École de technologie supérieure
Department of Software Engineering
and Information Technology
Neuro-iX

Presentation of Ten Papers

Towards Automated Neuroanatomy :
Segmentation and Landmark Localization

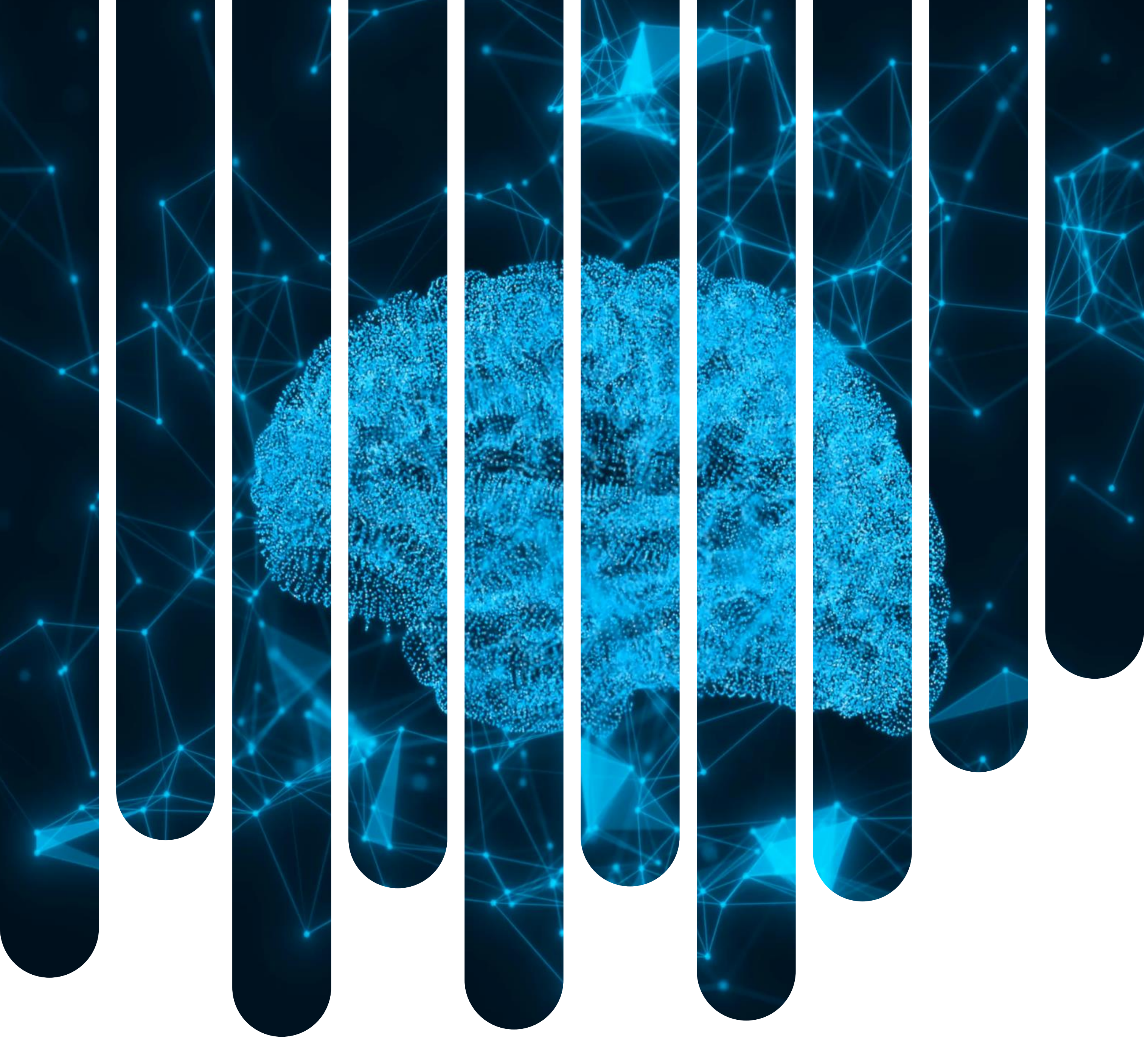
Prepared by:
Ahmed REKIK

Supervised by:
M.Sylvain BOUIX

Course MTR871: Directed Readings

Master's in Information Technology

Session Winter 2025



Contents

1

General Introduction

Overview of the Core Problem Explored in the 10 Articles

- **Anatomical Segmentation**:

Precise Identification of Structures in Medical Images

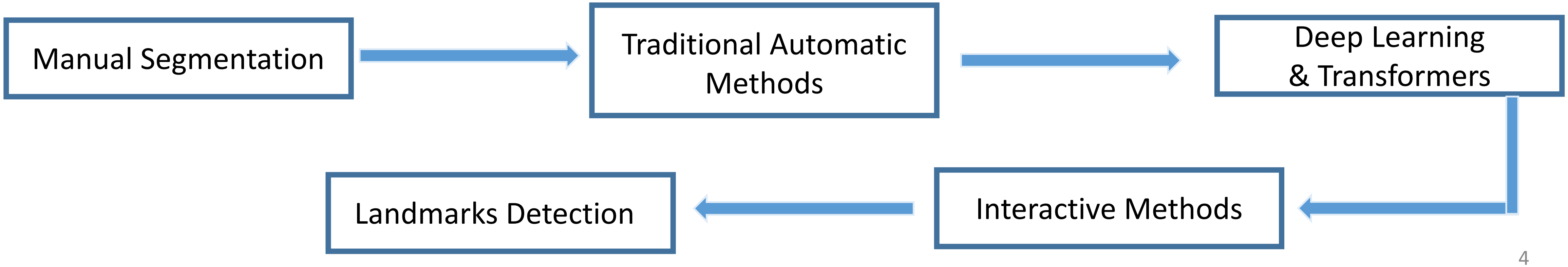
- **Importance**:

A fundamental step in many clinical analyses (e.g., volume and shape measurements), and essential for training AI models.

- **Challenges**:

Manual segmentation is time-consuming, subject to inter- and intra-rater variability, and lacks scalability — highlighting the urgent need for automation without compromising quality.

- The Path We Follow: Segmentation Methodologies Over Time



2

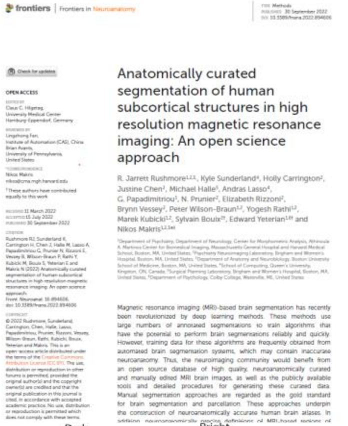
Manual Segmentation:

(Paper 1)

Manual Segmentation – Expert Protocol

Study by Rushmore et al., 2022:

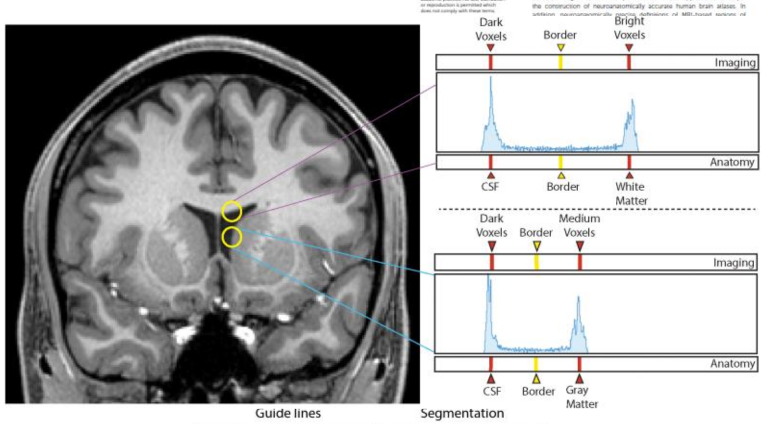
Manual segmentation of 28 brain structures on high-resolution MRI.


