Frame the adversary: a structure-aware attack methodology

Vicky Kouni
Paris Dauphine - PSL University
vasiliki.kouni@lamsade.dauphine.fr

Stelios Perrakis
CentraleSupélec - Paris-Saclay University
stelios.perrakis@centralesupelec.fr

Francis Bach
Inria - Ecole Normale Supérieure - PSL University
francis.bach@inria.fr

Pascal Frossard
EPFL
pascal.frossard@epfl.ch

Yann Chevaleyre
Paris Dauphine - PSL University
yann.chevaleyre@lamsade.dauphine.fr

Abstract

Frequency-based adversarial attacks have recently grown popular by exploiting spectral sensitivities shared across neural architectures. Unlike spatial perturbations, frequency-based attacks expose deeper vulnerabilities, making them especially valuable for robust evaluation of safety-critical and security-sensitive applications. Yet, existing approaches are typically not derived as solutions to an optimization problem that explicitly captures transform-domain structure. In this paper, we propose a methodology for crafting principled frequency-based adversarial attacks, via a dedicated optimization framework. A cornerstone of our method hinges on the introduction of a perturbation constraint set, tied to highly structured non-orthogonal transforms, well-known for their flexible, non-predefined frequency handling. We prove that the attacks emerge as weighted ℓ_2-projections onto this set, yielding a general and controlled attack generation mechanism. By this, we provide a clear geometric attack characterization, ensuring alignment between the optimization objective and the perturbation constraint. We assess our framework on standardized datasets, for pretrained and adversarially robust models. Results highlight that our attacks, being solutions to an optimization problem, over a structured perturbation set, are highly effective, even across different, unseen architectures. Our methodology could serve as a theoretical baseline for designing and analyzing transformed-based attacks, targeting fundamental model vulnerabilities, instead of mere architecture-specific artifacts typically studied in the robustness literature.

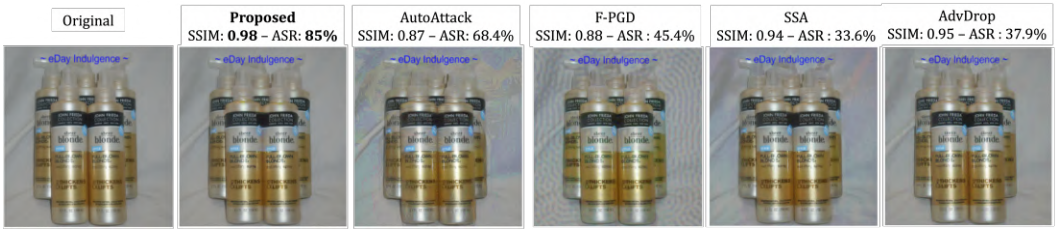
1 Introduction

Frequency-based attacks [58, 62, 65, 66] have recently emerged as a powerful paradigm in adversarial learning. By leveraging the sensitivity of models to frequency components of the data, these attacks reveal vulnerabilities that often remain hidden in the spatial domain. This helps prevent catastrophic failures in safety-critical vision systems or privacy protection, since frequency-based attacks are highly effective, even across unseen target models, thus implying high transferability [49, 58, 66]. The latter is often attributed to the ability of frequency-based attacks to exploit shared spectral biases inherent to neural architectures [26, 50], rather than overfitting to model-specific spatial features, thereby targeting common vulnerabilities that persist across diverse model designs.

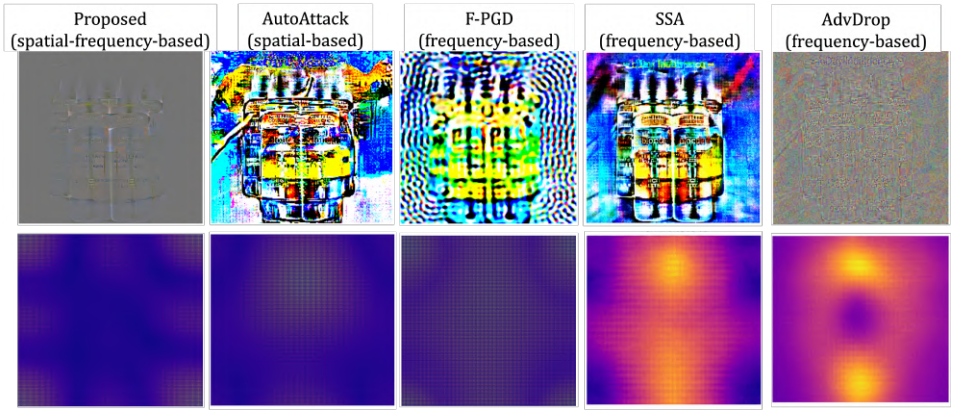
40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Most existing approaches rely on projected gradient descent (PGD) [44], by directly "plugging in" to the attack pipeline a frequency transform, e.g., Fourier [65], discrete cosine [30], or wavelet [43, 64], along with a selection mechanism on the transformed data. The motivation behind this mechanism hinges upon selectively amplifying or suppressing low-/high-frequency components, that carry the most structural information, thus exposing vulnerabilities that remain hidden in the pixel space. In this way, manipulating spectral information offers a more efficient attack regime, enabling perturbations that achieve higher attack success rates, with comparable computational effort.

Despite their success, state-of-the-art (SotA) frequency-based attacks are not accompanied by a theoretical justification of their design, thereby carrying a strong frequency bias. What is more, such attacks do not generally emerge as solutions to an open optimization problem, which, in turn, would reflect the true nature of adversarial learning in the presence of transforms, and help explaining the effectiveness and transferability of the attacks. Still, a few recent works have proposed optimization problems for creating frequency-based adversarial attacks [30, 67], but these are application-dependent, thus rendering the generalization to other settings unclear. The interaction between the attack constraint, the transform-domain representation, and the resulting perturbation structure is thus not fully understood yet. Without a principled methodology explaining the effectiveness and transferability of frequency-based attacks, it remains elusive whether attacks should rely on certain frequency components and in what way [27, 57, 65, 68].



(a) Example **ImageNet** adversarial images, with **SSIM** quality and **ASR** attack efficiency values. Bold letters indicate the method with the best performance in terms of both evaluation metrics.



(b) Spatial (top) and **frequency (bottom)** representation of the **proposed attack** and the four **baselines**, corresponding to the adversarial examples of Fig. 1a. In contrast to the baselines, our attack enjoys strategically-placed energy localization, combining the "best of both worlds". This allows it to align with the model's spectral sensitivities, exploiting critical feature extraction pathways to achieve the best attack success rate.

Fig. 1: Evaluation of the baselines ($\varepsilon = 8/255$) and our attack ($\varepsilon_{\text{DGF}} = \eta 8/255$), on an exemplary RobustBench model [2].

Motivated by recent advances in frequency-based adversarial attacks, and by the lack of a principled method and solid understanding of their underpinnings, in this paper, we aim to address the following core question:

What are the underlying mechanisms of success of frequency-based adversarial attacks? (†)

To that end, our main contributions are as follows:

- We propose *a new attack methodology, based on overcomplete (non-orthogonal) spatial-frequency (SF) transforms*, which are popular tools for image-processing tasks [40, 47]. The

overcompleteness enables flexible and structured representations that orthogonal transforms like Fourier or the discrete cosine cannot capture in baseline frequency-based attacks.

- Our framework consists of designing an optimization problem, with *a cornerstone being the introduction of a new perturbation set* for constraining the attacks, which satisfy a weighted SF norm constraint. The latter spreads attack energy across structured SF directions, aligned with models' spectral sensitivities, making the attack more effective and transferable. We employ PGD w.r.t. the proposed set and prove that its projection yields a closed-form attack, as the solution to our optimization problem. To the best of our knowledge, this is *the first principled methodology giving a convincing partial answer to* (†), by exploiting the geometry of the set constraining the attacks, through the structure of the SF transform.
- We evaluate the success of our proposed framework through a series of experiments. *Our findings are three-fold*: a) our attack is highly effective on a range of different source-target models, exposing vulnerabilities even of defended models, thereby indicating high transferability; b) comparisons with SotA representative spatial- and frequency-based attacks demonstrate that our method outperforms the baselines in terms of effectiveness and transferability, while preserving the underlying image structure. This behavior showcases that attacks resulting from a well-defined optimization problem, while respecting the structure of the transforms involved, can enjoy improved performance, without sacrificing visual quality; c) the observed transferability sheds light on the underpinnings of frequency-based attacks: structuring the perturbation set via geometry-aligned transforms, is key to capturing spectral sensitivities shared across architectures, bolstering our theoretical framework.

Overall, our principled methodology illuminates – beyond common high/low frequency assumptions – the effectiveness and transferability of frequency-based adversarial attacks, offering a flexible alternative to manually designed frequency selection strategies in the attacks. This enables a systematic study of how properties of involved transforms influence adversarial vulnerability. Our framework could be deployed as a testbed for developing more robust evaluation protocols and defenses that explicitly account for transform-domain structure, e.g., benchmarking robustness for structured, transformed-based perturbation sets. Finally, it could contribute to the development of a more holistic understanding of models' vulnerabilities to data perturbations, towards building safer AI systems.

2 Related work

Spatial-based attacks. For a loss function $\mathcal{L}(\cdot,\cdot)$ and input-output data pairs (x, y), a standard regime for creating adversarial attacks in the pixel space consists of solving

$$\max_{\delta\in\mathbb{R}^n} \mathcal{L}(x+\delta, y) \quad \text{such that} \quad \|\delta\|_p \leq \varepsilon. \tag{1}$$

The popular method of PGD [44] relies on the projection on the ℓ_p-ball to yield the solution of (1) after $k = 0, \ldots, K-1$ iterations, i.e.,

$$\delta^{k+1} = \Pi_{\|\cdot\|_p}(\delta^k + \gamma \cdot \mathrm{norm}(\nabla_\delta \mathcal{L}(x+\delta^k, y))). \tag{2}$$

Here, $\Pi_{\|\cdot\|_p}(\cdot)$ is the projection operator on the ℓ_p-ball, $\gamma > 0$ is the step size, and $\mathrm{norm}(\cdot)$ pertains to some form of normalization for the gradients, to ensure that each update moves optimally within the allowed ℓ_p-ball. For instance, if $p = 2$, then $\mathrm{norm}(\nabla_\delta \mathcal{L}(x+\delta^k, y)) = \nabla_\delta \mathcal{L}(x+\delta^k, y)/\|\nabla_\delta \mathcal{L}(x+\delta^k, y)\|_2$ while for $p = \infty$, it holds $\mathrm{norm}(\nabla_\delta \mathcal{L}(x+\delta^k, y)) = \mathrm{sign}(\nabla_\delta \mathcal{L}(x+\delta^k, y))$. This implies that the gradients are normalized by using the dual norm of $\|\cdot\|_p$. Finally, the adversarial image is $x_{\mathrm{adv}} = x + \delta^K$, clipped in $[0, 1]$ to fall within a valid image range.

