Medical/Bio Research Topics Ⅱ: Week 10 (07.11.2025)

Hands-on AI Regression Model Development (2): Model Architecture

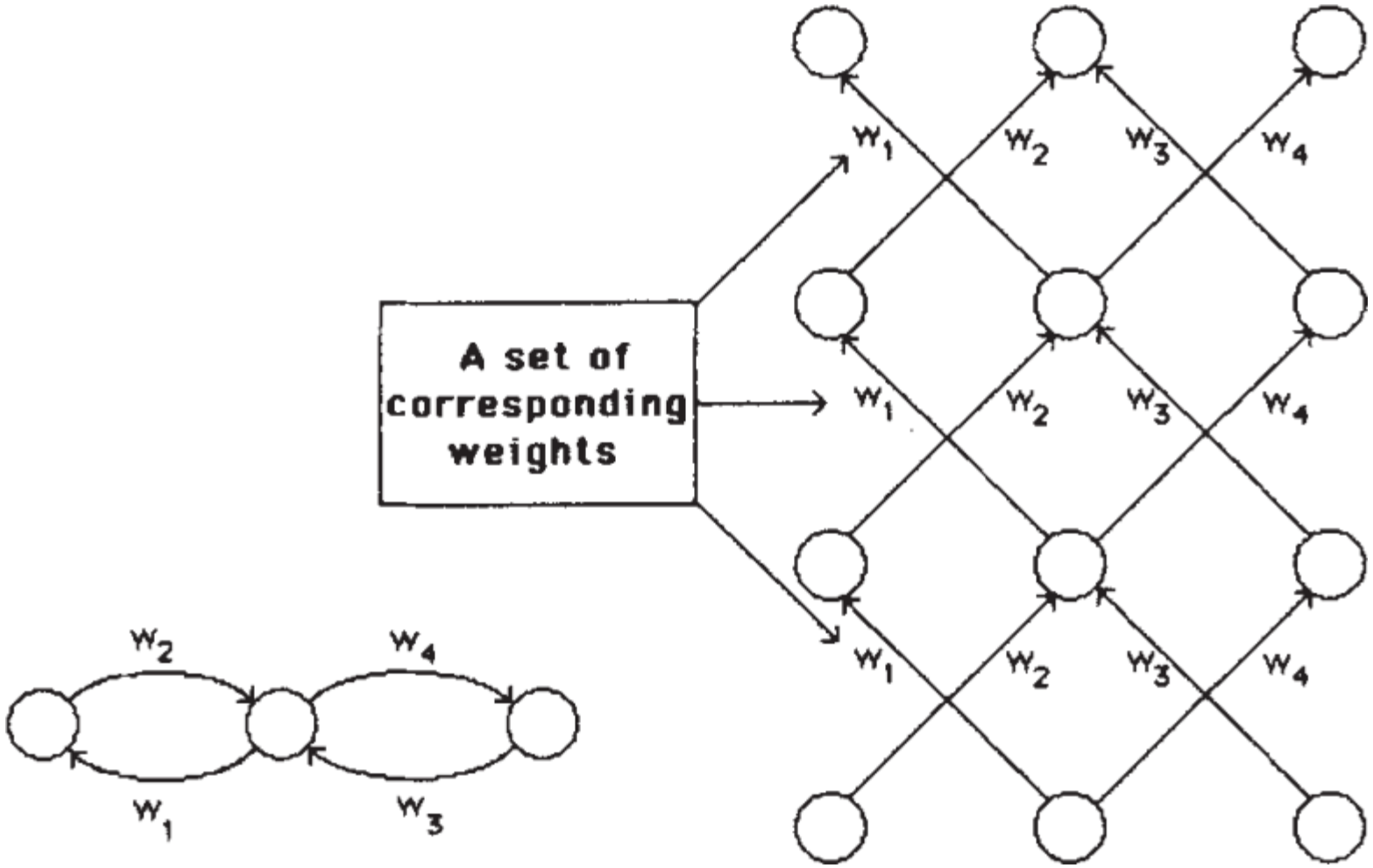
인공지능 회귀 모델 개발 실습 (2): 모델 구조

Deep Learning Architectures for Regression/Classification

- Regression vs. classification
  - Many architectures can be adapted for both regression and classification by modifying the output layer and loss function
  - Output layer
    - Regression: single neuron with linear activation
    - Classification: single/multiple neurons with sigmoid/softmax activation
  - Loss function
    - Regression: Mean absolute error (MAE), mean squared error (MSE)
    - Classification: cross-entropy loss, hinge loss

- Feedforward neural network (FNN) / multilayer perceptron (MLP)
  - "Learning Representations by Back-propagating Errors"

    [Rumelhart et al., 1986]
  - Most basic form of artificial neural networks
    - Composed of multiple layers of fully connected neurons
    - Foundation for all subsequent deep learning architectures
  - Can be used for image processing but inefficient due to:
    - Inability to effectively utilize spatial information
    - Large number of parameters

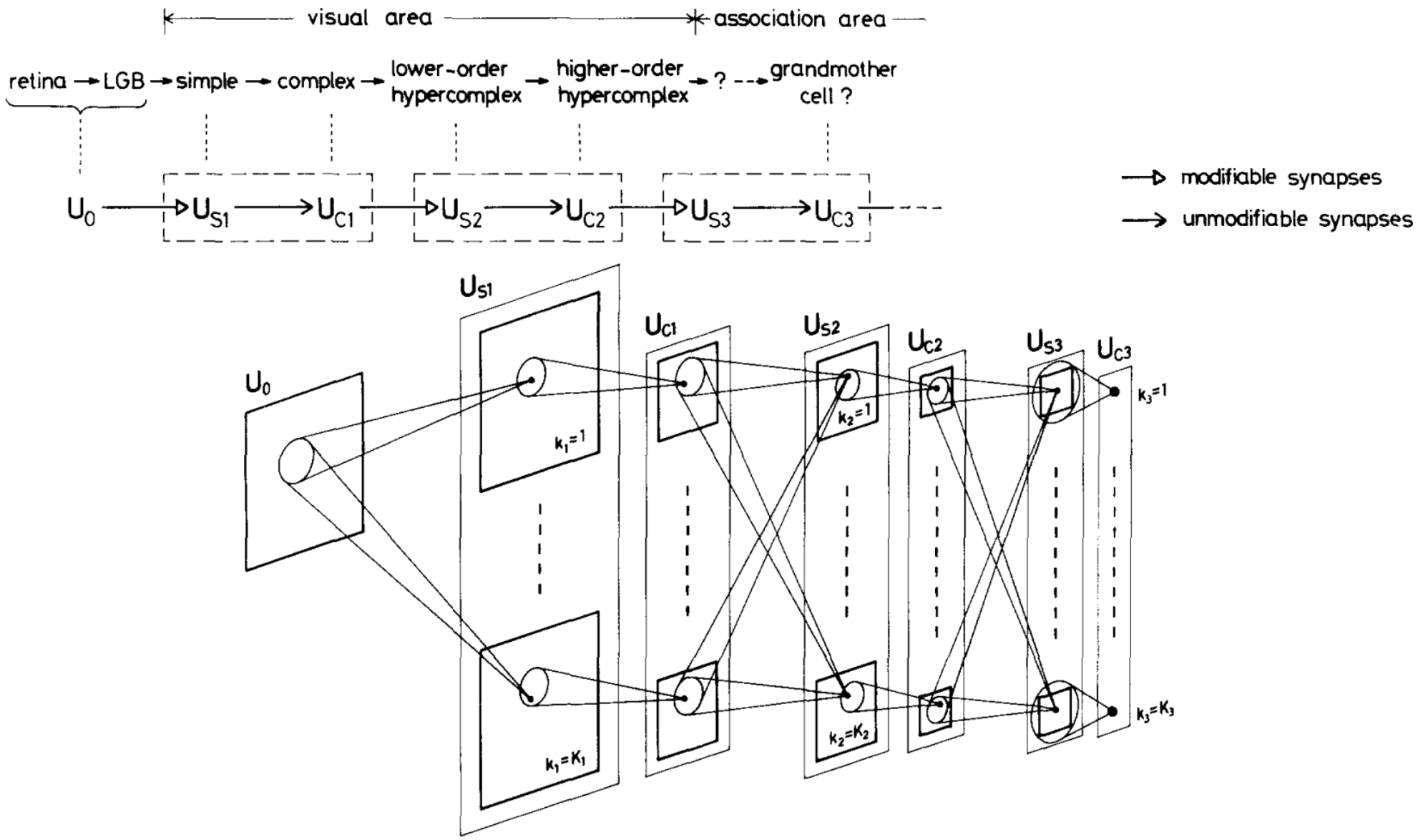


[Rumelhart et al., 1986]

Weight Sharing that Enables Efficient Backpropagation

- Neocognitron
  - "Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position"

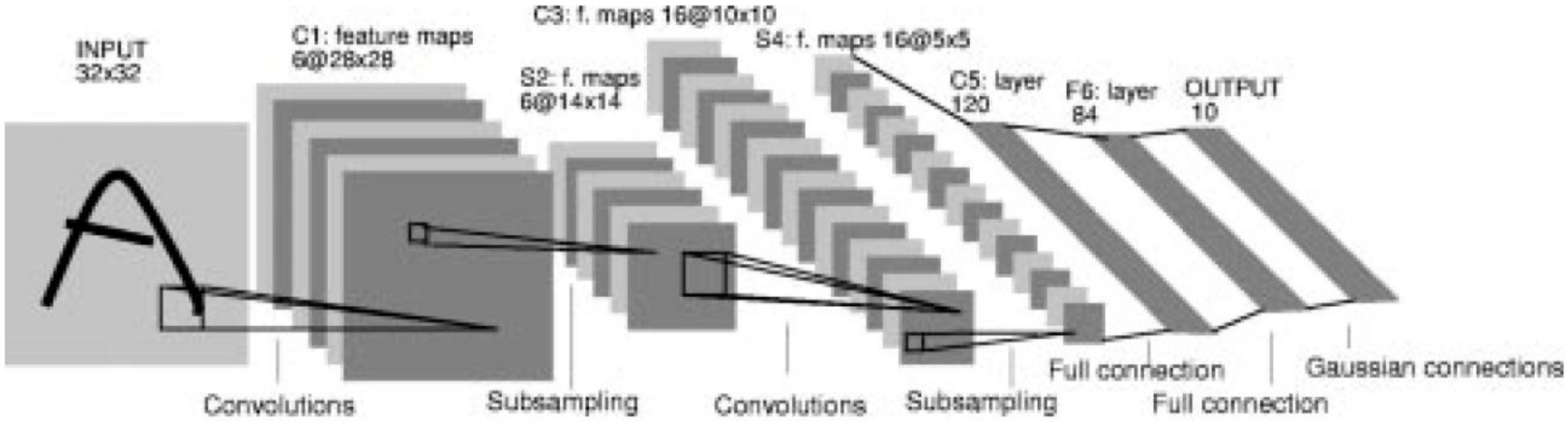
    [Fukushima, 1980]
  - Hierarchical, multi-layered neural network inspired by the visual cortex structure
    - Introduces concepts of simple and complex cells, similar to convolution and pooling
  - Precursor to convolutional neural networks (CNNs), establishing the biological motivation for hierarchical feature learning



[Fukushima, 1980]

Neocognition Architecture that Mimics the Hierarchical Structure of the Visual Cortex

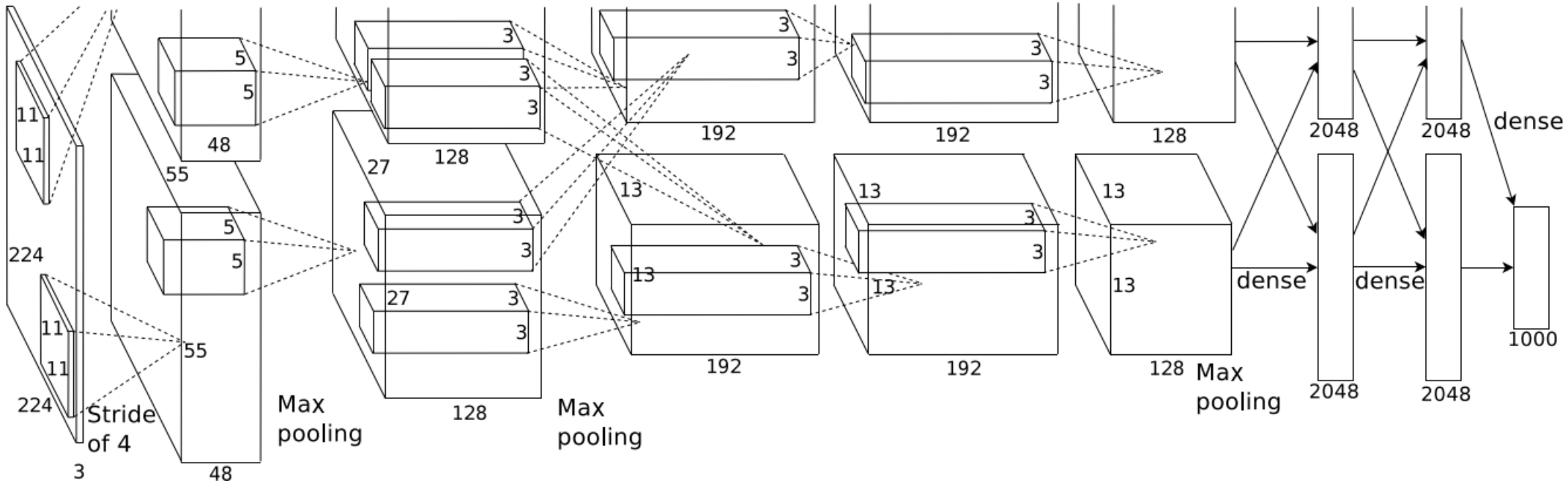
- LeNet-5 (named after Yann LeCun)
  - "Gradient-Based Learning Applied to Document Recognition" [LeCun et al., 1998]
  - Refinement of earlier LeNet series, including LeNet-1 through LeNet-4, developed from the late 1980s to 1998
  - First successful CNN demonstrating end-to-end learning with backpropagation
    - Core CNN structure: convolutional layers, pooling (subsampling) layers, and fully connected layers
  - Introduces larger and more complex architecture (5 layers) compared to its predecessors
    - Relatively shallow compared to modern CNNs: Input → Conv1 → Subsampling1 → Conv2 → Subsampling2 → FC1 → FC2 → Output (FC3)



[LeCun et al., 1998]

LeNet-5 Architecture

- AlexNet (named after Alex Krizhevsky)
  - "ImageNet Classification with Deep Convolutional Neural Networks" [Krizhevsky et al., 2012]
  - Marks the beginning of the deep learning revolution in computer vision
  - Demonstrates the power of deep CNNs (8 layers) for image classification
    - Significantly larger and more complex than LeNet-5: Input → Conv1 → MaxPool1 → Conv2 → MaxPool2 → Conv3 → Conv4 → Conv5 → MaxPool3 → FC1 → FC2 → Output (FC3)
  - Popularizes the use of rectified linear unit (ReLU) activation, dropout, and data augmentation in deep learning



[LeCun et al., 1998]

AlexNet Architecture

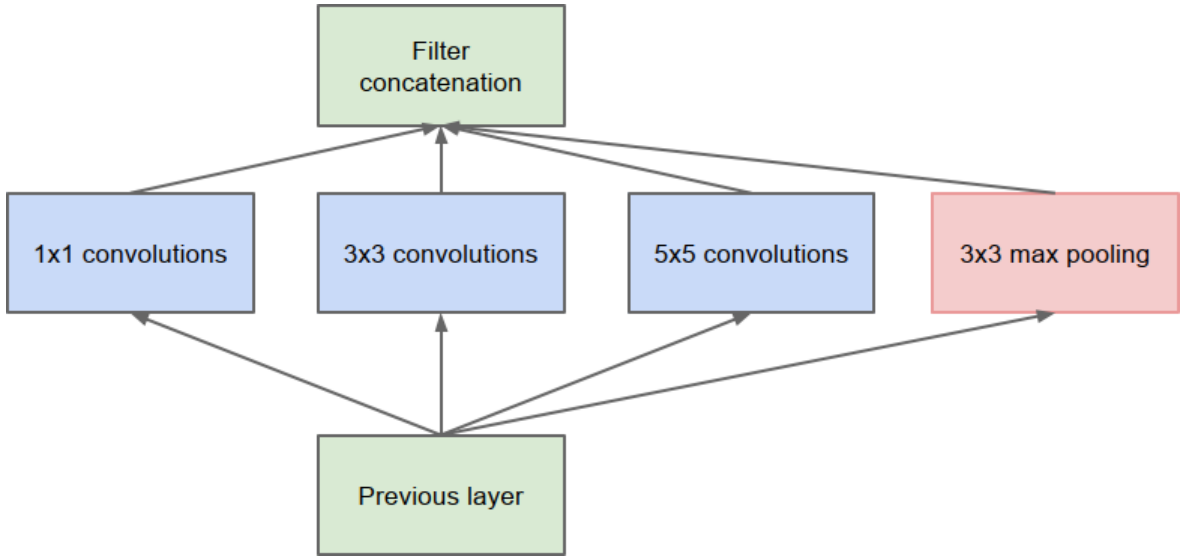
- VGGNet (named after the Visual Geometry Group at Oxford University)
  - "Very Deep Convolutional Networks for Large-Scale Image Recognition" [Simonyan & Zisserman, 2014]
  - Proves the importance of network depth using repeated small 3×3 convolution filters
    - Enables deeper networks while reducing parameters and increasing receptive field
  - Two main versions: VGG16 (13 convolutional layers + 3 fully connected layers), VGG19 (16 convolutional layers + 3 fully connected layers)

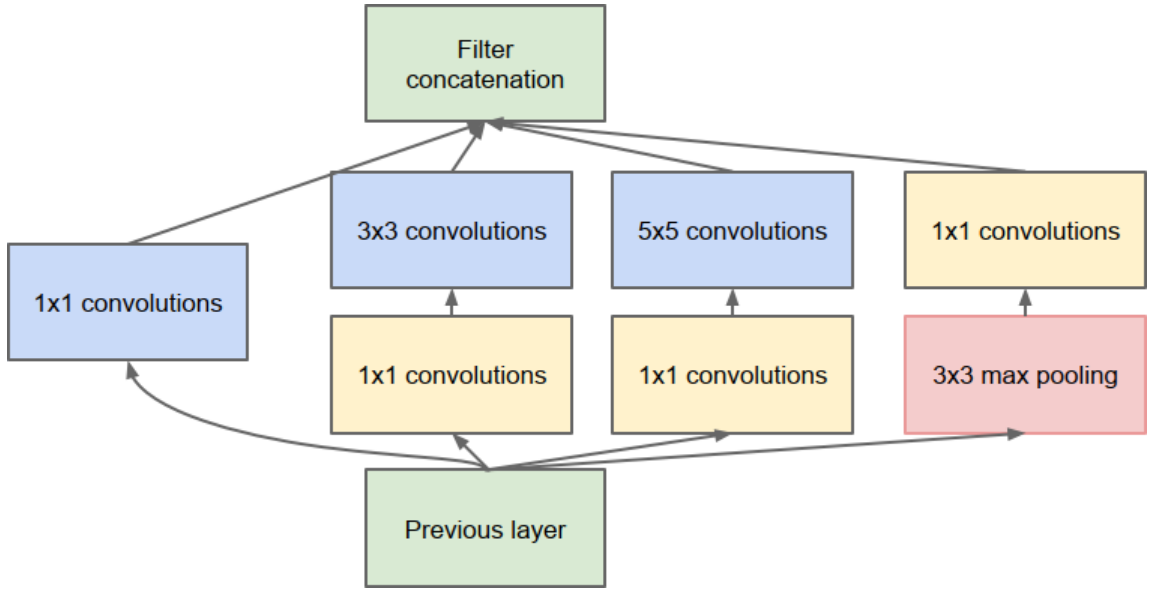
ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

[Simonyan & Zisserman, 2014]

VGGNet Architecture

- GoogLeNet (named after Google and inspired by LeNet)
  - "Going Deeper with Convolutions" [Szegedy et al., 2015]
  - Introduces multi-scale feature learning through parallel convolutional operations (inception modules)
  - Inception modules enable efficient computation while capturing features at different scales
    - 22 layers with 9 Inception modules, advancing both depth and computational efficiency
  - Various versions: Inception v1 (original GoogLeNet), Inception v2, Inception v3, Inception v4, Inception-ResNet

