsymbol{z}^2)|$ are true, we still have $\text{AFD}_{lf}(\boldsymbol{x}, \boldsymbol{z}^1) = \text{AFD}_{lf}(\boldsymbol{x}, \boldsymbol{z}^2)$ and $\text{AC}(\boldsymbol{x}, \boldsymbol{z}^1) = \text{AC}(\boldsymbol{x}, \boldsymbol{z}^2)$. The above analysis means that the existing AFD/AC measures cannot accurately distinguish between the different perturbations. □

B. Proof of Theorem 1

Proof. We consider a common constrained optimization problem as follows:

$$\min ||\boldsymbol{z}||_2^2/2 \ \textbf{s.t.}\ CT_j(\boldsymbol{z}) \geq 0;\ j = 1, \cdots, N_{ct}. \tag{29}$$

According to (29), the Lagrangian form in [9], [11] is $\min ||\boldsymbol{z}||_2^2 - \lambda \sum_{j=1}^{N_{ct}} CT_j(\boldsymbol{z})$, where λ is the Lagrange coefficient (greater than or equal to 0). The KKT condition requires $CT_j(\boldsymbol{z}) \geq 0;\ j = 1, \cdots, N_{ct}$. Therefore, we can divide all constraints into $N_{\mathcal{G}}$ groups and form $N_{\mathcal{G}}$ disjoint regions for $\boldsymbol{z}$, namely, $\mathcal{G}_{j'}, j' = 1, \cdots, N_{\mathcal{G}}$.

Firstly, we prove that $\mathcal{G}_j$s are the closed sets. Let $\boldsymbol{z}^*$ be a cluster point of $\mathcal{G}_{j'}$. Then there exists $\{\boldsymbol{z}^m\}_{m=1}^{+\infty}$ such that $\lim_{m\to+\infty} \boldsymbol{z}^m = \boldsymbol{z}^*$ holds. Since $CT_j(\boldsymbol{z})$ is continuous, we have $CT_j(\boldsymbol{z}^*) \geq 0$. Otherwise, there exists $\hbar$ that satisfies $CT_j(\boldsymbol{z}) < 0;\ \forall \boldsymbol{z} \in (\boldsymbol{z}^* - \hbar, \boldsymbol{z}^* + \hbar)$. Due to $\lim_{m\to+\infty} \boldsymbol{z}^m = \boldsymbol{z}^*$, there exists a positive integer m^* that satisfies $\boldsymbol{z}_m \in (\boldsymbol{z}^* - \hbar, \boldsymbol{z}^* + \hbar),\ \forall\ m > m^*$, and $CT_j(\boldsymbol{z}_m) < 0$. This contradicts $CT_j(\boldsymbol{z}^m) \geq 0$.

Secondly, since $\mathcal{G}_{j'} \cap \mathcal{G}_{k'} = \emptyset$ is true, we have $\inf_{\boldsymbol{z}^{m'} \in \mathcal{G}_{j'}, \boldsymbol{z}^{m''} \in \mathcal{G}_k} ||\boldsymbol{z}^{m'} - \boldsymbol{z}^{m''}||_2 > 0$. There exist $\lim_{m'\to+\infty} \boldsymbol{z}_{j'}^{m'} = \boldsymbol{z}_{j'}^*$ and $\lim_{m''\to+\infty} \boldsymbol{z}_{k'}^{m''} = \boldsymbol{z}_{k'}^*$ that satisfy $\inf_{\boldsymbol{z}_{j'} \in \mathcal{G}_{j'}, \boldsymbol{z}_{k'} \in \mathcal{G}_{k'}} ||\boldsymbol{z}_{j'} - \boldsymbol{z}_{k'}||_2 = \lim_{m'\to+\infty} ||\boldsymbol{z}_{j'}^{m'} - \boldsymbol{z}_{k'}^{m'}||_2 = ||\boldsymbol{z}_{j'}^* - \boldsymbol{z}_{k'}^*||_2$. If $||\boldsymbol{z}_{j'}^* - \boldsymbol{z}_{k'}^*||_2 = 0$ holds, then we have $\boldsymbol{z}_{j'}^* = \boldsymbol{z}_{k'}^*;\ \boldsymbol{z}_{j'}^*, \boldsymbol{z}_{k'}^* \in \mathcal{G}_{j'} \cap \mathcal{G}_{k'}$. This contradicts $\mathcal{G}_{j'} \cap \mathcal{G}_{k'} = \emptyset$. Let $\mathfrak{c}^*$ be the minimum value of $\inf\{||\boldsymbol{z}^1 - \boldsymbol{z}^2||_2 : \boldsymbol{z}^1 \in \mathcal{G}_{j'}, \boldsymbol{z}^2 \in \mathcal{G}_{k'}\}, \forall j', k' = 1, \cdots, N_{\mathcal{G}}$ and $j' \neq k'$. The proof is completed. □

C. Proof of Theorem 2

Proof. According to (18), the Jacobian matrix of $\boldsymbol{F}$ is

$$\mathcal{J}(\boldsymbol{F}) = \begin{bmatrix} -\sum_{j=1}^{2|Y|} \zeta_j w_j \nabla^2 CT_j(\boldsymbol{z}) & -\mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}))^T \mathfrak{D}(\boldsymbol{w}) & \boldsymbol{0} \\ \mathcal{J}(\boldsymbol{CT}(\boldsymbol{z})) & \boldsymbol{I} & -\boldsymbol{I} \end{bmatrix}, \tag{30}$$

Let $\Phi := \{(\boldsymbol{z}, \boldsymbol{\zeta}, \boldsymbol{v}) \in \Re^n \times \Omega : \boldsymbol{F}(\boldsymbol{z}, \boldsymbol{\zeta}, \boldsymbol{v}) = 0\}$. Following Proposition 2.1 of [40], the normal cone of Φ is

$$NC_\Phi((\boldsymbol{z}, \boldsymbol{\zeta}, \boldsymbol{v})) \subseteq \left\{ \mathcal{J}(\boldsymbol{F})^T \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{bmatrix} : \begin{matrix} (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \in \Re^{n+2|Y|} \\ (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \in NC_\Omega((\boldsymbol{\zeta}, \boldsymbol{v})) \end{matrix} \right\}, \tag{31}$$

where $NC_\Omega((\boldsymbol{\zeta}, \boldsymbol{v}) = \{(\hat{\boldsymbol{\zeta}}, \hat{\boldsymbol{v}}) : (\hat{\zeta}_j, \hat{v}_j) \in N_w(\zeta_j, v_j),\ j = 1, \cdots, 2|Y|\}$. If $\zeta_j > 0$ and $v_j = 0$, then $N_w(\zeta_j, v_j) = \{0\} \times \Re$; if $\zeta_j = 0$ and $v_j > 0$, then $N_w(\zeta_j, v_j) = \Re \times \{0\}$; else $N_w(\zeta_j, v_j) = \Re \times \{0\} \cup \{0\} \times \Re \cup \Re_- \times \Re_-$.

According to Theorem 6.12 in [41], we have

$$\boldsymbol{0} \in \boldsymbol{z}^* + NC_\Phi((\boldsymbol{z}^*, \boldsymbol{\zeta}^*, \boldsymbol{v}^*))). \tag{32}$$

With (31), there exist $(\boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^*) \in \Re^{n+2|Y|}$ and $(\boldsymbol{\xi}^*, \boldsymbol{\xi}^{**}) \in NC_\Omega((\boldsymbol{\zeta}^*, \boldsymbol{v}^*))$, we have

$$\begin{pmatrix} \boldsymbol{z}^* \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix} + \begin{pmatrix} -\sum_{j=1}^{2|Y|} \zeta_j w_j \nabla^2 CT_j(\boldsymbol{z}^*) \boldsymbol{\eta}_1^* + \mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}^*)) \boldsymbol{\eta}_2^* \\ -\mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}))^T \mathfrak{D}(\boldsymbol{w}) \boldsymbol{\eta}_1^* + \boldsymbol{\eta}_2^* + \boldsymbol{\xi}^* \\ -\boldsymbol{\eta}_2^* + \boldsymbol{\xi}^{**} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}. \tag{33}$$

Further, we have

$$\begin{pmatrix} \boldsymbol{z}^* - \sum_{j=1}^{2|Y|} \zeta_j w_j \nabla^2 CT_j(\boldsymbol{z}^*) \boldsymbol{\eta}_1^* + \mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}^*)) \boldsymbol{\xi}^{**} \\ -\mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}))^T \mathfrak{D}(\boldsymbol{w}) \boldsymbol{\eta}_1^* + \boldsymbol{\xi}^* + \boldsymbol{\xi}^{**} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}. \tag{34}$$

We also have $\boldsymbol{\xi}_{\iota_1}^{**} = \mathcal{J}(\boldsymbol{CT}(\boldsymbol{z})^T \mathfrak{D}(\boldsymbol{w})) \boldsymbol{\eta}_1^*$ from $\boldsymbol{\xi}_{\iota_3}^{**} = \boldsymbol{0}, \boldsymbol{\xi}_{\iota_1}^* = \boldsymbol{0}$ and the second line of Eq. (34). Then from the first line of Eq. (34), we have

$$\begin{aligned} &\boldsymbol{z}^* + \mathcal{J}(\boldsymbol{CT}_{\iota_2}(\boldsymbol{z}^*)) \boldsymbol{\xi}_{\iota_2}^{**} - \Big(\sum_{j=1}^{2|Y|} \zeta_j^* w_j \nabla^2 CT_j(\boldsymbol{z}^*) \\ &+ \mathcal{J}(\boldsymbol{CT}_{\iota_1}(\boldsymbol{z}^*))^T \mathcal{J}((\boldsymbol{CT}(\boldsymbol{z})^T \mathfrak{D}(\boldsymbol{w}))_{\iota_1} \Big) \boldsymbol{\eta}_1^* = \boldsymbol{0}. \end{aligned} \tag{35}$$

The proof is completed. □

D. Proof of Theorem 3

Proof. The Lagrangian form of (21) is $L_G = \frac{\phi_G}{2} ||\boldsymbol{z}||_2^2 + < \boldsymbol{G}(\boldsymbol{z}, \boldsymbol{\zeta}, \boldsymbol{v}), \boldsymbol{\lambda}_G >$, where $< \cdot, \cdot >$ is the inner product. Following [40], there exist $\boldsymbol{\lambda} \neq \boldsymbol{0},\ \phi_G \geq 0$, we have $\boldsymbol{0} \in \partial_c L_G$, where ∂_c is Clarke generalized Jacobian. And we have$\partial_c \min\{\boldsymbol{\zeta}^*, \boldsymbol{v}^*\} = [\mathfrak{D}(\boldsymbol{\mathfrak{v}}^a), \mathfrak{D}(\boldsymbol{\mathfrak{v}}^b)]$, where $\boldsymbol{\mathfrak{v}}^a, \boldsymbol{\mathfrak{v}}^b$ satisfy that if $i \in \iota_1$, then $\mathfrak{v}_i^a = 0, \mathfrak{v}_i^b = 1$; if $i \in \iota_3$, then $\mathfrak{v}_i^a = 1, \mathfrak{v}_i^b = 0$; and if $i \in \iota_2$, then $\mathfrak{v}_i^a = t, \mathfrak{v}_i^b = 1 - t,\ t \in [0, 1]$. Then there exist $\varsigma^*,\ \boldsymbol{\eta}_1^*,\ \boldsymbol{\eta}_2^*,\ \boldsymbol{\xi}^*, \boldsymbol{\mathfrak{v}}^a, \boldsymbol{\mathfrak{v}}^b$, such that

$$\begin{pmatrix} \varsigma^* \boldsymbol{z}^* - \sum_{j=1}^{2|Y|} \zeta_j^* w_j \nabla^2 CT_j(\boldsymbol{z}^*) \boldsymbol{\eta}_1^* + \mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}^*)) \boldsymbol{\eta}_2^* \\ -\mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}))^T \mathfrak{D}(\boldsymbol{w}) \boldsymbol{\eta}_1^* + \boldsymbol{\eta}_2^* + \mathfrak{D}(\boldsymbol{\mathfrak{v}}^a) \boldsymbol{\xi}^* \\ -\boldsymbol{\eta}_2^* + \mathfrak{D}(\boldsymbol{\mathfrak{v}}^b) \boldsymbol{\xi}^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}. \tag{36}$$

Further, according to Eq. (36), we have

$$\begin{pmatrix} \varsigma^* \boldsymbol{z}^* - \sum_{j=1}^{2|Y|} \zeta_j^* w_j \nabla^2 CT_j(\boldsymbol{z}^*) \boldsymbol{\eta}_1^* + \mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}^*)) \mathfrak{D}(\boldsymbol{\mathfrak{v}}^b) \boldsymbol{\xi}^* \\ -\mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}))^T \mathfrak{D}(\boldsymbol{w}) \boldsymbol{\eta}_1^* + \mathfrak{D}(\boldsymbol{\mathfrak{v}}^b) \boldsymbol{\xi}^* + \mathfrak{D}(\boldsymbol{\mathfrak{v}}^a) \boldsymbol{\xi}^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}. \tag{37}$$

