Advanced NS-Net: Improving Generalizable AI-Generated Image Detection via Learned Semantic Null-Space Projections

Sachi Deshmukh
Indian Institute of Technology Bombay
22b1213@iitb.ac.in
(22B1213)

Shiwani Mishra
Indian Institute of Technology Bombay
shiwani.mishra@iitb.ac.in
(22B3907)

Neel Rambhia
Indian Institute of Technology Bombay
neel.rambhia@iitb.ac.in
(22B1298)

***Abstract*—The rapid advancement of generative models—ranging from GANs to modern diffusion systems—has enabled the creation of photorealistic synthetic images that closely resemble real photographs. This unprecedented realism amplifies the risk of misinformation, digital forgery, and erosion of public trust, making reliable and generalizable AI-generated image (AIGI) detection increasingly critical. However, existing detectors suffer from sharp degradation when encountering unseen generators, largely because they inadvertently rely on the semantic content of images rather than the intrinsic artifacts left by generative models. To address this limitation, we build upon NS-Net [1] and propose Advanced NS-Net, an enhanced AIGI detector that learns a refined semantic representation and performs dynamic per-sample null-space projection to decouple semantic information. Experiments demonstrate substantial improvements, surpassing the baseline NS-Net model. This highlights the effectiveness of learned semantic refinement and dynamic orthogonalization for improving generalizable AI-generated image detection.**

I. Introduction

Generative models such as ProGAN (Karras et al. 2018) [2], GauGAN (Park et al. 2019) [3], StyleGAN (Karras et al. 2020) [4], Midjourney [5], DALL-E [6], and Stable Diffusion (Rombach et al. 2022) [7] have reached a level of realism where AI-generated images are nearly indistinguishable from real photographs. While this progress unlocks creative and industrial potential, it simultaneously introduces severe risks: manipulated media, misinformation at scale, identity fraud, and erosion of public trust in digital content (Li et al. 2025c) [8]. As a result, AI-generated image detection has become an essential research direction in trustworthy AI and digital forensics.

Authenticity detection—determining whether a media input is *real* or *AI-generated*— is formally, a binary classification problem [9]:

$$F_\theta : X \to \{\text{real}, \text{fake}\}. \tag{1}$$

Early AIGI detectors relied on supervised learning over handcrafted artifacts specific to a particular generator. Although effective in narrow settings, these methods fail under distribution shifts, especially when encountering unseen generative models. More recent approaches employ vision–language models such as CLIP (Radford et al. 2021) [10], leveraging multimodal alignment to capture high-level semantic cues. However, a growing body of evidence shows that CLIP visual embeddings contain strong semantic priors that entangle object semantics with generator-specific artifacts. When real and generated images share similar semantics—e.g., cats, cars, landscapes—these semantic components become confounding factors, leading to severe overfitting and limited generalization (Zheng et al. 2024) [11].

NS-Net (Yan et al. 2025) [1] addressed this issue by projecting CLIP visual features into the null-space of textual features, effectively removing global semantic information and exposing subtle distributional differences introduced by generative processes. Despite its success, NS-Net exhibits two critical limitations: (1) the null-space is constructed using raw CLIP text embeddings, which may not span a clean semantic subspace, and (2) the projection matrix is computed globally, ignoring sample-specific variations.

To resolve these constraints, we propose **Advanced NS-Net**, a refined architecture that enhances semantic disentanglement via two key contributions. First, a learnable *Semantic Mapper* reshapes CLIP's textual features into a more coherent semantic basis, allowing the null-space to better capture semantic directions. Second, a *dynamic per-sample null-space projection* is computed for each image–caption pair, enabling precise removal of semantic components while preserving artifact-specific signals. Experiments on the DALLE Recognition dataset reveal substantial gains over the baseline NS-Net, demonstrating that learning semantic structure and applying individualized orthogonalization are crucial for next-generation generalizable AIGI detectors.

The rest of the paper is organized as follows: Section II gives the background information regarding NS-NET. Section III talks about the limitations in NS-NET and Section IV deals with our proposed improvements for the same. The methodology is discussed in Section V followed by discussion on training and results in Section VI and Section VII respectively. We end the paper with Section VIII talking about Future Work and Section IX being the conclusion.

