Enhancing Adversarial Transferability with Checkpoints of a Single Model's Training

Shixin Li[1]*, Chaoxiang He[2]*, Xiaojing Ma[1], Bin Benjamin Zhu[3], Shuo Wang[2], Hongsheng Hu[4], Dongmei Zhang[3], Linchen Yu[1]†
[1]Huazhong University of Science and Technology [2]Shanghai Jiao Tong University
[3]Microsoft Corporation [4]University of Newcastle, Australia
{shixinli, lindahust, linchenyu}@hust.edu.cn, {hechaoxiang, wangshuosj}@sjtu.edu.cn,
{binzhu, dongmeiz}@microsoft.com, hongsheng.hu@newcastle.edu.au

Abstract

Adversarial attacks threaten the integrity of deep neural networks (DNNs), particularly in high-stakes applications. In this paper, we present a novel black-box adversarial attack that leverages the diverse checkpoints generated during a single model's training trajectory. Unlike conventional ensemble attacks that require multiple surrogate models with diverse architectures, our approach exploits the intrinsic diversity captured over different training stages of a single surrogate model. By decomposing the learned representations into task-intrinsic and task-irrelevant components, we employ an accuracy gap-based selection strategy to identify checkpoints that predominantly capture transferable, task-intrinsic knowledge. Extensive experiments on ImageNet and CIFAR-10 demonstrate that our method consistently outperforms traditional ensemble attacks in terms of transferability, even under resource-constrained and practical settings. This work offers a resource-efficient solution for crafting highly transferable adversarial examples and provides new insights into the dynamics of adversarial vulnerability.

1. Introduction

Adversarial attacks pose a serious threat to deep neural network (DNN) models, especially in high-stakes applications such as autonomous driving, healthcare, and financial systems [5, 6]. These attacks often involve subtle perturbations that can lead to erroneous predictions [7, 24]. The threat is exacerbated by the transferability property of adversarial examples: crafted on a surrogate model, these examples can deceive target models with the same or different architectures. This transferability enables black-box adversarial attacks, where adversaries can compromise models without any prior knowledge of their parameters or structures, thus posing an especially severe threat.

Crafting effective adversarial examples in a black-box setting is inherently more challenging than in white-box scenarios. To improve transferability, model ensemble attacks [1, 18, 29] use multiple surrogate models with diverse architectures to identify common vulnerabilities and generate adversarial examples that generalize to unseen models. However, ensemble attacks have notable limitations. For novel or specialized datasets, pre-trained models are often scarce, forcing the extensive time and computational resources required to train multiple models from scratch. This reliance can be impractical under limited resources or significantly drive up the cost of a black-box adversarial attack.

This raises an intriguing **research question**: *Can a single model's training trajectory capture the diversity of different architectures to match the transferability of ensemble attacks?* At first glance, this seems unlikely since a training process confined to one architecture may not reflect the structural differences across models, implying that adversarial examples from one model's checkpoints might not transfer effectively to models with different architectures.

However, checkpoints within a model's training trajectory can exhibit the diversity seen in ensemble attacks. The knowledge learned during training can be split into *task-intrinsic knowledge*, which is essential for solving the task and generalizes across models, and *task-irrelevant knowledge*, which is tied to specific training data or model structure and often fails to transfer. Ensemble attacks leverage task-intrinsic knowledge across different architectures to craft adversarial examples that target common vulnerabilities, thereby enabling their transferability to new, unseen architectures.

We hypothesize that task-intrinsic knowledge embedded in a model's checkpoints can be harnessed to produce adversarial examples that strongly transfer across architectures.

*Equal Contribution
†Corresponding Author

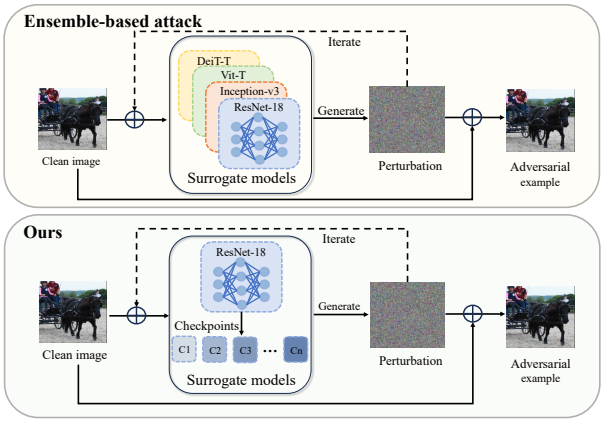


Figure 1. Comparison between our approach (using checkpoints from a single surrogate model's training) and traditional ensemble attacks (using surrogate models with diverse architectures).

As training progresses, a model advances from random initialization, through a phase of capturing task-intrinsic knowledge, to eventual overfitting. By generating adversarial examples from various checkpoints along this trajectory, we replicate the diversity effect of ensemble attacks without needing multiple surrogate models of different architectures.

