Distribution-aware Fairness Learning in Medical Image Segmentation From A Control-Theoretic Perspective

International Conference on Machine Learning (ICML) 2025, (Top-2.6% Spotlight Paper)

Yujin Oh*[1], Pengfei Jin*[1], Sangjoon Park*[2,3],
Sekeun Kim[1], Siyeop Yoon[1], Kyungsang Kim[1], Jin Sung Kim[2,4], Xiang Li†[1], Quanzheng Li†[1]

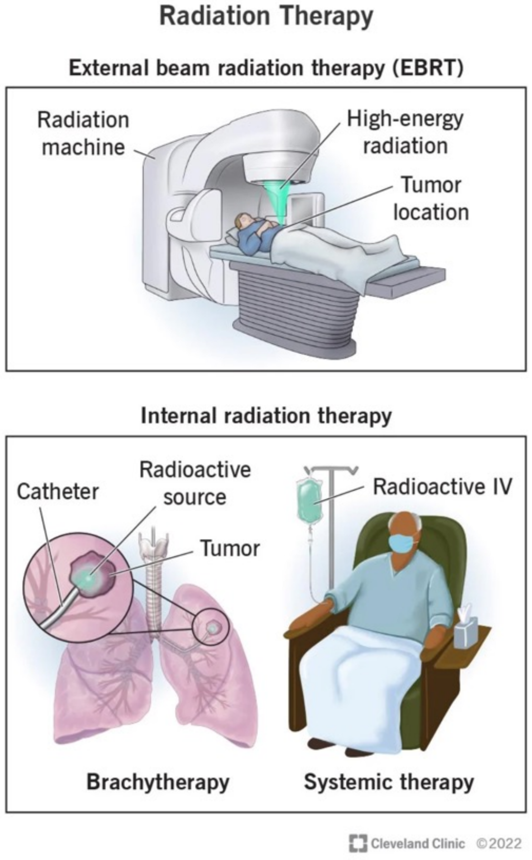
*Co-first authors, †Corresponding Authors

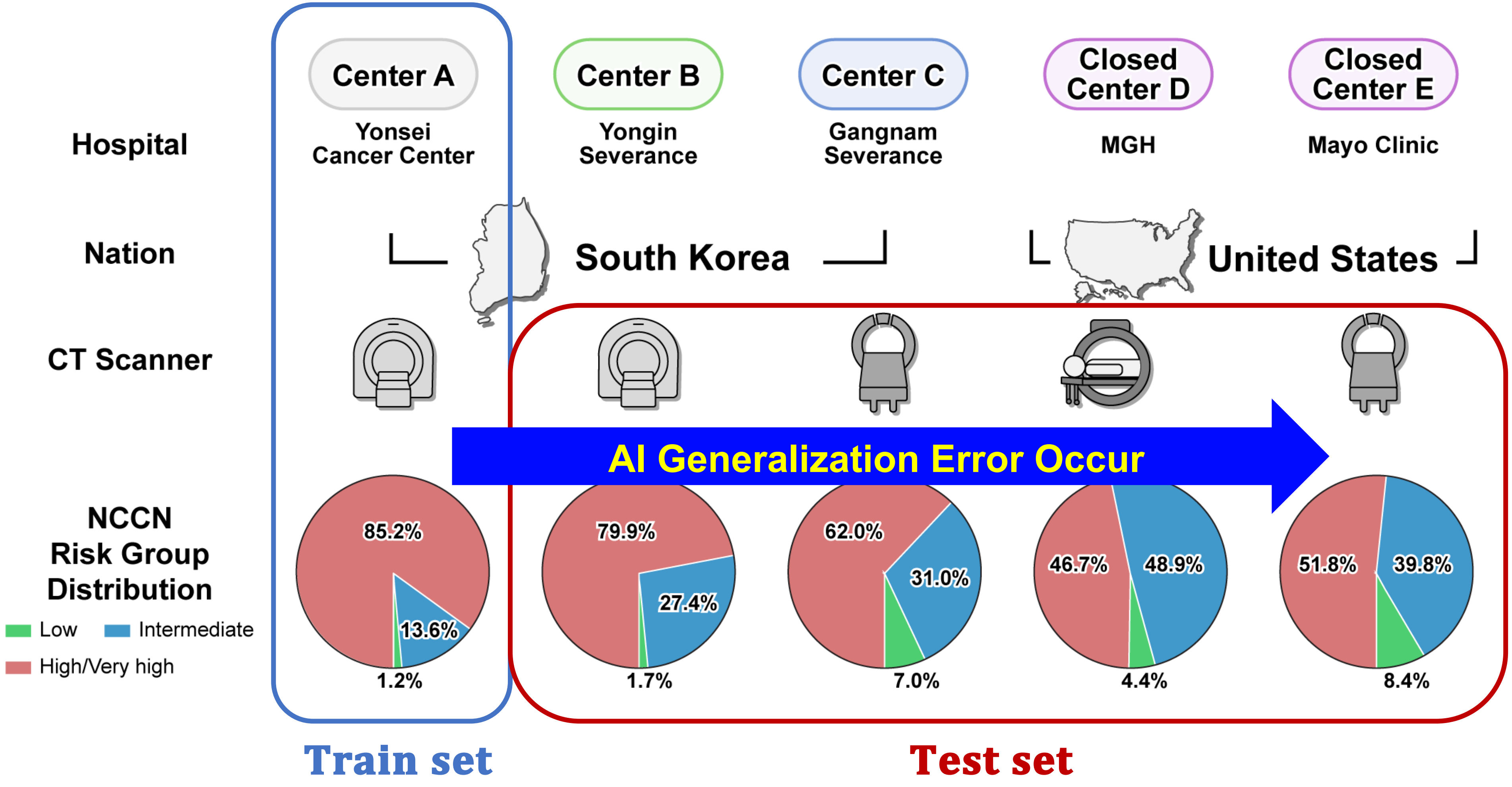
[1]Center for Advanced Medical Computing and Analysis (CAMCA), Harvard Medical School and Massachusetts General Hospital (MGH)
[2]Department of Radiation Oncology, Yonsei University College of Medicine, Yonsei University
[3]Institute for Innovation in Digital Healthcare, Yonsei University
[4]Oncosoft Inc.

