Contrastive Counterfactual Generation for Imperceptible Adversarial Attack

Anonymous authors
Paper under double-blind review

Abstract

Imperceptible adversarial attacks aim to mislead deep neural networks by adding signal-domain perturbations that induce misclassification while remaining visually indistinguishable from the original signal. Existing methods rely on untargeted loss maximization, producing perturbations poorly aligned with decision boundaries and providing limited control over locality and perceptual cost. To address these limitations, we propose **Contrastive Counterfactual Generation** (CoCoGen), a cross-domain adversarial attack framework that formulates perturbation synthesis as a constrained optimisation problem. CoCoGen explicitly targets the nearest decision boundary by minimising the *contrastive counterfactual margin* under a strict signal-energy budget. Perturbations are localised via gradient-based Top-k spatial projection and confined to the high-frequency subspace using a Fourier-domain projection operator, leveraging reduced human sensitivity to high spatial frequencies. The objective is optimised using masked gradient descent with momentum, while an adaptive sparsity grid search identifies minimal feasible signal support. Experiments across multiple architectures show that CoCoGen achieves 100% Attack Success Rate (vs. 80-100% for prior methods, with most below 99%) while maintaining a MUSIQ score of 61–63 (vs. 36-55), outperforming prior methods in both attack efficacy and visual quality.

1 Introduction

Deep Neural Networks (DNNs) are vulnerable to adversarial attacks Goodfellow et al. (2015), where carefully crafted signal-domain perturbations added to input signals induce incorrect predictions, posing serious security risks in real-world applications Yuan et al. (2019). Beyond exposing vulnerabilities, adversarial attacks are instrumental in evaluating model robustness and motivating defence strategies Lee & Kim (2023); Singh et al. (2024); Luo et al. (2023); Tramèr et al. (2018); Salman et al. (2020); Luo et al. (2021); Cohen et al. (2019). Many existing attacks Madry et al. (2018); Wei et al. (2023) maximise success under loose signal-energy budgets (*e.g.* ℓ_∞ or ℓ_2 norms), often producing perceptible signal artefacts detectable by the human visual system (HVS) Sharif et al. (2018). This has motivated growing interest in *imperceptible* adversarial attacks Carlini & Wagner (2017); Luo et al. (2018); Zhao et al. (2020); Laidlaw et al. (2021); Duan et al. (2021); Chen et al. (2023c); Jia et al. (2022), which seek to maintain attack efficacy while preserving the perceptual fidelity of the adversarial signal.

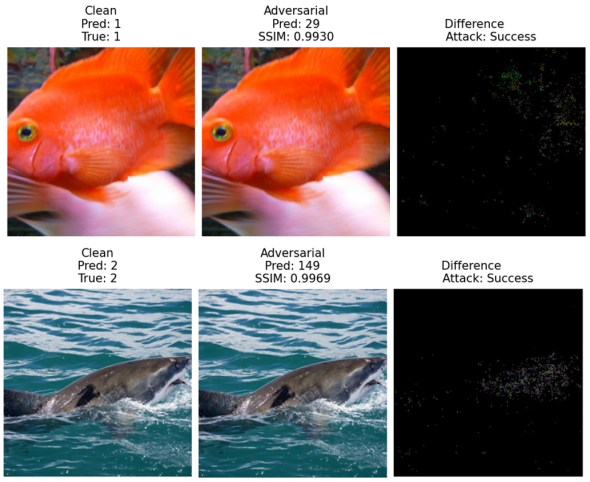

Figure 1: **Imperceptibility of CoCoGen.** Clean signal with labels (col. 1), adversarial signal (col. 2), and contrast-enhanced perturbation map (col. 3). The perturbations are sparse and confined to high-frequency, thus imperceptible.

Existing imperceptible attacks are broadly categorised into perturbation-constrained and unrestricted methods. Perturbation-based approaches exploit perceptual signal properties such as colour sensitivity Zhao et al. (2020), texture complexity Fang et al. (2026), and frequency characteristics Jia et al. (2022); Luo et al. (2022) to conceal adversarial signal components in perceptually insensitive regions. Unrestricted attacks modify semantic attributes via generative or diffusion-based models Song et al. (2018); Rombach et al. (2022); Chen et al. (2023b); Xue et al. (2023); Chen et al. (2023a; 2024), but often introduce unnatural signal distortions, particularly for complex scenes. More fundamentally, most methods rely on *untargeted loss maximisation*, showing no mechanism to identify the nearest decision boundary nor to concentrate signal energy on the components most responsible for the classification decision.

To address these limitations, we propose **Co**ntrastive **Co**unterfactual **Gen**eration (**CoCoGen**), a cross-domain imperceptible adversarial attack framework that formulates perturbation synthesis as a signal-domain constrained optimisation problem (Fig. 1). CoCoGen pushes samples toward the nearest decision boundary by minimising the *contrastive counterfactual margin* under a strict signal-energy budget. Imperceptibility is preserved by restricting perturbations to the high-frequency subspace via Fourier-domain projection and confining them to decision-critical signal components through gradient-based Top-k spatial selection. The optimisation is performed using masked projected gradient descent with momentum. Extensive experiments show that CoCoGen outperforms existing imperceptible attacks across multiple architectures in both effectiveness and visual quality. The contributions of this work are as follows:

- We propose CoCoGen, an attack that minimises the contrastive counterfactual margin, focusing perturbations on the most competitive incorrect class.
- We integrate Top-k spatial projection, Fourier high-frequency constraints, and masked momentum iterative update with adaptive sparsity to produce imperceptible perturbations.
- CoCoGen achieves **100%** ASR across all models, outperforming prior methods (mostly 80-99%) with lower LPIPS (≈ 0.01) and higher MUSIQ (61-63).

2 Proposed Method: CoCoGen

2.1 Problem Formulation

Let $\boldsymbol{x} \in \mathbb{R}^N$ denote a vectorised image, where $N = H \cdot W \cdot C$ is the total number of signal entries obtained by flattening the spatial dimensions H (height) and W (width) over C colour channels (note that this explanation is simplified; for convolutional neural networks, the representation differs). A neural classifier $f : \mathbb{R}^N \to \mathbb{R}^{|\mathcal{Y}|}$ produces a logit vector $f(\boldsymbol{x})$, with $f_c(\boldsymbol{x}) \in \mathbb{R}$ denoting the score for class $c \in \mathcal{Y}$. We write $y_{\text{true}} \in \mathcal{Y}$ for the ground-truth label of $\boldsymbol{x}$. We seek a perturbation $\boldsymbol{\delta} \in \mathbb{R}^N$ such that the adversarial example

$$\boldsymbol{x}_{\text{adv}} = \boldsymbol{x} + \boldsymbol{\delta} \tag{1}$$

is misclassified by f while remaining visually indistinguishable from $\boldsymbol{x}$. equation 1 is the standard additive perturbation model Goodfellow et al. (2015). Imperceptibility is enforced via an ℓ_∞ budget $\|\boldsymbol{\delta}\|_\infty \leq \epsilon$, where $\epsilon > 0$ is a user-specified tolerance (typically $\epsilon \in [4/255, 16/255]$ for 8-bit images; scaled accordingly for other bit depths).

2.2 Contrastive Counterfactual Margin

Naïvely minimising a cross-entropy loss diffuses perturbation energy across all N signal components, producing detectable artefacts. Instead, we target the *contrastive counterfactual margin* $\mathcal{M}(\cdot)$:

$$\mathcal{M}(\boldsymbol{x}) \;=\; f_{y_{\text{true}}}(\boldsymbol{x}) \;-\; \max_{c \neq y_{\text{true}}} f_c(\boldsymbol{x}), \tag{2}$$

where $f_{y_{\text{true}}}(\boldsymbol{x})$ is the logit of the correct class and $\max_{c \neq y_{\text{true}}} f_c(\boldsymbol{x})$ is the highest competing logit. The margin in equation 2 is positive for a correctly classified input and negative once misclassification is achieved. The optimisation problem is therefore

$$\min_{\boldsymbol{\delta}} \; \mathcal{M}(\boldsymbol{x} + \boldsymbol{\delta}), \qquad \text{s.t.} \quad \|\boldsymbol{\delta}\|_\infty \leq \epsilon, \tag{3}$$

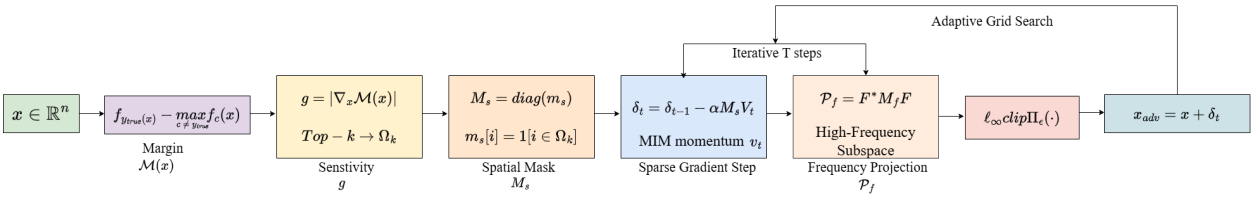

Figure 2: **CoCoGen attack pipeline.** The input signal $\boldsymbol{x}$ is processed through four stages: contrastive margin computation, gradient-based Top-k spatial projection $\boldsymbol{M}_s$, Fourier-domain frequency projection $\mathcal{P}_f$, and ℓ_∞ clipping Π_ϵ, iterated over T steps with an adaptive sparsity grid search over $k^\star$ to yield the adversarial signal $\boldsymbol{x}_{\text{adv}}$.

where the constraint $\|\boldsymbol{\delta}\|_\infty \leq \epsilon$ bounds the per-entry magnitude of the perturbation.

