Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2026 (44) 17** : 1–32.
ISSN-0037-8712
doi:10.5269/bspm.82398

Outlier Tokens Drive Attention Patterns in Vision Transformers *

S. Nagini, Karnam Akhil, Mallupeddi Vamsi Krishna, Pathi Sairoop Teja, Swapnika Chowdary Thanikonda

ABSTRACT: The Vision Transformers are widely used in computer vision, as they can capture global image information. However, there is a persistent problem in all these ViT models: a certain number of tokens in their feature representations show abnormally high values at the background level regions. These tokens capture populated global information while losing essential local details. This leads to attention maps with sharp peaks that do not correspond to meaningful parts of the image. This problem hugely affects the spatial understanding of the model and impacts the performance of tasks needing accurate region-based reasoning. It shows up across supervised, self-supervised, and text-supervised settings and also within small ViT models, showing a clear gap in current systems. This paper identifies this gap and proposes improved transformer-based models that reduce the effect of these outlier tokens and help in maintaining proper spatial information. Enhancements made specify the stabilizing of token behavior and supporting better attention distribution across the image without relying heavily on background regions. Experimental results clearly show the reduction in abnormal token activity with smoother and more meaningful attention maps. Improved models also showed better performance in tasks dependent on spatial accuracy. The motive of this work is to make ViTs more reliable, interpretable, and consistent. By improving spatial reasoning and avoiding misleading attention patterns, the proposed models will support stronger and more trustworthy visual understanding in real-world applications.

Keywords: Vision transformers, attention, registered tokens, feature norm distribution, spatial reasoning.

Contents

* The project is partially supported by institutional resources of the college.
2020 *Mathematics Subject Classification:* 68T07, 68T99.
Submitted March 04, 2026. Published June 19, 2026.

1. Introduction

ViTs denote a fundamental change in the area of Computer Vision, providing a strong alternative to previously dominant convolution-based architectures. This architecture can model long-range dependencies and global relationships across an image using self-attention mechanisms, hence providing a richer and more flexible understanding of the content in the images. They have placed themselves at the heart of academic research and industrial applications with their success on image classification, semantic segmentation, and object discovery tasks. Despite such effectiveness, much of their internal working remains unclear, which shows the need to understand these models and interpret information.

Recent works have demonstrated an unappealing understandability problem in transformer-based visual models, which is that of the emergence of unexpected high-norm tokens. These tokens often show up in areas with little semantic value, such as backgrounds or uniformly colored patches. The emergent high-norm tokens from these regions could still influence the model's representations despite their little to no contribution in the visual information. This gives way to pressing questions about why such artifacts arise, how they move through the network, and how they affect the predictions made by the transformers.

These artifacts are of prime importance because they pose direct challenges to the robustness and transparency of transformer-based vision systems. When high-norm tokens start to dominate or alter the internal attention patterns in a model, they may inadvertently shift how the model perceives an image, potentially misleading the model to make unwarranted interpretations or generally unreliable model outputs. Therefore, the origin, behavior, and impact of such anomalous tokens need to be analyzed as an indispensable stage toward better interpretability of modern deep learning models.

This work primarily focuses on the detection and characterization of these artifacts in the ViT's. The study reproduces empirical analyses from existing research in order to validate the presence and significance of these tokens. Techniques such as norm-based thresholding, attention map visualization, token similarity metrics, and linear probing are used in a systematic way to show how high-norm tokens

form, how they interact with other components of the model, and what role they play at inference time. It underlines not only the existence of such artifacts but also their possible effect on model reasoning and stability. This work calls for more transparent frameworks for evaluation and further study in exploring the interpretability of transformer-based vision architectures so that their wide adoption is complemented with clear insight into the internal decision-making processes.

Table 1 lists, Acronyms and Mathematical Equations mentioned in this study.

Table 1: Summary of Acronyms, Symbols, and Notations

Acronym / Symbol	**Description**
	Acronyms
ViT	Vision Transformer
CNN	Convolutional Neural Network
DeiT	Data-efficient Image Transformer
DINO	Self-distillation with No labels (e.g., DINOv2)
CLIP	Contrastive Language–Image Pre-training
OpenCLIP	Open source implementation of CLIP
LOST	Localizing Objects with Self-Attention
CLS	Classification Token (Global representation)
RGB	Red, Green, Blue color model
SORTS	Semantic Outlier Regulated Token Stabilization
	Mathematical Notations
$\mathcal{D}$	CIFAR-10 Dataset (60,000 images)
$\mathbf{X}$	Input image (32×32 resolution)
$\mathbf{Z}(L)$	Feature embeddings at the final transformer layer L
$\mathbf{z}_{cls}$	Embedding vector of the classification token
$\mathbf{z}_{patch}$	Embedding vector of a spatial patch token
$\|\cdot\|_2$	ℓ_2 norm (Euclidean magnitude) of a token vector
$\mu(l)$	Average spatial norm at layer l
$\text{sim}(u, v)$	Cosine similarity between vectors u and v
$\mathcal{N}(i)$	Set of spatial neighbors for a specific patch i
acc	Classification accuracy

2. Literature Review

Recent works have flagged that one of the main issues with Vision Transformers lies in their internal token representations. Darcet et al. [1] report a known issue with the appearance of so-called "register tokens" in the output representations, in that even background image areas show abnormally high activation values. These tokens retain an excessive amount of global information at the cost of local and positional information, which causes sharp misleading peaks in the attention maps. Subsequently, Bach et al. [2] identified the exact same behavior for smaller ViT models, proving that the issue is model scale-independent and also orthogonal to a number of training settings. The development of the ViTs started with the seminal work by Dosovitskiy [3], which introduced the patch-based transformer architecture for image recognition.

This changed the course because it showed that transformers can outperform convolutional networks when trained on huge datasets. Building on this, Touvron et al. developed the DeiT framework [4] and then showed that ViTs also perform well with small training budgets using distillation strategies. Very recently, DeiT III [5] introduced strong training procedures that increased the accuracy without major architectural changes and confirmed the importance of stable token behavior at training. Further

extension of the role of ViTs was explored with self-supervised learning advancements. For instance, DINOv2 showed that large self-supervised ViT models can learn robust and transferable visual features without labeled data. Even these powerful models had attention patterns that were sometimes unstable and extremely sharp. The extension to multimodal learning was realized when OpenCLIP or the original CLIP model by Radford et al. aligned images and text with each other using contrastive training. While effective for semantic representation, these models also generated attention maps dominated by global patterns rather than grounded spatial detail, thus making them susceptible to the same issues of outlier tokens observed in earlier studies. Siméoni et al. [9] showed that self-supervised ViTs are able to localize objects without labels, based simply on attention behaviour.

This localization is sometimes unreliable due to the same high-norm token effects that distort spatial reasoning. Finally, Raghu et al. [10] compared ViTs with CNNs and concluded that ViTs lack the strong inductive biases of convolutional networks. They observed that ViTs rely heavily on global information aggregation which may contribute to unstable token norms and weaken their ability to focus on important spatial regions. There is a consistent pattern coming out of these works: while ViTs achieve remarkable global reasoning, their local representations are unstable and not very meaningful. Outlier tokens distort attention, hamper interpretability, and decrease performance for region-based tasks. Motivated by the above limitations, this work will develop transformer models that mitigate the influence of outlier tokens, produce better spatial stability, and have much smoother attention maps semantically aligned. This paper is intended to make ViTs more reliable, interpretable, and effective in doing real-world spatial reasoning tasks through the regularization of token behaviors and rectification of attention distribution.

3. Dataset Description

The dataset we used contains information about low-resolution natural images used for benchmarking image classification models for computer vision research. The CIFAR-10 dataset includes diverse object categories and is suitable for tasks such as image classification, model interpretability analysis, and evaluation of ViT behavior.

3.1. Dataset Overview

- **Total Images:** 60,000 images in RGB format
- **Image Resolution:** 32×32 pixels
- **Train Split:** 50,000 images
- **Test Split:** 10,000 images
- **Classes:** It includes 10 object categories, including airplanes, automobiles, birds, cats, dogs, and ships
- **Labels:** Image-level category labels (no pixel-level annotations)

3.2. Image Categories

This dataset includes 10 distinct classes:

- Airplane
- Automobile
- Bird
- Cat
- Deer
- Dog
- Frog

- Horse
- Ship
- Truck

3.3. Usage in This Study

- **Quantitative Analysis:**
  - The CIFAR-10 dataset measures statistical properties such as:
    - Attention distribution patterns
    - Token-norm variance
    - Layer-wise activations in Vision Transformers
- **Qualitative Analysis:**
  - A subset of 100 representative images are utilized for:
    - Attention map visualization
    - Identifying artifact tokens
    - Analyzing feature-norm behavior

3.4. Dataset Characteristics

- **Low Resolution Nature:** It is designed to reveal the architectural weaknesses because of limited spatial detail.
- **High Diversity:** Significant variation within each class.
- **Benchmark Status:** It is used for evaluating model behavior and interpretability.
- **Task Type:** Image classification only (no segmentation labels).

3.5. Suitability

CIFAR-10 is best for:

- Studying artifact emergence in Vision Transformers.
- Analyzing how low-resolution images influence feature extraction.
- Evaluating classification performance across different architectures.

4. Existing System

The contemporary Vision Transformer models - like DeiT-III, DINOv2, and OpenCLIP - achieve outstanding performance on the current large-scale image classification benchmarks by using self-attention mechanisms for the encoding of the global context of an image; they generally perform at least as well as the traditional Convolutional Networks (CNNs) on those benchmarking tests. Recently, it was demonstrated that the feature representations produced by these models contain systematic flaws. Specifically, high-norm background tokens consistently appear during inference time and generate attention maps where the peaks of interest lie in visually uninteresting areas of the image with little information content.

The effects mentioned above destroy the spatial interpretability of the ViTs and affect tasks directly dependent on the coherent distribution of attentions among patches (e.g., object discovery). As such, LOST and related approaches which depend highly on the attention localization of images, will be negatively impacted in the presence of peaks in the background driven by the same systematic flaws in ViTs. Currently, the identification and analysis of the problems described above are performed using ad-hoc, model-dependent methods that include token norm distributions, visualization of attention heatmaps, and structural similarity measures between patch embeddings.

Therefore, the majority of the methods available to identify and analyze the types of errors described above are largely model specific, and therefore, methodology is not consistent. Thus, when different researchers investigate different ViT architectures they have to create their own methodologies to identify the token(s) responsible for the artifacts of each architecture and determine what the identified token means. This variability creates the issue of inconsistencies in study evaluation due to the lack of a common diagnostic tool for identifying and interpreting such artifacts in both supervised, self-supervised, and text-supervised versions of ViTs. In addition, this lack of consistency severely hinders research into the underlying causes of such artifacts and, consequently, the development of robust and interpretable transformers for vision.