Frequency-based attacks. The PGD framework can be extended to attacks relying on a frequency representation of x. SotA frequency-based attacks take into account the sensitivity of models to high/low frequencies, and create PGD-type attacks, which rely on some orthogonal frequency transform $D \in \mathbb{R}^{n\times n}$, e.g., wavelet [27, 41], discrete cosine [58, 62, 66], or Fourier [37, 49, 65]. Then, a frequency-based counterpart of (2) is generally formulated as

$$\delta^{k+1} = \Pi_{\|\cdot\|_p}(\delta^k + \gamma \cdot \mathrm{norm}(g^k)), \tag{3}$$

where

$$g^k = \nabla_\delta \mathcal{L}(D^{-1}\mathrm{mask}(D(x+\delta^k)), y), \tag{4}$$

with the resulting adversarial image being $x_{\text{adv}} = x + \delta^K$, clipped in $[0, 1]$, similarly to the spatial attacks. In (4), $\text{mask}(\cdot)$ denotes a mechanism selecting a particular frequency range to which models are sensitive, to increase the effectiveness of the attack, or make it easily transferable across target models. This design is motivated by observations that neural networks exhibit spectral biases [26, 50], making them more vulnerable to perturbations in certain frequency ranges. For instance, $\text{mask}(\cdot)$ can be a random multiplicative or binary mask [62, 65], or some nonlinear transform selecting the high-/low-frequency components of the input image [21, 58].

Despite their empirical success, frequency-based attacks of the form (4) are not accompanied by a theoretical justification of their design, since the data's representation w.r.t. D is directly plugged in to the gradient of the loss. Consequently, the resulting perturbation is not guaranteed to solve a maximization problem that explicitly incorporates transform-domain structure. Moreover, standard ℓ_p-constraints do not reflect the geometry induced by the transform, leading to a mismatch between the constraint set and the attack construction.

Recent works have taken steps towards addressing these limitations by formulating optimization problems that include structure in the frequency domain, via task-specific masking strategies. While effective in specific applications like face forgery [30] or object detection [67], these methods lack a general framework and do not clearly characterize the constraint sets onto which PGD projects. As a result, deriving frequency-based attacks as solutions to an appropriately defined optimization problem, satisfying a structured perturbation constraint, stays largely unexplored.

3 Background

Notation. We write $\mathbb{F}$ for $\mathbb{R}$ or $\mathbb{C}$. We denote with Ψ^* the conjugate transpose, i.e., $\Psi^* = \overline{\Psi}^T$, and with $\Psi^+ = (\Psi^*\Psi)^{-1}\Psi^*$ the pseudo-inverse (Moore Penrose inverse) of $\Psi \in \mathbb{C}^{N\times n}$; the real part of Ψ is $\Re(\Psi)$.

Frames vs. orthogonal bases. While orthogonal bases provide elegant data representations, overcomplete bases, dubbed frames [12], offer crucial data processing advantages. A frame is an overcomplete set that spans the ambient space with more vectors than the space's dimension (cf. Appendix A). This overcompleteness can be leveraged to design the frame in a flexible fashion, to align with data geometry. For instance, SF frames are often constructed to match data geometry, hence improving performance in feature extraction [42], since overcompleteness enables the frame to flexibly capture localized and oriented patterns [47]. Next, we present operators associated to a frame, and move on to the representative class of Gabor frames; these will constitute the basis of our adversarial framework.

Definition 1 ([12]). *Let $(\psi_i)_{i=1}^N$, $N \geq n$, be a frame in $\mathbb{F}^n$. Then, for $x \in \mathbb{F}^n$, the analysis operator $\Psi \in \mathbb{F}^{N\times n}$ associated to the frame is defined as $\Psi x := (\langle x, \psi_i\rangle)_{i=1}^N$. The Grammian operator associated to the frame is defined in matrix form as $G = \Psi\Psi^* \in \mathbb{F}^{N\times N}$.*

Remark 2. *The analysis operator is injective, while the Grammian is rank deficient whenever $N > n$, and reveals intriguing properties of the frame: it reflects the linear dependencies among the frame elements and thus informs on its algebraic structure. Frames enable lossless data expansions: for any $x \in \mathbb{F}^n$, it holds $x = \Psi^*(\Psi^*)^+x = \Psi^+\Psi x$, where $\Psi^+ = (\Psi^*\Psi)^{-1}\Psi^*$ is the pseudoinverse of Ψ, serving as its left-inverse, due to the injectivity of Ψ. Importantly, frame theory is extendable to more dimensions, e.g., for data $x \in \mathbb{R}^{n\times n}$, by using tensor algebra; we give more details on Appendix A.*

Given a row-normalized $\Psi \in \mathbb{F}^{N\times n}$, its *average coherence* [6] is defined as

$$\nu(\Psi) = \frac{1}{N-1} \max_i \left| \sum_{j=1, j\neq i}^{N} \langle \psi_i, \psi_j \rangle \right|, \qquad i = 1, \ldots, N. \tag{5}$$

Average coherence relies on the Grammian and reflects a frame's geometry, by measuring the spread of the rows within the n-dimensional unit ball: the smaller the coherence, the more spread out the row vectors. On the contrary, large values of $\nu(\Psi)$ pinpoint to geometric degeneracy, e.g., vectors are poorly distributed on the ambient space, due to the large number of linear dependencies among them.

Definition 3 ([47]). *A discrete Gabor frame (DGF) results from SF shifts of a nonzero vector $g \in \mathbb{F}^n$. The shifts are controlled by the Gabor SF parameters $\alpha, \beta > 0$, $\alpha\beta < n$, so that the frame elements are $\psi_{k,p,j} = \exp(2\pi i\beta pj/n)g(j - \alpha k)$, where $k = 0, \ldots, (n/\alpha) - 1$, $p = 0, \ldots, (n/\beta) - 1$, $j = 0, \ldots, n-1$. By creating a grid over p, k, and representing them as a single index i, we obtain*

Algorithm 1: DGF-PGD

Input : Original image x, DGF analysis operator Ψ, attack level $\varepsilon > 0$
Output : Adversarial image x_{adv}

Calculate $\Re(M_D)$ and $(\Re(M_D))^{-1}$; initialize δ_0
for $k = 0, \dots, K-1$ **do**
  $\tilde{g}^k = \nabla_\delta \mathcal{L}(x + \delta^k, y)$;
  $\delta^{k+1} = \Pi_{\mathcal{C}_{M_D}}\left(\delta^k + \gamma \frac{(\Re(M_D))^{-1}\tilde{g}^k}{\sqrt{(\tilde{g}^k)^T(\Re(M_D))^{-1}\tilde{g}^k}}\right)$;
$x_{\text{adv}} = \text{clip}(x + \delta^K, 0, 1)$

$i = 0, \dots, N-1$, where $N = n^2/\alpha\beta$, so that we can equivalently write $\psi_{k,p,j} = \psi_{i,j}$. Then, the associated analysis operator is $\Psi \in \mathbb{C}^{N\times n}$, whose row elements are $\psi_i^* \in \mathbb{C}^n$, $i = 0, \dots, N-1$.

We note here that the Gabor parameters α, β crucially affect the representation ability of Ψ [34], and choosing them respectively is a challenging task [54]. Moreover, by Remark 2, we obtain $x = \Psi^+\Psi x$.

Why Gabor frames? Unlike global orthogonal transforms like Fourier/discrete cosine, DGFs provide a flexible SF data representation [14], with overcompleteness enabling simultaneous localization across space, frequency, and orientation, hence rendering them as great candidates for representing edges and directional patterns that naturally arise in images [24, 42]. This adaptive representation to structured and overlapping spectral patterns makes overcompleteness particularly important to the adversarial setting: by expressing attacks in the DGF domain, one can shape them to align with spectral sensitivities shared across different network architectures. What is more, the close connection between Gabor representations and visual processing mechanisms [18, 32] helps explain why Gabor frames induce norms closely related to human perception [8], which spatial-ℓ_p norms do not; this further motivates the construction of a DGF-based attack methodology, which we present in Section 4.

4 Main results

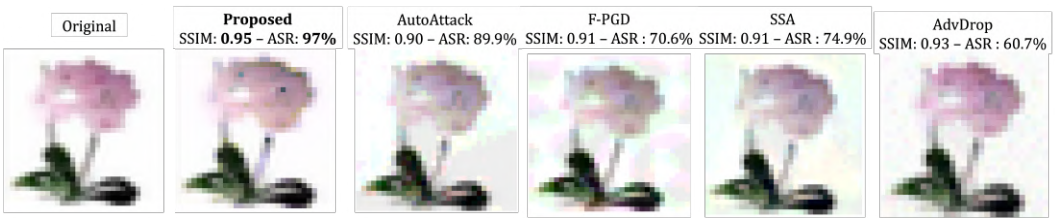
Definition 4. *For $\varepsilon > 0$ and $\Psi \in \mathbb{F}^{N\times n}$ being a DGF analysis operator, we say that $\delta \in \mathbb{F}^n$ is a spatial attack of level ε if $\|\delta\|_{\mathbb{F}^n} \leq \varepsilon$, w.r.t. some norm $\|\cdot\|_{\mathbb{F}^n}$ in $\mathbb{F}^n$. Additionally, we say that $w \in \mathbb{F}^N$ is an admissible frame attack of level ε if $\|w\|_{\mathbb{F}^N} \leq \varepsilon$ and $w = \Psi\delta$, for some $\delta \in \mathbb{F}^n$.*

Lemma 5 (Proof in Appendix B.1)**.** *For $\delta \in \mathbb{F}^n$ and $\varepsilon > 0$, let $w = \Psi\delta \in \mathbb{F}^N$, with $\Psi \in \mathbb{F}^{N\times n}$ being a DGF analysis operator. If $\delta \in \mathbb{F}^n$ is a spatial attack of level ε, then $w = \Psi\delta$ is an admissible frame attack of level $\|\Psi\|_{\text{op}}\varepsilon$, where $\|\Psi\|_{\text{op}} = \max_{\|\delta\|_{\mathbb{F}^n}\leq 1} \|\Psi\delta\|_{\mathbb{F}^N}$. Additionally, if $w \in \mathbb{F}^N$ is an admissible frame attack of level ε, then $\delta = \Psi^+ w$ is a spatial attack of level $\|\Psi^+\|_{\text{op}}\varepsilon$, where $\|\Psi^+\|_{\text{op}} = \max_{\|w\|_{\mathbb{F}^N}\leq 1} \|\Psi^+ w\|_{\mathbb{F}^n}$.*

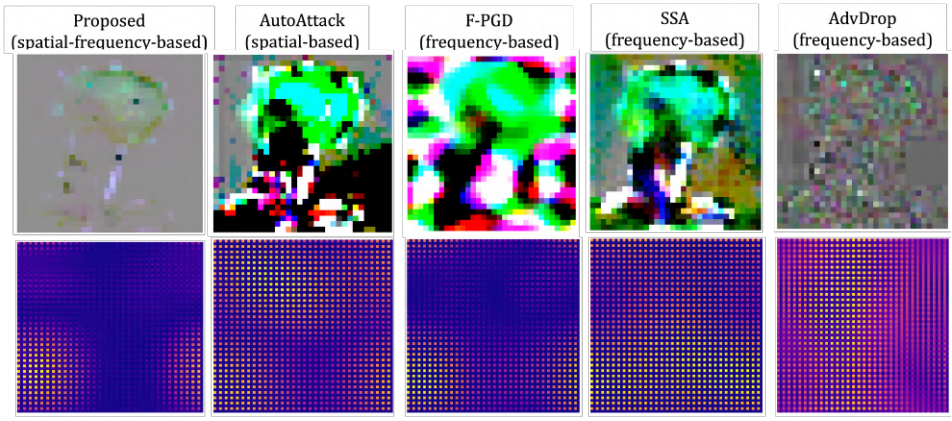
The purpose of Lemma 5 is to show that any admissible attack in the frame domain is induced by a suitable attack in the spatial domain and vice versa.