Motivation

Biased Clinical Data Distribution

Patient Distribution in Prostate Cancer Treated with **External Radiotherapy**





NCCN (National Comprehensive Cancer Network); Y. Oh et al., In Revision for *IEEE TPAMI*

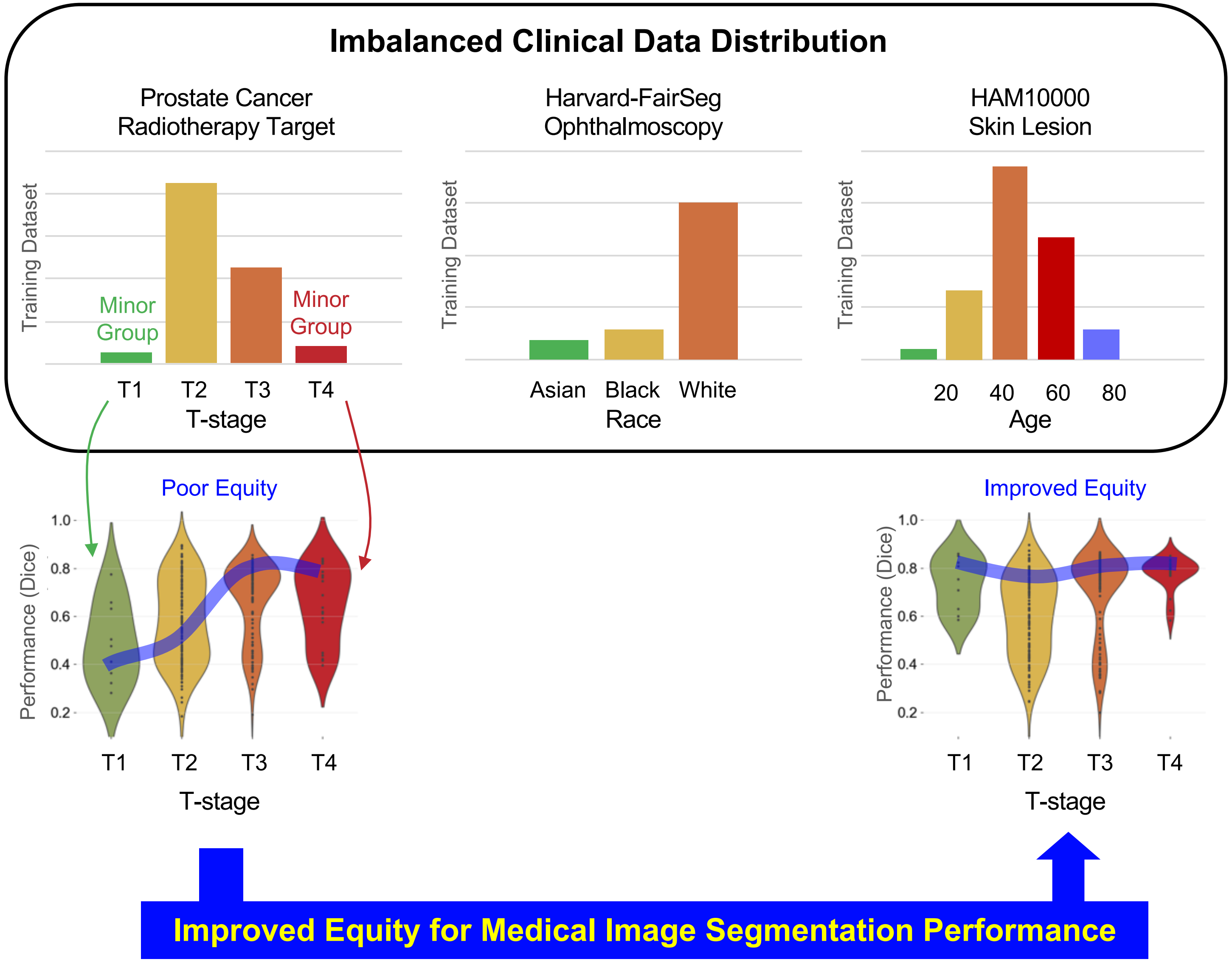
Fairness Learning in Medical AI Performance