To guide checkpoint selection, we use the accuracy gap between training and validation data as an indicator of overfitting and reliance on task-irrelevant knowledge. We hypothesize that, once a meaningful validation accuracy is reached, the transferability enhancement from checkpoints declines as the accuracy gap increases, with heavily overfitted models even reducing transferability. This hypothesis is empirically validated later.

Based on these hypotheses, we propose an innovative black-box adversarial attack that leverages a single surrogate model's training checkpoints to generate highly transferable adversarial examples. Our method employs an accuracy gap-based importance metric alongside an importance-based strategy to choose the most effective checkpoints. Figure 1 compares our approach with traditional ensemble attacks.

Extensive experiments on benchmark datasets, including ImageNet and CIFAR-10, show that our method consistently outperforms ensemble attacks in transferability. This is achieved by using checkpoints that are less overfitted and more reflective of task-intrinsic knowledge, unlike ensemble attacks that often rely on overfitted models. Moreover, our method remains robust under constraints such as limited model capacity, reduced training data, and non-overlapping or differently distributed surrogate datasets. For example, using only 10% (5,000 images) of CIFAR-10's training data, our method achieves an average attack success rate of 98.4% against target models.

Our major contributions are as follows:

1. **Novel Approach using Training Checkpoints**: We introduce a pioneering method that leverages multiple checkpoints from a single model's training trajectory, effectively replicating the diversity of traditional ensemble attacks without needing surrogate models of different architectures.
2. **Accuracy Gap-based Checkpoint Selection**: We propose an accuracy gap metric to quantify the extent of task-irrelevant knowledge and strategically select training checkpoints for generating highly transferable adversarial examples.
3. **Superior Transferability**: Extensive experiments on benchmark datasets show that our method consistently outperforms traditional model ensemble attacks, achieving higher success rates across diverse target models.
4. **Improved Practicality**: Our approach exhibits robust performance under real-world constraints—including limited model capacity, reduced training data, and variations in surrogate dataset distributions—ensuring its applicability in diverse settings.

2. Preliminaries

2.1. Background and Related Work

Adversarial Examples. Adversarial examples are inputs that mislead DNN models by adding small, often imperceptible perturbations to legitimate inputs. In white-box settings, where the target model's parameters are known, they are crafted by computing the gradient of the adversarial loss with respect to the input, as in *FGSM* [7] and *Iterative FGSM* (I-FGSM) [15].

Transferability of Adversarial Examples. Adversarial examples generated on white-box *surrogate models* can transfer to black-box *target models* [21] by exploiting their shared task-intrinsic knowledge. This enables effective attacks without accessing the target model's parameters or structure. However, overfitting to the surrogate model can reduce transferability. Techniques such as *Momentum Iterative FGSM* (MI-FGSM) [2], *Nesterov Iterative FGSM* (NI-FGSM) [17], *Diverse-Input* (DI) [28], and *Translation-Invariant* (TI) [3] improve transferability by incorporating momentum and input transformations.

Model Ensemble Attacks. To enhance transferability, model ensemble attacks combine multiple models of different architectures. In [18], for each model f_i in the ensemble, the gradient of the attack loss is computed as $\mathbf{g}_i = \nabla_{\mathbf{x}_t^{\text{adv}}} \mathcal{L}\left(f_i\left(\mathbf{x}_t^{\text{adv}}\right), y\right)$ and then aggregated, typically by averaging: $\mathbf{g}_{\text{ensemble}} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{g}_i$, where M is the number of models in the ensemble. *Adaptive Ensemble Attack* (AdaEA) [1] dynamically adjusts the averaging weights of surrogate models based on their contribution to the attack's effectiveness. *Stochastic Variance Reduced Ensemble* (SVRE) [29] uses SVRG to reduce gradi-

ent variance across surrogate models. Additionally, methods like *Ghost Networks* [16] and *Large Geometric Vicinity* (LGV) [8] create multiple variants from a single model—by perturbing dropout layers or applying high-learning-rate fine-tuning—to lessen reliance on diverse architectures.

2.2. Limitations of Model Ensemble Attacks

Existing ensemble attacks face two main limitations:

Limited Model Availability. These attacks often require diverse model architectures, but pre-trained models may be scarce, especially for niche datasets. Training multiple models from scratch is costly and time-intensive, making such attacks impractical in resource-constrained settings or cost-prohibitive.

Overfitting to Task-irrelevant Knowledge. Well-trained models often overfit to specific training data and architectural nuances. When used as surrogate models, this overfitting can reduce the transferability of adversarial examples, limiting their effectiveness against models with significantly different architectures.

3. Threat Model and Notation

Threat Model. We follow the standard approach for model ensemble attacks, assuming a black-box setting where adversaries cannot access the target model's parameters or architecture. Adversaries have an auxiliary dataset for training a surrogate model, which may be identical to the target model's data, share its distribution without overlapping, or follow a different distribution. Additionally, adversaries have sufficient computational resources to train a surrogate model and generate adversarial examples from multiple checkpoints along its training trajectory.

Notation. Let M be a model with initial parameters θ_0, trained on dataset D for T epochs. At each epoch t, its parameters θ_t are updated (e.g., via SGD or Adam) to improve performance. At predefined intervals, the state θ_t is saved as a checkpoint. The set of checkpoints, denoted by Θ, comprises N states collected during training.