2.3 Spatial Sparsity via a Diagonal Projection Operator

To localise the perturbation to the most decision-relevant signal entries, we compute the element-wise absolute gradient of $\mathcal{M}$ with respect to the input:

$$\boldsymbol{g} \;=\; \big|\nabla_{\boldsymbol{x}}\, \mathcal{M}(\boldsymbol{x})\big| \;\in\; \mathbb{R}^N_{\geq 0}, \tag{4}$$

where $|\cdot|$ is applied element-wise and $\boldsymbol{g}[i]$ quantifies the sensitivity of the margin to the i-th pixel. From $\boldsymbol{g}$ in equation 4 we identify the index set

$$\Omega_k \;=\; \text{Top-}k(\boldsymbol{g}) \;=\; \big\{i_1, i_2, \ldots, i_k\big\}, \tag{5}$$

where $\boldsymbol{g}[i_1] \geq \boldsymbol{g}[i_2] \geq \cdots \geq \boldsymbol{g}[i_k]$, and $k \in \{1, \ldots, N\}$ is the sparsity budget. Using Ω_k from equation 5, we form a binary mask vector $\boldsymbol{m_s} \in \{0,1\}^N$ with

$$\boldsymbol{m}_s[i] \;=\; \mathbb{1}[i \in \Omega_k], \qquad i = 1, \ldots, N, \tag{6}$$

where $\mathbb{1}[\cdot]$ is the indicator function. The mask in equation 6 is then lifted to a diagonal *spatial projection matrix*

$$\boldsymbol{M}_s \;=\; \text{diag}(\boldsymbol{m}_s) \;\in\; \{0,1\}^{N \times N}, \tag{7}$$

which satisfies $\boldsymbol{M}_s^2 = \boldsymbol{M}_s$ (idempotent projector) and $\boldsymbol{M}_s\boldsymbol{u}$ zeroes all entries of $\boldsymbol{u}$ outside Ω_k.

2.4 Frequency Subspace Projection

Spatial sparsity alone does not guarantee imperceptibility, since low-frequency perturbations are highly visible to the human visual system (HVS) Wang et al. (2004). We therefore further restrict $\boldsymbol{\delta}$ to the *high-frequency subspace* of $\mathbb{R}^N$ using the Discrete Fourier Transform (DFT).

The DFT is applied independently per colour channel on the 2D spatial grid $H \times W$. For channel $c \in \{1, \ldots, C\}$, let $\boldsymbol{F}_{HW} \in \mathbb{C}^{HW \times HW}$ be the unitary 2D DFT matrix (equivalently $\boldsymbol{F}_H \otimes \boldsymbol{F}_W$, the Kronecker product of the 1D DFT matrices along each spatial axis), defined entry-wise as

$$[\boldsymbol{F}_{HW}]_{p,q} \;=\; \frac{1}{\sqrt{HW}} \exp\!\Big(-2\pi\,\mathrm{i}\Big(\frac{p_h\, q_h}{H} + \frac{p_w\, q_w}{W}\Big)\Big), \tag{8}$$

where $p = (p_h, p_w)$ and $q = (q_h, q_w)$ index the 2D spatial and frequency grids respectively, and $\mathrm{i} = \sqrt{-1}$. The full DFT matrix acting on the vectorised C-channel signal is

$$\boldsymbol{F} \;=\; \boldsymbol{I}_C \otimes \boldsymbol{F}_{HW} \;\in\; \mathbb{C}^{N \times N}, \qquad N = H \cdot W \cdot C, \tag{9}$$

which applies $\boldsymbol{F}_{HW}$ independently to each channel and satisfies $\boldsymbol{F}\boldsymbol{F}^* = \boldsymbol{I}_N$.

The high-frequency binary mask $\boldsymbol{m}_f \in \{0,1\}^N$ is constructed as follows. Each vectorised index $i \in \{1,\dots,N\}$ maps to a triple (c_i, u_i, v_i) where $c_i \in \{1,\dots,C\}$ is the channel index and $(u_i, v_i) \in \{0,\dots,H-1\} \times \{0,\dots,W-1\}$ is the 2D frequency coordinate. The mask retains bins whose radial frequency exceeds a threshold $\tau_{\text{freq}} \geq 0$, selected by adaptive search:

$$\boldsymbol{m}_f[i] \;=\; \mathbb{1}\left[\sqrt{u_i^2 + v_i^2} > \tau_{\text{freq}}\right], \qquad i = 1,\dots,N, \tag{10}$$

where the threshold is applied identically across all C channels. The corresponding diagonal frequency-selection matrix is

$$\boldsymbol{M}_f \;=\; \operatorname{diag}(\boldsymbol{m}_f) \;\in\; \{0,1\}^{N\times N}. \tag{11}$$

The frequency projection operator is then

$$\mathcal{P}_f \;=\; \boldsymbol{F}^* \boldsymbol{M}_f \boldsymbol{F} \;\in\; \mathbb{C}^{N\times N}, \tag{12}$$

where $\boldsymbol{F}$ maps the signal to the frequency domain per channel, $\boldsymbol{M}_f$ retains only high-frequency components, and $\boldsymbol{F}^*$ maps back to the signal domain. One can verify that $\mathcal{P}_f^2 = \mathcal{P}_f$ (idempotent) and $\mathcal{P}_f^* = \mathcal{P}_f$ (self-adjoint), so $\mathcal{P}_f$ is an orthogonal projector.

2.5 Masked Momentum Iterative Update

We optimise equation 3 using a masked extension of the Momentum Iterative Method (MIM) Dong et al. (2018). Let $\boldsymbol{\delta}_0 = \mathbf{0}$ and $\boldsymbol{v}_0 = \mathbf{0}$ denote the initial perturbation and momentum buffer, respectively, and let $\alpha > 0$ be the step size. At iteration $t \in \{1,\dots,T\}$ the following three operations are applied in sequence.

Momentum accumulation. The gradient of the margin $\mathcal{M}$ (defined in equation 2) with respect to the current adversarial example is ℓ_1-normalised and accumulated into the momentum buffer:

$$\boldsymbol{v}_t \;=\; \boldsymbol{v}_{t-1} \;+\; \frac{\nabla_{\boldsymbol{x}}\, \mathcal{M}(\boldsymbol{x} + \boldsymbol{\delta}_{t-1})}{\left\|\nabla_{\boldsymbol{x}}\, \mathcal{M}(\boldsymbol{x} + \boldsymbol{\delta}_{t-1})\right\|_1}, \tag{13}$$

where $\|\cdot\|_1 = \sum_i |\,\cdot_i\,|$ normalises the update to unit scale across iterations.

Sparse gradient step. The spatial projection matrix $\boldsymbol{M}_s$ from equation 7 restricts the update to the k decision-critical entries:

$$\tilde{\boldsymbol{\delta}}_t \;=\; \boldsymbol{\delta}_{t-1} \;-\; \alpha\, \boldsymbol{M}_s\, \boldsymbol{v}_t, \tag{14}$$

so that only the components in Ω_k (equation 5) receive a non-zero gradient signal.

Composite projection. The intermediate perturbation $\tilde{\boldsymbol{\delta}}_t$ from equation 14 is first projected onto the high-frequency subspace via $\mathcal{P}_f$ (equation 12), and then clipped to the ℓ_∞ ball $\mathcal{B}_\epsilon = \{\boldsymbol{u} : \|\boldsymbol{u}\|_\infty \leq \epsilon\}$:

$$\boldsymbol{\delta}_t \;=\; \Pi_\epsilon\big(\mathcal{P}_f\, \tilde{\boldsymbol{\delta}}_t\big)\,, \tag{15}$$

where the ℓ_∞ projection is $\Pi_\epsilon(\boldsymbol{u}) = \operatorname{clip}(\boldsymbol{u}, -\epsilon, \epsilon)$, applied element-wise. Substituting equation 14 into equation 15 produces the *single closed-form iteration*:

$$\boldsymbol{\delta}_t \;=\; \Pi_\epsilon\Big(\boldsymbol{F}^* \boldsymbol{M}_f \boldsymbol{F}\big(\boldsymbol{\delta}_{t-1} - \alpha\, \boldsymbol{M}_s\, \boldsymbol{v}_t\big)\Big), \tag{16}$$

which involves three closed-form linear operators ($\boldsymbol{M}_s$, $\boldsymbol{F}/\boldsymbol{F}^*$, $\boldsymbol{M}_f$) plus an element-wise clip, and requires no iterative sub-solver.

2.6 Adaptive Sparsity Search

The optimal sparsity budget k is unknown *a priori*: a very small k may fail to achieve misclassification, while a large k wastes perceptual budget. We therefore perform a monotone search over a pre-specified candidate set $\mathcal{K} = \{k_1 < k_2 < \cdots < k_L\} \subset \{1,\dots,N\}$. For each $k \in \mathcal{K}$, the mask $\boldsymbol{m}_s$ in equation 6 is constructed with

$|\Omega_k| = k$ active entries, and T iterations of equation 16 are executed to yield $\boldsymbol{x}_{\text{adv}}^{(k)} = \boldsymbol{x} + \boldsymbol{\delta}_T^{(k)}$ (cf. equation 1). A candidate k is feasible only if it simultaneously satisfies the misclassification condition and two perceptual quality thresholds:

$$\frac{1}{B}\sum_{j=1}^{B} \text{SSIM}\Big(\boldsymbol{x}_{\text{adv}}^{(j,k)},\, \boldsymbol{x}^{(j)}\Big) \;\geq\; \tau_s, \qquad \text{FID}\Big(\Big\{\boldsymbol{x}_{\text{adv}}^{(j,k)}\Big\}_{j=1}^{B},\, \Big\{\boldsymbol{x}^{(j)}\Big\}_{j=1}^{B}\Big) \;\leq\; \tau_f, \tag{17}$$

where B is the number of evaluation images, $\tau_s \in (0, 1]$ is the minimum acceptable mean Structural Similarity Index Wang et al. (2004) across the batch, and $\tau_f \geq 0$ is the maximum tolerated Fréchet Inception Distance Heusel et al. (2017) computed over the full evaluation batch. The constraint in equation 17 gates candidates by both local fidelity (SSIM, averaged per image) and distributional realism (FID, computed over the batch).

