FadeFormer: Content-Adaptive Graph Diffusion for Medical Image Classification

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Vision Transformers (ViTs) have emerged as powerful backbones for medical imaging due to their ability to capture long-range semantic dependencies across anatomical structures. However, standard self-attention mixes information uniformly across all tokens, without explicitly modeling how features propagate across neighboring regions. To address this, we propose **FadeFormer**, a transformer architecture that augments self-attention with content-adaptive graph diffusion. We construct a learned similarity graph over patch tokens and perform spatially varying diffusion using a normalized graph Laplacian, enabling structured feature propagation conditioned on token similarity. Unlike attention-based aggregation or graph attention mechanisms, our formulation explicitly governs feature evolution through a diffusion process, encouraging information to spread across semantically related tokens rather than isolated discriminative patches. We evaluate FadeFormer on NIH ChestX-ray14, ISIC-2019, and MedMNIST v2, where it achieves improved or competitive classification performance and produces qualitatively broader activation coverage compared to standard ViT baselines, consistent with diffusion-driven feature propagation across anatomically related regions.

Keywords: Vision Transformer · Graph Diffusion · Medical Image Classification

1 Introduction

The introduction of transformers [1] marked a significant paradigm shift in deep learning, enabling models built around self-attention [2] that capture long-range dependencies beyond the capabilities of traditional CNNs. In medical imaging, such global context is essential, as pathological patterns often span multiple anatomical regions. However, transformers lack an explicit local inductive bias, limiting their ability to preserve fine-grained textures and spatial coherence. This often results in fragmented or noisy representations in tasks requiring precise localization. Prior work has attempted to address this limitation by introducing locality into transformer architectures. Hierarchical designs incorporate windowed attention [3], while hybrid CNN-Transformer models inject convolutional inductive biases [5, 6]. More recent methods explore diffusion-inspired and spectral

formulations [7, 11, 15, 12], but these rely on fixed or globally defined operators and do not explicitly model how neighboring image regions evolve together based on their content.

To address this, we propose **FadeFormer**, a transformer architecture that models token interactions as a content-adaptive graph diffusion operator. We construct a graph over tokens using learned pairwise feature similarity and define a diffusion process over it. Unlike attention-based methods, which aggregate features at each layer, our formulation governs feature evolution through this operator. The diffusion is spatially varying, with propagation strength conditioned on feature similarity, allowing information to diffuse within homogeneous regions while limiting propagation across structural discontinuities.

We evaluate FadeFormer on MedMNIST v2 [19, 18], ISIC-2019 [20, 21], and NIH ChestX-ray14 [22]. Our method achieves competitive classification performance while producing broader, more globally distributed activation patterns compared to ViT baselines, as evidenced by qualitative GradCAM [23] visualizations. The contributions of this work are threefold:

- We propose **FadeFormer**, a transformer architecture that models token interactions as a content-adaptive graph diffusion process, enabling structured and spatially continuous feature propagation.
- We introduce a spatially varying diffusion formulation based on a learned normalized graph Laplacian, where propagation is conditioned on feature similarity, allowing boundary-aware information flow across semantically related tokens.
- We evaluate FadeFormer across three medical imaging benchmarks, demonstrating competitive classification performance and qualitatively broader activation coverage compared to standard transformer baselines, consistent with the global propagation induced by graph diffusion.

2 Related Work

2.1 Vision Transformers and Hybrid Architectures

ViTs have become a prominent backbone for medical imaging [9] due to their ability to capture long-range dependencies through global self-attention [2], which is particularly relevant for modeling pathological patterns that occur across multiple regions. However, their lack of explicit local inductive bias remains a known limitation, especially in tasks requiring accurate spatial localisation. To address this limitation, several approaches introduce locality into transformer architectures, including hierarchical attention [3], multi-scale feature aggregation [10], domain-specific pretraining [8, 6], and hybrid CNN-Transformer architectures [4–6]. Hybrid models combine convolutional feature extraction with global attention to improve texture modeling and reduce data requirements [17, 16]. However, these approaches rely on fixed attention windows or convolutional kernels, limiting how features propagate across regions. As a result, the trade-off between global context and local precision remains unresolved [9].

