CargoScan: One-Shot Anomaly Detection in X-Ray Imaging

Anamika Pati
M. Tech CSE, IIIT Naya Raipur
anamika24300@iiitnr.edu.in

Avantika Singh
Assistant Professor DSAI, IIIT Naya Raipur
avantika@iiitnr.edu.in

***Abstract*—This project introduces a Siamese Neural Network (SNN) designed to detect anomalies in X-ray cargo images by distinguishing between two types of samples: positive images, which contain items of interest, and negative images, which are clear of such items. The model uses a pre-trained ResNet-18 as a shared feature extractor to process image pairs and learns to differentiate between them using contrastive loss.**

The dataset organized on Google Drive includes PNG images labeled as either positive or negative. After preprocessing the images and creating pairs, the model is trained for 20 epochs, reaching an accuracy of around 94%.

Performance is evaluated using a confusion matrix, which shows the model's strong ability to recognize both similar and dissimilar image pairs. This makes it especially useful for one-shot learning applications in cargo screening.

Overall, the approach shows promising potential for improving automated threat detection systems. Future improvements could come from better validation techniques, tuning hyperparameters and applying data augmentation.

***Index Terms*—Siamese Neural Network, Anomaly Detection, X-ray Cargo Images, Contrastive Loss, ResNet-18, One-shot Learning, Few-shot Learning, Feature Embedding, Image Similarity, Deep Learning, Cargo Inspection, Computer Vision, Confusion Matrix, PyTorch, Transfer Learning**

I. Introduction

The exponential growth in global trade has led to an increasing reliance on automated systems for cargo inspection and security screening. X-ray imaging has emerged as a key modality in the detection of contraband and prohibited items within cargo shipments. However, manual inspection of X-ray cargo images is labor intensive, error prone, and not scalable to the volume of containers processed at international ports. Traditional machine learning models for automated analysis often require large, balanced datasets and perform suboptimally when faced with class imbalance or limited samples of anomalous instances.

To address these limitations, this paper proposes an anomaly detection framework based on a *Siamese Neural Network (SNN)* architecture designed for *one-shot learning*. Unlike conventional classification networks, Siamese networks learn a similarity function between image pairs, enabling the detection of anomalies with minimal labeled data. This approach is particularly advantageous in real world scenarios where positive samples (i.e., images containing items of interest) are rare compared to abundant negative samples.

The proposed model employs a pre-trained *ResNet-18* as a shared feature extractor for processing input image pairs. Each input image is passed through the backbone to obtain a compact embedding vector. The *Euclidean distance* between these vectors is computed, and training is guided by a *contrastive loss function*, which minimizes the distance between similar pairs and maximizes the distance between dissimilar pairs. This allows the network to learn a discriminative embedding space where similar images are close together and dissimilar images are far apart.

The dataset comprises two classes of X-ray cargo images—positive samples containing anomalies and negative samples without anomalies—stored in separate Google Drive directories within `CargoX Dataset`. Data loading and preprocessing are handled by custom PyTorch dataset classes. During training, the model receives a balanced set of image pairs similar and dissimilar generated dynamically at each epoch. All images are resized to 224×224 pixels and normalized using the ImageNet mean and standard deviation.

The model is trained for 20 epochs using the Adam optimizer and evaluated in terms of *classification accuracy*, *contrastive loss*, and a *confusion matrix*. Experimental results demonstrate that the network converges effectively, with training accuracy improving from 50% in the initial epoch to approximately 93–94% by the final epoch. The confusion matrix reveals a high true positive and true negative rate, confirming the model's ability to discriminate between similar and dissimilar image pairs.

This work demonstrates the potential of Siamese networks for anomaly detection in X-ray cargo images, offering a data efficient alternative to traditional supervised learning. Future enhancements could include the introduction of a validation set, hyperparameter optimization, and advanced data augmentation techniques to improve performance. The proposed approach provides a scalable

and accurate foundation for enhancing automated cargo inspection systems.

II. Related Works

Gaikwad et al.[1] addressed the challenge of contraband detection in X-ray cargo images by proposing a self-supervised anomaly detection framework that eliminates the need for labeled anomaly data. They designed an encoder-decoder classifier architecture trained in two phases: first to reconstruct normal images while learning discriminative features using a modified triplet loss, and then to classify anomalies using binary cross entropy loss. To support their method, they introduced a physically realistic synthetic dataset generated from 3D cargo models and X-ray attenuation data. Their model, trained only on simple synthetic anomalies, demonstrated strong performance on complex, unseen anomalies across synthetic and real world dataset (CargoX), highlighting its robustness and generalization capabilities for real world cargo screening.

