

参赛学生姓名:

朱怡桦

中学:

南京外国语学校

省份:

江苏省

国家/地区:

中国

指导老师姓名:

李曙, 张亚

指导老师单位:

南京外国语学校, 东南大学

论文题目:

An Analysis of Visual Explanation for Segmentation Tasks:

Faithfully Presenting Semantic Quality in Vision Transformers

An Analysis of Visual Explanation for Segmentation Tasks: Faithfully Presenting Semantic Quality in Vision Transformers

Abstract

Semantic segmentation is the foundation of fine-grained perception. Pre-trained Vision Transformers (ViTs) have become popular backbones for segmentation, yet it remains challenging to evaluate their task-specific feature encoding capability (class-discriminable semantic quality) via feature maps. A major bottleneck is that the trustworthiness of CAM-based spatial attribution heatmaps for segmentation cannot be quantitatively measured. This prevents unified association between attribution results and feature semantic quality, hindering fair backbone comparison and limiting the development of interpretable research.

To address the inherent trustworthiness defects of existing CAM visualization schemes for dense segmentation tasks, we systematically analyze and optimize reliable attribution visualization and quantification pipelines. We first propose a novel evaluation paradigm termed **“feature PCA visualization supervising trustworthiness of attribution”**, which unifies intuitive feature PCA visualization (external feature manifestation) and rigorous Shapley Value reasoning (internal attribution essence) to build a complete, task-specific trustworthiness criterion for segmentation attribution maps. Built upon this paradigm, we further develop two core innovations to remedy the missing trustworthiness constraints in current CAM systems. First, we design Grad-CAM-See, a segmentation-friendly CAM visualization module that explicitly integrates Shapley Value reasoning to eliminate the inherent attribution unfaithfulness of conventional CAM variants. Second, we construct the T-PCA white-box quantification framework, which directly takes feature PCA visualization transformation as the core semantic inference line to regularize and measure attribution quality. Combined with the SSIM metric, T-PCA complements the defective attribution evaluation system of existing CAM methods, enabling systematic and fine-grained quantification of attribution reliability that is absent in current qualitative-only CAM evaluation pipelines.

Extensive experiments verify the effectiveness of our method. Grad-CAM-See generates more trustworthy heatmaps than existing CAM approaches. Our method accelerates denoising model convergence by 20% while maintaining semantic compactness and supports cross-backbone semantic quality comparison. This work demonstrates the benefits of integrating feature PCA visualization and attribution trustworthiness analysis, offering interpretable theoretical support and practical solutions for ViT backbone evaluation in segmentation.

Keywords: Vision Transformers, semantic segmentation, PCA visualization, attribution trustworthiness, semantic quality, backbone evaluation

Contents

Introduction	3
Related Work	5
PCA in ViTs: Reflecting Feature Quality	5
Why PCA Alone Is Not Enough– Bias from Downstream Tasks	6
Current Situation: Seg CAM Without Spatial Trustworthiness	7
Toward a Bridge: CAM+PCA as a Mirror of TaskOriented Quality	9
Methodology	10
Grad-CAM-See: Shapley-Augmented Seg-CAM	11
T-PCA: Transformed PCA for Spatial Semantic Quality	12
Denoising ViT Extension: CAM Augmented ViT Denoising	13
Experiments and Analysis	15
Baseline Evaluation with ROAD	16
Visual PCABased Attribution Evaluation	17
Quantitative Evaluation Based on TPCA Semantic Strength	18
Trustworthy CAM Heatmap Utility: Evaluation and Analysis	20
Conclusion	26
References	27
致谢	29

Introduction

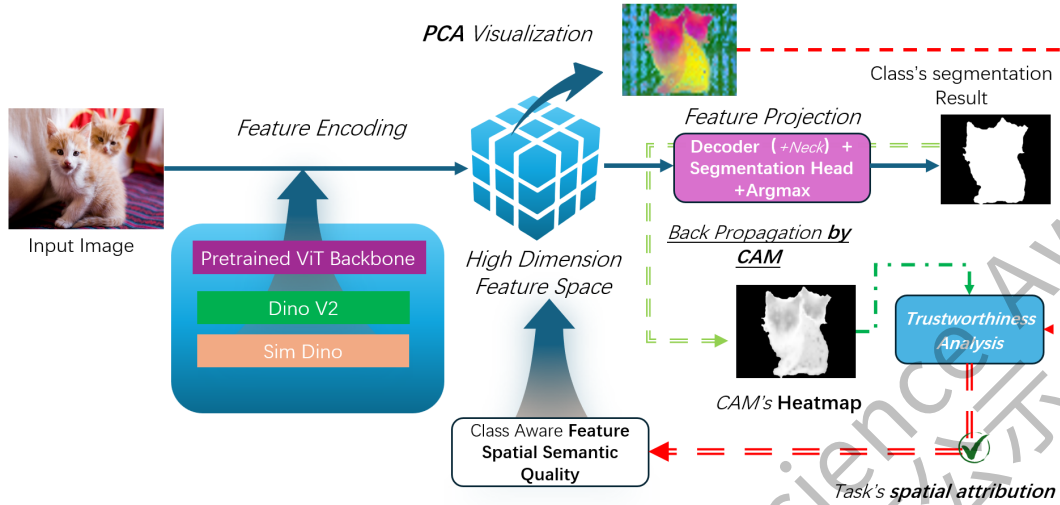


Figure 1: Illustration of the segmentation pipeline, feature visualization and attribution workflow: bridging backbone feature quality and spatial attribution trustworthiness

Semantic segmentation serves as a fundamental technique for fine-grained perception across diverse artificial intelligence scenarios. Vision Transformers (ViTs) [1], benefiting from outstanding scalability and powerful capabilities to model long-range contextual dependencies, have become the dominant backbone networks for semantic segmentation. To further exploit the representation potential of ViTs, researchers have proposed various training paradigms including supervised learning [2], self-supervised learning [3] and masked image modeling [4], which gradually evolve into two distinct branches of pre-training technologies. The first branch consists of black-box pre-trained models represented by DinoV2 [5] and the recent DinoV3 [6]. Trained on massive web datasets, these models feature opaque internal mechanisms for feature generation and constraint. The second branch refers to white-box frameworks exemplified by White-Box Transformers [7], [8], [9], whose feature learning and constraint processes follow explicit rules. Such frameworks, represented by the recent SimDino [10], effectively reduce hyperparameter dependence and improve model robustness as well as interpretability.

Currently, segmentation pipelines built upon pre-trained ViT backbones have emerged as a mainstream research direction. Nevertheless, with the rapid iteration of ViT architectures and pre-training strategies, how to quantitatively evaluate the category-discriminable semantic quality of backbones for segmentation tasks remains an open challenge. Specifically, standardized evaluation based on feature maps — the universal output of all backbone networks — has not been well addressed. The core bottleneck lies in the fact that spatial attribution heatmaps generated by Class Activation Mapping (CAM) [11] in segmentation scenarios lack quantifiable credibility criteria.

Existing CAM methods and evaluation frameworks still suffer from such notable limitations. The recently proposed Seg-Eigen-CAM [12] for segmentation suffers from incomplete spatial attribution coverage and obvious semantic localization bias. Meanwhile, prevailing evaluation protocols for CAM credibility (e.g., ROAD) [13] are not tailored to the spatial semantic attribution of segmentation tasks, and thus cannot reliably quantify the ground-truth credibility of attribution results in segmentation settings.

This deficiency leads to the absence of unified mapping among attribution heatmaps, spatial attri-

bution intensity, and the semantic quality encoded within backbone features. It not only impedes fair horizontal comparisons among different pre-trained backbones, but also hinders the advancement of segmentation interpretability research and its derivative applications.

To tackle the above issues, this paper conducts systematic analysis and optimization toward high-fidelity CAM visualization pipelines for segmentation tasks.

First, we propose a novel perspective for evaluating the credibility of interpretability results: utilizing denoised feature visualization to supervise the attribution fidelity of CAM heatmaps. This perspective leverages the neighborhood equivalence prior derived from feature visualization, and establishes computable credibility constraints based on Shapley Values.

Second, we propose Grad-CAM-See, a dedicated high-credibility CAM visualization method for semantic segmentation. Taking feature PCA visualization analysis as the entry point and Shapley Value consistency constraints as the theoretical optimization foundation, this method constructs a closed-loop interpretable system connecting backbone feature quality and downstream segmentation performance. The main contributions of this method are summarized as follows:

1. We verify the rationality and effectiveness of the perspective that feature visualization supervises attribution credibility;
2. We demonstrate that introducing Shapley Value constraints can substantially mitigate semantic bias inherent in conventional CAM and improve heatmap credibility;
3. We bridge the separation between task-level attribution and feature quality analysis, uncover the intrinsic correlation between feature PCA visualization and Shapley-based attribution constraints, and provide feasible solution to achieve consistent mapping between backbone feature quality and task attribution results.

Furthermore, to realize quantitative evaluation of attribution credibility, we extend the idea of supervision via feature visualization and innovatively propose T-PCA, a forward white-box semantic intensity representation and quantification framework for feature semantic spaces. T-PCA approximates the segmentation inference process into explicit transformation of backbone features PCA visualization map. Combined with the spatial consistency metric SSIM, T-PCA can quantitatively measure the spatial attribution credibility of different CAM methods.

Finally, we design two progressive experiments to demonstrate the derived value of our novel credibility-aware method and its outputs. Feature denoising experiments show that heatmaps generated by Grad-CAM-See can accelerate the convergence of the feature denoising model of the classical DVT paradigm [14] by 20% while preserving the compactness of semantic regions. Cross-backbone experiments demonstrate that our method enables fair comparisons of feature semantic quality among ViT backbones with different pre-training mechanisms (black-box and white-box), offering reliable support for backbone selection.

In summary, this paper carries out comprehensive and systematic analysis and methodological optimization for visual interpretability in segmentation tasks. We verify the core value of attribution credibility analysis guided by feature visualization under segmentation scenarios, and provide solid theoretical support and practical schemes for interpretable evaluation and real-world deployment of pre-trained ViT backbones.

Related Work

PCA in ViTs: Reflecting Feature Quality

Vision Transformers (ViTs) have become powerful backbones for dense prediction tasks such as semantic segmentation, depth estimation, and optical flow. Unlike image classification, which only predicts “*what is in the picture*,” dense prediction requires pixellevel understanding —answering “*where exactly is each object, and how do they relate spatially*.” This capability is crucial in applications like autonomous driving, medical imaging, and extended reality (AR/VR), where accurate spatial reasoning ensures safety and realism [15], [16].

Vision Transformers (ViTs) are particularly suited for semantic segmentation because their self-attention mechanism captures long-range dependencies across the entire image. This global contextual modeling enhances semantic consistency in segmentation outputs, helping ensure that spatially neighboring pixels are assigned coherent labels, which is critical for maintaining semantic quality. Recent selfsupervised approaches such as DINOv1, DINOv2, and DINOv3 show that ViTs can produce feature maps with emergent segmentation ability, even without explicit labels. Frameworks like SegDINO [17] further demonstrate that frozen backbones combined with simple heads can achieve strong segmentation, highlighting the backbone’s central role in dense prediction.

Within a segmentation pipeline [18], the backbone generates feature maps, which are then refined by a neck and projected by a head. The *feature map is the backbone’s only true product*, and its quality directly determines whether segmentation predictions are accurate and complete. Researchers therefore need diagnostic tools to evaluate feature maps before relying on downstream modules.