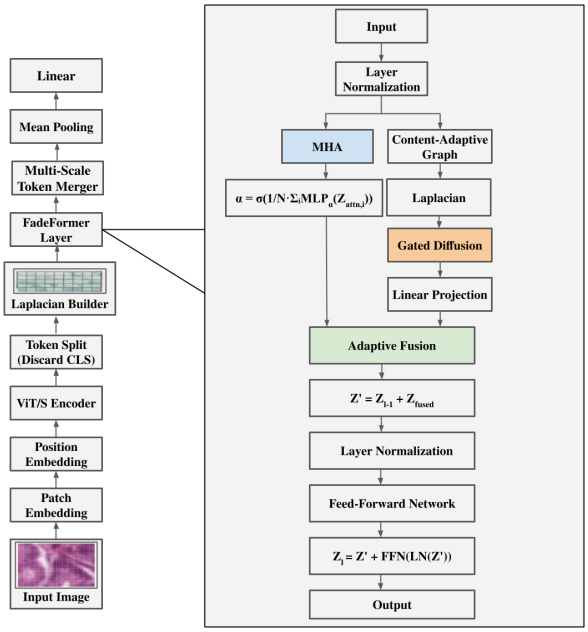

Fig. 1. Overview of the proposed FadeFormer architecture. Left: An input image is processed by a ViT-S backbone to obtain patch token embeddings. The patch tokens are passed through the proposed FadeFormer layer, fused with multi-scale token representations, and used for classification. **Right:** A FadeFormer layer combines global self-attention with content-adaptive graph diffusion. A normalized graph Laplacian governs feature diffusion across semantically related tokens, and the resulting attention and diffusion features are fused through a learned adaptive gate before the feed-forward network.

2.2 Diffusion and Graph-Based Methods

Recent work has explored diffusion-inspired mechanisms to enhance transformer representations [7]. Some approaches reinterpret attention as an unrolled diffusion process, while others perform iterative refinement in attention or latent feature space [7]. More recently, Laplacian-based transformers have incorporated spectral operators to enhance boundary awareness and texture preservation [13, 11, 14], while graph-based approaches model relationships through structured message passing. However, these methods typically operate in feature or attention space, rely on fixed or globally defined Laplacian operators, or perform discrete message passing using learned attention weights, without explicitly modeling the continuous evolution of features across local neighborhoods. In contrast, FadeFormer constructs a content-adaptive graph from token features and models feature propagation as a spatially varying diffusion process governed by a learned graph Laplacian, allowing information to diffuse within homogeneous regions while preserving structural boundaries.

3 Methodology

3.1 Overview

FadeFormer augments standard ViT backbones [2] with a content-adaptive graph diffusion module that operates directly on patch token embeddings. As illustrated in Figure 1, patch tokens extracted from a frozen, pretrained ViT-S encoder are processed by diffusion-augmented transformer layers. Multi-head attention produces $\mathbf{Z}_{\text{attn}}$, from which a content-adaptive graph is constructed. A normalized graph Laplacian governs feature diffusion, and the resulting attention and diffusion representations are fused through a learned gating mechanism before the feed-forward network, enabling FadeFormer to model feature evolution while producing broader, more globally distributed activation patterns across anatomically related regions.

Given an input image $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$, we extract a sequence of patch tokens using a ViT-S backbone. The image is partitioned into N non-overlapping patches, each projected into a d-dimensional embedding space. We discard the `[CLS]` token and retain only patch tokens:

$$\mathbf{Z}^{(0)} \in \mathbb{R}^{B \times N \times d}. \tag{1}$$

This ensures that the graph is defined only over patch-level token representations.

3.2 FadeFormer Layer

We employ the FadeFormer layer that integrates global self-attention with local graph diffusion.