4. Our Proposed Method

Our research question is: Can a single model's training trajectory capture the diversity of model ensemble attacks to enhance adversarial transferability? We propose leveraging a single model's training checkpoints to extract task-intrinsic knowledge—which generalizes across architectures—to craft transferable adversarial examples without requiring models of diverse architectures.

During training, a model acquires two types of knowledge: *task-intrinsic*, which transfers across models, and *task-irrelevant*, which is specific to the training data or architecture and limits transferability. In early training, the model learns foundational, transferable patterns; later, it captures more task-irrelevant details and overfits, reducing generalization.

Thus, checkpoints from the pre-overfitting phase should yield more transferable adversarial examples. We hypothesize (the **first hypothesis**) that adversarial examples derived from these checkpoints can exploit shared vulnerabilities across architectures. Accordingly, our method uses an overfitting indicator to measure task-irrelevant knowledge and a checkpoint selection strategy to identify the most effective checkpoints for transferability. These selected checkpoints are then used to generate adversarial examples, similar to model ensemble attacks.

4.1. Overfitting Indicator: Accuracy Gap

To measure overfitting, we use the accuracy gap between the training and validation sets as a metric. For each checkpoint f_{θ_t}, the accuracy gap is defined as

$$acc_t^G = acc_t^T - acc_t^V \tag{1}$$

where acc_t^T and acc_t^V are the training and validation accuracies, respectively. This gap reflects overfitting: it starts small and grows as the model shifts from learning task-intrinsic to task-irrelevant knowledge.

We hypothesize (the **second hypothesis**) that, once validation accuracy reaches a substantial level, checkpoints with smaller accuracy gaps yield more transferable adversarial examples. Lower overfitting reduces task-irrelevant knowledge, enhancing adversarial transferability by focusing on generalizable features that exploit shared vulnerabilities across different model architectures.

4.2. Importance of Checkpoints

The importance p_t of checkpoint t for transferability is defined based on its accuracy gap:

$$p_t = \frac{e^{-k \cdot acc_t^G}}{\sum_{t=1}^{N} e^{-k \cdot acc_t^G}} \tag{2}$$

where k is a scaling factor, acc_t^G is the accuracy gap of the t-th checkpoint, and N is the total number of checkpoints. This measure assigns higher weights to checkpoints with smaller accuracy gaps, as they likely capture more generalizable features and improve adversarial transferability.

To validate the second hypothesis, we examine the relationship between accuracy gap and transferability. We rank checkpoints by importance in descending order, $\{p_1, p_2, \ldots, p_N\}$, select subsets of the most important checkpoints, and evaluate the average attack success rate of adversarial examples generated from these subsets across multiple target models on ImageNet.[1]

[1]The surrogate model is ResNet-18 on ImageNet, with target models including 4 CNNs (ResNet-50, WRN101-2, BiT-101, Inception-v4) and 4 ViTs (LeViT, Visformer-S, PiT-S, Swin-S).

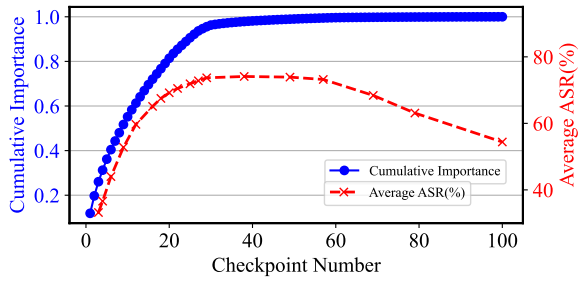


Figure 2. Cumulative importance and average attack success rates of using ensembles of varying numbers of top checkpoints.

Algorithm 1 Selecting Representative Checkpoints and Applying Multi-Iteration Ensemble Attack

Input: Set of checkpoints Θ, input $\mathbf{x}$, true label y, adversarial cross-entropy loss J, perturbation size ϵ, total iterations I, step size α.

Output: Adversarial example $\mathbf{x}^{\text{adv}}$ with $\|\mathbf{x}^{\text{adv}} - \mathbf{x}\|_\infty \leq \epsilon$.

Select a representative subset $\Theta' \subseteq \Theta$ using Eqs.1-3
Initialize $\mathbf{x}_1^{\text{adv}} \leftarrow \mathbf{x}$
for iteration $i = 1$ to I **do**
  $\mathbf{g}_{\text{ensemble}} \leftarrow 0$
  for each $\theta' \in \Theta'$ **do**
    Compute gradient $\mathbf{g}_{\theta'} = \nabla_{\mathbf{x}} J(\theta', \mathbf{x}_i^{\text{adv}}, y)$
    $\mathbf{g}_{\text{ensemble}} \leftarrow \mathbf{g}_{\text{ensemble}} + \mathbf{g}_{\theta'}$
  end for
  $\mathbf{g}_{\text{ensemble}} \leftarrow \mathbf{g}_{\text{ensemble}} / |\Theta'|$
  $\mathbf{x}_{i+1}^{\text{adv}} \leftarrow \text{clip}_{\mathbf{x}}^{\epsilon}(\mathbf{x}_i^{\text{adv}} + \alpha \cdot \text{sign}(\mathbf{g}_{\text{ensemble}}))$
end for
return $\mathbf{x}^{\text{adv}} = \mathbf{x}_I^{\text{adv}}$

