




Towards Real Unsupervised Anomaly Detection Via Confident Meta-Learning

Muhammad Aqeel¹ 

Shakiba Sharifi¹ 

Marco Cristani^{1,2} 

Francesco Setti^{1,2} 

¹ Dept. of Engineering for Innovation Medicine, University of Verona
Strada le Grazie 15, Verona, Italy

² Qualyco S.r.l., Strada le Grazie 15, Verona, Italy

Contact author: muhammad.aqeel@univr.it

Abstract

So-called unsupervised anomaly detection is better described as semi-supervised, as it assumes all training data are nominal. This assumption simplifies training but requires manual data curation, introducing bias and limiting adaptability. We propose Confident Meta-learning (CoMet), a novel training strategy that enables deep anomaly detection models to learn from uncurated datasets where nominal and anomalous samples coexist, eliminating the need for explicit filtering. Our approach integrates Soft Confident Learning, which assigns lower weights to low-confidence samples, and Meta-Learning, which stabilizes training by regularizing updates based on training-validation loss covariance. This prevents overfitting and enhances robustness to noisy data. CoMet is model-agnostic and can be applied to any anomaly detection method trainable via gradient descent. Experiments on MVTec-AD, VIADUCT, and KSDD2 with two state-of-the-art models demonstrate the effectiveness of our approach, consistently improving over the baseline methods, remaining insensitive to anomalies in the training set, and setting a new state-of-the-art across all datasets.

Keywords: Unsupervised Anomaly detection, Meta Learning, Soft Confident Learning

1. Introduction

In industrial manufacturing, real-time detection of defective products is critical to ensuring safety and quality standards, reducing waste, and optimizing production efficiency. The ability to detect defective products at an early stage prevents costly recalls and enhances customer satisfaction. Traditionally, quality control has relied on manual inspection by human operators. However, this process is expensive, produces inconsistent outputs, and is prone to errors due to fatigue and subjective judgment. Moreover, manual inspec-

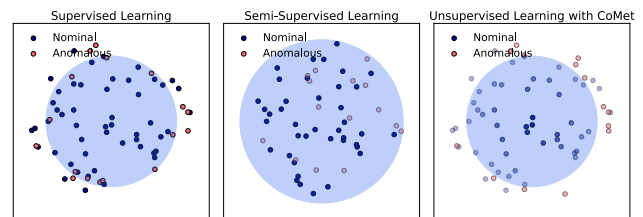


Figure 1. Impact of anomalies and near boundary samples in the training set. In Supervised Learning, positive and negative samples contribute to determining the boundary between the two data distributions. In Semi-Supervised Learning, often improperly called unsupervised, the lack of positive samples (*i.e.* anomalies) lead to overestimate the covariance of the nominal samples' distribution, resulting in a higher number of misdetections at inference time (red points lying inside the blue circle). Our Confident Meta-learning approach allows models to be learned properly in Unsupervised Learning by assigning low confidence weight to samples close to and beyond the decision boundary. Note that in Semi-Supervised Learning the positive samples (red dots) are transparent because they are not available at training time. In Unsupervised Learning instead, dots' transparency is proportional to the assigned confidence weights.

tions are difficult to scale in high-throughput industrial settings. As a result, there is a strong demand for automated defect detection systems that can operate with high accuracy, reliability, and efficiency.

Supervised learning approaches have shown strong performance in this task, but their effectiveness is highly dependent on the quantity and diversity of labeled defective and nominal samples available during training. They require a substantial number of defective samples with pixel-level annotations. Since defects are inherently rare and often subtle, collecting and annotating such data is both costly and time-consuming. Moreover, many defect types exhibit high intra-class variability, requiring a sufficiently large and diverse dataset to ensure robust generalization across different defect categories. To mitigate these challenges, recent

research focused on unsupervised anomaly detection, where models are trained exclusively on nominal data (negative samples) and learn to identify deviations from the learned distribution. While this approach alleviates the need for labeled defective samples, it still requires an operator to carefully curate the training dataset to ensure that no anomalous samples are present. This manual filtering step introduces a significant limitation: it is time-consuming and susceptible to human error and bias. Even a small number of defective samples in the training set can lead the model to misclassify these defects as nominal, reducing its ability to detect true anomalies. Additionally, the assumption that all training samples are nominal does not hold in many real-world industrial applications, where undetected anomalies may be inadvertently included in the dataset. A graphical representation of this phenomenon is shown in Figure 1.

In this paper, we argue that what is commonly referred to as *unsupervised anomaly detection* is more accurately described as semi-supervised anomaly detection, as it implicitly assumes that all training samples are nominal, and we propose a novel training strategy that eliminates the need for manually filtering training data. Our method enables deep learning models to learn from raw, uncurated datasets where nominal and anomalous samples may coexist, without requiring explicit labels. By relaxing the assumption that all training data are nominal, our approach allows anomaly detection models to operate in a truly unsupervised manner. This not only reduces the burden of data annotation but also improves the model’s adaptability to real-world conditions and enhances robustness against data distribution shifts. We achieve this by leveraging a *Soft Confident Learning* approach that dynamically estimates the reliability of training samples at each epoch based on the current model parameters. Samples exhibiting low confidence at a given epoch are assigned lower weights in subsequent gradient updates, reducing their influence on the learning process. Figure 2 shows some confidence weights at the last training epoch. This strategy alone may introduce instability and increase the risk of overfitting. To mitigate these issues, we incorporate a *Meta-Learning* framework that iteratively updates the model using partial subsets of training data while validating the learned representations on the remaining portion. The update step is regularized through a term inversely proportional to the covariance between training and validation losses, ensuring that large parameter updates are discouraged when the model exhibits overfitting tendencies. This combination of Soft Confident Learning and Meta-Learning enhances model robustness, allowing for more reliable learning in the presence of noisy or unfiltered training data. Our **Confident Meta-learning (CoMet)** approach can be used to extend to the truly unsupervised scenario any anomaly detection method that is trainable through gradient descent of a specific loss function, making it an extremely