One of the most effective tools is Principal Component Analysis (PCA). PCA projects highdimensional patch embeddings into a lowerdimensional space, revealing whether features are clustered and semantically coherent. In practice, PCA visualizations of ViT feature maps often show distinct clusters corresponding to object parts—such as wings, heads, or legs—colored differently. A smooth, unified PCA plot suggests strong semantic coherence, while scattered or noisy plots indicate unstable representations.

Recent studies confirm that PCA coherence is closely tied to downstream performance. DINOv3 discovered that maintaining finegrained spatial properties in feature maps improves generalization, with PCA projections aligning consistently with object boundaries. Denoising ViT [14] further showed that largescale training introduces noise, visible in PCA as fragmented clusters. By enforcing neighbor consistency and denoising, PCA plots become smoother, and segmentation results improve. As illustrated in Figure 2, PCA visualizations from Denoising ViT clearly demonstrate how denoising improves semantic quality: colors inside clusters become more consistent, and object boundaries align more clearly.

These findings highlight PCA visualization as more than a convenient picture—it has become an explicit *diagnostic lens* for judging feature quality. Moreover, PCA visualization’s color consistency within clusters and alignment with object edges are now recognized as important indicators of semantic quality. Preserving these attributes is essential for reliable segmentation, making PCA visualization a standard tool for analyzing ViT feature maps. As pretrained models grow larger and more complex, PCA provides a clear and accessible way to evaluate whether a backbone is truly ready to support accurate and complete dense prediction tasks.

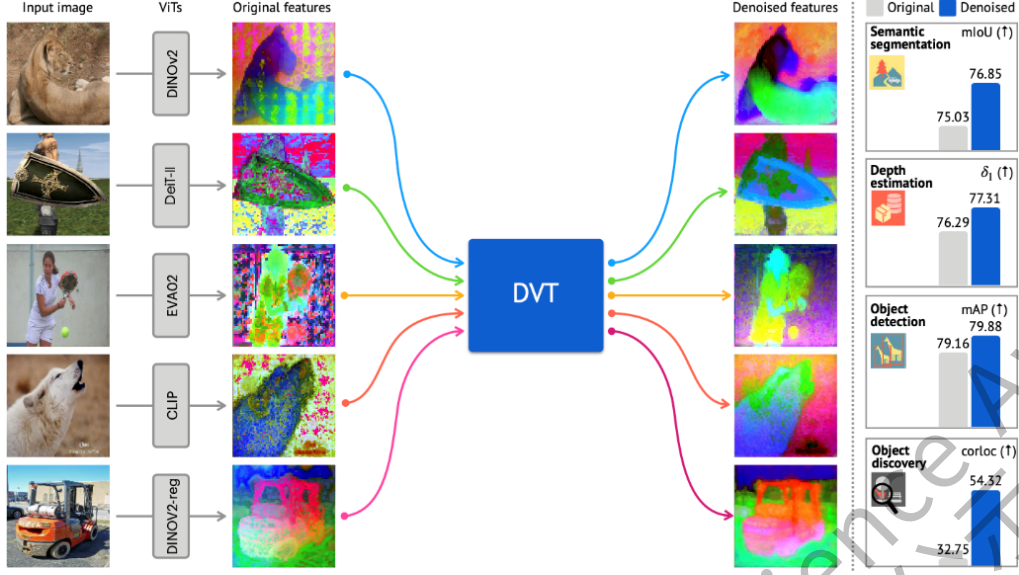


Figure 2: PCA visualization from Denoising Vision Transformers [14]. The figure shows how denoising improves feature map quality: colors inside clusters become more consistent, and object boundaries align more clearly. Such PCA visualizations provide an intuitive diagnostic lens for evaluating semantic quality of ViT feature maps

Why PCA Alone Is Not Enough– Bias from Downstream Tasks

PCA is powerful, but it is not sufficient. A critical challenge remains unresolved: it is difficult to directly assess the quality of backbone feature maps and their relationship to final segmentation results. Importantly, this “quality” is not simply the native PCA quality of frozen features, but rather the task-oriented quality represented by CAMs in class-aware semantic space.

Segmentation tasks ultimately produce class-aware semantic clusters. In contrast, PCA only reveals class-agnostic variance structures, capturing local semantic consistency without enforcing category unification. Within the prediction pipeline, the segmentation head and argmax operation play decisive roles: I) It enforces categorical uniformity across the mask; II) It suppresses minor inconsistencies in local features; Thus, while PCA provides a useful diagnostic of backbone embeddings, it cannot directly represent the task-level semantic quality required for segmentation.

This difference is illustrated clearly in Figure 3:

- In cats’ example, *PCA partitions show smooth transitions*, yet all regions belong to the same category. The final segmentation mask correctly unifies them.
- Also in cats’ example, *PCA visualizations contain artifacts or inconsistencies*, but the final mask remains accurate, showing that argmax ensures category consistency.

To bridge input features and task outputs, Explainable AI (XAI) provides a promising direction. Among these, Class Activation Maps (CAMs) have become widely studied visualization tools. Mature frameworks such as PyTorch-Grad-CAM [13] integrate several recent CAM variants and evaluation protocols, including: Grad-CAM [19], LayerCAM [20], HiResCAM [21], and the latest Seg-Eigen-CAM.

The basic formulation of CAM, distinguished from PCA, is given as follows:

$$M_c(x, y) = \sum_k w_k^c f_k(x, y) \quad (1)$$

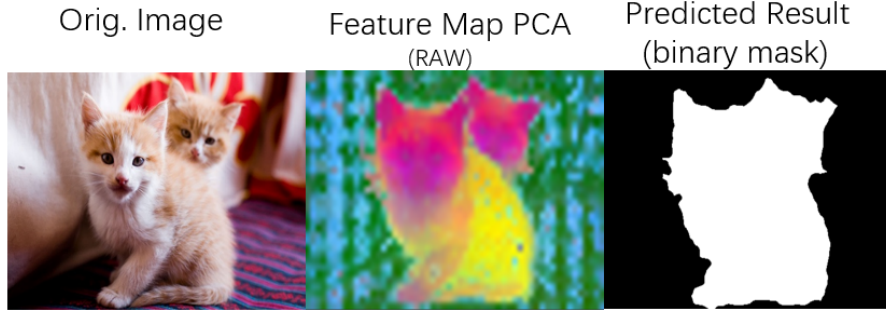


Figure 3: From the original image to feature PCA visualization and finally to the segmentation mask. The PCA visualization shows feature partitions with smooth transitions or local artifacts, but the final mask unifies categories and maintains consistency

Where $f_k(x, y)$ is the activation of feature map k at location (x, y) , and w_k^c is the weight for class c . CAM highlights regions most responsible for a class decision, producing taskdriven saliency maps.

- PCA emphasizes global variance and classagnostic clustering.
- CAM emphasizes taskspecific attribution strength and classaware saliency.

For the challenge, the quality of backbone feature maps and their relationship to final segmentation results, the segmentation pipelines (e.g., Denoising ViT) can be viewed as feature maps processed through BN decoders and CNN segmentation heads. Applying CAM variants such as Seg-Eigen-CAM to these feature maps enables classaware semantic quality analysis in attribution space.

Following this trajectory, CAMs provide a potential solution to the unresolved challenge: they connect backbone features to task outputs, complementing PCA’s diagnostic role with taskoriented attribution.

Current Situation: Seg CAM Without Spatial Trustworthiness

Existing CAM methods often struggle to faithfully reflect the spatial semantic quality in segmentation tasks. The Figure 4 clearly illustrates this point by comparing the original image, a denoised feature map PCA visualization, and CAM heatmaps (Grad-CAM and Seg-Eigen-CAM) in a segmentation pipeline. The original image shows two kittens. The denoised feature map PCA visualization reveals compact and consistent clusters, particularly in the head and body regions, indicating stable semantic structure. In contrast, the Grad-CAM and Seg-Eigen-CAM heatmaps fail to provide fair spatial semantic quality feedback: their attributions are fragmented and do not fully align with the compact PCA clusters. This highlights a *key finding*—while PCA denoising yields coherent feature representations, existing CAM methods lack trustworthiness in spatial attribution, underscoring the need for more faithful explanation frameworks.

The main reason is, firstly, that most existing CAM methods such as Grad-CAM and LayerCAM were originally designed for image-level classification (with only recent methods like Seg-Eigen-CAM targeting segmentation). They mainly focus on locating salient features. Although the PyTorch-Grad-CAM framework has introduced adaptations for segmentation, it has not deeply investigated the credibility of the adapted results. Secondly, including Seg-Eigen-CAM, these approaches remain largely heuristic visualizations. They can highlight contributions of certain regions but cannot guarantee fairness across critical pixels or neighborhood equivalence. In semantic segmentation, such fairness is crucial because

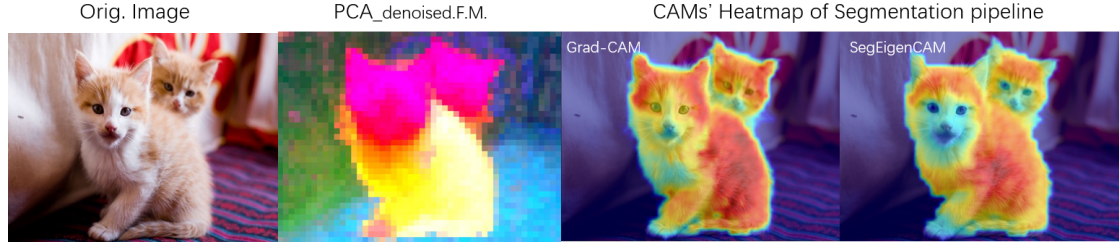


Figure 4. Comparison of original image, denoised feature map PCA visualization, and CAM heatmaps in a segmentation pipeline

the semantic quality of each pixel —namely, how each pixel contributes to the class boundary—is the foundation of an accurate mask. This also aligns with commonsense priors derived from visualization. Without neighborhood equivalence, CAMs may exaggerate or underrepresent certain regions, thereby distorting spatial semantics trustworthiness.

Recently, research on the computational completeness of CAM has sought to overcome the limitations of earlier heuristic approaches. Shapley-CAM [22] addresses these shortcomings by reframing CAM attribution within the perspective of cooperative game theory. Although it still focuses primarily on image classification, it provides a mathematically rigorous framework for attribution analysis. More importantly, it offers a principled way to fairly distribute contributions among all participating data points. According to its definition, the marginal contribution of each pixel is evaluated across all possible pixel combinations, ensuring that no pixel is overlooked.

However, the importance of Shapley Values in segmentation tasks has not yet been explicitly addressed in prior research. We find that this is partly because Shapley-CAM has focused on building mathematical completeness based on explicit fairness principles, rather than on how fairness should be manifested and assessed.

Moreover, current CAM computation frameworks lack mechanisms to evaluate attribution reliability in segmentation tasks. For example, the widely used PyTorch-Grad-CAM framework relies on ROAD (Remove and Debias) as an evaluation metric. If we examine ROAD more closely, its principle is to rank input data according to CAM heatmap intensity, then progressively mask or perturb these data while recording the decrease (deletion) or increase (insertion) of class logits. While this method is effective and widely applied in classification tasks, it is insufficient in segmentation scenarios.

