# ***CostFilter-AD***: Enhancing Anomaly Detection through Matching Cost Filtering

Zhe Zhang [1], Mingxiu Cai [1], Hanxiao Wang [2], Gaochang Wu [*,1], Tianyou Chai [1], Xiatian Zhu [*,3]

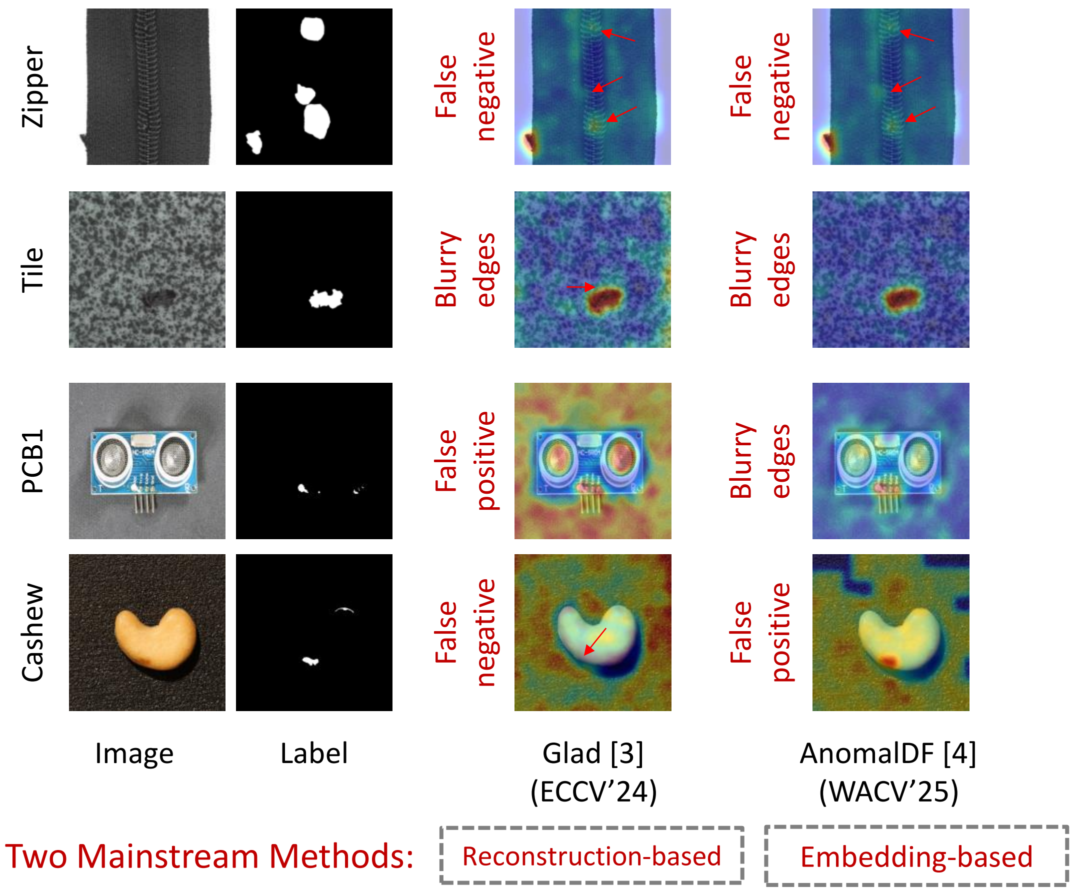
Reporter: Zhe Zhang

[1] Northeastern University, [2]Meta, [3]University of Surrey
[*] Corresponding author

1 July 2025

Northeastern University

## Background & Motivation: Unsupervised Anomaly Detection (UAD)[1] [2]




🔍 UAD is widely used in industrial inspection, where only normal data is available for training due to the scarcity of anomalies.

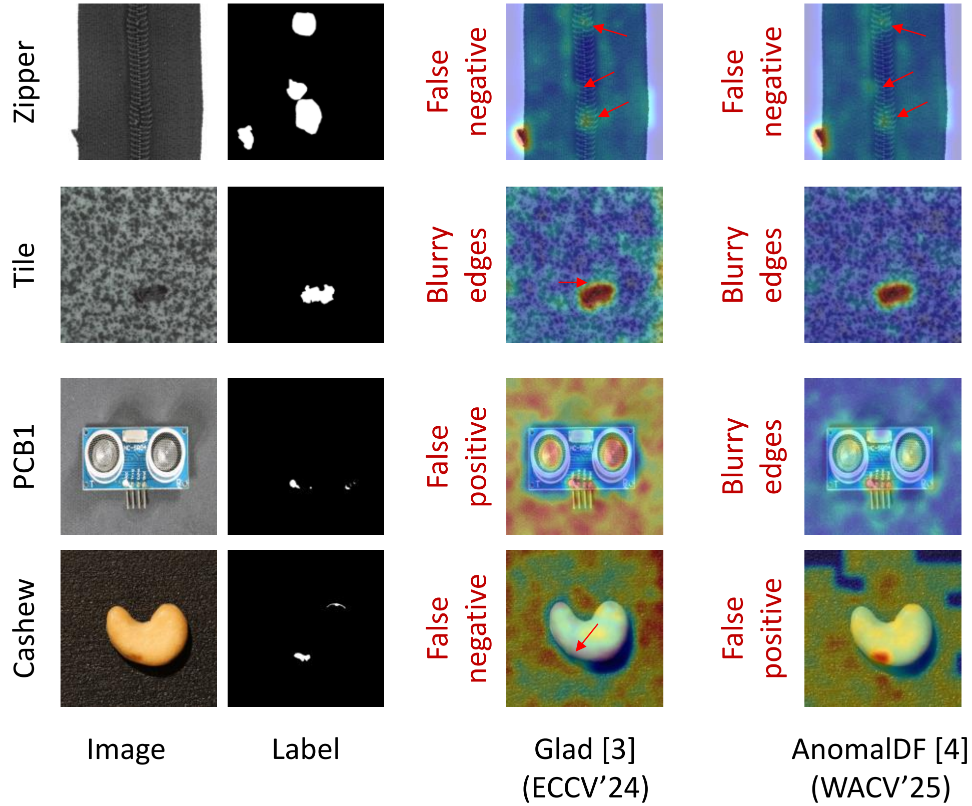
🔍 Existing UAD methods often emphasize *sample reconstruction, precise feature learning, or extensive feature banks*, whereas *we* study the UAD *from the perspective of matching*.

🔍 We find that matching noise often blurs the boundaries between normal and anomalous regions, which hampers anomaly detection accuracy, particularly for subtle anomalies.

[1] Zhao et al., OmniAL: A Unified CNN Framework for Unsupervised Anomaly Localization, CVPR 2023
[2] Guo et al., Dinomaly: The Less Is More Philosophy in Multi-Class Unsupervised Anomaly Detection, CVPR 2025.

# Introduction

## Background & Motivation: Matching Noise - Ubiquitous Yet Overlooked




Two Mainstream Methods: Reconstruction-based | Embedding-based

❗ Commonly, UAD relies on image- or feature-level matching, a process inherent to both reconstruction[3] - and embedding[4] -based methods.

❗ Such matching noise impairs the localization of subtle or boundary-adjacent anomalies.

❗ We address this via **cost volume filtering**, inspired by concepts in stereo and flow tasks.


[3] Yao et al., GLAD: Towards Better Reconstruction with Global and Local Adaptive Diffusion Models for Unsupervised Anomaly Detection, ECCV 2024
[4] Damm et al., AnomalyDINO: Boosting Patch-based Few-shot Anomaly Detection with DINOv2, WACV 2025.

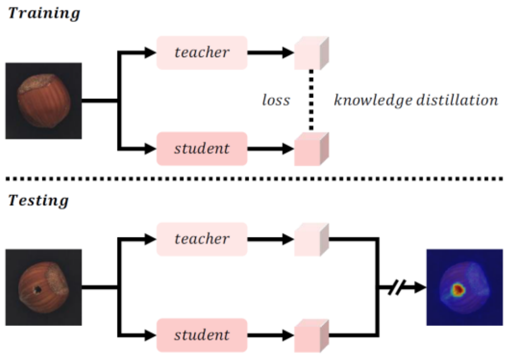
# Related work

## Unsupervised Anomaly Detection [5]

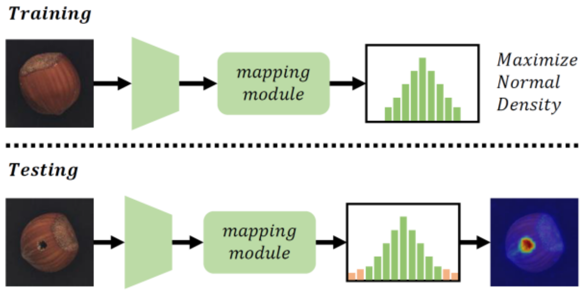
- **Embedding-based methods**
  Use pre-trained features to compare distributions,
  e.g., teacher–student networks, distribution modeling, memory banks.
- **Reconstruction-based methods**
  Rebuild normal patterns and detect anomalies via residuals,
  e.g., autoencoders, GANs, transformers, diffusion models, MoE.
- **Synthesis-based methods**
  Generate pseudo-anomalies to simulate real defects,
  e.g., pixel- or feature-level synthetic perturbations.

**Synthesis -based**

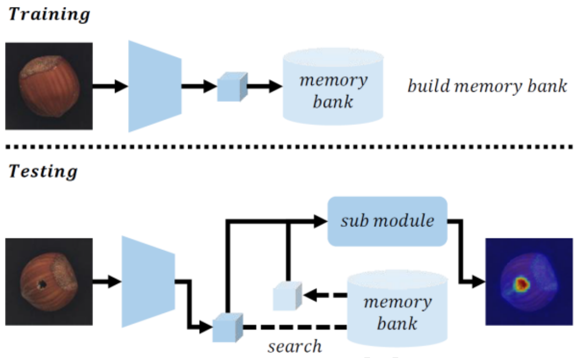
**Synthesis -based**




From paper [5]

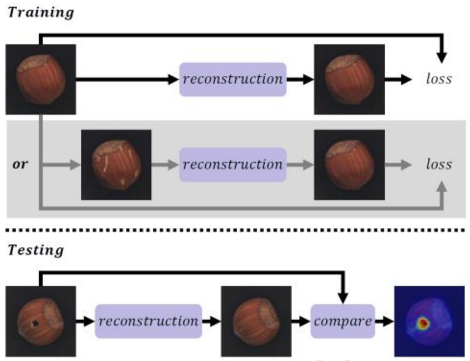



From paper [5]




From paper [5]

**Embedding-based**




From paper [5]

**Reconstruction-based**



[5] Lin Y et al., A survey on RGB, 3D, and multimodal approaches for unsupervised industrial image anomaly detection, Information Fusion, 2025.

Northeastern University

## Cost Volume Filtering in Vision Tasks

- **Stereo matching**

  Cost volumes correlate left and right image features along the disparity axis to capture pixel-wise similarity [6] [7] .

- **Depth estimation**

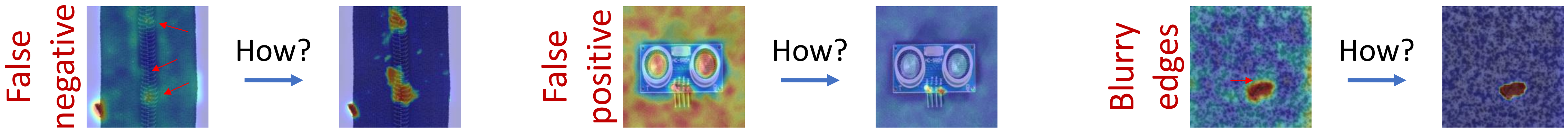
  Cost volumes model multi-view geometric relationships for precise depth estimation [8] [9] .

- **Motion analysis**

  Cost volumes refine pixel correspondences to improve optical flow accuracy [10] [11] .

[6] Kendall et al., End-to-End Learning of Geometry and Context for Deep Stereo Regression, ICCV 2017.
[7] Wang Y et al., Cost volume aggregation in stereo matching revisited: A disparity classification perspective, IEEE TIP 2024.
[8] Yang J et al., Self-supervised learning of depth inference for multi-view stereo, CVPR. 2021.
[9] Peng R et al., Rethinking depth estimation for multi-view stereo: A unified representation, CVPR. 2022.
[10] Zhang F et al., Separable flow: Learning motion cost volumes for optical flow estimation, ICCV 2021.
[11] Garrepalli R et al., Dift: Dynamic iterative field transforms for memory efficient optical flow, CVPR 2023.