Madan et al. [2] introduced the Self Supervised Masked Convolutional Transformer Block (SSMCTB), a novel architectural component designed to enhance anomaly detection across diverse domains, including industrial inspection, medical imaging, and both RGB and thermal video surveillance. SSMCTB incorporates masked 2D and 3D convolutions to reconstruct occluded regions using contextual clues, overcoming the limitations of traditional CNNs in modeling global feature arrangements. It integrates a multi-head self attention transformer for channel wise modulation of features and employs a robust Huber loss function to improve resilience against outliers. The block is plug and play and self contained, enabling easy integration into various CNN, transformer, and hybrid models. Experimental evaluations across benchmarks such as MVTec AD, BRATS, Avenue, and ShanghaiTech demonstrate that SSMCTB significantly boosts the performance of several baseline models, achieving cutting-edge results in many cases. Theoretical and empirical analyses confirm its generalizability, architectural efficiency, and utility in self supervised anomaly detection.

Abdelfatah et al.[3] led the core technical innovation of the study by architecting and implementing the contour driven pipeline that underpins both segmentation and classification. He first formulated the contour map extraction module leveraging structured tensors and multi-directional gradient smoothing to convert raw X-ray images into high fidelity edge representations that highlighted object boundaries. Building on this, Ahmed integrated these contour maps into an instance segmentation network, designing custom layers to detect and reduce individual threat items even under heavy occlusion. Recognizing the severe imbalance between harmless and dangerous samples, he co-formulated the Balanced Affinity Loss, a max margin clustering objective that enforces equi-spaced feature clusters for minority and majority classes, and fine-tuning its key parameter $\beta = 0.99$ through rigorous hyperparameter sweeps. He also established the end-to-end training choosing ADADELTA for segmentation, Adam for classification, and crafting an 80/20 train-test split with a dedicated validation subset and led the experimental evaluation across the SIXray datasets, demonstrating consistent state-of-the-art gains in IoU, DC, and mAP over existing methods.

Samet Akçay et al.[4] work on Skip-GANomaly advances a line of research that began with vanilla auto encoders for reconstruction-based anomaly detection and was transformed by the introduction of adversarial learning in AnoGAN and EGBAD. Recognizing the instability and high inference cost of those GAN-based methods. Akçay participated in developing GANomaly, which unified encoder–decoder reconstruction and adversarial objectives into a single forward pass but still suffered from sparse bottleneck representations. Drawing inspiration from U-Net architectures in medical imaging. Akçay then led the integration of skip-connections into the GANomaly framework, enabling richer multi-scale feature reuse and significantly sharper reconstructions of normal samples. He further refined the training process with a multi-loss objective combining adversarial, contextual, and latent losses to stabilize GAN training and enhance separation between normal and anomalous distributions in both image and latent spaces.

III. Proposed Methodology

This section contains the proposed framework which leverages a Siamese Neural Network (SNN) architecture to learn a discriminative similarity function capable of distinguishing between anomalous (positive) and normal (negative) X-ray cargo images. The design of the network, its training region, and the optimization strategy are collectively tailored to address the challenges of class imbalance and limited positive samples, which are common in real-world security screening scenarios.

A. Input Preprocessing and Embedding Extraction

Each input image is initially resized to a uniform resolution of 224×224 pixels to conform to the input dimensionality requirements of convolutional neural networks pre-trained on ImageNet. The images, originally grayscale X-ray scans, are transformed into three-channel RGB images by replicating the single channel across three channels to match the expected

