Cross-Modal Knowledge Distillation for PET-Free Amyloid-Beta Detection from MRI

Supplementary Material

Francesco Chiumento[1] Julia Dietlmeier[1,2] Ronan P. Killeen[3,4]
Kathleen M. Curran[2,4] Noel E. O'Connor[1,2] Mingming Liu[1,2]
[1]Dublin City University, Ireland [2]Insight Research Ireland Centre for Data Analytics, Ireland
[3]St. Vincent's University Hospital, Ireland [4]University College Dublin, Ireland
francesco.chiumento2@mail.dcu.ie

1. List of Abbreviations

Table 1 summarizes the main abbreviations used throughout the main paper and supplementary material.

Table 1. **Abbreviations used throughout the main paper and supplementary material.**

Abbr.	Full Form	Abbr.	Full Form
Medical & Imaging			
Aβ	Amyloid-β	OASIS-3	Open Access Series of Imaging Studies 3 [8]
AD	Alzheimer's disease	PET	Positron Emission Tomography
ADNI	Alzheimer's Disease Neuroimaging Initiative [1]	PiB	Pittsburgh Compound B [15]
CL	Centiloid	PUP	PET Unified Pipeline [12]
CSF	Cerebrospinal Fluid	SUVR	Standardized Uptake Value Ratio
FLAIR	Fluid-Attenuated Inversion Recovery	T1w	T1-weighted
T2w	T2-weighted	T2*	T2*-weighted
FWHM	Full Width at Half Maximum	NFTs	Neurofibrillary Tangles
MRI	Magnetic Resonance Imaging	APOE4	Apolipoprotein E ε4 allele
APOE	Apolipoprotein E	AV-45	Florbetapir (^{18}F-AV-45)
ATN	Amyloid–Tau–Neurodegeneration framework	SWI	Susceptibility-Weighted Imaging
ARIA	Amyloid-Related Imaging Abnormalities	ICBM152	MNI ICBM152 brain template [4]
DTI	Diffusion Tensor Imaging	EHR	Electronic Health Record
GRE	Gradient-Recalled Echo		
Deep Learning & Architecture			
BCE	Binary Cross-Entropy	MLP	Multi-Layer Perceptron
CLS	Classification Token	ReLU	Rectified Linear Unit
GELU	Gaussian Error Linear Unit	ViT	Vision Transformer
KD	Knowledge Distillation	MHA	Multi-Head Attention
LoRA	Low-Rank Adaptation [5]	CDKD	Cross-Dataset Knowledge Distillation
CLIP	Contrastive Language–Image Pre-training	BiomedCLIP	Biomedical CLIP vision–language model
Preprocessing & Tools			
ANTs	Advanced Normalization Tools [14]	MNI	Montreal Neurological Institute [4]
HD-BET	HD Brain Extraction Tool [6]	N4	N4 Bias Field Correction [13]
Evaluation Metrics			
Acc	Accuracy	F1	F1 score
AUC	Area Under the ROC Curve	Prec	Precision
Rec	Recall	NPV	Negative Predictive Value
CI	Confidence Interval	ROC	Receiver Operating Characteristic
Training			
AdamW	Adam with Weight Decay [9]	lr	Learning Rate
FP16	16-bit Floating Point	wd	Weight Decay

2. Supplementary Material Overview

In this section, we provide technical details and additional experiments that support the main paper. In Sec. 3, we present further implementation details, including augmentation strategies, hyperparameters for all three training phases, and architecture-specific formulas referenced in the main text. In Sec. 4, we report additional qualitative and quantitative results: radar charts to visualize performance trade-offs across metrics and to compare single-sequence models with multi-sequence distilled models (Fig. 1); ablation studies evaluating different Centiloid thresholds for negative patient selection during triplet mining (Sec. 4.2, Table 3); ROC curve evolution from the first to the last epoch (Sec. 4.3, Fig. 2); and interpretability analyses (Sec. 4.4) for both single- and multi-sequence models via gradient-based saliency [11] and HiResCAM [3], which confirm that the model's attention focuses on anatomically plausible regions (Figs. 3–4).

3. Implementation Details

3.1. Data Augmentation

To improve generalization while preserving PET–MRI spatial correspondence and to facilitate reproducibility, we apply the following synchronized augmentations during Phases 1–2 using shared random seeds for each PET–MRI pair:

- **Spatial transformations:** random affine (rotation $\pm 7°$), translation $\pm 5\%$, and isotropic scaling in $[0.95, 1.05]$.
- **Intensity modulations:** color jitter (brightness/contrast $\pm 10\%$), Gaussian blur (kernel size 3, $p = 0.3$), gamma correction $\gamma \in [0.9, 1.1]$ ($p = 0.5$), and Gaussian noise $\sigma \in [0.01, 0.03]$ ($p = 0.5$).
- **Random erasing:** $p = 0.25$, erase scale in $[0.05, 0.12]$.

The same augmentations are applied during Phase 3, with PET–MRI synchronization preserved (shared random seeds). At test time, only resizing and normalization are used.

Framework & hardware. We use PyTorch 2.x with CUDA 12.x on NVIDIA GeForce RTX 4090/5090 GPUs (24–32 GB VRAM).

3.2. Training Configuration

We use PyTorch with seed $= 42$, mixed-precision training (FP16), and gradient accumulation to increase the effective

batch size. Hyperparameters are tuned on the validation set and then kept fixed for all experiments. Early stopping is based on validation F1 (Phases 1 and 3) and on a combined score (triplet separation + F1) in Phase 2.

