Zero-Shot Attribute Attacks on Fine-Grained Recognition Models

Nasim Shafiee and Ehsan Elhamifar

Northeastern University, Boston, USA
{shafiee.n,e.elhamifar}@northeastern.edu

Abstract. Zero-shot fine-grained recognition is an important classification task, whose goal is to recognize visually very similar classes, including the ones without training images. Despite recent advances on the development of zero-shot fine-grained recognition methods, the robustness of such models to adversarial attacks is not well understood. On the other hand, adversarial attacks have been widely studied for conventional classification with visually distinct classes. Such attacks, in particular, universal perturbations that are class-agnostic and ideally should generalize to unseen classes, however, cannot leverage or capture small distinctions among fine-grained classes. Therefore, we propose a compositional attribute-based framework for generating adversarial attacks on zero-shot fine-grained recognition models. To generate attacks that capture small differences between fine-grained classes, generalize well to previously unseen classes and can be applied in real-time, we propose to learn and compose multiple attribute-based universal perturbations (AUPs). Each AUP corresponds to an image-agnostic perturbation on a specific attribute. To build our attack, we compose AUPs with weights obtained by learning a class-attribute compatibility function. To learn the AUPs and the parameters of our model, we minimize a loss, consisting of a ranking loss and a novel utility loss, which ensures AUPs are effectively learned and utilized. By extensive experiments on three datasets for zero-shot fine-grained recognition, we show that our attacks outperform conventional universal classification attacks and transfer well between different recognition architectures.

Keywords: Fine-grained recognition, Zero-shot models, Adversarial attacks, Attribute-based universal perturbations, Compositional model

1 Introduction

Despite the tremendous success of Deep Neural Networks (DNNs) for image recognition, DNNs have been shown to be vulnerable to attacks [8, 24]. Adversarial attacks are imperceptible image perturbations that result in incorrect prediction with high confidence and have highlighted the lack of robustness of DNNs. This has motivated a large body of research on generating small perturbations [8, 11, 16, 24, 28, 36, 49, 53, 58, 64, 68, 74, 81, 94], and subsequently using the attacks to design robust defense mechanisms, e.g., by detecting attacks or retraining the model using perturbed images. Motivated by the fact that gen-

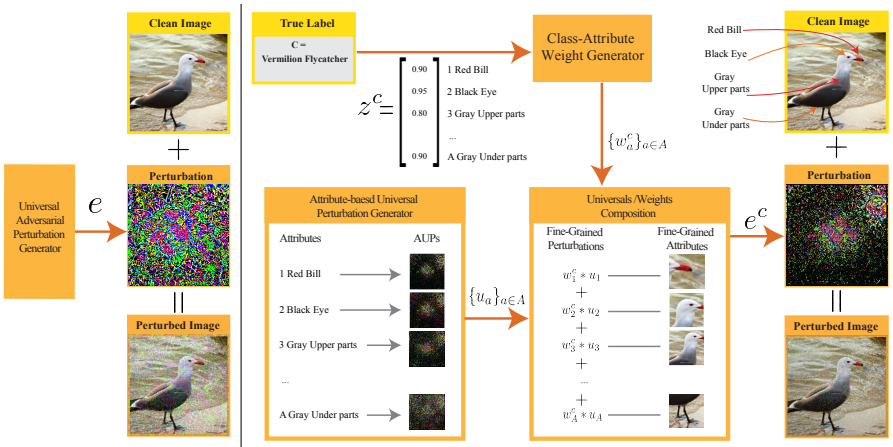


Fig. 1: (a) Traditional UAP [27, 57] generates one holistic perturbation for all classes, hence, does not efficiently capture fine-grained discriminative visual information. (b) We develop a compositional framework for generating robust and generalizable attacks for zero-shot fine-grained recognition. We generate attribute-based universal perturbations (AUPs) and learn to compose them for attacking fine-grained classes.

erating image-specific attacks are costly, especially when dealing with a large number of images, many works have studied finding universal attacks: image-agnostic perturbations that can change the ground-truth class of an arbitrary input image [37, 46, 56, 57, 59, 60, 70, 96], hence, ideally should generalize to previously unseen classes. The majority of existing works, however, have focused on coarse-level classification in which classes have wide visual appearance variations with respect to each other, e.g., 'person', 'car', 'building', etc. Zero-shot fine-grained recognition, on the other hand, is an important classification problem in which one has to distinguish visually very similar classes, e.g., clothing types [2, 50, 79], faces [51, 65, 84] or bird/plant species [14, 19, 44, 101, 102, 105], including classes without training images. The majority of successful methods for zero-shot fine-grained recognition have focused on identifying and leveraging small distinctions between classes using feature pooling [23,38,44], discriminative region localization [2, 14, 50, 79, 101, 102] and dense attribute attentions [31, 32]. This is done by describing each fine-grained class by a *class semantic vector* consisting of different attributes, such as 'bill color', 'belly pattern', 'wing color' for classification of birds. Despite its importance, however, robustness of (zero-shot) fine-grained recognition models to adversarial attacks has not received much attention in the literature.

As we show in the paper, conventional adversarial attacks for coarse-level classification do not work well for fine-grained recognition. This comes from the fact that generating holistic perturbations using existing methods fails to capture and leverage small distinctions between a specific class and others. Another lim-

itation of existing works is the lack of a principled method for crafting effective adversarial attacks for unseen classes that do not have training images.

Paper Contributions. We propose a new framework for generating effective and generalizable attacks on zero-shot fine-grained recognition models. To generate adversarial attacks that capture small differences between fine-grained classes, generalize well to unseen classes without training images, and can be applied in real-time, we develop a compositional attribute-based model. Leveraging class semantic vectors, we learn multiple attribute-based universal perturbations (AUPs), where each AUP corresponds to an image-agnostic perturbation on a specific attribute. To build a class-wise universal perturbation that changes the class of an input image, we propose to compose the AUPs with appropriate weights, which are obtained by learning a class-attribute compatibility function. To learn our AUPs and the parameters of the compositional weight generator, we minimize a ranking loss regularized by norms of AUPs, to keep them and the final perturbation imperceptible, as well as a novel utility loss, which ensures all AUPs are effectively trained and used. By extensive experiments on three datasets for zero-shot fine-grained recognition, we show that our method outperforms conventional classification attacks, with only half of the perturbation norm, and also can transfer well between different architectures.

2 Related Works

Adversarial Attacks. Adversarial attacks have been mainly studied for classification problems and can be categorized as targeted or non-targeted [97]. Targeted attacks perturb an image to change its ground-truth class into a desired secondary class [5, 8, 39], while non-targeted attacks focus on only altering the ground-truth class [16, 24, 58, 78]. From a security perspective, attacks can have different access levels to the model whose output needs to be modified. In the white-box attack, the adversary has access to the parameters of the target model [8, 16, 24, 39, 58], while, in the black-box attack, there is no access to the parameters of the target model [62, 90]. Earlier works proposed generating attacks using gradient-based and optimization-based approaches. Gradient-based schemes maximize the cross-entropy loss function to find adversarial examples [16,24,40,73]. Also, optimization-based methods obtain adversarial examples by solving a constrained optimization problem [8, 24, 40, 58, 72, 78].

