Enhancing GANomaly-Based Anomaly Detection for X-Ray Cargo Inspection

Kholoud Alotaibi, Nasser Nasrabadi
Computer Science and Engineering, Oakland University Rochester, MI, USA

Abstract—**Anomaly detection in X-ray cargo imagery is challenging due to complex scene structures, object overlap, and limited labeled abnormal data. Reconstruction-based methods address this problem by learning normal cargo patterns and identifying deviations during testing. This study investigates how feature-level reconstruction objective functions influence detection performance within the GANomaly framework. Five objective configurations are evaluated on the CargoX dataset: a pixel-based baseline and three perceptual loss variants using Visual Geometry Group 16-layer network (VGG16) feature supervision at different depths (i.e., Rectified Linear Unit layers ReLU2_2, ReLU3_3, ReLU4_3, and multi-scale), and an encoder replacement using a ResNet50 with and without perceptual supervision. Performance is assessed using Receiver Operating Characteristic Area Under Curve (ROC-AUC), precision, recall, and F1-score, supported by qualitative analysis of reconstructions and residual maps. Results show that mid-level perceptual supervision (ReLU3_3) achieves the best performance. It improves ROC-AUC from 0.7182 to 0.7548 and demonstrates enhanced sensitivity to structural anomalies. Replacing the original GANomaly encoder with ResNet50 increases ROC-AUC to 0.7312 and improves precision. Combining ResNet50 with perceptual supervision achieves a ROC-AUC of 0.7517. However, it does not surpass the original ReLU3_3 configuration in recall or F1-score. Shallow features (ReLU2_2) and multi-scale aggregation do not improve detection. Failure analysis highlights challenges with low-contrast anomalies and structurally complex normal cargo scenes. These findings show that anomaly detection performance depends on both reconstruction supervision and encoder design. Therefore, loss selection and feature extraction should be analyzed together in reconstruction-based models.**

Keywords—Anomaly detection; cargo X-ray imaging; GANomaly; perceptual loss; feature-level reconstruction; semi-supervised learning; generative adversarial networks; structural anomaly detection; security screening; reconstruction-based detection; deep learning for X-ray inspection; ResNet50

I. INTRODUCTION

Anomaly detection in cargo X-ray imagery is a critical task for border security and customs inspection. Automated screening systems must detect hidden or dangerous items in cluttered cargo scenes. These scenes include overlapping objects, occlusions, and complex materials. In this domain, anomalous instances are rare, diverse, and difficult to annotate, which limits the applicability of fully supervised learning approaches.

Reconstruction-based anomaly detection provides a practical alternative by learning the distribution of normal cargo patterns and identifying deviations at test time. Among such methods, GANomaly [2] has emerged as a widely adopted framework. It models normal data through an encoder-decoder-encoder architecture and detects anomalies using reconstruction discrepancies and latent feature inconsistencies. However, the original GANomaly [2] formulation relies on pixel-level reconstruction losses, which primarily enforce low-level intensity similarity. In complex X-ray cargo imagery, such losses often fail to capture structural and semantic deviations. This limitation reduces sensitivity to subtle but security-relevant anomalies.

To address this limitation, feature-level reconstruction objectives offer a promising direction. Perceptual loss is computed using intermediate features from a pretrained convolutional network and measures similarity in deep feature space instead of raw pixel space. These deep features capture structural information related to object layout and spatial organization. Such structural features are important in cargo inspection scenarios.

This work investigates how reconstruction objectives influence anomaly detection performance in cargo X-ray imagery. Five configurations within the GANomaly framework are evaluated under a unified protocol on the CargoX dataset [18], including a pixel-based baseline, perceptual loss variants, and an encoder replacement using ResNet50 with and without perceptual supervision. This design allows analysis of how reconstruction objectives and encoder representations jointly influence anomaly sensitivity and reconstruction behavior.

This study extends GANomaly in two directions. First, it evaluates perceptual reconstruction supervision at multiple VGG16 feature depths. Second, it examines encoder replacement using ResNet50 within the same reconstruction-based framework. These experiments are conducted on CargoX to analyze how reconstruction supervision and encoder design affect anomaly detection performance.

Experimental results show that mid-level perceptual supervision (e.g., ReLU3_3) significantly improves detection performance compared to pixel-wise reconstruction, with ROC-AUC increasing from 0.7182 to 0.7548 (+5.1%), F1-score improving by 12.7%, and recall improving by 23.2%. Surprisingly, multi-scale perceptual supervision that combines features from multiple depths does not improve performance beyond the best single-scale configuration. This result indicates that selecting the appropriate feature depth is more important than aggregating features from multiple depths. The ResNet50 encoder [15] provides moderate gains (+1.8% ROC-AUC) compared to the baseline GANomaly [2] Deep Convolutional Generative Adversarial Network (DCGAN) encoder. This shows that encoder capacity contributes to performance improvement. However, the reconstruction objective has a stronger impact on the final results.

These findings highlight the importance of selecting reconstruction objectives that balance structural consistency and semantic preservation. The reconstruction objective should maintain meaningful semantic information while ensuring stable structure. At the same time, it should allow sufficient latent-space discrimination between normal and abnormal patterns.

II. Related Work

A. Reconstruction-Based Anomaly Detection in X-ray Imaging

Automated anomaly detection in X-ray cargo screening aims to identify concealed or unauthorized items in visually complex environments. Traditional inspection systems often rely on human operators, where fatigue and subjective judgment can reduce detection reliability [1]. Early automated approaches used handcrafted features and rule-based methods, which struggled with clutter, occlusion, and unpredictable cargo configurations [4]. Classical machine learning models such as Support Vector Machines (SVMs) and Random Forests improved flexibility but remained limited by their dependence on manually designed features [6].

Deep learning approaches, particularly convolutional neural networks, significantly advanced anomaly detection by learning features directly from raw data [6]. When labeled anomaly samples are scarce, reconstruction-based methods offer a practical solution. Autoencoders and variational autoencoders learn representations of normal data and identify anomalies as deviations from the learned distribution [7]. GAN-based models [2]-[3] further improved this paradigm by learning complex data distributions.

GANomaly [2] is a widely adopted reconstruction-based model. It uses an encoder-decoder-encoder generator and detects anomalies using both reconstruction discrepancies and latent feature inconsistencies. However, the original GANomaly framework relies on pixel-based reconstruction losses such as mean absolute error (L1), which primarily enforce low-level visual similarity and may not capture higher-level structural deviations in complex X-ray cargo scenes.

Deep convolutional architectures have further improved feature representation in anomaly detection tasks. Residual networks (ResNet) introduce identity skip connections that facilitate stable optimization and enable deeper feature extraction without degradation during training [15]. The ResNet50 architecture has demonstrated strong performance in visual recognition scenarios due to its ability to capture hierarchical structural patterns [15]. Recent studies have explored ResNet-based encoders for anomaly detection in industrial and surveillance contexts [16], [17], showing improved representation learning compared to conventional convolutional encoders. In industrial defect detection, ResNet50 has shown superior performance in identifying surface anomalies such as broken components and contamination [16], while in video surveillance, combining ResNet50 with Bidirectional Long Short-Term Memory (BiLSTM) networks achieves high accuracy in abnormal activity detection [17]. Motivated by these findings, this work evaluates a ResNet50 encoder within the GANomaly framework to investigate whether stronger feature extraction improves reconstruction-based anomaly detection performance.

B. Perceptual Loss for Reconstruction and Anomaly Detection

Pixel-wise losses such as L1 and mean squared error (L2) are computationally simple but do not align well with human perception and often fail to represent semantic structure. Perceptual loss, introduced by Johnson et al. [8], measures similarity using deep feature representations extracted from pretrained convolutional networks, commonly VGG models. These features encode hierarchical structural information.

Zhang et al. [9] demonstrated that deep feature representations correlate more aligned with human perceptual evaluation than traditional metrics. Perceptual loss has shown strong performance in image reconstruction tasks such as super-resolution and style transfer [8].

In anomaly detection, perceptual supervision has been applied to improve sensitivity to find structural deviations. Shvetsova et al. [10] showed improved detection of subtle medical anomalies using perceptual autoencoders, while Tuluptceva et al. [11] introduced Relative Perceptual L1 loss for improved robustness to illumination changes. Applications in CT reconstruction [12], photovoltaic inspection [13], and embedding refinement [14] further demonstrate the value of perceptual features.

Despite these advances, the influence of perceptual feature depth and multi-scale supervision within GAN-based anomaly detection, particularly for X-ray cargo imagery, remains insufficiently explored.

