Learning with Bilevel-Minimax Optimization for Efficient and Reliable Transfer Attacks

Yaohua Liu[1], Yifan Guo[2], and Jiaxin Gao[3]*

[1] The University of Hong Kong, Hong Kong SAR, China
liuyaohua.918@gmail.com
[2] International School of Information Science & Engineering, Dalian University of Technology, Dalian, China
friscoguo@gmail.com
[3] The Hong Kong Polytechnic University, Hong Kong SAR, China
jiaxinn.gao@outlook.com

Abstract. Transfer-based adversarial attacks craft adversarial examples using surrogate models to mislead black-box victim models. Beyond perturbation generation, transferability is fundamentally governed by the coupling of initialization, surrogate adaptation, and gradient dynamics. We revisit this challenge from a `Bilevel-Minimax` perspective and instantiate it in `BMAT` (`Bilevel-Minimax Adversarial Transfer`). The bilevel formulation captures the dependency between initialization and perturbation, while the inner minimax problem promotes surrogate robustness for cross-architecture generalization. Algorithmically, we design an integrated bottom-up solver that combines a Soft Weight Modulator and an Implicit Gradient Approximator for ternary coupling interaction. We further provide theoretical insights into the optimization dynamics of the proposed bilevel-minimax framework. Extensive experiments on classification and segmentation benchmarks show that `BMAT` surpasses `10+` strong baselines across `30+` victim models, improving both intra- and cross-architecture transfer, and yielding up to `2×` mIoU reduction. Code is available at `https://github.com/callous-youth/BMAT`.

Keywords: Transfer attacks · Bilevel-Minimax Optimization· Classification · Semantic Segmentation

1 Introduction

Adversarial attacks [2,25,33,49,82] craft imperceptible perturbations to mislead deep neural networks, posing serious threats to real-world vision systems [1,5,23, 63]. Among these, *transfer-based adversarial attacks* are particularly concerning as they generate Adversarial Examples (AEs) against surrogate models that can transfer to unknown victim models without query access [12,71,79]. This black-box characteristic makes transfer attacks both highly practical and challenging to defend against [27,34,36,57].

* Corresponding author.

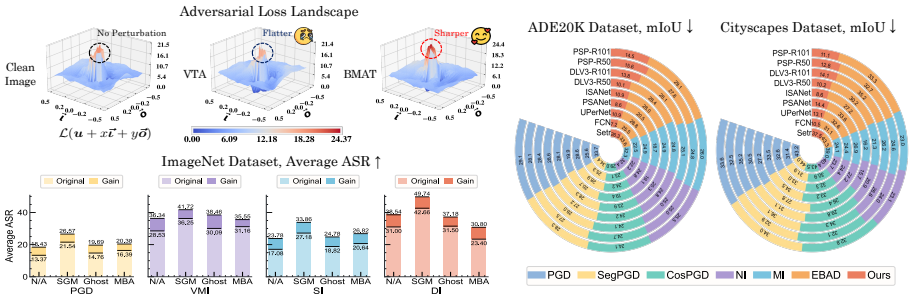

Fig. 1: Landscape (Top Left): We visualize the loss $\mathcal{L}(\boldsymbol{u} + x\boldsymbol{\iota} + y\boldsymbol{o})$ along sign-gradient direction $\boldsymbol{\iota}$ and random direction $\boldsymbol{o}$, centered at the original image $\boldsymbol{u}$. Compared with VTA, `BMAT` reaches a *higher loss region* with stronger cross-model transfer behavior. **Classification (Bottom Left):** `BMAT` achieves a **26.2**% ASR gain on ImageNet [64] across 10 victims with 16 base attackers. **Segmentation (Right):** `BMAT` reduces mIoU by **46.4**% on ADE20K [88] and **43.7**% on Cityscapes [11].

Prior methods have explored transferability from several perspectives, such as surrogate structure modifications [31, 60, 74, 78], input transformation techniques [42,52,79], and momentum-based strategies [12,40]. Some ensemble-based methods [3, 36] enhance transferability by leveraging ensemble gradients from multiple surrogate models. Other studies [53, 55, 70] also explore transferability by employing stronger surrogate models, often at the cost of additional training and task-specific loss designs. Several efforts [15,17] have explored the role of initialization through customized schemes, typically combining data augmentation and surrogate ensembles in a hand-crafted manner. However, most approaches still optimize isolated factors such as perturbation design or surrogate structure, while treating other variables as fixed or heuristic. This fragmented treatment fails to capture the interaction dynamics among initialization, perturbation, and surrogate adaptation, limiting both transferability and optimization efficiency.

We posit that transferability fundamentally arises from the *ternary coupling interaction* among: the **initialization perturbation (IP)**, which seeds the attack trajectory and determines the explored perturbation regions; the **adversarial perturbation**, which exploits model-specific vulnerabilities; and the **surrogate parameter**, which shapes the gradient landscape. When these variables are tuned independently, the attack dynamics often degenerate into a standard white-box procedure on the surrogate, overfitting surrogate-specific artifacts rather than promoting cross-model transfer. Existing methods largely adopt factor-wise and heuristic designs, resulting in fragmented optimization dynamics. This decoupled paradigm leads to two key limitations: (i) *misaligned optimization dynamics*, where separately tuned variables fail to produce coherent cross-model transfer behavior; and (ii) *the absence of a unified optimization formulation* to systematically capture and coordinate these interdependent factors. These observations prompt a fundamental question: *How can we develop*

a unified optimization framework that explicitly models and jointly coordinates these interacting factors to enhance transferability in a principled manner?

1.1 Contributions

To bridge these gaps, we formalize the above hierarchical dependencies in a unified `Bilevel-Minimax` optimization framework, casting the ternary interactions of transfer attacks into a mathematically principled formulation. `BMAT` (`Bilevel-Minimax Adversarial Transfer`) formulates transfer attacks as a bilevel-with-minimax problem: the inner minimax jointly adapts the perturbation and the surrogate's soft weights (one backward pass) to surface robust, transferable gradients, while the outer level learns an IP using a conjugate-gradient hypergradient that avoids unrolling. Algorithmically, we design a tailored bottom-up solver, where Soft Weight Modulator (SWM) performs single-step joint updates of perturbations and soft surrogate weights, and Implicit Gradient Approximator (IGA) refines IP using implicit feedback without incurring the overhead of nested gradient computations. Theoretically, we analyze the joint optimization dynamics of the regularized `BMAT` solver, offering insights into its stability and descent behavior. Fig. 1 shows `BMAT` consistently boosts AE transferability across attackers in classification and segmentation. The main contributions are summarized as follows:

- **Formulation.** `BMAT` casts transfer attacks into a unified `Bilevel-Minimax` framework that *explicitly couples* the initialization, perturbation, and surrogate, thereby enabling *hierarchical dynamic coordination*.
- **Algorithm.** Our bottom-up solver orchestrates a SWM for cross-architecture generalization and an IGA for adaptive initialization guidance, enabling efficient *trajectory seeding* and *warm-started fast transfer*.
- **Analysis.** We provide a stability-oriented analysis of the regularized bilevel-minimax solver, characterizing the joint optimization dynamics of these coupled variables.
- **Experiments.** Evaluations cross `30+` victim models and 2 tasks (classification and segmentation) show that `BMAT` consistently improves both intra- and cross-architecture transferability.

2 Related Work

Input- and Gradient-based Transfer Attacks. Most transfer attacks operate on a fixed surrogate and improve transferability by reshaping gradients and/or transforming the input during iteration [24, 39, 59]. Momentum-based methods such as MI [12], NI [40], and VMI [72] stabilize trajectories by accumulating or reweighting historical gradients. In parallel, input-transformation techniques, including DI [79], TI [13], SI [40], and Admix [73], inject diversity via random resizing, padding, translation, or patch mixing to reduce overfitting. Objective-level designs further encode architectural or task priors, e.g.,

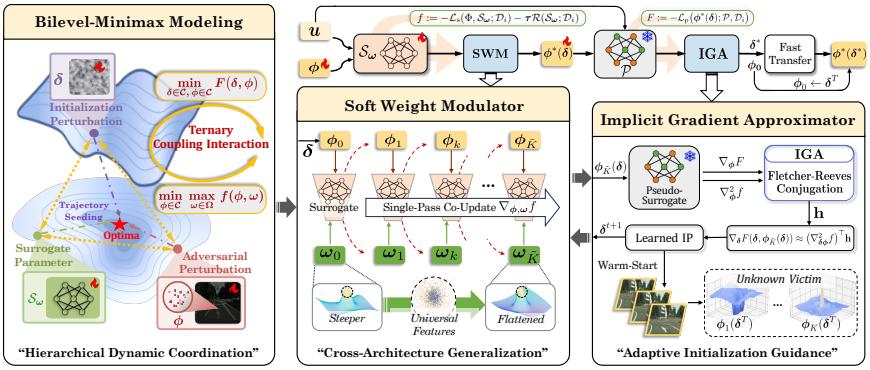

Fig. 2: BMAT pipeline. *Left*: Bilevel-Minimax formulation illustrating the ternary coupling interaction across three variables. *Top*: Overall data flow and supervision in `BMAT`. *Middle*: SWM showing inner-loop single-pass co-update of perturbation and soft surrogate weights to obtain cross-architecture gradients. *Right*: IGA depicting hypergradient calculation of IP using Fletcher-Reeves Conjugation.

