Dual-Stream Feature Synergy and Prior Fusion Method for Chest X-ray Anomaly Detection

Chenxi Gao

{23009290073@stu.xidian.edu.cn }

School of Artificial Intelligence, Xidian University, Xi'an, Shaanxi, China

Abstract. The current single-stream unsupervised anomaly detection algorithms focus on two problems which are: CNNs (Convolutional Neural Networks) ignoring global context and pre-trained ViT (Vision Transformer) models ignoring layered deep features and losing fine grained pathological textures. To remedy this, the study proposes a dual-stream synergy architecture framework which integrates both 'texture' and 'context' streams. More precisely, the texture branch employs ResNet50 local features at high frequencies and the context branch integrates anatomy prior filters, background noise suppression, and multi-level aggregation from DINOv3 to extract texture at shallow and middle levels. The Kermany Chest X-ray database experiments show that this supplementary feature synergy improves detection performance. The dual-stream model (AUC=0.8944) indeed outstrips DINOv3 (AUC=0.65) and ResNet50 (AUC=0.8728) standalone models, cementing the practicality of the CNNs inductive bias and the Transformers global semantics to achieve strong zero-shot medical diagnosis.

Keywords: Anomaly detection; Zero-shot learning; Dual-stream network; DINOv3; Feature fusion.

1 Introduction

One of the deadliest infectious diseases is pneumonia which is screenable via chest x-ray [1]. While fully supervised deep learning models score high recognition accuracy, they need large amount of labeled data which includes large-scale pixel-level annotation bringing huge costs [2]. In order to address these data-intensive constraints, the study propose unsupervised anomaly detection (UAD). Building models to understand normal anatomy is the only requirement of UAD, which, on the one hand, eliminates the need for data annotation, and on the other hand reduces the cost and time to complete data annotation [3].

Nevertheless, some unsupervised techniques, when applied to medical imaging, have certain obstacles. Convolutional neural networks (CNNs) PatchCore [4] and PaDiM [5] have local inductive bias that these approaches have to deal with. Without an understanding of macroscopic pulmonary anatomy, these methods could incorrectly mark abnormalities such as vascular patterns and rib overlaps. Methods using global views from Vision Transformers (ViT) [6], specifically those using the DINO series [7] for zero-shot detection, frequently fail

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365105

to overcome the "semantic adaptation" issue [8]. DINO series models are described as operating on deep levels of object-based semantics, which are, in this context, gross/specimen-level characteristics of the objects to the neglect of important fine-level characteristics, such as the glassy texture of certain types of pneumonia (i.e., ground glass opacities). The result of this is that when models are utilized without fine-tuning, the outputs from such models are frequently unreliably deficient in terms of an adequate level of detection.

This paper proposes a dual-stream feature collaboration model to address such limitations. It posits CNNs' local bias and ViT global attention as mutually beneficial. The model utilizes ResNet [9] to extract texture streams and DINOv3 [7] to build context streams. The features extracted by CNNs address textures, while ViT's global attention maps minimize the false positives related to the anatomy. The inaccurate understanding of background and detail texture, related to ViT, is mitigated by the inclusion of the context stream and the MLFA (Multi-Level Feature Aggregation) module, which reuses shallow Transformer streams and gets guided by anatomical prior filters, which mask out background noise that is not related to the lung region.

In the exhaustive Kermany chest X-ray dataset [10] experiments, the method outperformed the others and was highly clinically relevant. Also, the dual-stream architecture used in the method increased the AUC to 0.8944, which is an increase over the baseline achieved by ResNet50 [9] and single-stream ViT. Also, the method achieves lesion localization even without additional computational costs, thus, this output stands out to fully supervised object detection models like YOLO [11]. This confirms that the method is fully autonomous and can operate in a diagnostic capacity. This can be especially useful in low-resource settings.

