Reproducing INP-Former: Reconstructing Intrinsic Normal Prototypes for Industrial Anomaly Detection

Anonymous authors
Paper under double-blind review

Abstract

We reproduce the central claims of Luo et al. (2025), which introduces INP-Former, a reconstruction-based method for multi-class industrial anomaly detection. The method extracts Intrinsic Normal Prototypes (INPs) from test images themselves, avoiding the need for stored memory banks, and uses them to guide a decoder that reconstructs normal features. We verify the main performance claims on MVTec AD and VisA using the authors' released code, extend the evaluation to MVTec AD 2, and conduct ablation studies on the loss components, number of prototypes, input resolution, and backbone architecture. Beyond reproduction, we propose several extensions: (1) LoRA-based parameter-efficient transfer learning from Real-IAD to downstream benchmarks, (2) an analysis of how anomaly area size affects detection performance, (3) an inference speed benchmark, (4) a lighting robustness evaluation on MVTec AD 2, and (5) an epoch efficiency analysis showing convergence behavior. Our reproduction supports the central claims: INP-Former achieves state-of-the-art multi-class anomaly detection, and both the INP mechanism and the proposed loss terms contribute meaningfully to performance. We additionally find that the optimal number of prototypes varies across datasets, that LoRA adaptation can recover most of the fully-trained performance using only 5% of trainable parameters, that the model largely converges by epoch 50, that small anomalies are the primary failure mode, and that the method is robust to brightness changes but vulnerable to color shifts.

The code we used for reproducing the study can be found here: `https://anonymous.4open.science/r/Repro-INPFormer-D6D1/README.md`

1 Introduction

Industrial anomaly detection (AD) aims to identify defective products in manufacturing using only normal training samples. Recent reconstruction-based approaches learn to encode and reconstruct normal patterns; at test time, anomalies produce high reconstruction error and are thus detected. A key challenge in multi-class settings where a single model handles all product categories is that the model must capture diverse normal appearances without conflating them.

Luo et al. (2025) propose INP-Former, which addresses this by extracting *Intrinsic Normal Prototypes* (INPs) directly from test images via cross-attention aggregation, rather than relying on stored memory banks or class-specific models. The INPs capture the dominant normal patterns present in each test image and guide a prototype-guided decoder to reconstruct features. Anomalous regions, which deviate from the extracted normal prototypes, yield high reconstruction error. The method additionally introduces a soft-mining loss $\mathcal{L}_{sm}$ and a coherence loss $\mathcal{L}_c$ to improve training stability.

In this reproducibility study, we:

1. Verify the main performance claims on MVTec AD (Bergmann et al., 2019) and VisA (Zou et al., 2022) using the authors' released code and pretrained DINOv2 backbone (Oquab et al., 2024).
2. Extend the evaluation to MVTec AD 2 (Bergmann et al., 2024), a newer and more challenging benchmark.

3. Reproduce the loss component ablation (Table 5 of the original paper) with multiple random seeds on MVTec AD to assess statistical robustness.
4. Conduct additional ablations on the number of INP tokens, input resolution, and backbone size.
5. Propose extensions beyond the original paper: LoRA-based parameter-efficient transfer learning (Hu et al., 2022), an analysis of anomaly area versus detection performance, inference speed benchmarks, a lighting robustness evaluation, epoch efficiency analysis, and AU-PRO evaluation at a stricter FPR threshold.

2 Scope of Reproducibility

We investigate the following claims from Luo et al. (2025):

Claim 1: INP-Former achieves state-of-the-art multi-class anomaly detection performance, reporting 99.6% image-level AUROC on MVTec AD and 98.8% on VisA with a single unified model.

Claim 2: The INP mechanism is essential: removing it (replacing the prototype-guided decoder with a standard bottleneck + self-attention decoder) substantially degrades performance.

Claim 3: Both the soft-mining loss $\mathcal{L}_{sm}$ and coherence loss $\mathcal{L}_c$ contribute to the final performance. The full model with both losses outperforms ablated variants.

Claim 4: The number of INP tokens M affects performance, with $M = 6$ providing the best trade-off on MVTec AD and VisA.