As shown in Fig. 2, selecting the top few checkpoints achieves substantial cumulative importance. For instance, the top 40 checkpoints account for nearly 100% of the cumulative importance. We assess the attack success rate of adversarial examples generated using ensembles of varying numbers of top checkpoints. The results show that incorporating high-importance checkpoints significantly enhances transferability, whereas adding low-importance checkpoints (e.g., the last 40 checkpoints after reaching 98% cumulative importance) reduces transferability.

These findings confirm the second hypothesis that lower overfitting levels, as indicated by higher importance weights, enhance adversarial transferability, while high-overfitting checkpoints degrade it. Thus, selecting checkpoints with minimal overfitting is crucial for optimizing adversarial transferability.

4.3. Checkpoint Selection Strategy

As shown in Fig. 2, a checkpoint's importance score quantifies its task-intrinsic knowledge and overfitting, indicating its ability to generate transferable adversarial examples. Based on this, we propose the following selection strategy. First, we discard checkpoints with validation accuracy below a threshold α to ensure sufficient utility. Next, we rank the remaining checkpoints by importance and iteratively select the most important ones until their cumulative importance reaches a target $K\%$. Formally, the selected set is:

$$\Theta' = \{\theta_1, \theta_2, \ldots, \theta_i \mid \sum_{j=1}^{i} p_j \leq K\%,\ acc_j^V \geq \alpha, \forall j\} \quad (3)$$

where p_j is the importance of the j-th checkpoint (in descending order), and $K\%$ is the target cumulative importance. This strategy prioritizes checkpoints with smaller accuracy gaps to enhance adversarial transferability.

Once selected, these checkpoints are used in a multi-iteration ensemble attack to generate adversarial examples, as detailed in Algorithm 1.

5. Experiments

5.1. Experimental Setting

Datasets. We follow prior works [1, 29] and evaluate our method on two benchmark datasets: an ImageNet-compatible dataset[2] and CIFAR-10 [14].

Models. For adversarial transferability evaluation, we use diverse target models, including CNNs—ResNet-50 (RN-50) [9], WideResNet-101 (WRN-101) [31], BiT-M-R101 (BiT-101) [13], Inception-v4 (Inc-v4) [26], DenseNet-121 (DN-121) [11], VGG-16 [23], and GoogLeNet [25]—and vision transformers (ViTs): ViT-Base (ViT-B) [4], DeiT-Base (DeiT-B) [27], Swin-Base (Swin-B) [19], and PiT-Base (PiT-B) [10]. We evaluate ensemble-based defense targets built using standard training, ADP [20], GAL [12], and DVERGE [30] on an ensemble, denoted as *ens3-RN-20*, comprising three ResNet-20 models.

For ensemble attacks, we follow [1] with four surrogates on ImageNet—ResNet-18 (RN-18) [9], Inception-v3 (Inc-v3) [26], ViT-Tiny (ViT-T) [4], and DeiT-Tiny (DeiT-T) [27]—and four on CIFAR-10: ResNet-18, ResNet-34, Inception-v3, and MobileNet-V2 [22]. For single-model methods (ours and LGV [8]), we primarily use ResNet-18 but also ViT-Tiny. When both are used, checkpoints (ours) or variants (LGV) are selected independently and then combined to generate adversarial examples.

Attack Baselines. We compare our approach with three ensemble attacks: Ens [18], LGV [8], and AdaEA [1], using their original ensemble configurations as described above, unless stated otherwise. LGV originally employs a longitudinal ensemble that randomly samples one model per iteration. For fair comparison, we configure LGV to use all

[2] https://github.com/cleverhans-lab/cleverhans/tree/master/cleverhans_v3.1.0/examples/nips17_adversarial_competition/dataset

Table 1. ASRs (%) on ImageNet for CNN and ViT target models using different surrogate configurations for LGV and our method (CNN-only, ViT-only, and CNN+ViT), while Ens and AdaEA use both CNN and ViT surrogates. Adversarial examples are generated by I-FGSM. Bold indicates the best performance.