Due to $\mathfrak{D}(\boldsymbol{\mathfrak{v}}^a) + \mathfrak{D}(\boldsymbol{\mathfrak{v}}^b) = \boldsymbol{I}$, then $\boldsymbol{\xi}^* = \mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}) \mathfrak{D}(\boldsymbol{w}))^T \boldsymbol{\eta}_1^*$ holds. And bringing $\boldsymbol{\xi}^*$ into the first line of Eq. (37), then we have $\varsigma^* \boldsymbol{z}^* - \sum_{j=1}^{2|Y|} \zeta_j^* w_j \nabla^2 CT_j(\boldsymbol{z}^*) \boldsymbol{\eta}_1^* + \mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}^*)) \mathfrak{D}(\boldsymbol{\mathfrak{v}}^b) \mathcal{J}(\boldsymbol{CT}(\boldsymbol{z})^T \mathfrak{D}(\boldsymbol{w})) \boldsymbol{\eta}_1^* = \boldsymbol{0}$. According to the definition of $\boldsymbol{\mathfrak{v}}^a, \boldsymbol{\mathfrak{v}}^b$, Eq. (22) holds. □

E. Proof of Corollary 1 and Corollary 2

Proof. According to Eq. (22) in Theorem 3, if $\boldsymbol{w}_{\iota_1^c} = \boldsymbol{1}$ and $\boldsymbol{w}_{\iota_{j'}^c} = \boldsymbol{0}, \forall j' = 2, \cdots, l$ holds, then we have

$$\begin{aligned} &\Big[-\sum_{j \in \iota_1^c} \zeta_j^* \nabla^2 CT_j(\boldsymbol{z}^*) + \mathcal{J}(\boldsymbol{CT}_{\iota_1 \cap \iota_1^c}(\boldsymbol{z}^*))^T \mathcal{J}(\boldsymbol{CT}_{\iota_1 \cap \iota_1^c}(\boldsymbol{z}^*)) + \\ &\mathcal{J}(\boldsymbol{CT}_{\iota_2 \cap \iota_1^c}(\boldsymbol{z}^*)) \mathfrak{D}(\boldsymbol{v}^*_{\iota_2 \cap \iota_1^c}) \mathcal{J}(\boldsymbol{CT}_{\iota_2 \cap \iota_1^c}(\boldsymbol{z}^*)) \boldsymbol{v}^*_{\iota_2 \cap \iota_1^c} \Big] \boldsymbol{\eta}_1^* + \varsigma^* \boldsymbol{z}^* = \boldsymbol{0}. \end{aligned} \tag{38}$$

Obviously, Eq. (38) means that only the constraints in set ι_1^c participate in the optimization process. Due to $|\iota_1^c| \geq |\iota_2^c| \geq \cdots \geq |\iota_l^c|$ and $\{\boldsymbol{z} : CT_j(\boldsymbol{z}) \leq 0, j \in \iota_1^c\} \neq \emptyset$, Eq. (38) implies that the labels in ι_1^c are successfully attacked.

Further, if we set $\boldsymbol{w}_{\iota_1^c} = \boldsymbol{1}$ and $\boldsymbol{w}_{\iota_{j'}^c} = \boldsymbol{0}, \forall j' = 1, \cdots, l$, then Eq. (38) means the entire optimization process is performed only in a sub-region of $\boldsymbol{z}$. We get $\inf\left\{\inf_{\boldsymbol{z}^1 \in \mathcal{G}_{j'}, \boldsymbol{z}^2 \in \mathcal{G}_{k'}} ||\boldsymbol{z}_1 - \boldsymbol{z}_2||_2\right\} = 0$ because $j' = k' = 1$ holds. This means that the weighting strategy makes the optimization process converge. □

Supplementary Materials for Multi-label Adversarial Attack with New Measures and Self-paced Constraint Weighting

Fengguang Su, Ou Wu, and Weiyao Zhu

S.I. Supplementary Materials for Section III

A. Supplementary Materials for Section III-B

In this section, a more detailed statistical analysis of Fig. 1 is provided. Firstly, Fig. 1(a) presents a top-3 untargeted multi-label attack for the toy dataset when Eq. (1) is used. And a multi-label classifier based on multi-layer perceptron (MLP) is trained to analyze the defects of Eq. (1). According to the output of the MLP classifier, the top-3 prediction label set $\hat{Y}_3(\boldsymbol{x}_1+\boldsymbol{z}_1^1)$ of $\boldsymbol{x}_1^1$ $(=\boldsymbol{x}_1+\boldsymbol{z}_1^1)$ is $\{1,2,3\}$ and the top-3 prediction label set $\hat{Y}_3(\boldsymbol{x}_1+\boldsymbol{z}_1^2)$ of $\boldsymbol{x}_1^2$ $(=\boldsymbol{x}_1+\boldsymbol{z}_1^2)$ is $\{5,6,4\}$. Since $Y=\{2,3,4\}$, $\mathbb{I}(Y\subset\hat{Y}_k(\boldsymbol{x}_1+\boldsymbol{z}_1^1))=0$, $\mathbb{I}(\hat{Y}_k(\boldsymbol{x}_1+\boldsymbol{z}_1^1)\subset Y)=0$, $\mathbb{I}(\hat{Y}_k(\boldsymbol{x}_1+\boldsymbol{z}_1^1)=Y)=0$, and $\mathbb{I}(Y\subset\hat{Y}_k(\boldsymbol{x}_1+\boldsymbol{z}_1^2))=0$, $\mathbb{I}(\hat{Y}_k(\boldsymbol{x}_1+\boldsymbol{z}_1^2)\subset Y)=0$, $\mathbb{I}(\hat{Y}_k(\boldsymbol{x}_1+\boldsymbol{z}_1^2)=Y)=0$ hold. Then, according to the definition of Eq. (1), $\text{AFD}_{lc}(\boldsymbol{x}_1,\boldsymbol{z}_1^1)=\text{AFD}_{lc}(\boldsymbol{x}_1,\boldsymbol{z}_1^2)=0$ holds. If $||\boldsymbol{z}_1^1||\leq||\boldsymbol{z}_1^2||$, then $\boldsymbol{x}_1^1$ $(=\boldsymbol{x}_1+\boldsymbol{z}_1^1)$ rather than $\boldsymbol{x}_1^2$ $(=\boldsymbol{x}_1+\boldsymbol{z}_1^2)$ may be the solution. However, $\boldsymbol{x}_1^2$ affords a smaller failure degree than $\boldsymbol{x}_1^1$ because $1=|Y\cap\hat{Y}_k(\boldsymbol{x}_1^2)|<|Y\cap\hat{Y}_k(\boldsymbol{x}_1^1)|=2$. Hence, an Eq. (1)-based optimization goal is inappropriate.

Similarly, a multi-label classifier based on MLP is trained on the toy dataset shown in Fig. 1(b). Based on the predictions of this multi-label classifier on $Y=\{1,4,5\}$, $f_1(\boldsymbol{x}_2+\boldsymbol{z}_2^1)\geq 0.5$, $f_4(\boldsymbol{x}_2+\boldsymbol{z}_2^1)<0.5$, $f_5(\boldsymbol{x}_2+\boldsymbol{z}_2^1)<0.5$, and $f_1(\boldsymbol{x}_2+\boldsymbol{z}_2^1)\geq 0.5$, $f_4(\boldsymbol{x}_2+\boldsymbol{z}_2^1)<0.5$, $f_5(\boldsymbol{x}_2+\boldsymbol{z}_2^1)\geq 0.5$ hold. Then, according to $Y_{lf}(\boldsymbol{z})=\{j: f_j(\boldsymbol{x}+\boldsymbol{z})\geq 0.5,\forall j\in Y\}$, $Y_{lf}(\boldsymbol{z}_2^1)=\{1\}$ and $Y_{lf}(\boldsymbol{z}_2^2)=\{1,5\}$ also hold. Further, according to the definition of $\text{AFD}_{lf}(\boldsymbol{x},\boldsymbol{z})=\mathbb{I}(|Y_{lf}(\boldsymbol{z})|\neq 0)$, we have $\text{AFD}_{lf}(\boldsymbol{x}_2,\boldsymbol{z}_2^1)=\text{AFD}_{lf}(\boldsymbol{x}_2,\boldsymbol{z}_2^2)=1$. Although $\boldsymbol{x}_2^1$ and $\boldsymbol{x}_2^2$ have the identical AFD_{lf} value, $\boldsymbol{x}_2^1$ makes more labels (labels 4 and 5) with prediction confidence below 0.5 than the adversarial image $\boldsymbol{x}_2^2$ makes (label 4), and the measure in Eq. (2) cannot distinguish them.

Thirdly, the toy dataset in Fig. 1(c) is used to analyze $\text{AC}(\boldsymbol{x},\boldsymbol{z})$. Likewise, a multi-label classifier based on MLP is trained on this toy dataset. According to the trained MLP model, $\mathbb{I}(f_1(\boldsymbol{x}_3)\geq 0.5)=0$, $\mathbb{I}(f_4(\boldsymbol{x}_3)\geq 0.5)=1$, and $\mathbb{I}(f_5(\boldsymbol{x}_3)\geq 0.5)=1$ hold. For the data point $\boldsymbol{x}_3^1(=\boldsymbol{x}_3+\boldsymbol{z}_3^1)$, we have $\mathbb{I}(f_1(\boldsymbol{x}_3)\geq 0.5)=1$, $\mathbb{I}(f_4(\boldsymbol{x}_3)\geq 0.5)=0$, and $\mathbb{I}(f_5(\boldsymbol{x}_3)\geq 0.5)=0$. And for the data point, $\mathbb{I}(f_1(\boldsymbol{x}_3)\geq 0.5)=0$, $\mathbb{I}(f_4(\boldsymbol{x}_3)\geq 0.5)=1$, and $\mathbb{I}(f_5(\boldsymbol{x}_3)\geq 0.5)=0$ also hold. Then $B_I(\boldsymbol{x}_3,\boldsymbol{z}_3^1)=\emptyset$ and $B_I(\boldsymbol{x}_3,\boldsymbol{z}_3^2)=\{1,4\}$ are obtained. Further, according to the definition of $\text{AC}(\boldsymbol{x},\boldsymbol{z})=\mathbb{I}(|B_I(\boldsymbol{x},\boldsymbol{z})|\neq|B|)$, $\text{AC}(\boldsymbol{x}_3,\boldsymbol{z}_3^1)=1$ and $\text{AC}(\boldsymbol{x}_3,\boldsymbol{z}_3^2)=1$ hold. If the measure in Eq. (3) is used. However, the actual cost of $\boldsymbol{x}_3^2$ $(=\boldsymbol{x}_3+\boldsymbol{z}_3^2)$ is obviously smaller than that of $\boldsymbol{x}_3^1$ because $B\cap B_I(\boldsymbol{x}_3,\boldsymbol{z}_3^1)=\emptyset$ and $B\cap B_I(\boldsymbol{x}_3,\boldsymbol{z}_3^2)=\{1,4\}$. Therefore, the measure in Eq. (3) may impose a higher cost when no ideal attack exists.

B. Supplementary Materials for Section III-C

A video, named "Untargeted attack procedure.mp4", shows the optimization process of kFool, PGD, T_kML, and our method, which attack a pre-trained multi-layer perceptron (MLP) on a 2D toy example dataset. The results show that other methods are more likely to diverge during the attack process, while our method is more stable. The video can be available at https://github.com/ffgg11/MASW.

S.II. Supplementary Materials for Section IV

A. Supplementary Materials for Section IV-A-1

In this section, the reason for using $\min\{|Y|,|\hat{Y}_k(\boldsymbol{x}+\boldsymbol{z})|\}$ as the denominator is explained in detail. The first measure is the soft label consistency (SLC):

$$\text{AFD}_{slc}(\boldsymbol{x},\boldsymbol{z})=|Y\cap\hat{Y}_k(\boldsymbol{x}+\boldsymbol{z})|\big/\min\{|Y|,|\hat{Y}_k(\boldsymbol{x}+\boldsymbol{z})|\}. \tag{S.1}$$

In Eq. (S.1), we use $\min\{|Y|,|\hat{Y}_k(\boldsymbol{x}+\boldsymbol{z})|\}$ as the denominator. The reason is : *In experiments, we found that the attack performance was negatively affected when the maximum set $|Y\cup\hat{Y}_k(\boldsymbol{x}+\boldsymbol{z})|$ was used, and the evaluation metric, namely, soft attack success rate (SASR), derived from the maximum set was highly correlated with k, which meant that the actual attack success rate could not be reflected. Therefore, we just use $\min\{|Y|,|\hat{Y}_k(\boldsymbol{x}+\boldsymbol{z})|\}$ as the denominator of AFD_{slc}.*

B. Supplementary Materials for Section IV-A-4

Below, more results regarding the measure for AC are provided. A new measure based on the Jaccard index [8] for AC is defined as follows:

$$\text{AC}(\boldsymbol{x},\boldsymbol{z})=1-\frac{|B_I\cap B|}{|B_I\cup B|}=1-\frac{|B_I|}{|B|}. \tag{S.2}$$

Next, we formulate the following constrained optimization problem for targeted attacks

$$\min_{\boldsymbol{z}}\ [-\text{AFD}_{slc},\ \text{AFD}_{slf},\text{AC},\ ||\boldsymbol{z}||_2^2/2],\ \text{s.t. }\boldsymbol{x}+\boldsymbol{z}\in[-1,1]^n. \tag{S.3}$$

Therefore, a new optimization problem with the least constraint violation for targeted attack can be constructed as follows

$$\begin{aligned}\min_{\zeta_j',\zeta_j'',\boldsymbol{z}}\ & ||\boldsymbol{z}||_2^2+\frac{\lambda_1}{2}\boldsymbol{\zeta}'^T\mathfrak{D}(\boldsymbol{w}')\boldsymbol{\zeta}'+\frac{\lambda_2}{2}\boldsymbol{\zeta}''^T\mathfrak{D}(\boldsymbol{w}'')\boldsymbol{\zeta}''\\ \text{s.t. }& f_j(\boldsymbol{x}+\boldsymbol{z})\le f_{[k+1]}(\boldsymbol{x}+\boldsymbol{z})+\zeta_j';\ \forall j\in C\setminus Y_t,\\ & f_j(\boldsymbol{x}+\boldsymbol{z})+\zeta_j'\ge f_{[k+1]}(\boldsymbol{x}+\boldsymbol{z});\ \forall j\in Y_t,\\ & f_j(\boldsymbol{x}+\boldsymbol{z})\le 0.5+\zeta_j'';\ \forall j\in Y\cup B_0,\\ & f_j(\boldsymbol{x}+\boldsymbol{z})+\zeta_j''\ge 0.5;\ \forall j\in B_1,\\ & \boldsymbol{x}+\boldsymbol{z}\in[-1,1]^n;\end{aligned}\tag{S.4}$$

where $B_1 = j\in B: f(\boldsymbol{x}+\boldsymbol{z})>=0.5$ and $B_0 = j\in B: f(\boldsymbol{x}+\boldsymbol{z})<0.5$.