Anatomically curated segmentation of human subcortical structures in high resolution magnetic resonance imaging: An open science approach

Innovative tools : 3D Slicer Segmentation Tool:

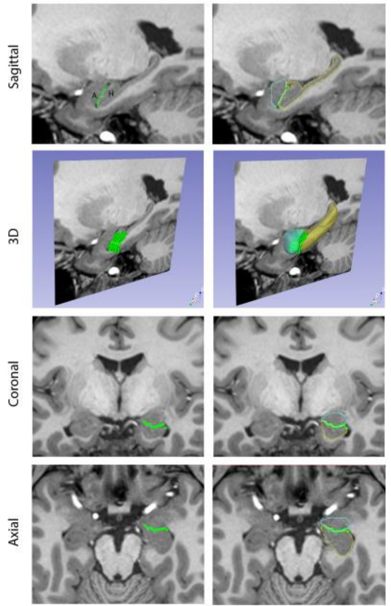
- <u>Intensity histograms</u> : guide threshold selection between regions.

- <u>Guide Markup Tool</u> : Sagittal landmarks guide coronal segmentation for better 3D consistency.



Standardized protocols :

NeuroNames ontology and Harvard-Oxford atlas adapted for high-resolution imaging.



Goal :

Create a robust, manually segmented database (50 brains) open to the research community.

Manual Segmentation – Reliability and Limitations

Results

- **High inter- and intra-rater reliability** (Dice Coefficient > 0.90 for most structures).

- Detected cerebral asymmetries:

Ex :larger nucleus accumbens on the left, hippocampus larger on the right.

- Sex-based anatomical differences: subcortical structures appear larger in males.

Limitations

- Extremely time- and labor-intensive.

- Difficult to scale for large datasets.

- Strong dependence on expert annotators.

Region of interest (ROI)	Mean Dice	SD	Min	Max
Lateral Ventricle Left	0.95	0.02	0.92	0.98
Lateral Ventricle Right	0.95	0.02	0.93	0.98
Third Ventricle	0.84	0.05	0.75	0.90
Fourth Ventricle	0.87	0.04	0.80	0.94
Nucleus Accumbens Left	0.84	0.04	0.78	0.89
Nucleus Accumbens Right	0.84	0.05	0.76	0.93
Caudate Left	0.93	0.01	0.91	0.96
Caudate Right	0.93	0.02	0.88	0.96
Putamen Left	0.93	0.02	0.91	0.99
Putamen Right	0.93	0.02	0.91	0.99
Globus Pallidus Left	0.83	0.04	0.76	0.90
Globus Pallidus Right	0.81	0.06	0.73	0.90
Brainstem	0.95	0.01	0.94	0.98
Thalamus Left	0.88	0.04	0.78	0.92
Thalamus Right	0.88	0.03	0.82	0.93
Ventral Diencephalon Left	0.88	0.02	0.84	0.92
Ventral Diencephalon Right	0.88	0.01	0.85	0.90
Inferior Horn of Lateral Ventricle Left	0.72	0.05	0.61	0.82
Inferior Horn of Lateral Ventricle Right	0.72	0.05	0.61	0.81
Hippocampal Formation Left	0.87	0.03	0.82	0.90
Hippocampal Formation Right	0.87	0.02	0.82	0.90
Amygdala Left	0.84	0.03	0.78	0.88
Amygdala Right	0.80	0.05	0.71	0.88
Fifth Ventricle	0.76	0.07	0.65	0.84
Optic Chiasm	0.74	0.15	0.54	0.95

Inter-rater reliability

Region of interest (ROI)	Mean Dice	SD	Min	Max
Lateral Ventricle Left	0.95	0.02	0.93	0.97
Lateral Ventricle Right	0.96	0.03	0.93	0.98
Third Ventricle	0.89	0.04	0.84	0.91
Fourth Ventricle	0.90	0.03	0.87	0.93
Nucleus Accumbens Left	0.87	0.04	0.82	0.89
Nucleus Accumbens Right	0.89	0.02	0.87	0.91
Caudate Left	0.93	0.03	0.91	0.96
Caudate Right	0.94	0.02	0.92	0.95
Putamen Left	0.94	0.02	0.92	0.95
Putamen Right	0.94	0.02	0.91	0.95
Globus Pallidus Left	0.81	0.05	0.77	0.86
Globus Pallidus Right	0.80	0.05	0.76	0.85
Brainstem	0.96	0.01	0.95	0.97
Thalamus Left	0.91	0.01	0.91	0.92
Thalamus Right	0.91	0.03	0.88	0.93
Ventral Diencephalon Left	0.90	0.01	0.89	0.91
Ventral Diencephalon Right	0.90	0.01	0.89	0.91
Inferior Horn of Lateral Ventricle Left	0.76	0.09	0.68	0.86
Inferior Horn of Lateral Ventricle Right	0.80	0.04	0.75	0.83
Hippocampal Formation Left	0.90	0.02	0.89	0.93
Hippocampal Formation Right	0.90	0.05	0.85	0.94
Amygdala Left	0.84	0.07	0.77	0.91
Amygdala Right	0.84	0.07	0.76	0.90
Fifth Ventricle	0.75	0.08	0.66	0.83
Optic Chiasm	0.87	0.18	0.66	0.99

Intra-rater reliability

3

Traditional Automatic Segmentation

(Papers 2 & 3)

Traditional Automatic Segmentation
- Probabilistic Atlas-


Neuron, Vol. 33, 341–355, January 31, 2002, Copyright ©2002 by Cell Press

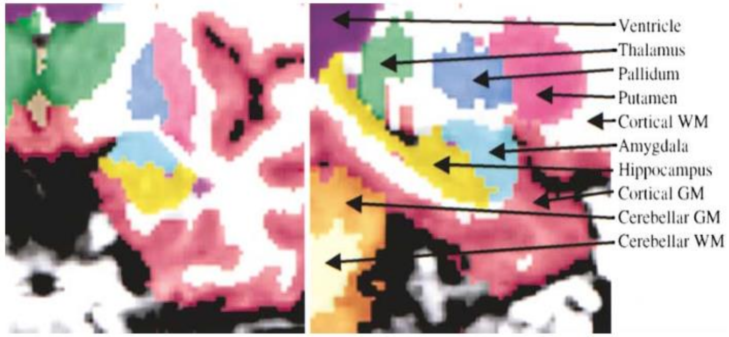
Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain

Neurotechnique

Bruce Fischl,[1] David H. Salat,[1] Evelina Busa,[1] Marilyn Albert,[2,3] Megan Dieterich,[1] Christian Haselgrove,[5] Andre van der Kouwe,[1] Ron Killiany,[4] David Kennedy,[5] Shuna Klaveness,[1] Albert Montillo,[6] Nikos Makris,[5] Bruce Rosen,[1] and Anders M. Dale[1,7]
[1] Massachusetts General Hospital
Nuclear Magnetic Resonance Center
Rm. 2328, Building 149
13th Street
Charlestown, Massachusetts 02129
[2] Department of Neurology
Massachusetts General Hospital
Harvard Medical School
55 Fruit Street, VBK 901
Boston, Massachusetts 02114

structural changes in the brain. These changes can cause alterations in the imaging properties of brain tissue, as well as changes in morphometric properties of brain structures. Morphometric changes may include variations in the volume or shape of subcortical regions, as well as alterations in the thickness, area, and folding pattern of the cortex. While surface-based analyses that depend on models of the position and orientation of the cortical ribbon can provide an accurate assessment of cortical variability, volumetric techniques are required to detect changes in noncortical structures. For example, changes in ventricular or hippocampal volume are frequently associated with a variety of diseases (e.g., Puri et al., 1999; Killiany et al., 2000; Wolf et al., 2001). This type of analysis has commonly been accomplished by a having a trained anatomist or technician manually label

Méthode de Fischl et al. (2002) :



- First algorithm for whole-brain automatic labeling.
- Implemented in FreeSurfer — a reference tool in neuroimaging