Direct comparison with other anomaly detection methods is limited because many studies report results on different datasets and application domains. Therefore, this study focuses on controlled evaluation within the GANomaly framework on the CargoX dataset. The experiments systematically examine perceptual supervision at multiple feature depths and evaluate the effect of encoder architecture. This allows the impact of each design choice on reconstruction-based anomaly detection to be examined.

C. Research Gap

Although perceptual loss has been applied in image synthesis and anomaly detection across various domains [8], [10], [11], [13], its systematic application to X-ray cargo screening remains unexplored. Prior work has not compared perceptual feature depths, evaluated multi-scale aggregation, or quantified the impact of deeper encoder architectures for this domain.

This work addresses these gaps by systematically evaluating perceptual supervision at three feature depths (ReLU2_2, ReLU3_3, ReLU4_3), multi-scale perceptual aggregation, and encoder capacity effects (DCGAN vs ResNet50), providing practical guidance for selecting reconstruction objectives in GANomaly-based cargo screening systems.

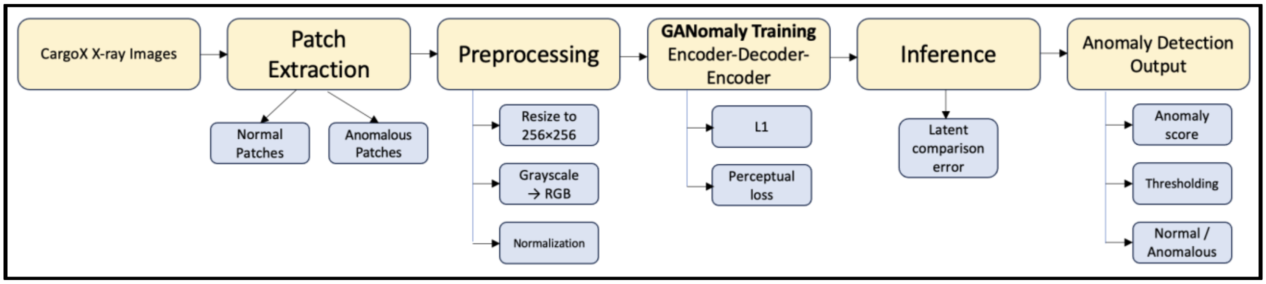

Fig. 1. Overview of the proposed anomaly detection pipeline, illustrating the main stages from CargoX X-ray image preprocessing through 256×256 patch extraction, GANomaly training with perceptual supervision, to final anomaly scoring using latent consistency error and decision-making via Youden's J threshold selection.

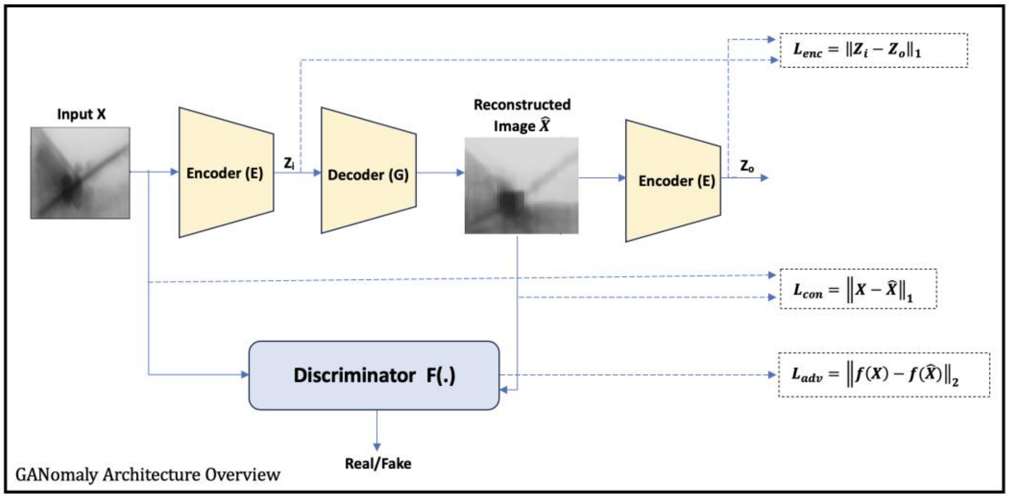
$L_{enc} = \|Z_i - Z_o\|_1$
Input X
Encoder (E)
Z_i
Decoder (G)
Reconstructed Image $\hat{X}$
Encoder (E)
Z_o
$L_{con} = \|X - \hat{X}\|_1$
Discriminator F(.)
$L_{adv} = \|f(X) - f(\hat{X})\|_2$
Real/Fake
GANomaly Architecture Overview

Fig. 2. Baseline GANomaly architecture for reconstruction-based anomaly detection. The model consists of an encoder-decoder-encoder generator and an adversarial discriminator. The first encoder maps the input image *X* to a latent representation Z_i, the decoder reconstructs the image $\hat{X}$, and the second encoder produces *Zo*. Training is guided by adversarial loss L_{adv}, pixel-wise contextual reconstruction loss L_{con}, and latent consistency loss L_{enc}. Anomaly detection is performed by measuring the discrepancy between Z_i and *Zo*.

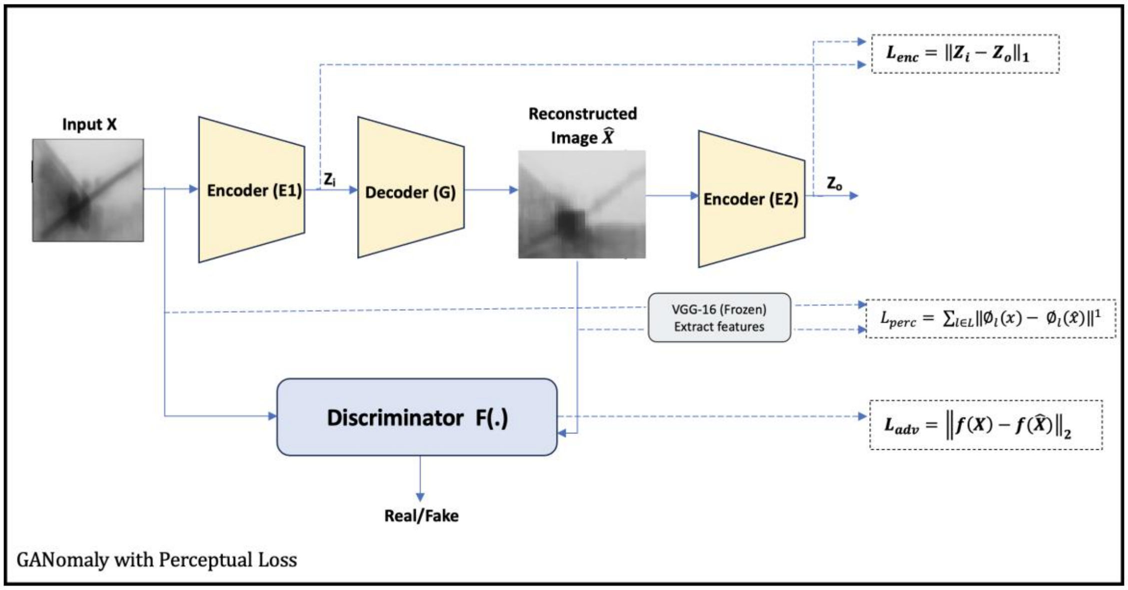
$L_{enc} = \|Z_i - Z_o\|_1$
Input X
Encoder (E1)
Z_i
Decoder (G)
Reconstructed Image $\hat{X}$
Encoder (E2)
Z_o
VGG-16 (Frozen) Extract features
$L_{perc} = \sum_{l \in L} \|\emptyset_l(x) - \emptyset_l(\hat{x})\|^1$
Discriminator F(.)
$L_{adv} = \|f(X) - f(\hat{X})\|_2$
Real/Fake
GANomaly with Perceptual Loss

Fig. 3. GANomaly with perceptual reconstruction supervision. The baseline GANomaly framework is modified by replacing the pixel-wise contextual reconstruction loss with the perceptual loss Lperc. Reconstruction quality is evaluated using deep feature representations extracted from a pretrained VGG-16 network. The adversarial loss Ladv and latent consistency loss Lenc remain unchanged. This modification shifts supervision from intensity-level similarity to feature-level structural similarity.

III. Methodology

A. Overview

Two controlled modifications are investigated: 1) replacing pixel-based reconstruction loss with perceptual supervision at different feature depths, and 2) replacing the original DCGAN encoder with a ResNet50 backbone to study the effect of encoder capacity. Fig. 2 illustrates the baseline architecture, and Fig. 3 shows the perceptual-loss modification. The full pipeline includes dataset preparation, patch extraction, preprocessing, GANomaly training with modified loss functions, and anomaly scoring at inference time, as shown in Fig. 1.