- Medical data is often ill-posed due to:
  - **Demographics** (age, gender, race)
  - **Clinical variability** (disease severity)
- Imbalanced data distribution during AI training leads to **biased model performance**
- Advanced fairness learning strategies:
  - FEBS (Y. Tian et al., ICLR 2024)
  - FairDiff (Li et al., *MICCAI* 2024)

  >> Demographic aspect
- Our goal:

  >> Both demographic & clinical aspect

  >> Account for distributional patterns

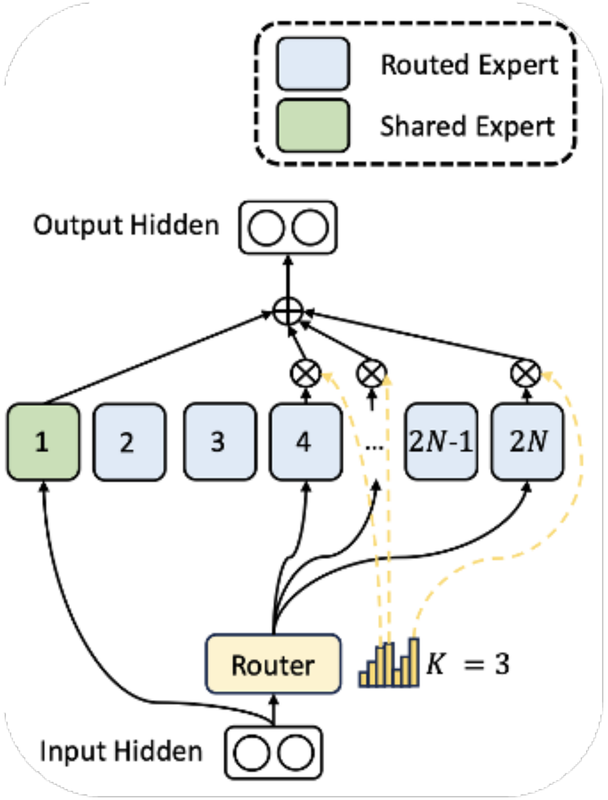


Motivated by Sparse Gating from Mixture of Expert

- Mixture of Expert (MoE)

