This CVPR 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.

Data-free Universal Adversarial Perturbation with Pseudo-semantic Prior

Chanhui Lee Yeonghwan Song Jeany Son
AI Graduate School, GIST
{as584868, yeonghwan.song}@gm.gist.ac.kr jeany@gist.ac.kr

Abstract

Data-free Universal Adversarial Perturbation (UAP) is an image-agnostic adversarial attack that deceives deep neural networks using a single perturbation generated solely from random noise without relying on data priors. However, traditional data-free UAP methods often suffer from limited transferability due to the absence of semantic content in random noise. To address this issue, we propose a novel data-free universal attack method that recursively extracts pseudo-semantic priors directly from the UAPs during training to enrich the semantic content within the data-free UAP framework. Our approach effectively leverages latent semantic information within UAPs via region sampling, enabling successful input transformations—typically ineffective in traditional data-free UAP methods due to the lack of semantic cues—and significantly enhancing black-box transferability. Furthermore, we introduce a sample reweighting technique to mitigate potential imbalances from random sampling and transformations, emphasizing hard examples less affected by the UAPs. Comprehensive experiments on ImageNet show that our method achieves state-of-the-art performance in average fooling rate by a substantial margin, notably improves attack transferability across various CNN architectures compared to existing data-free UAP methods, and even surpasses data-dependent UAP methods. Code is available at: https://github.com/ChnanChan/PSP-UAP.

1. Introduction

Deep neural networks (DNNs) have become widely used in computer vision, achieving remarkable performance across a diverse range of tasks, such as image classification [7, 34], object detection [25, 26], semantic segmentation [27], and visual tracking [1, 42]. Despite these successes, DNNs are vulnerable to carefully crafted, imperceptible perturbations in input data, causing the model to make highly confident yet incorrect predictions. This vulnerability poses significant challenges for deploying DNNs in critical applications, such as autonomous driving [5] and security systems [2],

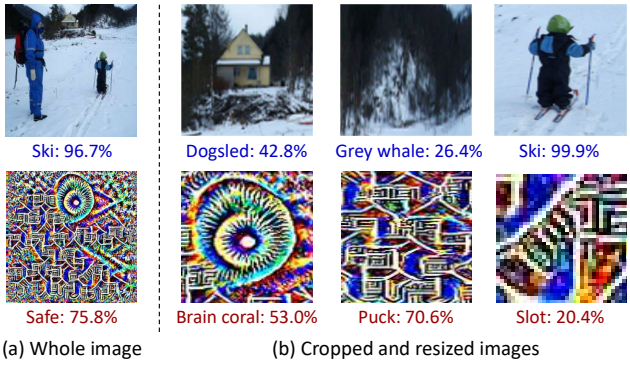

Figure 1. Diverse semantic contents in both a real-image and UAP: (a) Whole images from the ImageNet dataset (top) and our generated data-free UAP (bottom) using DenseNet-121, shown at iteration 900 during the training phase. The Top-1 class and its score are shown below each image. (b) Cropped regions from the whole image (top) and our UAP (bottom). Those regions contain diverse semantics that differ from the class of the original images.

and has led to increased research into adversarial attacks that generate adversarial examples.

To craft the adversarial examples with high transferability across various DNN architectures, there have been many adversarial attack methods using a specific target image [3, 6, 11, 18, 32, 39]. However, these methods generate a unique perturbation for each target image, which is time-consuming, impractical for real-world scenarios, and limits their generalization to other images. To tackle this limitation, the Universal Adversarial Perturbation (UAP) [19] introduced an image-agnostic attack that generates a single image-agnostic adversarial perturbation, which is capable of attacking a wide range of unknown images. Many studies [15, 22, 24, 30] focus on developing data-dependent UAPs that target unknown models, thereby enhancing transferability across diverse and unseen scenarios. While these UAPs deceive diverse categories of images with a single perturbation, they still rely on large-scale data samples and their labels from the target domain to capture diverse semantics, such as the ImageNet [28] dataset.

Accessing data priors from the target domain is often

impractical, leading to recent interest in data-free UAP methods [12, 14, 16, 20, 21, 23, 41]. Data-free UAP poses a greater challenge than conventional data-dependent UAP generation tasks, as it restricts the employment of any prior knowledge of the target domain dataset. Prior works [16, 20, 21] have attempted to craft UAPs from random noise without any dataset, by maximizing activations in convolutional neural networks (CNNs) layers. However, these methods solely rely on random priors, such as Gaussian noise or jigsaw patterns, which lack semantic information and thus offer limited transferability to unseen models. To overcome this limitation, several works [12, 23] utilize auxiliary data samples generated by optimizing against the outputs of a surrogate model. Although these methods allow the use of semantic information in synthetic data, the crafted UAPs often show inferior transferability due to overfitted data to the surrogate model.

In this paper, we explore how to leverage semantic information directly from the UAP itself, without any dataset priors, to address these challenges. Our approach is inspired by the observation that even a single generated UAP contains diverse semantic information as well as its dominant label, as shown in Figure 1. We find that the generated UAP encodes diverse semantic features, similar to a real-world image with various semantic contents across different regions. For example, in Figure 1, while the whole UAP is predicted as 'Safe' due to its tendency to have a dominant label, cropped regions within the UAP are predicted as classes like 'Brain coral,' 'Puck,' and 'Slot.' This observation motivates us to utilize UAP as a semantic prior for training within a data-free UAP framework.