Global Self-Attention Given input tokens $\mathbf{Z}^{(\ell-1)}$, we apply multi-head self-attention:

$$\mathbf{Z}_{\text{attn}} = \text{MHA}(\text{LN}(\mathbf{Z}^{(\ell-1)})). \tag{2}$$

Content-Adaptive Graph Construction We construct a graph over tokens based on feature similarity. Specifically, we compute query and key projections:

$$\mathbf{Q} = \mathbf{Z}_{\text{attn}}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{Z}_{\text{attn}}\mathbf{W}_K, \tag{3}$$

and define a similarity matrix:

$$\mathbf{S} = \frac{\mathbf{Q}\mathbf{K}^\top}{d^{1/2}}. \tag{4}$$

To enforce locality, we retain the top-k neighbors (k=10) (Table 4) for each token, yielding a sparse adjacency matrix $\mathbf{A}$. The matrix is symmetrized:

$$\mathbf{A} = \tfrac{1}{2}(\hat{\mathbf{A}} + \hat{\mathbf{A}}^\top). \tag{5}$$

The normalized graph Laplacian is then computed as:

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}. \tag{6}$$

The graph is content-adaptive and sparse. Computing the similarity matrix has the same $\mathcal{O}(N^2 d)$ asymptotic complexity as self-attention, while diffusion scales as $\mathcal{O}(kNd)$. Top-k selection and gating introduce modest additional overhead relative to this asymptotic cost.

Gated Graph Diffusion We model feature propagation as a single-step graph diffusion process:

$$\mathbf{Z}_{\text{diff}} = \mathbf{Z}_{\text{attn}} - \boldsymbol{\gamma} \odot (\mathbf{L}\mathbf{Z}_{\text{attn}}), \tag{7}$$

where

$$\boldsymbol{\gamma} = \sigma(\text{MLP}_{\gamma}(\mathbf{Z}_{\text{attn}})). \tag{8}$$

Using $\mathbf{L} = \mathbf{I} - \mathbf{A}_{\text{norm}}$, this becomes

$$\mathbf{Z}_{\text{diff}} = (1 - \boldsymbol{\gamma}) \odot \mathbf{Z}_{\text{attn}} + \boldsymbol{\gamma} \odot \mathbf{A}_{\text{norm}}\mathbf{Z}_{\text{attn}}. \tag{9}$$

Each token evolves by blending its current representation with information propagated from its symmetrically normalized neighborhood, where the learned gate $\boldsymbol{\gamma}$ controls the degree of diffusion at each spatial location. Over-smoothing is mitigated through single-step diffusion, learnable gating, and adaptive fusion.

Adaptive Fusion To balance global and local information, we compute

$$\boldsymbol{\alpha} = \sigma\left(\mathbf{W}_{\alpha} \cdot \frac{1}{N}\sum_{i=1}^{N} \mathbf{Z}_{\text{attn},i}\right), \tag{10}$$

and fuse the attention and diffusion features:

$$\mathbf{Z}_{\text{fused}} = \boldsymbol{\alpha} \odot \mathbf{Z}_{\text{attn}} + (1 - \boldsymbol{\alpha}) \odot \mathbf{Z}_{\text{diff}}. \tag{11}$$

The output is computed using residual connections:

$$\mathbf{Z}' = \mathbf{Z}^{(\ell-1)} + \mathbf{Z}_{\text{fused}}, \tag{12}$$

$$\mathbf{Z}^{(\ell)} = \mathbf{Z}' + \text{FFN}(\text{LN}(\mathbf{Z}')). \tag{13}$$

3.3 Multi-Scale Token Merging and Classification

To incorporate both fine-grained and coarse contextual information, the patch tokens are reshaped into a grid and downsampled by a factor of 2 using average pooling to produce coarse tokens at half the spatial resolution. The coarse tokens are linearly projected and concatenated with the original tokens:

$$\mathbf{Z}_{\text{merged}} = [\mathbf{Z}_{\text{fine}} \,||\, \mathbf{Z}_{\text{coarse}}]. \tag{14}$$