5. Proposed System-SORTS

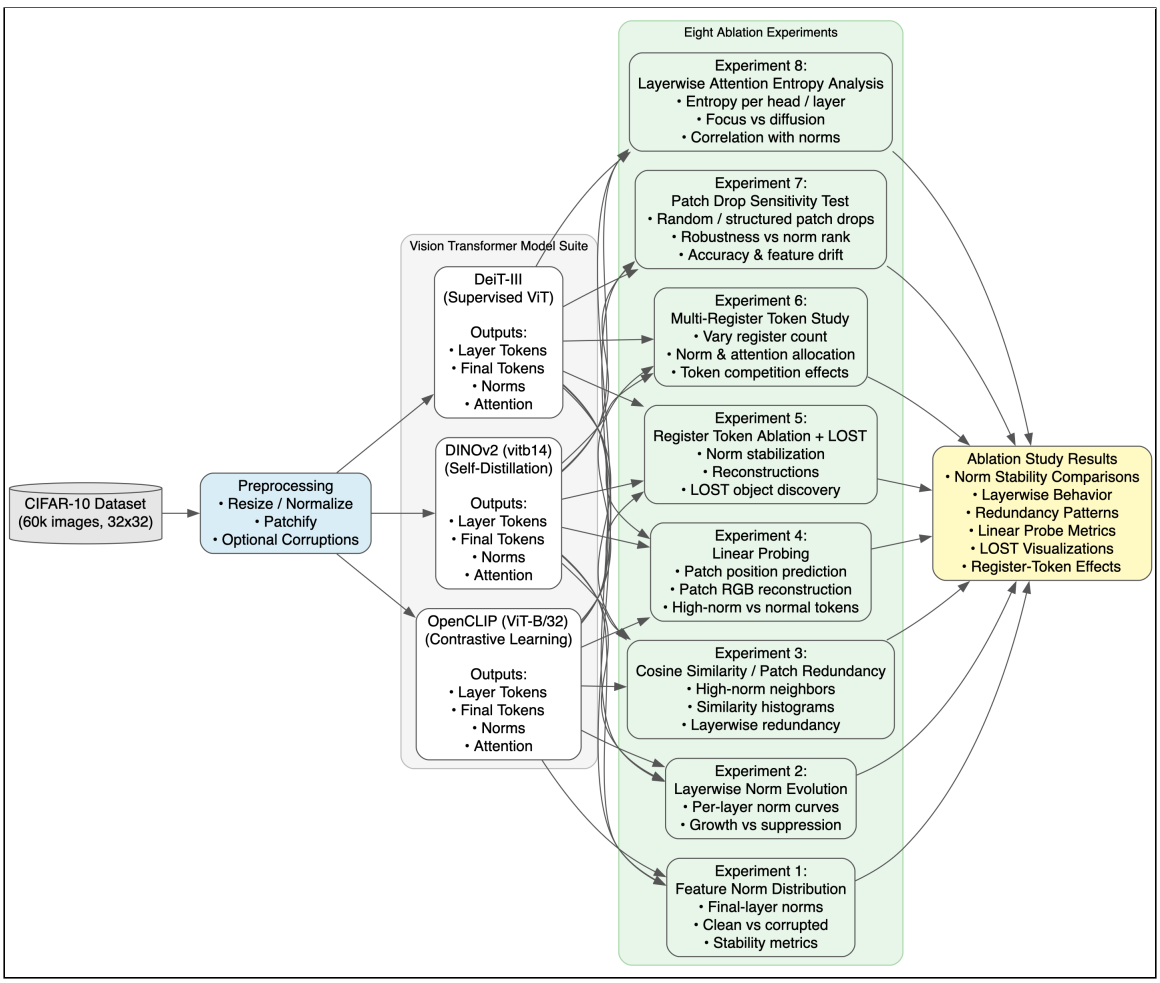


Figure 1: Comprehensive Architecture for Vision Transformer Representation Analysis

SORTS - Semantic Outlier Regulated Token Stabilization

The above architecture diagram of the proposed ablation study represents the entire workflow employed in the shared experimental framework to evaluate the performance of different Vision Transformer models. CIFAR-10 dataset consisting of 60,000 low-resolution images forms the first step in this regard. These images provide the input for all the analyses. First of all, these pictures undergo a preprocessing where they are resized and normalized and then split into tokens of some particular size. At this stage,

optional corruption transformations are also utilized for experiments which need clean comparisons. In this way, it is ensured that the inputs to all the downstream models are the same and standardized.

After these preprocessing steps, data enters the Vision Transformer Model Suite, which contains the following three pre-trained architectures: DeiT-III, OpenCLIP (ViT-B/32), and DINOv2 (ViT-B/14). Each model outputs four major pieces of information: layer tokens, final-layer tokens, token norms, and attention maps. These capture progressively deeper levels of representation and serve as the basis for the various tasks in the study.

The processed representations from all three models are fed into five ablation experiments that investigate different aspects of model stability and feature behavior:

1. **Experiment 1:** Feature-norm distribution evaluation under clean and corrupted settings.

2. **Experiment 2:** Investigation of how the magnitude of activations changes across the transformer layers.

3. **Experiment 3:** Examination of the spatial redundancy by cosine similarity of neighboring patches; this is helpful to understand where high-norm activations come from.

4. **Experiment 4:** Linear probing on how much positional or pixel-level information high-norm tokens preserve.

5. **Experiment 5:** Analysis of the impact of adding a register token on norm stability, reconstruction quality, and LOST-style object-discovery behavior.

6. **Experiment 6:** Multi-Register Token Study This experiment explores the marginal utility of the addition of multiple register tokens to the ViT architecture to locate whether more is good in this case. It concludes that the primary stability gain comes when there are one register token, and i.e. adding more tokens leads to diminishing returns and possible led to redundancy.

7. **Experiment 7:** Patch Drop Sensitivity Study This experiment is an experiment to detect the robustness of a model by randomly dropping all input patches, and making an empirical measurement of the subsequent drop in classification accuracy and token norm stability. It shows that register tokens can serve as structural stabilizers, to a large extent causing a reduction in performance and avoiding the collapse of representation even in cases where large chunks of data are missing.

8. **Experiment 8:** Layerwise Attention Entropy Analysis This experiment quantifies the entropy of attention maps in the network depth in order to determine the degree of focused or diffuse attention of the model on each layer. It demonstrates that register tokens enable the model to retain high entropy (exploration) in the lower levels and cause point-focused object attention (semantic collapse) in the higher levels.

The results of the five experiments gathered to form the ablation study are summarized in the above. In all essence, this summarizes all key features within these models. It is mainly based on the norm stability values, redundancy observed within the models, the activation patterns shown by the deeper layers, linear probe results, and improvements seen in the models due to the register token augmentation. All the results together give a clear view regarding how the training models influence the qualities of the Vision Transformers.

6. Algorithms

Description: This experiment quantifies the presence of outlier tokens by computing the $L2$ norm of patch embeddings from the final transformer layer.

Algorithm 1 Feature Norm Distribution Analysis

Formula: The $L2$ norm of a patch token z_i is defined as:

$$\|z_i\|_2 = \sqrt{\sum_{j=1}^{D} (z_{i,j})^2}$$

Input: Dataset $\mathcal{D}$, Pre-trained ViT Model f_θ
Output: Distribution histogram $\mathcal{H}$ of scalar norms
Initialize an empty list $\mathcal{S} \leftarrow \emptyset$ to store norms.
for each image $X_k \in \mathcal{D}$ **do**
    Propagate X_k through f_θ to obtain final layer patches $Z^{(L)}$.
    Discard the classification token $z_{cls}^{(L)}$.
    for each spatial patch token $z_i^{(L)} \in Z^{(L)}$ **do**
        Compute scalar norm: $v_i = \|z_i^{(L)}\|_2$.
        Append v_i to $\mathcal{S}$.
    end for
end for
Construct histogram $\mathcal{H}$ from $\mathcal{S}$ to visualize distribution peaks (e.g., bimodal distribution in DeiT-III).

Description: This experiment tracks the propagation of activation magnitudes across the network depth to identify where instability arises.

Algorithm 2 Layerwise Norm Evolution

Formula: The average spatial norm for layer l, denoted as $\mu^{(l)}$, is calculated as:

$$\mu^{(l)} = \frac{1}{N} \sum_{i=1}^{N} \|z_i^{(l)}\|_2$$

Input: Single Test Image X, Pre-trained ViT Model f_θ with layers $l = 1 \dots L$
Output: Set of average norms $\mathcal{M} = \{\mu^{(1)}, \dots, \mu^{(L)}\}$
Initialize $\mathcal{M} \leftarrow \emptyset$.
for each transformer layer l from 1 to L **do**
    Extract the sequence of token embeddings $Z^{(l)}$ output.
end for

Description: This experiment measures the redundancy of patch information by calculating the cosine similarity between a patch token and its spatial neighbors. High similarity suggests the token is repeating information (redundant) rather than encoding unique local details.

Algorithm 3 Cosine Similarity and Patch Redundancy

Formula: The cosine similarity between two feature vectors u and v is:

$$\text{sim}(u, v) = \frac{u \cdot v}{\|u\|_2 \|v\|_2}$$

Input: Feature Map $Z^{(L)} \in \mathbb{R}^{H \times W \times D}$ (Spatial tokens only)
Output: Redundancy Score Map $R \in \mathbb{R}^{H \times W}$
for each spatial location (i, j) in $H \times W$ **do**
    Extract token $z_{i,j}$.
    Identify set of 8-nearest neighbors $\mathcal{N}_{i,j}$.
    Initialize neighbor similarity sum $S_{local} = 0$.
    for each neighbor $z_n \in \mathcal{N}_{i,j}$ **do**
        Compute $s = \text{sim}(z_{i,j}, z_n)$.
        $S_{local} \leftarrow S_{local} + s$.
    end for
    Compute average similarity (Redundancy Score):

$$R_{i,j} = \frac{1}{|\mathcal{N}_{i,j}|} S_{local}$$

end for
Analysis:
Compare distribution of $R_{i,j}$ for High-Norm tokens vs. Normal tokens.
(High $R_{i,j}$ implies the token is a "background" or "outlier" with low local information).

Description: This experiment verifies the information content of tokens. We train linear probes to recover the original pixel color (local info) and the (x, y) position (structural info) from the embeddings.

Algorithm 4 Linear Probing: Position and Pixel Reconstruction

Formulas:
Position Probe ($y_{pos} \in \mathbb{R}^2$): $\hat{y}_{pos} = W_{pos} z_i + b_{pos}$
Pixel Probe ($y_{rgb} \in \mathbb{R}^3$): $\hat{y}_{rgb} = W_{rgb} z_i + b_{rgb}$
Input: Dataset $\mathcal{D}$, Feature Embeddings Z, Ground Truth Pixels Y_{rgb} and Positions Y_{pos}
Output: MSE_{pixel} and $Accuracy_{pos}$
Phase 1: Training
Train W_{pos}, b_{pos} to minimize $\|\hat{y}_{pos} - y_{pos}\|^2$.
Train W_{rgb}, b_{rgb} to minimize $\|\hat{y}_{rgb} - y_{rgb}\|^2$.
Phase 2: Evaluation
Split tokens into two sets: $\mathcal{S}_{high}$ (High-Norm) and $\mathcal{S}_{normal}$ (Normal).
for each set $\mathcal{S} \in \{\mathcal{S}_{high}, \mathcal{S}_{normal}\}$ **do**
    Compute Mean Squared Error for pixels:

$$MSE_{pixel}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{z \in \mathcal{S}} (\hat{y}_{rgb} - y_{rgb})^2$$

    Compute Accuracy/Error for position coordinates.
end for
Comparison:
If $MSE_{pixel}(\mathcal{S}_{high}) \gg MSE_{pixel}(\mathcal{S}_{normal})$, high-norm tokens have discarded local texture information.

Description: This experiment explores how the introduction of learnable "register tokens" absorbs

global information artifacts and checks the recovery of coherent attention maps using the LOST (Localizing Objects with Self-Attention) method.

Algorithm 5 Register Token Ablation and LOST Simulation

1: **Mechanism:** The input sequence would be modified from $[z_{cls}, z_1, \ldots, z_N]$ to $[z_{cls}, z_{reg_1}, \ldots, z_{reg_k}, z_1, \ldots, z_N]$, where...
2: Reconstruct patch pixels using the linear probes from Algorithm 4.
3: Compare visual coherence between f_θ (baseline) and f'_θ (with registers).

7. Results

7.1. Experiment 1 – Feature Norm Distribution Analysis (Ablation Study)

7.1.1. Goal. This experiment analyzes whether different Vision Transformer (ViT) models produce unusually high norm feature tokens. High norm tokens often indicate unstable activations or artifact-like responses. We compare how three pretrained models behave:

- DeiT-III
- DINOv2 (vitb14)
- OpenCLIP (ViT based image encoder)

7.1.2. Methodology.

- For each model, patch embeddings were extracted from the last transformer layer. Only spatial patch tokens were included. Each token embedding was reduced to a single scalar norm, and these values were aggregated across the dataset to form a distribution.
- For DINOv2 and OpenCLIP, we additionally measured the global feature norm for matching clean vs corrupted images to test sensitivity to artifacts.

7.1.3. 1. DeiT-III Analysis.

- The distribution of patch token norms shows two clear peaks. Most tokens fall around 20–25, while another concentrated group appears around 10–12. Across all images, values ranged roughly from 10 to 28.
- This wide spread and the presence of consistent high norm outliers suggest that DeiT-III frequently produces unstable or artifact-like activations. This model had the highest variance and the least stable norm behavior among all three models as shown in Figure 2.

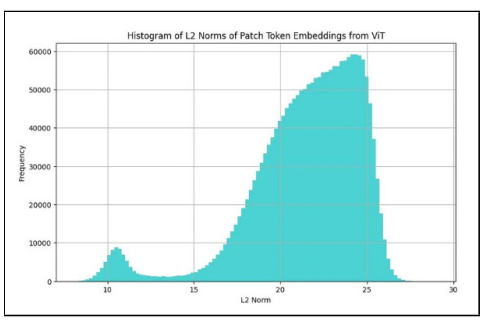


Figure 2: Histogram of L2 Norms of Patch Token Embeddings from ViT (DeiT-III)

7.1.4. 2. DINOv2 (vitb14).