SGM [76] and Ghost [38] for classification [7, 26, 60, 78], and SegPGD [25] or CosPGD [1] for dense prediction. Recent work such as RAP [62] encourages flatter loss landscapes via repeated explicit maximization, but remains within a single-level minimax framework. *Summary.* Despite their strong empirical performance, these attacks still treat the surrogate and initialization as fixed design choices and optimize only the perturbation in a single-level manner, lacking a principled formulation of why and how perturbations transfer across models.

Surrogate Ensembles and Adaptation. A complementary thread manipulates the surrogate side via ensembles or adaptive modeling. Ensemble-based attacks aggregate gradients from multiple architectures to reduce model bias [6,36], and dynamic reweighting further stabilizes multi-model signals [3]. Beyond ensembling, GFCS [53] exploits co-image sampling, undertrained surrogates flatten loss landscapes [55], DRA [89] regularizes perturbations distributionally, and FAUG [70] augments surrogate features. Meta-learning strategies [17, 83] learn cross-model update rules but usually rely on surrogate ensembles or tailored training pipelines. BETAK [50] introduces ensemble-guided initialization learning, yet still lacks a unified optimization formulation. *Summary.* While these methods emphasize the surrogate as a key component, surrogate adaptation is usually decoupled from perturbation generation and initialization, or achieved at high cost via large ensembles.

Bilevel Optimization in Adversarial Robustness. Bilevel optimization [44, 46, 47, 56] provides a principled framework for hierarchical learning [18, 22, 32]. A prominent line reformulates adversarial training as a bilevel problem, where the inner problem minimizes perturbed-data loss and the outer problem updates model parameters for robustness, leading to more stable and efficient training schemes [16, 45, 85]. Beyond adversarial training, bilevel formulations have also

been used to tune robustness-related hyperparameters [54] and adaptive attack budgets [37]. *Positioning.* In contrast, our work is, to our knowledge, among the first to systematically apply a `Bilevel-Minimax` formulation to *transfer-based black-box attacks*, explicitly coupling three variables within a single optimization scheme rather than treating them as fixed or independent components.

3 Methodology

3.1 Preliminary

We begin by formalizing the standard transfer attack paradigm. Let $\mathcal{D}_i = (\boldsymbol{u}_i, \boldsymbol{v}_i)$ denote the i-th data pair from dataset $\mathcal{D} = \{(\boldsymbol{u}_i, \boldsymbol{v}_i)\}_{i=1}^{M}$, and $\mathcal{S}_{\boldsymbol{\omega}}$ represent the surrogate model parameterized by $\boldsymbol{\omega} \in \Omega$. The surrogate loss on $\mathcal{D}_i$ is written as $\mathcal{L}_{\mathrm{s}}(\mathcal{S}_{\boldsymbol{\omega}}(\boldsymbol{u}_i), \boldsymbol{v}_i)$. In Vanilla Transfer Attack (VTA) [35], the objective is to maximize the surrogate loss under perturbation constraints:

$$\max_{\boldsymbol{\phi} \in \mathcal{C}} \big\{ \mathcal{L}_{\mathrm{s}}(\boldsymbol{\phi}; \mathcal{S}_{\boldsymbol{\omega}}, \mathcal{D}_i) := \mathcal{L}_{\mathrm{s}}(\mathcal{S}_{\boldsymbol{\omega}}(\boldsymbol{u}_i + \boldsymbol{\phi}), \boldsymbol{v}_i) \big\}, \tag{1}$$

where the perturbation $\boldsymbol{\phi}$ is constrained within an ϵ-bounded ℓ_q norm ball $\mathcal{C} = \{\boldsymbol{\phi} \mid \|\boldsymbol{\phi}\|_q \leq \epsilon\}$, and is typically generated via K-step projected gradient updates:

$$\boldsymbol{\phi}_{k+1} \leftarrow \mathit{\Pi}_{\mathcal{C}} \left(\boldsymbol{\phi}_k + \alpha \cdot \mathrm{sgn}(\nabla_{\boldsymbol{\phi}} \mathcal{L}_{\mathrm{s}}(\boldsymbol{\phi}_k; \mathcal{S}_{\boldsymbol{\omega}}, \mathcal{D}_i)) \right), \tag{2}$$

for $k = 0, 1, \cdots, K-1$, where α is the step size, $\mathit{\Pi}_{\mathcal{C}}$ denotes projection onto $\mathcal{C}$, and $\mathrm{sgn}(\cdot)$ is the sign operation. We denote IP by $\boldsymbol{\delta}$ (i.e., $\boldsymbol{\phi}_0 = \boldsymbol{\delta}$), which serves as the seed of the perturbation trajectory. Such *single-level* optimization neglects the interplay among different variables, limiting its generalization capability.

3.2 Bilevel-Minimax Adversarial Transfer

As shown in Fig. 2, `BMAT` departs from the VTA paradigm by explicitly modeling a ternary coupling among these variables under a bilevel-minimax hierarchy. We next formalize this formulation and derive the bottom-up solver.

Bilevel with Minimax Formulation We cast transfer attacks into a bilevel optimization problem whose inner subproblem adopts a minimax structure.

Inner Minimax Modeling. During attack iterations, perturbations that exploit universal gradient features of robust surrogates tend to exhibit enhanced transferability across victim models [61]. Motivated by this, `BMAT` first introduces an inner minimax objective, i.e., f over $\boldsymbol{\phi}$ and $\boldsymbol{\omega}$ to capture their interaction:

$$\min_{\boldsymbol{\phi} \in \mathcal{C}} \max_{\boldsymbol{\omega} \in \Omega} \big\{ f(\boldsymbol{\phi}, \boldsymbol{\omega}) := -\mathcal{L}_{\mathrm{s}}(\boldsymbol{\phi}, \mathcal{S}_{\boldsymbol{\omega}}; \mathcal{D}_i) - \tau \, \mathcal{R}(\mathcal{S}_{\boldsymbol{\omega}}; \mathcal{D}_i) \big\}, \tag{3}$$

where $\mathcal{R}(\mathcal{S}_{\boldsymbol{\omega}}; \mathcal{D}_i) := \mathcal{L}_{\mathrm{s}}(\mathcal{S}_{\boldsymbol{\omega}}(\boldsymbol{u}_i), \boldsymbol{v}_i)$ serves as a natural-accuracy regularizer with respect to $\boldsymbol{\omega}$, and $\tau > 0$ is the coefficient to balance robustness and natural accuracy. Unlike VTA, which optimizes $\boldsymbol{\phi}$ against a fixed pretrained surrogate,

this formulation jointly adapts both surrogate and perturbation to cultivate universal gradient features for cross-architecture transferability.

Bilevel Modeling. IP critically shapes the attack trajectory and the eventual transferability of $\boldsymbol{\phi}$. VTA only evaluates transferability implicitly through the surrogate loss, which can cause AEs to overfit surrogate-specific artifacts. Instead of choosing $\boldsymbol{\delta}$ heuristically, `BMAT` optimizes IP using feedback from a pseudo-surrogate model $\mathcal{P}$. In practice, $\mathcal{P}$ is instantiated by reusing the available white-box surrogate (or its Bayesian version [36]), without introducing any extra victim access, thus keeping the threat model identical to VTA. Let $\mathcal{L}_{\mathrm{p}}$ denote the loss w.r.t. $\mathcal{P}$. We then formulate the following bilevel–minimax problem:

$$\begin{aligned} &\min_{\boldsymbol{\delta}\in\mathcal{C}} \left\{F(\boldsymbol{\delta}, \boldsymbol{\phi}^*(\boldsymbol{\delta})) := -\,\mathcal{L}_{\mathrm{p}}\big(\boldsymbol{\phi}^*(\boldsymbol{\delta}); \mathcal{P}, \mathcal{D}_i\big)\right\}, \\ &\texttt{where } \boldsymbol{\phi}^*(\boldsymbol{\delta}) := \arg\min_{\boldsymbol{\phi}\in\mathcal{C}} \max_{\boldsymbol{\omega}\in\Omega} f(\boldsymbol{\phi}, \boldsymbol{\omega}), \text{s.t.,}\ \boldsymbol{\phi}_0 = \boldsymbol{\delta}. \end{aligned} \tag{4}$$

Here $\boldsymbol{\phi}^*(\boldsymbol{\delta})$ denotes the finite-step inner response induced by Eq. (3) when the attack trajectory is *seeded* at $\boldsymbol{\phi}_0 = \boldsymbol{\delta}$, rather than an exact global minimizer. Thus, the outer objective evaluates this local trajectory response on $\mathcal{P}$ and updates $\boldsymbol{\delta}$ accordingly. This view provides a principled mechanism for learning IP that consistently leads to stronger transfer attacks, with (i) the inner level performing robustness-aware adaptation of $(\boldsymbol{\phi}, \boldsymbol{\omega})$ and (ii) the outer level optimizing $\boldsymbol{\delta}$ for trajectory seeding and improved transferability.

Bottom-up Solver for `BMAT` We design a bottom-up solver comprising SWM and IGA, aligned with the hierarchical problem structure.