2 Data and Methodology

2.1 Dataset configuration and preprocessing

This research makes use of the Kermany Chest Radiograph Dataset [10], and for the One-Class Classification (OCC) protocol, the only training set scanned was 1,341 radiologist-validated normal chest X-ray images, allowing the model to learn the healthy anatomical features of the manifold. The test set was constructed by blending normal and abnormal samples. The abnormal samples are all attributed to the "pneumonia" class in the dataset, which are the pathological anomalies the model attempts to identify. The normal test samples are the remaining normal X-ray images. Because of this, no images from the training set are present in the test set, which guarantees that the evaluation of the model is based on its ability to handle new, unseen real test data.

I have a standard workflow for pre-processing, which starts by ensuring all the images are in RGB format and reducing the image size to 224×224 using bilinear interpolation and normalizing the images using the ImageNet standards [12].

2.2 Adaptability of ViT deep features

Before developing the dual-stream framework, this study explored the capabilities of the traditional Vision Transformer. An image-level AUC of 0.6552 was obtained using a KNN

model based on features from the 24th layer of a pre-trained DINOv3 (ViT-Large) [7]. While this is a case better than random predictions, it is still below the ResNet standard [9]. Also, from the feature analysis, one explanation may be invoked from the semantic blurring of the deep architectures. The deep layers of the ViT model focus on a 'macroseismic' level summary, capturing entities of an organ, and thus suffer from a weak mapping of deep-level semantics to high-frequency entities, such as ground-glass opacities and other pathological textures. Consequently, this study claims that, for medical detection tasks performed without training (zero-shot), features abstracted solely at the deep level, or the so-called deep level features, on their own are not sufficient. Added to this, shallow and intermediate levels of textures need to be considered, which is the guiding principle of the proposed MLFA (Multi-Level Feature Aggregation) approach.

2.3 Proposed method

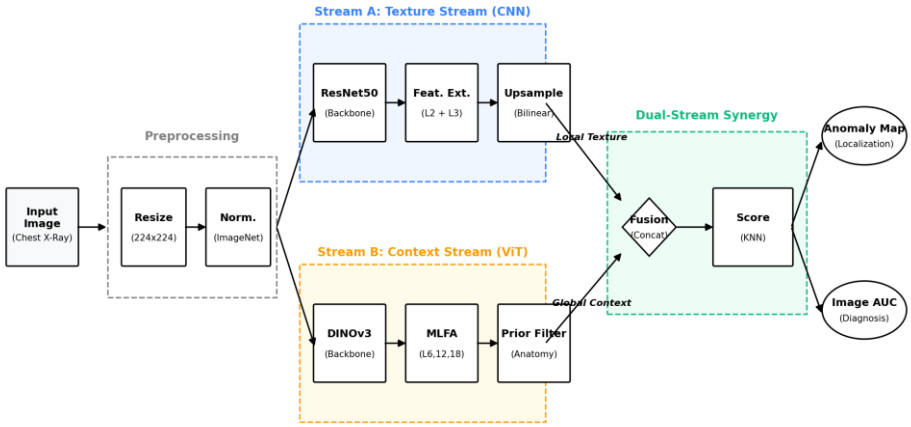

Fig. 1. Overview of Dual-Stream Feature Synergy with Anatomical Prior Network (Picture credit: Original)

Dual-Stream parallel processing architecture. The proposed system (Figure 1) has a dual-stream system architecture: one stream for capturing texture and the other for capturing contextual features. The Texture Stream (top stream, blue) uses a ResNet50 architecture to capture local high-frequency texture features. In contrast, the Context Stream (bottom stream, orange) uses DINOv3 with MLFA and the Anatomical Prior Filter to capture global contextual features and reduce distracting background noise. The Dual-Stream Synergy (DS-S) module (green) combines (or fuses) features from both streams to generate localization anomaly maps and image-level diagnostic scores.

Texture stream based on convolutional inductive bias. The overall design structure (Figure 1) includes the CNN texture stream branch, specifically created to detect subtle pathological textures present in the CXR images. Given the prominent local feature extraction abilities of the ResNet50 architecture [9] in models such as PatchCore [4], this architecture was used as the foundational structure.