Perturbations generated by adversarial attacks can be either image-specific or image-agnostic. While image-specific attacks generate a specific perturbation for each input image, image-agnostic attacks generate a unique perturbation, referred to as a universal adversarial perturbation, for all input images [4,57,70]. Another category of attacks focuses on semantic and perceptual properties of visual features to preserve the image concept while perturbing it. Therefore, [1, 6, 12, 20, 29, 41, 91] focused on generating a perturbation through geometric transformations, global color shifts, image filters. These approaches can craft more realistic adversarial examples, but their restricted perturbation space limits their strength. [13, 17, 25, 52, 55, 82, 82, 85, 98, 104] find perturbations that can

be hidden in the texture and regions of the image with drastic visual variations. Recently, [80] have decomposed perturbations into independent components and investigated the attribute to which each component belongs. Also, [61] has shown that maximally separating the discriminative latent features of classes can improve the robustness of fine-grained models.

Fine-Grained Recognition. DNNs have achieved significant improvement on fine-grained recognition, where the challenge lies in recognizing different classes with small but distinct visual appearance variations. To detect interactions between discriminative feature maps, [18, 79, 100] have employed pooling methods. [43, 101] and [14, 77, 105] have used part-based and weak supervision for localizing discriminative parts of an image, respectively. On the other hand, instead of considering all discriminative features with the same importance for recognition, [15, 30, 34, 45, 99, 106] have employed several attention mechanisms for better extraction of more critical discriminative visual features. To better capture relationships and interactions between discriminative features, [9, 103] have proposed to employ graph networks and multi-granularity label prediction, respectively. Despite advances in fine-grained recognition, many previous models cannot generalize to zero and few-shot learning, which we review next.

Zero-Shot Learning. The goal of zero-shot learning is to transfer the knowledge a model can gain from images of seen classes for recognition of unseen classes, given a shared semantic space for both seen and unseen classes [54, 95]. [76] proposed a class-wise normalization technique to maintain the variance of seen and unseen classes. [10, 32, 35, 48, 71] have learned to find an alignment between visual features and semantic vectors in an embedding space. [21, 75, 86, 88] have proposed to generate synthesized samples for unseen classes and reformulated the zero-shot problem as a fully supervised setting. Although the generative methods have improved the unseen classification performance, for fine-grained recognition, they often cannot generate enough discriminative unseen features, which has motivated recent works in [26, 31, 33]. Specifically, [26] proposed a hybrid generative and discriminative framework for fine-grained recognition, while [31] developed a compositional generative model and [33] used a few samples of unseen classes in training to generate more discriminative seen/unseen samples.

3 Fine-Grained Compositional Adversarial Attacks

3.1 Problem Setting

Assume we have two sets of fine-grained classes $\mathcal{C}_s$ and $\mathcal{C}_u$, where $\mathcal{C}_s$ denotes seen classes that have training images, $\mathcal{C}_u$ denotes unseen classes without training images and $\mathcal{C} \triangleq \mathcal{C}_s \cup \mathcal{C}_u$ denotes the set of all classes. Let $(I_1, y_1), \ldots, (I_N, y_N)$ be N training samples, where I_i is the i-th training image and $y_i \in \mathcal{C}_s$ corresponds to its fine-grained class. We denote the set of all training images by $\mathcal{I} = \{I_1, \ldots, I_N\}$. Let $\{\boldsymbol{z}^c\}_{c \in \mathcal{C}}$ denote *class semantic* vectors that provide descriptions of classes. More specifically, $\boldsymbol{z}^c = [z_1^c, \ldots, z_A^c]^\top$ is the semantic vector of the class c with A attributes, where z_a^c is the score of having the a-th attribute in the class c [3, 7, 42, 63, 71, 87]. Also, let $\{\boldsymbol{v}_a\}_{a=1}^A$ denote *attribute semantic* vectors, where $\boldsymbol{v}_a$ is the average of GloVe representation [67] of the

words describing attribute a, e.g., 'red beak'. Assume we have learned a zero-shot fine-grained classifier, using training samples from seen classes, which can classify a test image from a seen or an unseen class. Our goal is to generate a fine-grained adversarial perturbation for an image from a seen or an unseen class that results in misclassification.

As we show, our proposed compositional method for generating adversarial attacks *work with any zero-shot fine-grained recognition model*. In the paper, *we focus on the non-targeted attacks*, i.e., aim to misclassify an image without specifying the adversarial class. However, our formulation can be modified to the targeted setting, which we leave for future studies.

3.2 Compositional Attribute-based Universal Perturbations (CAUPs)

To generate an attack on an image I that belongs to a class y, we need to find a perturbation $\boldsymbol{e}$ that results in a higher prediction score for a class $c \neq y$ on the image. More specifically, our goal is to find an attack $\boldsymbol{e}$ such that

$$\exists\, c \in \mathcal{C} \backslash y \quad \text{s.t.} \quad s^c(I + \boldsymbol{e}) > s^y(I + \boldsymbol{e}). \tag{1}$$

There are *two conventional ways* of finding the attack. *i)* Generate an image-specific perturbation. However, this requires significant computation per image (more drastically if we want to misclassify multiple images, e.g., video frames) and does not allow generating attacks in real-time. *ii)* Generate a single universal perturbation that can change the class of any image. However, a single universal perturbation cannot incorporate small differences between every pair of fine-grained classes (as we show in the experiments, it does not work well).

Proposed Compositional Model. To obtain adversarial attacks that can be generated in real-time, capture small differences between fine-grained classes (where often only a few attributes of any two classes are different) and generalize well to unseen classes, we propose a compositional model. First, for each attribute a, we learn an attribute-based universal perturbation $\boldsymbol{u}_a$, which has the same size as the input image. We refer to $\{\boldsymbol{u}_a\}_{a=1}^{A}$ as attribute-based universal perturbations (AUPs). We compose the AUPs with learnable weights to build class-wise universal perturbations, denoted by $\{\boldsymbol{e}^c\}_{c=1}^{|\mathcal{C}|}$. More specifically, we propose

$$\boldsymbol{e}^c = \sum_{a=1}^{A} \omega_a^c\, \boldsymbol{u}_a, \quad \forall c \in \mathcal{C}, \tag{2}$$

in which we generate the universal attack on class c using linear combination of AUPs with weights $\{\omega_1^c, \ldots, \omega_A^c\}$. *Unlike prior works, instead of generating one attack per image or one attack for all images from all classes, we generate $|\mathcal{C}|$ universal perturbations, one per class.* Additionally, our class-wise attack is a composition of AUPs, which allows us to generate attacks for a seen or an unseen class as we show below. Given that the adversary knows the ground-truth class y of the image or obtains it from the output of the fine-grained classifier, it can attack the image using $\boldsymbol{e}^y$.

An important question is how to choose/find the composition weights $\{\omega_a^c\}$ that allows finding class-wise universal perturbations even for unseen classes. We propose to find the compositional weights using class-attribute compatibility

$$\omega_a^c = \tanh(\boldsymbol{v}_a^\top \boldsymbol{W}_a \boldsymbol{z}^c) \in [-1, +1], \tag{3}$$

where the term inside the hyperbolic tangent measures the compatibility between the attribute a and class c using a learnable matrix $\boldsymbol{W}_a$. We use GloVe vectors for $\boldsymbol{v}_a$'s to generate ω_a^c's that reflect similarity of attributes. Using tanh both normalizes the composition weights and assigns positive/negative values to them with the goal of perturbing visual features by adding/removing attributes.[1] Notice that using (3), once we learn $\boldsymbol{W}_a$'s from some training images, we can generate compositional weights for both seen and unseen classes, hence, obtain class-wise universal perturbations for any $c \in \mathcal{C}$ using (2).

