

PANC: Prior-Aware Normalized Cut for Object Segmentation

Juan Gutiérrez

Víctor Gutiérrez-García

José Luis Blanco-Murillo

Universidad Politécnica de Madrid

Av. Complutense 30, 28040 Madrid, Spain

juan.gutierrez@upm.es, v.garcia@upm.es, jl.blanco@upm.es

Abstract

Fully unsupervised segmentation pipelines naively seek the most salient object, should this be present. As a result, most of the methods reported in the literature deliver non-deterministic partitions that are sensitive to initialization, seed order, and threshold heuristics.

We propose PANC, a weakly supervised spectral segmentation framework that uses a minimal set of annotated visual tokens to produce stable, controllable, and reproducible object masks. From the TokenCut approach, we augment the token–token affinity graph with a handful of priors coupled to anchor nodes. By manipulating the graph topology, we bias the spectral eigenspace toward partitions that are consistent with the annotations. Our approach preserves the global grouping enforced by dense self-supervised visual features, trading annotated tokens for significant gains in reproducibility, user control, and segmentation quality.

Using 5 to 30 annotations per dataset, our training-free method achieves state-of-the-art performance among weakly and unsupervised approaches on standard benchmarks (e.g., DUTS-TE, ECSSD, MS COCO). Contrarily, it excels in domains where dense labels are costly or intra-class differences are subtle. We report strong and reliable results on homogeneous, fine-grained, and texture-limited domains, achieving 96.8% (+14.43% over SotA), 78.0% (+0.2%), and 78.8% (+0.37%) average mean intersection-over-union ($mIoU$) on CrackForest (CFD), CUB-200-2011, and HAM10000 datasets, respectively. For multi-object benchmarks, the framework showcases explicit, user-controllable semantic segmentation. The code is available at [this repository](#).

1. Introduction

Annotating per-pixel segmentation masks at scale is costly in both human labor and computation. Creating large supervised datasets is an extensive task that requires countless hours of manual effort and a substantial annotation infrastructure [1, 2]. This expensive need drives interest

toward methods that reduce annotation burden by relying on either purely unsupervised discovery or very sparse, weak supervision. Recent unsupervised pipelines based on self-supervised Vision Transformer (ViT) tokens, most notably TokenCut and related token-ranking heuristics, exploit dense, frozen patch embeddings to produce class-agnostic object masks without input labels [3–5]. These approaches demonstrate strong zero-shot performance on standard benchmarks, but are inherently underconstrained. Selection rules based on saliency, ranking, or heuristic thresholds can choose different entities in multi-object or ambiguous scenes and often fail on homogeneous or low-semantic-content images where the most “salient” region is not the desired target [3, 4].

Weak supervision provides a middle ground: the use of a small amount of targeted supervision (points, scribbles, or image-level cues) may solve ambiguities and inject semantic intent into propagation or grouping algorithms [2, 6, 7]. However, incorporating priors into global clustering is not trivial. Naïve constraints may either be ignored by global objectives or overly dominate local affinities. Learned affinity models introduce training overhead and dataset-specific tuning [8–10]. Quite recently, advances in self-supervised ViT encoders (notably the recent DINOv3) produce dense token embeddings with improved stability across resolutions and stronger geometric consistency (namely, Gram anchoring and multiresolution exposure). These representations are particularly suitable as the substrate for seed-driven propagation and constrained spectral objectives [11–13]. The encoder reduces sensitivity to token resolution and improves reliability for affinity-based grouping on limited numbers of priors.

We follow a weakly-supervised strategy to inject a compact bank of token-level priors into a training-free, spectral token-graph pipeline, built upon strong, self-supervised ViT features. Our method, illustrated in Figure 1, uses only a small number (tens) of manually annotated token exemplars drawn from a few selected images. These priors (bottom left) are integrated as anchor nodes in the affinity graph (bottom center). The resulting normalized-cut

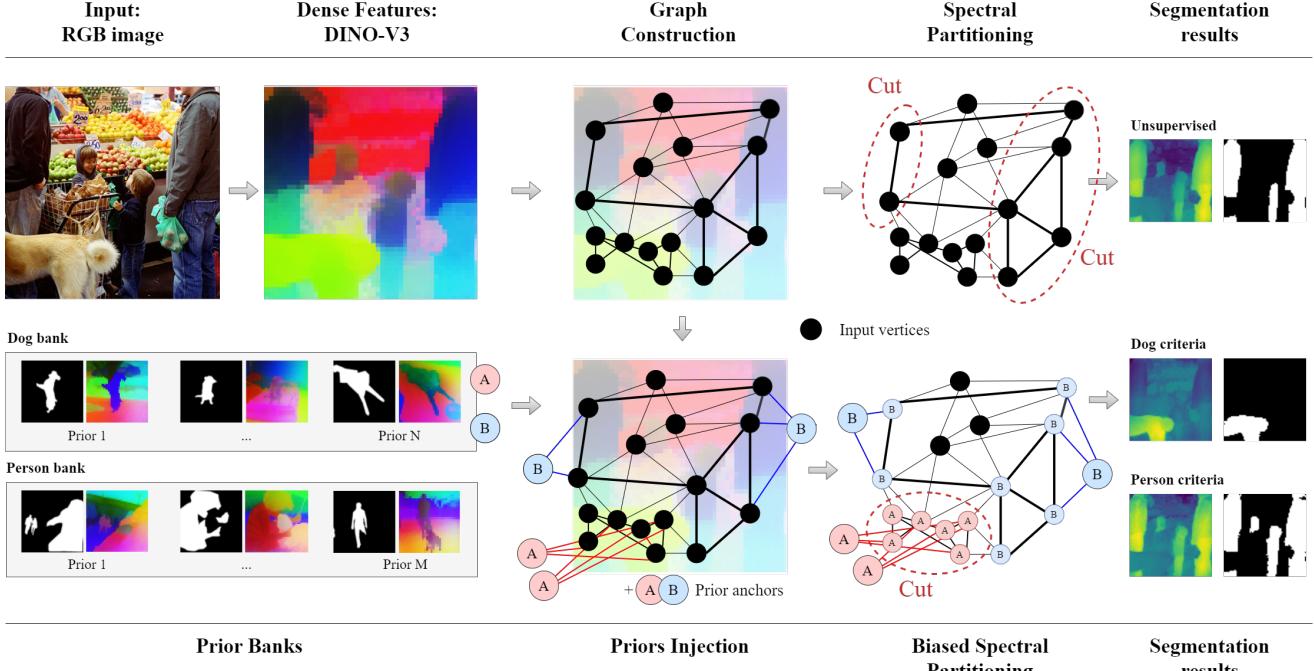


Figure 1. Schematic on the proposed PANC framework. To segment an input image (top left) NCut algorithm (center) relies on building the affinity graph. The result (top right) is largely uncontrolled in terms of the segmented “salient” objects and the labels assigned to these. Our solution (bottom center) introduces a minimal set labels (bottom left) as annotated priors directly into the affinity graph. Segmentation on the augmented graph focused on classes exemplified by the priors and fosters consistency in spectral partitioning.

eigenspace is biased toward partitions consistent with the supplied exemplars and preserves global grouping due to the dense features (bottom right). Our design provides explicit, user-controllable selection of class-instance to segment (critical in multi-object scenes or business-rule scenarios), and improves robustness on homogeneous or low-semantic-content images, where unsupervised saliency is unreliable. Moreover, since we rely on standard DINOv3-style tokens and a training-free spectral formulation, we avoid per-dataset affinity learning while still leveraging stability and cross-resolution consistency of recent encoders [12, 13]. Our main contributions include:

- We introduce **PANC**, a compact weakly-supervised framework that injects a small bank of token-level priors into a spectral token-graph pipeline. PANC enables explicit, user-controllable selection of the segmented entity and produces reproducible, dense, and accurate masks with low annotation costs.
- We release a GPU-accelerated implementation of the anchor-augmented graph partitioning pipeline, including efficient iterative eigensolvers, anchor injection, deterministic orientation and a stable score-to-mask conversion, to handle high-resolution inputs with predictable memory/runtime trade-offs. See Appendix A.
- We present an extensive evaluation across heterogeneous

and homogeneous domains, demonstrating improved segmentation quality per unit of supervision. We report ablations over injected vertices, anchor coupling, affinity temperature, image resolution, and thresholding strategies, and compare to state-of-the-art baselines on unsupervised and weakly-supervised methods.

2. Related Work

Self-supervised Vision Transformers with dense tokens. Patch-based Vision Transformers trained with self-supervision provide dense token embeddings. Similarities across these correlate with the extents of parts and objects, proving their value as a strong substrate for label-free grouping [11, 14–16]. DINOv2 demonstrated that scaling and careful data curation yield frozen features that transfer robustly to dense prediction [12]. At the same time, DINOv3 introduced training refinements (e.g., Gram anchoring, multiresolution exposure, long schedules) that improve per-token stability and geometric robustness, critical when building token graphs for segmentation [13]. In practice, token resolution is governed by patch stride and memory limits, demanding long-sequence encodings or multi-view/upsampling strategies to recover fine detail from frozen backbones [13, 17, 18].

Unsupervised object discovery. Class-agnostic discovery has evolved from co-localization/co-segmentation via discriminative clustering and region matching to single-image pipelines that rely on dense learned features [19–24]. Deep spectral analyses on feature affinities produce meaningful regions without labels [5], and ViT-token methods such as LOST and TokenCut build graphs from frozen self-supervised embeddings to extract objects using simple spectral/graph criteria, achieving strong zero-shot performance on VOC/COCO [1, 3, 4, 25]. These results suggest that high-quality self-supervised tokens, especially from DINO-style encoders, are well-suited to classical grouping formulations [11, 12].

Spectral clustering and graph-based segmentation. Spectral methods cast segmentation as balanced partitioning on an affinity graph, solved via Laplacian eigenvectors (e.g., normalized cuts) [26, 27]. Random-walk and harmonic-function formulations provide a probabilistic perspective to propagate sparse information across the graph [28, 29]. Practical systems leverage constrained variants (must-/cannot-link or linear constraints) and scalable approximations (sparse k NN graphs, Nyström/landmark methods, randomized/iterative eigensolvers) to handle dense images and large datasets [8, 9, 30–32]. When nodes are self-supervised ViT tokens, spectral relaxation yields stable global groupings that can be sharpened by standard edge-aware refinement [5, 33].

Seeded and weakly supervised segmentation. Weak cue—points, scribbles, boxes, or image-level tags are routinely converted into dense masks by seed generation and propagation with analytic or learned affinities, followed by boundary refinement [2, 6, 7, 33, 34]. Classical interactive approaches (GrabCut; random walks) already reported how a few seeds can drive high-quality segmentations via graph cuts or harmonic propagation [28, 29, 35]. Subsequent pipelines improved seed quality with CAMs and learned affinity networks, regularizing expansion with CRFs or graph objectives [7, 10, 33, 36]. Constrained spectral formulations and normalized-association losses provide alternative ways to encode sparse labels while preserving global consistency [8, 9, 37]. More recently, prototype/bank-based guidance aggregates representative features to steer pixel assignments under weak labels, improving pseudo-mask completeness without dense supervision [17, 38]. Within this landscape, one should benefit from self-supervised ViT tokens as graph nodes, as their dense, geometry-robust embeddings reduce the need for per-dataset, while retaining controllability [11–13].

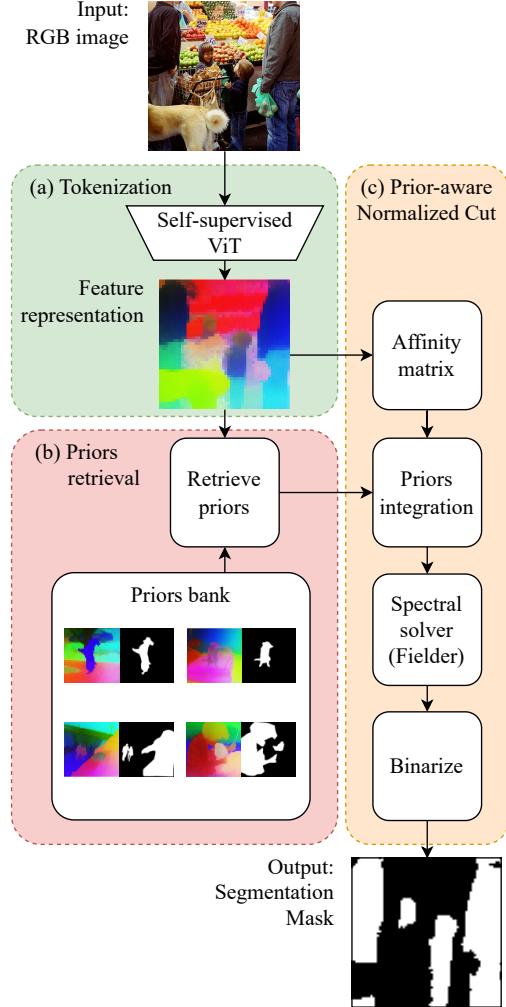


Figure 2. Overview of the proposed pipeline. The input image is tokenized using ViT. We use the resulting representation to retrieve suitable priors from the annotated bank, and to build the affinity matrix. The new extended affinity matrix is processed in the same was as in TokenCut, to produce the final segmentation mask.