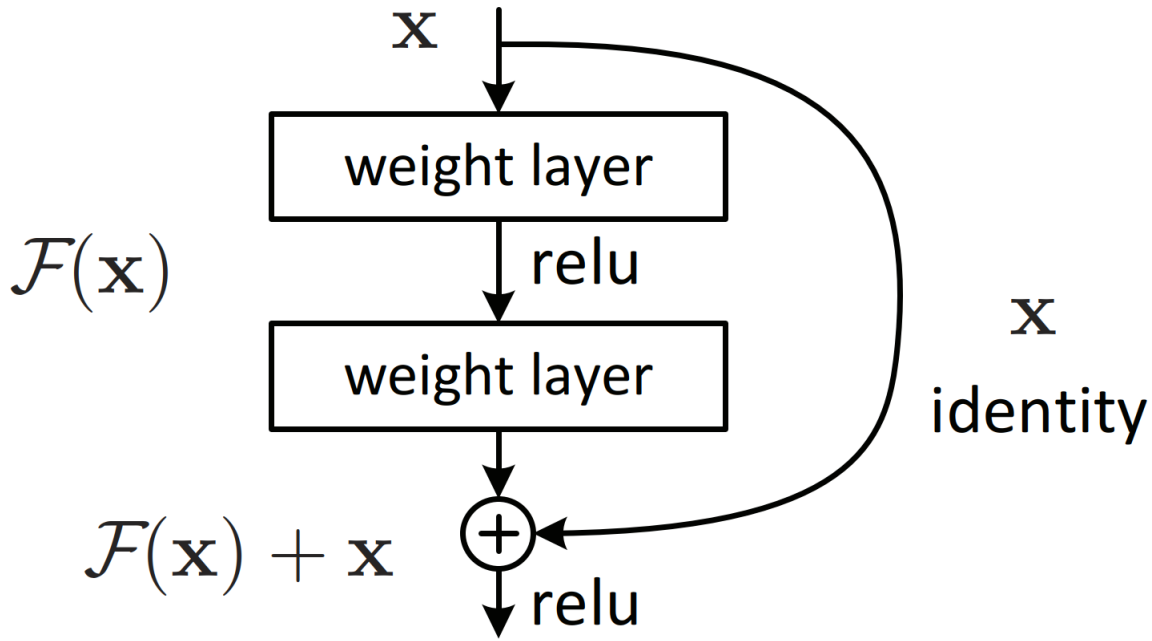




[LeCun et al., 1998]

Inception Module in GoogLeNet

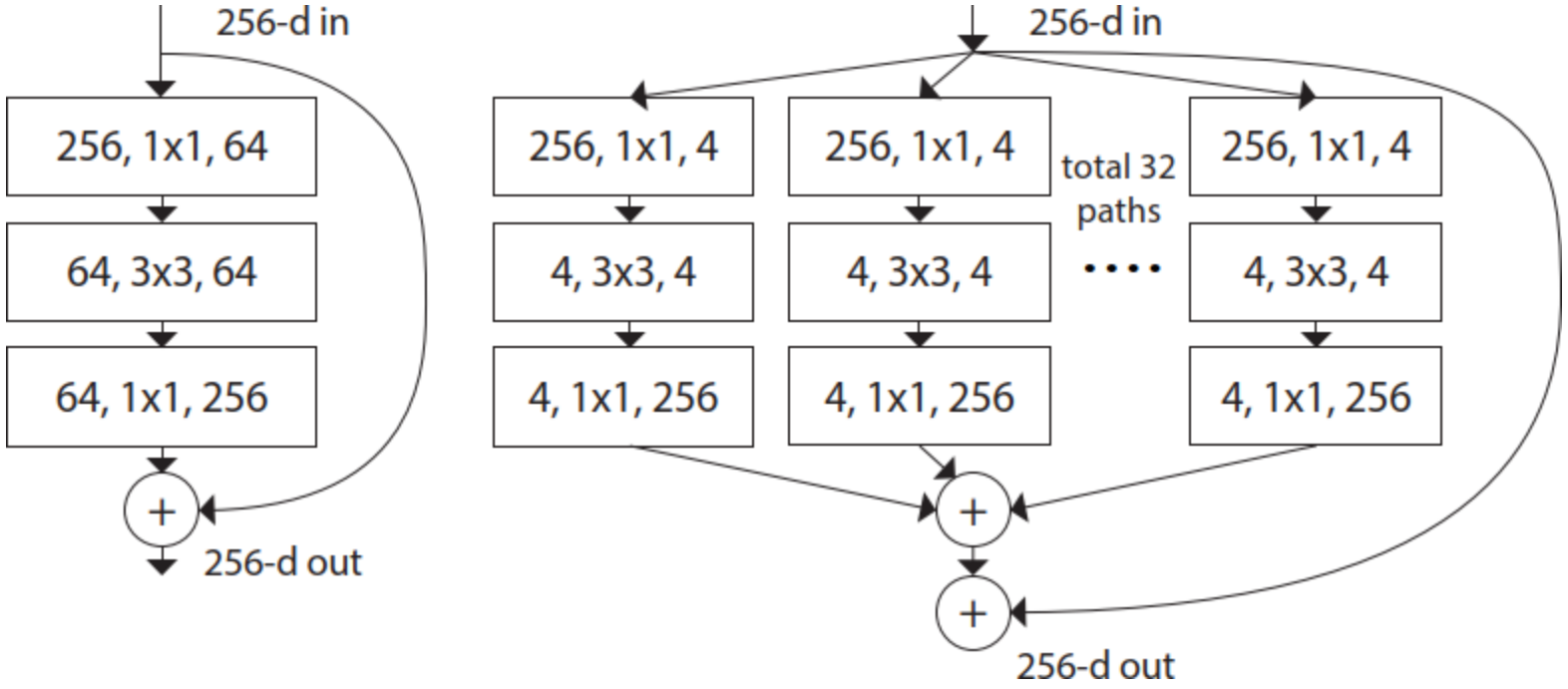
- ResNet (Residual Network)
  - "Deep Residual Learning for Image Recognition" [He et al., 2016]
  - Revolutionary skip connections solve the vanishing gradient problem in very deep networks
  - Became the backbone architecture for most subsequent CNN developments
  - Various versions: ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152



[He et al., 2015]]

Building Block for Residual Learning in ResNet

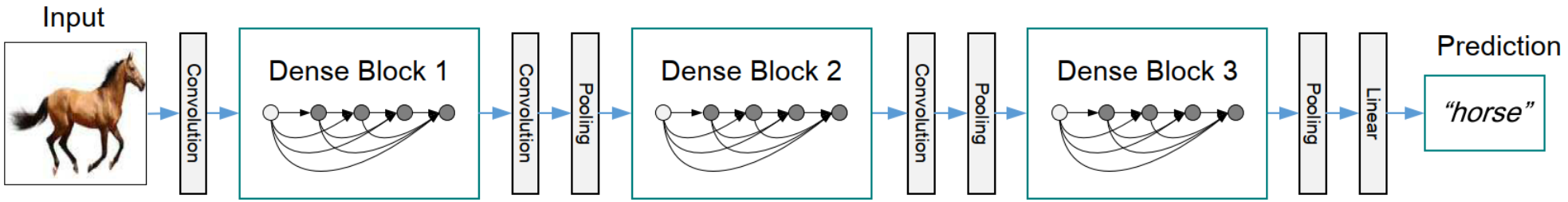
- ResNeXt (Residual Network Extended or Residual Network with Cardinality)
  - "Aggregated Residual Transformations for Deep Neural Networks" [Xie et al., 2017]
  - Introduces cardinality (number of parallel paths in a grouped convolution block) as a new dimension beyond depth and width
    - Demonstrates that increasing cardinality is more effective than going deeper or wider
    - Uses group convolutions for improved parameter efficiency while maintaining performance
  - Various versions: ResNeXt-50, ResNeXt-101, ResNeXt-152

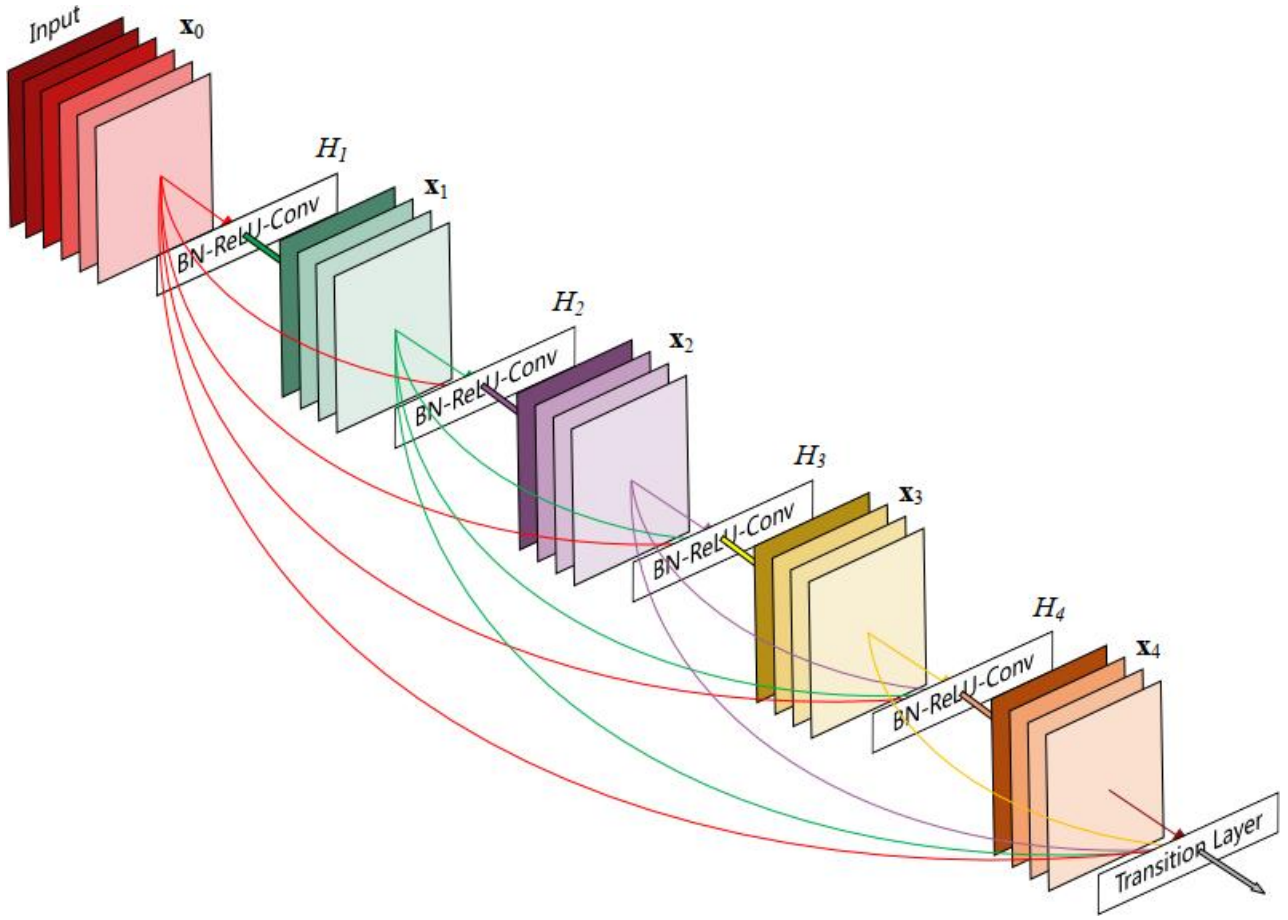


[Xie et al., 2017]]

Comparison of ResNet and ResNeXt (Cardinality = 32) Block Structure

- DenseNet (Densely Connected Convolutional Network)
  - "Densely Connected Convolutional Networks" [Huang et al., 2017]
  - Maximizes feature reuse through dense connectivity between all layers
    - Improves gradient flow and reduces parameters while enhancing feature propagation
  - Various versions: DenseNet-121, DenseNet-169, DenseNet-201, and DenseNet-264

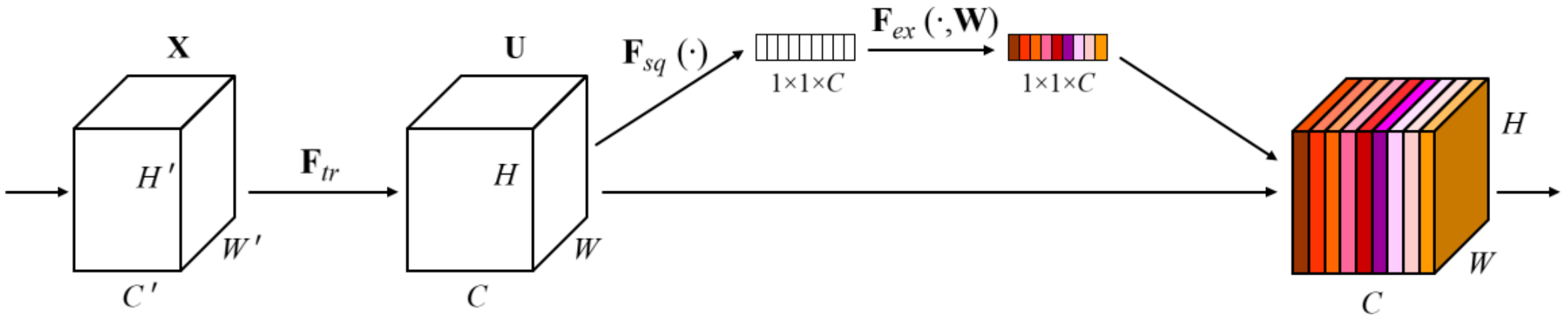




[Huang et al., 2017]

Dense Block in DenseNet

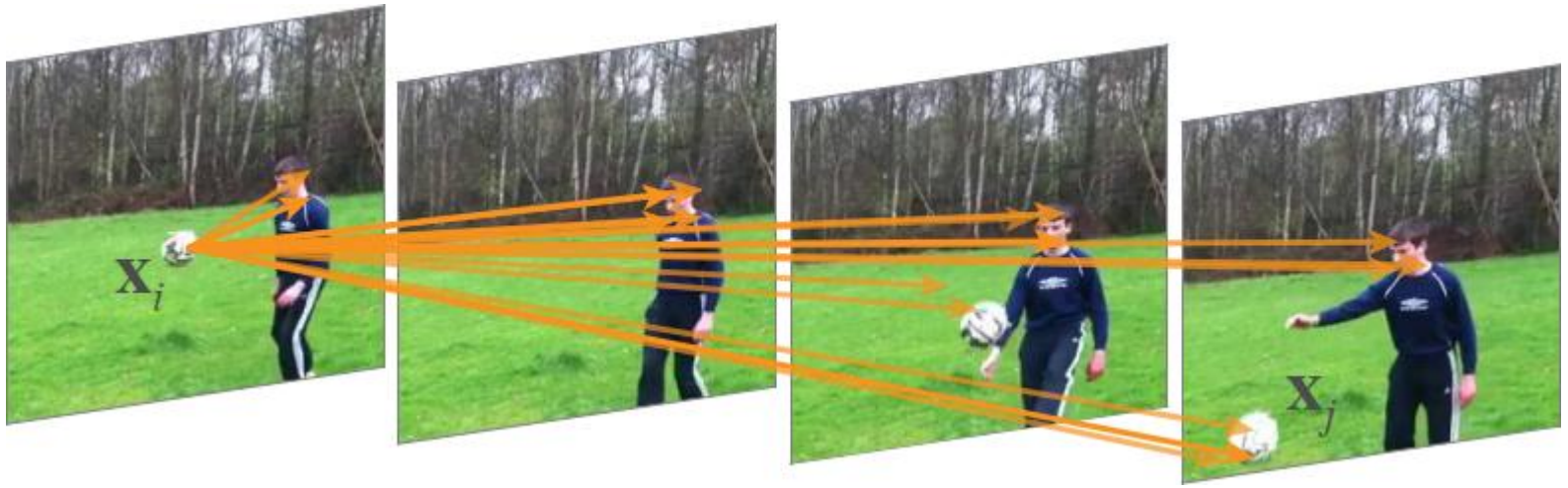
- SENet (Squeeze-and-Excitation Network)
  - "Squeeze-and-Excitation Networks" [Hu et al., 2018]
  - Introduces channel-wise attention mechanism in CNNs
    - Adaptively recalibrates channel-wise feature responses with minimal computational overhead
  - First major integration of attention mechanism into CNN architectures, establishing the foundation for attention-based feature refinement
  - Various versions: SE-ResNet-50/101/152, SE-ResNeXt-50/101, SENet-154



[Hu et al., 2018]]

Squeeze-and-Excitation Block in SENet

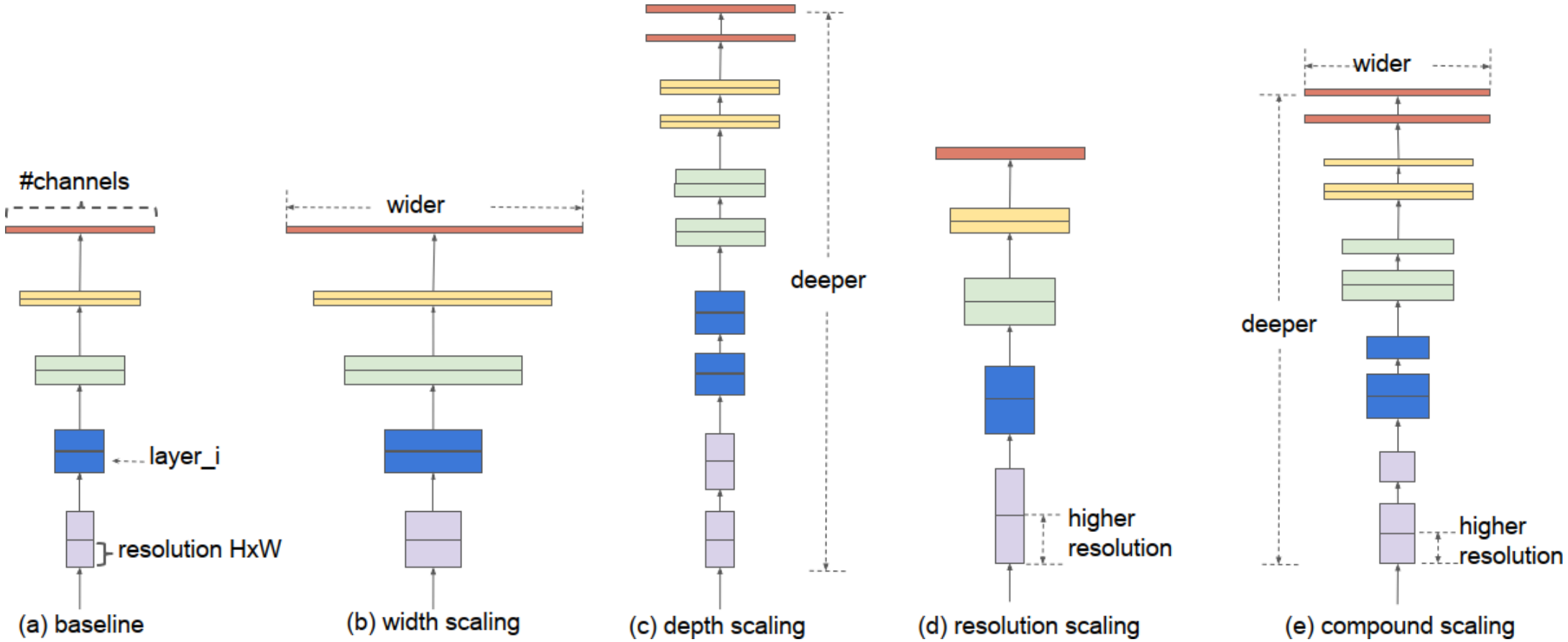
- Non-local Network
  - "Non-local Neural Networks" [Wang et al., 2018]
  - Introduces spatial self-attention mechanism in CNNs
    - Computes responses at each position as weighted sum of features at all spatial positions
  - Pioneering work that bridged CNN's local processing with attention's global modeling capability, directly inspiring Transformer architectures



[Wang et al., 2018]

Non-local Operation (Self-Attention Across Spatial-Temporal Dimensions) in Non-local Networks

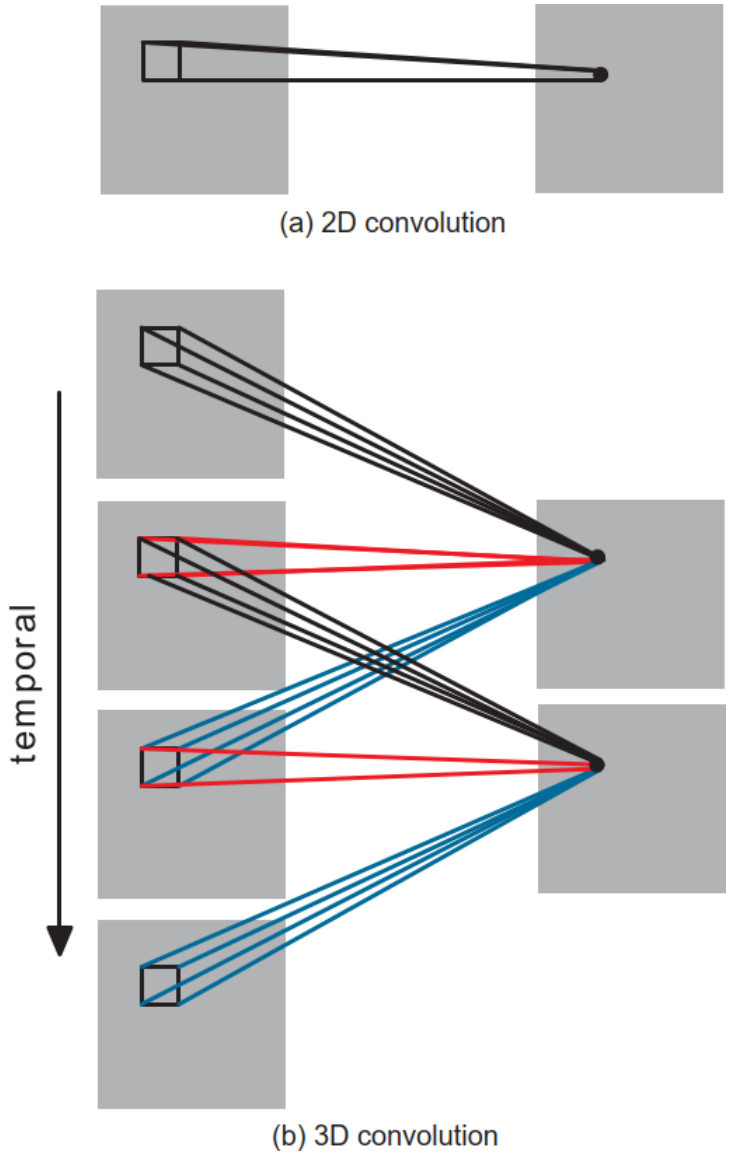
- EfficientNet (Efficient Network)
  - "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks" [Tan & Le, 2019]
  - Introduces compound scaling method for balancing network depth, width, and resolution
    - Achieves superior efficiency through systematic scaling rather than arbitrary increases
  - Various versions: B0 (baseline) through B7, with EfficientNet-L2 as the largest variant