The optimal sparsity level is the smallest feasible candidate:

$$k^\star \;=\; \min\Big\{k \in \mathcal{K} \;\Big|\; \underbrace{\frac{1}{B}\sum_{j=1}^{B} \mathbb{1}\Big[\operatorname*{argmax}_c\, f_c\Big(\boldsymbol{x}_{\text{adv}}^{(j,k)}\Big) \neq y_{\text{true}}^{(j)}\Big] = 1}_{\text{all } B \text{ images misclassified}} \;\wedge\; \underbrace{\textit{equation 17 holds}}_{\text{perceptually valid}}\Big\}, \tag{18}$$

where $\wedge$ denotes logical conjunction. The final adversarial examples are $\{\boldsymbol{x}_{\text{adv}}^{(j,k^\star)}\}_{j=1}^{B}$ from equation 1. If no $k \in \mathcal{K}$ satisfies equation 18, the attack is declared unsuccessful for this batch.

2.7 Complexity Analysis

Each iteration of equation 16 requires: (i) one forward–backward pass through f for the gradient in equation 13, costing $\mathcal{O}(\Phi(f))$ where $\Phi(f)$ denotes the number of floating-point operations (FLOPs) in a single forward pass of f; (ii) two DFTs ($\mathcal{O}(N \log N)$) from equation 12; and (iii) two diagonal matrix multiplications ($\mathcal{O}(N)$) from equation 7 and equation 11. The dominant cost is the backward pass, which by standard autodifferentiation requires at most $\mathcal{O}(\Phi(f))$ operations; the same asymptotic order as the forward pass. Over T iterations and $|\mathcal{K}|$ sparsity candidates, the total cost scales as $\mathcal{O}(|\mathcal{K}| \cdot T \cdot \Phi(f))$. Since the $|\mathcal{K}|$ sub-problems are independent, the search over equation 18 is embarrassingly parallel (one can run them parallely on different GPUs and CPUs).

3 Theoretical Analysis

We establish four formal results that justify the design of CoCoGen: (i) the contrastive counterfactual margin is a signed distance surrogate to the nearest decision boundary (Theorem 1); (ii) the composite signal projection $\boldsymbol{M}_s\mathcal{P}_f$ is a well-defined bounded linear operator with controlled spectral norm (Proposition 1); and (iii) CoCoGen strictly generalises the C&W Carlini & Wagner (2017) objective while producing a geometrically distinct solution path (Proposition 2).

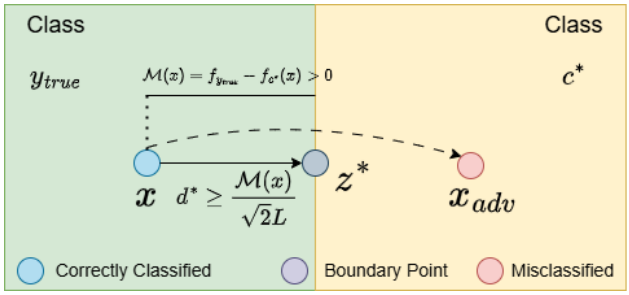

Figure 3: **Decision boundary geometry of CoCoGen.** The contrastive counterfactual margin $\mathcal{M}(\boldsymbol{x}) = f_{y_{\text{true}}}(\boldsymbol{x}) - f_{c^*}(\boldsymbol{x}) > 0$ provides a lower bound on the signed distance $d^*(\boldsymbol{x}) \geq \mathcal{M}(\boldsymbol{x})/(2L)$ to the nearest decision boundary $\mathcal{D}_{y_{\text{true}},c^*}$. Minimising $\mathcal{M}$ under the signal constraints $\boldsymbol{M}_s$ and $\mathcal{P}_f$ guides $\boldsymbol{\delta}$ toward z^*, crossing the boundary with $\mathcal{M}(\boldsymbol{x}+\boldsymbol{\delta}) \leq 0$ using minimal signal energy.

3.1 Contrastive Margin as a Decision-Boundary Surrogate

We first formalise the relationship between the contrastive counterfactual margin $\mathcal{M}(\boldsymbol{x})$ defined in equation 2 and the signed distance of $\boldsymbol{x}$ to the nearest decision boundary of f.

Assumption 1 (Lipschitz classifier) *The classifier $f : \mathbb{R}^N \to \mathbb{R}^{|\mathcal{Y}|}$ is L-Lipschitz continuous with respect to the ℓ_2 norm,*

$$\|f(\boldsymbol{x}) - f(\boldsymbol{x}')\|_2 \;\leq\; L\,\|\boldsymbol{x} - \boldsymbol{x}'\|_2, \qquad \forall\, \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^N, \tag{19}$$

where $L > 0$ is the global Lipschitz constant of f.

Assumption 2 (Unique runner-up class) *At the input $\boldsymbol{x}$ with true label y_{true}, the most competitive incorrect class $c^* = \operatorname{argmax}_{c \neq y_{\text{true}}} f_c(\boldsymbol{x})$ is unique, i.e., $f_{c^*}(\boldsymbol{x}) > f_c(\boldsymbol{x})$ for all $c \notin \{y_{\text{true}}, c^*\}$.*

Under Assumption 2, the margin in equation 2 reduces:

$$\mathcal{M}(\boldsymbol{x}) \;=\; f_{y_{\text{true}}}(\boldsymbol{x}) - f_{c^*}(\boldsymbol{x}). \tag{20}$$

Let $\mathcal{D}_{y,c} = \{\boldsymbol{z} \in \mathbb{R}^N : f_y(\boldsymbol{z}) = f_c(\boldsymbol{z})\}$ denote the decision boundary between classes y and c, and define the signed distance from $\boldsymbol{x}$ to the nearest decision boundary as

$$d^*(\boldsymbol{x}) \;=\; \min_{c \neq y_{\text{true}}} \inf_{\boldsymbol{z} \in \mathcal{D}_{y_{\text{true}},c}} \|\boldsymbol{x} - \boldsymbol{z}\|_2. \tag{21}$$

Theorem 1 (Margin–boundary relationship) *Under Assumptions 1 and 2, for any $\boldsymbol{x} \in \mathbb{R}^N$ correctly classified by f (refer to Fig. 3):*

(i) ***Lower bound:*** $d^*(\boldsymbol{x}) \;\geq\; \dfrac{\mathcal{M}(\boldsymbol{x})}{2L}$.

(ii) ***Monotonicity:*** *If perturbation $\boldsymbol{\delta}$ satisfies $\mathcal{M}(\boldsymbol{x} + \boldsymbol{\delta}) < \mathcal{M}(\boldsymbol{x})$, then $d^*(\boldsymbol{x} + \boldsymbol{\delta}) \leq d^*(\boldsymbol{x}) - \dfrac{\Delta}{2L} + \|\boldsymbol{\delta}\|_2$, where $\Delta := \mathcal{M}(\boldsymbol{x}) - \mathcal{M}(\boldsymbol{x} + \boldsymbol{\delta}) > 0$.*

(iii) ***Crossing condition:*** *$\mathcal{M}(\boldsymbol{x} + \boldsymbol{\delta}) \leq 0$ implies that $\boldsymbol{x} + \boldsymbol{\delta}$ lies on or beyond the decision boundary $\mathcal{D}_{y_{\text{true}},c^*}$.*

(i) Lower bound. Let $\boldsymbol{z}^* \in \mathcal{D}_{y_{\text{true}},c^*}$ be a minimiser of equation 21 for class c^*. By definition of $\mathcal{D}_{y_{\text{true}},c^*}$, $f_{y_{\text{true}}}(\boldsymbol{z}^*) = f_{c^*}(\boldsymbol{z}^*)$, so

$$\mathcal{M}(\boldsymbol{z}^*) \;=\; f_{y_{\text{true}}}(\boldsymbol{z}^*) - f_{c^*}(\boldsymbol{z}^*) \;=\; 0. \tag{22}$$

Define $h(\boldsymbol{x}) := f_{y_{\text{true}}}(\boldsymbol{x}) - f_{c^*}(\boldsymbol{x})$, so $\mathcal{M}(\boldsymbol{x}) = h(\boldsymbol{x})$. By the triangle inequality and the L-Lipschitz condition on f in equation 19,

$$\begin{aligned} |h(\boldsymbol{x}) - h(\boldsymbol{x}')| &\leq |f_{y_{\text{true}}}(\boldsymbol{x}) - f_{y_{\text{true}}}(\boldsymbol{x}')| + |f_{c^*}(\boldsymbol{x}) - f_{c^*}(\boldsymbol{x}')| \\ &\leq L\|\boldsymbol{x} - \boldsymbol{x}'\|_2 + L\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \;=\; 2L\,\|\boldsymbol{x} - \boldsymbol{x}'\|_2, \end{aligned} \tag{23}$$

so h is $2L$-Lipschitz. Applying equation 23 between $\boldsymbol{x}$ and $\boldsymbol{z}^*$, and using $h(\boldsymbol{z}^*) = 0$:

$$\mathcal{M}(\boldsymbol{x}) = |h(\boldsymbol{x}) - h(\boldsymbol{z}^*)| \;\leq\; 2L\,\|\boldsymbol{x} - \boldsymbol{z}^*\|_2 = 2L\,d^*(\boldsymbol{x}), \tag{24}$$

which rearranges to

$$d^*(\boldsymbol{x}) \;\geq\; \frac{\mathcal{M}(\boldsymbol{x})}{2L}. \tag{25}$$

