RETHINKING NORMAL MEMORY READOUT FOR TRAINING-FREE ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Normal patch memories support anomaly detection by comparing test features with stored normal references. Nearest-neighbor scoring measures proximity to one reference, whereas nearby references can jointly provide a local estimate of normal appearance. We propose ProCon, a training-free anomaly detector that retrieves normal neighbors, reconstructs a query by their distance-weighted average, and uses the reconstruction residual as its anomaly score. Independent memories at different DINOv2 layers preserve the normal neighborhood of each feature space; a median across memory banks followed by a mean across layers combines their residual maps. With the encoder and memory fixed, replacing nearest-neighbor distance with local reconstruction improves pixel AP by 1.4 and 2.5 percentage points on MVTec-AD and VisA. Under the Dinomaly evaluation protocol, ProCon achieves pixel AP of 75.3% and 57.0%, respectively, outperforming the compared Dinomaly and INP-Former configurations on this metric. At matched memory storage, independent layer reconstruction improves pixel AP and AUPRO over feature concatenation in all 27 categories. These results show that a normal patch memory can serve as an effective local reconstruction model without training a decoder.

1 INTRODUCTION

Industrial anomaly detection uses normal product images to identify defects and locate the affected regions. A patch memory offers a direct way to use pretrained visual representations for this task: extract normal features, retain representative references, and compare each test feature with the stored set. The quality of those references matters, but so does the operation that converts their relationship with a query into an anomaly score. A fixed memory can support different scores even when its encoder, feature transformation, and stored content are unchanged.

Nearest-neighbor scoring assigns a query the distance to its closest normal reference. This operation, used by PatchCore-style detectors (Roth et al., 2022), asks whether one stored feature lies close to the query. It leaves the arrangement of the other nearby references outside the score. Two queries can have the same nearest-reference distance while differing in how accurately their surrounding normal features reconstruct them. Proximity to one reference and agreement with a local normal estimate are therefore different quantities.

We investigate the question: **can a normal patch memory serve as a local reconstruction model?** We propose **ProCon**, which retrieves nearby normal references and forms a distance-weighted feature estimate. The distance between the query and this estimate is the reconstruction residual. Both retrieval and reconstruction use the stored normal features directly. This gives a reconstruction-based anomaly score with a frozen encoder and no learned decoder.

ProCon applies the same operation to several layers of DINOv2 (Oquab et al., 2023). Each layer constructs its own memory and retrieves neighbors in its own feature space. The residual maps are combined after reconstruction, allowing the same image location to draw on different normal references at different layers. Within each layer, repeated coreset banks provide several reference sets whose residuals are combined by a median. A mean across layers then yields the patch anomaly map. The default uses five banks per layer, while a one-bank configuration offers a lower-cost operating point.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

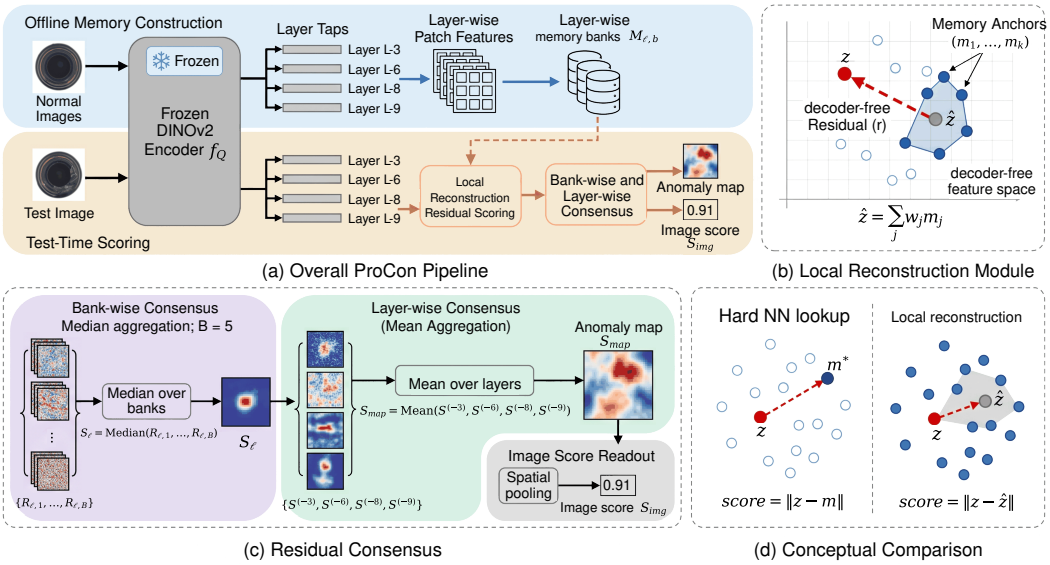

Figure 1: **ProCon overview.** Normal training features form separate memories at four frozen encoder layers. Each query is reconstructed from its nearest normal references, and the residual supplies its anomaly score. The default combines five bank residuals by a median within each layer, then averages the layer maps. The resulting map is resized and smoothed for evaluation. Panels (b) and (d) illustrate the distinction between distance to one reference and distance to a weighted reconstruction.

On MVTec-AD and VisA, ProCon reaches pixel AP of 75.3% and 57.0% under the Dinomaly evaluation protocol. It exceeds the pixel AP of the compared Dinomaly and INP-Former configurations, which we run with DINOv2 ViT-B/14 in the single-category setting. The same-memory comparison supplies direct evidence for the reconstruction score: replacing nearest-neighbor distance with the residual raises pixel AP by 1.4 and 2.5 points, and AUPRO by 1.0 and 4.1 points. A separate storage-matched comparison shows that independent layer reconstruction improves both pixel AP and AUPRO over feature concatenation in all 27 categories.

Our contributions are threefold:

- **Local reconstruction from normal memory.** ProCon estimates a query's normal feature using distance-weighted neighbors and scores the reconstruction residual without training a decoder.
- **Independent reconstruction across layers.** Layer-specific memories preserve separate normal neighborhoods and combine their residuals through bank medians and a layer mean.
- **Evidence for the readout and its operating points.** Same-memory Hard–Soft comparisons, matched memory budgets, and measured GPU inference costs explain the localization gains and the performance–cost trade-off.

2 Related Work

Normal feature memories. Memory-based anomaly detectors model normal appearance using pretrained image features. SPADE uses deep feature correspondences, PaDiM fits spatial feature distributions, and PatchCore retains a coreset of normal patch features (Cohen & Hoshen, 2020; Defard et al., 2021; Roth et al., 2022). These choices determine which normal information is available to a test query. ProCon changes how the stored information is read: several nearby references contribute to a normal feature estimate. The Hard–Soft comparison uses the same memory to measure the effect of this scoring operation.

Feature reconstruction. Reconstruction-based detection uses the discrepancy between an observed feature and a normal estimate. RD4AD reconstructs teacher features through reverse distillation (Deng & Li, 2022). Dinomaly uses Transformer-based feature reconstruction (Guo et al., 2025), and INP-Former extracts intrinsic normal prototypes from a test image (Luo et al., 2025). FastRecon demonstrates training-free feature reconstruction in the few-shot setting through distribution-regularized regression (Fang et al., 2023). ProCon computes its estimate from the query's retrieved neighbors using normalized distance weights. The residual connects normal-memory retrieval with local feature reconstruction through a direct, non-parametric operation.

Combining layers. Pretrained encoders such as DINOv2 provide patch descriptions at multiple layers. The stage at which these descriptions are combined determines the retrieval space. Concatenating features selects references and searches for neighbors in a single joint space. ProCon performs selection, retrieval, and reconstruction separately at each layer, then combines the spatial residual maps. Our matched-storage comparison evaluates these two ways of organizing the same collection of encoder layers.

3 PROCON

Figure 1 summarizes memory construction, local reconstruction, and residual aggregation.

3.1 FROZEN FEATURES AND NORMAL MEMORY

Let $\mathcal{D}_{\text{normal}}$ contain the normal training images of one category. A frozen DINOv2 ViT-B/14 encoder produces a patch feature $f_{\ell,p}(x) \in \mathbb{R}^{768}$ at layer ℓ and spatial location p of image x. The encoder input resolution is 392×392, and the CLS token is discarded. We normalize each patch feature and apply a fixed linear projection:

$$z_{\ell,p}(x) = G_\ell \frac{f_{\ell,p}(x)}{\|f_{\ell,p}(x)\|_2}, \qquad G_\ell \in \mathbb{R}^{512\times 768}. \tag{1}$$

The projection has no bias. Its entries are sampled from $\mathcal{N}(0,1)/\sqrt{768}$ with projection seed 42. Memory construction and inference use the same matrices, and neither the encoder nor the projection is trained.

An approximate greedy coreset selects normal reference features from the training set. For N candidate features and ratio ρ, the memory retains $\lfloor \rho N \rfloor$ references. Selection distances are computed in a separate 192-dimensional space; the selected 512-dimensional features are stored for retrieval and reconstruction. We call one such reference set a *memory bank*. Banks use different coreset starting points.

The default uses layers $\mathcal{L} = \{-3, -6, -8, -9\}$, with indices counted backward from the last Transformer block, and $B = 5$ banks per layer. Each layer–bank pair retains a 1% coreset. All selected layers are obtained in one encoder forward pass and share spatial coordinates. We write the memory of layer ℓ and bank b as $M_{\ell,b}$.

3.2 LOCAL RECONSTRUCTION FROM NORMAL NEIGHBORS

For a query feature z and memory M, the Hard readout returns the nearest-reference distance,

$$r_{\text{Hard}}(z; M) = \min_{m \in M} \|z - m\|_2. \tag{2}$$

The Soft readout reconstructs the query from several normal neighbors. Let $\mathcal{N}_k(z; M) = \{m_1, \ldots, m_k\}$ be its k nearest references. Their weights are

$$w_j(z; M) = \frac{\exp(-\|z - m_j\|_2^2/\tau)}{\sum_{t=1}^{k} \exp(-\|z - m_t\|_2^2/\tau)}. \tag{3}$$

The weighted estimate and anomaly score are

$$\hat{z} = \sum_{j=1}^{k} w_j m_j, \qquad r_{\text{Soft}}(z; M) = \|z - \hat{z}\|_2. \tag{4}$$

We use $k = 5$. Closer references receive larger weights, and $\hat{z}$ lies in the convex hull of the retrieved neighbors. The score measures how accurately their weighted estimate reconstructs the query.