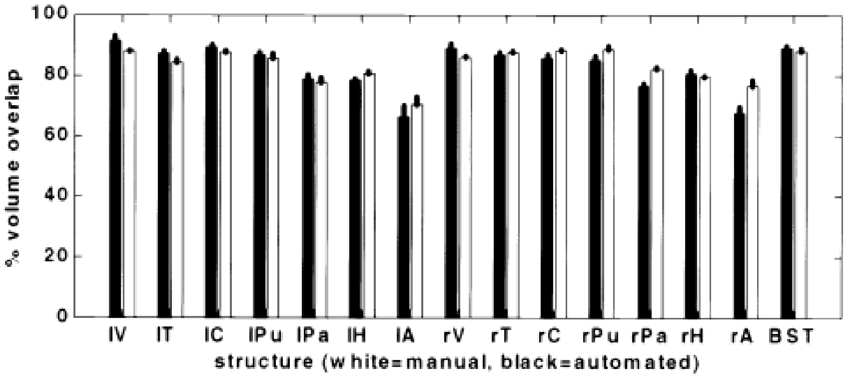
Principle of atlas-based segmentation :

1. Statistical Atlas Construction :
   - Each voxel in the probabilistic atlas encodes intensity, location, and spatial context for anatomical labeling.

2. MRI-to-Atlas Registration :
   - Each Aligns MRI scans to the atlas reference space.

$$L = \arg\min_{L} \int (T(r) - I(Lr))^2 \, dr$$

3. Bayesian Voxel-Wise Labeling :
   - Bayesian MAP estimation to assign the most probable tissue label
   - Markov Random Field modeling to enforce anatomical consistency across neighboring voxels



Results :

- High accuracy (~90% Dice Score) for subcortical structures; detects subtle disease-related changes (e.g., in Alzheimer's)
- Limitation: Sensitive to registration errors; lacks modeling of inter-individual anatomical variability (single average atlas)

Traditional Automatic Segmentation
-Multi-Atlas Label Fusion-


HHS Public Access

Author manuscript

IEEE Trans Pattern Anal Mach Intell. Author manuscript; available in PMC 2013 December 22.

Published in final edited form as:
IEEE Trans Pattern Anal Mach Intell. 2013 March ; 35(3): 611–623. doi:10.1109/TPAMI.2012.143.

Multi-Atlas Segmentation with Joint Label Fusion

Hongzhi Wang[1] [Member, IEEE], Jung W. Suh[2], Sandhitsu R. Das[1], John Pluta[3], Caryne Craige[4], and Paul A. Yushkevich[1] [Member, IEEE]

[1]Penn Image Computing and Science Lab, Department of Radiology, University of Pennsylvania

[2]HeartFlow, Inc

[3]Departments of Neurology and Radiology, University of Pennsylvania

[4]School of Medicine, Temple University

Abstract

Multi-atlas segmentation is an effective approach for automatically labeling objects of interest in biomedical images. In this approach, multiple expert-segmented example images, called *atlases*, are registered to a target image, and deformed atlas segmentations are combined using *label fusion*. Among the proposed label fusion strategies, weighted voting with spatially varying weight distributions derived from atlas-target intensity similarity have been particularly successful. However, one limitation of these strategies is that the weights are computed independently for each atlas, without taking into account the fact that different atlases may produce similar label errors. To address this limitation, we propose a new solution for the label fusion problem, in which weighted voting is formulated in terms of minimizing the total expectation of labeling error, and in which pairwise dependency between atlases is explicitly modeled as the joint probability of two atlases making a segmentation error at a voxel. This probability is approximated using intensity similarity between a pair of atlases and the target image in the neighborhood of each voxel. We validate our method in two medical image segmentation problems: hippocampus segmentation and hippocampus subfield segmentation in magnetic resonance (MR) images. For both problems, we show consistent and significant improvement over label fusion strategies that assign atlas weights independently.

Index Terms

multi-atlas label fusion segmentation; dependence; hippocampal segmentation

I. Introduction

Atlas-based segmentation is motivated by the observation that segmentation strongly correlates with image appearance. A target image can be segmented by referring to atlases, i.e. expert-labeled sample images. After warping the atlas to the target image via deformable registration, one can directly transfer labels from the atlas to the target image. As an extension, multi-atlas based segmentation makes use of more than one atlas to compensate

- ➢ Uses multiple manually segmented atlases to enhance segmentation accuracy.
- ➢ Each atlas is individually registered to the target MRI, and labels are fused to generate a consensus segmentation.

Selection and preprocessing of atlases → **Atlas registration to the target image** → **Fusion of labels from the aligned atlases**

$$M_x(i,j) = \mathbb{E}[\delta_i(x)\delta_j(x)] \propto \left[\sum_{y \in \mathcal{N}(x)} |F_T(y) - F_i(y)||F_T(y) - F_j(y)| \right]^{\beta} .$$

JLF - Joint Label Fusion (Method proposed by the paper)

$$\mathbf{w}_x = \arg\min_{\mathbf{w}_x} \mathbf{w}_x^T M_x \mathbf{w}_x$$

The main idea is that some atlases may have similar errors, and these dependencies must be taken into account.

$$\hat{S}_T(x) = \sum_{i=1}^{n} w_i(x) S_i(x)$$

Improvement with Multi-Atlas Label Fusion

Méthode	Average Dice Score (%)
Majority Voting (MV)	85.2
Joint Label Fusion (JLF - Proposé)	89.9

- Improved accuracy: +1.5% Dice Score for hippocampal segmentation compared to single-atlas methods.

- Key result: Multi-atlas fusion improves segmentation accuracy while reducing dependence on a single reference atlas.

4

Transformer-Based Segmentation

(Papers 4 & 5)

Transformer-Based Segmentation

Motivation in Medical Imaging

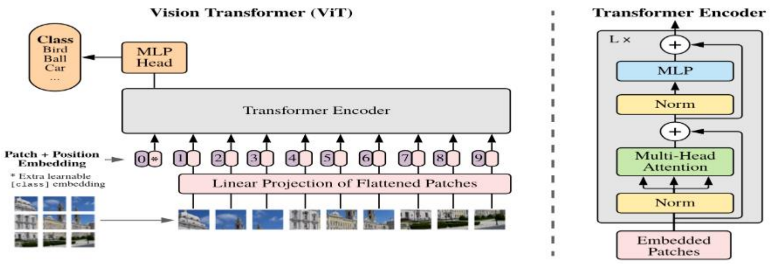
- Anatomy is contextual: Brain structures are interdependent and spatially organized
- Global context matters: A local anomaly is often meaningful only in relation to surrounding regions
- CNNs are limited by small receptive fields → Transformers provide a broader, global view

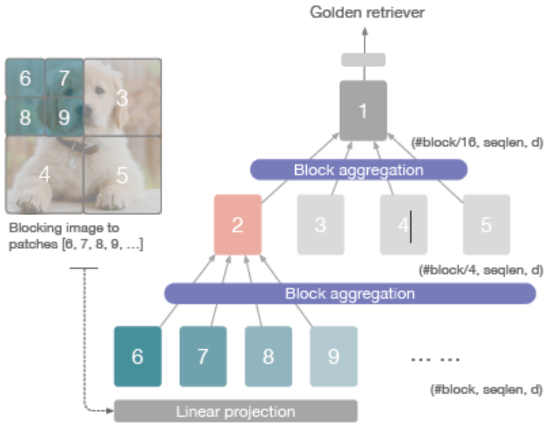
Advances in Transformer Architectures

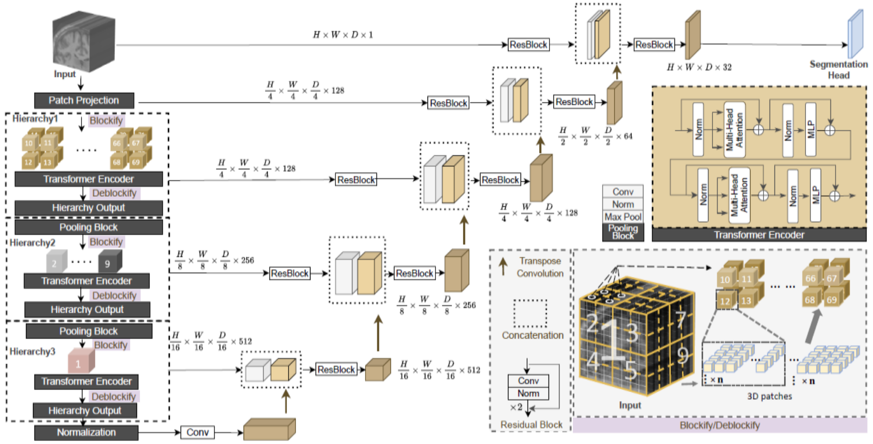
Vision Transformer (ViT)

Nested Transformer (NesT)

UNesT







Vision Transformer (ViT) - Principle and Architecture

- Inspired by Natural Language Processing (NLP) Transformers, ViT applies self-attention mechanisms to images without relying on convolutions.