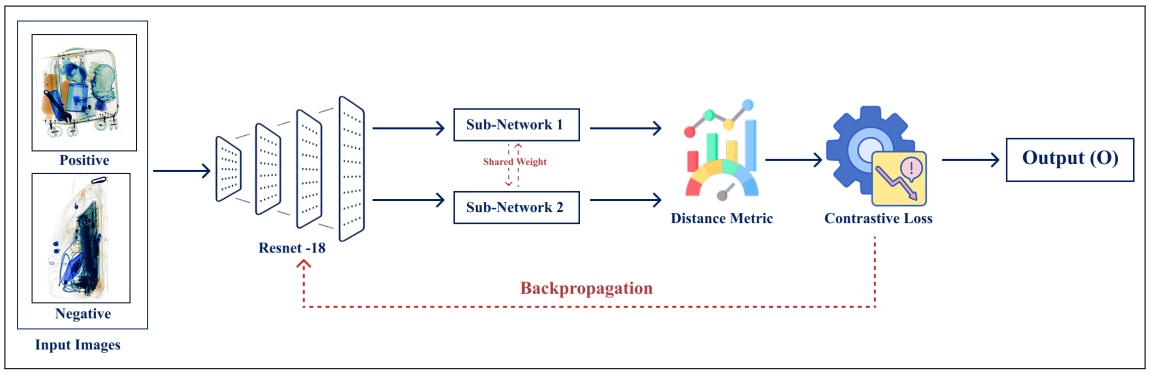

Fig. 1: High Level Architecture Diagram with workflow

input of ResNet architectures. These inputs are then normalized using the mean and standard deviation of the ImageNet dataset: $\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$.

Let x_1 and x_2 denote two input images. Both images are independently passed through a shared-weight convolutional backbone, specifically a ResNet-18 model truncated at its final classification layer. This results in two 512-dimensional embeddings, denoted as $f(x_1), f(x_2) \in \mathbb{R}^{512}$. The embedding function $f()$ maps raw image inputs to a latent space where Euclidean distances reflect semantic similarity:

$$D(x_1, x_2) = \|f(x_1) - f(x_2)\|_2 = \sqrt{\sum_{i=1}^{512} (f_i(x_1) - f_i(x_2))^2}. \quad (1)$$

B. Contrastive Loss for Metric Learning

To guide the network toward learning an effective similarity metric, we employ the contrastive loss function, which penalizes the network based on the similarity or dissimilarity of image pairs. Given a binary label $y \in \{0, 1\}$ where $y = 0$ indicates dissimilar and $y = 1$ indicates similar the contrastive loss L is defined as:

$$L(y, D) = (1 - y)\frac{1}{2}D^2 + y\frac{1}{2}\left[\max(0, m - D)\right]^2,$$

where m is a pre-defined margin that enforces a minimum distance between dissimilar pairs. In our implementation, we set $m = 0.5$. For similar pairs, the loss encourages embeddings to move closer in the latent space, while for dissimilar pairs, it pushes them apart, provided their distance is less than the margin.

C. Optimization and Training Procedure

The network is trained using the Adam optimizer with a learning rate of 1×10^{-4}. We perform mini-batch training over 20 epochs with a batch size of 32. In each iteration, the model receives a mixture of 50% similar pairs and 50% dissimilar pairs. Gradients of the loss function with respect to network parameters are computed via backpropagation and used to update the weights of the embedding extractor.

A pair is classified as similar if $D(x_1, x_2) \leq T$ and dissimilar otherwise, where T is a fixed distance threshold set to 0.5. This threshold was chosen tentatively based on the distribution of distances observed during training.

D. Evaluation Metrics

To evaluate the effectiveness of the model in distinguishing between similar and dissimilar image pairs, we employ accuracy as the primary evaluation metric. This metric quantifies the proportion of correctly classified image pairs relative to the total number of predictions. It is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. In addition, we compute the F_1-score, which offer insight into the model's performance on imbalanced data. These metrics are derived from the confusion matrix built during the final epoch of training, where the model achieves convergence.

The convergence of the contrastive loss, the increasing classification accuracy over epochs, and the final confusion matrix collectively demonstrate that the network successfully learns to encode the visual semantics of cargo images, distinguishing between anomalous and normal content with high fidelity. This metric-based

framework provides a scalable and generalizable approach to anomaly detection in security imaging environments.

IV. Dataset Description

The evaluation uses the publicly available CargoX dataset [5], created for identifying unusual items in cargo X-ray images. This dataset is useful for security tasks, such as spotting prohibited or hidden items in cargo containers, which is important for customs and border protection. The dataset includes grayscale X-ray images divided into two main groups: positive samples, which show cargo with prohibited or hidden items, and negative samples, which show normal cargo without such items. The images reflect real-world situations, capturing different types of cargo, object positions, and imaging conditions.

