Università
degli Studi
di Padova

CROSS-ATTENTION AND RECONSTRUCTION-REGULARIZED TRANSFORMERS FOR MULTI-LABEL CHEST X-RAY CLASSIFICATION

MACHINE LEARNING FOR HUMAN DATA
A.Y. 2025 - 2026

Ngoc Hoa Pham
Nhu Ngoc Hoang

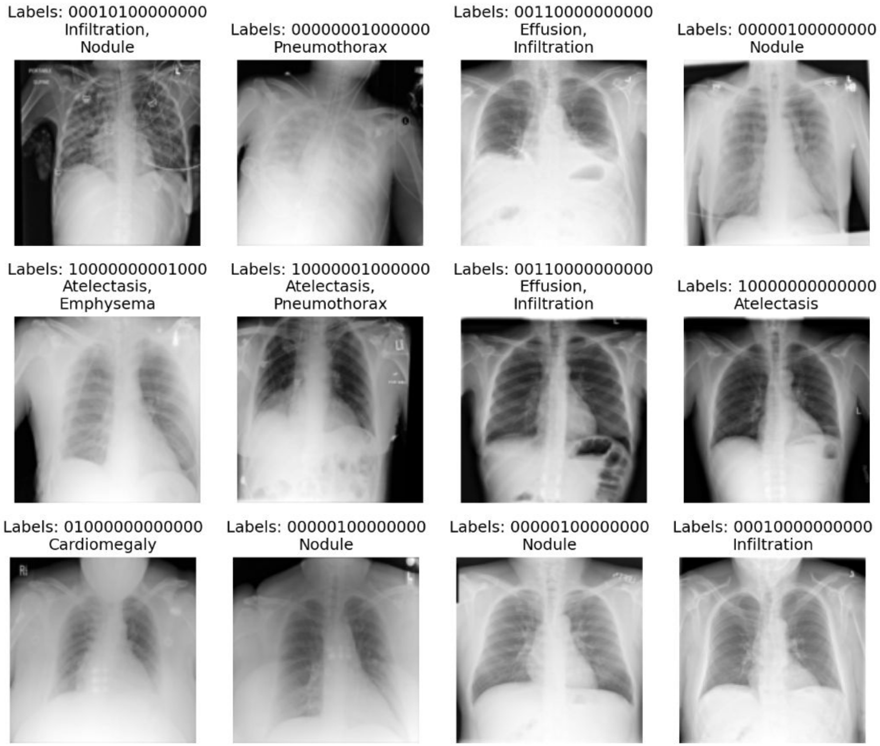
CONTENTS

PROBLEM OVERVIEW

Dataset
Processing pipeline

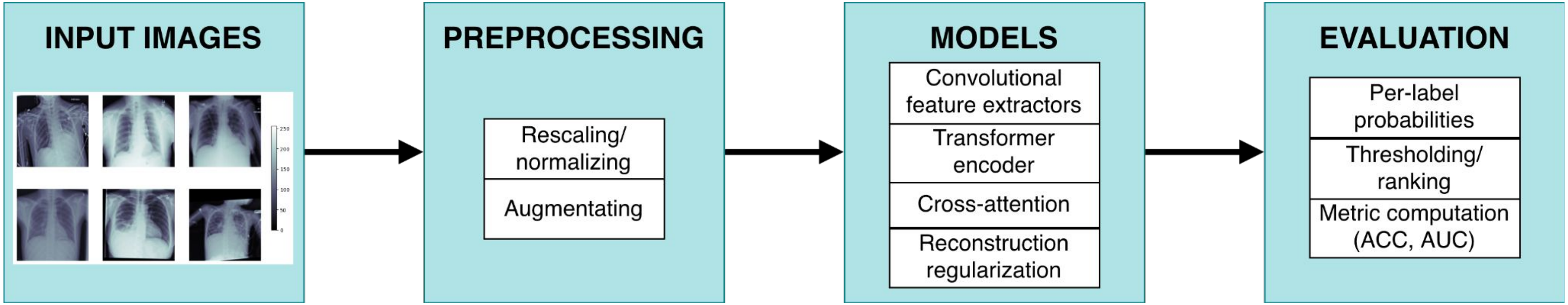
ChestMNIST dataset

- Task: multi-label lung disease classification
- Dataset: ChestMNIST [1], standardized dataset of 112,120 X-ray images from 30,805 patients
- 3 resolutions: 64×64, 128×128, 224×224
- Label: 14-D binary vector = 14 disease labels
- Possible to have no associated disease
- Pre-split train/validation/test sets (78,468/22,433/11,219 images)



[1] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification," Scientific Data, vol. 10, Jan. 2023.

Processing pipeline



$$H \times H \times 1$$

Grayscale images, each pixel = integer in range [0, 255]

Rescaling

Divide all by 255 so each pixel becomes a float in range [0, 1]

Augmentation

Lightweight geometric transformations: flipping, rotating, zooming, translating

Parameterized function mapping input image to a multi-label classification vector

Compare predicted probability vector with ground-truth vector

Measure performance (ACC, AUC), complexity (# parameters), and latency (inference time)