Inspired by this insight, we propose PSP-UAP, a novel data-free UAP method that generates pseudo-semantic priors from a UAP during training. To capture more diverse semantics in pseudo-semantic priors, we randomly crop and resize regions to extract semantic samples and treat them as images to be fooled. This approach effectively addresses the data-free constraint in UAP generation by leveraging richer semantic information inherent in the UAP, rather than relying solely on random noise. To improve attack transferability, we further incorporate input transformations [40, 43], commonly used in image-specific adversarial attacks, into our data-free UAP framework. This strategy has not been explored in existing data-free UAP methods, as random priors lack semantic information, limiting its effectiveness. In contrast, our pseudo-semantic prior contains richer semantic content, thereby enabling improved transferability to unknown models through input transformations. Moreover, since semantic samples obtained through random cropping and transformation vary in informativeness, we introduce a sample reweighting that prioritizes hard examples, which are less effectively deceived by the current UAP, to improve the overall effectiveness of the generated UAPs.

The main contributions can be summarized as follows:

- We propose PSP-UAP, a novel data-free universal attack method that generates pseudo-semantic priors from the UAP itself, using inherent semantic information of the UAP as an alternative data source during training.
- We are the first to incorporate input transformations into the data-free UAP framework by leveraging pseudo-semantic priors with diverse semantic cues, boosting transferability across various CNN architectures in black-box settings.
- Our sample reweighting prioritizes challenging examples during UAP training, by reducing the influence of uninformative samples produced by random sampling in our pseudo-semantic prior and input transformations.
- Our method achieves outstanding performance over the state-of-the-art data-free UAP methods by a substantial margin, and even outperforms existing data-dependent UAP methods.

2. Related Work

Data-dependent Universal Attack. Data-dependent UAPs aim to generate a single perturbation that misleads any image sample. UAP [19] first proposed finding the minimal universal adversarial perturbation at each step by DeepFool [18] method. SPGD-UAP [30] combined the stochastic gradient method with the projected gradient descent (PGD) [17] attack method. SGA-UAP [15] used stochastic gradient aggregation in mini-batch to address gradient vanishing and quantization errors. AT-UAP [12] integrated image-specific and image-agnostic attacks to improve the robustness of universal perturbation. NAG [22] and GAP [24] applied generative adversarial frameworks to craft perturbations. Although these works [15, 22, 24, 30] increase the transferability of black-box attacks, they require the training dataset, making it impractical when the adversary does not have any prior of the target domain.

Data-free Universal Attack. Data-free universal attack aims to craft UAPs without any dataset to be used for training, thereby alleviating the data access requirement presented in data-dependent universal attacks. Generating UAP without prior knowledge of the target domain is more practical and suitable for real-world applications. Fast Feature Fool (FFF) [20] first proposed a data-free universal adversarial attack by maximizing the feature activation values of all CNN layers. Generalizable Data-free UAP (GD-UAP) [21] improved FFF [20] attack method with saturation check strategy on the training stage. AT-UAP [12] also performed experiments in a data-free manner by applying adversarial attacks to random noise. Prior-Driven Uncertainty Approximation (PD-UA) [14] introduced an attack method to train UAP by maximizing the uncertainty approximation of the model with the prior patterns, and Cosine-

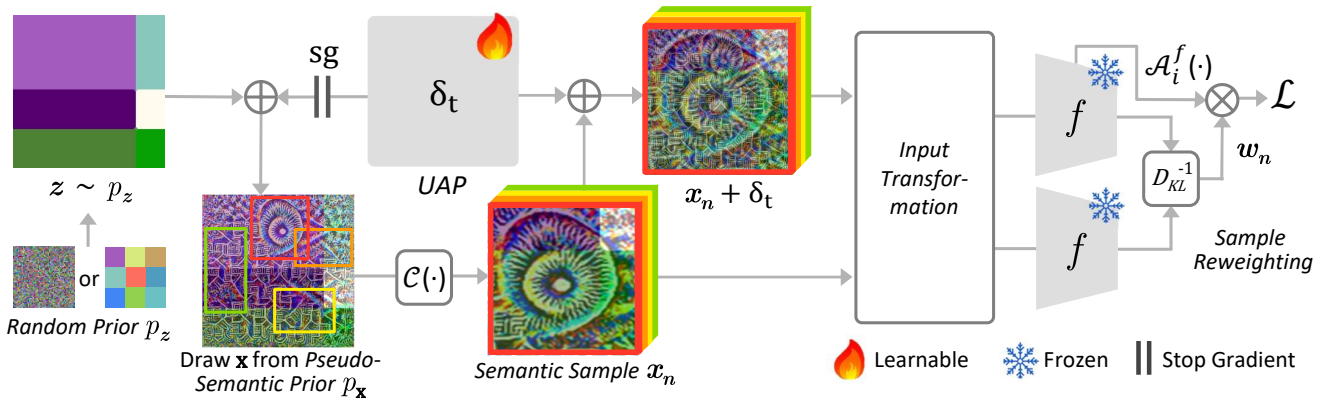

Figure 2. Overall pipeline of the proposed PSP-UAP. The pseudo-semantic prior is created by adding random noise to the UAP. Semantic samples are then generated by randomly cropping and resizing the pseudo-semantic prior. Input transformation is applied to both adversarial and clean versions of the semantic samples to calculate sample reweighting. Finally, the loss is defined as the product of sample reweighting and the activations of the semantic samples, from which gradients are computed to update the model.

UAP [41] proposed the minimizing cosine similarity to craft UAPs in a self-supervised manner. AAA [23] crafted class impressions with logits to train a generative model to optimize UAPs. TRM-UAP [16] increased the ratio of positive and negative activations on the shallow convolution layers and adapted curriculum learning to enhance attack transferability stably. Despite employing various methods to generate UAPs in a data-free setting, they face significant challenges due to the lack of information on both the target models and domains. Additionally, these works substantially rely on random priors to generate UAPs, and auxiliary data directly utilizes label information, leading to overfitting on surrogate models.

