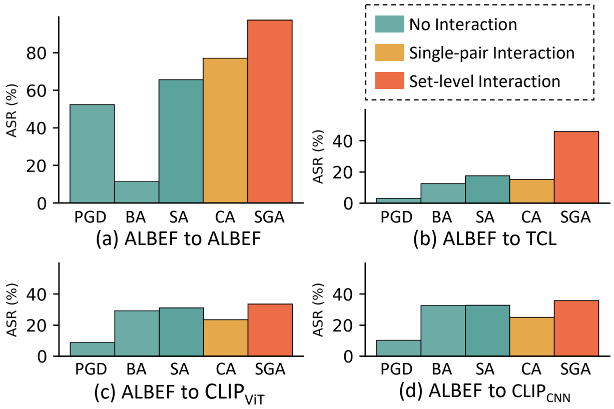
This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Set-level Guidance Attack: Boosting Adversarial Transferability of Vision-Language Pre-training Models

Dong Lu[1]*, Zhiqiang Wang[1]*, Teng Wang[1,2], Weili Guan[3], Hongchang Gao[4], Feng Zheng[1,5]†
[1]Southern University of Science and Technology [2]The University of Hong Kong
[3]Monash University [4]Temple University [5]Peng Cheng Laboratory
sammylu_@outlook.com wangzq_2021@outlook.com tengwang@connect.hku.hk
honeyguan@gmail.com hongchang.gao@temple.edu f.zheng@ieee.org

Abstract

Vision-language pre-training (VLP) models have shown vulnerability to adversarial examples in multimodal tasks. Furthermore, malicious adversaries can be deliberately transferred to attack other black-box models. However, existing work has mainly focused on investigating white-box attacks. In this paper, we present the first study to investigate the adversarial transferability of recent VLP models. We observe that existing methods exhibit much lower transferability, compared to the strong attack performance in white-box settings. The transferability degradation is partly caused by the under-utilization of cross-modal interactions. Particularly, unlike unimodal learning, VLP models rely heavily on cross-modal interactions and the multimodal alignments are many-to-many, e.g., an image can be described in various natural languages. To this end, we propose a highly transferable Set-level Guidance Attack (SGA) that thoroughly leverages modality interactions and incorporates alignment-preserving augmentation with cross-modal guidance. Experimental results demonstrate that SGA could generate adversarial examples that can strongly transfer across different VLP models on multiple downstream vision-language tasks. On image-text retrieval, SGA significantly enhances the attack success rate for transfer attacks from ALBEF to TCL by a large margin (at least 9.78% and up to 30.21%), compared to the state-of-the-art. Our code is available at https://github.com/Zoky-2020/SGA.



(a) ALBEF to ALBEF (b) ALBEF to TCL (c) ALBEF to CLIP$_{ViT}$ (d) ALBEF to CLIP$_{CNN}$

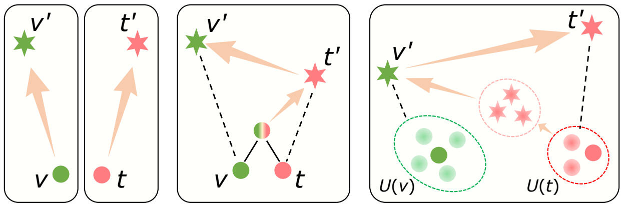
Figure 1: **Comparison of attack success rates (ASR) using five different attacks on image-text retrieval.** Adversarial examples are crafted on the source model (ALBEF) to attack the target white-box model or black-box models. The first three columns refer to the image-only **PGD** attack [24], text-only BERT-Attack [19] (**BA**), and the combined separate unimodal attack (**SA**), which all belong to the methods without cross-modal interactions. The fourth column is the state-of-the-art multimodal Co-Attack [41] (**CA**) that employs single-pair cross-modal interactions. The last column is the proposed Set-level Guidance Attack (**SGA**), which leverages multiple set-level cross-modal interactions, successfully attacking the white-box model and transferring to attack all black-box models with the highest ASR. More discussions are in Section 3.

1. Introduction

Recent work has shown that vision-language pre-training (VLP) models are still vulnerable to adversarial examples [41], even though they have achieved remarkable performance on a wide range of multimodal tasks [30, 15, 16]. Existing work mainly focuses on white-box attacks, where information about the victim model is accessible. However, the transferability of adversarial examples across VLP models has not been investigated, which is a more practical setting. It is still unknown whether the adversarial data generated on the source model can successfully attack another model, which poses a serious security risk to the deployment of VLP models in real-world applications.

This paper makes the first step to investigate the trans-

[*] Equal contribution. [†] Corresponding author.



(a) Sep-Attack (b) Co-Attack (c) SGA

Figure 2: **Comparison of cross-modal interactions.** To generate adversarial examples, existing methods either: **(a)** separately perturb unimodal data without any cross-modal interactions (**Sep-Attack**) or, **(b)** perturb multiple modalities but with single image-text pairs to model cross-modal interactions (**Co-Attack**). However, our method is capable of learning cross-modal interactions among multiple alignments through **(c)** set-level guidance (**SGA**). Specifically, v denotes the input image, and t is the paired caption. v' and t' represent the corresponding adversarial examples, respectively. $U(v)$ and $U(t)$ represent the scale-invariant image set and most matching caption set. Arrows indicate the guidance for generating adversarial examples.

ferability of adversarial samples within VLP models. Without loss of generality, most of our experiments are based on image-text retrieval tasks. We first empirically evaluate this attack performance with respect to different modalities on multimodal tasks across multiple datasets. Our results show that the adversarial transferability of attacking both modalities (image & text) consistently beats attacking unimodal data (image or text). Unfortunately, even though two modalities are allowed to be perturbed simultaneously, the attack success rates of existing methods [24, 19, 41] still significantly drops when transferring from the white-box to black-box settings, as shown in Figure 1.