The merged token sequence is globally average-pooled and passed through a linear classifier:

$$\hat{\mathbf{y}} = \text{Softmax}\left(\mathbf{W}_c \cdot \text{Pool}(\mathbf{Z}_{\text{merged}})\right). \tag{15}$$

Model	Atel	Card	Effu	Infi	Mass	Nodu	Pneu	Pntx	Cons	Edem	Emph	Fibr	P_T	Hern	Ave
U-DCNN [22]	0.700	0.810	0.759	0.661	0.693	0.669	0.658	0.799	0.703	0.805	0.833	0.786	0.684	0.872	0.745
LSTM [28]	0.733	0.856	0.806	0.673	0.718	0.777	0.684	0.805	0.711	0.806	0.842	0.743	0.724	0.775	0.761
AGCL [29]	0.756	0.887	0.819	0.689	0.814	0.755	0.729	0.850	0.728	0.848	0.906	0.818	0.765	0.875	0.803
CheXNet [30]	0.769	0.885	0.825	0.694	0.824	0.759	0.715	0.852	0.745	0.842	0.906	0.821	0.766	0.901	0.807
DNet [31]	0.767	0.883	0.828	0.709	0.821	0.758	0.731	0.846	0.745	0.835	0.895	0.818	0.761	0.896	0.807
CRAL [32]	0.781	0.880	0.829	0.702	0.834	0.773	0.729	0.857	0.754	0.850	0.908	0.830	0.778	0.917	0.816
CAN [33]	0.777	0.894	0.829	0.696	0.838	0.771	0.722	0.862	0.750	0.846	0.908	0.827	0.779	**0.934**	0.817
DualCXN [34]	0.784	0.888	0.831	0.705	0.838	0.796	0.727	0.876	0.746	0.852	0.942	**0.837**	0.796	0.912	0.823
LLAGnet [35]	0.783	0.885	0.834	0.703	0.841	0.790	0.729	0.877	0.754	0.851	0.939	0.832	0.798	0.916	0.824
CheXGCN [36]	0.786	0.893	0.832	0.699	0.840	**0.800**	0.739	0.876	0.751	0.850	**0.944**	0.834	0.795	0.929	0.826
InMerge [37]	0.805	0.895	0.881	0.710	**0.842**	0.750	**0.770**	0.878	0.801	0.888	0.915	0.818	0.773	0.903	0.830
ViT-S/16 [2]	0.831	0.911	0.867	0.728	0.835	0.751	0.765	0.871	0.816	0.891	0.891	0.829	0.781	0.895	0.833
Ours	**0.832**	**0.919**	**0.891**	**0.742**	0.833	0.721	**0.770**	**0.879**	**0.831**	**0.911**	0.899	0.829	**0.812**	0.918	**0.842**

Table 1. Per-pathology AUC comparison on NIH ChestXRay14. Bold indicates best per column. Columns follow the order: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, Hernia, and Average AUC.

Model	BMCA (%)	AUC (%)
ResNet-18 [27]	65.7	91.2
EfficientNet [25]	66.8	92.0
ConvNeXt [26]	68.4	93.3
MTA-Net [38]	66.7	–
ViT-B/16 [2]	75.6	95.3
Swin-B [3]	76.4	95.6
MNV2+GNet+VGG [39]	64.8	91.4
ViT-S/16 [2]	75.0	95.0
Ours	**77.9**	**96.4**

Table 2. Comparison with different methods on the ISIC-2019 dataset. Best result per metric is shown in bold.