(ii) Monotonicity. Let $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}$ and define $\Delta := \mathcal{M}(\boldsymbol{x}) - \mathcal{M}(\boldsymbol{x}') > 0$. By the $2L$-Lipschitz continuity of h from equation 23,

$$\Delta = |h(\boldsymbol{x}) - h(\boldsymbol{x}')| \;\leq\; 2L\,\|\boldsymbol{\delta}\|_2, \quad \text{so} \quad \|\boldsymbol{\delta}\|_2 \;\geq\; \frac{\Delta}{2L}. \tag{26}$$

Applying the lower bound equation 25 to $\boldsymbol{x}'$:

$$d^*(\boldsymbol{x}') \;\geq\; \frac{\mathcal{M}(\boldsymbol{x}')}{2L} = \frac{\mathcal{M}(\boldsymbol{x}) - \Delta}{2L} = \frac{\mathcal{M}(\boldsymbol{x})}{2L} - \frac{\Delta}{2L}. \tag{27}$$

By the triangle inequality,

$$d^*(\boldsymbol{x}') \;\le\; d^*(\boldsymbol{x}) + \|\boldsymbol{\delta}\|_2. \tag{28}$$

Combining equation 27 and equation 28 with equation 25:

$$\boxed{d^*(\boldsymbol{x}') \;\le\; d^*(\boldsymbol{x}) - \frac{\Delta}{2L} + \|\boldsymbol{\delta}\|_2 \;<\; d^*(\boldsymbol{x}) + \|\boldsymbol{\delta}\|_2,} \tag{29}$$

where the strict inequality holds since $\Delta > 0$. In particular, if the perturbation is margin-efficient, i.e., $\|\boldsymbol{\delta}\|_2 = \Delta/(2L)$, then $d^*(\boldsymbol{x}') \le d^*(\boldsymbol{x})$, with equality only if $\boldsymbol{x}'$ lies exactly on the boundary.

(iii) Crossing condition. If $\mathcal{M}(\boldsymbol{x}+\boldsymbol{\delta}) \le 0$, then $f_{y_{\text{true}}}(\boldsymbol{x}+\boldsymbol{\delta}) \le f_{c^*}(\boldsymbol{x}+\boldsymbol{\delta})$, so $\boldsymbol{x}+\boldsymbol{\delta}$ is misclassified or lies exactly on $\mathcal{D}_{y_{\text{true}},c^*}$. By Assumption 2, c^* is the unique runner-up, so the first boundary crossed is precisely $\mathcal{D}_{y_{\text{true}},c^*}$ — the nearest boundary.

Remark 1 *Theorem 1(i) shows that $\mathcal{M}(\boldsymbol{x})$ provides a computationally tractable lower bound on $d^*(\boldsymbol{x})$, tight when f is linear: for $f(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$, the boundary $\mathcal{D}_{y_{\text{true}},c^*}$ is a hyperplane with $d^*(\boldsymbol{x}) = \mathcal{M}(\boldsymbol{x})/\|\boldsymbol{w}_{y_{\text{true}}} - \boldsymbol{w}_{c^*}\|_2$. Since $\|\boldsymbol{w}_{y_{\text{true}}} - \boldsymbol{w}_{c^*}\|_2 \le \sqrt{2}\|\boldsymbol{W}\|_2 = 2L$ (taking $L = \|\boldsymbol{W}\|_2/\sqrt{2}$), the bound $d^*(\boldsymbol{x}) \ge \mathcal{M}(\boldsymbol{x})/(2L)$ is tight at equality for this choice of L. Minimising $\mathcal{M}$ therefore reduces d^* at a rate bounded by $1/(2L)$, providing a principled surrogate for boundary traversal without computing d^* explicitly, which is NP-hard in general Katz et al. (2017).*

3.2 Composite Signal Projection

Proposition 1 (Bounded composite projector) *Let $\boldsymbol{M}_s$ and $\mathcal{P}_f = \boldsymbol{F}^*\boldsymbol{M}_f\boldsymbol{F}$ be defined as in equation 7 and equation 12. Then:*

- *(i) $\boldsymbol{M}_s$ and $\mathcal{P}_f$ are each orthogonal projectors on $\mathbb{R}^N$, satisfying $\boldsymbol{M}_s^2 = \boldsymbol{M}_s$, $\boldsymbol{M}_s^\top = \boldsymbol{M}_s$, $\mathcal{P}_f^2 = \mathcal{P}_f$, and $\mathcal{P}_f^* = \mathcal{P}_f$.*
- *(ii) The composite operator $\boldsymbol{P} := \mathcal{P}_f\boldsymbol{M}_s$ is a bounded linear operator with spectral norm $\|\boldsymbol{P}\|_2 \le 1$.*
- *(iii) The image of $\boldsymbol{P}$ satisfies $\mathrm{Im}(\boldsymbol{P}) \subseteq \mathrm{Im}(\mathcal{P}_f)$, so all perturbations produced by CoCoGen lie in the high-frequency subspace of $\mathbb{R}^N$.*

(i) Orthogonal projectors. $\boldsymbol{M}_s = \mathrm{diag}(\boldsymbol{m}_s)$ with $\boldsymbol{m}_s \in \{0,1\}^N$, so $\boldsymbol{M}_s^2 = \boldsymbol{M}_s$ and $\boldsymbol{M}_s^\top = \boldsymbol{M}_s$ trivially. For $\mathcal{P}_f$: since $\boldsymbol{F}$ is unitary ($\boldsymbol{F}^*\boldsymbol{F} = \boldsymbol{I}$) and $\boldsymbol{M}_f = \mathrm{diag}(\boldsymbol{m}_f)$ with $\boldsymbol{m}_f \in \{0,1\}^N$,

$$\begin{aligned}
\mathcal{P}_f^2 &= (\boldsymbol{F}^*\boldsymbol{M}_f\boldsymbol{F})(\boldsymbol{F}^*\boldsymbol{M}_f\boldsymbol{F}) \\
&= \boldsymbol{F}^*\boldsymbol{M}_f(\boldsymbol{F}\boldsymbol{F}^*)\boldsymbol{M}_f\boldsymbol{F} \\
&= \boldsymbol{F}^*\boldsymbol{M}_f^2\boldsymbol{F} \\
&= \boldsymbol{F}^*\boldsymbol{M}_f\boldsymbol{F} \\
&= \mathcal{P}_f,
\end{aligned}$$

using $\boldsymbol{F}\boldsymbol{F}^* = \boldsymbol{I}$ and $\boldsymbol{M}_f^2 = \boldsymbol{M}_f$. Self-adjointness follows from $\mathcal{P}_f^* = (\boldsymbol{F}^*\boldsymbol{M}_f\boldsymbol{F})^* = \boldsymbol{F}^*\boldsymbol{M}_f^*\boldsymbol{F} = \boldsymbol{F}^*\boldsymbol{M}_f\boldsymbol{F} = \mathcal{P}_f$, since $\boldsymbol{M}_f$ is real and diagonal.

(ii) Spectral norm of $\boldsymbol{P}$. For any $\boldsymbol{u} \in \mathbb{R}^N$,

$$\begin{aligned}
\|\boldsymbol{P}\boldsymbol{u}\|_2^2 &= \|\mathcal{P}_f\boldsymbol{M}_s\boldsymbol{u}\|_2^2 = \langle \mathcal{P}_f\boldsymbol{M}_s\boldsymbol{u},\, \mathcal{P}_f\boldsymbol{M}_s\boldsymbol{u}\rangle \\
&= \langle \boldsymbol{M}_s\boldsymbol{u},\, \mathcal{P}_f^2\boldsymbol{M}_s\boldsymbol{u}\rangle = \langle \boldsymbol{M}_s\boldsymbol{u},\, \mathcal{P}_f\boldsymbol{M}_s\boldsymbol{u}\rangle \\
&\le \|\boldsymbol{M}_s\boldsymbol{u}\|_2 \|\mathcal{P}_f\boldsymbol{M}_s\boldsymbol{u}\|_2,
\end{aligned} \tag{30}$$

where we used $\mathcal{P}_f^2 = \mathcal{P}_f$ and Cauchy–Schwarz. Since $\boldsymbol{M}_s$ is a projection, $\|\boldsymbol{M}_s\boldsymbol{u}\|_2 \le \|\boldsymbol{u}\|_2$. Similarly, $\mathcal{P}_f$ is a projection so $\|\mathcal{P}_f\boldsymbol{v}\|_2 \le \|\boldsymbol{v}\|_2$ for any $\boldsymbol{v}$. Combining: $\|\boldsymbol{P}\boldsymbol{u}\|_2^2 \le \|\boldsymbol{u}\|_2\,\|\boldsymbol{P}\boldsymbol{u}\|_2$, giving $\|\boldsymbol{P}\boldsymbol{u}\|_2 \le \|\boldsymbol{u}\|_2$, i.e., $\|\boldsymbol{P}\|_2 \le 1$.

(iii) Image containment. For any $\boldsymbol{u}$, $\boldsymbol{P}\boldsymbol{u} = \mathcal{P}_f(\boldsymbol{M}_s\boldsymbol{u}) \in \mathrm{Im}(\mathcal{P}_f)$ by definition of the image.

