Analyzing Real and Generative Image Models Through Feature Maps and Embedding Spaces

Liran Azran

September 2025

1 Introduction

Modern generative image models have evolved from **Generative Adversarial Networks (GANs)** to **Diffusion Models**. This transition fundamentally changed the nature of synthetic images, influencing both their statistical fingerprints and the way they are represented inside deep neural networks.

In this study, we examine how images from several generative models are represented by three standard recognition backbones: **ResNet-50**, **Xception**, and **Vision Transformer (ViT)**. These architectures represent distinct paradigms of visual representation learning:

- **ResNet-50** – A classical convolutional neural network (CNN) that processes the image hierarchically using residual blocks, emphasizing local textures and hierarchical features.
- **Xception** – A depthwise separable CNN that improves representational efficiency and captures finer details by factorizing standard convolutions into spatial and channel-wise operations.
- **Vision Transformer (ViT)** – A transformer-based model that divides the image into patches and models long-range dependencies via self-attention, focusing on global context and semantic relations.

Each of these architectures provides a different lens into the internal representation of GenAI images – ResNet and Xception emphasize low- and mid-level visual patterns, while ViT encodes more global and semantic cues. This study explores how these representations vary when analyzing images produced by both GAN and diffusion models.

Representational Assumptions. We further consider the relationship between two key representational domains within each model: the **embedding space** and the **feature map space**. The embedding space, derived after global pooling or token aggregation, can be viewed as a compact representation that already contains a rich distribution of activations with sufficient diversity and sparsity to characterize differences between generation mechanisms. In contrast, the feature maps from earlier or intermediate layers contain a substantially higher-dimensional signal, preserving finer-grained spatial and structural information. While it might be expected that combining both representations could enhance discrimination between generative models, our analysis indicates that this relationship is more nuanced- in certain cases, embeddings alone capture relevant distinctions effectively, whereas in others, feature maps provide stronger separation. This observation motivates a deeper examination of how architectural biases and representational depth influence the distinguishability of generative image sources.

1.1 GAN vs Diffusion: Key Differences

- **Training Mechanism:** GANs employ adversarial training between a generator and a discriminator, while diffusion models iteratively learn to denoise data from Gaussian noise.
- **Artifacts:** GANs tend to produce high-frequency or periodic artifacts caused by upsampling and deconvolution. Diffusion models, in contrast, generate smoother and more natural structures, often preserving semantic consistency.

- **Feature Representation:** CNN-based models like ResNet and Xception are more sensitive to local texture artifacts typical of GANs, whereas ViT is more responsive to semantic and global cues, revealing differences in object-level coherence typical of diffusion-based images.

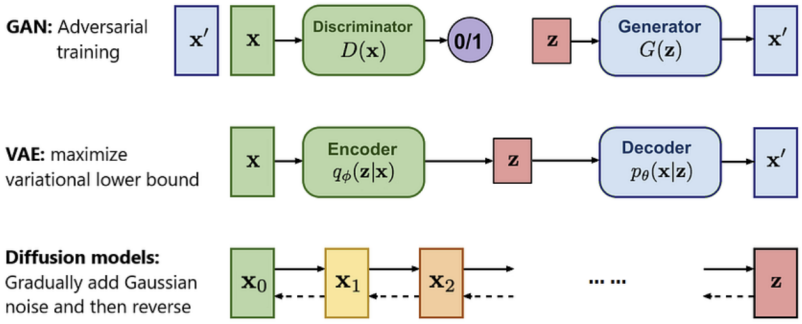

Figure 1: Comparison between images generated by **VAE**, **GAN**, and **Diffusion** models across architectures. The visual differences illustrate distinct reconstruction styles and consistency levels among the generative families.

2 Dataset and Generators

Our dataset includes both **GAN-based** and **Diffusion-based** models:

- GAN: BigGAN.
- Diffusion: ADM, GLIDE, Stable Diffusion (SDv5), VQDM, WuKong, and Midjourney.

The dataset follows a GenImage-like structure, where each subset represents images generated by a specific model. In addition, it includes **Real-world (REAL)** and **Forged (FAKE)** samples derived from the **AutoSplice dataset** for authenticity analysis. Throughout this work, the terms **REAL** and **FAKE** specifically refer to genuine and manipulated images from this dataset, respectively.

