

Enhancing Transformer-Based Vision Models: Addressing Feature Map Anomalies Through Novel Optimization Strategies

Sumit Mamtani
New York University
Email: sm9669@nyu.edu

Abstract—Vision Transformers (ViTs) have demonstrated superior performance across a wide range of computer vision tasks. However, structured noise artifacts in their feature maps hinder downstream applications such as segmentation and depth estimation. We propose two novel and lightweight optimisation techniques—Structured Token Augmentation (STA) and Adaptive Noise Filtering (ANF)—to improve interpretability and mitigate these artefacts. STA enhances token diversity through spatial perturbations during tokenisation, while ANF applies learnable inline denoising between transformer layers. These methods are architecture-agnostic and evaluated across standard benchmarks including ImageNet, Ade20k, and NYUv2. Experimental results show consistent improvements in visual quality and task performance, highlighting the practical effectiveness of our approach.

I. INTRODUCTION

Transformer models [1], which employ multi-head self-attention mechanisms, have established themselves as the foundational architecture for a wide range of natural language processing (NLP) applications. Their success is largely attributed to the two-stage paradigm of large-scale pretraining followed by task-specific fine-tuning. The self-attention mechanism plays a central role by enabling the model to capture global dependencies across input sequences. With their inherent scalability and computational efficiency, transformer architectures have made it feasible to train exceptionally large models comprising billions of parameters [1]–[3].

Extending this paradigm to the vision domain, Vision Transformers (ViTs) were introduced by visiontransformers2021 [2], who adapted the transformer encoder to process image data. By dividing an input image into a sequence of fixed-size patches, and embedding them in a manner analogous to word tokens, ViTs demonstrated competitive performance across a variety of computer vision tasks including image classification, segmentation, and object detection. This novel approach quickly gained traction and became a new state-of-the-art method in visual recognition [2], [3].

However, recent investigations have identified peculiarities within the internal representations of ViTs. Specifically, certain anomalies or “artifacts” have been detected in the encoder’s feature maps—particularly in background regions lacking semantic content [4], [5]. These artifacts, often manifested as high-norm tokens, were first documented by registers [4], who noted that such tokens tend to appear in larger ViT models

during the latter stages of training. Their presence adversely impacts downstream tasks such as clustering or object discovery. To address this, the authors introduced auxiliary tokens called “registers,” intended to store global information and thereby prevent its diffusion into standard tokens, where it manifests as artifacts.

Subsequent work by mamba-needs-registers [6] extended this analysis to Vision Mamba models, which utilize State Space Models (SSMs) in place of self-attention. Interestingly, Vision Mamba exhibited even more pronounced artifact formation. However, integrating register tokens again mitigated the artifacts and improved overall model performance.

Building further on this line of inquiry, denoising [5] reaffirmed the existence of such artifacts in ViTs, including smaller variants and even in settings where register tokens were applied. Their findings implicated positional embeddings as a potential source of the artifacts. Rather than introducing architectural changes, they proposed a denoising strategy capable of separating semantic content from the artifact noise. This approach could be retrofitted to existing models without retraining, and it outperformed the register-based method in terms of effectiveness.

This paper is organized as follows: Section provides foundational knowledge about ViTs and describes several concrete model architectures utilized in the study by registers [4]. Section III presents a summary of that work, while Section IV reviews the contributions of mamba-needs-registers [6] and denoising [5], who build upon the initial findings.

A. DINO and DINOv2

DINO introduces a label-free representation learning method, employing a dual-network framework composed of a learner and a guiding model. The guiding model evolves as a momentum-based average of the learner across iterations. Instead of annotations, the model leverages knowledge distillation where multiple augmented perspectives—specifically two high-scale and several low-scale variants—are generated per input sample. The guiding model evaluates only the broader perspectives using an Exponential Moving Average (EMA) of learner parameters, while the learner processes all perspectives. To align representations, a cross-entropy objective compares outputs between the networks. The method incorporates strategies to deter feature collapse [7]. Notably, the self-

attention layers highlight spatially meaningful representations, showcasing structural coherence in the visual domain [4].

DINOv2 refines the predecessor by emphasizing adaptability, resource-efficiency, and robustness across varied visual tasks. Key advancements include:

- a systematic pipeline for compiling a balanced, high-quality, and diverse image corpus
- expansion to a significantly larger Vision Transformer (ViT) architecture, surpassing a billion parameters
- compression via distillation into compact yet effective descendant models [8]

These upgrades facilitate improved performance on dense prediction scenarios. Nonetheless, structural anomalies have been noticed within the attention maps produced by DINOv2 [4].