3 Methodology

3.1 Model Description

INP-Former uses a frozen DINOv2 with registers (Darcet et al., 2024) encoder (ViT-Base/14 by default) to extract multi-layer patch features from input images. These features pass through a bottleneck MLP, after which M learnable prototype queries aggregate normal information from the encoded features via cross-attention (the INP Extractor). The resulting INP tokens then guide a Prototype-Guided Decoder consisting of 8 transformer blocks with cross-attention between decoded features and the INPs. The training loss combines cosine similarity reconstruction loss with two auxiliary terms: a soft-mining loss $\mathcal{L}_{sm}$ that upweights hard-to-reconstruct normal patches, and a coherence loss $\mathcal{L}_c$ that encourages consistency among the extracted prototypes. At inference, anomaly maps are computed as the cosine distance between encoder and decoder features, followed by Gaussian smoothing.

We use the authors' released code without modification to the core model architecture or training procedure.

3.2 Datasets

- **MVTec AD** (Bergmann et al., 2019): 15 categories of industrial products and textures. 3,629 training images (normal only), 1,725 test images.
- **VisA** (Zou et al., 2022): 12 categories of complex structures. 8,659 training images, 2,102 test images.
- **MVTec AD 2** (Bergmann et al., 2024): 8 categories with more challenging anomalies including subtle defects and lighting variations. Not evaluated in the original paper. Due to a technical issue with the download link being down we had to omit the fabric category altogether thus utilizing 7/8 of the classes.
- **Real-IAD** (Wang et al., 2024): 30 categories of real industrial products. Used as source domain for transfer learning experiments. We did not use it to train for multi-class anomaly detection as the dataset was far larger compared to other 2 (53 GB vs approximately 2 GB for ViSa and 5 GB for MVTEC AD)

Table 1: Multi-class anomaly detection results. "Original" values from Luo et al. (2025); "Ours" from our reproduction using the authors' code with the same hyperparameters.

Dataset	Source	I-AUROC	P-AUROC	P-AUPRO
MVTec AD	Original	99.6	98.5	95.0
	Ours	99.67	98.45	94.95
VisA	Original	98.8	99.0	95.1
	Ours	98.90	98.97	95.27
MVTec AD 2	Ours	69.29	88.87	60.29

3.3 Hyperparameters

We follow the exact hyperparameters reported in the original paper: input size 448×448, center crop to 392×392, batch size 16, 200 training epochs, Adam optimizer (lr $= 10^{-3}$, $\beta = (0.9, 0.999)$), $M = 6$ INP tokens, $\mathcal{L}_{sm}$ with $y = 3$, $\mathcal{L}_c$ with $\lambda = 0.2$. The encoder (DINOv2 ViT-Base/14 with registers) is frozen throughout training; only the bottleneck, INP extractor, and decoder are trained.

3.4 Experimental Setup

All experiments were conducted on an institutional HPC cluster using NVIDIA A100-SXM4-40GB and L4-24GB GPUs.

3.5 Evaluation Metrics

We report image-level AUROC (I-AUROC) for detection, pixel-level AUROC (P-AUROC) for segmentation, and per-region overlap AU-PRO at FPR≤0.3 (P-AUPRO) for localization quality. All metrics are computed per-category and averaged.

3.6 Computational Budget

The total computational budget for this study was approximately 450 GPU-hours on NVIDIA A100 and L4 GPUs, distributed across: baseline training (3 datasets × 15h ≈ 45h), seed ablation (17 runs × 7–15h ≈ 207h), INP count ablation (7 runs × 15h ≈ 105h), resolution/backbone ablation (4 runs × 15h ≈ 60h), LoRA transfer (6 runs × 3h ≈ 24h), epoch efficiency (9h), and evaluation-only experiments (∼4h).

4 Results

4.1 Claim 1: Main Performance Reproduction

Table 1 compares our reproduced results with those reported in the original paper. We find close agreement on both MVTec AD and VisA, supporting Claim 1.

Our MVTec AD results match the original to within 0.1 percentage points across all three metrics. VisA results are similarly close. We additionally report results on MVTec AD 2, which was not evaluated in the original paper. The substantially lower performance on MVTec AD 2 (69.3% I-AUROC vs. 99.7% on MVTec AD) reflects the increased difficulty of this benchmark, which features more subtle and diverse anomaly types.

Per-category results are provided in Appendix A.