B. Dataset

Experiments are conducted on the CargoX X-ray cargo screening dataset [18]. The dataset contains X-ray images of cargo containers and trucks with anomalous regions annotated using polygon-based segmentation masks in JSON format. The annotations indicate anomaly locations but do not provide explicit normal labels. This structure makes CargoX suitable for unsupervised and semi-supervised anomaly detection, where models are trained primarily on normal samples.

C. Patch Extraction and Preprocessing

To construct fixed-size inputs for reconstruction-based learning, a patch extraction pipeline is applied. Anomalous patches are generated by identifying the largest annotated polygon in each image and extracting a 256×256 patch centered on that region. This ensures inclusion of the most prominent anomalous object. Normal patches are initially extracted using a sliding window of size 256×256 with a stride of 32 pixels. Patches overlapping any anomaly mask are excluded. From the remaining candidates, one non-overlapping patch per image is retained to reduce redundancy and maintain diversity in the training distribution.

All patches are stored in grayscale format and replicated to three channels during training to match the GANomaly input configuration (nc = 3, where nc denotes the number of channels). Pixel intensities are normalized according to the GANomaly implementation. This representation also ensures compatibility with pretrained VGG16 networks used for the perceptual loss.

D. Baseline GANomaly Architecture

The baseline model follows the GANomaly architecture proposed by Akçay et al. [2] as shown in Fig. 2. It consists of a generator and a discriminator trained adversarially. The generator adopts an encoder-decoder-encoder structure. The first encoder maps the input image x to a latent vector z_i. The decoder reconstructs the image $\hat{x}$, which is then processed by a second encoder to produce z_o. The discriminator distinguishes real images from reconstructed images and provides adversarial feedback to improve reconstruction realism.

E. Encoder Replacement with ResNet50

To investigate the effect of encoder capacity on anomaly detection performance, the original DCGAN encoder in GANomaly is replaced with a ResNet50 backbone. The ResNet50 encoder follows a residual learning design that enables deeper feature extraction while maintaining stable optimization. The network consists of a convolutional stem followed by four residual stages and global average pooling to produce a compact latent representation. The final feature tensor is mapped to the GANomaly latent space using a 1×1 convolution layer. All remaining components of GANomaly, including the decoder, discriminator, loss functions, and anomaly scoring strategy, remain unchanged to ensure controlled comparison with the baseline architecture.

F. Baseline Training Objective

The baseline GANomaly generator is trained using a weighted combination of three objective functions. An adversarial loss L_{adv} is employed to encourage the reconstructed images to resemble realistic visual patterns. A contextual loss L_{con}, defined as the L1 distance between x and its reconstruction $\hat{x}$. In addition, a latent consistency loss L_{enc}, defined as the L2 distance between the latent representations z_i and z_o, ensures consistency between the encoded features of the input and reconstructed images. The total generator loss is:

$$L_G = w_{adv} . L_{adv} + w_{con} . L_{con} + w_{enc} . L_{enc} \qquad (1)$$

G. GANomaly with Perceptual Loss

To improve semantic reconstruction quality, the pixel-based contextual loss is replaced with perceptual loss computed using a pretrained VGG16 network [8]. Feature maps are extracted from selected VGG16 layers, including ReLU2_2, ReLU3_3, and ReLU4_3. These layers represent different semantic levels of visual representation. ReLU2_2 captures low-level textures, ReLU3_3 captures mid-level structural patterns, and ReLU4_3 captures higher-level semantic information. This hierarchical selection enables systematic evaluation of how different levels of semantic abstraction affect reconstruction-based anomaly detection in X-ray cargo imagery. All VGG16 parameters are frozen during training. The perceptual loss is defined as:

$$L_{perc} = \sum_{l\in L} \|\emptyset_l(x) - \emptyset_l(\hat{x})\|^1 , \qquad (2)$$

where $\phi l(\cdot)$ denotes the activation of layer l in the VGG16 network. When perceptual loss is used, the generator loss becomes:

$$L_G = w_{adv} . L_{adv} + w_{perc} . L_{perc} + w_{enc} . L_{enc} \qquad (3)$$

This modification is applied to both the original GANomaly encoder and the ResNet50 encoder variants.

H. Anomaly Scoring

At inference time, anomaly detection is performed using latent representation discrepancy. Given an input image x, the anomaly score is computed as:

$$A(x) = \|z_i - z_o\|_2^2 . \qquad (4)$$

Higher scores indicate greater deviation from the learned normal distribution. Final decisions are obtained by thresholding anomaly scores using Youden's J statistic [5] derived from the ROC curve.

IV. Experimental Design

This section describes the experimental protocol used to evaluate GANomaly-based anomaly detection on the CargoX dataset. The study follows a controlled ablation strategy to evaluate perceptual reconstruction loss at different feature depths while keeping the remaining components fixed.

A. Dataset Split and Protocol

All experiments use the CargoX X-ray cargo screening dataset. The training set contains normal patches only, while the test set contains a balanced set of normal and anomalous patches. The training set contains 40000 normal patches. The test set contains 12000 normal patches and 11264 anomalous patches. This setup reflects realistic screening conditions, where anomalous samples are typically unavailable during training but must be detected at inference time.

B. Training Configuration and Hardware

All models are trained under identical optimization settings to ensure fair comparison. Experiments are conducted on an NVIDIA A40 GPU. The input resolution is 256×256, and the latent dimension is 512. Training uses the Adam optimizer (β_1=0.5, β_2=0.999) with a learning rate of 0.00005, batch size of 16, and 150 epochs. The adversarial loss and latent consistency loss hyperparameters are fixed to w_{adv}=1 and w_{enc}=1 across all experiments. Dataset split, preprocessing, GANomaly architecture, anomaly scoring method, and threshold selection (Youden's J statistic [5]) remain unchanged for all configurations. The computational cost of perceptual supervision was also examined. The average training time was approximately 22 minutes per epoch. Processing time during training was measured using a batch size of 16 patches. The baseline GANomaly model processed batches in approximately 40 milliseconds. The ReLU3_3 perceptual configuration required approximately 44 milliseconds per batch. The ResNet50 configuration required approximately 50 milliseconds per batch. These results indicate that perceptual supervision and encoder replacement introduce only modest additional computational cost during training.

C. Evaluation Metrics and Thresholding

For each test sample, a continuous anomaly score is produced from the latent consistency error. These scores are used to generate the ROC curve and compute ROC-AUC as a threshold-independent measure of separability between normal and anomalous samples. In addition, precision, recall, and F1-score are computed after converting scores to binary predictions using a single operating threshold. The threshold is selected using Youden’s J statistic, defined as

$$J = TPR - FPR \,, \tag{5}$$

where TPR (True Positive Rate) represents sensitivity or recall, and FPR (False Positive Rate) represents 1 − specificity. The threshold maximizing J on the ROC curve is used as the decision point.

Qualitative evaluation is also performed using visual comparisons of the input image, reconstructed image, and residual map to assess reconstruction behavior and anomaly localization.

D. Model Configurations

1) Baseline GANomaly: The baseline configuration follows the original GANomaly formulation [2], which employs an encoder-decoder-encoder generator and an adversarial discriminator. Reconstruction supervision is provided using pixel-wise contextual loss (L1) with w_{con}=50. Anomaly detection is performed based on the discrepancy between latent representations produced by the two encoders. This baseline serves as the primary reference for all comparisons.

2) Perceptual loss experiments: As described in the GANomaly with Perceptual Loss subsection above, perceptual loss is computed from VGG16 features at different depths. The GANomaly generator and discriminator architectures remain unchanged, and anomaly scoring continues to rely on latent representation discrepancy. Three single-scale feature depths are evaluated: ReLU2_2, ReLU3_3, and ReLU4_3, representing increasing abstraction levels. A multi-scale configuration is also tested by combining feature differences across multiple VGG16 layers into a single perceptual reconstruction objective. All perceptual experiments are trained using the same GPU regime as in the baseline, with a loss hyperparameter w_{perc}=15. The contextual L1 loss is not used in these runs. Evaluation is performed on the same test set and metrics described in Section C to ensure controlled comparison.