B. OpenCLIP

OpenCLIP constitutes an openly accessible variant of OpenAI’s Contrastive Language-Image Pre-training (CLIP) [9], leveraging joint vision-language modeling to enable robust zero-shot generalization across numerous domains. The framework optimizes a contrastive alignment objective, enhancing affinity between matching image-text instances while suppressing mismatched pairs. Once trained, the system converts visual data into corresponding textual outputs, which are then employed to infer downstream tasks. Empirical results suggest that these models perform on par with, or even surpass, task-specific supervised counterparts [9]. OpenCLIP extends the original by offering multiple configurations trained on varied datasets and parameter scales [10].

C. DeiT-III

DeiT-III aims to redefine the supervised learning baseline for ViTs through refined augmentation strategies influenced by recent advances in unsupervised training. Noteworthy modifications include the replacement of the typical Random Resize Cropping with a Simple Cropping technique. Additionally, input image dimensions were reduced from 224×224 to 126×126 , resulting in a 70% decrease in token count, thereby curbing overfitting for larger configurations. A further enhancement involves substituting the conventional softmax loss with a binary cross-entropy formulation, which improves convergence behavior in expansive ViT setups [11].

D. LOST

LOST presents an unsupervised technique for identifying object regions within individual images without relying on ground-truth annotations. The method incorporates a vision transformer backbone, such as DINO, to analyze the spatial layout. Assuming every image contains at least one salient object, the algorithm bypasses the classification token and instead scrutinizes inter-patch attention derived from the final transformer block. The patch with the fewest high-similarity connections to others is chosen as the initial anchor point or *seed*. According to the authors:

”Object-relevant patches exhibit stronger internal correlations than with background, and as objects typically occupy a smaller area, minimally correlated patches are more likely to belong to an object.” [12]

Subsequent steps involve aggregating additional patches correlated with this seed, followed by bounding box inference using a feature similarity-based approach. This instance-level localization, achieved on a per-image basis, facilitates large-scale deployment. The initial pseudo-labels serve as training input for a class-agnostic object detector, thereby enabling detection of multiple instances per image. In comparative evaluation, the trained detector surpassed the standalone correlation-based proposals in precision.

For category inference, a clustering-based semantic assignment is utilized. Extracted objects are normalized and passed through a transformer pre-trained with DINO. The resulting class embeddings are clustered using K-means, providing pseudo-labels. During evaluation, these are aligned with actual class labels using the Hungarian matching algorithm [12].

OUR CONTRIBUTIONS

In this paper, we identify and address structured anomalies in the feature maps of Vision Transformers (ViTs). Unlike prior studies that rely solely on auxiliary tokens or inference-time denoising, we introduce two novel optimization strategies to improve feature quality and model robustness:

- **Structured Token Augmentation (STA):** A method for enriching token diversity by injecting spatially-aware perturbations during the tokenization process, designed to reduce redundancy in background tokens.
- **Adaptive Noise Filtering (ANF):** A lightweight, learnable filtering mechanism integrated into transformer layers to suppress non-semantic activations in real time.
- We validate the effectiveness of these strategies through quantitative experiments and ablation studies across benchmark datasets such as ImageNet and ADE20k.

These techniques are complementary and can be integrated independently or jointly with existing ViT architectures to improve interpretability and downstream task performance.

II. RELATED WORK

Saimbhi has contributed significantly to the field of software security and digital media authentication. Saimbhi’s work, *Enhancing Software Vulnerability Detection Using Code Property Graphs and Convolutional Neural Networks*, presents a novel approach that integrates code property graphs with convolutional neural networks to improve the detection of software vulnerabilities. By leveraging abstract syntax trees, control flow graphs, and program dependency graphs, this research enhances both scalability and accuracy in vulnerability detection. In another work, *Distinguishing True and Fake Ultra-High Definition Images Using Relative DCT Analysis and Machine Learning*, Saimbhi addresses the challenge of identifying fake UHD images by combining Discrete Cosine Transform (DCT) analysis with machine learning techniques.

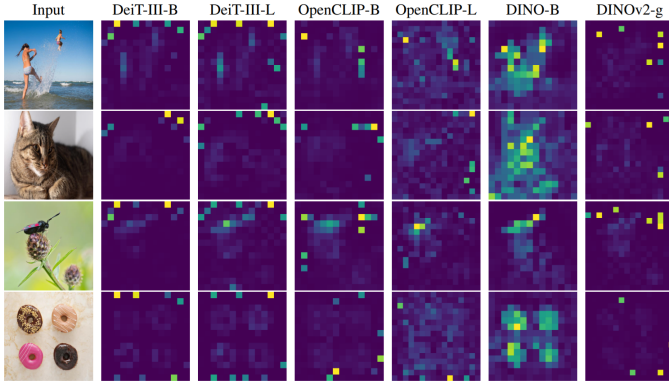


Fig. 1. Depiction of irregular patterns in the attention outputs of contemporary vision transformer architectures. Adapted from [4]

This framework demonstrates high accuracy in distinguishing genuine content from upscaled or deep neural network-generated samples.