Input Transformation Attack. Input transformation methods have emerged as one of the effective ways to improve attack transferability in image-specific adversarial attacks. Diverse input method (DIM) [40], translate invariant method (TIM) [4], and scale invariant method (SIM) [13] revealed the DNN models' invariant properties to transformations such as resizing, translation, and scaling before the gradient calculation. SIA [38] applied various transformations to the input image while maintaining its overall structure. *Admix* [37] created admixed images by blending a small fraction of images from different categories into the input image. Block shuffle and rotation (BSR) [36] randomly shuffled and rotated the sub-blocks of the input image to reduce the variance in attention heatmaps across different models. L2T [43] used reinforcement learning to increase the diversity of transformed images by selecting the optimal transformation combinations. To fully leverage our pseudo-semantic prior, we incorporate input transformations into our data-free UAP method to further enhance black-box transferability.

3. Methodology

In this section, we present our motivation and approach for generating the pseudo-semantic prior. We then describe the input transformation applied to semantic samples derived from the pseudo-semantic prior, followed by our sample reweighting strategy for optimizing UAP.

3.1. Preliminaries of data-free UAP

Universal adversarial attacks aim to optimize a single perturbation δ using a model f that effectively deceives most of the samples I in the target domain dataset, with the pixel intensities of δ restricted by a constraint parameter ϵ:

$$f(I+\delta) \neq f(I), \quad \text{s.t. } \|\delta\|_\infty \leq \epsilon. \tag{1}$$

However, in data-free settings where the target dataset is inaccessible, UAPs are typically trained using simple random priors, such as Gaussian noises or jigsaw images [16, 41]. Given these random priors p_z, GD-UAP [21] introduced an activation maximizing loss as follows:

$$\mathcal{L} = -\mathbb{E}_{z\sim p_z} \sum_{i=1}^{L} \log \|\mathcal{A}_i^f(z+\delta)\|_2, \tag{2}$$
$$\text{s.t. } \|\delta\|_\infty \leq \epsilon,$$

where $\mathcal{A}_i^f(\cdot)$ indicates the activation of the i-th layer of the surrogate network f, L denotes the number of layers in f, and z represents pseudo-data sampled from a simple random prior distribution, p_z. This loss is designed to overactivate features extracted from multiple convolutional layers of the surrogate model without input images. Consequently, the distorted activation interferes with feature extraction, leading CNN models to make incorrect predictions [20, 21].

3.2. Pseudo-Semantic Prior

Although data-free UAP methods [16, 21] use random priors, such as Gaussian noises or artificial jigsaw puzzles, to mimic the statistical properties of image datasets, they are still limited by a lack of semantic information. Furthermore, since UAPs are trained by maximizing activations in network layers, they tend to overfit to surrogate models. Optimizing UAPs without semantic content and relying on activation or outputs of surrogate networks, reduces their effectiveness in disrupting real images in unseen models, leading to degraded performances in black box transferability.

To address these issues, we aim to enhance the semantic content within a data-free UAP framework, inspired by previous works [15, 16, 19, 41] that demonstrate the presence of dominant labels in generated UAPs. For instance, early works [15, 19] observed that untargeted UAPs often cause misclassification toward a dominant label, a property that also holds in data-free settings [16]. Cosine-UAP [41] further showed that the logit distribution of UAPs tends to dominate that of the input data x. This suggests that, although the UAP is a subtle perturbation, it behaves like a single image with strong semantic information, guiding the model toward classification with a dominant label.

Inspired by this observation, we leverage the inherent semantic information in UAPs by treating the combination of UAP and random noise as a single image, termed the ***pseudo-semantic prior***, to resolve the lack of semantic content in data-free UAP training. As shown in Figure 1 and Figure 3, the generated UAPs exhibit diverse semantic labels across different regions. Although regions within the UAP are classified under the same label, the attention heatmaps generated by Grad-CAM [29] show distinct patterns, suggesting that diverse semantic patterns are embedded within the UAP.

Building on the above insight, we generate pseudo-data samples from the pseudo-semantic prior to enrich the semantic content, which we refer to as ***semantic samples*** x_n:

$$\mathbf{x} \sim p_{\mathbf{x}|p_z,\delta_t} = \{z + \delta_t | z \in p_z\}, \tag{3}$$

$$\{x_1, x_2, ..., x_N\} = \mathcal{C}(\mathbf{x}; N), \tag{4}$$

where z and δ_t denote random noise sampled from p_z and the UAP being trained in t-th iteration, respectively. The pseudo-semantic prior $p_{\mathbf{x}}$ denotes a set of adversarial examples with δ_t derived from the random prior distribution. $\mathcal{C}$ is a sampler that draws N numbers of semantic samples x_n from $p_{\mathbf{x}}$ by applying crop and resize operations. Specifically, we randomly crop a region of $\mathbf{x}$, the sum of random noise z and the UAP δ_t, and resize it to the original scale of the UAP size. We believe that leveraging the generated semantic information embedded in different regions of the UAP provides more effective guidance for successfully attacking target features than relying solely on random noise.

3.3. Input Transformation

To enhance black-box attack transferability, we incorporate input transformation techniques into our semantic samples. While input transformation is commonly used in image-specific adversarial attacks, it has been less explored in data-free UAPs due to the limited semantic information in random priors. In our method, however, the semantic samples generated from pseudo-semantic priors contain diverse semantic cues, making input transformations more effective in a data-free setting.