The core reason is that it does not reflect spatial faithfulness. Segmentation decisions are made pixel by pixel, and the results are expressed in the predicted foreground mask (with logits normalized). ROAD can verify whether high CAM values overlap with the predicted foreground, but it cannot distinguish whether attribution within the mask is uniformly reliable or scattered and noisy. Consequently, if the predicted segmentation mask is correct and overlaps with the CAM heatmap, the ROAD score may remain high despite noise or uneven distribution. This incompleteness explains why fairness and spatial consistency have long been overlooked: existing evaluation frameworks do not enforce them.

Our study reveals that uncovering the significance of Shapley Values in segmentation is far more critical than in image classification. In segmentation, the meaning is clear: every pixel’s role in shaping the class boundary is fairly considered, preventing attribution maps from exaggerating or underrepresenting certain regions. In this way, Shapley Values establish a principled mechanism that ensures fairness and neighborhood consistency among critical pixels. This also guarantees coherence between input features

and attribution results, allowing spatial semantics to be faithfully represented in the segmentation mask. At the same time, a new question arises: how can we intuitively demonstrate and evaluate this trustworthiness, and how can we further explore the role of Shapley Values in enhancing the trustworthiness and mathematical completeness of CAMs for segmentation?

Toward a Bridge: CAM+PCA as a Mirror of TaskOriented Quality

Recent research on feature map denoising (such as *Denoising Vision Transformers* [14]) has opened an important opportunity for this study. DVT shows that denoised feature maps, when visualized with PCA, exhibit water-tight, compact, and locally linear properties, which represent a set of participants with approximately equalstatus features. From another perspective, the PCA visualization of raw feature maps can be seen as clean features (true attribution signals) overlaid with noisy features (attribution noise), where the latter introduces imperfections. The PCA visualization of clean features directly reflects neighborhood semantic quality consistency and neighborhood equivalence. When combined with Shapley Values—which in segmentation scenarios ensure fairness and neighborhood consistency in spatial attribution—this becomes a crucial measure. It provides a trustworthy reflection of the alignment between CAM outputs (heatmaps) and true attribution maps, and serves as a reliable indicator of semantic quality in feature space.

“For clean and denoised features as revealed by PCA visualization, a trustworthy CAM heatmap should exhibit approximately consistent attribution intensity across all similar linear neighborhoods and fully cover these regions.”

The PCA visualization of denoised feature maps provides an excellent way to explicitly reflect the mechanism of Shapley Values. At the same time, when combined with the practical process of Shapley Value computation, the PCA visualization of raw feature maps can serve as a metaphor for Shapley Values, offering feedback on neighborhood feature/attribution consistency and acting as an important means of supervising the trustworthiness of CAM results. The specific analysis and derivation are as follows:

In practice, Shapley Value computation (as in Shapley-CAM) relies on approximations using Hessian Vector Products (HVPs) to capture second-order interactions between gradients and directions in the feature space. For a scalar loss function $L(\theta)$ and a vector v , *HVP* is defined as:

$$Hv = \nabla^2 L(\theta)v = \nabla_{\theta}(\nabla_{\theta} L(\theta)^T v) \quad (2)$$

Mathematically, HVP exhibits input and output additivity. For two vectors v_1 and v_2 :

$$H(v_1 + v_2) = \nabla^2 L(\theta) \cdot (v_1 + v_2) = Hv_1 + Hv_2 \quad (3)$$

This linear property allows feature map decomposition and correspondence (via PCA visualization, as discussed earlier) to be directly integrated into Shapley Value computation.

Within the Denoising ViT (DVT) setup, raw feature maps can be decomposed as:

$$F_{\text{raw}} = F_{\text{good}} + F_{\text{noisy}} \quad (4)$$

Where F_{good} represents clean components (dominant, task-aligned directions observable through PCA visualization), and F_{noisy} represents noisy components (variance from artifacts or disturbances, also

identifiable via PCA visualization). Using *HVP* additivity:

$$H(F_{\text{raw}}) = H(F_{\text{good}} + F_{\text{noisy}}) = H(F_{\text{good}}) + H(F_{\text{noisy}}) \quad (5)$$

Thus, Shapley attribution can be expressed as:

$$\phi(F_{\text{raw}}) = \phi(F_{\text{good}}) + \phi(F_{\text{noisy}}) \quad (6)$$

This decomposition shows that clean components strengthen fairness attribution, while noisy components reveal instability.

This further illustrates that PCA visualization of raw features can provide important cues: whether clustering compactness reflects strict spatial neighborhood consistency, and whether deviations may trigger attribution instability. In turn, PCA visualization becomes a crucial supervisory tool for assessing the trustworthiness of CAM interpretability results. It offers a direct way to judge whether CAM outputs in segmentation scenarios faithfully reflect attribution and semantic quality.

In summary: Leveraging PCA visualization of pretrained features as supervisory guidance for Shapley-based CAM computation is both natural and necessary. PCA visualization reveals the compactness and closeness of local neighborhoods in the feature space, while Shapley attribution evaluates how each PCA component contributes fairly to the final prediction. Together, they form a dual lens—PCA as unsupervised structure (clean vs. noisy directions), and Shapley as supervised fair attribution across those directions. Therefore, CAM feedback becomes more faithful to true spatial semantics.

Methodology

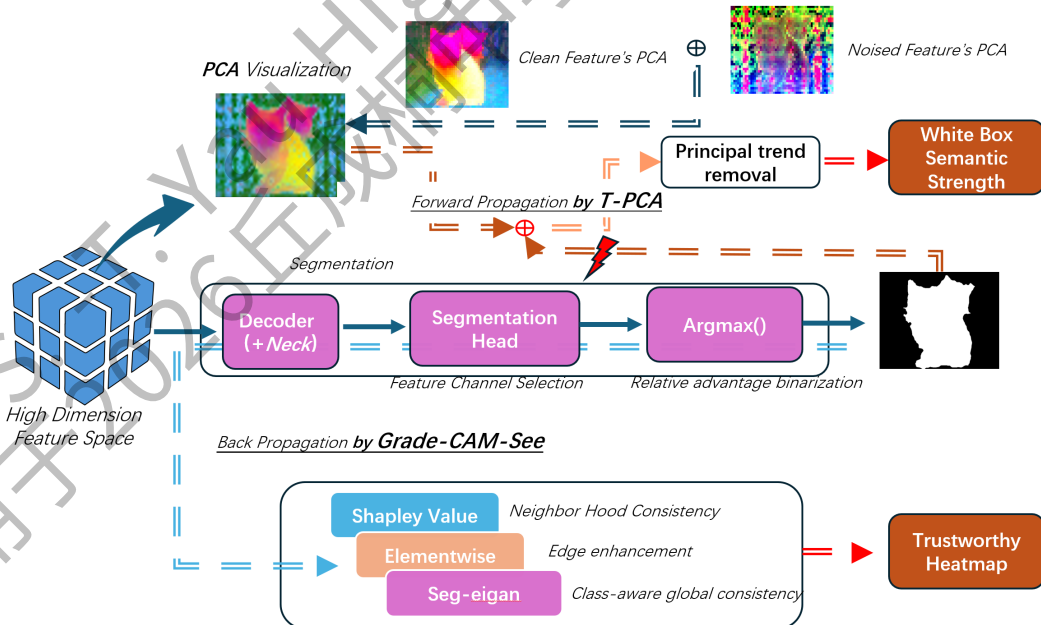


Figure 5. Pipeline of the proposed trustworthy analysis framework: the novel trustworthy CAM's heatmap generation and forward white-box semantic strength derivation via T-PCA

Grad-CAM-See: Shapley-Augmented Seg-CAM

To validate the rationality of the proposed perspective and demonstrate that integrating Shapley Value computation can significantly enhance the trustworthiness of existing CAM heatmaps, our goal is to create a novel CAM-based interpretable visualization method specifically designed for segmentation tasks. The starting point of our framework is to adopt Shapley Values as the theoretical foundation to ensure fair and consistent attribution across feature channels. On top of this, we integrate the strengths of classic CAM variants to construct a unified, robust attribution pipeline for semantic segmentation. The full computation proceeds in three fused and structure-preserving stages.

First, Shapley-aware gradient correction performs fairness-based adjustment on raw backpropagated gradient tensors. Let $G \in R^{C \times H \times W}$ denote the original backward gradient, $HVP(\cdot)$ denote the approximate Shapley estimation, and λ denote the correction coefficient. The corrected gradient $G^\sim$ is formulated as:

$$G^\sim = G - \lambda \cdot HVP(G) \quad (7)$$

This operation suppresses gradient components from irrelevant channels and mitigates abrupt noise, aligning with the neighborhood consistency prior for reliable segmentation attribution.

Second, we conduct element-wise multiplication between the corrected gradient and backbone feature activations to preserve dense pixel-level spatial attribution. Let $A \in R^{C \times H \times W}$ be the feature activations and $\odot$ denote element-wise multiplication. The weighted activation is further regularized by a *ReLU* truncation to filter negative and uninformative responses:

$$E = \max(G^\sim \odot A, 0) \quad (8)$$

Positive gradient entries highlight class-related decision boundaries and foreground support, while zeroed negative entries discard uninformative negative responses. This element-wise paradigm strengthens boundary details and ensures that the heatmap precisely encodes “where decisions are made”.

Finally, Seg-Eigen-CAM is adopted as the fundamental module to produce the final fused attribution heatmap. In the previous step, the gradient correction and elementwise weighting steps are structurepreserving transformations. Although *ReLU* truncation introduces nonlinear clipping, the intrinsic feature manifold of backbone outputs is largely retained. Following the Eigen paradigm, we extract the first principal component with sign correction from the refined weighted activation tensor E , and perform 2D projection to capture the dominant semantic activation mode, yielding the final attribution output.

Benefiting from approximateShapleybased cross-channel fairness constraint, element-wise spatial reinforcement, and eigen-based global semantic consistency extraction, the resulting heatmap achieves spatially continuous, semantically coherent, and gradient-noise-robust visual attribution for dense segmentation tasks.

Compared with Seg-Eigen-CAM, the proposed Grad-CAM-See is an extended version that incorporates Shapleyinspired weighting and element-wise gradient multiplication. It integrates local semantic consistency and global contextual coherence constraints to adapt to the characteristics of dense prediction tasks. This method generates high-fidelity activation heatmaps that accurately reflect the intrinsic semantic quality of input feature maps, transforming implicit feature semantics into interpretable and intercomparable visual and numerical representations.

T-PCA: Transformed PCA for Spatial Semantic Quality

Built upon the segmentation-oriented Grad-CAM-See for trustworthy attribution analysis, we propose T-PCA (Transformed PCA), a lightweight white-box evaluation metric tailored for segmentation tasks, to establish a complete closed-loop validation system.

A lightweight DVT-style segmentation pipeline is adopted for the target segmentation task, consisting of a ViT backbone, a *BN* decoder, a segmentation head with 1×1 convolutional layer and a *SoftMax*, and an additional *Argmax* prediction head. Rather than simple discrete channel selection, the 1×1 convolution implements task-driven channel fusion and feature projection via weighted linear combination of multi-channel inputs, formulated as $y = Wx + b$, which converts class-agnostic backbone features into category-aware semantic channels.

PCA visualization of ViT backbone features inherently encodes neighborhood equivalence, clustering compactness and object boundary alignment, faithfully reflecting the latent quality of segmentation input features. Our T-PCA method takes raw PCA visualizations as input and simulates the above segmentation inference via targeted semantic transformation to produce quantitative spatial intensity distributions.