4 Results and Discussion

4.1 Datasets and Experimental Setup

We evaluate on MedMNIST v2 [19, 18] across six datasets (OrganAMNIST, OrganCMNIST, RetinaMNIST, BloodMNIST, DermaMNIST, and ChestMNIST), NIH ChestX-ray14 [22] containing 112,120 frontal chest X-rays across 14 pathology labels, and ISIC-2019 [20, 21] using 25,331 training and 8,238 test dermoscopic images across eight skin lesion classes, using official train/validation/test splits throughout. Balanced Multi-Class Accuracy (BMCA) and AUC are reported for ISIC-2019 to account for class imbalance. Experiments are conducted on a single NVIDIA A100 (40 GB) GPU with a batch size of 32. FadeFormer is trained for 50 epochs using Adam with a learning rate and weight decay of 1×10^{-4}, together with a ReduceLROnPlateau scheduler and early stopping (patience = 5). Cross-entropy is used for single-label classification, binary cross-entropy with logits for multi-label classification. Data augmentation includes random horizontal flips, rotations, color jitter, and random erasing. All images were resized to 224×224. The complete model contains 24.04M parameters and requires 4.70 GFLOPs. The ViT-S/16 baseline and the ViT-S backbone used in FadeFormer are initialized from the same pretrained checkpoint (via `timm`) and kept frozen.

	OrganAMNIST		**OrganCMNIST**		**RetinaMNIST**		**BloodMNIST**		**DermaMNIST**		**ChestMNIST**	
Backbone	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc
ResNet18 [18, 19]	99.8	95.1	99.4	92.0	71.0	49.3	99.8	96.3	87.0	74.8	77.3	94.7
MedViT [5]	99.7	94.3	99.4	92.3	77.3	56.1	99.7	96.8	93.7	78.0	80.5	**95.9**
ViT-B/16 [24]	**99.9**	97.8	**99.9**	93.8	75.3	53.3	**99.9**	99.1	96.5	83.4	81.2	94.8
DINO ViT-B/16 [24]	**99.9**	97.5	**99.9**	**95.1**	82.6	**63.7**	**99.9**	98.9	97.1	**84.9**	76.1	94.8
SAM ViT-B/16 [24]	99.6	95.9	99.0	94.6	76.0	60.1	99.7	98.6	51.0	66.9	62.2	94.7
ViT-S/16 [2]	99.8	97.1	99.8	93.2	76.4	56.8	99.8	98.6	95.7	82.0	80.1	94.5
Ours	**99.9**	**98.5**	**99.9**	94.9	**83.9**	61.7	**99.9**	**99.2**	**97.2**	84.3	**81.9**	95.3

Table 3. AUC (%) and Accuracy (%) on MedMNIST v2. Best result per dataset per metric is shown in bold.

Variant	**AUC**	**BMCA (%)**
FadeFormer (Full)	**0.9641**	**77.93**
w/o Diffusion	0.9445	73.97
w/o Adaptive Fusion	0.9523	75.84
w/o Multi-Scale Merging	0.9578	76.41
$k = 5$	0.9595	75.09
$k = 20$	0.9574	76.13

Table 4. Ablation study on the ISIC-2019 dataset. The complete FadeFormer achieves the best classification performance among all variants.

4.2 Ablation Study

Table 4 presents an ablation study on ISIC-2019 evaluating the contribution of each component. Removing the graph diffusion module causes the largest drop in AUC, indicating that content-adaptive diffusion is the primary contributor. Replacing the learned adaptive fusion gate with a fixed equal-weight blend reduces AUC to 0.9523, demonstrating that dynamic regulation of the attention-diffusion trade-off is beneficial. Removing multi-scale token merging similarly reduces performance (AUC 0.9578), showing that coarse contextual tokens complement fine-grained spatial representations. Among the graph sparsity variants, FadeFormer (Full) achieves the best balance, with both k=5 and k=20 yielding lower performance.

4.3 Classification Performance

FadeFormer consistently achieves improved or competitive performance across classification tasks of varying complexity, from fine-grained skin lesion grading to multi-label pathology detection. On MedMNIST v2 (Table 3), FadeFormer achieves the highest or tied-highest AUC on all six datasets and the best accuracy on OrganAMNIST and BloodMNIST. On OrganCMNIST, RetinaMNIST, and DermaMNIST, FadeFormer trails DINO ViT-B/16 in accuracy by small margins, and on ChestMNIST it trails MedViT, despite using a smaller backbone.