Following L2T [43], which shows that rotation, scaling, and shuffling are particularly effective for improving attack transferability, we randomly select one of these transformations and apply it to each semantic sample. For rotation, the angle α is drawn from a truncated normal distribution within the range $-\theta \leq \alpha \leq \theta$. Scaling is applied with a uniform distribution within the bounds $\beta_{low} \leq \beta \leq \beta_{high}$. Shuffling involves randomly rearranging $m \times m$ blocks. During optimization, applying input transformations to our PSP-UAP increases the variation of semantic samples and enhances the black-box attack transferability.

3.4. Sample Reweighting

Random processes involved in drawing semantic samples from the pseudo-semantic prior and applying input transformations lead to an imbalance in difficulty due to variations in semantic content; some samples are easily fooled by the UAP, while others are more difficult to deceive. To tackle this, we propose a new sample reweighting method that prioritizes harder-to-fool samples.

Specifically, we compute a weight for each sample using the KL-divergence between the transformed input and its adversarial counterpart during training. We define the original distribution P and the adversarial distribution Q for each semantic sample and its corresponding adversarial example generated by the current UAP δ as follows:

$$P(x_n) = f(T(x_n)), \tag{5}$$

$$Q(x_n) = f(T(x_n + \delta_t)), \tag{6}$$

where $f(\cdot)$ denotes the temperature-scaled softmax output of a surrogate model and and $T(\cdot)$ represents a randomly selected input transformation as described in Sec. 3.3. We then compute the weights for each semantic sample using the KL-divergence:

$$w_n = D_{KL}(P(x_n) \| Q(x_n))^{-1}. \tag{7}$$

The large KL-divergence value indicates that δ_t has significantly altered the distribution of x_n, whereas a small value suggests an ineffective attack. Thus, we take the reciprocal of KL-divergence values to assign greater weights to semantic samples where the attack results in minimal distributional change. We reweight the semantic samples using

Algorithm 1 Pseudo-semantic Prior Universal Attack

Input: Surrogate model f, number of semantic samples N, maximum perturbation magnitude ϵ, learning rate η, maximum iteration number T, convergence threshold $F_{\max}$, validation test hyperparameter H, saturation threshold r.
Output: Universal adversarial perturbation δ.

Initialize $\delta_0 \sim \mathcal{U}(-\epsilon, \epsilon), t = 0, F = 0$
while $t < T$ and $F < F_{max}$ **do**
  $t = t + 1$
  Generate the random noise set $z \sim p_z$
  Update pseudo-semantic prior $p_{\mathbf{x}}$ with δ_t via Eq. (3)
  Sample N semantic samples x_n via Eq. (4)
  Select and apply transformation $T \in \{$*rotation*, *scaling*, *shuffling*$\}$
  Compute the weight w via Eq. (5), (6), and (7)
  Calculate the gradient $\nabla\mathcal{L}$ of the loss in Eq. (8)
  Update $\delta_t = \delta_{t-1} + \eta \cdot \nabla\mathcal{L}$
  Clip $\delta_t = \min(\epsilon, \max(\delta_t, -\epsilon))$
  Compute the saturation rate $\hat{r}$ and adjust δ_t if $r < \hat{r}$
  if $t\%H == 0$ **then**
    Conduct the fooling rate test FR
    if FR is not the best fooling rate **then**
      $F = F + 1$
    end if
  end if
end while
return δ_t

the weights generated from Eq. (7), considering the influence of the UAP on each sampled x_n, which enables us to optimize the UAP effectively.

3.5. Overall Loss

We integrate the pseudo-semantic prior, input transformation, and sample reweighting into Eq. (2) to optimize the proposed PSP-UAP. The final loss function of our PSP-UAP is defined as follows:

$$\mathcal{L} = -\mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}} \sum_{n=1}^{N} \sum_{i=1}^{l} \log\left(w_n \|\mathcal{A}_i^f(T(x_n + \delta_t))\|_2\right), \quad (8)$$

where $\mathcal{A}_i^f(\cdot)$ denotes the activation of the i-th layer in network f, x_n represent the n-th semantic sample extracted from the pseudo-semantic prior $p_{\mathbf{x}}$, δ_t denotes the UAP at the t-th iteration, w_n is the weight of x_n from Eq. (7), l indicates the number of the convolutional layers used in the activation sum, $T(\cdot)$ denotes a randomly selected transform for input transformation, and N is the number of semantic samples extracted from one pseudo-semantic prior.

By optimizing the overall loss, we fully leverage the generated semantic samples as data to produce a highly transferable UAP, even without prior knowledge of the target domain. The overall PSP-UAP framework and detailed algorithm are shown in Figure 2 and Algorithm 1, respectively.

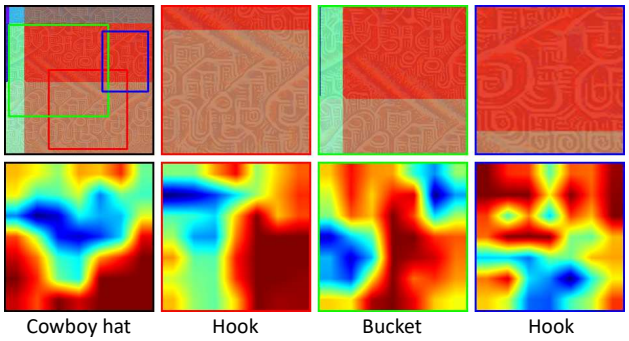

Figure 3. Semantic samples derived from an adversarial example during training, along with their predicted labels and GradCAM heatmaps from Dense-121. Despite originating from the same example, the variations in predicted labels and heatmaps indicate that these semantic samples capture diverse semantic features.