LEARNING FRAMEWORKS

Baseline models
Proposed models

Baseline models

CNN-BASED

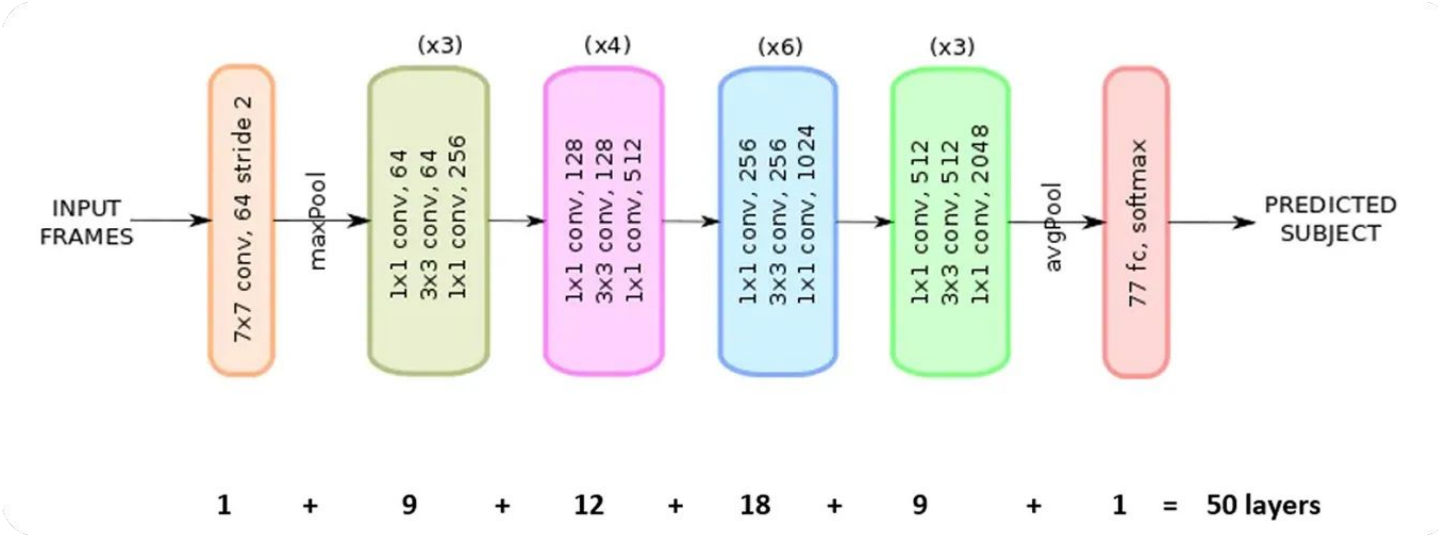
1. **ResNet50**: Residual connection
2. **DenseNet**: Dense features between layers
3. **EfficientNet-B0**: Compound scaling

HYBRID

4. **MedViT**: Vision Transformer

Baseline models - ResNet50

- 49 convolutional layers
- 1 fully connected layer.



ResNet50 Architecture [2]

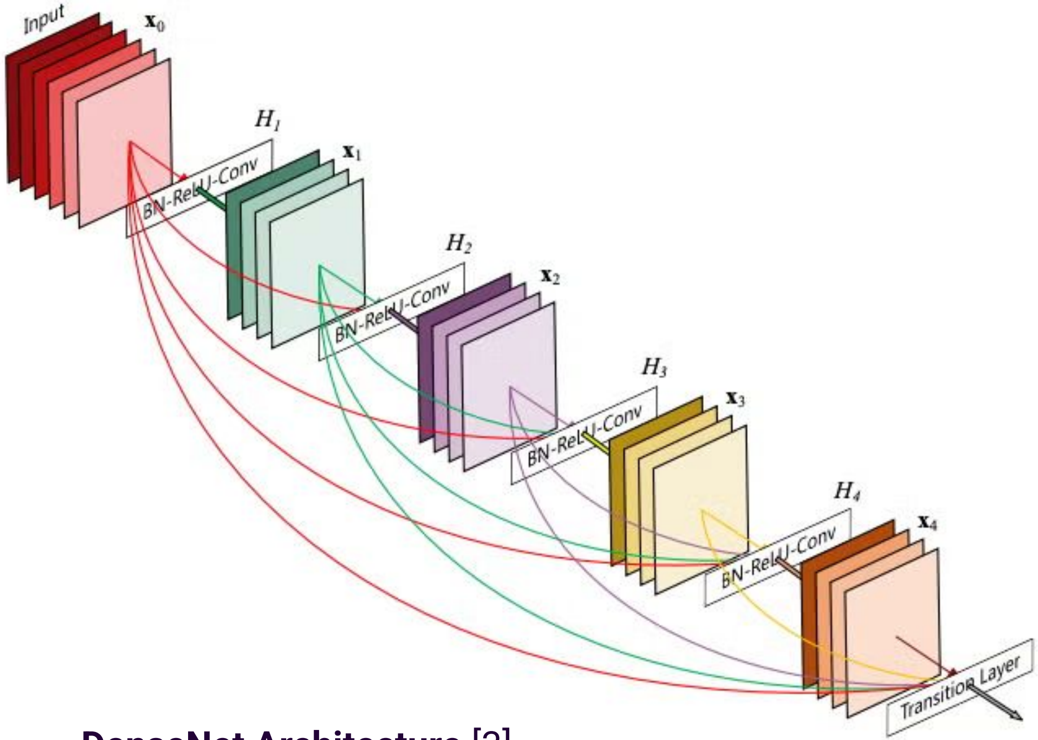
[2] ResNet-50 Architecture

Baseline models - DenseNet

Università degli Studi di Padova

Dense connectivity between layers:

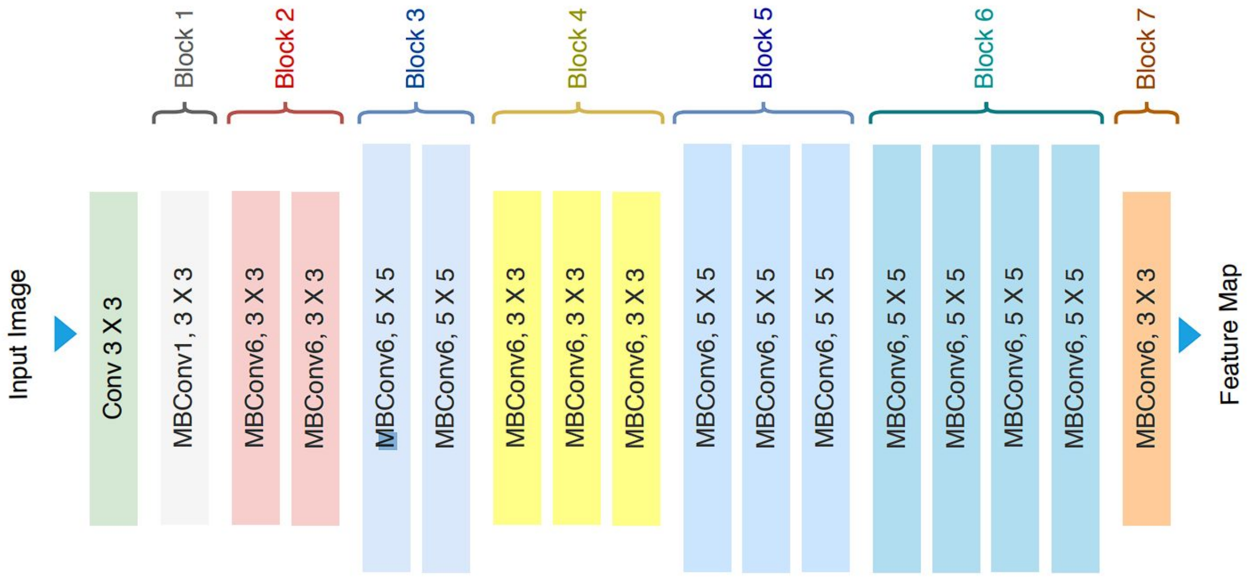
- Each layer receives feature maps from all previous layers
- Feature concatenation instead of summation