Remark 1. In the experiments, we show that our compositional model works significantly better than learning a single universal perturbation for all classes. We also show that using compositional weights as in (3) works better than combining AUPs with uniform weights.

Remark 2. Our method learns attribute-based universal perturbations and combines them to produce class-wise perturbations. This is different from UAP [57] and GAP [27] that generate a single class-agnostic universal perturbation.

3.3 Learning AUPs

Our goal is to learn $\{\boldsymbol{u}_a, \boldsymbol{W}_a\}_{a=1}^A$ from a set of training images, which correspond to samples from seen classes. As stated before, for a training image I from class y, we want to find $\boldsymbol{e}^y$ so that (1) is satisfied. Therefore, inspired by [8], we first use the ranking loss function,

$$\mathcal{L}_{rank} = \sum_{I \in \mathcal{I}} \max\{0, \delta + s^y(I + \boldsymbol{e}^y) - \max_{c \neq y} s^c(I + \boldsymbol{e}^y)\}, \tag{4}$$

whose minimization ensures that for each training image I, there exists a non-ground truth class ($c \neq y$) that obtains a higher score (by a margin $\delta > 0$) on the perturbed image than the ground truth class y. Notice that the ranking loss in [8] only optimizes image-specific perturbations, while our proposed method searches for image-agnostic perturbations. We use the ranking loss instead of the cross entropy loss since specifying the margin δ allows computing stronger perturbations (also empirically it works better). For our CAUP attack based on ℓ_2-norm and ℓ_∞-norm, we also use the following regularization loss

$$\mathcal{L}_{reg} = \begin{cases} \sum_{a=1}^A \|\boldsymbol{u}_a\|_2^2, & \ell_2 \ \ attack \\ \sum_{a=1}^A \sum_j (|u_{a,j}| - k)^+, & \ell_\infty \ attack \end{cases} \tag{5}$$

where the first (ℓ_2) regularization ensures that entries of AUPs are sufficiently small, while the second (ℓ_∞-type) regularization penalizes $\boldsymbol{u}_a$ entires that are more than a constant k and benefits from being differentiable.

[1] Empirically, we obtained better results using tanh than other activation functions.

Notice that by only minimizing these two losses, it is possible to never use some of the AUPs (i.e., for some attributes a, the weights ω_a^c will be zero across all classes), which is undesired. To prevent this, we define the *utility of each attribute* a as $\tau_a \triangleq \sum_c |\omega_{a,c}|$, which is the sum of the absolute compositional weight of using the attribute a to attack the class c. We further regularize learning AUPs by minimizing the utility loss,

$$\mathcal{L}_{util} = \sum_{a=1}^{A} \tau_a^2, \tag{6}$$

which ensures nonzero values for all attribute utilities, hence, each attribute will be used for generating attacks for at least some classes. As a result, to learn the AUPs and the composition weights $\{\boldsymbol{u}_a, \boldsymbol{W}_a\}_{a=1}^{A}$, we propose to minimize

$$\mathcal{L}_{rank} + \lambda_{reg}\mathcal{L}_{reg} + \lambda_{util}\mathcal{L}_{util}, \tag{7}$$

where $\lambda_{reg}, \lambda_{util} \geq 0$ are hyperparameters. We use stochastic gradient descent to optimize our proposed loss.

4 Experiments

We evaluate the performance of our proposed compositional attribute-based universal perturbations (CAUP) on fine-grained datasets. We show that CAUP works well for different fine-grained recognition architectures and also our attacks transfer well across architectures. We also demonstrate that CAUP works better than conventional universal attacks and that composition of attribute-based perturbations is important for generating effective attacks. We investigate the effectiveness of different components of our method and present qualitative results illustrating properties of our attacks.

Table 1: Statistics of the datasets used in our experiments.

Dataset	# attributes	# seen (val) / unseen classes	# training / testing samples
CUB	312	100 (50) / 50	7,057 / 4,731
AWA2	85	27 (13) / 10	23,527 / 13,795
SUN	102	580 (65) / 72	10,320 / 4,020

4.1 Experimental Setup

Datasets. We use three popular zero-shot fine-grained recognition datasets: Caltech-UCSD Birds-200-2011(CUB) [83], Animal with Attributes (AWA2) [89] and SUN Attribute (SUN) [66, 92]. Table 1 shows the statistics of the three datasets. CUB [83] contains 11,788 total images from fine-grained bird-species with 150 seen and 50 unseen classes. AWA2 [89] is an animal classification dataset with 40 seen and 10 unseen classes and has 37,322 samples in total. SUN [66] consists of different visual scenes with 14,340 images from 645 seen and 72 unseen classes. It has the largest number of classes among the datasets. However, it only contains 16 training images per class due to its small overall training set. Notice that these datasets include class semantic vectors, hence they are suitable for our zero-shot fine-grained attack model. For CUB, AWA2 and SUN, we follow the same training, validation and testing splits as in prior works [32, 87].

Evaluation Metrics. Following prior works on zero-shot fine-grained recognition [26, 32, 47, 76, 89], we only apply the attacks on images that the model correctly classfies and measure the top-1 accuracy of a classifier on these perturbed test images. We consider the challenging generalized zero-shot setting, in which test samples come from both seen and unseen classes (as opposed to the zero-shot setting, where test images come only from unseen classes). We report the fooling percentage on attacked testing images from seen classes, $fool_s = (1 - acc_s) * 100$, and from unseen classes, $fool_u = (1 - acc_u) * 100$. Since universal perturbations generated by different methods and across architectures can have different magnitudes, we ℓ_2 normalize the generated perturbations and, similar to prior works [8, 53, 57], report the performance as a function of the scaling of the perturbation magnitude.

Baselines. We investigate the effectiveness of our proposed attacks on four recent fine-grained recognition models: DAZLE [32], DCN [47], CNZSL [76] and CEZSL [26]. We also study the transferability properties of our attacks across these models. DAZLE is a discriminative fine-grained recognition model that extracts and uses dense attribute features for classification. It uses the class semantic vectors and attribute embeddings to learn A attention models for A attributes and to compute the final class score. On the other hand, DCN and CNZSL, are two discriminative methods that extract a holistic feature from an input image to learn a compatibility function between images and class semantic vectors. The generative CEZSL model is trained to capture the distribution of the images and their attributes. It augments the seen image features with both seen and unseen synthetically features to train a standard classifier. We chose these fine-grained architectures since they can handle the zero-shot setting, on which we can test our zero-shot fine-grained attack.

We compare CAUP with several baseline attacks: i) Universal Adversarial Perturbation (UAP) [57], which learns a single perturbation template. ii) Generative Adversarial Perturbation (GAP) [69], which is a network that generates a single universal adversarial perturbation. We trained both UAP and GAP on seen classes and test them on seen/unseen classes. iii) Uniform, where we simply combine all the AUPs, $\{\boldsymbol{u}_a\}_{a=1}^{A}$, with uniform weights of $1/A$ in (2). This allows us to investigate the effectiveness of our composition weights. iv) Random, where we randomly generate a perturbation vector whose entries come from a standard Normal distribution. Notice that since our attack is image-agnostic, we assume the attacker cannot eavesdrops the input, hence *the results are not comparable to image-specific attacks* [16, 22, 58, 93].