4.2 Claim 2: INP Mechanism is Essential

Table 2 shows the loss/module ablation with multiple random seeds on MVTec AD, alongside the values reported in the original paper (Table 5 of Luo et al. 2025). Removing the INP mechanism entirely ("No INP"

Table 2: Loss and module ablation on MVTec AD. "Original" values from Table 5 of Luo et al. (2025); "Ours" are mean ± std over 5 seeds. Rows marked with "—" were not reproduced due to computational constraints. Row 4† was not tested in the original paper; we add it to isolate $\mathcal{L}_{sm}$'s contribution.

				Original		Ours	
Configuration	INP	$\mathcal{L}_c$	$\mathcal{L}_{sm}$	I-AUROC	P-AUPRO	I-AUROC	P-AUPRO
No INP	×	×	×	98.59	92.73	98.19 ± 0.65	91.55 ± 1.19
INP only	✓	×	×	99.53	94.69	—	—
INP + $\mathcal{L}_c$	✓	✓	×	99.61	95.10	—	—
INP + $\mathcal{L}_{sm}$†	✓	×	✓	—	—	99.64 ± 0.04	94.85 ± 0.08
Full	✓	✓	✓	99.67	94.87	99.62 ± 0.01	94.92 ± 0.07

† Configuration not present in the original paper's ablation.

row) reduces I-AUROC by 1.4 points and P-AUPRO by 3.3 points compared to INP + $\mathcal{L}_{sm}$, confirming that the INP mechanism provides a meaningful contribution. Our No INP result (98.19) is slightly lower than the original's (98.59), but it's still within the range of variance for the no-INP setting. We observe that the experimental settings with INP enabled achieved lower variance overall and thus provided more stable inference results.

4.3 Claim 3: Loss Component Contributions

The ablation in Table 2 addresses Claim 3. Our Full configuration achieves $99.62 \pm 0.01\%$ I-AUROC and $94.92 \pm 0.07\%$ P-AUPRO, closely matching the original (99.67/94.87) results on MVTec AD.

The original paper tested INP + $\mathcal{L}_c$ (without $\mathcal{L}_{sm}$) but never isolated INP + $\mathcal{L}_{sm}$ (without $\mathcal{L}_c$). We fill this gap: our INP + $\mathcal{L}_{sm}$ achieves $99.64 \pm 0.04\%$ I-AUROC and $94.85 \pm 0.08\%$ P-AUPRO. Read alongside the Full configuration and the original's INP + $\mathcal{L}_c$ row, this exposes a pattern that the original paper's discussion does not separate: the two auxiliary losses contribute to different metrics.

$\mathcal{L}_{sm}$ primarily improves image-level detection. On MVTec AD I-AUROC, the original Table 5 reports 99.53 (INP only) → 99.61 (+ $\mathcal{L}_c$) → 99.67 (+ $\mathcal{L}_{sm}$): the $\mathcal{L}_{sm}$ step adds 0.06 points, comparable to the 0.08 from $\mathcal{L}_c$. Our reproduction sharpens this: INP + $\mathcal{L}_{sm}$ alone reaches $99.64 \pm 0.04\%$, statistically indistinguishable from the Full configuration's $99.62 \pm 0.01\%$. This aligns with the loss's stated motivation: upweighting hard-to-reconstruct normal patches widens the gap between normal and anomalous responses, and image-level AUROC (computed from max-pooled anomaly scores) rewards exactly that contrast.

$\mathcal{L}_c$ primarily improves pixel-level localization. The P-AUPRO trajectory in the original paper runs in the opposite direction: 94.69 (INP only) → 95.10 (+ $\mathcal{L}_c$) → 94.87 (+ $\mathcal{L}_{sm}$). Adding $\mathcal{L}_c$ produces a +0.41 jump, the largest single-loss contribution to either metric on MVTec AD, and adding $\mathcal{L}_{sm}$ on top *reduces* P-AUPRO by 0.23 points. Our reproduction is directionally consistent: Full (94.92 ± 0.07) edges past INP + $\mathcal{L}_{sm}$ (94.85 ± 0.08), though within seed noise. The mechanism is intuitive: $\mathcal{L}_c$ forces the extracted INPs to coherently represent normality, yielding cleaner reconstructions of normal regions and thus crisper anomaly-map boundaries: precisely what AU-PRO rewards.