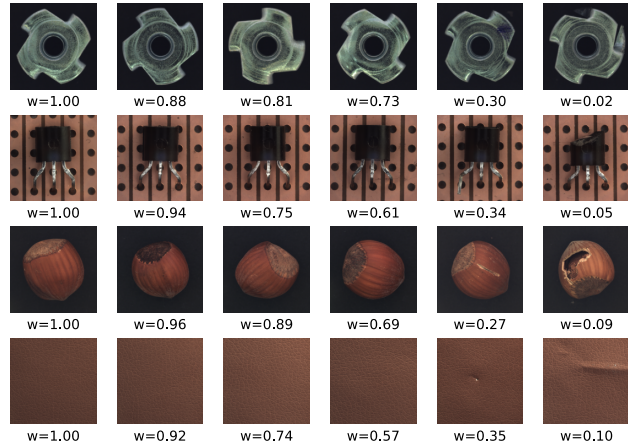


Figure 2. Training samples (nominal and anomalous) for some classes of the MVTec-AD dataset with the associated confidence weight w estimated by CoMet. Weights close to 1 indicate prototypical samples, while lower weights suggest samples close to (or beyond) the decision boundary.

versatile approach that can adapt seamlessly to new state-of-the-art models.

Extensive experiments conducted on the popular benchmarks MVTec-AD [5], VIADUCT [14] and KSDD2 [6], with two state-of-the-art anomaly detection models, namely DifferNet and SimpleNet, demonstrate the effectiveness of CoMet in learning robust parametrizations. These parametrizations not only avoid overfitting to nominal samples but are also largely insensitive to the presence of anomalies in the training set.

The main contributions of our paper can be summarized as follows:

- We propose CoMet, a novel training framework that allows anomaly detection models to learn more robust models by assigning low confidence scores to ambiguous samples near the decision boundary.
- Models trained with CoMet achieve higher performance in anomaly detection by significantly reducing undetected anomalies (false negatives) at the cost of slightly increasing false positives.
- Extensive experiments on three public benchmarks demonstrate that CoMet achieves state-of-the-art performances, effectively handling the presence of anomalous samples in the training set.

2. Related Work

Supervised learning methods have achieved remarkable results in industrial anomaly detection, mostly leveraging fully convolutional network architectures [3, 8]. However, collecting and annotating anomalous samples is a costly and labour-intensive operation that is also prone to er-

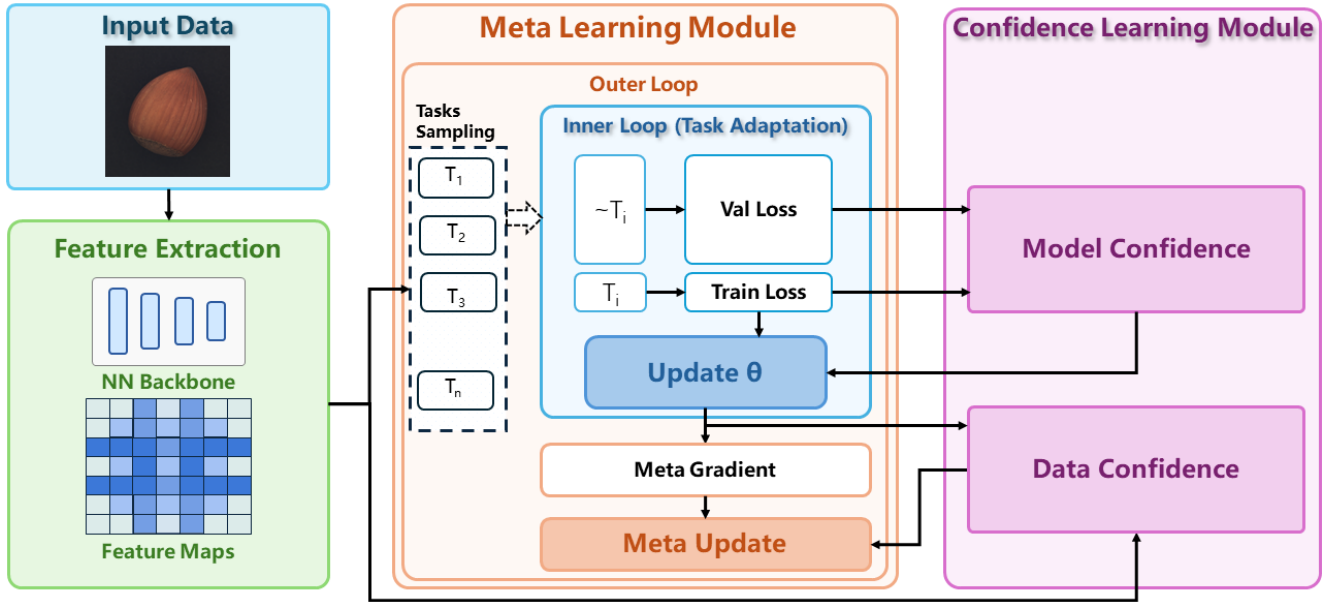


Figure 3. CoMet pipeline. A pretrained feature extractor transforms input images to feature maps if required by the anomaly detection model of choice. Features are then divided into disjointed tasks. In the meta-learning inner loop one task is used for training and all the others for testing. Losses for training and validation are passed to the Soft Confident Learning module to compute the model confidence that is used for regularization of the weights update. Once all tasks have been processed, the SCL module uses current parametrization to compute the confidence of each sample data and use it to weight data in the meta update step.