Soft Weight Modulator. To efficiently solve Eq. (3), we introduce SWM to jointly update $\boldsymbol{\phi}$ and $\boldsymbol{\omega}$ with a single backward pass. Each outer iteration begins with $\boldsymbol{\phi}_0 = \boldsymbol{\delta}^t$, seeding the perturbation trajectory with the current initialization estimate. SWM then performs $\tilde{K}$ inner steps to jointly optimize $(\boldsymbol{\phi}_k, \boldsymbol{\omega}_k)$ at low cost:

$$\begin{Bmatrix} \boldsymbol{\omega}_{k+1} \\ \boldsymbol{\phi}_{k+1} \end{Bmatrix} \leftarrow \begin{Bmatrix} \boldsymbol{\omega}_k + \gamma \nabla_{\boldsymbol{\omega}} f_k \\ \boldsymbol{\phi}_k - \beta \nabla_{\boldsymbol{\phi}} f_k \end{Bmatrix} \tag{5}$$

where $f_k := f(\boldsymbol{\phi}_k, \boldsymbol{\omega}_k)$. Here the pretrained surrogate $\boldsymbol{\omega}_0$ serves as hard weights, and $\boldsymbol{\omega}_k$ are soft, adapted weights used within the inner loop. At the beginning of outer iteration (each batch), we restore $\boldsymbol{\omega}_0$ from the clean surrogate, so that SWM performs only local, attack-specific adaptation without accumulating cross-batch drift away from the original model. This design strengthens

Algorithm 1: BMAT.

Input: Outer step T, inner step $\tilde{K}$, attack step K, surrogate $\mathcal{S}_{\boldsymbol{\omega}}$, pseudo-surrogate $\mathcal{P}$

// I: Learning IP
Initialize: IP $\boldsymbol{\delta}^0$, pretrained weights $\boldsymbol{\omega}_0$
for $t = 0$ **to** $T-1$ **do**
  $\boldsymbol{\phi}_0 \leftarrow \boldsymbol{\delta}^t$ // Trajectory Seeding
  // Soft Weight Modulator
  for $k = 0$ **to** $\tilde{K}-1$ **do**
    Compute $\nabla_{\boldsymbol{\phi}} f_k$, $\nabla_{\boldsymbol{\omega}} f_k$
    Update $\boldsymbol{\phi}_k$, $\boldsymbol{\omega}_k$ via Eq. (5)
  // Implicit Gradient Approximator
  $\mathbf{h} \leftarrow \texttt{IGA}\left(\nabla_{\boldsymbol{\phi}} F(\boldsymbol{\phi}_{\tilde{K}}),\ \nabla^2_{\boldsymbol{\phi}} f(\boldsymbol{\phi}_{\tilde{K}})\right)$
  $\boldsymbol{\delta}^{t+1} \leftarrow$
    $\Pi_{\mathcal{C}}\left(\boldsymbol{\delta}^t - \alpha \cdot \mathrm{sgn}((\nabla^2_{\boldsymbol{\delta}\boldsymbol{\phi}} f(\boldsymbol{\phi}_{\tilde{K}}))^\top \mathbf{h})\right)$
// II: Fast Transfer w/ Learned IP
$\boldsymbol{\phi}_0 \leftarrow \boldsymbol{\delta}^T$ // Warm-Start
for $k = 0$ **to** $K-1$ **do**
  $\boldsymbol{\phi}_{k+1} \leftarrow$
    $\Pi_{\mathcal{C}}(\boldsymbol{\phi}_k + \alpha \cdot \mathrm{sgn}(\nabla_{\boldsymbol{\phi}} \mathcal{L}_{\mathrm{s}}(\boldsymbol{\phi}_k; \mathcal{S}_{\boldsymbol{\omega}}, \mathcal{D}_i)))$
return Final attack $\boldsymbol{\phi}_K$

the perturbation using robustness-aware surrogate responses without incurring extra backward passes compared to VTA.

Implicit Gradient Approximator. Central to the outer-level IP optimization is the hypergradient calculation, i.e., $\nabla_{\boldsymbol{\delta}} F(\boldsymbol{\delta}, \boldsymbol{\phi}^*(\boldsymbol{\delta}))$ [20, 43, 48, 84]. Whereas, unrolling the full inner minimax trajectory and backpropagating through $\{\boldsymbol{\phi}_k, \boldsymbol{\omega}_k\}_{k=1}^{K}$ incurs prohibitive computational and memory costs [19, 21, 44]. Instead, `BMAT` employs an implicit-gradient approximation to compute the hypergradient w.r.t. $\boldsymbol{\delta}$. By the implicit function theorem [44], at the approximate inner optimum $\boldsymbol{\phi}^*(\boldsymbol{\delta})$, we have

Algorithm 2: IGA for Computing $\mathbf{h}$.

Function `IGA`$(\nabla_{\boldsymbol{\phi}} F,\ \nabla^2_{\boldsymbol{\phi}} f)$:
  Input: $\mathbf{h}_0 = 0$, tolerance ζ, max iterations N
  $r_0 \leftarrow \nabla_{\boldsymbol{\phi}} F$, $p_0 \leftarrow r_0$
  for $\nu = 0, 1, \ldots, N$ **do**
    $\eta_\nu \leftarrow \frac{r_\nu^\top r_\nu}{p_\nu^\top (\nabla^2_{\boldsymbol{\phi}} f \cdot p_\nu)}$
    $\mathbf{h}_{\nu+1} \leftarrow \mathbf{h}_\nu + \eta_\nu p_\nu$,
     $r_{\nu+1} \leftarrow r_\nu - \eta_\nu (\nabla^2_{\boldsymbol{\phi}} f \cdot p_\nu)$
    if $\|\nabla^2_{\boldsymbol{\phi}} f \cdot \mathbf{h}_{\nu+1} - \nabla_{\boldsymbol{\phi}} F\|_2 \le \zeta$ **then**
      return $\mathbf{h}_{\nu+1}$
    // Fletcher-Reeves Conjugation
    $\lambda_\nu \leftarrow \frac{r_{\nu+1}^\top r_{\nu+1}}{r_\nu^\top r_\nu}$,
     $p_{\nu+1} \leftarrow r_{\nu+1} + \lambda_\nu p_\nu$
  return $\mathbf{h}_N$

$$\nabla_{\boldsymbol{\delta}} F(\boldsymbol{\delta}, \boldsymbol{\phi}^*(\boldsymbol{\delta})) = \left(\nabla^2_{\boldsymbol{\delta}\boldsymbol{\phi}} f\right)^\top \left(\nabla^2_{\boldsymbol{\phi}\boldsymbol{\phi}} f\right)^{-1} \nabla_{\boldsymbol{\phi}} F, \tag{6}$$

which can be equivalently rewritten as

$$\nabla_{\boldsymbol{\delta}} F(\boldsymbol{\delta}, \boldsymbol{\phi}^*(\boldsymbol{\delta})) \approx \left(\nabla^2_{\boldsymbol{\delta}\boldsymbol{\phi}} f\right)^\top \mathbf{h}, \text{ where } \nabla^2_{\boldsymbol{\phi}\boldsymbol{\phi}} f\, \mathbf{h} = \nabla_{\boldsymbol{\phi}} F. \tag{7}$$

Here $\mathbf{h}$ is the solution of a linear system defined by $\nabla^2_{\boldsymbol{\phi}\boldsymbol{\phi}} f$ and $\nabla_{\boldsymbol{\phi}} F$. We denote the solver as `IGA`$(\nabla_{\boldsymbol{\phi}} F,\ \nabla^2_{\boldsymbol{\phi}} f)$, which returns $\mathbf{h}$ for approximating $\nabla_{\boldsymbol{\delta}} F(\boldsymbol{\phi}_{\bar{K}}(\boldsymbol{\delta}))$, with details provided in Alg. 2. `IGA` introduces a Fletcher-Reeves conjugate gradient [58],which avoids explicit Hessians and high-order backpropagation while respecting our bilevel-minimax coupling. $\mathbf{h}$ is then substituted into Eq. (7) to update $\boldsymbol{\delta}$. Note that Eq. (6) presents the idealized expression. In practice, we adopt a damped variant $(\nabla^2_{\phi\phi} f + \rho I)^{-1}$ to ensure local invertibility and numerical stability.

Fast Transfer with IP. To attack unknown victim models, `BMAT` performs a standard iterative attack initialized at $\boldsymbol{\phi}_0 = \boldsymbol{\delta}^T$. The learned IP accelerates convergence and consistently yields stronger black-box transfer than VTA. The complete procedure is summarized in Alg. 1.

Algorithm Analysis The optimization dynamics of such bilevel-minimax formulations remain less understood in transfer-based attacks. Inspired by recent progress on bilevel optimization [29,44,81], we provide a stability-oriented analysis of the regularized `BMAT` under the standard formulation: $\min_{\boldsymbol{\delta} \in \mathcal{C}} F\big(\boldsymbol{\delta}, \boldsymbol{\phi}^*(\boldsymbol{\delta}), \boldsymbol{\omega}^*(\boldsymbol{\delta})\big)$, where the inner response is defined by $(\boldsymbol{\phi}^*(\boldsymbol{\delta}), \boldsymbol{\omega}^*(\boldsymbol{\delta})) = \arg\min_{\boldsymbol{\phi} \in \mathcal{C}} \max_{\boldsymbol{\omega} \in \Omega} f(\boldsymbol{\phi}, \boldsymbol{\omega})$.