In order to maintain the required global contextual information that is omnipresent in the global average pooling and lesion localization the feature levels, various layers were used from varied feature levels. Basic information such as edges and color is found in the lower-level features (Layer 1) and high-level features (Layer 4) suffer from loss of resolution. Consequently, feature extraction was restricted to the middle layers, Layer 2 and Layer 3. This strategy is meant to achieve equilibrium between "semantic depth" and "spatial resolution." Thus, it was possible to capture texture patterns while spatial information was preserved at 1/8 and 1/16 dimensions relative to the input image.

The feature map $\phi_3(x)$ from Layer 3 is bilinearly interpolated and upscaled to match the spatial dimensions of Layer 2's $\phi_2(x)$, and then the two are concatenated along the channel dimension. For the input image $x \in \mathbb{R}^{H\times W\times C}$, the local feature representation$F_{tex}(x)$ extracted by the texture stream, can be defined as follows:

$$F_{tex}(x) = Concat\left(\phi_2(x), Upsample\left(\phi_3(x)\right)\right) \quad (1)$$

In(1), where$\phi_j(x)$ represents the feature map output from the j-th residual block of the ResNet [9]. This strategy of multi-scale fusion allows the model to capture finer local details as well as larger lesion areas.

Panoramic contextual flow structure based on DINOv3. Alongside texture flow, the Context Stream (bottom, orange in Figure 1) utilizes DINOv3 (ViT-Large) [7] for the enhancement of the global semantic understanding. As for the texture deficiency in the top layers of the ViT features highlighted in Section 2.2, a Multi-Level Feature Aggregation (MLFA) technique is put forward. Utilizing the hierarchical nature of ViT, this framework fuses components from the third level: Layer 6 captures shallow texture and edge details, Layer 12 seizes features of medium-scale anatomical sub-parts, while Layer 18 extracts global contextual semantics. This method captures macro-level semantics, while sidestepping the "smoothing" operation at Layer 24.

For an input image x , the output of layer l is a sequence of Patch Tokens $\psi_l(x) \in \mathbb{R}^{N\times D}$, where $N = 196$and $D = 1024$. The MLFA module concatenates features from the selected layer set $\mathcal{S} = \{6,12,18\}$ along the channel dimension:

$$F_{ctx}(x) = \text{Concat}_{l\in\mathcal{S}}\psi_l(x) = [\psi_6(x), \psi_{12}(x), \psi_{18}(x)] \quad (2)$$

The resulting contextual features $F_{ctx}(x) \in \mathbb{R}^{N\times 3D}$ form a high-dimensional dense descriptor.

Anatomy-Guided background suppression. Given that universal visual models apply the same treatment to all pixels, which, in the case of medical imaging, is not ideal since background regions have sensor noise, the study propose an "anatomical optical density priors" based filtering method. According to the physics of X-ray, pixels corresponding to air outside the body have lower radiation attenuation and thus exhibit lower intensity. This study an

optical density threshold τ to design a background suppression technique that removes the non-anatomical noise, thus centering the model on the biological tissue.

More precisely, the author partitions the image into patches P_i and calculates the mean optical intensity $\mu(P_i)$. A binary mask$M \epsilon \{0,1\}^N$ is constructed to demonstrate if a Patch is inside a Region of Interest (ROI):

$$M_i = \begin{cases} 1, & if \mu(P_i) > \tau \\ 0, & otherwise \end{cases} \tag{3}$$

The empirical threshold τ is defined as 30. Features falling below this threshold are eliminated, purging the feature space and securing that anomaly scores are purely related to lung tissue pathology, thus improving the robustness.

Anatomical prior filtering feature space is "cleaned" by "ignoring" the background, and this also enhances the robustness of the final anomaly scores by ensuring that they are solely based on pathological changes within the lung tissue itself.

Adaptive feature fusion and unified inference. Integrating features for final diagnosis is done through the Dual-Stream Synergy module (green) in Figure 1. Simply overlaying the unprocessed score values from the two streams is problematic because of \textit{"scale inconsistency"} attributed to the differing feature topologies. To prevent the higher valued flows from overshadowing other values, a recalibration of the distribution is necessary.