Attack	AlexNet	VGG16	VGG19	RN152	Google	Avg.
FFF [20]	80.92	47.10	43.62	-	56.44	-
AAA [23]	89.04	71.59	72.84	60.72	75.28	73.89
GD-UAP [21]	85.24	90.01	87.34	45.96	45.87	70.88
PD-UAP [14]	-	70.69	64.98	46.39	67.12	-
Cosine-UAP [41]	91.07	89.48	86.81	65.35	**87.57**	84.05
AT-UAP-U [12]	**96.66**	94.50	92.85	73.15	82.60	87.95
TRM-UAP [16]	93.53	94.30	91.35	67.46	85.32	86.39
PSP-UAP (Ours)	91.77	**96.26**	**94.65**	**85.65**	81.43	**89.95**

Table 1. FR (%) of our PSP-UAP and other data-free universal attack methods for white-box attacks.

4. Experiments

Experimental Setup. We follow the experiment setup in existing data-free universal attacks [16, 21] to evaluate the performance of our PSP-UAP. We evaluate the proposed method on ImageNet [28] validation set with five ImageNet pre-trained CNN models, AlexNet [10], VGG16 [31], VGG19 [31], ResNet152 (RN152) [7], and GoogleNet [33], which are commonly used in data-free UAP methods. We further explore four additional CNN models including DenseNet121 [9], MobileNet-v3-Large [8], ResNet50 [7], and Inception-v3 [35], pre-trained on ImageNet dataset.

Evaluation Metrics. To effectively evaluate the attack performance of our proposed method, we use the fooling rate (FR) which is widely used in universal attacks [16, 19]. FR indicates the proportion of samples with label changes when applying UAP.

Baselines. The proposed method is compared with the following data-free universal attacks, including FFF [20],

Model	Attack	AlexNet	VGG16	VGG19	ResNet152	GoogleNet	Average
AlexNet	AT-UAP-U	**96.66**±0.12*	72.33±0.50	67.24±0.18	43.63±0.29	62.01±0.32	68.37
	TRM-UAP	93.53±0.07*	60.10±0.24	57.08±0.15	27.31±0.30	32.70±0.22	54.14
	PSP-UAP (Ours)	91.77±0.32*	**76.56**±0.67	**74.07**±0.54	**49.20**±1.12	**66.00**±0.75	**71.52**
VGG16	AT-UAP-U	**54.15**±0.70	94.50±0.21*	86.65±0.70	36.96±1.03	48.53±1.32	64.16
	TRM-UAP	47.53±0.51	94.30±0.12*	89.68±0.14	61.43±0.40	53.95±0.59	69.38
	PSP-UAP (Ours)	50.40±0.53	**96.26**±0.21*	**92.60**±0.33	**74.10**±1.10	**64.89**±0.66	**75.65**
VGG19	AT-UAP-U	**62.05**±1.01	88.96±0.50	92.85±0.48*	42.72±0.51	**60.99**±1.41	69.51
	TRM-UAP	46.01±0.44	89.82±0.15	91.35±0.30*	47.19±0.46	46.48±0.78	64.17
	PSP-UAP (Ours)	48.93±0.72	**94.55**±0.14	**94.65**±0.10*	**67.13**±1.37	58.83±1.19	**72.81**
ResNet152	AT-UAP-U	49.78±0.68	62.78±0.71	60.54±0.49	73.15±1.15*	48.37±0.49	58.92
	TRM-UAP	53.56±0.75	77.20±0.35	73.30±0.41	67.46±0.35*	57.54±0.50	65.81
	PSP-UAP (Ours)	**58.82**±1.17	**88.59**±1.38	**87.35**±0.92	**85.65**±1.70*	**76.00**±1.33	**79.28**
GoogleNet	AT-UAP-U	55.65±0.37	71.38±0.83	68.25±0.59	43.03±0.42	82.60±0.72*	64.18
	TRM-UAP	60.10±1.16	**79.66**±0.95	**79.98**±1.06	**58.85**±1.94	**85.32**±0.04*	**72.78**
	PSP-UAP (Ours)	**65.22**±0.56	78.43±0.73	79.26±0.73	57.63±0.66	81.43±0.49*	72.39

Table 2. Black-box attack transferability of the UAP synthesized by our PSP-UAP method compared to other data-free universal attacks, AT-UAP-U [12] and TRM-UAP [16]. We show the mean and standard deviation of FR with five runs. Bold FR (%) denotes the best performance. The UAPs are crafted on AlexNet, VGG16, VGG19, ResNet152, and GoogleNet. * indicate FR of the white-box model.

Model	Attack	ResNet50	DenseNet121	MobileNet-v3-L	Inception-v3	Average
ResNet50	TRM-UAP	73.26±0.82*	54.42±1.23	61.25±1.48	37.36±0.69	56.57
	PSP-UAP (Ours)	**77.60**±0.42*	**66.11**±0.87	**70.50**±1.10	**42.32**±1.32	**64.13**
DenseNet121	TRM-UAP	35.24±2.55	70.10±2.07*	34.17±1.77	32.11±2.38	42.91
	PSP-UAP (Ours)	**53.03**±0.90	**85.81**±1.17*	**50.22**±0.58	**50.73**±0.78	**59.95**
MobileNet-v3-L	TRM-UAP	39.47±1.11	40.37±0.47	73.07±0.96*	30.11±0.81	45.76
	PSP-UAP (Ours)	**54.38**±1.40	**54.62**±1.82	**90.39**±0.23*	**46.29**±0.69	**61.42**
Inception-v3	TRM-UAP	53.53±0.57	54.93±0.54	67.16±0.60	64.22±0.33*	59.96
	PSP-UAP (Ours)	**57.60**±0.26	**57.50**±0.59	**70.20**±0.56	**65.38**±0.51*	**62.67**

Table 3. FR (%) for the UAPs crafted by TRM-UAP [16] and our PSP-UAP across additional CNN models. The UAPs are crafted on ResNet50, DenseNet121, MobileNet-v3-Large, and Inception-v3. * indicates the white-box model.