3. Method

The central idea behind PANC is to guide a spectral, token-based segmentation pipeline using a compact set of token-level priors that are manually annotated and selected from a small pool of representative images. As illustrated in Figure 2, the pipeline comprises three high-level stages.

First, an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ is tokenized by a frozen, self-supervised Vision Transformer into a regular grid of n patch embeddings. These token embeddings serve as the atomic units for the subsequent processing and capture rich semantic and appearance cues. produced by the self-supervised pretraining.

Second, we maintain a compact prior bank constructed from a small number of representative images selected via

clustering in the image-embedding space. We sparsely and manually annotate these representative images at the token level. Each prior token is labeled either as a positive example (the class-of-interest) or as a negative example (the complementary class). The prior bank is deliberately small (tens of token exemplars) to minimize annotation cost while preserving diversity across the domain. For each input image a small and specific subset of priors is retrieved based on a predefined affinity metric.

Third, we compute the affinity graph for the patched input image tokens and inject the priors into the reasoning stage as anchor nodes. This anchoring biases the downstream spectral solution toward partitions consistent with the selected priors while retaining global grouping induced by the dense token affinities. The method produces a continuous score per token, which is deterministically converted to a binary mask.

3.1. Prior bank construction

Representative image selection We construct a compact prior bank from a small set of representative images chosen by clustering image-level descriptors produced by a frozen, self-supervised Vision Transformer. Each image \mathbf{x} is encoded to obtain the ℓ_2 -normalized CLS token $\mathbf{c}(\mathbf{x}) \in \mathbb{R}^{d_c}$. Let $\mathbf{C} \in \mathbb{R}^{N \times d_c}$ stack all embeddings from the training set. We run k -means with K_{clusters} clusters on \mathbf{C} to obtain centroids $\{\boldsymbol{\mu}_k\}_{k=1}^{K_{\text{clusters}}}$. For each cluster k , the representative image is $\mathbf{x}_k^* = \arg \min_{\mathbf{x} \in \mathcal{S}_k} \|\mathbf{c}(\mathbf{x}) - \boldsymbol{\mu}_k\|_2$, where \mathcal{S}_k is the set of images assigned to cluster k . This yields K_{clusters} images that cover the dominant appearance modes with minimal redundancy. From each \mathbf{x}_k^* , we extract its token grid $\{\mathbf{f}_i\}_{i=1}^n$ as candidate priors.

Annotation protocol Token labels are produced by a lightweight, manual procedure that combines automatic proposals with human refinement. One can help labeling a required image \mathbf{x}_k^* , using, for example, the CVAT interactive labeling interface together with SAM, and make minor rectifications to correct errors in the SAM proposals. Our method assumes that we annotate positive points for foreground and negative points for background on the token grid. For full details of the procedure, see the Appendix B.

The process is intentionally minimal: creating and refining the sparse token annotations takes under one minute per selected image on average, which is a tiny fraction of the effort required for dense per-pixel labelling. The process yields high-quality priors that guide spectral grouping.

Tokens selection To obtain a compact and diverse set of priors for each input image from the bank, we apply a two-stage, diversity-aware selection. First, we compute a relevance score for every bank token by averaging its top K_{sim} cosine similarities to tokens in the current image: $r_j = \frac{1}{K_{\text{sim}}} \sum_{t \in \text{top}_{K_{\text{sim}}}(\langle \mathbf{p}_j, \mathbf{f}_t \rangle)} \langle \mathbf{p}_j, \mathbf{f}_t \rangle$, and prefilter the bank to the top M' candidates per label according to r_j . Second, from each prefiltered pool we greedily select a small set of priors using maximum marginal relevance. At each step, we pick the bank token that maximizes relevance minus a diversity penalty: $\text{score}_j = r_j - \lambda \max_{s \in S_{\text{sel}}} \langle \mathbf{p}_j, \mathbf{p}_s \rangle$, where S_{sel} is the set of indices already selected and $\lambda \in [0, 1]$ trades relevance for diversity. This procedure returns a small, label-balanced set of prior vertices that are both relevant to the current image and mutually diverse.

3.2. Affinity graph and spectral embedding

Graph construction Let the token set $\mathcal{V} = \{1, \dots, n\}$ index the n patch descriptors $\{\mathbf{f}_i\}_{i=1}^n$. We define an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where an edge $(i, j) \in \mathcal{E}$ indicates that a pairwise affinity is computed between tokens i and j . Cosine similarities $S_{ij} = \mathbf{f}_i^\top \mathbf{f}_j$ are converted to non-negative edge weights using a temperature-controlled kernel, $(\mathbf{W}_{\text{feat}})_{ij} = \exp(S_{ij}/\tau)$, $\tau > 0$, and self-loops are removed by setting $\text{diag}(\mathbf{W}_{\text{feat}}) = \mathbf{0}$. In practice, $\mathbf{W}_{\text{feat}} \in \mathbb{R}^{n \times n}$ therefore encodes the weighted adjacency of \mathcal{G} , defining the affinities between tokens in the feature-space in a way that reflects their visual and semantic proximity within the image.

Anchor-based prior integration For each input image, a small set of priors retrieved from the bank is injected into the affinity graph through two anchor vertices. These new vertices correspond to the two labels. We augment the initial graph with these anchors to form the block affinity

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{\text{feat}} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{0}_{2 \times 2} \end{bmatrix}, \quad (1)$$

where $\mathbf{C} \in \mathbb{R}^{n \times 2}$ links a token to its corresponding anchor with a uniform coupling α . We set the coupling strength proportional to the average affinity in the feature graph, $\alpha = \kappa \cdot \text{mean}\{(\mathbf{W}_{\text{feat}})_{ij} : i \neq j\}$, $\kappa > 0$, so anchors bias the solution while preserving the structure induced by \mathbf{W}_{feat} . Tokens indexed by the positive and complementary prior sets are connected to their respective anchors with weight α ; all other entries of \mathbf{C} are zero.

Spectral embedding On the augmented graph, we form $\mathbf{D} = \text{diag}(\mathbf{W})$ and solve the normalized-cut relaxation,

$$(\mathbf{D} - \mathbf{W}) \mathbf{v} = \lambda \mathbf{D} \mathbf{v}.$$

The second smallest eigenvector $\mathbf{v} \in \mathbb{R}^{n+2}$ (Fiedler vector) yields a globally consistent partition over tokens and anchors. Its token entries provide continuous scores that are oriented and binarized in Section 3.3. The presence of anchors modifies the eigenspace biasing the solution toward

cuts consistent with the injected priors, retaining the global grouping induced by the initial dense token affinities. See Appendix C for algorithm computational cost assessment.

3.3. Deterministic binarization

Prior-aware sign stabilization The Fiedler vector $\mathbf{v} \in \mathbb{R}^{n+2}$ from the augmented eigenproblem is sign-ambiguous. Let $\mathcal{L}_+, \mathcal{L}_- \subset \{1, \dots, n\}$ denote the token indices associated with the positive label (class of interest) and the complementary label, respectively. We orient \mathbf{v} using the priors by comparing class-wise means, $\mu_{\pm} = \frac{1}{|\mathcal{L}_{\pm}|} \sum_{i \in \mathcal{L}_{\pm}} v_i$, and applying a global flip $\mathbf{v} \leftarrow -\mathbf{v}$ if $\mu_+ < \mu_-$. For segmentation, we restrict to token entries and normalize to $[0, 1]$ via a min–max mapping, $\tilde{s}_i = \frac{v_i - \min_{j \leq n} v_j}{\max_{j \leq n} v_j - \min_{j \leq n} v_j}$, $i = 1, \dots, n$, yielding normalized scores $\tilde{\mathbf{s}} \in [0, 1]^n$.

Thresholding To obtain a binary mask, we estimate a data-driven threshold t^* from the priors labeled using a ROC-based criterion. Any candidate threshold t defines the empirical true- and false-positive rates in the prior sets

$$\text{TPR}(t) \text{ or } \text{FPR}(t) = \frac{1}{|\mathcal{L}_{\pm}|} \sum_{i \in \mathcal{L}_{\pm}} \mathbf{1}\{\tilde{s}_i > t\}.$$

We select the operating point that maximizes Youden’s J -statistic, the vertical distance to the ROC diagonal. $t^* = \arg \max_{t \in \mathcal{T}} (\text{TPR}(t) - \text{FPR}(t))$, where \mathcal{T} is the set of unique score values $\{\tilde{s}_i\}_{i=1}^n$ (or a dense grid on $[0, 1]$). The final token mask is $M_i = \mathbf{1}\{\tilde{s}_i > t^*\}$, $i = 1, \dots, n$.

4. Experiments

We evaluate PANC on three segmentation tasks to evaluate different capabilities: saliency detection guided by sparse priors (Section 4.1), class-aware segmentation (Section 4.2), and challenge-domain segmentation on homogeneous or texture-limited imagery (Section 4.3). Our evaluation focuses on the quality of segmentation per unit of supervision, controllability, and reproducibility introduced by anchor priors, as well as robustness across domain shifts. Ablation studies are reported in Section 4.4.

We use two frozen DINOv3 encoders [13] as feature backbones: ViT-H/16 (distilled) as the default for natural-image experiments and a satellite-pretrained ViT-L/16 (distilled) for low-diversity, texture-dominated domains. Evaluation is performed with per-image Intersection-over-Union (IoU) and aggregated as mean IoU (mIoU) per dataset. Unless otherwise stated, inputs are resized/padded to 1120 × 1120 pixels, experiments run on a single NVIDIA A100 (40 GB) GPU, and hyperparameters such as patch stride P , affinity temperature τ , and anchor coupling multiplier κ are chosen per-dataset. Reported runtimes reflect these choices.

4.1. Weakly-supervised saliency detection

On benchmarks of heterogeneous datasets for standard saliency detection, PANC automatically retrieves priors from the bank that are most representative of the input image’s content. In contrast with purely unsupervised methods, which rely on saliency heuristics, this process leverages our framework as a prior-assisted segmentation. The retrieved exemplars provide weak supervision to assist in spectral partitioning of the main subject. We compare PANC’s performance with the existing unsupervised and weakly supervised saliency detection methods.

Datasets. We evaluate on three standard heterogeneous saliency benchmarks: ECSSD [39], DUTS [40], and DUT-OMRON [41]. The ECSSD dataset contains 1000 images of complex scenes. For DUTS, we use the DUTS-TE test set (5,019 images) to validate comparisons against. Finally, DUT-OMRON contains 5168 images of everyday scenes.

Settings. Affinity temperature $\tau = 0.7$, anchor coupling $\kappa = 1.0$, prior bank size 30 images, 1500 vertices retrieved.

Results The benchmark against state-of-the-art in unsupervised and weakly-supervised methods, with the mIoU comparisons is summarized in Table 1. Our method consistently outperforms all unsupervised baselines across the three datasets. Compared to the weakly-supervised WSCUOD, PANC shows improved performance on ECSSD (+0.6%) and a solid gain on DUT-OMRON (+5.0%). The most significant result is on the DUTS dataset, where PANC achieves an mIoU of 74.8%, a +14.9% absolute improvement over the next-best method. This demonstrates the substantial benefit of our prior-assisted approach, as the retrieved priors help stabilize the spectral partition on this large and diverse dataset, a known challenge for purely unsupervised methods.

Figure 3 shows our method’s behaviour on sample ECSSD images; additional examples across other datasets are provided in Appendix D. The intermediate eigenvector attention (Eigen Attn.) map, derived from the Fiedler vector, clearly and smoothly localizes the object of interest. Deterministic thresholding, successfully converts this continuous map into a precise segmentation.

4.2. Controlled saliency detection

In contrast to general saliency detection, which targets the most prominent object, controlled, class-selective segmentation utilizes class-specific priors to guide the segmentation toward a particular object class.