Threat model. We operate in a white-box regime, where the adversary has access to the model, input data and gradients, and the loss function. In Section 5.2, we also evaluate transfer-based black-box attacks, where adversarial examples crafted on a source model are evaluated on an unseen one. This combination assesses attack effectiveness, even beyond the source model, addressing (†). Finally, we introduce a structured, DGF-based perturbation set, whose geometry promotes SF-localized attacks.

Proposed adversarial attack. We are inspired by Zhang et al. [65] and Duan et al. [21], who propose frequency-based adversarial attacks relying on orthogonal transforms. Similarly, we derive frequency-based adversarial attacks, yet differentiate our approach in the following ways: a) we operate in the DGF (non-orthogonal) domain, where the associated analysis operator possesses inherent structure and offers greater flexibility in the data representation, b) we introduce a maximization problem over a new set of perturbations, satisfying a weighted-DGF norm constraint. We employ a variant of projected gradient ascent w.r.t. that set and prove that the associated projection operator yields a solution to our maximization problem, leading to a principled attack methodology. By Lemma 5, the attack we derive in the spatial domain induces an attack in the frame domain. We note that classical optimization results are deployed as analytical tools, not as the contribution. Our novelty resides in the unified framework that connects Gabor frame geometry with adversarial optimization through frame-induced perturbation sets. Next, we detail our complete framework below.



(a) Example **CIFAR100** adversarial images, with **SSIM** quality and **ASR** attack efficiency values. Bold letters indicate the method with the best performance in terms of both evaluation metrics.



(b) Spatial (top) and **frequency (bottom)** representation of the **proposed attack** and the four **baselines**, corresponding to the adversarial examples of Fig. 2a. Our attack adapts to the SF patterns arising in the images, taking advantage of spectral biases that persist across different ambient dimensions (cf. Fig. 1b), thereby outperforming the baselines, which lack a similar double localization.

Fig. 2: Evaluation of the baselines ($\varepsilon = 16/255$) and our attack ($\varepsilon_{\mathrm{DGF}} = \eta\varepsilon$), on an exemplary RobustBench model [13].

Let a classification loss $\mathcal{L}(\cdot,\cdot)$, and choose a DGF-based analysis operator. Then, we transform any clean image $x \in \mathbb{R}^n$ in the DGF domain: $z = \Psi x \in \mathbb{C}^N$. For $\varepsilon > 0$, we propose to solve

$$\max_{w} \mathcal{L}(\Psi^{+}(z+w), y), \quad \text{such that} \quad w = \Psi\delta,\ \|w\|_D \leq \varepsilon, \tag{6}$$

where $w \in \mathbb{C}^N$ is an admissible frame attack of level ε (cf. Definition 4) w.r.t. $\|\cdot\|_D$ in $\mathbb{C}^N$, which we define to be a weighted Gabor-based norm:

$$\|w\|_D = \sqrt{w^* D w}. \tag{7}$$

$D \in \mathbb{R}^{N\times N}$ is a diagonal weighting matrix that constitutes a function of the average coherence (5):

$$d_i = \left(\tau + (N-1)^{-1}\left|\sum_{j=1}^{N}\langle\psi_i,\psi_j\rangle\right|\right)^{-1}, \qquad i = 1,\dots,N,\ \tau > 0. \tag{8}$$

Now, we rewrite (7) as $\|w\|_D = \|\Psi\delta\|_D = \sqrt{\delta^T M_D \delta}$, where

$$M_D = \Psi^* D \Psi \in \mathbb{C}^{n\times n}. \tag{9}$$

Why is (8) important? D should reflect frame geometry in a similar manner to the Grammian operator (cf. Definition 1). By defining D as a regularized inverse proxy of the average coherence, we perform downweighting – in a positive-definite and stable fashion due to the sufficiently small $\tau > 0$ – to prevent redundant frame elements from dominating the attack level, thereby enforcing a perturbation norm that better reflects intrinsic signal-space complexity.

By Remark 2, Ψ allows for a lossless expansion, so we get $\delta = \Psi^{+}w$. Since w is an admissible frame attack, by Lemma 5, δ is a spatial attack, and we can easily move our optimization problem to the spatial domain; thus, (6) is equivalently written as

$$\max_{\delta} \mathcal{L}(x+\delta, y) \quad \text{such that} \quad \delta^T M_D \delta \leq \varepsilon^2. \tag{10}$$

For M_D as in (9), we introduce the perturbation set

$$\mathcal{C}_{M_D} := \{\delta \in \mathbb{R}^n \mid \delta^T M_D \delta \leq \varepsilon^2\}. \tag{11}$$

To solve (10), we rely on projected gradient ascent – with direction dictated by $\mathcal{C}_{M_D}$ – which naturally follows as an appropriate first-order solver, similarly to how PGD is the standard tool[1] for approximately solving (1). Thus, we obtain (Proof in Appendix B.2)

$$\delta^{k+1} = \Pi_{\mathcal{C}_{M_D}}\left(\delta^k + \gamma \frac{(\Re(M_D))^{-1}\tilde{g}^k}{\sqrt{(\tilde{g}^k)^T(\Re(M_D))^{-1}\tilde{g}^k}}\right), \tag{12}$$

where $\tilde{g}^k = \nabla_\delta \mathcal{L}(x+\delta^k)$, and $\Pi_{\mathcal{C}_{M_D}}(\cdot)$ is the projection operator on $\mathcal{C}_{M_D}$ w.r.t. the ℓ_2-norm, given below.

Proposition 6 (Proof in Appendix B.3)**.** *For $\Psi \in \mathbb{F}^{N\times n}$ being the analysis operator associated to a DGF and $M_D = \Psi^* D \Psi$, the projection operator on $\mathcal{C}_{M_D}$ defined in (11), w.r.t. the ℓ_2-norm, is given for any $\delta \in \mathbb{R}^n$ as*

$$\Pi_{\mathcal{C}_{M_D}}(\delta) = \arg\min_{\delta' \in \mathcal{C}_{M_D}} \|\delta' - \delta\|_2 = \begin{cases} \delta, & \delta^T M_D \delta \leq \varepsilon^2 \\ (I + 2\lambda \Re(M_D))^{-1}\delta, & \delta^T M_D \delta > \varepsilon^2 \end{cases}, \tag{13}$$

for a uniquely determined $\lambda = \lambda(\varepsilon) > 0$.

Due to (12) and (13), we can also write

$$\delta^{k+1} = \begin{cases} \widetilde{\delta^k} & \text{if } \widetilde{\delta^k} \in \mathcal{C}_{M_D} \\ (I + 2\lambda\Re(M_D))^{-1} \cdot \widetilde{\delta^k} & \text{if } \widetilde{\delta^k} \notin \mathcal{C}_{M_D}, \end{cases} \qquad k = 0, \ldots, K-1, \tag{14}$$

where $\widetilde{\delta^k} = \delta^k + \gamma(\Re(M_D))^{-1}\tilde{g}^k \cdot \left((\tilde{g}^k)^T(\Re(M_D))^{-1}\tilde{g}^k\right)^{-1/2}$. The attack δ^K is obtained after $K \in \mathbb{N}$ iterations of the projected gradient ascent scheme (12), where each projection onto $\mathcal{C}_{M_D}$ is computed exactly according to Proposition 6. Since the objective in (10) is generally non-convex, δ^K constitutes an approximate solution to (10). The adversarial image is then defined as $x_{\text{adv}} = x + \delta_{\text{opt}}$, appropriately clipped to fall within $[0, 1]$. Overall, our proposed attack is given in Algorithm 1.

Why is Proposition 6 important? Our main result gives a convincing partial answer to (†) via the induced geometry of $\mathcal{C}_{M_D}$. By optimizing over $\mathcal{C}_{M_D}$, we enable attacks to align with structured SF components and exploit model sensitivities, while remaining distributed across correlated directions through frame overcompleteness, thus avoiding overfitting to pixel-space features, and improving both effectiveness on the source model and transferability to unseen models; we empirically validate this in Section 5.2. The inherent overcompleteness of frames also suggests our framework could generalize to other structured transform families, which we leave as future work (cf. Section 6).

5 Experiments

We complement our principled methodology for transform-based attacks with an empirical evaluation, designed to assess its effectiveness and transferability in practice. Section 5.1 features representative experimental settings across diverse pretrained and robust models, and standardized datasets; in Section 5.2, we discuss corresponding results (and refer to Appendix C for more experimental details). While not intended as an exhaustive benchmark, these experiments support our theoretical motivation and lay the groundwork for broader empirical investigation (cf. Section 6).

5.1 Settings

Datasets, models & baselines. We evaluate our attack on 224×224 (5000 validation) and 32×32 (10000 test) coloured images from ImageNet [20] and CIFAR100 [35], respectively, on SotA standard pretrained [33, 63], and on adversarially robust models taken from the standardized benchmark of RobustBench [16]. Both datasets are typical benchmarks for studying robustness across different data resolutions and complexities, and model families, making them suitable candidates for assessing the behavior of our proposed methodology. We select RobustBench models having undergone both

[1] In this sense, the optimization algorithm is a consequence of the proposed formulation, not our primary methodological contribution.

Attack	ImageNet		CIFAR100	
	Pretrained	RobustBench	Pretrained	RobustBench
Proposed (spatial-frequency-based)	$\mathbf{74.5 \pm 0.019}$	$\mathbf{63.4 \pm 0.025}$	$\mathbf{96.5 \pm 0.023}$	$\mathbf{85.1 \pm 0.023}$
AutoAttack (spatial-based)	33.6 ± 0.022	30.9 ± 0.025	90.1 ± 0.029	44.4 ± 0.021
F-PGD (frequency-based)	32.8 ± 0.021	32.1 ± 0.026	74.2 ± 0.023	46.0 ± 0.023
SSA (frequency-based)	39.9 ± 0.022	32.7 ± 0.022	59.4 ± 0.024	45.2 ± 0.026
AdvDrop (frequency-based)	30.1 ± 0.025	32.2 ± 0.021	46.9 ± 0.022	42.3 ± 0.020

Table 1: Average **ASR** (in %), $\pm$ average standard deviation (computed over 10 runs), across source-target model pairs, computed from the off-diagonal entries of the heatmaps of Fig. 3. **Bold letters indicate better cross-model transferability**. Results support our methodological contribution: attacks derived from a principled optimization framework, based on structured overcomplete transforms that align more naturally with data geometry, can exploit shared model vulnerabilities more effectively than purely spatial-/frequency-based methods.