Different from recent studies focusing on separate attacks on unimodal data [24, 19], multimodal pairs exhibit intrinsic alignment and complementarity to each other. The modeling of inter-modal correspondence turns out to be a critical problem for transferability. Considering that the alignments between image and text are many-to-many, for example, an image could be described to be with various human perspectives and language styles, a reasonable perturbation direction may be determined with diverse guidance from multiple labels in the other modality. However, recent adversarial attack methods for VLP models [41] usually employ a single image-text pair to generate adversarial samples. Although they exhibit strong performance in white-box settings, the poor diversity of guidance makes adversarial samples highly correlated with the alignment pattern of the white-box model, and therefore impedes generalization to black-box settings.

To address the weak transferability problem, we propose Set-level Guidance Attack (SGA), which leverages diverse cross-modal interactions among multiple image-text pairs (Figure 2). Specifically, we introduce alignment-preserving augmentation which enriches image-text pairs while keeping their alignments intact. The image augmentation is based on the scale-invariant property of deep learning models [22], thus we can construct multi-scale images to increase the diversity. For text augmentation, we select the most matching caption pairs from the dataset. More importantly, SGA generates adversarial examples on multimodal augmented input data with carefully designed cross-modal guidance. In detail, SGA iteratively pushes supplemental information away between two modalities with another modality as supervision to disrupt the interactions for better harmonious perturbations. Note that resultant adversarial samples could perceive the gradients originated from multiple guidance.

We conduct experiments on two well-established multimodal datasets, Flickr30K [27] and MSCOCO [23], to evaluate the performance of our proposed SGA across various Vision-and-Language (V+L) downstream tasks. The experimental results demonstrate the high effectiveness of SGA in generating adversarial examples that can be strongly transferred across VLP models, surpassing the current state-of-the-art attack methods in multimodal learning. In particular, SGA achieves notable improvements in image-text retrieval under black-box settings and also exhibits superior performance in white-box attack settings. Moreover, SGA also outperforms the state-of-the-art methods in image captioning and yields higher fooling rates on visual grounding.

We summarize our contributions as follows. **1)** We make the first attempt to explore the transferability of adversarial examples on popular VLP models with a systematical evaluation; **2)** We provide SGA, a novel transferable multimodal attack that enhances adversarial transferability through the effective use of set-level alignment-preserving augmentations and well-designed cross-modal guidance; **3)** Extensive experiments show that SGA consistently boosts adversarial transferability across different VLP models than the state-of-the-art methods.

2. Related Work

2.1. Vision-Language Pre-training Models

Vision-Language Pre-training (VLP) aims to improve the performance of downstream multimodal tasks by pre-training large-scale image-to-text pairs [17]. Most works are developed upon the pre-trained object detectors with region features to learn the vision-language representations [5, 20, 42, 34]. Recently, with the increasing popularity of Vision Transformer (ViT) [8, 31, 40], some other works propose to use ViT as an image encoder and transform the input into patches in an end-to-end manner [18, 38, 17, 9, 33].

According to the VLP architectures, VLP models can be classified into two typical types: fused VLP models and aligned VLP models [41]. Specifically, fused VLP models (*e.g.*, ALBEF [18], TCL [38]) first utilize separate unimodal encoders to process token embeddings and visual features, and further use a multimodal encoder to process image and text embeddings to output fused multimodal embeddings. Alternatively, aligned VLP models (*e.g.*, CLIP [28]) have only unimodal encoders with independent image and text modality embeddings. In this paper, we focus on popular architectures with fused and aligned VLP models.

2.2. Image-Text Retrieval Task

Image-Text Retrieval (ITR) aims to retrieve the relevant top-ranked instances from a gallery database with one modality, given an input query from another modality [35, 4, 43, 6]. This task can be divided into two subtasks, image-to-text retrieval (TR) and text-to-image retrieval (IR).

For ALBEF [18] and TCL [38], the semantic similarity score in the unimodal embedding space will be calculated for all image-text pairs to select top-k candidates. Then the multimodal encoder takes the top-k candidates and computes the image-text matching score for ranking. For CLIP [28], without the multimodal encoder, the final rank list can be obtained based on the similarity in the embedding space between image and text modalities.

2.3. Adversarial Transferability

Existing adversarial attacks can be categorized into two settings: white-box attacks and black-box attacks. In a white-box setting, the target model is fully accessible, but not in a black-box setting. In computer vision, many methods employ gradient information for adversarial attacks in white-box settings, such as FGSM [11], PGD [24], C&W [3], and MI [7]. In contrast, in the field of natural language processing (NLP), current attack methods mainly modify or replace some tokens of the input text [19, 29, 10, 13]. In the multimodal vision-language domain, Zhang *et al.* [41] proposed a white-box multimodal attack method with respect to popular VLP models on downstream tasks.

