Mutual-Modality Adversarial Attack with Semantic Perturbation

Jingwen Ye, Ruonan Yu, Songhua Liu, Xinchao Wang †

National University of Singapore
jingweny@nus.edu.sg, {ruonan,songhua.liu}@u.nus.edu, xinchao@nus.edu.sg

Abstract

Adversarial attacks constitute a notable threat to machine learning systems, given their potential to induce erroneous predictions and classifications. However, within real-world contexts, the essential specifics of the deployed model are frequently treated as a black box, consequently mitigating the vulnerability to such attacks. Thus, enhancing the transferability of the adversarial samples has become a crucial area of research, which heavily relies on selecting appropriate surrogate models. To address this challenge, we propose a novel approach that generates adversarial attacks in a mutual-modality optimization scheme. Our approach is accomplished by leveraging the pre-trained CLIP model. Firstly, we conduct a visual attack on the clean image that causes semantic perturbations on the aligned embedding space with the other textual modality. Then, we apply the corresponding defense on the textual modality by updating the prompts, which forces the re-matching on the perturbed embedding space. Finally, to enhance the attack transferability, we utilize the iterative training strategy on the visual attack and the textual defense, where the two processes optimize from each other. We evaluate our approach on several benchmark datasets and demonstrate that our mutual-modal attack strategy can effectively produce high-transferable attacks, which are stable regardless of the target networks. Our approach outperforms state-of-the-art attack methods and can be readily deployed as a plug-and-play solution.

Introductions

With the milestone performances of Deep Neural Networks (DNNs) in numerous computer vision tasks, the efficiency (Ma, Fang, and Wang 2023a; Fang et al. 2023; Liu et al. 2022; Yang et al. 2022) and reliability (Ye, Liu, and Wang 2023; Ye et al. 2022a,b) of these techniques become equally important when deployed in the real world. However, recent researches (Goodfellow, Shlens, and Szegedy 2014) have found that such DNNs are vulnerable to adversarial examples. This is, through only a small norm of perturbation applied on the original input, the maliciously crafted adversarial samples could cause misclassification or unexpected behavior to machine learning models.

† Corresponding author.

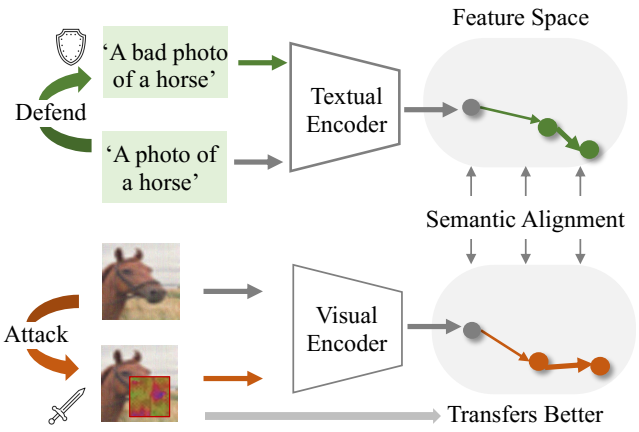


Figure 1: Joint attack and defense framework in the visual and textual modalities. The visual features are attacked to push away from the original one, while the textual features defend to pull back this similarity gap.

As a crucial assessment of the strength and security of DNNs, various attack algorithms (Hayes and Danezis 2017; Liu et al. 2019) have been proposed, achieving relatively high fooling rates. However, the effectiveness of these attacks is largely affected by different conditions, with the black-box setting being the most challenging yet realistic scenario. In the black-box setting, attackers cannot access the model's parameters and structure, leading to the need for improving the transferability of attacks to arbitrary target networks (Cheng et al. 2019; Dong et al. 2018). The corresponding methods include ensemble-model attacks (Liu et al. 2016), momentum-based attacks (Dong et al. 2018), input transformation-based attacks (Xie et al. 2019), and model-specific attacks (Wu et al. 2019). Such methods aim to enhance the transferability of attacks by either exploiting the inherent weaknesses of the target model or exploring the common vulnerabilities of a group of models.

The majority of current techniques striving to amplify the transferability of adversarial attacks predominantly hinge on the selection of surrogate models. However, these surrogate models often prove to be unstable and profoundly influenced by the architectural similarities between the surrogate model

itself and the target networks. Hence, the careful selection of an optimal surrogate model characterized by a robust feature extractor and exceptional generalizability emerges as a critical factor. Previous studies have used multiple ImageNet-pretrained networks as surrogate models, given the large number of images and object categories in the dataset. In this paper, we leverage the recent progress of the CLIP model in computer vision and natural language processing. Having been trained on over 400 million image pairs, CLIP can now serve as our surrogate model, enabling the generation of powerful and broadly effective perturbations.

In addition to its large-scale training data, we utilize the CLIP model as the surrogate model due to its ability to align visual and textual modalities in an aligned feature space. This is accomplished through a visual encoder and textual encoder pairing, allowing us to generate adversarial samples with semantic perturbations. Semantic perturbations differ from previous methods, which simply maximize feature differences. Instead, our approach maximizes semantic differences to ensure that the features after the attack retain explicit semantic information and do not fall into areas without clear semantic meaning, ensuring the effectiveness of the generated attacks.

In this paper, we propose integrating attack and defense into one framework, building upon the semantic perturbations obtained from the pre-trained CLIP model's aligned visual and textual embedding space. As shown in Fig. 1, we apply visual perturbations to clean images, increasing the semantic difference in the feature space and causing a contradiction with the textual embedding when given the input "A bad photo of a horse." We then defend against the attack by updating the text prompt template, eliminating this semantic gap and restoring entailment. This iterative attack and defense optimization strategy enhances the attack's transferability to target black-box networks.

To summarize, we make the following contributions:

- Firstly, we propose a method to generate reliable adversarial attacks by using the semantic consistency of pre-trained CLIP model to learn perturbations in the semantic feature embedding space. We ensure fidelity by constraining the perturbations with semantic consistency from the text input;
- Secondly, we propose an iterative optimization strategy to improve the transferability of the generated attack across different architectures and datasets, where we attack the visual input and defend in the textual one;
- Thirdly, we conduct extensive experiments to evaluate the transferability of our proposed approach in cross-dataset, cross-architecture, and cross-task settings. Our results demonstrate that our approach is efficient, effective, and can be used as a plug-and-play solution.