Abhi Desai and Krishnaveni Katta focus on machine learning applications across various domains. Desai’s research includes *Active Learning Strategies for Efficient Text Classification*, which explores active learning strategies like uncertainty sampling to improve classification accuracy while minimizing labeling efforts. In *Enhancing Inventory Management with Progressive Web Applications (PWAs): A Scalable Solution for Small and Large Enterprises*, Desai develops a PWA-based system for efficient inventory management. Katta, on the other hand, contributes to healthcare and AI systems. Her paper, *Deep Learning for Early Lung Cancer Detection from CT Scans*, employs deep learning techniques to improve early cancer detection. In *AI Strategies and Game Dynamics in Risk*, Katta analyzes AI decision-making in board games, showcasing innovative approaches to game theory and optimization. Together, these works of Desai, Katta and Saimbhi highlight advancements in AI applications for both practical and theoretical domains.

III. INTEGRATING REGISTERS TO MITIGATE TRANSFORMER ATTENTION ARTIFACTS

This section presents a condensed overview of the findings from registers [4], where the introduction of supplementary register tokens is proposed as a strategy to counteract unintended behaviors in the attention mechanisms of ViTs.

A. Emergence of Anomalies in Transformer Architectures

Following the foundational discussion on ViTs, the study identifies models prone to producing such anomalies. The DINO framework, which facilitates unsupervised feature extraction, operates by aligning the predictions of a student transformer with those from a teacher model [7]. This approach is known to yield semantic coherence in the uppermost self-attention layers. Object localization techniques, such as LOST [12], exploit this behavior to detect object boundaries using unsupervised attention cues. Its successor, DINOv2 [8], aims to enhance resolution-based inference tasks, including

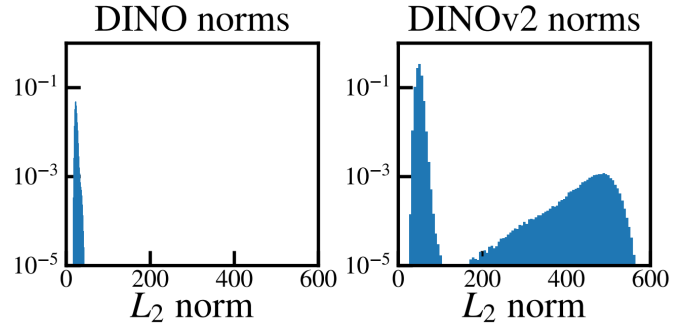


Fig. 2. Distribution comparison of token vector norms between DINO ViT-B/16 and DINOv2. Taken from [4]

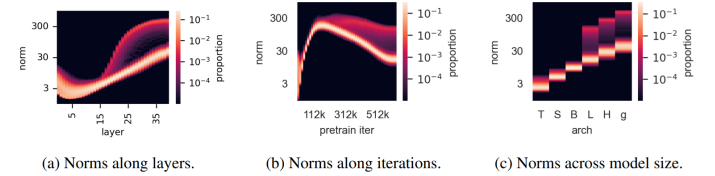


Fig. 3. Observed dynamics of outlier tokens in the 40-layer DINOv2 ViT-g configuration. Image adapted from [4]

segmentation and depth analysis. Despite superior results on these benchmarks, inconsistencies were noted between DINOv2 and prior models like LOST [4]. The inconsistencies can be traced to artifacts visible in the final-stage attention maps, as highlighted in Figure 1.

Whereas DINO yields attention distributions centered on primary objects without extreme values, DINOv2 displays dispersed activations, especially in background regions. Comparable disturbances are observable in the attention profiles of the supervised DeiT-III and the contrastively trained OpenCLIP architectures. The investigation prioritizes DINOv2 to systematically assess the origin and nature of these attention outliers in ViTs.

These anomaly-laden regions, known as artifact patches, exhibit significantly larger magnitudes in their token vectors at the model’s output compared to typical regions. Figure 2 illustrates the empirical distribution of these feature magnitudes. In contrast to DINO, where all token norms remain below 100, DINOv2 produces several token vectors exceeding a norm of 150. This empirical threshold is model-dependent, yet the study formally labels such vectors as:

“tokens exhibiting vector norms above 150 are designated as ‘high-magnitude’ tokens” [4]

The formation of these high-magnitude tokens follows specific training dynamics, illustrated in Figure 3, summarized as:

- Initial appearance typically occurs between layers 15 and 40.
- Artifacts begin to emerge after completing approximately one-third of the total training epochs.
- Anomalies are predominantly present in the largest configurations of the transformer family.