rors. For this reason, recent research has focused on unsupervised learning techniques. Reconstruction-based methods learn a model that can compress and reconstruct the training data well, relying on the assumption that anomalous regions should not be properly reconstructed because not present in the training data. These approaches can work at the image level [11, 17, 26, 28] or at the feature level [29]. Feature embedding-based methods use deep learning models pre-trained on large datasets to extract features from training samples. Different techniques are then applied to score anomalies by modeling data distributions [20, 21, 27, 30], storing knowledge in memory banks [7, 19], using teacher-student models [4, 22], or training one-class discriminators [15, 16]. Although commonly called unsupervised, a more accurate classification for these methods is semi-supervised, as they inherently assume that all training samples are nominal. While this assumption simplifies model training by eliminating the need for labeled anomalies, it still necessitates human intervention in selecting the training data, potentially introducing bias and limiting the model’s robustness to unseen anomalies. To overcome this issue, some recent works used statistical methods to refine training data iteratively removing samples that are anomalous or close to the decision boundary [1, 2]. In this work, we address these limitations by employing meta-learning alongside soft confident learning to mitigate overfitting of problematic samples.

Soft Confident Learning (SCL) have been explored to enhance the reliability of supervised systems by identifying and correcting mislabeled samples in training datasets [18]. Rather than assuming all training labels are accurate, SCL aims to detect and re-weight or correct mislabeled examples. We extend SCL to the unsupervised scenario to assign low confidence scores to near- and beyond-boundary samples.

To improve model adaptability in dynamic environments, meta-learning has been extensively applied in scenarios requiring rapid model refinement, such as few-shot learning and domain adaptation [9, 12]. Metric-based approaches, including prototypical networks [23] and matching networks [25], learn an embedding space where similar samples are closer, facilitating fast adaptation with minimal data. Optimization-based approaches learn an optimal initialization of the model that allows for quick adaptation to new tasks [9]. These second strategy is more suited to the anomaly detection problem, where meta-learning enables models to adapt decision boundaries dynamically, allowing for effective handling of data distribution shifts over time [24].

3. Confident Meta-Learning pipeline

The proposed CoMet approach leverages soft confident learning and meta-learning to perform anomaly detection

within an unsupervised framework. The pipeline of the proposed approach is shown in Fig. 3.

3.1. Anomaly Detection backbone

The CoMet framework can be coupled with any anomaly detection model that can be trained end-to-end via gradient descent. In these terms, our approach is agnostic with respect to the anomaly detection backbone, and we will show experiments using two very different approaches based on normalizing flows and feature embeddings.

Let $x_i \in \mathcal{D}_{train}$ be the set of training images consisting predominantly of nominal samples, with no explicit labels indicating anomalies. During training, the AD model learns a function $\phi_\theta : \mathcal{D} \rightarrow \mathbb{R}$ that maps an input sample x_i to an anomaly score $a_\theta(x_i) \in \mathbb{R}$, where higher scores indicate a higher likelihood of being anomalous. The parametrization θ is the one that minimized the method-specific loss function \mathcal{L}_{AD} : reconstruction-based methods typically minimize the Mean Squared Error (MSE), defined as $\mathcal{L}_{MSE} = \frac{1}{n} \sum (x_i - \hat{x}_i)^2$ with \hat{x}_i the reconstructed input; density-based methods rely on Negative Log-Likelihood (NLL), given by $\mathcal{L}_{NLL} = -\sum \log(p(x_i|\theta))$ where $p(x|\theta)$ models the normal data distribution; and feature-based methods often incorporate variations of the Structural Similarity Index (SSIM), like $\mathcal{L}_{feat} = \sum (\phi_\theta(x_i) - c)^2$ where c is the center of normal representations in feature space. At testing time, an anomaly score $a_\theta(x_j)$ is computed for each test sample $x_j \in \mathcal{D}_{test}$, and an anomaly is flagged if $a_\theta(x_j) > \tau$, where τ is a predefined or adaptive threshold.

3.2. Soft Confident Learning

The goal of the soft confident learning module is to allow the model to rely more on those samples that are more prototypical for the nominal class and less on those samples that are anomalous or close to the boundary. To this aim, we quantify both model uncertainty and data uncertainty within an unsupervised learning framework, where we do not have access to labels. We will then use these uncertainties to estimate coefficients used in the training phase.

3.2.1. Quantifying Data Uncertainty

To quantify data uncertainty, we adapt the concept of the *confident joint* from the Confident Learning framework to our unsupervised setting. In this context, we consider the relationships between data points and their confidence scores assigned by the model.

We notice that the anomaly score $a_\theta(x_i)$ reflects the confidence of a model parametrized by θ to predict x_i . Since we lack true labels, we use these confidence scores to assess the uncertainty associated with each data point. We map these scores to weights to be used in a soft confident learning framework using a saturated inverse function of the

confidence score as:

$$w_i = \min(1, t/a_\theta(x_i)) \quad (1)$$

where t is a threshold defined according to the Interquartile Range (IQR) statistical method to identify outliers [10]. If we define Q_1 and Q_3 the first and third quartiles of the distribution of the confidence scores $a_\theta(x)$ respectively, the threshold t is defined as:

$$t = Q_3 + \kappa(Q_3 - Q_1) \quad (2)$$

where κ is a parameter that can be tuned to best suit the data characteristics.