According to Eq. (S.2), a new evaluation metric, namely soft AC (SAC), is defined as follows:

$$\text{SAC}=1-\frac{\sum_{\boldsymbol{x}\in X_v}\text{AC}(\boldsymbol{x},\boldsymbol{z})}{N_v}.\tag{S.5}$$

C. Supplementary Materials for Section IV-B

In this section, the transformation process from (14) to (15) is provided first. Then, the mathematical transformations among optimization problems (15), (18), and (21), as well as their differences and similarities, are described.

Following [13], the optimization objective in (14) is minimized by first solving the constrained optimization problem as follows:

$$\begin{aligned}\min_{\boldsymbol{\zeta},\boldsymbol{z}}\ & \frac{1}{2}\boldsymbol{\zeta}^T\mathfrak{D}(\boldsymbol{w})\boldsymbol{\zeta}\\ \text{s.t. }& \boldsymbol{0}\le\boldsymbol{CT}(\boldsymbol{z})+\boldsymbol{\zeta}\end{aligned},\tag{S.6}$$

Then a solution set of (S.6) is $\mathcal{S}=\{(\boldsymbol{z},\boldsymbol{\zeta}): \min_{\boldsymbol{\zeta},\boldsymbol{z}}\frac{1}{2}\boldsymbol{\zeta}^T\mathfrak{D}(\boldsymbol{w})\boldsymbol{\zeta}, \text{s.t. } \boldsymbol{0}\le\boldsymbol{CT}(\boldsymbol{z})+\boldsymbol{\zeta}\}$. Second, a final adversarial attack with the smallest ℓ_2 norm can be obtained from $\mathcal{S}$. Then, (15) is obtained in our manuscript.

Below, more details on (15), (18) and (21) are provided: Following [13], the sub-optimization problem in optimization problem (15) is dealt with first, which is the constrained optimization problem (S.6). Then $\Theta(\boldsymbol{z})$ is introduced for (S.6) as follows:

$$\begin{aligned}\Theta(\boldsymbol{z})&=\min_{\boldsymbol{\zeta}}\{\frac{1}{2}\boldsymbol{\zeta}^T\mathfrak{D}(\boldsymbol{w})\boldsymbol{\zeta}:\boldsymbol{CT}(\boldsymbol{z})+\boldsymbol{\zeta}\ge 0\}\\&=\frac{1}{2}[\boldsymbol{CT}(\boldsymbol{z})]_-^T\mathfrak{D}(\boldsymbol{w})[\boldsymbol{CT}(\boldsymbol{z})]_-,\end{aligned}\tag{S.7}$$

where $[\boldsymbol{CT}(\boldsymbol{z})]_-=\min\{\boldsymbol{0},\boldsymbol{CT}(\boldsymbol{z})\}$. Then the following derivative of $\Theta(\boldsymbol{z})$ with respect to $\boldsymbol{z}$ is easily obtained:

$$\nabla\Theta(\boldsymbol{z})=\mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}))^T\mathfrak{D}(\boldsymbol{w})[\boldsymbol{CT}(\boldsymbol{z})]_-,\tag{S.8}$$

where $\mathcal{J}(\cdot)$ is the jacobian matrix.

Let $\boldsymbol{v}=\boldsymbol{CT}(\boldsymbol{z})+\boldsymbol{\zeta}$. Following [13], then (15) can be extended as follows:

$$\min_{\boldsymbol{z}}\ \frac{1}{2}||\boldsymbol{z}||_2^2 \text{ s.t. } \boldsymbol{F}(\boldsymbol{z},\boldsymbol{\zeta},\boldsymbol{v})=\boldsymbol{0},\ (\boldsymbol{\zeta},\boldsymbol{v})\in\Omega,\tag{S.9}$$

where $\Omega=\{(\boldsymbol{\zeta},\boldsymbol{v}):\boldsymbol{0}\le\boldsymbol{\zeta}\perp\boldsymbol{v}\ge\boldsymbol{0}\}$ and $F(\boldsymbol{z},\boldsymbol{\zeta},\boldsymbol{v})=\begin{bmatrix}-\mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}))\mathfrak{D}(\boldsymbol{w})\boldsymbol{\zeta}\\ \boldsymbol{CT}(\boldsymbol{z})+\boldsymbol{\zeta}-\boldsymbol{v}\end{bmatrix}$. Then, the conversion between (15) to (18) is completed.

Further, following [13], denote $\boldsymbol{G}(\boldsymbol{z},\boldsymbol{\zeta},\boldsymbol{v})$ as follows:

$$\boldsymbol{G}(\boldsymbol{z},\boldsymbol{\zeta},\boldsymbol{v})=[-\mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}))\mathfrak{D}(\boldsymbol{w})\boldsymbol{\zeta},\ \boldsymbol{CT}(\boldsymbol{z})+\boldsymbol{\zeta}-\boldsymbol{v},\ \min\{\boldsymbol{\zeta},\boldsymbol{v}\}]^T.\tag{S.10}$$

Assume $\boldsymbol{G}(\cdot,\cdot,\cdot)$ is a Lipschitz continuous mapping. Following [13], (15) can be extended as follows:

$$\min_{\boldsymbol{z}}\ ||\boldsymbol{z}||_2^2/2 \text{ s.t. } \boldsymbol{G}(\boldsymbol{z},\boldsymbol{\zeta},\boldsymbol{v})=\boldsymbol{0}.\tag{S.11}$$

Then, (21) is obtained. As mentioned in [13], when the optimization problem (15) is convex, the optimization problem (18) is the equivalent form of optimization problem (15). Further, if (15) is a nonlinear optimization problem with possible inconsistent constraints, then (21) is a mathematical program with complementarity constraints (MPCC), which is equivalent to the minimization problem with least constraint violation in (15). A deep learning model $\boldsymbol{f}(\cdot)$ may be convex or non-convex. Therefore, our theoretical analysis on weighting strategy is universally applicable to various situations.

D. Supplementary Materials for Section IV-C

In addition to the Lagrangian form (Eq. (23)), a innovative adversarial attack algorithm in this section is derived based on Theorem 3.

Eq. (22) in Theorem 3 gives the necessary condition that an optimal perturbation $\boldsymbol{z}^*$ will satisfy. Note that:

$$\begin{aligned}\mathfrak{L}(\boldsymbol{z},\mathfrak{v},\boldsymbol{\eta}_1,\varsigma)=\frac{1}{2}\Bigg|\Bigg|\Bigg(&\mathcal{J}(\boldsymbol{CT}(\boldsymbol{z}))^T\mathfrak{D}(\mathfrak{v})\mathcal{J}(\boldsymbol{CT}(\boldsymbol{z})^T\mathfrak{D}(\boldsymbol{w}))\\&+\sum_{j=1}^{2|Y|}\zeta_j^*w_j\nabla^2CT_j(\boldsymbol{z}^*)\Bigg)\boldsymbol{\eta}_1-\varsigma\boldsymbol{z}\Bigg|\Bigg|_2^2.\end{aligned}\tag{S.12}$$

Theorem 3 states that there is a non-zero solution $[\boldsymbol{z}^*,\mathfrak{v}^*,\boldsymbol{\eta}_1^*,\varsigma^*]^T$ that satisfies $\mathfrak{L}(\boldsymbol{z}^*,\mathfrak{v}^*,\boldsymbol{\eta}_1^*,\varsigma^*)=\boldsymbol{0}$. Therefore, we model the adversarial attack as a nonlinear system of equations solving problem. The damped least squares method (DLS) [11] is often used to solve nonlinear systems of equation. It updates variables as follows:

$$\mathcal{J}(\mathfrak{L})^T\mathcal{J}(\mathfrak{L})\boldsymbol{\delta}=-\mathcal{J}(\mathfrak{L})^T\mathfrak{L},\tag{S.13}$$

where $\mathcal{J}(\mathfrak{L})$ is the Jacobian matrix of $\mathfrak{L}$ at $(\boldsymbol{z},\mathfrak{v},\boldsymbol{\eta}_1,\varsigma)$, and $\boldsymbol{\delta}=\Delta[\boldsymbol{z},\mathfrak{v},\boldsymbol{\eta}_1,\varsigma]^T$ is the increment for $[\boldsymbol{z}^*,\mathfrak{v}^*,\boldsymbol{\eta}_1^*,\varsigma^*]^T$. A damping factor λ_D is introduced to make Eq. (S.13) more stable [11]. Thus, a novel updated version is as follows:

$$(\mathcal{J}(\mathfrak{L})^T\mathcal{J}(\mathfrak{L})+\lambda_D\boldsymbol{I})\boldsymbol{\delta}=-\mathcal{J}(\mathfrak{L})^T\mathfrak{L},\tag{S.14}$$

where I is an identity matrix. Then we have:

$$\boldsymbol{\delta}=-(\mathcal{J}(\mathfrak{L})^T\mathcal{J}(\mathfrak{L})+\lambda_D\boldsymbol{I})^{-1}\mathcal{J}(\mathfrak{L})^T\mathfrak{L}.\tag{S.15}$$

The (non-negative) damping factor λ_D is adjusted at each iteration. Generally, if $\mathfrak{L}$ decreases, then λ_D also decreases; conversely, λ_D increases. An increase by $\mathfrak{f}_{in}=2$ and a decrease by $\mathfrak{f}_{de}=3$ have been shown to be effective, while for large problems, $\mathfrak{f}_{in}=1.5, \mathfrak{f}_{de}=5$ can work better [11].

Solving $\mathfrak{L}(\boldsymbol{z}^*,\mathfrak{v}^*,\boldsymbol{\eta}_1^*,\varsigma^*)=\boldsymbol{0}$ also includes the following three main steps:

- Perform Eq. (S.15) induced by DLS to minimize (S.12) in the first iteration step.

Algorithm S.1 DSW-based Untargted Attack (**MADWU**).

Input: $\boldsymbol{x}$, Y, λ_1, λ_2, model $\boldsymbol{f}(\cdot)$, T, η, $\mathfrak{f}_{in}, \mathfrak{f}_{de}$, λ_D
Output: Adversarial perturbation $\boldsymbol{z}^*$

Initialize $\boldsymbol{z}^0, \mathfrak{v}^0, \boldsymbol{\eta}_1^0, \varsigma^0, \lambda_D^0, \boldsymbol{w}'^0, \boldsymbol{w}''^0$;
Denote $\mathfrak{L}^0 = \mathfrak{L}(\boldsymbol{z}^0, \mathfrak{v}^0, \boldsymbol{\eta}_1^0, \varsigma^0)$;
for t = 1 to T
  Update $\boldsymbol{\delta}^t$ by Eq. (S.15);
  $[\boldsymbol{z}^t, \mathfrak{v}^t, \boldsymbol{\eta}_1^t, \varsigma^t]^T = [\boldsymbol{z}^{t-1}, \mathfrak{v}^{t-1}, \boldsymbol{\eta}_1^{t-1}, \varsigma^{t-1}]^T + \boldsymbol{\delta}^t$;
  if $\mathfrak{L}^t < \mathfrak{L}^{t-1}$ **then**
    $\epsilon^t = \eta * \epsilon^{t-1}$;
    $\boldsymbol{z}^* = \boldsymbol{z}^t$;
    $\lambda_D^t = \lambda_D^{t-1} / \mathfrak{f}_{de}$;
  else
    $\epsilon^t = (\epsilon^{t-1} + \epsilon^{t-2})/2$;
    $\mathfrak{L}^t = \mathfrak{L}^{t-1}$;
    $\epsilon^{t-1} = \epsilon^{t-2}$;
    $\lambda_D^t = \lambda_D^{t-1} * \mathfrak{f}_{in}$;
  if $||\nabla \mathcal{J}(\mathfrak{L}^t)||_2 \leq 10^{-3}$ **then** break;
  Update $\boldsymbol{w}'^t$ and $\boldsymbol{w}''^t$ by using Eq. (24);
return $\boldsymbol{z}^*$.

Algorithm S.2 Universal untargted Attack (**MADWUv**).