Class Balancing and Sampling To mitigate label imbalance, we use per-class inverse-frequency weights and a weighted sampler. Let n_0, n_1 be the counts of negative/positive samples in the training set; we set per-class weights $w_c = 1/n_c$ and assign each sample the weight of its class. A `WeightedRandomSampler` (replacement) is used to draw mini-batches with balanced label proportions. We do not apply additional positive reweighting in the BCE (i.e., `pos_weight` $= 1.0$) to avoid double-counting the imbalance already handled by sampling.

Reproducibility. Fixed seed (42) for all stochastic operations (data loading, augmentation, model initialization, dropout); CuDNN in deterministic mode with benchmarking disabled.

Training: Mixed precision with `GradScaler` and gradient accumulation; batch sizes 6 (Phases 1–2), 10 (Phase 3).

Early Stopping:

- Phase 1 (Pre-training): validation F1, patience 5
- Phase 2 (Contrastive): validation combined score (triplet separation + F1), patience 3
- Phase 3 (Distillation): validation F1, patience 25

Data loading & checkpointing: 4 DataLoader workers with pin memory enabled. Phase 1 (pre-training): best validation F1. Phase 2 (teacher): best *combined score* (triplet separation + classification F1, weight 1:0.5). Phase 3 (student): best teacher–student similarity and best validation F1 (saved separately).

3.3. Architecture Implementation Details

3.3.1. LoRA Parameterization

Low-Rank Adaptation (LoRA) [5] reparameterizes each projection with pretrained weights $\mathbf{W}_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ as:

$$\mathbf{W}' = \mathbf{W}_0 + s\,\Delta\mathbf{W}, \quad \Delta\mathbf{W} = \mathbf{B}\mathbf{A}, \quad s = \alpha/r, \tag{1}$$

where $\mathbf{B} \in \mathbb{R}^{d_{\text{out}} \times r}$ and $\mathbf{A} \in \mathbb{R}^{r \times d_{\text{in}}}$ are the only trainable parameters, with rank $r = 32$ and scaling factor $\alpha = 32$ (so $s = \alpha/r = 1.0$), while $\mathbf{W}_0$ remains frozen. We rely on the standard LoRA initialization from the PEFT library: the up-projection $\mathbf{B}$ is initialized to zero so that $s\,\mathbf{B}\mathbf{A} = \mathbf{0}$ at the beginning of training, and $\mathbf{W}' = \mathbf{W}_0$. LoRA adapters are applied to attention projections (query, key, value, output) in transformer blocks 6–11 (0-indexed).

3.3.2. Projection Head Architecture

ViT CLS tokens (768D) are projected to 128D using the dropout rates specified in Table 2:

$$\begin{aligned} \mathbf{h}^{(1)} &= \text{Dropout}_1\big(\text{GELU}(\text{LN}_1(\mathbf{W}_1\mathbf{h}))\big), \\ \mathbf{e} &= \text{Dropout}_2\big(\text{LN}_2(\mathbf{W}_2\mathbf{h}^{(1)})\big). \end{aligned} \tag{2}$$

where $\mathbf{W}_1 : 768 \to 256$ and $\mathbf{W}_2 : 256 \to 128$.

3.3.3. Attention Pooling Formula

Given slice embeddings $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_S] \in \mathbb{R}^{S \times 128}$ after MHA (4 heads, $d_{\text{head}} = 32$), with S slices, we compute:

$$w_s = \mathbf{W}_{\text{pool}}\mathbf{A}_s + b, \quad \alpha_s = \text{softmax}(w_s/\tau), \tag{3}$$

where $\mathbf{W}_{\text{pool}} \in \mathbb{R}^{1 \times 128}$, $b \in \mathbb{R}$ is a scalar bias term, and $\tau = 2.0$. The patient representation is $\mathbf{e} = \sum_{s=1}^{S} \alpha_s\,\mathbf{A}_s$.

3.3.4. Classification Head Architecture

Both teacher and student use:

$$z = \mathbf{W}_2(\text{Dropout}(\text{ReLU}(\mathbf{W}_1\mathbf{e}))), \tag{4}$$

with $\mathbf{W}_1 : 128 \to 64$, $\mathbf{W}_2 : 64 \to 1$.

Weight initialization: the attention pooling weight uses Xavier-uniform (gain = 1.0) with bias 0.0; other linear layers use PyTorch defaults. Before student distillation, the classifier output layer ($64 \to 1$) is re-initialized with Xavier-uniform (gain = 0.5) and bias 0.0.

Table 2. **Dropout schedules.** Dropout rates for projection and classification heads in teacher (Phases 1–2) and student (Phase 3).

	Phases 1–2 (Teacher)	**Phase 3 (Student)**
Projection head (p_1, p_2)	$(0.5,\ 0.4)$	$(0.3,\ 0.2)$
Classification head p	0.6	0.4

Phase 1 - Complete Hyperparameters

- Epochs: 30
- Batch size: 6
- Optimizer: AdamW with lr $= 2 \times 10^{-5}$, wd $= 1 \times 10^{-3}$ (uniform)
- Slices per subject: 25 (uniformly spaced)
- Loss: BCE (no label smoothing)
- Gradient clipping: max_norm dynamically adjusted (1.0 for epochs 1–2, 2.0 for epochs 3–5, 5.0 thereafter)

Phase 2 - Complete Hyperparameters

- Epochs: 15
- Batch size: 6
- Optimizer: AdamW with component-specific rates:
  - Vision backbone, projection, attention: lr $= 5 \times 10^{-6}$, wd $= 10^{-2}$
  - Classification head: lr $= 2 \times 10^{-5}$, wd $= 5 \times 10^{-3}$
- Scheduler: CosineAnnealingWarmRestarts with $T_0 = 5$, $T_{\text{mult}} = 2$, $\eta_{\min} = 10^{-7}$

Phase 2 Regularization $\mathcal{L}_{\text{reg}}$ includes three components: (i) ℓ_2 penalty on anchor, positive, and negative embedding norms with coefficient 0.01; (ii) inter-anchor similarity

penalty, penalizing mean pairwise cosine similarity above 0.5; (iii) anchor-negative similarity penalty (weight 0.5), penalizing mean similarity above -0.1. Components (ii–iii) use progressive epoch-dependent scaling: $s = 0.1$ for epochs 1–3, then linearly increasing to 1.0 by epoch 13.