Related Work

Adversarial Attack

Adversarial attacks (Madry et al. 2017; Dong et al. 2017; Guo et al. 2019; Akhtar and Mian 2018; Zhang et al. 2021) are designed to deceive machine learning models by adding small, imperceptible perturbations to input data, causing the model to generate incorrect outputs or misclassify inputs. One of the traditional attack methods (Goodfellow, Shlens, and Szegedy 2014) is to use gradient information to update the adversarial example in a single step along the direction of maximum classification loss. Building on this work, GAMA (Yuan et al. 2021) is proposed as a plug-and-play method that can be integrated with any existing gradient-based attack method to improve cross-model transferability. Besides, many works (Xie et al. 2019; Wang and He 2021) have been proposed to improve the attack transferability.

To ensure the efficiency of generating attacks, generative model-based attack methods (Hayes and Danezis 2017; Poursaeed et al. 2017; Liu et al. 2019; Xiang et al. 2022; Qiu et al. 2019; Salzmann et al. 2021; Aich et al. 2022) have been extensively studied. For example, Aishan et al. (Hayes and Danezis 2017) train a generative network capable of generating universal perturbations to fool a target classifier. To generate patch-based attacks, PS-GAN (Liu et al. 2019) is utilized to simultaneously enhance the visual fidelity and attacking ability of the adversarial patch.

Other than designing attacks in the image domain, attacking NLP models (Morris et al. 2020; Boucher et al. 2022; Chen et al. 2021; Zhang et al. 2020; Perez and Ribeiro 2022) has become a popular research direction. Specifically, prompt learning attacks have attracted many researchers as a lighter method to tune large-scale language models, which can be easily attacked by illegally constructed prompts. Shi et al. (Shi et al. 2022) propose a malicious prompt template construction method to probe the security performance of PLMs. Du et al. (Du et al. 2022) propose obtaining poisoned prompts for PLMs and corresponding downstream tasks by prompt tuning.

Different from the above attack methods, we propose the plug and play dynamic updating method, which boosts the transferability.

Vision-and-Language Models

A vision-and-language model (Jia et al. 2021; Radford et al. 2021; Mu et al. 2022; Yao et al. 2021; Fang, Ma, and Wang 2023; Ma, Fang, and Wang 2023b) is a powerful learning model processing both images and text in a joint manner to align the vision and texture embeddings. As the most popular VL model, CLIP (Radford et al. 2021) learn SOTA image representations from scratch on a dataset of 400 million image-text pairs collected from the internet, which enables various tasks like zero-shot classification. To further boost the VL models' performance, Zhou et al. (Zhou et al. 2022b) propose CoOp to models the prompt's context words with learnable vectors for adapting CLIP-like vision-language models for downstream image recognition. And CoCoOp (Zhou et al. 2022a) extends CoOp by further learning a lightweight neural network to generate for each image an input-conditional token.

Target at such large-scale pre-trained VL models, there are a bulk of adversarial attack approaches (Noever and Noever 2021; Zhang, Yi, and Sang 2022) try to fool it. Noever et al. (Noever and Noever 2021) demonstrate adversarial attacks of spanning basic typographical, conceptual, and

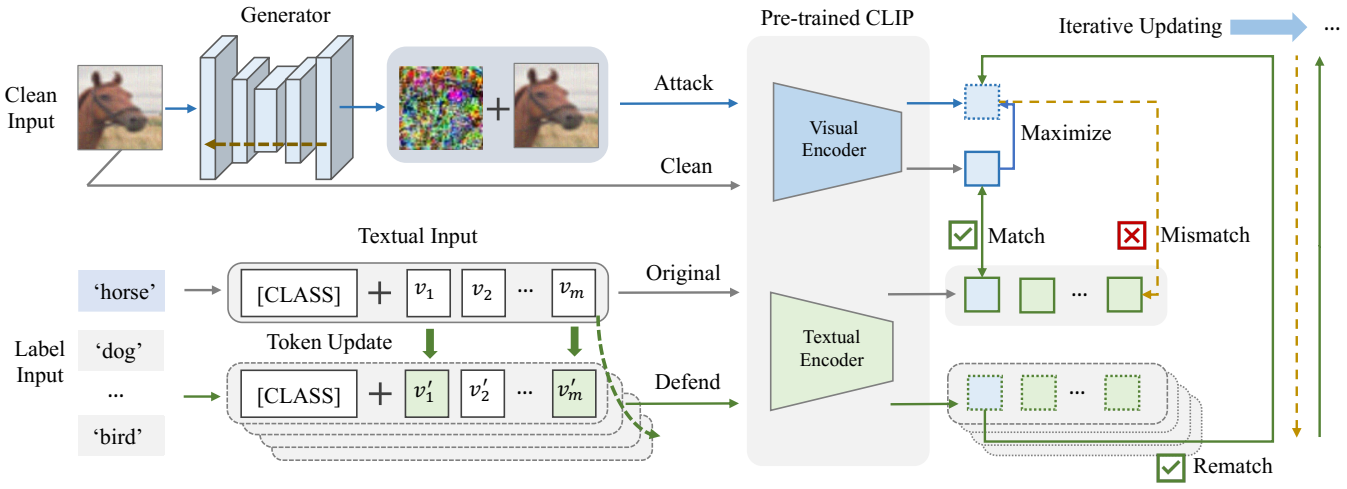


Figure 2: The framework of the joint Mutual-modality attack-defense method. We use the pre-trained CLIP as the surrogate model. The generator is optimized by maximizing the difference with the clean image as input. And the textual input is updated to re-match the features from the textual encoder and the visual encoder.

iconographic inputs generated to fool the model into making false or absurd classifications. Zhang et al. (Zhang, Yi, and Sang 2022) propose a multimodal attack method that collectively carries out the attacks on the image modality and the text modality. Hintersdorf et al. (Hintersdorf, Struppek, and Kersting 2022) introduce to assess privacy for multi-modal models and reveal whether an individual was included in the training data by querying the model with images of the same person. Jia et al. (Jia, Liu, and Gong 2022) injects backdoors into a pre-trained image encoder.