Lemma 1. *The following descent inequality holds:*

$$F(\boldsymbol{\delta}_{t+1}, \boldsymbol{\phi}_{t+1,\tilde{K}}, \boldsymbol{\omega}_{t+1,\tilde{K}}) \leq F(\boldsymbol{\delta}_t, \boldsymbol{\phi}_{t,\tilde{K}}, \boldsymbol{\omega}_{t,\tilde{K}}) - \frac{\alpha}{2}\left\|\nabla_{\boldsymbol{\delta}} F(\boldsymbol{\delta}_t, \boldsymbol{\phi}_{t,\tilde{K}}, \boldsymbol{\omega}_{t,\tilde{K}})\right\|^2 + \frac{\alpha}{2}\,\epsilon_{\mathrm{IGA}}^2 + \frac{L_F \alpha^2}{2}\big(G_F + \epsilon_{\mathrm{IGA}}\big)^2.$$

Lemma 2. *The inner updates in* ***BMAT*** *(i.e., SWM) satisfy the following bounds:*

$$\left\|\boldsymbol{\phi}_{t+1,\tilde{K}} - \boldsymbol{\phi}^*(\boldsymbol{\delta}_{t+1})\right\|^2 \leq \big(1+\beta^2 L_{\boldsymbol{\phi}}^2\big)^{\tilde{K}} \left\|\boldsymbol{\delta}_t - \boldsymbol{\phi}^*(\boldsymbol{\delta}_{t+1})\right\|^2 + \beta^2 \tilde{K}\,\epsilon_{\boldsymbol{\phi}}^2,$$
$$\left\|\boldsymbol{\omega}_{t+1,\tilde{K}} - \boldsymbol{\omega}^*(\boldsymbol{\delta}_{t+1})\right\|^2 \leq \big(1+\gamma^2 L_{\boldsymbol{\omega}}^2\big)^{\tilde{K}} \left\|\boldsymbol{\omega}_0 - \boldsymbol{\omega}^*(\boldsymbol{\delta}_{t+1})\right\|^2 + \gamma^2 \tilde{K}\,\epsilon_{\boldsymbol{\omega}}^2.$$

The following result characterizes the averaged stationarity behavior of the outer updates and provides insight into the stability of coupled optimization dynamics.

Theorem 1. *After running* ***BMAT*** *for T iterations with step sizes $\alpha = \beta = \gamma = \frac{c}{\sqrt{T}}$ (where $c > 0$ is a suitable constant), there exist a constant $C > 0$ and error terms $\epsilon_{\mathrm{IGA}}, \epsilon_{\boldsymbol{\phi}}^{(\tilde{K})}, \epsilon_{\boldsymbol{\omega}}^{(\tilde{K})}$ such that*

$$\frac{1}{T+1}\sum_{t=0}^{T}\left\|\nabla_{\boldsymbol{\delta}} F(\boldsymbol{\delta}_t, \boldsymbol{\phi}_{t,\tilde{K}}, \boldsymbol{\omega}_{t,\tilde{K}})\right\|^2 \leq \frac{C}{\sqrt{T}} + \epsilon_{\mathrm{IGA}}^2 + \big(\epsilon_{\boldsymbol{\phi}}^{(\tilde{K})}\big)^2 + \big(\epsilon_{\boldsymbol{\omega}}^{(\tilde{K})}\big)^2.$$

4 Experiments

Datasets and Models. Experiments are conducted on ImageNet (classification) and Cityscapes and ADE20K (segmentation). For classification, ResNet-50 serves as the surrogate model. The 10 victims include 4 CNNs ([1]-[4]: IncRes-v2 [67], MobileNet [65], PNASNet [41], and SENet [28]), 3 robust ensembles ([5]-[7]: Inc-v3$_{\mathrm{ens3}}$, Inc-v3$_{\mathrm{ens4}}$, IncRes-v2$_{\mathrm{ens}}$) [69], and 3 Transformers ([8]-[10]: ViT [14],Visformer [9], and Swin [51]). Inception-v3 [68] is used as the ensemble model for MBA and pseudo-surrogate for `BMAT`. For segmentation, we use MMSegmentation [10]. Each dataset includes 10 victims: 8 CNN-based models (FCN [66], UPerNet [77], PSANet [86], ISANet [30], DLV3-R50/R101 [8], and PSP-R50/R101 [75]) and 2 Transformer-based models (Segformer [80] and Setr [87]). GCNet [4] is used analogously for EBAD and `BMAT` in segmentation.
Baselines and Evaluation Metrics. We report ASR (↑) for image classification and mIoU (↓) for semantic segmentation tasks. We consider 12 mainstream attackers, including standard PGD [35], model-based SGM [76] and Ghost [38], input-transformation-based SI [40], DI [79] and TI [13], momentum-based MI [12], VMI [72], GMI [71], and RAP [62], ensemble-based MBA [36], initialization learning based BETAK [50], and surrogate-adaptation-based DRA [89] and FAUG [70]. For segmentation, we further include momentum-based NI [40], SegPGD [25], ensemble-based EBAD [3], and CosPGD [1]. These baselines span diverse categories to ensure comprehensive evaluation of `BMAT`.

Table 1: ASR results of combining BMAT with 9 mainstream attack methods. Surrogates [1]-[7] and[8]-[10] denote CNN- and transformer-based victims. Performance gains in $(\mathbf{0}, \mathbf{5}]$ and $> \mathbf{5}$ are marked with light blue and deep blue, respectively.

Image Classification, ResNet-50 backbone, ImageNet dataset, ASR ↑																						
Basic Attacker		CNN				CNN Ensemble			Transformer			Basic Attacker		CNN				CNN Ensemble			Transformer		
		[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]			[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
PGD	N/A	13.64	36.58	13.28	16.84	10.34	9.38	5.58	4.22	11.04	12.8	SI	N/A	17.08	46.26	17.76	22.84	13.28	11.42	6.84	5.56	15.26	16.32
	Ours	20.12	47.52	20.40	24.06	14.04	13.08	7.62	5.76	15.46	16.28		**Ours**	23.78	55.18	23.92	30.28	16.60	14.02	8.38	6.50	19.10	19.20
	SGM	20.06	55.74	21.92	29.94	13.86	12.68	7.62	8.06	21.36	24.18		SGM	27.18	66.34	30.34	38.68	18.16	15.38	9.92	9.92	28.78	30.50
	Ours	27.86	63.36	29.50	37.34	18.32	16.18	10.06	9.62	26.00	27.44		**Ours**	34.02	72.86	36.52	44.20	21.24	18.20	11.48	11.04	32.26	32.72
	Ghost	13.92	41.64	14.12	19.16	11.04	10.54	6.18	4.60	12.56	13.86		Ghost	18.82	50.94	19.44	25.50	13.86	12.48	7.34	5.18	16.62	16.54
	Ours	21.18	51.94	21.48	26.74	14.86	13.36	8.04	5.82	15.94	17.52		**Ours**	24.78	59.38	25.20	31.26	17.12	14.04	8.04	6.06	19.34	18.78
	MBA	15.34	57.22	14.04	20.40	11.10	10.26	6.26	3.78	12.22	13.28		MBA	20.64	64.18	19.18	25.48	15.88	12.90	8.74	4.14	14.62	14.66
	Ours	20.98	64.26	19.90	25.96	15.52	13.30	8.56	4.62	15.20	15.52		**Ours**	26.82	71.08	24.68	31.44	19.26	15.88	10.44	5.10	17.44	17.40
MI	N/A	22.02	49.70	22.92	30.24	15.88	14.46	8.82	7.54	18.40	18.92	TI	N/A	14.36	38.32	15.94	19.82	10.98	10.48	6.4	5.34	11.22	11.86
	Ours	34.78	61.50	34.84	38.66	25.92	22.62	14.72	10.14	23.98	24.34		**Ours**	19.54	47.96	21.90	25.24	14.90	13.00	8.26	5.86	13.54	13.98
	SGM	28.76	66.00	31.64	41.40	19.62	17.44	11.30	11.60	28.38	28.90		SGM	21.96	57.62	26.92	33.92	15.68	14.04	8.94	9.66	22.62	24.24
	Ours	38.74	72.96	40.36	47.30	27.44	23.96	16.02	13.14	32.22	32.64		**Ours**	28.24	64.48	32.56	39.46	19.00	16.76	10.62	10.26	25.32	25.96
	Ghost	23.60	56.20	24.70	34.10	17.24	15.24	9.62	7.44	20.30	20.04		Ghost	15.66	43.02	17.02	22.74	12.56	10.96	6.86	4.96	12.68	12.54
	Ours	38.72	68.58	38.66	43.82	28.40	24.36	16.14	10.12	26.34	26.44		**Ours**	19.78	52.18	22.16	27.70	15.02	12.74	7.78	5.64	14.18	13.96
	MBA	28.34	76.52	28.98	36.48	21.10	17.62	11.70	6.58	20.82	20.62		MBA	18.48	60.86	19.36	24.06	15.62	13.28	8.68	4.82	12.56	12.84
	Ours	39.54	80.66	37.90	43.02	30.74	26.84	17.36	9.18	25.52	24.66		**Ours**	24.92	67.28	26.94	30.46	19.84	16.66	11.24	5.68	15.76	15.00
VMI	N/A	31.04	60.16	33.62	39.94	23.12	20.88	14.02	10.74	25.62	26.14	DI	N/A	31.0	62.8	36.66	38.66	23.0	20.58	13.5	8.24	21.54	20.78
	Ours	44.06	69.46	44.46	47.02	33.28	29.42	20.14	13.08	31.42	31.10		**Ours**	38.54	70.12	41.84	45.06	26.88	22.82	14.8	9.06	24.84	23.3
	SGM	38.36	74.46	42.34	50.34	27.70	24.36	16.16	15.32	36.38	37.12		SGM	42.66	79.8	48.7	52.96	29.9	25.5	17.86	14.22	36.42	36.14
	Ours	48.70	78.80	49.88	54.74	36.16	31.08	21.42	16.60	39.56	40.22		**Ours**	49.74	84.48	55.4	58.6	34.74	28.5	19.98	15.72	38.96	37.96
	Ghost	32.50	65.82	35.10	42.54	24.86	21.42	14.56	10.38	26.74	27.02		Ghost	31.5	66.32	34.76	39.5	23.68	20.06	12.68	7.76	20.9	20.62
	Ours	47.16	73.72	47.34	50.44	35.86	30.44	20.26	12.86	33.48	33.00		**Ours**	37.18	71.74	40.36	44.68	25.36	21.44	13.6	7.9	22.96	22.74
	MBA	33.50	81.56	34.80	40.70	25.96	22.44	14.74	8.22	24.72	24.98		MBA	23.4	66.86	20.7	28.72	19.4	16.62	10.38	4.28	12.42	11.76
	Ours	42.50	82.26	40.84	44.94	33.52	29.12	18.70	10.00	27.10	26.56		**Ours**	30.8	71.88	26.74	34.2	24.18	20.66	13.36	5.1	15.4	13.82