However, white-box attacks are unrealistic due to the inaccessibility of model information in practical applications. In addition, there is no related work that systematically analyzes the adversarial transferability of multimodal attack methods on VLP models. Therefore, in this work, we mainly focus on generating highly transferable adversaries across different VLP models.

3. Analysis of Adversarial Transferability

In this section, we conduct an empirical study on VLP models to evaluate adversarial transferability using existing methods. A common approach to attack multimodal tasks is combining unimodal adversarial attacks [24, 7, 37, 19] of each modalities together. For instance, the separate unimodal attack (Sep-Attack) includes PGD [24] and BERT-Attack [19] for attacking image modality and text modality, respectively. Other recent multimodal adversarial attacks, such as Co-Attack [41], consider cross-modal interactions by perturbing image modalities and text modalities collectively.

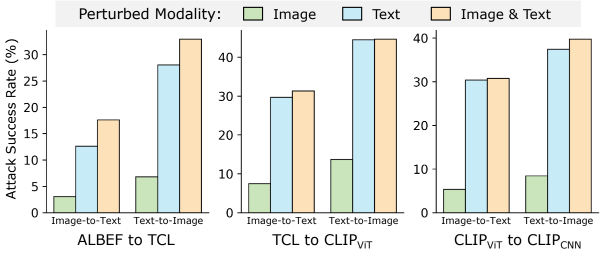


Figure 3: **Attack success rates (%) of different perturbed modalities on image-text retrieval.** Adversarial examples with respect to different modalities from the source model to attack the target model using Sep-Attack. Different colors represent different perturbed modalities.

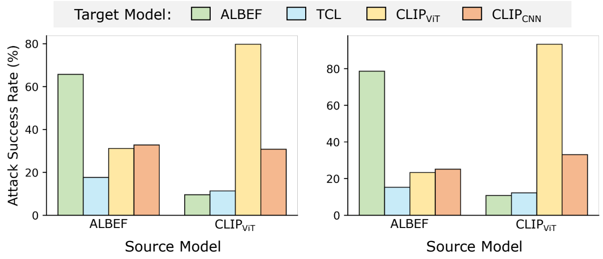


Figure 4: **Attack success rates (%) on white-box models and black-box models.** Both images and text are perturbed by Sep-Attack (**Left**) and Co-Attack (**Right**) on image-text retrieval. Colors indicate different target models.

We first present the observations regarding the adversarial transferability of VLP models. Then we discuss the limitations of the existing methods. By conducting this study, we aim to provide insights into the robustness of VLP models against adversarial attacks and the effectiveness of different attack strategies.

3.1. Observations

To investigate the adversarial transferability of perturbed inputs with respect to different modalities (*i.e.*, image, text, and image & text) and the effect of different VLP models on transferability, we conduct experiments and present the attack success rates of the adversarial examples generated by the source model to attack the target models in Figure 3 and Figure 4. The observations are summarized below:

- The adversarial transferability of attacking both modalities (image & text) is consistently more effective than attacking any unimodal data alone (image or text). As shown in Figure 3, transferring both adversarial image and text from ALBEF to TCL leads to a much higher attack success rate than transferring adversarial examples of any single modality. Notably, ALBEF and TCL are both fused VLP models. Similar observations also exist in the following settings: (1) The source and target models are different types of VLP models but have the same basic architectures (*e.g.*, TCL and CLIP_{ViT}). (2) The source and target models are the same types of VLP models, but with different basic architectures (*e.g.*, CLIP_{ViT} and CLIP_{CNN}).

- Adversarial multimodal data (*i.e.*, adversarial image & text), which have strong attack performance on the source model, can hardly maintain the same capability when transferring to target models. For example, as illustrated in Figure 4, even though ALBEF and TCL have the same model architecture, the attack success rate sees a significant drop when transferring adversarial examples generated on ALBEF to TCL. The phenomenon exists in both Sep-Attack and Co-Attack.

In summary, the attack methods can have stronger transferability in black-box settings if all the modalities are attacked simultaneously. However, even though two modalities are allowed to be perturbed, existing methods still exhibit much lower transferability. This suggests that attacks with higher transferability should be specifically designed, instead of directly utilizing the existing white-box attack methods.

3.2. Discussions

We posit that the degradation in transferability of adversarial examples is mainly due to the limitations of the existing attack methods:

- One major limitation of Sep-Attack is that it does not take into account the interactions between different modalities. As a combined independent attack method for each modality, it fails to model the inter-modal correspondence that is crucial for successful attacks in multimodal learning. This is particularly evident in multimodal tasks such as image-text retrieval, where the ground truth is not discrete labels (*e.g.*, image classification) but another modality data that corresponds to the input modality. The complete lack of cross-modal interactions in Sep-Attack severely limits the generalization of adversarial examples and reduces their transferability among different VLP models.

- While Co-Attack is designed to leverage the collaboration between modalities to generate adversarial examples, it still suffers from a key drawback that hinders its transferability to other VLP models. Unlike unimodal learning, multimodal learning involves multiple complementary modalities with many-to-many cross-modal alignments, which pose unique challenges to achieving sufficient adversarial transferability. However, Co-Attack only uses single image-text pairs to generate adversarial data, limiting the diversity of guidance from multiple labels in other modalities. This lack of diversity in cross-modal guidance makes adversarial samples highly correlated with the alignment pattern of the white-box model. Therefore, the generality of adversarial examples is restricted, and their effectiveness in transferring to other models drops.