MarginFocal Loss Details (Phase 3) Complete hyperparameters: $\gamma = 2.0$ (focal parameter), $w = 1.0$ (pos_weight for balanced batches), label smoothing $= 0$, $\varepsilon = 10^{-8}$ (numerical stability floor).

The positive-weighted BCE is: $\mathrm{BCE}_w(\tilde{z}, y') = -wy' \log \sigma(\tilde{z}) - (1 - y') \log(1 - \sigma(\tilde{z}))$.

Margin annealing schedule:

- Epochs 1–6: $m = 0.3$
- Epochs 7–20: m is linearly increased from 0.3 towards the final value
- Epochs 21+: $m = 1.2$ (fixed)

Gap deficit scaling: The term $[\, m - (\bar{z}_+ - \bar{z}_-) \,]_+$ is multiplied by 0.1 internally before adding to the loss, yielding effective $\lambda_{\text{gap}} \approx 0.01$ (epochs 1–10) and 0.03 (epochs 11+).

Phase 3 - Complete Hyperparameters

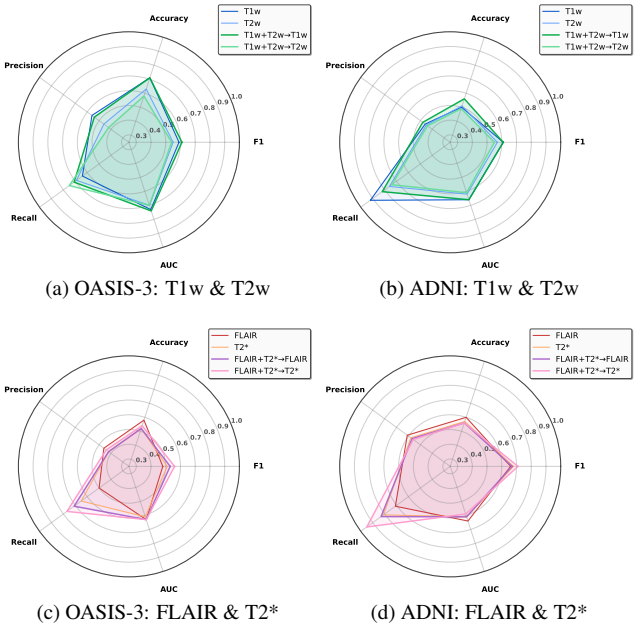
- Epochs: 100
- Batch size: 10
- Optimizer: AdamW with component-specific rates:
  - LoRA adapters: lr $= 2 \times 10^{-4}$, wd $= 0$ (preserve low-rank structure)
  - Projection modules: lr $= 1 \times 10^{-4}$, wd $= 1 \times 10^{-4}$
  - Attention modules: lr $= 1 \times 10^{-4}$, wd $= 1 \times 10^{-3}$
  - Classification head: lr $= 1 \times 10^{-4}$, wd $= 1 \times 10^{-3}$
- Scheduler: ReduceLROnPlateau (mode=max, factor=0.7, patience=15, threshold=0.005, `min_lr` $= 10^{-5}$, cooldown=2; monitoring validation F1)
- Temperature annealing: $T = 2.5$ (epochs 1–6), linearly decreased (epochs 7–20) towards $T = 1.0$, and fixed at $T = 1.0$ (epochs 21+)

Knowledge Distillation Loss Weights Warm-up During Phase 3, we linearly warm up the loss weights over the first 10 epochs from $(\lambda_{\text{cls}}, \lambda_{\text{feat}}, \lambda_{\text{logit}}) = (0.3, 0.5, 0.2)$ to $(0.4, 0.4, 0.2)$. After epoch 10, the weights are kept fixed at $(\lambda_{\text{cls}}, \lambda_{\text{feat}}, \lambda_{\text{logit}}) = (0.4, 0.4, 0.2)$.

4. Additional Experimental Results

4.1. Multi-Contrast Performance Analysis

Fig. 1 compares models that use a single sequence with models that use multiple sequences and are then tested on a single sequence using five metrics (F1, Accuracy, Precision, Recall, AUC). Multi-sequence models consistently improve recall (e.g., T1w on OASIS-3: 0.64 → 0.71, +10.9%), while also increasing AUC on OASIS-3 (0.73 → 0.74) and accuracy for T1w on ADNI (0.50 → 0.56), and maintaining comparable precision and AUC in both cohorts. On OASIS-3 (Fig. 1c), FLAIR+T2* → T2* achieves the largest F1 improvement (0.51 → 0.56, +9.8%) with substantial recall gains (0.65 → 0.77). In ADNI, the recall gain is larger; in particular, FLAIR+T2* → T2* reaches F1 0.71 with recall 0.95. Across all four charts, multi-sequence distillation yields higher recall while maintaining or improving the other metrics, suggesting that using multiple sequences provides the model with richer context when predicting from a single sequence and is therefore preferable in settings where minimizing false negatives is critical.