1. Splitting the image into N patches of size (P×P): $N = \frac{H \times W}{P^2}$.

2. **Patch Encoding**: each patch x_i is flattened and projected into an embedding vector.

3. Addition of the classification token and **positional** embeddings.

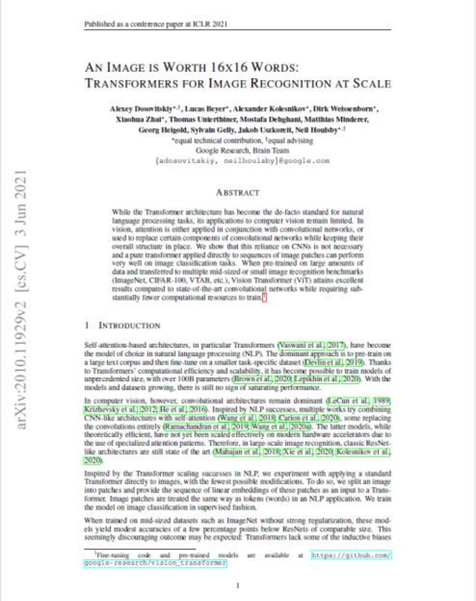
$$z_0 = [x_{\text{class}};\, x_p^1 E;\, x_p^2 E;\, \cdots;\, x_p^N E]. \qquad z_0 = z_0 + E_{\text{pos}}$$

4. **Transformer Encoder**: Stacked L layers of Multi-Head Self-Attention (MSA) and Multi-Layer Perceptron (MLP).

$$z'_\ell = \mathrm{MSA}(\mathrm{LN}(z_{\ell-1})) + z_{\ell-1}, \qquad z_\ell = \mathrm{MLP}(\mathrm{LN}(z'_\ell)) + z'_\ell.$$

5. Final representation of the classification token (first position of z_L) is normalized to obtain the **prediction**.

- Limitation:
  - Performs well on large datasets (e.g., ImageNet-21k, JFT-300M), but struggles on small datasets.



arXiv:2010.11929v2 [cs.CV] 3 Jun 2021

Published as a conference paper at ICLR 2021

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*, Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†
*equal technical contribution, †equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

1 INTRODUCTION

Self-attention-based architectures, in particular Transformers (Vaswani et al., 2017), have become the model of choice in natural language processing (NLP). The dominant approach is to pre-train on a large text corpus and then fine-tune on a smaller task-specific dataset (Devlin et al., 2019). Thanks to Transformers' computational efficiency and scalability, it has become possible to train models of unprecedented size, with over 100B parameters (Brown et al., 2020; Lepikhin et al., 2020). With the models and datasets growing, there is still no sign of saturating performance.

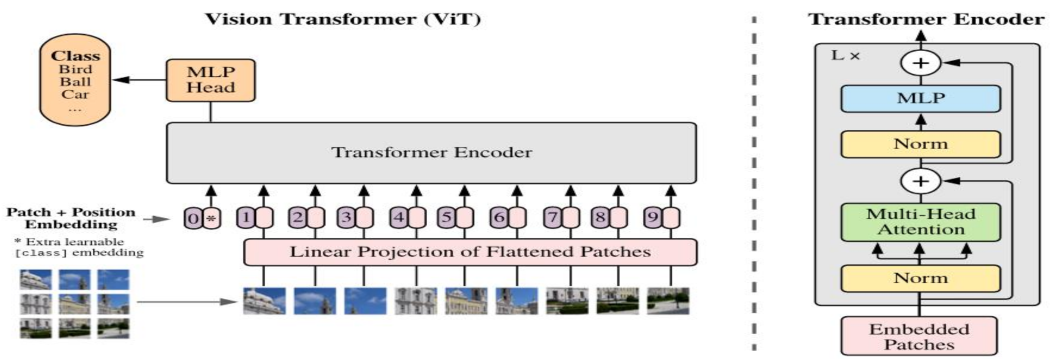
In computer vision, however, convolutional architectures remain dominant (LeCun et al., 1989; Krizhevsky et al., 2012; He et al., 2016). Inspired by NLP successes, multiple works try combining CNN-like architectures with self-attention (Wang et al., 2018; Carion et al., 2020), some replacing the convolutions entirely (Ramachandran et al., 2019; Wang et al., 2020a). The latter models, while theoretically efficient, have not yet been scaled effectively on modern hardware accelerators due to the use of specialized attention patterns. Therefore, in large-scale image recognition, classic ResNet-like architectures are still state of the art (Mahajan et al., 2018; Xie et al., 2020; Kolesnikov et al., 2020).

Inspired by the Transformer scaling successes in NLP, we experiment with applying a standard Transformer directly to images, with the fewest possible modifications. To do so, we split an image into patches and provide the sequence of linear embeddings of these patches as an input to a Transformer. Image patches are treated the same way as tokens (words) in an NLP application. We train the model on image classification in supervised fashion.

When trained on mid-sized datasets such as ImageNet without strong regularization, these models yield modest accuracies of a few percentage points below ResNets of comparable size. This seemingly discouraging outcome may be expected: Transformers lack some of the inductive biases

[1]Fine-tuning code and pre-trained models are available at https://github.com/google-research/vision_transformer

1



Nested Transformer (NesT) – A Hierarchical Improvement

➢ NesT introduces a nested hierarchical organization to better capture spatial relationships.

1. Image partitioning into blocks (instead of individual patches).

2. **Local processing**: Each block is independently analyzed by a local Transformer.

3. **Hierarchical block fusion**: Gradual aggregation of blocks to capture global context.

   - Progressive aggregation using **3×3 convolutions** and **max-pooling**.
   - Ensures better integration of local and global information.
   - Each set of 4 neighboring blocks is merged into a higher-level block.
   - Gradually reduces the total number of blocks.

➢ Advantages over ViT:

- Requires less data: Achieves better results on smaller datasets (CIFAR-10, ImageNet).

- Faster training: Thanks to its nested hierarchical structure.