Datasets The 2017 MS COCO validation set comprises 5,000 images and spans 80 object classes (e.g., person, dog,

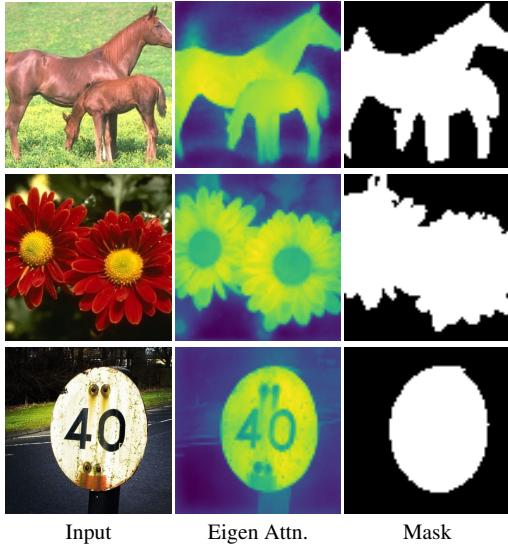


Figure 3. Examples on the ECSSD benchmark. Columns display the input image, the eigenvector-based attention map (normalized Fiedler scores visualized as a heatmap), and the binarized mask, for a different sample image (by rows).

airplane, handbag), resulting in a diverse collection of everyday real-world scenes. We construct three subsets: (1) 1042 images tagged with person class and at least one of its top-5 co-occurring classes (chair, car, handbag, dining table, cup); (2) 125 images containing dog and at least one of its top-5 co-occurring classes (person, car, couch, bed, frisbee); (3) 76 images containing both person and dog.

Settings. Increase hyperparameters with respect to previous task: anchor coupling $\kappa = 400$, 3500 vertices retrieved.

Results Figure 4 illustrates the core advantage of PANC: user-controllable, class-aware segmentation. It provides two examples (left and right) from the COCO dataset containing multiple co-occurring objects (people and dogs). By supplying priors for the ‘dog’ class (top rows), our method biases the spectral partitioning toward segmenting dogs. Crucially, when using ‘person’ priors on the *exact same* input image (bottom rows), the model correctly shifts its focus, as seen in the Eigen Attention map (center column) and only segments the people on the returned mask (right column), matching the ground truth (GT, left column). This highlights PANC’s ability to handle ambiguity in complex scenes and provide explicit control over the segmentation target, a feature absent in unsupervised methods.

4.3. Homogeneous and challenging-domain

Unlike diverse natural scenes, homogeneous and challenging-domain settings often lack texture or exhibit limited visual cues, causing conventional segmentation

Table 1. Comparison (mIoU %) for saliency detection. PANC is evaluated against unsupervised and weakly-supervised methods on three standard benchmarks.

| Method | ECSSD [39] | DUTS [40] | DUT-OMRON [41] |
|----------------------------------|-------------|-------------|----------------|
| <i>Unsupervised methods</i> | | | |
| BigBiGAN [42] | 67.2 | 49.8 | 45.3 |
| E-BigBiGAN [42] | 68.4 | 51.1 | 46.4 |
| FindGAN [43] | 71.3 | 52.8 | 50.9 |
| LOST [44] | 65.4 | 51.8 | 41.0 |
| DeepSpectral [5] | 64.5 | 47.1 | 42.8 |
| TokenCut [4] | 71.2 | 57.6 | 53.3 |
| <i>Weakly-supervised methods</i> | | | |
| WSCUOD [45] | 72.7 | 59.9 | 53.6 |
| PANC (ours) | 73.3 | 74.8 | 58.62 |

methods to struggle. In this context, the incorporation of prior knowledge plays a crucial role, enabling PANC to enhance segmentation quality substantially.

Datasets We selected three datasets that highlight different levels of visual complexity and domain difficulty. CUB-200-2011 [46] serves as a reference point for a homogeneous set. It comprises 5,794 images of bird species and displays consistent appearance and structural patterns. For a challenging and clinically sensitive domain, we utilize HAM10000 [47], comprising 10,015 dermoscopic images covering various types of skin lesions. Subtle textures and low inter-class contrast characterize these. Finally, the CrackForest Dataset (CFD) [48] is our textureless scenario, which contains 152 images of road surfaces with thin low-contrast cracks.

Settings Similar configuration used in the previous evaluations. We used the DINOv3 pretrained satellite VIT-L/16 as the main encoder on the HAM10000 and CFD datasets, where we found that it outperforms the embeddings extracted from the ViT-H/16 on low-diversity images. The prior bank is built from 5 images, we retrieve 2500 vertices per test sample. Anchor coupling multiplier $\kappa = 1000$.

Results The benchmark against state-of-the-art methods is summarized in Table 2. PANC demonstrates a strong advantage in these specialized domains. On the fine-grained CUB-200-2011 dataset, our method achieves 78.0 mIoU. Similarly, on the HAM10000 medical imaging dataset, PANC (78.8 mIoU) outperforms all baselines. The most significant improvement is on the CrackForest (CFD) dataset, which consists of low-texture, homogeneous images. Here, PANC achieves 96.8 mIoU, a massive +14.5% absolute improvement over the state-of-the-art.

Figure 5 illustrates the primary advantage of our approach: enhancing robustness on homogeneous and low-

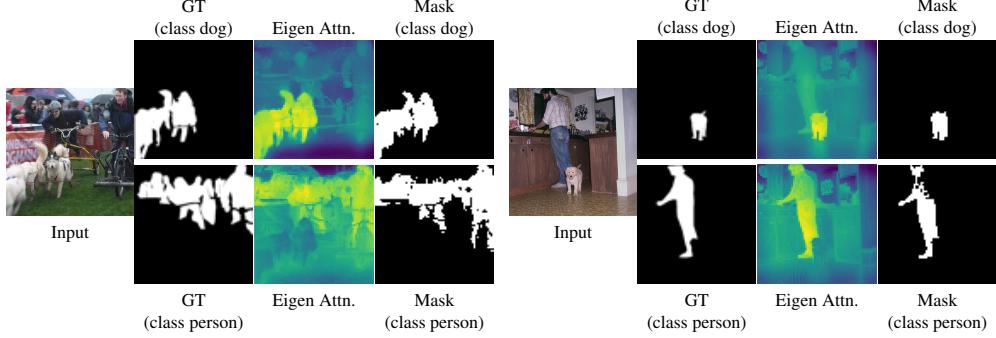


Figure 4. Visual examples of PANC’s class-selective controllability on MS COCO. For each input, we show two results: segmentation using ‘dog’ priors (top row) and ‘person’ priors (bottom row). The Eigen Attention map and final Mask correctly shift focus to the target class specified by the priors, demonstrating PANC’s ability to control segmentation in multi-object scenes.

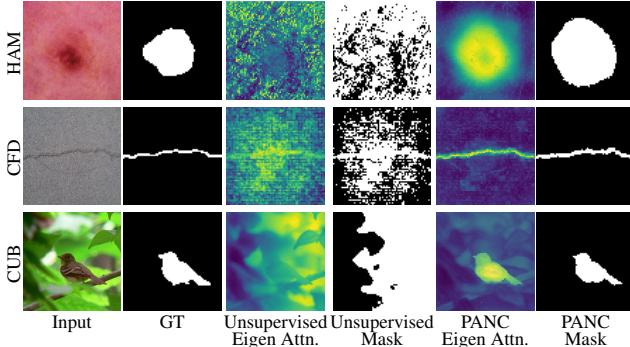


Figure 5. Qualitative comparison on challenging specialized datasets (HAM, CFD, CUB). This figure illustrates PANC’s robustness on homogeneous and low-semantic-content images where unsupervised baselines fail.

semantic-content images. In the first row, the dermoscopic lesion displays low contrast to the surrounding skin, along with surface artifacts such as faint scratches and hairs. In the third column, these artifacts cause the unsupervised baseline, TokenCut, to distribute attention across irrelevant regions rather than the lesion itself. After injecting the priors, attention concentrates on the lesion (fifth column), resulting in a more accurate segmentation mask (last column).

The second row displays a sample from the CrackForest dataset. The crack exhibits texture and intensity patterns nearly indistinguishable from the background. The unsupervised model fails to locate the crack, whereas our method, guided by prior information, yields a segmentation that closely matches the ground truth.

Finally, on the homogeneous dataset (third row), while unsupervised methods fail to segment the class, unable to direct attention to the salient object (third row, third column), the class vertex priors successfully redirect attention to the desired class.

4.4. Ablation studies

We assess two representative datasets for saliency detection and challenging-domain tasks addressed in this paper: DUTS and CrackForest (CFD).

Impact of injected vertices As shown in Figure 6, the number of injected vertices impacts performance differently depending on the dataset type, showing that homogeneous datasets benefit from adding more exemplars (vertices) to the graph. This is consistent with our approach, as a larger set of relevant representations strengthens the target cluster. In contrast, heterogeneous datasets show a detrimental effect as the vertex count grows. This is likely because a larger retrieval of vertices in a diverse domain increases the chance of pulling irrelevant or ambiguous exemplars, which inject noise into the graph affinities and degrade the partition quality.

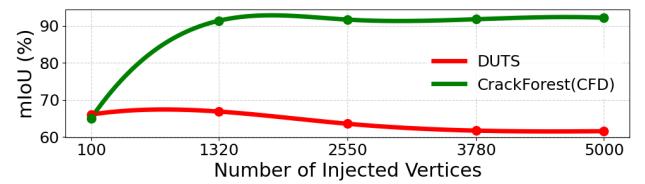


Figure 6. **Mean IoU as a function of the injected vertices.** Increasing the number of vertices introduces noise in heterogeneous datasets, reducing performance, whereas homogeneous datasets benefit from additional vertices, improving prediction quality.

Impact of anchor coupling The anchor coupling multiplier, κ , controls the influence of the prior anchors on the graph. As shown in Table 3, its impact differs by dataset. For the homogeneous CFD dataset, performance consistently improves as κ increases, with the gains saturating at high values. This is expected, as giving a stronger weight

Table 2. Comparison (mIoU %) for segmentation on homogeneous and challenging-domain datasets. PANC is evaluated against unsupervised and weakly-supervised methods on the CUB-200-2011, CFD, and HAM10000 datasets.

| Method | CUB-200-2011 [49] | Method | CrackForest (CFD) [51] | Method | HAM10000 [47] | |
|----------------------------------|-------------------|----------------------------------|-----------------------------|--------------------|----------------------------------|--|
| <i>Unsupervised methods</i> | | | <i>Unsupervised methods</i> | | | |
| BigBiGAN [42] | 68.3 | DeepSpectral [5] | 82.3 | DeepSpectral [5] | 78.4 | |
| E-BigBiGAN [42] | 71.0 | UP-CrackNet [52] | 30.5 | TokenCut [4] | 67.5 | |
| FindGAN [43] | 66.4 | <i>Weakly-supervised methods</i> | | | <i>Weakly-supervised methods</i> | |
| LOST [44] | 68.8 | UWSCS [53] | 74.5 | SG-MIAN [54] | 74.3 | |
| DeepSpectral [5] | 66.7 | PANC (ours) | 96.8 | TS-CAM [55] | 67.5 | |
| TokenCut [4] | 74.8 | | | PANC (ours) | 78.8 | |
| <i>Weakly-supervised methods</i> | | | | | | |
| PFENet [50] | 72.4 | | | | | |
| WSCUOD [45] | 77.8 | | | | | |
| PANC (ours) | 78.0 | | | | | |

Table 3. **Ablation study on PANC parameters** Segmentation (mIoU%) on heterogeneous (DUTS) and homogeneous (CFD) datasets.

| κ | Anchor coupling | | Affinity temperature | | | H, W | Resolution | | Thresholding strategies | | |
|----------|-----------------|-------------|----------------------|-------------|-------------|--------|-------------|-------------|-------------------------|-------------|-------------|
| | DUTS [40] | CFD [51] | τ | DUTS [40] | CFD [51] | | DUTS [40] | CFD [51] | Method | DUTS [40] | CFD [51] |
| 1 | 68.2 | 92.2 | 0.10 | 72.5 | 92.8 | 160 | 42.2 | 89.1 | Median | 70.1 | 91.6 |
| 10 | 68.5 | 92.8 | 0.40 | 70.3 | 92.0 | 480 | 74.8 | 96.5 | ROC | 70.1 | 92.3 |
| 100 | 69.4 | 96.2 | 0.70 | 70.1 | 91.6 | 880 | 61.3 | 95.7 | GMM | 66.5 | 91.0 |
| 1000 | 59.5 | 96.8 | 1.00 | 70.0 | 91.6 | 1120 | 66.5 | 91.0 | Platt | 68.8 | 92.30 |

to the connections between image tokens and the highly-relevant ‘crack’ anchors reinforces the correct partition. In contrast, for the heterogeneous DUTS dataset, performance peaks at a moderate κ and then drops significantly.

Impact of affinity temperature As shown in Table 3, the affinity temperature τ influences the segmentation quality. We found that a sharp kernel with $\tau = 0.1$ yields the best results. Performance seems robust to this hyperparameter, similar values do not significantly affect.