Input: X_t, λ_1, λ_2, E, $\boldsymbol{f}(\cdot)$, T, η
Output: Universal perturbation $\boldsymbol{z}^*$

Initialize $\boldsymbol{z}^* = \boldsymbol{0}$, $E_1, E_2 = 0$, $\tau_1, \tau_2 = 0$;
while true **do**
  for $(\boldsymbol{x}_i, Y_i) \in X_t$ **do**
    $\boldsymbol{z}_i^*$ = MADWU($\boldsymbol{x}_i$, Y_i, λ_1, λ_1, $\boldsymbol{f}$, T, η);
    $\boldsymbol{z}^* = \mathcal{P}_\tau(\boldsymbol{z}^* + \boldsymbol{z}_i^*)$;
  Calculate SASR and LFR on X;
  if SASR $< \tau_1$ **then** $E_1 = E_1 + 1$;
  else $\tau_1 =$ SASR; $E_1 = 0$;
  if LFR $< \tau_2$ **then** $E_2 = E_2 + 1$;
  else $\tau_2 =$ LFR; $E_2 = 0$;
  if $E_1 > E$ and $E_2 > E$ **then** break;
return $\boldsymbol{z}^*$.

- In the second iteration step, the threshold ϵ is updated.
- In the third iteration step, the weights for each constraint (or slack variable) are updated using Eq. (24). Thereafter, return to the first iteration step.

Similarly, ϵ is updated according to Eq. (S.12), denoted as $\mathfrak{L}^t$. If $\mathfrak{L}^t$ decreases compared with $\mathfrak{L}^{t-1}$, then ϵ is increased to allow more constraints to participate in the optimization process, i.e., $\epsilon^t = \eta \times \epsilon^{t-1}$ ($\eta > 1$); otherwise, ϵ is reduced to limit the participation of difficult constraints by $\epsilon^t = (\epsilon^{t-1} + \epsilon^{t-2})/2$. Algorithm S.1 reports the details of our implementation for the multi-label untargeted attack based on DLS and SPW, namely, **MADW**, for simplicity. Algorithm S.1 computes third-order derivatives and therefore requires a higher time cost than **MASW**. Algorithm S.2 outlines the calculation steps for universal untargeted attack, which is extended by Algorithm S.1. Additionally, following the same approach as Algorithm S.1, we can easily derive an algorithm for targeted attack, as presented in Algorithm S.3.

Algorithm S.3 DSW-based Targted Attack (**MADWT**).

Input: $\boldsymbol{x}$, Y, Y_t, λ_1, λ_2, $\boldsymbol{f}(\cdot)$, T, η
Output: Adversarial perturbation $\boldsymbol{z}^*$

Initialize $\boldsymbol{z}^0, \mathfrak{v}^0, \boldsymbol{\eta}_1^0, \varsigma^0, \lambda_D^0, \boldsymbol{w}'^0, \boldsymbol{w}''^0$;
Denote $\mathfrak{L}^0 = \mathfrak{L}(\boldsymbol{z}^0, \mathfrak{v}^0, \boldsymbol{\eta}_1^0, \varsigma^0)$ for targeted attack;
for t = 1 to T
  Update $\boldsymbol{\delta}^t$ by Eq. (S.15);
  $[\boldsymbol{z}^t, \mathfrak{v}^t, \boldsymbol{\eta}_1^t, \varsigma^t]^T = [\boldsymbol{z}^{t-1}, \mathfrak{v}^{t-1}, \boldsymbol{\eta}_1^{t-1}, \varsigma^{t-1}]^T + \boldsymbol{\delta}^t$;
  if $\mathfrak{L}^t < \mathfrak{L}^{t-1}$ **then**
    $\epsilon^t = \eta * \epsilon^{t-1}$;
    $\boldsymbol{z}^* = \boldsymbol{z}^t$;
    $\lambda_D^t = \lambda_D^{t-1} / \mathfrak{f}_{de}$;
  else
    $\epsilon^t = (\epsilon^{t-1} + \epsilon^{t-2})/2$;
    $\mathfrak{L}^t = \mathfrak{L}^{t-1}$;
    $\epsilon^{t-1} = \epsilon^{t-2}$;
    $\lambda_D^t = \lambda_D^{t-1} * \mathfrak{f}_{in}$;
  if $||\nabla \mathcal{J}(\mathfrak{L}^t)||_2 \leq 10^{-3}$ **then** break;
  Update $\boldsymbol{w}'^t$ and $\boldsymbol{w}''^t$ by using Eq. (24);
return $\boldsymbol{z}^*$.

S.III. Supplementary Materials for Section V

A. Supplementary Materials for Section V-A

Competing Methods. The comparison methods include: Fast gradient sign method (FGSM [1]), Momentum iterative fast gradient sign method (MFGSM [2]), Projected gradient descent (PGD [7]), Rank I [3], Multi-label DeepFool (MLDF [3]) and Carlini & Wagner (MLCW [3]) attacks, Multilabel attack by linear programming (MLALP [6]), Top-k universal untargeted attck (kUv [5]) and untarged attck by DeepFool (kFool [5]), and T$_k$ML [4], Generative Adversarial Multi object Attacks (GAMA [16]), Local Patch Difference (LPD [15]), Top-k Attack with Label Correlation (T$_k$ALC [18]) and Top-k Measure Imperceptible Attack (T$_k$MIA [17]). For these methods, we use the settings in [4], [10]. Three suffixes U, Uv and T mean the untargeted attack, universal untargeted attack, and targeted attack, respectively. Then our methods are denoted as MASWU, MASWUv and MASWT.

Setting. Next, we introduce the experimental setup of the proposed MADW. DLS is used to solve the nonlinear equation $\mathfrak{L}(\boldsymbol{z}^*, \mathfrak{v}^*, \boldsymbol{\eta}_1^*, \varsigma^*) = \boldsymbol{0}$. We set $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$, η is set to 1.015, and ϵ is initialized to 0.01. We record the results when $T = 300$. Furthermore, $\boldsymbol{w}'$, and $\boldsymbol{w}''$ are initialized to obey a uniform distribution of [0, 1]. 3000 images from the validation set of each benchmark dataset are selected to build X_t for universal untargeted attack. Additionally, we apply early stopping [9] on X_t to terminate Algorithm S.2. The patience of early stopping E is set to 40.

The damping factor λ_D in MADW is dynamically adjusted at each step. [11] suggested that for large-scale problems, $\mathfrak{f}_{in} = 1.5$ and $\mathfrak{f}_{de} = 5$ are used to adjust λ_D. We use this setting. These settings result in relatively good attack performance, shown in Tables S.I–S.VIII.

TABLE S.I
COMPARSION OF UNTARGETED ATTACK METHODS. BOLD NUMBERS HIGHLIGHT THE BEST RESULTS.

k	T	Method	VOC 2012 ASR	VOC 2012 SASR	VOC 2012 Pert	VOC 2012 LFR	COCO 2014 ASR	COCO 2014 SASR	COCO 2014 Pert	COCO 2014 LFR
3	1	FGSMU	23.00	31.06	3.99	47.30	18.10	51.62	7.31	40.55
	40	MFGSMU	17.20	24.70	**0.32**	53.74	22.20	56.87	0.56	37.35
	300	MLCWU	19.80	29.46	2.38	50.27	25.00	33.94	3.71	45.59
		MLDFU	18.00	29.17	1.74	45.28	23.60	31.19	2.76	44.37
		PGDU	85.17	93.21	3.67	85.42	90.43	97.72	5.71	91.75
		MLALP	43.90	57.48	0.68	79.75	54.70	66.47	0.56	82.39
		kFool	93.50	96.87	1.42	94.05	60.80	83.20	4.41	68.66
		T_kMLU	95.53	95.73	0.48	97.88	99.83	99.86	**0.51**	98.36
		LPD	89.90	90.11	0.91	90.42	92.30	93.86	0.73	91.97
		GAMA	91.30	92.57	0.73	91.05	94.40	95.39	0.69	93.78
		T_kALCU	95.97	96.11	0.52	98.18	99.40	99.52	0.56	98.01
		T_kMIAU	96.03	96.17	0.55	98.14	99.73	99.75	0.53	98.14
		MADWU	96.90	97.02	0.50	99.54	99.87	99.94	0.52	99.12
		MASWU	**97.13**	**97.34**	0.49	**99.83**	**99.90**	**99.95**	**0.51**	**99.46**
5	1	FGSMU	17.30	24.43	4.00	47.39	14.40	44.28	7.27	40.62
	40	MFGSMU	11.80	17.73	**0.33**	52.91	18.00	50.74	0.56	37.38
	300	MLCWU	18.00	26.95	2.45	51.74	23.70	31.29	3.86	48.74
		MLDFU	17.20	27.06	1.85	49.21	22.10	30.37	2.82	47.81
		PGDU	85.27	93.25	3.75	85.95	89.47	97.53	5.87	92.93
		MLALP	43.20	57.12	0.69	80.26	54.00	65.45	0.58	84.02
		kFool	93.60	95.78	2.35	95.81	65.10	84.69	7.81	76.91
		T_kMLU	93.33	93.76	0.52	98.06	99.67	99.71	**0.54**	98.67
		LPD	87.23	89.26	0.91	90.42	91.13	92.28	0.73	91.97
		GAMA	90.10	91.23	0.73	91.05	92.47	93.74	0.69	93.78
		T_kALCU	94.10	95.20	0.56	98.27	99.70	99.72	0.58	98.26
		T_kMIAU	94.37	95.20	0.57	98.36	99.53	99.60	0.55	98.39
		MADWU	96.00	96.02	0.52	99.44	99.73	99.74	0.55	99.07
		MASWU	**96.17**	**96.23**	0.52	**99.92**	**99.80**	**99.78**	**0.54**	**99.63**
10	1	FGSMU	9.90	14.81	3.98	47.30	11.20	35.34	7.29	40.55
	40	MFGSMU	6.60	10.12	**0.32**	53.24	14.30	41.83	**0.57**	37.01
	300	MLCWU	15.20	25.61	2.52	53.91	20.60	29.17	3.88	49.16
		MLDFU	17.10	26.54	1.87	52.38	20.00	27.97	2.86	48.14
		PGDU	85.00	92.76	3.85	86.20	87.67	97.08	6.08	93.58
		MLALP	42.10	56.52	0.69	81.27	52.40	64.28	0.58	84.95
		kFool	88.40	90.18	4.95	97.12	68.00	85.82	14.95	85.82
		T_kMLU	87.93	88.43	0.57	98.15	99.47	99.52	0.60	98.90
		LPD	86.73	88.09	0.91	90.42	90.10	91.12	0.73	91.97
		GAMA	88.07	89.18	0.73	91.05	90.17	91.06	0.69	93.78
		T_kALCU	89.17	90.29	0.60	98.38	99.40	99.51	0.64	98.48
		T_kMIAU	89.30	90.67	0.59	98.40	99.10	99.26	0.62	98.72
		MADWU	90.57	90.13	0.59	99.66	99.80	99.82	0.61	99.48
		MASWU	**91.40**	**91.89**	0.57	**99.97**	**99.93**	**99.98**	0.60	**99.89**

TABLE S.II
COMPARSION OF UNIVERSAL UNTARGETED ATTACK METHODS WITH $k = 2, 3, 5, 10$ ON VOC 2012 AND COCO 2014. BOLD NUMBERS MEAN THE BEST RESULTS. ✗ MEANS THAT THE METHOD CAN NOT OUTPUT THE RESULT.

k	Method	VOC 2012 ASR	VOC 2012 SASR	VOC 2012 Pert	VOC 2012 LFR	COCO 2014 ASR	COCO 2014 SASR	COCO 2014 Pert	COCO 2014 LFR
2	PGDUv	60.13	64.76	34.12	77.19	73.23	86.52	71.52	81.28
	kUv	✗	✗	✗	✗	72.90	86.18	51.38	79.56
	T_kMLUv	67.37	71.52	16.27	92.55	79.23	88.78	15.26	94.44
	T_kALCUv	66.30	69.12	18.41	90.15	78.20	86.82	17.06	92.41
	T_kMIAUv	66.60	70.29	17.21	91.59	78.50	87.18	16.46	92.64
	MADWUv	68.13	72.19	15.79	93.97	80.17	89.08	14.99	95.40
	MASWUv	**68.47**	**72.21**	**15.63**	**94.39**	**80.50**	**89.17**	**14.79**	**95.61**
3	PGDUv	57.73	63.41	39.54	78.92	71.57	76.29	74.11	83.95
	kUv	✗	✗	✗	✗	61.40	73.64	51.26	80.43
	T_kMLUv	64.43	68.11	17.28	96.11	78.07	87.24	16.36	94.82
	T_kALCUv	62.10	66.13	20.18	94.01	77.50	84.47	19.31	92.91
	T_kMIAUv	63.20	66.71	19.29	94.28	78.30	85.18	18.69	93.12
	MADWUv	64.80	68.89	16.95	96.36	80.17	87.17	15.97	95.42
	MASWUv	**64.90**	**69.13**	**16.59**	**96.41**	**80.33**	**87.34**	**15.62**	**95.71**
5	PGDUv	53.97	62.37	43.53	80.54	70.73	80.58	75.61	85.13
	kUv	✗	✗	✗	✗	69.80	78.82	52.34	85.51
	T_kMLUv	61.87	66.22	19.27	98.01	78.40	85.29	17.19	96.01
	T_kALCUv	61.00	65.21	23.73	95.16	76.50	84.09	21.09	94.03
	T_kMIAUv	61.80	65.12	23.15	95.26	77.10	84.49	20.17	94.51
	MADWUv	63.47	67.79	18.55	98.72	78.70	86.02	16.48	96.15
	MASWUv	**63.60**	**67.97**	**18.57**	**98.99**	**78.83**	**86.19**	**16.47**	**96.66**
10	PGDUv	35.33	38.94	49.55	81.02	63.83	66.79	79.62	85.98
	kUv	✗	✗	✗	✗	✗	✗	✗	✗
	T_kMLUv	41.63	47.93	22.84	98.69	74.07	83.49	18.61	97.92
	T_kALCUv	40.20	47.11	25.19	96.92	74.10	82.21	23.34	96.42
	T_kMIAUv	41.00	47.41	23.81	97.09	74.40	83.02	22.14	97.10
	MADWUv	42.63	48.92	22.59	99.02	75.90	85.12	18.26	98.04
	MASWUv	**42.97**	**49.53**	**22.56**	**99.39**	**76.07**	**85.79**	**18.25**	**98.39**