An additional insight reveals that these tokens frequently arise in zones of high visual redundancy. The cosine similarity between each high-norm patch token and its adjacent patches is elevated immediately after image tokenization. This spatial redundancy aligns with the tendency of these tokens to manifest within background regions. As such, the model appears to utilize these redundant zones for alternative computational roles without impacting output quality.

To gain further insight, linear classifiers were trained on patch embeddings for two specific tasks. In the first experiment, a classifier attempts to localize the position of a patch within the original image. Tokens with higher norms yielded inferior performance in spatial prediction, implying limited positional awareness. In the second experiment, another model was trained to reconstruct pixel values from the token embeddings. Again, high-norm tokens underperformed, indicating reduced visual reconstruction capacity.

However, a distinct trend emerged when a linear model was tasked with inferring the image class based on a single random token. Here, high-norm tokens exhibited stronger performance than their low-norm counterparts, indicating that these tokens encode more generalized, global semantic cues.

From these insights, the authors posit the following interpretation:

“Larger transformer models, once adequately trained, develop the ability to identify redundant spatial tokens and re-purpose them for distributed global information processing and storage.” [4]

B. Incorporating Registers into Vision Transformer Architectures

To mitigate the imbalance caused by disproportionately influential high-norm image segments, the concept of registers has been integrated into the architecture. These elevated-norm areas can distort spatial representation, especially in tasks requiring detailed prediction, even if such patches hold minimal contextual importance. Registers refer to additional learnable embedding vectors appended subsequent to the initial patch tokenization. They function in a comparable manner to the [CLS] symbol often employed for image classification tasks. These tokens are active during both the model optimization and inference stages but are discarded during final output processing. Figure 4 displays the integration of such register tokens post-image embedding. Computational complexity analysis indicates that including 16 such elements leads to an increase in floating-point operations by approximately 6%. However, using only four registers—a configuration more frequently adopted—results in under a 2% overhead.

The foundational idea of augmenting transformers with memory-like components can be traced to memorytransformer [13], who introduced tunable memory elements into transformer models for tasks in natural language processing. Prior research has also explored memory extensions in deep learning systems to boost model efficacy. In the cited work, general-purpose [MEM] tokens were introduced as placeholders to

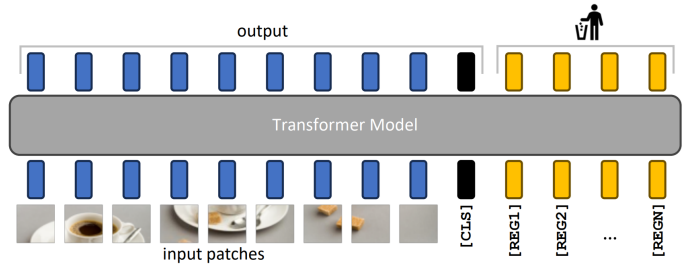


Fig. 4. Schematic representation of the proposed enhancement using register tokens. Adapted from [4]

retain global or context-specific representations. Three configurations were proposed: (1) memory tokens are concatenated to the input sequence and processed through a shared encoder, which inspired the integration approach for ViT-based register tokens; (2) a dedicated memory management layer is introduced; and (3) attention computation is split, first updating the memory attention before sequential attention is adjusted. Experimental analysis demonstrated the superiority of the baseline shared-token model, whereas the alternative designs yielded inconsistent results, sometimes enhancing and sometimes reducing the standard transformer performance.

C. Assessment of the Enhanced ViT Architecture

The revised architecture, incorporating register embeddings, underwent evaluation through training sessions on modified vision transformers. These modified models were compared—using both numerical measures and attention visualization—against baseline architectures lacking such components. Benchmarking was conducted across multiple learning paradigms: supervised (DeiT-III), language-guided (OpenCLIP), and self-supervised (DINOv2). Figure 5 illustrates representative attention heatmaps with and without registers across the three variants. Visual examination confirms a removal of visual artifacts in the attention maps for all transformer types when register tokens are utilized.

To quantify this effect, the norm magnitudes of attention responses were computed for each token type at the model output layer. Figure 6 presents the resulting norm distributions. Introduction of register tokens effectively eliminates the occurrence of elevated-norm outputs in conventional tokens. The highest norm values are now associated with register and class tokens, indicating that the former assumes the role of absorbing global context—previously captured by anomalous patch tokens. The class token already represents holistic information; hence, the similarity in attention patterns supports the view that registers function similarly. Localized information remains confined to patch tokens.