(a) OASIS-3: T1w & T2w (b) ADNI: T1w & T2w

(c) OASIS-3: FLAIR & T2* (d) ADNI: FLAIR & T2*

Figure 1. **Single vs Multi-Sequence Performance Comparison.** Spider charts comparing single-sequence and multi-sequence distilled models across metrics (F1, Accuracy, Precision, Recall, AUC). Top: T1w+T2w; bottom: FLAIR+T2*. Left: OASIS-3; right: ADNI. Multi-sequence distillation improves recall.

4.2. Ablation Study: Impact of Centiloid-Based Negative Mining

To assess the contribution of the Centiloid-guided negative mining strategy used in Phase 2, we vary the minimum amyloid burden difference threshold $\Delta_{\text{CL}}^{\min}$ required between anchor and negative samples (Table 3) and evaluate three configurations for both T1w and T2w sequences:

- **Uniform ($\Delta_{\text{CL}}^{\min} = 0$):** Negatives sampled uniformly, representing a baseline where triplet learning relies on visual similarity.
- **Moderate ($\Delta_{\text{CL}}^{\min} = 5.0$):** configuration used in the main experiments, requiring negatives to differ by at least 5 CL units from the anchor.
- **Strict ($\Delta_{\text{CL}}^{\min} = 10.0$):** more restrictive threshold enforcing larger amyloid burden differences, yielding harder

Table 3. **Ablation study: impact of Centiloid-based negative mining.** Test performance on OASIS-3 and ADNI for T1w and T2w models when varying the minimum Centiloid gap $\Delta_{\text{CL}}^{\min}$ between anchor and negative samples. We report metrics both at a fixed decision threshold of 0.5 and at the validation-optimized threshold θ^*.

	OASIS-3											ADNI										
Sequence	$\Delta_{\text{CL}}^{\min}$	@0.5				@θ^*				AUC	NPV	$\Delta_{\text{CL}}^{\min}$	@0.5				@θ^*				AUC	NPV
		F1	Acc	Prec	Rec	F1	Acc	Prec	Rec				F1	Acc	Prec	Rec	F1	Acc	Prec	Rec		
T1-weighted MRI																						
T1w	0.0	0.53	0.49	0.38	0.87	0.53	0.68	0.51	0.56	0.72	0.77	0.0	0.60	0.63	0.56	0.66	0.60	0.60	0.52	0.71	0.65	0.71
T1w	5.0	0.53	0.71	0.57	0.49	0.59	0.71	0.56	0.64	0.73	0.81	5.0	0.58	0.56	0.48	0.74	0.61	0.50	0.46	0.92	0.66	0.77
T1w	10.0	0.53	0.65	0.47	0.60	0.52	0.63	0.46	0.60	0.68	0.77	10.0	0.59	0.60	0.52	0.68	0.58	0.53	0.47	0.76	0.65	0.68
T2-weighted MRI																						
T2w	0.0	0.47	0.66	0.49	0.46	0.55	0.60	0.44	0.75	0.69	0.81	0.0	0.55	0.53	0.46	0.68	0.56	0.51	0.45	0.74	0.59	0.64
T2w	5.0	0.54	0.54	0.40	0.80	0.55	0.63	0.46	0.69	0.70	0.80	5.0	0.51	0.63	0.59	0.45	0.57	0.51	0.45	0.76	0.62	0.65
T2w	10.0	0.55	0.61	0.44	0.71	0.55	0.61	0.44	0.71	0.70	0.80	10.0	0.52	0.59	0.51	0.53	0.58	0.51	0.46	0.79	0.60	0.67

negatives but reducing the number of eligible triplets.

Analysis of Results Table 3 shows that using a specific Centiloid gap constraint value ($\Delta_{\text{CL}}^{\min} > 0$) is beneficial compared to using a uniformly selected negative patient. The choice $\Delta_{\text{CL}}^{\min} = 5.0$ provides the best trade-off between sensitivity and specificity across datasets, with better calibration of the metrics. For T1w on OASIS-3, the threshold of 5.0 improves the F1 score from 0.53 to 0.59 (+11.3%) and increases NPV from 0.77 to 0.81 compared to ($\Delta_{\text{CL}}^{\min} = 0$), reducing the number of missed amyloid-positive cases. The recall improvement (0.56→0.64) implies a relative reduction of approximately 18% in the number of false negatives in the test set. The **Strict** strategy ($\Delta_{\text{CL}}^{\min} = 10.0$) underperforms on OASIS-3 T1w (F1 = 0.52, AUC = 0.68), suggesting that using too high a Centiloid value prevents the network from learning smaller differences between patients that are not too dissimilar, as obtained with $\Delta_{\text{CL}}^{\min} = 5.0$. For T2w on OASIS-3, all three strategies lead to similar F1 scores, but using a moderate $\Delta_{\text{CL}}^{\min}$ achieves an AUC of 0.70 (similar to the strict configuration and higher than the uniform baseline at 0.69), indicating better calibration.