The core principle of T-PCA is “*PCA visualization + semantic mask constraint*”, and the full procedure follows a three-stage workflow as follows:

Step 1: Foreground Masking (Task-Related Channel Screening)

Following the segmentation paradigm, the 1×1 convolution optimizes features via cross-channel weighted fusion and task-oriented projection. we first adopt binary segmentation masks to extract task-relevant foreground regions from raw PCA visualizations, screening target semantic channels and eliminating background interference. The foreground feature set is defined as:

$$X_{fg} = \{p_i \in \text{pca}_{rgb} \mid \text{mask}(i) = 1\} \quad (9)$$

The masked foreground still contains heterogeneous class-agnostic feature clusters with inter-channel directional imbalance, requiring further statistical regularization for task-aligned semantic fusion.

Step 2: PCA Redundancy Removal and Target Class Channel Feature Construction

To mimic the linear channel fusion and projection of the 1×1 convolution, we compute the global foreground mean and remove redundant global trends, retaining only the dominant differences among heterogeneous feature clusters:

$$m = \frac{1}{|X_{fg}|} \sum_{p \in X_{fg}} p, \quad X'_{fg} = X_{fg} - m \quad (10)$$

We then project the refined foreground features onto the first principal component to extract dominant inter-channel directional imbalance:

$$\hat{X} = P_1(X'_{fg}), \quad X_{proj} = \hat{X} + m \quad (11)$$

Where, $\hat{X}$ denotes the PCA-reconstructed structural feature and X_{proj} denotes the mean- restored feature map.

Furthermore, cross-cluster directional imbalance is eliminated with residual fused features $X_{fg} - \hat{X}$, to achieve implicit cross-channel linear fusion. And also, we normalize the resulted feature range as below:

$$X_{reg} = (X_{fg} - \hat{X}) \cdot \frac{X_{proj} - \min(X_{proj})}{\max(X_{proj}) - \min(X_{proj})} \quad (12)$$

This recalibrates unbalanced channel responses and unifies scattered class-agnostic features into task-aligned representations. Only global redundant trends are removed, while fine-grained local neighborhood variations are fully preserved to maintain intra-region structural consistency.

With the above transforms, T-PCA and 1×1 convolution follow the same linear transformation formulation $y = Wx + b$, projecting features into a task-aligned channel space while preserving intrinsic spatial structural relationships. The key difference lies in parameter acquisition: convolutional weights are end-to-end task-learned, while T-PCA achieves channel fusion purely through intrinsic feature statistics without learnable parameters. Accordingly, T-PCA serves as a parameter-free statistical approximation of CNN-based semantic reconstruction in segmentation heads.

Step 3: Weighted Channel Distance Calculation and Grayscale Strength Map Generation

Based on the fully regularized feature map X_{reg} , we compute its global feature mean as the unified task-aligned class center c_{main} , which encodes the dominant targetclass semantics after statistical regularization:

$$c_{main} = \frac{1}{|X_{reg}|} \sum_{p \in X_{reg}} p \quad (13)$$

To achieve precise and unbiased quantification of spatial semantic discrepancies, we adopt an RGB-adaptive weighted channel distance function. This adaptive measurement mitigates inherent dimensional imbalance among RGB channels and provides more reliable evaluation than standard Euclidean distance. The weighted distance between each spatial feature and the unified class center is calculated to generate the final T-PCA strength map s_i :

$$s_i = \sqrt{2(R_i - R_{main})^2 + 4(G_i - G_{main})^2 + 3(B_i - B_{main})^2} \quad (14)$$

where (R_i, G_i, B_i) denote RGB values of individual spatial features presented in PCA visualization, and $(R_{main}, G_{main}, B_{main})$ correspond to channel values of the unified class center c_{main} .

Finally, by transforming unconstrained, class-agnostic PCA visualization into category-biased, numerically measurable semantic strength representations, T-PCA enables quantitative evaluation of CAM attribution trustworthiness for segmentation tasks.

Denoising ViT Extension: CAM Augmented ViT Denoising

To verify that the Grad-CAM-See-generated heatmaps are not merely visualization results but effective semantic supervisory signals for feature optimization, we propose a heatmap-augmented feature denoising framework based on the classic DVT paradigm. This framework introduces CAM heatmap as semantic guidance to assist feature denoising training, accelerating model convergence while maintaining stable semantic compactness and structural integrity of denoised features.

The original DVT framework decomposes noisy visual features into three components: clean semantic features, global bias noise, and position-variant local artifacts, following the formulation below:

$$F = F_{\text{clean}} + N_{\text{global}} + N_{\text{pos}}(x) \quad (15)$$

Where F_{clean} denotes noise-free target semantic features constructed by a NeRF network [23], N_{global} represents spatially uniform global bias noise that causes holistic semantic deviation, and N_{pos} indicates position-dependent artifacts that break local neighborhood consistency and fine-grained structural continuity.

DVT adopts a two-stage progressive denoising strategy to eliminate hierarchical noise components:

Stage 1 (Global Noise Removal, without residual predictor): This stage removes global bias noise, and the predicted feature is formulated as:

$$\hat{F}_{\text{raw}} = F_{\text{denoised}} + P_{\text{shared}} \quad (16)$$

Stage 2 (Local Artifact Removal, with residual predictor): Based on the stage 1 outputs, this stage further suppresses position-dependent residual noise, with the complete prediction formulated as:

$$\hat{F}_{\text{raw}} = F_{\text{denoised}} + P_{\text{shared}} + R_{\text{pred}} \quad (17)$$

Each decomposed component corresponds to a specific optimization target in staged denoising. F_{denoised} denotes valid semantic features with inherent secondary spatial continuity, characterizing compact and smooth foreground object structures. P_{shared} represents globally shared noise that induces holistic semantic bias, serving as the primary noise to be eliminated in stage 1. R_{pred} denotes position-variant residual noise that disturbs local spatial continuity, which is the core optimization target of stage 2.

The DVT framework adopts three standard and complementary loss functions for multi-dimensional denoising optimization, each undertaking a staged optimization task:

Patch L2 Loss is a standard numerical reconstruction loss that ensures basic feature amplitude fitting fidelity.

Cosine Similarity Loss constrains feature vector directional consistency to eliminate global semantic bias.

Residual Loss serves as a Stage-2-specific local constraint, eliminating position-dependent residual noise to preserve spatial structural continuity.

Based on the distinct noise characteristics and stage-specific loss properties, we design a two-stage CAM-enhanced DVT paradigm for the above denoising constraints.

Stage 1: CAM Weighted Cosine Similarity Loss

We integrate a CAM-weighted regularization term into the vanilla cosine similarity loss to enhance prior-assisted semantics-aware global alignment. The final loss function is defined as:

$$L_{\text{cos}}^{\text{final}} = L_{\text{cos}} + a(t) \cdot L_{\text{cam-weighted}} \quad (18)$$

where $a(t)$ denotes a time-varying progressive coefficient that adaptively regularizes training strength. The Stage-1 CAM-weighted loss is formulated as:

$$L_{\text{cam-weighted}} = \frac{1}{N} \sum_i D_i \cdot w_{\text{cam},i}, \quad D = (F_{\text{denoised}} - F_{\text{raw}})^2 \quad (19)$$

In Stage 1, guided by Grad-CAM-See heatmaps ($w_{\text{cam},i} \in G_{\text{map}}$), the composite loss $L_{\text{cos}}^{\text{final}}$ steers NeRF toward targetsemantic orientations for earlystage global semantic anchoring. The primary term L_{cos} corrects feature directional bias, while the CAMweighted loss $L_{\text{camweighted}}$, modulated by timedependent weight, imposes amplitude regularization over foreground regions to suppress uncontrolled feature magnitude drift during training.

Stage 2: CAM Smooth Loss

Upon the globally aligned features obtained from stage 1, we introduce a CAM smooth regularization term to optimize local semantic consistency and boundary accuracy. The overall Stage-2 loss is formulated as:

$$L = L_{\text{patch}_12} + L_{\text{cos}} + L_{\text{residual}} + L_{\text{sparsity}} + L_{\text{cam-smooth}} \quad (20)$$

where the CAM smooth term inherits the adaptive progressive regularization strategy:

$$L_{\text{cam-smooth}} = \alpha(t) \cdot L_{\text{cam-weighted}} \quad (21)$$

This stage also adopts stage-specific progressive $\alpha(t)$ regularization. Unlike Stage 1, the Stage-2 CAM-weighted loss penalizes residuals over the full predicted features containing position-dependent noise:

$$L_{\text{cam-weighted}} = \frac{1}{N} \sum_i D_i \cdot w_{\text{cam},i}, \quad D = (F_{\text{denoised}} + P_{\text{shared}} + R_{\text{pred}} - F_{\text{raw}})^2 \quad (22)$$

In stage 2, positionvariant residual noise disrupts local spatial continuity and object boundaries. In contrast to stage 1, which performs global semantic fitting anchoring, this stage concentrates on local structural refinement. Guided by G_{map} , positionaware semantic cues pinpoint spatialartifactprone regions. When combined with Residual Loss, such trainingtime constraints suppress NeRF-induced local overfitting and produce spatially smooth, structurally compact foreground features.

The proposed two-stage CAM supervision forms a hierarchical denoising paradigm tailored for dense segmentation tasks. Benefiting from the position and magnitude dual accuracy of G_{map} semantic feedback, our method accelerates training convergence, and produces denoised features with stable global semantics, consistent spatial distribution, and precise structural representation.

Experiments and Analysis

In the subsequent chapters, we design four progressive and complementary experiments to systematically validate the originality and rigor of this study’s trustworthiness analysis of CAM visualization results (semantic quality feedback) in segmentation scenarios.

All segmentation pipeline fine-tuning and inference experiments are implemented based on the MM-Segmentation framework [18], strictly following the training configurations and hyperparameter settings reported in the original DVT paper. We adopt the classic PASCAL VOC 2012 dataset for validation. In addition, CAM heatmap generation is embedded into the standard MM-Segmentation inference pipeline

with gradient backpropagation enabled throughout the attribution computation process.

Through carefully structured procedures, we reveal the limitations of current CAM frameworks, and demonstrate the effectiveness of the proposed systematic solution. Furthermore, we verify that hightrust CAM heatmaps can provide additional utility in downstream tasks.

The experiments are organized into four parts:

1. *Baseline evaluation*: Under the PyTorch-Grad-CAM framework, ROAD metrics are adopted to perform fundamental trustworthiness verification on heatmap outputs yielded by various CAM methods within the segmentation pipeline, revealing the absence of robust evaluation standards as a major gap.
2. *Qualitative evaluation*: Building on the unique perspective introduced earlier, denoised feature map PCA visualization is combined with CAM analysis. Supplemented by Shapley Value attribution principles, this process demonstrates the intuitive role of visualization in guiding trustworthiness analysis and qualitatively validates the trustworthiness of the proposed segmentation CAM method.
3. *Quantitative evaluation*: The rationality of the whitebox spatial semantic quality intensity representation (TPCA) is verified. Combined with quantitative trustworthiness metrics, this establishes a validation cycle that confirms the superiority of the proposed CAM method and TPCA representation.
4. *Utility demonstration*: Experiments further show that hightrust heatmaps ensure consistency among heatmaps, attribution maps, and semantic quality intensity in segmentation scenarios. Additional benefits include (i) accelerating denoising convergence without degrading the resulting feature quality, and (ii) enabling crossbackbone semantic feature quality comparison.