To prepare the images for the ResNet-18 model, a common image processing tool pretrained on a large image collection, each X-ray image is adjusted to fit the model's needs. Specifically, images are resized to 224x224 pixels, a standard size for this model, using a method that keeps the image clear. Since the X-ray images are grayscale, they are converted to a three-color format by copying the grayscale values three times to match the model's requirements. The images are also adjusted using standard values (mean: [0.485, 0.456, 0.406] and standard deviation: [0.229, 0.224, 0.225]) to make them similar to the images the model was trained on. Optional image tweaks, like flipping images horizontally or slightly rotating them (within ±10 degrees), are applied during training to help the model handle different cargo arrangements.

For training the Siamese network, a method that compares pair of images, pairs are created during each training round. A total of 1000 balanced pairs are made per round, with 500 similar pairs and 500 different pairs. Similar pairs include either two positive samples or two negative samples, helping the model recognize when images are alike. Different pairs include one positive and one negative sample, teaching the model to tell them apart. These pairs are randomly chosen each round to keep the training varied and prevent the model from learning the same examples repeatedly.

The CargoX dataset is challenging because the images within each group vary a lot and often show objects overlapping or hiding each other, as seen in real cargo scans. Positive samples might include different types of prohibited items, varying in size, shape, or material, while negative samples might show anything from tightly packed boxes to loosely arranged goods. These overlapping patterns and differences make it hard to spot unusual items, making the dataset a great choice for testing models that detect unusual cargo. To manage the images efficiently, custom tools built with PyTorch, a popular programming framework, are used to load and prepare the images. A batch size of 32 images is used to keep training smooth and efficient, and multiple processes handle image loading to avoid delays.

This detailed preparation process ensures the CargoX dataset is ideal for training and testing models to detect unusual items in cargo X-ray images, offering a realistic and tough challenge for improving security-focused image analysis.

V. Results and Discussion

To assess the effectiveness of the Siamese neural network in distinguishing between similar and dissimilar image pairs, a detailed evaluation was conducted using a confusion matrix and corresponding label distribution analysis, along with a comparative study against existing anomaly detection models in the literature.

A. Confusion Matrix Analysis

The classification performance of the Siamese neural network was evaluated using a confusion matrix computed at the twentieth epoch. The matrix quantifies the network's ability to distinguish between similar and dissimilar image pairs based on the learned embedding space. The results are as follows: a total of 497 image pairs with dissimilar ground truth labels were correctly classified as dissimilar, whereas 3 dissimilar pairs were misclassified as similar. For image pairs labeled similar, 443 were correctly identified, while 57 were incorrectly predicted as dissimilar.

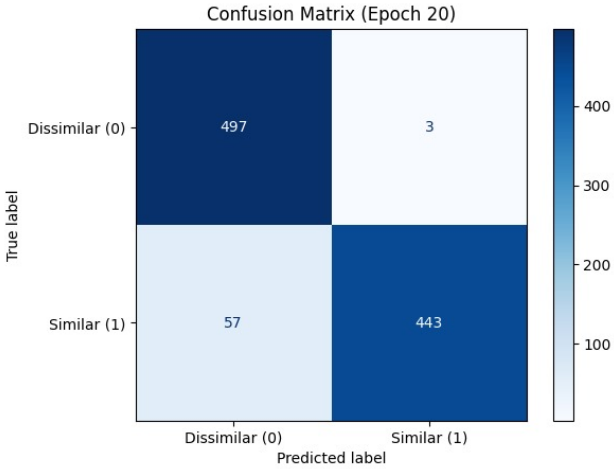

Fig. 2: Confusion Matrix showing prediction performance

These results indicate a strong discriminative capacity of the model, particularly in identifying dissimilar pairs, as evidenced by the low false positive rate for the similar class. The overall structure of the confusion matrix reveals a diagonally dominant pattern, affirming the network's effectiveness in minimizing intra-class misclassifications. The matrix is visualized using a color map ranging from white to dark blue, corresponding to frequency values from 0 to 500. The darkest cells represent the most frequent and accurately predicted classes, while the lighter regions indicate the distribution of misclassifications. This visual encoding provides an immediate understanding of model accuracy and error concentration across categories.

These values indicate that while the model performs well in identifying negative samples, it struggles with detecting positive samples, evident from the higher number of false negatives.