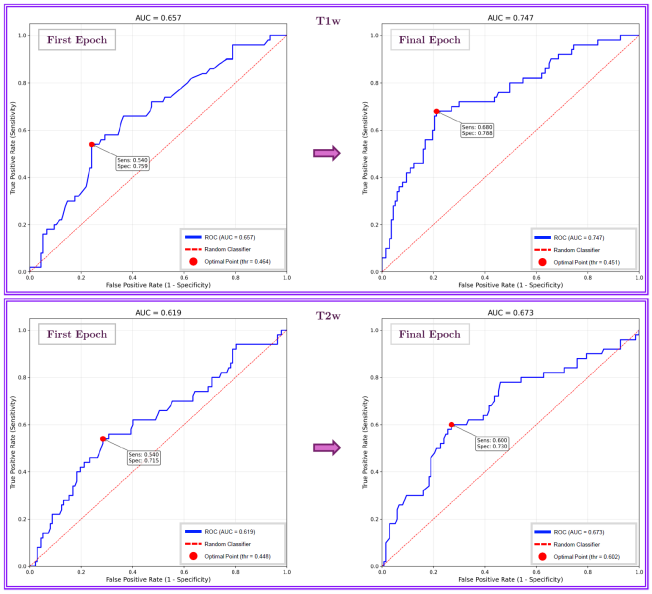
On ADNI, the trends are similar but less pronounced. For T1w, using the Moderate configuration (F1 = 0.61, AUC = 0.66) marginally outperforms both Uniform (F1 = 0.60, AUC = 0.65) and Strict (F1 = 0.58, AUC = 0.65). A similar pattern is observed for T2w, with Moderate achieving F1 = 0.57 and AUC = 0.62. We also note that the Uniform baseline remains competitive on ADNI (e.g., T1w: F1 = 0.60), likely due to its more balanced class distribution, which reduces the risk of trivial negatives. The smaller performance gap on ADNI suggests that CL-aware mining provides greater benefits in imbalanced settings. These results validate our choice of $\Delta_{\text{CL}}^{\min} = 5.0$ as the optimal balance: it enforces biochemically meaningful separation by requiring negatives to differ by at least 5 Centiloid units from the anchor, a threshold that exceeds both the test-retest measurement error (2.5–3.5 CL) and the reliable annual amyloid accumulation rate (3–5 CL/year) [2], while maintaining sufficient triplet diversity for effective contrastive learning. Across both datasets and contrasts, the Moderate setting either matches or outperforms the Uniform and Strict strategies, with the largest absolute gains observed on OASIS-3.

4.3. Training Convergence Analysis

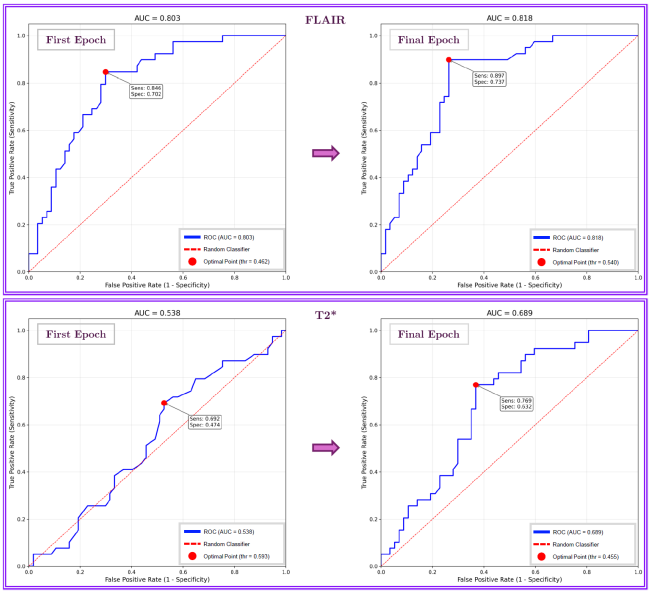
Fig. 2 visualizes the evolution of the ROC curves from the first epoch (Epoch 1) to the last epoch (Final Epoch) on the validation set. We plot the final training epoch rather than the early-stopping checkpoint. All models show positive ΔAUC improvements from the first to the final epoch, confirming effective knowledge distillation. OASIS-3 achieves superior results in terms of AUCs compared to ADNI, possibly due to differences in cohort composition, image acquisition protocols, or amyloid distribution across datasets [7, 10].

4.4. Interpretability

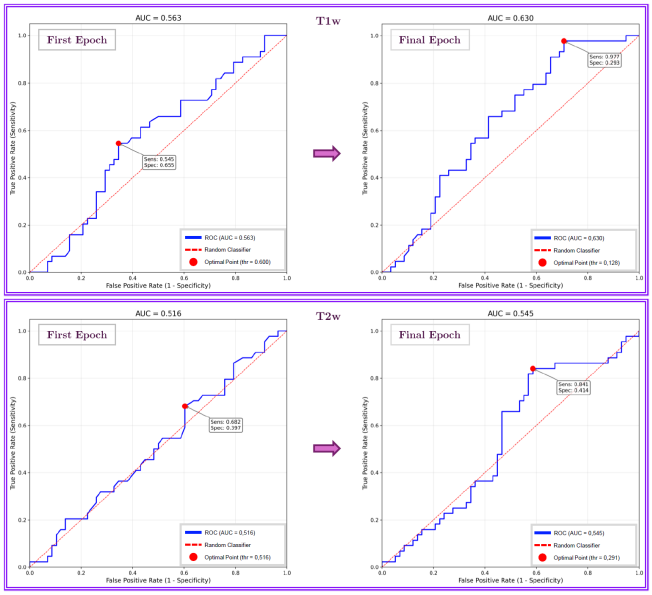
Figs. 3 and 4 show how the model's spatial attention evolves during training. We visualize three epochs (1, 8, and 25) for both single-sequence models (T1w, T2w, FLAIR, T2*) and multi-sequence models (T1w+T2w and FLAIR+T2*), each tested on an individual contrast. For each dataset (OASIS-3 and ADNI) and configuration, we display the target PET image with gradient-based saliency maps and HiResCAM explanations. At epoch 1, both saliency and HiResCAM are relatively diffuse across the brain volume, indicating that the networks initially rely on non-specific global patterns. By epoch 8, the model's attention becomes more structured and begins to concentrate on regions that more closely correspond to the reference PET signal. By epoch 25, the maps are more focal, highlighting neuroanatomically plausible regions. Qualitatively, the attention patterns are consistent between OASIS-3 and ADNI, suggesting that the learned features capture generalizable amyloid-related patterns rather than dataset-specific artifacts.