In conclusion, the analysis motivates our investigation into the adversarial transferability of VLP models. Moreover, it highlights the pressing need to explore transferable multimodal attacks for generating adversarial examples that can be effectively transferred across different VLP models.

4. Methodology

In this section, we propose a transferable multimodal adversarial attack, termed Set-level Guidance Attack (SGA). SGA is designed to enhance adversarial transferability across VLP models by leveraging multimodal interactions and incorporating diverse alignment-preserving inter-modal information with carefully designed cross-modal guidance. We provide our motivation first, followed by relevant notations, and finally present SGA in detail.

4.1. Motivation

To improve the transferability of multimodal attacks, we first conduct an investigation into the shortcomings of existing methods in black-box settings. Through a systematic analysis of failure cases, we observe that around half of these cases arise due to the presence of multiple matched captions of the image, as shown in Table 1. More specifically, our findings indicate that while the generated adversarial image may be significantly distant from the single supervised caption in the source model, it is prone to approaching other matched captions in the target model, where the alignments can be modeled and ultimately lead to attack failure.

Therefore, to maintain the attack ability of adversarial images when transferring to other models, it is crucial to consider multiple paired captions and push the adversarial image away from all the paired captions, thus preserving the attack ability when transferred to other black-box models. Crafting adversarial text for high transferability follows a similar approach, which can also benefit from more paired images. More discussions can be found in Appendix A.

4.2. Notations

Let (v, t) denote an image-text pair sampled from a multimodal dataset D. For VLP models, we denote f_I as the image encoder and f_T as the text encoder. The multimodal fusion module in fused VLP models is denoted by f_M. Specifically, $f_I(v)$ represents the image representation e_v encoded

Attack	Cross-modal	p (*Event A*) $\uparrow$	p (*Event B* $\mid$ *Event A*) $\downarrow$		
	Interaction	ALBEF	TCL	CLIP$_{\text{ViT}}$	CLIP$_{\text{CNN}}$
Sep-Attack	No	74.60%	46.78%	41.69%	41.82%
Co-Attack	Single-pair	80.60%	50.50%	40.82%	42.93%
SGA	Set-level	**97.20%**	**28.91%**	**34.67%**	**38.58%**

Table 1: **Adversarial images with insufficient cross-modal interactions may get closer to other paired captions when transferring.** ***Event A***: success cases in white-box settings. ***Event B***: failure cases in black-box settings influenced by other paired captions. Adversarial data are generated on Flickr30K from ALBEF to attack other target models.

by f_I taking an image v as input, $f_T(t)$ denotes the text representation e_t encoded by f_T taking a text t as input, and $f_M(e_v, e_t)$ denotes the multimodal representation encoded by f_M taking image and text representations as inputs.

We use $B[v, \epsilon_v]$ and $B[t, \epsilon_t]$ to represent the legal searching spaces for optimizing adversarial image and text, respectively. Specifically, ϵ_v denotes the maximal perturbation bound for the image, and ϵ_t denotes the maximal number of changeable words in the caption.

4.3. Transferable Set-Level Guidance Attack

Alignment-preserving Augmentation. The analysis presented in Section 3 highlights the key limitation of existing methods: the inter-modal information used to generate adversarial examples lacks diversity. The limitation will make the generated adversarial examples fail to generalize to other black-box models with strong attack performance, resulting in limited adversarial transferability.

To inject more diversity in the generation of generalizable adversarial examples, we propose using set-level alignment-preserving augmentation to expand multimodal input spaces while maintaining cross-modal alignments intact. Unlike previous methods that only consider a single image-text paired example (v, t) to generate adversarial data, we enlarge the input to a set level of images and captions. Specifically, we select the most matching caption pairs from the dataset of each image v to form an augmented caption set $\boldsymbol{t} = \{t_1, t_2, ..., t_M\}$, and resize each image v into different scales $S = \{s_1, s_2, ..., s_N\}$ and then add Gaussian noise to obtain a multi-scale image set $\boldsymbol{v} = \{v_1, v_2, ..., v_N\}$ based on the scale-invariant property. The enlarged input set $(\boldsymbol{v}, \boldsymbol{t})$ is then used to generate the adversarial data (v', t').

Cross-modal Guidance. Cross-modal interactions play a crucial role in multimodal tasks. For example, in image-text retrieval, the paired information from another modality provides unique annotation supervision for each sample. Similarly, in adversarial attacks, supervisory information is essential in guiding the search for adversarial examples.

To fully utilize the enlarged alignment-preserving multimodal input set $(\boldsymbol{v}, \boldsymbol{t})$ and further improve the transferability of the generated adversarial data, we propose cross-modal guidance to utilize interactions from different modalities. Specifically, we use the paired information from another modality as the supervision to guide the direction of optimizing the adversarial data. This guidance iteratively pushed away the multimodal information and disrupt the cross-modal interaction for better harmonious perturbations. Notably, the resultant adversarial examples can perceive the gradients originated from multiple guidance.