# Challenges




- **Matching Noise vs. Fine Anomalies**

  Suppressing matching noise while preserving subtle anomaly cues.

- **Subtle and Edge-bound Defects**

  Low-contrast or boundary-adjacent anomalies are easily confused with normal regions.

- **Identical Shortcut in Reconstruction-based or Embedding-based methods**

  The "identical shortcut" effect always replicates anomalies, hindering residual-based detection.

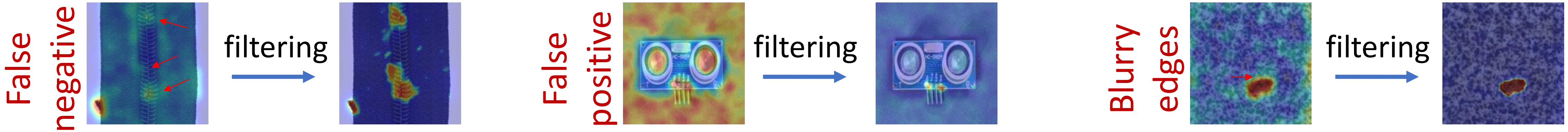
- **Category-wise Anomaly Diversity**

  Multi-class UAD must handle varying anomaly types across categories, increasing the complexity.

# Method



## Problem Reformulation




The task targets **image- and pixel-level anomaly detection** using only synthesized anomalies, without access to real defects during training.

**We reformulate multi-class UAD as a three-step process:**

1. **Feature extraction:** from input and template or reconstructed samples.
2. **Anomaly Cost Volume Construction:** modeling spatial anomaly patterns and channel-wise matching similarity.
3. **Cost Volume Filtering:** with dual-stream attention guidance for noise suppression and anomaly refinement.

# Method

## Our contribution

💡 **New Unsupervised Anomaly Detection Formulation**

We reinterpret anomaly detection as a cost filtering process to explicitly address matching noise.

🧩 **CostFilter-AD Method**

A plug-and-play filtering network guided by attention to refine cost volumes and suppress noise.

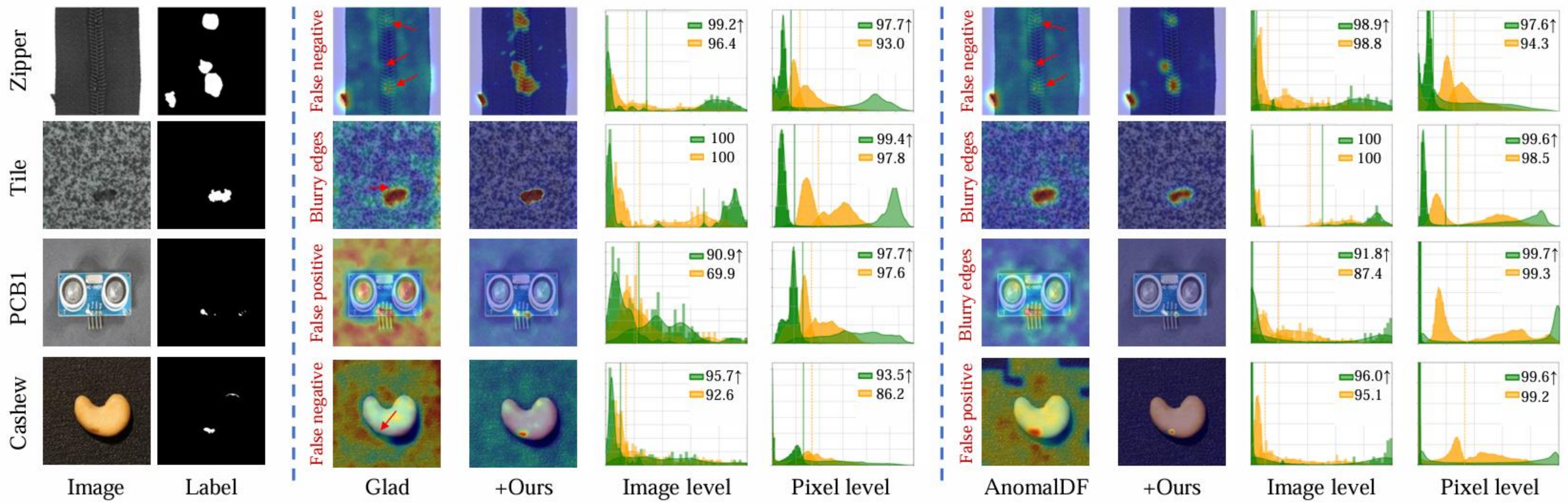
🧩 **Broad Compatibility**

Our method integrates seamlessly with both reconstruction- and embedding-based models.

🏆 **Strong Performance Gain**

We enhance 5 baselines across 7 metrics and achieve state-of-the-art results on 4 popular datasets.

## Analysis: From Heatmaps to Histograms -- Revealing Ubiquitous Matching Noise



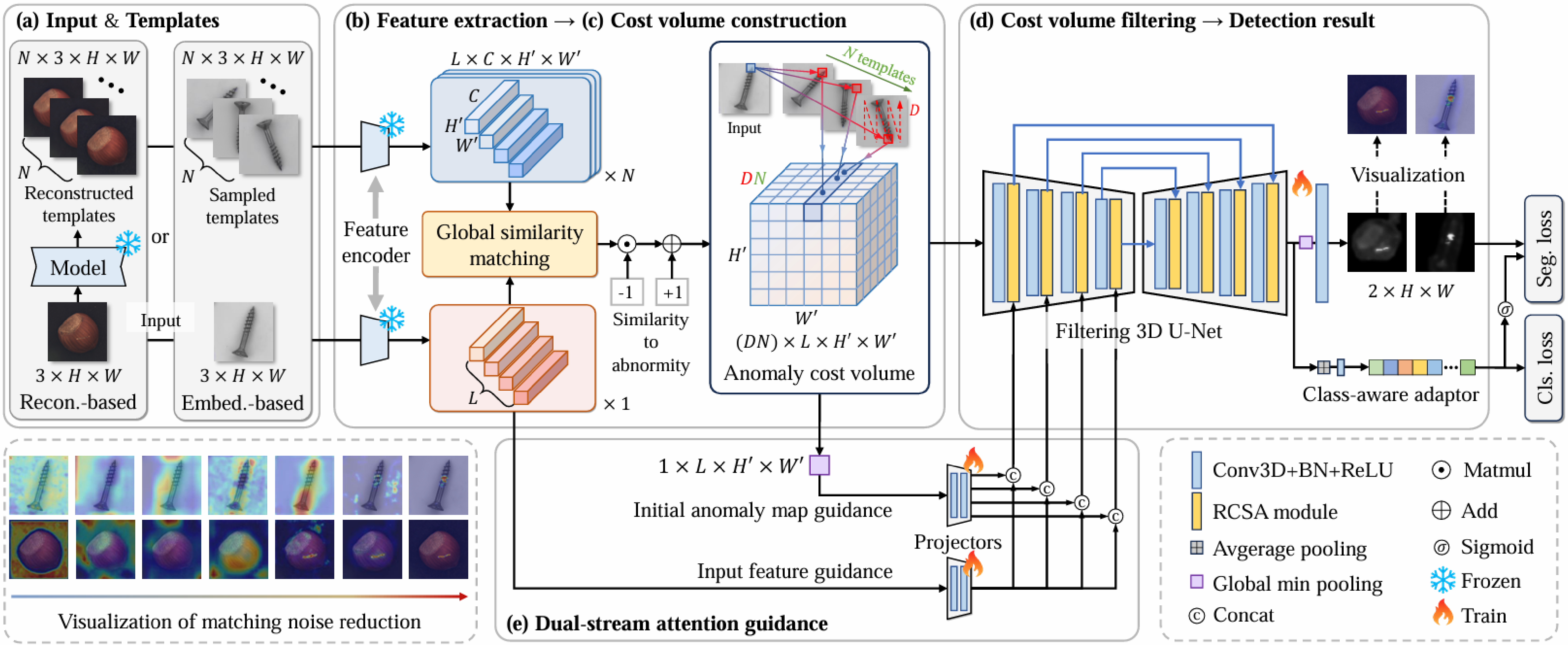

🔍 **Visualization and KDE curves** show image- and pixel-level logits.

🟡 **Baseline results** are highlighted in yellow, 🟢 **ours** in green.

✨ **Our method** yields less noisy detections and clearer normal–abnormal separation.