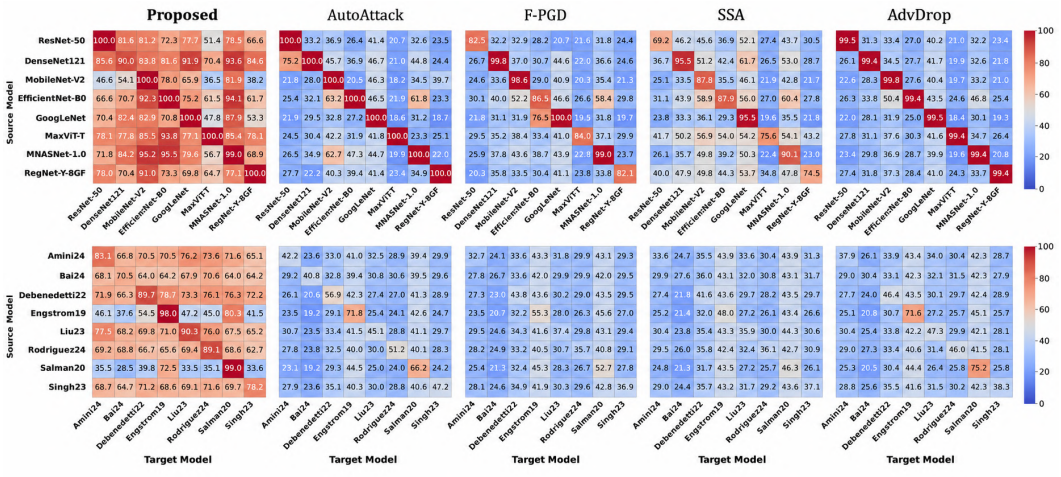
adversarial training and more elaborate defenses, i.e., sparsity-promoting [2] and stochastic smoothing strategies [5], and regularized adversarial training [4], as a means of examining the performance of our attack in more advanced settings. We compare our attack to SotA baselines, being indicative of attack classes: the spatial-based AutoAttack [15], and the frequency-based Fourier PGD (F-PGD) [65] and DCT-inspired SSA [41] and AdvDrop [21], all constrained in the ℓ_∞-norm ball. Since our attack results as a solution to a projection on $\mathcal{C}_{M_D}$, we aim to see how the structure induced by this closed-form solution relates to less structured attacks. For both baselines, we set the best hyperparameters proposed by the associated papers.

Evaluation metrics. We examine varying attack levels ε, and measure attack effectiveness via the common attack success rate (ASR), i.e., the rate of adversarial examples that successfully fooled the model; the higher the ASR, the more vulnerable a model is. All experiments are repeated at least 10 times with different random seeds, and results are reported as averages over these runs. What is more, we compute the standard deviation across runs, which we found to be consistently small; due to this low variability, we report their mean values. We also examine visual quality via the standard structural similarity index measure (SSIM) [59], ranging in $[0, 1]$; the higher the SSIM, the better the visual quality is.

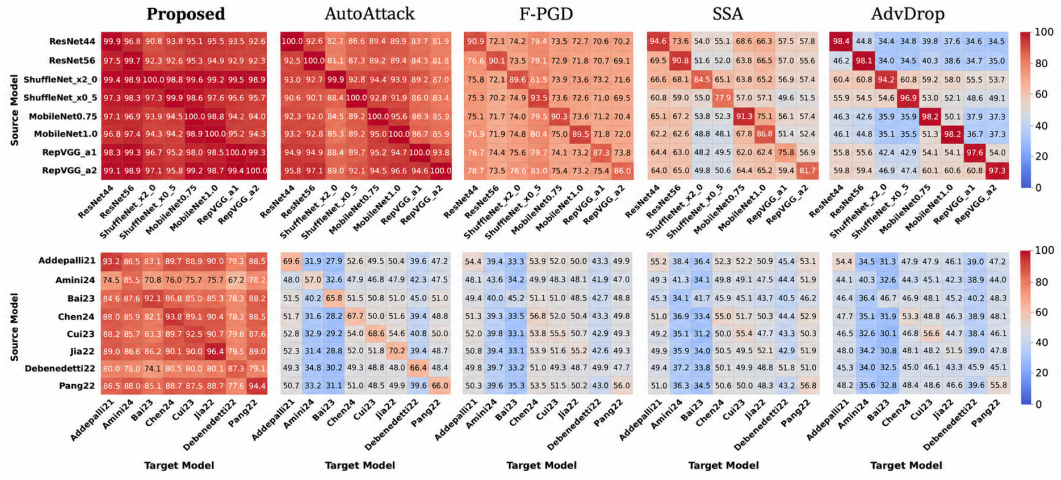
Attack parameters & volume-matching criterion. The perturbation set $\mathcal{C}_{M_D}$ of (11) is an ellipsoid, not a ball. To enable a fair baseline comparison, a common practice hinges on comparing volumes of perturbation sets [3, 38, 48], i.e., we require the ball volume (controlled by ε) to be equal to the ellipsoid volume (controlled by an attack level we denote as $\varepsilon_{\mathrm{DGF}}$) and solve for $\varepsilon_{\mathrm{DGF}}$. Then, by following standard arguments [7, 11] for the volumes of ℓ_p balls and ellipsoids, it can be proven that $\varepsilon_{\mathrm{DGF}} = \eta\varepsilon$, where $\eta = \sqrt{2n/(\pi e)}(\pi n \det(M_D))^{1/2n}$. That expression is exact up to the point where Stirling's formula [51] is invoked. Yet, for the ambient dimensions considered in our experiments, Stirling's approximation is extremely accurate, making the resulting discrepancy practically negligible. Thus, for simplicity of exposition, we write the final expression as an equality. We also emphasize here that the volume-matching criterion is intended as a natural geometric normalization – not equivalence – when comparing perturbation sets of different shapes, since the volume quantifies the overall size of the admissible perturbation region, independently of its specific shape. Furthermore, $\mathcal{C}_{M_D}$ redistributes the fixed attack budget to allocate more perturbation energy to informative directions and less to uninformative ones. Consequently, the volumetric comparison evaluates precisely the question addressed by our work: whether a transform-aware perturbation geometry can utilize the same overall attack budget more effectively than the geometry.

5.2 Results & discussion

The effect of $\mathcal{C}_{M_D}$ (and Gabor frames) on effectiveness and transferability. We evaluate the effectiveness of the proposed attack on pretrained and RobustBench models of ImageNet and CIFAR100, with fixed $\varepsilon_{\mathrm{DGF}} = \eta 4/255$ and $\varepsilon_{\mathrm{DGF}} = \eta 8/255$, respectively, and report the results in the diagonal of the left heatmaps of Fig. 3a and Fig. 3b, respectively, with varying architectures, e.g., from CNN to Transformers. As illustrated in the first column of Fig. 3, the proposed Gabor-based attack achieves nearly ideal ASR on the pretrained models, and consistently high ASR on the RobustBench models. This strong effectiveness can be owed to the structure of $\mathcal{C}_{M_D}$ (11), which departs fundamentally from ℓ_p balls. By encoding the geometry of the underlying DGF, $\mathcal{C}_{M_D}$ enables the attack to distribute energy across SF directions, being highly relevant to the models' internal representations. Overall results confirm that designing an optimization problem, endowed with a transform-aware perturba-



(a) ImageNet, with $\varepsilon = 4/255$ for the baselines and $\varepsilon_{\text{DGF}} = \eta 4/255$ for our attack, on **pretrained (top)** and **RobustBench (bottom)** models. **Left to right: proposed attack, AutoAttack, F-PGD, SSA, AdvDrop**.



(b) CIFAR100, with $\varepsilon = 8/255$ for the baselines and $\varepsilon_{\text{DGF}} = \eta 8/255$ for our attack, on **pretrained (top)** and **RobustBench (bottom)** models. **Left to right: proposed attack, AutoAttack, F-PGD, SSA, AdvDrop**.

Fig. 3: ASR (in %) for both datasets, for all attacks, evaluated on different models. **Color intensity indicates higher transferability**. Each row represents a different source model, based on which an attack is created. We observe that our proposed attack enjoys a high transferability across a variety of models, and outperforms the baselines, thereby highlighting the importance of deriving attacks from a principled methodology.

tion constraint set, is a key factor for a frequency-based attack to be highly effective on different architectures, extending naturally from small- to large-scale datasets.

The transferability of our attack is also reflected in the off-diagonal entries of the left heatmaps of Fig. 3a (ImageNet) and 3b (CIFAR100). For the sake of completeness, in Table 1, we also summarize the average off-diagonal ASR of Fig. 3 – a common practice when examining cross-model effectiveness [58, 60, 61]. An adversarial attack is transferable when generated on a source model, yet successfully deceives unseen target models. We observe that our attack enjoys a high transferability rate, across most of the source–target model combinations, with a mild reduction for the RobustBench models. For both datasets, color intensity variations indicate that our attack generalizes beyond the adversarial defense used for each source model, even under more advanced schemes than standard adversarial training. We attribute this effect to the representation power of DGFs (cf. Section 3), to handle SF patterns that naturally arise in images, so the attacks greatly exploit shared spectral biases of different models; this enhances transferability. Therefore, with a DGF-based attack, we hit models at their core: their sensitivity to SF-like features, being essential for effective feature extraction, becomes their weakness. All in all, Table 1, and Fig. 3a and 3b, demonstrate strong cross-model transferability, under both pretrained and robust regimes.

Baseline comparisons (and a note on visual fidelity). We compare the effectiveness and transferability of our attack against AutoAttack, F-PGD, SSA and AdvDrop, on pretrained and RobustBench models, and report the results in the heatmaps of Fig. 3a (ImageNet) and Fig. 3b (CIFAR100), with $\varepsilon = 4/255$ and $\varepsilon = 8/255$, respectively, for the baselines, and $\varepsilon_{\mathrm{DGF}} = \eta\varepsilon$ for our attack. For clarity, we also present the averaged off-diagonal ASR of each heatmap in Table 1. We observe that the proposed attack achieves an almost identical self-model ASR to AutoAttack, which solves (1), thus confirming our motivation for designing the optimization problem (10) to derive SF attacks. In contrast, F-PGD and SSA follow the formulation of (3) - (4), without resulting from an optimization problem. Their effectiveness is similar to that of AdvDrop, although the latter is inspired by JPEG quantization to cast an optimization problem for deriving DCT-based attacks. Our attack consistently outperforms the baselines by means of source-target ASR, across pretrained and robust models alike, implying higher transferability, even against strongly defended RobustBench models. Overall results demonstrate that structured, overcomplete SF transforms, by aligning more naturally with data geometry, enable attacks to capture shared spectral sensitivities of architectures, more effectively than purely spatial-/frequency-based attacks. These findings also pinpoint to a concrete path for creating more robust models: incorporate transform-domain priors, enjoying additional structure, in adversarial training. For instance, our attack is a solution to (10), so the latter can naturally replace the standard ℓ_p-constrained inner maximization in adversarial training, to yield $\min_\theta \max_{\delta \in \mathcal{C}_{M_D}} \mathcal{L}(x+\delta, y; \theta)$, for parameters θ; we leave this strategy for future work (cf. Section 6).

We also examine visual quality, via SSIM, of adversarial samples of RobustBench models, illustrated in Fig. 1a (ImageNet) and 2a (CIFAR100), with $\varepsilon = 8/255$ and $\varepsilon = 16/255$, respectively, for the baselines, and $\varepsilon_{\mathrm{DGF}} = \eta\varepsilon$ for our attack; we visualize all three attacks in the spatial and frequency domain, in Fig. 1b and 2b, respectively. While all methods retain a reasonable visual quality, our attack outperforms the baselines, in terms of ASR and SSIM. We observe that AutoAttack leads to pixel-aligned variations, reflecting the ℓ_∞-norm geometry, while the frequency-based baselines generally produce smoother spatial distortions, with energy concentrated in specific frequencies. On the contrary, our method distributes energy along structured SF orientations, confirming our motivation for designing a perturbation set to reflect the DGF geometry and overcompleteness. More broadly, Fig. 1 and 2 showcase that attacks resulting from a maximization problem, with a constraint set relying on Gabor frames being closely related to human perception (cf. Section 3), can be highly effective, while preserving visual fidelity. We pinpoint here that SSIM scores and visual comparisons serve as expository diagnostics of the induced geometry, instead of evidence that imperceptibility is necessary or a principal advantage of the proposed attack.