First, we generate corresponding adversarial captions for all captions in the text set $\boldsymbol{t}$, forming an adversarial caption set $\boldsymbol{t'} = \{t'_1, t'_2 ..., t'_M\}$. The process can be formulated as,

$$t'_i = \underset{t'_i \in B[t_i, \epsilon_t]}{\arg\max} - \frac{f_T(t'_i) \cdot f_I(v)}{\|f_T(t'_i)\| \|f_I(v)\|}. \tag{1}$$

The adversarial caption t'_i is constrained to be dissimilar to the original image v in the embedding space. Next, the adversarial image v' is generated by solving

$$v' = \underset{v' \in B[v, \epsilon_v]}{\arg\max} - \sum_{i=1}^{M} \frac{f_T(t'_i)}{\|f_T(t'_i)\|} \sum_{s_i \in S} \frac{f_I(g(v', s_i))}{\|f_I(g(v', s_i))\|}, \tag{2}$$

where $g(v', s_i)$ denotes the resizing function that takes the image v' and the scale coefficient s_i as inputs. All the scaled images derived from v' are encouraged to be far away from all the adversarial captions t'_i in the embedding space. Finally, the adversarial caption t' is generated as follows,

$$t' = \underset{t' \in B[t, \epsilon_t]}{\arg\max} - \frac{f_T(t') \cdot f_I(v')}{\|f_T(t')\| \|f_I(v')\|}, \tag{3}$$

in which t' is encouraged to be far away from the adversarial image v' in the embedding space. The detailed algorithm can be found in Appendix C.

5. Experiments

In this section, we present experimental evidence for the advantages of our proposed SGA. We conduct experiments on a diverse set of datasets and popular VLP models. First, we describe the experimental settings in Section 5.1. Then, in Section 5.2, we validate the immediate integration of transfer-based unimodal attacks into multimodal learning. Next, we provide the main evaluation results compared to the state-of-the-art method in Section 5.3. In Section 5.4, we analyze the cross-task transferability between different V+L tasks. Finally, we present ablation studies in Section 5.5.

5.1. Experimental Settings

Datasets. We consider two widely used multimodal datasets, Flickr30K [27] and MSCOCO [23]. Flickr30K consists of 31,783 images, each with five corresponding captions. Similarly, MSCOCO comprises 123,287 images, and each image is annotated with around five captions. We adopt the Karpathy split [14] for experimental evaluation.

Attack	**ALBEF***		**TCL**		**CLIP$_{\text{ViT}}$**		**CLIP$_{\text{CNN}}$**	
	TR R@1*	IR R@1*	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
Sep-Attack	65.69	73.95	17.60	32.95	31.17	**45.23**	32.82	45.49
Sep-Attack + MI	58.81(↓)	65.25(↓)	16.02(↓)	28.19(↓)	23.07(↓)	36.98(↓)	26.56(↓)	39.31(↓)
Sep-Attack + DIM	56.41(↓)	64.24(↓)	16.75(↓)	29.55(↓)	24.17(↓)	37.60(↓)	25.54(↓)	38.77(↓)
Sep-Attack + PNA_PO	40.56(↓)	53.95(↓)	18.44(↑)	30.98(↓)	22.33(↓)	37.02(↓)	26.95(↓)	38.63(↓)
Co-Attack	77.16	83.86	15.21	29.49	23.60	36.48	25.12	38.89
Co-Attack + MI	64.86(↓)	75.26(↓)	25.40(↑)	38.69(↑)	24.91(↑)	37.11(↑)	26.31(↑)	38.97(↑)
Co-Attack + DIM	47.03(↓)	62.28(↓)	22.23(↑)	35.45(↑)	25.64(↑)	38.50(↑)	26.95(↑)	40.58(↑)
SGA	**97.24**	**97.28**	**45.42**	**55.25**	**33.38**	44.16	**34.93**	**46.57**

Table 2: **Attack success rates (%) of R@1 of integrating transfer-based image attacks on image-text retrieval.** The adversarial data are generated on Flickr30K using the source model ALBEF to attack other target models. * indicates white-box attacks. A higher ASR indicates better adversarial transferability.

VLP Models. We evaluate two popular VLP models, the fused VLP and aligned VLP models. For the fused VLP, we consider ALBEF [18] and TCL [38]. ALBEF contains a 12-layer visual transformer ViT-B/16 [8] and two 6-layer transformers for the image encoder and both the text encoder and the multimodal encoder, respectively. TCL uses the same model architecture as ALBEF but with different pre-train objectives. For the aligned VLP model, we choose to evaluate CLIP [28]. CLIP has two different image encoder choices, namely, CLIP$_{\text{ViT}}$ and CLIP$_{\text{CNN}}$, that use ViT-B/16 and ResNet-101 [12] as the base architectures for the image encoder, respectively.

Adversarial Attack Settings. To craft adversarial images, we employ PGD [24] with perturbation bound $\epsilon_v = 2/255$, step size $\alpha = 0.5/255$, and iteration steps $T = 10$. For attacking text modality, we adopt BERT-Attack [19] with perturbation bound $\epsilon_t = 1$ and length of word list $W = 10$. Furthermore, we enlarge the image set by resizing the original image into five scales, $\{0.50, 0.75, 1.00, 1.25, 1.50\}$, using bicubic interpolation. Similarly, the caption set is enlarged by augmenting the most matching caption pairs for each image in the dataset, with the size of approximately five.