B. Supplementary Materials for Section V-B

Results. Table S.I reports the top-k attack performance for $k = 3, 5, 10$. The results infer that the proposed MADWU and MASWU achieve the best or comparable results. FGSMU uses single-step GD ($T = 1$) and a large learning rate, still requiring a larger Pert and performing poorly in terms of ASR, SASR, and LFR. For MFGSMU, which utilizes a few-step GD ($T = 40$) and the momentum, although it requires the smallest Pert, its performance is poor. Moreover, the multi-label attack methods MLCWU and MLDFU explicitly lower the prediction confidence of the ground truth below a certain threshold but only achieve a low LFR. Both PGDU and kFool are initially designed for single-label learning. They usually use larger perturbation bound Pert to attack model, but the attack performance is still inferior to MADWU and MASWU. MLALP uses the interior point method to solve linear programs, but its performance is still lower than that of MADWU and MASWU. Compared with the SOTA method T_kMLU, MADWU and MASWU achieve better results on ASR, SASR, and LFR for $T = 300$ while yielding similar Pert values. Same or smaller Perts imply that T_kMLU may just obtain suboptimal solutions. Both LPD and GAMA utilize complex generative models, yet rely on larger perturbation bounds (Pert), resulting in lower ASR, SASR, and LFR values compared to our method. Though T_kALCU and T_kMIAU use label correlation and design complex optimization problems in adversarial attacks respectively, they are still inferior to our method.

C. Supplementary Materials for Section V-C

Results. Table S.II presents the results of each universal untargeted attack method on VOC 2012 and COCO 2014. Our methods MADWUv and MASWUv use the smaller or smallest Pert to achieve the best ASR, SASR, and LFR. ✗ in Table S.II means that the method kUv takes an excessive amount of time (more than a week) but could not produce the result. Therefore we do not report this result. PGDUv usually utilizes the highest Pert but achieves lower ASR, SASR and LFR than our methods. Compared to T_kMLUv, T_kALCUv, and T_kMIAUv, the proposed methods achieves higher ASR, SASR, and LFR only with a lower Pert.

D. Supplementary Materials for Section V-D

Additional settings. Following [4], this paper considers three target types Y_t, namely, worst, random, and best cases. These cases mean that labels in Y_t have the lowest prediction scores, labels in Y_t are selected randomly, and labels in Y_t have the largest prediction scores, respectively.

Results. Table S.III reports the results for top-k targeted attacks under the best case. When similar Pert values are achieved on two datasets, MADWT and MASWT outperform all competitor methods considering the ASR, SASR, and LFR metrics for different k values. Although T_kMLT and Rank I

TABLE S.III
COMPARISON OF THE TARGETED ATTACK METHODS UNDER THE BEST CASE SCENARIO.

k	T	Method	VOC 2012 ASR	SASR	Pert	LFR	COCO 2014 ASR	SASR	Pert	LFR
3	1	FGSMT	5.70	62.83	0.80	23.41	6.70	52.00	1.48	29.68
	40	MFGSMT	11.50	66.19	**0.20**	26.68	16.30	59.13	**0.34**	35.18
		MLCWT	82.70	93.86	0.47	86.48	82.10	92.43	0.56	88.51
		MLDFT	52.30	78.13	0.87	53.53	58.30	72.20	1.42	65.83
		PGDT	37.17	76.83	0.81	49.19	37.37	62.16	1.43	71.47
		Rank I	92.10	95.86	0.43	93.44	99.10	99.21	0.56	90.04
	300	T_kMLT	93.13	97.39	0.43	93.35	99.03	99.11	0.58	90.78
		T_kALCT	93.17	97.41	0.45	93.70	99.10	99.20	0.61	92.29
		T_kMIAT	93.20	97.47	0.47	95.29	99.17	99.23	0.63	93.45
		MADWT	93.80	97.61	0.44	97.95	99.43	99.78	0.60	95.76
		MASWT	**93.97**	**97.79**	0.43	**98.25**	**99.50**	**99.86**	0.60	**96.37**
5	1	FGSMT	4.30	65.52	0.78	23.26	5.80	60.05	1.51	28.61
	40	MFGSMT	3.50	73.05	**0.23**	26.26	7.10	67.88	**0.38**	33.28
		MLCWT	38.00	84.14	0.55	84.88	77.20	93.45	0.78	91.65
		MLDFT	13.80	74.53	0.92	21.38	39.90	70.54	1.45	56.03
		PGDT	21.37	78.04	0.84	44.18	26.67	69.54	1.44	69.07
		Rank I	84.20	93.16	0.49	96.56	98.80	99.34	0.67	97.28
	300	T_kMLT	86.03	96.55	0.49	96.62	99.03	99.20	0.69	96.83
		T_kALCT	86.27	95.19	0.53	96.71	99.07	99.12	0.73	97.11
		T_kMIAT	86.40	95.25	0.52	97.10	99.10	99.20	0.71	97.34
		MADWT	87.93	96.77	0.50	98.14	99.50	99.79	0.69	98.11
		MASWT	**88.17**	**96.83**	0.49	**98.81**	**99.67**	**99.94**	0.69	**98.56**
10	1	FGSMT	3.70	66.28	0.81	24.15	3.70	61.89	1.69	28.31
	40	MFGSMT	2.70	80.07	**0.20**	22.88	4.40	74.71	**0.38**	30.51
		MLCWT	10.00	74.92	0.64	73.02	62.60	94.19	1.17	90.14
		MLDFT	7.80	82.41	0.99	40.56	16.90	73.82	1.47	37.03
		PGDT	18.03	77.88	0.85	36.56	6.37	70.49	1.43	57.43
		Rank I	69.90	91.11	0.52	97.09	97.10	99.02	0.81	97.79
	300	T_kMLT	75.33	96.27	0.53	97.13	98.83	99.47	0.85	97.47
		T_kALCT	75.60	97.13	0.55	97.40	99.07	99.50	0.89	98.01
		T_kMIAT	75.77	97.89	0.56	97.84	99.10	99.54	0.89	98.18
		MADWT	77.83	98.19	0.54	98.91	99.17	99.87	0.86	98.74
		MASWT	**77.97**	**99.65**	0.53	**99.16**	**99.40**	**99.91**	0.86	**99.16**

TABLE S.IV
COMPARISON OF TARGETED ATTACK METHODS UNDER RANDOM CASE.

k	T	Method	VOC 2012 ASR	SASR	Pert	LFR	COCO 2014 ASR	SASR	Pert	LFR
3	1	FGSMT	5.10	21.23	0.68	16.84	5.30	7.68	0.82	39.13
	40	MFGSMT	10.70	26.43	**0.23**	20.96	8.70	16.33	**0.45**	51.99
		MLCWT	45.90	74.43	0.60	89.70	46.80	75.19	1.04	90.06
		MLDFT	6.80	17.14	0.86	13.58	7.80	19.21	1.45	17.36
		PGDT	14.17	49.73	0.82	49.82	4.83	31.96	1.47	85.47
		Rank I	68.20	71.73	0.55	98.48	98.40	99.06	0.95	99.17
	300	T_kMLT	79.03	88.80	0.57	97.09	98.77	99.28	1.00	99.27
		T_kALCT	79.27	89.11	0.59	97.34	98.90	99.31	1.03	99.30
		T_kMIAT	79.70	89.20	0.60	97.83	99.03	99.39	1.02	99.27
		MADWT	81.57	89.11	0.58	98.89	99.47	99.47	1.01	99.56
		MASWT	**81.73**	**89.65**	0.57	**99.53**	**99.73**	**99.87**	1.00	**99.95**
5	1	FGSMT	4.30	30.89	0.70	18.04	4.70	8.78	0.84	39.63
	40	MFGSMT	8.70	32.55	**0.34**	21.15	7.60	15.46	**0.47**	52.47
		MLCWT	8.30	61.00	0.67	85.51	21.80	72.24	1.29	91.39
		MLDFT	5.50	20.79	0.90	7.98	6.50	15.64	1.46	14.01
		PGDT	12.73	43.99	0.76	44.88	4.57	28.33	1.46	87.56
		Rank I	53.10	63.44	0.58	98.01	89.00	91.66	1.10	99.29
	300	T_kMLT	68.77	86.69	0.60	96.28	95.47	97.54	1.17	99.25
		T_kALCT	69.20	87.06	0.61	97.20	95.50	97.60	1.20	99.29
		T_kMIAT	70.07	87.20	0.62	97.72	95.73	97.64	1.19	99.30
		MADWT	71.27	87.71	0.60	98.46	96.10	97.94	1.17	99.62
		MASWT	**72.37**	**88.19**	0.60	**99.29**	**96.57**	**98.35**	1.17	**99.98**
10	1	FGSMT	3.50	53.64	0.75	18.95	4.50	12.45	0.85	40.43
	40	MFGSMT	7.30	54.61	**0.35**	22.74	6.20	16.97	**0.48**	47.83
		MLCWT	6.70	61.46	0.75	73.40	11.10	48.78	1.31	92.28
		MLDFT	3.00	51.71	0.92	6.33	4.70	12.65	1.46	12.60
		PGDT	9.77	57.19	0.79	35.30	3.27	25.84	1.47	81.34
		Rank I	38.30	69.02	0.57	98.08	37.10	59.22	1.17	99.27
	300	T_kMLT	59.07	90.09	0.61	95.31	87.53	94.89	1.28	99.44
		T_kALCT	60.30	90.59	0.64	96.00	87.60	95.08	1.30	99.47
		T_kMIAT	60.90	90.64	0.63	96.20	87.63	95.10	1.29	99.40
		MADWT	62.77	90.03	0.63	98.51	87.73	95.61	1.29	99.69
		MASWT	**63.40**	**91.52**	0.61	**98.91**	**87.80**	**95.90**	1.28	**99.96**

are designed specifically for top-k targeted attacks, their attack performance is still lower than our methods regarding ASR, SASR, and LFR. Moreover, the FGSMT, MFGSMT, MLCWT, MLDFT, and PGDT attain a much lower performance than our methods by a large margin in terms of ASR, SASR, and LFR. Similar Perts imply that Rank I and T_kMLT may just obtain suboptimal solutions. Although MLCWT and MLDFT explicitly lower the prediction confidence of the ground truth below a certain threshold, they achieve low LFR values. Besides, FGSMT uses a single-step GD ($T = 1$), but it requires larger bounds and achieves poor ASR, SASR, and LFR values for targeted attacks. Although MFGSMT obtains the smallest Pert, its attack performance is still unacceptable. Although using complex methods, both T_kALCT and T_kMIAT are inferior to our methods.

Tables S.IV and S.V present the results under the random and worst cases. Similar to Table S.III, our methods achieve a better attack performance in terms of ASR, SASR, and LFR with comparable Pert. As k increases, the larger Pert is required, but the ASR and SASR generally decrease due to the increased difficulty of the attack.

E. Supplementary Materials for Section V-E

Table S.VI reports the results of untargeted attacks on two large datasets, NUS-WIDE and Open Images, revealing that our methods, MADWU and MASWU, obtain the best results for $k = 3, 5, 10$. Regarding the SOTA methods MLCWU, MLDFU, PGDU, kFool, and T_kMLU, although these require a larger perturbation bound Pert, they still achieve lower ASR, SASR and LFR values. Besides, MLCWU and MLDFU explicitly lower the confidence of the true label below a certain threshold but still achieve low LFR. Using label correlation and constructing complex constrained optimization problems respectively, yet, T_kALCT and T_kMIAT are inferior to our methods.

Table S.VII reports the results of universal untargeted attacks. Similarly, MADWUv and MASWUv obtain the lower Pert and the higher ASR, SASR, and LFR.

Table S.VIII presents the results of targeted attack for $k = 3, 5, 10$ under best, random, and worst cases. MADWT and MASWT still outperform other SOTA methods in ASR, SASR, and LFR by utilizing a lower or comparable Pert.

Table S.IX shows the results about AC, indicating that our method is superior to the existing methods.