Min-Max normalization Strategy. To avoid the issues of dimensionality and bring alignment to all outputs of the streams, Min-Max Normalization is applied. Anomaly scores from each stream, $S_{tex}(ResNet)$ and $S_{ctx}(DINO)$, are calculated for each image of the test set. Using the mean and variance of the test batch, the scores are adjusted to the range $[0,1]$:

$$\hat{S}_k = \frac{S_k - \min(S_k)}{\max(S_k) - \min(S_k)}, k \in \{tex, ctx\} \tag{4}$$

Texture and context scores are represented by $\hat{S}_{tex}$ and $\hat{S}_{ctx}$, respectively. This guarantees uniform numerical scales for fusion and maintains the order of samples in each of their streams.

Weighted Linear Fusion. After normalization, a weighted linear combination is used to obtain the final image-level anomaly score S_{final} :

$$S_{final} = \lambda \cdot \hat{S}_{tex} + (1 - \lambda) \cdot \hat{S}_{ctx} \tag{5}$$

Here, $\lambda \epsilon [0, 1]$ is a balancing parameter. The study focused on the value obtained from the validation set, setting λ to 0.5. Thus, the local texture anomalies (identified by ResNet) and the global structural anomalies (identified by DINO) have the same impact on the pneumonia

detection level. This shows that the fusion mechanism is working as a "dual validation", in the sense that when both streams produce high scores, fusion improves the confidence. On the other hand, for the areas that ResNet flagged as pathological but that DINO considers normal including the background, fuse scores are low, thus alleviating the problem of false positives.

3 Experiments

3.1 Experimental setup

Implementation details. The experiments were run using the PyTorch 2.0 framework on an NVIDIA RTX 4090 (24GB) machine. For image preprocessing, images were resized to 224× 224. For model architecture, frozen pre-trained weights were loaded for both streams: ResNet50 (ImageNet) [9], [12] and DINOv3 (ViT-Large) [7] and for semi-supervised anomaly detection, Scikit-learn's KNN (K=50, Euclidean, fusion weight = 0.5) was used. In the course of the experiments, to maximize efficiency, only the forward inference was run, without any gradient calculations or updates, and without fine-tuning.

Evaluation metrics. This edited text adds more detail on the methodology used to test the performance of the proposed strategy in unsupervised anomaly detection. The method includes the Step-wise Evaluation Strategy that integrates local and global assessment in terms of discrimination and localisation. For global discrimination, besides AUC, other metrics are used which are derived from the confusion matrix. These metrics include Accuracy, Positive Predictive Value (PPV), and the clinically valuable Negativie Predictive Value (NPV), which are complemented by the F1 measure to assess the Discriminate Power (DP).

3.2 Main results

The outcomes related to the Kermany Chest X-ray dataset [10] along with the comparison to other contemporary unsupervised methods can be found in Table 1. The Dual-Stream framework exhibits notable advancements when compared to both PatchCore [4] and Vanilla ViT (SVT) [8] baselines across all key metrics. The framework reaches the highest Image-level AUC at 0.8944 which illustrates its ability to integrate both texture and contextual data for zero-shot screening.

Table 1. Main Results Comparison

Method	Backbone	Pre-training	AUC	ACC	Precision	Recall	F1-score
Baselines							
PatchCore (Re-impl) [4]	ResNet50	ImageNet	0.8719	0.8937	0.9041	0.9800	0.9405
Vanilla ViT [8]	DINOv3-L	LVD-142M	0.6552	0.8428	0.8521	1.0000	0.9259
Ablations (Ours)							
Context Stream	DINOv3-L (with MLFA)	LVD-142M	0.8016	0.8611	0.8606	1.0000	0.9251

Method	Backbone	Pre-training	AUC	ACC	Precision	Recall	F1-score
Only Texture Stream	ResNet50 (Layers 2+3)	ImageNet	0.8728	0.8954	0.9042	0.9820	0.9415
Only Dual-Stream	ResNet50+DINOv3	Hybrid	0.8944	0.8851	0.9108	0.9600	0.9348

The performance ranking of the evaluated components is outlined in Table 1. The performance of Vanilla ViT (0.6552) supports the validation of the "semantic gap" hypothesis; conversely, the CNN-based Texture Stream (PatchCore) sets a robust baseline (0.8728).The suggested Dual-Stream technique combines MLFA with anatomical priors for the first time, merging the global context of Vision Transformers with the local textural features of Convolutional Neural Networks. Consequently, it outperforms every single-stream model with an AUC of 0.8944.