Metrics. We employ Attack Success Rate (ASR) as the metric for evaluating the adversarial robustness and transferability in both white-box and black-box settings. Specifically, ASR evaluates the percentage of attacks that only produce successful adversarial examples. A higher ASR indicates better adversarial transferability.

5.2. Transferability Analysis

In this paper, we present a systematic study of the adversarial transferability of VLP models, which has not been explored. As demonstrated in Section 3, existing methods, including the separate unimodal adversarial attack (Sep-Attack) and the multimodal adversarial attack (Co-Attack), exhibit limited transferability to other VLP models.

To improve transferability in multimodal learning, we intuitively investigate the adoption of transfer-based attacks from unimodal learning such as image classification. Specifically, we consider MI [7], DIM [37], and PNA_PO [36]. However, this approach can be problematic if cross-modal interactions and the unique many-to-many alignments in multimodal learning are not taken into account.

Table 2 illustrates that multimodal attack methods that incorporate transfer-based image attacks exhibit minimal improvement in transferability while compromising white-box performance. Specifically, when integrated with MI, Co-Attack drops significantly by 12.3% in white-box settings, while only maintaining 25.40% ASR in transferability (ALBEF to TCL). However, our SGA shows superior performance in both white-box and black-box settings. Notably, Sep-Attack combined with transfer-based attacks not only reduces the effectiveness of white-box attacks but also fails to improve adversarial transferability in almost all black-box settings. The results provide empirical evidence that directly combining unimodal adversarial attacks in multimodal learning without considering cross-modal interactions and alignments can be problematic, even when using transfer-based unimodal attacks. Additional discussion is provided in Appendix B.2.

5.3. Experimental Results

Multimodal Fusion Modules. First, we investigate VLP models with different fusion modules, namely, fused VLP models and aligned VLP models. We generate adversarial examples on both types of models and evaluate their attack performance when transferred to other VLP models while ensuring consistency in the input size of images. For example, adversarial images generated by ALBEF or TCL are resized to 224×224 before performing transfer attacks on CLIP, and adversarial examples generated on CLIP are resized to 384×384 before being transferred to ALBEF or TCL.

As shown in Table 3, experimental results demonstrate

Source	Attack	ALBEF		TCL		CLIP$_{\text{ViT}}$		CLIP$_{\text{CNN}}$	
		TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
ALBEF	PGD	52.45*	58.65*	3.06	6.79	8.96	13.21	10.34	14.65
	BERT-Attack	11.57*	27.46*	12.64	28.07	29.33	43.17	32.69	46.11
	Sep-Attack	65.69*	73.95*	17.60	32.95	31.17	**45.23**	32.82	45.49
	Co-Attack	77.16*	83.86*	15.21	29.49	23.60	36.48	25.12	38.89
	SGA	**97.24±0.22***	**97.28±0.15***	**45.42±0.60**	**55.25±0.06**	**33.38±0.35**	44.16±0.25	**34.93±0.99**	**46.57±0.13**
TCL	PGD	6.15	10.78	77.87*	79.48*	7.48	13.72	10.34	15.33
	BERT-Attack	11.89	26.82	14.54*	29.17*	29.69	44.49	33.46	46.07
	Sep-Attack	20.13	36.48	84.72*	86.07*	31.29	44.65	33.33	45.80
	Co-Attack	23.15	40.04	77.94*	85.59*	27.85	41.19	30.74	44.11
	SGA	**48.91±0.74**	**60.34±0.10**	**98.37±0.08***	**98.81±0.07***	**33.87±0.18**	**44.88±0.54**	**37.74±0.27**	**48.30±0.34**
CLIP$_{\text{ViT}}$	PGD	2.50	4.93	4.85	8.17	70.92*	78.61*	5.36	8.44
	BERT-Attack	9.59	22.64	11.80	25.07	28.34*	39.08*	30.40	37.43
	Sep-Attack	9.59	23.25	11.38	25.60	79.75*	86.79*	30.78	39.76
	Co-Attack	10.57	24.33	11.94	26.69	93.25*	95.86*	32.52	41.82
	SGA	**13.40±0.07**	**27.22 ±0.06**	**16.23±0.45**	**30.76±0.07**	**99.08±0.08***	**98.94±0.00***	**38.76±0.27**	**47.79±0.58**
CLIP$_{\text{CNN}}$	PGD	2.09	4.82	4.00	7.81	1.10	6.60	86.46*	92.25*
	BERT-Attack	8.86	23.27	12.33	25.48	27.12	37.44	30.40*	40.10*
	Sep-Attack	8.55	23.41	12.64	26.12	28.34	39.43	91.44*	95.44*
	Co-Attack	8.79	23.74	13.10	26.07	28.79	40.03	94.76*	96.89*
	SGA	**11.42±0.07**	**24.80±0.28**	**14.91±0.08**	**28.82±0.11**	**31.24±0.42**	**42.12±0.11**	**99.24±0.18***	**99.49±0.05***

Table 3: **Comparison with state-of-the-art methods on image-text retrieval.** We report the attack success rate (%) of R@1 for both IR and TR. The source column indicates the source models used to generate the adversarial data on Flickr30K. * indicates white-box attacks. A higher ASR indicates better adversarial transferability.