Remark 2 *Proposition 1(ii) guarantees that no iteration of equation 16 can amplify the perturbation signal energy beyond its current level, since* $\|\boldsymbol{P}\|_2 \leq 1$. *This provides a stability certificate for the update rule that is absent in unconstrained gradient ascent methods such as PGD.*

3.3 Relationship to C&W and PGD Objectives

Proposition 2 (Generalisation of C&W) *The C&W objective Carlini & Wagner (2017) for untargeted attacks is*

$$\ell_{\mathrm{CW}}(\boldsymbol{x}+\boldsymbol{\delta}) \;=\; \max\Big(\max_{c\neq y_{\mathrm{true}}} f_c(\boldsymbol{x}+\boldsymbol{\delta}) - f_{y_{\mathrm{true}}}(\boldsymbol{x}+\boldsymbol{\delta}),\; -\kappa\Big), \tag{31}$$

where $\kappa \geq 0$ *is a confidence margin. Then:*

(i) $\ell_{\mathrm{CW}} = \max(-\mathcal{M}, -\kappa)$. *The two objectives relate across three regimes:*

- ***Pre-crossing*** *(*$\mathcal{M} > 0$*): with* $\kappa = 0$, $\ell_{\mathrm{CW}} = 0$ *and its gradient vanishes, providing no optimisation signal. By contrast,* $\mathcal{M} > 0$ *retains a non-zero gradient, actively driving* $\boldsymbol{\delta}$ *toward the decision boundary.*
- ***At the boundary*** *(*$\mathcal{M} = 0$*): both objectives vanish simultaneously,* $\ell_{\mathrm{CW}} = \mathcal{M} = 0$.
- ***Post-crossing*** *(*$\mathcal{M} \leq 0$*): with* $\kappa = 0$, ℓ_{CW} *saturates at* 0 *and* $\nabla_{\boldsymbol{\delta}}\,\ell_{\mathrm{CW}} = \mathbf{0}$, *so C&W ceases to optimise.* $\mathcal{M}$ *continues to decrease below zero, driving the example deeper into the misclassified region:*

$$\mathcal{M}(\boldsymbol{x}+\boldsymbol{\delta}) \leq 0 \;\;\Longrightarrow\;\; \ell_{\mathrm{CW}}(\boldsymbol{x}+\boldsymbol{\delta}) = 0, \quad \nabla_{\boldsymbol{\delta}}\,\ell_{\mathrm{CW}} = \mathbf{0}, \tag{32}$$

whereas $\nabla_{\boldsymbol{\delta}}\,\mathcal{M}$ *remains well-defined and non-zero in general.*

(ii) Under the signal constraints $\boldsymbol{M}_s$ *and* $\mathcal{P}_f$, CoCoGen *minimises* $\mathcal{M}$ *over the restricted feasible set*

$$\mathcal{F} \;:=\; \big\{\boldsymbol{\delta} \;:\; \|\boldsymbol{\delta}\|_\infty \leq \epsilon,\; \boldsymbol{\delta} \in \mathrm{Im}(\mathcal{P}_f),\; \mathrm{supp}(\boldsymbol{\delta}) \subseteq \Omega_k\big\}, \tag{33}$$

which is a strict subset of the C&W feasible set $\mathcal{B}_\epsilon$ *whenever* $k < N$ *or* $\boldsymbol{M}_f \neq \boldsymbol{I}$.

(iii) The CoCoGen *solution* $\boldsymbol{\delta}^*_{\mathrm{CoCoGen}} = \mathrm{argmin}_{\boldsymbol{\delta}\in\mathcal{F}}\,\mathcal{M}(\boldsymbol{x}+\boldsymbol{\delta})$ *satisfies* $\|\boldsymbol{\delta}^*_{\mathrm{CoCoGen}}\|_0 \leq k$ *and* $\boldsymbol{\delta}^*_{\mathrm{CoCoGen}} \in \mathrm{Im}(\mathcal{P}_f)$, *whereas no such guarantees hold for the C&W solution in general.*

(i) Three-regime analysis. Under Assumption 2, the margin reduces to $\mathcal{M}(\boldsymbol{x}+\boldsymbol{\delta}) = f_{y_{\mathrm{true}}}(\boldsymbol{x}+\boldsymbol{\delta}) - f_{c^*}(\boldsymbol{x}+\boldsymbol{\delta})$, so

$$\begin{aligned}\ell_{\mathrm{CW}}(\boldsymbol{x}+\boldsymbol{\delta}) &= \max\big(f_{c^*}(\boldsymbol{x}+\boldsymbol{\delta}) - f_{y_{\mathrm{true}}}(\boldsymbol{x}+\boldsymbol{\delta}),\; -\kappa\big)\\ &= \max\big(-\mathcal{M}(\boldsymbol{x}+\boldsymbol{\delta}),\; -\kappa\big),\end{aligned} \tag{34}$$

establishing the identity $\ell_{\mathrm{CW}} = \max(-\mathcal{M}, -\kappa)$. We now analyse each regime with $\kappa = 0$.

Pre-crossing ($\mathcal{M} > 0$). Since $-\mathcal{M} < 0$,

$$\ell_{\mathrm{CW}} = \max(-\mathcal{M}, 0) = 0. \tag{35}$$

The objective is identically zero and its subgradient satisfies $\partial_{\boldsymbol{\delta}}\,\ell_{\mathrm{CW}} = \{0\}$, so C&W provides no gradient signal in this regime. $\mathcal{M}$ is positive and its gradient $\nabla_{\boldsymbol{\delta}}\mathcal{M} = \nabla_{\boldsymbol{\delta}} f_{y_{\mathrm{true}}} - \nabla_{\boldsymbol{\delta}} f_{c^*}$ is non-zero in general (it vanishes only at saddle points of the logit difference, which occur on a set of measure zero for smooth f). Hence CoCoGen retains a non-trivial gradient signal throughout the pre-crossing phase, whereas C&W does not.

At the boundary ($\mathcal{M} = 0$). $\ell_{\mathrm{CW}} = \max(0, 0) = 0 = \mathcal{M}$, so both objectives coincide.

Post-crossing ($\mathcal{M} \leq 0$). Since $-\mathcal{M} \geq 0$,

$$\ell_{\mathrm{CW}} = \max(-\mathcal{M}, 0) = 0, \tag{36}$$

and again $\partial_{\boldsymbol{\delta}} \ell_{\mathrm{CW}} = \{0\}$. C&W therefore ceases to update $\boldsymbol{\delta}$ once misclassification is achieved. By contrast, $\mathcal{M}(\boldsymbol{x}+\boldsymbol{\delta}) \leq 0$ implies $f_{c^*}(\boldsymbol{x}+\boldsymbol{\delta}) \geq f_{y_{\text{true}}}(\boldsymbol{x}+\boldsymbol{\delta})$, and continuing to minimise $\mathcal{M}$ further increases $f_{c^*} - f_{y_{\text{true}}}$, i.e., the confidence of misclassification. Formally, for any $\boldsymbol{\delta}$ with $\mathcal{M}(\boldsymbol{x}+\boldsymbol{\delta}) \leq 0$ and any $\boldsymbol{\delta}'$ with $\mathcal{M}(\boldsymbol{x}+\boldsymbol{\delta}') < \mathcal{M}(\boldsymbol{x}+\boldsymbol{\delta})$,

$$f_{c^*}(\boldsymbol{x}+\boldsymbol{\delta}') - f_{y_{\text{true}}}(\boldsymbol{x}+\boldsymbol{\delta}') \;>\; f_{c^*}(\boldsymbol{x}+\boldsymbol{\delta}) - f_{y_{\text{true}}}(\boldsymbol{x}+\boldsymbol{\delta}), \tag{37}$$

so CoCoGen strictly increases misclassification confidence at each iteration after boundary crossing, while C&W stagnates. This establishes equation 32 and the post-crossing divergence of the two objectives.

(ii) Strict subset. $\mathcal{F}$ in equation 33 imposes two constraints beyond $\mathcal{B}_\epsilon$:

(a) $\boldsymbol{\delta} \in \mathrm{Im}(\mathcal{P}_f)$: by Proposition 1(iii), every iterate of CoCoGen lies in the high-frequency subspace $\mathrm{Im}(\mathcal{P}_f) \subsetneq \mathbb{R}^N$ whenever $\boldsymbol{M}_f \neq \boldsymbol{I}$, i.e., at least one frequency bin is suppressed.

(b) $\mathrm{supp}(\boldsymbol{\delta}) \subseteq \Omega_k$: the spatial mask $\boldsymbol{M}_s$ from equation 6 zeroes all entries outside the top-k index set Ω_k, so $\|\boldsymbol{\delta}\|_0 \leq k < N$ whenever $k < N$.

Since neither constraint is imposed by C&W, $\mathcal{F} \subsetneq \mathcal{B}_\epsilon$ strictly, and the C&W solution $\boldsymbol{\delta}^*_{\mathrm{CW}} = \mathrm{argmin}_{\boldsymbol{\delta} \in \mathcal{B}_\epsilon} \ell_{\mathrm{CW}}(\boldsymbol{x}+\boldsymbol{\delta})$ need not lie in $\mathcal{F}$. Hence the two optimisation problems are defined over geometrically distinct feasible sets, and their solutions are distinct in general.

(iii) Signal support guarantees. $\boldsymbol{\delta}^*_{\mathrm{CoCoGen}} \in \mathcal{F}$ by construction, so:

- $\|\boldsymbol{\delta}^*_{\mathrm{CoCoGen}}\|_0 \leq k$ follows directly from constraint (b) in part (ii).
- $\boldsymbol{\delta}^*_{\mathrm{CoCoGen}} \in \mathrm{Im}(\mathcal{P}_f)$ follows from constraint (a) and Proposition 1(iii).

For the C&W solution, the feasible set $\mathcal{B}_\epsilon$ imposes no sparsity or spectral constraints, so $\boldsymbol{\delta}^*_{\mathrm{CW}}$ is dense in general and not confined to any frequency subspace, and neither guarantee holds.