[Tan & Le, 2019]

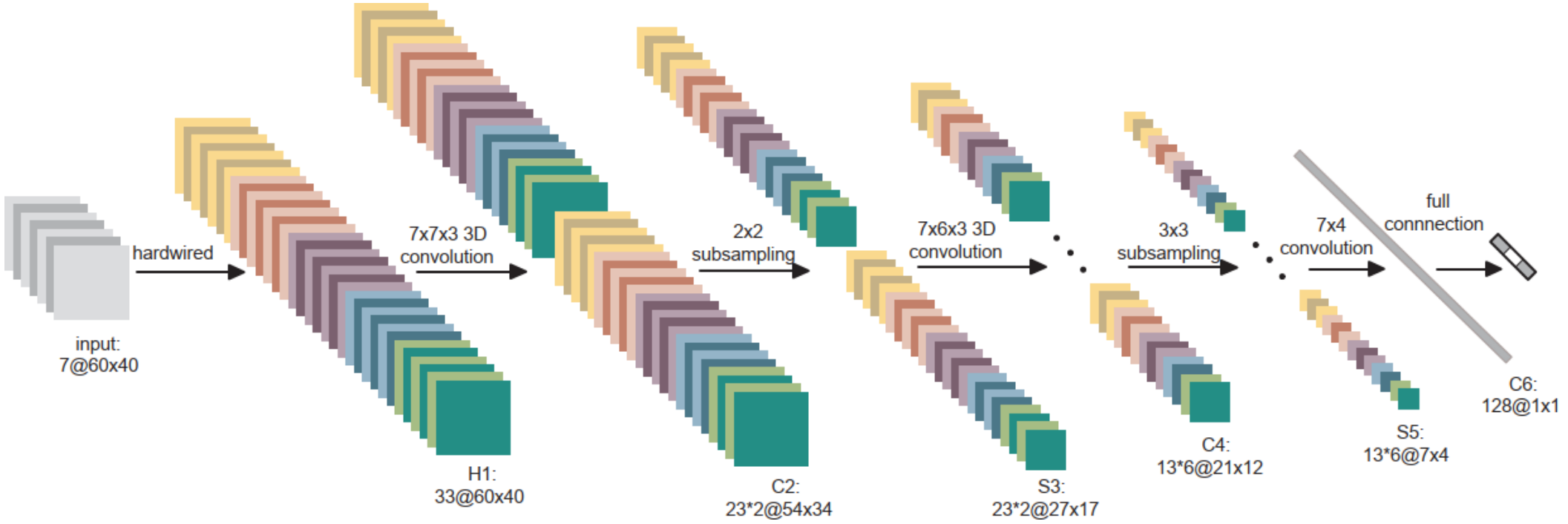
Model Scaling for EfficientNet

- Early 3D CNNs: extension of 2D CNNs into the 3D domain
  - Early works: "3D Convolutional Neural Networks for Human Action Recognition" [Ji et al., 2013]
  - Medical applications: "3D Deep Learning for Multi-modal Imaging-Guided Survival Time Prediction of Brain Tumor Patients" [Nie et al., 2016]
    - Extends CNN capabilities to volumetric data, essential for 3D medical imaging



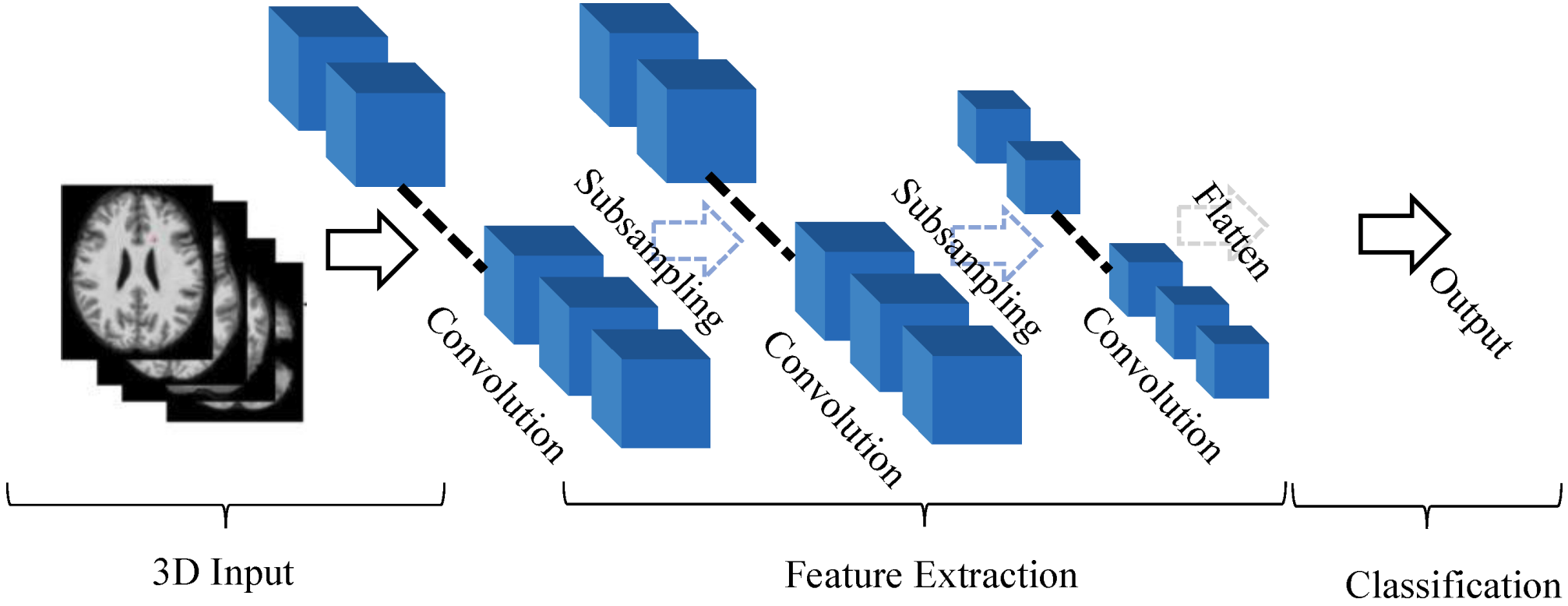
[Ji et al., 2013]

Comparison of 2D and 3D (Including the Temporal Dimension) Convolutions



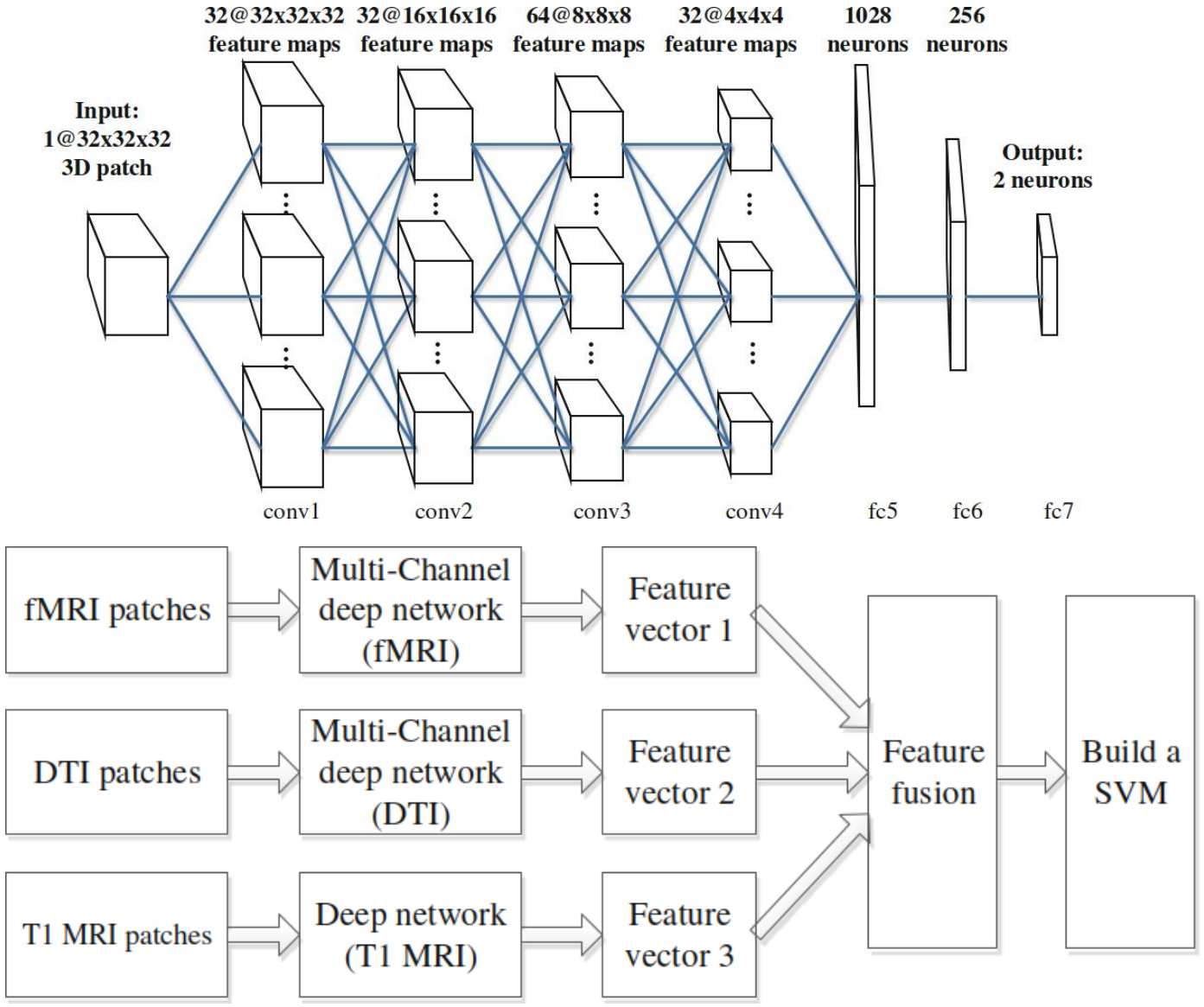
[Ji et al., 2013]

3D CNN Architecture for Human Action Recognition



[Singh et al., 2020]

Typical 3D CNN Architecture



[Nie et al., 2016]

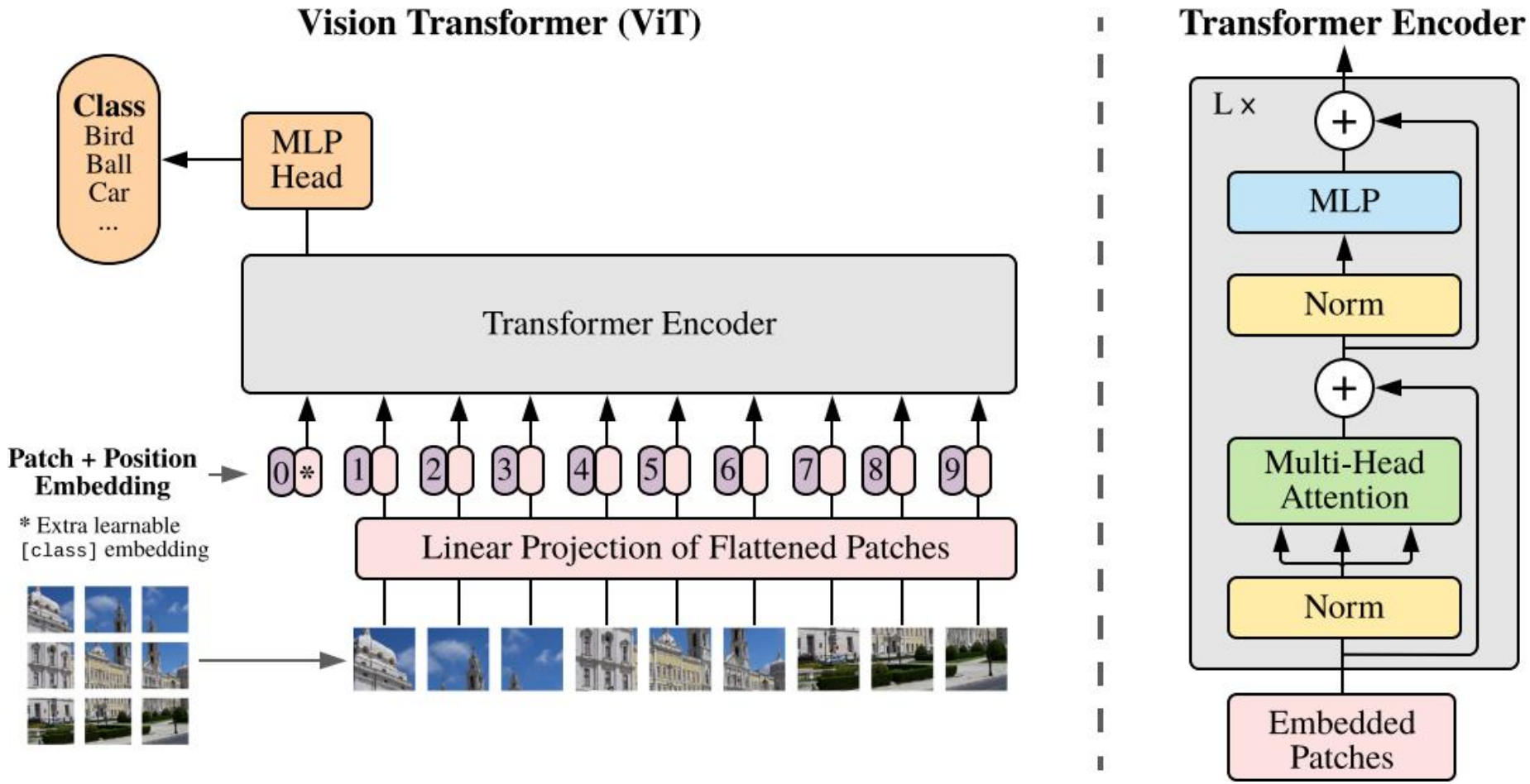
Multi-channel Feature Extraction from 3D Patches for Survival Prediction

Ref.	Task	Model	Data	Performance Measures
Yang et al. [39]	AD classification	3D VggNet, 3D Resnet	MRI scans from ADNI dataset (47 AD, 56 NC)	86.3% AUC using 3D VggNet and 85.4% AUC using 3D ResNet
Kruthika et al. [75]	–do–	3D capsule network, 3D CNN	MRI scans from ADNI dataset (345 AD, NC, 605, and 991MCI)	Acc. for AD/MCI/NC 89.1%
Feng et al. [76]	–do–	3D CNN + LSTM	PET + MRI scans from ADNI dataset (93 AD, 100 NC)	Acc. 65.5% (sMCI/NC), 86.4% (pMCI/NC), and 94.8 % (AD/NC)
Wegmayr et al. [77]	–do–	3D CNN	ADNI and AIBL data sets, 20000 T1 scans	Acc. 72% (MCI/AD), 86 % (AD/NC), and 67 % (MCI/NC)
Oh et al. [84]	–do–	3D CNN +transfer learning	MRI scans from the ADNI dataset (AD 198, NC 230, pMCI 166, and sMCI 101) at baseline.	74% (pMCI/sMCI), 86% (AD/NC), 77% (pMCI/NC)
Parmar et al. [10]	–do–	3D CNN	fMRI scans from ADNI dataset (30 AD, 30 NC)	Classification acc. 94.85 % (AD/NC)
Nie et al. [79]	Brain tumor	3D CNN with learning supervised features	Private, 69 patient (T1 MRI, fMRI, and DTI)	Classification acc. 89.85 %
Amidi et al. [85]	Protein shape	2-layer 3D CNN	63,558 enzymes from PDB datasets	Classification acc. 78%
Zhou et al. [80]	Breast cancer	Weakly supervised 3D CNN	Private, 1537 female patient	Classification acc. 78% 83.7%

[Singh et al., 2020]

Applications of 3D CNNs to Classification with Medical Imaging (2016-2020)

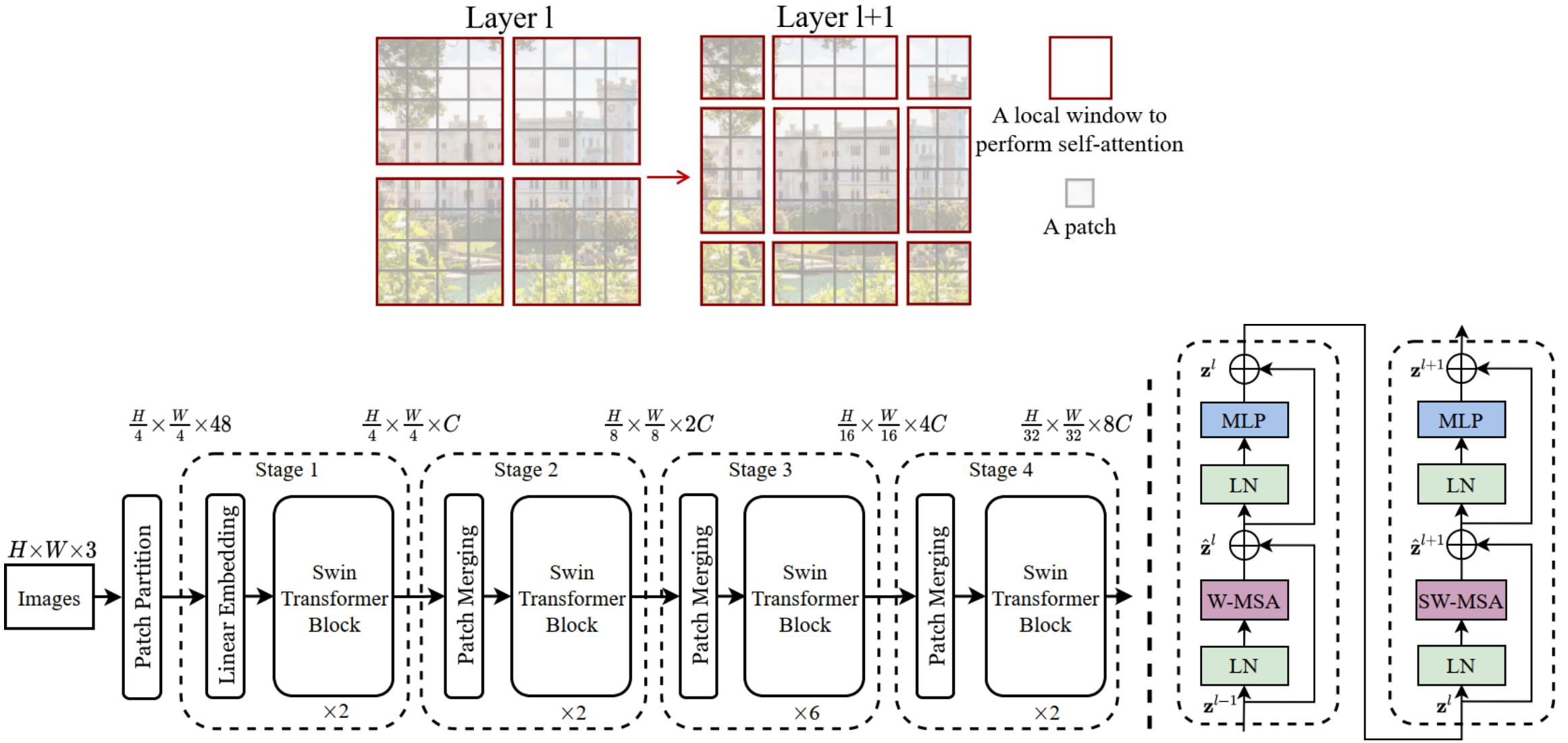
- ViT (Vision Transformer)
  - "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" [Dosovitskiy et al., 2020]
  - First pure transformer architecture for image classification, treating image patches as tokens
    - Paradigm shift from convolution-based to attention-based image processing
  - Proved transformers can work effectively in computer vision without any convolutional components
    - Demonstrates competitive performance with CNNs when trained on large-scale datasets
  - Various versions: ViT-Base (12 layers), ViT-Large (24 layers), ViT-Huge (32 layers)



[Singh et al., 2020]

ViT Architecture Including Transformer Encoder Blocks

- Swin (Shifted Window) Transformer
  - "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows" [Liu et al., 2021]
  - Hierarchical transformer architecture with shifted window attention for computational efficiency
    - Combines transformer's global modeling with CNN-like hierarchical features
  - Bridged the gap between ViT's flat structure and CNN's multi-scale processing
  - Various versions: Swin-T (tiny, 12 layers), Swin-S (small, 24 layers), Swin-B (base, 24 layers), and Swin-L (large, 24 layers)



[Singh et al., 2020]