In this work, we utilize the VL models to establish the attack frameworks. Previous works, however, either focus on one modality or treat the two modalities separately, which constrains transferring to target networks or other datasets.

Proposed Method

Framework Overview

In the proposed framework, we utilize the generator-orientated adversarial attack method, which integrates CLIP to enable the transferability. That is, from a training distribution of images, we train the generator $\mathcal{G}$ to generate universal perturbations applied on the input clean image x_i. The corresponding adversarial sample x_i' can be thus obtained as:

$$x_i' = \min\big(x_i + \epsilon, \max(\mathcal{G}(x_i), x_i - \epsilon)\big), \tag{1}$$

where ϵ is the predetermined maximum perturbation on the input. Each applied perturbations are bounded with $[-\epsilon, +\epsilon]$, in the rest of the paper, we simplify this process and use $\mathcal{G}(x_i)$ to denote the bounded adversarial sample.

Denote the groundtruth label of the clean image x_i as y_{true}, then the attack goal is to obtain the universal perturbations enabling cross-architecture and cross-dataset transferring. The goal of attacking a group of black-box models $\mathcal{M} = \{\mathcal{M}^0, \mathcal{M}^1, ..., \mathcal{M}^{P-1}\}$ pre-trained on various datasets $\mathcal{D} = \{\mathcal{D}^0, \mathcal{D}^1, ..., \mathcal{D}^{P-1}\}$ is to maximize:

$$\sum \mathbf{1}_{y' \neq y_{\text{true}}} | y' \leftarrow \mathcal{M}^p\big[\mathcal{G}(x_i)\big], \quad x_i \in \mathcal{D}^p, \mathcal{M}^p \in \mathcal{M} \tag{2}$$

where $\mathbf{1}$ is the indicator function.

To achieve it, we utilize CLIP as the surrogate model to train the powerful generator $\mathcal{G}$, as is shown in Fig. 2. Considering that CLIP is a powerful model capable of learning the relationship between different modalities, we utilize CLIP as the surrogate model to generate the perturbations. Denote the textual encoder to be E_i and the visual encoder to be E_t, then CLIP is capable of zero-shot predictions by matching the most similar textual embedding as:

$$\begin{aligned} p(y|x_i) &= Clip\big(x_i, X_t\big) \\ &= \frac{\exp[\text{sim}\big(E_i(x_i), E_t(x_t^y)\big)/\tau]}{\sum_{c=0}^{C-1} \exp[\text{sim}\big(E_i(x_i), E_t(x_t^c)\big)/\tau]}, \end{aligned} \tag{3}$$

where $\text{sim}(\cdot)$ is the cosine similarity, τ is the temperature parameter. And $X_t = \{x_t^0, x_t^1, ..., x_t^{C-1}\}$ is text modality input that generates for each classification label c ($c \in \{0, 1, ..., C-1\}$).

As the CLIP takes both two modalities as input to make predictions as $y' = \arg\max_y p(y|x_i)$, we construct the attack and defense into one framework on these two modalities. Specifically, producing attack with CLIP can be treated as a kind of adversarial training process:

$$\begin{aligned} &\max_{\mathcal{P}} \min_{\mathcal{G}} V(\mathcal{P}, \mathcal{G}) \\ &= \mathbb{E}_{x_i \sim p_{X_i}} \text{sim}\big[Clip\big(\mathcal{G}(x_i), \mathcal{P}(X_t)\big), Clip\big(x_i, \mathcal{P}(X_t)\big)\big] \\ &+ \mathbb{E}_{x_t \sim p_{X_t}} \text{sim}\big[Clip\big(\mathcal{G}(X_i), x_t\big), Clip\big(\mathcal{G}(X_i), \mathcal{P}(x_t)\big)\big], \end{aligned} \tag{4}$$

where $\mathcal{P}(\cdot)$ is the prompt tuning function for the textual input, the details of which are in Sec. Similar as the GAN training, the optimization of Eq. 4 is to iteratively update $\mathcal{G}$ and $\mathcal{P}$. The whole process is,

- for optimizing $\mathcal{G}$ to **generate the perturbations** on the clean image x_i, we minimize the output similarity be-

tween the clean input x_i and the adversarial input x_i', which is in *'Visual Attack with Semantic Perturbation'*;

- for optimizing $\mathcal{P}$ to **defend the attack** on the image modality input, we tune the prompt as $\mathcal{P}(x_t)$ to match the image-text embedding again, which is in *'Textual Defense with Prompt Updating'*;
- with the **iterative training** of $\mathcal{G}$ and $\mathcal{P}$, we obtain the final generative perturbation network $\mathcal{G}$, the whole algorithm is given in the supplementary.

Visual Attack with Semantic Perturbation

Recall that the visual attack x_i' generated from $\mathcal{G}$ is bounded with ϵ (in Eq. 1), it is assumed to be assigned with the wrong prediction by the target network $\mathcal{M}$. Considering the fact that $\mathcal{M}$ keeps as the black-box, we turn to attack on the feature embedding space of the CLIP model.

The CLIP model's pre-trained image encoder E_i is a powerful feature extractor that has high transferability. To ensure the transferability of adversarial samples, we aim to maximize the distance between the feature representation of the adversarial input x_i', denoted as $E_i(x_i')$, and the feature representation of the clean input x_i, denoted as $E_i(x_i)$. This is achieved by minimizing the loss function ℓ_{feat}, which is calculated as:

$$\begin{aligned} \ell_{feat} &= -\|\mathcal{F}_i - \mathcal{F}_i'\|^2 \\ \mathcal{F}_i = \frac{E_i(x_i)}{\|E_i(x_i)\|}, &\mathcal{F}_i' = \frac{E_i(x_i')}{\|E_i(x_i')\|}, \end{aligned} \tag{5}$$

where ℓ_{feat} is calculated based on the MSE loss.