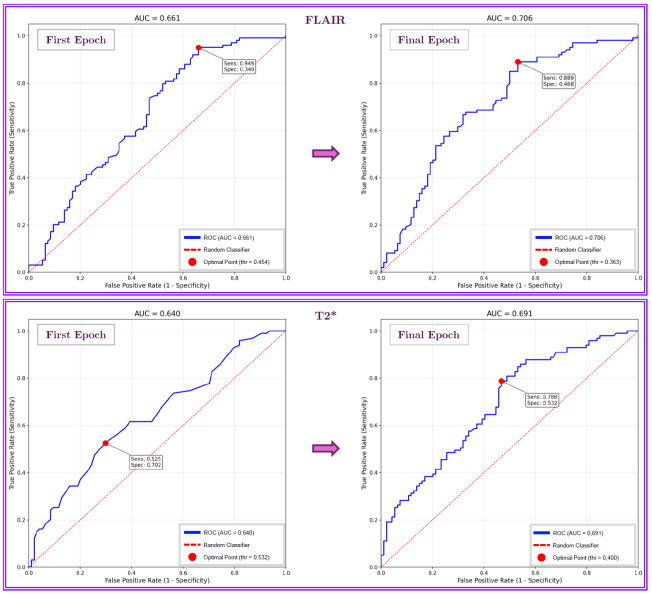
(a) OASIS-3: T1w and T2w



(b) OASIS-3: FLAIR and T2*

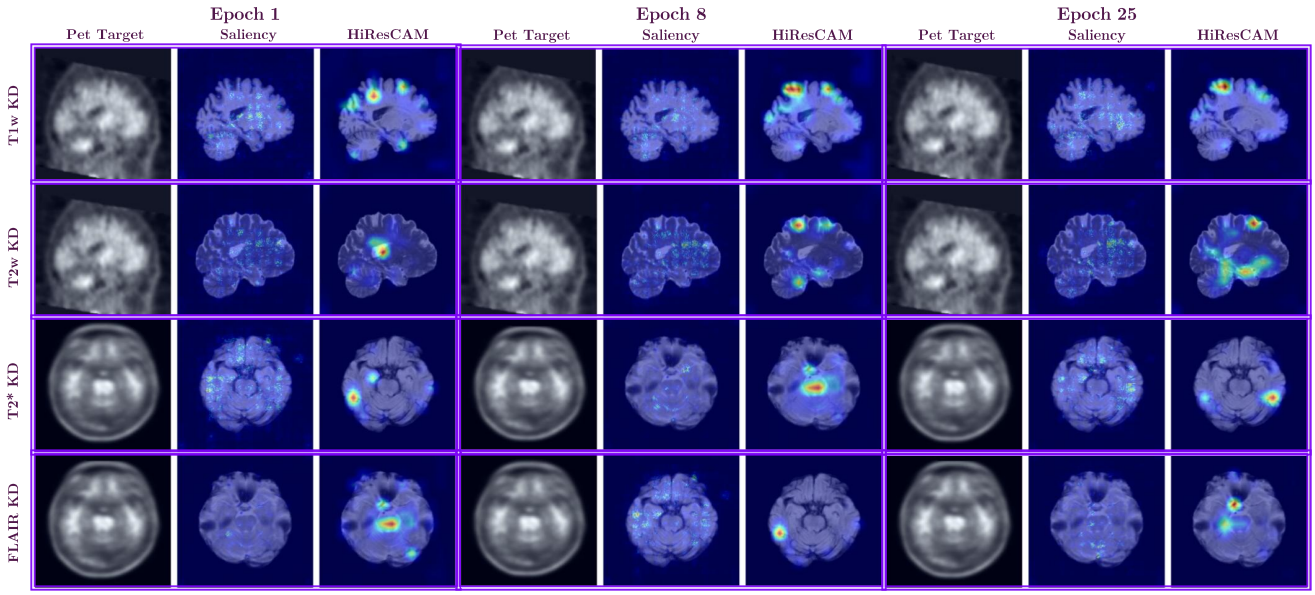


(c) ADNI: T1w and T2w

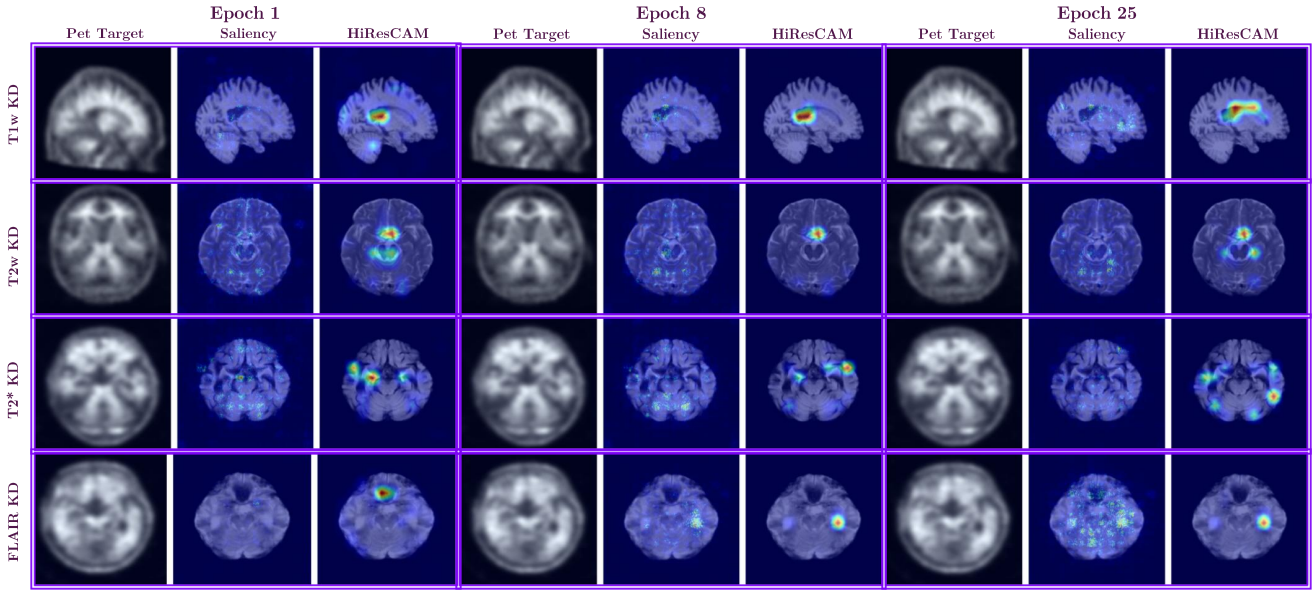


(d) ADNI: FLAIR and T2*

Figure 2. **ROC curve evolution across datasets and sequences.** Validation performance at Epoch 1 (initialization) vs. Final Epoch (convergence). Top row: OASIS-3 dataset for T1w+T2w training (left) and FLAIR+T2* training (right). Bottom row: ADNI dataset with same training configurations. All models show substantial AUC improvements demonstrating effective knowledge distillation.