Resolutions In Table 3, we report different input image resolutions. The maximum resolution is constrained by GPU memory. Although higher resolutions would intuitively yield more accurate masks, we observe that intermediate resolutions perform best. We suspect this behavior is linked to the DINoV3 training strategy, which relies on 256 and 112 pixel global and local crops, and later adapted to larger image sizes.

Thresholding strategies We compared four different strategies for binarizing the continuous Fiedler vector score into a final mask. As shown in Table 3, the method based on finding the best threshold via a Receiver Operating Characteristic (ROC) curve analysis yielded the best performance on the homogeneous CFD dataset (92.3 mIoU) and matched

the top performance on the heterogeneous DUTS dataset (70.1 mIoU). Given these results, we selected the ROC-based strategy as our default binarization method.

5. Conclusions

Inspired by the intrinsic limitations of the normalized cut and the structure of affinity graphs built on patched visual embeddings, we successfully formulated and tested a compact, weakly supervised framework for image segmentation. PANC leverages manual annotation and embeddings, manipulating the affinity graph, to produce dense segmentation of visually and semantically consistent classes.

The frozen DINoV3 proved to be a reliable feature extractor, providing multi-resolution capabilities. Nevertheless, other backbone models might be equally suitable. The PANC method outperformed the state-of-the-art in weakly supervised and class-specific saliency detection, and excelled in annotating homogeneous and challenging domain datasets, where few-shot capabilities are highly valued.

Future contributions shall expand on how to efficiently build, search, and maintain the banks of priors and affinity graphs. The selection of images for annotation may be extended beyond our initial proposal for clustering of the embedded tokens. When annotating large datasets, one may expect an increase in the number of manual annotations, but reusing know-how as possible due to redundancy.

The computation and operation of the affinity graph

is the bottleneck for all normalized cut methods to operate at scale. Simplifications of the adjacency matrix based on the similarity metric and local affinities shall help reduce complexity and lighten computational burden.

References

- [1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. [1](#), [3](#), [14](#)
- [2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What’s the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. [1](#), [3](#), [14](#)
- [3] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels, 2021. [1](#), [3](#)
- [4] Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut, 2023. [1](#), [3](#), [6](#), [8](#), [12](#), [14](#), [15](#)
- [5] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8364–8375, 2022. [1](#), [3](#), [6](#), [8](#), [14](#)
- [6] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016. [1](#), [3](#)
- [7] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4981–4990, 2018. [1](#), [3](#)
- [8] Xiang Wang, Buyue Qian, and Ian Davidson. On constrained spectral clustering and its applications. *Data Mining and Knowledge Discovery*, 28(1):1–30, 2014. [1](#), [3](#)
- [9] Linli Xu, Wenye Li, and Dale Schuurmans. Fast normalized cut with linear constraints. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2866–2873. IEEE, 2009. [3](#)
- [10] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015. [1](#), [3](#)
- [11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. [1](#), [2](#), [3](#)
- [12] Maxime Oquab, Timothée Darzet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. [2](#), [3](#), [15](#)
- [13] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darzet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprise, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. [1](#), [2](#), [3](#), [5](#)
- [14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. [2](#)
- [15] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer, 2022.
- [16] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. [2](#)
- [17] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution, 2024. [2](#), [3](#)
- [18] Ronan Docherty, Antonis Vamvakeros, and Samuel J. Cooper. Upsampling dinov2 features for unsupervised vision tasks and weakly supervised materials segmentation, 2025. [2](#)
- [19] Armand Joulin, Francis Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1943–1950. IEEE, 2010. [3](#)
- [20] Armand Joulin, Francis Bach, and Jean Ponce. Multi-class cosegmentation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 542–549. IEEE, 2012.
- [21] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov. Object cosegmentation. In *CVPR 2011*, pages 2217–2224. IEEE, 2011.
- [22] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1201–1210, 2015.
- [23] Huy V Vo, Francis Bach, Minsu Cho, Kai Han, Yann LeCun, Patrick Pérez, and Jean Ponce. Unsupervised image matching and object discovery as optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8287–8296, 2019.
- [24] Huy V Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *European Conference on Computer Vision*, pages 779–795. Springer, 2020. [3](#)

- [25] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 3
- [26] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000. 3, 14
- [27] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007. 3
- [28] Leo Grady. Random walks for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 28(11):1768–1783, 2006. 3
- [29] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003. 3
- [30] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001. 3
- [31] Xinlei Chen and Deng Cai. Large scale spectral clustering with landmark-based representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 313–318, 2011. 14
- [32] Farhad Pourkamali-Anaraki. Scalable spectral clustering with nyström approximation: Practical and theoretical aspects. *IEEE Open Journal of Signal Processing*, 1:242–256, 2020. 3, 14
- [33] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015. 3
- [34] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015. 3
- [35] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. “ grabcut” interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004. 3
- [36] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision*, pages 695–711. Springer, 2016. 3
- [37] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1818–1827, 2018. 3
- [38] Qingchen Tang, Lei Fan, Maurice Pagnucco, and Yang Song. Prototype-based image prompting for weakly supervised histopathological image segmentation, 2025. 3
- [39] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):717–729, 2015. 5, 6
- [40] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 136–145, 2017. 5, 6, 8, 14
- [41] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013. 5, 6
- [42] Andrey Voynov, Stanislav Morozov, and Artem Babenko. Object segmentation without labels with large-scale generative models. In *International Conference on Machine Learning*, pages 10596–10606. PMLR, 2021. 6, 8
- [43] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Finding an unsupervised image segmenter in each of your deep generative models. *arXiv preprint arXiv:2105.08127*, 2021. 6, 8
- [44] Xi Shen, Alexei A Efros, Armand Joulin, and Mathieu Aubry. Learning co-segmentation by segment swapping for retrieval and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5082–5092, 2022. 6, 8
- [45] Yunqiu Lv, Jing Zhang, Nick Barnes, and Yuchao Dai. Weakly-supervised contrastive learning for unsupervised object discovery. *IEEE Transactions on Image Processing*, 33:2689–2702, 2024. 6, 8
- [46] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6
- [47] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 6, 8
- [48] Yong Shi, Limeng Cui, Zhiqian Qi, Fan Meng, and Zhen-song Chen. Automatic road crack detection using random structured forests. *IEEE Transactions on Intelligent Transportation Systems*, 17(12):3434–3445, 2016. 6
- [49] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds-200-2011. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 8
- [50] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):1050–1065, 2020. 8
- [51] Yong Shi, Limeng Cui, Zhiqian Qi, Fan Meng, and Zhen-song Chen. Automatic road crack detection using random structured forests. *IEEE Transactions on Intelligent Transportation Systems*, 17(12):3434–3445, 2016. 8
- [52] Nachuan Ma, Rui Fan, and Lihua Xie. Up-cracknet: Unsupervised pixel-wise road crack detection via adversarial image restoration. *IEEE Transactions on Intelligent Transportation Systems*, 25(10):13926–13936, 2024. 8
- [53] Chao Xiang, Vincent JL Gan, Lu Deng, Jingjing Guo, and Shaopeng Xu. Unified weakly and semi-supervised crack

- segmentation framework using limited coarse labels. *Engineering Applications of Artificial Intelligence*, 133:108497, 2024. 8
- [54] Zhixun Li, Nan Zhang, Huijing Gong, Ruiyun Qiu, and Wei Zhang. Sg-mian: Self-guided multiple information aggregation network for image-level weakly supervised skin lesion segmentation. *Computers in Biology and Medicine*, 170:107988, 2024. 8
- [55] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2886–2895, 2021. 8

A. GPU-Accelerated Spectral Partitioning

This appendix details the implementation of the PANC framework on a fully device-resident design. While spectral clustering is mathematically elegant, standard implementations (e.g., TokenCut [4]) rely on CPU-bound solvers like ‘scipy.sparse’. The cubic complexity of full eigendecomposition and the latency of moving dense tensors across the PCI-E bus cause a significant bottleneck.

We developed an implementation that mitigates these limitations by managing all tensors—features, affinities, and eigenvectors—strictly on the VRAM. We use iterative solvers and vectorized matrix operations to reduce effective computational complexity; and process high-resolution token grids near real time. Hereafter, we describe the algorithmic logic, the details, and the scalability of our implementation. The repository with the source code is available in [URL omitted for anonymity].

A.1. Algorithm Overview

Algorithm 1 outlines the core logic. The pipeline avoids explicit token loops, using batched tensor operations to maximize GPU occupancy. For $Z \in \mathbb{R}^{N \times d}$ being the matrix containing N staked embeddings coming from the patched input images (to be segmented) and $Z_P \in \mathbb{R}^{P \times d}$ the matrix concatenation of embeddings corresponding to the P patches of annotation priors, all with dimension d , we define y the label corresponding to each prior patch—i.e., background or foreground—, τ the affinity temperature hyperparameter, and the concatenation of inputs and priors, F :

Algorithm 1: PANC Pipeline

- 1: **input:** $Z \in \mathbb{R}^{N \times d}$, $Z_P \in \mathbb{R}^{P \times d}$, $y \in \{0, 1\}^P$, τ
 - 2: $F \leftarrow [Z; Z_P] \in \mathbb{R}^{(N+P) \times d}$
 - 3: $S \leftarrow FF^\top$
 - 4: $W_{\text{feat}} \leftarrow \text{affinity}(S, \tau)$
 - 5: add fore- and back-ground anchors to W_{feat} (1).
 - 6: compute $D = \text{diag}(W\mathbf{1})$
 - 7: $L \leftarrow I - D^{-1/2}WD^{-1/2}$
 - 8: compute second eigenvalue and eigenvector of L :
 $\lambda_1 \leq \lambda_2 \leq \dots$ and v_1, v_2, \dots
 - 9: $v \leftarrow v_2$
 - 10: if $\text{median}(v_{y=1}) < \text{median}(v_{y=0})$: $v \leftarrow (1 - v)$
 - 11: $s \leftarrow \text{min-max normalize } 0 \leq v \leq 1$
 - 12: $t \leftarrow \text{threshold}(s_{N+1:N+P}, y)$
 - 13: **output:** mask = $\mathbf{1}\{s_{1:N} > t\}$
-

A.2. Implementation and Complexity

Affinity Construction. Constructing the fully-connected affinity matrix requires computing pairwise similarities between all tokens. Although theoretically $O(N^2)$, this operation maps perfectly to GPU-optimized General Matrix

Multiplication (GEMM) kernels using PyTorch. By operating entirely in the VRAM, we avoid the iterator overhead of CPU construction.

```
# Complexity: O(N^2 * d) - Dominated by GEMM
# Executed as a single block-wise GPU kernel
S = feat_nodes @ feat_nodes.T
A_feat = torch.exp(S / (tau + eps))
```

Iterative Spectral Solver. Standard eigensolvers compute the full spectrum with $O(N^3)$ complexity. However, PANC requires only the Fiedler vector (the second smallest eigenvector). We employ the Locally Optimal Block Preconditioned Conjugate Gradient (LOBPCG) algorithm. This iterative approach reduces complexity to approximately $O(I \cdot N^2)$, where I is the number of iterations (typically < 100), making it feasible for dense graphs.

```
# Complexity: O(I * k * N^2)
# Significantly faster than full decomposition O(N^3)
# k=2 eigenvectors, I=iterations
init = torch.randn(M, k, device=device)
_, vecs = torch.lobpcg(L_sym, k=k, X=init,
                      largest=False)
v = vecs[:, 1]
```

Deterministic Orientation and Thresholding. We use the sparse priors to deterministically orient the eigenvector. For binarization, we vectorize the ROC analysis. Instead of iterating through thresholds (which is slow in Python), we use tensor broadcasting to compute the J-statistic for all candidate thresholds in parallel on the GPU.

```
# Sign stabilization
# Complexity: O(P) - Negligible cost
if (fg_med - bg_med) < 0: v = -v

# ROC Thresholding: Maximize Youden's J (TPR - FPR)
# Complexity: O(T * P) where T is num steps, P is priors
# Vectorized implementation avoids CPU loops
steps = torch.linspace(0, 1, 200, device=dev)
preds = prior_scores.unsqueeze(0) > steps.
    unsqueeze(1) #(T,P)

# Compute True Positives and False Positives in parallel
tps = (preds & (labels==1)).sum(1)
fps = (preds & (labels==0)).sum(1)

# Calculate J-statistic and select best threshold
tpr = tps / (labels==1).sum()
fpr = fps / (labels==0).sum()
thresh = steps[(tpr - fpr).argmax()]
```

A.3. Scalability

By removing the CPU-GPU transfer bottleneck, the scalability of PANC is primarily constrained by VRAM capacity rather than compute time. A standard NVIDIA A100 (40GB) can accommodate the dense $O(N^2)$ affinity matrix for image resolutions up to 1120×1120 pixels (approx. 5,000 tokens using DINOv3). For applications requiring higher resolutions, the modular design allows the dense solver to be swapped for a Nyström approximation with minimal code changes.