GD-UAP [21], PD-UA [14], Cosine-UAP [41], AT-UAP [12], and TRM-UAP [16]. Since AT-UAP includes both data-free (AT-UAP-U) and data-dependent (AT-UAP-S) versions, we evaluate both in our experiments. We also compared our method with SGA-UAP [15] which is one of the state-of-the-art data-dependent universal attacks.

Implementation Details. Our experiments are implemented on PyTorch with a single NVIDIA A6000 GPU. We set $\epsilon = 10/255$ to restrict ℓ_∞-norm, the maximum iteration T as $10,000$, and the saturation threshold r to 0.001%, following the setting in TRM-UAP [16, 21]. For input transformations, we set $\theta = 6$ for rotation, $\beta_{low} = 0.8$ and $\beta_{high} = 4$ for scaling, and $m = 2$ for random shuffling. Moreover, the number of semantic samples N is set to 10. We define the temperature parameter $\tau \in \{1.0, 5.0, 5.0, 3.0, 5.0\}$ corresponding to AlexNet, VGG16, VGG19, ResNet152, GoogleNet. For the additional CNN models, we set $\tau \in \{3.0, 10.0, 2.0, 3.0\}$ corresponding to ResNet50, DenseNet121, MobileNet-v3-Large, Inception-v3. We follow the same curriculum learning and saturation check strategy of TRM-UAP. We set different ratios to use the activations of intermediate layers across different models. To ensure a fair evaluation, we find the optimal parameters for TRM-UAP on RN50, DN121, MB-v3-L, and Inc-v3 through our best efforts.

4.1. Evaluation on White-Box Attack

We first evaluate UAPs generated by our PSP-UAP on various CNN models under the white-box setting. We compare the attack performance of our UAPs with other data-free universal attacks on the ImageNet validation set, as shown in Table 1. While the FR on AlexNet and GoogleNet is slightly lower than other methods, our approach achieves the highest average FR across all universal attack methods. Notably, the improvement on ResNet152 is substantial, with a 12.5% increase in the FR, demonstrating that our PSP-UAP performs exceptionally well, particularly on deeper and more complex CNN models.

4.2. Evaluation on Black-Box Attack

Comparison with SoTA Data-free UAPs. We evaluate the transferability of our UAPs on commonly used CNN models in the black-box scenario. Table 2 represents the attack performance across different settings, with columns representing target models and rows indicating surrogate

Model	Data	Attack	AlexNet	VGG16	VGG19	ResNet152	GoogleNet	Average
AlexNet	✓	SGA-UAP	**97.43***	66.41	60.96	35.76	49.71	62.05
		AT-UAP-S	97.01±0.11*	62.37±1.37	57.72±0.62	33.40±0.77	47.31±1.65	59.56
	✗	**PSP-UAP (Ours)**	91.77±0.32*	**76.56**±0.67	**74.07**±0.54	**49.20**±1.12	**66.00**±0.75	**71.52**
VGG16	✓	SGA-UAP	49.02	**98.36***	**94.17**	49.02	55.78	69.27
		AT-UAP-S	45.58±0.29	97.51±0.08*	91.53±0.22	47.16±0.95	53.63±0.90	67.08
	✗	**PSP-UAP (Ours)**	**50.40**±0.53	96.26±0.21*	92.60±0.33	**74.10**±1.10	**64.89**±0.66	**75.65**
VGG19	✓	SGA-UAP	**50.67**	**95.52**	**97.69***	51.08	56.87	70.37
		AT-UAP-S	46.04±0.58	93.49±0.17	97.56±0.04*	43.53±0.57	52.58±0.81	66.64
	✗	**PSP-UAP (Ours)**	48.93±0.72	94.55±0.14	94.65±0.10*	**67.13**±1.37	**58.83**±1.19	**72.81**
ResNet152	✓	SGA-UAP	51.59	81.77	79.01	**94.04***	64.05	74.09
		AT-UAP-S	47.33±0.89	81.93±0.94	78.72±0.91	91.52±0.78*	61.32±0.98	72.16
	✗	**PSP-UAP (Ours)**	**58.82**±1.17	**88.59**±1.38	**87.35**±0.92	85.65±1.70*	**76.00**±1.33	**79.28**
GoogleNet	✓	SGA-UAP	62.56	**83.62**	**82.11**	**59.09**	**92.12***	**75.90**
		AT-UAP-S	55.90±0.62	78.71±0.67	76.01±0.45	54.49±0.29	90.82±0.29*	71.19
	✗	**PSP-UAP (Ours)**	**65.22**±0.56	78.43±0.73	79.26±0.73	57.63±0.66	81.43±0.49*	72.39

Table 4. FR (%) of our PSP-UAP and data-dependent UAPs. SGA-UAP [15] and AT-UAP-S [12]. The "data" column indicates whether a dataset was used to train UAPs (data-dependent UAP, ✓) or not (data-free UAP, ✗). * indicates the white-box model.

models to craft the UAPs. As shown in Table 2, PSP-UAP achieves superior results than other data-free attack methods across most models, with performance comparable to, but slightly below, TRM-UAP on GoogleNet. Notably, in strictly black-box settings (excluding the white-box scenario where GoogleNet attacks itself), PSP-UAP achieves an average FR of 70.1% compared to TRM-UAP's 69.6%, demonstrating better transferability.

