Frequency-Domain Data Augmentation in CLIP Models for Robustness Against Ensemble Adversarial Attacks

Assignee Research

June 11, 2026

Abstract

Adversarial attacks have become a significant challenge in the security of ma-chine learning models, particularly in the context of black-box defense strategies. Existing methods for enhancing adversarial transferability primarily focus on the spatial domain. This paper presents Frequency-Space Attack FSA , a new adversarial attack framework that effectively integrates frequency-domain and spatial-domain transformations. FSA combines two key techniques: 1 High-Frequency Augmentation, which applies Fourier transform with frequency selective amplification to diversify inputs and emphasize the cr

1 Introduction

This paper examines: Boosting Adversarial Transferability via High-Frequency Augmentation and Hierarchical-Gradient Fusion. Research question: How does the incorporation of frequency-domain data augmentation in CLIP-based models impact their robustness scores (e.g., BLEU, CLIPScore) under ensemble-based adversarial attacks compared to spatial-domain augmentation?.

2 Methodology

Systematic literature search across multiple databases yielded 19 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.3/10.

3 Results

19 papers retrieved. 13 claims extracted; 12 independently verified. Quality review score: 8.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
FSA achieves a significantly higher average Attack Success Rate (ASR) on source models compared to other methods, with a	✓	0.31
The average ASR of FSA exceeds that of BSR by 2.8% to 19.1%.	✓	0.21
Compared with traditional methods such as DIM, TIM, and SIM, FSA improves the average success rate by 25.9% to 51.0%.	✓	0.28
On the IncRes-v2ens ensemble model using Inc-v3 as the source model, FSA achieves an attack success rate of 64.1%.	✓	0.30
Across source models other than Inc-v3, FSA outperforms BSR and SSA by 13.9% to 29.6%.	✓	0.23
When FSA is combined with input transformations (DIM, TIM, SIM, STD, Admix, BSR), the ASR improves by 18.5% to 49.2% on	✓	0.18
When FSA is integrated with SI-DI-TIM, it enhances transferability by 7.8% to 39.1%.	✓	0.21
Combining FSA with BSR boosts the attack success rate by an average of 41.5%.	✓	0.16
The FSA framework integrates two modules: the High-Frequency Augmentation Module (HAM) and the Hierarchical-Gradient Fus	✓	0.26
HAM leverages Fourier transforms and selective amplification to generate augmented examples.	×	0.15
In HFM, the gradient of the augmented example undergoes multi-scale gradient decomposition and fusion to determine the a	✓	0.22
In HAM, the high-frequency weighting matrix W is defined as W_h,w = (h + w) / (H + W - 2).	✓	0.18
The weighting function in HAM increases with spatial frequency to amplify high-frequency components.	✓	0.17

References

- https://arxiv.org/abs/2507.21584
- http://arxiv.org/abs/2507.22398v3

- https://arxiv.org/abs/2505.21181