E. Comparative Protocol

All configurations are compared under identical conditions, including dataset split, preprocessing, architecture, anomaly scoring definition, threshold selection, and evaluation metrics. The ablation study includes: 1) the baseline GANomaly, 2) single-scale perceptual loss variants at three depths (ReLU2_2, ReLU3_3, ReLU4_3), and 3) multi-scale perceptual loss. The ablation study evaluates two types of architectural modification: reconstruction supervision (pixel-wise versus perceptual loss at different feature depths) and encoder replacement using a ResNet50 backbone, while all other training settings remain fixed. This design ensures that performance differences can be attributed directly to the proposed modifications.

V. Experimental Results

Table I summarizes the quantitative performance of all evaluated configurations, including the original GANomaly baseline, perceptual loss variants, and encoder replacement using ResNet50. The baseline GANomaly model achieves an ROC-AUC of 0.7182, which serves as the primary reference. Replacing the original DCGAN encoder with the ResNet50 improves representation quality and increases ROC-AUC to 0.73. This indicates stronger structural feature extraction compared to the baseline encoder. Replacing pixel-wise reconstruction with perceptual supervision leads to notable performance differences depending on the selected feature depth.

Among all models, the ReLU3_3 perceptual configuration achieves the best performance, with ROC-AUC of 0.7548 (5.1% improvement over baseline), recall of 0.7707, and F1-score of 0.7127 (12.7% improvement over baseline). When perceptual supervision is combined with the ResNet50 encoder, performance improves compared to ResNet50 alone but remains slightly below the DCGAN-based ReLU3_3 model. The ResNet50 + ReLU3_3 configuration produces clearer residual responses and improves anomaly separation compared to ResNet50 alone. This demonstrates that mid-level perceptual features significantly improve anomaly detection compared to pixel-wise reconstruction. The ReLU4_3 model also improves over the baseline but does not exceed ReLU3_3. The shallow ReLU2_2 configuration shows weaker discrimination performance. This suggests that low-level texture features are less effective for separating anomalies. Such features do not capture enough structural information in complex cargo X-ray images. The multi-scale perceptual model (ROC-AUC 0.7157) does not outperform the best single-scale configuration (ReLU3_3: 0.7548) and performs slightly below the baseline (0.7182). Combining features from multiple depths spreads residual responses across both normal and anomalous structures. This reduces the model's ability to focus on abnormal regions, which weakens discrimination between normal structural variation and true anomalies.

TABLE. I. QUANTITATIVE PERFORMANCE ON CARGOX

Model	ROC-AUC	Precision	Recall	F1-score	Improvement vs Baseline
Baseline (DCGAN)	0.7182	0.6391	0.6254	0.6322	
Perceptual (ReLU2_2)	0.6983	0.6401	0.5950	0.6167	- 2.8%
Perceptual (ReLU3_3)	**0.7548**	**0.6628**	**0.7707**	**0.7127**	**+ 5.1%**
Perceptual (ReLU4_3)	0.7220	0.6470	0.7167	0.6800	+ 0.5%
Perceptual (Multi-scale)	0.7157	0.6563	0.6045	0.6293	- 0.3%
Baseline (ResNet50)	0.7312	0.7023	0.6085	0.6521	+ 1.8 %
ResNet50 + perceptual (Relu3_3)	0.7517	0.6690	0.6847	0.6768	+ 4.7%

Overall, Perceptual ReLU3_3 (DCGAN encoder) achieves the best overall performance with ROC-AUC 0.7548 (+5.1% vs baseline 0.7182), recall 0.7707 (+23.2% vs 0.6254), and F1 0.7127 (+12.7% vs 0.6322). This demonstrates that mid-level perceptual features significantly improve anomaly detection compared to pixel-wise reconstruction. Replacing the encoder with ResNet50 improves the baseline ROC-AUC from 0.7182 to 0.7312 and increases precision to 0.7023, but recall decreases to 0.6085. Adding ReLU3_3 perceptual loss with ResNet50 increases ROC-AUC to 0.7517, but it does not surpass the DCGAN-based ReLU3_3 model in recall (0.6847) or F1-score (0.6768). The ResNet50 backbone improves ROC-AUC and precision compared to the original baseline. However, the perceptual ReLU3_3 configuration using the original encoder remains the strongest model in recall and F1-score.

Training stability was monitored throughout the experiments. Convergence behavior varied significantly across configurations. The ReLU3_3 configuration converged rapidly at epoch 7, while ReLU2_2 required 100 epochs, ReLU4_3 required 21 epochs, and ResNet50 required 149 epochs.

Qualitative analysis supports these observations (Fig. 4). The baseline model (top row) produces visually realistic reconstructions. However, subtle anomalies are often only partially reconstructed, which results in weak residual responses. The ReLU3_3 model (middle row) generates more abstract reconstructions that emphasize structural differences. This leads to stronger and more concentrated residual activation near anomalous regions. In contrast, the multi-scale model (bottom row) exhibits diffuse residual responses distributed across multiple structural patterns that reduce anomaly separability.

With the ResNet50 encoder (Fig. 5), reconstructions appear smoother and more stable, and residual maps emphasize structural differences. When combined with perceptual supervision (ReLU3_3), residual responses become more structured, but they are less selective than the DCGAN-based ReLU3_3 model, which aligns with the lower recall and F1-score. Additional reconstruction examples are provided in the Appendix.

To analyze model behavior across different operating thresholds, the Precision–Recall curve of the best perceptual configuration is shown in Fig. 6. The curve illustrates how precision changes as recall increases. Precision remains high at low recall levels and gradually decreases as more samples are classified as anomalies. The PR-AUC value of 0.711 indicates consistent anomaly discrimination across a wide range of threshold values.

Fig. 7 shows the ROC curves for all seven configurations. The curves confirm that Perceptual ReLU3_3 achieves the highest true positive rate at any given false positive rate. This shows superior discrimination capability. The optimal operating points (marked in red) are determined by Youden's J statistic, balancing sensitivity and specificity for each model. In contrast, ReLU2_2 falls below the baseline across all operating points. This confirms that shallow features are unsuitable for this task.

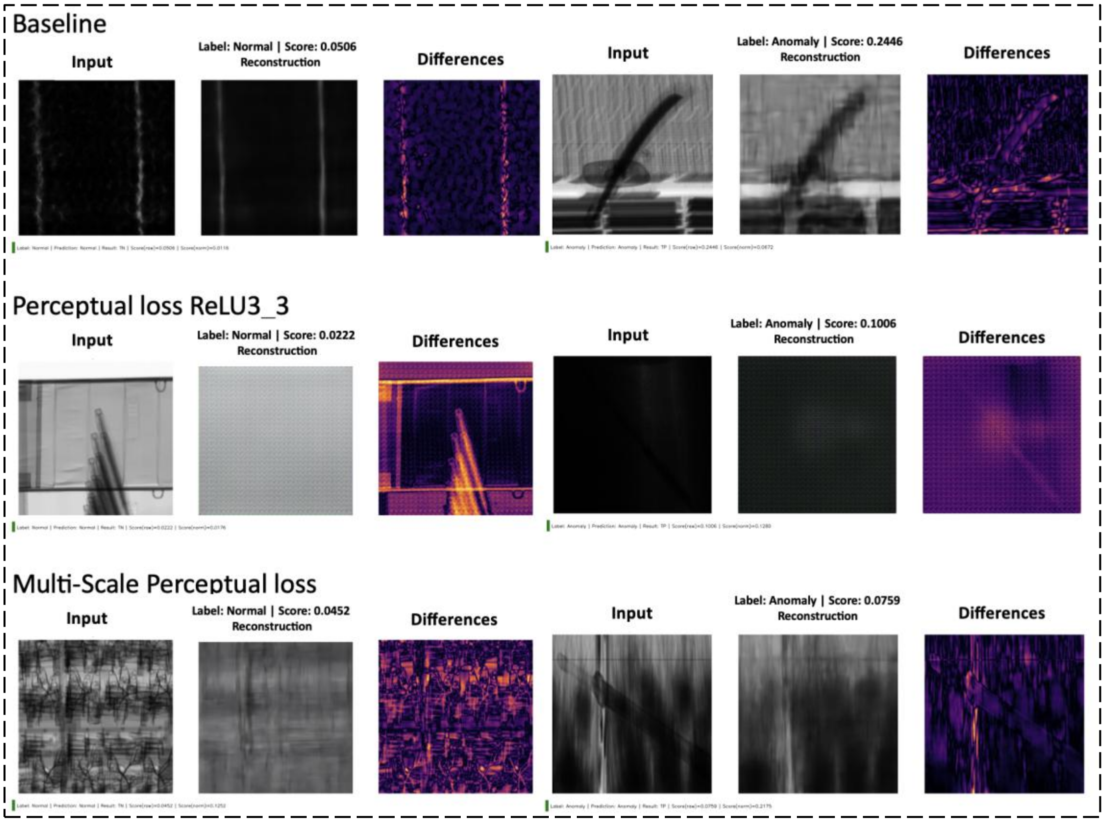
Baseline
Input
Label: Normal | Score: 0.0506
Reconstruction
Differences
Input
Label: Anomaly | Score: 0.2446
Reconstruction
Differences
Label: Normal | Prediction: Normal | Result: TN | Score(raw)=0.0506 | Score(norm)=0.0116
Label: Anomaly | Prediction: Anomaly | Result: TP | Score(raw)=0.2446 | Score(norm)=0.0672
Perceptual loss ReLU3_3
Input
Label: Normal | Score: 0.0222
Reconstruction
Differences
Input
Label: Anomaly | Score: 0.1006
Reconstruction
Differences
Label: Normal | Prediction: Normal | Result: TN | Score(raw)=0.0222 | Score(norm)=0.0176
Label: Anomaly | Prediction: Anomaly | Result: TP | Score(raw)=0.1006 | Score(norm)=0.1280
Multi-Scale Perceptual loss
Input
Label: Normal | Score: 0.0452
Reconstruction
Differences
Input
Label: Anomaly | Score: 0.0759
Reconstruction
Differences
Label: Normal | Prediction: Normal | Result: TN | Score(raw)=0.0452 | Score(norm)=0.1252
Label: Anomaly | Prediction: Anomaly | Result: TP | Score(raw)=0.0759 | Score(norm)=0.2175