Extended Evaluation with Additional CNN Models. To further explore the effectiveness of the proposed PSP-UAP, we conduct additional experiments on widely used CNN models, including ResNet50, DenseNet121, MobileNet-v3-Large, and Inception-v3. We compare the attack performance of PSP-UAP with TRM-UAP as shown in Table 3. Note that we limit the comparison to TRM-UAP since the public code for AT-UAP has not yet been released. The results demonstrate that our PSP-UAP consistently outperforms TRM-UAP in terms of FR with a substantial margin. The substantial improvement demonstrates PSP-UAP's strong generalization across diverse CNN models, highlighting the robustness of our approach.

Comparison with Data-dependent UAPs. To verify whether our method effectively alleviates the lack of prior knowledge, we compare our method to state-of-the-art data-dependent universal attacks in the black-box scenario. As shown in Table 4, the FR of white-box attacks is inevitably higher for SGA-UAP and AT-UAP-S, as they fully utilize the target domain dataset. On the other hand, PSP-UAP exhibits superior transferability, as observed in the black-box setting. Our method outperforms by achieving a higher average FR across most models, surpassing data-dependent approaches. Furthermore, even in cases where some FR results are lower, they do not fall significantly behind the data-dependent universal methods. These results indicate that our semantic samples prove to be effective as data for training the UAP, successfully overcoming the limitation posed by the absence of a target domain.

4.3. Ablation Study

We conduct ablation experiments to explore the effectiveness of the proposed pseudo-semantic prior, sample reweighting, and applying input transformation techniques on the semantic samples. In the following experiments, we generate UAPs on ResNet152 and evaluate them across various CNN models (*e.g.*, AlexNet, VGG16, VGG19, ResNet152, and GoogleNet).

Impact of Each Proposed Component. We investigate how each component of PSP-UAP such as pseudo-semantic prior, sample reweighting, and input transformation affects attack performance. For a fair comparison, we assign the number of random noises and semantic samples to 10. As shown in Figure 4, our pseudo-semantic prior method achieves a higher FR rather than random prior when used as an input prior. Additionally, applying sample reweighting to the semantic samples improves the FR in both white-box and black-box attacks. Incorporating input transformation into semantic samples enhances the attack performance. Remarkably, combining both sample reweighting and input transformation yields the greatest overall performance improvement.

Impact of Input Transformation. We evaluate the impact of applying input transformations to semantic samples in our method versus random noise on two data-free UAP methods (*e.g.*, TRM-UAP [16] and GD-UAP [21]), to show that input transformations are effective when semantic information is present. For a fair evaluation, we set the number of samples $N = 10$ for both random noise and semantic samples. As shown in Figure 5, applying input transforma-

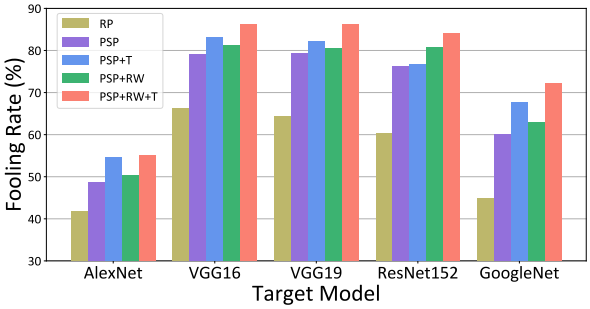

Figure 4. Ablation study on each proposed component in PSP-UAP. RP and PSP refer to training a UAP using random noises and semantic samples drawn from pseudo-semantic prior, respectively. RW and T denote the use of sample reweighting, and input transformation, respectively.

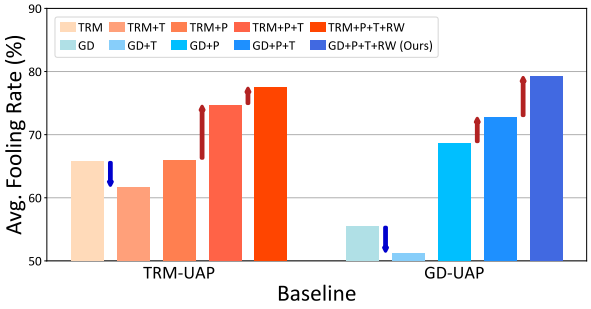

Figure 5. Demonstrating that PSP serves as a universal strategy to other data-free UAP methods. Avgerage fooling rate (%) refers average FR (%) on AlexNet, VGG16, VGG19, ResNet152, and GoogleNet, with UAP crafted on ResNet152. P, T, and RW denote PSP, input transformation, and sample reweighting, respectively.

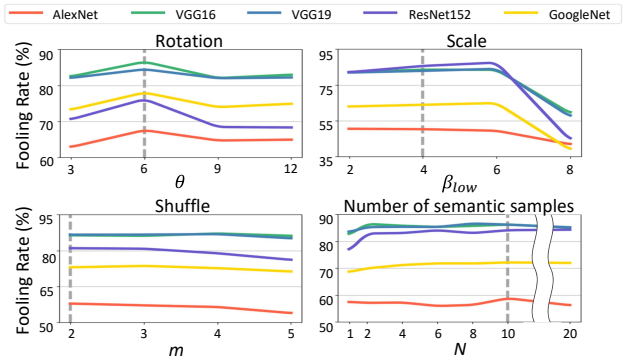

Figure 6. Ablation study on the hyperparameters for input transformation and the number of semantic samples. The hyperparameters used in our experiments are marked with gray dashed line.