II. Background: NS-Net

The emergence of large vision–language models such as CLIP has demonstrated that their embedding spaces encode rich, high-level semantic information. While this semantic alignment is beneficial for downstream multimodal tasks, it

poses a fundamental challenge for AI-generated image detection. Real and synthetic images often share nearly identical global semantics (e.g., "a white horse in a field"), causing CLIP's image features to cluster according to semantic similarity rather than their authenticity. As a result, classifiers trained directly on CLIP embeddings frequently fail to capture subtle forensic cues that differentiate real from generated images.

NS-Net [1] addresses this challenge by introducing the core hypothesis that ***forensic artifacts lie in a subspace orthogonal to the semantic subspace encoded by CLIP***. Thus, removing semantic components from CLIP visual embeddings should amplify authenticity-relevant signals.

Let $v_i \in R^d$ denote the CLIP textual feature associated with the caption of the i-th training image. These features span the **semantic subspace**

$$V = \text{span}\{v_1, v_2, \ldots, v_n\} \subset R^d. \tag{2}$$

To isolate this semantic subspace, NS-Net constructs the matrix

$$\mathbf{V} = [v_1, v_2, \ldots, v_n]^\top,$$

and computes its singular value decomposition (SVD):

$$\mathbf{V} = U\Sigma W^\top. \tag{3}$$

The columns of W corresponding to zero singular values form an orthonormal basis of the **null-space** of the semantic subspace:

$$\mathcal{N}(V) = \{x \in R^d \mid v_i^\top x = 0,\ \forall i\}. \tag{4}$$

If $R \in R^{d\times k}$ contains these null-space basis vectors, then the orthogonal projection matrix onto the null-space is:

$$P = RR^\top, \qquad P = P^\top,\ \ P^2 = P. \tag{5}$$

Given a CLIP visual embedding f_{img}, NS-Net computes the **decoupled visual representation**:

$$f_{\text{null}} = f_{\text{img}}P, \tag{6}$$

which removes task-irrelevant semantic information while preserving residual components associated with synthesis artifacts (e.g., texture irregularities, spectral distortions, and generator-specific signals).

The projected features are then optimized under a contrastive learning objective that enforces:

- *intra-class compactness* among real images,
- *inter-class separation* between real and synthetic images.

This yields a representation space governed primarily by authenticity cues rather than semantics, enabling strong generalization to unseen generative models and diverse image distributions.

III. Limitations of Original NS-Net

From our analysis and empirical experimentation, the original NS-Net has three key limitations:

- CLIP textual embedding is *not* the true semantic basis; it is only one component of the multimodal space.
- Null-space projection matrix P is precomputed once from all captions, making it static and inflexible.
- Projection is *global*, whereas each image may require a different semantic basis for accurate orthogonalization.

These issues motivated our improved design, proposed in the next section.

IV. Proposed Improvements

A. Learned Semantic Mapper

We introduce a 2-layer MLP with residual skip and LayerNorm:

$$f_{\text{out}} = \text{LayerNorm}(v + \text{MLP}(v)). \tag{7}$$

This module learns a refined representation of semantic features.

With the presence of skip connections, at worst our model learns the identity mapping itself, which is the original NS-Net implementation. In practice it normalizes semantics and improves null-space quality.

The Semantic Mapper configuration used:

- Input dimension: 768
- Hidden layer: 1024
- Output dimension: 768

B. Per-Sample Null-Space Projection

Instead of static P, we compute:

$$P_i = I - \frac{v_i v_i^\top}{\|v_i\|^2}, \tag{8}$$

for each image-caption pair. This ensures the semantic removal is tailored to each sample.

C. Improved Loss Function

We combine:

$$\mathcal{L} = \mathcal{L}_{\text{contrastive}} + \alpha\mathcal{L}_{\text{BCE}} + \beta\mathcal{L}_{\text{align}}, \tag{9}$$

where alignment penalizes divergence between mapped text and visual features.