Attack	Surrogate Models	CNN					ViT				
		RN-50	WRN-101	BiT-101	Inc-v4	Average	ViT-B	DeiT-B	Swin-B	PiT-B	Average
Ens	RN-18, Inc-v3,	54.0	52.0	46.6	56.9	52.4	38.3	59.2	21.9	30.8	37.6
AdaEA	ViT-T, DeiT-T	55.9	52.9	46.1	53.5	52.1	39.8	58.8	22.6	32.4	38.4
LGV	RN-18	86.2	92.8	78.2	83.4	85.2	21.7	33.6	25.7	35.3	29.1
	ViT-T	31.9	40.5	36.0	43.7	38.0	28.0	48.2	15.7	22.4	28.6
	Res-18, ViT-T	68.9	80.6	66.0	71.5	71.8	24.5	40.9	20.0	29.7	28.8
Ours	RN-18	90.0	94.9	83.4	88.2	89.1	26.0	41.4	26.8	39.1	33.3
	ViT-T	35.8	41.8	39.5	45.6	40.7	49.8	78.8	22.3	32.5	45.9
	RN-18, ViT-T	**91.4**	**95.5**	**87.7**	**89.6**	**91.1**	**60.2**	**84.1**	**45.1**	**57.8**	**61.8**

Table 2. ASRs (%) on CIFAR-10 target CNN models using ResNet-18 as the surrogate for LGV and ours, and four CNN surrogates (see Section 5.1) for Ens and AdaEA. Adversarial examples are generated with I-FGSM. Bold indicates best performance.

Attack	Target Model				
	RN-50	VGG-16	DN-121	GoogleNet	Average
Ens	93.1	92.8	91.6	89.1	91.7
AdaEA	93.8	94.4	92.4	91.1	92.9
LGV	59.4	65.2	55.5	69.1	62.3
Ours	**99.9**	**100**	**100**	**99.8**	**99.9**

40 model variants per iteration, resulting in an average attack success rate 3.3% higher on ImageNet than its original longitudinal ensemble.

Other Configurations. We follow the baseline methods to evaluate the adversarial transferability using untargeted black-box attacks. Adversarial examples are generated under an L_∞ bound of $\epsilon = 16/255$ over $T = 10$ iterations (step size ϵ/T). Surrogate models are trained using SGD for 100 epochs, starting with an initial learning rate of 0.1, which is decayed by a factor of 0.1 every 30 epochs. The importance scaling factor, k, is set to 1, with a target cumulative importance of $K = 98\%$ and a validation accuracy threshold, α, set at 60% of the final performance. All experiments are implemented in PyTorch and executed on two NVIDIA GeForce RTX 3090 GPUs.

5.2. Adversarial Transferability

We assess adversarial transferability by measuring the attack success rate (ASR) on ImageNet and CIFAR-10 target models, using surrogate models trained on the same data as their respective targets. We evaluate both standard and robustness-enhanced targets as detailed in Section 5.1.

5.2.1. Standard Target Models

Tables 1 (ImageNet) and 2 (CIFAR-10) report our results. Key observations include:

1. **High Transferability within Homogeneous Architectures.** Our method achieves high transferability when the surrogate and target share the same architecture (CNNs or ViTs), i.e., a *homogeneous architecture*. For example, with ResNet-18 as the surrogate on ImageNet, our method outperforms LGV by 3.9% and both Ens and AdaEA by 36.7% and 37%, respectively, on CNN targets. Similarly, using ViT-Tiny as the surrogate yields superior performance on ViT targets. A comparable trend is observed on CIFAR-10, where our method surpasses the three baseline attacks by 8.2%, 7.0%, and 37.6% in average ASR on four target models. These results demonstrate that our method effectively captures generalizable, task-intrinsic features within homogeneous architectures.

2. **Lower Transferability across Heterogeneous Architectures.** Transferability between heterogeneous architectures (e.g., CNNs to ViTs or vice versa) is notably lower due to differences in gradient flow [1]. For example, using ResNet-18 as the surrogate for ViT targets leads to a performance drop—our method falls behind ensemble approaches. Similarly, when ViT-Tiny serves as the surrogate for CNN targets, our method underperforms compared to ensemble methods. Specifically, with ViT-Tiny as the surrogate, our method outperforms Ens and AdaEA by 8.3% and 7.5% on ViT targets but performs slightly worse on CNN targets. Note that both Ens and AdaEA employ both CNN and ViT surrogates.

3. **Combining Heterogeneous Surrogate Models Enhances Transferability.** Aggregating heterogeneous surrogates significantly boosts our method's transferability. With both ResNet-18 and ViT-Tiny, our method achieves an average ASR of 91.1% on CNN targets and 61.8% on ViT targets on ImageNet—outperforming Ens by 38.7% and 24.2%, and AdaEA by 39.0% and 23.4%, respectively. Notably, LGV shows no improvement under this configuration, reinforcing that our method more effectively captures task-intrinsic knowledge across diverse architectures.

5.2.2. Robustness-Enhanced Target Models

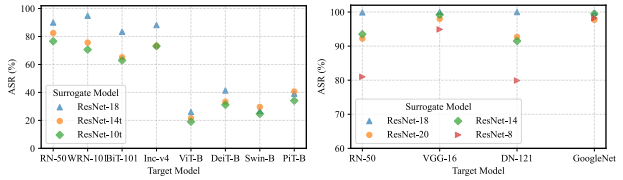
Table 3 summarizes ASRs on CIFAR-10 for four ensemble-based defense target models. When averaged over these de-

Table 3. ASRs (%) on CIFAR-10 against ensemble-based defense models. Bold indicates the best performance.