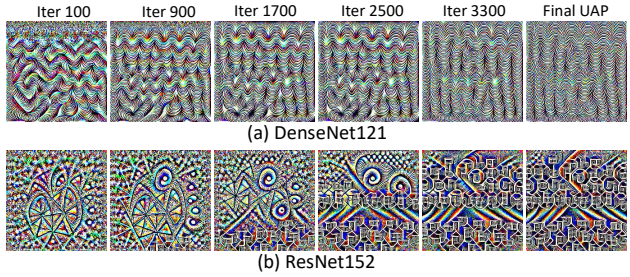

Figure 7. Visualization of UAPs crafted by PSP-UAP during training on DenseNet121 and ResNet152. From left to right, the UAPs are shown at iterations 100, 900, 1700, 2500, 3300, and the final UAP after training. Pixel values are scaled to [0, 255].

tion directly to random noise leads to performance degradation, whereas integrating our pseudo-semantic prior (PSP) significantly boosts performance. Moreover, each baseline achieves its highest FR when combined with our full approach, including sample reweighting. We believe that this improvement results from the richer semantic content provided by our PSP, which allows the UAP to learn a broader range of patterns through input transformation. This experiment also validates our approach as a universal strategy.

Hyperparameter Analysis. We conduct a hyperparameter analysis to evaluate how different settings for input transformation and the number of semantic samples affect our method's attack performance on the ImageNet validation set (see Figure 6). With the exception of extreme scaling values (*e.g.*, 8), the proposed method demonstrates stable performance across various hyperparameter values. Furthermore, performance remains consistent across different the number of semantic samples; notably, even with $N = 1$, method outperforms other data-free approaches. Note that all hyperparameters used are chosen based on the ImageNet train set (See Appendix for further details).

UAP Visualization. In Figure 7, we visualize the UAPs at each training iteration to verify that the pseudo-semantic prior, constructed from UAPs in the training phase, captures a variety of inherent patterns. We observe that the UAPs contain more diverse patterns in the early stages of training. We believe this variability allows us to obtain semantic samples with a broad range of patterns, through which the UAP learns diverse semantic representations.

5. Conclusion

In this paper, we proposed a novel data-free universal attack method that leverages UAPs as a prior enriched with semantic information. This approach allows us to directly draw semantic samples from the pseudo-semantic prior, overcoming the lack of target domain knowledge. To further enhance transferability, we applied input transformation methods to these semantic samples. Additionally, we introduced sample reweighting to ensure a balanced attack across semantic samples. We demonstrated the exceptional transferability of our method by comparing PSP-UAP with both data-free and data-dependent universal attack approaches across various CNN models on the ImageNet validation dataset.

Acknowledgements. This work was supported by the IITP grants (RS-2019-II191842 (3%), RS-2021-II212068 (2%), RS-2022-II220926 (50%)) funded by MSIT, and the GIST-MIT Research Collaboration grant (45%) funded by GIST, Korea.

References

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *ECCV Workshops*, 2016.

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, 2017.

[3] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018.

[4] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*, 2019.

[5] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *CVPR*, 2018.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[8] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *CVPR*, 2019.

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[11] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018.

[12] Maosen Li, Yanhua Yang, Kun Wei, Xu Yang, and Heng Huang. Learning universal adversarial perturbation by adversarial example. In *AAAI*, 2022.

[13] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *ICLR*, 2020.

[14] Hong Liu, Rongrong Ji, Jie Li, Baochang Zhang, Yue Gao, Yongjian Wu, and Feiyue Huang. Universal adversarial perturbation via prior driven uncertainty approximation. In *ICCV*, 2019.

[15] Xuannan Liu, Yaoyao Zhong, Yuhang Zhang, Lixiong Qin, and Weihong Deng. Enhancing generalization of universal adversarial perturbation through gradient aggregation. In *ICCV*, 2023.

[16] Yiran Liu, Xin Feng, Yunlong Wang, Wu Yang, and Di Ming. Trm-uap: Enhancing the transferability of data-free universal adversarial perturbation via truncated ratio maximization. In *ICCV*, 2023.

[17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[18] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016.

[19] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, 2017.

[20] KR Mopuri, U Garg, and R Venkatesh Babu. Fast feature fool: A data independent approach to universal adversarial perturbations. In *BMVC*, 2017.

[21] Konda Reddy Mopuri, Aditya Ganeshan, and R Venkatesh Babu. Generalizable data-free objective for crafting universal adversarial perturbations. *TPAMI*, 41(10), 2018.

[22] Konda Reddy Mopuri, Utkarsh Ojha, Utsav Garg, and R Venkatesh Babu. Nag: Network for adversary generation. In *CVPR*, 2018.

[23] Konda Reddy Mopuri, Phani Krishna Uppala, and R Venkatesh Babu. Ask, acquire, and attack: Data-free uap generation using class impressions. In *ECCV*, 2018.

[24] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *CVPR*, 2018.

[25] J Redmon. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 39(6):1137–1149, 2016.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115, 2015.

[29] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[30] Ali Shafahi, Mahyar Najibi, Zheng Xu, John Dickerson, Larry S Davis, and Tom Goldstein. Universal adversarial training. In *AAAI*, 2020.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[32] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

[33] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[36] Kunyu Wang, Xuanran He, Wenxuan Wang, and Xiaosen Wang. Boosting adversarial transferability by block shuffle and rotation. In *CVPR*, 2024.

[37] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *ICCV*, 2021.

[38] Xiaosen Wang, Zeliang Zhang, and Jianping Zhang. Structure invariant transformation for better adversarial transferability. In *ICCV*, 2023.

[39] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In *IJCAI*, 2018.

[40] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, 2019.

[41] Chaoning Zhang, Philipp Benz, Adil Karjauv, and In So Kweon. Data-free universal adversarial perturbation and black-box attack. In *ICCV*, 2021.

[42] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, 129:3069–3087, 2021.

[43] Rongyi Zhu, Zeliang Zhang, Susan Liang, Zhuo Liu, and Chenliang Xu. Learning to transform dynamically for better adversarial transferability. In *CVPR*, 2024.