DenseNet Architecture [3]

[3] *DenseNet Architecture*

Baseline models - EfficientNetB0



EfficientNetB0 Architecture [4]

- Mobile Inverted Bottleneck (MBConv) blocks
- Squeeze-and-excitation (SE) modules

[4] *EfficientNetB0 Architecture*

Baseline models - MedViT

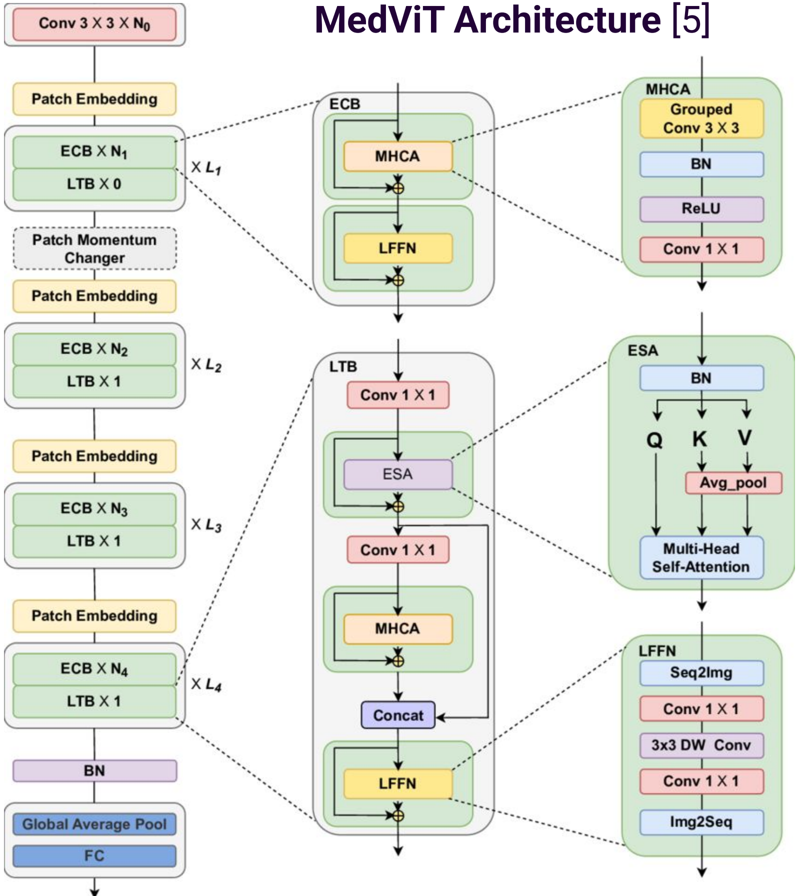
Università degli Studi di Padova

4 stages:
- Patch Embedding layer
- Efficient Convolution Block (ECB):
  - MHCA as token mixer
  - LFFN as depth-wise convolution
- Local Transformer Block (LTB):
  - Captures low-frequency signals with ESA
  - Captures parallel information with MHCA

- MHCA: Multi-Head Convolutional Attention
- LFFN: Locally Feed Forward Network
- ESA: Efficient Self-Attention

[5] MedViT Architecture

MedViT Architecture [5]



Proposed models

Università degli Studi di Padova

1. Dual Branch Cross-Attention Transformer (DBCT)

Structured cross-attention framework that explicitly models local - global features

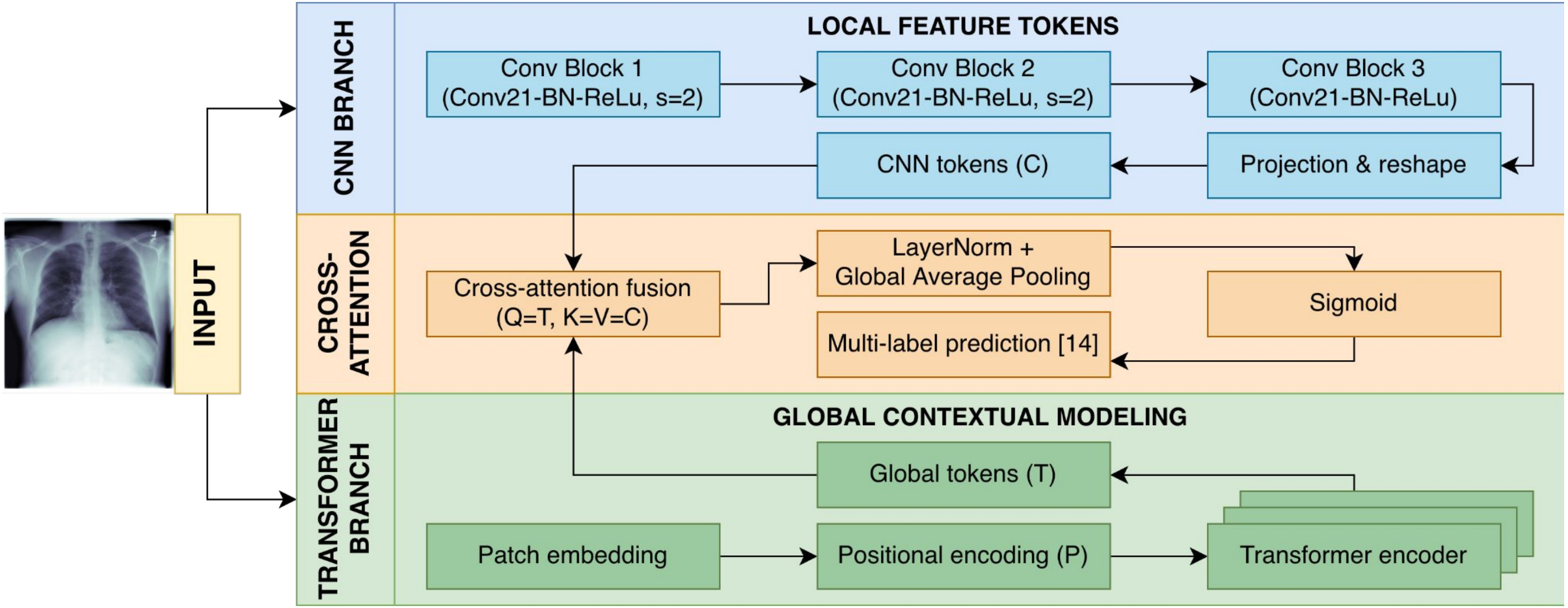
2. Reconstruction-Regularized Vision Transformer (RR- ViT)

