Advanced NS-Net: Model for Generalizable AI Generated Image Detection

Course Project for EE782

Sachi Deshmukh, Shiwani Mishra, Neel Rambhia

Indian Institute of Technology, Bombay

Problem Statement

Authenticity Detection:

- Binary Classification task to determine whether a given media X is authentic (real) or AI Generated
- Formal Definition: Build the detection model $F_\theta : X \rightarrow \{real, fake\}$

Explainability

- Provide humanly interpretable reasoning for detection decision based on salient features

Problem Statement

- Developing a **robust, generalizable detection model** is difficult
- Current detectors are highly dependent on **semantic information**
- Utilize CLIP (Radford et al. 2021), a multimodal model pre-trained on large-scale image-text dataset
- However, CLIP based detectors inadvertently use semantic correlation to predict "fake"

NS-Net Architecture

- Reference: Yan et al. (2025). NS-Net: Decoupling CLIP Semantic Information through NULL-Space for Generalizable AI-Generated Image Detection. arXiv:2508.01248
- Yan et. al. show that **semantic information** embedded in the visual features, extracted by CLIP, **negatively impacts** performance of AI-generated Image detection
- Key Idea: The artifact information is **orthogonal** to the semantic information in embedding space

NS-Net Architecture

- NS-Net uses NULL-Space projection to **decouple** semantic information from CLIP's visual features
- Then **contrastive learning** is used to capture intrinsic distributional differences between real and generated images

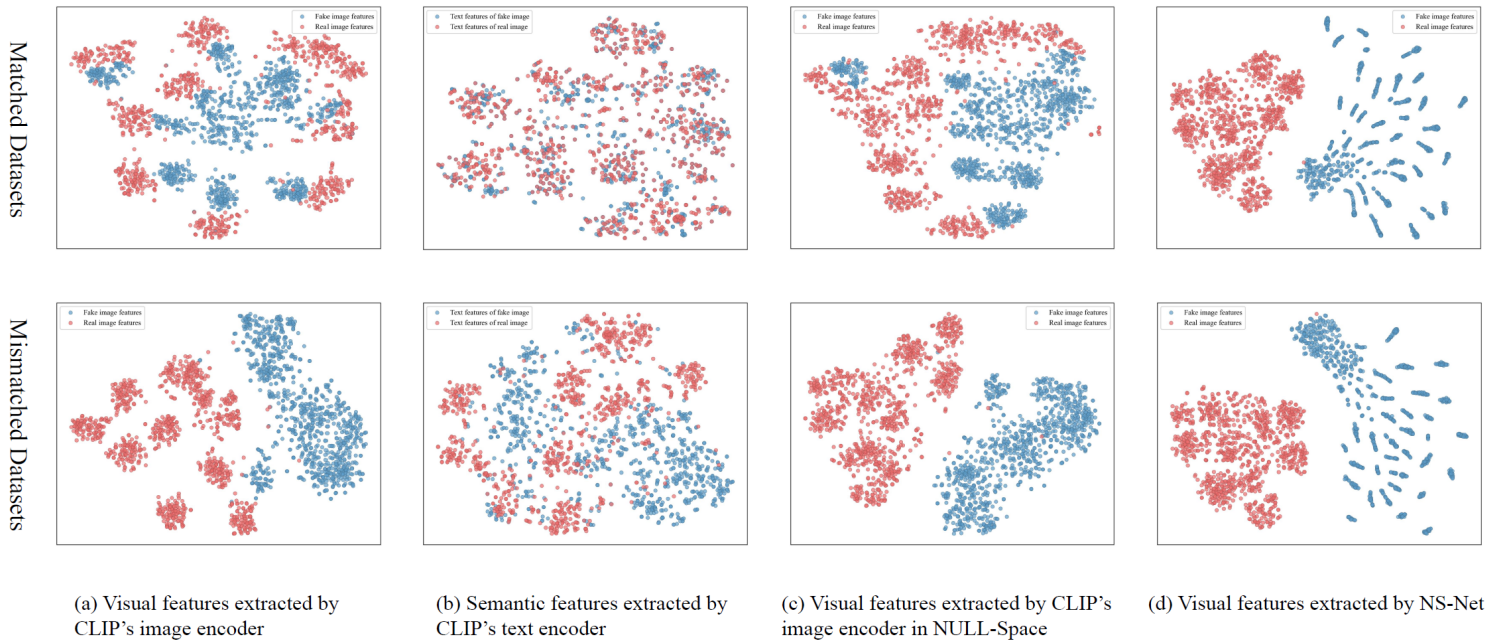

Figure 1: T-SNE Visualization of Features. Image adapted from Yan et al., NS-Net, arXiv 2025