V. Methodology

This section describes the improved NS-Net pipeline we implemented. We first summarize notation and then present the detailed mathematical operations for Patch Selection, feature extraction (with LoRA), semantic mapping, NULL-space computation (global and per-sample variants), projection, and the combined training objective used in our experiments.

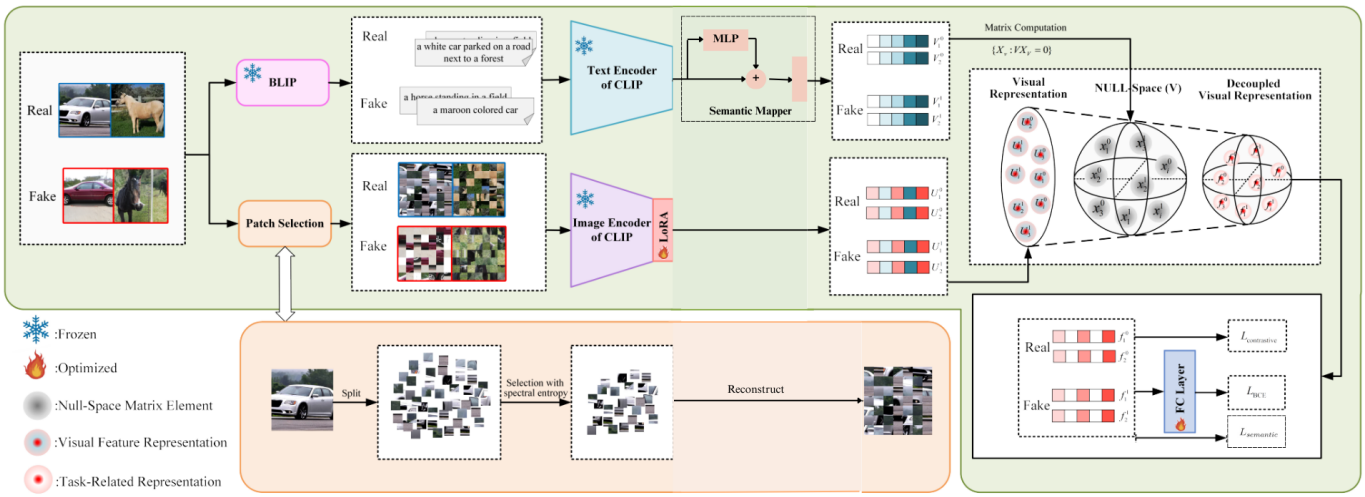

Fig. 1. Overall architecture of Advanced NS-Net framework. Image adapted and modified from [1].

1) Notation: Let a training dataset be

$$\tilde{\mathcal{D}} = \{(\text{img}_j, \text{text}_j, y_j)\}_{j=1}^{N}, \qquad y_j \in \{0, 1\}, \tag{10}$$

where $y = 0$ denotes real and $y = 1$ denotes AI-generated images. We denote CLIP image encoder outputs by $E_I(\cdot)$ and text encoder outputs by $E_T(\cdot)$. After applying LoRA adapters to the CLIP image encoder (ViT-L variant), image features are written as

$$U = E_I^{\text{LoRA}}(\text{img}) \in R^{B \times d_I}, \tag{11}$$

and text (semantic) features as

$$V = E_T(\text{text}) \in R^{B \times d_T}, \tag{12}$$

where B is batch size and d_I, d_T are feature dimensions (for ViT-L / CLIP we use $d_I = d_T = 768$ in our implementation).

2) Patch Selection (spectral entropy): To preserve fine-grained artifacts and reduce semantic dominance, each input image is partitioned into non-overlapping patches $\{x_i\}_{i=1}^{n}$, $x_i \in R^{N \times N \times 3}$ (we use $N = 32$). For each patch compute the FFT magnitude spectrum and derive its *spectral entropy*:

$$S(x_i) = -\sum_k p_{i,k} \log(p_{i,k}), \quad p_{i,k} = \frac{|\mathcal{F}(x_i)_k|}{\sum_k |\mathcal{F}(x_i)_k| + \epsilon}. \tag{13}$$

We select the top-K patches by spectral entropy and bottom-K patches (to preserve both texture–rich and texture–poor regions), shuffle them and reconstruct an $M \times M$ input for CLIP. This step reduces semantic bias from global layout while preserving traces left by generators.