4.1 Experimental Results

Image Classification. Tab. 1 presents the ASR results by enhancing 9 base attackers with BMAT, resulting in 24 variants to evaluate its flexibility. As shown, whether built on CNN or Transformers, or ensemble-based robust models, the AEs generated with BMAT exhibit consistently stronger generalizability across 10 victims, yielding an average ASR gain of 23.28% across 24 combinations.

Comparison with Initialization and Adaptation Techniques. Vanilla momentum-based attackers use zero initialization, while GMI adds a global momentum warm start on a fixed surrogate. As shown in Tab. 2, across all victims, BMAT consistently yields higher ASR than both the vanilla and GMI variants. We also compare with surrogate-adaptation attacks, i.e., DRA and FAUG. BMAT also achieves uniformly better transfer results, demonstrating the effectiveness of the bilevel-minimax coordination.

Comparison with Bilevel and Minimax-based Attacks. We further compare BMAT with RAP and BETAK, which are more closely related in formulation. RAP enhances transferability via repeated explicit maximization within a single-level minimax framework. In contrast, BMAT models the interaction among initialization, perturbation, and surrogate adaptation in a unified bilevel-

Table 2: Analysis of different initialization types and surrogate adaptation techniques.

ImageNet dataset, ResNet-50 backbone, ASR ↑											
Basic Attacker		CNN				CNN Ensemble			Transformer		
		[1] IncRes-V2	[2] MobileNet	[3] PNASNet	[4] SENet	[5] Inc-v3$_{ens3}$	[6] Inc-v3$_{ens4}$	[7] IncRes-v2$_{ens}$	[8] ViT	[9] Visformer	[10] Swin
MI	N/A	22.02	49.70	22.92	30.24	15.88	14.46	8.82	7.54	18.40	18.92
	GMI	25.70	57.84	26.02	35.92	17.16	15.56	9.76	7.84	21.02	21.86
	Ours	35.78	64.38	35.40	42.62	24.84	21.58	14.26	10.38	25.32	25.92
VMI	N/A	31.04	60.16	33.62	39.94	23.12	20.88	14.02	10.74	25.62	26.14
	GMI	32.98	64.92	35.10	43.22	23.74	20.92	13.66	10.46	26.80	26.64
	Ours	45.36	72.78	45.84	50.74	33.62	28.86	19.86	13.24	32.98	32.30
PGD	DRA	20.34	56.68	17.66	15.54	25.04	26.16	17.60	5.32	9.86	10.48
	Ours	21.84	58.38	19.06	16.62	26.14	27.5	18.80	5.62	10.30	10.70
	FAUG	21.70	32.34	22.96	21.02	15.42	14.20	10.60	8.24	18.38	19.62
	Ours	25.48	37.34	27.52	25.06	19.60	17.74	12.42	9.06	21.12	22.02

Table 3: Comparative results under normalized budgets, i.e., Backward Passes (BP). The best results are denoted with **boldface**.

Method	CNN	CNN Ensemble	Transformer	Avg. ASR	Memory (GB)	Runtime (Sec)
PGD (BP=10)	10.93	5.73	7.16	8.24	**3.21**	**2.68**
PGD (BP=40)	11.05	5.25	6.70	8.00	3.22	3.15
RAP (BP=40)	7.31	4.67	3.71	5.44	3.75	3.07
RAP (BP=400)	13.74	6.46	7.47	9.68	5.46	6.91
BETAK (BP=40)	17.16	8.07	9.25	12.06	22.69	6.37
Ours (BP=40)	**22.52**	**8.75**	**11.34**	**15.03**	7.89	4.64

minimax formulation. As shown in Tab. 3, under normalized computational budgets (BP=40), `BMAT` consistently achieves higher ASR than RAP, and remains superior even when RAP increases its budget by 10×. Compared with BETAK, which relies on ensemble-guided initialization updates, `BMAT` achieves stronger transferability with substantially lower memory overhead, indicating the benefit of unified coordination without ensemble dependence.

Semantic Segmentation. In Tab. 4, we present mIoU comparisons based on 3 surrogate structures. Momentum-based methods notably improve AE transferability. In comparison, although EBAD incorporates gradients from additional surrogate models, it struggles to produce perturbations with stronger generalizability. In contrast, `BMAT` consistently enhances transferability across diverse models. Notably, with Segformer as the surrogate, `BMAT` surpasses other attackers with nearly 2× higher transferability.

Discussion on Transferability. The segmentation results further show that stronger white-box optimization does not necessarily yield stronger black-box transfer. Momentum-based attacks already perform well when the surrogate and victim share similar CNN structures, but their gains are less stable on transformer victims. `BMAT` improves this harder cross-architecture regime by adapting

Table 4: Comparison with 6 mainstream attacks on segmentation tasks. †, ‡, and ∗ denote the white-box surrogate models used for generating AEs. The best and second-best results are designated with **boldface** and underline, respectively.

Semantic Segmentation, Cityscapes dataset, mIoU ↓											
Surrogate Model	Basic Attacker	CNN								Transformer	
		FCN†	UPerNet	PSANet	ISANet	DLV3-R50‡	DLV3-R101	PSP-R50	PSP-R101	Segformer*	Setr
Clean Data		72.25	77.10	77.63	78.49	79.09	80.20	77.85	78.34	76.54	78.10
FCN†	PGD	1.97†	3.74	3.43	3.31	3.52	8.09	3.61	6.83	33.96	42.09
	SegPGD	2.02†	3.60	3.33	3.05	3.09	10.04	3.20	7.98	36.65	44.73
	CosPGD	2.63†	5.74	5.10	4.50	4.96	13.42	5.37	10.40	35.30	42.79
	NI	1.89†	3.92	3.76	2.77	3.58	7.23	3.28	5.79	29.40	40.43
	MI	2.40†	3.57	3.38	3.17	3.55	6.52	3.51	5.39	27.81	38.75
	EBAD	2.00†	3.83	3.55	3.34	3.54	8.40	3.81	7.05	33.91	42.10
	Ours	**1.75**†	**2.74**	**2.60**	**2.42**	**2.81**	**5.30**	**2.62**	**4.44**	**26.58**	**38.35**
DLV3-R50‡	PGD	7.55	4.85	2.74	4.78	0.81‡	15.02	3.04	11.43	41.04	48.37
	SegPGD	10.43	6.20	3.74	4.43	1.08‡	18.02	4.40	14.45	43.57	49.74
	CosPGD	10.57	6.34	3.94	5.23	**0.64**‡	20.31	4.99	14.75	41.78	48.73
	NI	5.88	4.80	5.41	4.34	2.11‡	9.88	4.14	6.77	33.82	45.19
	MI	5.57	4.29	2.95	4.05	0.99‡	9.55	3.77	6.52	32.99	43.38
	EBAD	7.42	4.90	2.62	4.83	0.77‡	15.45	3.20	11.50	41.16	48.45
	Ours	**2.60**	**1.96**	**2.16**	**2.31**	1.34‡	**4.54**	**1.86**	**3.57**	**28.53**	**42.61**
Segformer*	PGD	31.43	32.61	33.51	27.22	30.49	35.22	32.47	33.56	1.85*	43.16
	SegPGD	31.85	32.98	34.47	27.79	31.11	36.09	32.86	34.05	3.20*	44.05
	CosPGD	30.94	32.26	33.25	26.39	29.79	34.30	32.11	32.84	**1.70***	43.47
	NI	23.68	24.44	27.23	15.67	20.90	26.59	24.01	23.12	2.84*	40.43
	MI	24.09	24.37	26.93	16.23	21.28	26.21	24.60	23.03	2.09*	38.95
	EBAD	31.12	32.59	33.63	27.16	30.21	35.19	32.70	33.27	1.89*	43.27
	Ours	**10.48**	**13.08**	**14.36**	**8.63**	**10.25**	**14.10**	**12.78**	**11.11**	2.70*	**37.52**

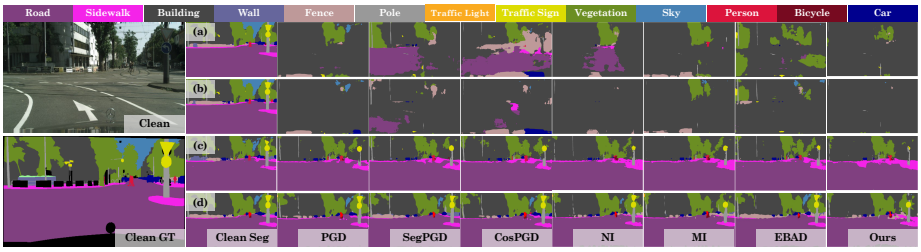

Fig. 3: Visualization of the attack results generated by 4 victim models on the Cityscapes dataset, including (a) DLV3-R101, (b) PSP-R101, (c) Segformer, and (d) Setr. We adopt Segformer as the surrogate model.

the attack trajectory instead of merely strengthening perturbation updates on a fixed surrogate.