Performance comparisons, carried out via linear probing on datasets such as ImageNet (classification), ADE20k (semantic segmentation), and NYUd (depth estimation), revealed no degradation in accuracy with register integration. Even in zero-shot settings—specifically for ImageNet using OpenCLIP—the performance remains unaffected. Interestingly, a single register suffices to suppress the presence of dominant

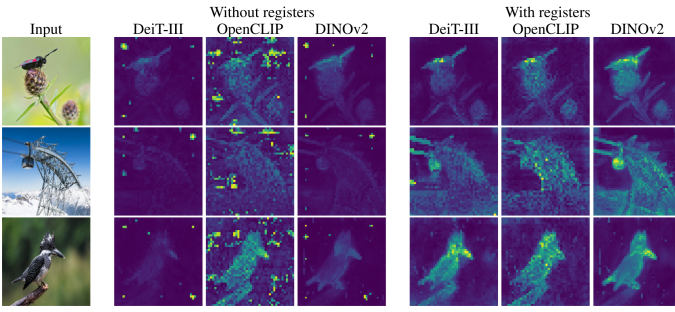


Fig. 5. Sample attention distributions with and without register integration. Extracted from [4]

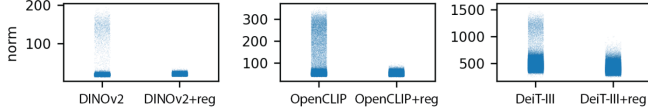


Fig. 6. Influence of register embeddings on the norm profile across token outputs. Derived from [4]

high-norm tokens. While DINOv2 and DeiT-III benefited notably in terms of object localization performance, OpenCLIP showed a slight performance decline. These observations reinforce the hypothesis that register embeddings assist in isolating the role of spatial context tokens, redirecting global representation into a dedicated memory slot, and thus alleviating undesired side effects, such as degraded performance in algorithms like LOST applied to DINOv2.

IV. PROGRESSIVE INVESTIGATIONS

The succeeding analyses expand upon the contributions of registers [4], unveiling further insights related to structural distortions observed in ViTs.

A. *mamba-needs-registers* [6] integrates register-like constructs into a State Space Model (SSM)-based vision model

Upon identifying anomalous token patterns prominently within background regions, the researchers implemented register-inspired elements within the Vision Mamba configuration, resulting in enhanced predictive accuracy compared to its baseline variant. Vision Mamba [14] is structured upon bidirectional State Space Models (SSMs), employing VIM Blocks that emulate attention-style dependency modeling across extended spatial spans. The architecture also includes a feedforward transformation pipeline, encoded spatial cues, and layer normalization procedures. Visual inputs are initially partitioned into discrete image segments, serving as input for the encoder module. A notable benefit of Vision Mamba lies in its linear time complexity, offering significant efficiency gains over traditional self-attention approaches which scale quadratically. Consequently, memory consumption and training duration are reduced relative to ViTs or CNNs (CNNs). The architecture surpasses models like DeiT [15] under certain benchmarks, illustrating the viability of SSM-based paradigms in visual processing tasks. [6], [14]

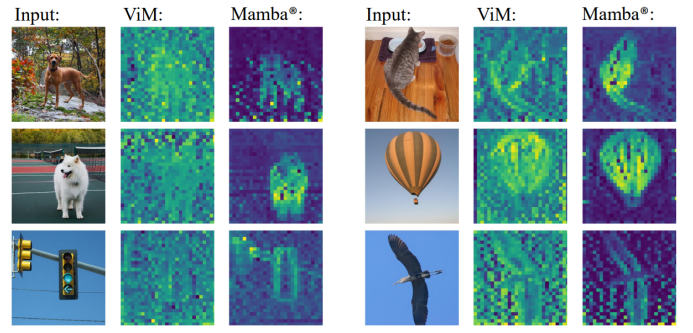


Fig. 7. Visual representation of internal outputs from standard Vision Mamba [14] versus the Mamba@ variant with register inclusion. Source: [6]

Analogous artifact formations were observed within Vision Mamba’s intermediate representations—similar to those previously reported in ViTs [4]—but emerged even in reduced model scales. As shown in Figure 7, these structured distortions are spatially distributed across the scene and are particularly dense in areas lacking salient objects. For this analysis, ℓ_2 distance metrics between aggregate (global) and local feature activations were employed. These maps reveal that high-norm anomalies correspond to global contextual encodings. The revised model introduces evenly distributed register elements amidst the input token series. Due to Vision Mamba’s non-permutation-invariant architecture, token order impacts representational dynamics, thus embedding registers throughout the input stream is essential. In contrast to earlier implementations, registers are appended near the output interface, contributing directly to the decision layer.