The detailed processes and corresponding results of these experiments will be elaborated in the following subsections.

Baseline Evaluation with ROAD

In the first experiment, we employ the ROAD (Remove and Debias) metric from the PyTorch-Grad-CAM framework, together with its standard computation procedure, to evaluate the trustworthiness of classical CAM methods as well as the proposed Grad-CAM-See in segmentation scenarios. ROAD is formally defined as follows: its principle is to mask or perturb input pixels in descending/ascending order of CAM saliency values and record the corresponding drop-in class logits. The area under the resulting curve reflects the quality of pixel ranking. A higher score indicates that the highlighted regions are more causally important to the prediction, thereby reflecting greater trustworthiness.

$$\text{ROAD}(M) = \sum_{p \in S} (f(x) - f(x \setminus p)) \quad (23)$$

where M is the CAM method, S is the set of pixels ranked by saliency, and $f(x)$ is the model confidence on input x .

The test image input adopts the typical “dog and cat” example recommended by the CAM framework. The chosen ROAD evaluation parameters and final ROAD scores given CAMs are listed in Table 1. And also, we overlay the CAM heatmaps for both dog and cat semantics directly on the original image for intuitive visualization, as shown in Figure 6, together with the chosen ROAD evaluation parameters and

final ROAD scores.

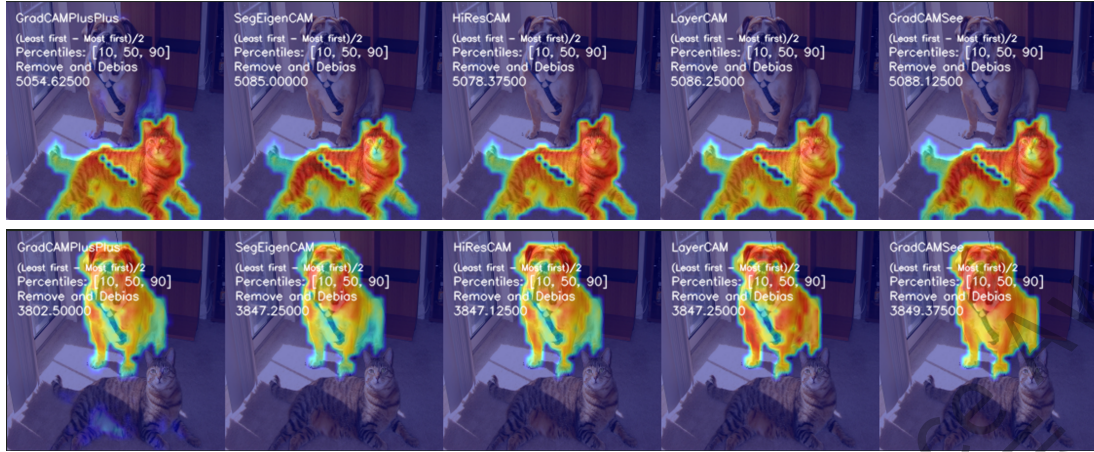


Figure 6. ROAD evaluation with the classical dog and cat tutorial image

Table 1 ROAD scores evaluation with the classical dog and cat tutorial image

	Grad-CAM++ [24]	Seg-Eigen CAM	HiResCAM	Layer CAM	Grad-CAM-See
Cat	5054	5085	5078	5086	5088
Dog	3802	3847	3847	3847	3849

Configuration: Least first – Most first; Percentiles [10, 50, 90] Remove and Debias

As shown in Table 1, under the adapted PyTorch-Grad-CAM segmentation framework, the ROAD scores of different classical classification CAM methods, recent segmentation CAM methods, and our proposed CAM method all fall within a narrow range (5054–5088 for Cat, 3802–3849 for Dog), although our method achieves the highest score. On the one hand, this confirms that our CAM method successfully passes the basic trustworthiness check. On the other hand, it highlights the limitation of ROAD in segmentation tasks, as previously discussed. As shown in Figure 6, such small numerical differences fail to reveal the clear attribution distinctions among CAM methods—differences that should be evident when considering the CAM heatmap visualizations and the underlying attribution mechanisms.

Visual PCABased Attribution Evaluation

In the subsequent experiment, denoised feature maps are used as inputs, with their PCA visualizations serving as auxiliary references. This design highlights the rationality of our analytical perspective and qualitatively demonstrates the advantages of our segmentation-specific CAM method in terms of trustworthiness.

As shown in Figure 7, the VOC2012 dataset is adopted for segmentation finetuning and semantic target evaluation. Six representative samples (airplane, cat, car, person, sheep, bird) are selected as raw inputs (first column). In the latter part of the segmentation pipeline, raw ViT backbone feature maps are denoised (third column) and then used as inputs for CAM methods. We compare Seg-Eigen-CAM, classical CAM methods adapted for segmentation (HiResCAM, LayerCAM), and our proposed Grad-CAM-See. The corresponding CAM heatmaps are overlaid for visualization.

The denoised PCA visualizations clearly reveal compact and domain-consistent feature clusters. For example, in the airplane sample (first row), the denoised PCA visualization shows a clean and complete airplane cluster with uniform feature quality. Based on the earlier analysis of CAM trustworthiness and

ShapleyValue principles, attribution intensity within the predicted foreground should be approximately consistent, especially in the cluster’s interior area with similar feature value. Crossmethod comparisons confirm that our segmentation CAM method, explicitly incorporating ShapleyValue, produces heatmaps that more uniformly and completely cover the foreground. In contrast, other methods either highlight only partial discriminative regions or produce fragmented activations, reflecting insufficient spatial trustworthiness when supervised by PCA visualization.

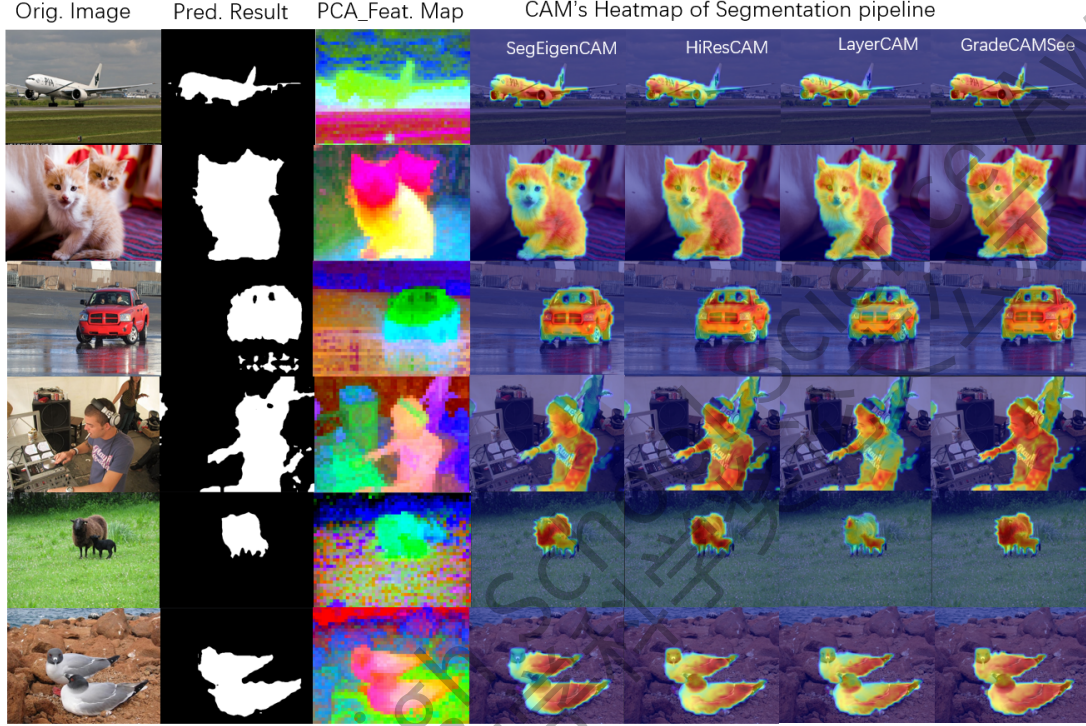


Figure 7. Visual PCAInspired Evaluation on VOC2012 samples

Quantitative Evaluation Based on TPCA Semantic Strength

In the subsequent experiment, we first design a scenario to validate the rationality of the TPCA semantic strength map. Since TPCA reasoning is based on segmentation mask regions, we begin by visually comparing the grayscale representation of the final strength map with the segmentation mask to check logical consistency—specifically, whether the mask decision is supported by reasonable grayscale trends. At the same time, because the purpose of the strength map is to reflect local linear inconsistencies revealed in PCA visualization, we also compare the grayscale strength map with the original PCA visualization to assess consistency.

This comparison covers a wide range of sample types, as shown in Figure 8. The first column presents the input category. For each sample, we display the original image, the raw feature map PCA visualization, the grayscale TPCA semantic strength map (with red circles marking critical regions), and the segmentation mask. These critical regions are where the logical consistency between grayscale trends and mask areas is most evident.

As shown in Figure 8, particularly in the critical regions of the first row (airplane) and the last row (bird), the background misclassification in the segmentation results originates from the gradual weakening of semantic strength (reflected by the transition from bright to dark in TPCA), which eventually

surpasses the threshold. This not only demonstrates the logical consistency between the TPCA grayscale representation and the segmentation mask, but also aligns with the spatial artifact distribution and intensity revealed in PCA visualization.

At the same time, when focusing on samples with balanced and clean foreground feature quality (such as the airplane in the second row, sheep, and car), as well as samples with more subtle artifacts (such as cat), it is evident that the TPCA grayscale representation accurately reflects the neighborhood linear consistency revealed by PCA visualization. These experiments collectively validate the rationality of the TPCA semantic strength map.

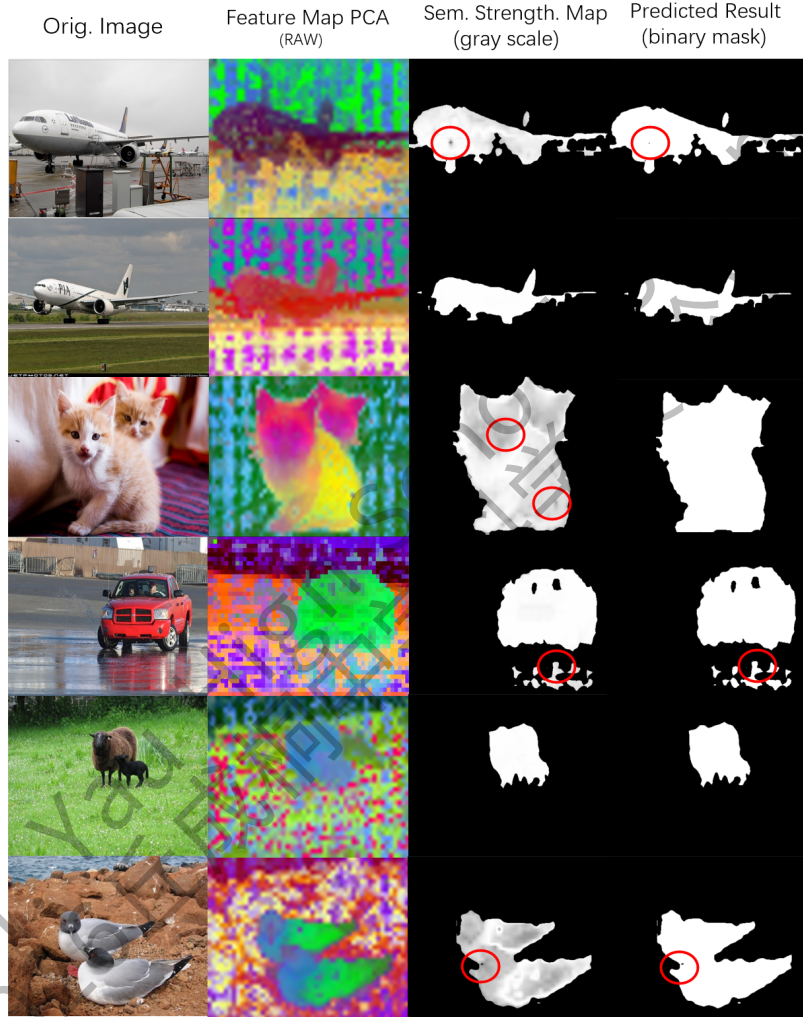


Figure 8. TPCA semantic strength maps across categories

After verifying the rationality of the whitebox semantic strength method TPCA and its results, we further designed experiments to cyclically validate our proposed approach and demonstrate its superiority within the TPCA quantitative trustworthiness framework. Specifically, following the procedure shown in Figure 8, we computed the Structural Similarity Index (SSIM) between the heatmaps of four CAM methods and the whitebox semantic strength maps. The CAM methods include: Grad-CAM-See, LayerCAM, HiResCAM, and Seg-Eigen -CAM.