The temperature τ is a single scalar for each image, layer, and bank. We sample at most 512 queries from the current image, retrieve the k neighbors of each sampled query, and pool their squared query-to-neighbor distances. Their median defines τ. Thus, each query retrieves its own neighbors, while all queries in the corresponding image–layer–bank map share the temperature. Sampling uses a fixed evaluation seed. This calculation depends on image features and normal references.

The shared temperature calibrates how sharply the reconstruction favors closer neighbors within that image and memory. For a fixed neighbor set, a small positive temperature concentrates weight on the nearest reference; a large temperature moves the estimate toward the neighbors' arithmetic mean. Using a median distance ties this scale to the current query–memory distances. If queries and references undergo the same positive rescaling, their squared distances and the median temperature scale together, preserving the reconstruction weights. Appendix G gives the corresponding derivation.

3.3 What the Residual Measures

The reconstruction residual depends on the arrangement of references as well as their distance from the query. Since $\sum_j w_j = 1$,

$$\sum_{j=1}^{k} w_j \|z - m_j\|_2^2 = \|z - \hat{z}\|_2^2 + \sum_{j=1}^{k} w_j \|m_j - \hat{z}\|_2^2. \tag{5}$$

The weighted squared distance on the left separates into the squared reconstruction residual and the weighted dispersion of references around their mean. Consequently, comparable neighbor distances can produce different residuals when the references occupy different directions around a query. ProCon uses this local geometry in its normal estimate.

A simple geometric example makes this distinction explicit. Place a query at the center of five normal references arranged on a regular pentagon. Every query-to-reference distance is the same, so all five weights are equal. Their weighted estimate coincides with the query, giving a zero reconstruction residual even though the nearest-reference distance is positive. If normal references instead occupy one side of the query, their weighted center lies on that side and produces a nonzero residual. The operation thus asks whether the selected normal features jointly reconstruct the query.

For $k = 1$, the estimate is the nearest reference and Soft reduces exactly to Hard. With several neighbors, the score measures agreement with their weighted center. Appendix G derives Equation 5, gives an equivalent pairwise-distance expression, and describes the behavior under changes in temperature and feature scale.

3.4 Independent Layers and Bank Aggregation

Each layer searches its own memory and computes a residual map. A location can therefore use different normal references at different layers. This is the operational difference from concatenation: joint-space retrieval selects one neighborhood according to a combined distance, whereas independent retrieval allows each feature space to select its own normal explanation. Combining residuals after this step retains those separate retrieval decisions. Because the selected Transformer layers share spatial coordinates, the scalar residuals can then be combined directly at corresponding locations. Let $r_{\ell,b,p}$ be the residual at layer ℓ, bank b, and position p. We first compute the median within a layer:

$$R_\ell(p) = \mathrm{median}_{b=0,\ldots,B-1}\, r_{\ell,b,p}. \tag{6}$$

This operation reduces the influence of an extreme residual from an individual coreset. We then average the layer maps:

$$S(p) = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} R_\ell(p). \tag{7}$$

We combine residuals without per-image normalization, retaining the distance scale of each normal estimate. The final map is resized and smoothed following the Dinomaly evaluation protocol. The lightweight setting uses $B = 1$: bank 0 supplies each layer map, and the same layer mean is applied.

Inputs: normal training images, frozen encoder, fixed projections, layer set $\mathcal{L}$, bank count B, coreset ratio ρ, and neighbor count k.

Memory construction

1. Extract features from the selected layers for each normal training image.
2. L2-normalize each feature and project it to 512 dimensions.
3. At each layer and bank, select a ρ-fraction coreset and store the selected features as $M_{\ell,b}$.

Inference for an image x

1. Extract all selected layer features in one encoder forward pass and apply the fixed transformations.
2. For each layer and bank, retrieve the k nearest normal references and estimate $\tau_{\ell,b}(x)$.
3. Reconstruct each query with its distance-weighted references and compute $r_{\ell,b,p} = \|z_{\ell,p} - \hat{z}_{\ell,b,p}\|_2$.
4. Take the bank median within each layer and then the mean across layers.
5. Resize and smooth the resulting map under the evaluation protocol; obtain pixel and image scores from the processed map.

Algorithm 1: ProCon: normal-memory construction and inference

3.5 Computation and Implementation

Algorithm 1 separates memory construction from test-time scoring. The memory is built once from the normal training set and remains fixed during inference. A test image changes the retrieved neighbors and their weights, but does not update the stored references. The normal estimate is therefore local to each query: two nearby image locations can select different reference sets and obtain different reconstructions from the same bank.

The reconstruction requires only the k retrieved features and their distances. Once these are available, Equation 3 normalizes the distance weights, Equation 4 forms a 512-dimensional weighted feature, and its Euclidean residual gives one scalar per location. The reconstructed feature is transient; it is not added to the memory. Repeated banks apply this same computation to different coreset reference sets, and all layers use the same reconstruction rule.

For P queries, feature dimension D, and n references per layer and bank, exhaustive retrieval costs $O(|\mathcal{L}|BPnD)$ and stores $O(|\mathcal{L}|BnD)$ feature values. Reconstruction adds $O(kD)$ operations per query and memory. All banks share one encoder forward pass. Bank count thus controls memory storage and retrieval cost while preserving the reconstruction rule.

4 Experiments

4.1 Experimental Setup

Datasets and comparisons. We evaluate on the 15 categories of MVTec-AD (Bergmann et al., 2019) and the 12 categories of VisA (Zou et al., 2022). Each category uses its normal training images to construct a separate memory. Dinomaly and INP-Former are run in their single-category DINOv2 ViT-B/14 configurations, and we confirm reproduction of the corresponding results reported in their papers.

ProCon settings. We use four layers, 512-dimensional stored features, $k = 5$, and a 1% coreset for each layer and bank. The default is $B = 5$; the lightweight setting is $B = 1$. The layer set and residual aggregation rule were selected using MVTec-AD results and transferred unchanged to VisA. Appendix C describes the candidate layers and aggregation rules.

Evaluation. We follow the Dinomaly evaluation protocol (Guo et al., 2025). We report image-level AUROC, AP, and maximum F1, together with pixel-level AUROC, AP, maximum F1, and AUPRO. I and P denote image and pixel metrics, and all F1 values are maxima over thresholds. Dataset scores are unweighted category averages. Performance, Hard–Soft, and bank-count comparisons report mean and sample standard deviation over memory-construction seeds 0, 1, and 2. Memory-budget and neighbor-count/weighting comparisons use seed 0. Percentages and standard

	MVTec-AD			VisA		
Method	I-AUROC	P-AP	AUPRO	I-AUROC	P-AP	AUPRO
Dinomaly	**99.7**	68.9	95.0	**98.9**	50.7	**95.1**
INP-Former	**99.7**	70.2	**95.4**	98.5	49.2	93.8
ProCon	**99.7**	**75.3**	95.2	98.5	**57.0**	94.0

Table 1: **Comparison under Dinomaly-based evaluation.** Methods use DINOv2 ViT-B/14 and single-category modeling. ProCon uses four independent layers, $B = 5$, and a 1% coreset per layer and bank. Values are percentages; bold indicates the best value at displayed precision.

Dataset	B	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	AUPRO
MVTec-AD	1	99.7±0.0	99.9±0.0	99.1±0.0	98.2±0.0	75.0±0.0	71.9±0.0	95.1±0.0
	5	99.7±0.0	99.9±0.0	99.2±0.1	98.3±0.0	75.3±0.0	72.0±0.0	95.2±0.1
VisA	1	98.5±0.0	98.5±0.0	96.1±0.0	98.8±0.0	56.6±0.2	58.2±0.1	93.8±0.1
	5	98.5±0.0	98.6±0.0	96.2±0.1	98.9±0.0	57.0±0.0	58.5±0.1	94.0±0.1

Table 2: **Complete ProCon metrics.** All rows use 4L Soft (Independent), with a 1% coreset per layer and bank. Values are mean ± sample standard deviation across three memory seeds. F1 means max-F1; ±0.0 results from rounding.

deviations are displayed to one decimal place; differences are calculated before this display rounding. Appendix B specifies map processing and metric aggregation.

Controlled configurations. Hard and Soft share the encoder, projection, and normal memory. The one-bank configuration uses bank 0 of the five-bank group. In configuration names, 1L and 4L denote the number of selected layers. *4L Soft (Independent)* is ProCon's layer-wise reconstruction followed by residual aggregation. *4L Soft (Concat)* joins the four layer features before memory construction and reconstruction. B is the number of banks per layer for Independent and per joint feature space for Concat.

4.2 Comparison with Existing Methods

ProCon has the highest pixel AP among the compared methods on both datasets (Table 1). On MVTec-AD, it reaches 75.3%, exceeding Dinomaly by 6.4 points and INP-Former by 5.1 points. On VisA, it reaches 57.0%, with respective margins of 6.3 and 7.8 points. Local reconstruction from stored normal references therefore provides strong pixel precision–recall performance on both benchmarks.

Image AUROC is 99.7% on MVTec-AD and 98.5% on VisA, while AUPRO is 95.2% and 94.0%. Table 2 reports all seven metrics for both bank settings. The five-bank configuration combines the pixel-AP improvement with high image-level detection performance. The metric-specific comparison is useful here: pixel AP evaluates the ranking of all defect pixels, while AUPRO averages overlap across connected defect regions.

4.3 Local Reconstruction with the Same Memory

Table 3 starts from a single-layer Hard readout and introduces four layers, Soft reconstruction, and five banks. All rows use the same evaluation protocol and a 1% coreset per layer and bank.

With four layers and one bank fixed, changing Hard to Soft increases MVTec-AD pixel AP from 73.6% to 75.0% and AUPRO from 94.1% to 95.1%. On VisA, pixel AP rises from 54.1% to 56.6% and AUPRO from 89.7% to 93.8%. Calculated before rounding, the improvements are 1.4/2.5 points in pixel AP and 1.0/4.1 points in AUPRO. The normal references are unchanged, so the improved localization comes from how those references produce the query score.