6 Conclusion & future work

In this paper, we introduced a principled methodology for explaining the success of frequency-based adversarial attacks. Our analysis relied on introducing an optimization problem over a newly-defined perturbation set, incorporating structured, non-orthogonal transforms, allowing for flexible data representation. We proved that the solution to the optimization problem is acquired through a projection on the aforesaid set, resulting in an adversarial attack of increased effectiveness and transferability, as supported by relevant experiments. Particularly, our proposed method exploited model vulnerabilities beyond traditional high-/low-frequency assumptions, highlighting the pivotal role of transform geometry in adversarial robustness.

Due to its grounded justification and preliminary empirical success, our work opens promising future directions. For instance, it would be intriguing to develop a larger empirical evaluation, with more datasets, e.g., CIFAR10, TinyImageNet and medical/satellite images, attacking settings, and different reweighting strategies for the DGF. Additionally, there is currently no standardized benchmark, analogous to RobustBench, comprising models robust against frequency-based attacks. This would motivate us to couple our method with adversarial training, to establish a complete benchmark for transform-aware defenses that align with perceptual and representational structure. Finally, it would be fruitful to extend our framework to other families of structured transforms. For instance, while alternative frame families correspond to different geometric frameworks, it would be interesting to examine those in the context of transform-aware attacks, e.g., w.r.t. shearlets [36] or curvelets [56], as well as build upon learnable adaptive transforms [45], to examine how these transforms could relate to shared spectral biases of diverse architectures.

Acknowledgments and Disclosure of Funding

This research was supported in part by the French National Research Agency (ANR) in the framework of the PEPR IA FOUNDRY project (ANR-23-PEIA-0003). Moreover, V. Kouni would like to thank P. Mertikopoulos for insightful comments and fruitful discussions.

References

[1] Addepalli, S., Jain, S., Sriramanan, G., and Venkatesh Babu, R. Scaling adversarial training to large perturbation bounds. In *Eur. Conf. Comput. Vis.*, pp. 301–316. Springer, 2022.

[2] Amini, S., Teymoorianfard, M., Ma, S., and Houmansadr, A. Meansparse: Post-training robustness enhancement through mean-centered feature sparsification. *arXiv preprint arXiv:2406.05927*, 2024.

[3] Araujo, A., Meunier, L., Pinot, R., and Negrevergne, B. Advocating for multiple defense strategies against adversarial examples. In *Joint Eur. Conf. Mach. Learn. & Knowl. Discovery Databases*, pp. 165–177, 2020.

[4] Bai, Y., Anderson, B. G., Kim, A., and Sojoudi, S. Improving the accuracy-robustness trade-off of classifiers via adaptive smoothing. *SIAM J. Math. Data Sci.*, 6(3):788–814, 2024.

[5] Bai, Y., Zhou, M., Patel, V. M., and Sojoudi, S. Mixednuts: Training-free accuracy-robustness balance via nonlinearly mixed classifiers. *arXiv preprint arXiv:2402.02263*, 2024.

[6] Bajwa, W. U., Calderbank, R., and Jafarpour, S. Why Gabor frames? two fundamental measures of coherence and their role in model selection. *J. Commun. & Networks*, 12(4):289–307, 2010.

[7] Ball, K. et al. An elementary introduction to modern convex geometry. *Flavors of Geometry*, 31(1-58):26, 1997.

[8] Ballé, J. Subjective evaluation of texture similarity metrics for compression applications. In *Picture Coding Symp.*, pp. 241–244, 2012.

[9] Ben-Tal, A. and Nemirovski, A. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

[10] Bertsekas, D. P. Nonlinear programming. *J. Oper. Res. Soc.*, 48(3):334–334, 1997.

[11] Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

[12] Casazza, P. G. and Kutyniok, G. *Finite Frames: Theory and Applications*. Springer Science & Business Media, 2012.

[13] Chen, E. C. and Lee, C. R. Data filtering for efficient adversarial training. *Patt. Recognit.*, 151: 110394, 2024.

[14] Christensen, O. *An Introduction to Frames and Riesz Bases*, volume 7. Springer, 2003.

[15] Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Int. Conf. Mach., Learn.*, pp. 2206–2216. PMLR, 2020.

[16] Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. RobustBench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

[17] Cui, J., Tian, Z., Zhong, Z., Qi, X., Yu, B., and Zhang, H. Decoupled kullback-leibler divergence loss. *Adv. Neural Inf. Process. Syst.*, 37:74461–74486, 2024.

[18] Daugman, J. G. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Amer. A*, 2(7):1160–1169, 1985.

[19] Debenedetti, E., Sehwag, V., and Mittal, P. A light recipe to train robust vision transformers. In *Conf. Secure & Trustworthy Mach. Learn.*, pp. 225–253. IEEE, 2023.

[20] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *Proc. Conf. Comput. Vision & Patt. Recognit.*, pp. 248–255, 2009.

[21] Duan, R., Chen, Y., Niu, D., Yang, Y., Qin, A. K., and He, Y. Advdrop: Adversarial attack to DNNs by dropping information. In *Proc. Int. Conf. Comput. Vis.*, pp. 7506–7515, 2021.

[22] Engstrom, L., Ilyas, A., Salman, H., Santurkar, S., and Tsipras, D. Robustness (python library), 2019. URL https://github.com/MadryLab/robustness.

[23] Feichtinger, H. G. and Strohmer, T. Gabor analysis and algorithms. 1998.

[24] Filus, K., Domańska, J., and Klamka, J. Evaluating convolutional neural networks via measuring similarity of their first-layer filters and Gabor filters. In *Int. Conf. Model., Anal. & Simul. Comput. & Telecommun. Syst.*, pp. 1–8, 2024.

[25] Folland, G. B. and Sitaram, A. The uncertainty principle: a mathematical survey. *J. Fourier Anal. & Appl.*, 3(3):207–238, 1997.

[26] Fridovich-Keil, S., Gontijo Lopes, R., and Roelofs, R. Spectral bias in practice: The role of function frequency in generalization. *Adv. Neural Inf. Process. Syst.*, 35:7368–7382, 2022.

[27] Fu, Y., Liu, Z., and Lyu, J. Transferable adversarial attacks for remote sensing object recognition via spatial-frequency co-transformation. *IEEE Trans. Geosci. & Remote Sens.*, 2024.

[28] Gröchenig, K. *Foundations of time-frequency analysis*, volume 359. Springer, 2001.

[29] Harris, F. J. On the use of windows for harmonic analysis with the discrete fourier transform. *Proc. IEEE*, 66(1):51–83, 1978.

[30] Jia, S. et al. Exploring frequency adversarial attacks for face forgery detection. In *Proc. Conf. Comput. Vision & Patt. Recognit.*, pp. 4103–4112, 2022.

[31] Jia, X., Zhang, Y., Wu, B., Ma, K., Wang, J., and Cao, X. Las-at: adversarial training with learnable attack strategy. In *Conf. Comput. Vis. & Patt. Recognit.*, pp. 13388–13398. IEEE, 2022.

[32] Jones, J. P. and Palmer, L. A. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiology*, 58(6):1233–1258, 1987.

[33] Ketkar, N. Introduction to pytorch. In *Deep learning with python*, pp. 195–208. Springer, 2017.

[34] Kouni, V., Rauhut, H., and Theoharis, T. Star DGT: a robust Gabor transform for speech denoising. *Sampl. Theory, Signal Process., & Data Anal.*, 21(1):14, 2023.

[35] Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. 2009.

[36] Kutyniok, G. and Labate, D. Introduction to shearlets. In *Shearlets: Multiscale analysis for multivariate data*, pp. 1–38. Springer, 2012.

[37] Li, X.-C., Zhang, X.-Y., Yin, F., and Liu, C.-L. F-mixup: Attack CNNs from Fourier perspective. In *25th Int. Conf. Patt. Recognit. (ICPR)*, pp. 541–548, 2021.

[38] Liu, C., Tomioka, R., and Cevher, V. On certifying non-uniform bounds against adversarial attacks. In *Int. Conf. Mach. Learn.*, pp. 4072–4081. PMLR, 2019.

[39] Liu, C., Dong, Y., Xiang, W., Yang, X., Su, H., Zhu, J., Chen, Y., He, Y., Xue, H., and Zheng, S. A comprehensive study on robustness of image classification models: Benchmarking and rethinking. *Int. J. Comput. Vis.*, 133(2):567–589, 2025.

[40] Liu, S., Cheng, S., and Alkhalifah, T. A. Gabor-based learnable sparse representation for self-supervised denoising. *IEEE Trans. Geosci. & Remote Sens.*, 62:1–16, 2024.

[41] Long, Y., Zhang, Q., Zeng, B., Gao, L., Liu, X., Zhang, J., and Song, J. Frequency domain model augmentation for adversarial attack. In *Eur. Conf. Comput. Vis.*, pp. 549–566, 2022.

[42] Luan, S., Chen, C., Zhang, B., Han, J., and Liu, J. Gabor convolutional networks. *IEEE Trans. Image Process.*, 27(9):4357–4366, 2018.

[43] Luo, C. et al. Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proc. Conf. Comput. Vision & Patt. Recognit.*, pp. 15315–15324, 2022.

[44] Madry, A. et al. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[45] Mairal, J., Ponce, J., Sapiro, G., Zisserman, A., and Bach, F. Supervised dictionary learning. *Adv. Neural Inf. Process. Syst.*, 21, 2008.

[46] Pang, T., Lin, M., Yang, X., Zhu, J., and Yan, S. Robustness and accuracy could be reconcilable by (proper) definition. In *Int. Conf. Mach. Learn.*, pp. 17258–17277. PMLR, 2022.

[47] Paukner, S. *Foundations of Gabor Analysis for Image Processing*. Master's thesis, 2007.

[48] Pfrommer, S., Anderson, B. G., and Sojoudi, S. Projected randomized smoothing for certified adversarial robustness. *Trans. Mach. Learn. Rese.*, 2023. ISSN 2835-8856.

[49] Qian, Y., Chen, K., Wang, B., Gu, Z., Ji, S., Wang, W., and Zhang, Y. Enhancing transferability of adversarial examples through mixed-frequency inputs. *IEEE Trans. Inf. Forensics & Secur.*, 2024.

[50] Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. On the spectral bias of neural networks. In *Int. Conf. Mach, Learn.*, pp. 5301–5310. PMLR, 2019.

[51] Robbins, H. et al. A remark on stirling's formula. *Amer. Math. Monthly*, 62(1):26–29, 1955.

[52] Rodríguez-Muñoz, A., Wang, T., and Torralba, A. Characterizing model robustness via natural input gradients. In *Eur. Conf. Comput. Vis.*, pp. 161–178. Springer, 2024.

[53] Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., and Madry, A. Do adversarially robust imagenet models transfer better? *Adv. Neural Inf. Process. Syst.*, 33:3533–3545, 2020.

[54] Scherzer, O. *Handbook of Mathematical Methods in Imaging*. Springer Science & Business Media, 2010.

[55] Singh, N. D., Croce, F., and Hein, M. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. *Adv. Neural Inf. Process. Syst.*, 36:13931–13955, 2023.

[56] Starck, J.-L., Candès, E. J., and Donoho, D. L. The curvelet transform for image denoising. *Trans. Image Process.*, 11(6):670–684, 2002.

[57] Wang, H., Wu, X., Huang, Z., and Xing, E. P. High-frequency component helps explain the generalization of convolutional neural networks. In *Proc. Conf. Comput. Vision & Patt. Recognit.*, pp. 8684–8694, 2020.