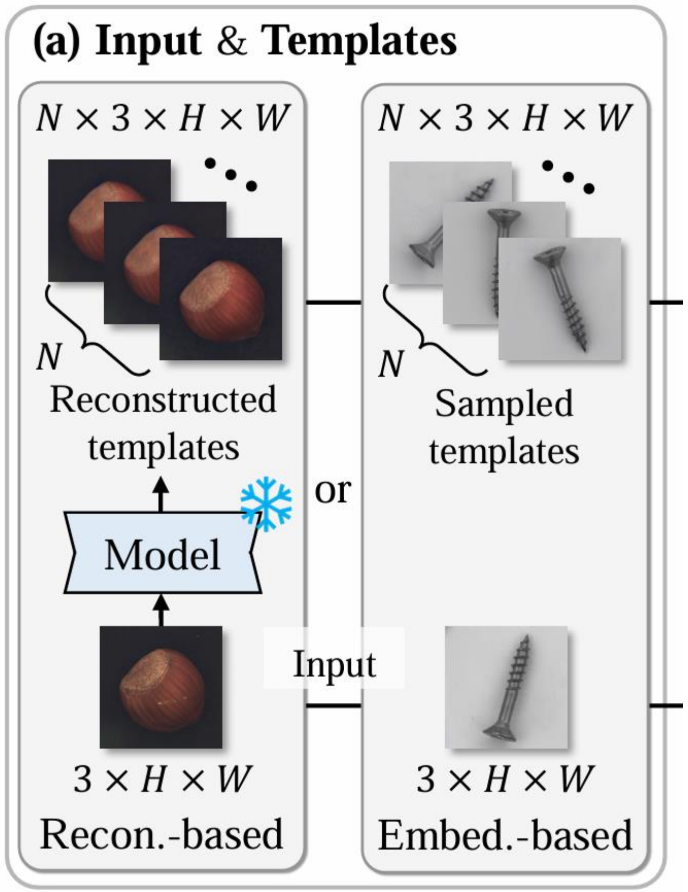
# Method

## Overview: Architecture of Costfilter-AD




1. Feature Extraction  2. Anomaly Cost Volume Construction  3. Cost Volume Filtering

## Image & Templates in CostFilter-AD

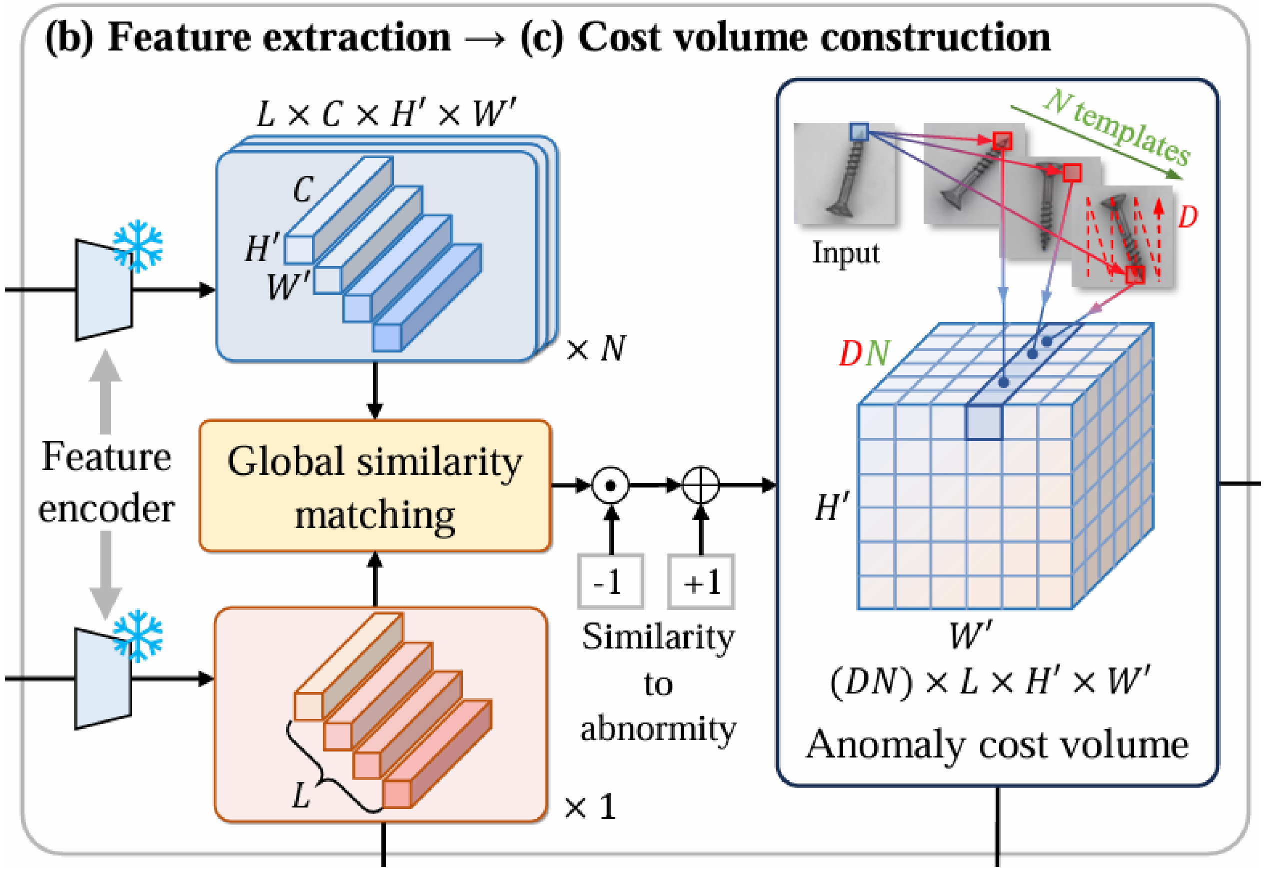



- **Reconstruction-based (e.g., HVQ-Trans, GLAD)**
  - **Image**: Original input image
  - **Template**: Reconstructed normal image from model
    - -*HVQ-Trans*: Multi-scale features via vector quantization (N = 1)
    - -*GLAD:* Multi-step reconstruction via adaptive diffusion

      (1 ≤ N ≤ total steps) $I_{t\to 0} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(I_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(I_t, t)\right)$

- **Embedding-based (e.g., AnomalDF)**
  - **Image**: Features from pre-trained encoder
  - **Template**: Normal features from memory bank
    - –*AnomalDF:* Randomly sampled normal features (N = 3)

## Extract Features & Construct Anomaly Cost Volume



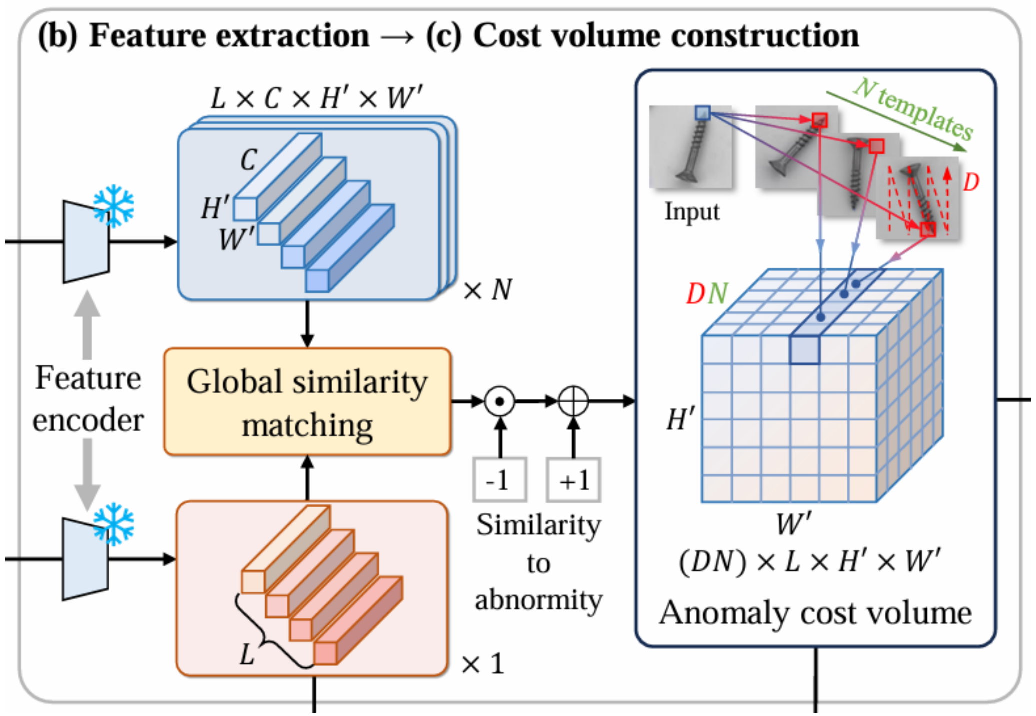

For reconstruction- and embedding-based piplines, we perform global spatial matching over input and template features:

$$\mathcal{V}(j,n,l,i) = \frac{f_{\mathcal{S}}^{i,l} \cdot f_{\mathcal{T}}^{n,j,l}}{\|f_{\mathcal{S}}^{i,l}\| \cdot \|f_{\mathcal{T}}^{n,j,l}\|},$$

Lower similarity implies higher anomaly likelihood, thus forming the anomaly cost volume.

$$\mathcal{C}(j,n,l,i) = 1 - \mathcal{V}(j,n,l,i)$$

## Extract Features & Construct Anomaly Cost Volume

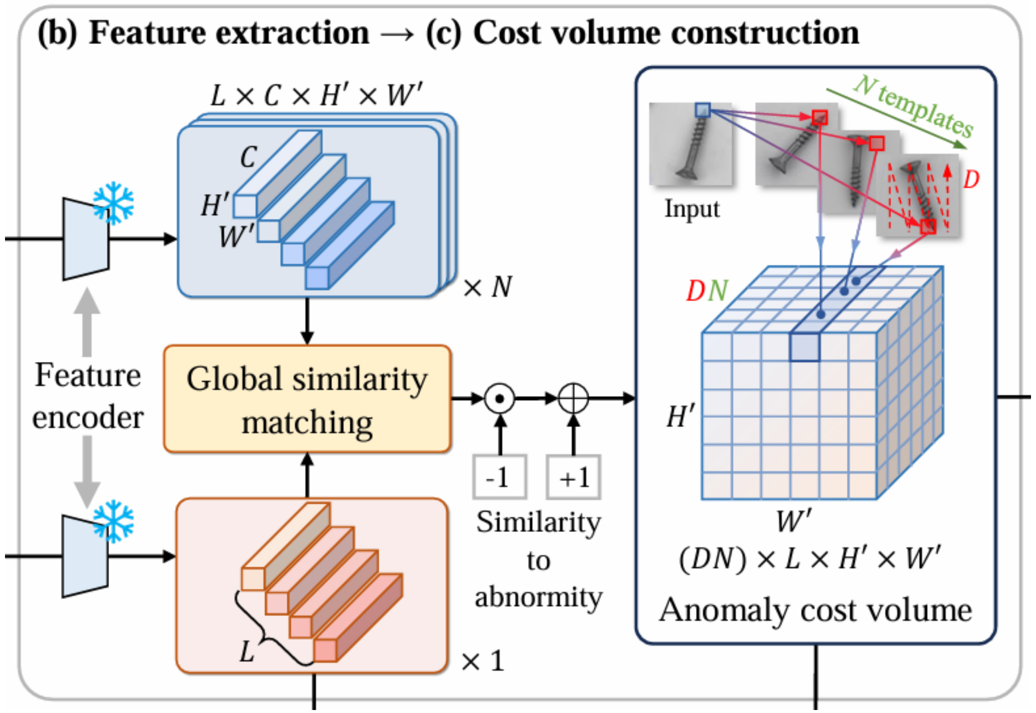



### Notations & Dimensions

- $f_i^l \in \mathbb{R}^C$: Feature vector at spatial index $i$ from the input image at layer $l \in \{1, 2, \ldots, L\}$
- $f_{n,j,T}^l \in \mathbb{R}^C$: Feature vector at spatial index $j$ of the $n$-th template at layer $l$
- $V \in \mathbb{R}^{D \times N \times L \times (H'W')}$: Similarity volume
  - $D = H' \times W'$: matching dimension (from template features)
  - $N$: number of templates
  - $L$: number of layers
  - $H'W'$: flattened spatial positions of the input
- $C \in \mathbb{R}^{(DN) \times L \times H' \times W'}$: Anomaly cost volume (after merging D and N, and reshaping)
- $\bar{M} \in \mathbb{R}^{L \times H' \times W'}$: Initial multi-layer anomaly map from global min-pooling over matching dimension

# Method

## Extract Features & Construct Anomaly Cost Volume




🔍 **Physical Meaning in Anomaly Detection**

- **Matching Dimension (DN):**

  Represents *what to match* — all candidate positions in templates for similarity comparison.

- **Spatial Dimension (H′ × W′):**

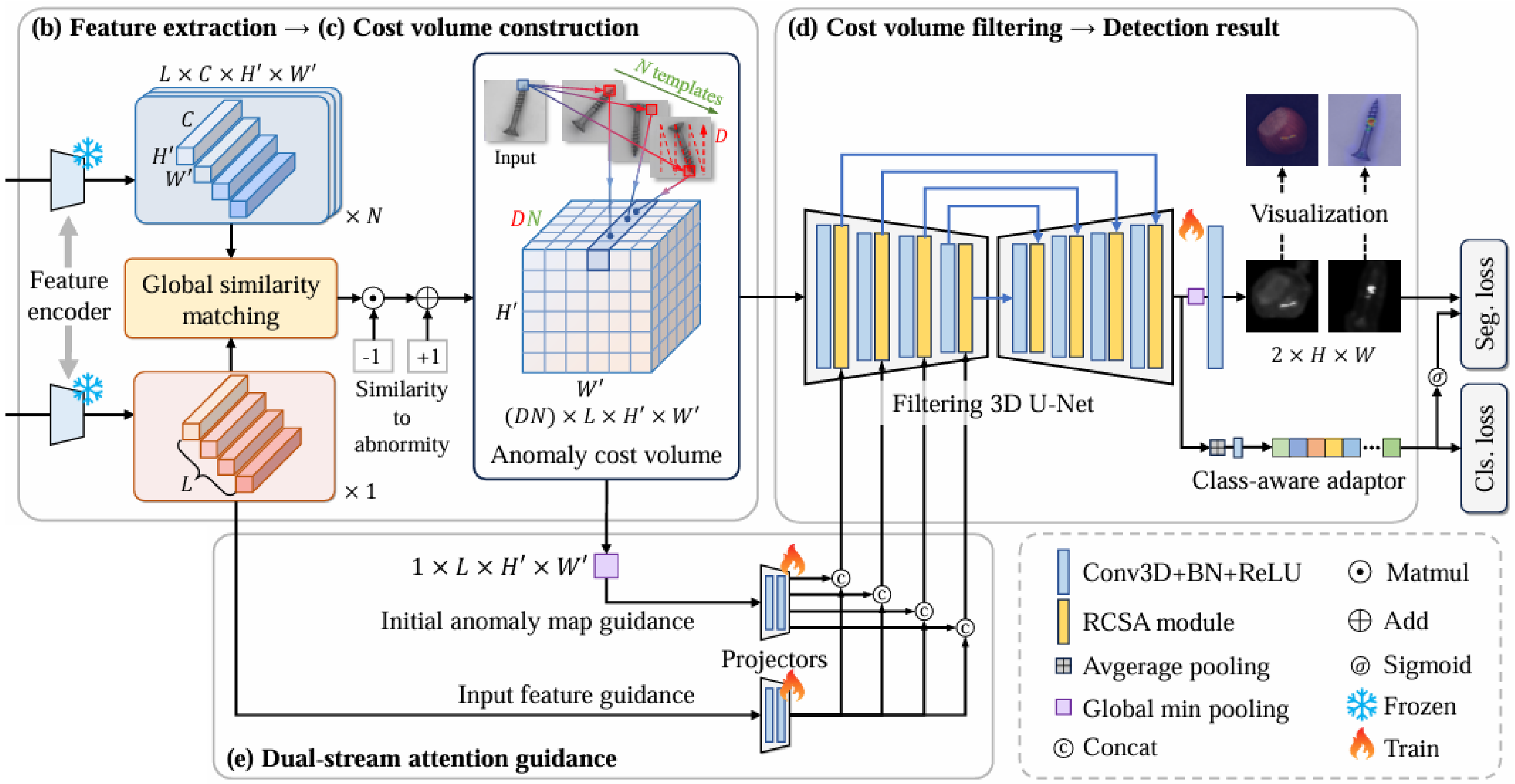
  Represents *where to detect* — pixel locations in the input image being evaluated.