3.3 Ablation study

Table 2: Examination of the impact of different components of the network. A gradual ablation study shows that the addition of each extra component (i.e., Texture Stream, Context Stream, MLFA, and Anatomical Prior Filter) to the complete dual-stream framework (Experiment 5) results in successive enhancements in performance. The final dual-stream configuration (Exp. 5) results in the optimal performance.

Table 2. Ablation Study

Exp.ID	Stream A (CNN)	Stream B (ViT)	MLFA (Layers 6, 12, 18)	Anatomy Filter	AUC	Improvement
1	×	√	×(Layers24 only)	×	0.6552	Baseline
2	×	√	×	×	0.7630	+10.78%
3	×	√	√	√	0.8016	+3.86%
4	√	-	-	-	0.8728	+7.12%
5(Ours)	√	√	√	√	0.8944	+2.16%

This section uses a stepwise ablation study (Table 2) to assess the impact of each individual component. The AUC with MLFA is around 10% higher which shows the MLFA helps the model capture more shallow features and texture more effectively. Likewise, Experiments 2 and 3 show that with the Anatomical Prior Filter, there is an additional 4% improvement which suggests that this method is effective for feature space noise reduction. Experiment 4 shows us that the single-stream CNN has a solid level of robustness and sets a very strong baseline. Notably, the comparison of Experiments 4 and 5 shows that the dual-stream fusion architecture gives an additional 2.16% improvement which demonstrates that while the CNN is good at local texture feature extraction, the ViT is good at global context feature extraction and the two together are more effective.

3.4 Qualitative analysis of localization performance in clinical screening scenarios

Even though the proposed model can barely do the object detection tasks, some "soft localization" hints can be generated, which are useful for the clinicians. Though the model proposed in this paper cannot generate bounding boxes as some fully-supervised object detection models do (e.g., YOLO series [11]), the anomaly detection heatmaps do pinpoint the centres of most core im regions for pulmonary consolidation and ground-glass opacities. P. Just as can be seen in figure2, while the proposed dual-stream model is not able to generate precise bounding boxes, the generated anomaly heatmaps do. This "soft localization" can be useful for screening, balancing the review by the physicians to augment the model's screening efficiency, and interpreting the diagnosis, all at no annotation costs.