[58] Wang, Y., Wu, Y., Wu, S., Liu, X., Zhou, W., Zhu, L., and Zhang, C. Boosting the transferability of adversarial attacks with frequency-aware perturbation. *IEEE Trans. Inf. Forensics & Secur.*, 19:6293–6304, 2024.

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

[60] Wei, Z., Chen, J., Goldblum, M., Wu, Z., Goldstein, T., and Jiang, Y.-G. Towards transferable adversarial attacks on vision transformers. In *Proc. Conf. Artif. Intell.*, volume 36, pp. 2668–2676, 2022.

[61] Yang, X., Lin, J., Zhang, H., Yang, X., and Zhao, P. Improving the transferability of adversarial examples via direction tuning. *Inf. Sci.*, 647:119491, 2023.

[62] Yao, J., Luo, H., Qi, J., and Zhang, X.-L. Interpretable spectrum transformation attacks to speaker recognition systems. *IEEE/ACM Trans. Audio, Speech, & Language Process.*, 32: 1531–1545, 2024.

[63] Yaofo, C. PyTorch CIFAR models. https://github.com/chenyaofo/pytorch-cifar-models, 2025.

[64] Yi, H., Ming, Y., Liu, D., and Kong, W. Time-frequency jointed imperceptible adversarial attack to brainprint recognition with deep learning models. In *IEEE Int. Conf. Multimedia & Expo (ICME)*, pp. 1–6. IEEE, 2024.

[65] Zhang, L., Luo, Y., Shen, H., Wang, T., and Larson, K. A Fourier perspective of feature extraction and adversarial robustness. In *Int. Joint Conf. Artif. Intell. Organiz.*, pp. 1715–1723, 2024.

[66] Zheng, Y., Wan, C., Guo, Z., Kuang, H., and Lu, X. Boosting adversarial transferability via high-frequency augmentation and hierarchical-gradient fusion. *arXiv preprint arXiv:2505.21181*, 2025.

[67] Zhou, Z., Li, B., Song, Y., Yu, Z., Hu, S., Wan, W., Zhang, L. Y., Yao, D., and Jin, H. Numbod: A spatial-frequency fusion attack against object detectors. In *Proc. Conf. Artif. Intell.*, volume 39, pp. 1201–1209, 2025.

[68] Zhu, H., Zhu, Y., Zheng, H., Ren, Y., and Jiang, W. LIGAA: Generative adversarial attack method based on low-frequency information. *Comput. & Secur.*, 125:103057, 2023.

Appendices

A Details of Section 3

We present below the mathematical definition of a frame.

Definition 7 ([12]). *A set $(\psi_i)_{i=1}^N$ of N vectors in $\mathbb{F}^n$, $N \geq n$, is a frame for $\mathbb{F}^n$ if and only if it is a spanning set for $\mathbb{F}^n$. Equivalently, if the inequalities*

$$A\|x\|_2^2 \leq \sum_{i=1}^{N} |\langle x, \psi_i \rangle|^2 \leq B\|x\|_2^2 \tag{15}$$

hold for all $x \in \mathbb{F}^n$, for some $0 < A \leq B < \infty$ (frame bounds), then $(\psi_i)_{i \in N}$ is a frame for $\mathbb{F}^n$.

While frames are usually studied in the 1D case, e.g., for $x \in \mathbb{R}^n$, they can be naturally extended to 2D data, by relying on tensor algebra of frames, since the tensor product of two frames is a frame itself [47]. As a result, for $x \in \mathbb{R}^{n \times n}$ and a DGF with associated analysis operator $\Psi \in \mathbb{C}^{N \times n}$ as in Definition 3, we have

$$(\Psi \otimes \Psi)\text{vec}(x) = \Psi x \Psi^T = w \in \mathbb{C}^{N \times N}, \tag{16}$$

with $\Psi \otimes \Psi$ being the analysis operator associated to a DGF in $\mathbb{C}^{N \times n} \otimes \mathbb{C}^{N \times n} \cong \mathbb{C}^{N^2 \times n^2}$. In this case, we call Ψ a 2D separable DGF analysis operator. Similarly to Remark 2, 2D frames enable lossless expansions [47]:

$$x = \Psi^+ w (\Psi^+)^T, \tag{17}$$

for $\Psi^+ \in \mathbb{C}^{n \times N}$ being the pseudo-inverse of $\Psi \in \mathbb{C}^{N \times n}$.

Based on the afore-described framework, we can readily extend the methodology we propose in Section 4, to account for images $x \in \mathbb{R}^{n \times n}$, with adversarial attacks $\delta \in \mathbb{R}^{n \times n}$, so $w = \Psi \delta \Psi^T \in \mathbb{C}^{N \times n}$ is an admissible frame attack, due to Lemma 5; we give more details in Appendix B.4.

Table 2: List of referenced evaluated models taken from RobustBench, on all datasets, as these are depicted on the heatmap tables of Figs. 3, 5, 6.

ImageNet	Amini24 [2]	Bai24 [4]	Debenedetti22 [19]	Engstrom19 [22]	Liu23 [39]	Rodriguez24 [52]	Salman20 [53]	Singh23 [55]
CIFAR100	Addepalli23 [1]	Amini24 [2]	Bai24 [5]	Chen24 [13]	Cui23 [17]	Debenedetti22 [19]	Jia22 [31]	Pang22 [46]

B Proofs of Section 4

B.1 Proof of Lemma 5

Proof. For the first bullet-point, let us assume that δ is a spatial attack of level ε in $\mathbb{F}^n$, so that $\|\delta\|_{\mathbb{F}^n} \leq \varepsilon$. Then, for $w = \Psi\delta$, we have

$$\|w\|_{\mathbb{F}^N} = \|\Psi\delta\|_{\mathbb{F}^N} \leq \|\Psi\|_{\mathrm{op}}\varepsilon.$$

By definition of Ψ, we conclude that w is an admissible frame attack of level $\|\Psi\|_{\mathrm{op}}\varepsilon$.
For the second bullet-point, let us suppose that $w = \Psi\delta$ is an admissible frame attack of level ε, so it holds $\|w\|_{\mathbb{F}^N} \leq \varepsilon$. Then, due to the lossless expansion offered by the frame, we get $\|\delta\|_{\mathbb{F}^n} = \|\Psi^+ w\|_{\mathbb{F}^n} \leq \|\Psi^+\|_{\mathrm{op}}\varepsilon$. The proof is complete. □

B.2 Proof of derivation of (12)

Proof. We aim to solve (10) by a variant of projected gradient ascent. To that end, we seek the steepest ascent direction under the geometry induced by the perturbation set (11). At iteration k, we set $\tilde{g}^k = \nabla_\delta \mathcal{L}(x + \delta^k, y)$ and solve

$$u^\star = \arg\max_{u\in\mathbb{R}^n} (\tilde{g}^k)^T u \quad \text{such that} \quad u^T M_D u \leq 1. \tag{18}$$

By definition of (8) and due to the injectivity of Ψ, M_D is Hermitian and positive definite, and thus invertible. Moreover, it holds $u^T M_D u = u^T \Re(M_D) u$, for any $u \in \mathbb{R}^n$.

We set the Langrangian of (18):

$$L(u;\lambda) = (\tilde{g}^k)^T u - \lambda(u^T \Re(M_D) u - 1), \qquad \lambda \geq 0. \tag{19}$$

By the KKT conditions, we have

$$\nabla_u L(u;\lambda) = 0 \implies \tilde{g}^k = 2\lambda\Re(M_D) \implies u = \frac{(\Re(M_D))^{-1}\tilde{g}^k}{2\lambda}. \tag{20}$$

Imposing the active constraint $u^T\Re(M_D)u = 1$ yields

$$\frac{(\tilde{g}^k)^T(\Re(M_D))^{-1}\tilde{g}^k}{(2\lambda)^2} \implies 2\lambda = \sqrt{(\tilde{g}^k)^T(\Re(M_D))^{-1}\tilde{g}^k}. \tag{21}$$

Hence, the steepest ascent direction is given by

$$u^\star = \frac{(\Re(M_D))^{-1}\tilde{g}^k}{\sqrt{(\tilde{g}^k)^T(\Re(M_D))^{-1}\tilde{g}^k}}. \tag{22}$$

Then, deploying projected gradient ascent in the direction of (22) yields (12), which complements the proof. □

B.3 Proof of Proposition 6

Proof. We are inspired by theoretical results relevant to projections on ellipsoids [10] and Mahalanobis-type distances [9], and use standard optimization theory [11] to set the Langrangian

$$\begin{aligned} L(\delta';\lambda) =& \frac{1}{2}\|\delta' - \delta\|_2^2 + \lambda((\delta')^T M_D \delta' - \varepsilon^2) \\ =& \frac{1}{2}\|\delta' - \delta\|_2^2 + \lambda((\delta')^T \Re(M_D) \delta' - \varepsilon^2), \end{aligned} \tag{23}$$

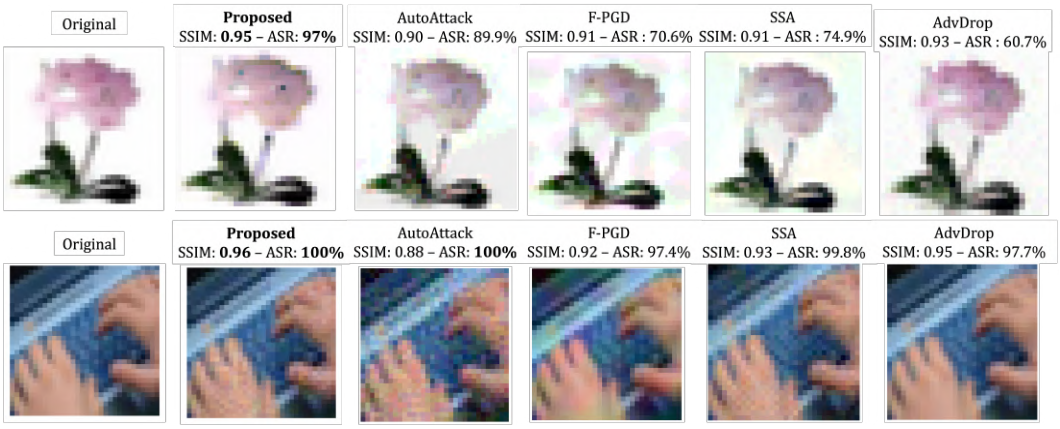


Fig. 4: Examples of images and their adversarial counterparts, with **SSIM** quality and **ASR** attack efficiency values, evaluated on a pretrained version of RegNet for **ImageNet (top)** and on a pretrained version of ResNet for **CIFAR100 (bottom)**, for $\varepsilon = 8/255$ and $\varepsilon = 16/255$, respectively, on the baselines, and $\varepsilon_{\text{DGF}} = \eta 8/255$ $\varepsilon_{\text{DGF}} = \eta 16/255$, respectively, on our attack. Bold letters indicate the method with the best performance in terms of both evaluation metrics.