The prescribed defaults are not optimal on either metric. The authors present the Full configuration ($\lambda = 0.2$, $\gamma = 3$) as the recommended setting and apply identical hyperparameters across MVTec AD, VisA, and Real-IAD. But by their own MVTec AD numbers, no single configuration is best on both metrics simultaneously: INP + $\mathcal{L}_c$ achieves the highest P-AUPRO (95.10) while Full achieves the highest I-AUROC (99.67) at the cost of 0.23 P-AUPRO points. The gap between Full and each metric's optimum is of similar magnitude to the gain Full claims over INP-only, suggesting the two losses target different failure modes rather than jointly improving the same one. Practitioners with a clear primary objective should consider tuning λ and γ accordingly: an image-level pass/fail use case can likely drop $\mathcal{L}_c$ at no measurable cost, while a localization-critical use case (e.g., downstream defect classification or root-cause analysis) may benefit from reducing or removing $\mathcal{L}_{sm}$. We note that this metric specialization is most apparent on MVTec AD; on VisA,

Table 3: Effect of INP count M on MVTec AD 2. All runs use ViT-Base/14 at 448/392 resolution.

M	I-AUROC	P-AUROC	P-AUPRO
1	70.57	88.46	63.03
2	**72.43**	88.03	**62.98**
4	69.85	**89.23**	61.54
6	69.29	88.87	60.29
8	68.70	88.99	60.73
12	69.27	88.64	60.02
16	68.27	88.13	57.60

Table 4: Resolution and backbone ablation on MVTec AD.

Backbone	Input/Crop	I-AUROC	P-AUROC	P-AUPRO
ViT-Base/14	224/196	99.21	97.97	92.07
ViT-Base/14	336/294	99.63	98.28	94.04
ViT-Base/14	448/392	99.67	98.45	94.95
ViT-Small/14	448/392	99.20	98.18	94.39
ViT-Base/14	448/392	99.67	98.45	94.95

Table 5 of the original paper shows $\mathcal{L}_{sm}$ improving both metrics simultaneously (P-AUPRO 93.71 $\rightarrow$ 94.36, I-AUROC 98.16 $\rightarrow$ 98.90 when added on top of $\mathcal{L}_c$), so the prescription happens to be reasonable there.

Due to computational constraints (each configuration requires $\sim$75 GPU-hours for 5 seeds), we reproduced 3 of the 5 configurations on MVTec AD only, choosing the endpoints (No INP, Full) and one novel intermediate (INP + $\mathcal{L}_{sm}$) to maximize the insight per GPU-hour. Reproducing INP + $\mathcal{L}_c$ and INP-only under our seeding protocol, to confirm the $\mathcal{L}_c$ localization benefit reproduces with variance estimates: is the most informative follow-up experiment we did not run.

4.4 Claim 4: Effect of INP Count

The original paper reports an INP count ablation on MVTec AD and VisA, finding $M = 6$ to be optimal. We extend this analysis to MVTec AD 2 (Table 3).

Interestingly, the optimal M on MVTec AD 2 is lower ($M = 2$ for I-AUROC, $M = 4$ for P-AUROC) than on MVTec AD/VisA ($M = 6$). Performance degrades monotonically for $M \geq 4$ on I-AUROC. This suggests that the ideal number of prototypes is dataset-dependent: we hypothesize that given MVTec AD 2 has fewer categories (7 vs. 15) and different anomaly characteristics, it potentially requires fewer prototypes to capture the normal distribution. With too many prototypes, the model may overfit to noise or redundantly represent similar normal patterns.

5 Additional Experiments Beyond the Original Paper

5.1 Resolution and Backbone Ablation

Table 4 shows the effect of input resolution and backbone size on MVTec AD.

Higher resolution provides consistent improvements, with P-AUPRO benefiting most (+2.9 points from 224 to 448). ViT-Small/14 incurs only a modest drop (-0.47 I-AUROC, -0.56 P-AUPRO), suggesting it may be a viable option when computational resources are limited.

Table 5: Inference speed on NVIDIA L4 GPU (batch size 1, 50 warmup + 200 timed runs).

Config	Setting	FPS	Latency (ms)
Resolution (ViT-B, M=6)	224/196	85.3	11.7
	336/294	50.1	20.0
	448/392	25.0	40.0
Backbone (448/392, M=6)	ViT-Small/14	72.6	13.8
	ViT-Base/14	25.0	40.0
INP count (ViT-B, 448/392)	M=2–16	25.0	40.0

Table 6: LoRA transfer learning from Real-IAD to target datasets. Zero-shot applies the Real-IAD model directly without finetuning. "Full train" trains from scratch on the target dataset (200 epochs). LoRA trainable params: 5.1% of total (adapter params: 0.3–0.6%).