We can now define the data-weighted loss function as:

$$\mathcal{L}_{data}(\theta) = \sum_{i=1}^N w_i \cdot \mathcal{L}_{AD}(x_i|\theta) \quad (3)$$

where $\mathcal{L}_{AD}(x_i|\theta)$ is the individual loss function of the anomaly detection model of choice.

3.2.2. Quantifying Model Uncertainty

We quantify the model's uncertainty by calculating the determinant of the covariance matrix Σ formed from the training and validation loss distributions. This approach aligns with statistical methods that interpret the determinant of a covariance matrix as a measure of the volume of spread in multivariate space, with larger values indicating greater variability or uncertainty [13]. Let \mathcal{L}_{train} and \mathcal{L}_{val} represent vectors of the model's training and validation losses, respectively. The covariance matrix Σ is defined as:

$$\Sigma = \begin{bmatrix} \text{Cov}(\mathcal{L}_{train}, \mathcal{L}_{train}) & \text{Cov}(\mathcal{L}_{train}, \mathcal{L}_{val}) \\ \text{Cov}(\mathcal{L}_{val}, \mathcal{L}_{train}) & \text{Cov}(\mathcal{L}_{val}, \mathcal{L}_{val}) \end{bmatrix}. \quad (4)$$

The determinant of Σ serves as a scalar measure of the overall variability between training and validation losses: High values of $\det(\Sigma)$ indicate significant variability, suggesting high model uncertainty. Low values instead reflect consistent performance, indicating higher confidence in the learning process.

To incorporate this measure into the training procedure, we introduce an adaptive regularization term λ that adjusts dynamically:

$$\lambda(\Sigma) = \lambda_0 \cdot (1 + \gamma \cdot \det(\Sigma)) \quad (5)$$

where λ_0 is the base regularization coefficient and γ is a scaling hyperparameter controlling sensitivity to model uncertainty. This adaptive regularization allows the model to impose stronger regularization when uncertainty is high, promoting better generalization, and to relax regularization when the model is learning confidently.

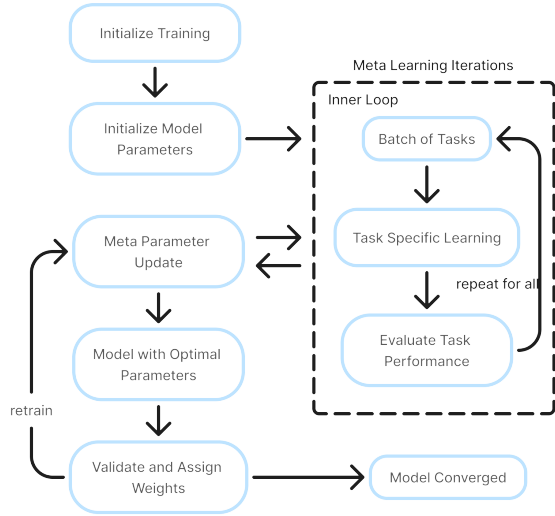


Figure 4. Meta-learning process with outer and inner loops. In the inner loop, each task batch T_1 to T_n updates task-specific parameters by optimizing θ . After completing inner loop updates, the outer loop aggregates these adjustments to refine the global θ , enhancing the model’s ability to generalize and adapt quickly to new tasks.

Combining both model and data uncertainty, our soft confident learning loss function becomes:

$$\mathcal{L}_{SCL}(\theta) = \sum_{i=1}^N w_i \cdot \mathcal{L}_{AD}(x_i|\theta) + \lambda(\Sigma) \cdot \|\theta\|_2^2 \quad (6)$$

3.3. Meta Learning

To enhance the model’s ability to efficiently adapt to new data with sample-specific confidence, we incorporate Model-Agnostic Meta-Learning (MAML) [9] into our unsupervised framework. Unlike standard regularization that treats parameters uniformly, MAML enables covariance-based uncertainty quantification and identifies which parameter updates generalize well across tasks. We randomly divide training samples into n disjoint tasks, where each iteration uses task T_i for training and the remaining tasks for validation, creating the structured splits necessary for our adaptive regularization term $\lambda(\Sigma)$.

The MAML algorithm, shown in Figure 4, consists of two optimization loops. In the inner loop, we update the model parameters to fit a specific task using gradient descent steps with learning rate $\alpha = 10^{-4}$:

$$\theta' = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\text{train}}(\theta), \quad (7)$$

where $\mathcal{L}_{\text{train}}(\theta)$ is the training loss on the current task computed as in (6).

In the outer loop, we update the original model parameters θ using validation performance across the remaining

tasks with meta-learning rate $\beta = 2 \times 10^{-4}$:

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \mathcal{L}_{\text{meta}}(\theta'), \quad (8)$$

This two-level optimization enables the model to learn which parameter updates generalize well, effectively learning sample-specific weights while balancing confidence-weighted updates with generalization performance.

We integrate the reweighted loss function from Soft Confident Learning into the meta-objective:

$$\mathcal{L}_{\text{meta}}(\theta') = \sum_{i=1}^N w_i \cdot \mathcal{L}_{AD}(x_i|\theta') + \lambda(\Sigma') \cdot \|\theta'\|_2^2, \quad (9)$$

where θ' are the adapted parameters from the inner loop, $\mathcal{L}_{AD}(x_i|\theta')$ is the individual loss for data point x_i using parameters θ' , w_i are the weights based on confidence scores calculated using θ' , and $\lambda(\Sigma')$ is the adaptive regularization term as previously defined in (5).

4. Experiments

We present experiments integrating CoMet with two distinct approaches, enhancing their performance. Using the standard protocol with only nominal training data, our method outperforms baselines by isolating boundary-close nominal samples. We also demonstrate robustness to mislabelled anomalies in training, and an ablation study confirms the impact of our proposed modules.