for $\lambda \geq 0$. By KKT conditions, we get

$$\begin{aligned} \nabla_{\delta'} L = 0 &\implies (\delta' - \delta) + 2\lambda \Re(M_D)\delta' = 0 \\ &\implies \delta' = (I + 2\lambda \Re(M_D))^{-1}\delta, \end{aligned} \tag{24}$$

for $(\delta')^T \Re(M_D)\delta' \leq \varepsilon^2$, $\lambda((\delta')^T \Re(M_D)\delta' - \varepsilon^2) = 0$. Since $\Re(M_D)$ is invertible, so is $I + 2\lambda\Re(M_D)$, for all $\lambda \geq 0$.
If $\delta \in \mathcal{C}_{M_D}$, then it satisfies the constraint and $\lambda = 0$, yielding $\delta' = \delta$. If $\delta \notin \mathcal{C}_{M_D}$, then the constraint is active, $\lambda > 0$ and $\delta' = (I + 2\lambda\Re(M_D))^{-1}\delta$, with λ chosen such that $(\delta')^T \Re(M_D)\delta' = \varepsilon^2$. Overall, we have

$$\Pi_{\mathcal{C}_{M_D}}(\delta) = \begin{cases} \delta & \text{if } \delta^T M_D \delta \leq \varepsilon^2 \\ (I + 2\lambda\Re(M_D))^{-1}\delta & \text{if } \delta^T M_D \delta > \varepsilon^2, \end{cases} \tag{25}$$

which completes the first part of the proof.
Due to the positive definitiveness of M_D (and $\Re(M_D)$), the set $\mathcal{C}_{M_D}$ is non-degenerate and the projection is unique for all δ. To compute λ, we diagonalize $\Re(M_D)$, so that $\Re(M_D) = U\Lambda U^T$, where Λ is a diagonal matrix with entries $\mu_i > 0$, $i = 1, \ldots, n$, and set $\zeta = U^T\delta$. We define

$$\phi(\lambda) := \sum_{i=1}^{n} \frac{\mu_i \zeta_i^2}{(1 + 2\lambda\mu_i)^2} - \varepsilon^2; \tag{26}$$

it can be proven that ϕ is a strictly decreasing, continuous function. Moreover, $\phi(0) > 0$ and

$$\phi(\lambda) \overset{\lambda \to \infty}{\longrightarrow} -\varepsilon^2 < 0. \tag{27}$$

By the intermediate value theorem and strict monotonicity, there is a unique $\lambda^* > 0$ satisfying $\phi(\lambda^*) = 0$; we find λ^* by solving $\phi(\lambda) = 0$ using the bi-section method. □

B.4 Extension to a 2D setting

In Section 4, our proposed methodology considers 1D data; this is essential, in order to scale to 2D data like the ImageNet and CIFAR100 coloured images, which we process batch- and channel-wise. As such, we follow the 2D extension we describe in Appendix A. We capitalize all vectors to shift to matrix notation, use a 2D separable DGF analysis operator, and perform separable, i.e., row- and column-wise, weighting, using D (8). By turning to a Frobenius $\|\cdot\|_F$-norm instead of the ℓ_2-norm for (7), we extend the latter in the 2D case as $\|\Psi\Delta\Psi^T\|_D^2 = \|W\|_D^2 = \text{tr}(W^T DWD)$, or equivalently $\|W\|_D = \sqrt{\text{vec}(W)^T(D \otimes D)\text{vec}(W)} = \sqrt{\text{vec}(\Delta)^T(M_D \otimes M_D)\text{vec}(\Delta)}$, where $\text{tr}(\cdot)$ denotes the trace of a matrix, and M_D is as in (9). Then, by slightly abusing notation and

Table 3: Average **ASR** (in %), ± average standard deviation (computed over 10 runs), across source-target model pairs, computed from the off-diagonal entries of the heatmaps of Fig. 5 and Fig. 6. **Bold letters indicate better cross-model transferability**.

	ImageNet		CIFAR100	
Attack	Pretrained	RobustBench	Pretrained	RobustBench
Proposed (spatial-frequency-based)	**75.4 ± 0.023**	**63.5 ± 0.021**	**96.9 ± 0.025**	**83.8 ± 0.024**
AutoAttack (spatial-based)	44.0 ± 0.025	41.6 ± 0.027	89.6 ± 0.026	64.9 ± 0.025
F-PGD (frequency-based)	42.7 ± 0.027	39.7 ± 0.024	77.6 ± 0.026	58.5 ± 0.026
SSA (frequency-based)	56.6 ± 0.021	41.5 ± 0.024	92.12 ± 0.020	63.8 ± 0.020
AdvDrop (frequency-based)	31.5 ± 0.025	35.3 ± 0.026	48.8 ± 0.019	47.7 ± 0.024

denoting the perturbation set for both 1D and 2D cases as $\mathcal{C}_{M_D}$, we obtain for (11) and (12)

$$\mathcal{C}_{M_D} = \{\Delta \in \mathbb{R}^{n\times n} : \operatorname{tr}(\Delta^T M_D \Delta M_D) \leq \varepsilon^2\}, \tag{28}$$

$$\Delta^{k+1} = \Pi_{\mathcal{C}_{M_D}}\left(\Delta^k + \gamma \frac{(\Re(M_D))^{-1}\widetilde{G}^k(\Re(M_D))^{-1}}{\sqrt{(\widetilde{G}^k)^T(\Re(M_D))^{-1}\widetilde{G}^k(\Re(M_D))^{-1}}}\right), \tag{29}$$

respectively, where $\widetilde{G}^k = \nabla_\Delta \mathcal{L}(X + \Delta^k, Y)$, and in (29) we have used the identity $(\Re(M_D) \otimes \Re(M_D))^{-1} = (\Re(M_D))^{-1} \otimes (\Re(M_D))^{-1}$. Then, the corresponding projection of Proposition 6 is given by

$$\Pi_{\mathcal{C}_{M_D}}(\Delta) = \arg\min_{\Delta' \in \mathcal{C}_{M_D}} \|\Delta' - \Delta\|_F = \begin{cases} \Delta, & \Delta \in \mathcal{C}_{M_D} \\ (I + 2\lambda\Re(M_D))^{-1}\Delta(I + 2\lambda\Re(M_D))^{-1}, & \Delta \notin \mathcal{C}_{M_D} \end{cases}, \tag{30}$$

which easily follows by mutatis mutandis adapting the proof presented in Appendix B.3, based on the 2D separability of all associated matrices.

C Additional experimental details & results

Following standard principles from the signal-processing and Gabor analyses [23, 25, 28, 29], we outline below the choices – fixed throughout all experiments – of α, β, g; these quantities do not serve as attack-specific hyperparameters. To construct the DGF, we need to use an adequate so-called window vector g (cf. Definition 3), on which we perform the SF shifts. We select the Hann window vector $g \in \mathbb{R}^n$, defined as

$$g_j = 0.5\left(1 - \cos\left(\frac{2j\pi}{n-1}\right)\right), \tag{31}$$

$j = 0, \ldots, n-1$, commonly used in Gabor analysis and for image-processing applications, offering favorable SF localization and low spectral leakage [29]. The DGF parameters determine the SF resolution of the representation, and the frame overcompleteness, ensuring $\alpha\beta < n$. Choosing $\alpha = 1$ enables maximal spatial sampling, allowing the representation to capture fine localized structures, and $\beta = 16$ (for CIFAR100) and $\beta = 112$ (for ImageNet) provides a balanced frequency resolution, that preserves relevant spectral information. In fact, the choice of $\alpha < \beta$ corresponds to dense spatial sampling and comparatively coarser frequency sampling, a standard design choice for localized image representations [23]. Our configuration yields a stable, overcomplete representation, maintaining a constant overcompleteness $n/\alpha\beta = 2$ across datasets, thereby preserving the same degree of oversampling, while adapting naturally to the image resolution [28]. Importantly, our overall selections are guided by the uncertainty principle [25], according to which SF localization cannot be simultaneously optimized, so that α, β, g determine the particular localization trade-off of the representation. Finally, we set $\tau = 10^{-6}$ for the constant ensuring the positive-definitiveness of the reweighting matrix D defined in (8). We also calculate the scaling factor η separately for each dataset, as it is dimension-dependent, due to the appearance of M_D.

All experiments are implemented in PyTorch [33], with a batch size 8 for ImageNet and 32 for CIFAR100, and conducted on two high-performance computing clusters, using NVIDIA A100 (40GB VRAM) and H100 (80GB VRAM) GPUs. Under these configurations, the average execution time was consistently under 20 minutes. An exception was the transferability analysis of AutoAttack, which required exceedingly more computational time, which is natural, by definition of the AutoAttack's

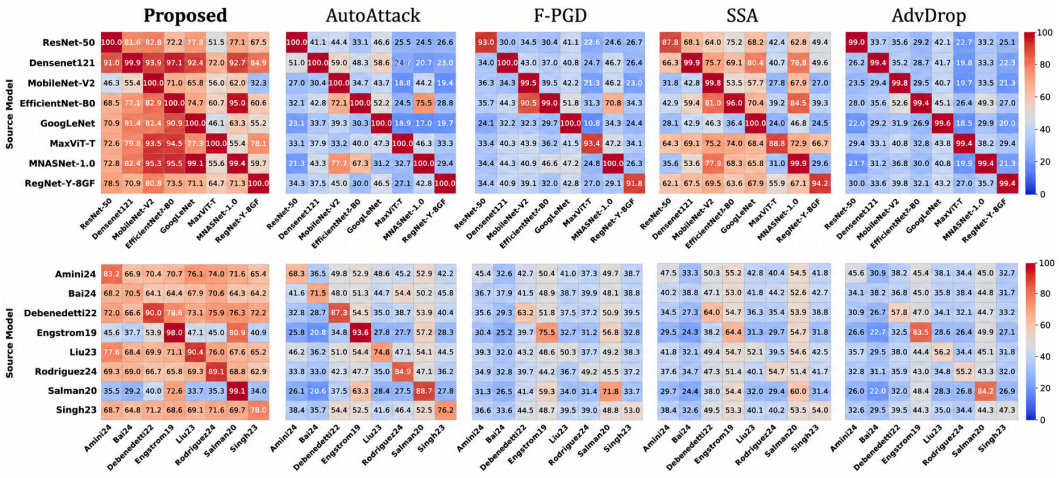


Fig. 5: ASR (in %) with $\varepsilon = 8/255$ and $\varepsilon_{\text{DGF}} = \eta 8/255$, on **ImageNet**, for different **pretrained (top)** and **RobustBench (bottom)** models. **Left to right: proposed attack, AutoAttack, F-PGD, SSA, AdvDrop**. **Color intensity indicates higher transferability**. We observe that our proposed attack is consistently more transferable than the baselines, as we double the attack level (cf. Fig. 3a), across a variety of models. These results conform with our motivation deriving attacks from a principled methodology, based on structured transforms.

scheme. While AutoAttack has a predefined number of iterations for each of the attacks involved in its design (e.g., it may be $K = 100$), we set the number of iterations to $K = 20$ for our proposed attack and the frequency-based baselines, to enable a fair comparison. Relative to F-PGD, the additional cost of the proposed attack arises from the implementation of $\Re(M_D)$, which requires applying its inverse on the gradient of the loss and projecting onto the perturbation set at each iteration. All computations exploit the separable structure of the transforms and are implemented via forward and adjoint operations. The projection step reduces to applying $(I + 2\lambda\Re(M_D))^{-1}$ from the left and right via structured linear solves. The scalar parameter λ is computed using a bisection procedure, which has logarithmic cost and introduces only a negligible overhead compared to the cost of the transform operations. In contrast, AutoAttack combines multiple iterative attacks and requires substantially more gradient evaluations, leading to significantly higher runtime. Similarly, SSA averages gradients over multiple spectrum transformations at each iteration, requiring more gradient evaluations than our proposed attack, while AdvDrop incurs additional cost through blockwise DCT/IDCT operations and differentiable quantization, without requiring the structured linear solves of our method. Compared to SSA, our method requires fewer gradient evaluations per iteration, while its structured solves differ from the blockwise transforms and differentiable quantization underlying AdvDrop's computational cost. All in all, any additional computational costs are compensated by substantial improvements in attack quality. Despite preserving the same iterative PGD framework as in F-PGD, our method consistently achieves markedly higher transferability across pretrained and robustified models (cf. Tables 1 - 3 and Figs. 3, 5, 6). Overall results indicate a favorable trade-off between computational cost and attack performance. For a complete list of all models deployed in our experiments, as these are taken from RobustBench, we refer to Table 2. For completeness purposes, we present below complementary experiments to those of Section 5.2.