Attack	B@4	METEOR	ROUGE_L	CIDEr	SPICE
Baseline	39.7	31.0	60.0	133.3	23.8
Co-Attack	37.4	29.8	58.4	125.5	22.8
SGA	**34.8**	**28.4**	**56.3**	**116.0**	**21.4**

Table 4: **Cross-Task Transferability: ITR → IC.** Adversarial data generated from Image-Text Retrieval (**ITR**) to attack Image Captioning (**IC**) on MSCOCO. The source model and target model are ALBEF and BLIP, respectively. The Baseline represents the original performance of IC on clean data. Lower values indicate better adversarial transferability.

the superiority of our proposed SGA over existing multimodal attack methods in all black-box settings. Specifically, our SGA achieves significant improvements in adversarial transferability when the source and target models are of the same type. For instance, SGA outperforms Co-Attack by approximately 30% in terms of attack success rate when transferring adversarial data from ALBEF to TCL. Moreover, in the more challenging scenario where the source and target models are of different types, SGA also surpasses Co-Attack with higher attack success rates. More experiments on the MSCOCO dataset are provided in Appendix B.3.

Model Architectures. Then, we explore VLP models with respect to different model architectures. Many VLP models commonly use ViTs as the vision encoder, where images are segmented into patches before being processed by the transformer model. However, in the case of CLIP, the image encoder consists of two distinct architectures: conventional Convolutional Neural Networks (CNNs) and ViTs. The transferability between CNNs and ViTs has been well-studied in unimodal learning. Therefore, we also investigate the adversarial transferability of CNN-based and ViT-based CLIP models in multimodal learning.

As shown in Table 3, we observe a similar phenomenon as unimodal learning that compared to CNNs, ViTs show better robustness against adversarial perturbations [25]. Specifically, for all attack methods, the same adversarial multimodal data have a stronger white-box attack effect on CLIP$_{\text{CNN}}$ compared to CLIP$_{\text{ViT}}$. Moreover, the adversarial examples generated on CLIP$_{\text{ViT}}$ are found to be more transferable to CLIP$_{\text{CNN}}$ than vice versa (38.76% vs. 31.24%).

Furthermore, our proposed SGA consistently improves transferability on both CNN-based CLIP and ViT-based CLIP compared to other attacks. For instance, SGA increases the adversarial transferability for CLIP$_{\text{ViT}}$ by 5.83% and 6.24% compared to Co-Attack under the white-box setting and black-box setting, respectively.

5.4. Cross-Task Transferability

Cross-modal interactions and alignments are the core components of multimodal learning regardless of the task.

Attack	Val	TestA	TestB
Baseline	58.46	65.89	46.25
Co-Attack	54.26	61.80	43.81
SGA	**53.55**	**61.19**	**43.71**

Table 5: **Cross-Task Transferability: ITR → VG.** Adversarial data generated from Image-Text Retrieval (**ITR**) to attack Visual Grounding (**VG**) on RefCOCO+. The source model and target model are both ALBEF. The Baseline represents the original performance of VG on clean data. Lower values indicate better adversarial transferability.

Therefore, we conduct extensive experiments to explore the effectiveness of our proposed SGA on two additional V+L tasks: Image Captioning (IC) and Visual Grounding (VG).

Image Captioning. Image captioning is a generation-based task, where an input image is encoded into a feature vector and then decoded into a natural language sentence. In our experiments, we craft adversarial images using the source model (ALBEF) with an image-text retrieval objective and then directly attack the target model (BLIP [17]) on image captioning. We employ the MSCOCO dataset, which is suitable for both two tasks, and utilize various evaluation metrics to measure the quality of the generated captions, including BLEU [26], METEOR [2], ROUGE [21], CIDEr [32], and SPICE [1].

We present the performance on image captioning of BLIP after being attacked in Table 4. Experimental results demonstrate clear improvements in the adversarial transferability of the proposed SGA compared to Co-Attack. Specifically, our SGA improves the BLEU score by up to 2.6% and the CIDEr score by up to 9.5%.

Visual Grounding. Visual grounding is another V+L task that aims to localize the region in an image based on the corresponding specific textual description. Similarly, we generate adversarial images using the source model (ALBEF) from image-text to attack the target model (ALBEF) on visual grounding. Table 5 shows the results on RefCOCO+ [39], where our SGA still outperforms Co-Attack.

5.5. Ablation Study

To systematically investigate the impact of our set-level alignment-preserving augmentations, we conducted ablation experiments on image-text retrieval to evaluate the effect of varying the number of augmented image sets N with multi-scale transformation and the number M of augmented caption sets. Specifically, we employed ALBEF as the source model and $\mathrm{CLIP}_{\mathrm{ViT}}$ as the target model on Flickr30K. More details are provided in Appendix B.4.

Multi-scale Image Set. We propose the use of multiple scale-invariant images to generate diverse adversarial data in SGA. Results in the left of Figure 5 reveal that the transferability significantly increases as we introduce more diverse images with different scales, peaking when the scale range is set to $[0.50, 1.50]$ with a step of 0.25. We set the scale range $S = \{0.50, 0.75, 1.00, 1.25, 1.50\}$ for optimal performance.