Swin Transformer Architecture Including Swin Transformer Blocks

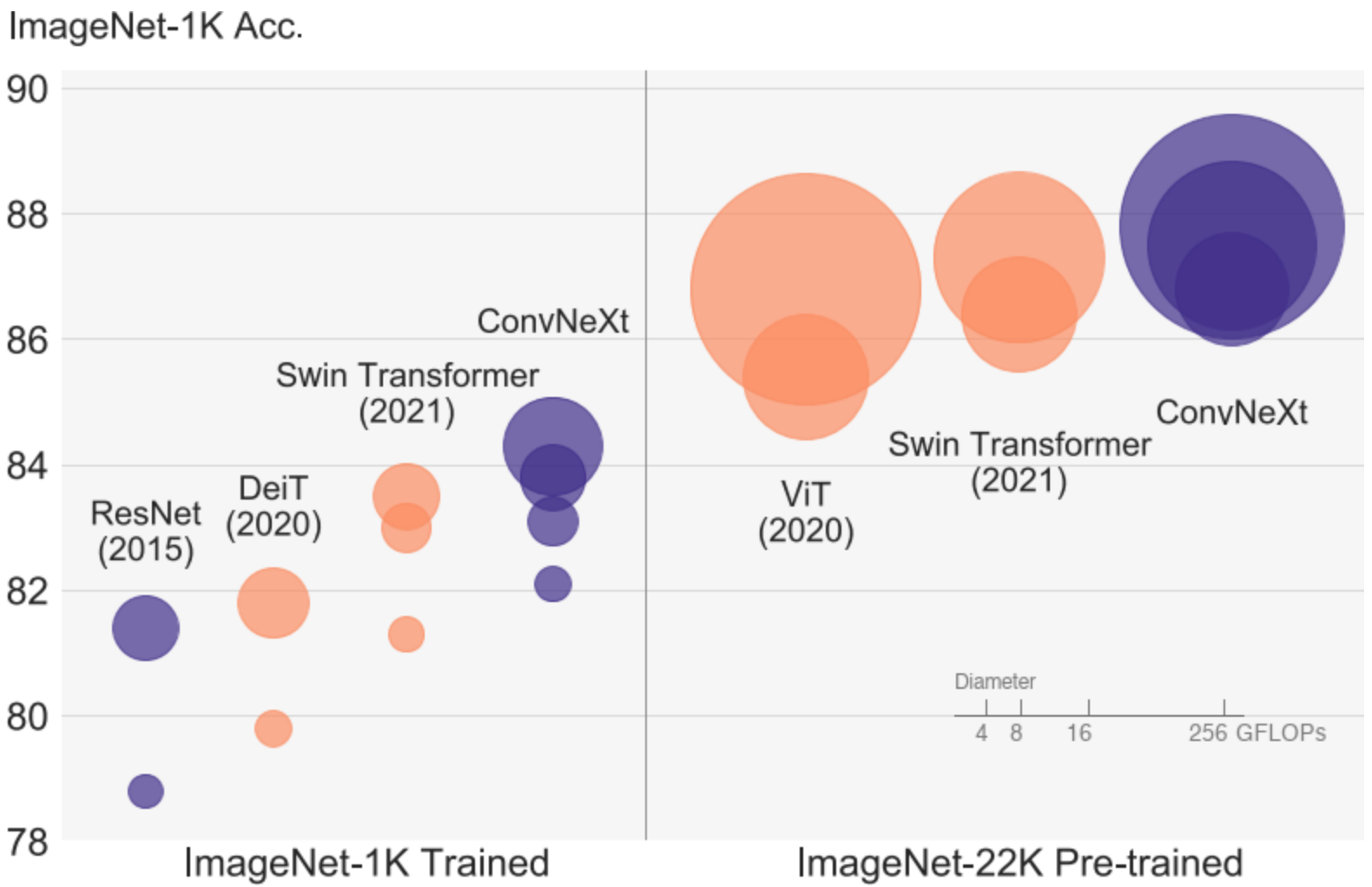
- CoAtNet (Convolution + Attention Network)
  - "CoAtNet: Marrying Convolution and Attention for All Data Sizes" [Dai et al., 2021]
  - First systematic hybrid design combining CNN's inductive bias with transformer's capacity
    - Vertical fusion of convolution and self-attention through principled design
  - Demonstrated that CNN-Transformer hybrids outperform pure architectures across all data regimes

Properties	Convolution	Self-Attention
Translation Equivariance	✓	
Input-adaptive Weighting		✓
Global Receptive Field		✓

[Dai et al., 2021]

CoAtNet Design Principle: Complementary Properties of Convolution and Self-Attention

- ConvNeXt (Next-generation Convolutional Network)
  - "A ConvNet for the 2020s" [Liu et al., 2022]
  - Modernized CNN architecture incorporating transformer design principles
    - Systematic modernization of ResNet using transformer-inspired components
  - Proved that properly designed CNNs can compete with transformers, challenging the transformer supremacy narrative



[Dai et al., 2021]

ConvNeXt Achievement: Modernized CNNs Competing with Transformers

- DINOv2 (Distillation with No Labels version 2)
  - "DINOv2: Learning Robust Visual Features without Supervision" [Oquab et al., 2023]
  - Self-supervised foundation model trained on 142M curated images without labels
    - Shifts focus from architecture design to scale and data quality
    - Achieves strong performance across classification, regression, and dense prediction tasks through learned representations
  - Marks transition from architecture wars to foundation model era, proving that scale and data matter more than architectural choices

ImageNet and Deep Learning Evolution

- ILSVRC (ImageNet Large Scale Visual Recognition Challenge)
  [https://www.image-net.org/challenges/LSVRC/]
  - Annual competition (2010-2017) for computer vision algorithms on large-scale image dataset
  - Dataset scale: ~1.4 million images, 1,000 object categories
  - Primary tasks: Image classification, object detection, object localization
  - Though officially ended in 2017, ImageNet remains the standard benchmark for evaluating new vision models

Top-5 error rate was the official metric for ILSVRC competition

Year	Model	Top-5 error rate	Key innovation
2010-11	Traditional computer vision methods	~25%	Hand-crafted features
2012	AlexNet	16.4%	First CNN winner
2013	ZFNet	11.7%	Refined CNN architecture
2014	GoogLeNet	6.7%	Inception modules
2014	VGGNet	7.3%	Deeper networks
2015	ResNet	3.6%	Residual connections
2016	ResNeXt	~3%	Cardinality
2017	SENet	2.3%	Squeeze-and-excitation blocks

ILSVRC Winning Models (2010-2017)

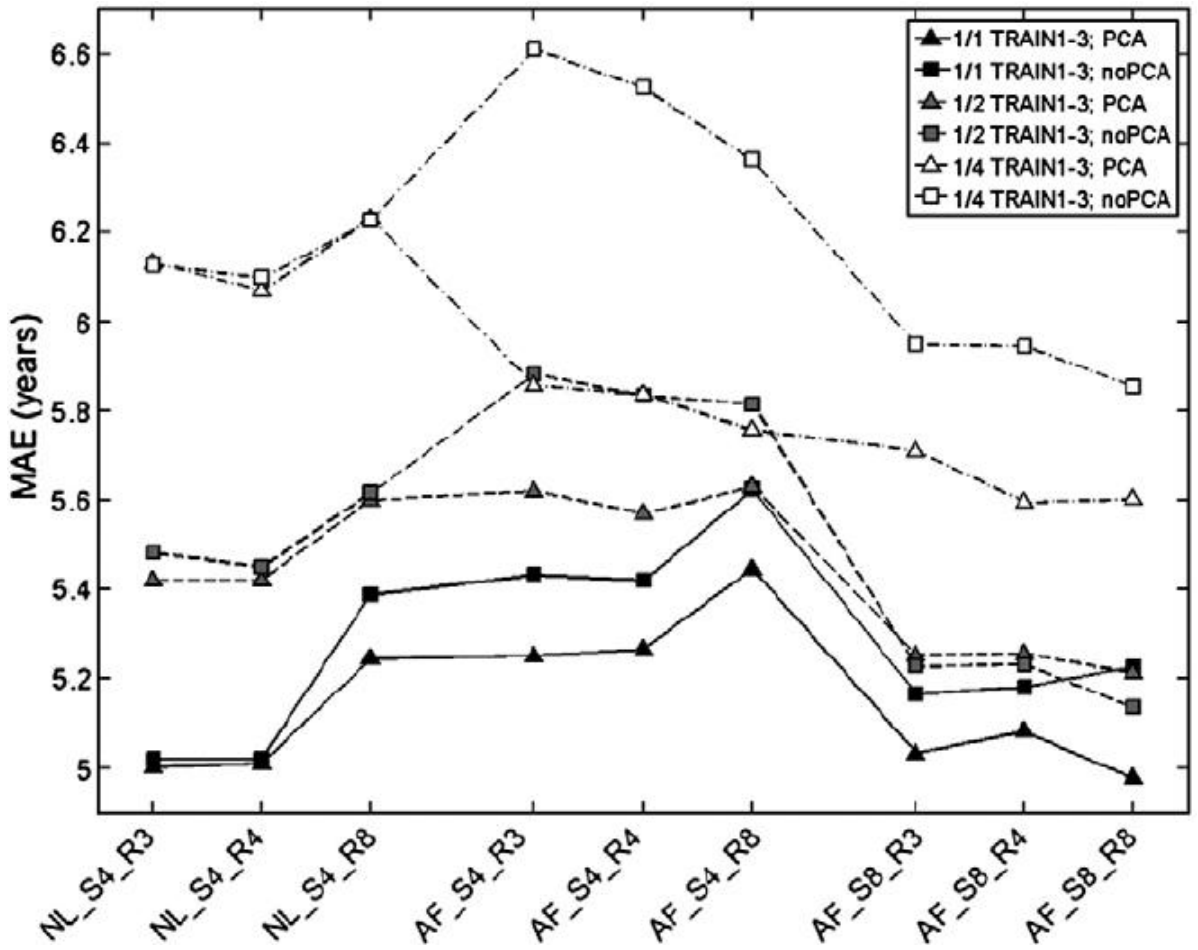
Post-competition research shifted to reporting Top-1 accuracy as models approached human-level performance on Top-5 metrics.

Year	Model	Top-1 accuracy	Key innovation
2019	EfficientNet	84.4%	Compound scaling
2020	ViT	88.55%	First pure transformer architecture for vision
2021	CoAtNet	90.45%	Convolution-attention hybrid
2022-23	ConvNeXt	91.1%	Modernized CNN design with transformer techniques
2023-24	DINOv2	91.5%+	Self-supervised representation learning

Post-ILSVRC Advancements

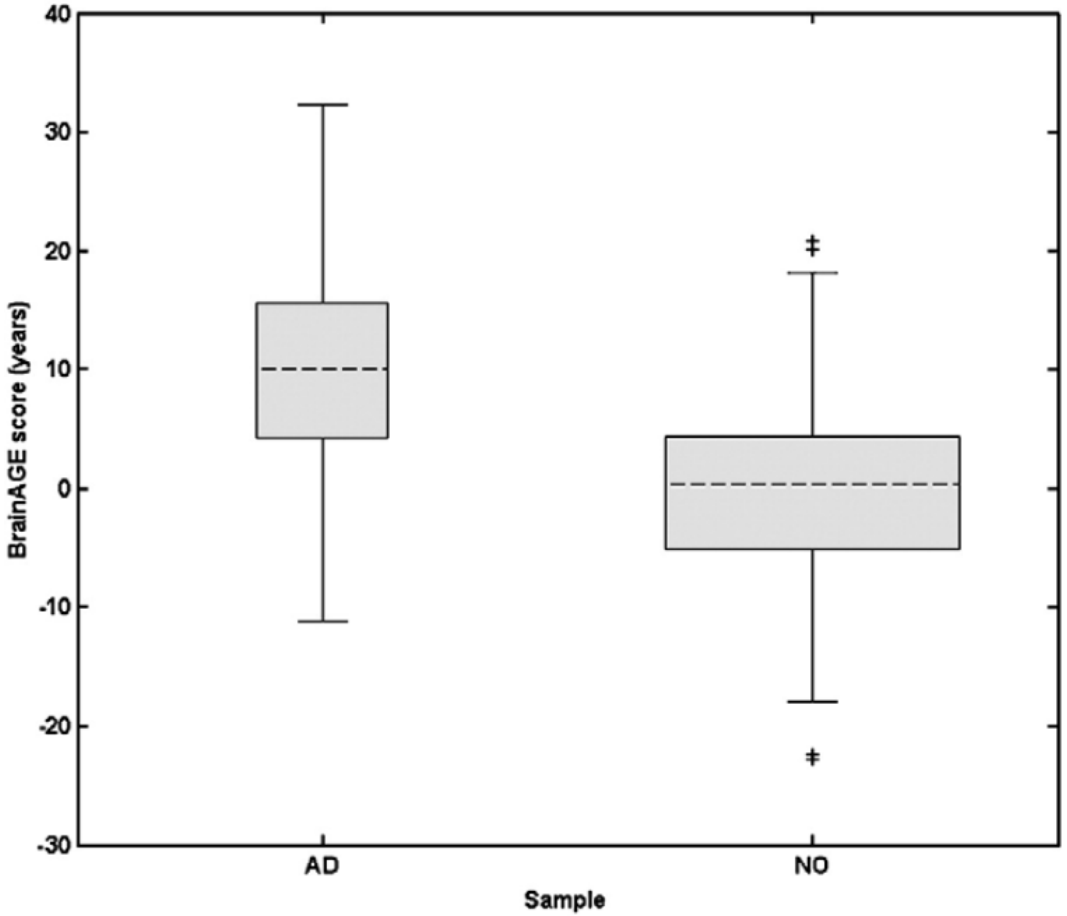
Age Prediction Performance

- Relevance vector regression [Franke et al., 2010]
  - GM map values → principal component analysis
  - Datasets:
    - Training: n = 410 (20 ~ 86 years)
    - Test: n = 137 (19 ~ 83 years)
    - External test: n = 108 (20 ~ 59 years)
  - MAE:
    - Test: 4.61 years
    - External test: 5.44 years



[Franke et al., 2010]

Influences of Data Processing Options on Age Prediction Performance

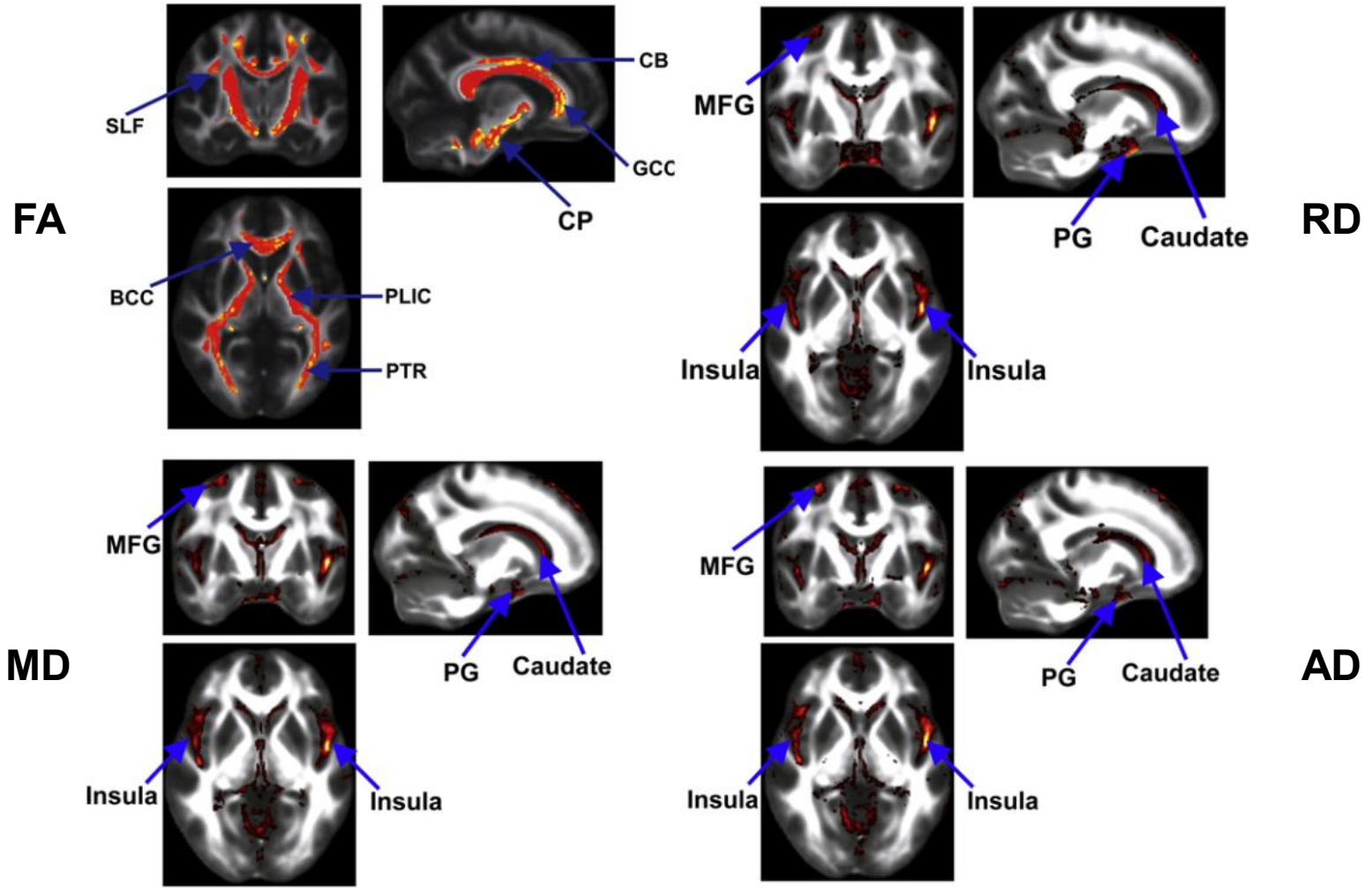


AD, Alzheimer's disease

[Franke et al., 2010]

Comparison of BAG

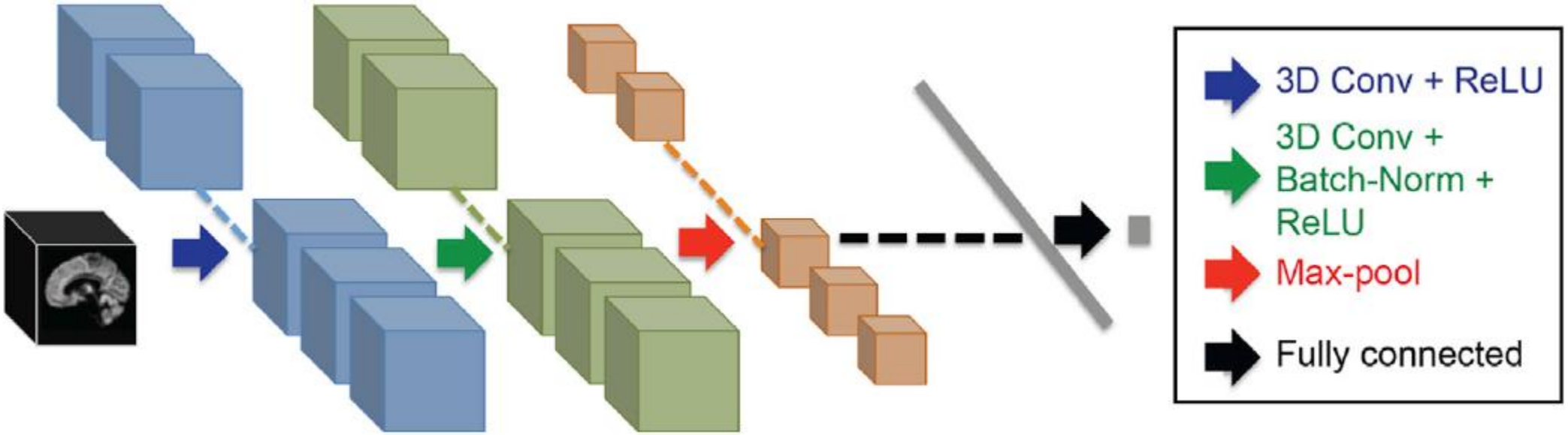
- Relevance vector regression [Mwangi et al., 2013]
  - Voxel-wise FA | MD | AD | RD map values → recursive feature elimination
  - Datasets: n = 188 (4 ~ 85 years)
    - Leave-one-out cross validation
  - MAE:
    - 6.9 years (RD)
    - 7.1 years (MD)
    - 7.2 years (AD)
    - 8.2 years (FA)



[Mwangi et al., 2013]