Nested Hierarchical Transformer: Towards Accurate, Data-Efficient and Interpretable Visual Understanding

Zizhao Zhang[1] Han Zhang[2] Long Zhao[2] Ting Chen[2] Sercan Ö. Arık[1] Tomas Pfister[1]

[1]Google Cloud AI [2]Google Research

arXiv:2105.12723v4 [cs.CV] 30 Dec 2021

Abstract

Hierarchical structures are popular in recent vision transformers, however, they require sophisticated designs and massive datasets to work well. In this paper, we explore the idea of nesting basic local transformers on non-overlapping image blocks and aggregating them in a hierarchical way. We find that the block aggregation function plays a critical role in enabling cross-block non-local information communication. This observation leads us to design a simplified architecture that requires minor code changes upon the original vision transformer. The benefits of the proposed judiciously-selected design are threefold: (1) NesT converges faster and requires much less training data to achieve good generalization on both ImageNet and small datasets like CIFAR; (2) when extending our key ideas to image generation, NesT leads to a strong decoder that is 8× faster than previous transformer-based generators; and (3) we show that decoupling the feature learning and abstraction processes via this nested hierarchy in our design enables constructing a novel method (named GradCAT) for visually interpreting the learned model. Source code is available https://github.com/google-research/nested-transformer

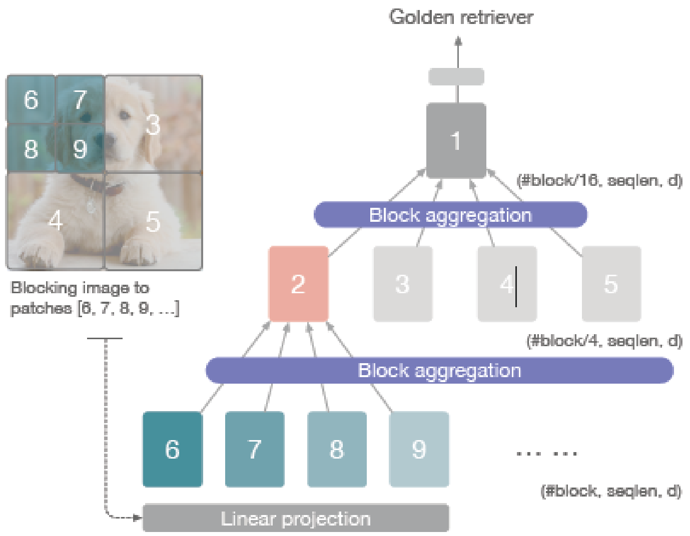
Introduction

Vision Transformer (ViT) (Dosovitskiy et al. 2021) model and its variants have received significant interests recently due to their superior performance on many core visual applications (Cordonnier, Loukas, and Jaggi 2020; Liu et al. 2021). ViT first splits an input image into patches, and then patches are treated in the same way as tokens in NLP applications. Following, several self-attention layers are used to conduct global information communication to extract features for classification. Recent work (Dosovitskiy et al. 2021; Cordonnier, Loukas, and Jaggi 2020) shows that ViT models can achieve better accuracy than state-of-the-art convnets (Tan and Le 2019; He et al. 2016) when trained on datasets with tens or hundreds of millions of labeled samples. However, when trained on smaller datasets, ViT usually underperforms its counterparts based on convolutional layers. Addressing this data inefficiency is important to make ViT applicable to other application scenarios, e.g. semi-supervised learning (Sohn et al. 2020) and generative modeling (Goodfellow et al. 2014; Zhang et al. 2019).

Lack of inductive bias such as locality and translation equivariance, is one explanation for the data inefficiency of ViT models. Cordonnier, Loukas, and Jaggi (2020) discovered that transformer models learn locality behaviors in a deformable convolution manner (Dai et al. 2017): bottom layers attend locally to the surrounding pixels and top layers favor long-range dependency. On the other hand, global self-attention between pixel pairs in high-resolution images is computationally expensive. Reducing the self-attention range is one way to make the model training more computationally efficient (Beltagy, Peters, and Cohan 2020). These type of insights align with the recent structures with local self-attention and hierarchical transformer (Han et al. 2021; Vaswani et al. 2021; Liu et al. 2021). Instead of holistic global self-attention, these perform attention on local image patches. To promote information communication across patches, they propose specialized designs such as the "haloing operation" (Vaswani et al. 2021) and "shifted window" (Liu et al. 2021). These are based on modifying the self-attention mechanism and often yields in complex architectures. Our design goal on the other hand keeping the attention as is, and introducing the design of the aggregation function, to improve the accuracy and data efficiency, while bringing interpretability benefits.

The proposed NesT model stacks canonical transformer blocks to process non-overlapping image blocks individually. Cross-block self-attention is achieved by nesting these transformers hierarchically and connecting them with a proposed aggregation function. Fig. 1 illustrates the overall architecture and the simple pseudo code to generate it. Our contributions can be summarized as:

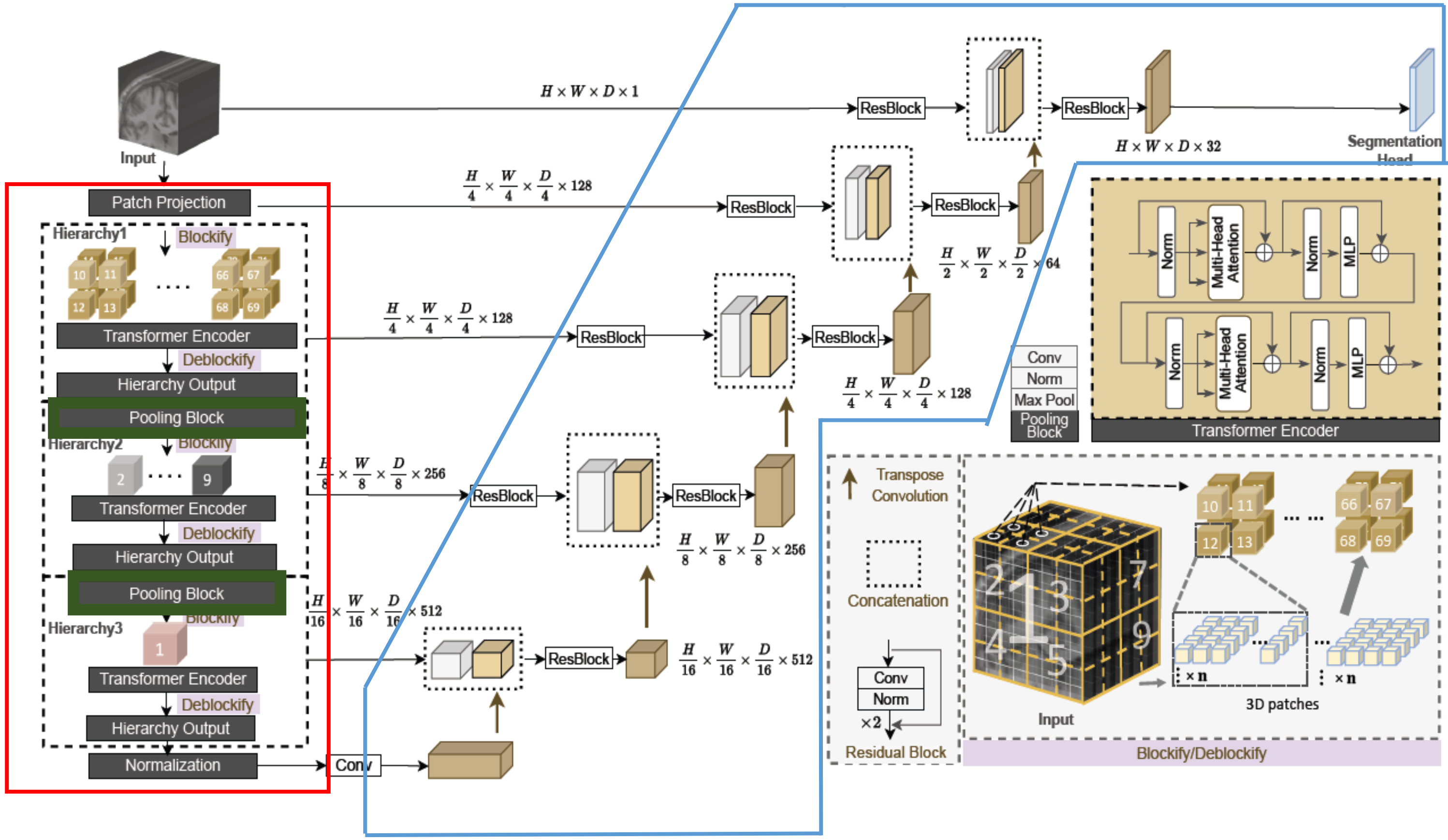
1. We demonstrate integrating hierarchically nested transformers with the proposed block aggregation function can outperform previous sophisticated (local) self-attention variants, leading to a substantially-simplified architecture and improved data efficiency. This provides a novel perspective for achieving effective cross-block communication.
2. NesT achieves impressive ImageNet classification accuracy with a significantly simplified architectural design. E.g., training a NesT with 38M/68M parameters obtains 83.3%/83.8% ImageNet accuracy. The favorable data efficiency of NesT is embodied by its fast convergence, such as achieving 75.9%/82.3% training with 30/100 epochs. Moreover, NesT achieves matched accuracy on small datasets compared with popular convolutional



UNesT – A U-Net with Nested Transformers for 3D Segmentation

Hierarchical Encoder :

- 3D input volume is split into 4×4×4 patches, each flattened into a 128-dimensional vector.
- Patches are grouped into non-overlapping blocks for independent local processing.
- Each block is processed by a mini-transformer (MSA, MLP, LN) with skip connections between layers.