Implementation Details. We attached RESNET backbone to each fine-grained model and retrained them to achieve similar performance as they reported in their paper. For all models, we follow the exact experimental setting as in their reported work [26, 32, 47, 76]. *We compute all attacks in the image-space and not in the feature-space*, which is conventional in the fine-grained recognition. We perform the experiments in the generalized zero-shot setting, where testing images come from both seen and unseen classes. For efficiency, similar to [57],

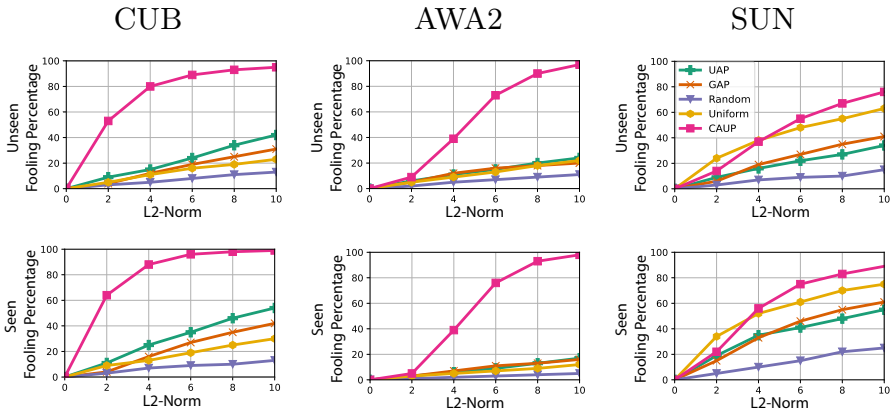


Fig. 2: Unseen (top) and seen (bottom row) fooling percentage of different universal attacks on DAZLE as a function of the magnitude (ℓ_2-norm) of the perturbation.

we randomly use one-third of the training images to generate different types of universal attacks. In our experiments, using more samples for training did not change the performances. We implement all attacks in Pytorch on a server running Ubuntu 18.04 with an Intel Xeon Gold CPU and eight NVIDIA Quadro RTX 6000 GPUs and optimized with the default setting of ADAM optimizer with 0.001 learning rate and batch size of 50. To avoid overfitting, we employ an early stopping method with the patience of 20 (one average it stops at the 10-th epoch). To select λ_{reg} and λ_{util} in our method, we perform hyperparameter tuning over the validation sets. The optimal hyperparameter values of our attack are reported in the supplementary materials.

4.2 Experimental Results

4.2.1 Quantitative Analysis

Effectiveness of Compositionality and Attribute-based Attacks. We start by comparing the effectiveness of the CAUP against other possible universal attacks, which are not attribute-based or not truly compositional. To do so, we learn attacks by different methods for the DAZLE model on the three datasets. Figure 2 shows the unseen fooling percentage(top row) and seen fooling percentage(bottom row) as a function of the magnitude (ℓ_2-norm) of the perturbation (due to space limitation, the results for the harmonic mean and the ℓ_∞-norm are reported in the supplementary materials). Notice that on all datasets, CAUP achieves higher fooling percentage (hence, more effective attack) for different values of the perturbation magnitude. In particular, on CUB, with ℓ_2-norm of perturbation being 2, CAUP increases the fooling percentage to more than 50% while other attacks only increase the fooling percentage to about 15%. We also obtain a similar 50% gap between CAUP and (UAP,GAP) fooling percentages for ℓ_2-norm of perturbation being 6 on CUB and AWA2. As expected, when the perturbation norm increases, all attacks perform bet-

Table 2: Fooling Percentage (seen/unseen) of our CAUP attack (perturbation ℓ_2-norm of 6) on fine-grained recognition models on test images of three datasets.

Fooling Percentage	CUB			AWA2			SUN		
Seen/Unseen	UAP	GAP	CAUP	UAP	GAP	CAUP	UAP	GAP	CAUP
DAZLE	35/ 24	27/ 19	**96/ 89**	09/ 15	11/ 16	**89/ 76**	41/ 22	46/ 27	**75/ 55**
DCN	51/ 46	21/ 21	**70/ 70**	02/ 05	07/ 10	**08/ 15**	41/ 21	30/ 15	**59/ 33**
CNZSL	26/ 25	18/ 21	**96/ 91**	13/ 12	09/ 12	**54/ 42**	**29/ 16**	27/ 20	22/ 12
CEZSL	39/ 40	35/ 35	**99/ 95**	24/ 21	12/ 15	**81/ 75**	29/ 28	26/ 26	**92/ 89**

Table 3: Fooling Percentage (seen/unseen) of our CAUP attack (perturbation ℓ_∞-norm of 0.06) on fine-grained recognition models on test images of three datasets.

Fooling Percentage	CUB			AWA2			SUN		
Seen/Unseen	UAP	GAP	CAUP	UAP	GAP	CAUP	UAP	GAP	CAUP
DAZLE	14/ 11	89/ 77	**98/ 91**	04/ 08	43/ 38	**82/ 77**	21/ 10	**90/ 78**	85/ 71
DCN	05/ 06	**74/ 73**	70/ 66	01/ 02	15/ 19	**47/ 48**	09/ 05	**81/ 63**	66/ 41
CNZSL	16/ 18	61/ 55	**97/ 93**	06/ 08	57/ 73	**86/ 79**	20/ 11	**75/ 55**	48/ 19
CEZSL	18/ 24	79/ 77	**99/ 95**	02/ 04	55/ 43	**97/ 96**	26/ 15	81/ 72	**92/ 86**

ter. While on SUN, CAUP still performs best, the gap is smaller than CUB and AWA2. This comes from the fact that some of the attributes in SUN are abstract concepts, such as research/vacation, which are harder to visually attack, while CUB and AWA2 have physical attributes, such as red wing/spotted belly and gray/stripes. Notice that Uniform in all datasets underperforms CAUP which demonstrates that attribute-based attacks without properly composing them with appropriate weights are not effective. In fact, using compositional weights (3), by weighting attribute-based universal perturbations based on the class to be attacked, significantly improves the efficacy of the perturbations.

More generally, Table 2 shows the fooling percentage of UAP, GAP and CAUP ℓ_2-norm attacks on four fine-grained models on perturbed test images from seen/unseen classes. The results are for ℓ_2-norm of the perturbation being 6. Each box is an attack that is trained and tested on the corresponding fine-grained model. Notice that in almost all cases, for both seen and unseen classes and all three datasets, CAUP extremely outperforms UAP and GAP, which shows the effectiveness of the compositional model not only for attacking seen classes but for generalization to classes without any training images. In particular, on CUB, our attack improves over UAP and GAP by at least 19% on seen and 24% on unseen images. On AWA2 and SUN, DAZLE and CEZSL are easier architectures to attack while DCN and CNZSL are harder to fool. Even for these robust models, CAUP outperforms on fooling percentage for 3 out of four cases. Also, on SUN, the effectiveness of three attack strategies is lower than the other two datasets. We believe this is due to the very small number of training samples (16) per class, which makes learning perturbations, in particular, multiple attribute-based attacks, more difficult. Despite this difficulty, CAUP performs significantly better on SUN when using DCN, CEZSL and DAZLE, while UAP performs only 7% better than CAUP on DCN. Similar to Table 2, Table 3 includes the fooling percentage of UAP, GAP and CAUP ℓ_∞-norm attacks on four

Table 4: **a)** Transferability (seen/unseen fooling percentage) of CAUP attacks across different fine-grained models. The ℓ_2-norm of perturbation is 6 for all cases. **b)** Ablation study to investigate the effectiveness of different loss functions on the DAZLE.