Fig. 4. Qualitative comparison of reconstruction behavior across different training objectives. Each row corresponds to a different model configuration: baseline GANomaly (top row), perceptual loss using VGG16 ReLU3_3 features (middle row), and multi-scale perceptual loss (bottom row). The baseline produces visually realistic reconstructions, but with limited contrast for anomalies. The ReLU3_3 model generates more abstract reconstructions while emphasizing structural differences in the residual map. The multi-scale model shows diffuse residual responses distributed across multiple structural regions.

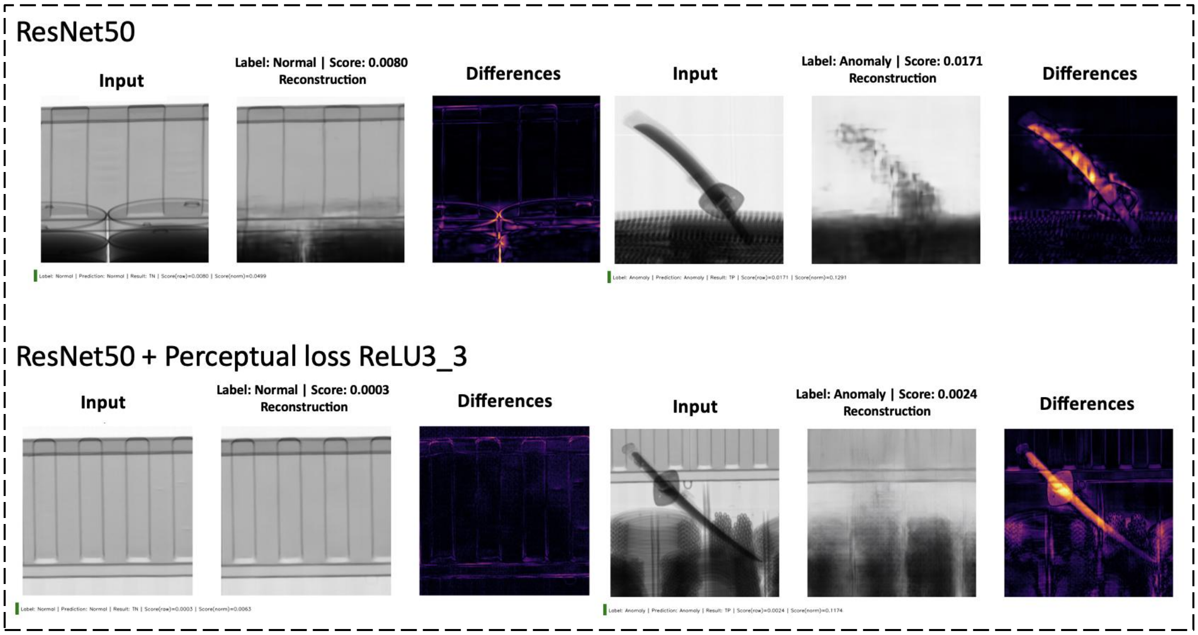
ResNet50
Input
Label: Normal | Score: 0.0080
Reconstruction
Differences
Input
Label: Anomaly | Score: 0.0171
Reconstruction
Differences
Label: Normal | Prediction: Normal | Result: TN | Score(raw)=0.0080 | Score(norm)=0.0499
Label: Anomaly | Prediction: Anomaly | Result: TP | Score(raw)=0.0171 | Score(norm)=0.1291
ResNet50 + Perceptual loss ReLU3_3
Input
Label: Normal | Score: 0.0003
Reconstruction
Differences
Input
Label: Anomaly | Score: 0.0024
Reconstruction
Differences
Label: Normal | Prediction: Normal | Result: TN | Score(raw)=0.0003 | Score(norm)=0.0063
Label: Anomaly | Prediction: Anomaly | Result: TP | Score(raw)=0.0024 | Score(norm)=0.1174

Fig. 5. Qualitative reconstruction comparison using the ResNet50 encoder. The top row shows the baseline ResNet50 model, while the bottom row shows ResNet50 combined with perceptual supervision using VGG16 ReLU3_3 features. For each configuration, a normal sample (left tri plet) and an anomalous sample (right triplet) are presented, including the input image, reconstructed image, and residual map. The ResNet50 backbone produces smoother reconstructions and highlights structural differences, while perceptual supervision improves residual organization around anomalous regions.

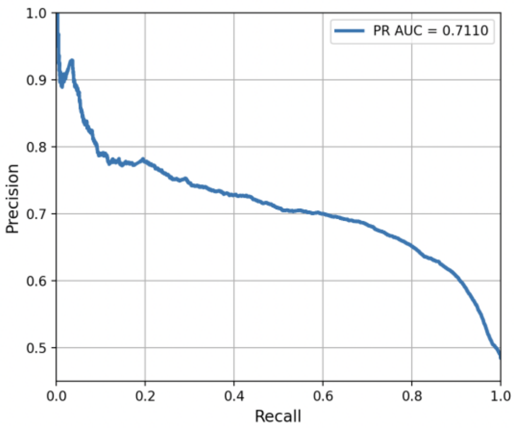

Fig. 6. Precision–Recall curve of the best perceptual configuration (ReLU3_3) evaluated on the CargoX test set.

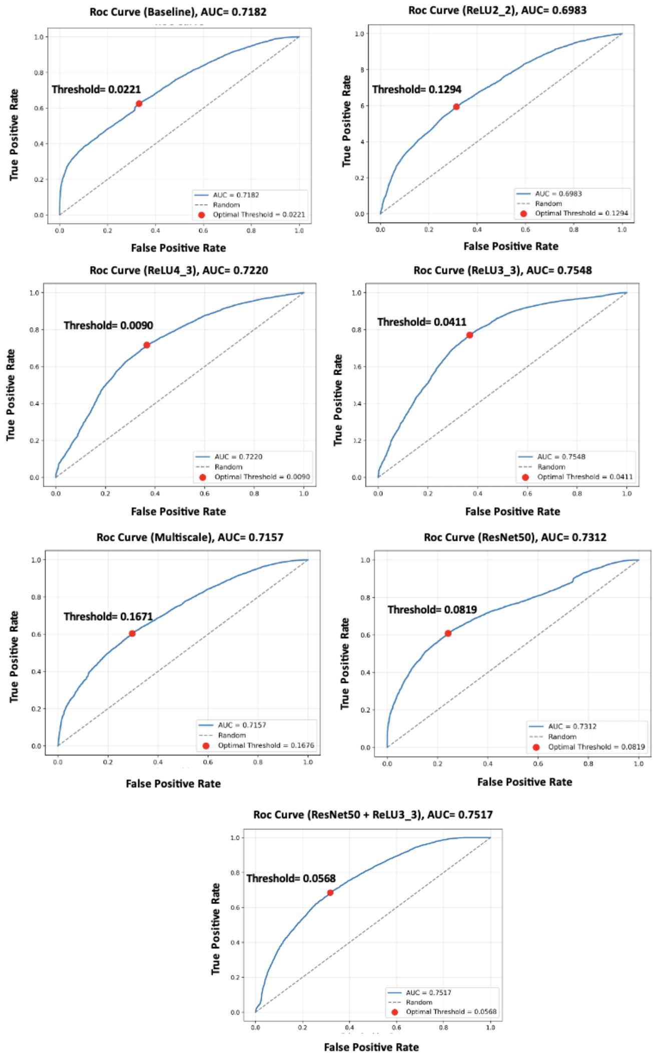

Fig. 7. ROC curves for all evaluated configurations on the CargoX test set. The optimal operating point for each model (red dot) is determined by Youden's J statistic. Perceptual ReLU3_3 achieves the highest AUC (0.7548), while ReLU2_2 (0.6983) and Multi-scale (0.7157) perform below the baseline (0.7182). ResNet50 variants show improved performance in low false positive regions.