3D aggregation :

- Merges neighboring blocks to capture global spatial context with low memory cost.

Hierarchical Convolutional Decoder :

- At each ascending level, features are upsampled, merged with encoder outputs, and refined toward final segmentation.
- Processes and fuses features at every level to progressively reconstruct the segmentation map.

Comparative Results of Transformer Architectures

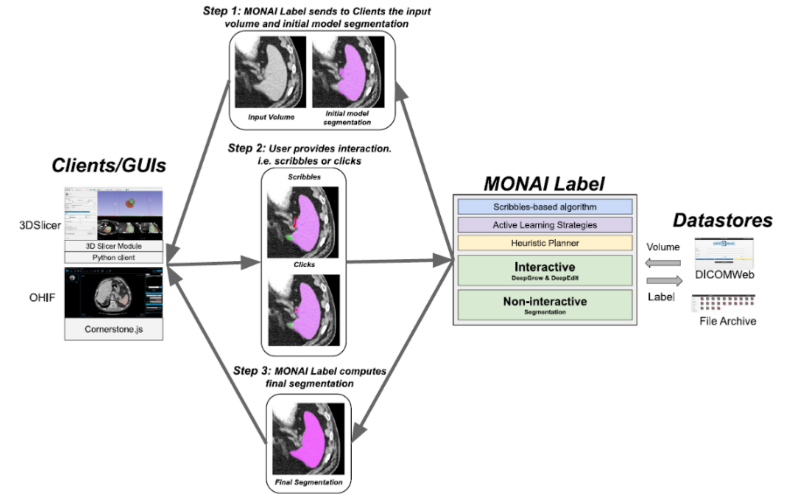
Criterion	ViT (2020)	NesT (2022)	UNesT (2023)
Structure	Flat, non-hierarchical Transformer	Hierarchical nested Transformers (2D)	Hierarchical 3D Transformers with convolutional decoder
Global context	Fully captured from the beginning	Progressively captured across levels	Captured at all hierarchical levels
Local detail fidelity	Low (no pyramid structure)	Moderate (nested aggregation)	Very high (fine resolution, skip connections)
Data requirements	Very high (requires massive pretraining)	Moderate (pretraining helpful but optional)	Moderate (hierarchical structure performs well with limited data)
Evaluated tasks	Classification, basic segmentation	2D/3D multi-organ segmentation	Fine volumetric segmentation (133 structures)
Performance (Dice)	Around 83%	86–88% depending on task	91–93% on complex brain structures
Advantages	Simplicity, global view, few parameters	Good global/local balance, generalizable	High accuracy, context + detail, native 3D support
Limitations	Low spatial resolution, high data demand	Not natively 3D, nested complexity	High computational cost, long training time

5

Interactive Segmentation

(Papers 7 & 8)

Monai Label



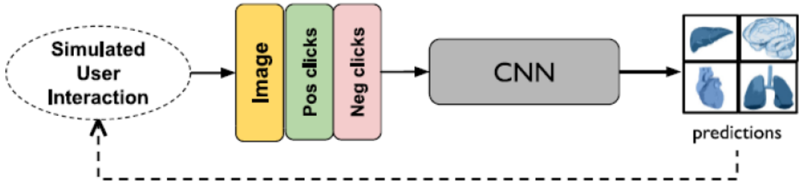
Open-source client-server platform for AI-assisted medical annotation

- Clients (3D Slicer, OHIF web): for expert interaction
- AI Server (MONAI Label): hosts pre-trained segmentation models and supports online learning strategies

Non-interactive approach
(U-Net 2D/3D, HighResNet ou DynUNet)

Interactive approach

❖ DeepGrow
(clicks)



- User clicks (positive/negative) are converted into binary maps added to the input image
- The model refines segmentation dynamically based on these contextual cues

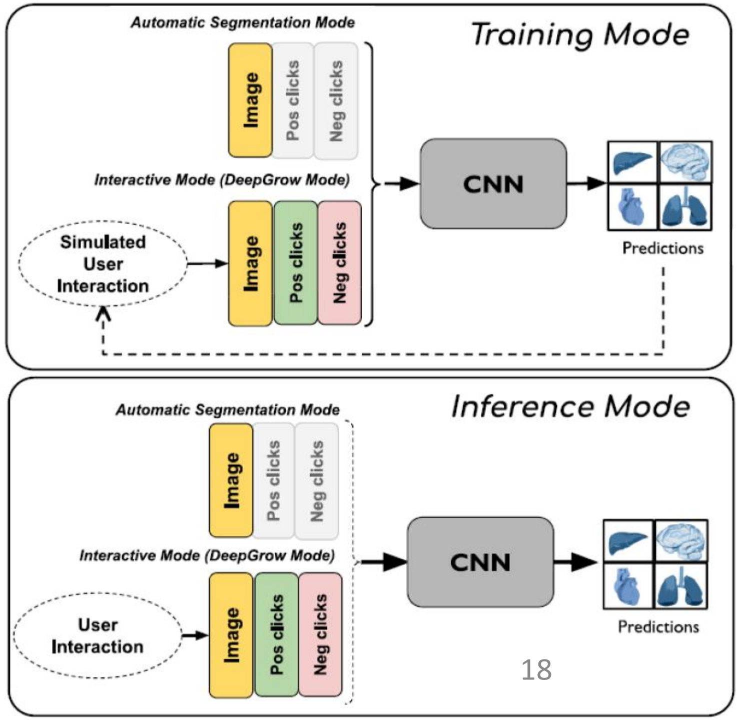
<u>Two-phase process:</u>

1. Auto mask on load (no clicks)
2. User refines with corrective clicks

<u>Dual-mode design:</u>

- Trained with/without clicks
- Switches between auto and interactive modes
- Enables active learning with initial masks

❖ DeepEdit
(auto+clicks)



SimpleClick

Interactive segmentation method based on Vision Transformer without hierarchical backbone

- User Click Encoding: Generate two binary maps from clicks (positive on object, negative on background) plus a prediction map from the previous segmentation.