For each CAM method, the SSIM values of its heatmaps were plotted as line charts, as shown in Figure 9, to illustrate spatial consistency with the whitebox semantic strength maps—namely, whether the

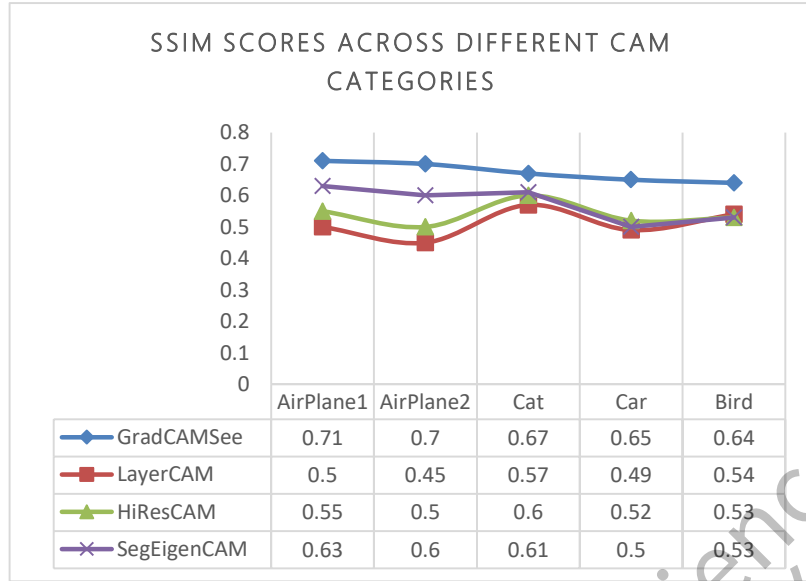


Figure 9. SSIM scores across different CAM categories

CAM heatmaps accurately reflect the neighborhood linear consistency revealed by PCA visualization. From Figure 9, it can be observed that our segmentation-oriented Grad-CAM-See consistently achieves the highest SSIM values across all input categories (0.71, 0.70, 0.67, 0.65, 0.64). This not only cyclically validates the advantage of our Grad-CAM-See design (supported by ShapleyValue) in faithfully reflecting spatial semantic, but also confirms the trustworthiness of TPCA in representing semantic strength. The consistent trends between the two further highlight the important role of PCA in guiding CAM trustworthiness evaluation in segmentation scenarios.

Figure 10 visualizes the advantages of the Grad-CAM-See method in reflecting spatial semantic trustworthiness, as well as its trend consistency with the whitebox PCA-derived TPCA semantic strength representation. This serves as a complementary visualization and validation process for the numerical results shown in Figure 9. Together with Figure 10, we can clearly observe the importance and rationality of the proposed approach—“using feature visualization to supervise interpretability trustworthiness.”

Trustworthy CAM Heatmap Utility: Evaluation and Analysis

CAM augmented DVT’s Convergence Analysis

After verifying the effectiveness of our novel evaluation perspective in systematically analyzing the trustworthiness of CAM results, as well as the improvements introduced to the CAM framework (including both the new updated CAM method and the original spatial evaluation framework), the subsequent experiments examine the auxiliary role of high-fidelity segmentation attribution maps (trustworthy CAM’s heatmap) in downstream feature processing tasks such as denoising. As designed in Method 3, these maps help ensure final quality and accelerate training speed in related denoising models.

To validate the general applicability of this approach, we tested multiple representative input categories. We compared the loss curves between the standard DVT feature denoising process and DVT training augmented by segmentation CAM heatmaps proposed in this work. The comparison includes convergence curves of Patch Loss, CosSim Loss, and Residual Loss under both settings (Figure 11), as well as the final training outcomes (Figure 12). For clarity, Figure 11 presents only the representative bird

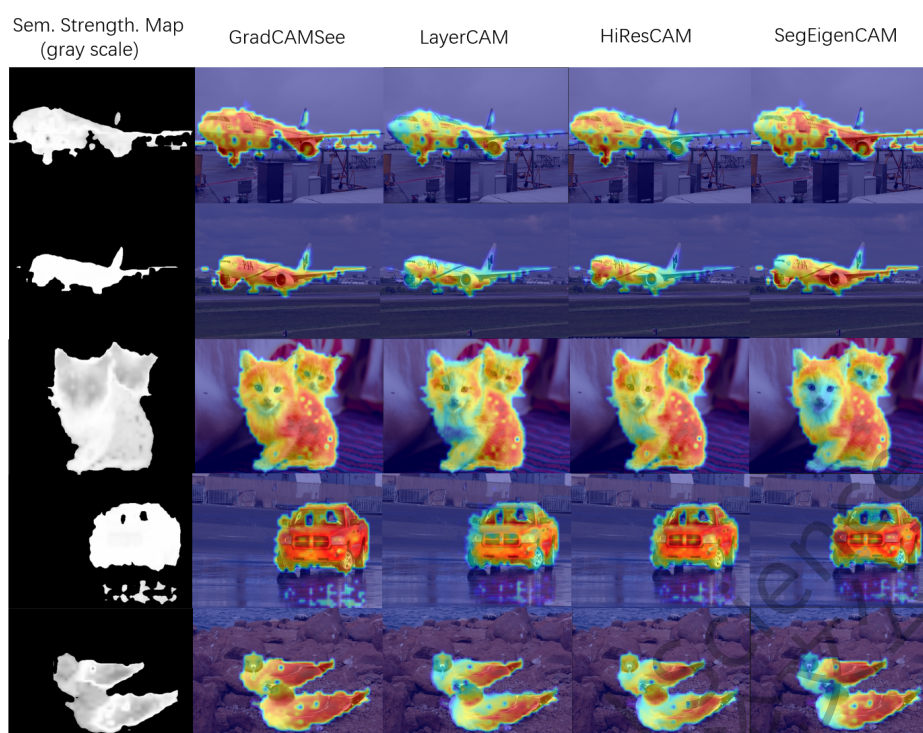


Figure 10. Visual comparison of CAM heatmaps and TPCA semantic strength maps

sample data, while results for other categories are provided in the appendix (all showing similar trends to the bird sample).

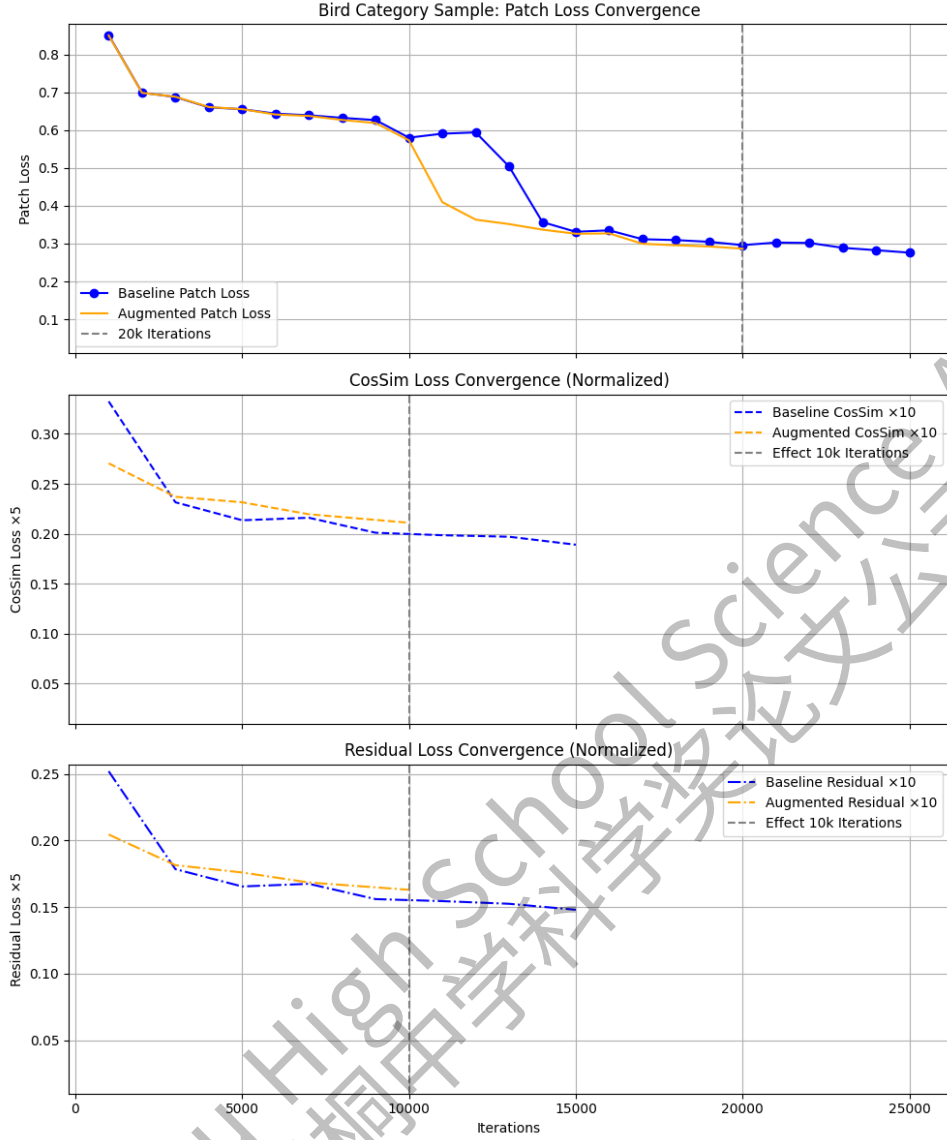


Figure 11. Convergence analysis of Patch Loss, CosSim Loss, and Residual Loss

As shown in Figure 11, and consistent with the description in the methods section, the standard DVT offline denoising process consists of two stages. The original training cycle was set to 25,000 iterations. Throughout the process, Patch Loss serves as the primary supervisory signal, while CosSim Loss and Residual Loss provide supervision in the second stage, starting from 10,000 iterations.