Defense	ens3-RN-20				
	Standard	ADP	GAL	DVERGE	Average
Ens	85.9	80.5	90.3	82.3	84.8
AdaEA	86.0	81.9	90.4	82.3	85.2
LGV	87.7	83.3	89.7	86.2	86.7
Ours	**89.7**	**89.9**	**90.2**	**91.2**	**90.3**

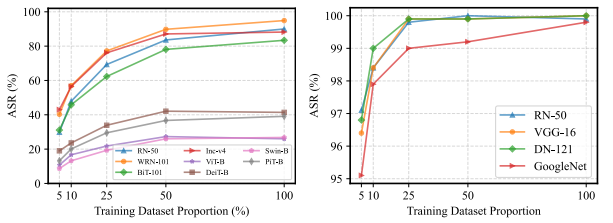
fenses, our method outperforms Ens, AdaEA, and LGV by 5.5%, 5.1%, and 3.6%, respectively. These results underscore the strong capability of our approach to bypass diverse defensive mechanisms.

Takeaway 1. Our method achieves superior adversarial transferability on both standard and robustness-enhanced target models, particularly within homogeneous architectures. Transferability across heterogeneous architectures (e.g., CNNs to ViTs and vice versa) remains challenging but improves significantly with heterogeneous surrogate aggregation. Our approach leverages heterogeneous surrogates more effectively than traditional model ensemble attacks. By prioritizing checkpoints that capture task-intrinsic knowledge and reduce task-irrelevant information, our method significantly enhances adversarial transferability across diverse target models.



(a) ImageNet (b) CIFAR-10

Figure 3. Impact of surrogate model capacity on ASRs.



(a) ImageNet (b) CIFAR-10

Figure 4. Impact of training dataset size on attack effectiveness.

5.3. Performance Under More Practical Scenarios

We assess our method's efficiency under more practical conditions, examining how surrogate model capacity, training dataset size, and dataset discrepancies between surrogate and target models impact attack effectiveness. This analysis provides insights into our method's practicality in resource-constrained settings.

Table 4. ASRs (%) for surrogate and target models trained on non-overlapping datasets or datasets with different distributions.

Training Dataset	Target Model				
	RN-50	VGG-16	DN-121	GoogleNet	Average
Overlap	99.9	100	100	99.8	99.9
Non-Overlap	98.4	99.0	98.4	97.9	98.4
Different Distribution	98.2	98.8	97.8	97.9	98.2

5.3.1. Surrogate Model Capacity

We assess the impact of surrogate model capacity on attack performance using ResNet models of varying complexities. For CIFAR-10, we compare ResNet-8, ResNet-14, ResNet-20, and ResNet-18, while for ImageNet, we use ResNet-10t, ResNet-14t, and ResNet-18. As shown in Fig. 3, overly simple models like ResNet-8 and ResNet-10t degrade attack performance. However, moderately larger models, such as ResNet-14 and ResNet-14t, maintain strong attack effectiveness without a significant performance drop.

5.3.2. Training Dataset Size of the Surrogate Model

We investigate the feasibility of training the surrogate model with a limited amount of data. To assess how the amount of training data affects attack performance, we vary the training set size from 5% to 100% of the full dataset. For CIFAR-10, the surrogate model is trained with between 500 and 50,000 images, whereas for ImageNet, it is trained with between 60,000 and 1,200,000 images. As illustrated in Fig. 4, even when using only 10% of CIFAR-10's training data (5,000 images), our approach achieves an impressive average ASR of 98.4%. On ImageNet, training with 600,000 images yields an average ASR of 58.9%, substantially outperforming baseline methods.

These findings demonstrate that high attack performance can be maintained while reducing both the size of the training dataset and the complexity of the surrogate model. Consequently, our method offers a cost-efficient and practical solution that minimizes computational requirements without sacrificing effectiveness.

5.3.3. Distinct Training Datasets

In Section 5.2, we assumed that the target model's dataset was used to train the surrogate models. Here, we investigate the impact of using different training data for surrogate model training by considering two experimental setups:

1. **Same Distribution, Non-Overlapping.** In this setup, both the surrogate and target models are trained on data drawn from the same distribution, but the datasets do not overlap. For this experiment, the surrogate model is trained on 5,000 images from the CIFAR-10 training set, while the remaining 45,000 images are used to train the target model.

Table 5. ASRs (%) on ImageNet using various adversarial example generation methods with ResNet-18 as the surrogate model. Bold numbers indicate the best performance, and Δ represents the improvements over I-FGSM.

Base	CNN					ViT				
	RN-50	WRN-101	BiT-101	Inc-v4	Average (Δ)	ViT-B	DeiT- B	Swin-B	PiT-B	Average (Δ)
I-FGSM	90.0	94.9	83.4	88.2	89.1	26.0	41.4	26.8	39.1	33.3
MI-FGSM	95.9	97.9	91.9	94.9	95.2 (+6.1)	46.2	62.1	45.6	59.0	53.2 (+19.9)
DIM-FGSM	96.8	98.8	95.6	98.2	97.4 (+8.3)	55.2	71.7	50.5	66.6	61.0 (+27.7)

Table 6. ASRs (%) on ImageNet using various checkpoint selection strategies with ResNet-18 as the surrogate model. Adversarial examples are generated using I-FGSM. Bold numbers highlight the best results.