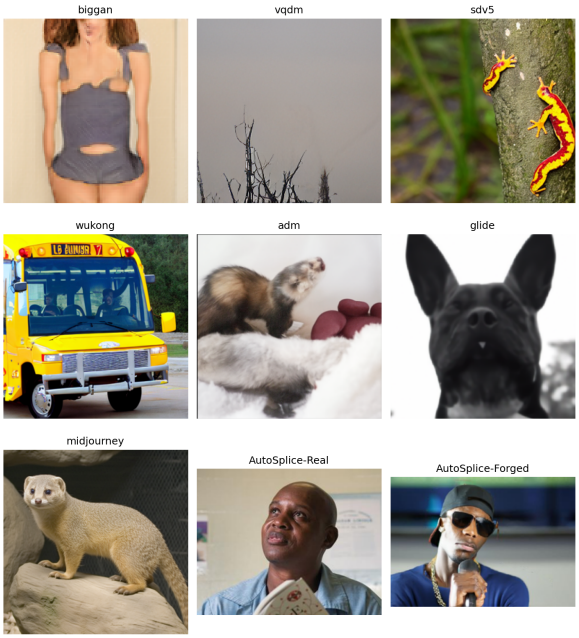

Figure 2: Sample images from each generator and the AutoSplice dataset.

3 Architectures and Feature Layers

Each architecture used in this study contributes complementary representational properties:

- **ResNet-50:** Employs residual connections to stabilize gradient flow, enabling deep hierarchical feature extraction. We extract representations from layers C2, C3, and C4.
- **Xception:** Uses depthwise separable convolutions that better preserve spatial information while reducing redundancy, allowing more precise detection of local anomalies. We focus on stages 1, 2, and 3.
- **ViT:** Divides the input image into patches (e.g., 16x16), encodes them into tokens, and processes them via multi-head self-attention, allowing global modeling of relationships between image regions.

3.1 Feature Map Comparison Across Architectures

Overview. To understand how early layers represent forgeries, we visualize stage-1 feature maps from three backbones: ViT-Base, Xception, and ResNet-50.