F. Supplementary Materials for Section V-F

In this section, the experimental setups for ablation experiments and sensitivity tests are provided firstly. Further details of the ablation study and sensitivity test are provided. Similar to [19], [20], λ_2 is firstly fixed at 0.5, while λ_1 varies within the set $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ in our sensitivity test. And, as λ_1 changes, the values of ASR and LFR are recorded separately. Then Fig. 10(a) is drawn. As λ_1 increases, ASR also increases, and LFR decreases. λ_1 and λ_2 should be fine-tuned according to the application scenario of multi-label adversarial attack. From Fig. 10(a), if the optimization goal AFD_{slc} is given more weight than AFD_{slf}, then the value of λ_1 should be greater than λ_2. Otherwise the value of λ_2 should be greater than λ_1. Although the attack performance can be

TABLE S.V
COMPARISON OF TARGETED ATTACK METHODS UNDER WORST CASE.

k	T	Method	VOC 2012 ASR	SASR	Pert	LFR	COCO 2014 ASR	SASR	Pert	LFR
3	1	FGSMT	4.20	6.19	0.72	17.21	5.20	6.98	0.83	42.48
	40	MFGSMT	7.30	10.26	**0.31**	20.51	7.30	14.38	**0.49**	36.22
		MLCWT	12.70	40.06	0.66	89.30	23.40	57.93	1.13	92.35
	300	MLDFT	5.70	9.88	0.78	7.63	5.40	10.75	1.56	12.69
		PGDT	12.30	22.23	0.86	50.13	4.30	26.61	1.45	91.12
		Rank I	38.20	41.09	0.61	98.19	78.10	83.47	1.05	99.24
		T_kMLT	56.83	72.23	0.64	96.65	82.63	93.06	1.12	99.24
		T_kALCT	57.10	72.68	0.65	97.01	82.70	92.74	1.14	98.67
		T_kMIAT	58.30	73.00	0.66	97.46	82.93	92.90	1.15	98.90
		MADWT	59.77	73.82	0.64	98.54	82.73	93.11	1.13	99.42
		MASWT	**60.60**	**74.85**	0.64	**99.06**	**83.13**	**93.27**	1.13	**99.89**
5	1	FGSMT	3.90	5.62	0.73	15.74	4.80	6.26	0.84	41.61
	40	MFGSMT	6.50	9.83	**0.33**	18.29	6.60	10.25	**0.51**	34.33
		MLCWT	7.30	28.66	0.74	86.38	9.90	62.06	1.28	93.17
	300	MLDFT	4.50	8.74	0.79	7.12	4.10	9.52	1.57	13.46
		PGDT	10.70	18.71	0.94	44.11	3.87	26.58	1.58	90.73
		Rank I	22.20	28.72	0.61	97.83	62.40	71.72	1.15	99.29
		T_kMLT	45.30	70.65	0.66	95.40	73.73	91.16	1.24	99.25
		T_kALCT	45.50	71.01	0.68	96.10	73.80	91.20	1.25	99.30
		T_kMIAT	45.70	71.28	0.68	96.45	73.77	91.25	1.25	99.27
		MADWT	46.53	71.82	0.67	98.61	74.43	92.01	1.24	99.61
		MASWT	**46.97**	**72.58**	0.66	**98.81**	**74.60**	**92.31**	1.24	**99.97**
10	1	FGSMT	2.90	4.31	0.73	15.35	4.10	5.34	0.82	41.47
	40	MFGSMT	4.30	6.34	**0.35**	17.54	5.50	8.79	**0.51**	31.06
		MLCWT	5.30	37.55	0.78	77.38	6.70	58.68	1.31	93.79
	300	MLDFT	2.40	6.77	0.79	6.54	3.50	7.66	1.57	12.58
		PGDT	8.33	28.51	0.97	34.58	2.37	21.79	1.61	86.88
		Rank I	14.70	40.18	0.59	97.93	25.20	44.73	1.22	99.37
		T_kMLT	36.37	76.83	0.65	93.98	56.23	88.01	1.29	99.07
		T_kALCT	36.40	76.91	0.66	94.37	56.30	88.19	1.30	99.10
		T_kMIAT	36.57	77.10	0.66	95.10	56.43	88.21	1.31	99.06
		MADWT	37.13	78.05	0.65	98.10	57.00	88.19	1.29	99.63
		MASWT	**37.50**	**78.29**	0.65	**98.24**	**57.23**	**88.65**	1.28	**99.96**

TABLE S.VI
COMPARISON OF UNTARGETED ATTACK METHODS ON TWO LARGE SETS.

k	T	Method	NUS-WIDE ASR	SASR	Pert	LFR	Open Images ASR	SASR	Pert	LFR
3	300	MLCWU	14.70	29.65	0.87	35.84	27.60	78.17	1.39	25.22
		MLDFU	10.70	19.64	1.56	30.38	23.10	50.64	1.97	22.89
		PGDU	85.23	94.15	1.68	63.91	92.13	93.46	1.65	64.45
		kFool	69.80	88.03	1.51	36.33	87.10	93.37	2.73	40.35
		T_kMLU	96.33	97.01	0.15	79.75	97.27	97.99	0.14	85.34
		LPD	92.50	93.18	0.81	91.09	93.10	94.26	0.71	93.43
		GAMA	93.10	94.07	0.80	92.00	94.10	95.27	0.68	94.16
		T_kALCU	96.50	97.20	0.25	85.71	97.40	98.11	0.23	90.17
		T_kMIAU	96.73	97.45	0.24	86.79	97.57	98.20	0.21	91.34
		MADWU	97.73	99.11	0.16	96.29	99.40	99.16	0.15	97.14
		MASWU	**98.17**	**99.67**	**0.14**	**98.83**	**99.87**	**99.90**	**0.13**	**98.45**
5	300	MLCWU	10.00	25.47	0.62	37.29	14.20	66.11	1.35	27.46
		MLDFU	8.60	17.62	1.58	35.69	13.30	45.48	1.99	25.86
		PGDU	84.37	93.92	1.69	68.18	90.80	91.21	1.69	74.61
		kFool	72.30	88.22	5.28	47.49	82.70	92.18	5.71	51.53
		T_kMLU	95.77	96.54	0.16	81.29	96.20	97.01	0.15	86.36
		LPD	91.40	92.87	0.81	91.09	92.80	93.77	0.71	93.43
		GAMA	92.17	93.15	0.80	92.00	93.17	94.35	0.68	94.16
		T_kALCU	95.80	96.60	0.26	87.79	96.40	97.62	0.25	92.72
		T_kMIAU	95.90	96.78	0.25	89.56	96.93	98.00	0.24	93.45
		MADWU	97.07	98.48	0.15	96.45	98.77	98.76	0.15	97.24
		MASWU	**97.73**	**99.33**	**0.14**	**98.86**	**99.67**	**99.63**	**0.14**	**98.97**
10	300	MLCWU	8.50	17.81	0.59	37.64	9.10	50.84	1.09	29.02
		MLDFU	6.90	15.64	1.61	37.58	11.50	37.95	2.07	27.73
		PGDU	82.77	93.57	1.70	73.59	85.23	96.72	1.71	81.83
		kFool	71.70	87.06	11.59	60.54	81.50	91.75	14.98	63.03
		T_kMLU	94.10	95.43	0.16	85.69	94.73	95.48	0.17	87.12
		LPD	90.70	91.04	0.81	91.09	90.90	91.76	0.71	93.43
		GAMA	89.50	90.12	0.80	92.00	90.43	90.94	0.68	94.16
		T_kALCU	94.23	95.61	0.27	90.61	95.10	95.78	0.26	93.10
		T_kMIAU	94.40	95.78	0.26	91.19	95.33	96.00	0.25	94.28
		MADWU	95.03	96.56	0.17	96.57	98.43	98.27	0.16	97.43
		MASWU	**95.97**	**97.69**	**0.15**	**98.83**	**99.47**	**99.53**	**0.15**	**99.14**

TABLE S.VII
COMPARSION OF UNIVERSAL UNTARGETED ATTACK METHODS ON NUS-WIDE AND OPEN IMAGES. PERT IS IN 10^{-2}.

k	Method	NUS-WIDE ASR	SASR	Pert	LFR	Open Images ASR	SASR	Pert	LFR
2	PGDUv	69.90	76.18	8.91	57.64	75.53	80.71	7.45	66.48
	kUv	65.80	69.34	7.98	53.91	73.60	78.62	6.24	61.46
	T_kMLUv	77.93	87.07	3.93	63.08	82.70	94.19	3.41	81.17
	T_kALCUv	77.30	87.15	4.41	70.21	82.10	93.79	3.89	82.08
	T_kMIAUv	77.60	87.56	4.13	73.31	82.40	94.01	3.72	83.07
	MADWUv	79.40	89.68	3.92	79.39	84.17	96.02	3.27	84.10
	MASWUv	**79.93**	**90.31**	**3.89**	**80.79**	**84.97**	**97.24**	**3.23**	**85.93**
3	PGDUv	65.47	73.46	8.96	60.93	70.47	76.58	7.49	67.93
	kUv	60.70	65.96	8.04	55.78	67.80	73.48	6.63	65.28
	T_kMLUv	71.57	84.29	4.20	74.39	80.27	94.89	3.25	82.37
	T_kALCUv	72.70	83.28	4.91	77.47	79.10	94.25	4.11	83.62
	T_kMIAUv	73.10	84.11	4.57	80.25	80.20	94.35	4.05	84.17
	MADWUv	74.73	85.91	4.18	85.99	82.00	95.68	3.19	86.69
	MASWUv	**75.47**	**86.87**	**4.17**	**86.79**	**82.43**	**96.62**	**3.12**	**87.69**
5	PGDUv	60.43	67.65	9.04	64.56	64.53	71.45	7.91	70.91
	kUv	57.50	63.48	8.11	57.59	62.10	69.48	6.77	67.94
	T_kMLUv	69.03	82.07	4.61	80.13	77.03	95.46	3.56	83.07
	T_kALCUv	69.00	81.24	5.91	80.03	77.10	93.59	5.26	84.67
	T_kMIAUv	69.40	82.31	5.56	80.74	77.50	94.59	4.96	85.32
	MADWUv	71.50	83.47	4.59	91.92	80.17	95.72	3.34	86.78
	MASWUv	**72.27**	**84.35**	**4.55**	**93.39**	**80.87**	**96.27**	**3.29**	**89.26**
10	PGDUv	55.73	61.47	9.15	67.64	60.27	70.64	7.93	73.75
	kUv	54.30	60.38	8.75	65.69	57.90	67.26	6.91	70.72
	T_kMLUv	59.37	76.51	6.52	93.19	70.47	93.42	3.71	83.56
	T_kALCUv	58.70	75.17	6.61	94.07	71.70	91.22	6.56	85.79
	T_kMIAUv	59.00	75.47	6.64	94.64	72.40	92.16	4.79	88.28
	MADWUv	60.27	77.79	6.52	96.10	73.97	93.78	3.68	87.69
	MASWUv	**60.93**	**78.59**	**6.51**	**97.33**	**74.87**	**94.27**	**3.67**	**91.01**

improved under any experimental setup when λ_1 and λ_2 are well fine-tuned, it can result in a significant increase in time cost. Therefore, after a sensitivity test is simply conducted once, the values of λ_1 and λ_2 are fixed as two constant values in all experiments. And in all experiments, $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$ also achieve the best or competitive performance. Furthermore the values of λ_1 and λ_2 are also set to 0.5 in the ablation studies.

Visual comparison. We perform more visual comparisons. Fig. S.1 illustrates some examples of top-k untargeted attacks on VOC 2012, where our SPW-based method makes the top-k outputs of the adversarial image have higher confidence than the competitor methods, and the ground truth labels have lower prediction confidence than the other methods. Fig. S.2 visualizes some examples for kUv, T_kMLUv, and MASWUv on COCO 2014, revealing that MASWUv uses a smaller Pert to make a successful attack, affording higher confidence than the competitor methods. The ground truth labels have a lower prediction confidence acquired by MASWUv than kUv and T_kMLUv, and the top-3 outputs have a higher prediction confidence than the other two methods. For kUv, the perturbation is more visible than T_kMLUv and MASWUv.

Fig. S.3 illustrates a targeted attack example under the best-case scenario, where MASWT predicts the targeted labels as top-k outputs with higher confidence than the other methods. Moreover, MASWT affords the ground truth labels to have lower confidence than the other methods. When $k = 3, 5$, all methods successfully attack. However, for $k = 10$, Rank I and T_kMLT fail, but our MASWT successfully attacks.

Ablation study. We perform more ablation experiments on the proposed SPW strategy. Fig. S.4 shows the attack results of removing SPW from MASW and the original MASW. Removing SPW from MASW results in lower ASR and LFR, and higher Pert, which illustrates the effectiveness of the proposed method.

Sensitivity test. Three important hyperparameters, namely, λ_1,

TABLE S.VIII
COMPARISON OF TARGETED ATTACK METHODS ON TWO LARGE SETS.