Other than maximizing the features similarity with the perturbed and the clean input, an extra triplet loss is applied to ensure that the perturbed features $E_i(x_i')$ fool the downstream networks with a wrong prediction. To calculate the triplet loss, the original textual embedding should be precomputed as $\mathcal{F}_t^c = E_i(x_t^c)/\|E_i(x_t^c)\|$ ($c \in \{0, 1, ..., C-1\}$). Each x_t^c is composed of a prompt template and a label object, i.e. 'dog' + 'A clean photo of {}'. Thus, for each label $c \in \{0, 1, ..., C-1\}$, we pre-compute the corresponding textual embeddings as $\mathcal{F}_t = \{\mathcal{F}_t^0, \mathcal{F}_t^1, ..., \mathcal{F}_t^{C-1}\}$. In this way, for each clean image x_i with the groudtruth label y_{true}, we use the the triplet loss (Aich et al. 2022) ℓ_{tri} to mislead the match of the features from the two modalities. Specifically, ℓ_{tri} is calculated as:

$$\begin{aligned} \ell_{tri} = \|\mathcal{F}_i' - \mathcal{F}_t^{y'}\|^2 + \max\big(0, \alpha - \|\mathcal{F}_i' - \mathcal{F}_t^{y_{\text{true}}}\|^2\big), \\ where \quad y' = \arg\min_c \text{sim}(\mathcal{F}_i, \mathcal{F}_t^c), \end{aligned} \tag{6}$$

where α is the margin of the triplet loss. $\mathcal{F}_t^{y'}$ is the textual embedding that is least similar to that of the clean input. This triplet loss forces the perturbed features away from its groundtruth textual embedding $\mathcal{F}_t^{y_{\text{true}}}$ while minimizing the distance with the textual embedding $\mathcal{F}_t^{y'}$ that is originally least related to the clean image features $\mathcal{F}_i$.

Finally, following the previous adversarial attack methods, we utilize an extra classification loss:

$$\ell_{cls} = \frac{1}{\sigma + H_{CE}\big(Clip(x_i', X_t), y\big)}, \tag{7}$$

where we set $\sigma = 0.1$ to prevent gradient explosion. $H_{CE}(\cdot)$ is the standard cross-entropy loss, and $Clip(\cdot)$ is the output probabilities after softmax.

Thus, the final learning objective for visual attack is:

$$\arg\min_{\mathcal{G}} \ell_{feat} + \ell_{tri} + \ell_{cls}, \tag{8}$$

with which, the optimized $\mathcal{G}$ is capable of generating perturbations that could fool the textual input in the form of X_t and have a certain degree of transferablity.

Textual Defense with Prompt Updating

For a clean image x_i, its feature embedding can be denoted as $\mathcal{F}_i$. For each label $c \in \{0, 1, ..., C-1\}$, the text input for each label can be organized by $m+1$ text tokens as: $x_t^c = [< \text{CLASS}(c) >, v_1, v_2, ..., v_m]$. Then the textual embedding for each label c is calculated as $\mathcal{F}_t^c \leftarrow E_t(x_t^c)$. And the CLIP model tends to output the probabilities for each label as:

$$\begin{aligned} p(y|x_i, X_t) = Clip(x_i, X_t) \leftarrow softmax(\mathcal{F}_i * \mathcal{F}_t) \\ where \quad X_t = X_l + X_p. \end{aligned} \tag{9}$$

Here, we separate the text input X_t into the fixed label token $X_l =< \text{CLASS}(c) >$ and the dynamic prompt token $X_p = \{v_1, v_2, ..., v_m\}$. Previous work on prompt learning (Zhou et al. 2022b) has indicated that the position of the label token wouldn't make big difference on the final results, we intuitively put the label token at the beginning.

Supposing that based on the current text input X_t the CLIP model makes the right prediction on the clean input x_i with the groundtruth label y_{true}, and the attack generator has successfully attacked it by predicting it as a wrong label y'. The attack (A) and the defense (D) process could be formulated as:

$$\begin{aligned} &[A]: \ \arg\max_y p\big(y|\mathcal{G}(x_i), X_t\big) \neq \arg\max_y p\big(y|x_i, X_t\big). \\ &[D]: \ \arg\max_y p\big(y|\mathcal{G}(x_i), \mathcal{P}(X_t)\big) = \arg\max_y p\big(y|x_i, X_t\big), \end{aligned} \tag{10}$$

where in [A], we learn the generator $\mathcal{G}$ for generating adversarial perturbations and in [D], we update the text input X_t to X_t' with the prompt tuning function $\mathcal{P}$ to guide the right prediction on CLIP again.

During the prompt tuning, we fix the label tokens X_l, and only update the prompt template X_p' by maximizing the semantic similarity comparing with the former visual embeddings:

$$\begin{aligned} X_t' = \Big\{ x_t^c | c \in \{0, 1, .., C-1\}, \\ x_t^{c\prime} \leftarrow \arg\min_{x_t^{y_{\text{true}}}} \text{Sim}[E_i(x_i'), E_t(x_t^{y_{\text{true}}})] \Big\} \end{aligned} \tag{11}$$

where X_l^c is the c-th label token. However, due to the fact that it is impractical to directly learn each optimal text token, we turn to modify the Probability Weighted Word Saliency (Ren et al. 2019) method for each word token by randomly replacing each word token $v_n \subset X_p$ as:

$$\begin{aligned} S(v_n) = \max\big[p(y'|x', X_p) - p(y'|x', X_p^n), 0\big], \\ where \quad X_p^n = \{v_1, ...v_{n-1}, < \text{MASK} >, ..., v_m\}, \end{aligned} \tag{12}$$