(a)

Attack \Train		DAZLE	CEZSL	CNZSL	DCN
CUB	DAZLE	96/ 89	97/ 91	89/ 83	74/ 61
	CEZSL	84/ 81	99/ 95	92/ 93	77/ 70
	CNZSL	86/ 81	96/ 94	96/ 91	79/ 66
	DCN	81/ 82	97/ 96	88/ 90	70/ 70
AWA2	DAZLE	89/ 76	66/ 55	87/ 89	20/ 19
	CEZSL	73/ 63	81/ 75	50/ 52	37/ 28
	CNZSL	54/ 56	63/ 39	54/ 42	23/ 18
	DCN	47/ 52	62/ 49	32/ 28	08/ 15
SUN	DAZLE	75/ 55	84/ 65	23/ 16	59/ 42
	CEZSL	59/ 55	92/ 89	17/ 18	53/ 42
	CNZSL	66/ 46	85/ 65	22/ 12	58/ 37
	DCN	61/ 44	89/ 80	18/ 09	59/ 33
	AVG	67/ 62	**82/ 70**	55/ 53	53/ 42

(b)

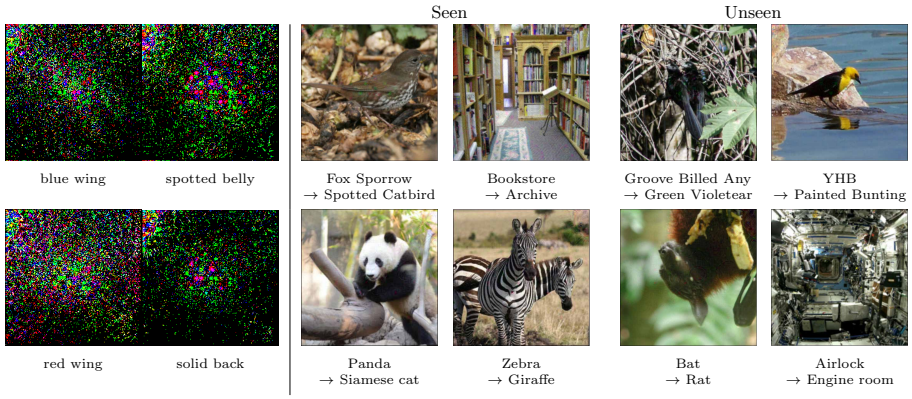
Seen/Unseen		ℓ_2
CUB	$\mathcal{L}_{rank}$	0.24/0.18
	$\mathcal{L}_{rank} + \mathcal{L}_{reg}$	0.53/0.40
	$\mathcal{L}_{rank} + \mathcal{L}_{util}$	0.90/0.82
	$\mathcal{L}_{rank} + \mathcal{L}_{reg} + \mathcal{L}_{util}$	**0.96/0.89**
AWA2	$\mathcal{L}_{rank}$	0.05/0.11
	$\mathcal{L}_{rank} + \mathcal{L}_{reg}$	0.18/0.21
	$\mathcal{L}_{rank} + \mathcal{L}_{util}$	0.36/0.33
	$\mathcal{L}_{rank} + \mathcal{L}_{reg} + \mathcal{L}_{util}$	**0.76/0.73**
SUN	$\mathcal{L}_{rank}$	0.38/0.22
	$\mathcal{L}_{rank} + \mathcal{L}_{reg}$	0.51/0.31
	$\mathcal{L}_{rank} + \mathcal{L}_{util}$	0.30/0.26
	$\mathcal{L}_{rank} + \mathcal{L}_{reg} + \mathcal{L}_{util}$	**0.75/0.55**

fine-grained models for both seen and unseen classes. The perturbations applied with scaled ℓ_∞-norm of 0.06. CAUP outperforms on eight out of twelve cases while GAP outperforms on the other four cases. In general DCN is a more robust model against all attacks, while DAZLE, CEZSL and CNZSL are easier to fool. Notice that GAP outperforms UAP in ℓ_∞-norm although it underperforms UAP and CAUP in ℓ_2-norm attack.

Limitation. Notice that to generate our CAUP attacks, we assumed access to class-semantic vectors. While in the datasets above, class semantics are available, some fine-grained recognition problems may not have such vectors. This, in fact, could be a limitation of our method technique, and extension to settings without semantic vectors is an interesting avenue of future research.

Transferability of CAUP across Recognition Models. Table 4a shows the transferability of our proposed CAUP attacks learned by one fine-grained model to other models. For each row, we learn the attribute-based universal perturbations and compositional weights using the indicated fine-grained recognition method and test the accuracy of each model against test data perturbed by our CAUP. Notice that, as expected, in almost all cases, the attack generated by each model works best on that model compared to others. On the other hand, CAUP attacks transfer quite well across different architectures. For example, on CUB, our attacks learned from CEZSL lead to $(97\%, 91\%)$ fooling percentage for (seen, unseen) using DAZLE, while leading to $(96\%, 94\%)$ fooling percentage using CNZSL and $(97\%, 96\%)$ fooling percentage using DCN. Similarly, the performance of CAUP attacks learned from DCN, CNZSL or DAZLE transfer well to other models. As the average transfer rate shows, the attacks learned on CEZSL transfer better than other attacks to other fine-grained models.

Ablation Study on Objective Function. In Table 4b, we investigate the effectiveness of each term of our proposed loss in (7) when using DAZLE as the fine-grained recognition model. We show the seen and unseen fooling percentages over all the three datasets, for perturbation magnitude being 6. Notice that only



a) AUPS **b) Adversarial Examples generated by CAUP**

Fig. 3: **a)** AUPs on CUB. The first/second column shows the perturbation for different wing colors and belly/back patterns, respectively. **b)** Successful adversarial examples generated with CAUP (perturbation norm of 6). We use the convention (ground-truth class → adversarial class). YHB means Yellow Headed Blackbird.

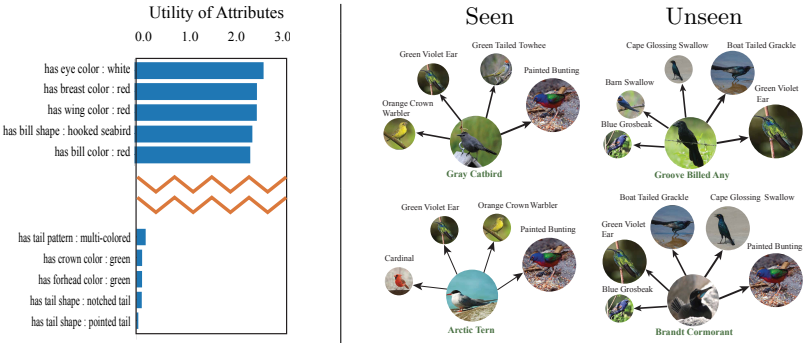
using $\mathcal{L}_{rank}$ does not lead to very effective attacks. Once we add $\mathcal{L}_{reg}$ or $\mathcal{L}_{util}$ to the ranking loss, the attack effectiveness improves. Generally, using $\mathcal{L}_{util}$ lead to more improvement compared to $\mathcal{L}_{reg}$. This shows the importance of ensuring that every attribute-based universal perturbation must be used across some classes. Finally, adding both $\mathcal{L}_{reg}$ and $\mathcal{L}_{util}$ to the ranking loss leads to significant improvement of the attack efficacy.