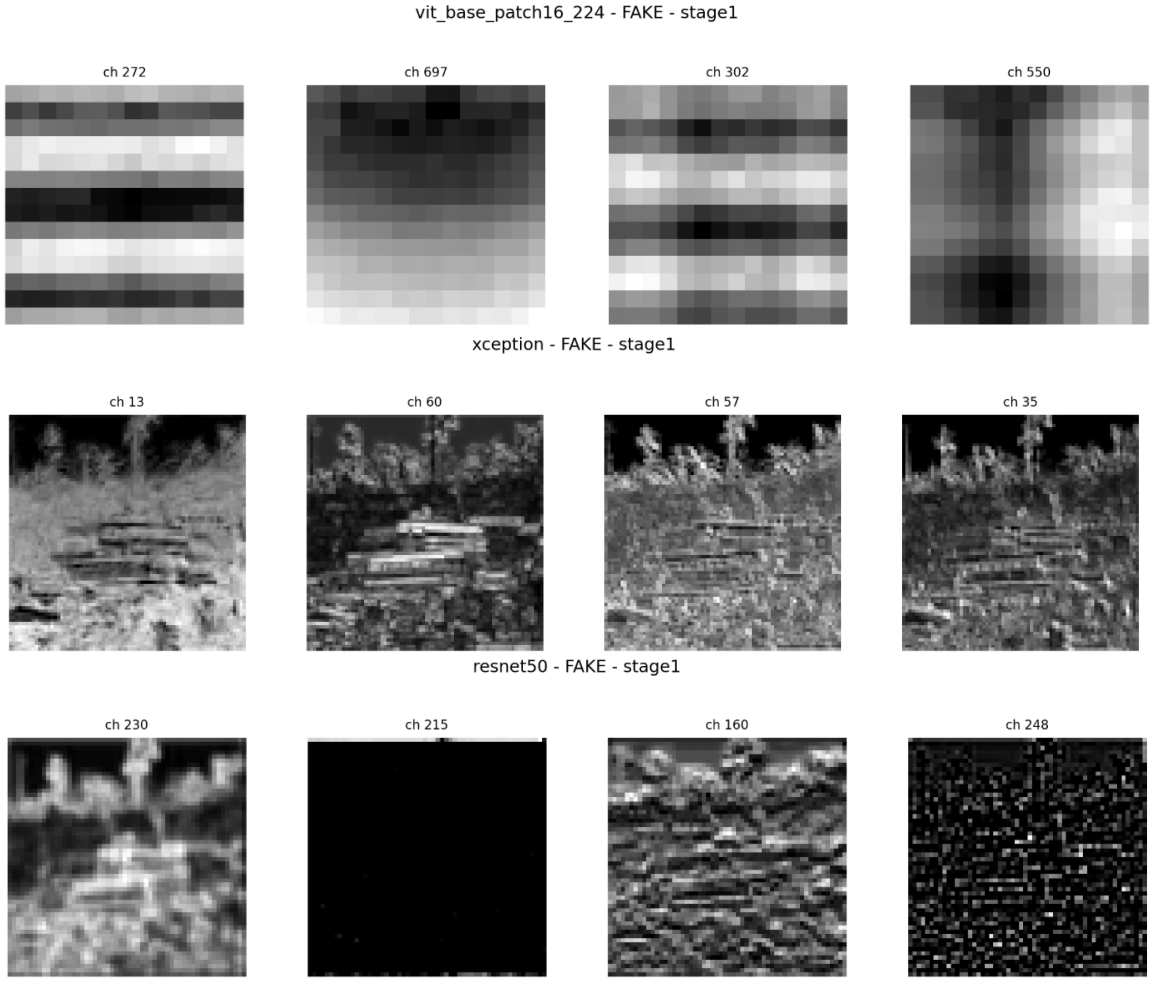

Figure 3: Stage-1 feature maps for a forged image across **ViT-Base**, **Xception**, and **ResNet-50**. ViT displays coarse, patch-aligned activations, while Xception and ResNet-50 capture local edges and textures, exposing fine-grained spatial irregularities typical of manipulations.

4 Experimental Framework and Studies

4.1 Study A: Embedding Space of Generators

Goal. Analyze the embedding distributions of multiple generative models by projecting their representations into a shared 2D latent space (t-SNE or PCA). The aim is to assess clustering, overlap, and separability among generator families.

Setup. We use the **Xception** backbone to extract global embeddings from generated images. For each generator, **100 samples** are randomly selected to ensure balanced comparison across model families.

All-Models View

The visualization below displays all generators together in a shared latent space. Each color represents a distinct generator family, revealing their structural and stylistic relations.

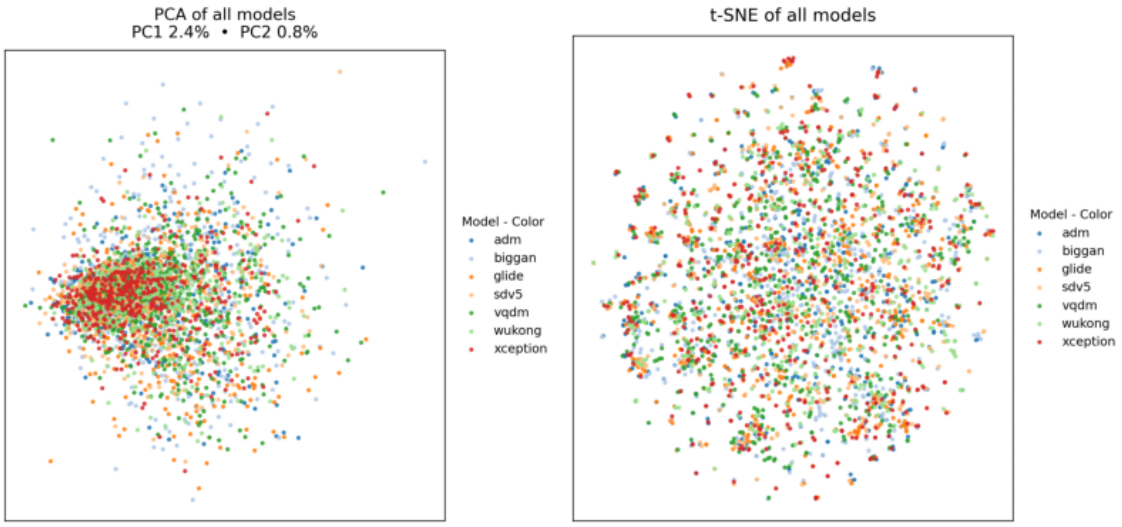

Figure 4: Embeddings of all generator families projected into a common 2D latent space. Colors denote models (**BigGAN**, **ADM**, **GLIDE**, **SDv5**, **VQDM**, **WuKong**, **Midjourney**). The overlap between clusters suggests shared high-level image representations across different generative mechanisms.