B. Comparative Analysis

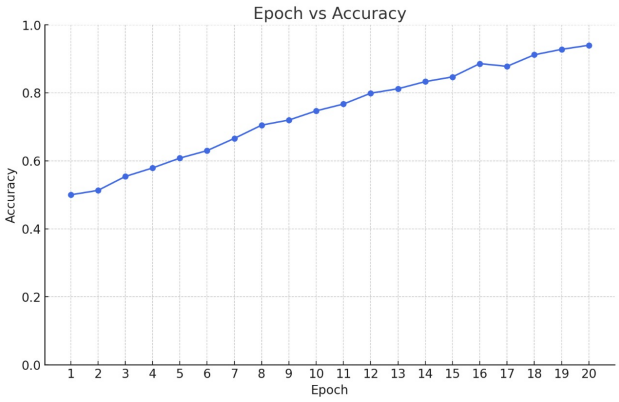

Fig. 3: Training accuracy across epochs for the proposed Siamese Neural Network.

To evaluate the learning behavior of the Siamese Neural Network, the accuracy trend over 20 training epochs was recorded. Fig.3 illustrates the model's increasing classification accuracy, starting from 50% in the initial epoch and reaching 94.00% by epoch 20. The curve reflects stable convergence and effective optimization of the contrastive loss.

To benchmark the proposed method, a comparative analysis was performed against recent models developed for anomaly detection. Table 1 presents the performance of each model in terms of classification accuracy. The proposed approach achieves competitive results while employing a pairwise learning strategy optimized through contrastive loss.

The proposed Siamese Neural Network with a ResNet-18 backbone achieved an accuracy of 94.00%, outperforming several established models in the domain of X-ray anomaly detection. Specifically, it surpasses the Encoder-Decoder-Classifier, which achieved 88.00%, and Skip-GANomaly, which recorded 86.00%. These results underscore the effectiveness of the Siamese framework in capturing subtle semantic differences between normal and anomalous image pairs.

TABLE I: Comparison with Existing Methods

Model	**Accuracy (%)**
Encoder-Decoder-Classifier [1]	88.00
Skip-GANomaly [4]	86.00
Proposed SNN (ResNet-18)	**94.00**

C. Discussion

The Siamese neural network demonstrates strong discriminative capability for dissimilar image pairs, correctly identifying 497 out of 500 such instances. However, it shows moderate difficulty in accurately recognizing similar pairs, as reflected by the 57 false negatives. This difference in performance may be attributed to complex intra-class variations and a limited diversity of positive pair examples within the training data. Despite this, the proposed model achieves 94.00% accuracy, outperforming several existing approaches such as Skip-GANomaly and encoder-decoder-based architectures, thereby confirming its effectiveness in learning pairwise representations for anomaly detection in cargo X-ray imagery.

VI. Conclusion and Future Work

The experimental evaluation validates the effectiveness of the Siamese Neural Network architecture in learning a discriminative embedding space for X-ray cargo image analysis. By utilizing a ResNet-18 encoder and contrastive loss, the network successfully distinguishes between semantically similar and dissimilar image pairs. After 20 training epochs, the model achieved an overall accuracy of 94.00%, with a final loss of 0.1103. The confusion matrix recorded 497 true negatives and 443 true positives, alongside only 3 false positives and 57 false negatives, illustrating the model's robustness in pairwise classification. When compared with existing methods in the literature, the proposed approach demonstrates competitive performance, outperforming several baseline models in terms of classification accuracy. These findings affirm the model's applicability in security-critical environments, particularly for the visual inspection of cargo, where reliable identification of concealed items is essential.

Future work may explore the integration of attention-based mechanisms to improve the model's sensitivity to

subtle intra-class variations. Additionally, strategies such as hard negative mining, advanced data augmentation, and the use of ensemble architectures could further enhance the generalization ability and robustness of the network in more diverse or real-time screening environments.

REFERENCES

[1] Gaikwad, Bipin, et al. "Self-Supervised Anomaly Detection and a New Benchmark for X-Ray Cargo Images." 2024 IEEE International Conference on Image Processing (ICIP). IEEE, 2024.

[2] Madan, Neelu, et al. "Self-supervised masked convolutional transformer block for anomaly detection." IEEE Transactions on Pattern Analysis and Machine Intelligence 46.1 (2023).

[3] Ahmed, Abdelfatah, et al. "Enhancing security in X-ray baggage scans: A contour-driven learning approach for abnormality classification and instance segmentation." Engineering Applications of Artificial Intelligence 130 (2024).

[4] Akçay, Samet, et al. "Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection." 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019.

[5] CargoX Dataset. [Online]. Available: magentahttps://github.com/Mbwslib/CargoX