Consensus Sensitivity Maps Showing Relevant Anatomical Regions

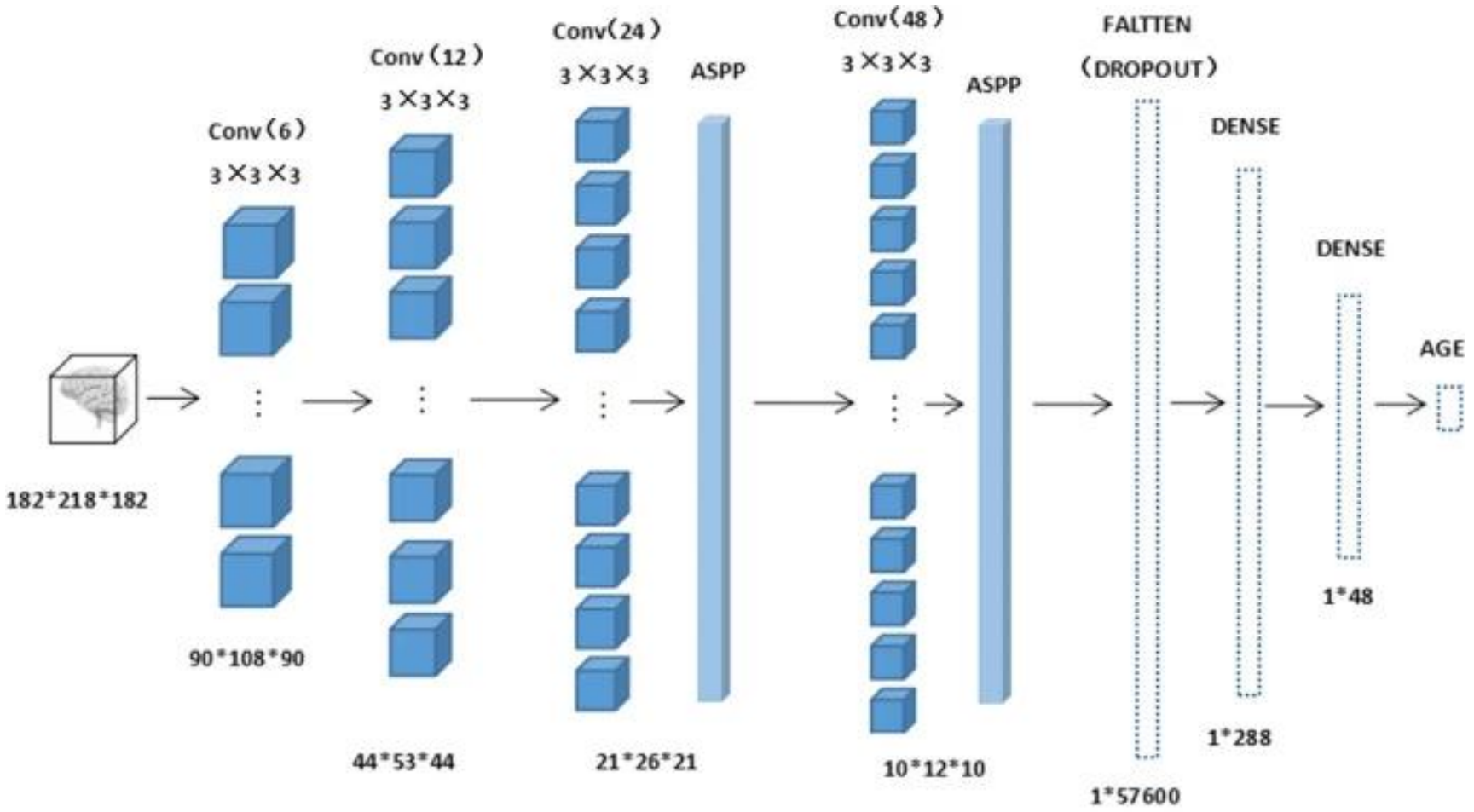
- 3D CNN [Cole et al., 2017]
  - Brain | GM | WM maps
  - Datasets (18 ~ 90 years):
    - Training: n = 1,601
    - Validation: n = 200
    - Test: n = 200
  - MAE:
    - 4.16 years (GM)
    - 4.34 years (concatenated GM and WM)
    - 4.65 years (Brain)
    - 5.14 years (WM)



[Cole et al., 2017]

Proposed 3D CNN Architecture

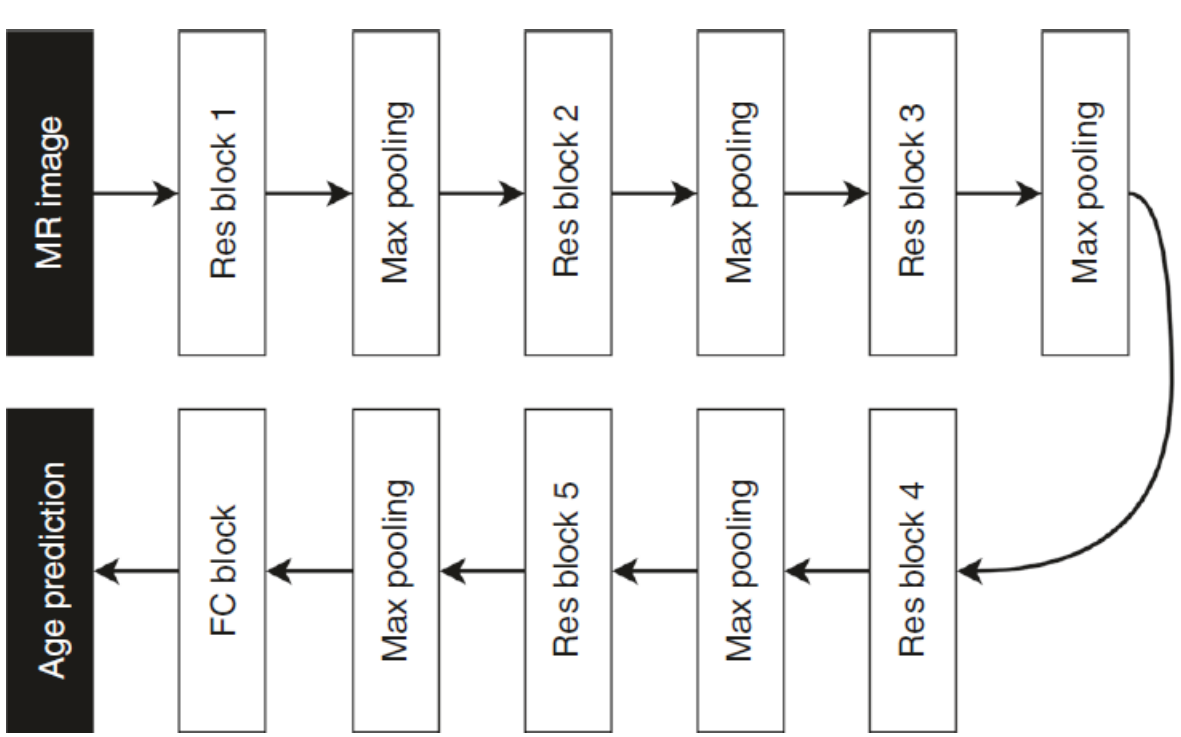
- 3D CNN [Wang et al., 2023]
  - WM-masked FA map (ICBM-DTI-81 atlas)
  - Datasets: n = 2,406 (17 ~ 60 years)
    - 10-fold cross validation
      - Training: 80%
      - Validation: 10%
      - Test: 10%
  - MAE:
    - 2.785 years

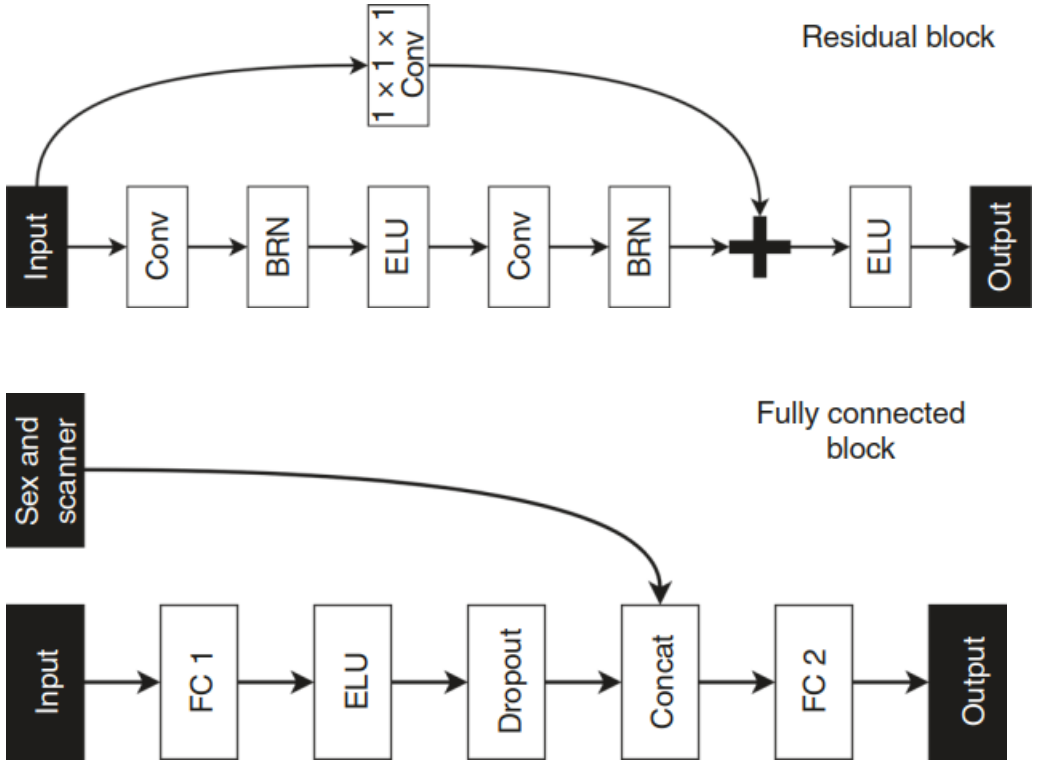


[Wang et al., 2023]

Proposed 3D CNN Architecture

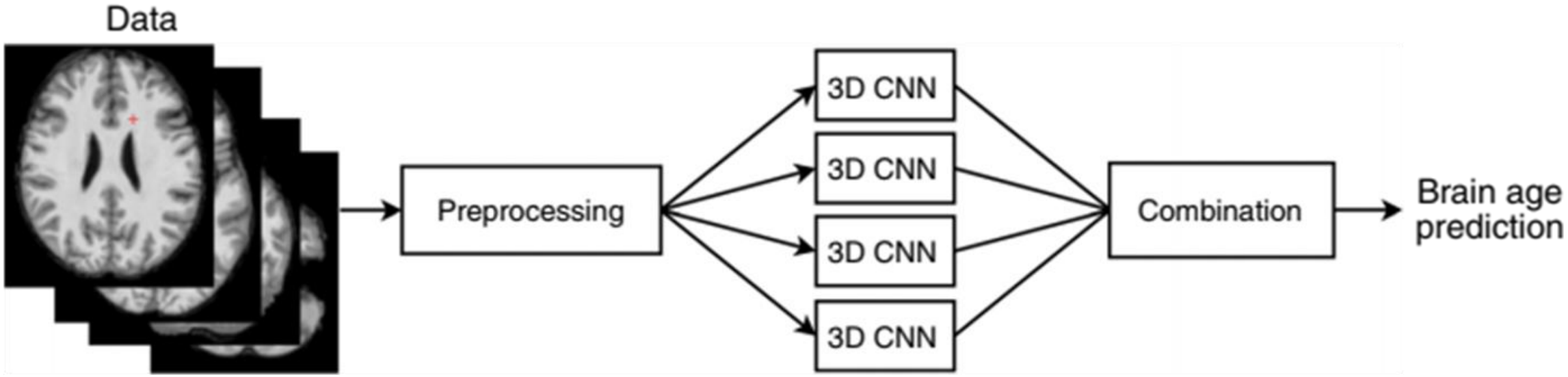
- ResNet [Jonsson et al., 2019]
  - Brain | Jacobian | GM | WM maps + sex and MRI scanner type
  - Datasets (18 ~ 75 years):
    - Training: *n* = 809 (1,171 images)
    - Validation: *n* = 202 (298 images)
    - Test: *n* = 253 (346 images)
  - MAE:
    - 3.388 years (T1, Jacobian, GM, WM)
    - 4.006 years (Brain)
    - 4.189 years (WM)
    - 4.641 years (GM)
    - 4.804 years (Jacobian)





[Jonsson et al., 2019]

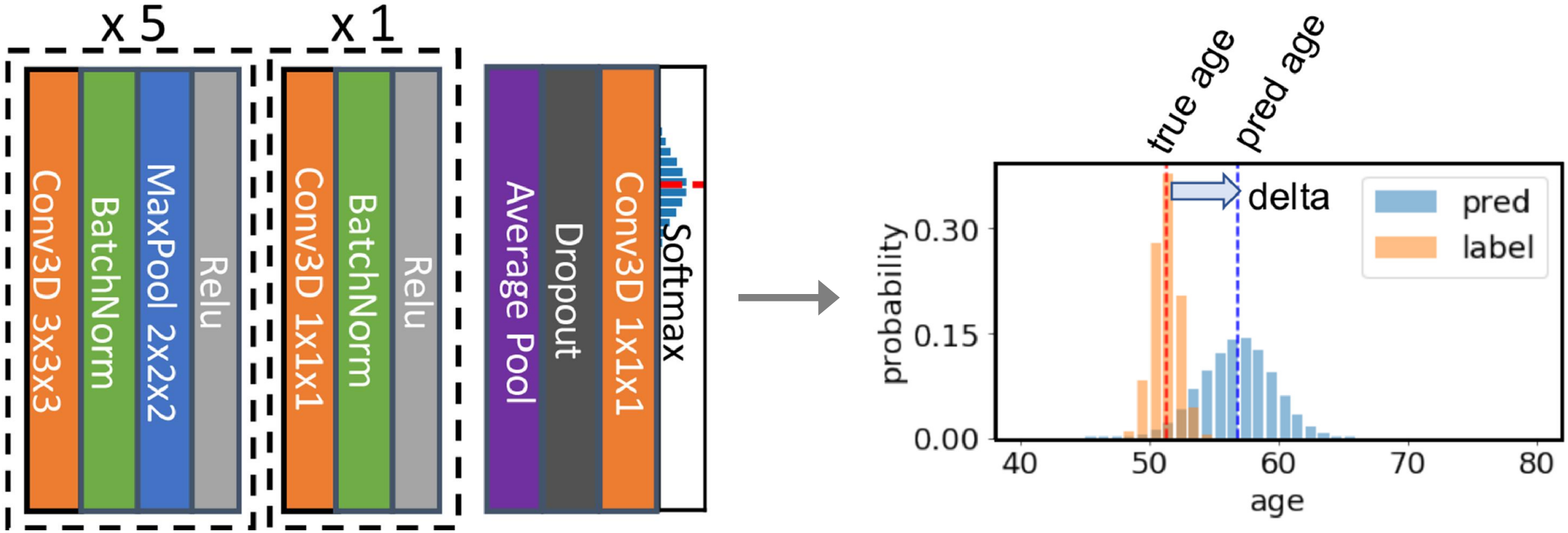
Proposed ResNet Architecture



[Jonsson et al., 2019]

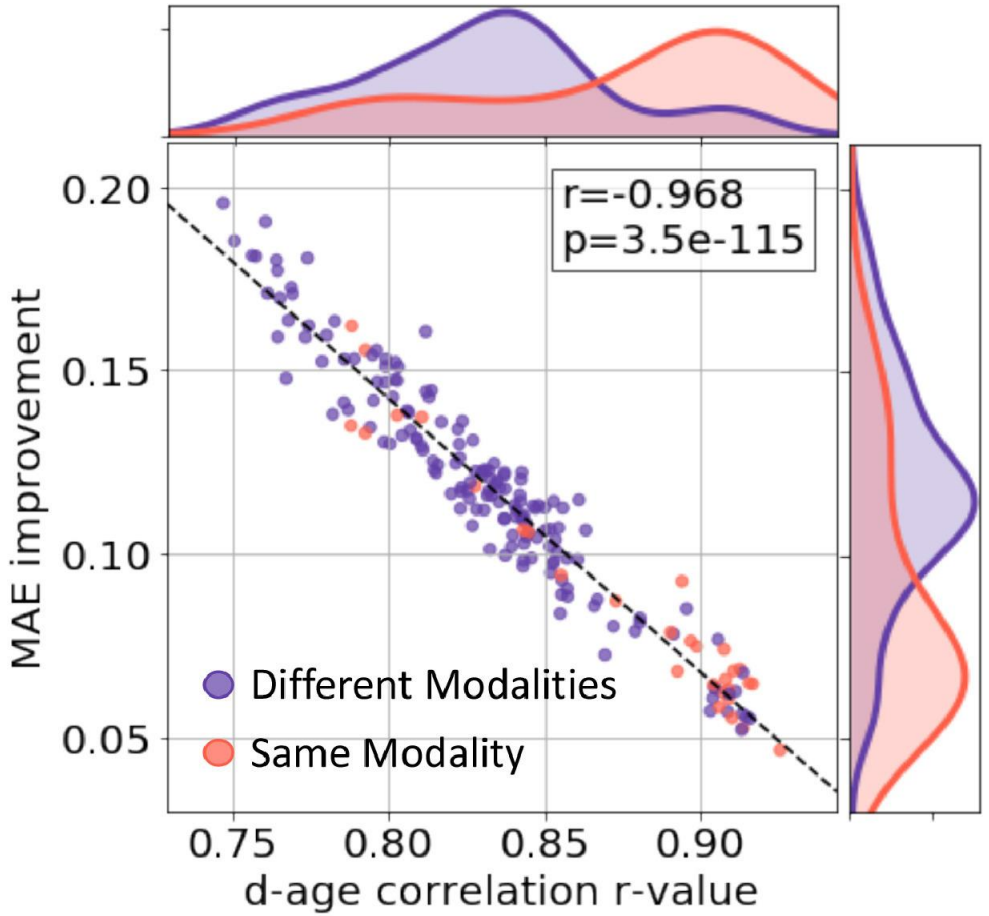
Combination of Predictions from Multiple CNNs

- SFCN (Simple Fully Convolutional Network) [Peng et al., 2021]
  - Brain_Lin (linear normalization) | Brain_Nonlin (nonlinear normalization) | GM | WM maps
  - Datasets (44 ~ 80 years):
    - Training: n = 12,949
    - Validation: n = 518
    - Test: n = 1,036
  - Data augmentation
    - Randomly shifted by 0, 1, or 2 voxels along every axis
    - Mirrored with a probability of 50% about the sagittal plane
  - MAE:
    - 2.14±0.05 years (Brain_Lin)



[Peng et al., 2021]

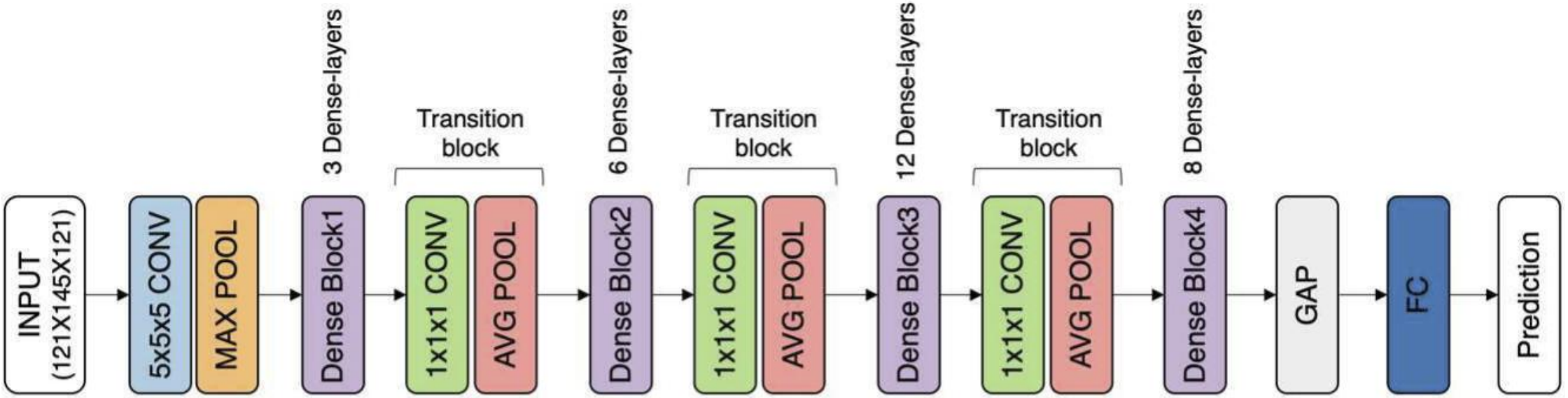
SFCN Architecture Using Soft-classification Loss



[Peng et al., 2021]

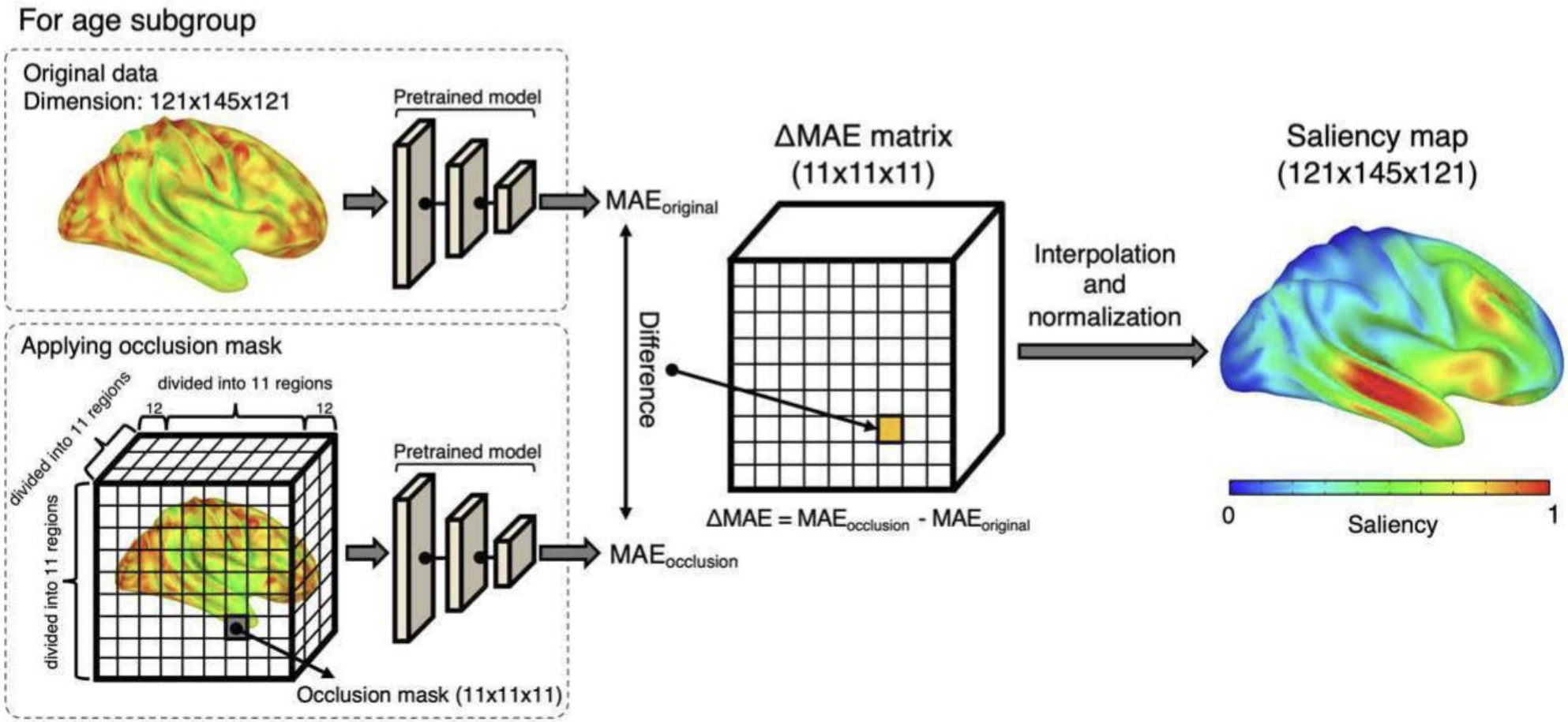
Ensemble Performance Improvement in Relation to BAG Correlation of Any Two Models