Pairwise Comparisons

To further explore model relationships, we visualize embeddings for generator pairs using three backbone types. Each comparison shows how different architectures encode and separate generation styles.

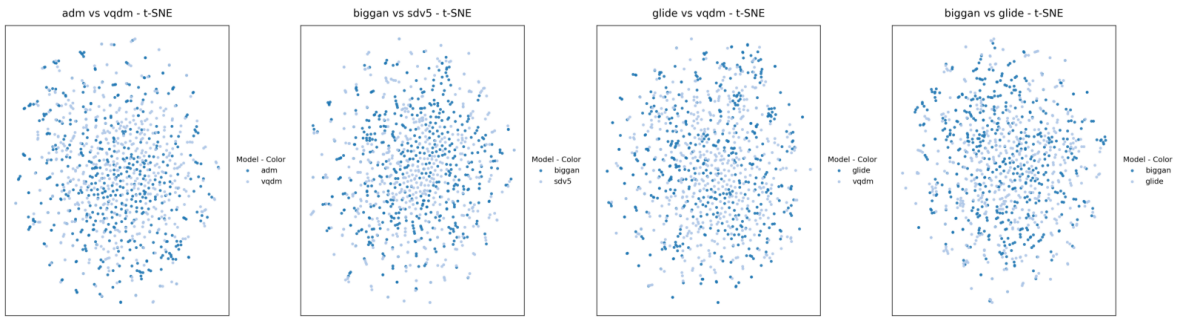

Figure 5: t-SNE visualizations of **Xception** embeddings for selected generator pairs. The strong overlap between clusters reflects limited separability across generators in global embedding space.

Xception Backbone.

Figure 6: t-SNE visualizations of **ResNet-50** embeddings. Similar to Xception, generator clusters are largely intertwined, suggesting that CNN-based embeddings capture shared visual priors.

ResNet-50 Backbone.

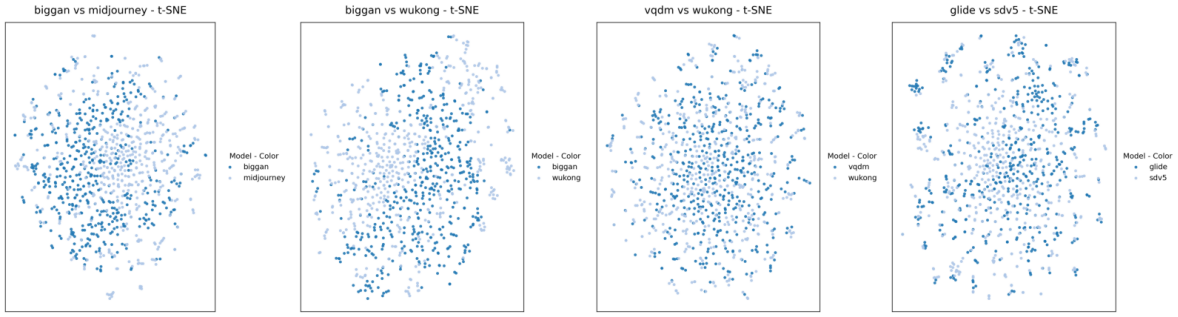

Figure 7: t-SNE visualizations of **ViT-Base** embeddings for the same generator pairs. Compared to CNNs, ViT exhibits higher inter-model separation, particularly between BigGAN and GLIDE, indicating that transformer-based representations encode more distinctive global and structural cues.

ViT-Base Backbone.

4.2 Study B: Feature Map Space

4.2.1 Feature Similarity Analysis Across Xception Stages Using CKA

To better understand how feature representations evolve within the Xception backbone, we conducted a **Centered Kernel Alignment (CKA)** analysis across different stages of the network. The analysis compares the representational similarity between feature maps extracted from various image sets – including *BigGAN, ADM, SDv5, VQDM, Glide, Midjourney, WuKong*, as well as **REAL** and **FAKE** samples. CKA serves as a robust, scale and rotation-invariant measure of similarity between internal neural representations, allowing a fair comparison of how distinct generators are encoded at different network depths.