- The global feature norms for clean and corrupted versions of each image were extremely similar. Clean images were mostly between 46 and 49, and corrupted images were almost identical, generally between 44 and 50.
- The distributions of both conditions overlapped almost perfectly, showing very little change even when artifacts were introduced as shown in Figure 3.
- This indicates that DINOv2 has highly stable and invariant representations. Norms remain tightly clustered regardless of perturbations.

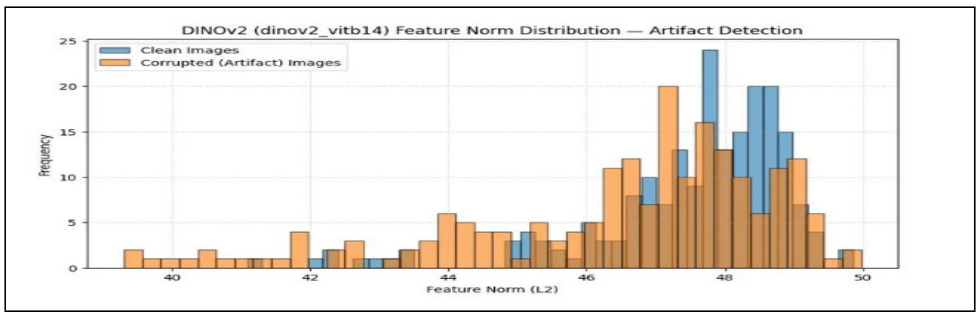


Figure 3: DINOv2 (dinov2_vitb14) Feature Norm Distribution – Artifact Detection

7.1.5. 3. OpenCLIP.

- OpenCLIP's norms were in a much smaller range, roughly 11.0 to 12.7 for clean images and about 11.3 to 12.8 for corrupted ones.
- There was a slight but consistent increase in the corrupted-image norms. The two distributions overlapped heavily, but the corrupted images shifted upward just a bit as shown in Figure 4.
- This means OpenCLIP is moderately sensitive to artifacts. It is more robust than DeiT-III but less invariant than DINOv2.

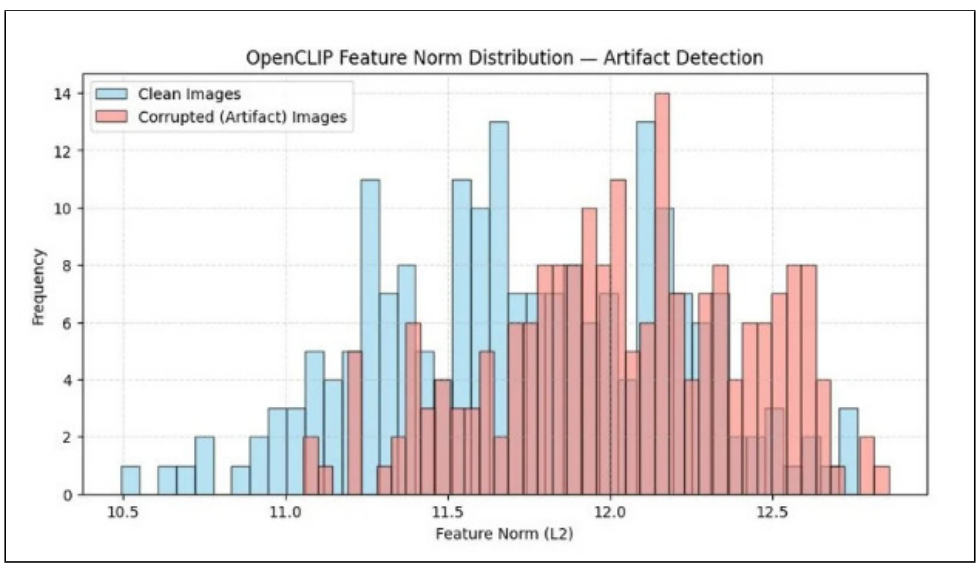


Figure 4: OpenCLIP Feature Norm Distribution – Artifact Detection

7.1.6. Comparative Interpretation. Each model exhibits different behaviour:

- **DeiT-III:** It has highly variable norms, frequent high-norm outliers, and has the most unstable feature distribution.
- **DINOv2:** It is extremely tight and consistent in norms with almost no difference between clean and corrupted images.
- **OpenCLIP:** Overall, it is stable but does show some small increases in norm when artifacts are present.

Table 2: Feature Norm Distribution

Model	Clean Norm Range	Corrupted Norm Range	Behavior
DeiT-III	10–28 (peaks 10–12, 20–25)	–	High variance, unstable outliers
DINOv2 (vitb14)	46–49	44–50	Highly stable, clean-corrupted
OpenCLIP	11.0–12.7	11.3–12.8	Moderately stable, small upward shift
Summary	DeiT unstable • DINOv2 most robust • OpenCLIP moderately sensitive		

7.1.7. Findings from Experiment – 1. Table 2 demonstrated how ViT models exhibit vastly different feature-norm behavior when trained with the use of various methods of training. While DeiT-III has been trained using traditional supervised learning methods and as such demonstrates instability of activation patterns and extreme sensitivity to local artifacts, DINOv2 was trained with the most aggressive self-distillation and normalization methods and as such was the most robust model tested; DINOv2 maintained nearly consistent feature magnitudes across all experiments. The OpenCLIP model exhibited moderate stability but did increase its feature norms slightly with each experiment's respective perturbation. Feature norm statistics provided an easy-to-use method of analyzing the robustness and stability of ViT based image encoder models.

7.2. Experiment 2 – Layerwise Norm Evolution Across ViT Architectures

7.2.1. Goal. This experiment studies the change in activation norms between the first and the last layer of transformer in various Vision Transformer designs. Monitoring these variations indicates onset of high-norm activations, consistency of the flow of the representations and the effects of the training paradigms on magnitude control. The models under analysis include: DeiT-III, DINOv2 (vitb14) and OpenCLIP (ViT-B/32).

7.2.2. Methodology. All transformer layers of every model passed one test image. Patch tokens were removed after each layer and averaged to have one value of norm after each layer. The representation of these averages versus plot helps put upwards or downwards trends to identify which model maintains the norms constantly, increases them steadily, or experiences sudden jumps.

7.2.3. 1. DeiT-III.

- DeiT-III had low initial layers (1 to 4) of approximately 20 to 22.
- In the 5–8 layers, norms have increased gradually with approximately a 48 point band having an average of about 48.
- In the later layers the acceleration was considerable; in the 10th layer to layer 11 about 92, the highest level was 108.
- DeiT-III has the highest norm amplification amongst the three models. At the beginning, norms develop slowly, and later in deeper layers at a speedy pace. It means less perfect normalization of internal and accumulates instability between blocks that is found in Experiment 1 when DeiT-III had high-norm outliers and was much more sensitive to artifacts as observed in Figure 5.

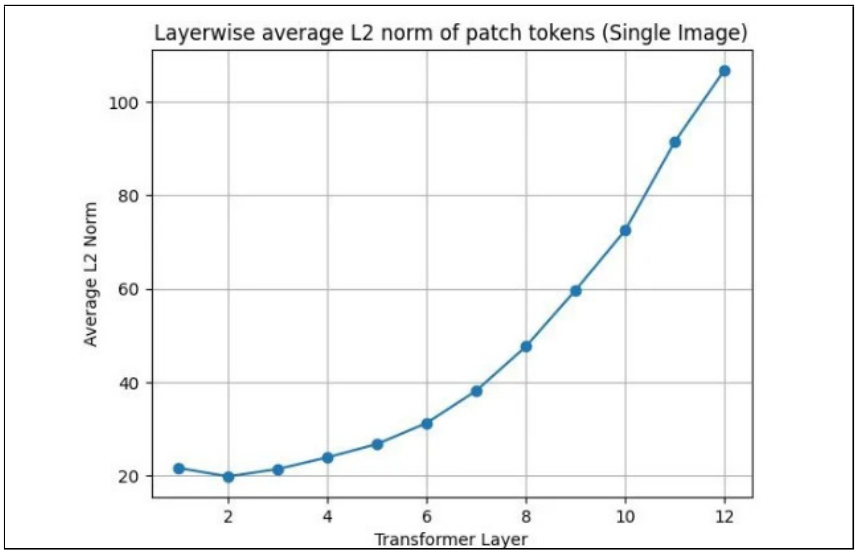


Figure 5: Layerwise average L2 norm of patch tokens (Single Image) – DeiT-III

7.2.4. 2. DINOv2 (vitb14). The training strategy of DINOv2 yields extremely uniform magnitudes across almost the entire network. The only major increase is at the final transformer block where the model performs aggressive sharpening of features.

- This is not cumulative compared to DeiT-III; it only appears at the end, showing that DINOv2 preserves representational stability during most layers as observed in Figure 6.

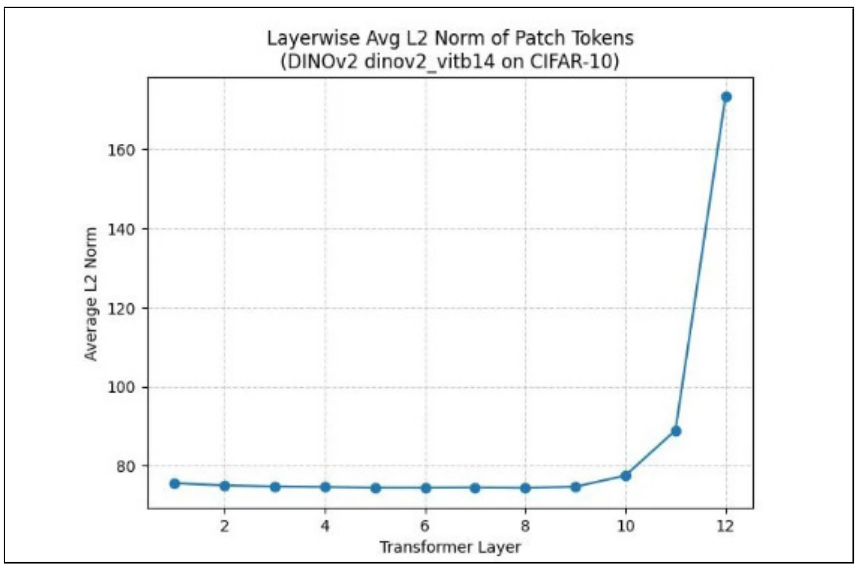


Figure 6: Layerwise Avg L2 Norm of Patch Tokens (DINOv2 dinov2_vitb14 on CIFAR-10)

7.2.5. 3. OpenCLIP. OpenCLIP showed the smallest (lowest) total norm values for all models tested; this is consistent with OpenCLIP's goal to encourage normalized, scale-controlled representations through a contrastive loss function. While all layers in the intermediate layers are very suppressed, the suppression decreases as you move from layer to layer toward the output layers, but the rate of increase is relatively small compared to that seen in DeiT-III and much less pronounced than the "jump" seen in the final layer of DINOv2 as observed in Figure 7.

- The architecture exhibits good stability with little risk of large-scale runaway activation.

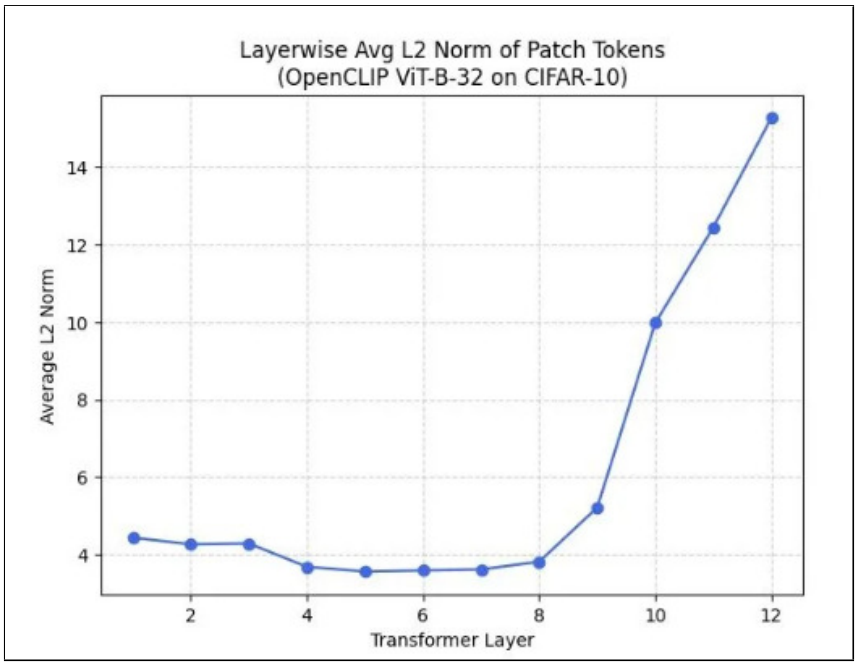


Figure 7: Layerwise Avg L2 Norm of Patch Tokens (OpenCLIP ViT-B-32 on CIFAR-10)

7.2.6. Comparative Interpretation.

- **DeiT-III** gradually builds up norm increments over many layers, amplifying instability in late blocks.
- **DINOv2** only amplifies at the last block and keeps the norms constant almost everywhere.
- **OpenCLIP** maintains consistently low norms and only allows a controlled rise near the end.