On NIH ChestX-ray14 (Table 1), FadeFormer achieves the highest average AUC of 0.842 and obtains the best performance on nine of the fourteen disease categories, including Atelectasis, Cardiomegaly, Effusion, Infiltration, Consolidation, Edema, Pneumothorax, Pleural Thickening, and Pneumonia. On ISIC-2019 (Table 2), FadeFormer achieves the highest AUC (96.4%) and balanced

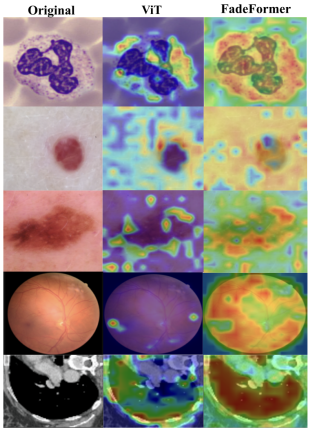

Fig. 2. GradCAM Visualizations. FadeFormer (right) produces activation maps that follow lesion and cellular boundaries more continuously compared to ViT (middle), consistent with diffusion constrained by learned token similarity rather than unconstrained global mixing.

multiclass accuracy (77.9%), outperforming all compared methods across both metrics.

4.4 Qualitative Analysis

Figure 2 presents GradCAM visualizations comparing ViT-S and FadeFormer. FadeFormer produces activation maps that follow anatomical structures more continuously, consistent with diffusion constrained by the learned content-adaptive graph rather than unconstrained global mixing. This is further supported by the ablation study, where removing the diffusion module reduces AUC from 0.9641 to 0.9445 on ISIC-2019 (Table 4), suggesting that the diffusion component contributes meaningfully to classification performance.

5 Conclusion and Future Work

We presented **FadeFormer**, a transformer architecture that augments self-attention with content-adaptive graph diffusion for structured feature propagation across semantically related tokens. Experiments on NIH ChestX-ray14, ISIC-2019, and MedMNIST v2 demonstrate competitive classification performance and qualitatively broader activation coverage than the standard ViT-S baseline. Limitations include graph construction overhead and the lack of clinical validation. Future work will extend FadeFormer to dense prediction tasks such as medical image segmentation.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS (2017)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
3. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV, pp. 10012–10022 (2021)
4. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, Y., Zhang, L.: CvT: Introducing convolutions to vision transformers. In: ICCV, pp. 22–31 (2021)
5. Manzari, O.N., Ahmadabadi, H., Kashiani, H., Shokouhi, S.B., Ayatollahi, A.: MedViT: A robust vision transformer for generalized medical image classification. Comput. Biol. Med. **157**, 106791 (2023)
6. Xia, Z., Li, H., Lan, L.: MedFormer: Hierarchical medical vision transformer with content-aware dual sparse selection attention. IEEE Trans. Med. Imaging (2025). arXiv:2507.02488
7. Wu, Q., et al.: DIFFormer: Scalable (graph) transformers induced by energy constrained diffusion. In: ICLR (2023)
8. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Xu, A.Y., Zhou, Q.: TransUNet: Transformers make strong encoders for medical image segmentation. In: MICCAI (2021)
9. Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, S.A., Zheng, L.: Transformers in medical imaging: A survey. Med. Image Anal. **88**, 102839 (2023)
10. Zhou, Y., et al.: TransFuse: Fusing transformers and CNNs for medical image segmentation. In: MICCAI, pp. 14–23. Springer (2021)
11. Bo, D., Shi, C., Wang, L., Liao, R.: SpecFormer: Spectral graph neural networks meet transformers. In: ICLR (2023)
12. Li, X.J., Yang, J., Zhang, F.L.: Laplacian mesh transformer: Dual attention and topology aware network for 3D mesh classification and segmentation. In: ECCV, p. 532. Springer (2022)
13. Azad, R., Kazerouni, A., Azad, B., Aghdam, E.K., Velichko, Y., Bagci, U., Merhof, D.: Laplacian-Former: Overcoming the limitations of vision transformers in local texture detection. In: MICCAI, pp. 608–618. Springer (2023)
14. Chen, Y., Su, D., Luo, J.: Laplacian-guided hierarchical transformer: A network for medical image segmentation. Comput. Methods Programs Biomed. **260**, 108526 (2025)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
16. Tu, Z., et al.: MaxViT: Multi-axis vision transformer. In: ECCV (2022)
17. Dai, Z., Liu, H., Le, Q.V., Tan, M.: CoAtNet: Marrying convolution and attention for all data sizes. NeurIPS (2021)
18. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: MedMNIST v2 – A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Sci. Data **10**(1), 41 (2023)
19. Yang, J., Shi, R., Ni, B.: MedMNIST classification decathlon: A lightweight AutoML benchmark for medical image analysis. In: ISBI, pp. 191–195 (2021)
20. Codella, N., Rotemberg, V., Gutman, P., et al.: Skin lesion analysis toward melanoma detection 2019: A challenge hosted by the ISIC. arXiv:1902.03368 (2019)

21. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**, 180161 (2018)
22. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: Hospital-scale chest x-ray database and benchmarks. In: CVPR, pp. 2097–2106 (2017)
23. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: ICCV, pp. 618–626 (2017)
24. Wu, F., Papież, B.W.: Rethinking foundation models for medical image classification through a benchmark study on MedMNIST. arXiv:2501.14685 (2025)
25. Tan, M., Le, Q.V.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: ICML, pp. 6105–6114. PMLR (2019)
26. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: CVPR, pp. 11976–11986 (2022)
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
28. Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., Lyman, K.: Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint arXiv:1710.10501 (2017)
29. Tang, Y., Wang, X., Harrison, A.P., Lu, L., Xiao, J., Summers, R.M.: Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In: MLMI Workshop, MICCAI, pp. 249–258. Springer (2018)
30. Rajpurkar, P., et al.: CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
31. Guendel, S., Grbic, S., Georgescu, B., Liu, S., Maier, A., Comaniciu, D.: Learning to recognize abnormalities in chest x-rays with location-aware dense networks. In: CIARP, pp. 757–765. Springer (2019)
32. Guan, Q., Huang, Y.: Multi-label chest x-ray image classification via category-wise residual attention learning. Pattern Recognit. Lett. **130**, 259–266 (2020)
33. Ma, C., Wang, H., Hoi, S.C.H.: Multi-label thoracic disease image classification with cross-attention networks. In: MICCAI, pp. 730–738. Springer (2019)
34. Chen, B., Li, J., Guo, X., Lu, G.: DualCheXNet: Dual asymmetric feature learning for thoracic disease classification in chest x-rays. Biomed. Signal Process. Control **53**, 101554 (2019)
35. Chen, B., Li, J., Lu, G., Zhang, D.: Lesion location attention guided network for multi-label thoracic disease classification in chest x-rays. IEEE J. Biomed. Health Inform. **24**(7), 2016–2027 (2019)
36. Chen, B., Li, J., Lu, G., Yu, H., Zhang, D.: Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. IEEE J. Biomed. Health Inform. **24**(8), 2292–2302 (2020)
37. Wang, H., Almakky, I., Ma, C., Saeed, N., Yaqub, M.: In-model merging for enhancing the robustness of medical imaging classification models. arXiv preprint arXiv:2502.20516 (2025)
38. Adebiyi, A., et al.: MTA-Net: Multi-scale triplet attention-aware network for multiclass skin lesion classification. Comput. Biol. Med. (2025)
39. Gessert, N., Nielsen, M., Shaikh, M., Werner, R., Schlaefer, A.: Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. Med. Image Anal. **66**, 101797 (2020)