Method	Surrogate	CIFAR-10 (Train/ Val)						ImageNet (Train/ Val)					
		CLIP		ResNet-50		Overall		CLIP		ResNet-50		Overall	
Clean	-	88.3	88.5	100.0	94.6	94.2	91.6	59.1	59.0	75.9	76.5	67.5	67.8
White-box Attack	ResNet-50	64.2	64.2	13.7	13.4	39.0	38.8	42.0	41.9	5.7	6.0	23.9	24.0
w/o ℓ_{feat}	CLIP	12.1	12.4	53.2	54.7	32.7	33.6	9.6	9.7	55.4	54.7	32.5	32.2
w/o ℓ_{tri}	CLIP	11.9	11.9	51.0	52.0	31.5	32.0	9.7	8.9	54.2	55.7	32.0	32.3
w/o ℓ_{cls}	CLIP	10.1	10.6	52.8	53.1	31.5	31.9	10.6	12.2	59.2	59.8	34.9	36.0
Visual Attack w/o Iter	CLIP	8.1	8.9	54.5	55.2	31.3	32.1	8.8	9.7	53.9	54.3	31.4	32.0
Random Prompt	CLIP	8.4	8.6	55.2	56.4	31.8	32.5	8.3	9.1	41.3	39.7	24.8	24.4
Ours(Full)	CLIP	7.9	7.2	41.3	41.8	**24.6**	**24.5**	7.5	7.8	25.0	25.3	**16.3**	**16.6**

Table 1: Ablation study on attacking CLIP. The experiments are conducted on both CIFAR-10 and ImgaeNet datasets. The attacks are obtained by the CLIP model and are tested with the CLIP model and a target pre-trained ResNet-50.

where we mask each word token to calculate the saliency score that fools the model to make the wrong prediction y' on the adversarial sample x'. We set the threshold value to be ρ, meaning that only the word tokens $X_{\text{update}} = \{v_n | S(v_n) > \rho, 1 \leq n \leq m\}$ are set to be updated.

Thus, we update each word token in X_{update} from a set of candidates, the updating process is formulated as:

$$\begin{aligned} v_n* = \arg\max_{v'_n} & p(y_{\text{true}}|x', X_p(v'_n) - p(y_{\text{true}}|x', X_p^n)), \\ & X_p(v'_n) = \{v_1, ...v_{n-1}, \mathbf{v'_n}, ..., v_m\}, \\ & v_n \in X_{\text{update}} \quad and \quad v'_n \in \Gamma(v_n), \end{aligned} \tag{13}$$

where $\Gamma(v_n)$ is the candidate word set generated by GPT-2 (Radford et al. 2019). And each candidate word token is updated to correct the semantic consistency again by ensuring most of the perturbed samples are re-matched with its groundtruth related word embedding.

As a whole, the prompt tuning function can be donated as: $\mathcal{P}(X_t) = X_l + X_p(\cup_n v_n*)$. And the full algorithm can be found in the supplementary.

Experiments

In our experiments, we evaluated the attack performance of our proposed framework on several publicly available benchmark datasets. As our framework aims to generate highly transferable attacks, we focused on evaluating the transferability of the attacks in cross-dataset and cross-architecture settings.

Settings

Datasets. Followed previous work (Hayes and Danezis 2017), We evaluate attacks using two popular datasets in adversarial examples research, which are the CIFAR-10 dataset (Krizhevsky 2009) and the ImageNet dataset (Russakovsky et al. 2014). For testing the cross-dataset transferability, we follow previous works (Naseer et al. 2019; Salzmann et al. 2021) and use the Comics and Paintings (Kaggle 2017) or ChestX datasets as source domain, and evaluate on the randomly selected 5000 images from the ImageNet.

Implemental Details. We used PyTorch framework for the implementation. In the normal setting of using the pre-trained CLIP as the surrogate model, we choose the 'ViT/32' as backbone. As for the generator, we choose to use the ResNet backbone, and set the learning rate to be 0.0001 with Adam optimizer. All images are scaled to 224×224 to train the generator. For the ℓ_∞ bound, we set $\epsilon = 0.04$. A total of 10 iterations (we set the NUM_G to be 2) are used to train the whole network, which costs about 8 hours on one NVIDIA GeForce RTX 3090 GPUs.

Inference Metrics. We evaluate the performance of the generated attack by the mean accuracy of the classifier, which is better with lower values. In addition, as we aim at generating the high-transferable attacks, which is evaluated with various down-stream networks and datasets. We get the overall accuracy be the group-wise average, where the architectures with similar architectures are calculated once, i.e. we average the accuracies on ResNet-based architecture.

Experimental Results

Ablation study on the proposed framework. We conduct the ablation study on the proposed framework in Table 1, where the methods listed for comparison are:

- Clean: clean input without any attack;
- White-box Attack: we generate the attack directly on the target network ResNet-50, which serives as the upper-bound;
- w/o ℓ_{feat}/ℓ_{tri}/ℓ_{cls}: the visual attack optimized without the loss item ℓ_{feat}/ℓ_{tri}/ℓ_{cls};
- Visual Attack w/o Iter: the visual attack with semantic consistency optimized with loss $\ell_{feat} + \ell_{tri} + \ell_{cls}$;
- Random Prompt: we use GPT-2 to randomly generate X_p in each iteration.

As can be observed from the table: (1) The proposed method achieves the best attack performance in 'Overall', which fools the classifier by decreasing the accuracies more than 60% on CIFAR10 and more than 50% on ImageNet. (2) Only applying the visual attack on CLIP ('Visual Attack w/o Iter') can successfully attack the CLIP Model, but transfers bad on the down stream network. (3) Randomly update the prompt templates ('Random Prompt') can't improve the attack's transferablity a lot, indicating the effectiveness of the proposed prompt tuning $\mathcal{P}(\cdot)$.