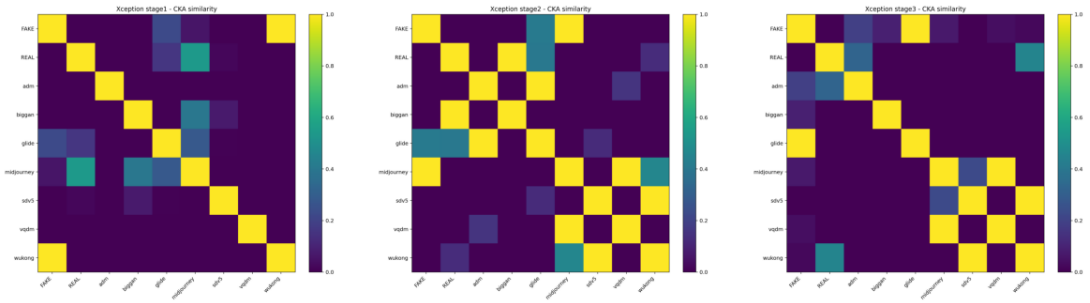

Figure 8: **CKA similarity across feature stages of Xception.** Each matrix visualizes the pairwise representational similarity between image sets at: (left) **Stage 1** – early convolutional features, (middle) **Stage 2** – mid-level representations, and (right) **Stage 3** – higher-level semantic features. Brighter regions (yellow) denote stronger representational similarity, while darker regions (purple) indicate greater dissimilarity. The gradual changes from Stage 1 to Stage 3 reveal that, in the **first stage**, there is almost no overlap or similarity between the different generator families, indicating that early convolutional features already capture highly discriminative patterns unique to each model type. In contrast, **later stages** exhibit higher similarity and partial overlap across generators, suggesting that deeper representations become more abstract and less model-specific. This pattern implies that the **first feature stage** may contain the most relevant information for distinguishing between image sources.

4.2.2 Discriminative Analysis Between Feature Maps and Embeddings

This study aims to compare the discriminative structure of **feature maps** extracted from intermediate Xception layers with the corresponding **global embeddings** obtained after the final pooling layer. We focus on assessing how well generative models separate in these two spaces.

Findings. A clear separation emerges in the **feature map space** (Xception stage1), where local texture and structural cues enable almost perfect linear discrimination between generators. In contrast, the **embedding space** exhibits considerable overlap, suggesting that high-level compression discards essential forgery-related information.

Quantitatively, a linear classifier achieves:

- **Feature map space (stage1)** – accuracy: **0.99**
- **Embedding space** – accuracy: **0.80**

This gap confirms that **early convolutional representations** (stage1 feature maps) are more effective at preserving generator-specific artifacts than the globally averaged embeddings, which tend to discard fine-grained spatial cues.

Visualization. The following t-SNE plots illustrate this distinction between the two representational spaces.

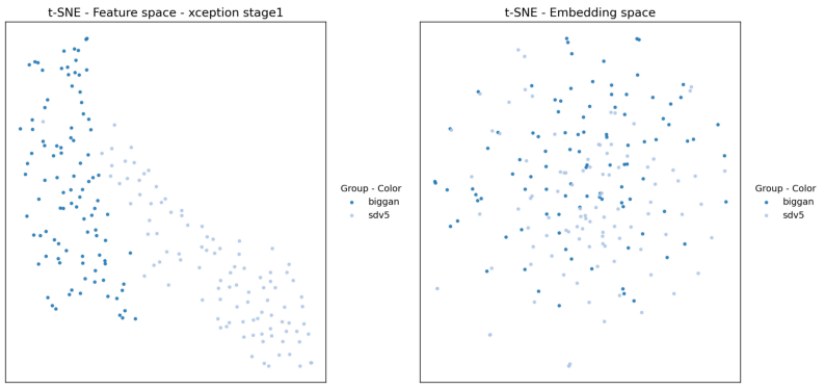

Figure 9: t-SNE comparison between the **feature space** (left) and the **embedding space** (right) for **Xception**, evaluated on **BigGAN** and **SDv5** images. The feature map space shows more distinct clustering patterns, while the embedding space appears more entangled, suggesting that spatial features may better reflect differences between generative mechanisms.

Observation. In the feature space projections, **GAN-based models** (e.g., BigGAN, StyleGAN) tend to appear farther from **diffusion-based models** (e.g., SDv5, ADM, Glide). This pattern suggests that GAN-generated images may produce somewhat more distinct feature representations, possibly reflecting their texture-oriented synthesis process, whereas diffusion-based models exhibit smoother and more globally coherent patterns. However, the observed separation should be interpreted cautiously, as overlaps and local variations still occur within both model families.