4.2.2 Qualitative Analysis

Adversarial Examples. In Figure 3, we show several examples of perturbed test images by our method that successfully fool the fine-grained DAZLE classifier. Notice that CAUP attacks are imperceptible or quasi-perceptible, and can not only work well for test images from seen classes but also from unseen classes. This is done by composing different attribute-based universal perturbations using class-attribute weights ω_a^c that specify how much to weight an attribute perturbation $\boldsymbol{u}_a$ for attacking a class c.

In Figure 5, we show a class and the most important attributes for changing the class label. In other words, we show the attributes with the largest composition weights ω_a^c, where thicker edges mean larger absolute weight, hence, more contribution from the associated attribute perturbation. Notice that our method automatically learns to give higher composition weights to 'purple forehead', 'purple crown' and 'red breast' attribute perturbations in order to misclassify the 'painted bunting' class or gives higher composition weights to 'red wing' and 'orange wing' to misclassify the 'red-winged black bird' class, which intuitively are also the most discriminating features of these classes.

Correlation of Attribute-based Universal Perturbations. To demonstrate the similarity among learned AUPs, in Figure 5, we show the cosine similar-



a) Utility of Attributes b) Associated Adversarial Classes

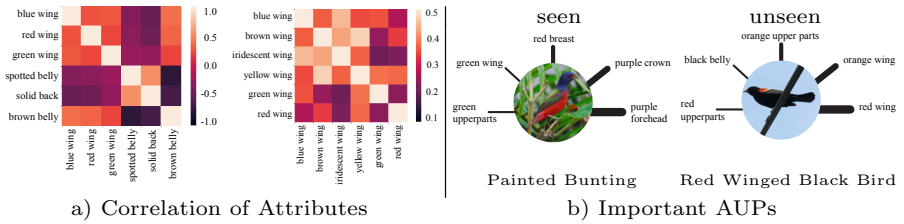
Fig. 4: a) Sorted utilities (5 largest and 5 smallest ones are shown) of AUPs trained for DAZLE on the CUB dataset. b)Visualization of four ground-truth classes and the associated adversarial classes on CUB. The node at the center of each component is a ground-truth class (green) and the surrounding nodes are the adversarial classes (black) after attack. The size of each adversarial class node represents the portion of samples that have been fooled into to that class by CAUP.

ities between pairs of AUPs learned on the CUB dataset. Notice that as the block-diagonal structure in the first plot shows, our method learns more similar AUPs for similar attributes (e.g., among wing colors) and more dissimilar AUPs for semantically different attributes (e.g., wing colors and belly/back patterns). On the other hand, as the second plot shows, within different color wings, the learned AUPs for more semantically related colors are more similar (e.g., 'brown wing' and 'yellow wing'), while being more distinct for distinct colors (e.g., 'blue wing' and 'red wing'). In Figure 3, for several attributes, we show the learned perturbations, which clearly have different patterns.

Utilities of AUPs. In Figure 4, we show the utilities of AUPs, i.e., how much each attribute contributes to misclassifying some classes. We show the top 5 attributes with the highest utilities, defined in (6), and the top 5 attributes with the lowest utilities. Notice that all attributes contribute for attacking some classes, as they have non-zero utilities. In addition, the utilities of attributes are not uniform, which means that our method successfully learns to use more discriminative attributes for generating attacks. For example, 'has tail shape: pointed tail' is very common across different bird species, therefore the associated AUP has a small utility, while 'has breast/wing color: red' are more specific to some bird species, therefore, they have higher utilities for attack (we can change many classes that do not have red breast/wing, by attacking these attributes).

Dominancy in Fine-Grained Classes. Given that our attacks are non-targeted, we investigate how CAUP on images from a specific fine-grained class misclassify it into other classes. To do so, for a fixed class and for all test images that belong to it, we count the number of images that have been misclassified by the CAUP as belonging to each particular/adversarial class. In Figure 4, we

show four classes c in CUB (each c being at the center of a connected graph) and the adversarial classes to which the perturbed test images of c have been misclassified (connected to the center node by edges). The size of images representing each adversarial class corresponds to the number of samples fooled into that adversarial class. First, notice that there are only a few adversarial classes for each ground-truth class. For example, images from 'gray catbird' get fooled into one of the four classes shown in the figure. On the other hand, many images from 'gray catbird', 'arctic tern' and 'Brandt cormorant' classes are misclassified as 'painted bunting', making this class a dominant adversarial class. The reason for this dominancy is the major contribution (i.e., high utility) of attribute 'has breast/wing color: red', which is more specific to some species of birds including 'painted bunting.' Additionally, for similar reasons, 'green violet ear' and 'orange crown warbler' are other examples of the dominant adversarial classes. These results indeed confirm the existence of dominant adversarial classes in the fine-grained settings (both seen and unseen), where CAUP frequently misclassifies images into such classes. We believe this could be used to investigate defense mechanisms for fine-grained recognition, which we leave for future studies.



a) Correlation of Attributes b) Important AUPs

Fig. 5: a) Correlations among some learned attribute-based universal perturbations. Lighter color means being more correlated. b) We visualize the attributes whose perturbation has the largest absolute weights for changing a given class. Thicker edges mean higher attribute-class weights, ω_a^c. YHB means Yellow Headed Black Bird.

5 Conclusions

We proposed a compositional method for generating effective, generalizable and real-time adversarial attacks on fine-grained recognition models, by learning attribute-based universal perturbations and a model for composing them. By extensive experiments on multiple fine-grained datasets and using several fine-grained recognition models, we showed that our attacks are significantly more effective than conventional universal perturbations and generalize well from seen to unseen classes and across different architectures.

6 Acknowledgements

This work is sponsored by NSF (IIS-2115110), DARPA (HR00112220001) and ARO (W911NF2110276). Content does not necessarily reflect the position/policy of the Government. No official endorsement should be inferred.