Method	Dataset	Transfer to Target Networks									Overall
		CLIP	Res18	Res34	Res50	VGG11	VGG19	Shufflev2	Mobilev2	SimpleViT	
Clean	ImageNet	59.0	70.3	73.6	76.5	69.2	72.5	69.8	72.1	80.9	71.0
	CIFAR-10	88.5	94.6	94.7	94.9	92.2	93.1	90.2	93.9	81.7	89.9
UAN-Res	ImageNet	41.9	18.5	20.1	6.0	16.4	15.6	32.4	11.3	60.5	31.0
	CIFAR-10	64.2	19.6	23.9	13.4	71.7	38.7	58.2	15.3	31.7	41.4
UAN-Clip	ImageNet	17.3	29.8	35.3	37.3	21.8	22.5	34.1	28.5	53.9	31.8
	CIFAR-10	10.6	50.2	52.9	54.5	70.8	40.5	55.1	30.9	25.5	37.5
BIA-VGG	ImageNet	45.9	23.6	23.7	25.4	9.0	3.6	30.4	20.6	62.2	32.8
	CIFAR-10	-	-	-	-	-	-	-	-	-	-
Ours-Clip	ImageNet	7.8	17.3	22.8	25.3	15.6	19.9	23.4	25.7	44.8	**23.3**
	CIFAR-10	7.2	38.5	40.2	41.8	64.5	38.9	45.2	20.0	20.6	**30.5**

Table 2: Comparative results on the cross-architectures transferability. We evaluate the classification accuracy (lower is better) and show the results on both CIFAR-10 and ImageNet datasets. Our proposed method achieves the best in the 'overall' metric.

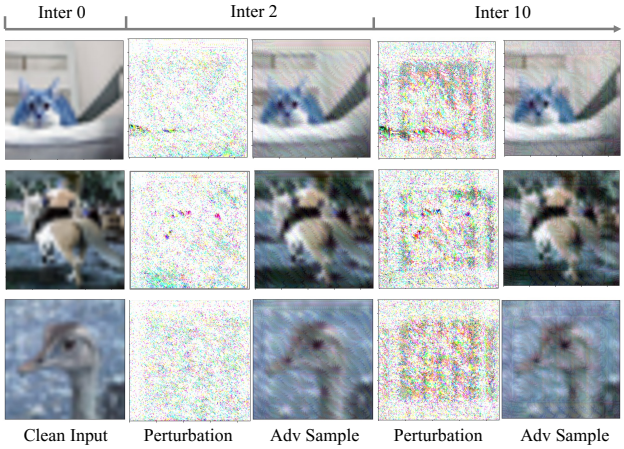
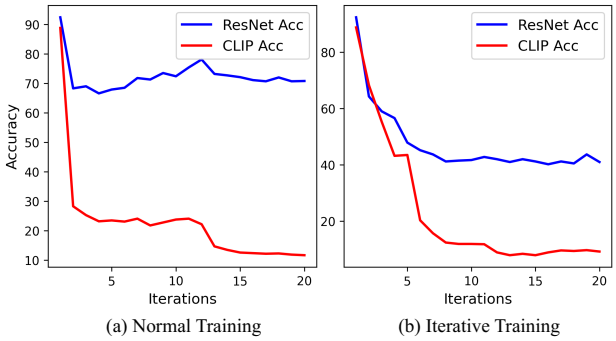


Figure 3: Visualizations on CIFAR-10 dataset on the 2-nd and the 10-th iterations.



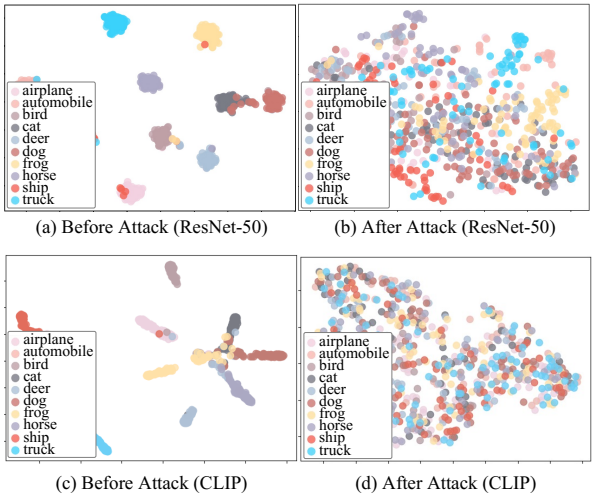
(a) Normal Training (b) Iterative Training

Figure 4: The attack's performance from the generative perturbation network in each iteration training on CIFAR10.

Visualization the generated adversarial samples. We show the generated adversarial samples in Fig. 3. In the figure, we sample the perturbation generator at the 2-nd iteration and the 10-th, respectively. It can be observed from the figure that the perturbations generated from the pre-trained CLIP model tend like some regular patterns. And this perturbation patterns are strengthened with the iterative training. It is worth noting that when generating the targeted perturbations, we find more interesting patterns, which could be found in the supplementary.

The performance during the iterative training. The main idea of the proposed framework is to utilize the iterative training strategy. Here, we depict the generated attack's performance while the iterative training in Fig. 4, where we compare the performance on normal training the generator $\mathcal{G}$ (a) and the proposed iteratively training the generator (b). As can be seen in the figure, we evaluate the attack capacity on both the surrogate network to train the generator (CLIP) and the target network (ResNet50). Thus, the observations are: (1) In the normal training scheme, the generator convergences faster, but finally the attack success rate is lower than the proposed iterative training. (2) In the normal training scheme, the attack's transferablity performance on the ResNet improves at first, but fails in the afterward iterations. While in our proposed framework, the transferablity improves stably, which is mainly due to the updating on the other modality. (3) About 10-iteration training would optimize an optimal perturbation generator, thus, we set the total iteration number to be 10 in the rest of experiments.

Analysis on the embedding visualization. We visualize the embedding features in Fig. 5. The visualization includes the feature space before and after attack and is conducted on the CIFAR-10 dataset, where the following observations can be made: (1) The features belonging to the same category are grouped together, making it easy for the classifier to recognize them. However, the feature space after attack is mixed together, which fools the classifier. (2) Comparing the visualized features after attack of CLIP (d) and ResNet-50 (c), the adversarial features generated by our work are mixed together more evenly. This makes it much more difficult to



(a) Before Attack (ResNet-50) (b) After Attack (ResNet-50)

(c) Before Attack (CLIP) (d) After Attack (CLIP)

Figure 5: TSNE visualization on the features with the clean/adversarial images as the input.

defend against and indicates a higher attack capability.