3) Caption Generation and Dataset Construction: To enable NULL-space construction from semantic features, we generate captions automatically using the BLIP model (*Bootstrapping Language-Image Pretraining*) [12].

Given an input image img, BLIP produces a natural-language caption which aims to summarize high-level semantic content of the image. After caption generation, each dataset sample becomes a triplet:

$$(\text{img}_j,\ \text{text}_j,\ y_j),$$

where $y_j \in \{0, 1\}$ indicates real or AI-generated. This expanded dataset is denoted:

$$\tilde{\mathcal{D}} = \left\{(\text{img}_j, \text{text}_j, y_j)\right\}_{j=1}^{N}.$$

We feed the image and generated caption to CLIP to extract the Visual and Text features.

4) Semantic Mapper (residual MLP): We introduce a small trainable Semantic Mapper $\mathcal{M}_\theta$ that maps text features into a *semantic proxy* used to compute the NULL-space. The mapper is a residual MLP:

$$\mathcal{M}_\theta(v) = \text{LayerNorm}\left(v + W_2\phi(W_1 v + b_1) + b_2\right), \tag{14}$$

where $v \in R^{d_T}$, $W_1 \in R^{d_T \times h}, W_2 \in R^{h \times d_T}$, hidden size $h = 1024$, and $\phi(\cdot)$ is ReLU. In our experiments we use $d_T = 768$, $h = 1024$ (matching your implementation).

This mapper allows the model to refine raw CLIP text embeddings into semantic vectors better aligned with detection-relevant semantics (or to be learned when pretrained text features are suboptimal).

5) NULL-Space computation (global and per-sample):

a) Global NULL-space (training captions): Given a set of semantic vectors (e.g., all mapped training captions) stacked into matrix $V \in R^{N_{\text{text}} \times d_T}$, compute SVD:

$$V = U\Sigma N^\top. \tag{15}$$

Let $N_{\text{null}} \in R^{d_T \times r}$ be the right singular vectors corresponding to (near) zero singular values (we threshold singular values by τ_{sv}). The NULL-space projection matrix is

$$P_V = N_{\text{null}} N_{\text{null}}^\top \in R^{d_T \times d_T},$$

which satisfies $VP_V \approx 0$. Projecting visual features onto the NULL-space removes components aligned with text semantics:

$$\widetilde{U} = UP_V. \tag{G}$$

This is the approach proposed in the original NS-Net (SVD + global projection) [1].

b) Per-sample NULL projection (our improvement / alternative): To further decouple semantics at the sample level and avoid relying on a single global null-basis, we implemented a per-sample rank-1 orthogonal projection that removes the semantic direction v_i for each example. Given a mapped semantic vector $v_i \in R^{d_T}$ (row of $\mathcal{M}_\theta(V)$) we form the normalized projection:

$$P_i \;=\; I - \frac{v_i^\top v_i}{\|v_i\|_2^2} \in R^{d_T \times d_T}, \qquad \widetilde{u}_i = u_i P_i, \tag{16}$$

where u_i is the image feature for the same sample (note shape conventions: u_i is row vector, v_i is row vector). This per-sample projection removes the principal semantic direction contributed by the paired text prompt and is numerically stable by clamping denominator with ϵ. The per-sample variant works well when batch captions are heterogeneous or when a global nullspace is noisy.

6) Contrastive learning and auxiliary classification: To encourage features of the same label to cluster while pushing different labels apart, we adopt a supervised contrastive loss (NT-Xent style) as the primary objective and a binary cross-entropy (BCE) classifier loss as an auxiliary supervision.