4.1. Datasets

We extensively validated our proposed approach on three challenging public benchmarks. We report results on the **MVTec-AD** dataset [5], which consists of 5,354 high-resolution color images spanning over 10 object and 5 texture categories, each containing multiple defect types. While this dataset remains the most widely used benchmark in the field, it is now considered largely solved, with several methods achieving AUROC scores surpassing 99%. A more challenging scenario is introduced by the recently released **VIADUCT** dataset [14], a benchmark consisting of 10,986 high-resolution images across 49 categories from 11 distinct sectors. Each object is captured from five different perspectives, offering a comprehensive view of the dataset’s variability. Finally, the **KSDD2** dataset [6] is a surface anomaly detection dataset containing 2,085 anomaly-free images and 246 anomalous images. The dataset is particularly challenging due to the presence of near-in-distribution surface anomalies, making it difficult to distinguish true anomalies from typical surface variations.

4.2. Evaluation Metrics

We evaluate image-level anomaly detection performance using the standard Area Under the Receiver Operating

Characteristic Curve (AUROC), denoted as I-AUROC, based on the anomaly detection scores $a_\theta(x)$, equations (12) and (16). We also report pixel-level AUROC (P-AUROC).

4.3. Implementation Details

We implemented our model using the PyTorch framework and trained it on an NVIDIA RTX 4090 GPU for efficient training and inference. Input images were resized to 256×256 pixels with optional rotation augmentations. Regularization was achieved through weight decay to prevent overfitting. Training was conducted over 180 epochs with a batch size of 192 and a learning rate of 2×10^{-4} . This setup enables our model to effectively learn and generalize complex data distributions, achieving robust performance in density estimation and generative tasks.

4.4. CoMet with Normalizing Flows

Normalizing flows (NF) are known to be effective in mapping input data $x \in X$ sampled from a complex distribution $p(x)$ to a latent space z with a simpler, *e.g.* gaussian, distribution $p(z)$. This is achieved by a series of transformations parametrized by θ . The mapping of data distributions is then defined by:

$$p_\theta(x) = p_\theta(z) \left| \det \frac{\partial z}{\partial u} \right| \quad (10)$$

In our work we incorporate the DifferNet model proposed by [20]: we use a pre-trained AlexNet to map training images $x_i \in X$ to a multi-scale feature space $u_i \in U$, capturing both fine and coarse details. Each input distribution is split into two parts, u_1 and u_2 , which interact with each other through alternating translation (τ) and scale (σ) functions. During training, the normalizing flow model is optimized to find the parametrization θ that minimizes the probability distribution of the inputs' negative log-likelihood of nominal samples. The loss function is:

$$\mathcal{L}_{NF}(u|\theta) = \frac{\|z\|^2}{2} - \log \left| \det \frac{\partial z}{\partial u} \right| \quad (11)$$

where the first term encourages features to map close to $z=0$ in the latent space, and the second term, involving the Jacobian's log determinant, penalizes trivial scaling solutions, thereby promoting meaningful transformation. By applying multiple transformations of each input during training, the model learns a robust mapping that generalizes well.

At inference time, the NF model evaluates the likelihood of image features. Features with low likelihood are flagged as anomalies. The scoring function is defined as:

$$a_\theta(x_i) = \mathbb{E}_{S_i} [-\log p(f_\theta(f_\phi(S_i(x_i))))] \quad (12)$$

where S_i is a transformation applied to the input x_i (*e.g.*, rotations, translations, flips), allowing for a more stable

anomaly score across variations. f_ϕ is the pre-trained feature extractor that maps the input images to a multi-scale feature space, and f_θ is the Normalizing Flow model. The anomaly score $a_\theta(x_i)$ is finally the expected value of the negative log-likelihood over multiple transformations S_i , ensuring robustness to small changes in the input.

4.5. CoMet with SimpleNet

SimpleNet [16] is a lightweight anomaly detection method that operates directly in the feature space. We use the original formulation of this method that employs a wideResNet-50 architecture pre-trained on Imagenet to map input images x_i to local features $o_i = F_*(x_i)$, where the * symbol indicates that the parametrization of the feature extractor is frozen. A shallow neural network G_{θ_1} , usually a single fully-connected layer, adapts the extracted features to a lower-dimensional space, also reducing domain bias and producing task-specific features q_i :

$$q_i = G_{\theta_1}(o_i) = G_{\theta_1}(F_*(x_i)) \quad (13)$$

At training time, SimpleNet generates synthetic anomalies by perturbing normal features with Gaussian noise such as:

$$q_i^- = q_i + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \quad (14)$$

where σ controls the perturbation intensity.

Finally, a binary discriminator D_{θ_2} is trained to act as a normality scorer, estimating the normality of each pixel in the image. This is a 2-layer MLP structure that learns to output positive scores for normal features, and negative scores for generated anomalous features. The loss function is based on the truncated L_1 norm as follows:

$$\mathcal{L}(x|\theta) = \max(0, th - D_{\theta_2}(q_i)) + \max(0, -th + D_{\theta_2}(q_i) + \epsilon) \quad (15)$$

with th a saturation threshold set to 0.5 by default, and $\theta = \{\theta_1, \theta_2\}$ is the complete parametrization of the model accounting for the feature adapter and the discriminator.