Case	k	T	Method	NUS-WIDE ASR	NUS-WIDE SASR	NUS-WIDE Pert	NUS-WIDE LFR	Open Images ASR	Open Images SASR	Open Images Pert	Open Images LFR
Best	3	300	MLCWT	18.20	48.43	0.17	46.23	27.00	45.26	0.37	57.98
			MLDFT	14.50	35.39	0.45	44.43	23.70	38.92	0.41	50.52
			PGDT	38.47	67.96	0.34	28.28	32.63	50.97	0.35	57.24
			Rank I	92.60	94.23	**0.13**	30.59	90.10	92.35	0.16	25.37
			T_kMLT	95.27	96.71	0.14	75.27	95.07	96.72	**0.15**	82.14
			T_kALCT	95.50	96.93	0.21	78.73	95.60	97.14	0.20	85.63
			T_kMIAT	95.93	97.42	0.20	80.54	96.20	97.89	0.21	87.79
			MADWT	97.73	98.01	0.14	92.03	99.03	99.05	0.16	94.97
			MASWT	**98.13**	**98.85**	**0.13**	**93.61**	**99.93**	**99.75**	**0.15**	**95.67**
		–	MLCWT	15.10	58.18	0.19	43.34	20.30	40.66	0.34	61.82
	5	300	MLDFT	13.70	34.56	0.41	35.28	18.90	34.72	0.39	45.86
			PGDT	31.07	70.26	0.33	22.41	23.97	47.79	0.35	59.15
			Rank I	91.50	94.19	**0.14**	31.62	89.60	91.13	0.18	28.69
			T_kMLT	94.07	95.94	0.15	77.39	94.83	96.01	**0.16**	83.56
			T_kALCT	94.30	96.11	0.23	80.59	95.10	96.73	0.21	86.14
			T_kMIAT	94.57	96.67	0.22	83.56	95.67	97.14	0.22	89.79
			MADWT	95.73	97.96	0.15	94.15	99.03	98.89	0.17	94.78
			MASWT	**96.53**	**98.69**	0.15	**95.18**	**99.80**	**99.74**	**0.16**	**95.51**
		–	MLCWT	14.80	70.27	0.25	40.14	14.60	35.64	0.30	55.29
	10	300	MLDFT	10.40	29.73	0.43	33.37	15.70	30.23	0.40	44.19
			PGDT	25.47	73.20	0.35	16.02	17.47	46.03	0.36	49.22
			Rank I	77.40	81.48	0.21	37.85	87.20	89.65	0.20	28.81
			T_kMLT	93.63	94.56	0.17	79.25	93.53	94.49	**0.18**	85.69
			T_kALCT	93.70	94.94	0.24	85.43	94.80	95.30	0.23	88.70
			T_kMIAT	94.10	95.30	0.24	88.79	95.10	96.78	0.23	91.35
			MADWT	94.70	96.99	0.17	96.02	96.97	98.16	0.19	94.96
			MASWT	**95.50**	**98.13**	**0.16**	**96.77**	**99.37**	**99.75**	**0.18**	**95.21**
Random	3	300	MLCWT	16.50	40.92	0.21	65.98	24.60	40.19	0.39	61.17
			MLDFT	13.20	33.39	0.40	47.71	20.90	35.65	0.44	55.87
			PGDT	25.77	50.29	0.32	50.39	25.63	43.38	0.41	50.51
			Rank I	80.90	82.62	0.19	33.69	78.50	79.49	0.26	68.66
			T_kMLT	88.93	91.28	0.20	77.25	89.47	92.75	0.29	74.64
			T_kALCT	89.10	92.56	0.23	80.11	90.40	93.34	0.31	86.23
			T_kMIAT	89.50	93.10	0.24	83.26	90.13	93.05	0.30	88.37
			MADWT	90.97	93.48	0.21	92.18	90.67	93.34	0.27	93.47
			MASWT	**91.73**	**95.12**	**0.19**	**94.01**	**91.57**	**94.21**	**0.25**	**96.32**
		–	MLCWT	12.50	32.26	0.24	64.41	18.50	31.28	0.35	54.82
	5	300	MLDFT	12.60	30.38	0.42	53.32	16.20	25.78	0.45	53.26
			PGDT	19.97	40.24	0.35	42.81	17.73	35.66	0.40	55.54
			Rank I	48.30	54.74	0.21	47.71	38.70	43.24	0.28	70.23
			T_kMLT	80.17	86.52	0.24	78.98	79.30	83.58	0.31	78.65
			T_kALCT	80.30	87.11	0.27	84.45	79.57	84.71	0.33	88.41
			T_kMIAT	80.47	88.02	0.25	86.35	79.60	85.28	0.33	90.39
			MADWT	81.73	90.77	0.22	92.39	80.77	87.01	0.26	93.16
			MASWT	**82.33**	**91.27**	**0.19**	**92.99**	**81.17**	**87.79**	**0.25**	**94.02**
		–	MLCWT	10.40	22.57	0.25	58.13	16.50	27.84	0.36	46.31
	10	300	MLDFT	10.10	25.37	0.44	54.72	15.20	22.17	0.46	52.69
			PGDT	15.43	33.41	0.37	32.96	15.63	33.29	0.42	55.14
			Rank I	45.60	50.28	**0.23**	49.96	30.00	38.73	0.30	69.31
			T_kMLT	67.63	71.52	0.26	81.09	41.47	61.44	0.31	71.69
			T_kALCT	68.10	72.67	0.30	85.10	43.70	64.51	0.35	82.53
			T_kMIAT	68.37	73.41	0.29	87.14	44.50	65.36	0.34	83.41
			MADWT	69.27	74.57	0.24	90.51	47.93	68.19	0.29	84.69
			MASWT	**70.63**	**76.16**	**0.23**	**91.01**	**48.77**	**69.99**	**0.28**	**86.42**
Worst	3	300	MLCWT	10.80	27.91	0.23	72.42	14.20	26.61	0.34	66.67
			MLDFT	9.70	23.46	0.39	65.47	12.30	23.79	0.46	55.75
			PGDT	19.20	31.28	0.29	55.14	17.27	36.19	0.39	54.28
			Rank I	23.10	24.88	0.22	49.33	75.20	78.62	**0.26**	58.87
			T_kMLT	50.33	62.48	0.23	71.43	87.17	87.94	0.27	74.12
			T_kALCT	50.40	63.18	0.25	73.56	86.20	86.04	0.33	80.31
			T_kMIAT	50.67	63.29	0.24	79.57	86.47	86.73	0.32	84.48
			MADWT	52.03	64.97	0.16	86.21	88.43	88.68	0.27	95.19
			MASWT	**52.50**	**65.70**	**0.14**	**88.23**	**88.70**	**88.93**	**0.26**	**95.74**
		–	MLCWT	7.80	19.71	0.25	71.82	10.40	23.19	0.35	62.38
	5	300	MLDFT	6.90	14.28	0.38	64.19	9.10	17.26	0.45	55.46
			PGDT	10.73	15.85	0.31	45.94	12.33	30.75	0.37	43.39
			Rank I	14.90	18.61	0.21	48.91	34.70	36.02	0.28	60.71
			T_kMLT	30.33	52.31	0.25	81.19	74.20	85.64	0.28	76.53
			T_kALCT	31.20	53.42	0.28	82.34	73.50	84.21	0.34	79.72
			T_kMIAT	32.70	54.84	0.27	84.56	73.77	85.38	0.33	83.59
			MADWT	33.70	55.41	0.16	90.16	74.97	86.93	0.26	92.96
			MASWT	**34.27**	**56.55**	**0.14**	**91.04**	**75.63**	**87.49**	**0.25**	**94.29**
		–	MLCWT	5.40	16.84	0.28	70.02	7.40	24.91	0.36	64.51
	10	300	MLDFT	4.50	14.72	0.43	67.85	6.70	18.84	0.49	56.74
			PGDT	8.43	11.28	0.33	48.59	9.97	26.71	0.38	46.78
			Rank I	12.10	19.86	**0.24**	46.64	57.30	66.80	0.31	71.56
			T_kMLT	14.33	46.06	0.29	83.42	84.00	92.54	0.32	79.24
			T_kALCT	14.50	46.46	0.29	84.72	84.40	93.46	0.36	79.02
			T_kMIAT	14.70	47.21	0.28	85.52	85.13	94.25	0.35	80.48
			MADWT	16.03	48.98	0.27	90.79	86.67	95.98	0.31	85.16
			MASWT	**16.50**	**49.72**	0.27	**91.23**	**87.50**	**96.89**	**0.29**	**86.43**

TABLE S.IX
COMPARISON OF AC OF TARGETED ATTACKS WITH $k = 3, 5, 10$. THE BEST RESULTS ARE IN BOLD.

	Method	NUS-WIDE $k=3$ SAC	NUS-WIDE $k=5$ SAC	NUS-WIDE $k=10$ SAC	Open Image $k=3$ SAC	Open Image $k=5$ SAC	Open Image $k=10$ SAC
Best Case	PGDT	94.82	95.09	95.53	97.54	97.52	97.31
	Rank I	94.96	94.64	94.72	98.34	98.01	97.54
	T_kMLT	95.01	94.79	96.86	98.56	98.60	98.42
	MASWT	**97.79**	**98.23**	**99.69**	**99.91**	**99.96**	**99.95**
Random Case	PGDT	93.13	94.38	91.07	97.31	97.05	96.46
	Rank I	92.91	94.31	93.02	97.54	97.51	97.63
	T_kMLT	93.18	94.51	92.37	98.43	98.39	98.23
	MASWT	**95.83**	**96.77**	**95.11**	**99.69**	**99.65**	**99.28**
Worst Case	PGDT	92.87	92.21	88.48	96.54	96.07	95.19
	Rank I	93.33	92.72	89.39	96.40	95.77	93.80
	T_kMLT	93.64	93.86	92.25	96.59	95.76	93.35
	MASWT	**95.89**	**94.97**	**94.87**	**99.47**	**99.13**	**98.08**

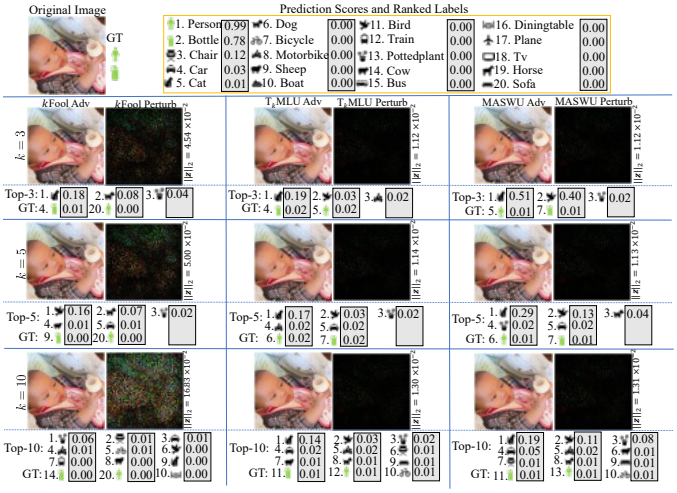


Fig. S.1. Illustrative images in VOC 2012 for untargeted attacks with $k = 3, 5, 10$. Adv and Perturb mean adversarial image and perturbation, respectively.

λ_2 and η, are used in MADW. Fig. S.5(a) presents the results of top-10 targeted attack under the worst case on VOC 2012 with varied λ_1 (λ_2 remains 0.5). As λ_1 increases, ASR also increases, and LFR decreases. λ_1 and λ_2 should be tuned according to the application scenario. Moreover, Fig. S.5(b) presents the effect of η on ASR and LFR, achieving the best value for $\eta = 1.015$. Based on this sensitivity test, we conclude that $\lambda_1, \lambda_2 = 0.5$, and $\eta = 1.015$ attain the best possible performance. Besides, Fig. S.5(c) shows the impact of the damping factor update method on attack performance. When $\mathfrak{f}_{in} = 1.5$ and $\mathfrak{f}_{de} = 5$, ASR and LFR achieve larger values than those when $\mathfrak{f}_{in} = 2$ and $\mathfrak{f}_{de} = 3$. Therefore we set $\mathfrak{f}_{in} = 1.5$ and $\mathfrak{f}_{de} = 5$ in our experiments.

Convergence Comparison. Fig. S.6 presents more results on the fluctuating Lagrangian values of MLCWU, T_kMLU, and MASWU on VOC 2012 and COCO 2014. The flat curves of MASWU rather than the undulatory curves of MLCWU and T_kMLU infer that MASWU is more stable than MLCWU and T_kMLU. We statistically analyzed 200 random images in VOC 2012/COCO 2014 for a top-10 untargeted attack and a targeted attack. The results on VOC 2012/COCO 2014 reveal that the proportions of fluctuations exceeding 0.3 of MLCWU, T_kALCU, T_kMIAU, T_kMLU, MADWU

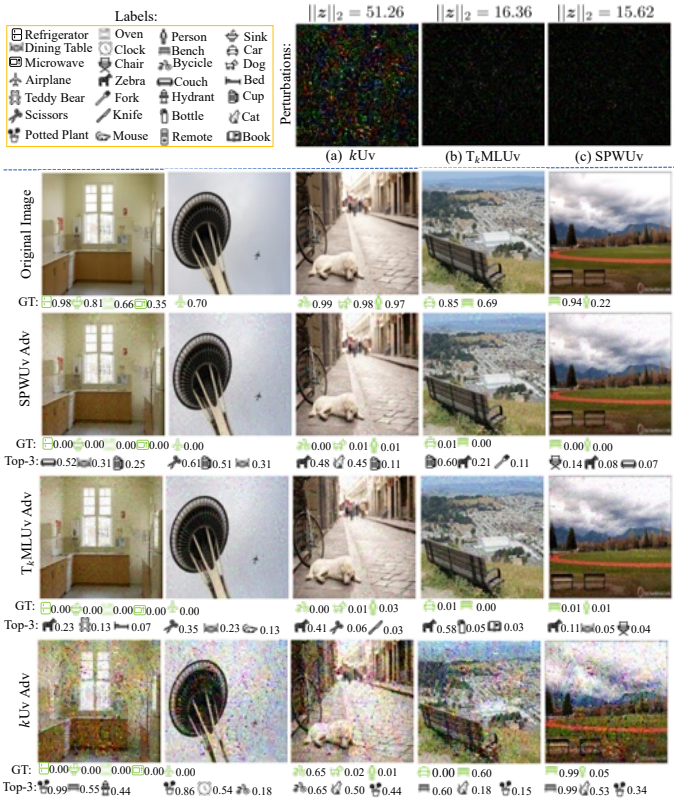


Fig. S.2. Illustrative images in COCO 2014 for top-3 universal untargeted attack. Adv means adversarial image.