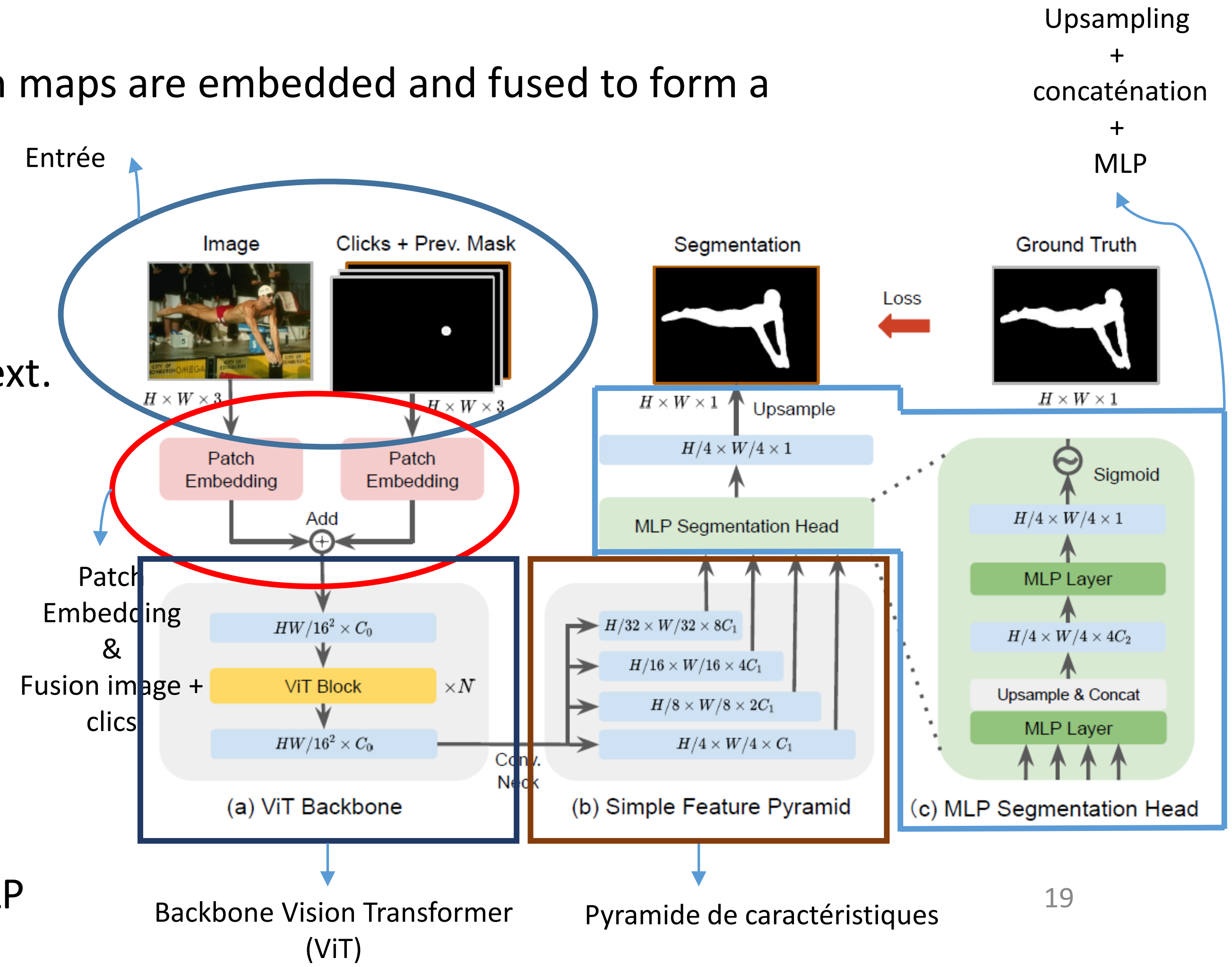
- Symmetric Patch Embedding: Both image and interaction maps are embedded and fused to form a unified ViT input.

- Window-based local attention: Attention is restricted to non-overlapping windows, reducing computational cost while preserving local context.

- Global attention blocks: Introduced at selected layers to model long-range dependencies and ensure global context understanding.

- Multi-scale features generated via parallel convolutions with different strides, followed by upsampling and concatenation.

- Final segmentation map predicted through a unified MLP and sigmoid activation.



Comparative Analysis: MONAI Label (DeepGrow, DeepEdit) vs SimpleClick

Criterion	DeepGrow (MONAI Label)	DeepEdit (MONAI Label)	SimpleClick (2023)
Core principle	Learns from a single seed point	Multiple user-guided corrections	ViT interprets positive/negative clicks
Interaction type	One central click	Multiple clicks (include/exclude)	Positive/negative clicks only
Spatial precision	High, depends on seed accuracy	Good, refined with interaction	Very high (ViT-based attention)
AI architecture	CNN encoder-decoder	CNN with correction-aware decoder	Vision Transformer pre-trained with MAE
Training data	Requires manual segmentations	Same, with local interaction examples	General pretraining + minimal tuning
Clicks vs. quality	3–6 clicks for decent quality	¡5 clicks for 90% Dice	4.15 clicks for 90% IoU
2D / 3D support	Full (2D and 3D)	Full (with Slicer/MITK/OHIF)	Mainly 2D (applied per slice in 3D)
Interaction speed	Good (depends on server)	Fast and interactive	Very fast (lightweight ViT)
Deployment complexity	Medium to high (server-client)	Same as DeepGrow	Moderate (standalone, portable)

6

Automatic Landmark Detection

(Papers 9 & 10)

Automatic Landmark Detection – Why It Matters

Anatomical landmark : A key point defined by the position of a structure (e.g., anterior commissure, etc)

Use cases :
- Image registration (align scans via shared points)
- Biometric measurements (e.g., distances)
- Extraction of standard planes in 3D imaging

Problem : Manual localization is tedious and prone to high inter-operator variability.

Challenges in automation :
- Few annotated datasets (point labeling is as laborious as segmentation)
- Large 3D volumes
- Multiple landmarks to detect simultaneously, requiring anatomical relationship modeling.

Solution :

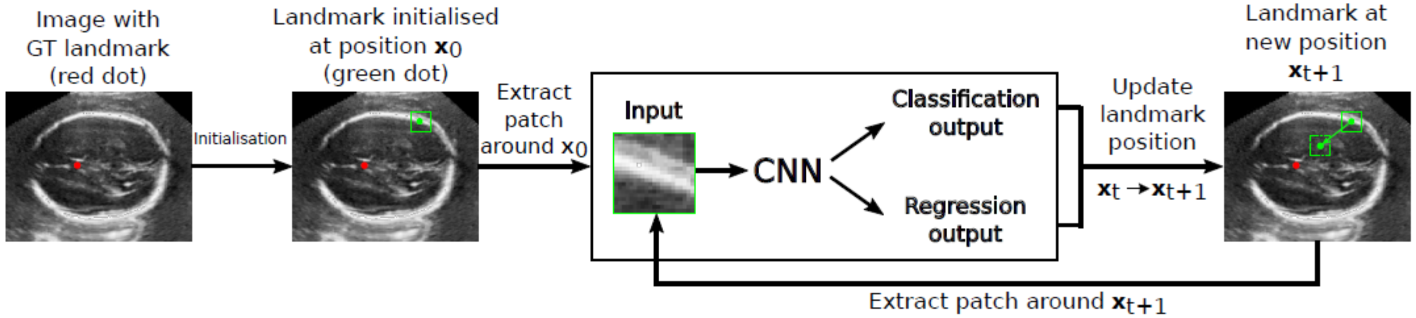
Dedicated deep learning approaches formulating landmark detection as a combined regression + classification problem for each point

PIN – Patch-based Iterative Network

Global-to-Local Landmark Localization

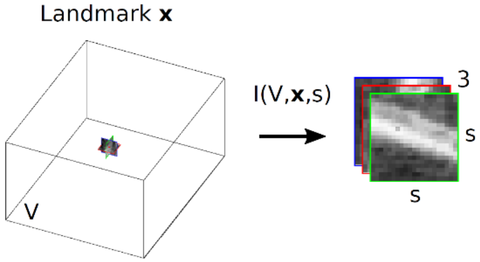
PIN – Patch-based Iterative Network

Formulates landmark localization as an iterative patch-based search.