Let $\{f_i\}_{i=1}^B$ be the *normalized* decoupled (projected) features:

$$f_i = \text{Normalize}\left(\, \mathcal{H}(\widetilde{u}_i) \,\right) \in R^{d_f}, \tag{17}$$

where $\mathcal{H}$ is a small projection head (two-layer MLP) producing d_f-dim features and Normalize means ℓ_2 normalization. For anchor i define positives $P_i = \{j \neq i : y_j = y_i\}$. The supervised contrastive loss used is:

$$\mathcal{L}_{\text{ctr}} = -\frac{1}{|P_i|} \sum_{p \in P_i} \log \frac{\exp(f_i \cdot f_p/\tau)}{\sum_{j \neq i} \exp(f_i \cdot f_j/\tau)} \quad \text{(averaged over } i\text{)}. \tag{18}$$

Additionally, we include classification logits $s_i = \text{linear}(f_i)$ and BCE loss:

$$\mathcal{L}_{\text{BCE}} = \frac{1}{B} \sum_{i=1}^{B} \text{BCE}(s_i, y_i). \tag{19}$$

a) Alignment loss (our introduced term).: To stabilize training and enforce correspondence between text semantics and image features prior to null projection, we also include an *alignment* mean squared error term between the mapped text vectors and image features (detached as required):

$$\mathcal{L}_{\text{align}} = \frac{1}{B} \sum_{i=1}^{B} \|\mathcal{M}_\theta(v_i) - u_i^{\text{detach}}\|_2^2. \tag{20}$$

This term encourages $\mathcal{M}_\theta$ to produce semantic proxies that are consistent with the CLIP image embedding space (useful when the mapper is jointly trained). In your experiments you used a weight for this MSE term to balance how strongly semantics influence training.

7) Final training objective: We combine the three terms into a single objective:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{ctr}} \;+\; \alpha\, \mathcal{L}_{\text{BCE}} \;+\; \beta\, \mathcal{L}_{\text{align}}, \tag{21}$$

where α and β are hyperparameters. In our current runs we used $\alpha = 0.2$ and $\beta = 0.3$ (empirically chosen; these can be tuned on a validation set). The original NS-Net used a weighted contrastive + BCE formulation; we extend it by adding the alignment MSE term that improves mapper training and numeric stability.

8) Training and Inference:

- **Training:** For each mini-batch, perform patch selection reconstruction, encode images with CLIP (LoRA-augmented) to get U, tokenize captions and encode text to get V, pass V through $\mathcal{M}_\theta$; compute either global P_V (from a larger caption pool) or per-sample P_i and obtain projected features $\widetilde{U}$; pass $\widetilde{U}$ through projection head and compute $\mathcal{L}$; update the projection head and $\mathcal{M}_\theta$ (keep CLIP backbone frozen except LoRA adapters).
- **Inference:** Use the learnt mapper to compute per-sample or global projection and classify using the trained linear head on projected features. Per-sample projection has the advantage of being robust to caption heterogeneity.

9) Implementation details and stability:

- For SVD we threshold singular values by $\tau_{\text{sv}} = 1\text{e}{-6}$ to determine the null basis.
- For per-sample projection we clamp denominator $\|v_i\|_2^2 \geq \epsilon$ with $\epsilon = 1\text{e}{-8}$ to avoid division by zero.
- Dimension choices: CLIP ViT-L features $d = 768$, SemanticMapper hidden $h = 1024$, projection head $d_f = 512$ (these match your current code).
- CLIP backbone weights are frozen; LoRA adapters are trained/finetuned (as you applied) to adapt image encoder representations without full fine-tuning.

The architecture is summarized and illustrated in Fig. 1.

VI. Dataset

Due to limited compute resources, we used a subset of the DALLE Recognition Dataset from Kaggle:

- **2000** real images for training
- **2000** fake images for training
- **100** real + **100** fake for testing

The original NS-Net paper trains on AIGIBench (Li et al. 2025d) [13] consisting of 144k images , but this was computationally infeasible in Colab.

VII. Experimental Results

Table I compares the original NS-Net implementation vs. our improved version.

Our improvements yield large gains, especially in fake accuracy, aligning with the goal of detecting generated images.