The ResNet50-based models show steeper slopes in the low false positive region. This indicates better precision at conservative thresholds and aligns with the results in Table I. The multi-scale curve closely tracks the baseline. This shows that feature aggregation does not improve discrimination beyond careful depth selection.

VI. Discussion

The results demonstrate that perceptual supervision improves anomaly detection performance compared to pixel-based reconstruction. Mid-level perceptual features from ReLU3_3 provide the most effective representation by improving ROC-AUC by 5.1% (from 0.7182 to 0.7548) and F1-score by 12.7% (from 0.6322 to 0.7127). They capture the structural organization of objects and layouts. At the same time, they are not dominated by fine texture noise. They also avoid being too abstract or overly semantic. This balance enables clearer separation between normal structural variation and anomalous deviations. The improvement is obtained under the same dataset split, training protocol, and evaluation procedure. This ensures a fair comparison between models.

A quantitative analysis of classification errors is also conducted. For the best perceptual configuration (ReLU3_3), the model produces 8680 true positives and 7585 true negatives. The model also produces 4415 false positives and 2583 false negatives on the test set. False positives mainly occur in visually complex cargo regions where structural patterns resemble anomalous objects. False negatives occur when anomalies are small, partially occluded, or visually similar to normal cargo structures.

Multi-scale perceptual supervision (ROC-AUC 0.7157) performs slightly below the baseline (0.7182, -0.35% degradation) and substantially below the best single-scale configuration ReLU3_3 (0.7548). Combining features from multiple depth levels (ReLU2_2, ReLU3_3, ReLU4_3) appears to weaken the discriminative power of mid-level features rather than enhancing it. This suggests that feature depth selection is more critical than multi-scale aggregation for this task.

The failure of multi-scale supervision can be explained by the characteristics of features at different depths. Early layers (ReLU2_2) capture low-level textures that vary strongly even in normal cargo. This variation is caused by differences in materials and orientations in X-ray imagery. As a result, these features introduce noise rather than providing a discriminative signal. Deep layers (ReLU4_3) encode high-level semantic concepts that may be too abstract for precise anomaly localization. Mid-level features (ReLU3_3) capture structural patterns at an optimal level of abstraction. They are specific enough for anomaly discrimination while remaining robust to normal cargo variations.

When these levels are combined through averaging, the discriminative mid-level signal is reduced by noisy shallow features and overly abstract deep features. This aligns with Johnson et al. [8], who showed that optimal perceptual feature depth depends on the specific visual task. For X-ray cargo anomaly detection, ReLU3_3 alone provides the best representation. This shows that careful depth selection is more effective than multi-scale aggregation.

Introducing the ResNet50 encoder increases ROC-AUC from 0.7182 to 0.7312 (+1.8%) and improves precision from 0.6391 to 0.7023 (+9.9%). However, recall decreases from 0.6254 to 0.6085 (-2.7%). This shows that while ResNet50 produces fewer false positives, it also misses more true anomalies. The F1-score increases from 0.6322 to 0.6521 (+3.1%). This indicates that the precision gain outweighs the recall loss for this application.

Combining ResNet50 with perceptual supervision (ReLU3_3) achieves ROC-AUC of 0.7517, precision of 0.6690, and recall of 0.6847. This configuration provides a balanced trade-off. It maintains precision improvements from ResNet50 while recovering recall performance through perceptual supervision. However, it does not surpass the DCGAN-based ReLU3_3 model in overall recall (0.7707) or F1-score (0.7127). Qualitative results show clearer structural reconstructions and more concentrated residual responses around anomalous regions. This reflects improved localization while maintaining stable reconstruction behavior.

Failure analysis (Fig. 8) shows two main challenges. Some anomalies have structures that look similar to normal cargo. These cases are difficult for the model to detect. In addition, some normal scenes are visually complex. They contain many overlapping structures and edges. These scenes can produce strong residual responses, even though they are not anomalous. These observations suggest that structural similarity and scene variability remain fundamental challenges for reconstruction-based anomaly detection in X-ray cargo imagery.

Failure cases observed with the ResNet50 encoder (Fig. 9) follow patterns similar to the baseline models. Some anomalies produce weak residual responses, while structurally complex normal scenes generate strong activation. Increasing encoder capacity alone does not fully resolve these limitations.

Overall, the findings show that mid-level feature supervision significantly improves detection compared to pixel-wise reconstruction, but the choice of feature depth is critical. Mid-level features (ReLU3_3) capture structural information at the optimal abstraction level for cargo anomaly detection, while shallow features (ReLU2_2) and multi-scale aggregation do not provide the same benefit. Encoder replacement with ResNet50 improves ROC-AUC and precision but does not surpass the DCGAN-based ReLU3_3 configuration in recall or F1-score.

VII. Conclusion

This study investigated reconstruction-based anomaly detection for X-ray cargo imagery using the GANomaly framework. Through systematic evaluation of perceptual supervision at different feature depths and encoder architectures, it was demonstrated that mid-level perceptual features (VGG16 ReLU3_3) significantly outperform pixel-wise reconstruction. This improves ROC-AUC by 5.1%, F1-score by 12.7%, and recall by 23.2%.

Four key findings emerge: 1) Feature-level reconstruction objectives substantially improve anomaly detection compared to pixel-level losses; 2) Mid-level features (ReLU3_3) provide the optimal balance between semantic structure and fine-grained detail for cargo X-ray anomalies; 3) The ResNet50 encoder improves ROC-AUC (+1.8%) and precision (+9.9%) but reduces recall (-2.7%). This reflects a precision–recall trade-off that favors fewer false alarms; 4) Aggregating features from multiple depths does not improve performance beyond the best single depth, suggesting that depth selection is more critical than multi-scale aggregation.

Overall, these findings establish feature-level reconstruction as a promising direction for X-ray cargo anomaly detection and provide practical guidance for selecting perceptual supervision depths in reconstruction-based anomaly detection frameworks.

VIII. Limitations and Future Work

Despite the improvements achieved through perceptual supervision, several limitations remain. First, reconstruction-based anomaly detection relies on differences between input and reconstruction, which may be weak when anomalous objects share structural similarity with normal cargo. Thin, low-contrast, or partially occluded anomalies remain difficult to detect. Second, the perceptual features used in this study are extracted from a VGG16 network pretrained on natural images (ImageNet). These features may not fully capture domain-specific structures present in X-ray imagery. X-ray images differ from natural images in several ways. The intensity distribution and texture patterns differ from those in natural images. Edges in X-ray images represent material density rather than surface boundaries. Despite this domain mismatch, the experimental results show that perceptual supervision still improves anomaly detection performance. This suggests that mid-level features learned from natural images can transfer to X-ray reconstruction tasks. However, training feature extractors directly on X-ray images may further improve performance. Third, anomaly scoring is based solely on latent representation discrepancy, which may not fully exploit spatial information available in residual maps. Fourth, the latent dimensionality was fixed at 512 in all experiments. The effect of different latent representation sizes was not evaluated in this study. Fifth, experiments were conducted using single training runs due to computational constraints. However, the consistent performance patterns across configurations (Table I) and substantial 5.1% improvement suggest robust findings.

Future work may explore domain-adaptive perceptual feature extractors trained directly on X-ray data. Incorporating attention mechanisms could help focus reconstruction learning on structurally important regions. Wavelet decomposition can also be examined to perform anomaly detection at different frequency sub-band levels.