Remark 3 *Proposition 2 reveals a fundamental asymmetry between CoCoGen and C&W that goes beyond the choice of feasible set. In the pre-crossing regime, C&W's gradient vanishes identically (equation 35), making it reliant on the confidence parameter $\kappa > 0$ to maintain a non-zero signal — yet large κ risks over-shooting the boundary and inflating perceptual cost. CoCoGen sidesteps this trade-off entirely by minimising $\mathcal{M}$ directly, which provides a non-zero gradient at every point where f is smooth, and continues to increase misclassification confidence after crossing (equation 37). The combination of a strictly smaller feasible set $\mathcal{F} \subsetneq \mathcal{B}_\epsilon$ and a gradient signal that is active across all three regimes explains the empirical observation that CoCoGen achieves 100% ASR with substantially lower perceptual cost (LPIPS ≈ 0.01 vs. 0.02-0.74) than C&W and other baselines.*

4 Experimental Setup

Dataset. Following prior work Zhao et al. (2020); Yuan et al. (2022); Wei et al. (2023); Chen et al. (2024), we evaluate on a 1,000-image subset of ImageNet Russakovsky et al. (2015), originally 299×299 pixels and resized to 224×224.

Models. We evaluate on four architectures spanning CNNs and vision transformers: ResNet-50 He et al. (2016), EfficientNet-B0 Tan & Le (2019), ConvNeXt-Base Liu et al. (2022), and ViT-Base Dosovitskiy (2020).

Metrics. Attack effectiveness is measured by Attack Success Rate (ASR). Perceptual fidelity is assessed via five complementary metrics: PSNR and SSIM Wang et al. (2004) measure pixel-level and structural similarity; LPIPS Zhang et al. (2018) captures deep perceptual distance; FID Heusel et al. (2017) quantifies distributional divergence between clean and adversarial images; and MUSIQ Ke et al. (2021) provides a no-reference image quality score, allowing quality assessment without access to a clean reference.

Table 1: **Results.** Attack Success Rate (ASR) and imperceptibility metrics for all evaluated attacks and target architectures. Runtime is measured on a single NVIDIA T4 GPU. ↑ higher is better; ↓ lower is better.

Model	Attack Method	Time (s)↓	ASR (%)↑	PSNR↑	SSIM↑	FID↓	LPIPS↓	MUSIQ↑
ResNet-50 He et al. (2016)	PGD Madry et al. (2018)	462.65	**100.00**	33.51	0.91	52.57	0.10	40.38
	DeepFool Moosavi-Dezfooli et al. (2016)	480.20	99.00	38.00	0.94	15.80	0.04	45.10
	NCF Yuan et al. (2022)	523.70	80.07	28.54	0.71	98.72	0.26	39.80
	ACA Chen et al. (2024)	601.30	97.34	8.19	0.62	93.05	0.14	36.43
	DiffPGD Xue et al. (2023)	1105.00	27.74	10.11	0.94	14.59	0.03	42.44
	AdvDrop Duan et al. (2021)	520.30	82.72	41.91	0.98	**4.43**	0.02	44.77
	PerC-AL Zhao et al. (2020)	1640.00	99.80	42.53	0.98	10.44	0.02	43.18
	SSAH Luo et al. (2022)	500.10	97.00	44.04	**0.99**	11.06	**0.01**	46.98
	CoCoGen (Ours)	1225.41	**100.00**	**44.67**	**0.99**	10.18	**0.01**	**61.44**
EfficientNet -B0 Tan & Le (2019)	PGD	293.71	98.50	32.99	0.89	69.00	0.12	40.76
	DeepFool	305.60	99.00	37.90	0.95	18.20	0.05	44.20
	NCF	298.20	44.91	28.33	0.70	98.68	0.27	38.74
	ACA	322.50	93.17	8.15	0.57	92.35	0.13	35.80
	DiffPGD	6499.00	69.10	31.10	0.90	16.19	0.03	55.25
	AdvDrop	295.50	83.00	41.43	0.97	5.09	0.02	44.76
	PerC-AL	1640.00	99.80	**42.53**	0.98	10.44	0.02	43.18
	SSAH	282.40	99.00	42.40	0.98	31.28	0.02	45.65
	CoCoGen (Ours)	1523.63	**100.00**	41.67	**0.99**	**4.05**	**0.01**	**62.63**
ConvNeXt Base Liu et al. (2022)	PGD	2416.89	99.70	34.04	0.92	44.34	0.08	40.84
	DeepFool	2350.50	99.00	38.40	0.96	16.75	0.04	45.00
	NCF	2253.80	80.07	28.34	0.73	98.00	0.28	38.77
	ACA	2771.40	68.40	8.19	0.62	93.05	0.14	35.95
	DiffPGD	1297.00	9.61	31.68	0.94	14.67	0.03	42.52
	AdvDrop	2256.40	82.40	41.69	**0.99**	5.40	0.02	44.70
	PerC-AL	1908.00	97.40	42.15	0.98	8.11	0.02	43.07
	SSAH	2214.00	98.00	43.33	0.99	13.31	**0.01**	47.01
	CoCoGen (Ours)	4111.54	**100.00**	**44.16**	**0.99**	**4.59**	**0.01**	**61.23**
ViT Base Dosovitskiy (2020)	PGD	4255.00	99.70	34.09	0.92	41.99	0.08	40.76
	DeepFool	1600.30	99.00	37.85	0.95	17.90	0.05	44.50
	NCF	1546.00	57.52	28.31	0.72	98.07	0.26	38.74
	ACA	1878.90	55.92	8.49	0.61	72.32	0.13	36.68
	DiffPGD	1325.70	7.19	31.69	0.94	13.70	0.04	42.49
	AdvDrop	1538.90	87.00	40.30	**0.99**	4.21	0.02	44.80
	PerC-AL	1920.00	**100.00**	**42.11**	0.98	8.35	0.02	43.21
	SSAH	1426.00	95.00	41.78	0.98	20.76	0.03	45.39
	CoCoGen (Ours)	4458.13	**100.00**	41.98	**0.99**	**3.56**	**0.01**	**63.70**

4.1 Result Comparison with State-of-the-Art Attacks

We evaluate CoCoGen against a comprehensive suite of untargeted white-box adversarial attacks (Table 1), spanning several methodological families: gradient-based attacks (PGD Madry et al. (2018), DeepFool Moosavi-Dezfooli et al. (2016)), perceptual and frequency-constrained attacks (PerC-AL Zhao et al. (2020), SSAH Luo et al. (2022), NCF Yuan et al. (2022)), content-aware attacks (ACA Chen et al. (2024)), stochastic feature-dropping attacks (AdvDrop Duan et al. (2021)), and diffusion-guided attacks (DiffPGD Xue et al. (2023)). We benchmark across four architecturally diverse target models—ResNet-50 He et al. (2016), EfficientNet-B0 Tan & Le (2019), ConvNeXt-Base Liu et al. (2022), and ViT-Base Dosovitskiy (2020)—to assess generalization across convolutional and attention-based inductive biases.

Attack Success Rate. CoCoGen achieves 100% Attack Success Rate (ASR) across all four target architectures, the only method to do so uniformly. Among baselines, PGD comes closest, reaching 100% on ResNet-50 but degrading to 98.50%, 99.70%, and 99.70% on EfficientNet-B0, ConvNeXt-Base, and ViT-Base, respectively. Diffusion-guided methods fare considerably worse: DiffPGD attains only 27.74% ASR on ResNet-50, collapsing further to 9.61% and 7.19% on ConvNeXt-Base and ViT-Base, suggesting that diffusion-manifold projections introduce constraints that are too conservative to reliably cross decision boundaries, particularly for architectures with more complex decision geometries. NCF similarly struggles beyond ResNet-50, dropping to 44.91% on EfficientNet-B0 and 57.52% on ViT-Base, indicating that its natural colour-flow constraints are insufficiently expressive to generalise across diverse feature spaces. ACA exhibits the sharpest degradation across architectures, falling from 97.34% on ResNet-50 to just 55.92% on ViT-Base, reflecting the limits of purely content-aware perturbation strategies when confronted with global attention mechanisms. In contrast, CoCoGen's consistent 100% ASR across all architectures demonstrates that targeting decision-critical, high-frequency spectral components provides a universally effective perturbation pathway regardless of the underlying architectural inductive bias.

Signal Fidelity. Beyond attack effectiveness, CoCoGen achieves substantial improvements in perceptual and signal-level fidelity across all architectures and metrics. On ResNet-50, PSNR increases from 33.51 dB

under PGD to 44.67 dB under CoCoGen, a gain of over 11 dB, while SSIM improves from 0.91 to 0.99, approaching the theoretical maximum. This is particularly significant because PGD, despite matching CoCoGen's 100% ASR, does so at a considerable perceptual cost, injecting broadband gradient energy indiscriminately across the signal spectrum. On EfficientNet-B0, CoCoGen improves PSNR from 32.99 dB (PGD) to 41.67 dB, and on ConvNeXt-Base from 34.04 dB (PGD) to 44.16 dB, with SSIM reaching 0.99 in both cases. On ViT-Base, PSNR reaches 41.98 dB with SSIM of 0.99, outperforming all baselines on the structural similarity axis while matching the only competing method (PerC-AL) that achieves 100% ASR on this architecture. Perceptual quality metrics confirm these findings. CoCoGen achieves the lowest LPIPS scores (0.01) across all four architectures, matching or outperforming SSAH, the next-best method on this metric, and attains the lowest FID scores on EfficientNet-B0 (4.05), ConvNeXt-Base (4.59), and ViT-Base (3.56), indicating that the distributional shift induced by CoCoGen's perturbations is minimal relative to the clean image manifold. On ResNet-50, FID is 10.18, competitive with SSAH (11.06) and PerC-AL (10.44), though slightly above AdvDrop's 4.43, the one setting where AdvDrop's frequency-dropping strategy proves competitive on distributional fidelity, albeit at the cost of a 17.28 percentage point reduction in ASR relative to CoCoGen. Most strikingly, CoCoGen achieves MUSIQ perceptual quality scores of 61.44, 62.63, 61.23, and 63.70 across ResNet-50, EfficientNet-B0, ConvNeXt-Base, and ViT-Base respectively, substantially outperforming all baselines, whose MUSIQ scores peak at 47.01 (SSAH on ConvNeXt-Base). This ∼14-17 point absolute improvement on the no-reference perceptual quality axis indicates that CoCoGen's adversarial examples are not merely metrically similar to clean images, but are perceived as higher quality in their own right, a property that no baseline attack approaches.