	MVTec AD			VisA		
Method	I-AUROC	P-AUROC	P-AUPRO	I-AUROC	P-AUROC	P-AUPRO
Zero-shot	69.15	79.92	57.17	51.47	72.60	31.60
LoRA $r = 2$	97.22	96.21	91.30	91.79	96.29	85.11
LoRA $r = 4$	97.81	96.62	91.87	93.17	96.79	87.55
LoRA $r = 8$	98.16	96.96	92.47	94.44	97.25	89.91
Full train	99.67	98.45	94.95	98.90	98.97	95.27

5.2 Inference Speed

We benchmark inference speed on an NVIDIA L4 GPU (Table 5). Notably, the number of INP tokens has virtually no impact on throughput: all values of M from 2 to 16 yield identical FPS ($\sim$25). This is because the INP cross-attention operates on M tokens against N_{patches} features, and at $M \leq 16$ the overhead is negligible compared to the encoder forward pass. Input resolution is the dominant factor, with a 3.4$\times$ slowdown from 224 to 448. ViT-Small/14 is 2.9$\times$ faster than ViT-Base/14.

5.3 LoRA Transfer Learning

We investigate whether INP-Former's learned representations can be efficiently transferred across datasets using Low-Rank Adaptation (LoRA) (Hu et al., 2022). Starting from a model pretrained on Real-IAD (Wang et al., 2024), we inject LoRA adapters (rank $r \in \{2, 4, 8\}$) into the decoder and bottleneck Linear layers (42 adapters total), freeze all other parameters, and finetune for 50 epochs on the target dataset.

LoRA finetuning with only 5% trainable parameters recovers the majority of fully-trained performance. At $r = 8$, MVTec AD reaches 98.2% I-AUROC (vs. 99.7% full train) and VisA reaches 94.4% (vs. 98.9%). Performance scales consistently with rank: each doubling of r improves I-AUROC by 0.3–1.4 points, with diminishing returns. The improvement over zero-shot transfer is dramatic: +29.0 I-AUROC on MVTec AD and +43.0 on VisA for $r = 8$. P-AUPRO shows a similar trend, with LoRA $r = 8$ reaching 92.5% vs. 95.0% full train on MVTec AD. Our experiments suggest that although INP-former is unable to learn generic representations that would enable it to perform accurate anomaly detection zero-shot, the representations learned from industrial datasets are largely transferrable between different datasets and only minor low-rank adjustments can help the model adapt to a new domain.

Table 7: Detection and localization by anomaly area quintile. Anomalous images are sorted by defect area and split into 5 equal-sized bins, each evaluated against all normal images.

	MVTec AD		VisA	
Area quintile	I-AUROC	P-AUPRO	I-AUROC	P-AUPRO
Q1 (smallest 20%)	90.7	96.0	89.1	95.3
Q2	98.7	95.8	94.4	93.1
Q3	99.6	92.1	97.6	90.4
Q4	99.5	90.5	99.7	89.9
Q5 (largest 20%)	100.0	83.0	99.9	87.5

5.4 Anomaly Area vs. Detection and Localization Performance

We investigate how anomaly size affects detection performance. Our initial hypothesis was that large anomalies would degrade INP extraction, fewer normal patches available for prototype learning. INP-Former's original premise was that it would be able to learn normality within images since most defective images would be dominated by the distribution of normal data, leading to worse detection when there are larger percent of anomalous areas. However, the results show the opposite trend.

On MVTec AD (1,258 anomalous images), anomaly score correlates positively with ground-truth anomaly area (Spearman $\rho = +0.53$, $p < 10^{-90}$). VisA shows an even stronger correlation ($\rho = +0.79$). Table 7 shows per-bin I-AUROC and P-AUPRO for both datasets, where each bin's anomalous images are evaluated against all normal images. Both datasets exhibit the same dual trend.

The two metrics reveal opposite trends with anomaly size. **Image-level detection** (I-AUROC) improves monotonically: small defects (<0.5% area) are hardest to detect (90.7%), while large defects (>6.2%) are trivially detected (100.0%). This is expected, larger defects produce more patches with high reconstruction error, yielding a stronger aggregate anomaly signal.