B. Annotation Protocol Details

This appendix details the Human-in-the-Loop (HITL) annotation protocol used to generate the sparse supervision signals required by PANC. We describe an interactive workflow that produces high-quality token-level labels using simple point prompts, trading negligible human effort for precise semantic control.

B.1. Annotation Interface

We utilize the Computer Vision Annotation Tool (CVAT) integrated with the Segment Anything Model 2 (SAM 2) backend, using GPU support. This setup replaces the laborious process of manual polygon tracing with a rapid “prompt-and-verify” workflow. The interface allows annotators to interact directly with the image using sparse clicks, which are immediately translated into dense pixel masks by the underlying foundation model.

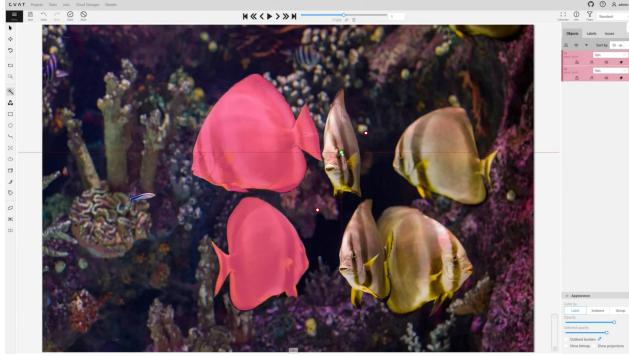


Figure 7. Sparse Point Annotation Interface. The annotator defines the target object using positive clicks (green) and defines the background or distracting regions using negative clicks (red). SAM 2 instantly visualizes the resulting segmentation mask (blue overlay) for verification.

B.2. Positive and Negative Point Protocol

The annotation strategy is strictly defined by positive (foreground) and negative (background) point interactions. This binary distinction is crucial for properly grounding the spectral affinity graph used in PANC.

- Positive Clicks (Target Definition):** The annotator places one or more clicks on the semantic object of interest. This establishes the core “positive” anchor nodes in the graph.
- Negative Clicks (Context Separation):** Unlike bounding box supervision, our protocol explicitly requires negative clicks on:
 - Immediate background regions to define boundaries.
 - Semantically distinct adjacent objects (e.g., clicking a person as negative when the target is a dog).
 These negative points are essential for resolving ambiguity in complex scenes, serving as “repulsive” anchors in the spectral embedding.
- SAM 2 Inference & Refinement:** SAM 2 predicts a dense mask based on current points. If the mask “bleeds” into the background, the user adds a negative point. If it misses part of the object, the user adds a positive point.
- Tokenization:** Once the user accepts the mask, it is downsampled to the ViT patch grid. Patches with $> 50\%$ overlap are assigned the positive label, while user-identified background areas are assigned the negative label.

B.3. Advantages over Dense Manual Annotation

The proposed point-based protocol offers distinct advantages over traditional fully supervised manual annotation (pixel-level painting or polygon tracing):

- Motor Efficiency:** Drawing a precise polygon around a complex object (e.g., a tree or a bicycle) requires fine motor control and hundreds of vertices. In contrast, placing a point is an $O(1)$ operation. Our protocol typically requires only 3–5 clicks to segment complex objects.
- Ambiguity Resolution:** Simple bounding boxes often contain background noise or overlapping objects. By selecting explicit positive and negative points, the annotator provides a stricter definition of the decision boundary, which is critical for the graph partitioning stage.
- Speed:** The integration of SAM 2 allows the user to operate at the speed of verification rather than generation. This is, masks are automatically generated and annotators merely review the segmentation results.

B.4. Time Cost Analysis

We evaluated the annotation cost by measuring the average interaction time per image and extrapolating these metrics to the construction of full Prior Banks. The efficiency of our approach is driven by the reduced physical interaction required to define complex shapes compared to traditional methods.

Per-Image Annotation Time. Our point-based protocol averages approximately 45 seconds per image. This duration includes the entire interaction loop:

- **Image Loading & Inspection:** ≈ 10 seconds
- **Point Prompting (Pos/Neg):** ≈ 15 seconds
- **Mask Verification & Saving:** ≈ 20 seconds

In contrast, fully manual polygon annotation for semantic segmentation typically requires 5 to 10 minutes per object, depending on boundary complexity and even for expert annotators [1, 2]. Consequently, our protocol yields a speedup factor of roughly $6\times$ to $10\times$ per image while retaining human semantic oversight; in specific segmentation tasks, this factor increases up to $100\times$.

Total Prior Bank Setup Time. The total time investment depends on the diversity of the target domain, which dictates the required size of the Prior Bank (K).

- **Homogeneous Domains ($K = 5$):** For specialized tasks like surface crack detection (CFD) or skin lesion segmentation (HAM10000), where visual variance is low, a bank of just 5 images is sufficient.

$$\text{Total Time} \approx 5 \text{ images} \times 45 \text{ sec/image} \approx \mathbf{3.75} \text{ minutes}$$

This minimal setup time (< 5 mins) enables rapid deployment of the model to new industrial or medical domains.

- **Heterogeneous Domains ($K = 30$):** For diverse natural scene benchmarks (e.g., DUTS, ECSSD, MS COCO) where objects exhibit significant variance in appearance and pose, we utilize a larger bank of 30 images.

$$\text{Total Time} \approx 30 \text{ images} \times 45 \text{ sec/image} \approx \mathbf{22.5} \text{ minutes}$$

Even for these complex general domains, the entire supervision set is generated in under half an hour—orders of magnitude faster than the hundreds of hours required to compile dense supervised training sets.

C. Performance Assessment

This appendix focuses on the computational capabilities and requirements of the PANC framework. We provide a detailed breakdown of the memory footprint (MB) and floating-point operations (FLOPs) required for the dense spectral partitioning and analyze the trade-offs when employing sparse approximations [31, 32]. We also evaluate the impact of pipeline components—such as backbone selection, prior selection strategies, and image resolution—on both computational demand and on spectral localization.

C.1. Experimental Setup and Metrics

Hardware and Profiling. All computational benchmarks were performed on a single NVIDIA A100 GPU with 40GB of VRAM. Profiling focused on the inference stage of the pipeline, which includes feature extraction, affinity graph construction, eigensolving, and mask binarization.

Evaluation Metrics. To assess efficiency and spectral quality, we utilize the following metrics:

- **FLOPs (Floating Point Operations):** Measures the total computational cost of the inference pass. This ought to be dominated by the dense matrix multiplication for affinity construction and the iterative matrix-vector products in the eigensolver [26].
- **Peak Memory (MB):** The maximum VRAM allocated during the forward pass. The peak ought to match the storage of the $N \times N$ affinity matrix, which scales quadratically with the number of tokens [4].
- **Inverse Participation Ratio (IPR):** To quantify the quality of the spectral embedding without relying on ground truth labels, we compute the IPR of the Fiedler vector v . The IPR measures the localization of the eigenvector: a higher value indicates that the energy of the vector is concentrated on a specific region (the object); a low value suggests delocalized noise [5]. The metric is defined as:

$$\text{IPR} = \frac{\sum_i v_i^4}{(\sum_i v_i^2)^2} \quad (2)$$

C.2. Components' Computational Performance

We evaluate how different hyperparameters and architectural choices affect computational cost and spectral quality.

Selection Strategy. We assess the impact of the priors' selection mechanism. For all experiments in this ablation, we used a fixed configuration derived from the optimal hyperparameters for DUTS-TE: a prior bank size of 30 and 1250 retrieved tokens injected. As shown in Table 4, we observe a clear trade-off. While stochastic selection (Random) minimizes computational overhead, it fails to provide semantic guidance. Our diversity-aware strategy (PANC) incurs higher GFLOPs and memory usage due to similarity search and diversity checks, but this computational investment is necessary to optimize spectral localization (IPR) compared to the naïve baselines.

| Strategy | Mechanism | GFLOPs | Peak Mem (MB) | Avg. IPR ($\times 10^{-4}$) |
|----------|----------------------------|--------|---------------|-------------------------------|
| Random | Stochastic selection | 97 | 6182 | 3.24 |
| Nearest | Top- K cosine similarity | 3170 | 10499 | 4.57 |
| PANC | Nearest + MMR Diversity | 3970 | 10493 | 5.34 |

Table 4. Selection Strategy Ablation. Evaluated on DUTS-TE [40], bank size of 30.

Component Cost & Scalability Analysis. Table 5 presents a comprehensive ablation study quantifying the

Table 5. Extended evaluation of computational resources (GFLOPs) and peak memory usage on an NVIDIA A100.

| Method | Backbone | Resolution | Tokens (N) | # Priors (M) | GFLOPs | Peak Mem (MB) |
|---|---------------|--------------------|----------------|------------------|--------|---------------|
| <i>Impact of Injected Vertices (M)</i> | | | | | | |
| TokenCut [4] | DINOv3-H+ | 480×480 | 1,156 | 0 | 567 | 6,705 |
| PANC | DINOv3-H+ | 480×480 | 1,156 | 10 | 567 | 6,705 |
| PANC | DINOv3-H+ | 480×480 | 1,156 | 100 | 572 | 6,705 |
| PANC | DINOv3-H+ | 480×480 | 1,156 | 1,000 | 1,086 | 6,705 |
| PANC | DINOv3-H+ | 480×480 | 1,156 | 5,000 | 12,916 | 16,218 |
| <i>Impact of Resolution</i> | | | | | | |
| PANC | DINOv3-H+ | 224×224 | 256 | 1,000 | 677 | 6,217 |
| PANC | DINOv3-H+ | 480×480 | 1,156 | 1,000 | 1,085 | 6,705 |
| PANC | DINOv3-H+ | 896×896 | 4,096 | 1,000 | 2,522 | 8,823 |
| PANC | DINOv3-H+ | 1120×1120 | 6,400 | 1,000 | 3,674 | 10,493 |
| PANC | DINOv3-H+ | 1344×1344 | 9,216 | 1,000 | 5,103 | 12,534 |
| <i>Impact of Backbone Architecture</i> | | | | | | |
| PANC | DINOv2-L [12] | 480×480 | 1,156 | 1,000 | 306 | 1,849 |
| PANC | DINOv3-S | 480×480 | 1,156 | 1,000 | 122 | 603 |
| PANC | DINOv3-S+ | 480×480 | 1,156 | 1,000 | 115 | 632 |
| PANC | DINOv3-B | 480×480 | 1,156 | 1,000 | 223 | 948 |
| PANC | DINOv3-L | 480×480 | 1,156 | 1,000 | 306 | 1,849 |

computational overhead of the PANC framework. We analyze three critical scaling dimensions:

1. **Number of Injected Priors (M):** We evaluate the impact of augmenting the graph with an increasing number of annotated vertices, scaling from $M = 0$ (unsupervised baseline [4]) up to $M = 5,000$. For typical usage ($M \leq 100$), the overhead is negligible. However, injecting a massive number of priors ($M = 5,000$) drastically expands the graph connectivity, resulting in a $\approx 20\times$ spike in GFLOPs. This confirms that PANC is most efficient when operating with a compact set of highly relevant priors.
2. **Resolution Scaling:** We evaluate input resolutions from 224^2 up to 1344^2 . As shown in Table 5, the transition from 480^2 to 1120^2 results in a $\approx 3.4\times$ increase in GFLOPs and a $\approx 1.6\times$ increase in memory. While memory growth is sub-quadratic due to fixed backbone overheads, computational demand scales sharply with token count.
3. **Backbone Efficiency:** We also benchmark the DINOv3 family against the DINOv2-L standard. Our results indicate that DINOv3-L matches the computational footprint of legacy DINOv2-L (306 GFLOPs). In resource-constrained scenarios, DINOv3-S+ offers a lightweight alternative, reducing GFLOPs by 62% compared to the Large model while maintaining modern architectural benefits.

As observed, the injection of priors introduces negligible computational overhead compared to the unsupervised baseline [4] when $M \leq 100$. High-resolution processing increases memory demand, yet it remains feasible on modern hardware (see Section C.3 for optimization strategies).

C.3. Scalability: Sparse vs. Dense Approximation

The primary computational bottleneck for high-resolution inputs ($N \geq 6400$ tokens) is the storage and processing of the affinity matrix $W \in \mathbb{R}^{N \times N}$. To mitigate the quadratic memory cost, we implement a Top- ξ sparsification strategy within the PANC pipeline.