Interpretation. These gains are not incidental; they follow directly from CoCoGen's perturbation construction principle. Rather than injecting perturbation energy uniformly across all frequency bands, as gradient-based methods inherently do, CoCoGen confines $\boldsymbol{\delta}$ to decision-critical, high-frequency spectral components identified via the adaptive grid search over $|\mathcal{K}|$ sub-problems. By aligning gradient updates with the most competitive incorrect class and restricting support to spectrally coherent signal components, CoCoGen crosses the decision boundary through perceptually minimal signal modifications that avoid contaminating the low-frequency luminance and chrominance channels most sensitive to human visual perception. This spectral coherence explains why PSNR and SSIM, both of which penalise broadband signal deviation, improve so markedly over gradient-based baselines: perturbation energy is concentrated where the model is most discriminatively sensitive, not where the pixel-level ℓ_p budget is easiest to spend. The low FID and LPIPS scores further confirm that this frequency-selective strategy keeps adversarial examples on or near the natural image manifold, avoiding the distributional drift characteristic of methods such as NCF, ACA, and DiffPGD that impose strong but misaligned constraints on the perturbation structure.

Baseline Comparison. Among the baselines, SSAH is the closest competitor on perceptual fidelity for convolutional architectures, achieving PSNR values of 44.04 dB and 43.33 dB on ResNet-50 and ConvNeXt-Base respectively, with SSIM of 0.99 in both cases. However, SSAH's ASR degrades on ViT-Base (95.00%) and its MUSIQ scores consistently lag CoCoGen by a wide margin, suggesting that frequency-domain regularisation alone is insufficient without the decision-boundary-aligned spectral targeting that CoCoGen provides. AdvDrop achieves the lowest FID on ResNet-50 (4.43) and competitive LPIPS scores, but at ASR values well below 100% across all architectures, a fundamental trade-off that CoCoGen resolves by demonstrating that high distributional fidelity and high ASR are simultaneously achievable through spectrally informed perturbation design.

Runtime Considerations. CoCoGen's runtime ranges from 1,225.41 s on ResNet-50 to 4,458.13 s on ViT-Base, reflecting the overhead of the adaptive grid search over $|\mathcal{K}|$ independent spectral sub-problems. Compared to baselines, this exceeds single-pass gradient methods such as PGD (462.65-4,255.00 s depending on architecture) but remains comparable in order of magnitude to other iterative or diffusion-based approaches. Critically, the $|\mathcal{K}|$ sub-problems are embarrassingly parallel (Sec. 2.7), meaning that wall-clock time reduces proportionally with the number of available GPU cores. In multi-GPU or distributed settings, CoCoGen's effective runtime becomes competitive with or faster than several baselines, and the quality,efficiency trade-off is therefore a function of hardware provisioning rather than an intrinsic algorithmic limitation.

4.2 Ablation Study

4.2.1 Effect of Contrastive Counterfactual Guidance

To isolate the contribution of contrastive counterfactual guidance, we compare two variants that both achieve 100% ASR but differ only in whether the margin objective $\mathcal{M}$ is used to steer the perturbation (Table 2).

Without guidance, the gradient signal is undifferentiated across all input dimensions, so the optimiser distributes perturbation energy broadly across the image rather than concentrating it where the classifier is most sensitive. This diffuse energy placement is the direct cause of the degraded fidelity observed: SSIM falls to ~ 0.86 and PSNR to $\sim 34\,\text{dB}$, meaning the adversarial image departs visibly from the original despite carrying no more total perturbation energy. The elevated FID scores (22-38) further indicate that the distribution of adversarial images has drifted substantially from the clean image distribution, a sign that structural artefacts have been introduced across the batch. With counterfactual guidance, the margin $\mathcal{M}$ explicitly identifies the most competitive incorrect class and directs gradient updates toward the signal dimensions that most influence the logit gap between the true class and that competitor. Because only a small subset of pixels drives the classification decision, this targeting concentrates the perturbation into a sparse, semantically meaningful support rather than spreading it uniformly. The effect is substantial: SSIM rises uniformly to 0.99 across all architectures, PSNR increases by 8-11 dB, LPIPS drops by an order of magnitude (from ~ 0.10 to 0.01), and FID falls to single digits, indicating that the adversarial distribution is nearly indistinguishable from the clean one. The consistency of these gains across CNN and transformer architectures suggests the benefit is not model-specific but reflects a general property of margin-directed perturbation: attacking the decision boundary directly requires far less signal energy than naive loss maximisation, leaving the remainder of the image untouched.

Table 2: **Effect of Contrastive Counterfactual Guidance.** All experiments maintain 100% ASR. **W/** and **W/O** denote with and without counterfactual guidance, respectively. ↑ higher is better; ↓ lower is better.

Metric	Guidance	ResNet-50	EffNet-B0	ConvNeXt	ViT-B
SSIM ↑	W/O	0.866	0.861	0.864	0.862
	W/	**0.990**	**0.990**	**0.990**	**0.990**
PSNR ↑	W/O	34.02	33.72	33.88	33.65
	W/	**44.67**	**41.67**	**44.16**	**41.98**
FID ↓	W/O	22.11	34.47	27.56	38.16
	W/	**10.18**	**4.05**	**4.59**	**3.56**
LPIPS ↓	W/O	0.105	0.121	0.092	0.142
	W/	**0.010**	**0.010**	**0.010**	**0.010**
MUSIQ ↑	W/O	47.08	48.15	48.65	47.50
	W/	**61.44**	**62.63**	**61.23**	**63.70**

4.2.2 Impact of High-Frequency Signal Components

To isolate the contribution of the Fourier-domain projection operator $\mathcal{P}_f$, we compare variants that apply versus omit frequency masking while keeping all other components fixed and maintaining 100% ASR throughout (Table 3).

Without $\mathcal{P}_f$, the optimiser is free to place perturbation energy at any spatial frequency, and in practice it exploits low-frequency components because they carry broad, spatially coherent gradient signal that efficiently reduces the margin $\mathcal{M}$. Low-frequency modifications, however, correspond to global changes in luminance, colour, and coarse structure — precisely the components to which the human visual system (HVS) is most sensitive Wang et al. (2004). The result is a perturbation that, while still technically bounded by $\|\boldsymbol{\delta}\|_\infty \leq \epsilon$, introduces visible blurring or tinting artefacts that degrade SSIM to ~ 0.96 and suppress PSNR to $\sim 39\,\text{dB}$. The no-reference MUSIQ scores (46-48) further confirm that the images are perceived as lower quality, since MUSIQ integrates multi-scale texture and sharpness cues that are disrupted by low-frequency contamination.

Applying $\mathcal{P}_f$ constrains the perturbation to the high-frequency subspace by zeroing all Fourier coefficients below a radial threshold τ_{freq}, as defined in equation 10. High-frequency components correspond to fine edges, textures, and noise-like patterns — regions where the HVS has markedly reduced contrast sensitivity, meaning larger perturbation amplitudes can be tolerated before the modification becomes visible.

Table 3: **Effect of High-Frequency Signal Masking on Perceptual Quality.** All experiments maintain 100% ASR. **W/** and **W/O** denote with and without applying the high-frequency projector $\mathcal{P}_f$ to the perturbation. ↑ higher is better; ↓ lower is better.

Metric	Masking	ResNet-50	EffNet-B0	ConvNeXt	ViT-B
SSIM ↑	W/O	0.961	0.963	0.961	0.967
	W/	**0.990**	**0.990**	**0.990**	**0.990**
PSNR ↑	W/O	39.49	38.49	39.26	39.68
	W/	**44.67**	**41.67**	**44.16**	**41.98**
FID ↓	W/O	11.64	25.08	11.27	8.71
	W/	**10.18**	**4.05**	**4.59**	**3.56**
LPIPS ↓	W/O	0.03	0.03	0.02	0.02
	W/	**0.01**	**0.01**	**0.01**	**0.01**
MUSIQ ↑	W/O	48.07	46.59	47.98	48.32
	W/	**61.44**	**62.63**	**61.23**	**63.70**

By routing all adversarial energy into this perceptually insensitive subspace, $\mathcal{P}_f$ breaks the coupling between attack effectiveness and perceptual cost: the classifier boundary can still be crossed because modern deep networks are known to rely heavily on high-frequency texture cues, yet the modification remains imperceptible to a human observer. The gains are consistent across all four architectures: SSIM reaches 0.99, PSNR increases by 4-6 dB, LPIPS halves, and MUSIQ rises by 13-16 points. The FID improvement is particularly pronounced for EfficientNet-B0 (25.08 → 4.05) and ViT-Base (8.71 → 3.56), suggesting that those architectures respond to adversarial updates in frequency bands that, when unconstrained, introduce distributional artefacts visible at the batch level. The orthogonal projector structure of $\mathcal{P}_f$ (established in Proposition 1) ensures that the frequency constraint is enforced exactly at every iteration rather than softly penalised, which prevents low-frequency energy from leaking back in through momentum accumulation.