During inference, the anomaly feature generator is removed, and the discriminator directly outputs the anomaly score:

$$a_\theta(x_i) = D_{\theta_2}(G_{\theta_1}(F_*(x_i))). \quad (16)$$

4.6. Results

Quantitative results on each class of MVTecAD, VIADUCT, and KSDD2 datasets at image- (I-AUROC) and pixel-level (P-AUROC) are reported in Tables 1, 2, and 3 respectively. We report the results of the four best-performing state-of-the-art methods, and a direct comparison with DifferNet and SimpleNet with our CoMet training procedure.

Our method demonstrates competitive performance across all the benchmarks. On MVTec-AD dataset (Table 1), we achieve the best performance of 99.7% using

Table 1. Comparison results on the MVTEC-AD dataset of anomaly detection performance (I-AUROC/P-AUROC in %).

Method	PaDiM[7]	DRÆM[28]	CSFlow[21]	PatchCore[19]	DifferNet[20]	CoMet-NF	SimpleNet[16]	CoMet-SN
Carpet	99.8/99.1	97.0/95.5	100/-	98.7/99.0	92.9/90.7	98.2/96.5	99.7/98.2	99.7/98.3
Grid	96.7/97.3	99.9/97.7	99.0/-	98.2/98.7	84.0/89.9	97.5/96.6	99.7/98.8	100/99.0
Leather	100/99.2	100/98.6	100/-	100/99.3	97.1/98.1	100/98.2	100/99.2	100/99.8
Tile	98.1/94.1	99.6/99.2	100/-	98.7/95.6	99.4/96.3	100/99.2	99.8/97.0	100/94.1
Wood	99.2/94.9	99.1/96.4	100/-	99.2/95.0	99.8/98.2	100/99.8	100/94.5	100/94.4
Bottle	99.1/98.3	99.2/99.1	99.8/-	100/98.6	99.0/98.3	100/99.8	100/96.9	100/98.6
Cable	97.1/96.7	91.8/94.7	99.1/-	99.5/98.4	95.9/94.6	98.9/97.6	99.9/97.6	100/98.1
Capsule	87.5/98.5	98.5/94.3	97.1/-	98.1/98.8	86.9/96.7	98.1/96.4	97.7/98.9	98.1/99.6
Hazelnut	99.4/98.2	100/99.7	99.6/-	99.9/98.7	99.3/98.4	100/99.6	100/97.9	100/98.3
Metal Nut	96.2/97.2	98.7/99.5	99.1/-	100/98.4	96.1/97.9	99.7/98.9	100/98.8	100/99.4
Pill	90.1/95.7	98.9/97.6	98.6/-	96.6/97.4	88.8/95.3	98.6/96.8	99.0/95.1	99.0/98.9
Screw	97.5/98.5	93.9/97.6	97.6/-	98.1/99.4	96.3/96.7	98.3/96.6	98.2/99.3	98.7/99.8
Toothbrush	100/98.8	100/98.1	91.9/-	100/98.7	98.6/99.0	100/100	99.7/98.5	100/99.1
Transistor	94.4/97.5	93.1/90.9	99.3/-	100/96.3	91.1/93.2	99.1/98.6	100/97.6	100/97.5
Zipper	98.6/98.5	100/98.8	99.7/-	99.4/98.8	95.1/96.4	99.6/96.8	99.9/98.9	100/99.4
<i>Average</i>	95.8/97.5	98.0/97.3	98.7/-	99.1/98.1	94.9/96.0	99.2/98.1	99.6/98.1	99.7/98.3

Table 2. Results of anomaly detection on VIADUCT dataset (I-AUROC/P-AUROC in %).

Method	EfficientAD[4]	DRÆM[28]	MSFlow[30]	PatchCore[19]	DifferNet[20]	CoMet-NF	SimpleNet[16]	CoMet-SN
Shredded CR.	84.1/-	51.4/-	94.1/-	89.1/-	80.9/92.6	85.2/88.6	95.3/96.4	98.5/99.3
Encoder	51.2/-	63.0/-	51.9/-	70.9/-	51.5/53.6	64.3/61.2	80.2/79.9	84.7/97.3
Raspberry	94.7/-	90.0/-	92.8/-	98.9/-	91.0/92.1	93.6/92.2	97.6/99.0	98.6/99.1
Device Box	87.5/-	50.8/-	79.6/-	75.3/-	55.8/60.7	72.7/68.4	89.0/97.0	95.0/98.2
L-Fitting	88.8/-	55.7/-	69.6/-	80.6/-	80.8/85.4	82.4/80.2	85.4/96.9	95.4/99.3
Threaded Fitting	59.6/-	60.5/-	58.6/-	55.0/-	56.5/64.7	64.9/64.7	58.9/70.3	61.3/79.8
Redon needle	98.1/-	66.8/-	100/-	97.2/-	93.2/92.7	100/95.8	99.9/99.9	100/99.1
Aluminium Plate	89.4/-	56.7/-	96.2/-	86.6/-	85.9/98.8	92.3/93.5	92.7/94.5	96.0/98.9
PaperClip	97.2/-	69.5/-	99.2/-	99.8/-	90.8/92.6	97.9/94.2	99.5/95.1	100/97.6
Air Muffler Large	89.1/-	55.4/-	94.7/-	99.8/-	89.9/91.6	94.5/91.6	99.0/99.3	99.7/99.4
Saw Blade	69.2/-	58.2/-	70.2/-	61.0/-	61.8/64.7	70.3/66.5	60.8/89.7	63.9/98.1
<i>Average</i>	82.6/-	61.6/-	82.4/-	83.1/-	76.2/80.9	83.5/81.5	87.1/92.5	90.3/97.0

Table 3. Comparison results on the KSDD2 dataset of anomaly detection performance (I-AUROC/P-AUROC in %).