Transferability Evaluation We have compared with other methods on the cross-architecture, cross-dataset (domain) and cross-task transferability.

Evaluate the cross-architecture transferability We form a set of pre-trained networks in various architectures, which are divided into 5 groups: (1) CLIP, (2) ResNet18, ResNet34, ResNet50, (3) VGG11, VGG19, (4) ShuffleNetv2, MobileNetv2 and (5) SimpleViT. Based on these grouping strategy, we calculate the overall accuracy by group-average accuracy. The experimental results are compared in Table 2. We have compared the proposed method with the other generator-oriented methods, which are UAN (Hayes and Danezis 2017) ('UAN-Res'), modified UAN that uses CLIP as the surrogate model ('UAN-CLIP') and BIA (Zhang, Yi, and Sang 2022).From the table, we observe that: (1) When transferring the generated attack to the target networks, the attacks perform better between the networks in similar architecture as the surrogate model; (2) We propose to generate the attacks with high transferability, which decrease the overall accuracies most on both ImageNet and CIFAR-10 datasets; (3) The other attack methods ('UAN' and 'BIA') perform unevenly according to the target networks, which could be easily defended by the ensemble attack; while **the proposed mutual-modality attack is stable whatever the target networks**, making it more difficult to defend. The corresponding experiment against the ensemble defense is included in the supplementary.

Evaluate the Cross-dataset Transferability. We evaluate the cross-data transferability by training the generator on the source dataset and test on the target dataset. Following the previous setting, we train the generator on Comics/Paintings/ChestX datasets with ChestXNet as the discriminator and evaluate the attack performance on ImageNet.

As we propose a plug-and-play method, we test the effectiveness of our method by integrating it into the existing methods, including: GAP (Poursaeed et al. 2018), CDA (Naseer et al. 2019) and LTAP (Salzmann et al. 2021). As can be observed from the table: (1) Enabling the cross-dataset transferability of the attacks is much more difficult than the cross-architecture one, and our method also shows satisfying results; (2) **Our method (' Curr. + Ours') improves the cross-dataset attack success rates when integrated into the current methods** especially on the cases with CLIP and SimpleViT as target networks.

Mtd	Datasets	Res152	CLIP	SimpleViT
		Curr. / Curr. + Ours		
GAP	Cosmics	50.3/51.2	20.3/30.8	27.6/35.7
	Paintings	52.9/53.0	36.7/37.7	37.8/46.9
	ChestX	29.2/32.8	19.8/36.4	19.4/38.7
CDA	Cosmics	38.8/39.6	36.8/46.6	37.6/42.6
	Paintings	41.7/41.8	38.6/43.5	39.0/46.2
	ChestX	23.7/30.3	16.7/29.3	19.2/27.9
LTAP	Cosmics	55.2/56.5	43.6/45.5	48.4/54.0
	Paintings	59.9/60.7	44.8/53.5	48.6/48.7
	ChestX	49.5/50.3	28.9/34.2	21.9/24.6

Table 3: Extreme cross-domain (dataset) transferability analysis evaluated by attack success rate.

Method	VGG16	Res50	SimpleViT	Overall
Clean	51.3	69.9	54.7	58.6
GAMA	**3.1**	22.3	38.3	21.2
GAMA+Ours	3.4	**20.4**	**35.9**	**19.9**

Table 4: The cross-task transferability evaluation.

Evaluate the Cross-task Transferability. Following previous work (Zhang et al. 2022) we conduct the cross-task tranferability evaluation in Table 4. We integrate the proposed framework into GAMA (Aich et al. 2022) ('GAMA+Ours'). We train the generator with Pascal-VOC dataset (Everingham et al. 2010), and then test on the ImageNet classification task. As can be observed from the figure, our proposed framework could be integrated into any adversarial attack framework, which could further enhance the attack transferability.

Conclusion

Overall, our proposed approach demonstrates promising results in improving the transferability and stability of adversarial attacks by generating perturbations in the semantic feature embedding space using the pre-trained CLIP model. By optimizing the attack iteratively from both image and text modalities, our method achieves improved transferability across different architectures and datasets, as demonstrated in our experiments on several benchmark datasets. We believe that our work provides a valuable contribution to the field of adversarial attacks and could have important implications for improving the security and reliability of machine learning systems in real-world applications.

Acknowledgements

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award Number: MOE-T2EP20122-0006).

References

Aich, A.; Ta, C.-K.; Gupta, A. A.; Song, C.; Krishnamurthy, S.; Asif, M. S.; and Roy-Chowdhury, A. 2022. GAMA: Generative Adversarial Multi-Object Scene Attacks. In *Advances in Neural Information Processing Systems*, volume 35, 36914–36930.

Akhtar, N.; and Mian, A. S. 2018. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access*, 6: 14410–14430.

Boucher, N.; Shumailov, I.; Anderson, R.; and Papernot, N. 2022. Bad characters: Imperceptible nlp attacks. In *2022 IEEE Symposium on Security and Privacy (SP)*, 1987–2004. IEEE.

Chen, X.; Salem, A.; Chen, D.; Backes, M.; Ma, S.; Shen, Q.; Wu, Z.; and Zhang, Y. 2021. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In *Annual Computer Security Applications Conference*, 554–569.

Cheng, S.; Dong, Y.; Pang, T.; Su, H.; and Zhu, J. 2019. Improving black-box adversarial attacks with a transfer-based prior. *Advances in neural information processing systems*, 32.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2017. Boosting Adversarial Attacks with Momentum. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9185–9193.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Du, W.; Zhao, Y.; Li, B.; Liu, G.; and Wang, S. 2022. PPT: Backdoor Attacks on Pre-trained Models via Poisoned Prompt Tuning. In *International Joint Conference on Artificial Intelligence*.

Everingham, M.; Gool, L. V.; Williams, C. K. I.; Winn, J. M.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88: 303–338.