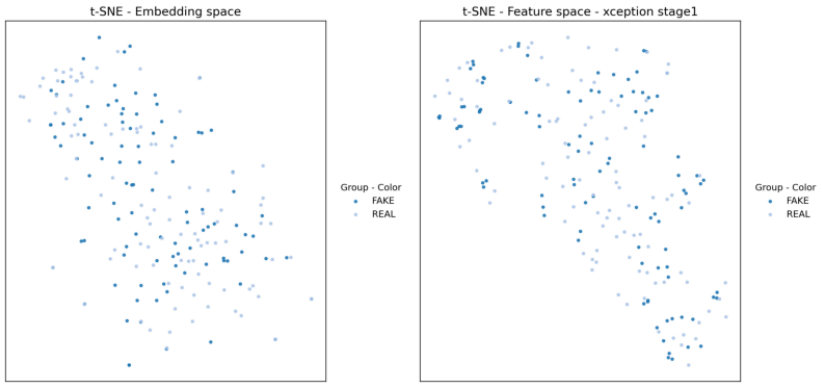

Figure 10: Comparison between the **embedding space** (left) and the **feature map space** (right) for **REAL** and **FAKE** images using the Xception model. The **FAKE** images in this analysis correspond to *partially forged samples*, where only specific regions of the image were generated or edited using GenAI methods, while the remaining areas remain authentic. In the embedding space, there is a strong overlap between REAL and FAKE samples (*logistic regression accuracy: 0.515*), whereas the feature map space shows slightly improved separation (*accuracy: 0.655*; with XGBoost: **0.745**). This moderate separability likely stems from the **mixed nature of forged images** - where genuine and synthetic regions coexist - resulting in shared textural and structural characteristics across both classes.

4.3 Study C: Combined Representation

Goal We tested whether concatenating z-scored embeddings and feature vectors could enhance class separability compared to using either representation alone.

Overall Findings Across all experiments, feature maps provided notably stronger discriminative power than embeddings. For REAL vs. FAKE images, feature-based representations achieved the highest separability, while concatenation offered only a minor improvement. In the Glide vs. Mid-

journey comparison, feature maps again outperformed embeddings substantially, with concatenation yielding a small decrease compared to pure feature-based separation.

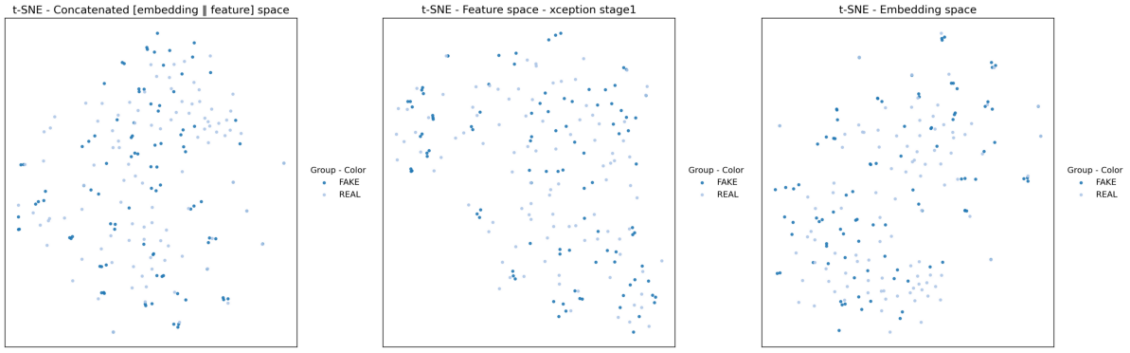

Figure 11: t-SNE projection for embeddings, feature maps, and concatenated space (REAL vs. FAKE).