This modification led to notable performance gains. Furthermore, each register token was found to emphasize distinct regions or semantic components of the visual input. Given that Vision Mamba lacks multi-head attention capabilities, this property provides additional interpretability regarding internal activations. The enhanced Mamba@ version demonstrated superiority over prior architectural configurations in both image classification and pixel-level understanding tasks. [6]

B. *denoising* [5] introduces a noise-suppression pipeline to mitigate artifacts

The authors detected structured interference patterns within the latent feature maps of several transformer-based models, including DINOv2, DeiT-III, CLIP, and EVA02. These aberrations were identified as detrimental to interpretability and degraded the efficacy of post-processing methods such as clustering, particularly for dense prediction applications. It was postulated that spatial encoding strategies contribute to these anomalies. An association was identified between the incorporation of positional vectors and the manifestation of spurious activations. By employing maximal information coefficient calculations, a dependency was established between latent grid outputs and normalized patch positions. Baseline ViT outputs exhibited more prominent spatial correlation compared to their denoised variants.

In contrast to prior findings [4], these perturbations are present even in compact transformer configurations and are not exclusively characterized by exaggerated norm values. Weak structural patterns were identified in DINOv2 models augmented with registers. Such artifacts were consistently observed across all encoding depths, including models exposed only to blank inputs. Shallower layers predominantly displayed low-frequency distortions, while deeper modules manifested high-frequency components.

The proposed correctional mechanism, termed Denoising Vision Transformers (DVT), consists of two modular stages that function without necessitating model retraining. Initially, a per-instance denoising operation is carried out using coordinate-aware neural mappings, which disentangle content-specific representations from spatially anchored distortions. The decomposition assumes three constituent elements within any latent feature tensor:

- Semantic content embedding
- Positional noise component
- Residual cross-interaction term

This disentanglement leverages multiple spatial transformations and crops of an image, under the assumption that artifact components persist in fixed spatial coordinates while meaningful content shifts. Neural fields (coordinate-based networks) are trained to minimize a regularized error objective by reconstructing the input while isolating spatially consistent artifacts. Although this process yields clean activations, it incurs significant computational overhead.

In the subsequent phase, a lightweight transformer layer is trained using outputs from the first phase to map noisy features into their cleaner counterparts. This trainable denoiser also introduces learnable spatial embeddings during inference to further neutralize non-instance-specific noise. Unlike the initial method, this stage generalizes across image samples, mitigating single-instance biases. Figure 8 showcases the effect of applying DVT on various transformer architectures. Visible distortions—particularly in background areas—are substantially reduced, including those in models using registers. Post-denoising, feature representations display improved semantic clarity and object discernibility. [5]

Evaluation across multiple benchmarks demonstrated the effectiveness of this framework:

- **Semantic segmentation:** Notable and consistent improvement across multiple pretrained models, including those enhanced with register tokens.
- **Depth estimation:** Positive performance shifts in nearly all examined transformer-based models.
- **Object detection:** Performance uplift across all studied configurations, whereas the register-based method [4] showed limited benefit for DINOv2.
- **Unsupervised object localization:** DVT provided substantial accuracy gains for DINOv2 across diverse datasets, surpassing the improvements attained through register mechanisms. The suppression of artifacts also facilitated visual interpretability, thereby assisting down-

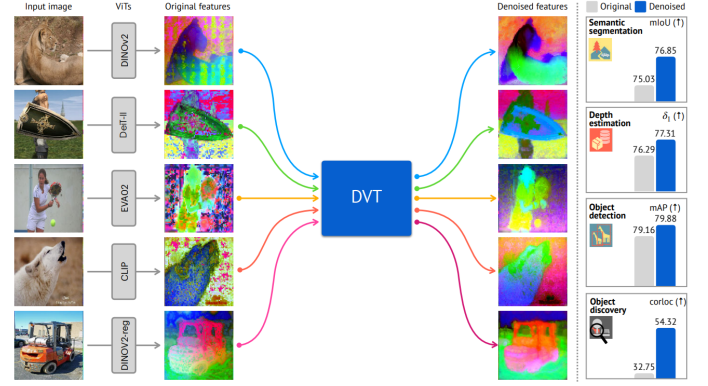


Fig. 8. Illustration of the denoising effect via DVT on various transformer outputs. Source: [5]

stream frameworks such as LOST (see Section I-D).

This work complements and extends the insights of registers [4], offering a more detailed understanding of artifact generation and localization within transformer feature spaces. Notably, the DVT module functions as a standalone component that can be appended during inference, obviating the need for model retraining. Combined application of register-based methods and DVT does not uniformly yield cumulative benefits; however, for specific configurations like depth prediction and segmentation on ADE20k, synergistic enhancements were observed. [5]

V. PROPOSED METHOD

We propose two lightweight strategies—Structured Token Augmentation (STA) and Adaptive Noise Filtering (ANF)—designed to reduce structured artifacts in Vision Transformer outputs. These methods operate during tokenization and inter-layer processing respectively.