The improvement extends to the other pixel metrics. Soft raises pixel AUROC by 0.9 and 1.3 points and pixel max-F1 by 0.8 and 1.5 points on MVTec-AD and VisA. Thus, the same-memory change

	MVTec-AD		VisA	
Configuration	P-AP	AUPRO	P-AP	AUPRO
1L Hard, $B = 1$	69.2±0.2	89.9±0.1	45.2±0.3	82.3±0.2
4L Hard, $B = 1$	73.6±0.0	94.1±0.1	54.1±0.2	89.7±0.3
4L Soft, $B = 1$	75.0±0.0	95.1±0.0	56.6±0.2	93.8±0.1
4L Soft, $B = 5$	75.3±0.0	95.2±0.1	57.0±0.0	94.0±0.1

Table 3: **Component comparison.** Four-layer rows use independent memories. The $B = 1$ Hard–Soft pair shares identical references. Moving from 1L to 4L increases both the layer count and the total memory; Section 4.4 matches the memory budget. Values are three-seed mean ± sample standard deviation.

		MVTec-AD		VisA	
Configuration	Coreset	P-AP	AUPRO	P-AP	AUPRO
1L Soft	4%	74.0	94.0	50.4	90.8
4L Soft (Independent)	1%/layer	**75.0**	**95.1**	**56.4**	**93.9**
4L Soft (Concat)	1% joint	73.2	94.4	54.3	92.4

Table 4: **Matched memory budgets.** Seed 0, $B = 1$, and Dinomaly-based evaluation. The one-layer comparison matches total stored references; Concat matches memory-feature storage. Values are category means from one seed. Complete metrics and category results appear in Appendices E and F.

improves all four pixel metrics on both datasets. Appendix F provides the complete image and pixel results for the Hard and Soft configurations.

Four layers also improve Hard scoring: pixel AP rises from 69.2% to 73.6% on MVTec-AD and from 45.2% to 54.1% on VisA. AUPRO rises from 89.9% to 94.1% and from 82.3% to 89.7%. This configuration draws on normal references from several feature spaces; Soft reconstruction provides an additional gain on top of it.

4.4 Reconstruction at Matched Memory Budgets

We compare the independent four-layer memory against two alternatives using seed 0 and $B = 1$. The first places a 4% coreset in one layer, matching the total stored-reference count of four 1% coresets up to integer selection. The second concatenates the four layer features before selecting a 1% coreset. Concat and Independent store the same number of memory-feature bytes in every category. Table 4 shares the Independent row between the two comparisons.

At matched reference count, Independent improves pixel AP over 1L Soft by 1.0 points on MVTec-AD and 6.1 points on VisA, with AUPRO gains of 1.1 and 3.1 points. At matched storage, it improves pixel AP over Concat by 1.8 and 2.2 points and AUPRO by 0.7 and 1.5 points. Independent has higher category means for all seven metrics in both comparisons.

The category results show a particularly consistent difference from Concat: Independent has higher pixel AP and AUPRO in all 15 MVTec-AD categories and all 12 VisA categories. The median category-level pixel-AP gains are 1.8 and 2.0 points. Against 1L Soft, pixel AP improves in 9/15 and 9/12 categories, respectively. These results support layer-specific retrieval followed by residual aggregation under the matched memory budgets.

4.5 Bank Count, Performance, and Variation

The $B = 1$ and $B = 5$ configurations share bank 0, the coreset ratio, and the reconstruction rule. Five banks retrieve additional normal reference sets and aggregate their residuals by the median within each layer.

Increasing the bank count raises pixel AP from 75.0% to 75.3% on MVTec-AD and from 56.6% to 57.0% on VisA. The gains before rounding are 0.3 and 0.4 points; AUPRO increases by 0.1 and

Dataset	Banks (B)	Latency (ms/image)	Peak VRAM (MiB)
MVTec-AD	1	20.9 ± 3.9	397.1
	5	39.8 ± 9.3	445.1
VisA	1	22.2 ± 4.2	437.4
	5	45.4 ± 12.0	548.2

Table 5: **GPU inference cost.** Both settings use 4L Soft (Independent) with a 1% coreset per layer and bank. The $\pm$ values summarize timing variation, separate from memory-seed variation in the performance tables.

0.2 points. All seven metric means increase on both datasets (Table 2). On VisA, pixel AP also has lower variation across memory seeds, indicating that bank aggregation reduces coreset sensitivity for this metric. We use $B = 5$ for the default performance setting and retain $B = 1$ as a lightweight configuration of the same method.

4.6 Inference Efficiency

We measure the GPU inference path on an NVIDIA GeForce RTX 4090. The timed scope includes host-to-device image transfer, DINOv2 forward, projection, memory scoring, and map resizing and smoothing. CPU image decoding and preprocessing, and offline memory construction, are outside this scope. Latency is aggregated by category, and peak VRAM is the maximum category value.

The five-bank configuration takes 39.8 ms per image on MVTec-AD and 45.4 ms on VisA. One bank reduces this to 20.9 and 22.2 ms, saving 18.9 and 23.2 ms. Peak VRAM decreases by 48.0 and 110.8 MiB. Bank count changes memory retrieval and residual aggregation while the encoder pass remains shared. The two settings therefore offer a direct performance–cost trade-off: five banks provide the higher metric means, and one bank runs in approximately half the time.

4.7 Reading the Component Results

The comparisons assign a distinct role to each operation. With four layers and the stored references held fixed, the Hard–Soft pair measures the effect of replacing nearest-reference distance by reconstruction residual. Both datasets improve on all four pixel metrics. The matched-budget comparisons then evaluate where reconstruction takes place: four independent feature spaces give higher dataset means than either a larger single-layer memory or a concatenated feature memory. These comparisons retain the Soft rule and change the organization of its normal references.

The category results also distinguish these two alternatives. Independent reconstruction improves pixel AP over the storage-matched concatenated memory in every category. Against the larger single-layer memory, the improvement varies across categories, with 9 of 15 MVTec-AD categories and 9 of 12 VisA categories gaining pixel AP. Separate layer memories therefore provide a consistent advantage over the tested joint retrieval space, while the benefit over one layer depends on the product category. The complete category tables make these differences visible alongside the dataset averages.

Reconstruction from several neighbors also improves localization (Appendix D.4). At seed 0 and $B = 1$, increasing k from 1 to 5 raises pixel AP by 1.4 and 2.6 points on MVTec-AD and VisA, with AUPRO gains of 1.0 and 3.9 points. On MVTec-AD, $k = 5$ and $k = 10$ give nearly equal pixel AP and AUPRO; VisA favors $k = 10$. Uniform five-neighbor reconstruction gives nearly the same pixel AP as distance-weighted reconstruction on both datasets. These results support reconstructing from several normal references, with the weighting rule making a smaller difference in this comparison.

4.8 Qualitative Results

Figure 2 shows localization on a transistor, zipper, and tile. The transistor map responds near the damaged lead. The zipper map highlights both marked defect locations, while the tile map follows the crack. These examples show residual responses to defects in object parts and repeated textures. Layer-wise maps and examples from every category are provided in Appendix H.

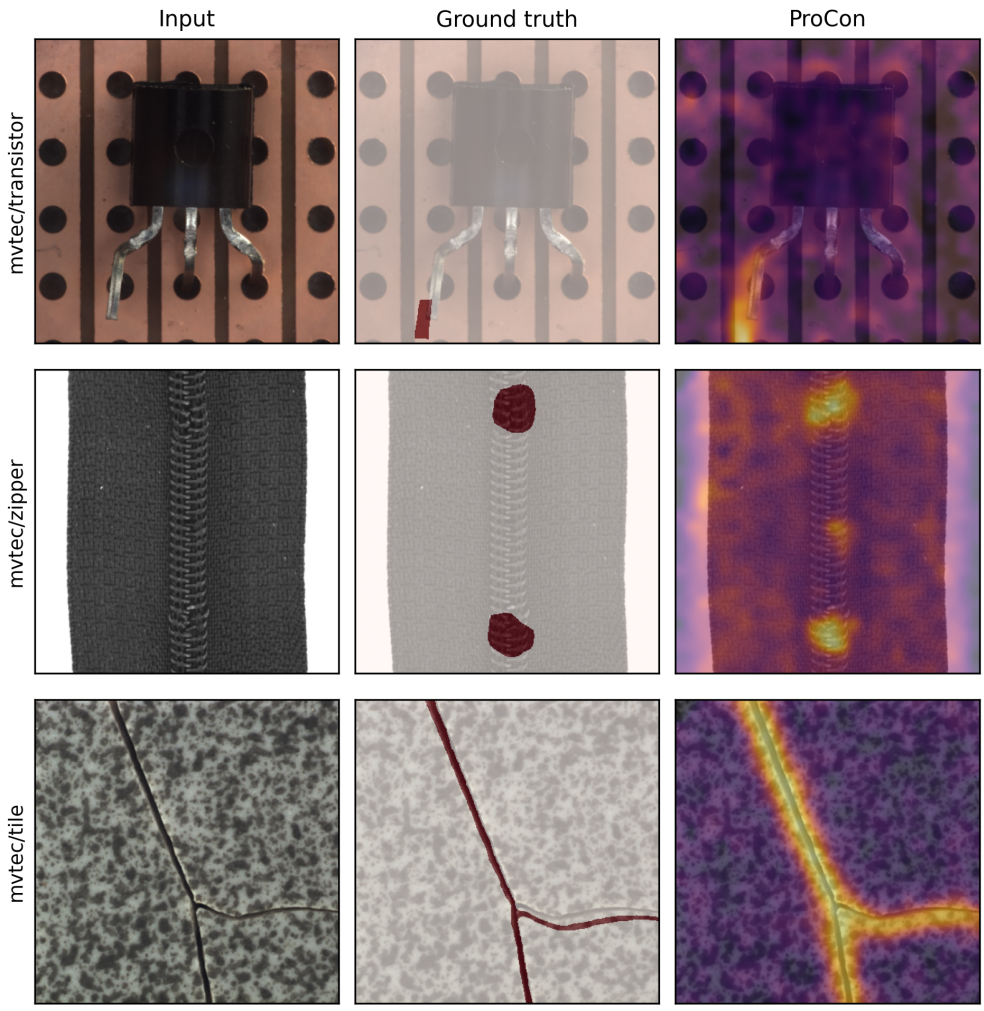

Figure 2: **Defect localization with ProCon.** Rows show transistor, zipper, and tile examples from MVTec-AD. Columns show the input, ground-truth defect region, and ProCon anomaly map.