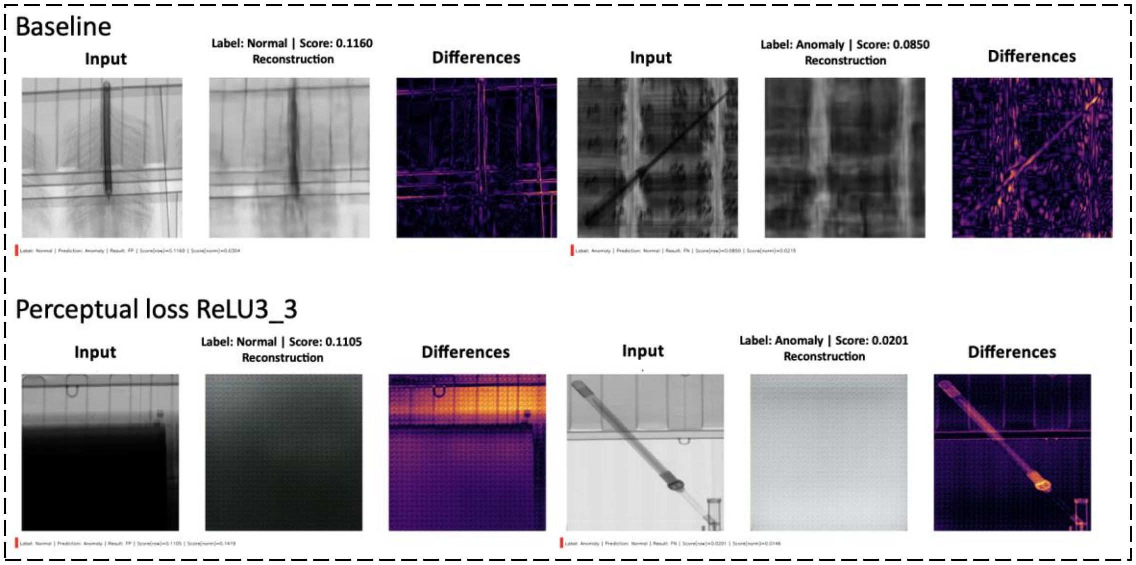

Fig. 8. Failure case comparison between the baseline GANomaly model (top row) and the perceptual loss model using VGG16 ReLU3_3 features (bottom row). For each model, a false positive example (left triplet) and a false negative example (right triplet) are shown, including the input image, reconstructed image, and residual map. In the false positive cases, structurally complex but normal regions produce strong residual responses, leading to incorrect anomaly predictions. In the false negative cases, anomalous structures generate weak or diffuse residual activation, resulting in low anomaly scores and missed detections.

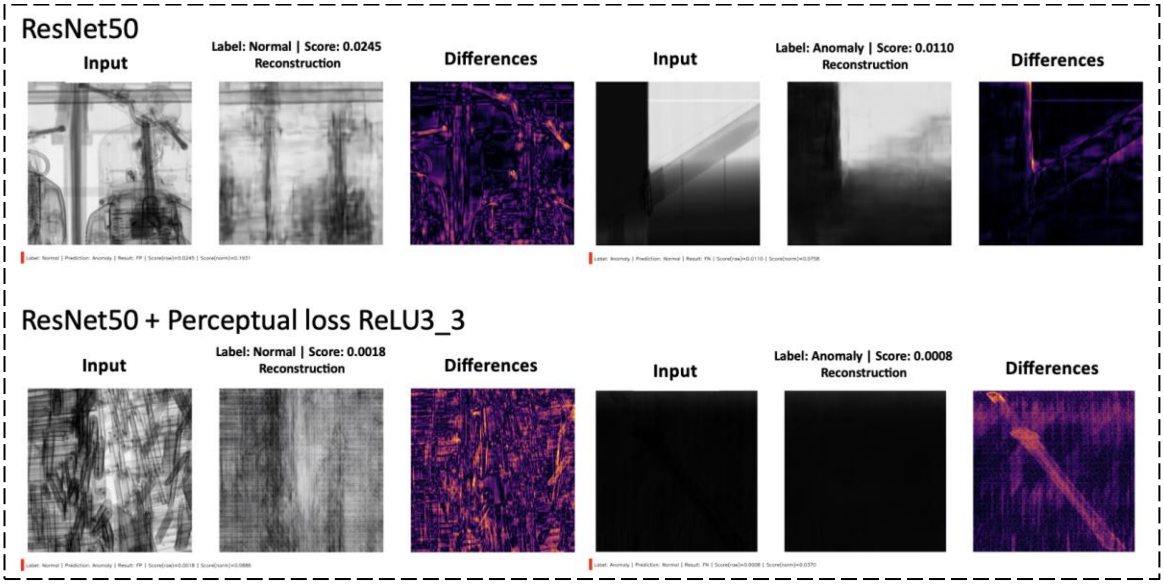
ResNet50
Input
Label: Normal | Score: 0.0245
Reconstruction
Differences
Input
Label: Anomaly | Score: 0.0110
Reconstruction
Differences
ResNet50 + Perceptual loss ReLU3_3
Input
Label: Normal | Score: 0.0018
Reconstruction
Differences
Input
Label: Anomaly | Score: 0.0008
Reconstruction
Differences

Fig. 9. Failure cases observed with the ResNet50 encoder configurations. The top row corresponds to the ResNet50 baseline, and the bottom row shows ResNet50 with perceptual supervision. Each row includes a false positive example (left triplet) and a false negative example (right triplet). Structurally complex normal scenes can generate strong residual activation leading to false positives, while subtle or low-contrast anomalies may produce weak responses resulting in missed detections.

REFERENCES

[1] N. Jaccard et al., "Detection of concealed cars in complex cargo X-ray imagery using deep learning," J. X-Ray Sci. Technol., vol. 25, no. 3, pp. 323–339, 2017. https://doi.org/10.48550/arXiv.1606.08078

[2] S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in Proc. Asian Conf. Comput. Vis., 2018. [Online]. Available: https://arxiv.org/abs/1805.06725

[3] T. Schlegl et al., "Unsupervised anomaly detection with GANs to guide marker discovery," in Proc. Med. Imag. Deep Learn. (MIDL), 2017. [Online]. Available: https://arxiv.org/abs/1703.05921

[4] E. Esme et al., "A literature review on deep learning algorithms for analysis of X-ray images," Int. J. Mach. Learn. Cybern., 2024, doi: 10.1007/s13042-023-01961-z

[5] W. J. Youden, “Index for rating diagnostic tests,” Cancer, vol. 3, no. 1, pp. 32-35, Jan. 1950, doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

[6] B. Gaikwad et al., "Self-supervised anomaly detection and localization for X-ray cargo images: Generalization to novel anomalies," Eng. Appl. Artif. Intell., vol. 140, p. 109675, 2025. https://doi.org/10.1016/j.engappai.2024.109675

[7] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," arXiv preprint arXiv:1901.03407, 2019. https://arxiv.org/abs/1901.03407

[8] Johnson, J., Alahi, A., Fei-Fei, L. (2016). “Perceptual losses for real-time style transfer and super-resolution”. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9906. Springer, Cham. https://doi.org/10.1007/978-3-319-46475-6_43

[9] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 586-595, doi: 10.1109/CVPR.2018.00068.

[10] N. Shvetsova, B. Bakker, I. Fedulova, H. Schulz and D. V. Dylov, "Anomaly detection in medical imaging with deep perceptual autoencoders," in IEEE Access, vol. 9, pp. 118571-118583, 2021, doi: 10.1109/ACCESS.2021.3107163.

[11] Tuluptceva, N., Bakker, B., Fedulova, I., Konushin, A. (2020). “Perceptual image anomaly detection”. In: Palaiahnakote, S., Sanniti di Baja, G., Wang, L., Yan, W. (eds) Pattern Recognition. ACPR 2019. Lecture Notes in Computer Science(), vol 12046. Springer, Cham. https://doi.org/10.1007/978-3-030-41404-7_12

[12] Théo Leuliet, Voichiţa Maxim, Françoise Peyrin, Bruno Sixou. “Combining conditional GAN with VGG perceptual loss for bones CT image reconstruction”. 16th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine (Fully3D), Jul 2021, Leuven, Belgium. ⟨hal-03706167⟩

[13] N. Drir, A. Mellit, M. Bettayeb and M. Dhimish, "Enhanced photovoltaic defect detection Using perceptual loss in DCGAN and VGG16-Integrated models on electroluminescence images," in IEEE Journal of Photovoltaics, vol. 15, no. 6, pp. 759-769, Nov. 2025, doi: 10.1109/JPHOTOV.2025.3542829.

[14] Pihlgren, G.G., Sandin, F., & Liwicki, M. (2020). “Improving image autoencoder embeddings with perceptual Loss. 2020 International Joint Conference on Neural Networks (IJCNN), 1-7.