References

1. Afifi, M., Brown, M.S.: What else can fool deep learning? addressing color constancy errors on deep neural network performance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 243–252 (2019)
2. Ak, K.E., Kassim, A.A., Lim, J.H., Tham, J.Y.: Learning attribute representations with localization for flexible fashion search. IEEE Conference on Computer Vision and Pattern Recognition (2018)
3. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2016)
4. Benz, P., Zhang, C., Imtiaz, T., Kweon, I.S.: Double targeted universal adversarial perturbations. Asian Conference on Computer Vision (2020)
5. Benz, P., Zhang, C., Karjauv, A., Kweon, I.S.: Universal adversarial training with class-wise perturbations. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
6. Bhattad, A., Chong, M.J., Liang, K., Li, B., Forsyth, D.A.: Unrestricted adversarial examples via semantic manipulation. arXiv preprint arXiv:1904.06347 (2019)
7. Bucher, M., Herbin, S., Jurie, F.: Generating visual representations for zero-shot classification. IEEE International Conference on Computer Vision Workshops (2017)
8. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. IEEE Symposium on Security and Privacy (2017)
9. Chang, D., Pang, K., Zheng, Y., Ma, Z., Song, Y.Z., Guo, J.: Your" flamingo" is my" bird": Fine-grained, or not. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11476–11485 (2021)
10. Changpinyo, S., Chao, W., Gong, B., Sha, F.: Synthesized classifiers for zero-shot learning. IEEE Conference on Computer Vision and Pattern Recognition (2016)
11. Chen, P.Y., Sharma, Y., Zhang, H., Yi, J., Hsieh, C.J.: Ead: Elastic-net attacks to deep neural networks via adversarial examples. AAAI Conference on Artificial Intelligence (2018)
12. Choi, J., Larson, M., Li, X., Li, K., Friedland, G., Hanjalic, A.: The geo-privacy bonus of popular photo enhancements. In: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. pp. 84–92 (2017)
13. Croce, F., Hein, M.: Sparse and imperceivable adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4724–4732 (2019)
14. Ding, Y., Zhou, Y., Zhu, Y., Ye, Q., Jiao, J.: Selective sparse sampling for fine-grained image recognition. IEEE International Conference on Computer Vision (2019)
15. Ding, Y., Ma, Z., Wen, S., Xie, J., Chang, D., Si, Z., Wu, M., Ling, H.: Ap-cnn: weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. IEEE Transactions on Image Processing **30**, 2826–2836 (2021)
16. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. IEEE Conference on Computer Vision and Pattern Recognition (2018)
17. Duan, R., Chen, Y., Niu, D., Yang, Y., Qin, A., He, Y.: Advdrop: Adversarial attack to dnns by dropping information. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7506–7515 (2021)

18. Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R., Naik, N.: Pairwise confusion for fine-grained visual classification. European Conference on Computer Vision (September 2018)
19. Elhoseiny, M., Zhu, Y., Zhang, H., Elgammal, A.M.: Link the head to the "beak": Zero shot learning from noisy text description at part precision. IEEE Conference on Computer Vision and Pattern Recognition pp. 6288–6297 (2017)
20. Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., Madry, A.: A rotation and a translation suffice: Fooling cnns with simple transformations (2018)
21. Felix, R., Kumar, B.G.V., Reid, I.D., Carneiro, G.: Multi-modal cycle-consistent generalized zero-shot learning. European Conference on Computer Vision (2018)
22. Gao, L., Zhang, Q., Song, J., Liu, X., Shen, H.T.: Patch-wise attack for fooling deep neural network. In: European Conference on Computer Vision. pp. 307–322. Springer (2020)
23. Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. IEEE Conference on Computer Vision and Pattern Recognition (2016)
24. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. International Conference on Learning Representations (2015)
25. Gragnaniello, D., Marra, F., Verdoliva, L., Poggi, G.: Perceptual quality-preserving black-box attack against deep learning image classifiers. Pattern Recognition Letters **147**, 142–149 (2021)
26. Han, Z., Fu, Z., Chen, S., Yang, J.: Contrastive embedding for generalized zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2371–2381 (2021)
27. Hayes, J., Danezis, G.: Learning universal adversarial perturbations with generative models. IEEE Security and Privacy Workshops (2018)
28. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15262–15271 (2021)
29. Hosseini, H., Poovendran, R.: Semantic adversarial examples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1614–1619 (2018)
30. Huang, S., Wang, X., Tao, D.: Stochastic partial swap: Enhanced model generalization and interpretability for fine-grained recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 620–629 (2021)
31. Huynh, D., Elhamifar, E.: Compositional zero-shot learning via fine-grained dense feature composition. Neural Information Processing Systems (2020)
32. Huynh, D., Elhamifar, E.: Fine-grained generalized zero-shot learning via dense attribute-based attention. IEEE Conference on Computer Vision and Pattern Recognition (2020)
33. Huynh, D., Elhamifar, E.: Compositional fine-grained low-shot learning. arXiv preprint arXiv:2105.10438 (2021)
34. Ji, R., Wen, L., Zhang, L., Du, D., Wu, Y., Zhao, C., Liu, X., Huang, F.: Attention convolutional binary neural tree for fine-grained visual categorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10468–10477 (2020)
35. Jiang, H., Wang, R., Shan, S., Chen, X.: Transferable contrastive network for generalized zero-shot learning. IEEE International Conference on Computer Vision (2019)
36. Kariyappa, S., Prakash, A., Qureshi, M.K.: Maze: Data-free model stealing attack using zeroth-order gradient estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13814–13823 (2021)

37. Khrulkov, V., Oseledets, I.: Art of singular vectors and universal adversarial perturbations. IEEE Conference on Computer Vision and Pattern Recognition (2018)
38. Kong, S., Fowlkes, C.C.: Low-rank bilinear pooling for fine-grained classification. IEEE Conference on Computer Vision and Pattern Recognition (2017)
39. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. ArXiv preprint, arXiv:1607.02533 (2016)
40. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. International Conference on Learning Representations (2017)
41. Laidlaw, C., Feizi, S.: Functional adversarial attacks. arXiv preprint arXiv:1906.00001 (2019)
42. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. IEEE Conference on Computer Vision and Pattern Recognition (2009)
43. Lin, D., Shen, X., Lu, C., Jia, J.: Deep lac: Deep localization, alignment and classification for fine-grained recognition. IEEE Conference on Computer Vision and Pattern Recognition (2015)
44. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. IEEE International Conference on Computer Vision (2015)
45. Liu, C., Xie, H., Zha, Z., Yu, L., Chen, Z., Zhang, Y.: Bidirectional attention-recognition model for fine-grained object classification. IEEE Transactions on Multimedia **22**(7), 1785–1795 (2019)
46. Liu, H., Ji, R., Li, J., Zhang, B., Gao, Y., Wu, Y., Huang, F.: Universal adversarial perturbation via prior driven uncertainty approximation. International Conference on Computer Vision (2019)
47. Liu, S., Long, M., Wang, J., Jordan, M.I.: Generalized zero-shot learning with deep calibration network. Neural Information Processing Systems (2018)
48. Liu, Y., Guo, J., Cai, D., He, X.: Attribute attention for semantic disambiguation in zero-shot learning. IEEE International Conference on Computer Vision (2019)
49. Liu, Y., Zhang, W., Wang, J.: Zero-shot adversarial quantization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1512–1521 (2021)
50. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. IEEE Conference on Computer Vision and Pattern Recognition (2016)
51. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. IEEE International Conference on Computer Vision (2015)
52. Luo, B., Liu, Y., Wei, L., Xu, Q.: Towards imperceptible and robust adversarial example attacks against neural networks. In: Thirty-second aaai conference on artificial intelligence (2018)
53. Maho, T., Furon, T., Le Merrer, E.: Surfree: a fast surrogate-free black-box attack. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10430–10439 (2021)
54. Mall, U., Hariharan, B., Bala, K.: Field-guide-inspired zero-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9546–9555 (2021)
55. Mao, C., Chiquier, M., Wang, H., Yang, J., Vondrick, C.: Adversarial attacks are reversible with natural supervision. arXiv preprint arXiv:2103.14222 (2021)
56. Metzen, J.H., Kumar, M.C., Brox, T., Fischer, V.: Universal adversarial perturbations against semantic image segmentation. International Conference on Computer Vision (2019)

57. Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. IEEE Conference on Computer Vision and Pattern Recognition (2017)
58. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. IEEEConference on Computer Vision and Pattern Recognition (2016)
59. Mopuri, K.R., Ganeshan, A., Babu, R.V.: Generalizable data-free objective for crafting universal adversarial perturbations. IEEE transactions on pattern analysis and machine intelligence **41**(10), 2452–2465 (2018)
60. Mopuri, K.R., Garg, U., Babu, R.V.: Fast feature fool: A data independent approach to universal adversarial perturbations. arXiv preprint arXiv:1707.05572 (2017)
61. Nakka, K.K., Salzmann, M.: Towards robust fine-grained recognition by maximal separation of discriminative features. Asian Conference on Computer Vision (2020)
62. Narodytska, N., Kasiviswanathan, S.: Simple black-box adversarial attacks on deep neural networks. IEEE Conference on Computer Vision and Pattern Recognition Workshops (2017)
63. Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G.S., Dean, J.: Zero-shot learning by convex combination of semantic embeddings. International Conference on Learning Representations (2014)
64. Park, S.M., Wei, K.A., Xiao, K., Li, J., Madry, A.: On distinctive properties of universal perturbations. arXiv preprint arXiv:2112.15329 (2021)
65. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. British Machine Vision Conference (2015)
66. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. IEEE Conference on Computer Vision and Pattern Recognition (2012)
67. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. Empirical Methods in Natural Language Processing (EMNLP) (2014)
68. Pony, R., Naeh, I., Mannor, S.: Over-the-air adversarial flickering attacks against video recognition networks. IEEE Conference on Computer Vision and Pattern Recognition (2021)
69. Poursaeed, O., Katsman, I., Gao, B., Belongie, S.: Generative adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4422–4431 (2018)
70. Rampini, A., Pestarini, F., Cosmo, L., Melzi, S., Rodola, E.: Universal spectral adversarial attacks for deformable shapes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3216–3226 (2021)
71. Romera-Paredes, B., Torr, P.H.: An embarrassingly simple approach to zero-shot learning. International Conference on Machine learning (2015)
72. Rony, J., Granger, E., Pedersoli, M., Ben Ayed, I.: Augmented lagrangian adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7738–7747 (2021)
73. Rony, J., Hafemann, L.G., Oliveira, L.S., Ayed, I.B., Sabourin, R., Granger, E.: Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4322–4330 (2019)
74. Sayles, A., Hooda, A., Gupta, M., Chatterjee, R., Fernandes, E.: Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14666–14675 (2021)
75. Schönfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., Akata, Z.: Generalized zero- and few-shot learning via aligned variational autoencoders. IEEE Conference on Computer Vision and Pattern Recognition (2019)
76. Skorokhodov, I., Elhoseiny, M.: Class normalization for (continual)? generalized zero-shot learning. arXiv preprint arXiv:2006.11328 (2020)
77. Sun, M., Yuan, Y., Zhou, F., Ding, E.: Multi-attention multi-class constraint for fine-grained image recognition. European Conference on Computer Vision (2018)
78. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. International Conference on Learning Representations (2014)
79. Wang, W., Xu, Y., Shen, J., Zhu, S.C.: Attentive fashion grammar network for fashion landmark detection and clothing category classification. IEEE Conference on Computer Vision and Pattern Recognition (2018)
80. Wang, X., Lin, S., Zhang, H., Zhu, Y., Zhang, Q.: Interpreting attributions and interactions of adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1095–1104 (2021)
81. Wang, X., Zhang, Z., Wu, B., Shen, F., Lu, G.: Prototype-supervised adversarial network for targeted attack of deep hashing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16357–16366 (2021)
82. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
83. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD Birds 200. Tech. Rep. CNS-TR-2010-001, California Institute of Technology (2010)
84. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. European Conference on Computer Vision (2016)
85. Wong, E., Schmidt, F., Kolter, Z.: Wasserstein adversarial examples via projected sinkhorn iterations. In: International Conference on Machine Learning. pp. 6808–6817. PMLR (2019)
86. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 5542–5551 (2018)
87. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning — the good, the bad and the ugly. IEEE Conference on Computer Vision and Pattern Recognition (2017)
88. Xian, Y., Sharma, S., Schiele, B., Akata, Z.: f-vaegan-d2: A feature generating framework for any-shot learning. IEEE Conference on Computer Vision and Pattern Recognition (2019)
89. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
90. Xiao, C., Li, B., Zhu, J.Y., He, W., Liu, M., Song, D.: Generating adversarial examples with adversarial networks. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. pp. 3905–3911 (2018)
91. Xiao, C., Zhu, J.Y., Li, B., He, W., Liu, M., Song, D.: Spatially transformed adversarial examples. arXiv preprint arXiv:1801.02612 (2018)
92. Xiao, J., Ehinger, K.A., Hays, J., Torralba, A., Oliva, A.: Sun database: Exploring a large collection of scene categories. International Journal of Computer Vision **119**(1), 3–22 (2016)

93. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2730–2739 (2019)
94. Xu, K., Liu, S., Zhao, P., Chen, P.Y., Zhang, H., Fan, Q., Erdogmus, D., Wang, Y., Lin, X.: Structured adversarial attacks: Towards general implementation and better interpretability. International Conference on Learning Representations (2019)
95. Yue, Z., Wang, T., Sun, Q., Hua, X.S., Zhang, H.: Counterfactual zero-shot and open-set visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15404–15414 (2021)
96. Zhang, C., Benz, P., Imtiaz, T., Kweon, I.S.: Cd-uap: Class discriminative universal adversarial perturbation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 6754–6761 (2020)
97. Zhang, C., Benz, P., Lin, C., Karjauv, A., Wu, J., Kweon, I.S.: A survey on universal adversarial attack. arXiv preprint arXiv:2103.01498 (2021)
98. Zhang, H., Avrithis, Y., Furon, T., Amsaleg, L.: Smooth adversarial examples. EURASIP Journal on Information Security **2020**(1), 1–12 (2020)
99. Zhang, L., Huang, S., Liu, W.: Intra-class part swapping for fine-grained image classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3209–3218 (2021)
100. Zhang, L., Huang, S., Liu, W., Tao, D.: Learning a mixture of granularity-specific experts for fine-grained categorization. IEEE International Conference on Computer Vision (2019)
101. Zhang, Z., Saligrama, V.: Zero-shot learning via joint latent similarity embedding. IEEE Conference on Computer Vision and Pattern Recognition (2016)
102. Zhao, X., Yang, Y., Zhou, F., Tan, X., Yuan, Y., Bao, Y., Wu, Y.: Recognizing part attributes with insufficient data. IEEE International Conference on Computer Vision (2019)
103. Zhao, Y., Yan, K., Huang, F., Li, J.: Graph-based high-order relation discovery for fine-grained recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15079–15088 (2021)
104. Zhao, Z., Liu, Z., Larson, M.: Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1039–1048 (2020)
105. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. IEEE International Conference on Computer Vision (2017)
106. Zheng, H., Fu, J., Zha, Z.J., Luo, J.: Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. IEEE Conference on Computer Vision and Pattern Recognition (2019)