- DenseNet [Lee et al., 2022]
  - Brain map
  - Datasets: n = 1,805 (20 ~ 98 years)
    - 5-fold cross validation
      - Training: 60%
      - Validation: 20%
      - Test: 20%
  - MAE:
    - 4.2055±0.2241 years
  - Model explanations through occlusion sensitivity analysis with occlusion masks of 11^3 mm^3



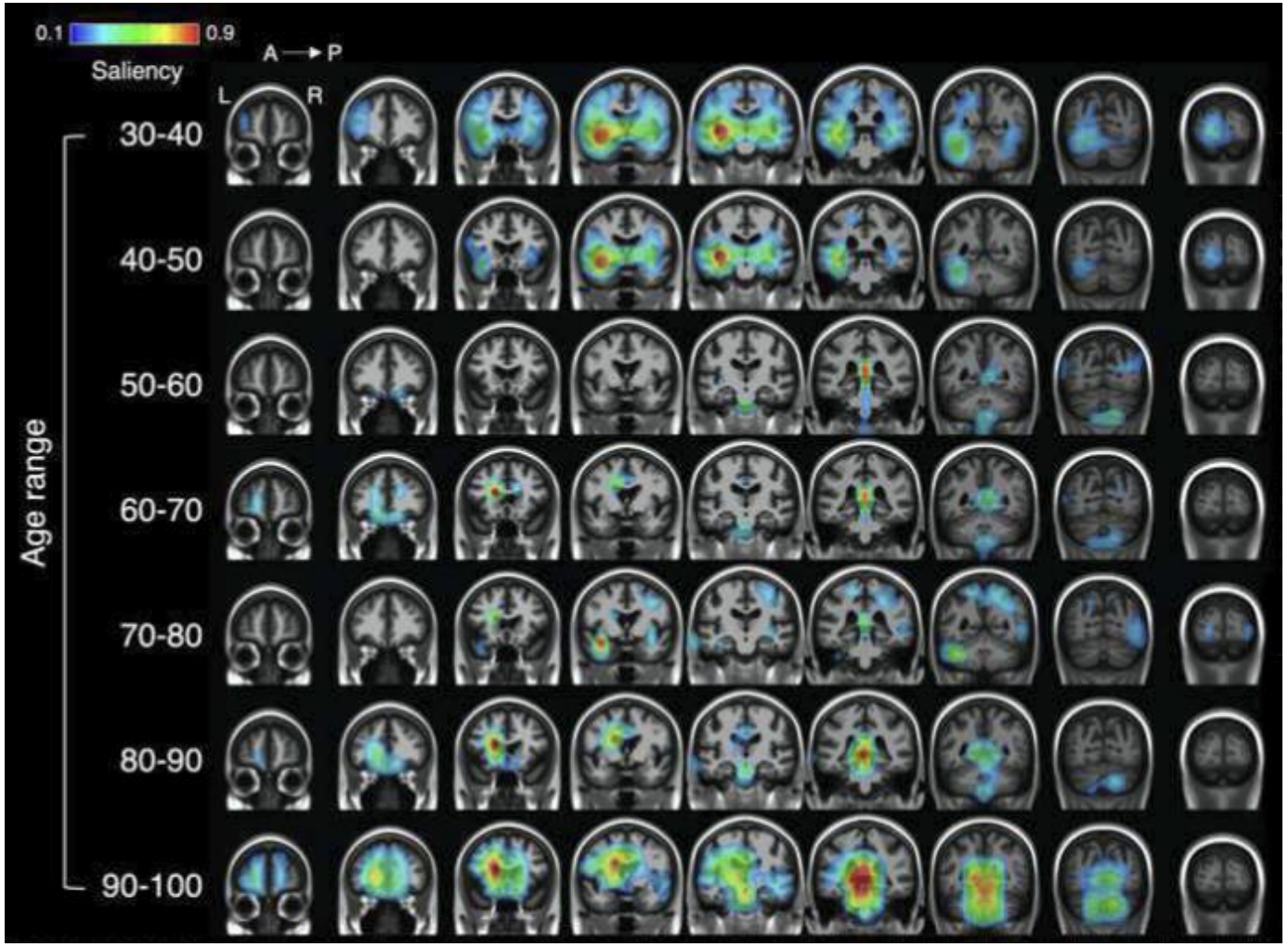
[Lee et al., 2022]

Proposed DenseNet Architecture



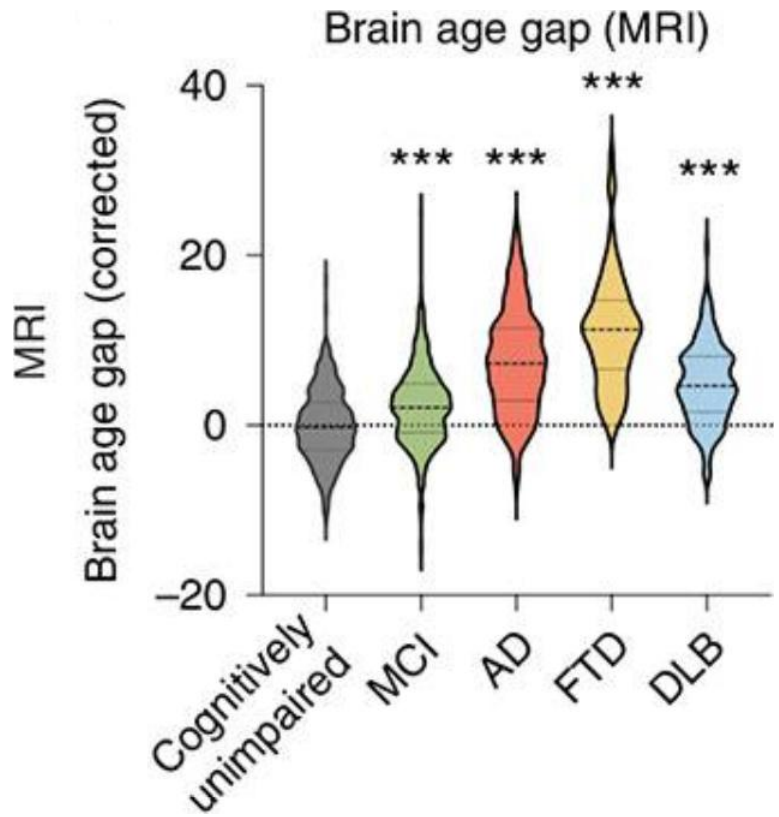
[Lee et al., 2022]

Occlusion Sensitivity Analysis



[Lee et al., 2022]

Saliency Maps Across Age Range Groups

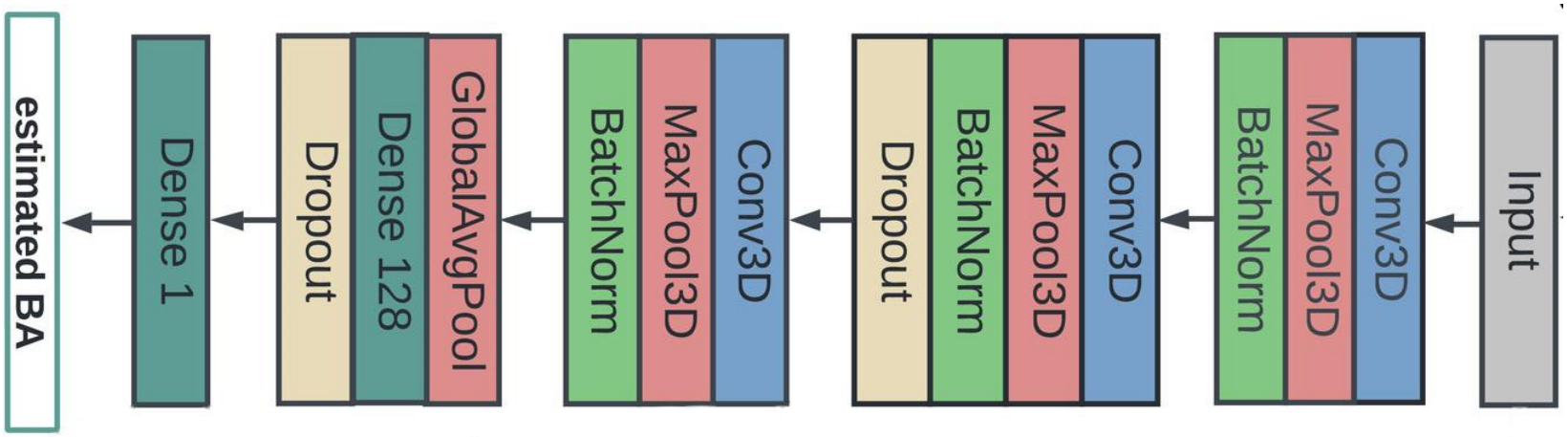


MCI, mild cognitive impairment
AD, Alzheimer's disease
FTD, frontotemporal dementia
DLB, dementia with Lewy bodies

[Lee et al., 2022]

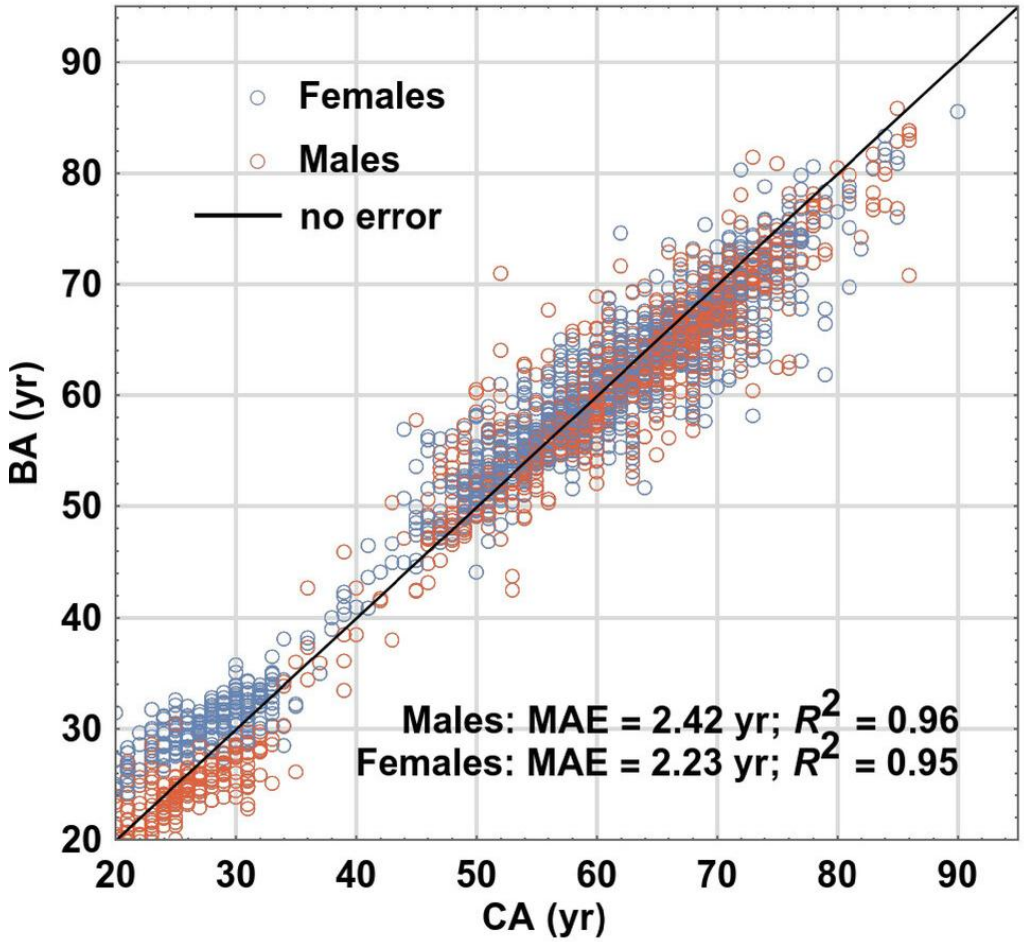
Comparison of BAG

- Sex-specific 3D CNN [Yin et al., 2023]
  - Brain map
  - Datasets:
    - Training: n = 4,681 (22 ~ 95 years)
    - Test: n = 1,170 (22 ~ 95 years)
    - External test: n = 650 (18 ~ 88 years)
  - MAE:
    - Test: 2.23 years (females) / 2.41 years (males)
    - External test: 4.71 years (females) / 3.01 years (males)
  - Model explanations through occlusion sensitivity analysis with occlusion masks of 1 mm^3



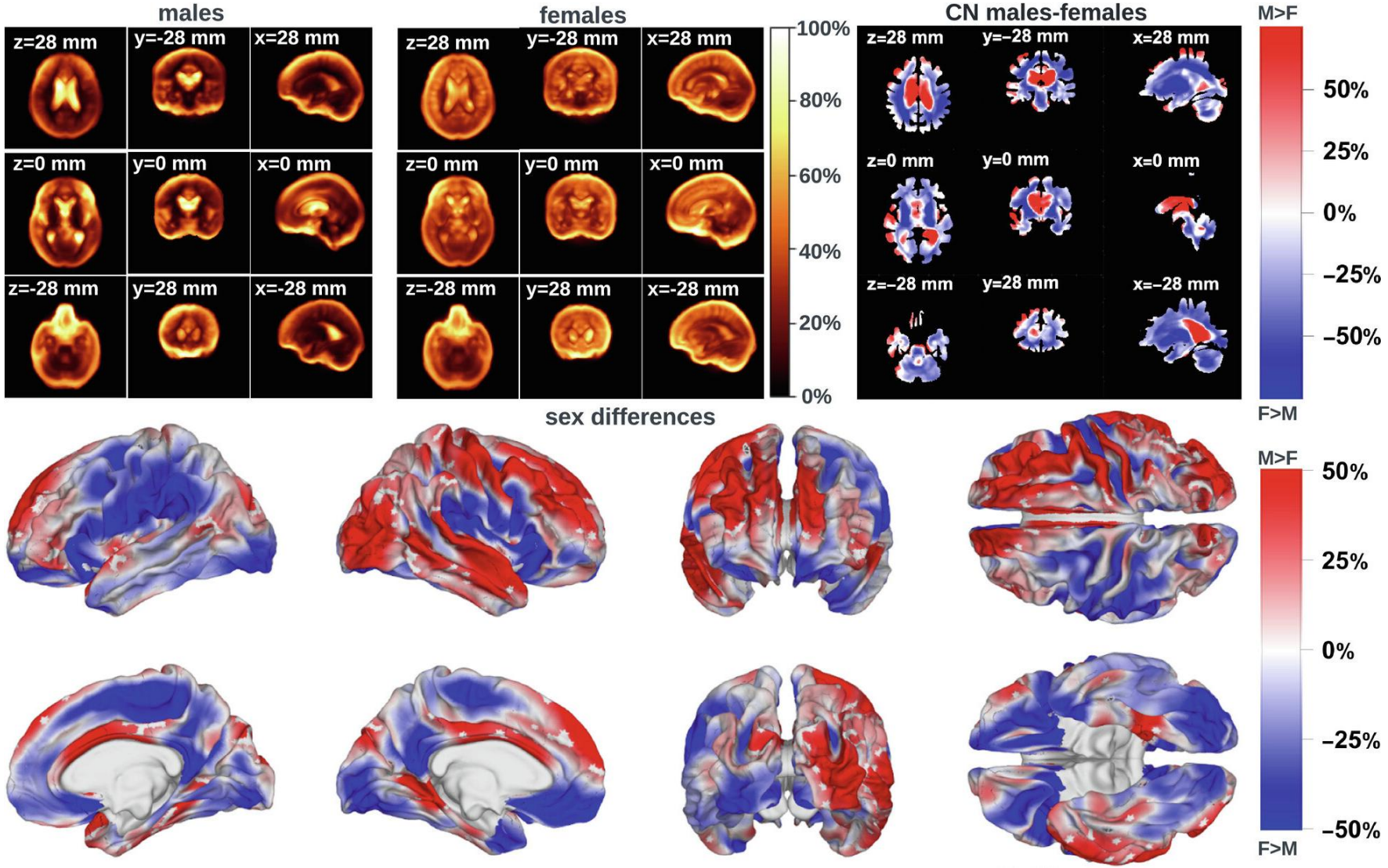
[Yin et al., 2023]

Proposed 3D CNN Architecture



[Yin et al., 2023]

Sex-specific Age Prediction



[Yin et al., 2023]

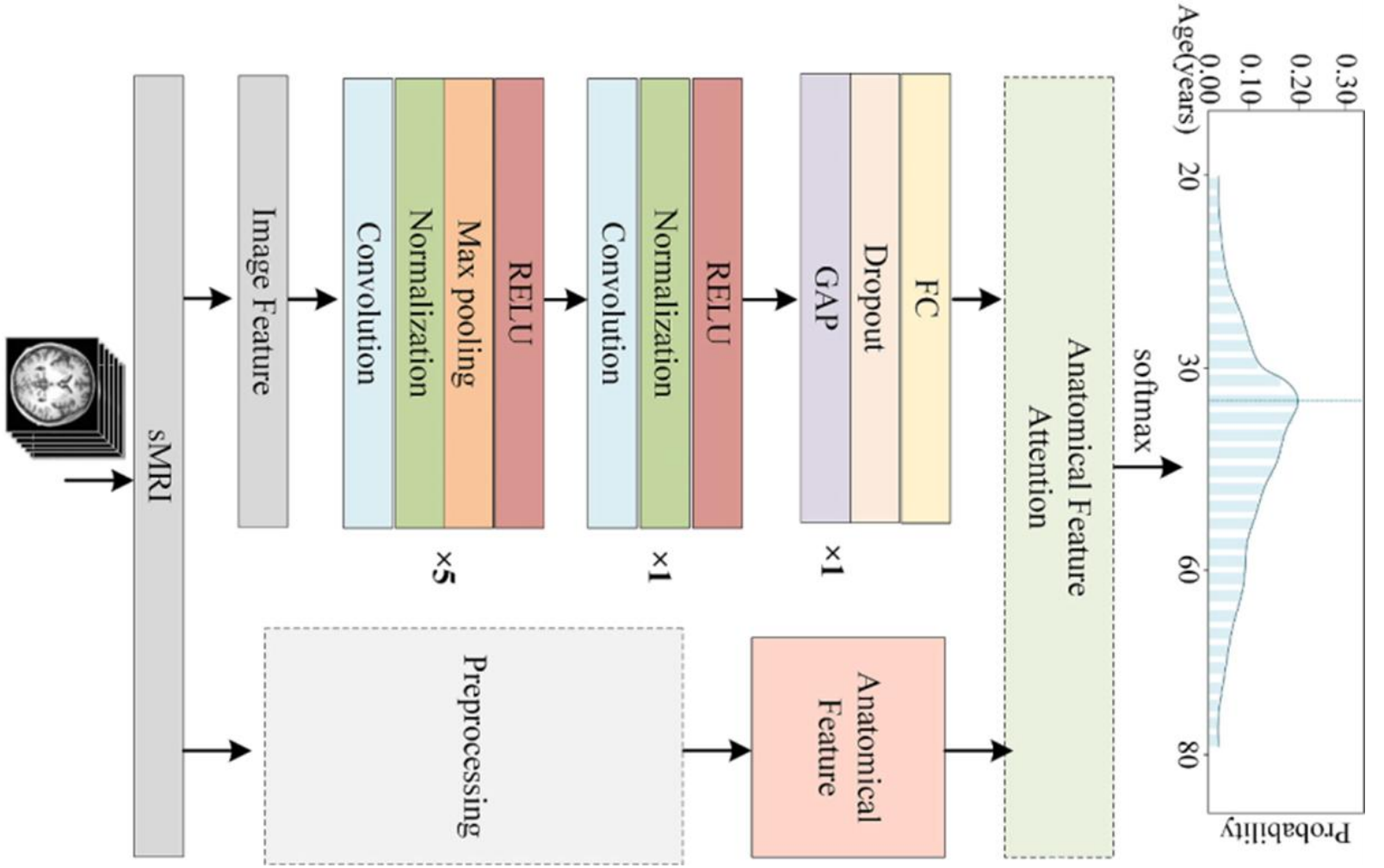
Comparison of Saliency Maps between Sexes