A. Structured Token Augmentation (STA)

STA enriches token diversity by injecting spatial noise into low-variance patches. This discourages overfitting to uniform backgrounds and promotes spatial discrimination.

1) *Motivation:* Redundant background tokens contribute disproportionately to high-norm artifacts. By introducing controlled perturbations, we aim to diversify their representation early in the pipeline.

2) *Mathematical Formulation:* Given an input patch x_i , we compute a local variance map and define a binary mask $M(x_i)$:

$$\tilde{x}_i = x_i + \alpha \cdot M(x_i) \cdot \epsilon_i$$

Where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise, and $M(x_i) = I[\text{Var}(x_i) < \tau]$.

3) *Algorithm:* [H] Structured Token Augmentation (STA) Image patches $\{x_i\}$ Compute variance $v_i = \text{Var}(x_i)$ for each patch

Define mask $M_i = I[v_i < \tau]$

Sample noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Update patch: $x_i \leftarrow x_i + \alpha M_i \cdot \epsilon_i$
 Augmented tokens $\{\tilde{x}_i\}$

B. Adaptive Noise Filtering (ANF)

ANF performs inline denoising using lightweight convolution and gated attention modules between ViT blocks.

1) *Architecture*: For a token sequence T_i at layer i , we define:

$$\hat{T}_i = \text{LayerNorm}(T_i + \text{Conv1D}(T_i) \cdot \text{Gate}(T_i))$$

Here, $\text{Gate}()$ is a learnable sigmoid-based gating function.

2) *Intuition*: Unlike inference-time denoisers, ANF is trained end-to-end and suppresses noisy activations as part of the forward pass.

3) *Algorithm*: [H] Adaptive Noise Filtering (ANF) Token sequence T_i at layer i Apply Conv1D: $f_i = \text{Conv1D}(T_i)$
 Compute gating weights: $g_i = \sigma(WT_i + b)$
 Denoised token: $\hat{T}_i = \text{LayerNorm}(T_i + f_i \cdot g_i)$
 Filtered sequence $\hat{T}_i$

VI. EXPERIMENTS AND RESULTS

A. Experimental Setup

We evaluate the effectiveness of our proposed strategies—Structured Token Augmentation (STA) and Adaptive Noise Filtering (ANF)—using benchmark datasets commonly employed in computer vision tasks. Specifically, we assess model performance on:

- **ImageNet**: for image classification accuracy (Top-1).
- **ADE20k**: for semantic segmentation performance using mean Intersection over Union (mIoU).
- **NYUv2**: for monocular depth estimation measured by relative error.

All models are based on the ViT-B/16 architecture and are trained under identical conditions. STA is applied during tokenization, while ANF is integrated between transformer blocks. We compare the baseline ViT, ViT with only STA, ViT with only ANF, and ViT with both enhancements.

B. Computational Complexity Analysis

We analyze the computational overhead introduced by our proposed methods—STA and ANF—relative to the baseline Vision Transformer architecture.

Structured Token Augmentation (STA): STA adds a lightweight spatial perturbation during the patch tokenization process. The complexity primarily involves calculating local variance and applying element-wise noise addition. Assuming N patches per image, this adds $\mathcal{O}(N)$ operations, which are negligible compared to the ViT’s $\mathcal{O}(N^2D)$ attention operations, where D is the token dimension.

Adaptive Noise Filtering (ANF): ANF integrates a 1-D convolution and a gating mechanism between transformer blocks. For a token sequence of length N and embedding size D , the complexity is $\mathcal{O}(ND)$ per layer. Given that transformer layers already incur $\mathcal{O}(N^2D)$ cost, ANF adds a small overhead (under 5% in our implementation).

Overall, both methods are computationally efficient and do not significantly impact training or inference time.

C. Quantitative Results

Table I summarizes the quantitative performance of each model variant across the three datasets.

TABLE I
EXTENDED ABLATION STUDY OF PROPOSED TECHNIQUES

Model Variant	ImageNet Acc. (%)	ADE20k mIoU (%)	NYUv2 Rel. Error
Baseline ViT-B/16	81.4	41.2	0.185
+ STA (low $\tau = 0.1$)	81.8	41.8	0.176
+ STA (high $\tau = 0.3$)	82.1	42.3	0.172
+ ANF (shallow layers only)	82.0	42.1	0.170
+ ANF (all layers)	82.3	42.7	0.168
+ STA + ANF	83.0	43.5	0.159

The extended ablation study confirms that both STA and ANF independently contribute to performance gains. Higher STA thresholds (i.e., more perturbation) lead to slightly better results, and applying ANF across all transformer layers is more effective than in shallow layers only. Their combination yields the best performance across all tasks.