Specifically, after computing token similarities, we instantiate the affinity matrix W as a sparse tensor by retaining only the ξ strongest connections per row. This sparse structure is propagated to the normalized Laplacian L . Con-

Table 6. Scalability Analysis at High Resolution (1120×1120). We compare the Dense baseline against Sparse topologies with varying connectivity (ξ). We report Graph Density (percentage of non-zero entries) to highlight the significant reduction in stored values between $\xi = 20$ and the dense baseline.

| Graph Topology | Density (%) | GFLOPs | Peak Mem (MB) |
|-----------------------|---------------|-------------|---------------|
| Dense (Default) | 100.00% | 3674 | 10493 |
| Sparse ($\xi = 80$) | $\sim 1.25\%$ | 3180 | 6520 |
| Sparse ($\xi = 20$) | $\sim 0.31\%$ | 2915 | 6150 |

| Input | GT | Eigen Attn | Prediction | Priors Bank |
|-------|----|------------|------------|-------------|
| | | | | |
| | | | | |

| Input | GT | Eigen Attn | Prediction | Priors Bank |
|-------|----|------------|------------|-------------|
| | | | | |
| | | | | |

Figure 8. Controllability via Prior Banks. We demonstrate how the model resolves semantic ambiguity in the input image by conditioning on different priors banks. For the same input scene containing multiple objects, changing the support bank shifts the eigen attention maps to the corresponding target, leading to a prediction that aligns with the specific Ground Truth (GT).

sequently, the matrix multiplication steps within the iterative LOBPCG eigensolver operate directly on sparse tensors, reducing memory demand from $O(N^2)$ to $O(N \cdot \xi)$.

Table 6 compares the dense baseline against sparse variants at high resolution (1120×1120). By reducing the graph density to just 1.25% ($\xi = 80$), we achieve a 38% reduction in peak memory. Aggressive sparsification to $\xi = 20$ (density 0.31%) yields further savings, reducing memory footprint to 6.15 GB, nearly half that of the dense baseline.

D. Extended examples

D.1. Multi-object controllability segmentation

Finally, we extend our class controllability tests on the MS COCO dataset, demonstrating that efficient prior bank generation enables high-quality segmentation of a class object in never-seen images.

We divide the examples of the validation set into two main categories based on the geometric deformability of the target class: rigid classes that mainly change their form with the perspective (see Figs. 9 and 10) and non-rigid classes with a deformable structure (see Figs. 11, 12, and 13). Despite the variety of airplanes and boats, the results are consistent from all points of view. A similar situation was identified on bananas, ties, and suitcases. In contrast, among the

bananas, PANC was able to identify a printed one that was not included in the ground truth—see Fig. 11, second row from the bottom.

D.2. On priors selection and effective representation

To illustrate the role of prior banks, we computed segmentation results for different numbers of priors. In Fig. 8, we include input images (left) with their corresponding ground truth segmentation for two different classes, along with the eigen-attention map and the result of the PANC segmentation. Each of the latter was computed using the tokens from a different prior bank. To the right, we include the manually annotated images from which the tokens in the prior bank were computed. Essentially, varying priors can improve segmentation quality and detail, but at the cost of increasing the computational load as the bank size increases. Examples illustrate variations across classes, demonstrating the adaptability of priors to bias results toward the segmentation targets.

We also tested priors derived from another data set to demonstrate transfer learning across domains. Using priors from the CUB data set on COCO birds, compared with segmentation results obtained on the original priors from the COCO dataset, produced visually similar results, as illustrated in Figure 14.