Fang, G.; Ma, X.; Song, M.; Mi, M. B.; and Wang, X. 2023. DepGraph: Towards Any Structural Pruning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Fang, G.; Ma, X.; and Wang, X. 2023. Structural Pruning for Diffusion Models. In *Advances in neural information processing systems*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and Harnessing Adversarial Examples. *International Conference on Learning and Representations*.

Guo, C.; Gardner, J.; You, Y.; Wilson, A. G.; and Weinberger, K. 2019. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, 2484–2493. PMLR.

Hayes, J.; and Danezis, G. 2017. Learning Universal Adversarial Perturbations with Generative Models. *2018 IEEE Security and Privacy Workshops (SPW)*, 43–49.

Hintersdorf, D.; Struppek, L.; and Kersting, K. 2022. CLIPping Privacy: Identity Inference Attacks on Multi-Modal Machine Learning Models. *arXiv preprint arXiv:2209.07341*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

Jia, J.; Liu, Y.; and Gong, N. Z. 2022. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, 2043–2059. IEEE.

Kaggle. 2017. Painter by Number. https://www.kaggle.com/c/painter-by-numbers/data.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Liu, A.; Liu, X.; Fan, J.; Ma, Y.; Zhang, A.; Xie, H.; and Tao, D. 2019. Perceptual-Sensitive GAN for Generating Adversarial Patches. In *AAAI Conference on Artificial Intelligence*.

Liu, S.; Wang, K.; Yang, X.; Ye, J.; and Wang, X. 2022. Dataset Distillation via Factorization. In *Advances in neural information processing systems*.

Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2016. Delving into Transferable Adversarial Examples and Black-box Attacks. In *International Conference on Learning Representations*.

Ma, X.; Fang, G.; and Wang, X. 2023a. DeepCache: Accelerating Diffusion Models for Free. *arXiv preprint arXiv:2312.00858*.

Ma, X.; Fang, G.; and Wang, X. 2023b. LLM-Pruner: On the Structural Pruning of Large Language Models. In *Advances in neural information processing systems*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards Deep Learning Models Resistant to Adversarial Attacks. *International Conference on Learning Representations*.

Morris, J. X.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. *EMNLP 2020*, 119.

Mu, N.; Kirillov, A.; Wagner, D.; and Xie, S. 2022. Slip: Self-supervision meets language-image pre-training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, 529–544. Springer.

Naseer, M. M.; Khan, S. H.; Khan, M. H.; Shahbaz Khan, F.; and Porikli, F. 2019. Cross-domain transferability of adversarial perturbations. *Advances in Neural Information Processing Systems*, 32.

Noever, D. A.; and Noever, S. E. M. 2021. Reading Isn't Believing: Adversarial Attacks On Multi-Modal Neurons. *arXiv preprint arXiv:2103.10480*.

Perez, F.; and Ribeiro, I. 2022. Ignore Previous Prompt: Attack Techniques For Language Models. *ArXiv*, abs/2211.09527.

Poursaeed, O.; Katsman, I.; Gao, B.; and Belongie, S. 2018. Generative adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4422–4431.

Poursaeed, O.; Katsman, I.; Gao, B.; and Belongie, S. J. 2017. Generative Adversarial Perturbations. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4422–4431.

Qiu, H.; Xiao, C.; Yang, L.; Yan, X.; Lee, H.; and Li, B. 2019. SemanticAdv: Generating Adversarial Examples via Attribute-conditional Image Editing. *European Conference on Computer Vision*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Ren, S.; Deng, Y.; He, K.; and Che, W. 2019. Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency. In *Annual Meeting of the Association for Computational Linguistics*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Fei-Fei, L. 2014. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115: 211–252.

Salzmann, M.; et al. 2021. Learning transferable adversarial perturbations. *Advances in Neural Information Processing Systems*, 34: 13950–13962.

Shi, Y.; Li, P.; Yin, C.; Han, Z.; Zhou, L.; and Liu, Z. 2022. PromptAttack: Prompt-Based Attack for Language Models via Gradient Search. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I*, 682–693. Springer.

Wang, X.; and He, K. 2021. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1924–1933.

Wu, D.; Wang, Y.; Xia, S.-T.; Bailey, J.; and Ma, X. 2019. Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets. In *International Conference on Learning Representations*.

Xiang, T.; Liu, H.; Guo, S.; Gan, Y.; and Liao, X. 2022. EGM: An Efficient Generative Model for Unrestricted Adversarial Examples. *ACM Transactions on Sensor Networks (TOSN)*.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2730–2739.

Yang, X.; Zhou, D.; Liu, S.; Ye, J.; and Wang, X. 2022. Deep Model Reassembly. In *Advances in neural information processing systems*.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. FILIP: Fine-grained Interactive Language-Image Pre-Training. In *International Conference on Learning Representations*.

Ye, J.; Fu, Y.; Song, J.; Yang, X.; Liu, S.; Jin, X.; Song, M.; and Wang, X. 2022a. Learning with recoverable forgetting. In *European Conference on Computer Vision*, 87–103. Springer.

Ye, J.; Liu, S.; and Wang, X. 2023. Partial network cloning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20137–20146.

Ye, J.; Mao, Y.; Song, J.; Wang, X.; Jin, C.; and Song, M. 2022b. Safe distillation box. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3117–3124.

Yuan, Z.; Zhang, J.; Jia, Y.; Tan, C.; Xue, T.; and Shan, S. 2021. Meta Gradient Adversarial Attack. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7728–7737.

Zhang, C.; Benz, P.; Lin, C.; Karjauv, A.; Wu, J.; and Kweon, I. S. 2021. A Survey On Universal Adversarial Attack. *International Joint Conference on Artificial Intelligence*.

Zhang, J.; Yi, Q.; and Sang, J. 2022. Towards Adversarial Attack on Vision-Language Pre-training Models. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5005–5013.

Zhang, Q.; Li, X.; Chen, Y.; Song, J.; Gao, L.; He, Y.; and Xue, H. 2022. Beyond imagenet attack: Towards crafting adversarial examples for black-box domains. *arXiv preprint arXiv:2201.11528*.

Zhang, W. E.; Sheng, Q. Z.; Alhazmi, A.; and Li, C. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3): 1–41.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.