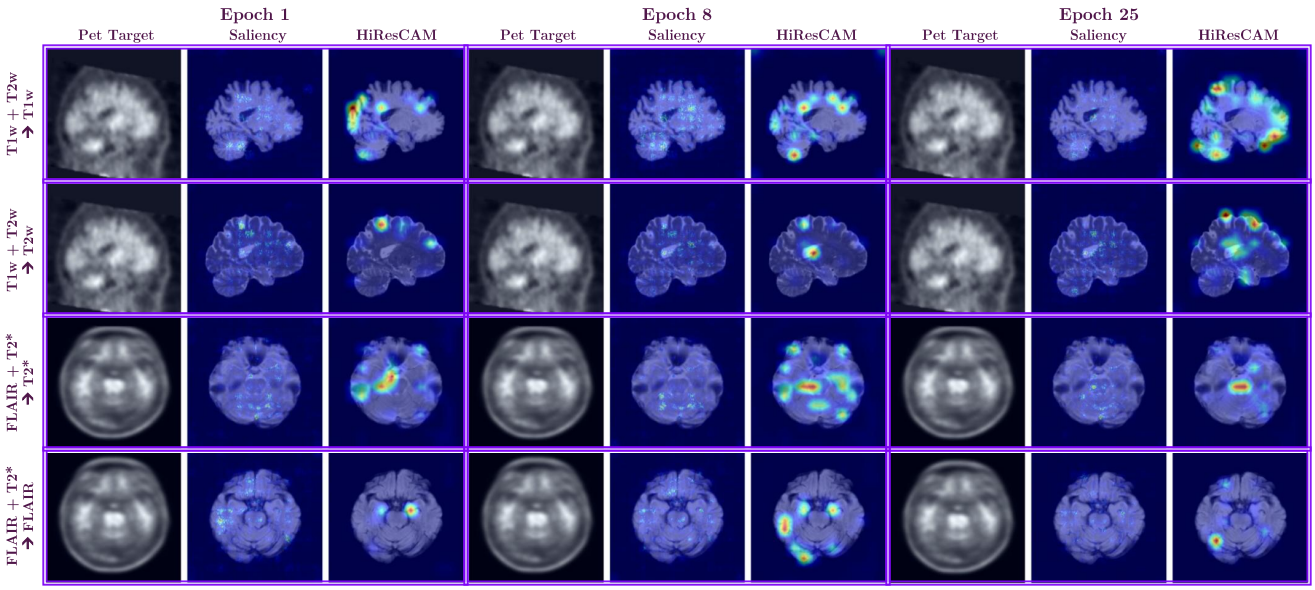


(a) OASIS-3 dataset

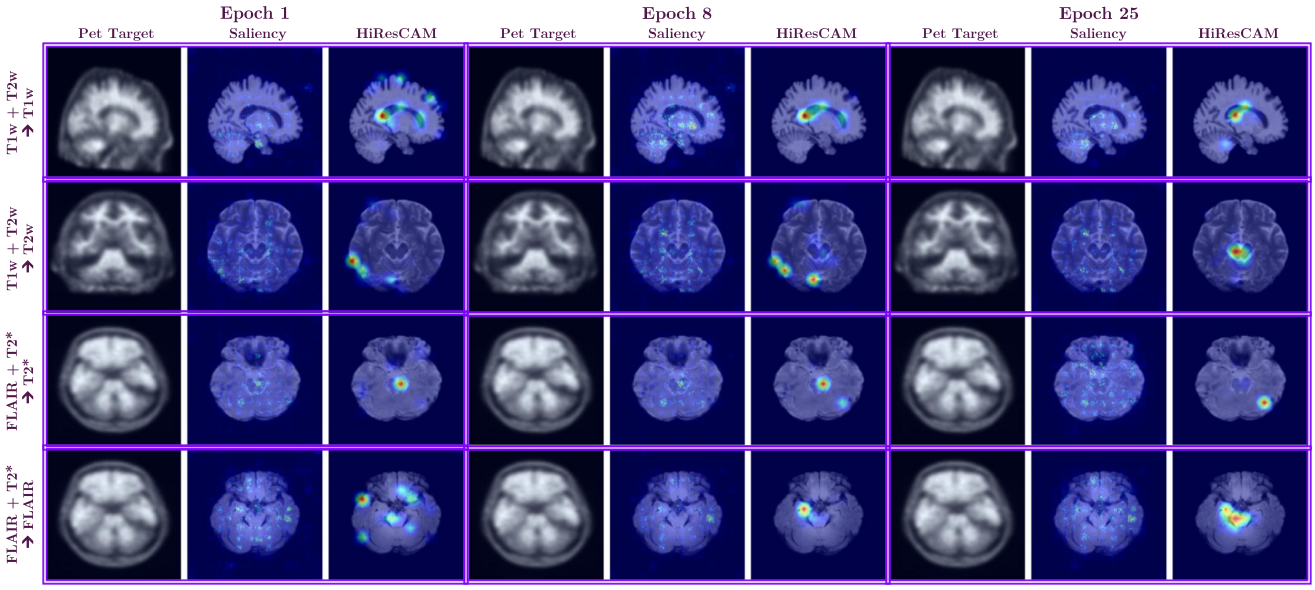


(b) ADNI dataset

Figure 3. **Single-sequence training: attention evolution across datasets.** Training progression (epochs 1, 8, 25) for models trained on individual MRI contrasts. Each row displays PET reference, target MRI, gradient saliency, and HiResCAM maps. The network progressively focuses on anatomically relevant brain structures, with consistent patterns across OASIS-3 and ADNI datasets demonstrating robust generalization.



(a) OASIS-3 dataset



(b) ADNI dataset

Figure 4. **Multi-sequence training with single-sequence inference.** Saliency/HiResCAM evolution (epochs 1, 8, 25) for models trained on paired sequences (T1w+T2w, FLAIR+T2*) and tested on individual contrasts, showing consistent spatial attention patterns across modalities.

References

[1] Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimer's disease neuroimaging initiative. Public dataset, https://adni.loni.usc.edu/, 2025. 1

[2] Lyduine E. Collij, Ariane Bollack, Renaud La Joie, Mahnaz Shekari, Santiago Bullich, Núria Roé-Vellvé, Norman Koglin, Aleksandar Jovalekic, David Valléz Garciá, Alexander Drzezga, Valentina Garibotto, Andrew W. Stephens, Mark Battle, Christopher Buckley, Frederik Barkhof, Gill Farrar, Juan Domingo Gispert, and Amypad Consortium. Centiloid recommendations for clinical context-of-use from the AMYPAD consortium. *Alzheimer's & Dementia*, 20(12):9037–9048, 2024. 4

[3] Rachel Lea Draelos and Lawrence Carin. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020. 1

[4] Vladimir S. Fonov, Alan C. Evans, Robert C. McKinstry, C. Robert Almli, and D. Louis Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102, 2009. 1

[5] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. 1, 2

[6] Fabian Isensee, Marianne Schell, Irada Tursunova, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, Martin Bendszus, Klaus Hermann Maier-Hein, and Philipp Kickingereder. Automated brain extraction of multi-sequence MRI using artificial neural networks. *Human Brain Mapping*, 40(17):4952–4964, 2019. 1