Attack Efficiency. High-frequency projection also improves attack efficiency (Fig. 4): ResNet-50 achieves 100% ASR with only 2,040 perturbed pixels (vs. 9,700 without projection), and ConvNeXt-Base with 4,310 (vs. 8,700). This reduction arises because constraining the perturbation to the high-frequency subspace fundamentally changes how optimisation allocates signal energy.

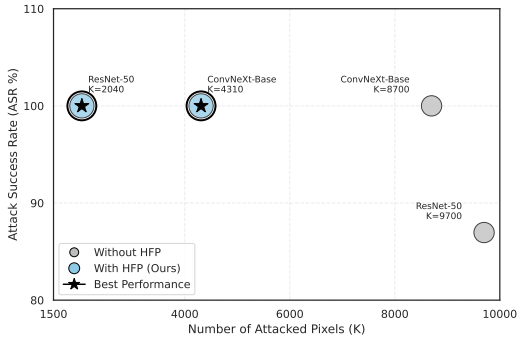

Figure 4: **Effect of high-frequency subspace projection on attack efficiency.** Fewer perturbed pixels indicate higher efficiency; 100% ASR is targeted in all cases. HFP denotes high-frequency projection, and K denotes the number of perturbed pixels.

Without $\mathcal{P}_f$, gradient updates are dominated by low-frequency components, which have large spatial support and therefore distribute perturbation energy broadly across the image. While such components are effective at reducing the margin $\mathcal{M}$, they do so inefficiently: many pixels are modified slightly, resulting in dense perturbations with high cardinality.

In contrast, applying $\mathcal{P}_f$ suppresses all low-frequency directions and forces the optimiser to operate exclusively in the high-frequency subspace. These components correspond to localised edges and fine textures, which have spatially concentrated support. As a result, each update affects a smaller subset of pixels but with greater directional alignment to features that strongly influence the classifier's decision boundary. This increases the *per-pixel effectiveness* of the perturbation: fewer pixels are required to achieve the same reduction in $\mathcal{M}$.

Moreover, modern deep networks are known to rely heavily on high-frequency texture cues for classification, so restricting the attack to this subspace preserves access to decision-critical features while eliminating redundant low-frequency directions. The combination of directional filtering and feature alignment therefore

yields perturbations that are both sparser and more targeted, achieving boundary crossing with significantly fewer modified pixels.

4.2.3 Impact of Adaptive Grid Search

We analyze the importance of adaptive grid search (Table 4). Without it, ASR slightly drops (99%–99.8%) and perceptual quality degrades, reflecting inefficient allocation of perturbation energy across the feasible set.

The role of adaptive grid search is to resolve the intrinsic trade-off between attack strength and perceptual fidelity by dynamically selecting the most effective constraint configuration from the candidate set $\mathcal{K}$. In the absence of this search, a fixed configuration (e.g., a single choice of frequency threshold or sparsity level) may be suboptimal: if the constraint is too weak, perturbations spread across redundant dimensions, leading to dense, visually noticeable modifications; if too strong, the feasible set becomes overly restrictive, limiting the optimiser's ability to reduce the margin $\mathcal{M}$ and achieve consistent misclassification.

Adaptive grid search mitigates this by evaluating multiple constraint settings and selecting the one that yields the greatest margin reduction under the feasibility constraints. This effectively performs a discrete optimisation over the geometry of the feasible set itself, rather than optimising solely within a fixed set. As a result, the method identifies configurations in which the perturbation is concentrated on the most decision-critical signal components, avoiding both under- and over-constrained regimes.

Table 4: **Effect of Adaptive Grid Search.** In all experiments **W/** and **W/O** denote with and without adaptive grid search, respectively. ↑ higher is better; ↓ lower is better.

Metric	Search	ResNet-50	EffNet-B0	ConvNeXt	ViT-B
ASR ↑	W/O	100.00	99.80	100.00	99.00
	W/	**100.00**	**100.00**	**100.00**	**100.00**
PSNR ↑	W/O	39.49	38.49	39.26	39.68
	W/	**44.67**	**44.16**	**41.67**	**41.98**
FID ↓	W/O	11.64	25.08	11.27	8.71
	W/	**10.18**	**4.59**	**4.05**	**3.56**
LPIPS ↓	W/O	0.03	0.03	0.02	0.02
	W/	**0.01**	**0.01**	**0.01**	**0.01**
MUSIQ ↑	W/O	48.07	46.59	47.98	48.32
	W/	**61.44**	**61.23**	**62.63**	**63.70**

This improved alignment has two key effects. First, it ensures that a valid decision-boundary crossing is achieved in all cases, restoring ASR to 100% across architectures. Second, it significantly improves perceptual quality: PSNR increases by 4-6 dB, LPIPS drops to 0.01, and MUSIQ improves by more than 10 points. These gains indicate that the selected configurations concentrate perturbation energy into a minimal, highly effective support, rather than distributing it across the image.

Consistent with this interpretation, adaptive grid search reduces the number of perturbed pixels by up to 9×, demonstrating that it does not merely improve attack success, but does so through more efficient and localised signal modification. In this sense, adaptive search acts as a higher-level control mechanism that selects the most favourable optimisation landscape for each input, enabling CoCoGen to operate near the optimal trade-off between attack efficacy and perceptual cost.

5 Conclusion

We proposed **Co**ntrastive **Co**unterfactual **Gen**eration (**CoCoGen**), an adversarial attack that minimises the contrastive counterfactual margin under explicit spatial and spectral constraints. By directly optimising the decision boundary geometry, CoCoGen provides a consistent gradient signal across all optimisation regimes while enforcing a strict energy budget. More specifically, adversarial vulnerability is concentrated in a small set of decision-critical signal components. By combining Top-k spatial masking with high-frequency Fourier projection, the method confines perturbations to a structured, low-dimensional subspace aligned with these components. This alignment enables highly efficient optimisation: perturbation energy is not spread across the image but concentrated where it most effectively reduces the margin.
Our experimental results demonstrate 100% ASR across diverse architectures while maintaining near-imperceptible distortion (SSIM ≈ 0.99, LPIPS ≈ 0.01, MUSIQ 61-63). These results show that reliable

misclassification does not require large or diffuse perturbations, but rather precise, spectrally coherent modifications targeted at the classifier's decision boundary.
Future work will explore adaptive and learned support selection, extend the framework to targeted and black-box settings, and generalise the signal-domain formulation to other modalities such as audio and video.

References

N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57. IEEE, 2017.

J. Chen, H. Chen, K. Chen, Y. Zhang, Z. Zou, and Z. Shi. Diffusion models for imperceptible and transferable adversarial attack. *arXiv preprint arXiv:2305.08192*, 2023a.

X. Chen, X. Gao, J. Zhao, K. Ye, and C. Xu. Advdiffuser: Natural adversarial example synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4562–4572, 2023b.

Z. Chen, Z. Wang, J. Huang, W. Zhao, X. Liu, and D. Guan. Imperceptible adversarial attack via invertible neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 414–424, 2023c.

Z. Chen, B. Li, S. Wu, K. Jiang, S. Ding, and W. Zhang. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems*, 36, 2024.

J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

R. Duan, Y. Chen, D. Niu, Y. Yang, A. Qin, and Y. He. Advdrop: Adversarial attack to dnns by dropping information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7506–7515, 2021.

Xiang Fang, Wanlong Fang, and Changshuo Wang. Unveiling the fragility of vision-language models: Multi-modal adversarial synergy via texture-constrained perturbations and cross-modal optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3867–3875, 2026.

I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer VVision and pattern recognition*, pp. 770–778, 2016.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

S. Jia, C. Ma, T. Yao, B. Yin, S. Ding, and X. Yang. Exploring frequency adversarial attacks for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4103–4112, 2022.

Guy Katz, Clark Barrett, David Dill, Kyle Julian, and Mykel Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pp. 3–21. Springer, 2017.

J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

C. Laidlaw, S. Singla, and S. Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *International Conference on Learning Representations*, 2021.

M. Lee and D. Kim. Robust evaluation of diffusion-based adversarial purification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 134–144, 2023.

Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

A. Luo, E. Li, Y. Liu, X. Kang, and Z J. Wang. A capsule network based approach for detection of audio spoofing attacks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6359–6363. IEEE, 2021.

A. Luo, C. Kong, J. Huang, Y. Hu, X. Kang, and A. C Kot. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security*, 19:1168–1182, 2023.

B. Luo, Y. Liu, L. Wei, and Q. Xu. Towards imperceptible and robust adversarial example attacks against neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

C. Luo, Q. Lin, W. Xie, B. Wu, J. Xie, and L. Shen. Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15315–15324, 2022.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Cision and Pattern Recognition*, pp. 10684–10695, 2022.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

H. Salman, A. Ilyas, L. Engstrom, A. Kapoor, and A. Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

M. Sharif, L. Bauer, and M. Reiter. On the suitability of lp-norms for creating and preventing adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1605–1613, 2018.

N. D. Singh, F. Croce, and M. Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. *Advances in Neural Information Processing Systems*, 36, 2024.

Y. Song, R. Shu, N. Kushman, and S. Ermon. Constructing unrestricted adversarial examples with generative models. *Advances in neural information processing systems*, 31, 2018.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Z. Wei, J. Chen, Z. Wu, and Y. Jiang. Enhancing the self-universality for transferable targeted attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12281–12290, 2023.

H. Xue, A. Araujo, B. Hu, and Y. Chen. Diffusion-based adversarial sample generation for improved stealthiness and controllability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

S. Yuan, Q. Zhang, L. Gao, Y. Cheng, and J. Song. Natural color fool: Towards boosting black-box unrestricted attacks. *Advances in Neural Information Processing Systems*, 35:7546–7560, 2022.

X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2805–2824, 2019.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Z. Zhao, Z. Liu, and M. Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1039–1048, 2020.