- **Depth Dimension (L)**

  Represents *how to represent* — multi-level features from different encoder layers.

# Method

## Cost volume filtering & Anomaly Output Generation




### Network Input

Combines the anomaly cost volume, input features, and initial anomaly map as inputs to the 3D U-Net.

### Dual-Stream Attention Guidance

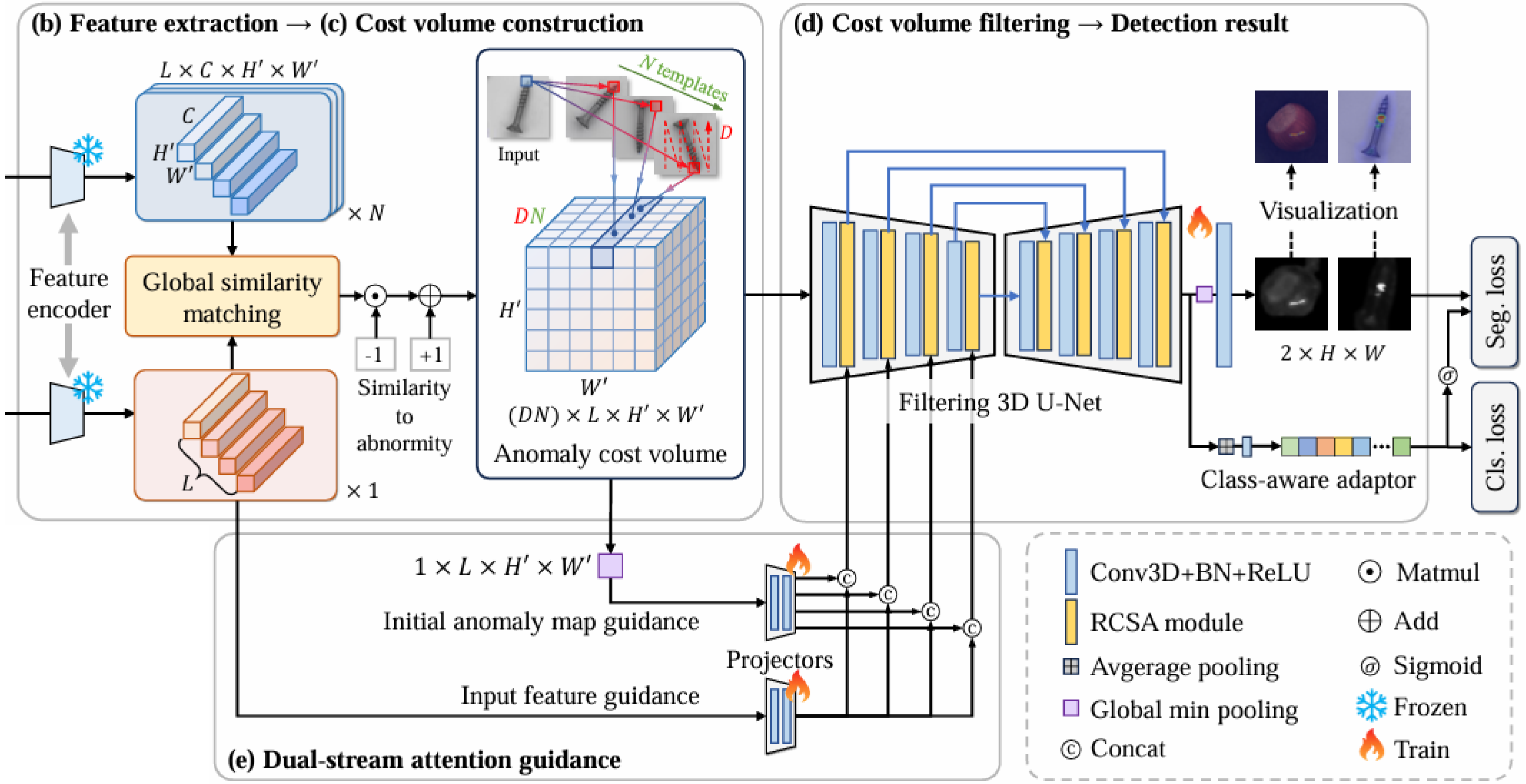
**1. Spatial Guidance (SG)**: Preserves fine details using input features

**2. Matching Guidance (MG)**: Focuses attention using initial anomaly maps

**3. Both are fused with U-Net features:** via residual channel-spatial attention for robust refinement.

Northeastern University

## Cost volume filtering & Anomaly Output Generation

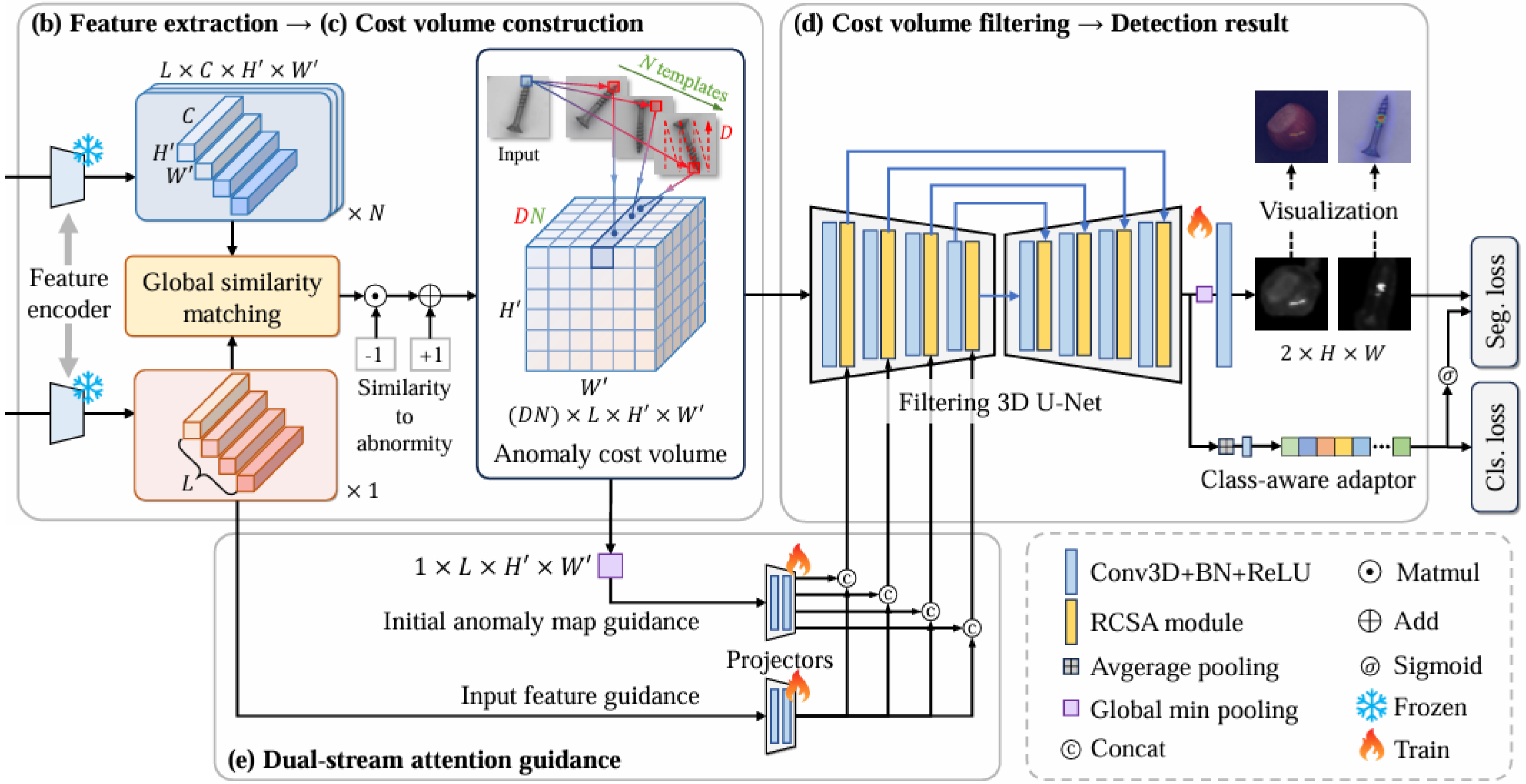



🌈 **Filtering Network Architecture**
Uses RCSA modules with **residual connections**, **3D convolutions**, and **dual attention** to enhance filtering across feature layers.

🧬 **Class-Aware Adaptor**
Learns class-aware soft logits via spatially pooled features, guiding the segmentation loss to improve detection across diverse anomalies.

Northeastern University

## Cost volume filtering & Anomaly Output Generation



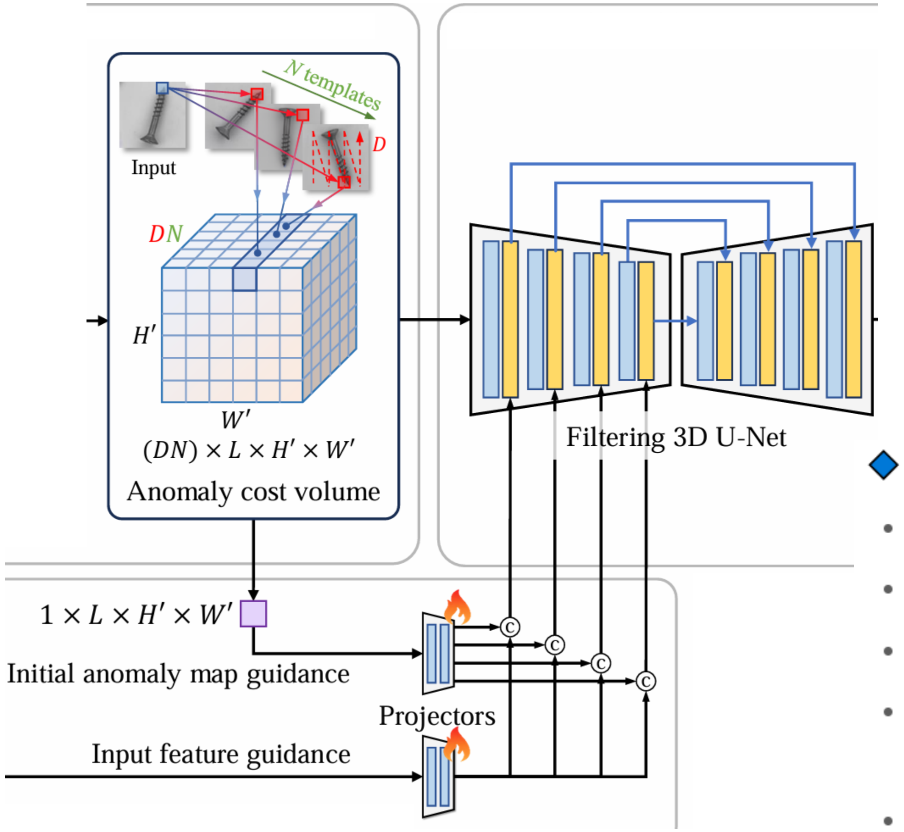

### Anomaly Output Generation

Performs global min-pooling → convolution → softmax

Outputs:

- **Pixel-level anomaly map** $M \in \mathbb{R}^{H' \times W'}$
- **Image-level score** from the average of top-250 values in the map

# Method




## Notation Explanation for filtering Network

$$x'_l = \text{cat}(x_l, h(\bar{M}), h(f_s^l)),$$

$$x_l^{ca} = \sigma\left(\text{conv}(\text{MP}(x'_l)) + \text{conv}(\text{AP}(x'_l))\right) * x'_l + x'_l,$$

$$x_l^{sa} = \sigma\left(\text{conv}(\text{cat}(\mu(x_l^{ca}), \max(x_l^{ca})))\right) * x_l^{ca} + x_l^{ca},$$

### Input & Intermediate Variables

- $x_l$ Anomaly cost volume feature at layer *l*
- $\bar{M}$ Initial anomaly map (guidance signal)
- $f_s^l$ Input image feature at layer *l*
- $x'_l$ Concatenated feature: $x'_l = \text{cat}(x_l, h(\bar{M}), h(f_s^l))$
- $x_l^{ca}$ Channel-attended feature
- $x_l^{sa}$ Spatial-attended feature (RCSA output)

### Functions & Operators

- $h(\cdot)$ Guidance projector (adjusts channel & resolution)
- **cat(•)** Concatenation along channel dimension
- **conv(•)** 3D convolution
- $\sigma(\cdot)$ Sigmoid activation
- **MP(•)** Global Max Pooling (spatial)
- **AP(•)** Global Average Pooling (spatial)
- $\mu(\cdot)$ Channel-wise mean
- **max(•)** Channel-wise max
- $*$ Element-wise multiplication
- $+$ Residual addition (skip connection)

# Method

## Training Procedure

**Plug-in Design**

Used as a generic plug-in for both reconstruction-based and embedding-based methods.