Qualitative Comparison. In Fig. 3, we present the attack results of 4 segmentation models using DLV3-R50 as the surrogate model. As shown, all attack methods reduce visual quality, while `BMAT` produces the most degraded segmen-

Table 5: Ablation analysis of IGA and SWM by employing MI as the base attacker.

Semantic Segmentation, Cityscapes dataset, mIoU ↓											
Surrogate Model	Basic Attacker	CNN								Transformer	
		FCN†	UPerNet	PSANet	ISANet	DLV3-R50‡	DLV3-R101	PSP-R50	PSP-R101	Segformer	Setr
FCN†	MI	2.40†	3.57	3.38	3.17	3.55	6.52	3.51	5.39	27.81	38.75
	MI+IGA	**1.56**†	**2.40**	**2.20**	**1.98**	**2.39**	**4.64**	**2.32**	**4.29**	27.16	40.07
	MI+IGA+SWM	1.75†	2.74	2.60	2.42	2.81	5.30	2.62	4.44	**26.58**	**38.35**
DLV3-R50‡	MI	5.57	4.29	2.95	4.05	**0.99**‡	9.55	3.77	6.52	32.99	43.38
	MI+IGA	**1.92**	**1.56**	**1.79**	**1.70**	1.20‡	**3.96**	**1.48**	**3.11**	29.78	44.42
	MI+IGA+SWM	2.60	1.96	2.16	2.31	1.34‡	4.54	1.86	3.57	**28.53**	**42.61**

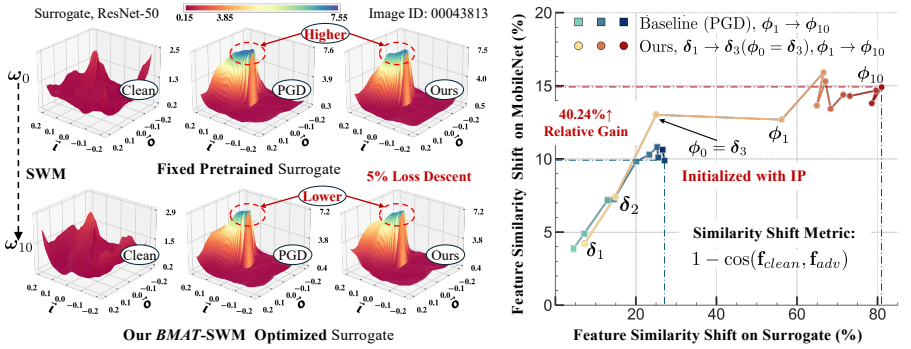

Fig. 4: We visualize the tri-coupled coordination process of BMAT. **Left Panel:** It shows that SWM flattens the loss landscape of surrogate within ∼10 steps. **Right Panel:** It demonstrates that δ guides the trajectory toward a better feature-shift region.

tation outputs, rendering CNN-based victims nearly ineffective. Additionally, for transformer victims, `BMAT` impairs the segmentation of key objects and regions.

4.2 Mechanism Analysis

Tri-Coupled Coordination Dynamics. We visualize the interaction among IP (δ), perturbation (ϕ), and surrogate adaptation (ω) in Fig. 4. *Left.* SWM reshapes the surrogate loss landscape within ∼10 steps, yielding ∼5% loss descent and a smoother optimization region than the fixed pretrained surrogate. *Right.* The learned IP evolves from noise (δ_1) to structured patterns (δ_3), steering the trajectory toward a higher feature shift, quantified as $1-\cos(f_{clean}, f_{adv})$, where $f(\cdot)$ denotes the normalized deep feature before the classification head. The induced perturbation initialized with $\phi_0 = \delta_3$ produces a larger shift than PGD and yields a 40.24% relative gain, indicating that transferability stems from coordinated variable evolution rather than isolated perturbation updates.

IGA and SWM Modules. As shown in Tab. 5, IGA mainly strengthens *within-architecture* transfer, consistently improving ASR on CNN victims.

Table 6: Ablation results of BMAT under the single-surrogate zero-prior setting.

Image Classification, Ablation No additional victim model supervision, ASR ↑										
Attacker	CNN				CNN Ensemble			Transformer		
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
PGD	13.64	36.58	13.28	16.84	10.34	9.38	5.58	4.22	11.04	12.8
Ours (Single-surrogate)	**17.10**	**63.16**	**17.04**	**25.98**	**13.30**	**11.54**	**6.48**	**4.28**	**14.28**	**15.68**
PGD+MBA	15.34	57.22	14.04	20.40	11.10	10.26	6.26	3.78	12.22	13.28
Ours (Single-surrogate)	**26.08**	**76.20**	**25.56**	**36.54**	**19.30**	**15.46**	**9.58**	**5.02**	**19.98**	**19.28**

Table 7: Ablation analysis by employing different pseudo-surrogate structures. †, ‡, and $*$ denote the corresponding pseudo-surrogate.

Semantic Segmentation, Cityscapes dataset, mIoU ↓										
Surrogate Model	Pseudo-surrogate Model	CNN							Transformer	
		FCN†	UPerNet	PSANet	ISANet	DLV3-R101	PSP-R50‡	PSP-R101	Segformer*	Setr
DLV3-R50	FCN†	2.74†	2.11	2.49	2.60	4.42	1.97	3.49	29.92	45.50
	PSP-R50‡	**2.67**	**1.76**	**2.33**	**2.19**	**2.85**	**1.77**‡	**2.68**	31.01	45.37
	Segformer*	3.51	2.85	2.55	2.50‡	4.76	2.21	4.09	**12.32***	**41.86**

Adding SWM brings complementary gains on transformer victims, confirming that surrogate adaptation is especially useful for *cross-architecture* transfer.

Pseudo-surrogate Model. Our main experiments use the auxiliary-surrogate setting, where $\mathcal{P}$ is instantiated by Inc-v3. To test the effectiveness of BMAT without extra victim access or architectural priors, we also instantiate $\mathcal{P}$ with a Bayesian version of the white-box surrogate via weight sampling [36]. In this single-surrogate, zero-prior setting, `BMAT` still improves average ASR by 30.17% and 58.34% over the baselines (Tab. 6), showing that the gain mainly comes from its internal coordination. Using stronger CNN or transformer pseudo-surrogates further shifts transferability toward the corresponding victim families (Tab. 7). This separation clarifies that the auxiliary pseudo-surrogate is not required for `BMAT`; it only provides an optional source of surrogate diversity, while the single-surrogate setting already verifies the bilevel trajectory effect.

Analysis of Attack Iteration Fairness. Tab. 8 reports a detailed ablation on the number of attack iterations. PGD almost saturates around $K = 10$, and increasing K to 20 or 40 brings only marginal or even degraded gains. In contrast, `BMAT` with $(T, \tilde{K}, K) = (2, 5, 10)$ and $(3, 10, 10)$ consistently outperforms PGD even when PGD is allowed to use more iterations; for example, at $K = 20$ and 40, `BMAT` improves ASR by +25.87% and +41.45% over PGD on average. These results suggest that transferability is not improved by stronger surrogate over-optimization alone. Instead, `BMAT` improves trajectory-level generalization by separating the task-specific perturbation ϕ from the task-agnostic seed δ.

Fast and Full `BMAT`. To clarify the default implementation choice behind our main results, we compare the fast `BMAT` variant with a full variant in Tab. 9.

Table 8: Ablation results of the attack iterations on the image classification tasks.

Image Classification, Ablation of Attack Iterations, ASR ↑										
Attacker	CNN				CNN Ensemble			Transformer		
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
PGD ,$K = 10$	13.64	36.58	13.28	16.84	10.34	9.38	5.58	4.22	11.04	12.80
PGD, $K = 20$	13.92	38.72	13.48	19.50	10.10	9.54	5.46	4.48	12.56	13.70
Ours, $(T, \tilde{K}, K) = (2, 5, 10)$	18.62	43.40	18.38	22.04	13.70	13.02	7.78	5.82	13.66	15.10
PGD, $K = 40$	13.22	38.40	13.56	19.10	9.32	8.68	4.90	4.28	12.18	13.68
Ours, $(T, \tilde{K}, K) = (3, 10, 10)$	**20.26**	**47.86**	**20.46**	**24.72**	**14.58**	**13.32**	**8.02**	**6.04**	**15.18**	**16.00**

Table 9: Fast/Full `BMAT` compared with strong combined baselines on ImageNet. Runtime and peak memory are measured with batch size 1. "Single-Surrogate" uses the Bayesian version of the same ResNet-50 surrogate as $\mathcal{P}$; "Auxiliary-Surrogate" uses Inc-v3 as $\mathcal{P}$; "Full" retains SWM in Phase-II.