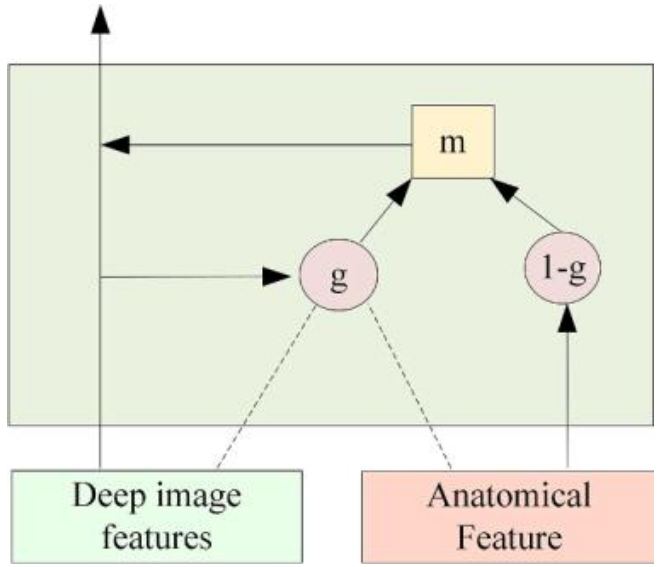
- AFAC (Anatomical Feature Attention-enhanced 3D CNN)

[Zhang et al., 2024]

  - Brain map + brain anatomical features
    - Curvature index, lateral ventricular volume, local gyrification index, cortical thickness, folding index, and surface area
  - Anatomical feature attention module
  - Datasets: n = 2,382 (20 ~ 80 years)
    - Training: 70%
    - Validation: 15%
    - Test: 15%
  - Data augmentation
    - Randomly shifted by 2 voxels along every axis
    - Mirrored with a probability of 50% about the sagittal plane

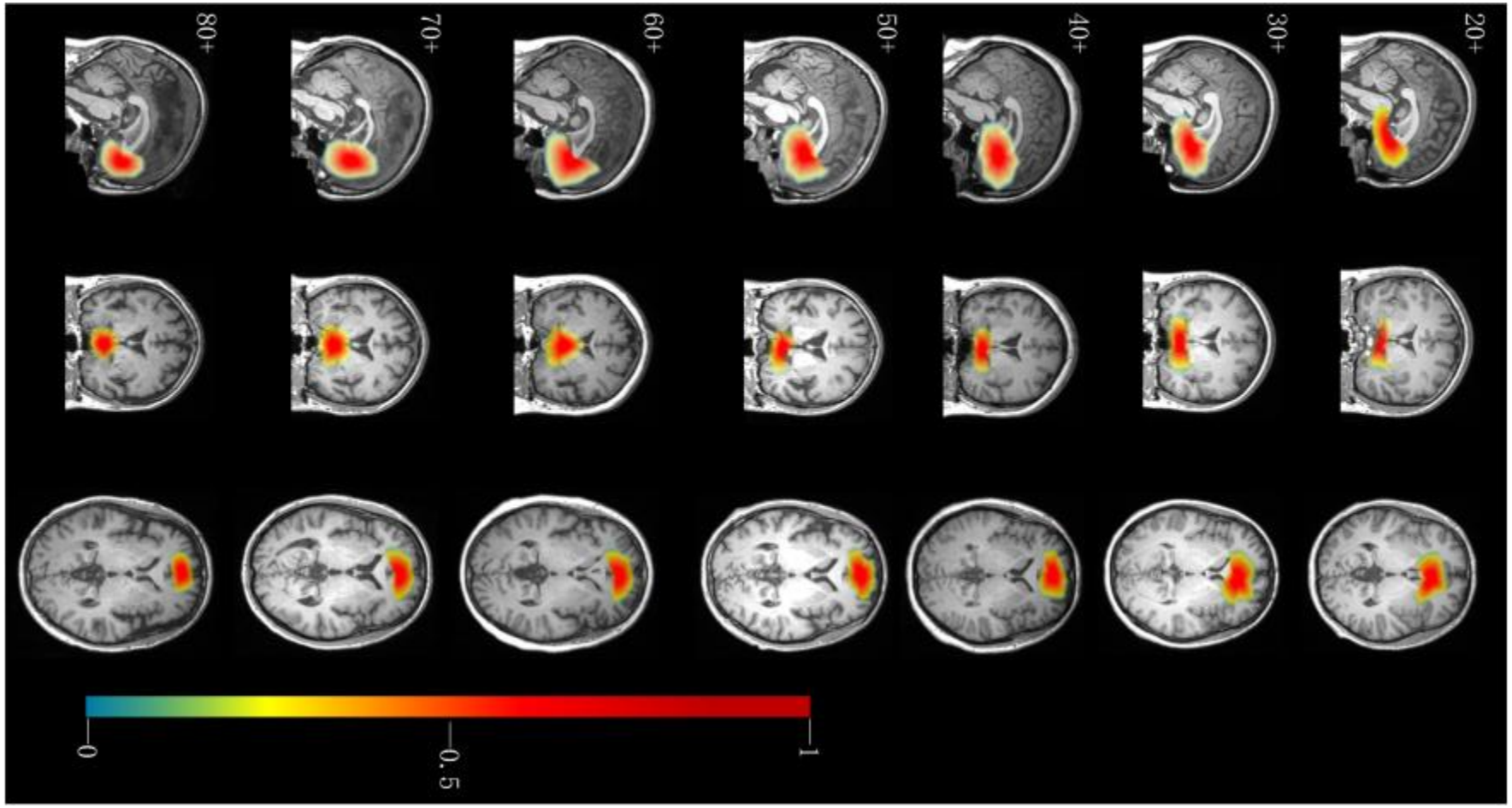
- MAE:
  - 2.20 years
- Model explanations through gradient-weighted class activation mapping (GradCAM)





[Zhang et al., 2024]

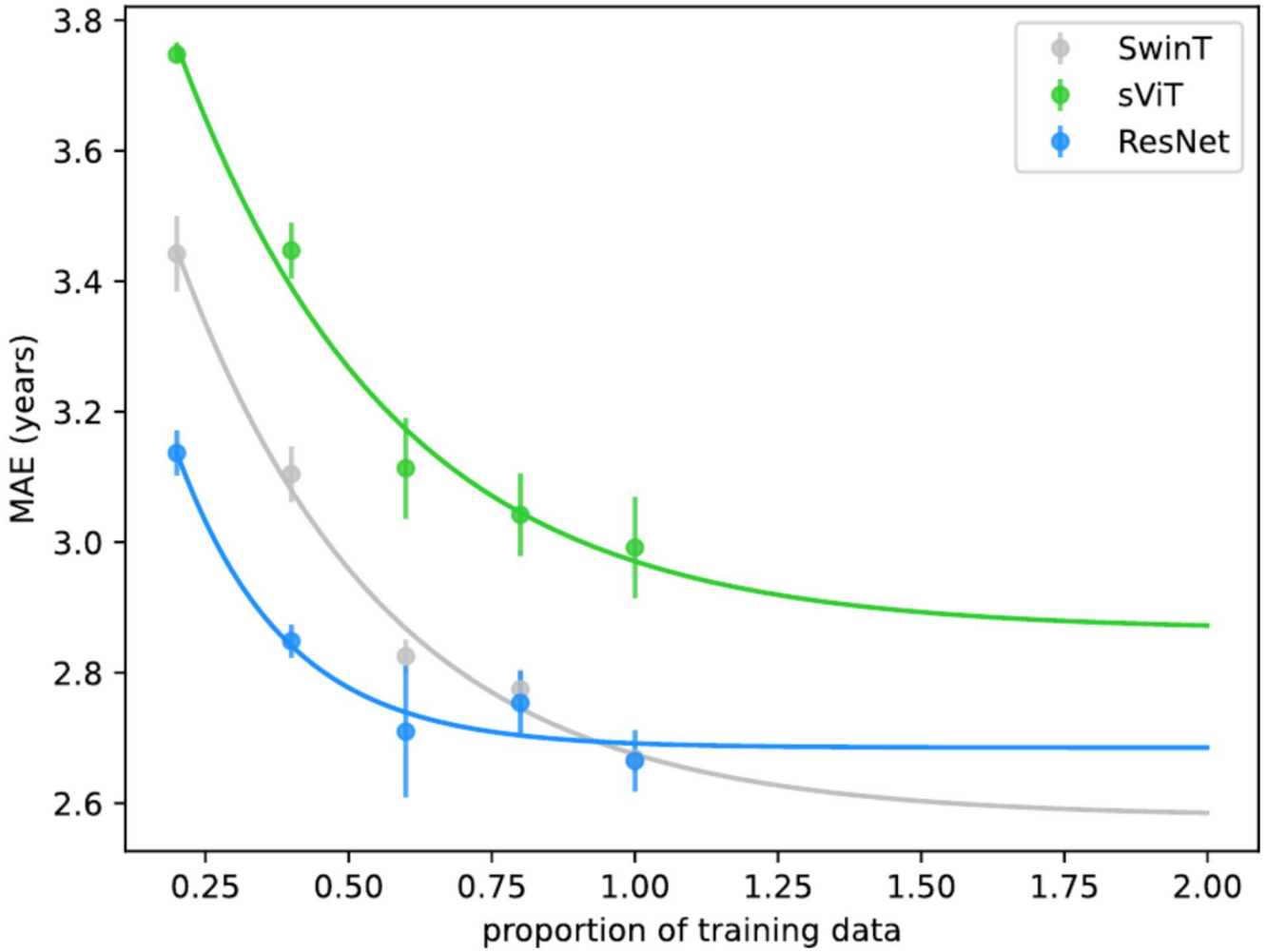
Proposed 3D CNN Architecture Including the Anatomical Feature Attention Module



[Zhang et al., 2024]

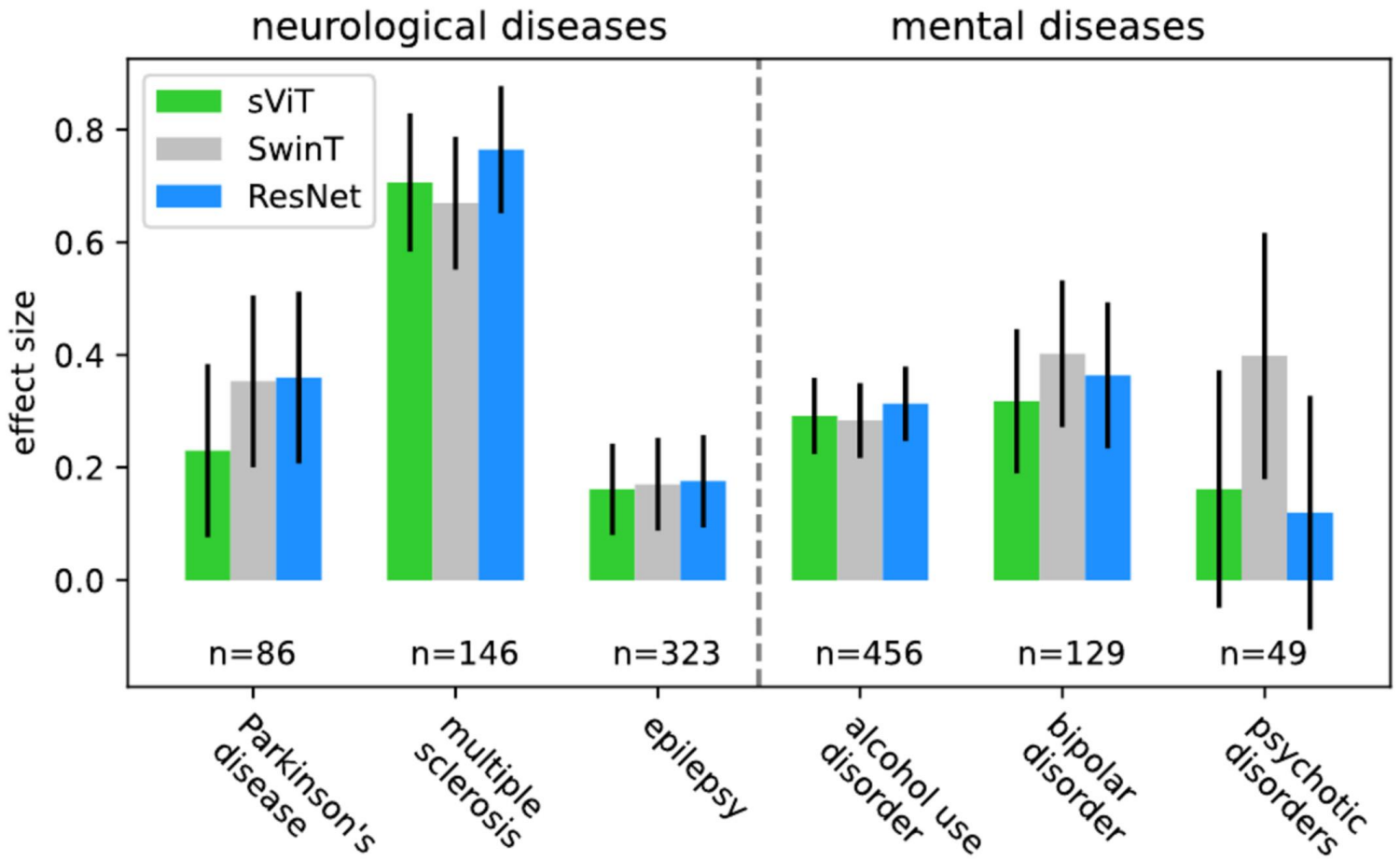
Heatmaps Across Age Range Groups

- Transformer [Siegel et al., 2025]
  - Brain map
  - Datasets: n = 46,381 (44 ~ 83 years)
    - Training: n = 27,538
    - Validation: n = 16,499
    - Test: n = 1,172
  - MAE:
    - 2.66±0.05 years (ResNet)
    - 2.67±0.02 years (Swin Transformer)
    - 3.02±0.08 years (simplified ViT)
  - Model explanations through IxG (Input X Gradient)



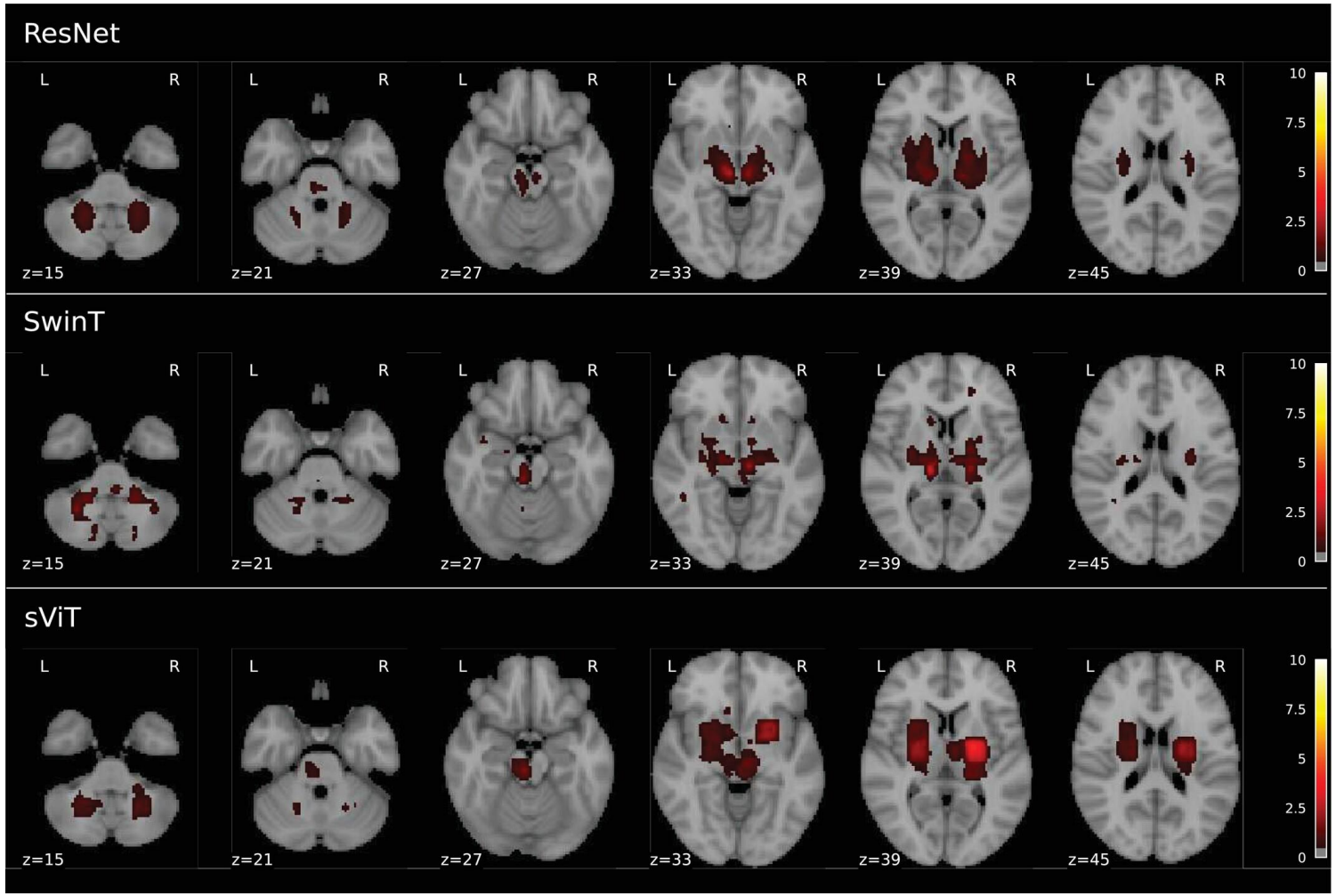
[Siegel et al., 2025]

Data Efficiency Comparison: CNN vs. Transformer Architectures



[Siegel et al., 2025]

Clinical Utility Assessment: CNN vs. Transformer Architectures



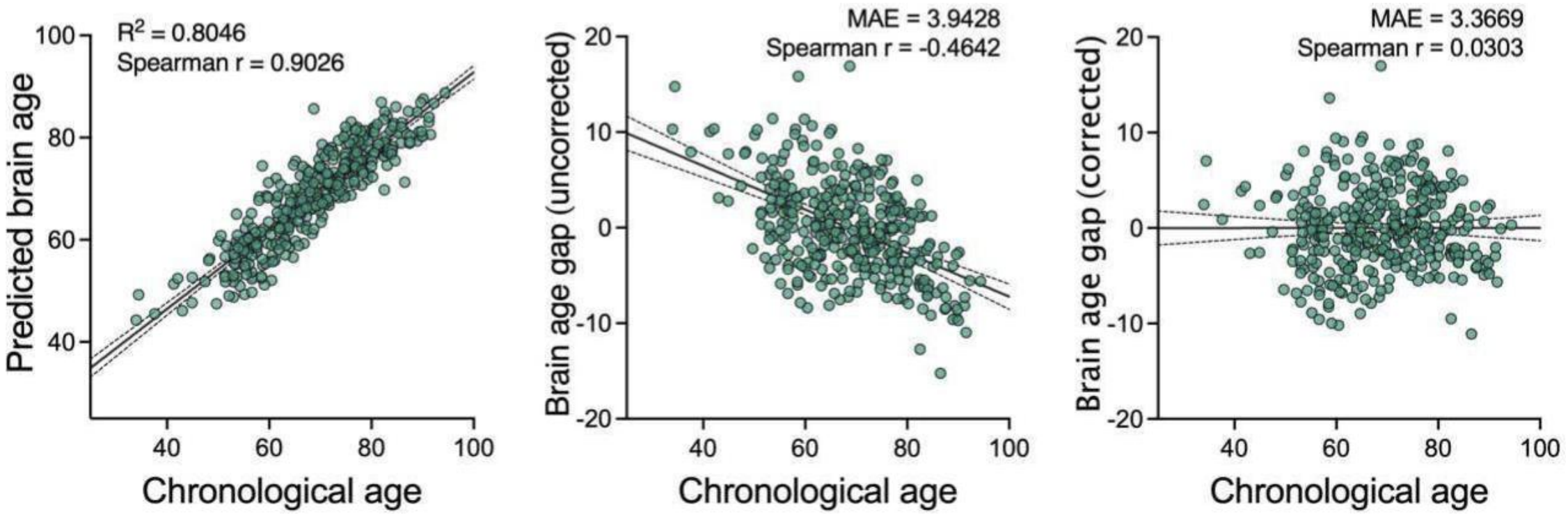
[Siegel et al., 2025]