The layer-wise visualization also illustrates the role of combining scores after reconstruction. In the tile example, individual layers respond with different spatial spread around the crack, while the combined map preserves its elongated structure. In the zipper example, the two marked defects remain visible across the layer responses. Each layer produces a complete spatial map, so their combination retains a score at every location. This connects the feature-space reconstruction to the dense localization evaluated by pixel AP and region overlap.

5 Conclusion

ProCon reconstructs a query from nearby normal memory features and uses the residual to detect anomalies. With the memory fixed, this readout improves pixel AP and AUPRO on MVTec-AD and VisA. Independent layer memories further improve both metrics over feature concatenation in every evaluated category at matched storage. The default and lightweight bank settings preserve the reconstruction rule while trading a small performance gain for additional inference cost. Normal patch memory can thus act as a training-free local reconstruction model, with the readout determining how stored references contribute to localization.

AI USE STATEMENT

Generative AI tools assisted with English translation, manuscript drafting and editing, interpretation and organization of author-supplied results, mathematical exposition, and preparation of LaTeX tables and figure annotations. Experimental measurements were supplied by the authors. The authors are responsible for the final text, claims, figures, and results.

REPRODUCIBILITY STATEMENT

Section 3 defines the feature transformation, reconstruction score, and aggregation. Appendices A–B detail the implementation and evaluation protocol. Appendices E–F provide complete comparison and category results, and Appendix G derives the residual identities.

REFERENCES

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9584–9592, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-7281-3293-8. doi: 10.1109/CVPR.2019.00982.

Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences, 2020.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges*, volume 12664 of *Lecture Notes in Computer Science*, pp. 475–489. Springer, 2021. doi: 10.1007/978-3-030-68799-1_35.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9737–9746, 2022.

Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. FastRecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17481–17490, 2023.

Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20405–20415, 2025.

Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9974–9983, 2025.

Maxime Oquab, Timothée Darcet, 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 Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14318–14328, 2022.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. SPot-the-Difference self-supervised pre-training for anomaly detection and segmentation. In *Computer Vision – ECCV 2022*, volume 13690 of *Lecture Notes in Computer Science*, pp. 392–408. Springer, 2022. doi: 10.1007/978-3-031-20056-4_23.

A Implementation Details

A.1 Feature Extraction and Fixed Projection

ProCon constructs a normal memory from the normal training images of each category. The encoder is frozen DINOv2 ViT-B/14 with a 392×392 input. Patch features are extracted at layers $\{-3, -6, -8, -9\}$, counted backward from the final Transformer block; the CLS token is omitted. The four layer features are obtained in one forward pass and share spatial coordinates.

Each 768-dimensional feature is L2-normalized and multiplied by a fixed 512×768 Gaussian matrix, as defined in Equation 1. Entries follow $\mathcal{N}(0,1)/\sqrt{768}$, with no bias and projection seed 42. Normal training references and test queries use the same transformation. Consequently, retrieval and reconstruction take place in the same 512-dimensional feature space.

Component	Setting
Encoder	Frozen DINOv2 ViT-B/14
Input resolution	392×392
Selected layers	$\{-3, -6, -8, -9\}$
Encoder feature dimension	768
Stored and retrieved feature dimension	512
Coreset selection dimension	192
Projection seed	42
Banks per layer	5 by default; 1 for the lightweight setting
Coreset ratio	1% per layer and bank
Retrieved neighbors	$k = 5$
Temperature estimation	At most 512 queries per image
Within-layer aggregation	Median across banks
Across-layer aggregation	Mean of residual maps
Repeated-run memory seeds	0, 1, 2

Table 6: **Default implementation.** Memory-budget comparisons change the coreset organization as specified in Appendix E.

A.2 Coreset Memories

For each layer, normal training features form the candidate set. Approximate greedy coreset selection retains $\lfloor \rho N \rfloor$ references from N candidates. The default ratio is $\rho = 1\%$. A separate linear transformation maps candidate features to 192 dimensions for selection distances. The initial distance estimate uses 10 randomly selected auxiliary points, determined by the bank seed. The selected references are stored as their full 512-dimensional features.

A memory bank is one selected reference set. The default constructs $M_{\ell,b}$ for $b \in \{0, 1, 2, 3, 4\}$ at each layer ℓ. Banks differ in their coreset starting points. The one-bank configuration uses bank 0 from the same group. Bank count B specifies the number of reference sets used within a run; the memory-construction seed specifies the random initialization when repeating the configuration.

A.3 Adaptive Temperature and Reconstruction

For each image, layer, and bank, temperature estimation samples at most 512 query features using a fixed evaluation seed. Each sampled query retrieves its own $k = 5$ nearest normal references. Let $\mathcal{Q}_{\ell,b}(x)$ denote the sampled positions. The shared temperature is

$$\tau_{\ell,b}(x) = \operatorname{median}\left\{ \|z_{\ell,p}(x) - m\|_2^2 : p \in \mathcal{Q}_{\ell,b}(x),\ m \in \mathcal{N}_k(z_{\ell,p}(x); M_{\ell,b}) \right\}. \tag{8}$$

Each query then uses Equation 3 to weight its retrieved neighbors. The weighted feature estimate and its Euclidean residual are computed using Equation 4. All positions in an image–layer–bank map share the same temperature. The computation uses only the current image features and the normal reference memory.

A.4 Residual Aggregation

The bank median is taken independently at each spatial position and layer. The resulting layer maps are averaged without per-image normalization. A query thus receives a score from each feature space before those scores are combined. Algorithm 1 lists memory construction and inference.

B Experimental and Evaluation Protocol

B.1 Datasets and Configurations

MVTec-AD contains 15 evaluated categories and VisA contains 12. Memories are constructed separately from each category's normal training images. Dinomaly and INP-Former are run in the single-category DINOv2 ViT-B/14 setting, with reproduction of the corresponding published results confirmed. The encoder weights, feature projection, selected layers, reconstruction rule, and evaluation remain fixed across ProCon's repeated memory-construction runs.

The default ProCon setting is 4L Soft (Independent), $B = 5$, with a 1% coreset per layer and bank. Its lightweight version uses $B = 1$. Hard and Soft refer to the score functions in Equations 2 and 4. Independent means that normal memories and reconstruction are separate at each layer; Concat means that the layer features are joined before memory construction and reconstruction.

B.2 Map Processing

We follow the Dinomaly evaluation protocol (Guo et al., 2025). The anomaly map is bilinearly resized to 256×256 with unaligned corners and smoothed by a 5×5 Gaussian kernel with standard deviation 4. The kernel weights sum to one, and convolution uses zero padding of width 2. Ground-truth masks are resized to the same resolution by nearest-neighbor interpolation. Nonzero pixels are treated as defects; multiple mask channels are combined by their union.

The complete processed map supplies the pixel scores. Image scores are obtained by spatial pooling of this map. This places image and pixel evaluation after the same resizing and smoothing operations.

B.3 Image and Pixel Metrics

Image AUROC, AP, and maximum F1 use the category's test-image labels and image scores. Pixel AUROC, AP, and maximum F1 use all pixels from that category's processed anomaly maps and binary masks. Maximum F1 is the largest F1 value over thresholds on the precision–recall curve. For precision $P(t)$ and recall $R(t)$, the evaluated F1 values use $2P(t)R(t)/(P(t) + R(t) + 10^{-7})$, excluding the terminal precision–recall endpoint without a score threshold.

AUPRO uses overlap with connected ground-truth defect regions. For threshold t and the set $\mathcal{R}$ of such regions in a category,

$$\mathrm{PRO}(t) = \frac{1}{|\mathcal{R}|} \sum_{R \in \mathcal{R}} \frac{\sum_{p \in R} \mathbf{1}[A_p > t]}{|R|}. \tag{9}$$

Each connected region receives equal weight. The false-positive rate is the fraction of all normal pixels classified as anomalous. Thresholds are spaced by one two-hundredth of the score range, starting at the minimum score and excluding the maximum endpoint. Points with FPR strictly below 0.3 are retained. Their FPR values are divided by the maximum retained FPR, and trapezoidal integration gives AUPRO.

B.4 Category and Seed Aggregation

Each metric is computed within a category and then averaged with equal category weights. For C categories and metric $v_{c,s}$ at memory seed s,

$$\bar{v}_s = \frac{1}{C} \sum_{c=1}^{C} v_{c,s}. \tag{10}$$

For performance, Hard–Soft, and bank-count comparisons, we report the mean and sample standard deviation of $\bar{v}_s$ over seeds 0, 1, and 2. Memory-budget and neighbor-count/weighting comparisons report the seed-0 category average. All performance values are percentages rounded to one decimal place. Differences and category-level comparisons use the values before this display rounding.

C Layer and Aggregation Selection

C.1 Selection on MVTec-AD

The layer set and residual aggregation rule were selected using MVTec-AD results. We first compared individual layers, then compared aggregation rules with a common candidate layer set, and finally selected a layer combination. The chosen layer set and aggregation rule were transferred unchanged to VisA.

The individual-layer candidates were $\{-1, -2, -3, -4, -6, -8, -9, -10, -12\}$. Aggregation comparisons used the common set $\{-3, -4, -6, -8, -9\}$ and considered mean, median, maximum, and per-image normalization of residual maps. Layer combinations were then compared with the aggregation rule fixed. Image detection and pixel localization informed the selection. The resulting configuration uses $\{-3, -6, -8, -9\}$ and the mean of residual maps without per-image normalization.

Comparison	Fixed components	Changed component
Individual layer	Encoder, feature transformation, scoring	Feature extraction layer
Aggregation	Candidate layer set, residual maps	Mean, median, maximum, normalization
Layer set	Reconstruction and aggregation rules	Layers included in the final map

Table 7: **Layer and aggregation candidates.** The selected configuration is used unchanged on VisA.

C.2 Feature Concatenation and Residual Aggregation

Concatenation forms a joint feature space before selecting references and retrieving neighbors. A query then uses a single normal neighborhood in that space. Independent reconstruction retrieves a separate neighborhood at each layer and combines the resulting residuals at the same image location. A patch can therefore be reconstructed by different normal references at different layers.