[15] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[16] A. V and L. Shine, "Advancing anomaly detection in industrial systems: a comparative study of autoencoder, CNN, and ResNet50," 2025 6th International Conference on Control, Communication and Computing (ICCC), Thiruvanathapuram, India, 2025, pp. 1-6, doi: 10.1109/ICCC64910.2025.11077266.

[17] Ramoliya, D., & Ganatra, A. (2025). “Enhancing anomaly detection performance using ResNet50 and BiLSTM networks on benchmark datasets”. International Journal of Electrical and Computer Engineering (IJECE), 15(4), 3727-3736. doi:http://doi.org/10.11591/ijece.v15i4.pp3727-3736

[18] T. Viriyasaranon, S.-H. Chae, and J.-H. Choi, "MFA-net: object detection for complex X-ray cargo and baggage security imagery," PLOS ONE, vol. 17, no. 9, pp. 1–19, 2022. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0272961

APPENDIX

A. Additional Reconstruction Examples

Fig. 10 shows representative reconstruction results of the perceptual loss ReLU3_3 model. Each example includes the input image, reconstructed image, and residual map.

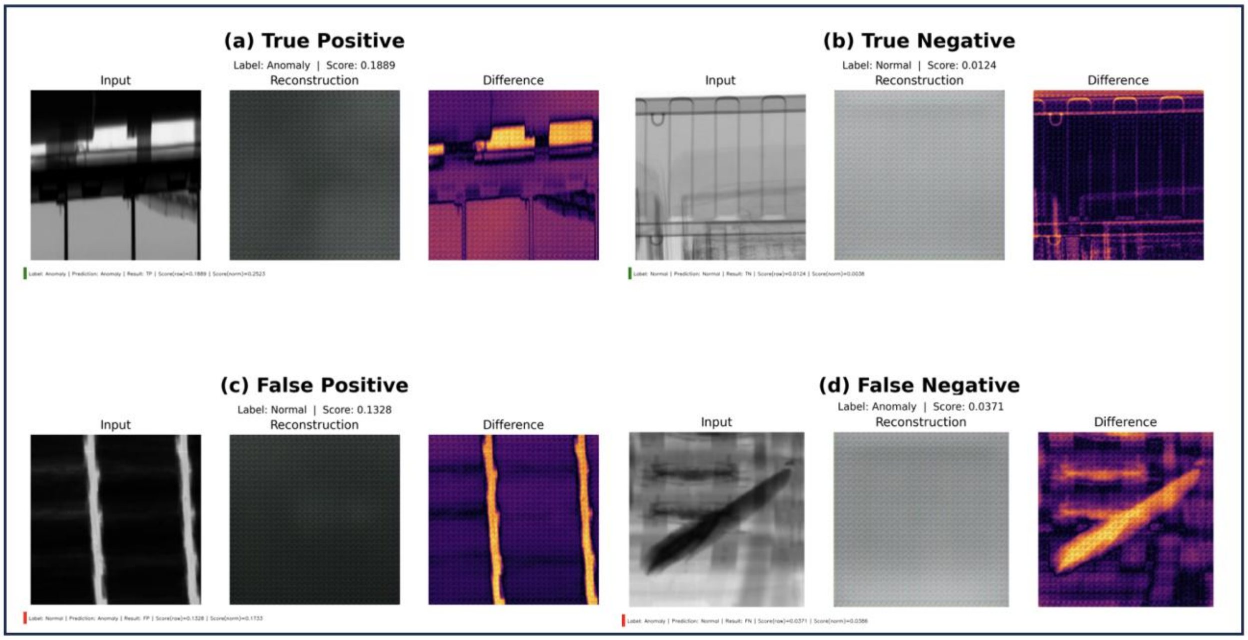
(a) True Positive
Label: Anomaly | Score: 0.1889
Input
Reconstruction
Difference
Label: Anomaly | Prediction: Anomaly | Result: TP | Score(raw)=0.1889 | Score(norm)=0.2523
(b) True Negative
Label: Normal | Score: 0.0124
Input
Reconstruction
Difference
Label: Normal | Prediction: Normal | Result: TN | Score(raw)=0.0124 | Score(norm)=0.0038
(c) False Positive
Label: Normal | Score: 0.1328
Input
Reconstruction
Difference
Label: Normal | Prediction: Anomaly | Result: FP | Score(raw)=0.1328 | Score(norm)=0.1733
(d) False Negative
Label: Anomaly | Score: 0.0371
Input
Reconstruction
Difference
Label: Anomaly | Prediction: Normal | Result: FN | Score(raw)=0.0371 | Score(norm)=0.0386

Fig. 10. Qualitative reconstruction results of the ReLU3_3 perceptual model. Each case shows the input image, reconstructed image, and Residual map for representative true positive, true negative, false positive, and false negative predictions.

Fig. 11 shows representative reconstruction results of the ResNet50 encoder configuration. Each example includes the input patch, reconstructed image, and residual map.

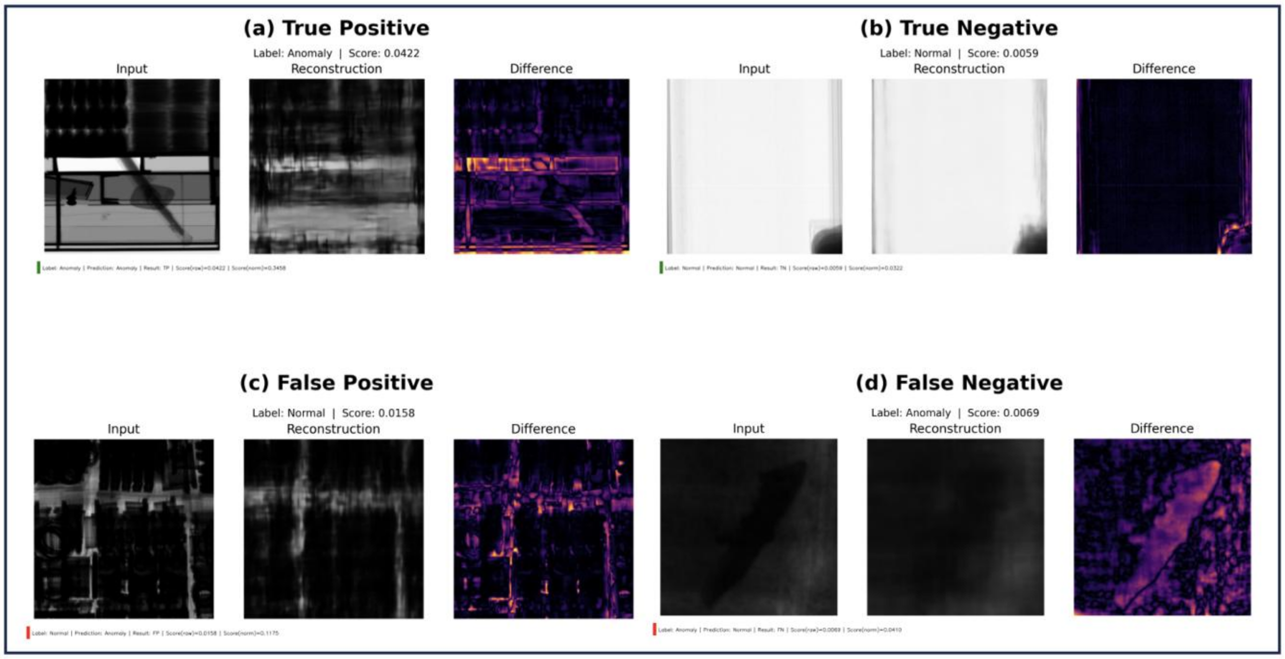
(a) True Positive
Label: Anomaly | Score: 0.0422
Input
Reconstruction
Difference
Label: Anomaly | Prediction: Anomaly | Result: TP | Score(raw)=0.0422 | Score(norm)=0.3458
(b) True Negative
Label: Normal | Score: 0.0059
Input
Reconstruction
Difference
Label: Normal | Prediction: Normal | Result: TN | Score(raw)=0.0059 | Score(norm)=0.0322
(c) False Positive
Label: Normal | Score: 0.0158
Input
Reconstruction
Difference
Label: Normal | Prediction: Anomaly | Result: FP | Score(raw)=0.0158 | Score(norm)=0.1175
(d) False Negative
Label: Anomaly | Score: 0.0069
Input
Reconstruction
Difference
Label: Anomaly | Prediction: Normal | Result: FN | Score(raw)=0.0069 | Score(norm)=0.0410

Fig. 11. Qualitative reconstruction results of the ResNet50 encoder configuration. The figure illustrates representative true positive , true negative, false positive, and false negative cases.