Similarly to Fig. 3a and Fig. 3b, we present transferability heatmaps, serving as comparisons between the proposed attack and the baselines, on pretrained and RobustBench models. We illustrate the results in Fig. 5 for ImageNet, with $\varepsilon = 8/255$ for the baselines and $\varepsilon_{\text{DGF}} = \eta\varepsilon$ for our attack, and in Fig. 6 for CIFAR100, with $\varepsilon = 16/255$ for the baselines and $\varepsilon_{\text{DGF}} = \eta\varepsilon$ for our attack. To better reflect transferability, we also present in Table 3, the average cross-model ASR, calculated according to each heatmap of Fig. 5 and Fig. 6. We observe that, for the pretrained models, on both datasets, our attack is almost identically effective to AutoAttack, in terms of self-model ASR, demonstrating that attacks derived from an appropriate optimization problem, over a specified perturbation set, can outperform frequency-based attacks that only follow the formulation of (3) and (4), without being accompanied by a grounded maximization problem. Interestingly, while effectiveness and transferability of the baselines increase as we transfer to a more noisy regime than that of Fig. 3a and Fig. 3b, our attack still outperforms the baselines, consistently for both datasets, in terms of the transferability rate. This further justifies our motivation for designing an optimization problem whose

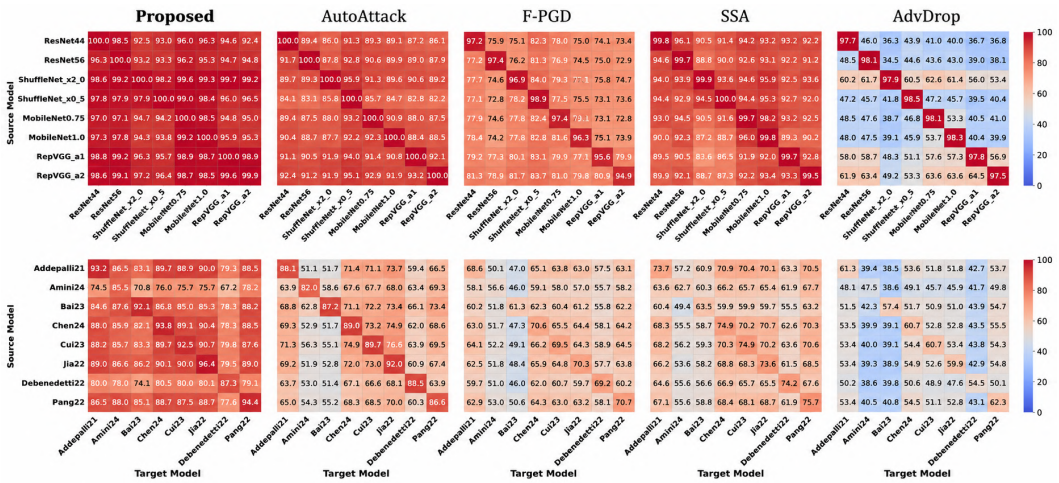


Fig. 6: ASR (%) on **CIFAR100**, with $\varepsilon = 16/255$ for the baselines and $\varepsilon_{\mathrm{DGF}} = \eta\varepsilon$ for the proposed attack, on different **pretrained (top)** and **RobustBench (bottom)** models. **Left to right: proposed attack, AutoAttack, F-PGD, SSA, AdvDrop**. **Color intensity indicates higher transferability**. We notice that our attack (left) is highly transferable, even more than on ImageNet (cf. Fig. 5), outperforming the baselines (middle and right), for both pretrained and robust models.

perturbation constraint set is endowed with a DGF, which better aligns with inherent SF patterns of natural images, thus the proposed attack captures more easily than purely spatial-/frequency-based attacks the shared spectral sensitivities of a range of models, both pretrained and defended.

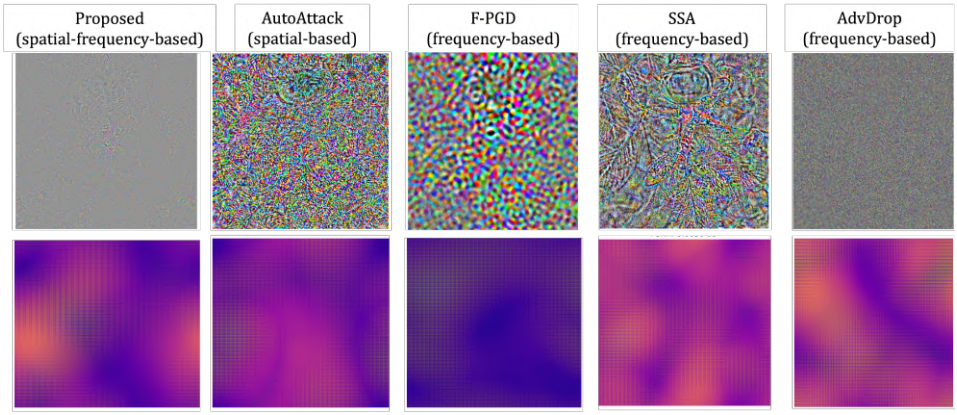
Finally, we illustrate original and adversarial image samples from ImageNet and CIFAR100, in Fig. 4 top and bottom, respectively, pertaining to the proposed attack, AutoAttack, F-PGD, SSA and AdvDrop, on pretrained models, as a means of examining whether our proposed methodology has the same effect depicted in Fig. 1 and Fig. 2, which relied on RobustBench models evaluations. For these image samples, we separately present in Fig. 7 the attacks illustrated in the spatial and frequency domain, so as to elaborate on the representation ability of each of the three attack methods. Firstly, we observe that our attack outperforms the baselines, by scoring a mildly higher SSIM value, for both datasets, albeit scoring the highest possible ASR, along with AutoAttack. Additionally, while all images retain a reasonable visual quality, they exhibit structural differences in both spatial- and frequency-based representations, highlighting that the attack effectiveness is not solely determined by operating in the frequency domain, but rather by how perturbations are organized and constrained during optimization. This phenomenon is further supported by Fig. 7, where we notice that the SF localization of the proposed attack is more evident, spread strategically in the transformed-domain, than that of the baselines. This observation conforms with our intuition of deploying Gabor frames, which induce norms that align better with natural image patterns than standard-ℓ_p balls, thereby leading to an attack that more flexibly handles image geometry to be highly effective, without sacrificing image quality.

D Impact Statement

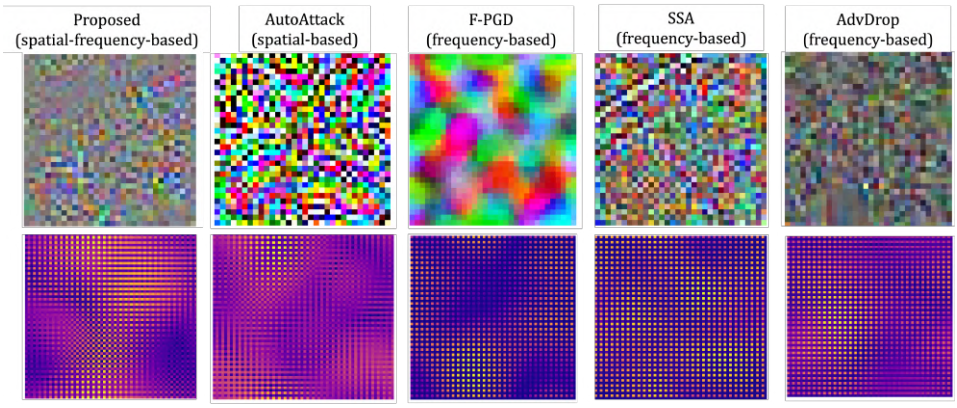
This work advances the understanding of adversarial vulnerabilities in deep learning models, by introducing a principled framework for generating frequency-based attacks, with our primary goal being the improvement of robustness evaluation and creation of better defenses.

Specifically, our framework can help researchers and practitioners better understand fundamental weaknesses in neural networks, particularly those that persist across different architectures. By demonstrating that attacks constrained in structured transform domains can be highly effective – even against adversarially trained models – we highlight the need for more comprehensive defense mechanisms, accounting for highly structured frequency-based attacks. This could lead to more robust models in safety-critical applications.

Like other adversarial attack research, this work could potentially be misused to compromise vision systems. However, we believe the benefits of openly studying these vulnerabilities outweigh the risks, as understanding attack mechanisms is essential for developing effective defenses. Our work does not introduce fundamentally new attack capabilities, but rather provides a more principled understanding



(a) ImageNet



(b) CIFAR100

Fig. 7: Exemplary **spatial (top)** and **frequency (bottom)** representation of the **proposed attack and the four baselines**, corresponding to the adversarial images of Figure 4. Similarly to Fig. 1b and Fig. 2b, we observe that the SF ability of our attack, for both datasets, provides a balanced representation in the transformed space, while the baselines rely mostly on spatial and frequency coordinates, thus missing the flexibility offered by the DGF.

of existing frequency-based attacks. By grounding adversarial behavior in a mathematically principled framework, our study provides tools for interpreting model vulnerabilities in spectral terms. We anticipate that these insights can encourage the development of robustness benchmarks and training schemes that better account for structure in real-world data representations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: All the claims found in the abstract and introduction are properly supported in the main results of Sec. 4 and the experimental results of Sec. 5. Additionally, proofs for our derived theory can be found in Appendix B and experimental extensions in Appendix C.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: Throughout the paper, we state the limitations of our work, among which the fact that a more thorough empirical evaluation is needed in the future for our proposed methodology (cf. Sec. 5).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For the course of our theoretical results presented in Sec. 4, we state minimal and justified assumptions, to clarify every possible dependency of the problem. All associated proofs can be found in a complete form at Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In Sec. 5.1, we provide details on the experimental setup of the paper leading to the corresponding experimental results. We also employ example datasets used in papers that are close to our work, while in Sec. C, we outline more experimental details, including choice of hyperparameters for reproducibility purposes.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Upon acceptance, we will provide a link to a public github repository with pytorch code, and sufficient documentation for reproducibility of all the experimental results that accompany the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We briefly describe the main experimental setup in Section 5.1, and then further elaborate on all details in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: All experiments are repeated at least 10 times with different random seeds, and results are reported as averages over these runs. Additionally, we compute the standard deviation across runs, which we found to be consistently small (typically below 0.05 in

absolute accuracy), and thus corresponding error bars are omitted from the figures for the sake of readability.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: For the course of our experiments, we present detailed descriptions on the computer resources in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We preserved anonymity in our submission. Our submission abides by the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss potential societal impacts of our work in Appendix D.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: : Our paper does not pose any such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All original owners of assets are properly credited, e.g., we cite the original papers of the baseline attacks we investigate for comparison with our proposed attack.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human subjects were involved in the experiments conducted for this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No IRB, crowd-sourcing, or research with human subjects were involved in the experiments conducted for this paper.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core methods developed during the present research do not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.