Per-image normalization rescales a residual map using its own score distribution. ProCon averages the unnormalized residuals, retaining the distances between queries and their local normal estimates. The direct comparison between independent reconstruction and concatenation is reported in Appendix E.

D Component Comparisons

D.1 Hard and Soft with Identical References

Hard and Soft use the same encoder, projection, four layers, bank 0, and evaluation protocol. Only the query scoring operation changes. Hard returns nearest-reference distance; Soft returns the reconstruction residual from five distance-weighted normal neighbors.

Dataset	Δ P-AUROC	Δ P-AP	Δ P-F1	Δ AUPRO
MVTec-AD	+0.9	+1.4	+0.8	+1.0
VisA	+1.3	+2.5	+1.5	+4.1

Table 8: **Soft minus Hard at 4L, $B = 1$.** Differences in percentage points are calculated from the three-seed means before rounding. F1 means max-F1.

Soft improves all four pixel metrics on both datasets. Pixel AP rises from 73.6% to 75.0% on MVTec-AD and from 54.1% to 56.6% on VisA. AUPRO rises from 94.1% to 95.1% and from 89.7% to 93.8%. These changes occur with the normal reference set fixed.

D.2 One Layer and Four Layers

Dataset	Layers	I-AUROC	P-AP	AUPRO
MVTec-AD	1	99.0	69.2	89.9
	4	99.6	73.6	94.1
VisA	1	96.8	45.2	82.3
	4	97.8	54.1	89.7

Table 9: **Layer count with Hard scoring.** Each layer uses one bank and a 1% coreset. Values are the three-seed means.

Four layers improve pixel AP over one layer by 4.4 points on MVTec-AD and 8.9 points on VisA, with AUPRO gains of 4.2 and 7.4 points. This configuration uses normal references in several feature spaces and increases total memory. Appendix E compares memory organizations at matched budgets.

D.3 Bank Aggregation

The five banks of a layer are constructed from the same normal training feature set, and their reconstruction residuals are combined by a spatial median. The one-bank setting uses bank 0 of that group. Five banks raise pixel AP from 75.0% to 75.3% on MVTec-AD and from 56.6% to 57.0% on VisA. AUPRO also increases on both datasets. VisA pixel AP has a higher mean and lower standard deviation across memory seeds with bank aggregation. The corresponding inference cost is reported in Table 5.

D.4 Neighbor Count and Reconstruction Weights

How many normal references should reconstruct a query, and how much do their weights affect localization? We compare distance-weighted reconstruction with $k \in \{1, 3, 5, 10\}$ and uniform reconstruction with $k = 5$. Weighted uses the distance-based rule in Equation 3; Uniform assigns weight $1/5$ to each of the five retrieved neighbors. All configurations use four independent layers, $B = 1$, seed 0, and Dinomaly-based evaluation. Tables 10 and 11 report the seven category-mean metrics. The weighted $k = 5$ configuration is the common setting used in the main comparisons.

Readout	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	AUPRO
Weighted $k = 1$	**99.7**	**99.9**	**99.2**	97.3	73.6	71.1	94.1
Weighted $k = 3$	**99.7**	**99.9**	**99.2**	98.0	74.7	71.7	94.8
Weighted $k = 5$ (default)	**99.7**	**99.9**	99.1	98.2	**75.0**	**71.9**	**95.1**
Weighted $k = 10$	99.5	99.8	99.0	**98.3**	**75.0**	**71.9**	95.0
Uniform $k = 5$	99.6	**99.9**	99.1	98.2	**75.0**	**71.9**	95.0

Table 10: **Neighbor count and weighting on MVTec-AD.** Four independent layers, $B = 1$, seed 0, and Dinomaly-based evaluation. Values are category means (%) over 15 categories. This is a single-seed comparison, with no seed standard deviation. F1 means max-F1. Bold marks the highest value at displayed precision.

Number of neighbors. Moving from $k = 1$ to $k = 5$ raises pixel AP from 73.6% to 75.0% on MVTec-AD and from 53.9% to 56.4% on VisA. The gains calculated before rounding are 1.4 and 2.6 points; AUPRO increases by 1.0 and 3.9 points. All four pixel metrics improve on both datasets for each tested $k > 1$ relative to $k = 1$. On MVTec-AD, $k = 5$ and $k = 10$ differ by less than 0.1 point in pixel AP and AUPRO. On VisA, $k = 10$ exceeds $k = 5$ on all seven metrics, reaching 57.0% pixel AP and 94.5% AUPRO. The best tested neighbor count therefore depends on the dataset.

Readout	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	AUPRO
Weighted $k = 1$	97.8	97.9	94.8	97.5	53.9	56.5	90.0
Weighted $k = 3$	98.2	98.4	95.7	98.5	55.9	57.8	92.8
Weighted $k = 5$ (default)	98.5	98.5	96.1	98.8	56.4	58.1	93.9
Weighted $k = 10$	**98.6**	**98.7**	**96.2**	**99.0**	**57.0**	**58.4**	**94.5**
Uniform $k = 5$	98.5	98.6	**96.2**	98.8	56.4	58.0	94.2

Table 11: **Neighbor count and weighting on VisA.** Four independent layers, $B = 1$, seed 0, and Dinomaly-based evaluation. Values are category means (%) over 12 categories. This is a single-seed comparison, with no seed standard deviation. F1 means max-F1. Bold marks the highest value at displayed precision.

Reconstruction weights. At $k = 5$, weighted and uniform reconstruction differ by less than 0.1 point in pixel AP on both datasets. Their MVTec-AD AUPRO values are 95.1% and 95.0%, respectively. On VisA, uniform reconstruction reaches 94.2% AUPRO, compared with 93.9% for weighted reconstruction. Both weighting rules improve localization over the single-reference readout; using several normal references makes the larger difference in this single-seed comparison.

E Memory Budgets and Computational Cost

E.1 Matched Memory Comparisons

The 1L Soft comparison places a 4% coreset in one feature space, while 4L Soft (Independent) places a 1% coreset in each of four spaces. Total stored-reference counts match up to integer selection. Both use $B = 1$ and 512-dimensional stored features.

The Concat comparison joins the four layer features before coreset selection and retrieval. Independent selects and retrieves references separately at each layer and then averages the residual maps. The two configurations store identical memory-feature bytes in every category. This storage refers to the reference features themselves; peak VRAM includes the other tensors present during inference.

Configuration	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	AUPRO
MVTec-AD							
1L Soft	99.5	99.8	98.7	97.5	74.0	71.2	94.0
4L Soft (Independent)	99.7	99.9	99.1	98.2	75.0	71.9	95.1
4L Soft (Concat)	99.6	99.8	99.1	97.8	73.2	70.8	94.4
VisA							
1L Soft	97.8	97.8	95.7	97.0	50.4	53.2	90.8
4L Soft (Independent)	98.5	98.5	96.1	98.8	56.4	58.1	93.9
4L Soft (Concat)	98.1	98.3	95.3	98.4	54.3	56.8	92.4

Table 12: **Complete matched-budget results.** Seed 0, $B = 1$, and Dinomaly-based evaluation. Coreset ratios are 4% for 1L Soft, 1% per layer for Independent, and 1% in the joint feature space for Concat. Values are category means in percent, with no seed standard deviation. F1 means max-F1.

Independent has higher category means for all seven metrics in both comparisons. Against 1L Soft, pixel AP improves by 1.0 and 6.1 points on MVTec-AD and VisA, and AUPRO by 1.1 and 3.1 points. Against Concat, the respective gains are 1.8 and 2.2 points in pixel AP and 0.7 and 1.5 points in AUPRO.

E.2 Category-Level Differences

Independent improves pixel AP and AUPRO over Concat in all 15 MVTec-AD categories and all 12 VisA categories. The median category-level pixel-AP gains are 1.8 and 2.0 points. This consistency shows that the aggregate difference from Concat extends across the evaluated product categories.

Relative to 1L Soft, pixel AP improves in 9/15 MVTec-AD categories and 9/12 VisA categories. AUPRO improves in 11/15 and 12/12 categories. Leather in MVTec-AD favors the one-layer memory in pixel AP, whereas cashew and chewinggum in VisA have large gains from the independent four-layer memory. The size of the benefit from distributing references across layers thus varies with the category. Appendix F.2 contains every category and all seven metrics.

For example, at matched reference count, the MVTec-AD screw category increases from 54.3% to 64.4% pixel AP, and hazelnut increases from 75.9% to 84.3%. Leather favors the one-layer configuration, at 74.8% versus 63.9%. The storage-matched Concat values for these categories are 62.7%, 83.1%, and 60.6%, respectively, all below Independent. The distinction between a larger single-layer neighborhood and a joint four-layer neighborhood is therefore consequential even when the final score is a reconstruction residual in both cases.

E.3 Inference Cost

Table 5 measures host-to-device transfer, the encoder forward pass, projection, memory scoring, and map processing on an RTX 4090. Offline memory construction and CPU image decoding and preprocessing are excluded. The one-bank setting saves 18.9 ms per image on MVTec-AD and 23.2 ms on VisA relative to five banks, with peak-VRAM reductions of 48.0 and 110.8 MiB. The corresponding pixel-AP means are 75.0%/56.6% for one bank and 75.3%/57.0% for five banks.

For P queries, D feature dimensions, and n references per layer–bank pair, exhaustive distance computation costs $O(|\mathcal{L}|BPnD)$. Weighted reconstruction adds $O(|\mathcal{L}|BPkD)$, and feature storage requires $O(|\mathcal{L}|BnD)$ values. Increasing bank count increases retrieval and reconstruction work, while all banks and layers share the encoder forward pass.

F Complete Metrics and Category Results

F.1 Three-Seed Performance

Tables 13 and 14 report all seven metrics for the Hard–Soft and bank-count comparisons. Each row uses a 1% coreset per layer and bank. Hard uses nearest-reference distance; Soft uses distance-weighted reconstruction. Four-layer configurations use independent memories. Values are percentages, reported as mean ± sample standard deviation over memory seeds 0, 1, and 2. F1 means max-F1 throughout.