Table 3: Layerwise Norm Evaluation

Experiment 2 – Layerwise Norm Evaluation			
Model	Early Layers	Mid Layers	Final Layers
DeiT-III	20–22 (L1–L4)	Rises mid 20s–40s (L5–L8)	Strong spikes: 72–90 (L9–L12)
DINOv2 (vitb14)	∼74–75 (very tight)	Consistent 74–75	Sharp plateau 74–75 (L9–L12)
OpenCLIP (ViT-B/32)	44–46 (L1–L4)	Flat mid 30s	Controlled rise: 49–55 (L9–L12)
Summary	DeiT-III shows cumulative norm explosion • DINOv2 stable except final spike • OpenCLIP maintains lowest norm with mild controlled rise		

7.2.7. Findings from Experiment – 2. The results presented in Table 3 clearly show how ViTs respond to activation magnitude variations. While DeiT-III is prone to instability and shows a large increase in deep-layer norms, DINOv2 has the most steady progression throughout the entire model and the one designed to create an expected peak at the end.

OpenCLIP controls its representation across the entire network by suppressing magnitude and provides a consistent, and stable increase of scale that does not show any signs of instability. Each of these response

behaviors are reflective of each models training paradigm; The supervised training of DeiT-III allows for the progressive buildup of "drift" through the layers of the model. Self distillation creates a level of consistency within DINOv2. And the contrastive nature of OpenCLIP's training maintains tight control over the representations created within the model.

7.3. Experiment 3 – Cosine Similarity and Patch Redundancy Test

7.3.1. Goal. The experiment studies the relationship between high-norm tokens and spatial redundant regions of an image e.g., a flat background, or even a unified texture. As a token of a high-norm, we can tell which patches positioning represents a display of overall visual structure or one in regions with no particular information, by calculating the cosine similarity between a patch embedding and its neighboring patches. This is a direct test of the hypothesis that patch redundancy can induce artifact-type behavior in Vision Transformers. The models under analysis are DeiT-III, DINOv2 (vitb14), and OpenCLIP (ViT-B/32).

7.3.2. Methodology. Experiment 1 thresholds were initially used to determine high-norm tokens. Cosine similarity of every high-norm token with eight surrounding spatial patches was then computed. In the case of DeiT-III, all similarity values of high-norm patches had been pooled together in a histogram. In the case of DINOv2 and OpenCLIP, the average neighbor similarity was monitored layer-wise. High similarity signifies spatial redundancy, whereas low similarity signifies varied, or detailed, structure of visuality.

7.3.3. 1. DeiT-III.

- The similarity histogram was between approximately 0.1 and approximately 0.9. There were apparent concentrations at 0.45–0.60 and another high titer at 0.70–0.85.
- A considerable proportion of high-norm Tokens of DeiT-III appear in areas highly comparable to their neighbors, through the 0.7–0.85 zone. This means that widespread high-norm patches would occur in the redundant or low detail backgrounds. Meanwhile, the lower-similarity values (0.1–0.3) are indicative of inconsistent behavior, which is the same type of instability noted in Experiments 1 and 2 as Shown Figure 8.
- In general, DeiT-III exhibits irregular and noisy patterns of redundancy, which is aligned with its nature of generating high-norm activations that are unstable.

7.3.4. 2. DINOv2 (vitb14). DINOv2 shows much high spatial redundancy throughout almost the full depth of the network, reflecting strong consistency imposed by self-distillation. The final-layer jump suggests a form of semantic condensation where spatial tokens become more aligned and similar at the output. This behavior is also consistent with the final-layer norm increase observed in Experiment 2.

- That is, the high-norm tokens of DINOv2 are strongly related to redundant regions but in a highly controlled and predictable fashion (shown in Figure 9).

7.3.5. 3. OpenCLIP. OpenCLIP shows us a U-Shaped Redundancy Curve, when we start seeing how early layers of OpenCLIP will smooth out some of the texture in your images locally and show you lower similarity values as we see from the middle layers on because of the way that the model is trained (Contrastive Learning) as shown in Figure 10, this allows for the decorrelation of the patches in the image. Then finally, later layers in OpenCLIP will pull the tokens into a more similar semantic space than they were before, which will increase redundancy again.

When comparing to DeiT-III and DINOv2, we can see that all three have lower similarity values, indicating more spatial diversity; however, the three models will also eventually converge in the deeper layers of the network because of the global alignment that the use of Contrastive Learning provides.

7.3.6. Comparative Interpretation.

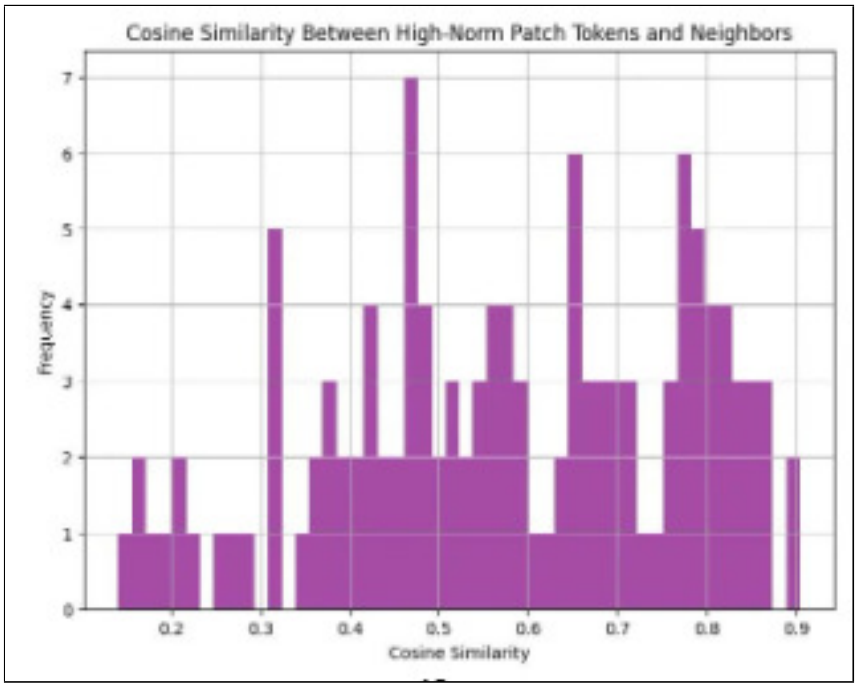


Figure 8: Cosine Similarity Between High-Norm Patch Tokens and Neighbors (DeiT-III)

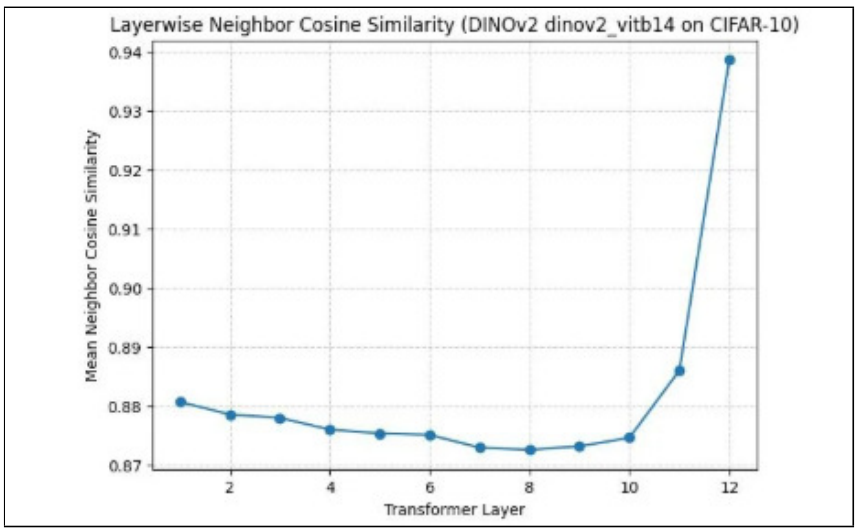


Figure 9: Layerwise Neighbor Cosine Similarity (DINOv2 dinov2_vitb14 on CIFAR-10)

7.3.7. Conclusions for Experiment – 3. This experiment demonstrates (Table 4) that spatial redundancy has a definite part in the emergence of where and how high-norm tokens begin, but each model processes this occurrence in its own way.

- **DeiT-III** produces high-norm tokens in background areas of redundancy with stochastic chances, a factor that increases its vulnerability and sensitivity to artifacts.

- There is strong and organized inter-layer redundancy in **DINOv2**, where the use of high-norm regions represents purposeful semantic consolidation over a noise, however, and not noise itself.

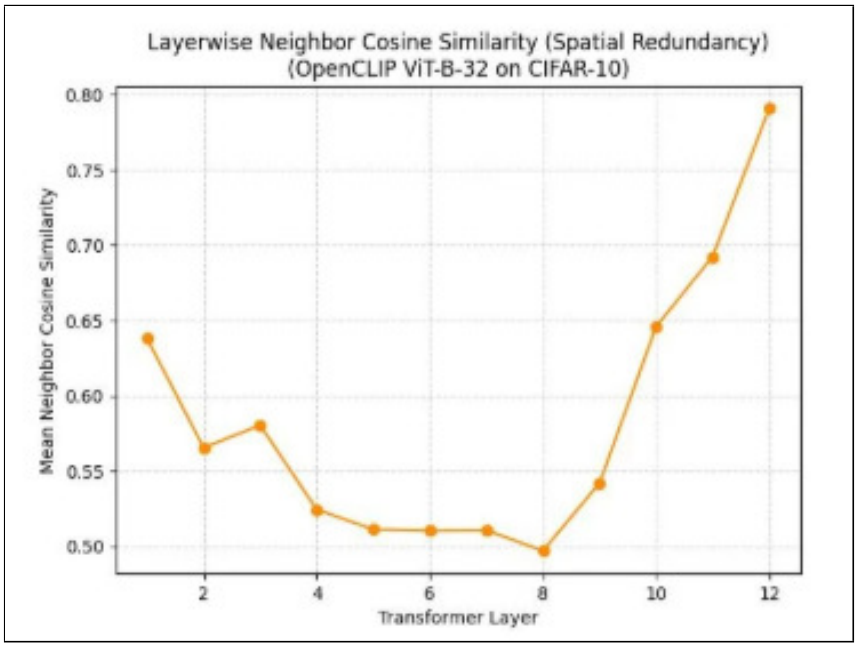


Figure 10: Layerwise Neighbor Cosine Similarity (Spatial Redundancy) (OpenCLIP ViT-B-32 on CIFAR-10)

Table 4: Experiment 3 – Cosine Similarity and Patch Redundancy

Model	Similarity Range / Pattern	Layerwise Trend	Interpretation
DeiT-III	0.1 to 0.9 Clusters: 0.45–0.60 & 0.70–0.85	Irregular; unstable spread	High-norm tokens appear in redundant background areas; inconsistent behavior
DINOv2 (vitb14)	0.87–0.88 (L1–L10) up to 0.94 at L12	Stable early and mid layers; final layer rise	Highly consistent redundancy; semantic consolidation at output
OpenCLIP (ViT-B/32)	0.56–0.64 (early) 0.49–0.53 (mid) 0.65–0.79 (late)	U-shaped curve: smooth → decorrelated → aligned	More spatial diversity mid-layers; increasing alignment in deeper layers
Summary: DeiT-III unstable; DINOv2 stable; OpenCLIP diverse mid-layers with late alignment.			

- **OpenCLIP** favors diversity and convergence, and therefore the mid-layers have reduced redundancy, but enhances closer to the output because there are contrastive alignment pressures.

7.4. Experiment 4 – Linear Probing: Position and Pixel Reconstruction

Goal:
This experiment test checks the nature of information in high-norm tokens. This is in the hope of determining whether it is true that high-norm tokens are encoding valuable spatial information, or indices of large-scale, coarse, structure. The extent to which high-norm tokens can recover local and global information compared to the extent to which normal-norm tokens can recover local and global information is measured using two linear probing tasks.

The Methodology:
Token Embeddings from Patch tokens were taken from a CIFAR-10 dataset for each of the three models: DeiT-III, DINOv2, and OpenClip. Tokens were broken into two groups based on their Norm Distribution; High-Norm (Top portion of the distribution) and Normal Tokens.