NS-Net Architecture

Methodology:

1. Patch Selection of Input Images
2. Caption the input image using BLIP
3. Extract Visual and Textual Features of Image and Caption from CLIP
4. Semantic subspace $V \approx$ Textual Feature vector of CLIP
5. Null Space $Null(V)$ is calculated using SVD
6. Visual features are projected on this Null Space
7. Contrastive learning used to capture intrinsic distributional differences between real, fake images
8. $Loss = (1 - \lambda)\ Contrastive + \lambda\ BCE$

NS-Net Architecture

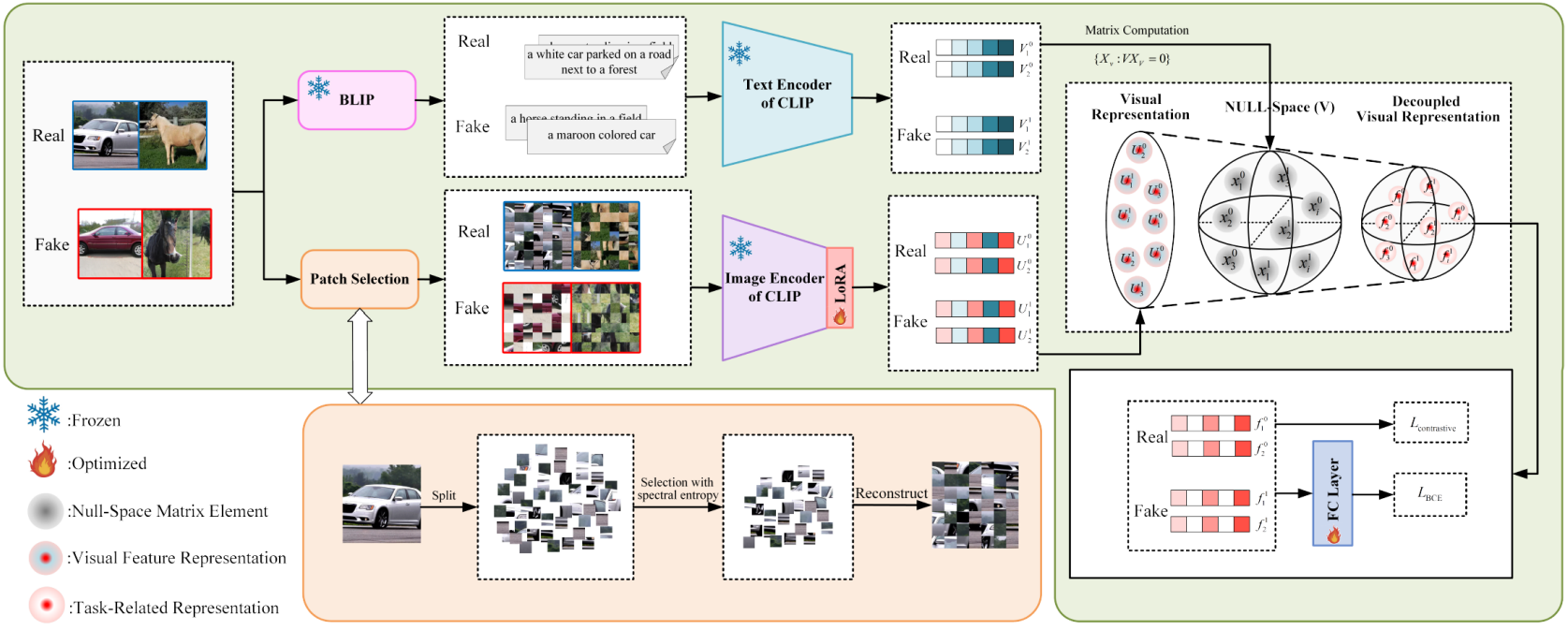

Figure 2: Architecture of NS-Net Model. Image adapted from Yan et al., NS-Net, arXiv 2025

Improvements to NS-Net Architecture

Potential limitations of the original model:

- Textual features from CLIP $\neq V$ but rather a component of it
- Additionally, it uses pre-computed null projection P based on all captions

Improvements:

- Learn a semantic mapping: $Text_{CLIP} \rightarrow Null(V)$
- Null space becomes dynamic and semantically normalized
- Semantic Mapper module (2-layer MLP + Residual Skip connection): $f_{out} = LayerNorm(V + MLP(V))$
- At worst, it learns the identity mapping $f_{out} = V$
- Projection Matrix per sample $P_i = I - \frac{f_{out} f_{out}^T}{||f_{out}||}$

Improvements to NS-Net Architecture

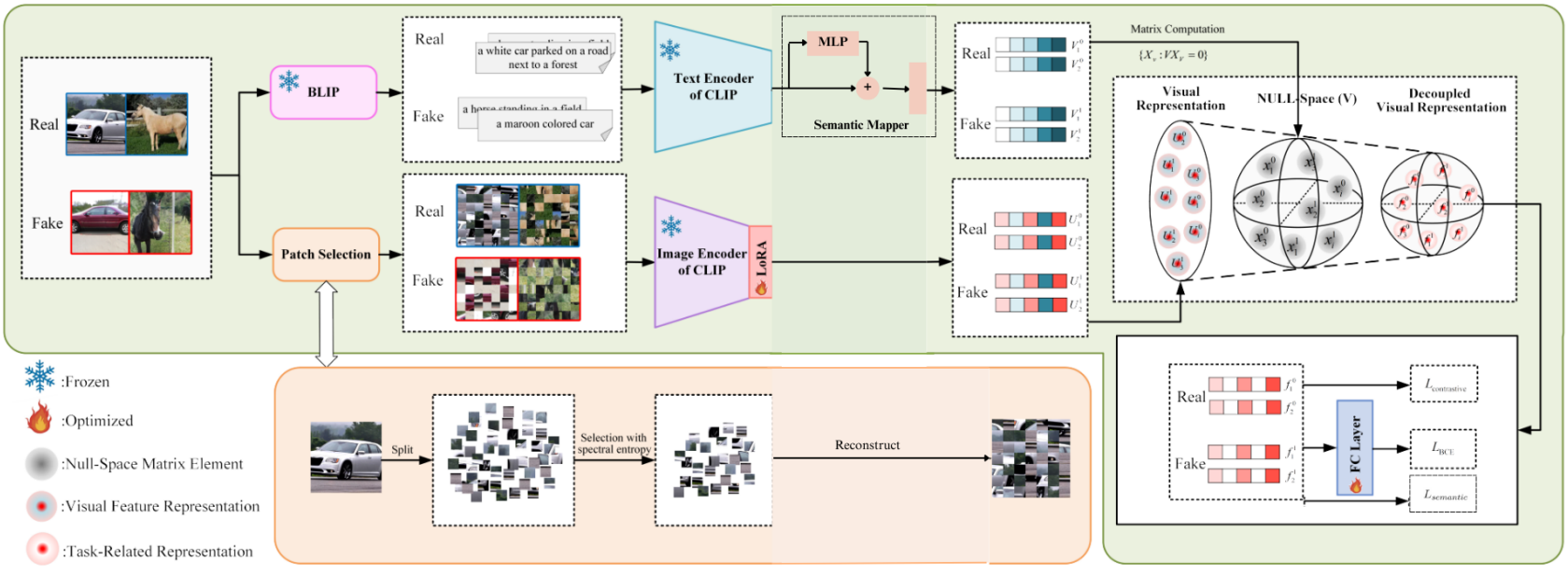

Figure 3: Architecture of Advanced NS-Net Model. Image adapted and modified from Yan et al., NS-Net, arXiv 2025

Evaluation

- We trained the models on Dalle Recognition Dataset from Kaggle
- Due to limited compute resources, we used 2000 Real and Fake images in training with 2 epochs and 200 images for testing

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

Table 1: Comparison of the models' performance

Further Tasks

- We have to benchmark our model against NS-Net and other SOTA models
- Original paper used AIGIBench (Li et al. 2025d) which involved training 144k images spanning: car, cat, chair and horse
- Testing was done on GenImage (Zhu et al. 2024), UniversalFakeDetect (Ojha, Li, and Lee 2023) AIGIBench (Li et al. 2025d)
- Their primary evaluation metric is **Fake Accuracy**

Thank You