[7] Clifford R. Jack, Arvin Arani, Bret J. Borowski, Dave M. Cash, Karen Crawford, Sandhitsu R. Das, Charles DeCarli, Evan Fletcher, Nick C. Fox, Jeffrey L. Gunter, Ranjit Ittyerah, Danielle J. Harvey, Neda Jahanshad, Pauline Maillard, Ian B. Malone, Talia M. Nir, Robert I. Reid, Denise A. Reyes, Christopher G. Schwarz, Matthew L. Senjem, David L. Thomas, Paul M. Thompson, Duygu Tosun, Paul A. Yushkevich, Chadwick P. Ward, and Michael W. Weiner. Overview of ADNI MRI. *Alzheimer's & Dementia*, 20(10):7350–7360, 2024. 4

[8] Pamela J. LaMontagne, Tammie LS Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G. Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. *medRxiv*, 2019. 1

[9] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. 1

[10] Jorge Samper-González, Ninon Burgos, Simona Bottani, Sabrina Fontanella, Pascal Lu, Arnaud Marcoux, Alexandre Routier, Jérémy Guillon, Michael Bacci, Junhao Wen, Anne Bertrand, Hugo Bertin, Marie-Odile Habert, Stanley Durrleman, Theodoros Evgeniou, and Olivier Colliot. Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. *NeuroImage*, 183:504–521, 2018. 4

[11] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 1

[12] Yi Su. Ysu001/PUP. GitHub repository, https://github.com/ysu001/PUP, 2025. 1

[13] Nicholas J. Tustison, Brian B. Avants, Philip A. Cook, Yuanjie Zheng, Alexander Egan, Paul A. Yushkevich, and James C. Gee. N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, 2010. 1

[14] Nicholas J. Tustison, Philip A. Cook, Andrew J. Holbrook, Hans J. Johnson, John Muschelli, Gabriel A. Devenyi, Jeffrey T. Duda, Sandhitsu R. Das, Nicholas C. Cullen, Daniel L. Gillen, Michael A. Yassa, James R. Stone, James C. Gee, and Brian B. Avants. The ANTsX ecosystem for quantitative biological and medical imaging. *Scientific Reports*, 11(1):9068, 2021. 1

[15] Ghiam Yamin and David B. Teplow. Pittsburgh Compound-B (PiB) binds amyloid β-protein protofibrils. *Journal of Neurochemistry*, 140(2):210–215, 2017. 1