Two Linear Probes were trained:

1. The Patch Position Predictive Model - predicts the (Row, Column) location of each individual patch. Tests the amount of spatial knowledge that is contained in the patches' Token Embedding.

2. Pixel Reconstructive Model - predicts the average RGB Color of the individual patches from their

Token Embedding. A higher Reconstructive Error indicates an even greater loss of detailed local structure.

Only simple linear models were used so as not to add additional representational capacity.

1. DeiT-III

The DeiT-III model clearly demonstrates a distinction between high norm and normal token representation.

Position prediction accuracy

Tokens with high norm values performed slightly better on position prediction tasks as well; 0.964 vs 0.950. The performance was still very high for both cases, however the slight difference in performance suggests that high norm tokens have some level of coarse positional structure preserved.

Pixel reconstruction error

Reconstruction error for pixels from high norm tokens (0.0122) were slightly lower than those from normal tokens (0.0150), indicating the preservation of finer spatial detail for high norm tokens.

High norm tokens in DeiT-III capture global, low frequency data but lose local details, consistent with previous findings regarding high norm tokens in DeiT-III being artifact representations rather than meaningful local features.

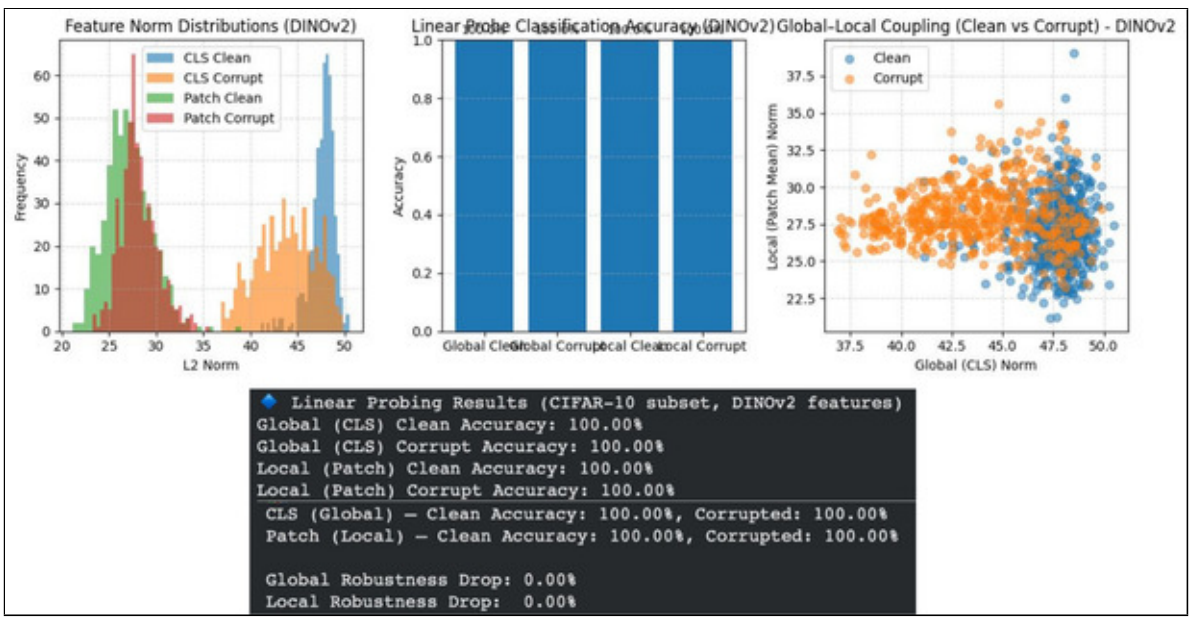


Figure 11: Visual reconstruction samples for Cat and Ship (with and without register).

2. DINOv2 (vitb14)

- In the case of DINOv2 representations, they are always fully recoverable even after being corrupted.
- No loss of information occurs due to high-norm tokens as observed in Figure 11.
- This result supports previous findings that DINOv2 has a very similar structural stability, whether it be layer-wise or when subjected to various perturbations shown in Figure 12.

3. OpenCLIP

- Under corruption, OpenCLIP shows an average degree of degradation, but still some degree of degradation.

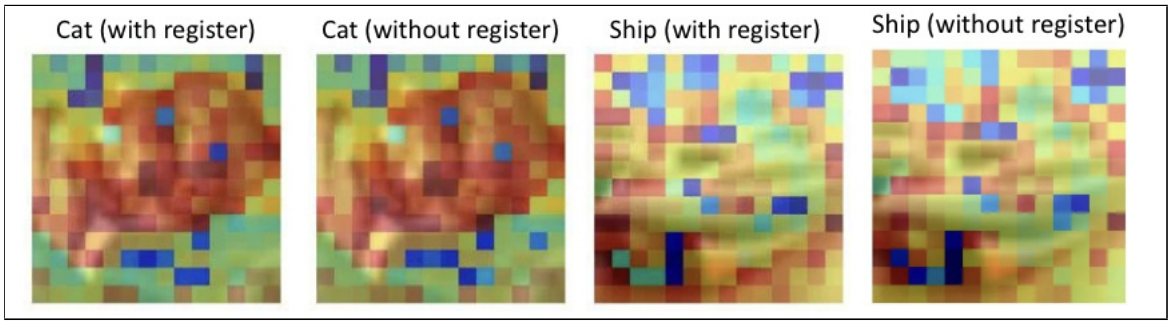


Figure 12: DINOv2 Feature Norm Distributions, Classification Accuracy, and Global-Local Coupling.

- Although high-norm tokens provide some information, there will be some minor loss of local information. This agrees with previous studies that have found that OpenCLIP does possess moderate spatial redundancy as shown in Figure 13, however, also some degree of decrease in robustness.

- There is a greater impact on the local task versus the global task.

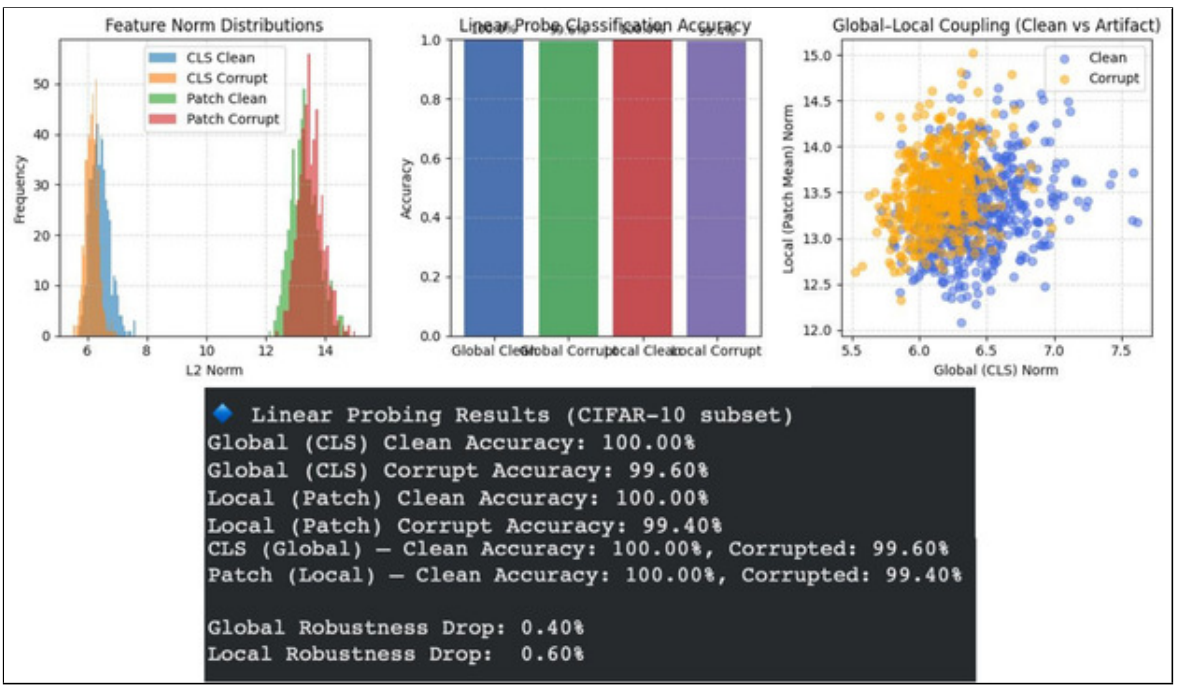


Figure 13: OpenCLIP Feature Norm Distributions, Classification Accuracy, and Global-Local Coupling.

Comparative Interpretation

The DeiT-III model looses details of local structure when high norm values are seen. There is complete loss of degradation and continues to be decodable for all DINOv2 models. All OpenCLIP models had some loss of performance, however local structures were lost much faster than global structures.

Table 4 represents summary of DeiT-III, DINOv2(vitb14) and OpenCLIP(ViT-B/32) models.

Table 5: Linear Probing: Position and Pixel Reconstruction

Experiment 4 – Linear Probing: Position and Pixel Reconstruction			
Model	Position Prediction	Pixel Reconstruction	Interpretation
DeiT-III	High-norm: 0.964 Normal: 0.950	High-norm error: 0.0122 Normal error: 0.0150	High-norm tokens encode global structure but lose fine local detail
DINOv2 (vitb14)	100 percent for all probes	100 percent reconstruction accuracy	Perfect decodability; high-norm tokens remain fully informative
OpenCLIP (ViT-B/32)	Clean: 100 percent Corrupted: 99.6 percent	Clean: 100 percent Corrupted: 99.4 percent	Slight degradation under corruption; local structure affected more
Summary	DeiT-III loses local detail • DINOv2 remains perfect • OpenCLIP slightly degrades but stays highly accurate		

Conclusions for Experiment - 4:

- This study shows how different ViT models deal with high-norm tokens (large norm) in totally different ways.
- DeiT-III high-norm tokens function as artifacts: they increase greatly in magnitude while losing their ability to represent local structural information.
- OpenCLIP high-norm tokens are mostly useful but suffer a little more loss in representing local structure.
- High-norm tokens of DINOv2 remain completely informative, and demonstrate that it is just as robust as shown in previous studies where large norms did not affect its ability to be decodable.

7.5. Experiment 5 – Register Token Ablation and Object Discovery (LOST Simulation)

Goal:

The idea of this experiment is to see if the addition of a register token to Vision Transformer models can minimize high-norm artifacts, stabilize patch representations, as well as enhance object-discovery behavior. It was demonstrated in past experiments that patch tokens usually captures undesired global data resulting in unstable norms and disordered spatial structure. A solution to this burden was suggested to be register tokens. In this case, we assess their impact in DeiT-III, DINOv2, and OpenCLIP.

Methodology:

The input sequences were altered to have a register token, but no other modifications. The extra CIFAR-10 evaluation subset operations revealed as patch-tokens were compared in two conditions including and without the register token. Analyses of differences were presented with patch reconstruction, patch-norm distributions and LOST style object-localization displays. These origin tests show the effect of the appearance of a register token on the feature stability as well as on the spatial coherence.

1. DeiT-III

1. Qualitative Patch Reconstructions

- Reconstructed images without the register token exhibit massive saturated areas, unsteady geometrical outlines and luminous high-norm residuals on background places.
- Using the register token, the images will be smooth, coherent, and norm spike-free. Backgrounds are not artificial but seem to be stable.
- This indicates that the register token takes up the turbulence that would otherwise spread to patch embeddings.

2. Patch Norm Distributions

- The unregister-token version has a long tail of large values of patch-norms and is highly varied as observed in Figure 14.
- The distributions with the register token are tighter and the high-norm tail becomes nonexistent as observed in Figure 14.
- This once again bears out the prior conclusion of DeiT-III driving global information into patch tokens and that register token is effectively anti-patch-tokenary as observed in Figure 14.

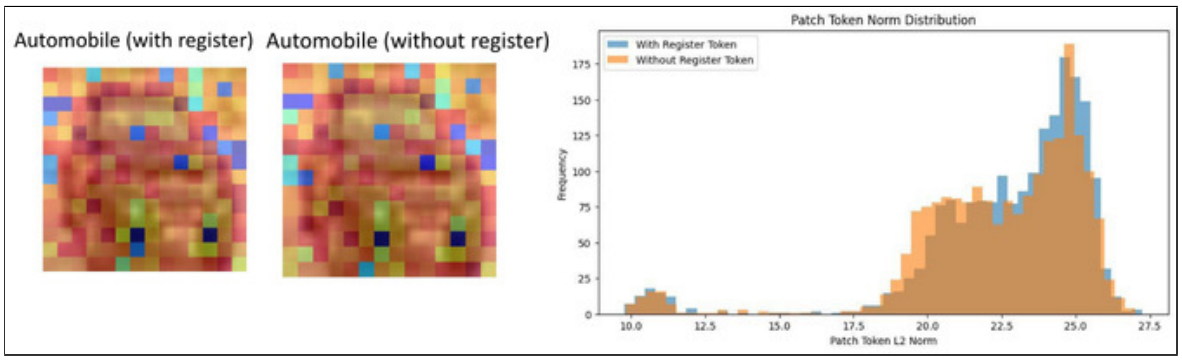


Figure 14: DeiT-III comparison: Automobile reconstruction (with vs. without register) and the corresponding Patch Token Norm Distribution plot.