Configuration	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	AUPRO
1L Hard, $B = 1$	99.0±0.1	99.6±0.0	98.3±0.1	95.2±0.1	69.2±0.2	67.7±0.2	89.9±0.1
4L Hard, $B = 1$	99.6±0.1	99.9±0.0	99.2±0.1	97.3±0.0	73.6±0.0	71.1±0.0	94.1±0.1
4L Soft, $B = 1$	99.7±0.0	99.9±0.0	99.1±0.0	98.2±0.0	75.0±0.0	71.9±0.0	95.1±0.0
4L Soft, $B = 5$	99.7±0.0	99.9±0.0	99.2±0.1	98.3±0.0	75.3±0.0	72.0±0.0	95.2±0.1

Table 13: **MVTec-AD complete metrics.** Mean ± sample standard deviation across three memory seeds. All four-layer rows use independent memories and all rows use a 1% coreset per layer and bank. F1 means max-F1.

Configuration	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	AUPRO
1L Hard, $B = 1$	96.8±0.3	96.9±0.3	93.9±0.2	93.5±0.1	45.2±0.3	49.8±0.3	82.3±0.2
4L Hard, $B = 1$	97.8±0.0	97.9±0.0	95.1±0.3	97.5±0.0	54.1±0.2	56.7±0.2	89.7±0.3
4L Soft, $B = 1$	98.5±0.0	98.5±0.0	96.1±0.0	98.8±0.0	56.6±0.2	58.2±0.1	93.8±0.1
4L Soft, $B = 5$	98.5±0.0	98.6±0.0	96.2±0.1	98.9±0.0	57.0±0.0	58.5±0.1	94.0±0.1

Table 14: **VisA complete metrics.** Mean ± sample standard deviation across three memory seeds. All four-layer rows use independent memories and all rows use a 1% coreset per layer and bank. F1 means max-F1.

Four-layer Hard and Soft both achieve high image-level scores on MVTec-AD, while Soft improves its pixel metrics. On VisA, the same readout change also raises image AUROC from 97.8% to 98.5%. On both datasets, pixel AP and AUPRO improve together.

F.2 Category Results at Matched Memory Budgets

The following tables give the category results for the three configurations in Table 4. All use seed 0, $B = 1$, and the Dinomaly evaluation protocol. Values are percentages. Coreset ratios are 4% for 1L Soft, 1% per layer for 4L Soft (Independent), and 1% of the joint feature set for 4L Soft (Concat). The final rows are unweighted category means. Category-level comparisons use values before the one-decimal display rounding; single-seed rows have no seed standard deviation.

Table 15: **MVTec-AD category results at matched memory budgets.** Seed 0, $B = 1$, and Dinomaly-based evaluation. All metrics are percentages; F1 means max-F1.

Category	Configuration	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	AUPRO
Bottle	1L Soft	100.0	100.0	100.0	99.1	91.1	85.3	97.2
	4L Soft (Independent)	100.0	100.0	100.0	99.1	90.3	83.5	97.2
	4L Soft (Concat)	100.0	100.0	100.0	99.0	89.7	83.2	96.5
Cable	1L Soft	99.9	100.0	98.9	97.2	75.0	70.4	91.9
	4L Soft (Independent)	99.9	99.9	98.9	98.1	77.9	72.7	94.3
	4L Soft (Concat)	99.7	99.8	98.3	97.7	76.4	72.5	93.8
Capsule	1L Soft	98.9	99.8	98.6	97.2	59.3	57.3	96.4
	4L Soft (Independent)	98.4	99.6	99.1	98.8	64.0	59.8	97.8
	4L Soft (Concat)	98.6	99.7	99.1	98.3	63.2	59.1	97.2
Carpet	1L Soft	99.3	99.8	98.9	98.0	76.5	74.0	95.3
	4L Soft (Independent)	99.9	100.0	99.4	99.4	81.2	77.2	98.4
	4L Soft (Concat)	100.0	100.0	99.4	99.3	79.0	76.2	98.1
Grid	1L Soft	99.4	99.8	99.1	98.8	67.6	64.2	96.2
	4L Soft (Independent)	99.8	99.9	99.1	99.0	69.4	66.2	96.0
	4L Soft (Concat)	99.8	99.9	99.1	99.0	67.5	65.7	95.8
Hazelnut	1L Soft	100.0	100.0	100.0	98.9	75.9	73.7	97.4
	4L Soft (Independent)	100.0	100.0	100.0	99.5	84.3	79.9	97.8
	4L Soft (Concat)	100.0	100.0	100.0	99.5	83.1	79.5	97.6
Leather	1L Soft	100.0	100.0	100.0	99.4	74.8	69.0	99.1
	4L Soft (Independent)	100.0	100.0	100.0	99.4	63.9	62.0	98.6
	4L Soft (Concat)	100.0	100.0	100.0	99.3	60.6	59.9	98.1
Metal nut	1L Soft	100.0	100.0	100.0	97.5	88.3	87.4	94.4
	4L Soft (Independent)	100.0	100.0	100.0	97.3	83.4	86.7	94.8
	4L Soft (Concat)	100.0	100.0	100.0	96.8	81.6	85.2	94.5
Pill	1L Soft	98.9	99.8	97.6	96.5	80.8	74.9	95.7
	4L Soft (Independent)	99.4	99.9	99.3	96.6	75.5	70.3	96.8
	4L Soft (Concat)	99.4	99.9	98.9	95.4	72.8	69.1	96.3
Screw	1L Soft	97.1	99.1	94.8	97.7	54.3	56.5	92.3
	4L Soft (Independent)	98.3	99.4	95.7	98.6	64.4	60.8	95.7
	4L Soft (Concat)	96.8	99.0	94.1	98.2	62.7	60.7	93.6
Tile	1L Soft	99.9	100.0	99.4	97.7	81.6	82.8	90.5
	4L Soft (Independent)	100.0	100.0	100.0	97.7	78.0	79.9	89.9
	4L Soft (Concat)	100.0	100.0	100.0	97.5	75.9	78.7	88.9
Toothbrush	1L Soft	100.0	100.0	100.0	99.0	66.3	69.4	96.4
	4L Soft (Independent)	100.0	100.0	100.0	99.1	70.2	72.4	96.7
	4L Soft (Concat)	100.0	100.0	100.0	99.0	67.7	70.2	96.0
Transistor	1L Soft	99.0	98.6	95.0	91.9	62.9	59.0	79.4
	4L Soft (Independent)	99.6	99.4	97.6	95.0	70.3	65.7	82.7
	4L Soft (Concat)	99.6	99.4	97.6	94.2	68.9	65.3	82.3
Wood	1L Soft	99.6	99.9	98.4	96.4	77.0	70.5	93.7
	4L Soft (Independent)	99.5	99.8	98.4	97.1	79.1	72.4	94.9
	4L Soft (Concat)	99.7	99.9	99.2	96.9	79.0	72.3	93.9

Continued on the next page

MVTec-AD category results (continued)

Category	Configuration	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	AUPRO
Zipper	1L Soft	100.0	100.0	99.6	97.3	78.0	74.4	94.1
	4L Soft (Independent)	100.0	100.0	99.6	97.9	73.2	68.5	94.7
	4L Soft (Concat)	100.0	100.0	100.0	97.4	69.6	65.1	93.1
Mean	1L Soft	99.5	99.8	98.7	97.5	74.0	71.2	94.0
	4L Soft (Independent)	99.7	99.9	99.1	98.2	75.0	71.9	95.1
	4L Soft (Concat)	99.6	99.8	99.1	97.8	73.2	70.8	94.4

Table 16: **VisA category results at matched memory budgets.** Seed 0, $B = 1$, and Dinomaly-based evaluation. All metrics are percentages; F1 means max-F1.

Category	Configuration	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	AUPRO
Candle	1L Soft	98.6	98.5	95.2	98.6	46.3	49.8	93.5
	4L Soft (Independent)	98.6	98.5	94.6	99.3	51.4	52.8	94.9
	4L Soft (Concat)	98.6	98.6	96.0	99.2	50.7	52.9	94.7
Capsules	1L Soft	98.4	98.8	97.0	98.4	63.9	63.4	97.3
	4L Soft (Independent)	98.8	99.1	97.5	98.9	69.6	66.5	98.0
	4L Soft (Concat)	98.7	99.1	97.1	98.6	67.0	64.7	97.0
Cashew	1L Soft	98.8	99.4	96.6	84.7	41.2	49.3	79.8
	4L Soft (Independent)	99.1	99.6	96.6	98.0	69.3	67.6	93.9
	4L Soft (Concat)	98.7	99.4	95.6	96.3	65.0	64.8	89.3
Chewinggum	1L Soft	97.6	99.0	96.4	98.0	60.4	57.2	79.5
	4L Soft (Independent)	99.1	99.7	98.5	99.0	74.3	71.8	86.6
	4L Soft (Concat)	98.9	99.6	98.0	98.6	73.1	70.0	85.0
Fryum	1L Soft	98.6	99.4	96.9	91.9	39.2	45.9	90.7
	4L Soft (Independent)	98.7	99.4	96.5	95.7	48.7	53.1	92.1
	4L Soft (Concat)	98.6	99.4	96.1	95.0	47.6	52.0	90.4
Macaroni 1	1L Soft	95.3	94.8	91.3	99.3	40.5	45.3	95.2
	4L Soft (Independent)	97.4	97.3	93.6	99.4	40.0	44.5	95.7
	4L Soft (Concat)	96.4	96.3	91.0	99.3	37.1	43.2	94.6
Macaroni 2	1L Soft	91.1	89.4	87.3	99.3	35.5	43.6	96.8
	4L Soft (Independent)	94.4	93.8	90.5	99.7	34.1	43.1	98.6
	4L Soft (Concat)	92.1	91.9	86.6	99.6	32.1	41.7	98.1
PCB1	1L Soft	98.7	98.6	96.0	99.1	82.9	76.9	93.1
	4L Soft (Independent)	98.8	98.8	95.4	99.6	87.3	79.8	95.7
	4L Soft (Concat)	98.7	98.8	95.0	99.6	86.2	79.2	94.5
PCB2	1L Soft	97.7	97.2	96.0	98.4	36.0	42.4	86.1
	4L Soft (Independent)	97.6	96.9	95.5	98.9	40.8	48.4	90.7
	4L Soft (Concat)	97.5	96.9	95.0	98.6	38.3	48.4	87.6
PCB3	1L Soft	99.0	98.8	97.5	99.1	38.8	41.4	92.5
	4L Soft (Independent)	99.5	99.6	96.9	99.2	43.0	46.4	92.9
	4L Soft (Concat)	99.4	99.4	96.5	99.1	37.9	44.2	90.9
PCB4	1L Soft	99.7	99.6	98.5	98.4	54.3	55.8	91.6
	4L Soft (Independent)	99.9	99.9	99.0	98.8	54.5	54.3	92.6
	4L Soft (Concat)	99.9	99.9	99.0	98.4	52.4	53.2	91.9
Pipe fryum	1L Soft	99.5	99.8	99.5	98.3	65.8	67.0	93.4
	4L Soft (Independent)	99.4	99.8	98.5	99.1	64.5	68.5	95.3
	4L Soft (Concat)	99.4	99.7	98.0	98.9	63.6	67.8	95.3
Mean	1L Soft	97.8	97.8	95.7	97.0	50.4	53.2	90.8
	4L Soft (Independent)	98.5	98.5	96.1	98.8	56.4	58.1	93.9
	4L Soft (Concat)	98.1	98.3	95.3	98.4	54.3	56.8	92.4