Model Interpretability: CNN vs. Transformer Architectures

Bias in Age Prediction

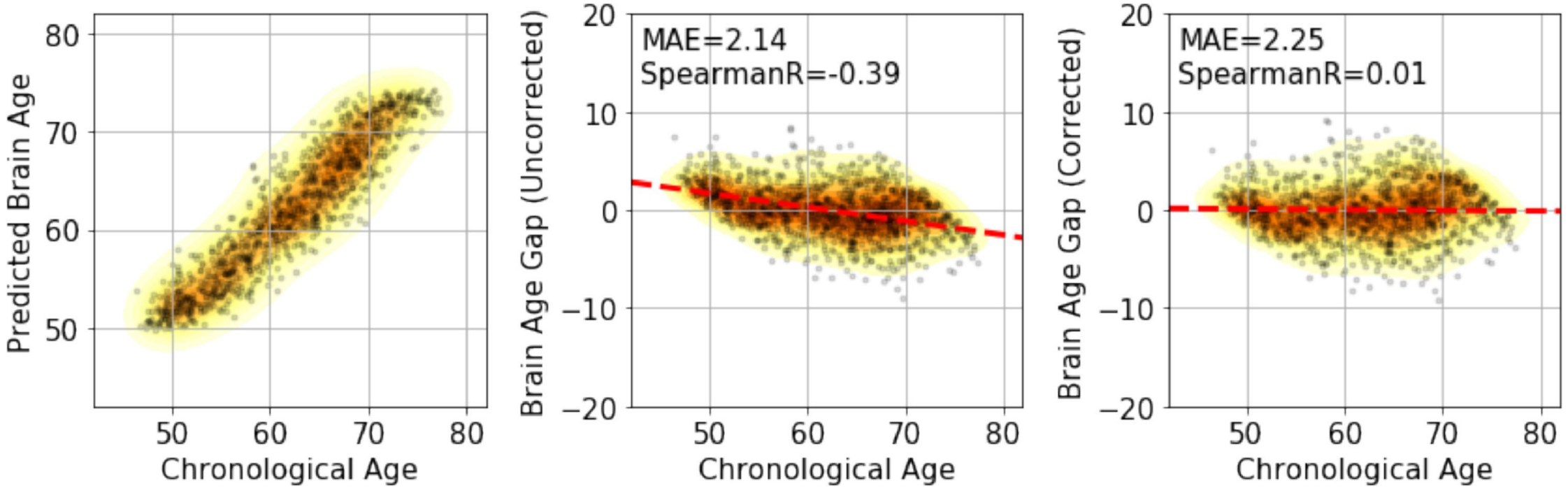
- Tendency to be biased towards the mean age of the total cohort
  - Overestimated brain age in younger individuals, but underestimated brain age in older individuals
- Induces the correlation between chronological age and BAG
  - Impacts the relationship between BAG and other variables of interest when they are also related to age

- Explained by the concept of regression to the mean (RTM) in statistics
  - For values observed with random error
  - Neither data-dependent nor specific to particular methods including deep learning
- Needs to be adjusted by regressing chronological age on brain age or brain age gap to provide corrected brain age gap
  - (brain age) ~ a × (chronological age) + b [Liang et al., 2019]
  - (BAG) ~ a × (chronological age) + b [Le et al., 2018]



[Lee et al., 2022]

Correction of BAG (1)

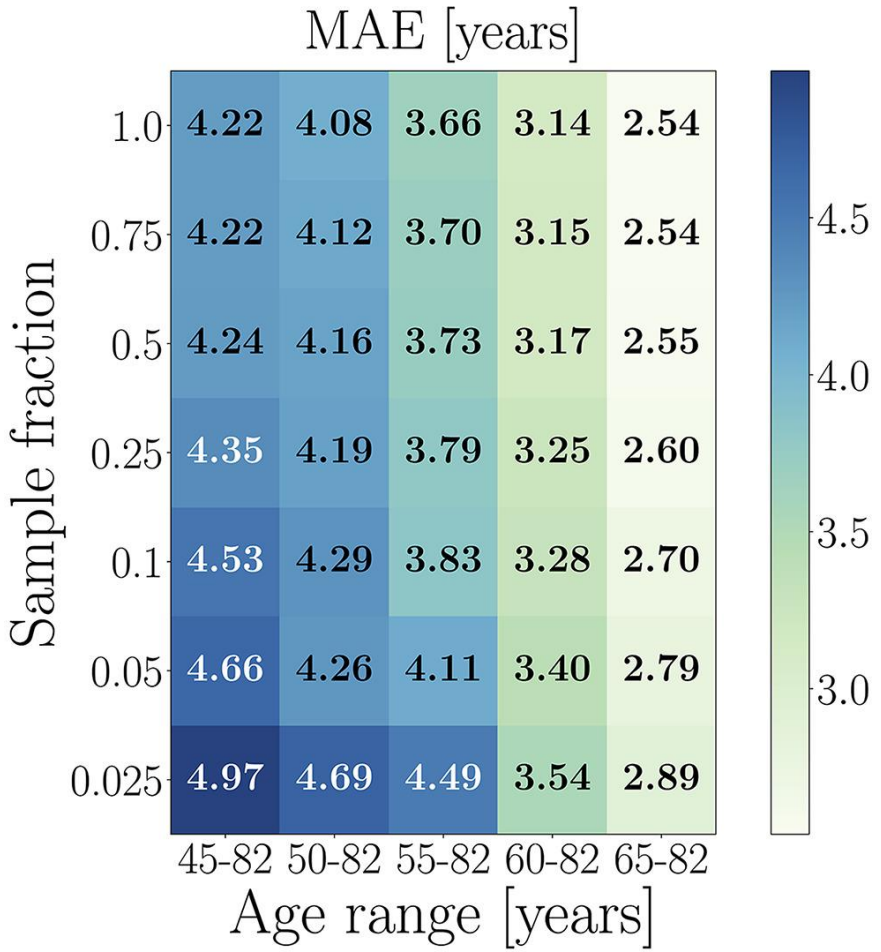


[Peng et al., 2021]

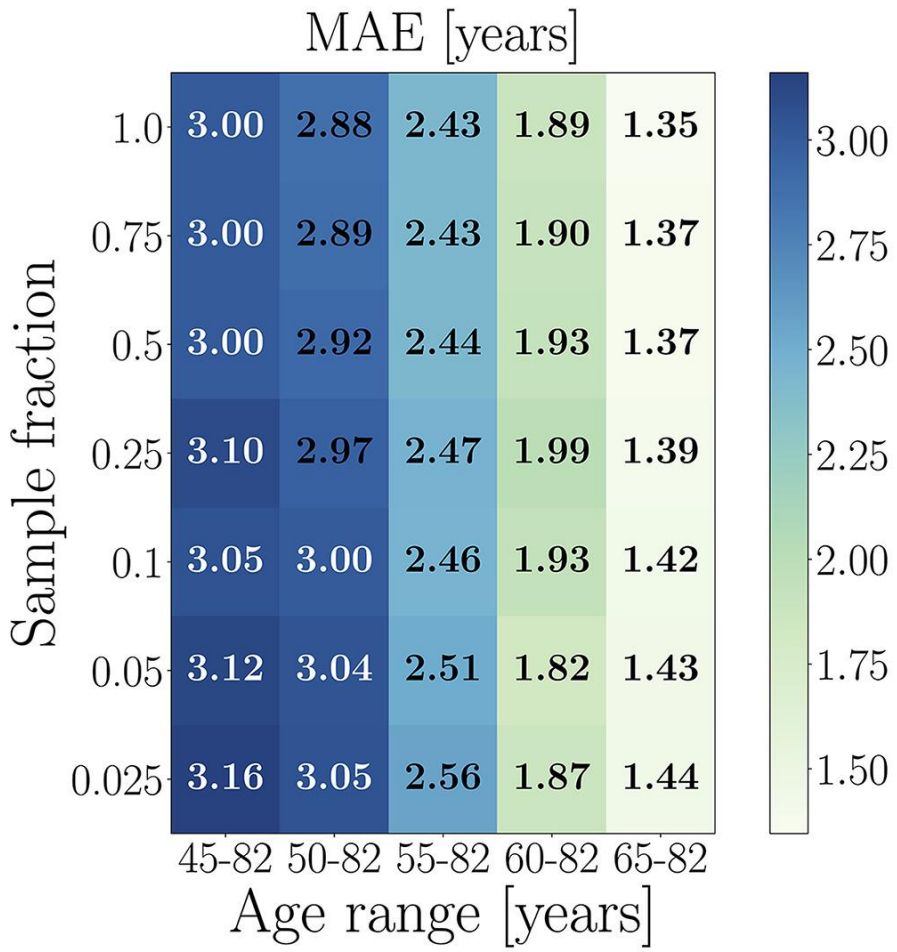
Correction of BAG (2)

Effects of Age Range and Sample Size on Age Prediction

- Better performance in samples with a narrower age range
  - Due to smaller error when predictions are closer to the mean age of the total cohort
- Better performance for larger sample sizes across different age ranges



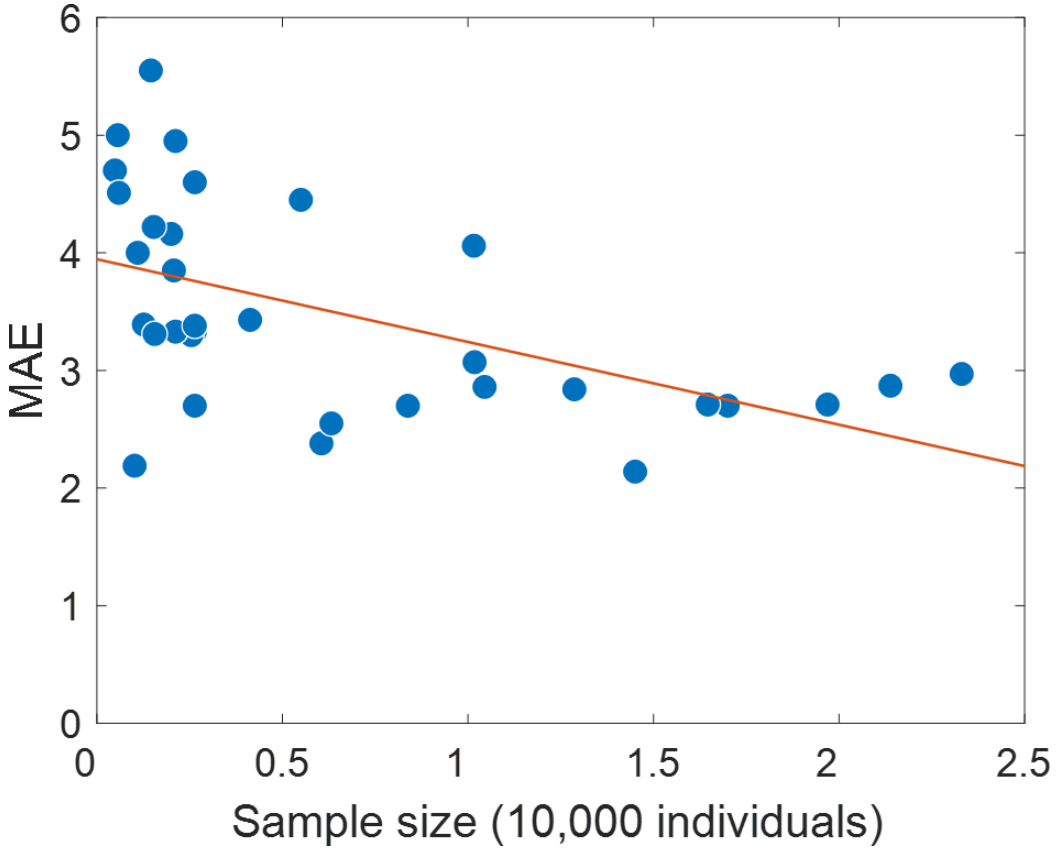
Without bias adjustment



With bias adjustment

[de Lange et al., 2022]

Comparison of Performance for Different Age Ranges and Sample Sizes



[Adapted from Tanveer et al., 2023]

Relationship between Sample Size and Performance

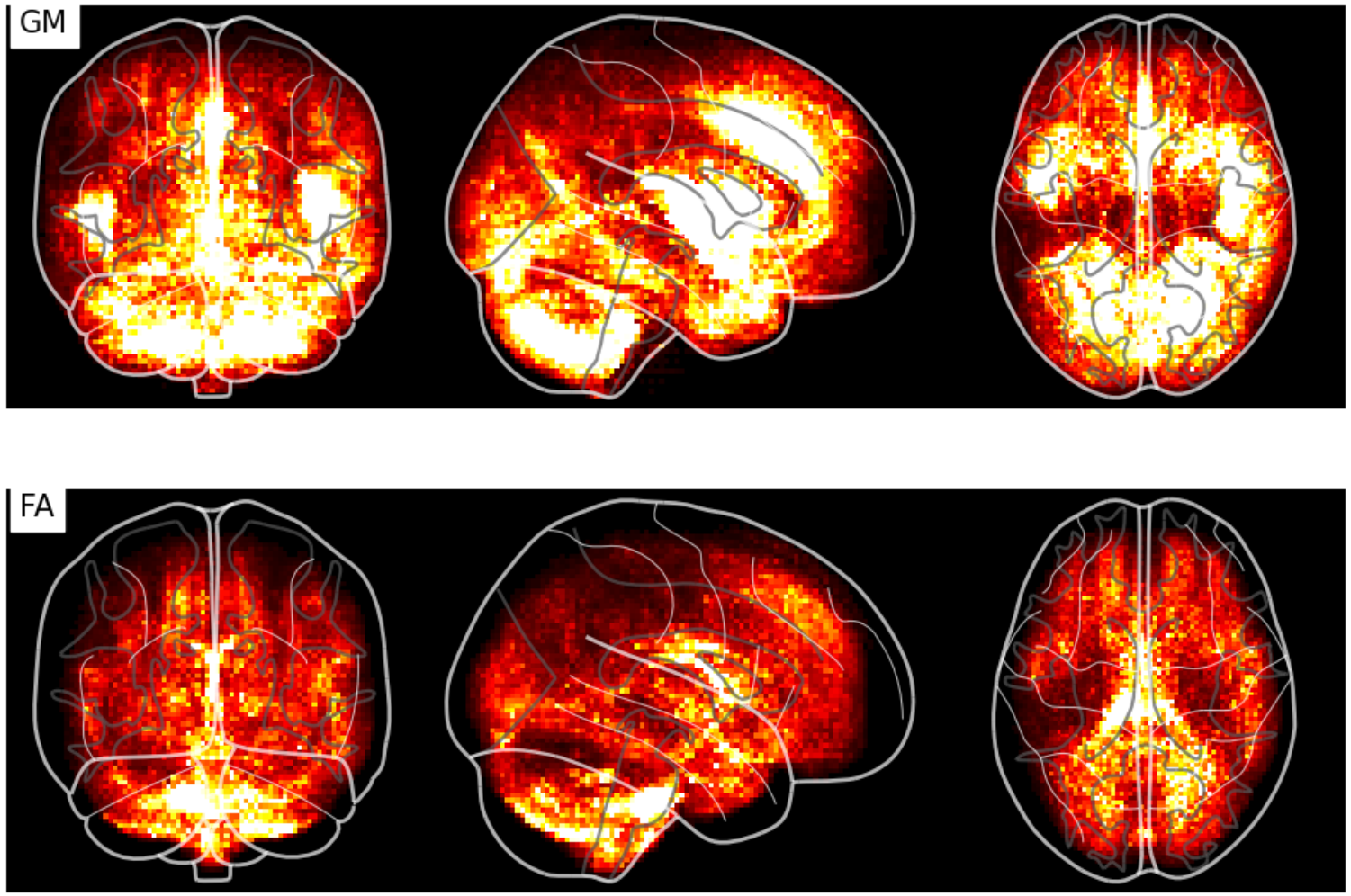
Model Explanations: SHAP Techniques

- SHAP (Shapley Additive Explanations) [Lundberg and Lee, 2017]
  - Game-theoretic approach to explain the output of machine learning models
    - Particularly grounded in the Shapley value that distributes the total gain generated by the coalition of all players (features) fairly among them
  - SHAP values provide a unified measure of feature importance by distributing the prediction among the features, indicating how much each feature contributes to a model's output

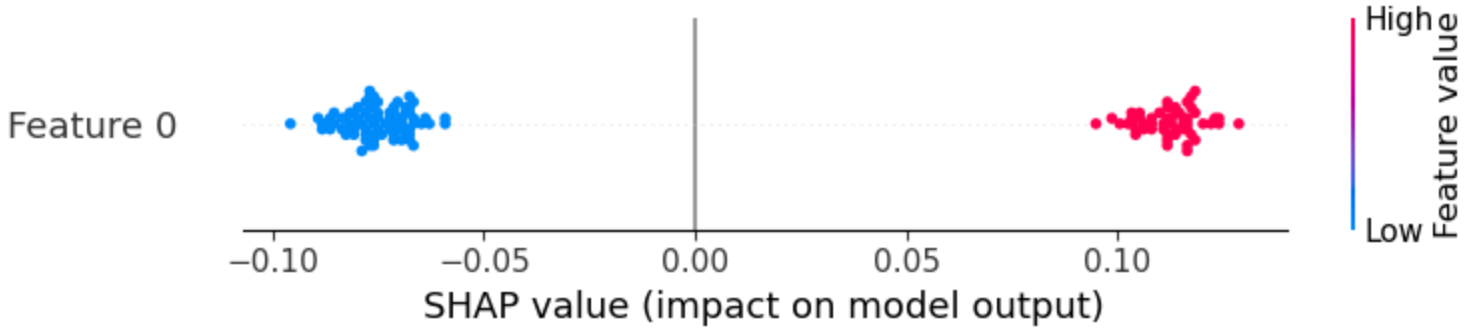
- Key features
  - Interpretability
    - SHAP values provide consistent and intuitive explanations for model predictions across different model types
  - Additivity
    - The sum of all feature contributions equals the difference between the prediction and the expected model output
  - Efficiency
    - Features that do not contribute to the model receive zero attribution, ensuring computational and conceptual efficiency
  - Symmetry
    - Features with identical contributions across all possible coalitions receive equal SHAP values, regardless of order

- SHAP variants
  - Model-agnostic:
    - KernelSHAP
      - Uses coalition sampling and weighted regression to estimate SHAP values for any model type
    - PermutationSHAP
      - Employs feature permutation approach to compute SHAP values for complex models
  - For deep learning models:
    - DeepSHAP
      - Combines gradient-based attribution with Shapley value principles for neural networks
    - GradientSHAP
      - Integrates gradients with reference baselines to approximate SHAP values efficiently

- For tree-based models:
  - TreeSHAP
    - Uses exact algorithms to compute SHAP values for decision trees and ensemble methods
- For linear models:
  - LinearSHAP
    - Provides exact SHAP computation through direct coefficient analysis



SHAP Magnitude Maps Displayed Using `nilearn.plotting.plot_glass_brain`



SHAP Value Plot Displayed Using `shap.summary_plot`

Demonstration Experiments

- `monai.networks.nets`
  - `Regressor`
    - Basic 3D CNN with sequential convolution-pooling layers and fully connected output
    - Age prediction applicability: Direct volumetric brain feature extraction to continuous age value regression
    - Advantages: Minimal architectural complexity, fast training convergence, and clear baseline for comparison

- **`ResNet`** (ResNet-50)
  - Residual learning with skip connections for deep 3D brain feature extraction
  - Age prediction applicability: Effective capture of complex aging patterns through deep hierarchical features
  - Advantages: Proven architecture stability, excellent feature learning capability, and robust performance
- **`DenseNet: DenseNet121`** (DenseNet-121)
  - Dense connectivity with feature reuse for efficient brain representation learning
  - Age prediction applicability: Comprehensive age-related feature utilization through dense connections
  - Advantages: Reduced overfitting through feature reuse, memory-efficient design, and strong gradient flow

- **`SENet: SEResNet50`** (SE-ResNet-50)
  - Squeeze-and-Excitation attention mechanism for channel-wise feature recalibration
  - Age prediction applicability: Adaptive emphasis on age-discriminative features through channel attention
  - Advantages: Minimal computational overhead, improved feature selectivity, and enhanced representation quality
- **`EfficientNet: EfficientNetBN`** (EfficientNet-B0)
  - Compound scaling approach balancing depth, width, and resolution for optimal efficiency
  - Age prediction applicability: Optimal resource utilization for age prediction with scalable design
  - Advantages: Superior accuracy-efficiency trade-off, systematic scaling methodology, and reduced computational requirements

– ViT

- Pure transformer architecture with global self-attention for brain age modeling
- Age prediction applicability: Global brain connectivity modeling through attention mechanisms across 3D volumes
- Advantages: Capture of long-range dependencies, attention-based interpretability, and state-of-the-art representation learning

- SFCN
  - Motivation for inclusion
    - Winner of PAC (Predictive Analytic Challenge) 2019
    - Specifically designed and validated for age prediction [Peng et al., 2021]
    - Represents domain-specific architectural optimization for neuroimaging
  - Architecture characteristics
    - Lightweight convolutional pathway with sequential pooling
    - Uses 1×1 convolution instead of traditional fully connected layer (conceptually cleaner but functionally equivalent)
  - Age prediction applicability
    - Optimized specifically for 3D brain MRI analysis
    - Demonstrated competitive performance on large-scale benchmarks

- Advantages
  - Significantly reduced parameter count (~3M parameters)
  - Faster training and inference compared to general-purpose deep networks
  - Practical baseline demonstrating that simpler architectures can be competitive for specialized tasks
- Why SFCN succeeds
  - Brain aging involves primarily global, gradual changes rather than fine-grained local patterns
  - Appropriate model capacity for the task complexity prevents overfitting
  - Demonstrates that deeper is not always better when features are relatively simple

Category	Parameters	Models
Small	< 10M	Regressor, SFCN, EfficientNetB0
Medium	10-50M	DenseNet121, ResNet50, SEResNet50
Large	50-100M	-
Very large	> 100M	ViT

Architecture Scale Comparison

- Implementation of SFCN-based age prediction
  - Input: GM + WM + CSF
  - Number of trainable parameters: 2,960,449
  - Validation set: MAE = 4.231 years
  - Test set: MAE = 4.170 years (r = 0.931)