**Anomaly Cost Volume Construction**

Matching between input image features and:

- Reconstructed outputs (reconstruction-based), or
- Randomly sampled normal templates (embedding-based).

**Supervised Learning Objective**

Trained as a **normal-vs-anomaly segmentation** task using synthesized masks $M_s$.

**Loss Function**

$$L = \mathrm{Focal}(M, M_s, \sigma(\hat{Y}_c)) + \mathrm{CE}(\hat{Y}_c, Y) + \alpha \cdot (\text{Soft-IoU}(M, M_s) + \mathrm{SSIM}(M, M_s))$$

- *Focal Loss*: Handles foreground–background imbalance
- *Soft-IoU*: Sharpens anomaly boundary localization
- *SSIM*: Preserves structural consistency
- *Cross-Entropy*: For multi-class classification

**Class-Aware γ Modulation**

If the adaptor predicts correctly:

$$\gamma = \gamma_0 - \sigma(\hat{Y}_c)$$

Otherwise:

$$\gamma = \gamma_0$$

## Inference Procedure

**Matching & Filtering**

Construct cost volume and apply the filtering network as in training.

**Final Anomaly Map Generation**

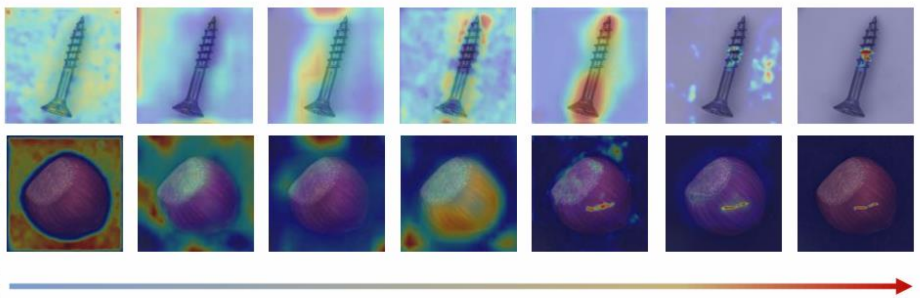
Produces refined anomaly score map $M$.

**Fusion with Baseline**

Anomaly map blended with baseline output:

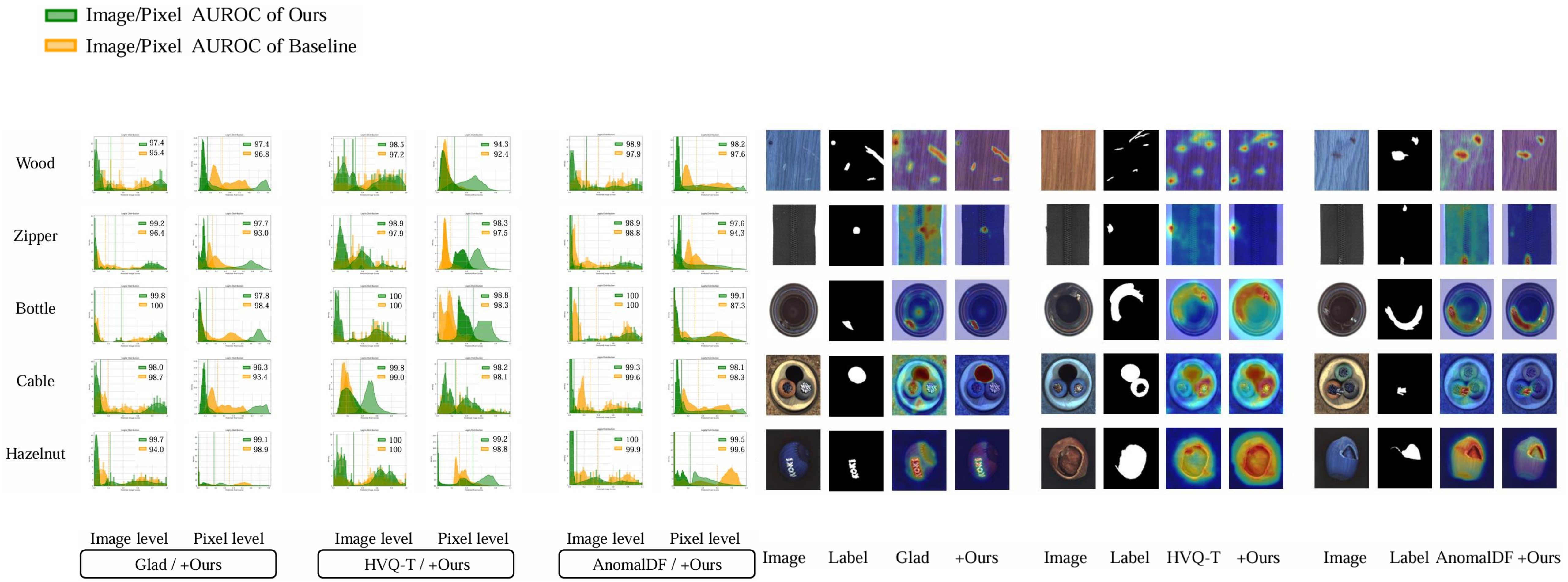
$$M_{\text{final}} = \lambda \cdot M + (1 - \lambda) \cdot M_{\text{baseline}}, \quad \lambda \in [0, 1]$$

→ Compensates for scale differences between components

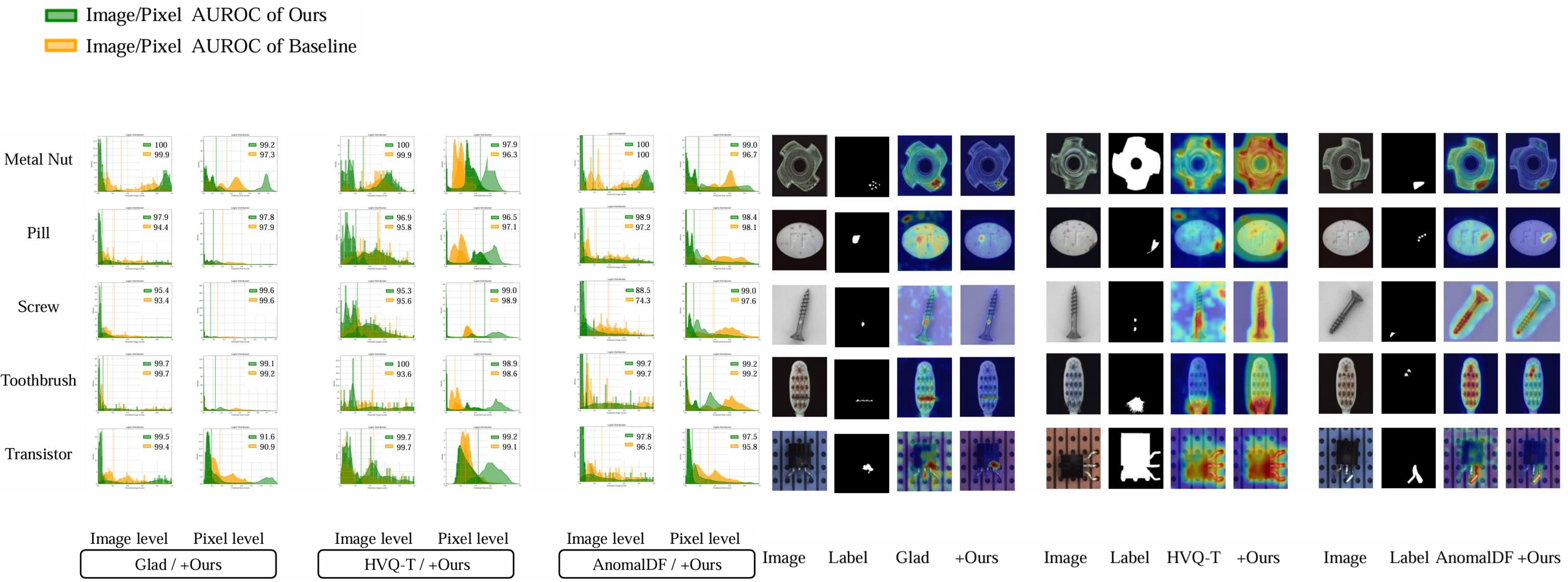


Visualization of matching noise reduction

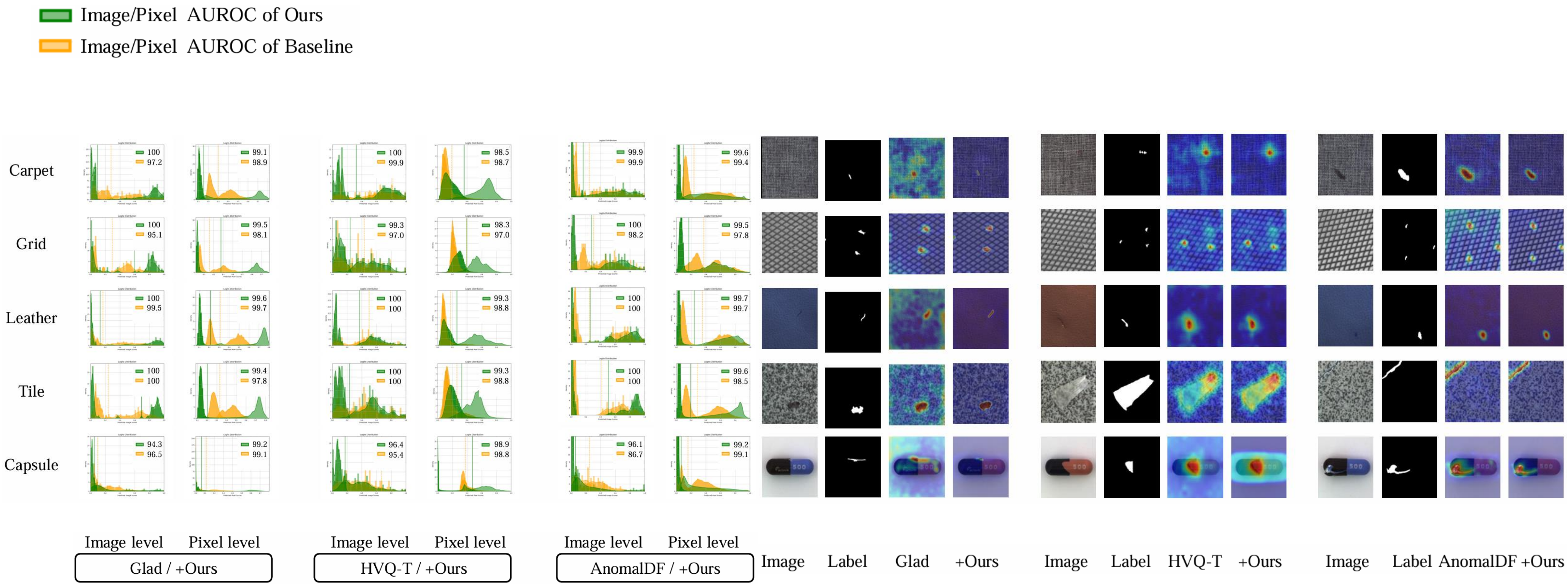
# Evaluation: qualitative results on **Mvtec-AD**