G Properties of Local Reconstruction Residuals

G.1 Residual and Reference Dispersion

Let $m_1, \ldots, m_k$ be the retrieved normal features, $w_j \geq 0$, and $\sum_j w_j = 1$. Their weighted estimate is $\hat{z} = \sum_j w_j m_j$. Writing $z - m_j = (z - \hat{z}) + (\hat{z} - m_j)$, then squaring, weighting, and summing gives

$$\begin{aligned}\sum_j w_j \|z - m_j\|_2^2 = \|z - \hat{z}\|_2^2 + \sum_j w_j \|m_j - \hat{z}\|_2^2 \\ + 2(z - \hat{z})^\top \sum_j w_j (\hat{z} - m_j).\end{aligned} \tag{11}$$

The last sum is $\hat{z} - \sum_j w_j m_j = 0$, so the cross term vanishes:

$$\boxed{\|z - \hat{z}\|_2^2 = \sum_j w_j \|z - m_j\|_2^2 - \sum_j w_j \|m_j - \hat{z}\|_2^2.} \tag{12}$$

G.2 Pairwise Reference Distances

The weighted dispersion satisfies $\sum_j w_j \|m_j - \hat{z}\|_2^2 = \frac{1}{2} \sum_i \sum_j w_i w_j \|m_i - m_j\|_2^2$. Substituting into Equation 12,

$$\|z - \hat{z}\|_2^2 = \sum_j w_j \|z - m_j\|_2^2 - \frac{1}{2} \sum_i \sum_j w_i w_j \|m_i - m_j\|_2^2. \tag{13}$$

G.3 Relation to Nearest-Neighbor Scoring

When $k = 1$, the single reference has weight one and $r_{\mathrm{Soft}}(z; M) = \|z - m_1\|_2 = r_{\mathrm{Hard}}(z; M)$. For $k > 1$ and a unique nearest reference, taking $\tau \to 0$ concentrates the weight on that reference, and Soft approaches Hard. As $\tau \to \infty$, all retrieved neighbors receive weight $1/k$, and the estimate approaches their arithmetic mean. The finite-temperature score interpolates through weighted local estimates determined by the neighbor distances.

G.4 Feature Scale and Adaptive Temperature

Apply a common transformation $z' = az + c$ and $m'_j = am_j + c$, where $a > 0$. Neighbor ordering is preserved, and squared distances are multiplied by a^2. With the same sampled queries, the median-based temperature becomes $\tau' = a^2 \tau$. Therefore,

$$\frac{\|z' - m'_j\|_2^2}{\tau'} = \frac{a^2 \|z - m_j\|_2^2}{a^2 \tau} = \frac{\|z - m_j\|_2^2}{\tau}. \tag{14}$$

For $\tau > 0$, reconstruction weights remain unchanged and the residual scales as $r' = ar$. The adaptive temperature thus preserves relative neighbor weights under a common positive rescaling of the feature space.

G.5 A Symmetric Reference Arrangement

Consider query $z = 0$ and five references at the vertices of a regular pentagon of radius $R > 0$:

$$m_j = R \left(\cos \frac{2\pi j}{5}, \sin \frac{2\pi j}{5} \right), \qquad j = 0, \ldots, 4. \tag{15}$$

All references are equally distant from the query, so each has weight $1/5$. Symmetry gives $\hat{z} = 0$. The Hard score is R, whereas the reconstruction residual is zero. This example isolates the distinction between distance to an individual normal reference and agreement with the weighted normal estimate. An anomalous query that is well reconstructed by the distance-weighted normal neighbors can therefore receive a small residual and be missed by Soft scoring.

H Qualitative Analysis

H.1 Layer Residuals

Layer maps place R_{-3}, R_{-6}, R_{-8}, and R_{-9} at corresponding spatial coordinates. Their mean provides the combined map. Figure 3 shows how individual feature spaces contribute to the response. In the displayed block numbering, blocks 4, 5, 7, and 10 correspond to layers $-9, -8, -6, -3$.

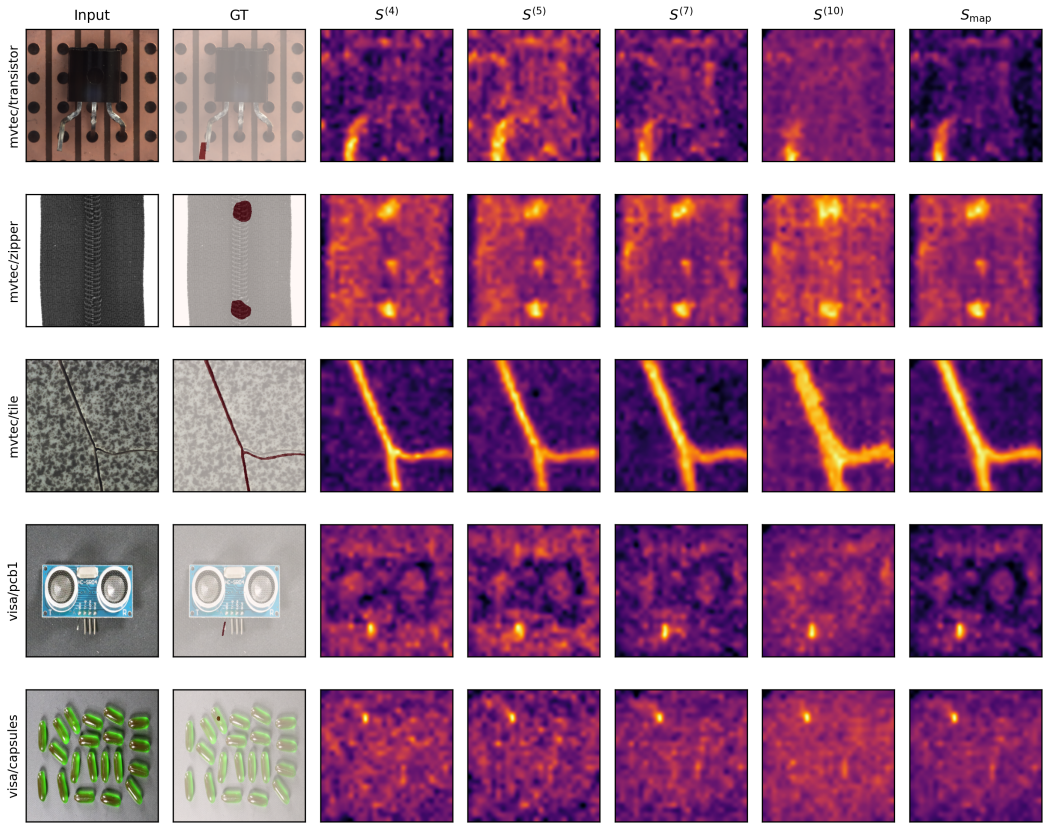

Figure 3: **Layer-wise residual maps.** The visualization places the input, mask, individual layer responses, and combined response side by side. Each layer searches and reconstructs from its own normal memory.

H.2 Examples Across Categories

Figures 4–7 show examples spanning the 15 MVTec-AD and 12 VisA categories. Each row aligns an input image, ground-truth defect region, nearest-neighbor memory map, soft reconstruction memory map, and ProCon map. The main-text examples show responses near a damaged transistor lead, two zipper defects, and a tile crack. The full set extends this visual inspection to the remaining object and texture categories.