Method	DRAEM[28]	MAD[17]	DSR[29]	MLD-IR[2]	DifferNet[20]	CoMet-NF	SimpleNet[16]	CoMet-SN
<i>Average</i>	77.8/-	79.3/-	87.2/-	94.3/92.6	91.5/92.1	94.9/93.4	91.7/93.0	92.2/93.4

CoMet-SN, the best performance on this dataset. The improvement on the original SimpleNet model is just 0.1%, yet a great achievement considering how close we are to the perfect performance. Moreover, our method consistently outperforms the baseline on all 15 classes, with a perfect score on 11 of them. CoMet-SN also sets the new state-of-the-art in terms of P-AUROC with 98.3%. As for the normalizing flows, CoMet-NF is performing very well with 99.2% I-AUROC on average and an improvement over the DifferNet of 4.3%. The performance is largely due to poor performance on GRID, CAPSULE, and PILL, which are also problematic for the DifferNet method, suggesting that the challenge may lie in fitting the reconstruction using normalizing flows.

On the VIADUCT dataset (Table 2), CoMet-SN establishes a new state-of-the-art performance, outperforming all methods in 6 out of 11 classes. Among the remaining 5

classes, in 2 of them the best performance is achieved by CoMet-NF, witnessing the effectiveness of our proposed training strategy; for the last 3 classes, our CoMet-SN is always the runner-up with an average gap of only 0.2%. With an overall average I-AUROC of 83.5% for CoMet-NF (7.3% better than the DifferNet) and 90.3% for CoMet-SN (3.2% better than SimpleNet) across 11 different sectors, our method demonstrates strong consistency in handling diverse anomalies and adaptability to challenging scenarios.

Finally, on the KSDD2 dataset (Table 3), our approach sets the new state-of-the-art I-AUROC with 94.9% using normalizing flows model (CoMet-NF), while the performance improvement of CoMet-SN is more limited in this case. This is possibly due to the fact that DifferNet is particularly suited for this kind of texture, as proved by the high results achieved by the baseline method. Both variants are comparable to the state of the art.

Table 4. Average Precision and Recall for baseline and CoMet models on MVTec AD, VIADUCT and KSDD2 datasets.

Dataset	DifferNet [20]			CoMet-NF		
	Precision	Recall	F ₁ -score	Precision	Recall	F ₁ -score
MVTec AD	95.6	76.4	84.9	92.5	93.4	92.9
VIADUCT	79.1	70.3	74.4	77.1	90.8	83.4
KSDD2	90.9	87.5	89.2	87.4	94.3	90.7

Dataset	SimpleNet [16]			CoMet-SN		
	Precision	Recall	F ₁ -score	Precision	Recall	F ₁ -score
MVTec AD	98.1	98.9	98.5	97.8	99.8	98.8
VIADUCT	84.1	93.6	88.6	83.4	97.5	89.9
KSDD2	96.2	70.0	81.0	96.0	70.3	81.2

Table 5. Average I-AUROC on MVTec-AD dataset for different component removals in the ablation study.

Configuration	CoMet-NF	CoMet-SN
CoMet w/o SCL and ML (baseline)	94.9	99.4
CoMet w/o ML	96.8	99.4
CoMet w/o SCL on Data & Model	97.2	99.5
CoMet w/o SCL on Data	97.9	99.5
CoMet (full)	99.2	99.7

Across all datasets, our method consistently establishes a new state-of-the-art in both image-level and pixel-level AUROC scores, demonstrating robust performance across diverse object types and anomaly complexities. Additionally, it significantly outperforms baseline methods across both metrics, achieving substantial improvements in accuracy and reliability. We attribute this strong performance to our framework’s ability to prioritize real prototypical samples while reducing the influence of nominal samples near the decision boundary. This leads to fewer false negatives (undetected anomalies) at the cost of a slight increase in false positives, *i.e.* nominal samples incorrectly flagged as anomalous. This is supported by the results in Table 4, which show CoMet significantly improves recall while maintaining a controlled decrease in precision.

4.7. Ablation Study

To assess the contribution of individual components in our framework, we conducted an ablation study on the MVTec-AD dataset, systematically removing Meta-Learning (ML) and Soft Confident Learning (SCL) from our model. The results, presented in Table 5, highlight the significant impact of these components on overall performance. Removing the Meta-Learning component (**w/o ML**) results in a significant decrease in I-AUROC of about 2.5% on NF model and 0.3% for SN model. Indeed, without meta-learning, the training and validation sets coincide at each step, making the model susceptible to overfitting when computing the confidence weights. Nevertheless, this model achieves a substantial improvement over the baseline for NF, demonstrating the effectiveness of the Soft Confident Learning module on its own. Removing Soft Confident Learning on Data (**w/o SCL on Data**) lowers I-AUROC, highlighting its role in handling ambiguous samples. Omitting both Soft Confident Learn-

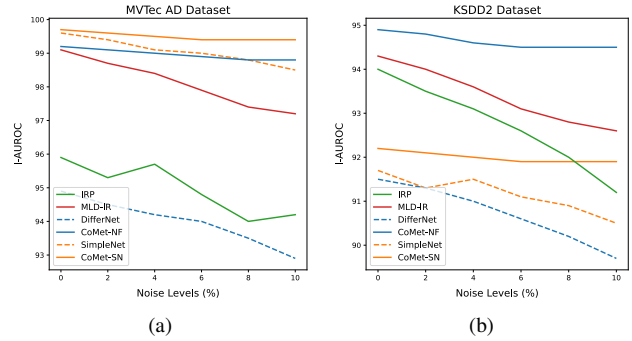


Figure 5. Comparison of anomaly detection methods on the MVtec-AD (a) and KSDD2 (b) datasets, showing I-AUROC values across noise levels from 0% to 10%.

ing on Data and Model (**w/o SCL on Data & Model**) further reduces performance, demonstrating the added benefit of integrating SCL at both levels.