However, **pixel-level localization** (P-AUPRO) shows the reverse: it *degrades* from 96.0% for tiny defects to 83.0% for large ones. This reflects the per-region nature of AU-PRO: large anomalies span more pixels and have more complex boundaries, making precise localization harder. The model correctly identifies the image as anomalous but struggles to delineate the exact defect boundary. Small defects, when detected, are localized more precisely because the anomalous region is compact and well-defined.

This implies 2 things: INP-Former's detection failure mode lies in small, subtle defects, precisely the scenario where industrial inspection is most challenging; for large defects, detection is reliable but localization quality drops, which may matter for downstream repair or root-cause analysis.

5.5 Epoch Efficiency

We evaluate how quickly INP-Former converges by training the full configuration ($y = 3$, $\lambda = 0.2$) on MVTec AD and evaluating at intermediate checkpoints (Table 8).

The model exhibits rapid early convergence: at epoch 25, I-AUROC already reaches 99.3% and P-AUPRO reaches 94.6%: a difference of 0.4 points from the final value. The loss drops 68% in the first 25 epochs (0.395→0.163) but only 22% over the remaining 175 epochs. From epoch 100 onward, improvements are marginal (<0.1 P-AUPRO points per 50 epochs). This suggests that 50–100 epochs may be sufficient for deployment under constrained computate, halving the training cost with <0.2% metric degradation.

5.6 AU-PRO at Strict FPR Threshold

The standard AU-PRO metric integrates the per-region overlap curve up to FPR≤0.3. We additionally evaluate at a stricter threshold of FPR≤0.05, which better reflects industrial inspection requirements where

Table 8: Training convergence on MVTec AD. All metrics are means across 15 categories. Δ shows relative change from epoch 200 (final).

Epoch	Loss	I-AUROC	P-AUROC	P-AUPRO	ΔP-AUPRO
25	0.163	99.29	98.07	94.57	−0.39
50	0.145	99.50	98.26	94.76	−0.20
100	0.133	99.55	98.40	94.90	−0.06
150	0.128	99.59	98.45	94.95	−0.01
200	0.126	99.61	98.48	94.96	—

Table 9: AU-PRO at standard (FPR≤0.3) vs. strict (FPR≤0.05) thresholds. MVTec AD 2 produced harsher degradations in performance

Dataset	P-AUPRO$_{0.3}$	P-AUPRO$_{0.05}$	Δ
MVTec AD	93.05	72.14	−20.9
MVTec AD 2	59.04	32.27	−26.8

false positives are costly (Table 9). We're also motivated by the reasoning provided by MVTEC AD2's paper Bergmann et al. (2024), which points out that the commonly used FPR value of 0.3 biases in favor of allowing false positives to pass. Thus reducing the FPR value will allow us to evaluate how the model performs when the cost of faulty results is expensive.

Tightening the FPR threshold from 0.3 to 0.05 causes a substantial drop: −20.9 points on MVTec AD and −26.8 on MVTec AD 2. This indicates that while INP-Former achieves good localization on average, the anomaly maps still contain significant false-positive regions that only become apparent under strict thresholds. The degradation is more severe on MVTec AD 2, where the model already struggles with subtle defects. Per-category analysis shows that transistor (36.4%) and tile (45.2%) suffer the most on MVTec AD, while wallplugs (14.5%) and can (4.5%) are nearly unusable at FPR≤0.05 on MVTec AD 2.

5.7 Robustness to Lighting Variations (MVTec AD 2)

MVTec AD 2 includes test images under different lighting conditions, enabling a robustness analysis. Table 10 shows that over- and underexposure cause minimal performance degradation ($<$ 1.5 points I-AUROC), but color shifts severely impact detection (up to −13.7 points for shift_3).

This is expected: the DINOv2 encoder is trained on natural images and can handle brightness variations, but color shifts alter the feature distribution more fundamentally. Per-category analysis (Appendix B) reveals that categories with uniform textures (wallplugs, walnuts) are most sensitive to lighting shifts, while structurally distinct categories (vial) remain robust.

6 Discussion

Reproducibility. The original paper's central claims are well-supported by our reproduction. Using the authors' released code and the same hyperparameters, we obtained results within 0.1 percentage points on both MVTec AD and VisA. The code is well-structured and the key components (INP extraction, prototype-guided decoding, loss computation) match the paper's descriptions.