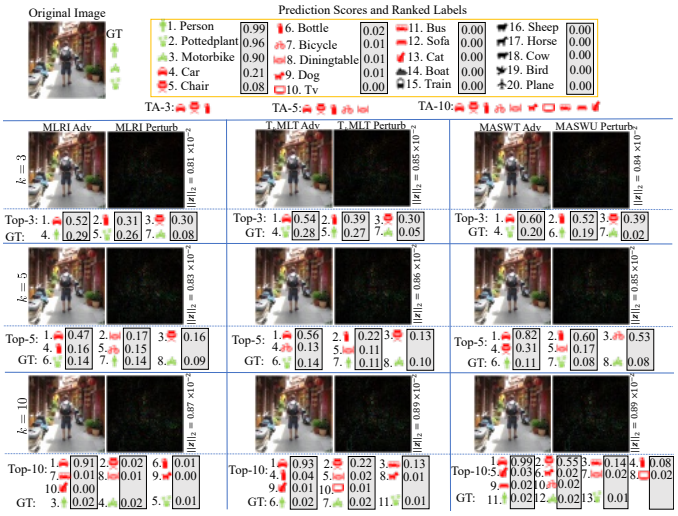


Fig. S.3. Illustrative images in VOC 2012 for targeted attacks under the best case scenario with $k = 3, 5, 10$.

and MASWU are 17.0%/21.5%, 9.0%/8.5%, 11.5%/10.0%, 10.5%/9.5% , 7.0%/5.0% and 6.5%/5.0%, respectively. And for targeted attack, the proportions on VOC 2012/COCO 2014 of MLCWT, T$_k$ALCT, T$_k$MIAT, T$_k$MLT, MADWT and MASWT are 12.5%/18.0%, 8.5%/9.0%, 7.0%/8.5%, 9.0%/11.5%, 6.0%/7.0% and 6.0%/7.5%, respectively. From the statistical analysis, our methods are more stable than these existing methods.

Time Cost Comparison: Tables S.X and S.XI report the runtime of the targeted and untargeted attack methods. Except

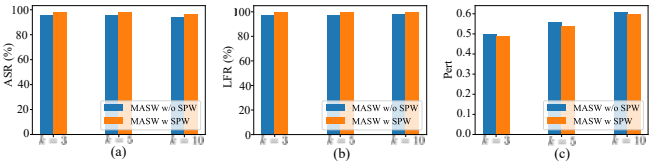


Fig. S.4. Ablation study on SPW on VOC 2012.

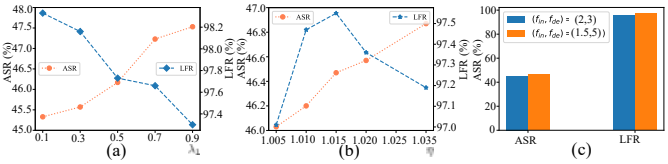


Fig. S.5. (a), (b) and (c) effect of λ_1, η and $(\mathfrak{f}_{in}, \mathfrak{f}_{de})$ on ASR and LFR, respectively.

for FGSM and MFGSM, our methods generally require the lowest time cost. Moreover, FGSM requires only one gradient ascent, so its time cost is sharply reduced. However, the time costs of our methods are comparable or even significantly lower than the SOTA methods, namely T$_k$ML, kFool, Rank I, MLDF, and MLCW, T$_k$ALC and T$_k$MIA. Although the generation of adversarial attacks by LPD and GAMA is rapid after the completion of the generative model training, the training of the generative model still requires a significant amount of time cost (i.e., exceeding three hours).

Table S.XII shows the time cost of each universal untargeted attack method. Specifically, kUv requires more than a week (168.00 hours) for an execution without eventually producing a result. Our method, MASWUv achieves the lowest time cost. What's more, MADWUv has comparable performance.

TABLE S.XI
RUN TIME (S) OF EACH UNTARGETED ATTACK METHOD ON VOC 2012 AND COCO 2014.

Dataset	VOC 2012			COCO 2014		
Method	$k=3$	$k=5$	$k=10$	$k=3$	$k=5$	$k=10$
FGSMU	0.06	0.07	0.07	0.04	0.05	0.05
MFGSMU	1.63	1.64	1.69	1.01	1.03	1.06
MLCWU	19.34	19.57	20.01	11.97	12.14	12.67
MLDFU	23.92	24.25	25.12	29.16	29.41	30.03
PGDU	**2.83**	**2.97**	**3.23**	1.22	1.34	1.57
kFool	4.18	5.04	13.12	13.91	15.57	22.99
T$_k$MLU	3.67	4.36	5.69	0.43	0.51	0.63
T$_k$ALCT	4.21	5.07	6.01	0.67	0.73	0.98
T$_k$MIAT	4.06	4.78	5.93	0.61	0.69	0.81
MADWU	4.31	5.34	9.57	1.01	1.14	1.48
MASWU	3.04	3.79	5.21	**0.41**	**0.46**	**0.61**

TABLE S.X
RUN TIME (S) OF EACH TARGETED ATTACK METHOD UNDER BEST CASE ON VOC 2012 AND COCO 2014.

Dataset	VOC 2012			COCO 2014		
Method	$k=3$	$k=5$	$k=10$	$k=3$	$k=5$	$k=10$
FGSMT	0.07	0.07	0.08	0.04	0.04	0.04
MFGSMT	1.60	1.66	1.66	0.96	0.99	1.02
MLCWT	5.46	13.61	19.88	2.38	3.16	5.77
MLDFT	67.66	95.06	118.31	116.51	147.92	174.50
PGDT	14.17	20.31	24.44	7.14	8.56	10.74
Rank I	3.83	5.87	9.02	0.56	1.06	2.29
T$_k$MLT	3.67	5.51	8.22	0.53	0.76	1.23
T$_k$ALCT	4.56	6.04	9.35	0.89	0.93	1.46
T$_k$MIAT	4.23	5.56	8.59	0.78	0.84	1.41
MADWT	5.21	5.56	10.26	1.16	1.42	2.01
MASWT	**3.11**	**4.64**	**7.07**	**0.54**	**0.75**	**1.21**

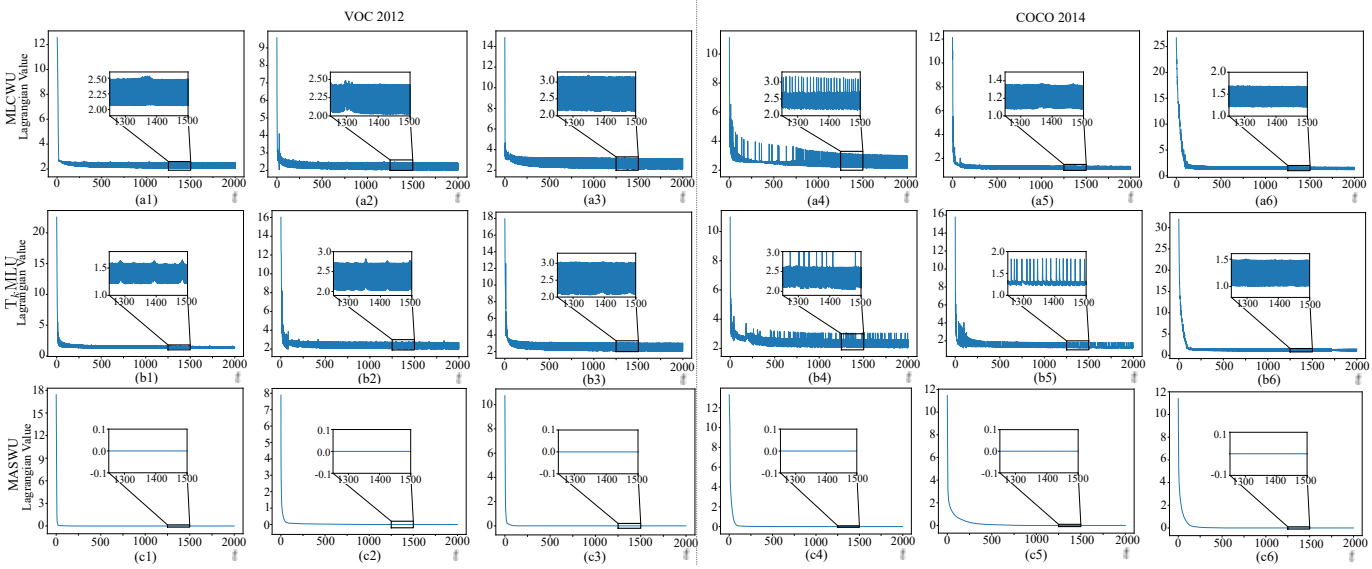


Fig. S.6. Variations of the Lagrangian values of MLCWU, T_kMLU and our MASWU on VOC 2012 and COCO 2014 during optimizing. The top-10 untargeted attack is considered.

TABLE S.XII
RUN TIME (H) OF EACH UNIVERSAL UNTARGETED ATTACK METHOD ON VOC 2012 AND COCO 2014.

Dataset	VOC 2012			COCO 2014		
Method	$k=3$	$k=5$	$k=10$	$k=3$	$k=5$	$k=10$
PGDUv	1.03	1.98	3.79	0.83	1.23	1.75
kUv	168.00	168.00	168.00	11.49	90.44	168.00
T_kMLUv	0.68	1.08	3.12	0.11	0.13	0.17
T_kALCUv	0.70	1.15	3.50	0.16	0.23	0.26
T_kMIAUv	0.71	1.16	3.52	0.15	0.18	0.23
MADWUv	1.01	1.64	3.97	0.21	0.22	0.22
MASWUv	**0.59**	**0.99**	**2.97**	**0.10**	**0.12**	**0.14**

REFERENCES

[1] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.

[2] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *CVPR*, 2018, pp. 9185–9193.

[3] Q. Song, H. Jin, X. Huang, and X. Hu, "Multi-label adversarial perturbations," in *IEEE ICDM*. IEEE, 2018, pp. 1242–1247.

[4] S. Hu, L. Ke, X. Wang, and S. Lyu, "Tkml-ap: Adversarial attacks to top-k multi-label learning," in *ICCV*, 2021, pp. 7649–7657.

[5] N. Tursynbek, A. Petiushko, and I. Oseledets, "Geometry-inspired top-k adversarial perturbations," in *WACV*, 2022, pp. 3398–3407.

[6] N. Zhou, W. Luo, X. Lin, P. Xu, and Z. Zhang, "Generating multi-label adversarial examples by linear programming," in *IJCNN*, IEEE, 2020, pp. 1–8.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[8] P. Jaccard, "The distribution of the flora in the alpine zone. 1," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[9] L. Prechelt, "Early stopping-but when?" in *Neural Networks: Tricks of the Trade*, Springer, 1998, pp. 55–69.

[10] N. Zhou, W. Luo, J. Zhang, L. Kong, and H. Zhang, "Hiding all labels for multi-label images: An empirical study of adversarial examples," in *IJCNN*. IEEE, 2021, pp. 1–8.

[11] M. K. Transtrum and J. P. Sethna, "Improvements to the levenberg-marquardt algorithm for nonlinear least-squares minimization," *arXiv preprint arXiv:1201.5885*, 2012.

[12] G. Debreu, "Valuation equilibrium and pareto optimum," *Proceedings of the national academy of sciences*, vol. 40, no. 7, pp. 588–592, 1954.

[13] Y.-H. Dai and L. Zhang, "Optimization with least constraint violation," *CSIAM Trans. on Appl. Math.*, vol. 2, pp. 551—-584, 2021.

[14] M. Everingham, S. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *IJCV*, vol. 111, no. 1, pp. 98–136, 2015.

[15] A. Aich, S. Li, C. Song, M. S. Asif, S. V. Krishnamurthy, and A. K. Roy-Chowdhury, "Leveraging local patch differences in multi-object scenes for generative adversarial attacks," in *WACV*, 2023, pp. 1308–1318.

[16] A. Aich, C.-K. Ta, A. Gupta, C. Song, S. Krishnamurthy, S. Asif, and A. Roy-Chowdhury, "Gama: Generative adversarial multi-object scene attacks," in *NeurIPS*, vol. 35, pp. 36914–36930, 2022.

[17] Y. Sun, Q. Xu, Z. Wang, and Q. Huang, "When measures are unreliable: Imperceptible adversarial perturbations toward top-k multi-label learning," in *ACM MM*, 2023, pp. 1515–1526.

[18] M. Ma, W. Zheng, W. Lv, L. Ren, H. Su, and Z. Yin, "Multi-label adversarial attack based on label correlation," in *ICIP*. IEEE, 2023, pp. 2050–2054.

[19] J. K J, R. Vamshi Teja, S. Krishnakant, and N. B. Vineeth, "Submodular batch selection for training deep neural networks," in *IJCAI*, 2019.

[20] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in *ICCV*, 2021, pp. 82–91.