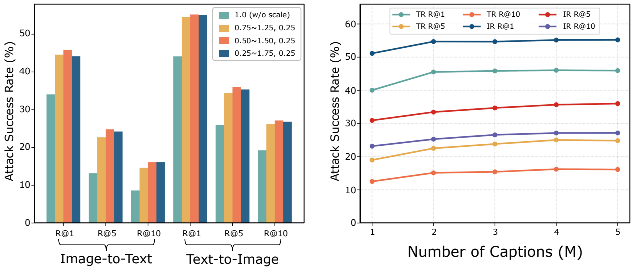


Figure 5: **Ablation Study:** Attack success rates (%) on set-level augmentations, **Left** for image set and **Right** for caption set. The performance of SGA is consistently improved by using larger augmented image and caption sets.

Multi-pair Caption Set. We also conduct experiments to investigate the impact of an enlarged caption set on adversarial transferability. The number of additional captions ranged from 1 to M, where M represents the most matching caption pairs from the dataset for each image. Results presented in the right panel of Figure 5 indicate that if $M > 1$, the black-box performance increases significantly but eventually plateaus. These results demonstrate the effectiveness of using multiple alignment-preserving inter-modal information to enhance adversarial transferability. Furthermore, we observed that the performance is relatively insensitive to the number of extra captions, but adding more captions can improve the overall adversarial transferability.

6. Conclusion

In this paper, we make the first attempt to investigate the adversarial transferability of typical VLP models. We systematically evaluate the existing attack methods and reveal that they still exhibit lower transferability, despite their impressive performance in white-box settings. Our investigation highlights the need for specially designed transferable attacks in multimodal learning that can model the many-to-many cross-modal alignments and interactions. We propose SGA, a highly transferable multimodal attack, which leverages set-level alignment-preserving augmentations through cross-modal guidance to thoroughly exploit multimodal interactions. We hope that this work could inspire further research to evaluate and enhance the adversarial robustness of VLP models.

Acknowledgments. This work was supported by the National Key R&D Program of China (Grant NO. 2022YFF1202903) and the National Natural Science Foundation of China (Grant NO. 62122035).

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 8

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*, 2005. 8

[3] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017. 3

[4] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12652–12660, 2020. 3

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 2

[6] Mengjun Cheng, Yipeng Sun, Long Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Vista: Vision and scene text aggregation for cross-modal retrieval. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5174–5183, 2022. 3

[7] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018. 3, 6

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021. 2, 6

[9] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. *ArXiv*, abs/2111.02387, 2021. 2

[10] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56, 2018. 3

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015. 3

[12] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[13] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI*, 2020. 3

[14] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:664–676, 2017. 5

[15] Zaid Khan and Yun Raymond Fu. Exploiting bert for multimodal target sentiment classification through input space translation. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 1

[16] Chenyi Lei, Shixian Luo, Yong Liu, Wanggui He, Jiamang Wang, Guoxin Wang, Haihong Tang, Chunyan Miao, and Houqiang Li. Understanding chinese video and language via contrastive multimodal pre-training. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 1

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *ArXiv*, abs/2201.12086, 2022. 2, 8

[18] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 2, 3, 6

[19] Linyang Li, Ruotian Ma, Qipeng Guo, X. Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert. *ArXiv*, abs/2004.09984, 2020. 1, 2, 3, 6

[20] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 2

[21] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*, 2004. 8

[22] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv: Learning*, 2019. 2

[23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 5

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2018. 1, 2, 3, 6

[25] Muzammal Naseer, Kanchana Ranasinghe, Salman Hameed Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *Neural Information Processing Systems*, 2021. 7

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 8

[27] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, J. Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123:74–93, 2015. 2, 5

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen

Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 6

[29] Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. Generating natural language adversarial examples through probability weighted word saliency. In *ACL*, 2019. 3

[30] Lei Shi, Kai Shuang, Shijie Geng, Peng Gao, Zuohui Fu, Gerard de Melo, Yunpeng Chen, and Sen Su. Dense contrastive visual-linguistic pretraining. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 1

[31] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herv'e J'egou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 2

[32] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 8

[33] Teng Wang, Yixiao Ge, Feng Zheng, Ran Cheng, Ying Shan, Xiaohu Qie, and Ping Luo. Accelerating vision-language pretraining with free language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23161–23170, 2023. 2

[34] Teng Wang, Wenhao Jiang, Zhichao Lu, Feng Zheng, Ran Cheng, Chengguo Yin, and Ping Luo. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix. In *International Conference on Machine Learning*, pages 22680–22690. PMLR, 2022. 2

[35] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5763–5772, 2019. 3

[36] Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards transferable adversarial attacks on vision transformers. *ArXiv*, abs/2109.04176, 2021. 6

[37] Cihang Xie, Zhishuai Zhang, Jianyu Wang, Yuyin Zhou, Zhou Ren, and Alan Loddon Yuille. Improving transferability of adversarial examples with input diversity. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2725–2734, 2018. 3, 6

[38] Jinyu Yang, Jiali Duan, S. Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul M. Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. *ArXiv*, abs/2202.10401, 2022. 2, 3, 6

[39] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. *ArXiv*, abs/1608.00272, 2016. 8

[40] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis E. H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 538–547, 2021. 2

[41] Jiaming Zhang, Qiaomin Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 1, 2, 3

[42] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5575–5584, 2021. 2

[43] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and S. Li. Context-aware attention network for image-text retrieval. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3533–3542, 2020. 3