MVTec AD 2 Challenge. Extending INP-Former to MVTec AD 2 reveals significant performance degradation (69.3% vs. 99.7% I-AUROC on MVTec AD). This is consistent with the known difficulty of MVTec AD 2, which features more subtle defects and lighting variations. The INP count ablation on this dataset further suggests that the method's hyperparameters (particularly M) may need dataset-specific tuning.

Table 10: Lighting robustness on MVTec AD 2 (mean across 7 categories). Δ relative to regular lighting.

Condition	I-AUROC	P-AUROC	ΔI-AUROC
Regular	74.59	91.65	—
Overexposed	73.18	90.84	-1.4
Underexposed	75.42	89.16	$+0.8$
Shift 1	63.56	90.06	-11.0
Shift 2	72.44	86.95	-2.2
Shift 3	60.92	88.28	-13.7

Statistical Robustness. Our multi-seed ablation reveals that the INP mechanism substantially reduces variance across seeds (± 0.04 vs. ± 0.65 for I-AUROC), in addition to improving mean performance. This is a positive finding not highlighted in the original paper.

Practical Transfer with LoRA. The LoRA extension demonstrates that INP-Former's decoder learns transferable normal-pattern representations. With 5% of parameters trainable and $4\times$ fewer training epochs, LoRA recovers the majority of fully-trained performance. This has practical implications for deployment scenarios where labeled normal data is scarce or where quick adaptation to new product lines is needed.

Anomaly Size and Detection. Our area analysis reveals that INP-Former's primary failure mode is small anomalies ($< 0.5\%$ image area, 63.6% detection rate). This is not specific to the INP mechanism, any reconstruction-based method will struggle when the anomalous region produces minimal reconstruction error. However, it suggests that complementary approaches (e.g., memory bank methods that compare patch-level features) could address this weakness.

Lighting Robustness. The lightshift analysis on MVTec AD 2 shows that INP-Former is robust to brightness changes but vulnerable to color shifts. Since the DINOv2 encoder is frozen, the model cannot adapt to out-of-distribution color statistics at test time.

Inference Efficiency. The number of INP tokens has no measurable impact on inference speed, meaning practitioners can tune M for accuracy without latency concerns. Resolution is the main speed lever: dropping from 448 to 224 yields a $3.4\times$ speedup at the cost of -2.9 P-AUPRO points.

Epoch Efficiency. The convergence analysis reveals that 200 epochs may be more than necessary: the model reaches 99.5% of its final P-AUPRO by epoch 50. Combined with the inference speed results, this suggests a practical "fast mode": train for 50 epochs at 336/294 resolution for a $5\times$ reduction in training cost with minimal quality loss.

Localization Under Strict Thresholds. Our AU-PRO evaluation at FPR$\leq$0.05 reveals that INP-Former's pixel-level localization degrades significantly under strict false-positive constraints (-20.9 points on MVTec AD). This is relevant for industrial deployment where false positives trigger costly manual inspection. The gap between standard and strict AU-PRO suggests that anomaly map post-processing (e.g., adaptive thresholding, morphological filtering) may be necessary for production use.

Limitations of This Study. Due to computational constraints, we could not reproduce the full 5-configuration loss ablation on both MVTec AD and VisA as in the original paper, run multi-seed experiments on VisA, or compare against all baselines reported in the original paper. We focused our computational budget on the central claims and the most informative ablations.

7 Conclusion

In conclusion, our study successfully validates the central claims of INP-Former, confirming its robust multi-class anomaly detection performance on MVTec AD and VisA, and demonstrating the essential role of its Intrinsic Normal Prototype mechanism. Our deeper analysis reveals critical nuances not highlighted in the original work: the auxiliary losses (L_{sm} and L_c) independently optimize detection and localization, respectively, and the optimal number of prototypes is highly dataset-dependent.

Through our novel extensions, we demonstrate that INP-Former is highly practical for real-world use, as it converges rapidly and supports highly efficient domain transfer via LoRA using only 5% of trainable parameters. However, we also identify notable limitations for rigorous industrial deployment: the model struggles to detect subtle, small-area defects, is vulnerable to out-of-distribution color shifts, and exhibits significant localization degradation under strict false-positive thresholds (FPR 0.05). Ultimately, while INP-Former provides a highly capable and efficient architecture, deploying it in production environments will require targeted hyperparameter tuning and careful consideration of domain-specific constraints.