With the integration of trustworthy heatmaps and the enhanced training strategy, the augmented denoising process allows the model to enter the stable regime earlier—requiring only about 20,000 iterations, approximately 20% faster. Moreover, the loss functions descend more smoothly compared to the baseline, demonstrating clear advantages.

As shown in Figure 12, when comparing the final denoising quality between the two training configurations—visualized through clustering maps after feature denoising—the results are essentially equivalent. This fully validates and illustrates that the augmented configuration achieves the same outcome while providing a more efficient training process.

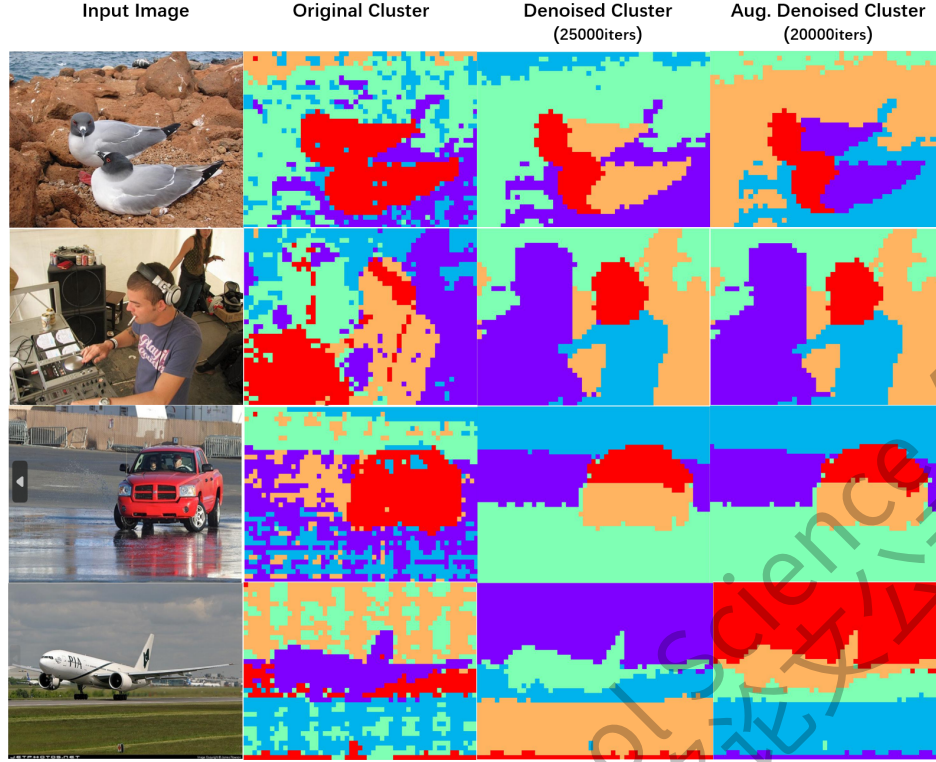


Figure 12. Cluster visualization results of baseline and augmented process

ViT Backbone Analysis under Segmentation Tasks

After verifying the role of Grad-CAM-See heatmaps in enhancing the original DVT denoising process, this section further expands the experimental scope to address the “Bridge: Mirror of TaskOriented Semantic Quality” raised in related work. The objective is to examine how Grad-CAM-See can support the selection and analysis of pretrained backbones in segmentation tasks by faithfully reflecting classaware semantic feature quality.

Under a unified segmentation framework and dataset protocol, four representative pretrained DI-NOv2 ViT base backbones were adopted:

- vit_base_patch16_reg4_dinov2.imagenet
- vit_base_patch16_reg4_sim_dinov2.imagenet
- vit_base_patch14_dinov2.lvd142m
- vit_base_patch14_reg4_dinov2.lvd142m

These models cover structural differences (with [25]/without register), variations in pretraining dataset scale (ImageNet vs. LVD142M), and distinct selfsupervised strategies (blackbox vs. whitebox, i.e., the SimDino proposed by Professor Ma), representing mainstream ViT paradigms.

Since all models are embedded into the same downstream segmentation pipeline (ViT backbone + BN decoder + Conv_seg head), performance differences originate mainly from the semantic quality of the backbone feature maps.

Exploration of highthreshold heatmap regions:

1. Figure 13 illustrates the airplane category from VOC2012, where strong threshold (>0.6) CAM masks are used as the analysis entry point to compare Patch14 and Patch14 Reg4 backbones. As

shown, high threshold CAM regions can reveal local semantic feature quality and noise characteristics: when noise is evident, the corresponding regions tend to exhibit more fragmentation and holes, as observed in the Reg4 model. This not only aligns with the final segmentation accuracy measured by mIoU (Patch14 achieves 0.9015 vs. Patch14 Reg4's 0.8751), but also corresponds to the feature properties revealed in PCA visualizations.

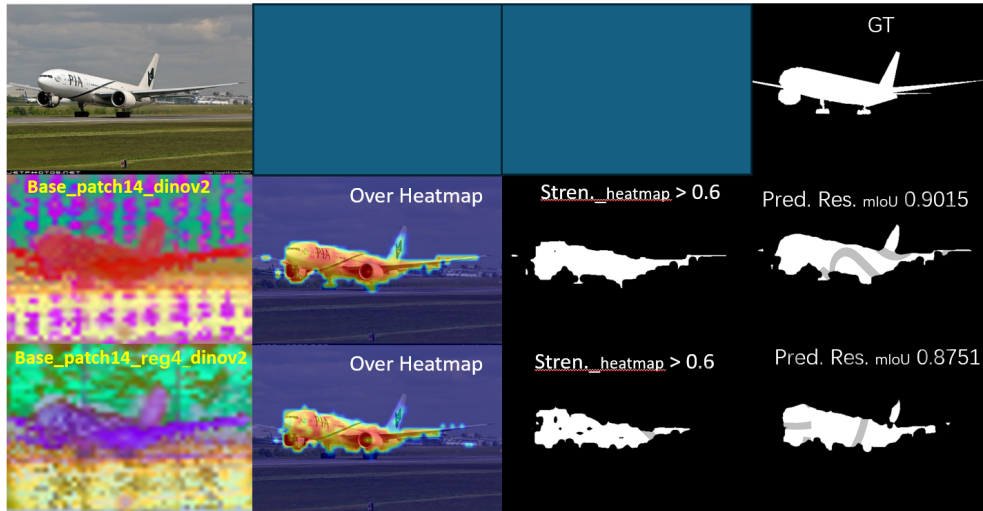


Figure 13. Airplane category test result comparing Patch14 Dinov2 and Patch14 Reg4 Dinov2 backbones.

- Figure 14, which tests multiobject image samples (containing both person and airplane), further validates how strongthreshold heatmap regions provide feedback on the noise quality of feature maps. For both Patch14 and Patch14 + Reg4 backbones, within the strong threshold regions, lower semantic quality (Patch14) corresponds to more and larger fragmented holes (i.e., more severe artifacts), which ultimately leads to reduced segmentation accuracy in the complete pipeline as measured by mIoU.

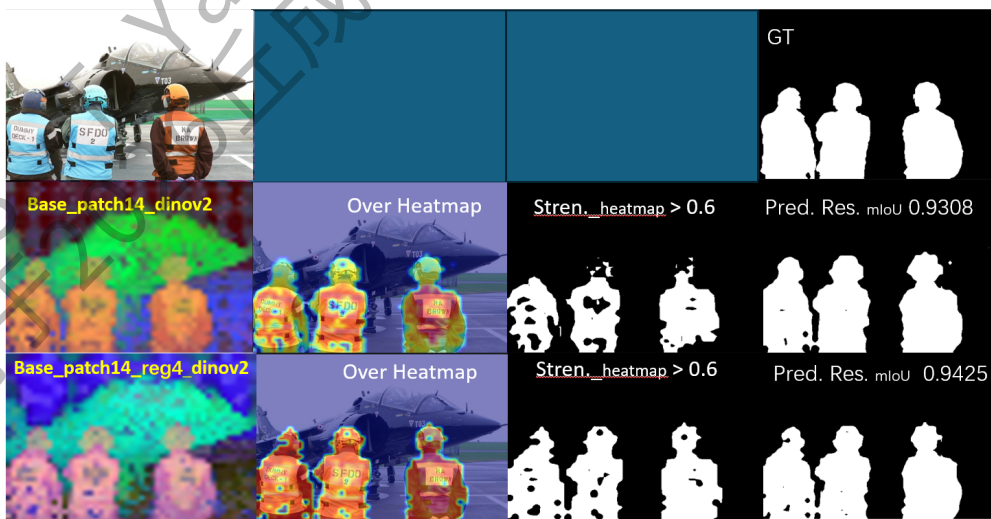


Figure 14. Person category test result comparing Patch14 Dinov2 and Patch14 Reg4 Dinov2 backbones.

Exploration of lowthreshold heatmap regions:

3. Figure 15 illustrates the airplane category from VOC2012, where lowthreshold (>0.3) CAM masks are used as the entry point for analysis of Patch16 Reg4 and Patch16 Reg4 Sim backbones. This analysis is particularly representative for two reasons: first, the Sim backbone represents the ‘black’ and ‘white’ training framework; second, as shown in the figure, because both backbones were pre-trained on ImageNet, noise in the feature maps is relatively less pronounced. This is why we introduce low threshold heatmap regions, to reveal whether the pretrained feature space aligns with the task dataset.

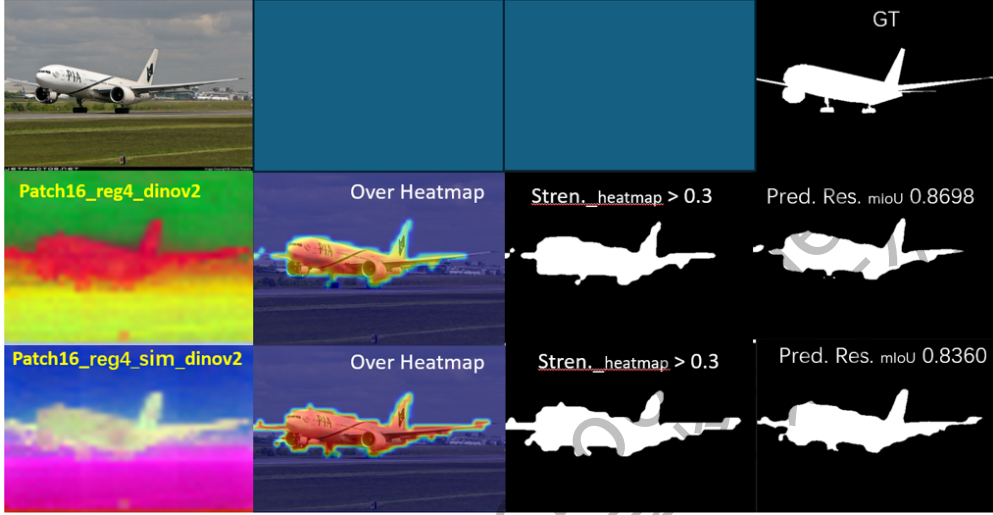


Figure 15. Airplane category test result comparing “Black” Dinov2 and “White” Dinov2 backbones.

When alignment differences are significant, it indicates that the feature space produced under certain training configurations deviates substantially from the feature space required by the target dataset, as exemplified by the Reg4 Sim model. This observation is consistent with the final segmentation accuracy measured by mIoU (Reg4 achieves 0.8698, while Reg4 Sim reaches only 0.8360), and also aligns with PCA visualizations that reveal feature deviations (blurred edges and misalignment with actual targets).