Image Classification, ResNet-50 surrogate, ASR ↑						
Method	CNN	CNN Ensemble	Transformer	Avg. ASR	Runtime	Peak Memory
DI-MI	55.95	29.57	27.60	39.53	**0.347**s	**399**MB
`BMAT` (Fast, Single-Surrogate)	66.75	37.97	28.13	46.53	1.489s	1201MB
`BMAT` (Full, Single-Surrogate)	**68.63**	42.00	30.67	49.25	1.616s	1290MB
`BMAT` (Auxiliary-Surrogate)	65.73	**44.50**	**32.43**	**49.37**	1.556s	1157MB
DI-MI-MBA	61.48	37.73	26.50	43.86	**0.348**s	**992**MB
`BMAT` (Fast, Single-Surrogate)	65.23	42.13	26.77	46.76	1.509s	1203MB
`BMAT` (Full, Single-Surrogate)	**67.05**	44.50	**28.57**	**48.74**	1.598s	1295MB
`BMAT` (Auxiliary-Surrogate)	65.23	**46.40**	27.83	48.36	1.583s	1153MB

In the fast variant, Phase-I learns the trajectory seed through SWM-based perturbation-surrogate adaptation, and Phase-II uses standard sign/projection updates for efficient deployment. This also clarifies that Phase-I is not a plain warm-up: the SWM inner response directly shapes the IGA-based IP update, so the learned seed already encodes perturbation-surrogate joint adaptation before the fast Phase-II attack. The full variant retains SWM in Phase-II, leading to higher average ASR, especially on ensemble and transformer victims, but with additional runtime and memory under batch size 1. These results show that the fast variant offers a practical accuracy-efficiency tradeoff, while the full variant further validates the benefit of maintaining joint adaptation throughout the attack trajectory.

4.3 Ablation Study

Hyperparameters. As shown in Fig. 5, using larger T leads to better transferability particularly from 1 to 3, while the larger T also means more running cost. Using larger $\tilde{K}$ yields only marginal performance gains. As for $\boldsymbol{\tau}$, relatively better performance is observed when $\tau = 0.1$ or 0.5.

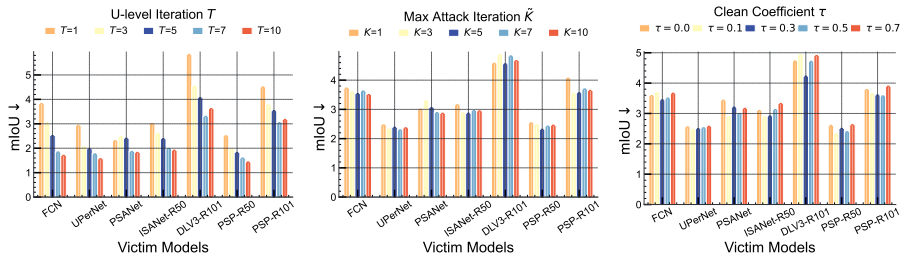

Fig. 5: Ablation results of 3 key hyperparameters across 7 CNN-based victim models on the semantic segmentation task. We adopt DLV3-R50 as the surrogate.

Convergence Behavior. In Fig. 6, we further analyze the optimization dynamics as T and K increase. Even a single IGA step ($T = 1$) already leads to a clear loss decrease. Increasing T from 1 to 3 brings the most noticeable improvements, consistent with the hyperparameter trends as shown in Fig. 5.

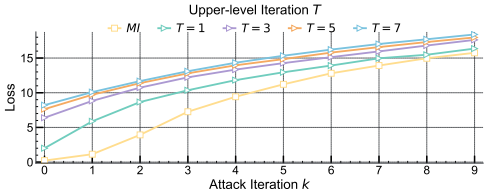

Fig. 6: Illustrating the loss convergence behavior. We employ MI as the base attacker.

Table 10: Runtime analysis as T or $\tilde{K}$ increases. Note that VTA is implemented as PGD when $(T, \tilde{K}) = 0$.

$\tilde{K}$ / T	0 (PGD)	1	2	3	4	5
0 (PGD)	6.82	N/A	N/A	N/A	N/A	N/A
1	N/A	8.78	8.81	8.90	8.86	8.88
2	N/A	9.06	9.13	9.20	9.23	9.30
3	N/A	9.37	9.48	9.73	9.70	10.11
4	N/A	9.67	9.73	10.24	10.04	10.16
5	N/A	10.00	10.21	10.19	10.36	10.57

Computational Efficiency. In Tab. 10, when $(T, \tilde{K})$ increase up to 3, the average ASR gain over 10 victims rises from 21.2% to 21.54% and 35.27%, while the average runtime only grows from 1.96s to 2.31s and 2.91s. Overall, the extra overhead introduced by `BMAT` optimization remains moderate and practically acceptable. We further verify in Tab. 8 that simply increasing PGD iterations ($K = 40$) leads to overfitting and performance degradation, whereas BMAT achieves a principled 41.45% gain under a comparable budget.

5 Conclusion

We propose `BMAT`, a bilevel-minimax transfer attack framework that explicitly encodes the ternary coupling among IP, perturbation, and surrogate adaptation, and solves it via SWM and IGA with stability-oriented optimization dynamics. **Limitation.** `BMAT` incurs extra adaptation and hypergradient computation, suggesting future exploration of more lightweight first-order variants.

References

1. Agnihotri, S., Jung, S., Keuper, M.: Cospgd: an efficient white-box adversarial attack for pixel-wise prediction tasks. In: Forty-first International Conference on Machine Learning (2024)
2. Bai, F., Liu, R., Du, Y., Wen, Y., Yang, Y.: Rat: Adversarial attacks on deep reinforcement agents for targeted behaviors. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 15453–15461 (2025)
3. Cai, Z., Tan, Y., Asif, M.S.: Ensemble-based blackbox attacks on dense prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4045–4055 (2023)
4. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
5. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. Ieee (2017)
6. Chen, H., Zhang, Y., Dong, Y., Yang, X., Su, H., Zhu, J.: Rethinking model ensemble in transfer-based adversarial attacks. arXiv preprint arXiv:2303.09105 (2023)
7. Chen, J., Feng, Z., Zeng, R., Pu, Y., Zhou, C., Jiang, Y., Gan, Y., Li, J., Ji, S.: Enhancing adversarial transferability with adversarial weight tuning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2061–2069 (2025)
8. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
9. Chen, Z., Xie, L., Niu, J., Liu, X., Wei, L., Tian, Q.: Visformer: The vision-friendly transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 589–598 (2021)
10. Contributors, M.: Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark (2020)
11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
12. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193 (2018)
13. Dong, Y., Pang, T., Su, H., Zhu, J.: Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4312–4321 (2019)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Du, J., Zhang, H., Zhou, J.T., Yang, Y., Feng, J.: Query-efficient meta attack to deep neural networks. arXiv preprint arXiv:1906.02398 (2019)
16. Fan, X., Wang, X., Gao, J., Wang, J., Luo, Z., Liu, R.: Bi-level learning of task-specific decoders for joint registration and one-shot medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11726–11735 (2024)
17. Fang, S., Li, J., Lin, X., Ji, R.: Learning to learn transferable attack. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 571–579 (2022)

18. Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., Pontil, M.: Bilevel programming for hyperparameter optimization and meta-learning. arXiv preprint arXiv:1806.04910 (2018)
19. Gao, J., Liu, X., Liu, R., Fan, X.: Learning adaptive hyper-guidance via proxy-based bilevel optimization for image enhancement. The Visual Computer **39**(4), 1471–1484 (2023)
20. Gao, J., Liu, Y.: Enhancing images with coupled low-resolution and ultra-dark degradations: A tri-level learning framework. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 8642–8651 (2024)
21. Gao, J., Liu, Y., Cui, D., Zhao, Z.: Snoc: Subtle nested objective configuration for joint ultra-low-light enhancement and super-resolution. In: Proceedings of the 2026 International Conference on Multimedia Retrieval. pp. 2172–2181 (2026)
22. Gao, J., Liu, Y., Yue, Z., Fan, X., Liu, R.: Collaborative brightening and amplification of low-light imagery via bi-level adversarial learning. Pattern Recognition **154**, 110558 (2024)
23. Gao, J., Yue, Z., Liu, Y., Xie, S., Fan, X., Liu, R.: A dual-stream-modulated learning framework for illuminating and super-resolving ultra-dark images. IEEE transactions on neural networks and learning systems **36**(4), 7500–7513 (2024)
24. Gu, J., Jia, X., de Jorge, P., Yu, W., Liu, X., Ma, A., Xun, Y., Hu, A., Khakzar, A., Li, Z., et al.: A survey on transferability of adversarial examples across deep neural networks. arXiv preprint arXiv:2310.17626 (2023)
25. Gu, J., Zhao, H., Tresp, V., Torr, P.H.: Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In: European Conference on Computer Vision. pp. 308–325. Springer (2022)
26. Guo, Y., Li, Q., Chen, H.: Backpropagating linearly improves transferability of adversarial examples. Advances in neural information processing systems **33**, 85–95 (2020)
27. He, M., Zhang, J., Yang, Z., He, M., Barnes, N., Dai, Y.: Transferable attack for semantic segmentation. arXiv preprint arXiv:2307.16572 (2023)
28. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
29. Hu, Q., Wang, B., Yang, T.: A stochastic momentum method for min-max bilevel optimization. In: Proc. 13th Annu. Workshop Optim. Mach. Learn (2021)
30. Huang, L., Yuan, Y., Guo, J., Zhang, C., Chen, X., Wang, J.: Interlaced sparse self-attention for semantic segmentation. arXiv preprint arXiv:1907.12273 (2019)
31. Huang, Z., Zhang, T.: Black-box adversarial attack with transferable model-based embedding. arXiv preprint arXiv:1911.07140 (2019)
32. Ji, K., Yang, J., Liang, Y.: Bilevel optimization: Convergence analysis and enhanced design. In: International conference on machine learning. pp. 4882–4892. PMLR (2021)
33. Jiao, X., Liu, Y., Gao, J., Chu, X., Fan, X., Liu, R.: Pearl: Preprocessing enhanced adversarial robust learning of image deraining for semantic segmentation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8185–8194 (2023)
34. Klingner, M., Bär, A., Fingscheidt, T.: Improved noise and attack robustness for semantic segmentation by using multi-task training with self-supervised depth estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1299–1309 (2020)
35. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: Artificial intelligence safety and security, pp. 99–112. Chapman and Hall/CRC (2018)