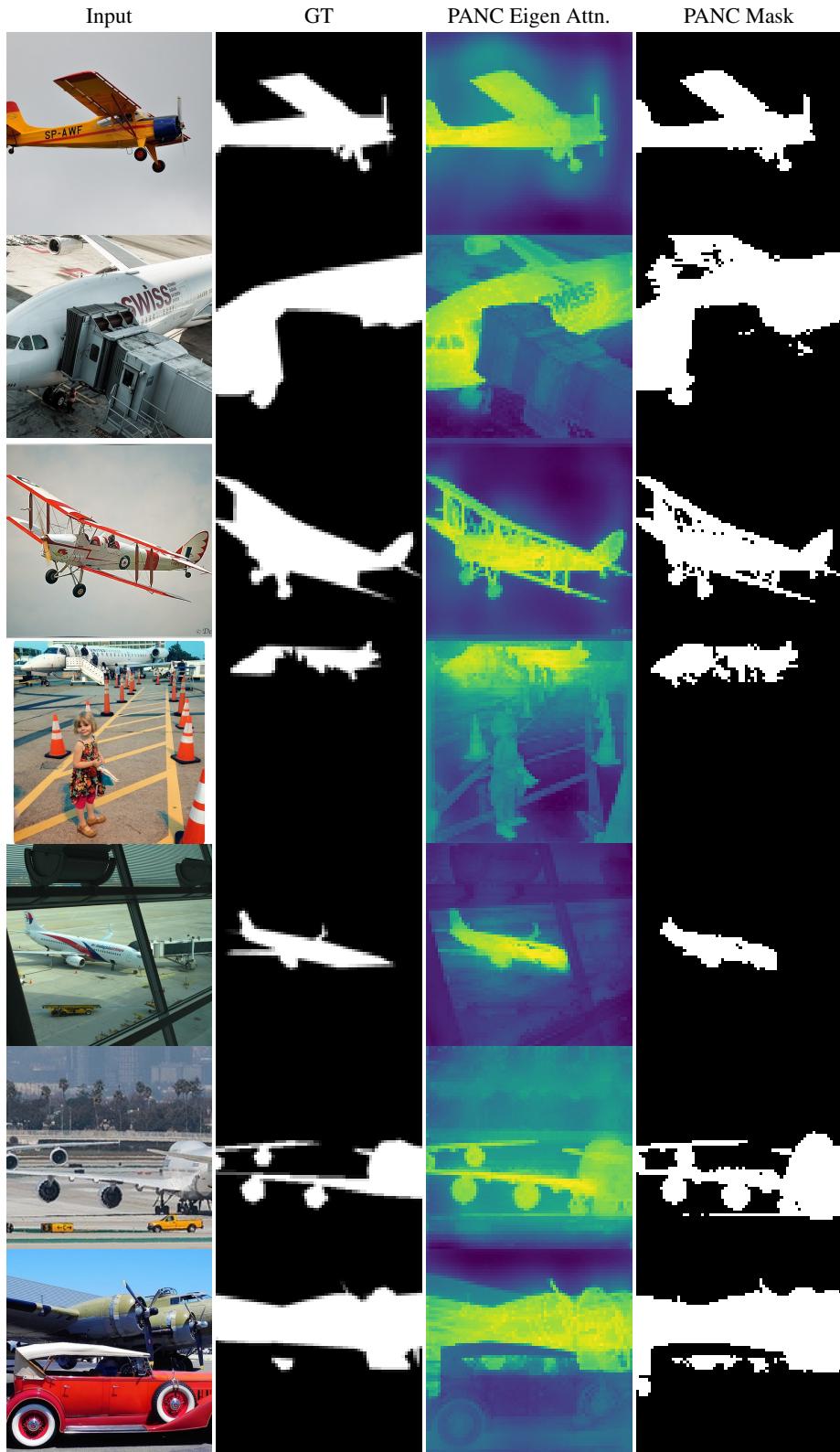


Figure 9. Additional qualitative comparison on rigid MSCOCO **Airplane** class

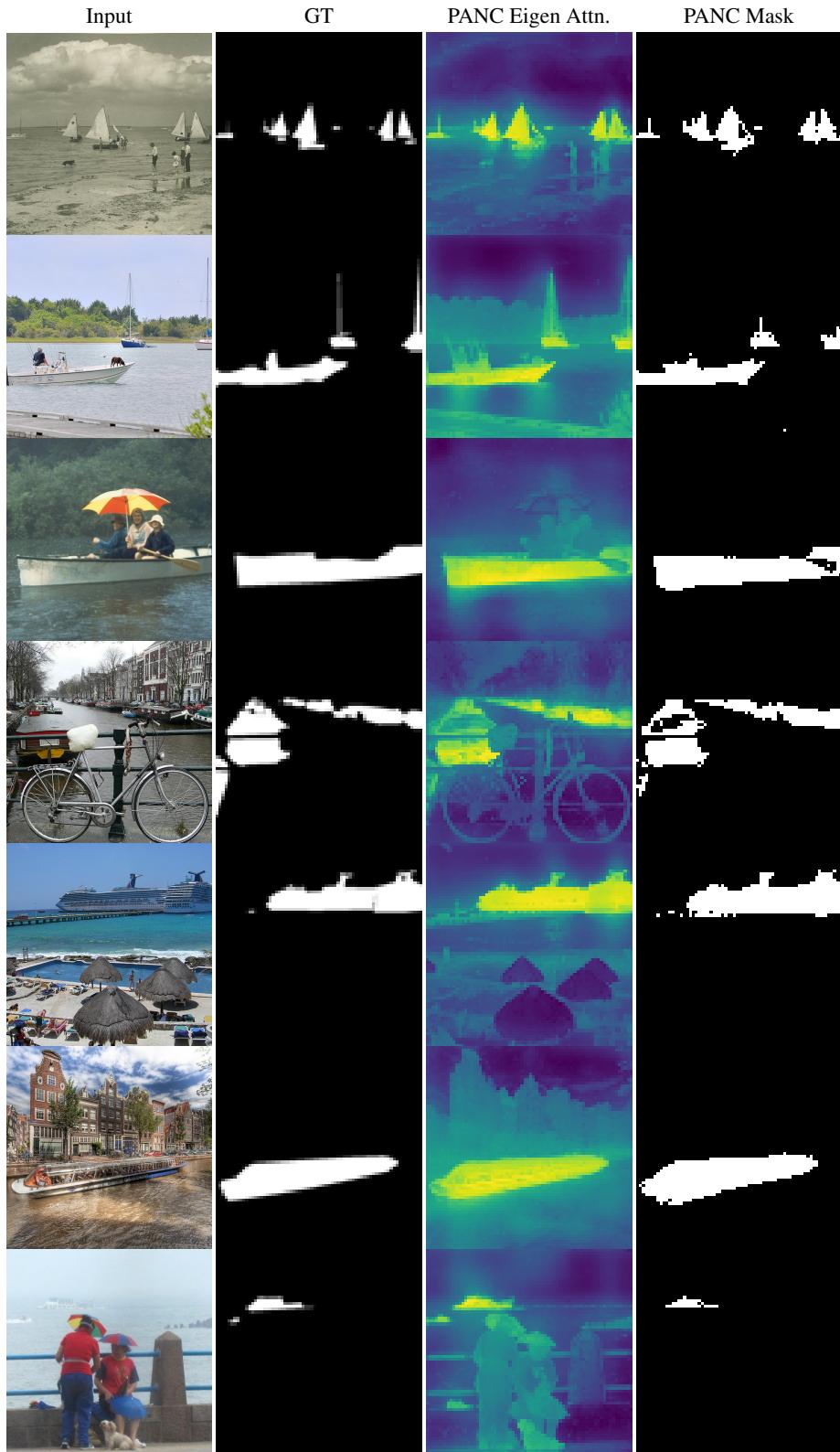


Figure 10. Additional qualitative comparison on rigid MSCOCO **Boat** class

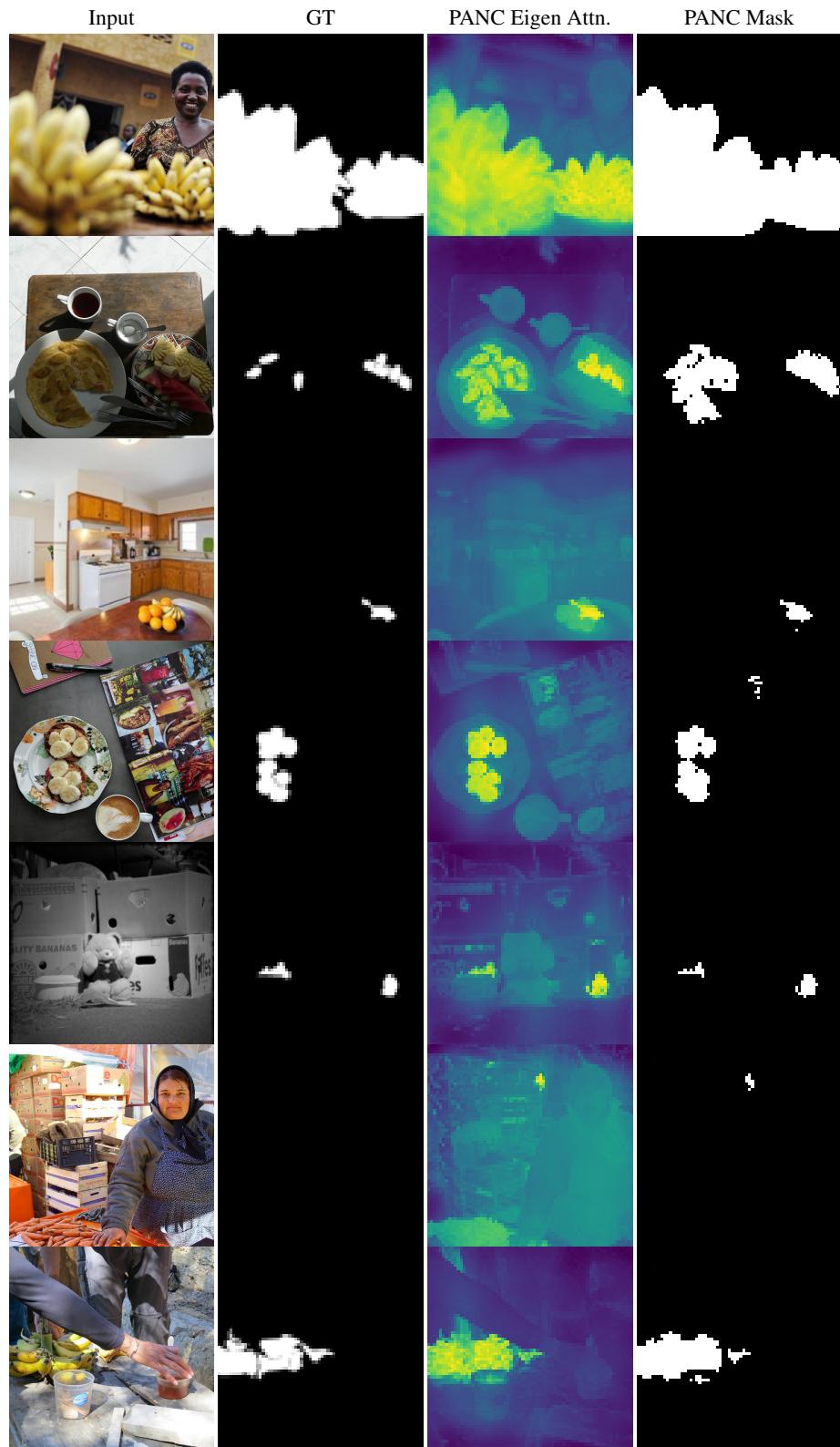


Figure 11. Additional qualitative comparison on non-rigid MSCOCO **Banana** class

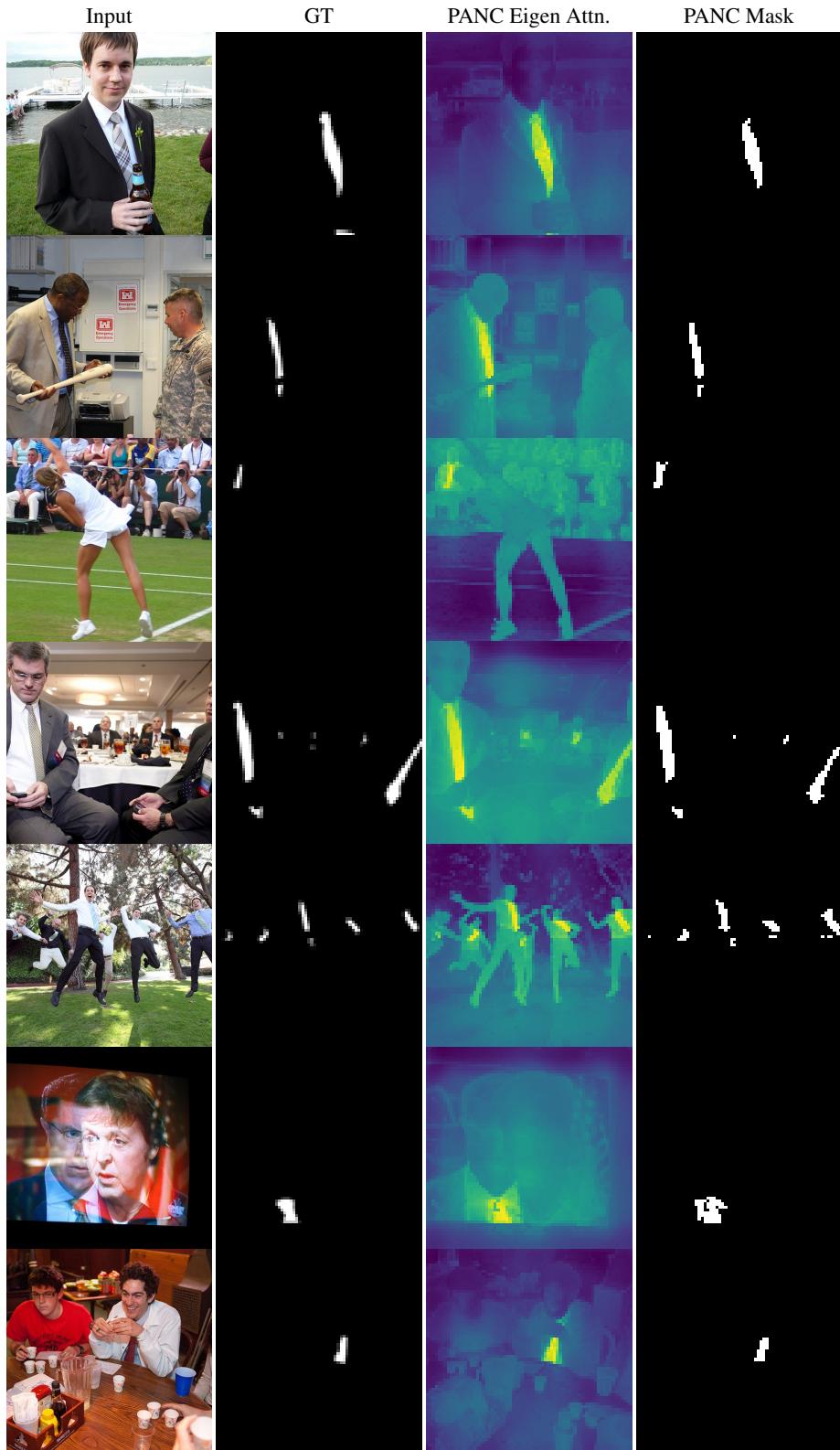


Figure 12. Additional qualitative comparison on non-rigid MSCOCO Tie class

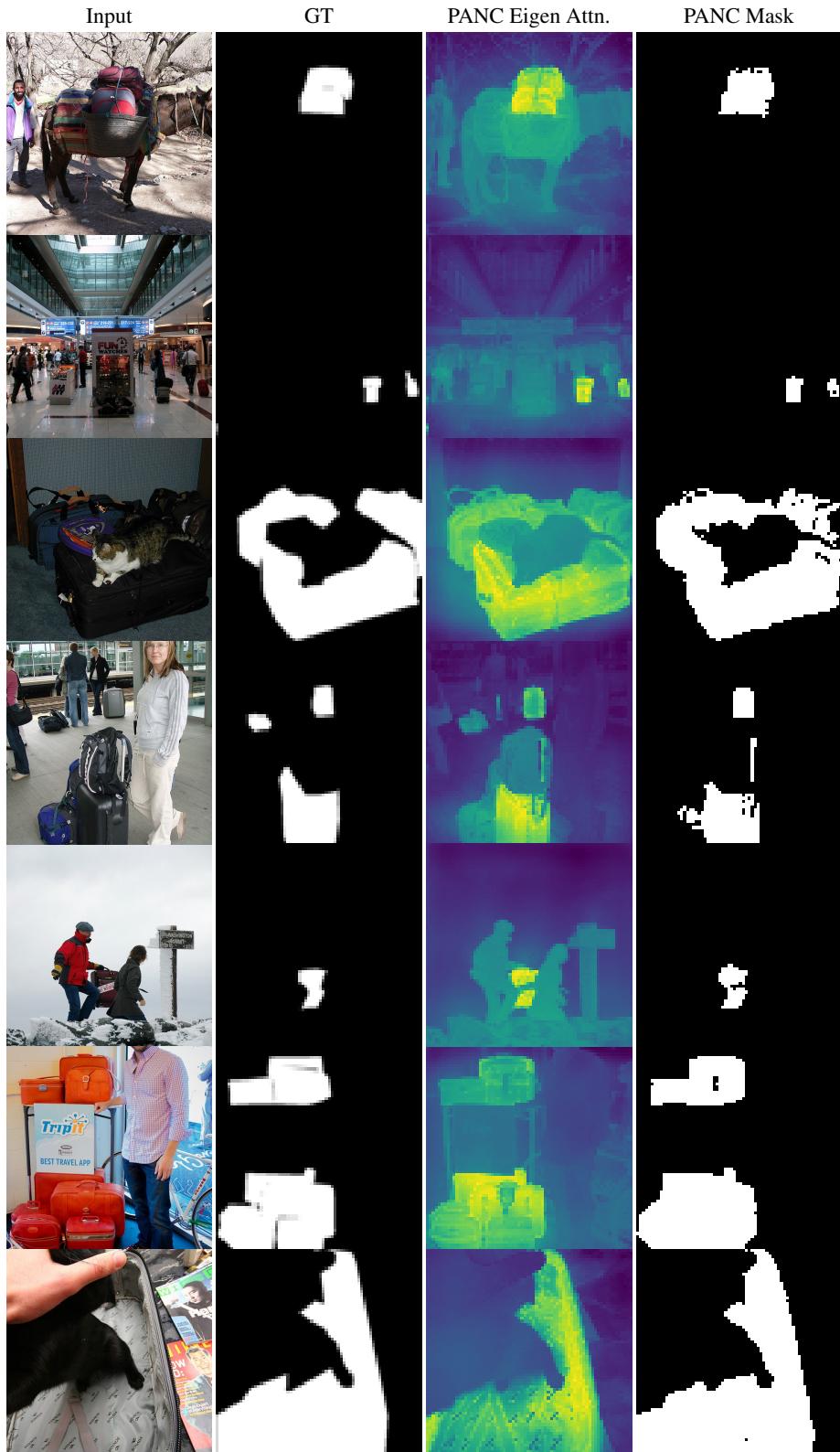


Figure 13. Additional qualitative comparison on non-rigid MSCOCO **Suitcase** class

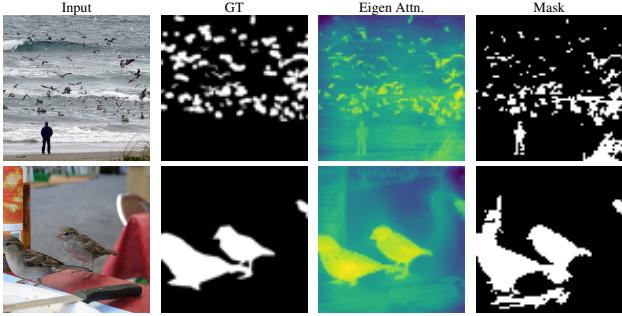


Figure 14. Transfer learning from the CUB-200-2011 dataset enables bird segmentation in the MS COCO dataset.

Tests effectively showcase the relevance of filter banks in our results, their ability to generalize across datasets, and the effectiveness of transfer learning through priors.

D.3. Additional examples

To illustrate our results, we provide further evidence on the advantages highlighted in Section 4.3. As mentioned above, the PANC, powered by a high-quality backbone such as DINOv3, outperforms previous unsupervised and weakly supervised methods. On homogeneous and challenging-domain datasets, injecting a small set of handmade annotations drastically improves performance on low-semantic-content images.

Examples of the HAM10000, CrackForest (CFD), and CUB-200-2011 datasets can be found in Fig. 16, Figure 17, and Figure 18, respectively, evidenced the expected train. We observe how injection priors greatly improve results for the unsupervised mask, yielding consistent results on skin lesions, surface cracks, and birds.

D.4. Known weaknesses and limitations

Throughout the paper, we have outlined the main weaknesses and limitations of our method in light of the segmentation task at hand. As mentioned in Section 4, a key challenge of our weakly supervised approach lies in selecting appropriate priors for heterogeneous datasets with high data variability. Since these datasets contain multiple, diverse objects, it may not be feasible to obtain a sufficient number of representative high-quality samples for every possible object that may appear.

If our prior bank is constructed from 10 images from the most representative classes $\{1, 2, \dots, 10\}$ of a heterogeneous dataset, these representations will serve as anchors to guide attention in subsequent segmentations. Suppose an outlier in the dataset appears, whose target class is not well represented in the prior bank previously generated. The algorithm will focus on the prior bank’s class rather than the target class, leading to incorrect segmentation of the target object. A naïve example is shown in Figure 15.