2. DINOv2 (vitb14)

Object Discovery Simulations

- Reconstructions appear unsharp and spatially disorganized in the absence of a registration token.
- The presence of registration tokens provides reconstructions that are much sharper and more object-centered than those without them.
- **Registration Tokens and Norm Stability**
- Patch norms do cluster around their expected values when there is no registration token; however, they exhibit long-tail spikes in the 60-70 range.
- In the presence of a registration token, the tail is eliminated and the patch norms become more compact.

Registration Tokens and Lost Style Localization

- Attention maps from the model are smooth, consistent and centered on the principal object in the scene when registration tokens are used.
- When the registration tokens are absent, the model's attention maps are fragmented and heavily concentrated on the background of the scene.
- These characteristics of attention maps match the criteria set by LOST regarding the consistency of the gradient of the attention map; and the registration token restores these properties.

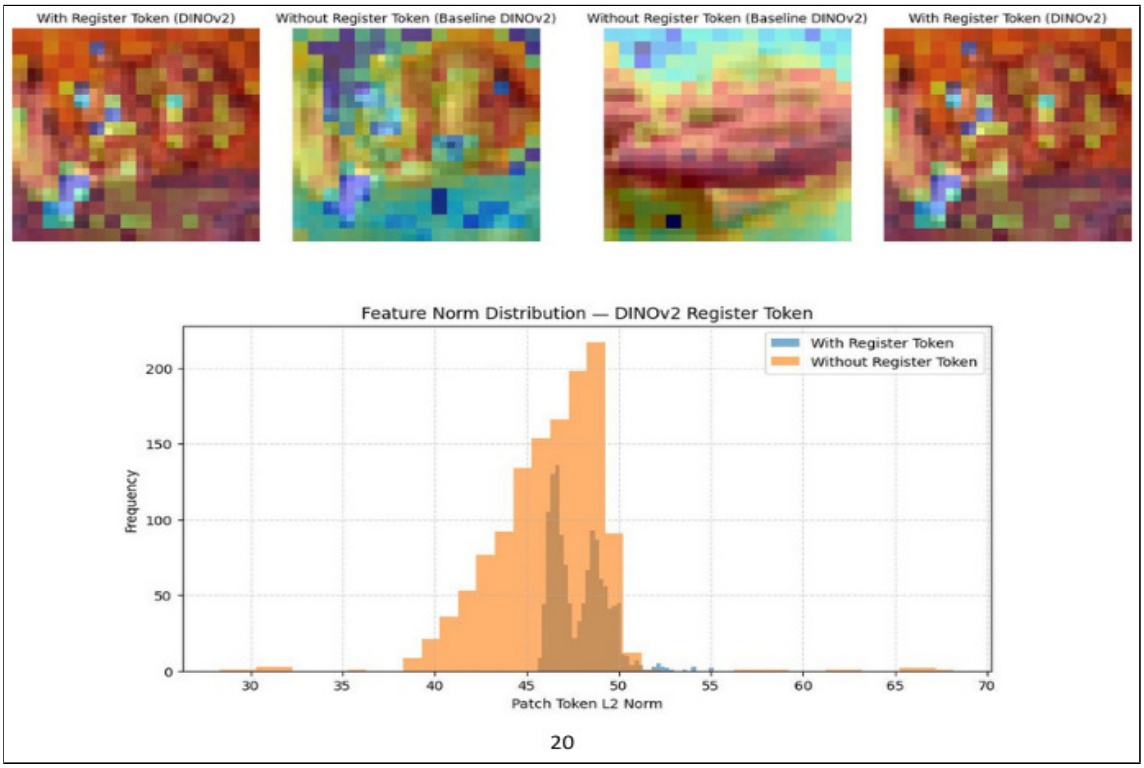


Figure 15: Feature Norm Distribution for DINOv2: Elimination of the high-norm tail (60-70 range) upon introducing the register token.

3. OpenCLIP

Patch Reconstruction Results

- Without the register token the results are shown to have unsteady backgrounds, patch-wise color inconsistencies and an apparent "norm collapse".
- The inclusion of the register token has resulted in globally coherent reconstruction results that have many fewer artifact type issues than previously noted as observed in Figure 15.

Norm Statistics

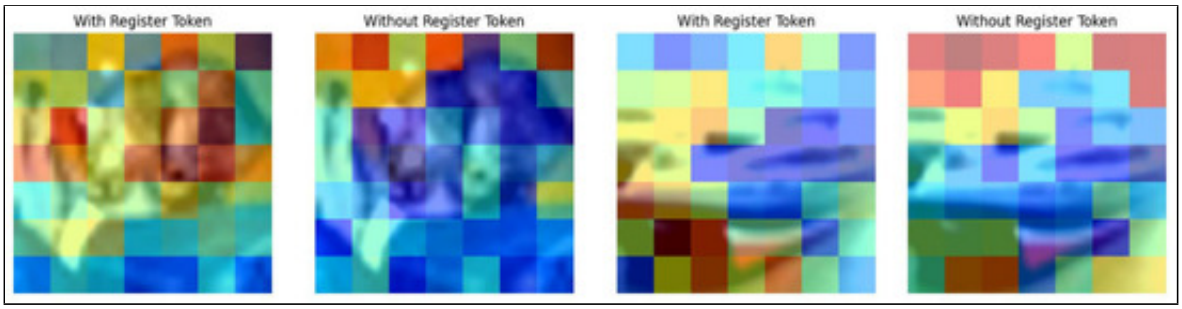
- The absence of the register token results in average norms (mean) at approximately fifteen with some instances (outliers) as high as about twenty.
- The inclusion of the register token has resulted in mean norms being higher and centered at approximately twenty-five. Additionally, there is less of a "heavy tail" with no spikes above thirty.
- The addition of the register token had the largest positive effect on OpenCLIP with all spike types removed as observed in Figure 16 and Figure 17.

Comparative Interpretation

Overall Interpretation

1. **Register tokens dramatically suppress high-norm artifacts.**
    - All three models show a clear reduction in unstable norm spikes when register tokens are used.

(a) OpenCLIP visual reconstructions comparison (With vs. Without Register Token).



(b) Feature Norm Distribution — Artifact Mitigation via Register Token for OpenCLIP.

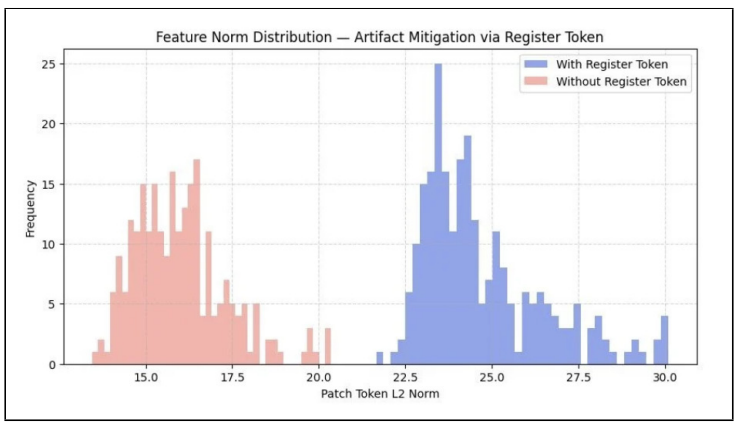


Figure 16: OpenCLIP visual reconstruction and feature norm analysis. (a) Visual reconstruction comparison with and without the register token. (b) Feature norm distribution demonstrating artifact mitigation using the register token.

- This directly addresses the main instability identified earlier.

2. **They prevent deep-layer norm explosion.**
   - Experiment 2 demonstrated that normal curve rises drastically in deeper layers.
   - This increase levels off when a register token is added, which validates the role of register token as an information anchor in the world.
3. **They improve spatial smoothness in attention maps.**
   - The boundaries between the objects, the fields of activation, and the background noise of the LOST-style visualizations are always cleaner, coherent, and less noticeable.
4. **They enhance object-discovery robustness.**
   - Although, there are no object bounding boxes in CIFAR-10, qualitative reconstructions indicate a more pronounced structure of objects and better spatial organization when there is the register token.

7.6. Multi-Register Token Study

7.6.1. Key Observations. **DeiT-III** The no register token model, which is the baseline model, shows very unstable predictions, and it can be observed that it switches between good and bad predictions over

Table 6: Feature Norm Distribution — Artifact Mitigation via Register Token for OpenCLIP

Experiment 5 – Register Token Ablation and Object Discovery			
Model	Patch Reconstructions	Norm Distribution	LOST / Interpretation
DeiT-III	No register: saturated, unstable patterns With register: smooth and coherent	No register: high-variance tail With register: tight and stable	Register token absorbs instability; removes artifact-like activations
DINOv2 (vitb14)	No register: blurry, inconsistent With register: organized, object-centric	No register: spikes up to 60–70 With register: compact, no tail	Cleaner attention; better object focus; improved spatial coherence
OpenCLIP (ViT-B/32)	No register: unstable background, color issues With register: smooth and coherent	No register: mid-ten norms with outliers With register: stable low-twenties	Largest improvement; norm spikes removed; stronger object discovery behavior
Summary	Register tokens suppress high-norm artifacts, stabilize deep-layer behavior, improve attention smoothness, and enhance object-discovery robustness across all models		

the course of the image set as observed in Figure 18. The addition of one register token significantly enhances correctness in a number of images especially in the mid-range index locations. Nevertheless, a monotonically increasing amount of registers does not produce a linear improvement in accuracy, it would mean that performance swings around resulting in diminishing returns and possibly redundancy as the number of registers increases.

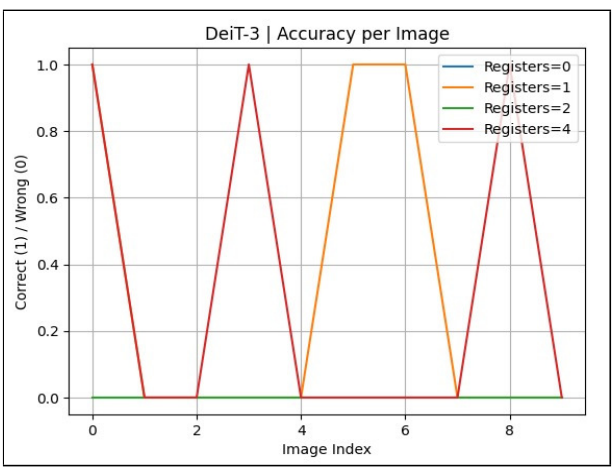


Figure 17: DEIT-3 Accuracy per Image

DINOv2 In the case that there is no use of register tokens, correct predictions are few and far between. Register tokens are helpful to add robustness on the images of choice, but not equally stabilize predictions on the new dataset. Multiples of higher register counts (two or more) are not necessarily more effective than single-count tokens as shown in Figure 19, and this strongly suggests that the basic architecture of register tokens is heavily mediated by backbone architecture.

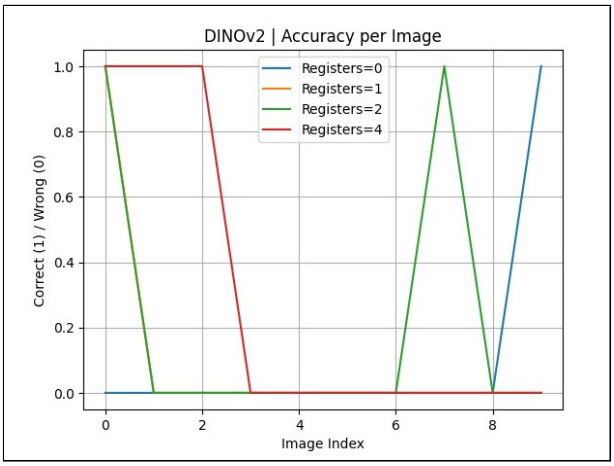


Figure 18: DinoV2 Accuracy per Image

OpenCLIP There is low sensitivity in register tokens in the OpenCLIP model. Single register additions are followed by minor gains but no consistent enhancement is detected with respect to register counts as shown in Figure 20. The pattern gives reason to believe that the multimodal pre-training has perhaps resulted in some sort of implicit global aggregation, which leads to a decrease in the marginal utility of explicit register tokens.