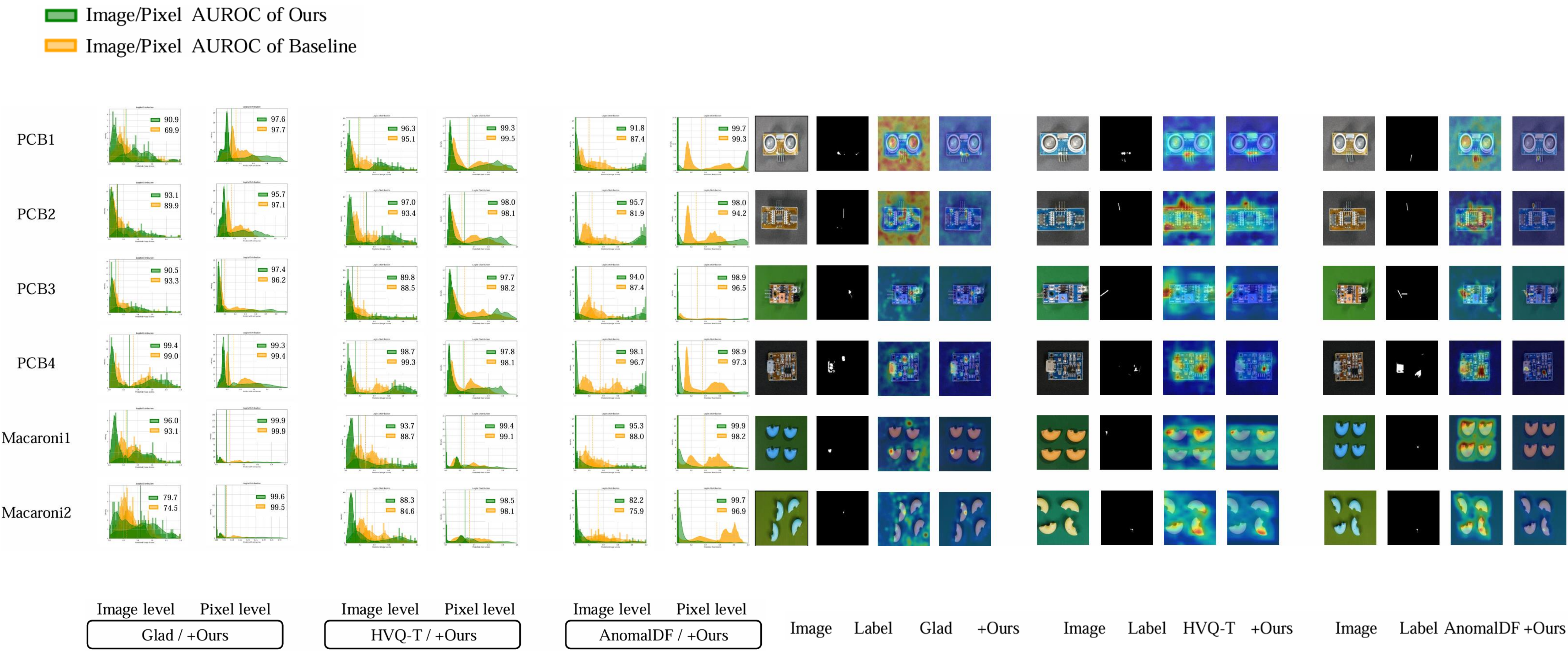
# Evaluation: qualitative results on **Mvtec-AD**



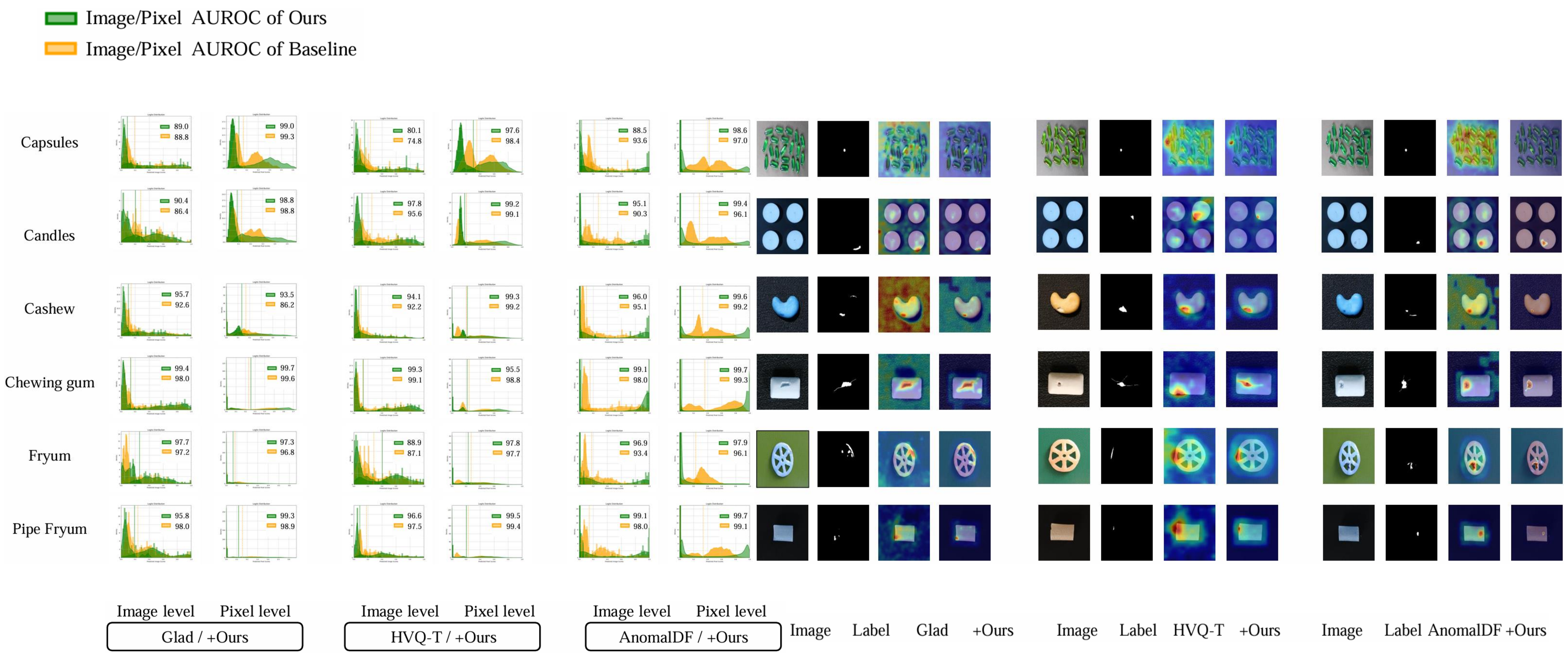
# Evaluation: qualitative results on **Mvtec-AD**



# Evaluation: qualitative results on VisA



# Evaluation: qualitative results on VisA



# Evaluation: quantitative results

## Plug-and-Play Boosting of Multi-class UAD on **Mvtec-AD**

*Table 1.* Multi-class anomaly detection/localization results (image AUROC/pixel AUROC) on MVTec-AD. Models are evaluated across all categories without fine-tuning, with the best results highlighted in bold.

| | Category | PatchCore | OmniAL | DiAD | VPDM | MambaAD | GLAD | GLAD+Ours | HVQ-Trans | HVQ-Trans+Ours | AnomalDF | AnomalDF+Ours |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Object | Bottle | **100** / **99.2** | **100** / **99.2** | 99.7 / 98.4 | **100** / 98.6 | **100** / 98.7 | **100** / 98.4 | 99.8 / 97.8 | **100** / 98.3 | **100** / 98.8 | **100** / 87.3 | **100** / 99.1 |
| | Cable | 95.3 / 93.6 | 98.2 / 97.3 | 94.8 / 96.8 | 97.8 / 98.1 | 98.8 / 95.8 | 98.7 / 93.4 | 98.0 / 96.3 | 99.0 / 98.1 | **99.8** / 98.2 | 99.6 / **98.3** | 99.3 / 98.1 |
| | Capsule | 96.8 / 98.0 | 95.2 / 96.9 | 89.0 / 97.1 | **97.0** / 98.8 | 94.4 / 98.4 | 96.5 / 99.1 | 94.3 / **99.2** | 95.4 / 98.8 | 96.4 / 98.9 | 89.7 / 99.1 | 96.1 / **99.2** |
| | Hazelnut | 99.3 / 97.6 | 95.6 / 98.4 | 99.5 / 98.3 | 99.9 / 98.7 | **100** / 99.0 | 97.0 / 98.9 | 99.4 / 99.1 | **100** / 98.8 | **100** / 99.2 | 99.9 / **99.6** | **100** / 99.5 |
| | Metal Nut | 99.1 / 96.3 | 99.2 / 99.1 | 99.1 / 97.3 | 98.9 / 96.0 | 99.9 / 96.7 | 99.9 / 97.3 | **100** / **99.2** | 99.9 / 96.3 | **100** / 97.9 | **100** / 96.7 | **100** / 99.0 |
| | Pill | 86.4 / 90.8 | 97.2 / **98.9** | 95.7 / 95.7 | 97.9 / 96.4 | 97.0 / 97.4 | 94.4 / 97.9 | 97.9 / 97.8 | 95.8 / 97.1 | 96.9 / 96.5 | 97.2 / 98.1 | **98.9** / 98.4 |
| | Screw | 94.2 / 98.9 | 88.0 / 98.0 | 90.7 / 97.9 | 95.5 / 99.3 | 94.7 / 99.5 | 93.4 / **99.6** | 95.4 / **99.6** | **95.6** / 98.9 | 95.3 / 99.0 | 74.3 / 97.6 | 88.5 / 99.0 |
| | Toothbrush | **100** / 98.8 | **100** / 99.0 | 99.7 / 99.0 | 94.6 / 98.8 | 98.3 / 99.0 | 99.7 / **99.2** | 99.7 / 99.1 | 93.6 / 98.6 | **100** / 98.9 | 99.7 / **99.2** | 99.7 / **99.2** |
| | Transistor | 98.9 / 92.3 | 93.8 / 93.3 | 99.8 / 95.1 | 99.7 / 97.9 | **100** / 97.1 | 99.4 / 90.9 | 99.5 / 91.6 | 99.7 / 99.1 | 99.7 / **99.2** | 96.5 / 95.8 | 97.8 / 97.5 |
| | Zipper | 97.1 / 95.7 | **100** / **99.5** | 95.1 / 96.2 | 99.0 / 98.0 | 99.3 / 98.4 | 96.4 / 93.0 | 99.2 / 97.7 | 97.9 / 97.5 | 98.9 / 98.3 | 98.8 / 94.3 | 98.9 / 96.7 |
| Texture | Carpet | 97.0 / 98.1 | 98.7 / 99.4 | 99.4 / 98.6 | **100** / 98.8 | 99.8 / 99.2 | 97.2 / 98.9 | **100** / 99.1 | 99.9 / 98.7 | **100** / 98.5 | 99.9 / 99.4 | 99.9 / **99.6** |
| | Grid | 91.4 / 98.4 | 99.9 / 99.4 | 98.5 / 96.6 | 98.6 / 98.0 | **100** / 99.2 | 95.1 / 98.1 | **100** / **99.5** | 97.0 / 97.0 | 99.3 / 98.3 | 98.2 / 97.8 | **100** / **99.5** |
| | Leather | **100** / 99.2 | 99.0 / 99.3 | 99.8 / 98.8 | **100** / 99.2 | **100** / 99.4 | 99.5 / **99.7** | **100** / 99.6 | **100** / 98.8 | **100** / 99.3 | **100** / **99.7** | **100** / **99.7** |
| | Tile | 96.0 / 90.3 | 99.6 / 99.0 | 96.8 / 92.4 | **100** / 94.5 | 98.2 / 93.8 | **100** / 97.8 | **100** / 99.4 | 99.2 / 92.2 | **100** / 95.0 | **100** / 98.5 | **100** / **99.6** |
| | Wood | 93.8 / 90.8 | 93.2 / 97.4 | **99.7** / 93.3 | 98.2 / 95.3 | 98.8 / 94.4 | 95.4 / 96.8 | 97.4 / 97.4 | 97.2 / 92.4 | 98.5 / 94.3 | 97.9 / 97.6 | 98.9 / **98.2** |
| | Mean | 96.4 / 95.7 | 97.2 / 98.3 | 97.2 / 96.8 | 98.4 / 97.8 | 98.6 / 97.7 | 97.5 / 97.3 | 98.7 / 98.2 | 98.0 / 97.3 | **99.0** / 98.0 | 96.8 / 98.1 | 98.5 / **98.8** |

✅ Our method consistently improves image- and pixel-level AUROC, outperforming GLAD, HVQ-Trans, and AnomalDF across benchmarks.