TABLE I
PERFORMANCE COMPARISON

Metric	Original	Ours
Overall Accuracy	0.56	**0.93**
Real Accuracy	0.55	**0.86**
Fake Accuracy	0.56	**1.00**
Average Precision	0.57	**0.99**

VIII. FUTURE WORK

Following the original NS-Net paper, future steps include:

- Training on large-scale AIGIBench (Li et al. 2025d) [13] dataset.
- Benchmarking on:
  - GenImage (Zhu et al.2024) [14]
  - UniversalFakeDetect (Ojha, Li, and Lee 2023) [15]
  - AIGIBench (Li et al. 2025d) [13]
- Evaluating cross-generator generalization on unseen models.
- Replacing CLIP ViT-L/14 with CLIP-LLaVA or SigLIP for improved embedding structure.

IX. CONCLUSION

We successfully improved NS-Net by introducing:

- a learnable Semantic Mapper,
- per-sample null-space projections,
- refined loss functions.

These modifications greatly enhance the model's ability to detect AI-generated images while reducing semantic dependence. The project demonstrates that learned semantic normalization can significantly advance generalizable detection of synthetic content.

REFERENCES

[1] D. Yan et al., "NS-Net: Decoupling CLIP Semantic Information through Null-Space for Generalizable AI-Generated Image Detection," arXiv:2508.01248, 2025.

[2] Karras, T., Aila, T., Laine, S., and Lehtinen, J. 2018. "Progressive Growing of GANs for Improved Quality, Stability, and Variation," In International Conference on Learning Representations, 2018

[3] Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. 2019. "Semantic image synthesis with spatially-adaptive normalization," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2337–2346.

[4] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T., "Analyzing and improving the image quality of StyleGAN". CVPR. pp. 8110–8119 (2020)

[5] Midjourney: https://www.midjourney.com/home (2023)

[6] OpenAI: https://openai.com/dall-e-3 (2023)

[7] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: https://github.com/CompVis/stable-diffusion (2022)

[8] Li, Y., Shao, S., He, Y., Guo, J., Zhang, T., Qin, Z., Chen, P.-Y., Backes, M., Torr, P., Tao, D., et al. 2025c. "Rethinking data protection in the (generative) artificial intelligence era," arXiv preprint arXiv:2507.03034.

[9] Y. Zou, P. Li, Z. Li, H. Huang, X. Cui, X. Liu, C. Zhang, and R. He, "Survey on AI-Generated Media Detection: From Non-MLLM to MLLM," 2024.

[10] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML, 2021.

[11] Zheng, C., Lin, C., Zhao, Z., Wang, H., Guo, X., Liu, S., Shen, C. 2024. "Breaking semantic artifacts for generalized aigenerated image detection," Advances in Neural Information Processing Systems, 37: 59570–59596.

[12] J. Li et al., "BLIP: Bootstrapped Language-Image Pre-training," NeurIPS, 2022.

[13] Z. Li, J. Yan, Z. He, K. Zeng, W. Jiang, L. Xiong, and Z. Fu, "Is artificial intelligence generated image detection a solved problem?" arXiv preprint arXiv:2404.01337, 2024.

[14] Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y. 2024. "Genimage: A millionscale benchmark for detecting ai-generated image," Advances in Neural Information Processing Systems, 36.

[15] Ojha, U., Li, Y., Lee, Y. J. 2023. "Towards universal fake image detectors that generalize across generative models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 24480–24489.

[16] D. Cozzolino, G. Poggi, M. Nießner, and L. Verdoliva, "Zero-Shot Detection of AI-Generated Images," in Proc. European Conf. Computer Vision (ECCV), Milan, Italy, 2024, pp. 54–72.

[17] J. Chen, J. Yao, and L. Niu, "A Single Simple Patch is All You Need for AI-generated Image Detection," arXiv:2406.02580, 2024.

[18] A. Laević, T. Kramberger, R. Kramberger, and D. Vlahek, "Detection of AI-generated synthetic images with a lightweight CNN," AI, vol. 5, no. 3, pp. 1575–1593, 2024, doi: 10.3390/ai5030076.

[19] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," IEEE Access, vol. 10, pp. 25494–25513, 2022, doi: 10.1109/ACCESS.2022.3154404.