Reconstruction-based regularization strategy

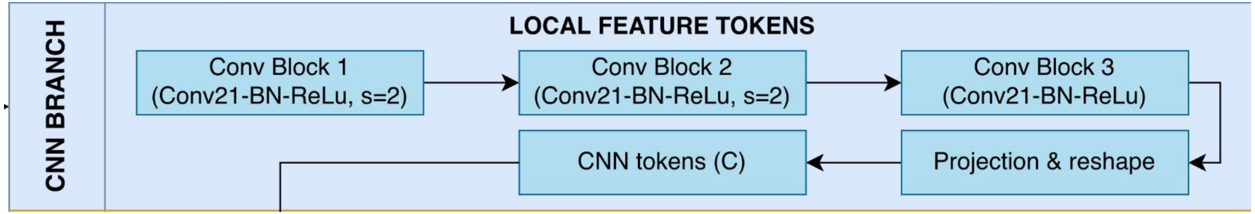
3. Hybrid Reconstruction-Regularized Vision Transformer (Hybrid RR- ViT)

Enhanced reconstruction-based regularization with convolutional stem

Dual Branch Cross-Attention Transformer (DBCT)



Dual Branch Cross-Attention Transformer (DBCT)



- Two convolutional layers use stride 2, progressively reducing spatial resolution:

$$H \times W \to \frac{H}{2} \times \frac{W}{2} \to \frac{H}{4} \times \frac{W}{4}$$

- A 1 × 1 convolution then projects the channel dimension
- The spatial grid is reshaped into a sequence of CNN tokens:

$$\mathbf{C} \in \mathbb{R}^{N_c \times D}, \quad N_c = \frac{H}{4} \cdot \frac{W}{4}$$

Dual Branch Cross-Attention Transformer (DBCT)

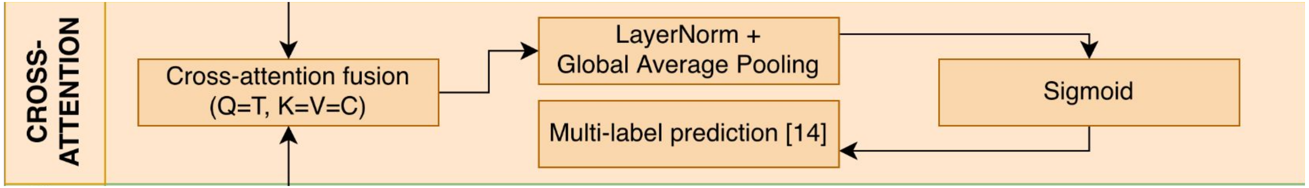
- The input image is partitioned into non-overlapping patches of size p × p:

$$\mathbf{T}_0 \in \mathbb{R}^{N_t \times D}, \quad N_t = \frac{H}{p} \cdot \frac{W}{p}$$

- Positional information is encoded into the token sequence
- The Transformer encoder consists of stacked blocks comprising: multi-head self-attention, feed-forward network (MLP), and layer normalization
- Self-attention is computed as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$$

Dual Branch Cross-Attention Transformer (DBCT)



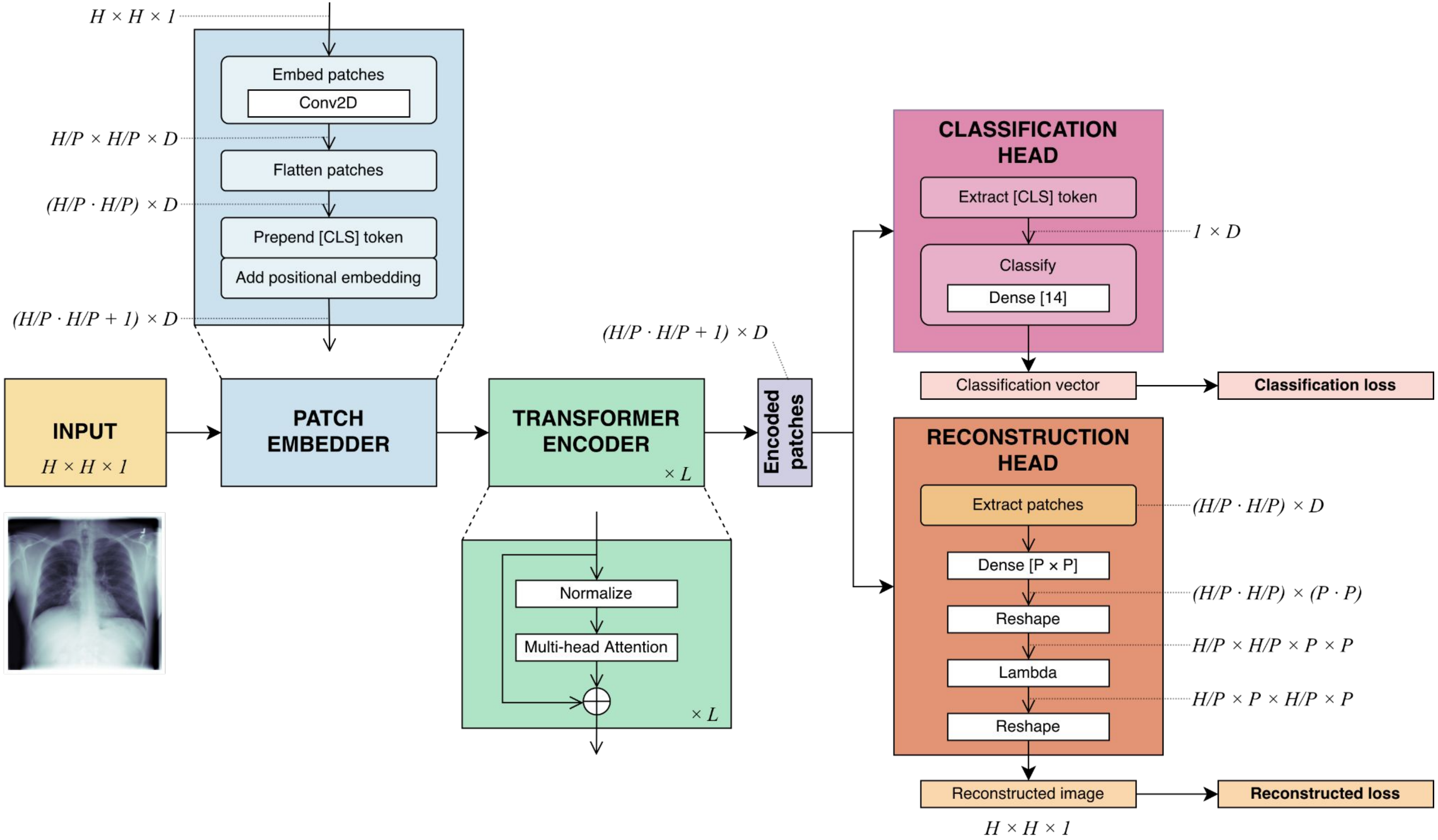
- Integrate global and local information by a cross-attention, the fused representation is computed as:

$$\mathbf{T}' = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

- The fused token sequence is normalized and aggregated via global average pooling
- A fully connected layer followed by a sigmoid activation produces multi-label predictions
- Focal Loss is used as loss function:

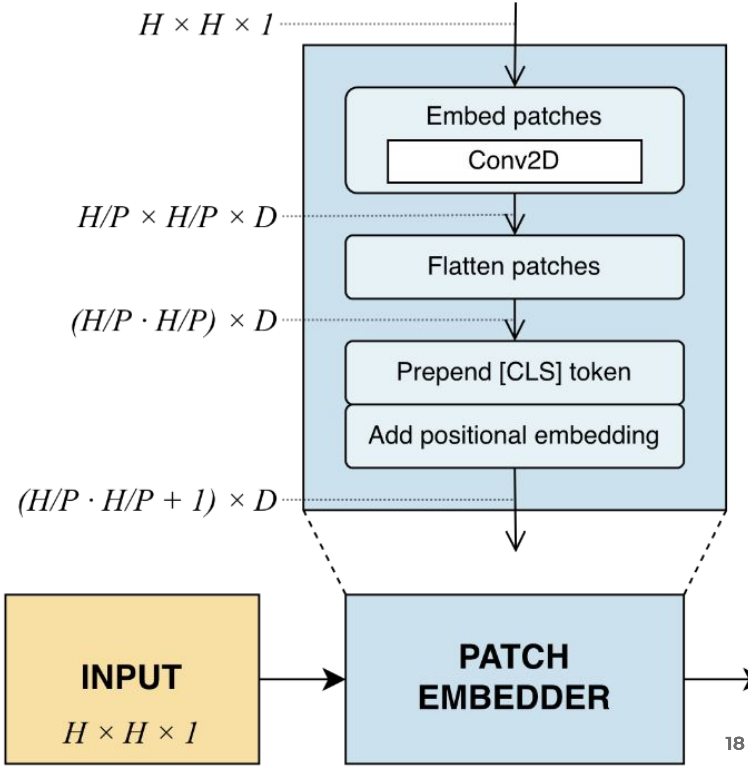
$$\mathcal{L}_{\text{focal}} = -\sum_{k=1}^{14} \alpha(1-\hat{y}_k)^\gamma y_k \log(\hat{y}_k) - (1-\alpha)\hat{y}_k^\gamma(1-y_k)\log(1-\hat{y}_k)$$

Reconstruction-Regularized Vision Transformer (RR- ViT)



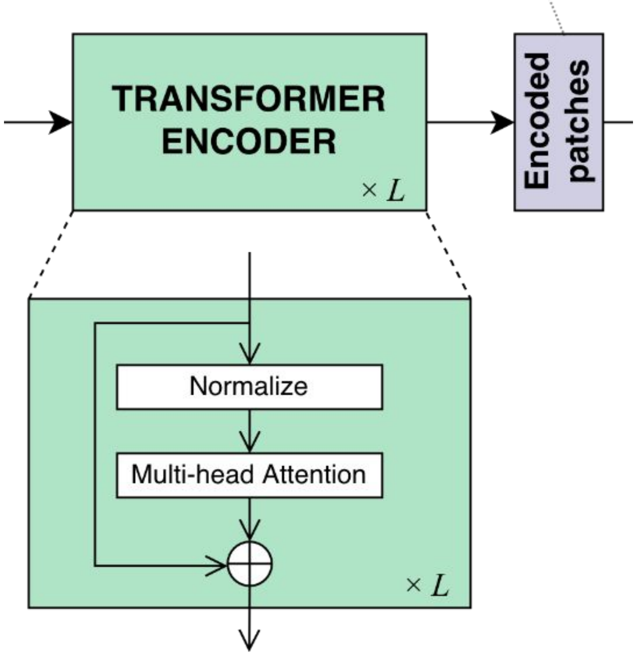
RR-ViT - Patch-based encoder

- Partitions the image into non-overlapping patches
- Each patch is mapped to a D-dimensional latent space using a learnable linear projection
- Prepend learnable [CLS] token to enable global aggregation for classification
- Add learnable 1-dimensional positional embedding to preserve spatial ordering



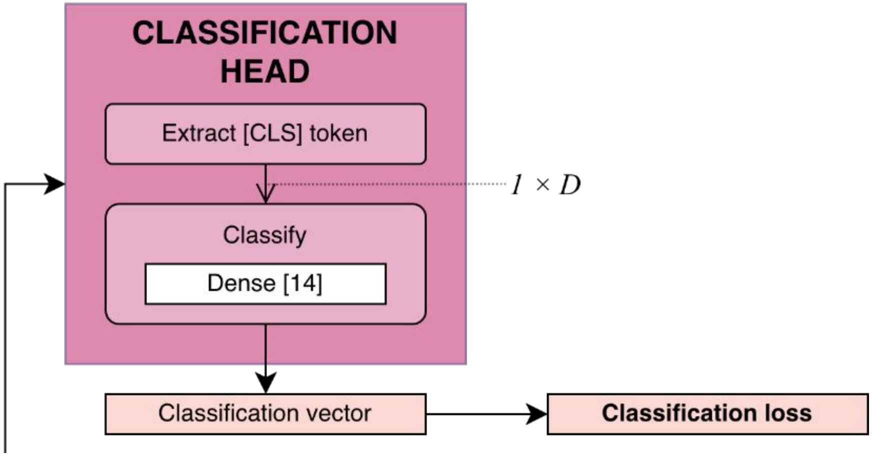
RR-ViT - Patch-based encoder

- L=4 stacked Transformer layers
- Minimal block with pre-normalization and multi-head self-attention (4 attention heads) with residual connections
- Lightweight design and simplified attention mechanism
- Encoded patches with dimension: $\left(\frac{H}{P} \cdot \frac{H}{P} + 1\right) \times D$