# Evaluation: quantitative results

Northeastern University

## Plug-and-Play Boosting of Multi-class UAD on **VisA**

*Table 2.* Multi-class anomaly detection/localization results (image AUROC/pixel AUROC) on VisA. Models are evaluated across all categories without fine-tuning, with the best results highlighted in bold.

| | Category | JNLD | OmniAL | DiAD | VPDM | MambaAD | GLAD | GLAD+Ours | HVQ-Trans | HVQ-Trans+Ours | AnomalDF | AnomalDF+Ours |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Complex Structure | PCB1 | 82.9 / 98.9 | 77.7 / 97.6 | 88.1 / 98.7 | **98.2** / 99.6 | 95.4 / **99.8** | 69.9 / 97.6 | 90.9 / 97.7 | 95.1 / 99.5 | 96.3 / 99.3 | 87.4 / 99.3 | 91.8 / 99.7 |
| | PCB2 | 79.1 / 95.0 | 81.0 / 93.9 | 91.4 / 95.2 | **97.5** / 98.8 | 94.2 / **98.9** | 89.9 / 97.1 | 93.2 / 95.7 | 93.4 / 98.1 | 97.0 / 98.0 | 81.9 / 94.2 | 95.7 / 98.0 |
| | PCB3 | 90.1 / 98.5 | 88.1 / 94.7 | 86.2 / 96.7 | **94.5** / 98.7 | 93.7 / **99.1** | 93.3 / 96.2 | 90.5 / 97.4 | 88.5 / 98.2 | 89.8 / 97.7 | 87.4 / 96.5 | 94.0 / 98.9 |
| | PCB4 | 96.2 / 97.5 | 95.3 / 97.1 | 99.6 / 97.0 | **99.9** / 97.8 | **99.9** / 98.6 | 99.0 / **99.4** | 99.4 / 99.3 | 99.3 / 98.1 | 98.7 / 97.8 | 96.7 / 97.3 | 98.1 / 98.9 |
| Multiple Instances | Macaroni1 | 90.5 / 93.3 | 92.6 / 98.6 | 85.7 / 94.1 | **97.5** / 99.6 | 91.6 / 99.5 | 93.1 / **99.9** | 96.0 / **99.9** | 88.7 / 99.1 | 93.7 / 99.4 | 88.0 / 98.2 | 95.3 / **99.9** |
| | Macaroni2 | 71.3 / 92.1 | 75.2 / 97.9 | 62.5 / 93.6 | 85.7 / 99.0 | 81.6 / 99.5 | 74.5 / 99.5 | 79.7 / 99.6 | 84.6 / 98.1 | **88.3** / 98.5 | 75.9 / 96.9 | 82.2 / **99.7** |
| | Capsules | 91.4 / **99.6** | 90.6 / 99.4 | 58.2 / 97.3 | 79.5 / 99.1 | 91.8 / 99.1 | 88.8 / 99.3 | 89.1 / 99.0 | 74.8 / 98.4 | 80.1 / 97.6 | **93.6** / 97.0 | 88.5 / 98.6 |
| | Candles | 85.4 / 94.5 | 86.8 / 95.8 | 92.8 / 97.3 | 97.2 / **99.4** | 96.8 / 99.0 | 86.4 / 98.8 | 90.5 / 98.8 | 95.6 / 99.1 | **97.8** / 99.2 | 90.3 / 96.1 | 95.1 / **99.4** |
| Single Instance | Cashew | 82.5 / 94.1 | 88.6 / 95.0 | 91.5 / 90.9 | 90.0 / 98.0 | 94.5 / 94.3 | 92.6 / 86.2 | 95.7 / 93.5 | 92.2 / 98.7 | 94.1 / 99.3 | 95.1 / 99.2 | **96.0** / **99.6** |
| | Chewing gum | 96.0 / 98.9 | 96.4 / 99.0 | 99.1 / 94.7 | 99.0 / 98.6 | 97.7 / 98.1 | 98.0 / 99.6 | **99.4** / **99.7** | 99.1 / 98.1 | 99.3 / 99.5 | 98.0 / 99.3 | 99.1 / **99.7** |
| | Fryum | 91.9 / 90.0 | 94.6 / 92.1 | 89.8 / 97.6 | 92.0 / **98.6** | 95.2 / 96.9 | 97.2 / 96.8 | **97.7** / 97.3 | 87.1 / 97.7 | 88.9 / 97.8 | 93.4 / 96.1 | 96.9 / 97.9 |
| | Pipe Fryum | 87.5 / 92.5 | 86.1 / 98.2 | 96.2 / 99.4 | 98.8 / 99.4 | 98.7 / 99.1 | 98.0 / 98.9 | 95.8 / 99.3 | 97.5 / 99.4 | 96.6 / 99.5 | 98.0 / 99.1 | **99.1** / **99.7** |
| | Mean | 87.1 / 95.2 | 87.8 / 96.6 | 86.8 / 96.0 | 94.2 / 98.9 | **94.3** / 98.5 | 90.1 / 97.4 | 93.2 / 98.1 | 91.3 / 98.5 | 93.4 / 98.6 | 90.5 / 97.5 | **94.3** / **99.2** |

✅ Our method consistently improves image- and pixel-level AUROC, outperforming GLAD, HVQ-Trans, and AnomalDF across benchmarks.

# Evaluation: quantitative results

Northeastern University

## Class-Aware Average Results Across **More** Datasets and Metrics

*Table 1.* Multi-class UAD evaluation on MVTec-AD and MPDD, reporting category-wise mean results for each benchmark.

| Benchmark | Method | Image-level | | | Pixel-level | | | |
|---|---|---|---|---|---|---|---|---|
| | | AU-ROC | AP | F1max | AU-ROC | AP | F1max | AUPRO |
| MVTec-AD | UniAD (NeurIPS'22) | 97.5 | 99.1 | 97.0 | 96.9 | 44.5 | 50.5 | 90.6 |
| | UniAD+Ours | **99.0** | **99.7** | **98.1** | **97.5** | **60.5** | **59.9** | **91.3** |
| | HVQ-Trans (NeurIPS'23) | 97.9 | 99.3 | 97.4 | 97.4 | 49.4 | 54.3 | 91.5 |
| | HVQ-Tran+Ours | **99** | **99.7** | **98.6** | **97.9** | **58.1** | **61.2** | **93.2** |
| | Glad (ECCV'24) | 97.5 | 98.8 | 96.8 | 97.3 | 58.8 | 59.7 | 92.8 |
| | Glad+Ours | **98.7** | **99.6** | **97.8** | **98.2** | **66.8** | **64.4** | **94.1** |
| | AnomalDF (WACV'25) | 96.8 | 98.6 | 97.1 | 98.1 | 61.3 | 60.8 | 93.6 |
| | AnomalDF+Ours | **98.5** | **99.4** | **97.8** | **98.8** | **67.8** | **64.9** | **94.1** |
| | Dinomaly (CVPR'25) | 99.6 | 99.8 | 99.0 | 98.3 | 68.7 | 68.7 | 94.6 |
| | Dinomaly+Ours | **99.7** | 99.8 | **99.1** | **98.4** | **68.9** | **68.9** | **94.8** |
| MPDD | HVQ-Trans (NeurIPS'23) | 86.5 | 87.9 | 85.6 | 96.9 | 26.4 | 30.5 | 88.0 |
| | HVQ-Tran+Ours | **93.1** | **95.4** | **90.3** | **97.5** | **34.1** | **37.0** | 82.9 |
| | Dinomaly (CVPR'25) | 97.3 | 98.5 | 95.6 | 99.1 | 60.0 | 59.8 | 96.7 |
| | Dinomaly+Ours | **97.5** | 98.5 | **95.8** | **99.2** | **60.2** | **59.9** | 96.7 |

*Table 2.* Multi-class UAD evaluation on VisA and BTAD, reporting category-wise mean results for each benchmark.

| Benchmark | Method | Image-level | | | Pixel-level | | | |
|---|---|---|---|---|---|---|---|---|
| | | AU-ROC | AP | F1max | AU-ROC | AP | F1max | AUPRO |
| VisA | UniAD (NeurIPS'22) | 91.5 | 93.6 | 88.5 | 98.0 | 32.7 | 38.4 | 76.1 |
| | UniAD+Ours | **92.1** | **94.0** | **88.9** | **98.6** | **34.0** | **39.0** | **86.4** |
| | HVQ-Trans (NeurIPS'23) | 91.5 | 93.4 | 88.1 | 98.5 | 35.5 | 39.6 | 86.4 |
| | HVQ-Tran+Ours | **93.4** | **95.2** | **89.3** | **98.6** | **41.4** | **45.0** | **86.8** |
| | Glad (ECCV'24) | 90.1 | 91.4 | 86.7 | 97.4 | 33.9 | 39.4 | 91.5 |
| | Glad+Ours | **93.2** | **94.1** | **89.2** | **98.1** | **40.7** | **43.7** | 91.5 |
| | AnomalDF (WACV'25) | 90.5 | 91.4 | 86.2 | 97.4 | 39.6 | 40.4 | 86.3 |
| | AnomalDF+Ours | **94.3** | **95.1** | **90.6** | **99.2** | **44.6** | **45.5** | 86.3 |
| | Dinomaly (CVPR'25) | 98.7 | 98.9 | 96.1 | 98.7 | 52.5 | 55.4 | 94.5 |
| | Dinomaly+Ours | 98.7 | **99.0** | **96.3** | **98.8** | **53.2** | **55.8** | **94.7** |
| BTAD | HVQ-Trans (NeurIPS'23) | 90.9 | 97.8 | 94.8 | 96.7 | 43.2 | 48.7 | 75.6 |
| | HVQ-Tran+Ours | **93.3** | **98.6** | **96.0** | **97.3** | **47.0** | **50.2** | **76.2** |
| | Dinomaly (CVPR'25) | 95.4 | 98.5 | 95.5 | 97.9 | 70.1 | 68.0 | 76.5 |
| | Dinomaly+Ours | **95.5** | **98.6** | **95.8** | **98.1** | **74.3** | **69.8** | **77.5** |

✅ **Our method consistently boosts multi-class UAD performance across diverse baselines and datasets by effectively filtering matching noise and preserving subtle anomaly details.**

# Evaluation: quantitative results

## Ablation Studies and Further Analysis

*Table 3.* Ablation studies of Glad+Ours on MVTec-AD. "$DN\rightarrow$ depth/channel" refers to mapping the matching dimension into the depth/channel dimension of the 3D U-Net. $\mathcal{C}_0$ denotes the volume uisng the final denoising step, $\mathcal{C}_{N-1}$ indicates uisng $N-1$ intermediate steps. SG and MG denote dual-stream attention guidance. $\mathcal{L}_{\mathrm{F}}$ is focal loss, $\mathcal{L}_{\mathrm{CE}}$ corresponds to the class-aware adaptor, and $\mathcal{L}_{\mathrm{S}}$ is the combination of $\mathcal{L}_{\mathrm{SSIM}}$ and $\mathcal{L}_{\text{Soft-Iou}}$.

| $DN\rightarrow$ depth | $DN\rightarrow$ channel | | | | $\mathcal{L}_{\mathrm{F}}$ | $\mathcal{L}_{\mathrm{CE}}$ | $\mathcal{L}_{\mathrm{S}}$ | Results |
|---|---|---|---|---|---|---|---|---|
| | $\mathcal{C}_0$ | $\mathcal{C}_{N-1}$ | SG | MG | | | | |
| ✓ | - | - | - | - | ✓ | - | - | 87.8/89.0 |
| - | ✓ | - | - | - | ✓ | - | - | 96.2/96.8 |
| - | ✓ | ✓ | - | - | ✓ | - | - | 96.7/97.3 |
| - | ✓ | ✓ | ✓ | - | ✓ | - | - | 97.8/97.5 |
| - | ✓ | ✓ | ✓ | ✓ | ✓ | - | - | 98.3/97.8 |
| - | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | - | 98.5/98.0 |
| - | ✓ | - | ✓ | ✓ | ✓ | ✓ | ✓ | 98.4/97.6 |
| - | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | **98.7/98.2** |

*Table 4.* Extended studies on single-class UAD with our models.

| Benchmark | Method | Image-level | | | Pixel-level | | | |
|---|---|---|---|---|---|---|---|---|
| | | AU-ROC | AP | F1max | AU-ROC | AP | F1max | AUPRO |
| MVTec-AD | Glad | 99.0 | 99.7 | 98.2 | 98.7 | 63.8 | 63.7 | 95.2 |
| | +Ours | **99.3** | 99.7 | **98.3** | **98.9** | **66.2** | **65.0** | **96.4** |
| VisA | Glad | 99.3 | 99.6 | 97.6 | 98.3 | 35.8 | 42.4 | 94.1 |
| | +Ours | **99.5** | **99.7** | **98.1** | **98.6** | **37.3** | **45.3** | **94.5** |