Selection Strategy	CNN					ViT				
	RN-50	WRN-101	BiT-101	Inc-v4	Average	ViT-B	DeiT-B	Swin-B	PiT-B	Average
Random	68.9	80.9	59.5	64.8	68.5	15.0	24.7	16.8	24.4	20.2
Uniform	85.2	89.6	73.0	75.2	80.8	20.4	32.6	22.0	36.0	27.8
Importance-based	**90.0**	**94.9**	**83.4**	**88.2**	**89.1**	**26.0**	**41.4**	**26.8**	**39.1**	**33.3**

Table 7. ASRs (%) on ImageNet using the vertical ensemble method with 4 models per iteration and ResNet-18 as the surrogate model. Adversarial examples are generated using I-FGSM.

Attack	CNN					ViT				
	RN-50	WRN-101	BiT-101	Inc-v4	Average	ViT-B	DeiT-B	Swin-B	PiT-B	Average
LGV	77.3	86.9	68.6	73.8	76.7	16.1	28.1	18.4	26.9	22.4
Ours	80.6	90.4	74.2	83.2	82.1	20.2	33.0	21.3	32.5	26.8

As discussed in Section 5.3.2 and shown in Fig. 4, 5,000 images are sufficient for our method to train a surrogate model that achieves strong attack performance on CIFAR-10.

2. **Different Distributions.** In this case, a distribution shift exists between the datasets used for the surrogate and target models. Specifically, the target model is trained on CIFAR-10, while the surrogate model is trained on the ImageNet portion of the CINIC-10 dataset, resulting in a significant distribution difference between the two datasets.

The results for both scenarios are presented in Table 4. They demonstrate that using non-overlapping datasets, or datasets from different distributions, has a minimal impact on attack performance. This finding underscores the robustness of our method, indicating that it remains effective even when the surrogate and target models are trained on distinct datasets.

Takeaway 2. Our method proves both efficient and effective in practical, real-world scenarios. By employing surrogate models with reduced complexity, training on smaller datasets, or using non-overlapping and differently distributed datasets, our approach consistently maintains high attack performance. This robustness makes the technique well-suited for deployment in settings with limited computational resources or data, further enhancing its real-world applicability.

5.4. Further Analysis

In this subsection, we assess the attack performance of our method under additional variations of settings.

5.4.1. Adversarial Example Generation Methods

We have evaluated our method using I-FGSM for generating adversarial examples. Notably, MI-FGSM [2] and DI [28] are known to generate more transferable adversarial examples than I-FGSM. To assess the benefits of these stronger methods within our approach, we compare the performance of MI-FGSM and the combination of MI-FGSM and DI (DIM-FGSM) against I-FGSM on the ImageNet dataset, using ResNet-18 as the surrogate model. As shown in Table 5, replacing I-FGSM with MI-FGSM significantly boosts our method's ASR, yielding an average improvement of 6.1% on CNN target models and 19.9% on ViT target models. When DIM-FGSM is used in place of I-FGSM, the average ASR improvements are even more significant, reaching 8.3% on CNN target models and 27.7% on ViT target models. These results demonstrate that integrating stronger adversarial example generation methods can further enhance the transferability of our adversarial attacks.

5.4.2. Checkpoint Selection Strategies

We also assess the impact of different checkpoint selection strategies, including our proposed importance-based approach alongside random selection using Gaussian and

uniform distributions. Experiments using ResNet-18 as the surrogate model on the ImageNet dataset (see Table 6) reveal that the average ASRs for the Gaussian-random and uniform-random strategies across the eight target models are 48.9% and 54.3%, respectively, while our importance-based strategy achieves 61.2%. This result clearly demonstrates the advantage of our selection strategy in enhancing overall attack performance.

5.4.3. Reduced Memory Usage

Ensembling multiple checkpoints to generate adversarial examples generally increases memory consumption. To address this issue, we propose a vertical ensemble method that significantly reduces memory requirements. Leveraging an importance-based selection strategy, we randomly select only 4 checkpoints per iteration—aligning with the number of models used in the Ens and AdaEA methods in Table 1. Table 7 presents the ImageNet results for both LGV and our method when applying this vertical ensemble setting with ResNet-18 as the surrogate model. Although the average ASRs for both LGV and our method decrease—with our method achieving an average ASR of 82.1% on CNN target models and 26.8% on ViT target models—our approach still outperforms baseline methods for homogeneous target models, as shown in Table 7 (for LGV and our method) and Table 1 (for Ens and AdaEA). This demonstrates a favorable trade-off between reduced memory usage and overall performance.