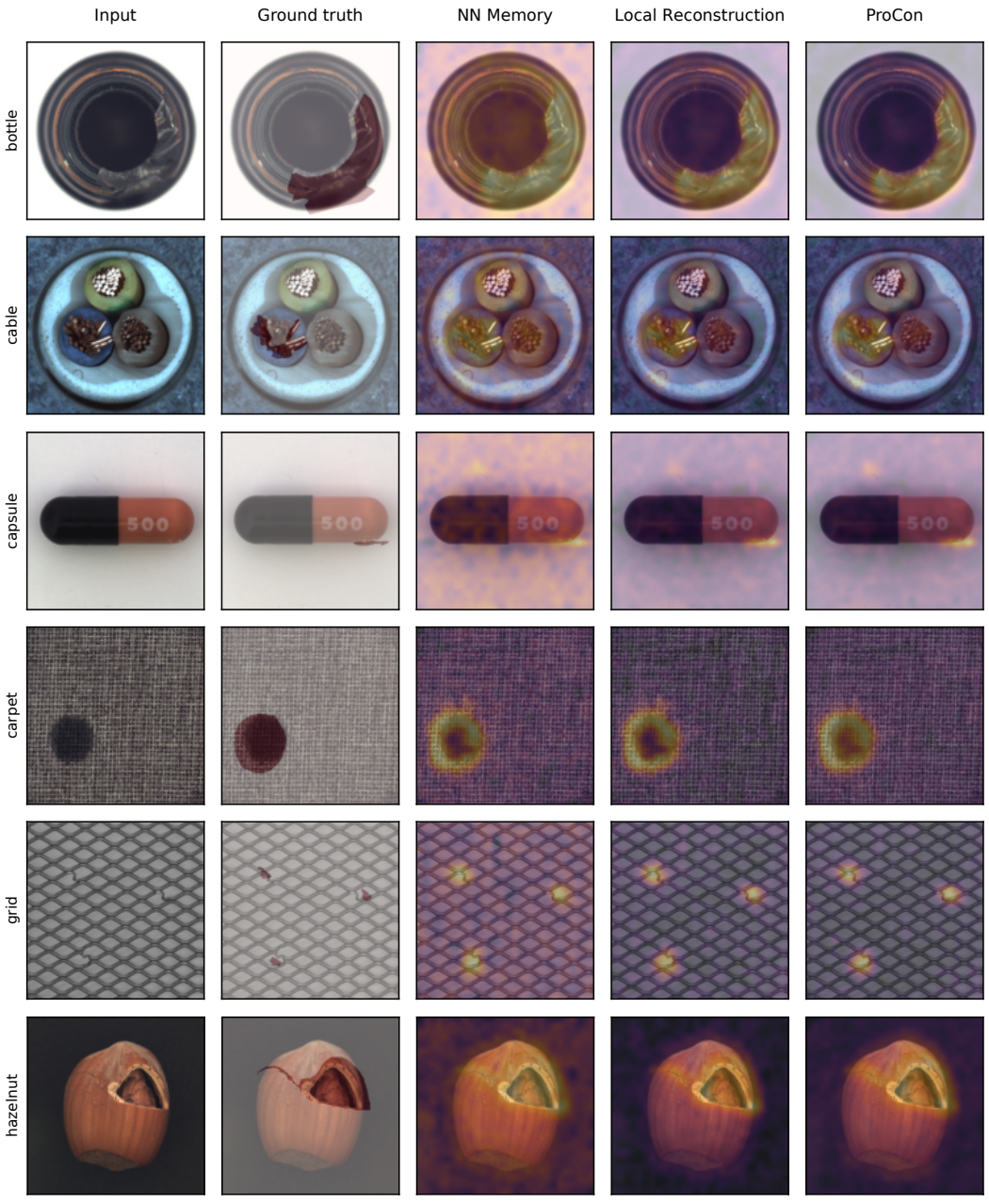

Figure 4: **Category examples, part 1.** Original ProCon qualitative examples, continued in Figures 5–7.

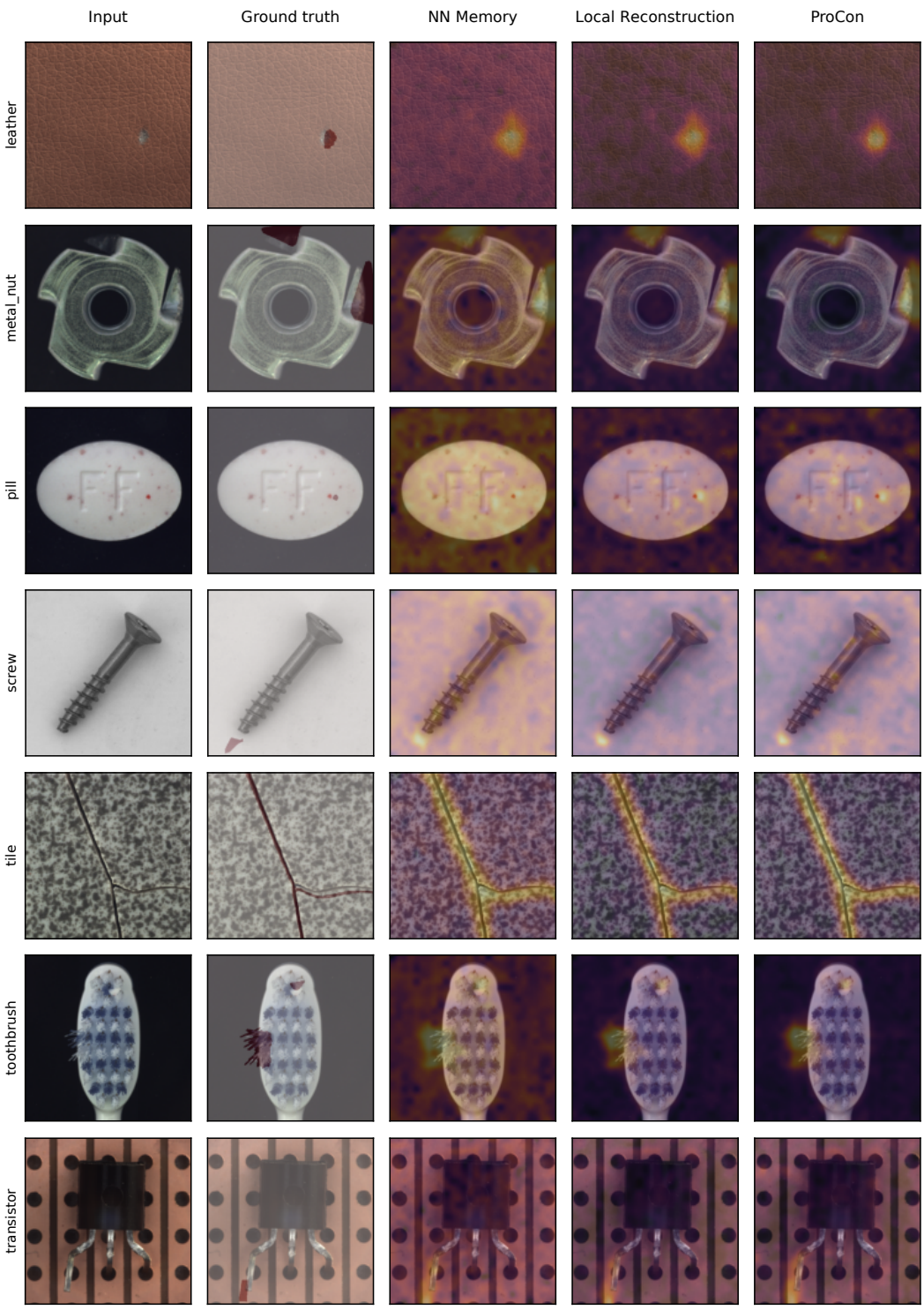

Figure 5: **Category examples, part 2.** Continuation of the MVTec-AD and VisA examples.

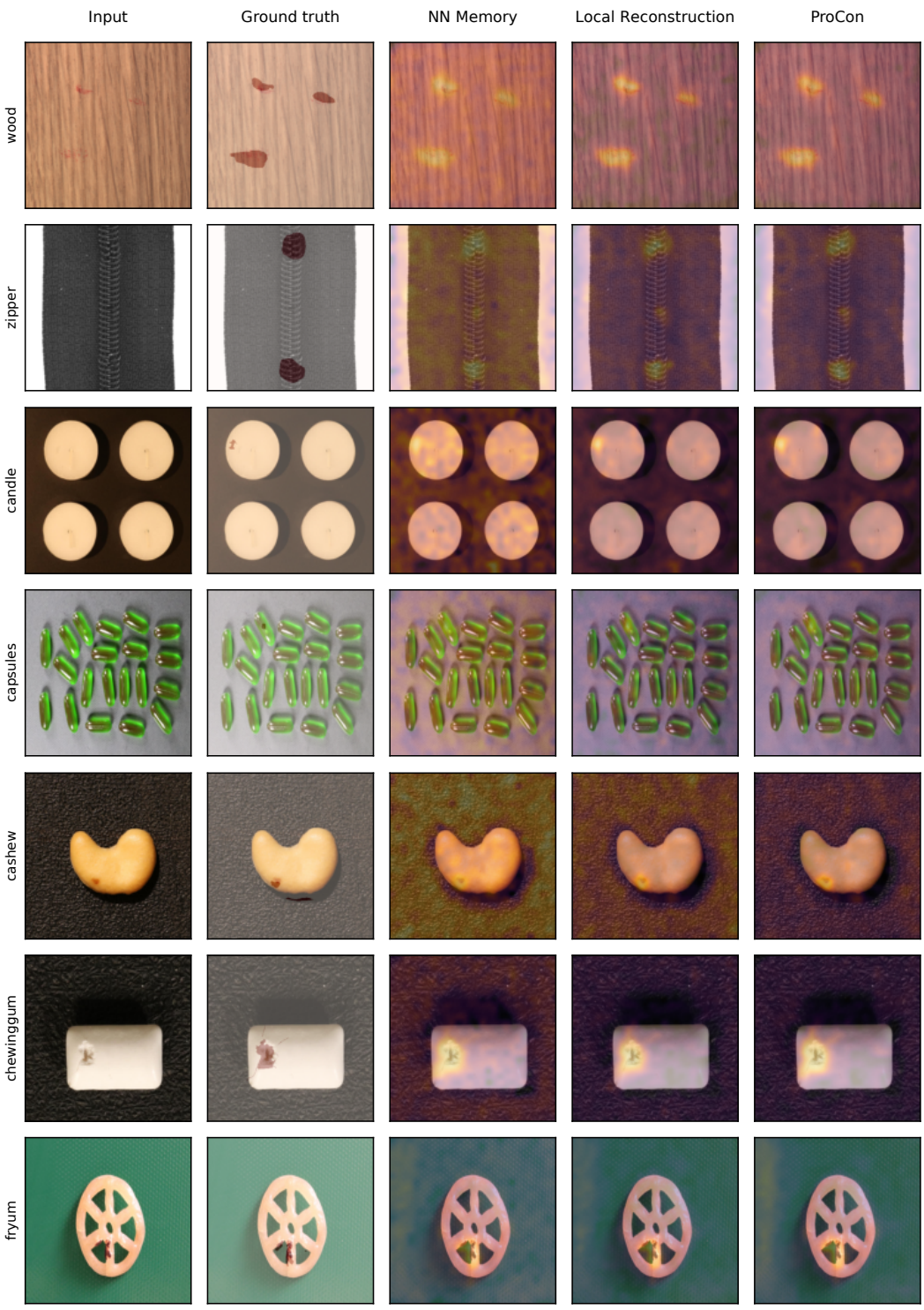

Figure 6: **Category examples, part 3.** Continuation of the MVTec-AD and VisA examples.



Figure 7: **Category examples, part 4.** Final examples completing all 27 categories.