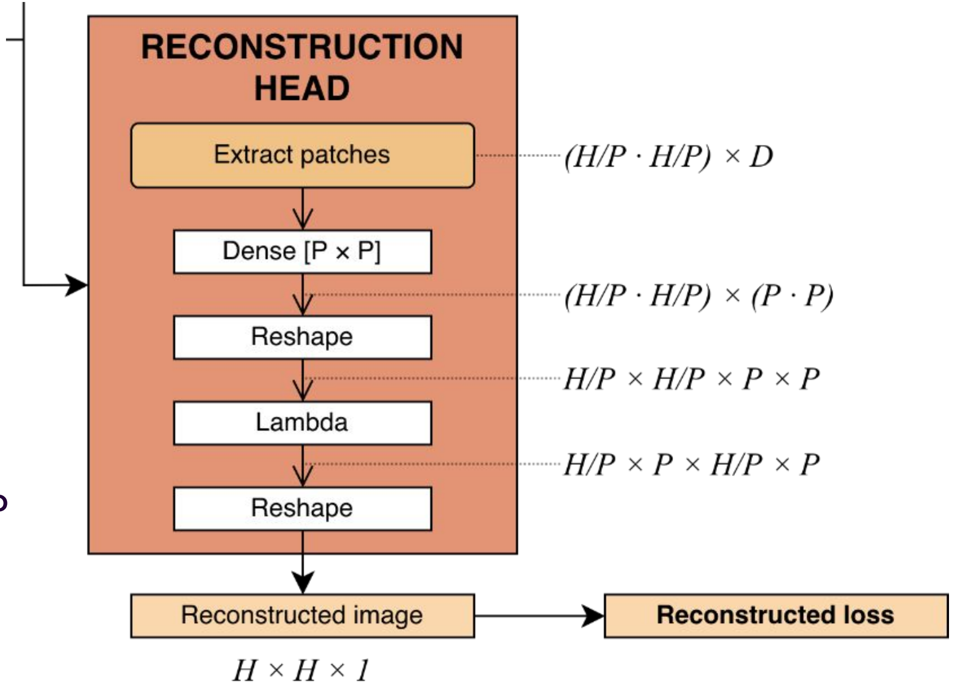
RR-ViT - Classification head

- Uses the final state of the [CLS] token; global representation of the image after having passed through attention mechanism
- Token is passed through a dropout & a dense layer with K=14 units with sigmoid activation
- Output: multi-label binary classification vector
- Classification loss: binary cross-entropy (BCE)