Author Contributions

Acknowledgments

References

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD – A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD 2 – A Larger, More Challenging Benchmark for Anomaly Detection and Localization in the Manufacturing Domain. *arXiv preprint*, 2024.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. *International Conference on Learning Representations (ICLR)*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shanen Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations (ICLR)*, 2022.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9974–9983, 2025. doi: 10.1109/CVPR52734.2025.00932.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Featured Certification.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-IAD: A Real-World Multi-View Dataset for Benchmarking Versatile Industrial Anomaly Detection. pp. 22883–22892, 06 2024. doi: 10.1109/CVPR52733.2024.02159.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. *SPot-the-Difference Self-supervised Pre-training for Anomaly Detection and Segmentation*, pp. 392–408. 11 2022. ISBN 978-3-031-20055-7. doi: 10.1007/978-3-031-20056-4_23.

A Per-Category Results

Table 11: Per-category results on MVTec AD (ViT-Base/14, 448/392, M=6).

Category	I-AUROC	P-AUROC	P-AUPRO	P-AUPRO$_{0.05}$
carpet	99.88	99.35	97.89	83.44
grid	99.92	99.42	97.75	81.69
leather	100.0	99.42	98.13	84.61
tile	100.0	97.68	88.36	45.22
wood	99.91	97.54	93.84	65.36
bottle	100.0	99.06	96.90	77.32
cable	100.0	98.80	94.78	69.17
capsule	99.00	98.74	97.02	83.03
hazelnut	100.0	99.41	96.69	70.38
metal_nut	100.0	97.35	95.24	68.16
pill	99.10	97.73	97.37	83.32
screw	97.46	99.58	98.00	86.85
toothbrush	100.0	98.98	95.75	64.97
transistor	99.71	94.73	79.65	36.37
zipper	100.0	98.98	96.86	82.24
Mean	**99.67**	**98.45**	**94.95**	**72.14**

Table 12: Per-category results on MVTec AD 2 (ViT-Base/14, 448/392, M=6).

Category	I-AUROC	P-AUROC	P-AUPRO	P-AUPRO$_{0.05}$
can	51.00	69.35	28.56	4.47
fruit_jelly	83.83	96.63	83.57	51.57
rice	63.36	93.28	60.10	31.97
sheet_metal	75.46	92.58	50.60	29.55
vial	87.24	90.27	87.18	62.22
wallplugs	46.06	84.08	37.57	14.53
walnuts	78.06	95.93	65.70	31.58
Mean	**69.29**	**88.87**	**59.04**	**32.27**

B Per-Category Lighting Robustness

Table 13: Per-category I-AUROC on MVTec AD 2 under different lighting conditions.

Category	Regular	Overexp.	Underexp.	Shift 1	Shift 2	Shift 3
can	52.2	52.2	61.1	48.3	48.3	51.7
fruit_jelly	85.3	84.0	84.0	84.0	—	—
rice	61.9	59.1	62.9	61.9	74.3	69.5
sheet_metal	83.3	81.7	83.3	73.3	76.7	68.3
vial	97.3	93.3	93.3	94.7	97.3	92.0
wallplugs	44.7	44.7	46.0	35.3	40.7	25.3
walnuts	97.3	97.3	97.3	47.3	97.3	58.7
Mean	**74.6**	**73.2**	**75.4**	**63.6**	**72.4**	**60.9**

Table 14: Per-category P-AUPRO on MVTec AD 2 under different lighting conditions.

Category	Regular	Overexp.	Underexp.	Shift 1	Shift 2	Shift 3
can	31.7	31.5	28.3	36.1	24.2	28.1
fruit_jelly	84.1	84.5	83.3	84.5	—	—
rice	66.0	66.2	63.2	60.0	61.5	63.4
sheet_metal	55.8	55.6	55.9	51.3	51.1	50.8
vial	94.4	90.7	84.2	91.8	93.9	93.5
wallplugs	64.8	63.0	65.7	51.9	38.0	47.2
walnuts	75.5	75.4	75.7	58.8	74.9	62.9
Mean	**67.5**	**66.7**	**65.2**	**62.1**	**57.3**	**57.6**