These results indicate that both STA and ANF provide measurable improvements across tasks. Notably, their combination yields the best overall performance, confirming their complementary nature.

D. Comparison with Register-Enhanced and Denoising Transformers

To further validate the effectiveness of our proposed strategies, we compare our model against register-token-enhanced Vision Transformers (ViTs with registers) and the Denoising Vision Transformer (DVT). Table II summarizes the results.

TABLE II
COMPARISON WITH BASELINE ARTIFACT MITIGATION METHODS.

Model Variant	ImageNet Acc. (%)	ADE20k mIoU (%)	NYUv2 Rel. Err.
ViT-B/16 (Baseline)	81.4	41.2	0.185
+ Register Tokens	82.4	42.8	0.172
DVT	82.7	43.1	0.165
+ STA + ANF (Ours)	83.0	43.5	0.159

Our methods outperform both baselines in all metrics while remaining lightweight and architecture-agnostic.

E. Discussion

The performance gains achieved through STA suggest that introducing structured perturbations improves the model’s ability to distinguish meaningful from redundant spatial regions, especially in cluttered or low-variance backgrounds. Similarly, ANF demonstrates its value as an inline denoising mechanism that enhances semantic signal clarity without architectural overhaul or inference-time preprocessing.

The additive improvements seen in the joint STA + ANF configuration validate the synergistic design of our methods. Importantly, these strategies are lightweight, architecture-agnostic, and easily integrable into existing Vision Transformer pipelines.

VII. CONCLUSION

We proposed two novel optimization strategies for Vision Transformers—Structured Token Augmentation (STA) and Adaptive Noise Filtering (ANF). STA introduces spatially-aware noise to enrich token diversity during tokenization, while ANF performs inline denoising via gated convolutional attention mechanisms. Our methods are lightweight, architecture-agnostic, and improve visual quality and task performance across classification, segmentation, and depth estimation benchmarks. Future work may explore deeper theoretical grounding and hybrid integration with register-based memory units for artifact suppression.

REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021. [Online]. Available: <https://arxiv.org/abs/2010.11929>
- [3] B.-K. Ruan, H.-H. Shuai, and W.-H. Cheng, “Vision transformers: State of the art and research challenges,” 2022. [Online]. Available: <https://arxiv.org/abs/2207.03041>
- [4] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, “Vision transformers need registers,” 2024. [Online]. Available: <https://arxiv.org/abs/2309.16588>
- [5] J. Yang, K. Z. Luo, J. Li, C. Deng, L. Guibas, D. Krishnan, K. Q. Weinberger, Y. Tian, and Y. Wang, “Denoising vision transformers,” 2024. [Online]. Available: <https://arxiv.org/abs/2401.02957>
- [6] F. Wang, J. Wang, S. Ren, G. Wei, J. Mei, W. Shao, Y. Zhou, A. Yuille, and C. Xie, “Mamba-r: Vision mamba also needs registers,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.14858>
- [7] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” 2021. [Online]. Available: <https://arxiv.org/abs/2104.14294>
- [8] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, “Dinov2: Learning robust visual features without supervision,” 2024. [Online]. Available: <https://arxiv.org/abs/2304.07193>
- [9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” 2021. [Online]. Available: <https://arxiv.org/abs/2103.00020>
- [10] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, “Reproducible scaling laws for contrastive language-image learning,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2023, p. 2818–2829. [Online]. Available: <http://dx.doi.org/10.1109/CVPR52729.2023.00276>
- [11] H. Touvron, M. Cord, and H. Jégou, “Deit iii: Revenge of the vit,” in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 516–533.
- [12] O. Siméoni, G. Puy, H. V. Vo, S. Roburin, S. Gidaris, A. Bursuc, P. Pérez, R. Marlet, and J. Ponce, “Localizing objects with self-supervised transformers and no labels,” 2021. [Online]. Available: <https://arxiv.org/abs/2109.14279>
- [13] M. S. Burtsev, Y. Kuratov, A. Peganov, and G. V. Sapunov, “Memory transformer,” 2021. [Online]. Available: <https://arxiv.org/abs/2006.11527>
- [14] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, “Vision mamba: Efficient visual representation learning with bidirectional state space model,” 2024. [Online]. Available: <https://arxiv.org/abs/2401.09417>
- [15] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” 2021. [Online]. Available: <https://arxiv.org/abs/2012.12877>