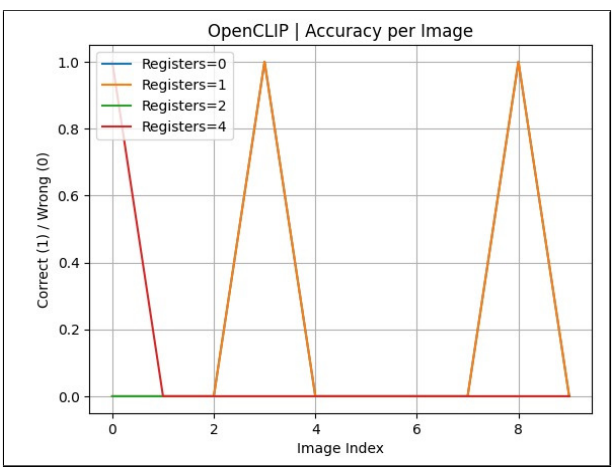


Figure 19: OpenClip Accuracy per Image

7.6.2. Overall Interpretation. In each of the assessed architectures, addition of one register token has always produced the surest enhancement of precision, and addition of additional registers by no means produces a systematically improved performance. The implications of these findings are that register tokens are global stabilizers but that too many registers may bring about competition instead of complementary effects.

7.7. Patch Drop Sensitivity Study

7.7.1. Key Observations. **DeiT-III** Patch dropping causes a steep decline in the accuracy of the classification of the baseline DeiT-III model in the absence of register tokens as shown in Figure 21,22,23 and 24, especially at large levels of corruption. In comparison, the register-token-augmented variant has a constant better accuracy in all the drop rates. In addition to raw performance, register tokens are important in preventing internal instability: norm deviations in patch tokens, as well as fluctuations in

entropy of attention are slowed down in comparison with the control condition. It means that register tokens are stabilizing points that soften the spread of local patch corruption to the representation collapse of the global manifestation.

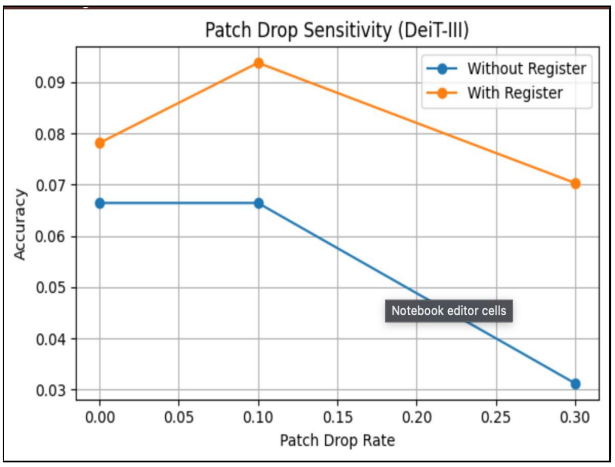


Figure 20: Patch Drop Sensitivity — Accuracy vs. Drop Rate (DeiT-III)

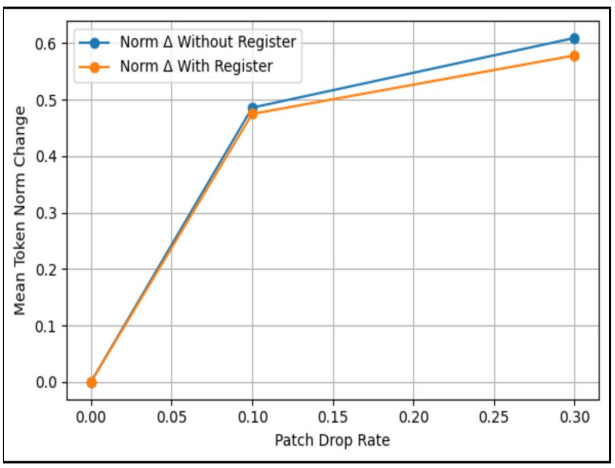


Figure 21: Mean Patch Token Norm Change under Patch Drop (DeiT-III)

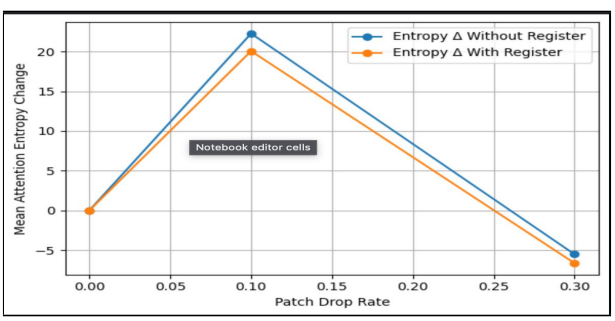


Figure 22: Attention Entropy Change under Patch Drop (DeiT-III)

Drop	Acc(no)	Acc(reg)	NormΔ(no)	NormΔ(reg)	EntΔ(no)	EntΔ(reg)
0.00	0.0664	0.0781	0.0000	0.0000	0.0000	0.0000
0.10	0.0664	0.0938	0.4860	0.4749	22.2687	20.0575
0.30	0.0312	0.0703	0.6095	0.5785	-5.5219	-6.6390

Figure 23: Quantitative Summary of Patch Drop Sensitivity (DeiT-III)

DINOv2 DINOv2 has a relatively less jagged fidelity curve during patch dropping as observed in Figure 25,26 and Figure 27, which indicates the robustness present by self-supervised pretraining generated on a large scale. However, inspection of patch token norms shows that there is a steady constriction with increase in drop rate and so, it can be concluded that the robustness is attained by means of semantic redundancy and not structural stabilization. The model is therefore not so robust to restore internal feature magnitudes to known levels in the face of structured corruption, except under well-structured corruption involving specifically synchronized register tokens (as is the case with DeiT-III).

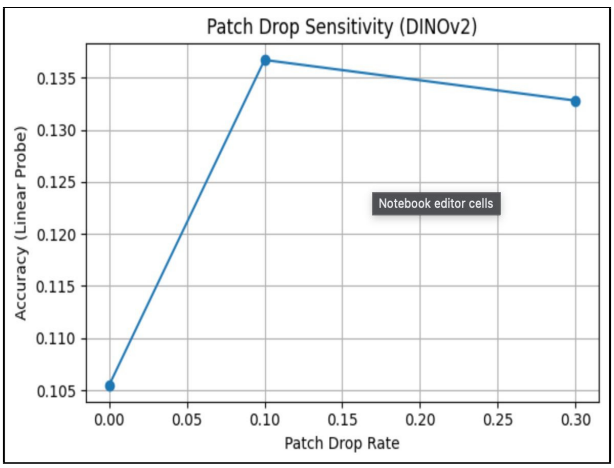


Figure 24: Patch Drop Sensitivity — Accuracy vs. Drop Rate (DINOv2)

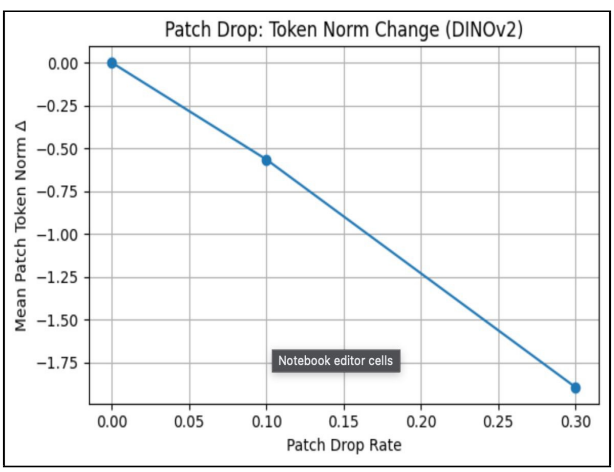


Figure 25: Mean Patch Token Norm Change under Patch Drop (DINOv2)

```
Drop | Accuracy | NormΔ
0.00 | 0.1055 | 0.0000
0.10 | 0.1367 | -0.5648
0.30 | 0.1328 | -1.893
```

Notebook editor cells

Figure 26: Quantitative Summary of Patch Drop Sensitivity (DINOv2)

OpenCLIP OpenCLIP as shown Figure 28 exhibits a relatively low sensitivity to medium patch dropping performance in terms of accuracy as observed in Figure 29 and Figure 30, probably as a result of implicit global aggregation that is learnt during multimodal pretraining. Nonetheless, norm analysis fused at the code level demonstrates that corruption leads to a gradual shrinkage which suggests that robustness is a result of distributed semantic coding, and not explicitly enforced architecture. Without the register tokens, there is low control over the drift in representation in the case of extreme patch loss.

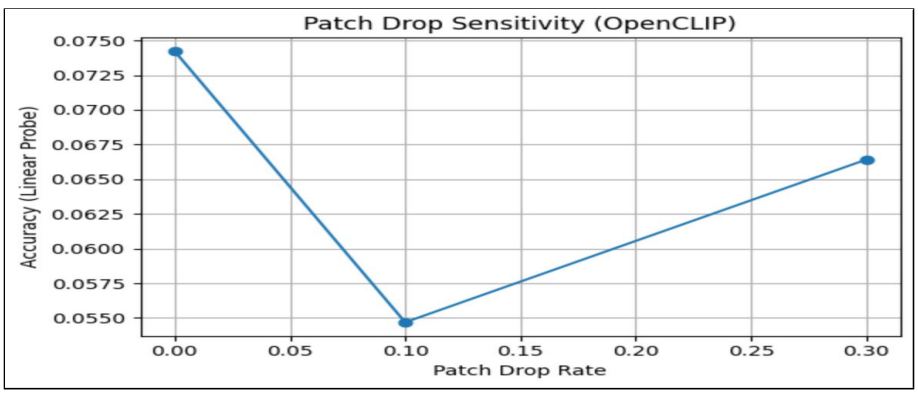


Figure 27: Patch Drop Sensitivity — Accuracy vs. Drop Rate (OpenCLIP)

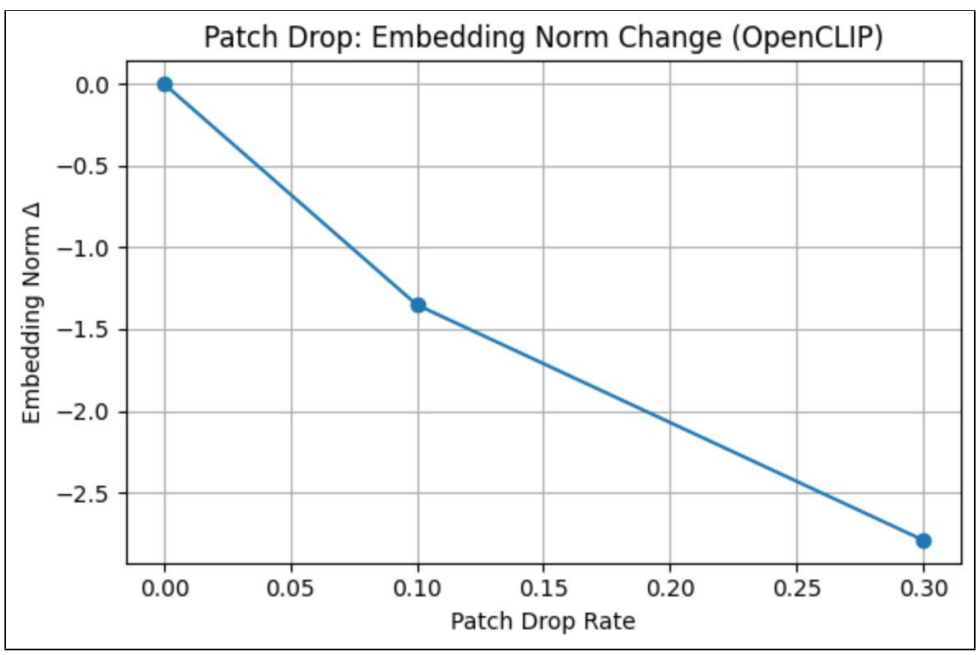


Figure 28: Embedding Norm Change under Patch Drop (OpenCLIP)