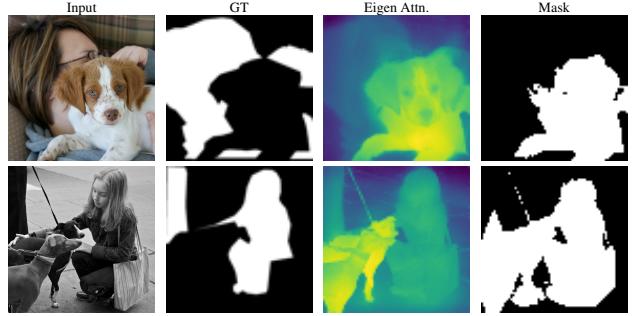


Figure 15. Examples of the missed prior selection leading to inaccurate segmentation of the target class.

In conclusion, the effectiveness of our method depends strongly on the quality of the priors used in segmentation, which is why we opted for highly reliable manual annotations to build them. Moving forward, a practical assessment of the proposed approach for assisted annotation should include quantitative evaluations of the segmentation time and the unit cost per correctly segmented item. Such evaluations will be essential to fully understand the trade-offs and practical benefits of this method in real-world annotation workflows, ultimately guiding future improvements and optimizations. This will help ensure that the method is both efficient and cost-effective for large-scale applications.

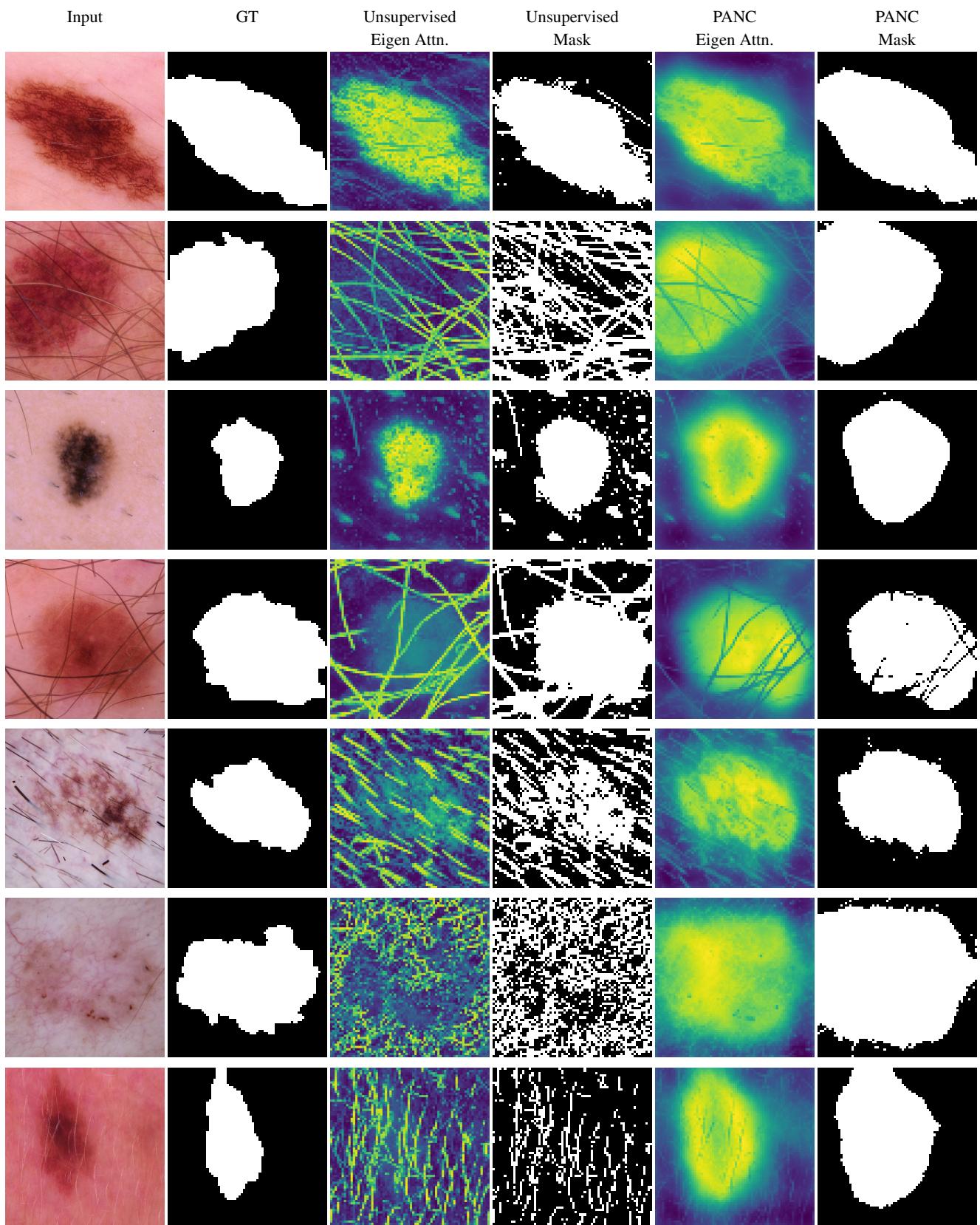


Figure 16. Additional qualitative comparison on HAM10000 dataset.

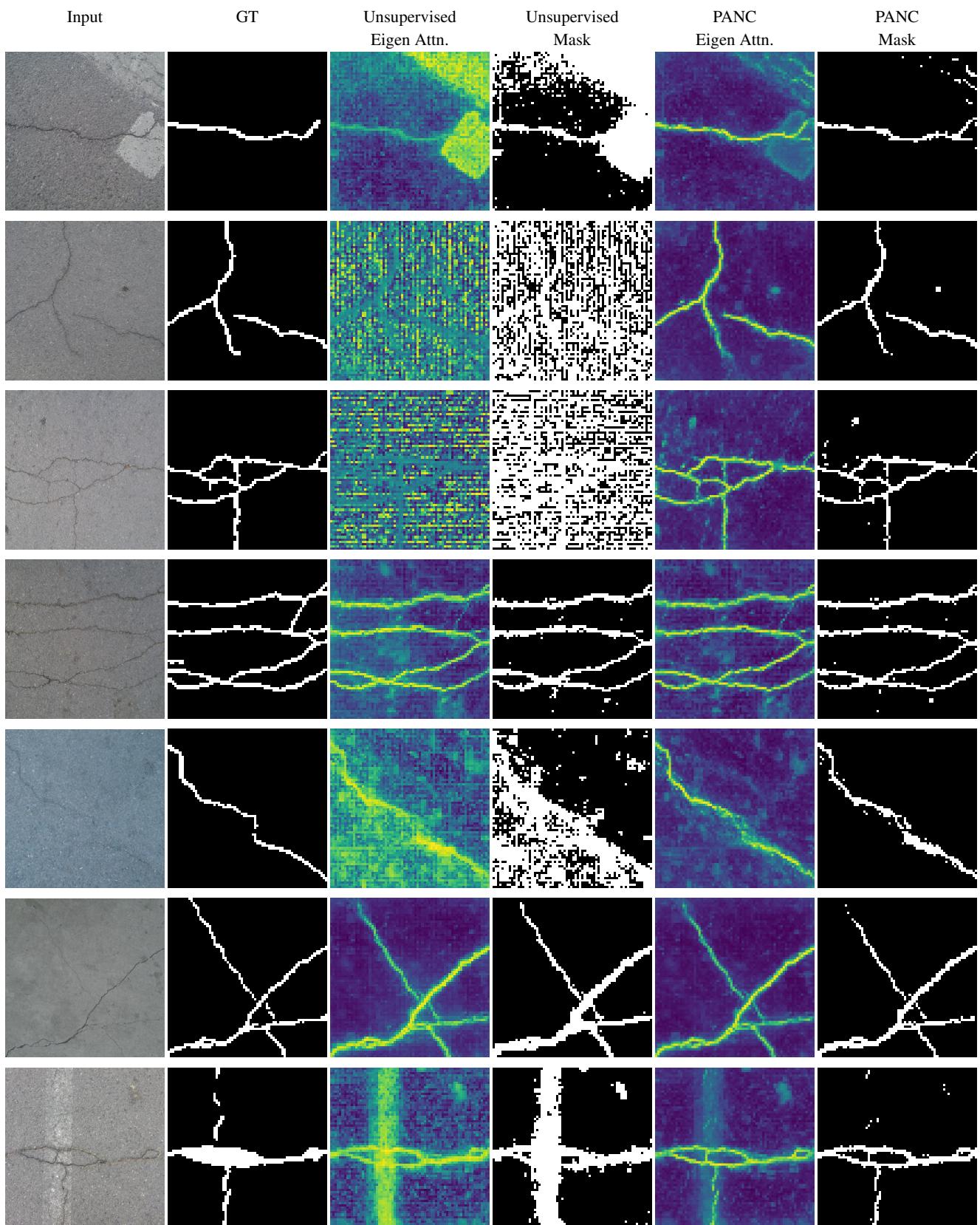


Figure 17. Additional qualitative comparison on CrackForest dataset.

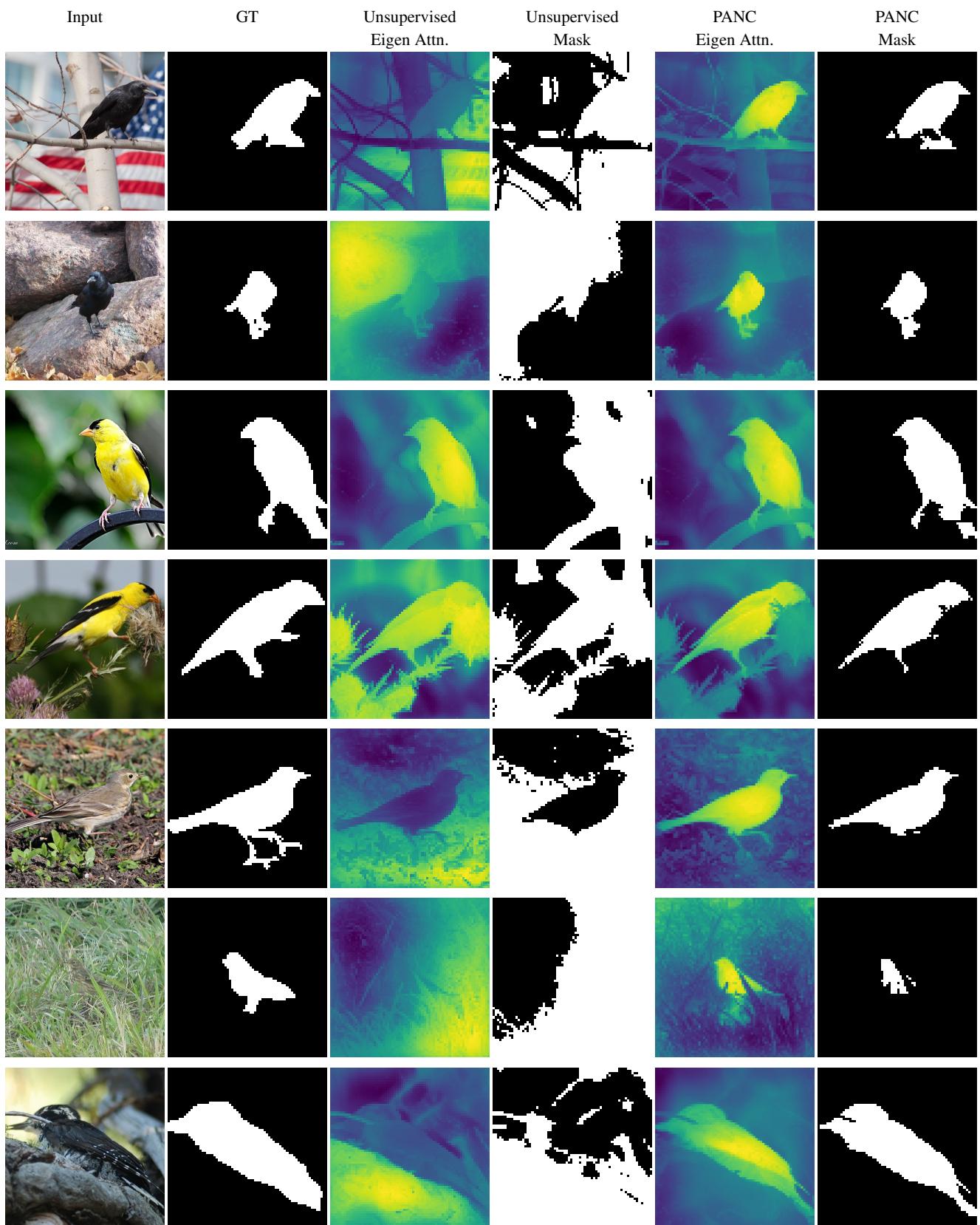


Figure 18. Additional qualitative comparison on CUB-200-2011 dataset.