4.8. Noise Robustness

We finally tested the robustness of the model when subject to increasing levels of noise in the input data. Noise is defined as the number of anomalous samples included in the training set. Figure 5 reports I-AUROC performance for noise levels ranging from 0% to 10%. We report comparison against both DifferNet and SimpleNet baselines, and two Robust Anomaly Detection methods IRP [1] and MLD-IR [2] that are specifically designed to address this kind of scenario. Besides being the best performing among the different alternatives, CoMet also sustains high AUROC with minimal degradation when noise intensity increases up to 10%. This experiment is crucial to prove the suitability of our proposed approach to handle training datasets containing both nominal and anomalous samples, moving from semi-supervised to real unsupervised anomaly detection.

5. Conclusion

We targeted the problem of unsupervised anomaly detection, where unlabelled nominal and anomalous samples are available at training time. We presented CoMet, an innovative framework for training anomaly detection models that integrates soft confident learning with meta learning to iteratively refine decision boundaries by dynamically identifying and downweighting ambiguous boundary samples. Extensive experiments on industrial datasets demonstrate that CoMet achieves state-of-the-art performance, significantly improving recall without compromising precision. Moreover, its ability to operate effectively on noisy, uncurated datasets makes it highly adaptable to real-world industrial applications. These results highlight the potential of CoMet to advance unsupervised anomaly detection, paving the way for more reliable and scalable defect detection.

Acknowledgements

This study was carried out within the PNRR research activities of the consortium iNEST (Interconnected North-East Innovation Ecosystem) funded by the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza (PNRR) – Missione 4 Componente 2, Investimento 1.5 – D.D. 1058 23/06/2022, ECS_00000043).

References

- [1] Muhammad Aqeel, Shakiba Sharifi, Marco Cristani, and Francesco Setti. Self-supervised iterative refinement for anomaly detection in industrial quality control. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2025. 3, 8
- [2] Muhammad Aqeel, Shakiba Sharifi, Marco Cristani, and Francesco Setti. Meta learning-driven iterative refinement for robust anomaly detection in industrial inspection. In *European Conference on Computer Vision*, pages 445–460. Springer, 2025. 3, 7, 8
- [3] Aimira Baitieva, David Hurych, Victor Besnier, and Olivier Bernard. Supervised anomaly detection for complex industrial images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17754–17762, 2024. 2
- [4] Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 128–138, 2024. 3, 7
- [5] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9592–9600, 2019. 2, 5
- [6] Jakob Božič, Domen Tabernik, and Danijel Skočaj. Mixed supervision for surface-defect detection: from weakly to fully supervised learning. *Computers in Industry*, 2021. 2, 5
- [7] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition (ICPR)*, pages 475–489. Springer, 2021. 3, 7
- [8] Hongwen Dong, Kechen Song, Yu He, Jing Xu, Yunhui Yan, and Qinggang Meng. PGA-Net: Pyramid feature fusion and global context attention network for automated surface defect detection. *IEEE Transactions on Industrial Informatics*, 16(12):7448–7458, 2019. 2
- [9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, pages 1126–1135, 2017. 3, 5
- [10] Alejandro C Frery. Interquartile range. In *Encyclopedia of Mathematical Geosciences*, pages 664–666. Springer, 2023. 4
- [11] Matic Fučka, Vitjan Zavrtanik, and Danijel Skočaj. Transfusion—a transparency-based diffusion model for anomaly detection. In *European Conference on Computer Vision (ECCV)*. Springer, 2025. 3
- [12] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3
- [13] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 4
- [14] Jan Lehr, Jan Philipps, Alik Sargsyan, Maximilian Botschen, Shoghik Gevorgyan, Anna-Maria Paust, Martin Pape, and Viet Nguyen Hoang. Viaduct: Multisector data set for visual industrial anomaly detection. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 5
- [15] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9664–9674, 2021. 3
- [16] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20402–20411, 2023. 3, 6, 7, 8
- [17] Neelu Madan, Nicolae-Cătălin Ristea, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised masked convolutional transformer block for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3, 7
- [18] Curtis G Northcutt, Lu Jiang, and Isaac L Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021. 3
- [19] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14318–14328, 2022. 3, 7
- [20] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but different: Semi-supervised defect detection with normalizing flows. In *Winter Conference on Applications of Computer Vision (WACV)*, 2021. 3, 6, 7, 8
- [21] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1088–1097, 2022. 3, 7
- [22] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2592–2602, 2023. 3
- [23] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 3
- [24] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. Hospedales. Learning to compare: Relation network for few-

- shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. [3](#)
- [25] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016. [3](#)
- [26] Hang Yao, Ming Liu, Zhicun Yin, Zifei Yan, Xiaopeng Hong, and Wangmeng Zuo. GLAD: towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection. In *European Conference on Computer Vision*. Springer, 2024. [3](#)
- [27] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021. [3](#)
- [28] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem—a discriminatively trained reconstruction embedding for surface anomaly detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8330–8339, 2021. [3](#), [7](#)
- [29] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dsr—a dual subspace re-projection network for surface anomaly detection. In *European Conference on Computer Vision (ECCV)*, pages 539–554. Springer, 2022. [3](#), [7](#)
- [30] Yixuan Zhou, Xing Xu, Jingkuan Song, Fumin Shen, and Heng Tao Shen. Msflow: Multiscale flow-based framework for unsupervised anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. [3](#), [7](#)