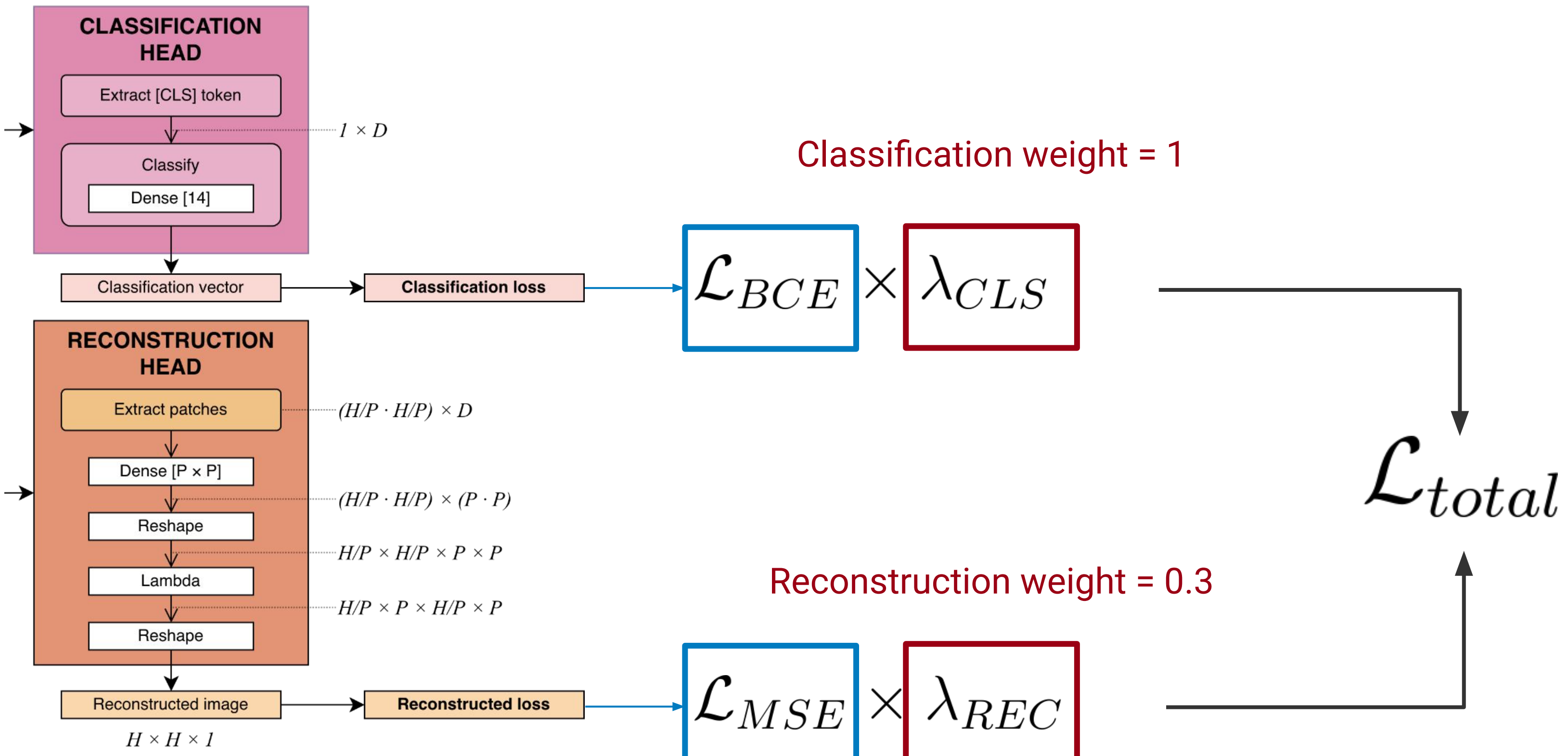


RR-ViT - Reconstruction head

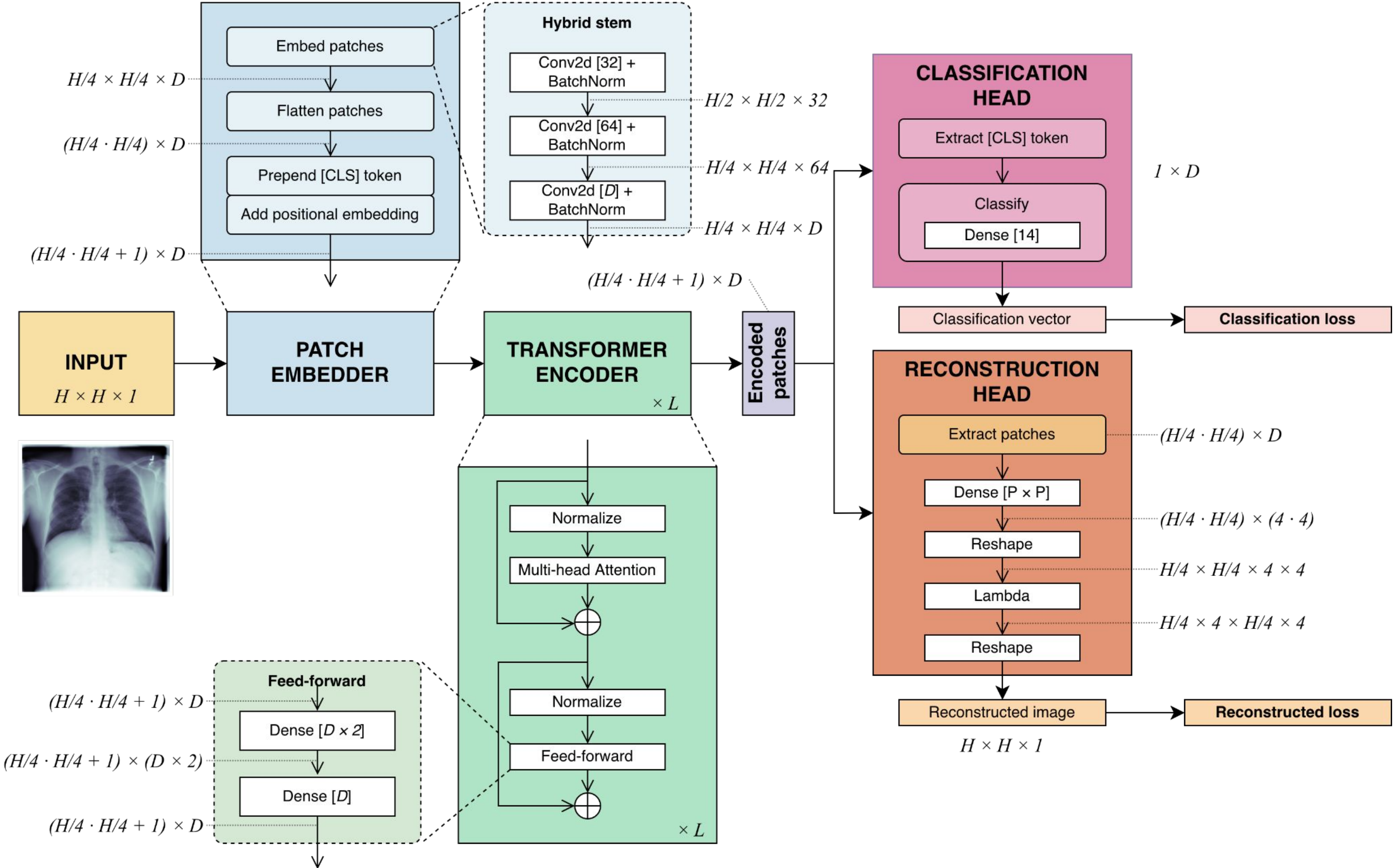
- Use patch tokens discarding [CLS] token to reconstruct input image
- Auxiliary reconstruction objective encourages the encoder to preserve meaningful structure in its latent space
- Lightweight decoder to ensure representation learning happens in the shared encoder
- Each D-d patch token is projected back to P x P pixels using a dense layer
- Reconstruction loss: mean square error (MSE)



RR-ViT - Joint optimization



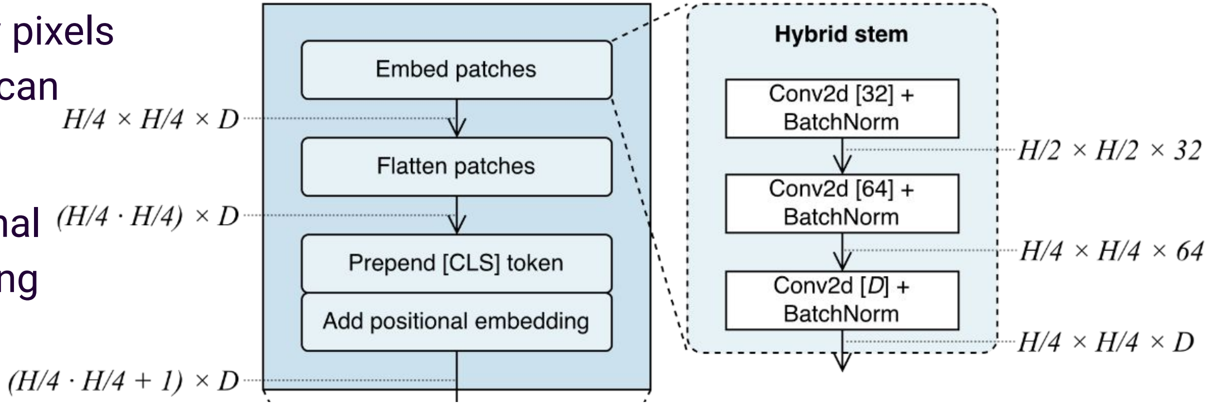
Hybrid Reconstruction-Regularized Vision Transformer (Hybrid RR- ViT)



Hybrid RR-ViT - Convolutional hybrid stem

Università degli Studi di Padova

- **RR-ViT** drawback: mapping raw pixels to patches via linear projection can discard fine-grained details
- **Hybrid RR-ViT** uses convolutional feature maps as tokens, following hybrid tokenization idea [6]