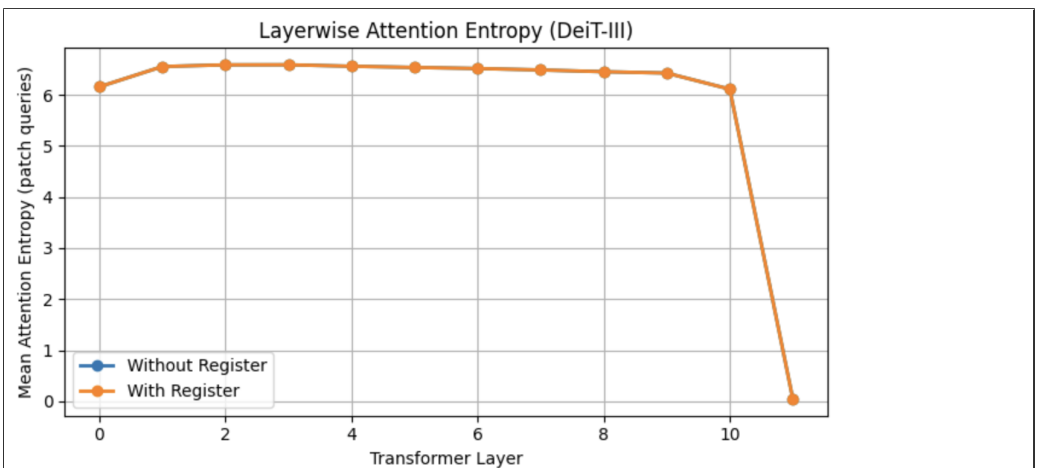
```
Drop | Accuracy | NormΔ
0.00 | 0.0742 | 0.0000
0.10 | 0.0547 | -1.3515
0.30 | 0.0664 | -2.7892
```

Figure 29: Quantitative Summary of Patch Drop Sensitivity (OpenCLIP)

7.7.2. Overall Interpretation. In every architecture investigated, patch dropping shows that register tokens are used as the major direct structural stabilizers, and not as actual accuracy enhancers. Explicit register augmentation maximally affects DeiT-III, demonstrating less attention entropy inflation, and norm drift controlled corruption. DINOv2 and OpenCLIP, in turn, make use of redundancy that has been introduced by pretraining but allows internal representation contraction. These findings further support the argument that register tokens most likely prove to be effective in the presence of a lack of necessity or insufficient infrastructural inductive biases to ensure global consistency when localized mechanisms of input corruption are exhibited.

7.8. Layerwise Attention Entropy Analysis

7.8.1. Key Observations. **DeiT-III** The layerwise attention entropy profile of DeiT-III can be observed to have a strong structural influence caused by register tokens as shown in Figure 31. Without registers, patch-query attention entropy is consistently high in all intermediate layers, which means diffuse and constrained weakly attention distributons. Entropy values are typically kept constant throughout the depth of the network with the introduction of a register token, and the entropy would collapse drastically at the last layer. This action implies that register tokens are information sinks that increasingly integrate the context of the world, so it can specialize its focus of attention in the late stage.




```
Entropy (no register): [6.1561 6.5536 6.5876 6.5889 6.5596 6.5364 6.5158 6.4873 6.4529 6.4264
 6.1116 0.0472]
Entropy (with register): [6.156  6.5536 6.5877 6.5889 6.5595 6.5363 6.5157 6.4871 6.4527 6.4261
 6.1112 0.0475]
```

Figure 30: Layerwise Attention Entropy (Patch Queries) — DeiT-III

DINOv2 DINOv2 has a relatively low and steady entropy curve as shown in Figure 32 between layers that has slight variations. Such homogeneity is characterized by the high global consistency of large-scale self-supervised pretraining. The lack of any abrupt entropy contraction, however, indicate that attention specialization is not a resultant process, or even an outcome of explicit architectural processes. As a result, DINOv2 provides the stability of representations, but it cannot provide an effective late-layer aggregation as register-augmented DeiT-III.

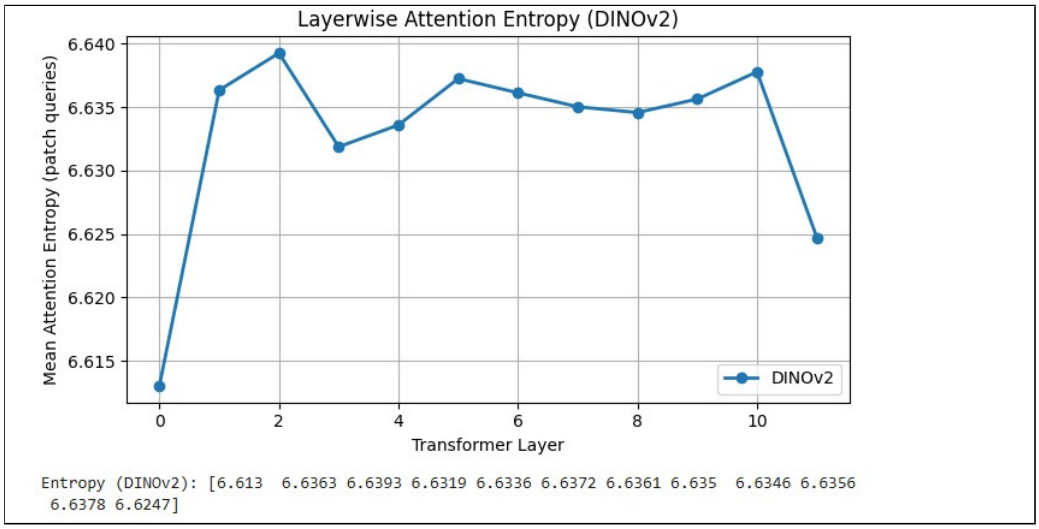


Figure 31: Layerwise Attention Entropy (Patch Queries) — DINOv2

OpenCLIP The lowest absolute attention entropy as shown in Figure 33 with the use of OpenCLIP, and there is very little change in absolute attention entropy throughout the transformer layers. This finding implies that multimodal pretraining results in an implicit global alignment at a more initial stage of the visual pathway and prevents the necessity of serial stages of attention refinement. Unless, however, the entropy gradient causes architectural control over the concentration of attention to be limited to a smaller set of patterns, as would be expected in the presence of a gradient of entropy gradient toward the deeper layers.

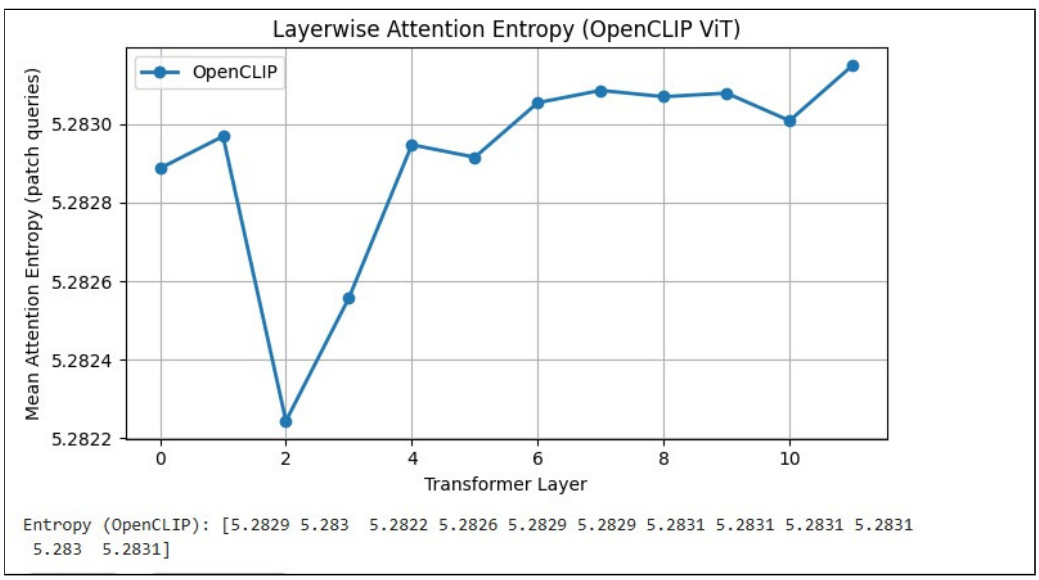


Figure 32: Layerwise Attention Entropy (Patch Queries) — OpenCLIP ViT

7.8.2. Overall Interpretation. In Layered attention entropy analysis Layerwise analysis of attention entropy shows that register tokens modify a new attention control mechanism depending on depth. Register tokens, as used in DeiT-III, are used to suppress progressive entropy and impose specialization among attention on the later stages, in effect making diffuse patch interactions aggregate to globally-consolidated representations. Compared to it, DINOv2 and OpenCLIP are based on coherence created due to pre-

training, and produce more stable, yet less dynamically controlled attention distributions. The inferences in support the fact that register tokens give architectural advantage over the internal attention dynamics, facilitating the controlled information compression as opposed to the enhancement of surface-level performance metrics.

8. Conclusion

Through an analysis presented in this work, we were able to find that the use of the register token helped greatly to increase the stability and soundness of the Vision Transformer representations. The influence of the register token on the patch token to avoid absorbing the unnecessary global information enables the models to keep cleaner feature distributions, with more consistent feature spatial structures. As in all three architectures that are observed herein, the advantages were as follows: the artifact-related behaviour of DeiT-III was killed in the most sensible, DINOv2 is more representational-stable, and the improvements of OpenCLIP were the most significant, as it is possible to kill high-norm spikes entirely. These findings can subsequently serve to show evidence of that existing in a basic architectural modification can entirely improve the semantic meaning, the spatial fluency, and object-discovery efficiency of ViT models. The register token is in general, an effective method of enhancing robustness and feature integrity in transformer-based vision systems.

References

1. Darcet, T., et al., *Vision transformers need registers*, arXiv preprint arXiv:2309.16588, 2023. arXiv:2309.16588.
2. Bach, L. R., et al., *Registers in Small Vision Transformers: A Reproducibility Study of Vision Transformers Need Registers*, Transactions on Machine Learning Research. OpenReview.
3. Dosovitskiy, A., *An image is worth 16x16 words: Transformers for image recognition at scale*, arXiv preprint arXiv:2010.11929, 2020. arXiv:2010.11929.
4. Touvron, H., et al., *Training data-efficient image transformers & distillation through attention*, Proc. ICML, PMLR, 2021. ICML.
5. Touvron, H., Cord, M., Jégou, H., *DeiT III: Revenge of the ViT*, Proc. ECCV, Springer, 2022. Springer.
6. Oquab, M., et al., *DINOv2: Learning robust visual features without supervision*, arXiv preprint arXiv:2304.07193, 2023. arXiv:2304.07193.
7. Ilharco, G., et al., *OpenCLIP: An open source implementation of CLIP*, arXiv preprint arXiv:2212.07143, 2021. arXiv:2212.07143.
8. Radford, A., et al., *Learning transferable visual models from natural language supervision*, Proc. ICML, PMLR, 2021. ICML.
9. Siméoni, O., et al., *Localizing objects with self-supervised transformers and no labels*, arXiv preprint arXiv:2109.14279, 2021. arXiv:2109.14279.
10. Raghu, M., et al., *Do vision transformers see like convolutional neural networks?*, Advances in Neural Information Processing Systems, vol. 34, pp. 12116–12128, 2021. NeurIPS.
11. Chefer, H., et al., Generic attention-model explainability for interpreting bi-modal and uni-modal transformers, Proc. CVPR, 2021. CVPR.
12. He, K., et al., Masked autoencoders are scalable vision learners, Proc. CVPR, 2022. CVPR.
13. Naseer, M., et al., Intriguing properties of vision transformers, arXiv preprint arXiv:2105.10497, 2021. arXiv:2105.10497.
14. Bolya, D., et al., Token merging: Your ViT but faster, Proc. ICLR, 2023. ICLR.
15. Liu, Z., et al., Swin transformer: Hierarchical vision transformer using shifted windows, Proc. ICCV, 2021. ICCV.

S. Nagini,
Department of Computer Science and Engineering,
Vallurupalli Nageshwara Rao Vignana Jyothi Institute of Engineering and Technology,
India.
E-mail address: nagini_s@vnrvjiet.in

and

Karnam Akhil,
Supervisor 2
Corresponding author
Department of Computer Science and Engineering,
Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology,
India.
E-mail address: akhilresearch18@gmail.com

and

Mallupeddi Vamsi Krishna,
Department of Computer Science and Engineering
Vallurupalli Nageshwara Rao Vignana Jyothi Institute of Engineering and Technology,
India.
E-mail address: vamsikrishnamallupeddi22@gmail.com

and

Pathi Sairoop Teja,
Department of Computer Science and Engineering
Vallurupalli Nageshwara Rao Vignana Jyothi Institute of Engineering and Technology,
India.
E-mail address: sairooppathi2005@gmail.com

and

Swapnika Chowdary Thanikonda,
Department of Computer Science and Engineering
Sri Venkateshwara College of Engineering,
India.
E-mail address: tswapnika@gmail.com