3.5 Visual analysis

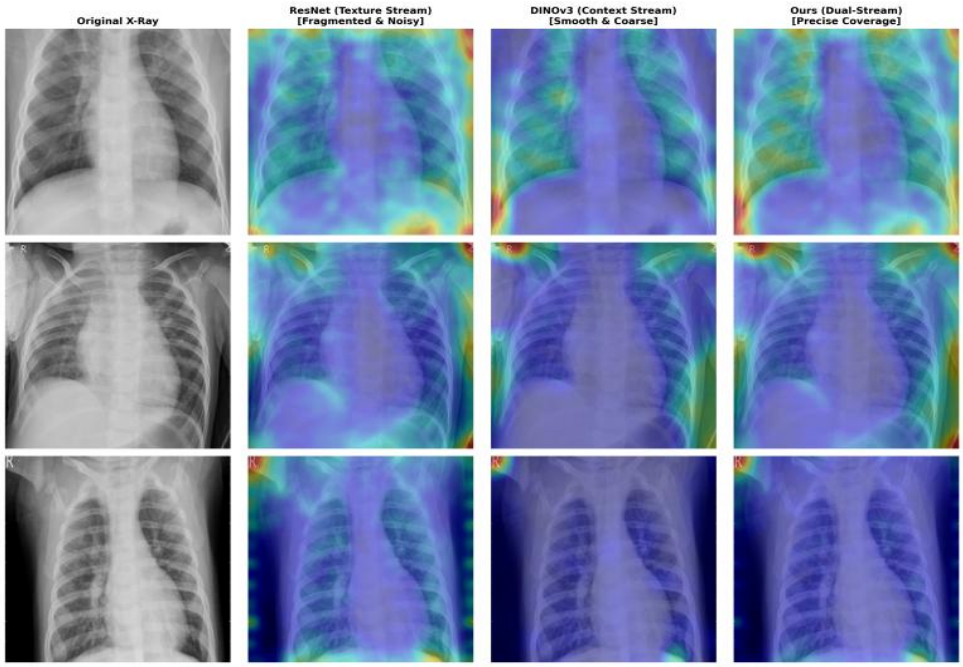

Fig. 2. Visualization of Heatmaps. (Picture credit: Original)

Qualitative comparison of the localization of anomalies in pneumonia cases. Columns from left to right show: (a) Original Chest X-Ray image; (b) texture stream (ResNet) maps with fragmented high-frequency noise; (c) context stream (DINOv3) maps with an over-smooth and coarse localization; (d) dual stream proposed maps with better background suppression and accurate lesion mapping. (Picture credit: Original)

The original X-ray images are in the first column; the second column has the visualizations of the Texture Stream (ResNet); the third column has the visualizations of the Context Stream (DINOv3); and the fourth column has the visualizations of the Dual-Stream Synergy model.

The coverage of lesions in the Dual Stream model is the best of all the models. (Picture credit: Original)

4 Conclusion

In the context of sparse data and expensive annotation in medical image analysis, the study proposes a framework for unsupervised anomaly detection using Dual-Stream Feature Synergy, which combines the local texture bias of CNNs and the global semantic understanding of Vision Transformers. Utilizing the MLFA strategy and combining it with an Anatomical Prior Filter, it overcomes the restrictions of single-model techniques due to texture or noise. On the Kermany chest X-ray dataset, experiments showed an image-level AUC of 0.8944, exceeding single-stream benchmarks and confirming the " texture-context " complementary mechanism.

While the absence of bounding box supervision results in a modest decrease in localization accuracy compared to fully supervised models and dual-stream computation exceeds single-stream benchmarks, the proposed zero-annotation dual-stream paradigm presents a cost-effective solution to assist diagnosis in resource-limited areas. This model creates useful lesion heatmaps with complete coverage of the ground glass opacities and pulmonary consolidation. Utilizing benefits like no need for annotation, no training parameters, and minimal computational training requirements, future studies will aim to lightweight structures (e.g., MobileViT) combined with knowledge distillation. Such approaches will be focused on optimizing for speed (latency) and will performance to meet the thresholds for real-time screening on mobile medical devices.

References

[1] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D.: Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine. pp. 195 (2019)

[2] Litjens, G., et al.: A survey on deep learning in medical image analysis. Medical Image Analysis. pp. 60–88 (2017)

[3] Ruff, L., et al.: A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE. pp. 756–795 (2021)

[4] Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14318–14328 (2022)

[5] Defard, T., Setkov, A., Loesch, A., Audigier, R.: PaDiM: A patch distribution modeling framework for anomaly detection and localization. International Conference on Pattern Recognition. pp. 475–489 (2021)

[6] Dosovitskiy, A., et al.: An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations. (2021)

[7] Siméoni, O., et al.: DINOv3. arXiv preprint. (2025)

[8] Matsoukas, C., Haslum, J. F., Sorkhei, M., Söderberg, M., Smith, K.: Is it time to replace CNNs with transformers for medical images? Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 2265–2276 (2021)

[9] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
[10] Kermany, D. S., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. pp. 1122–1131 (2018)
[11] Jocher, G., et al.: Ultralytics YOLOv5. GitHub. (2022). Available: https://github.com/ultralytics/yolov5.
[12] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)