5.4.4. Using the Same Number of Models in the Ensemble

In previous experiments, Ens and AdaEA used significantly fewer models in their ensembles than LGV and our method. For example, as shown in Table 2 for CIFAR-10, Ens and AdaEA utilized four models—one from each architecture—while LGV and our method employed 40 variants or checkpoints from a single surrogate model. Here, we compare their performance when the ensemble sizes are equal. In this experiment, we collected 10 well-trained models for each of the four architectures, expanding the ensembles for Ens and AdaEA to 40 models, denoted as Ens+ and AdaEA+. Table 8 presents the experimental results on CIFAR-10. We observe that Ens+ and AdaEA+ improve their average ASR over the original Ens and AdaEA by 6.4% and 0.2%, respectively, yet our method still outperforms Ens+ and AdaEA+ by 1.7% and 6.8%, respectively. These findings indicate that adversarial examples generated from a single model's checkpoints are more generalizable than those produced by a traditional ensemble of equally well-trained models.

5.4.5. Training Hyperparameters

Finally, we examine the impact of using different optimizers and learning rate schedulers in training the surrogate model on our method's performance. We evaluate three experimental setups with varying hyperparameters on CIFAR-10, and Table 9 presents the average ASR on the four target models under these configurations. The results indicate that our method maintains consistent attack performance across different setups, underscoring the robustness of our approach to variations in hyperparameter configurations.

Table 8. Comparison of ASRs (%) on CIFAR-10. Methods labeled with a '+' (Ens+ and AdaEA+) use the same number of ensemble models as our method, while those without (Ens and AdaEA) use the original settings (4 ensemble models). Bold indicates the best results.

Attack	Target Model				
	RN-50	VGG-16	DN-121	GoogleNet	Average
Ens	93.1	92.8	91.6	89.1	91.7
Ens+	98.1	98.8	97.8	97.9	98.2
AdaEA	93.8	94.4	92.4	91.1	92.9
AdaEA+	93.3	94.2	91.7	93.3	93.1
Ours	**99.9**	**100**	**100**	**99.8**	**99.9**

Table 9. Average ASRs (%) on the four CIFAR-10 target models under different hyperparameter settings, using ResNet-18 as the surrogate model.

Experiment	Hyperparameter			Average
	Learning Rate Schedule	Optimizer	Learning Rate	
E1	Step size decay ×0.1 each 30 epochs	SGD	0.01	99.7
E2	Cosine Annealing	SGD	0.01	99.8
E3	Step size decay ×0.1 each 30 epochs	Adam	0.001	98.3

6. Conclusion

We have introduced a novel black-box adversarial attack that leverages training checkpoints of a single surrogate model to replicate the diversity typically achieved using multiple surrogate models with diverse architectures. By employing an accuracy gap-based checkpoint selection strategy, our method focuses on those checkpoints that capture task-intrinsic knowledge rather than overfitting to specific training data or architectural nuances, thereby enhancing the transferability of generated adversarial examples. Extensive experiments on ImageNet and CIFAR-10 demonstrate that our approach consistently outperforms traditional ensemble attacks, achieving higher attack success rates across various conditions—including scenarios with limited surrogate capacity, reduced training data, and differing surrogate dataset distributions. Overall, our findings highlight the practical advantages of leveraging a single model's training trajectory for generating transferable adversarial examples and deepen our understanding of the intrinsic mechanisms underlying adversarial transferability.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China under Grant No. 62272175.

References

[1] Bin Chen, Jiali Yin, Shukai Chen, Bohao Chen, and Ximeng Liu. An adaptive model ensemble adversarial attack for boosting adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4489–4498, 2023. 1, 2, 4, 5

[2] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 2, 7

[3] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4312–4321, 2019. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018. 1

[6] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433): 1287–1289, 2019. 1

[7] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1, 2

[8] Martin Gubri, Maxime Cordy, Mike Papadakis, Yves Le Traon, and Koushik Sen. Lgv: Boosting adversarial example transferability from large geometric vicinity. In *European Conference on Computer Vision*, pages 603–618. Springer, 2022. 3, 4

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[10] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11936–11945, 2021. 4

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 4

[12] Sanjay Kariyappa and Moinuddin K Qureshi. Improving adversarial robustness of ensembles with diversity training. *arXiv preprint arXiv:1901.09981*, 2019. 4

[13] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020. 4

[14] Alex Krizhevsky. Learning multiple layers of features from tiny images. In *University of Toronto*, 2009. 4

[15] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. 2

[16] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. Learning transferable adversarial examples via ghost networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11458–11465, 2020. 3

[17] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 2

[18] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 1, 2, 4

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[20] Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, pages 4970–4979. PMLR, 2019. 4

[21] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv preprint arXiv:1602.02697*, 1 (2):3, 2016. 2

[22] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 4

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 4

[24] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. 2014. 2nd International Conference on Learning Representations (ICLR) 2014. 1

[25] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 4

[26] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4

[27] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 4

[28] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019. 2, 7

[29] Yifeng Xiong, Jiadong Lin, Min Zhang, John E Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14983–14992, 2022. 1, 2, 4

[30] Huanrui Yang, Jingyang Zhang, Hongliang Dong, Nathan Inkawhich, Andrew Gardner, Andrew Touchet, Wesley Wilkes, Heath Berry, and Hai Li. Dverge: diversifying vulnerabilities for enhanced robust generation of ensembles. *Advances in Neural Information Processing Systems*, 33: 5505–5515, 2020. 4

[31] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016. 4