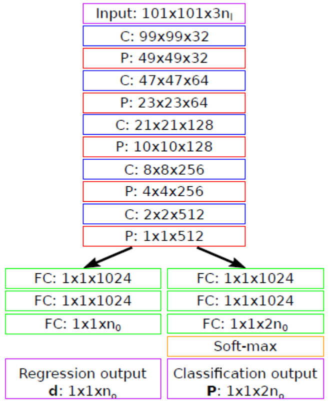


Start from a coarse initial position, then:

- Extract a 2.5D local patch (axial, sagittal, coronal slices centered on current point)



- A CNN predicts:
  - A displacement vector d=(Δx,Δy,Δz) toward the true point
  - A direction class P_{max}(one of ±X, ±Y, ±Z) for the main movement axis
  - Update position : $x_{t+1} = x_t + P_{\max} \cdot d$,repeat until convergence



Multi-landmark extension

- All landmark coordinates are projected into a lower-dimensional space (PCA)

$$b = W^T(X - \bar{X}) \qquad b_{t+1} = b_t + P_{\max} \cdot d_b$$

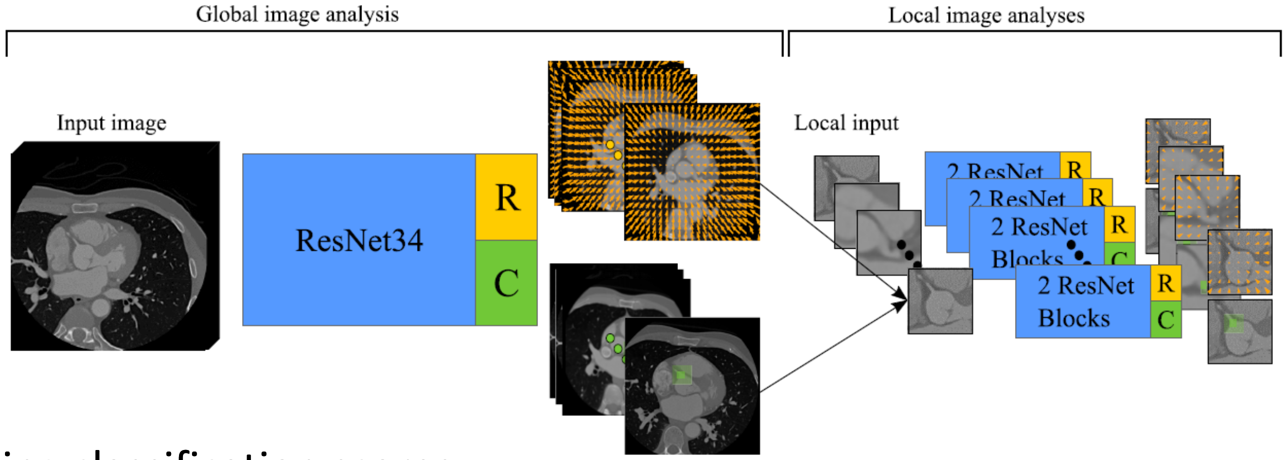
Enables the CNN to predict a **compact global vector** for all points

Global-to-Local Landmark Localization

Formulates landmark localization as an iterative patch-based search.

Global Stage :

- A fully convolutional neural network (FCNN) analyzes the entire image in a patch-based manner and predicts:
  - A displacement vector from the patch center to the landmark
  - A presence probability for each landmark (classification)



Final position: weighted average of displacement vectors using classification scores

Local Stage :

$$\hat{p} = \frac{\sum_{i=1}^{N} s_i \cdot (c_i + d_i)}{\sum_{i=1}^{N} s_i}$$

A specialized local FCNN analyzes a subvolume around each landmark to refine its position, again combining regression and classification

- For each landmark, extract a local region around the global estimate
- Apply a small local network to refine the coordinates through fine-grained search

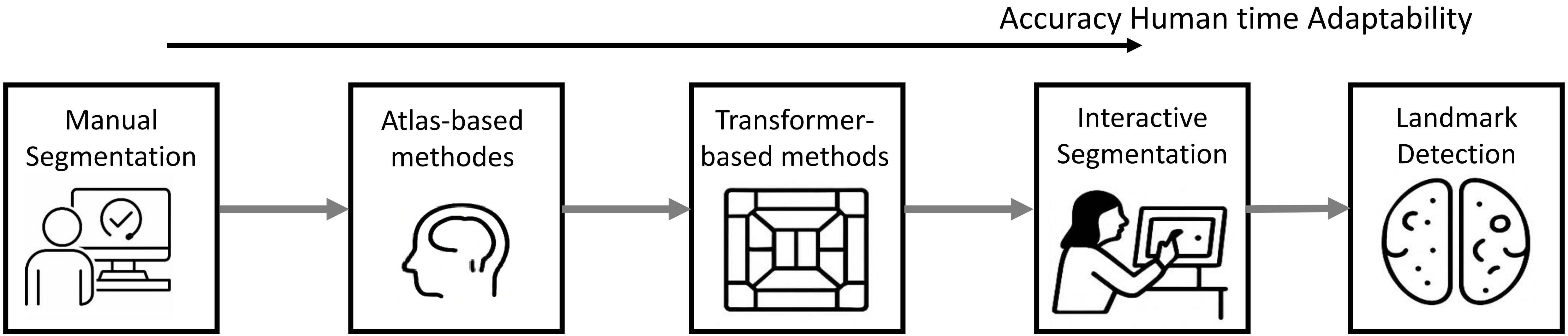
Performance Metrics for PIN vs Global-to-Local Approaches

Metric / Criterion	PIN (Li et al., 2018)	Global-to-Local (Noothout et al., 2020)
Mean localization error	5.59 mm (10 fetal brain landmarks)	2.0 mm (2D/3D datasets, expert-level)
Number of landmarks	10 localized jointly	Up to 19 localized jointly
Success rate (¡3 mm)	Not reported	95% of landmarks within 3 mm
Inference time per volume	0.44 seconds (entire 3D volume)	Slower (global + local passes per point)
Initialization requirement	Requires approximate patch-centered input	No manual initialization needed
Computational cost	Low (few patches per point)	High (full volume + refinements)

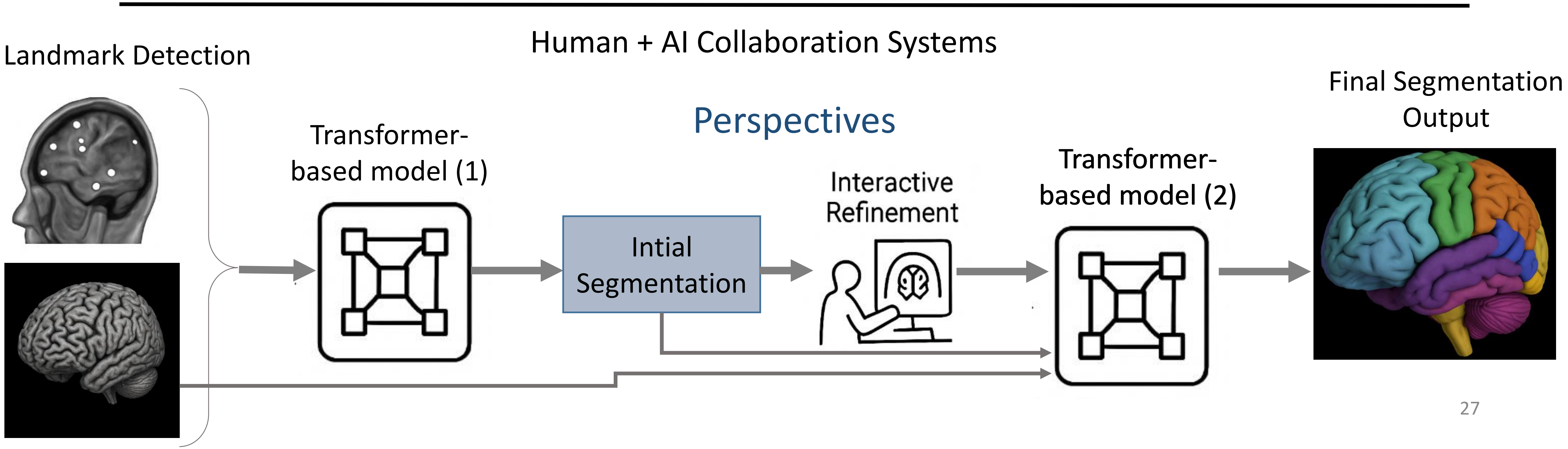
7

Conclusion

Conclusion



Perspectives



Thank you for your attention!