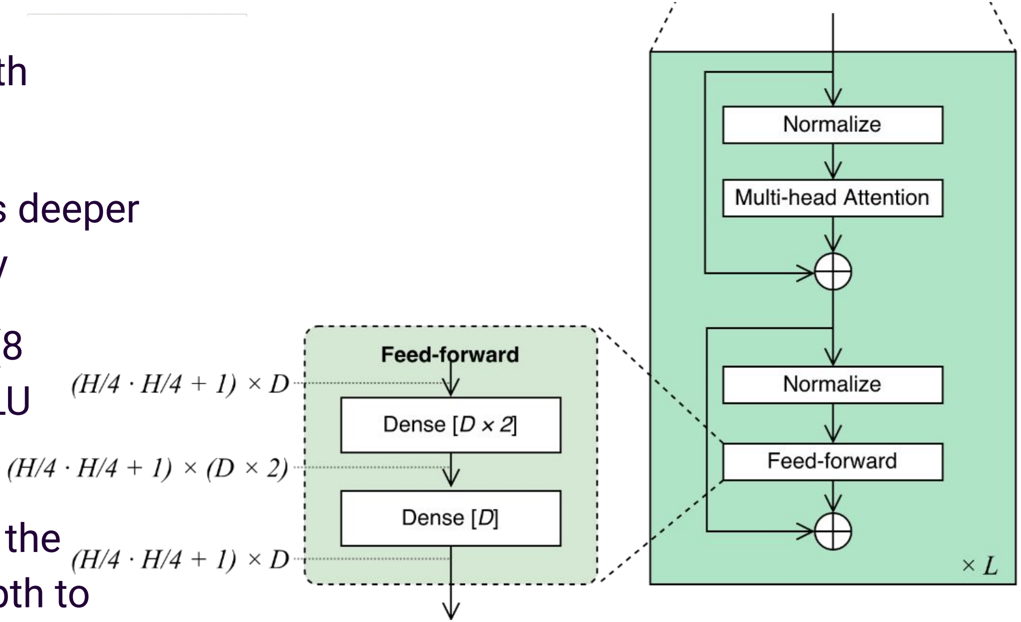


- Hybrid stem: 3 consecutive 3x3 convolutions, allowing extraction of hierarchical representation, ensuring tokens can capture complex local contexts
- Finer tokenization: e.g., 64 ✕ 64 input image
  - RR-ViT with P=8 yields (64/8)^2 = 64 tokens
  - Hybrid RR-ViT yields (64/4)^2 = 256 tokens

[6] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

Hybrid RR-ViT - Upgraded Transformer architecture

- Full standard Transformer architecture with feed-forward network
- Increases representational power, enables deeper stacks without sacrificing training stability
- 2 sublayers: (1) multi-head self-attention (8 attention heads), (2) 2-layer MLP with GELU activation
- L=8 stacked Transformer layers: ensuring the model possesses sufficient non-linear depth to process high-resolution tokens



RESULTS

Evaluation metrics
Experimental setup
Performance comparison
Efficiency considerations

Evaluation metrics

Accuracy (ACC)

- Proportion of sample for which the predicted 14-dimensional label vector exactly matches the ground-truth vector
- Binary prediction obtained by thresholding per-class sigmoid probabilities (0.5 threshold)

$$\mathrm{ACC} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\{\hat{\mathbf{y}}_i = \mathbf{y}_i\}$$

Area under the ROC curve (AUC)

- Capture ranking quality better under class imbalance
- Aggregate class-wise AUC values using a macro-average

$$\mathrm{AUC} = \frac{1}{14} \sum_{k=1}^{14} \mathrm{AUC}_k$$

Experimental setup

Setup

- TensorFlow implementation
- Experiments run on Kaggle's NVIDIA Tesla P100 GPUs
- Experiments on 3 dimensions: 64 × 64, 128 × 128, 224 × 224

Models

- CNN baseline models: ResNet50, DenseNet121, EfficientNetB0
  - TensorFlow implementation
  - Trained from scratch
- MedViT: best reported ChestMNIST performance from original paper
- Hybrid RR-ViT: no results on 224 × 224 due to memory constraints

Performance comparison - 64 × 64

UNIVERSITÀ DEGLI STUDI DI PADOVA

Model	ACC	AUC	# params	Inference
ResNet50	0.947	0.686	24.12M	1.848s
DenseNet121	0.948	0.726	7.30M	3.546s
EfficientNetB0	0.948	0.743	4.38M	2.480s
MedViT	N/A	N/A	N/A	N/A
DBCT	0.947	0.747	2.41M	**0.193s**
RR-ViT	0.947	0.722	**0.29M**	0.339s
Hybrid RR-ViT	0.948	**0.755**	4.46M	1.357s

Performance comparison - 128 × 128

Università degli Studi di Padova

Model	ACC	AUC	# params	Inference
ResNet50	0.948	0.706	24.12M	1.968s
DenseNet121	0.948	0.758	7.30M	3.118s
EfficientNetB0	0.948	0.740	4.38M	2.068s
MedViT	N/A	N/A	N/A	N/A
DBCT	0.947	0.736	2.41M	**0.190s**
RR-ViT	0.948	0.733	**0.32M**	0.351s
Hybrid RR-ViT	0.948	**0.761**	4.66M	10.33s

Performance comparison - 224 × 224

Model	ACC	AUC	# params	Inference
ResNet50	0.947	0.702	24.12M	2.237s
DenseNet121	0.948	0.701	7.30M	3.385s
EfficientNetB0	0.948	0.729	4.38M	1.982s
MedViT	0.959	**0.805**	57.69M	N/A
DBCT	0.947	0.747	2.45M	**0.176s**
RR-ViT	0.948	0.738	**0.38M**	1.323s
Hybrid RR-ViT	N/A	N/A	N/A	N/A

Efficiency considerations

UNIVERSITÀ DEGLI STUDI DI PADOVA

DBCT

- Consistently lightweight & lowest inference times across all dimensions
- Performance comparable/better than CNN baselines at low resolutions
- Room for performance improvement at higher resolution

RR-ViT

- Very parameter efficient: <0.4M parameters
- Performance comparable/close to CNN baselines
- Strong candidate when memory footprint is a primary constraint
- Simplified Transformer might not be enough

Hybrid RR-ViT

- Highest AUC among trained models
- High computational costs: notably slower at 128×128
- High memory costs: cannot be trained at 224×224
- Performance improvements may come at the cost of increased latency and hardware requirements

DEMO