4. Figure 16 presents the interpretability visualization results for the sheep category in VOC2012, using lowthreshold (>0.3) CAM masks as the entry point to analyze all four ViT backbones within the segmentation pipeline. This constitutes the most comprehensive crossbackbone analysis with the richest elements.

As shown in Figure 16, due to differences in pretraining datasets, the noise characteristics of the feature maps vary (foreground signals are less distinct). When combined with the lowthreshold heatmap regions, the comparison reveals clear differences in how the pretrained feature spaces align with the semantic space of the task dataset.

We find that the final segmentation accuracy measured by mIoU is consistent with the degree of semantic space deviation (as visualized in Figure 16). This demonstrates the critical role of alignment between the pretrained feature space and the task dataset’s semantic space in determining the accuracy of the complete segmentation pipeline.

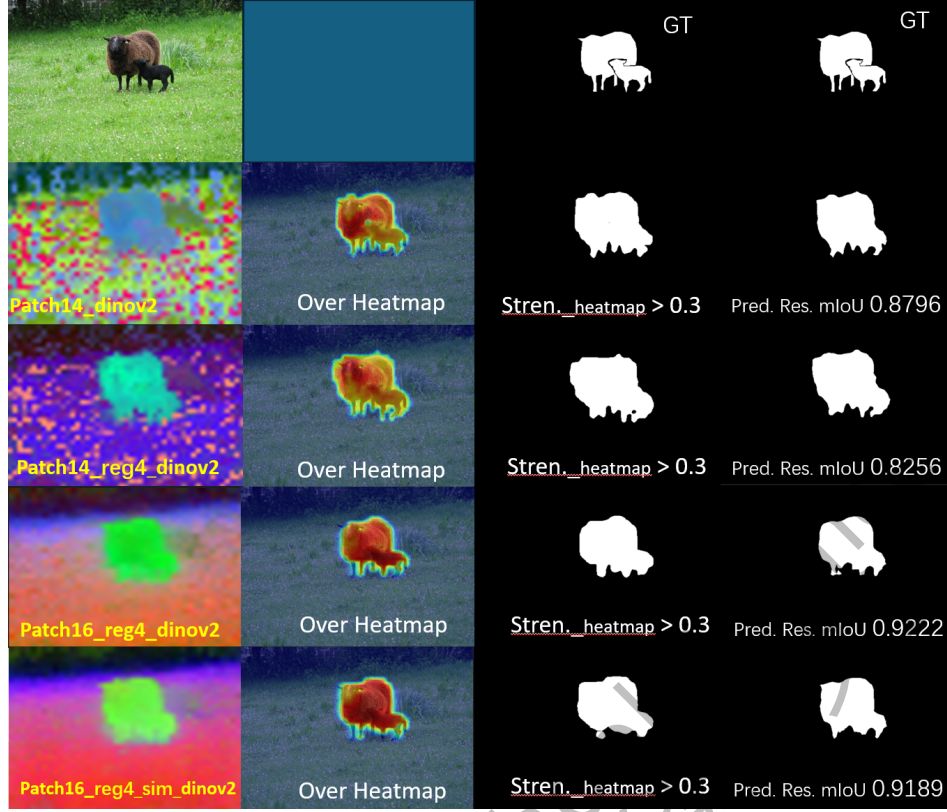


Figure 16. Sheep category test result across four backbones

Conclusion

This study focuses on a longneglected but crucial issue in the field of visual foundation models: how to credibly and fairly evaluate the feature semantic quality of different pretrained Vision Transformer (ViT) backbones in specific dense prediction tasks.

Core Inspiration — *Judgement of Semantic Quality through Trustworthy Task Interpretability Attribution via Feature PCA Visualization*: Visualizing the feature maps output by pre trained ViT backbones can effectively reveal the core structure and main characteristics of the representation space. Inspired by classical research on Denoising Vision Transformers (DVT), clean features not only exhibit local linear consistency and compactness but also correspond to high-quality semantic features, thus provides an excellent perspective for verifying the trustworthiness of CAM attribution in segmentation scenarios, i.e., the semantic quality of feature maps.

Core Theoretical Foundation—*Shapley Value for Mathematical Completeness and Trustworthiness of CAM*: One of the axioms of Shapley value—*Symmetry*—offers rigorous theoretical support for attribution analysis, especially when extended to segmentation scenarios.

Core Contributions—*PCA Visualization + Shapley Value to Construct and Enhance Trustworthy Attribution Analysis of CAM in Segmentation*: We propose Grad-CAM-See, a trustworthy attribution interpretability analysis method specifically designed for semantic segmentation; and TPCA, a whitebox semantic quality intensity representation method and its extension to work as a trustworthiness evaluation metric. Together, these new methods and quantifiable evaluation mechanisms complete the previously missing framework for semantic quality analysis of features in segmentation scenarios, providing stronger support for subsequent feature quality improvement and fair crossbackbone comparison of ViTs.

References

- [1] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv:2010.11929, 2020.
- [2] H. Touvron et al., “Training data-efficient image transformers & distillation through attention,” arXiv:2012.12877, 2020.
- [3] M. Caron et al., “Emerging properties in self-supervised vision transformers,” arXiv:2104.14294, 2021.
- [4] Y. Fang et al., “EVA: Exploring the Limits of Masked Visual Representation Learning at Scale,” arXiv:2211.07636, 2023.
- [5] M. Oquab et al., “DINOv2: Learning robust visual features without supervision,” arXiv:2304.07193, 2023.
- [6] O. Siméoni et al., “DINOv3,” arXiv:2508.10104, 2025.
- [7] Y. Yu et al., “White-Box Transformers via Sparse Rate Reduction,” arXiv:2306.01129, 2023.
- [8] Y. Yu et al., “White-Box Transformers via Sparse Rate Reduction: Compression Is All There Is?,” arXiv:2311.13110, 2023.
- [9] J. Yang et al., “Scaling White-Box Transformers for Vision,” arXiv:2405.20299, 2024.
- [10] J. Wu et al., “Simplifying DINO via Coding Rate Regularization,” arXiv:2502.10385, 2025.
- [11] B. Zhou et al., “Learning deep features for discriminative localization (CAM),” arXiv:1512.04150, 2015.
- [12] C. T. Chung and J. J. C. Ying, “SegEigenCAM: EigenValueBased Visual Explanations for Semantic Segmentation Models,” *Applied Sciences*, vol. 15, no. 13, p. 7562, 2026.
- [13] J. Gildenblat et al., “PyTorch library for CAM methods,” GitHub: [GitHub–jacobgil/pytorch-grad-cam](https://github.com/jacobgil/pytorch-grad-cam), 2021.
- [14] J. Yang et al., “Denoising Vision Transformers,” in *Computer Vision–ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Lecture Notes in Computer Science, vol. 15143. Cham: Springer, 2024.
- [15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, pp. 248–255, 2009.
- [16] O. Russakovsky et al., “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
- [17] S. Yang et al., “SegDINO: An Efficient Design for Medical and Natural Image Segmentation with DINOv3,” arXiv:2509.00833, 2025.
- [18] K. Chen et al., “MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark,” GitHub: [open-mmlab/mmdetection](https://github.com/open-mmlab/mmdetection), 2020.
- [19] R. R. Selvaraju et al., “Grad-CAM: Visual Explanations from Deep Networks via Gradientbased Localization,” arXiv:1610.02391, 2017.
- [20] P. T. Jiang et al., “LayerCAM: Exploring Hierarchical Class Activation Maps for Localization,” *IEEE Transactions on Image Processing*, vol. 30, pp. 5875–5888, 2021.

- [21] R. L. Draelos et al., “Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks,” arXiv:2011.08891, 2020.
- [22] H. Cai, “CAMs as Shapley Value-based Explainers,” arXiv:2501.06261, 2025.
- [23] B. Mildenhall et al., “NeRF: Representing scenes as neural radiance fields for view synthesis,” in *ECCV 2020*, 2020.
- [24] A. Chattopadhyay et al., “Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks,” arXiv:1710.11063, 2018.
- [25] T. Darcet et al., “Vision Transformers Need Registers,” arXiv:2309.16588, 2023.

致谢

本人人工智能视觉方向创新研究课题历时八个月，全程由本人独立主导完成，实验操作、数据分析、论文撰写等全部核心工作均为本人自主完成。本研究在南京外国语学校信息学教研组组长李曙老师、东南大学未来技术学院副院长张亚教授的无偿公益指导下完成，两位老师均利用课余及科研空余时间提供学术帮扶，未收取任何酬劳，为纯粹的科创育人指导关系。其中，李曙老师深耕中小学人工智能拔尖人才培养，为本研究提供落地性实践指导；张亚教授深耕人工智能算法科研领域，为课题提供专业理论支撑。

本次研究分工清晰明确，本人为核心执行负责人，全权推进各项研究落地工作，两位老师仅负责学术思路指导、研究方向迭代与内容答疑审视，具体分工与研究推进过程如下。

本人核心分工：自主构思并细化课题研究思路与整体框架，独立完成文献检索研读、开源数据集下载与预处理工作；和ViT骨干网络集成、CAM方法对比分析和创新、T-PCA语义强度设计、归因图辅助去噪等仿真实验，独立完成数据计算、结果分析与成果总结；自主完成论文撰写，结合老师指导建议打磨完善，最终敲定论文终稿。

指导老师分工：李曙老师主要把控课题研究进度与逻辑完整性，针对性指导模型集成优化、研究闭环搭建与创新点落地，全程审视研究内容与论文整体框架。张亚老师负责理论指导，指导我梳理领域文献、搭建基础技术框架，系统学习CAM方法、ROAD评测、PCA语义表征与Shapley归因理论，规范论文论证逻辑。

本研究循序渐进、迭代完善，各阶段研究工作均在两位老师的悉心指导下有序推进。研究初期，经李曙老师研讨指导，我明确了课题核心研究目标与整体方向；张亚老师则指导我梳理研读专业文献，搭建起图像分割与去噪的基础技术框架，筑牢研究基础。

在ViT骨干网络集成阶段，针对模型融合落地的思路局限问题，李曙老师多次提供优化建议，帮助我迭代完善模型集成策略。

在CAM方法研究阶段，面对Grad-CAM主流框架的多样可视化算法体系，我对各类方法优劣特性认知不足，张亚老师指导我深入学习，厘清PCA、Shapley归因理论的融合逻辑，通过对照实验完成系统性分析，助力我构思适配分割任务的优化思路，并在师生定期研讨中持续完善方案。

研究后期，在李曙老师的启发指导下，我设计实现归因图辅助去噪的创新方法，补齐研究逻辑闭环，完善整体研究体系。在此，我向两位老师无私尽心、耐心细致的无偿指导，致以最诚挚的感谢与崇高的敬意！

使用人工智能工具申明：在论文撰写完善阶段，我使用了字节跳动的“豆包”桌面工具(版本信息：2.11.6_win)。相关完整过程如附件，由工具完整导出和总结，包括语法润色，科研逻辑校验，公式文本一致性核对和术语确认等(可参看附件中总结部分)。其目的为参照人工智能通识形成文稿核对和写作逻辑复盘。结合过程，人工智能普通工具尚无法构建创新思维和严谨逻辑，但对于知识面和考虑视角有很好的提示。