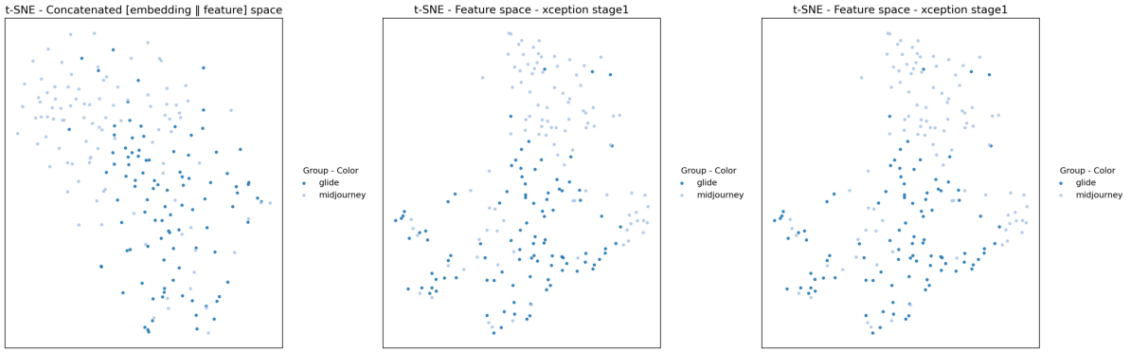

Figure 12: t-SNE projection for embeddings, feature maps, and concatenated space (Glide vs. Midjourney).

Quantitative Summary. Table 1 reports the linear (LR) and nonlinear (XGBoost) classification accuracies across the **embedding**, **feature map**, and **combined** spaces for two representative comparisons: REAL vs. FAKE and Glide vs. Midjourney.

Comparison	Space	Backbone + Layer	LR Acc	XGBoost Acc
REAL vs. FAKE	Embeddings	`xception`	0.475	0.575
	Feature maps	`xception stage1`	**0.615**	**0.765**
	Combined	`xception stage1 + emb`	0.580	0.650
Glide vs. Midjourney	Embeddings	`xception`	0.645	0.680
	Feature maps	`xception stage1`	**0.940**	**0.905**
	Combined	`xception stage1 + emb`	0.860	0.895

Table 1: Linear (LR) and nonlinear (XGBoost) separability across embeddings, feature maps, and combined spaces. Feature maps consistently yield higher separability, suggesting that spatial-level representations preserve more discriminative cues than global embeddings.

Interpretation. The quantitative results reveal a clear advantage for **feature-based representations** over global embeddings across both global (REAL vs. FAKE) and inter-model (Glide vs. Midjourney) comparisons. For the global classification task, feature maps from `xception stage1` achieved accuracies of **0.615** (LR) and **0.765** (XGBoost), outperforming the embedding space by approximately **15–20%**. In the inter-model comparison, the difference was even more pronounced: feature maps reached **0.940** (LR) and **0.905** (XGBoost), whereas embeddings achieved only **0.645** and **0.680**, respectively. The combined representation (`stage1 + embedding`) showed only marginal improvement,

with accuracies of **0.580–0.650** for REAL vs. FAKE and **0.860–0.895** for Glide vs. Midjourney. These results suggest that most discriminative information present in the embedding space is already encoded within the spatially rich feature maps, which retain finer textural and local structural details crucial for distinguishing between different generative models.

Discussion and Future Directions. These findings highlight the importance of exploring **intermediate convolutional features** rather than relying solely on compact embeddings when analyzing or detecting generative image models. Since feature maps preserve richer spatial and textural cues, they appear more sensitive to generator-specific artifacts, such as local inconsistencies or subtle frequency patterns that are often smoothed out through global pooling.

The relatively small performance gap observed in the **REAL vs. FAKE(forged images)** comparison suggests that, while current feature-based representations capture some discriminative cues, additional processing steps may be required to achieve reliable differentiation between authentic and synthetic content. Future research should therefore explore more **localized analysis strategies**, such as pixel- or region-level prediction, attention-based artifact localization, or patch-wise feature aggregation. Such approaches could help uncover fine-grained structural discrepancies that are not globally separable in embedding space.

Future work could also investigate **alternative convolutional formulations**, such as **Wavelet-Based Convolutions (WTConv)** or other multi-scale operators that explicitly capture both spatial and frequency-domain information. Integrating frequency-aware layers into the backbone architecture may enhance the model's ability to separate closely related generators and generalize across unseen architectures. Moreover, examining representations from **deeper or later stages** could shed light on how semantic abstraction interacts with artifact-level cues and whether hierarchical fusion of early and late features provides complementary benefits.

Overall, the results suggest that understanding and detecting generative model artifacts may require looking *beyond the embedding space*. While embeddings remain useful for global semantic reasoning, **the discriminative structure of feature maps - especially when analyzed at multiple scales or localized regions : offers a more informative and robust foundation** for distinguishing between images originating from different generative families, including **GANs, VAEs, and Diffusion models**.