*Table 5.* Evaluation of our models on various anomaly volumes.

| Train \ Test | MVTec-AD | | VisA | |
|---|---|---|---|---|
| | Recon. | Embed. | Recon. | Embed. |
| Recon. | 98.7 / **98.2** | 97.5↓ / 97.1↓ | **93.2** / 98.1 | 92.6↓ / 98.0↓ |
| Embed. | 94.5↓ / 98.0↓ | 98.5 / 98.8 | 85.6↓ / 96.9↓ | **94.3** / 99.2 |
| Hybrid | **98.8**↑ / 98.1 | **98.6**↑ / **98.9**↑ | 93.1 / **98.2**↑ | 92.9 / **99.3**↑ |

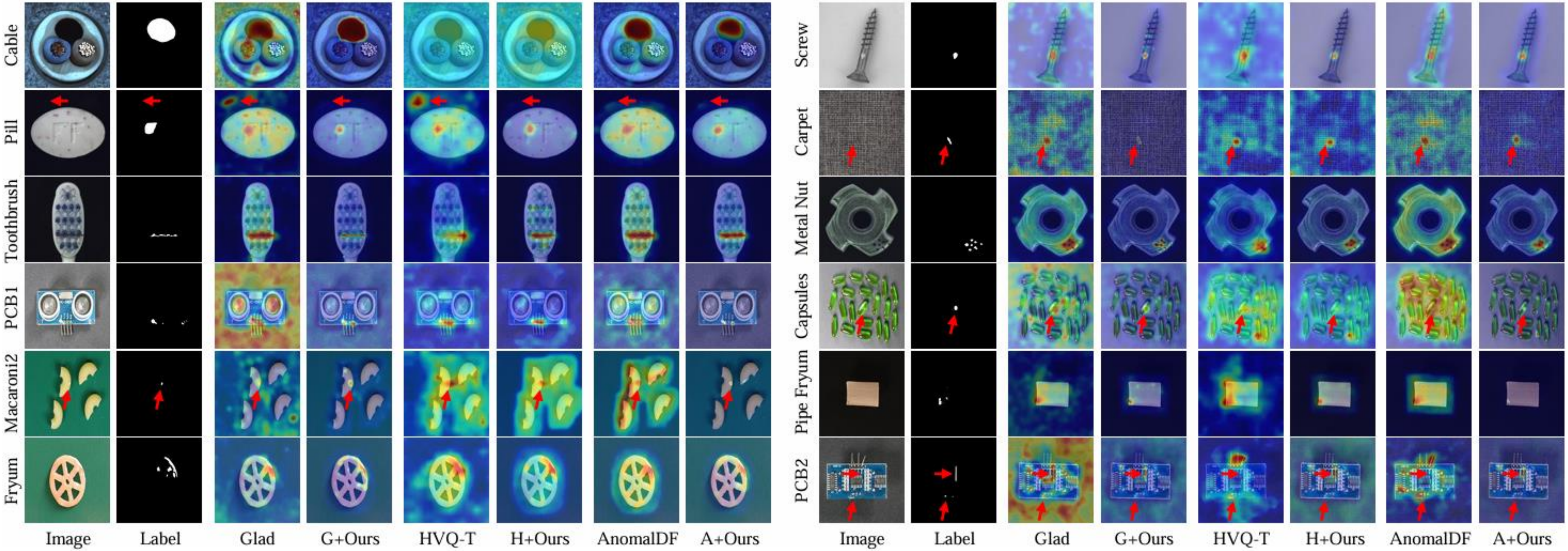
*Table 6.* Computational efficiency of baselines vs. + Ours.

| Method | #Params | FLOPs | Mem. (GB) | Inf. (s/image) |
|---|---|---|---|---|
| UniAD / +Ours | 7.7M / +43.0M | 198.0G / 207.8G | 4.53 / +0.56 | 0.01 / +0.04 |
| Glad/+Ours | 1.3B / +43.8M | >2.2T / 261.3G | 8.79 / +2.07 | 3.96 / +0.37 |
| HVQ-Trans/+Ours | 18.0M / +43.0M | 7.4G / 207.8G | 4.78 / +0.94 | 0.05 / +0.07 |
| AnomalDF/+Ours | 21.0M / +43.8M | 4.9G / 261.3G | 3.25 / +0.82 | 0.31 / +0.32 |
| Dinomaly/+Ours | 132.8M / +43.6M | 104.7G / 114.6G | 4.32 / +1.11 | 0.11 / +0.05 |

✅ CostFilter-AD demonstrates superior performance, effective ablations, strong generalization, and minimal computational overhead across Unsupervised Anomaly Detection tasks.

# Evaluation: precise localization of subtle anomalies

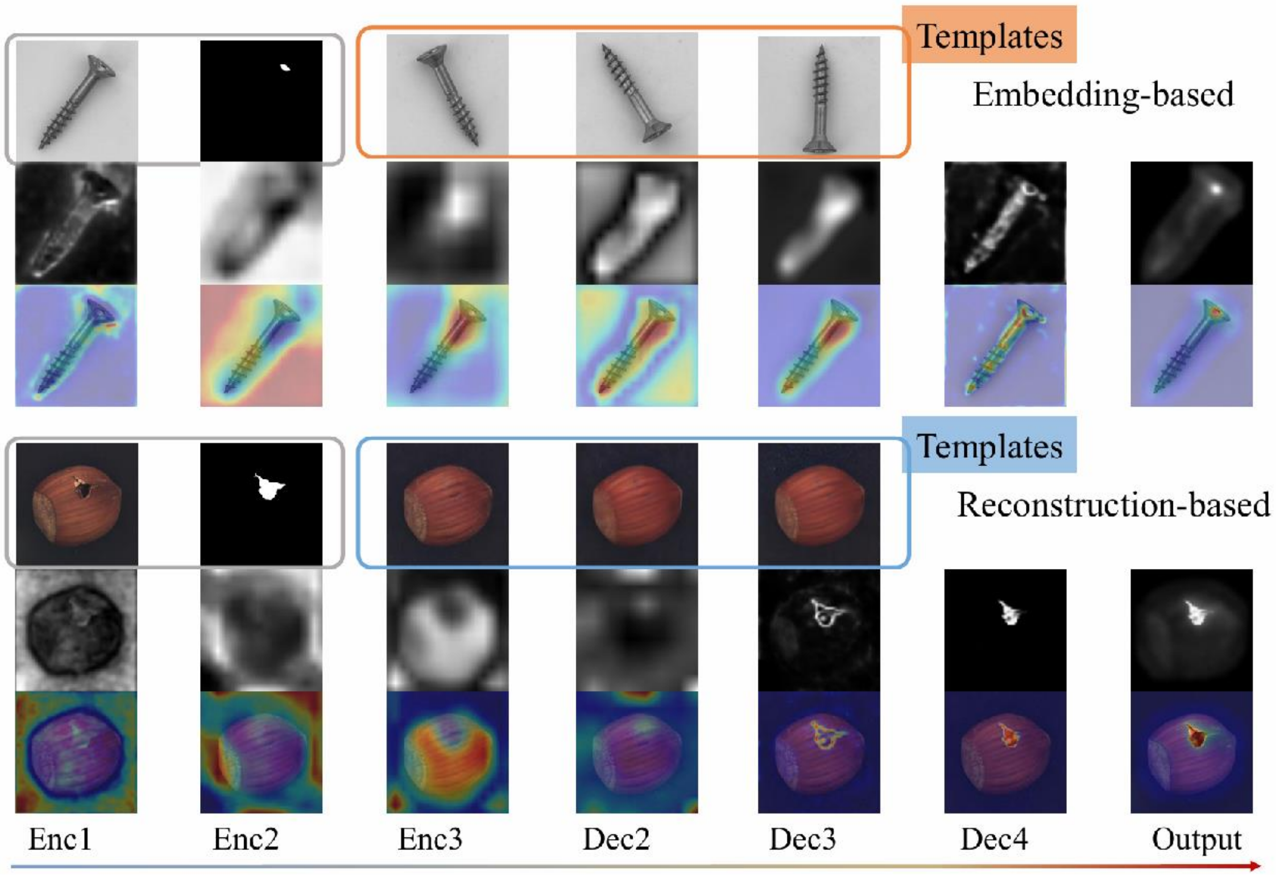
## Ours vs. GLAD, HVQ-Trans, and AnomalDF: Localization Visualization




✅ **Qualitative results show that our method reduces matching noise and improves anomaly localization over GLAD, HVQ-Trans, and AnomalDF on MVTec-AD and VisA.**

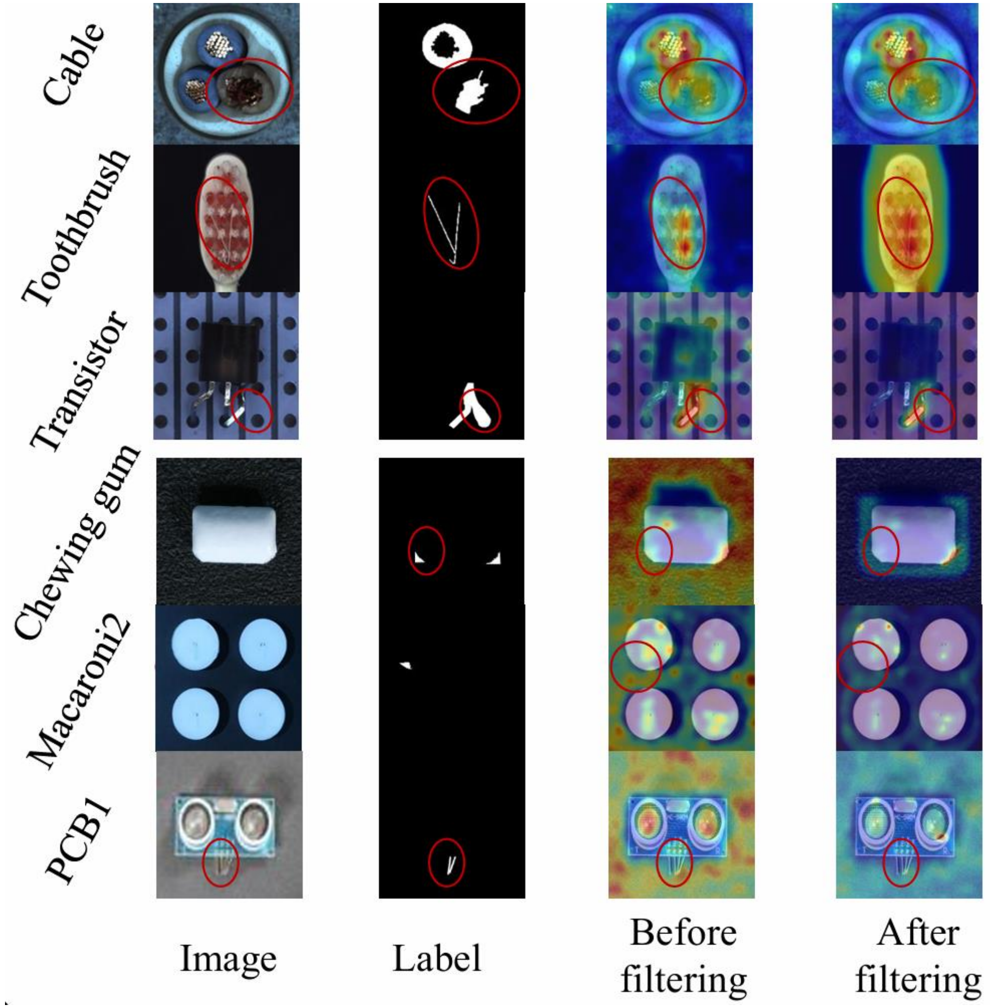
# Evaluation: qualitative results

## Progressive and Fine-grained Denoising




✅ Progressively refines spatial anomaly features across encoder and decoder layers, generating layer-wise heatmaps via attention-driven channel selection and aggregation.

# Failure Cases and Future Direction




## Failure Cases

- **Subtle Anomalies:** Fails on low-contrast or highly localized anomalies unseen during training.
- **Template Sensitivity:** Relies on representative templates; poor quality can degrade detection performance.

## Future Directions

- **Adaptive Cost Modeling:** Refine matching precision through improved or learned cost functions.
- **Spatiotemporal & Multi-modal Extension:** Extend to video or multi-modal inputs for broader applications.
- **Hard Negative Mining:** Incorporate challenging normal cases to enhance model robustness.

東北大學
Northeastern University

# Thank you!

State Key Laboratory of Synthetical Automation for Process Industries