36. Li, Q., Guo, Y., Zuo, W., Chen, H.: Making substitute models more bayesian can enhance transferability of adversarial examples. In: The Eleventh International Conference on Learning Representations (2023)
37. Li, Y., Li, L., Wang, L., Zhang, T., Gong, B.: Automa: A bayesian automation for adversarial attacks. In: International Conference on Machine Learning. pp. 5799–5808. PMLR (2020)
38. Li, Y., Bai, S., Zhou, Y., Xie, C., Zhang, Z., Yuille, A.: Learning transferable adversarial examples via ghost networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11458–11465 (2020)
39. Li, Z., Li, Q., Ren, M., Ru, Y., Sun, Z.: Enhancing adversarial transferability with alignment network. IEEE Transactions on Information Forensics and Security (2025)
40. Lin, J., Song, C., He, K., Wang, L., Hopcroft, J.E.: Nesterov accelerated gradient and scale invariance for adversarial attacks. arXiv preprint arXiv:1908.06281 (2019)
41. Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L.J., Fei-Fei, L., Yuille, A., Huang, J., Murphy, K.: Progressive neural architecture search. In: Proceedings of the European conference on computer vision (ECCV). pp. 19–34 (2018)
42. Liu, J., Lyu, X.: Boosting the transferability of adversarial examples via local mixup and adaptive step size. arXiv preprint arXiv:2401.13205 (2024)
43. Liu, R., Gao, J., Liu, X., Fan, X.: Learning with constraint learning: New perspective, solution strategy and various applications. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(7), 5026–5043 (2024)
44. Liu, R., Gao, J., Zhang, J., Meng, D., Lin, Z.: Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 10045–10067 (2021)
45. Liu, R., Liu, X., Yuan, X., Zeng, S., Zhang, J.: A value-function-based interior-point method for non-convex bi-level optimization. In: International Conference on Machine Learning. pp. 6882–6892. PMLR (2021)
46. Liu, R., Liu, Y., Yao, W., Zeng, S., Zhang, J.: Averaged method of multipliers for bi-level optimization without lower-level strong convexity. In: International Conference on Machine Learning. pp. 21839–21866. PMLR (2023)
47. Liu, R., Liu, Y., Zeng, S., Zhang, J.: Towards gradient-based bilevel optimization with non-convex followers and beyond. Advances in Neural Information Processing Systems **34**, 8662–8675 (2021)
48. Liu, R., Liu, Y., Zeng, S., Zhang, J.: Augmenting iterative trajectory for bilevel optimization: Methodology, analysis and extensions. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
49. Liu, Y., Gao, J.: Past as prior: Reweighted proxy guidance for stable adversarial training. In: ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3356–3360. IEEE (2026)
50. Liu, Y., Gao, J., Liu, X., Jiao, X., Fan, X., Liu, R.: Advancing generalized transfer attack with initialization derived bilevel optimization and dynamic sequence truncation. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. pp. 1137–1145 (2024)
51. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)

52. Long, Y., Zhang, Q., Zeng, B., Gao, L., Liu, X., Zhang, J., Song, J.: Frequency domain model augmentation for adversarial attack. In: European Conference on Computer Vision. pp. 549–566. Springer (2022)
53. Lord, N.A., Mueller, R., Bertinetto, L.: Attacking deep networks with surrogate-based adversarial black-box methods is easy. arXiv preprint arXiv:2203.08725 (2022)
54. MacDonald, J., Wäldchen, S., Hauch, S., Kutyniok, G.: Adversarial training and robustness for multiple perturbations. Advances in Neural Information Processing Systems **32** (2019)
55. Miller, C., Vosoughi, S.: Query-free adversarial transfer via undertrained surrogates. arXiv preprint arXiv:2007.00806 (2020)
56. Mu, P., Liu, Z., Liu, Y., Liu, R., Fan, X.: Triple-level model inferred collaborative network architecture for video deraining. IEEE Transactions on Image Processing **31**, 239–250 (2021)
57. Nakka, K.K., Salzmann, M.: Indirect local attacks for context-aware semantic segmentation networks. In: European Conference on Computer Vision. pp. 611–628. Springer (2020)
58. Nocedal, J., Wright, S.J.: Numerical optimization. Springer (2006)
59. Ododo, F.R., Sadiq, R.R.: Adversarial attacks in cybersecurity: A machine learning perspective. Journal of Science Innovation and Technology Research (2025)
60. Papernot, N., McDaniel, P., Goodfellow, I.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277 (2016)
61. Pedraza, A., Deniz, O., Bueno, G.: On the relationship between generalization and robustness to adversarial examples. Symmetry **13**(5), 817 (2021)
62. Qin, Z., Fan, Y., Liu, Y., Shen, L., Zhang, Y., Wang, J., Wu, B.: Boosting the transferability of adversarial attacks with reverse adversarial perturbation. Advances in neural information processing systems **35**, 29845–29858 (2022)
63. Rebuffi, S.A., Croce, F., Gowal, S.: Revisiting adapters with adversarial training. arXiv preprint arXiv:2210.04886 (2022)
64. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
65. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
66. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(4), 640–651 (2017)
67. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31 (2017)
68. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
69. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204 (2017)
70. Wang, D., Yao, W., Jiang, T., Zheng, X., Wu, J.: Improving the transferability of adversarial examples by feature augmentation. IEEE Transactions on Neural Networks and Learning Systems (2025)

71. Wang, J., Chen, Z., Jiang, K., Yang, D., Hong, L., Guo, P., Guo, H., Zhang, W.: Boosting the transferability of adversarial attacks with global momentum initialization. Expert Systems with Applications **255**, 124757 (2024)
72. Wang, X., He, K.: Enhancing the transferability of adversarial attacks through variance tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1924–1933 (2021)
73. Wang, X., He, X., Wang, J., He, K.: Admix: Enhancing the transferability of adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16158–16167 (2021)
74. Wang, Z., Guo, H., Zhang, Z., Liu, W., Qin, Z., Ren, K.: Feature importance-aware transferable adversarial attacks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7639–7648 (2021)
75. Wightman, R., Touvron, H., Jégou, H.: Resnet strikes back: An improved training procedure in timm. arXiv preprint arXiv:2110.00476 (2021)
76. Wu, W., Su, Y., Chen, X., Zhao, S., King, I., Lyu, M.R., Tai, Y.W.: Boosting the transferability of adversarial samples via attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1161–1170 (2020)
77. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 418–434 (2018)
78. Xiaosen, W., Tong, K., He, K.: Rethinking the backward propagation for adversarial transferability. Advances in Neural Information Processing Systems **36**, 1905–1922 (2023)
79. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2730–2739 (2019)
80. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. arXiv preprint arXiv:2105.15203 (2021)
81. Yao, W., Yin, H., Zeng, S., Zhang, J.: Overcoming lower-level constraints in bilevel optimization: A novel approach with regularized gap functions. In: International Conference on Learning Representations. vol. 2025, pp. 55516–55549 (2025)
82. Yin, J.L., Wang, W., Lin, W., Liu, X., et al.: Adversarial-inspired backdoor defense via bridging backdoor and adversarial attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9508–9516 (2025)
83. Yuan, Z., Zhang, J., Jia, Y., Tan, C., Xue, T., Shan, S.: Meta gradient adversarial attack. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7748–7757 (2021)
84. Yue, Z., Gao, J., Su, Z.: Unveiling details in the dark: Simultaneous brightening and zooming for low-light image enhancement. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6899–6907 (2024)
85. Zhang, Y., Zhang, G., Khanduri, P., Hong, M., Chang, S., Liu, S.: Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In: International Conference on Machine Learning. pp. 26693–26712. PMLR (2022)
86. Zhao, H., Zhang, Y., Liu, S., Shi, J., Change Loy, C., Lin, D., Jia, J.: Psanet: Pointwise spatial attention network for scene parsing. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 267–283 (2018)

87. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. arXiv preprint arXiv:2012.15840 (2020)
88. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)
89. Zhu, Y., Chen, Y., Li, X., Chen, K., He, Y., Tian, X., Zheng, B., Chen, Y., Huang, Q.: Toward understanding and boosting adversarial transferability from a distribution perspective. IEEE Transactions on Image Processing **31**, 6487–6501 (2022)