  : leverages sparse gating for computational efficiency in large neural networks

$$y = \sum_{i}^{k} G(x)_i E_i(x)$$

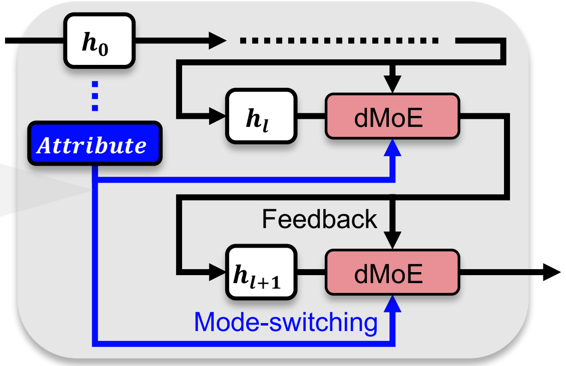


Explain MoE from

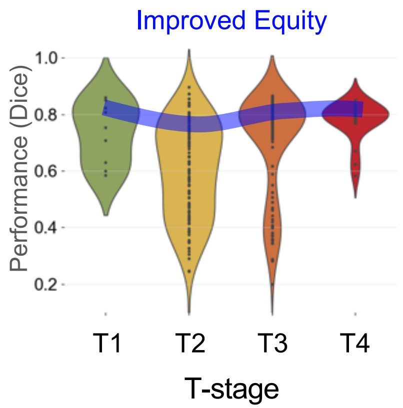
A Control-Theory

- Distribution-aware MoE (dMoE)

$$h_{l+1} = h_l + \sum_{i}^{k} G_i^{attr}(h_l) E_i(h_l)$$



Distribution-aware
Mixture of Experts



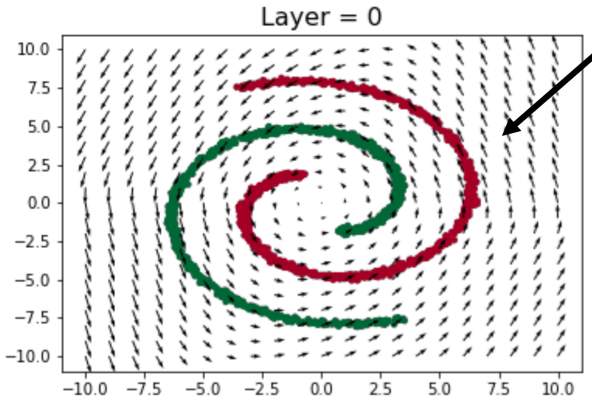
MoE, N. Shazeer et al., *ICLR* 2017

Theory

Explain MoE from A Control-Theoretic Perspective

- Neural Residual Network $$h_{l+1} = h_l + f(h_l, \theta_l).$$
- Forward Euler Scheme of Ordinary Differential Equation $$\frac{dh_t}{dt} = f(h_t, u_t),$$

t u_t



ResNet-50 for 2D classification

Neural ODE, R. Chen, *NeurIPS* 2018; LM-Resnet, Y. Lu et al., *PMLR* 2018

- Non-feedback Control $$\frac{dh_t}{dt} = f(h_t, \underbrace{u_t}_{\text{Neural Parameters}}),$$
- Feedback Control $$dh_t = f\big(h_t, \underbrace{u_t(h_t)}_{\text{Parameters governed by real-time state}}\big)dt,$$
- Mixture of Expert (MoE)

discretization

$$h_{l+1} = h_l + \underbrace{\sum_i^k G(h_l)_i}_{\text{Sparse gating}} \underbrace{E_i(h_l)}_{\text{Experts}}$$

$$u_t(h_t) \approx \underbrace{\sum_i K\big(h_t, h_t^i\big)}_{\text{Kernel method}} u_t(h_t^i),$$

MoE, N. Shazeer et al., *ICLR* 2017
Kernel method, B. Schölkopf *et al., MIT press 2002*

Explain MoE from A Control-Theoretic Perspective

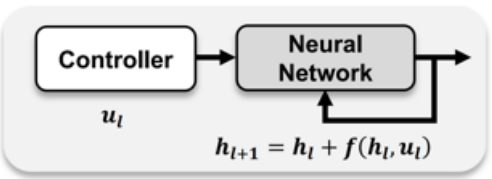
- Non-feedback Control

$$\frac{dh_t}{dt} = f(h_t, \underline{u_t}),$$

Neural Parameters

discretization

Neural Network



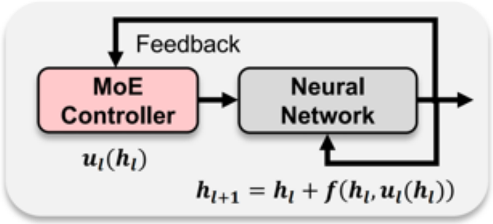
- Feedback Control

$$dh_t = f\big(h_t, \underline{u_t(h_t)}\big)dt,$$

Parameters governed by real-time state

discretization

MoE



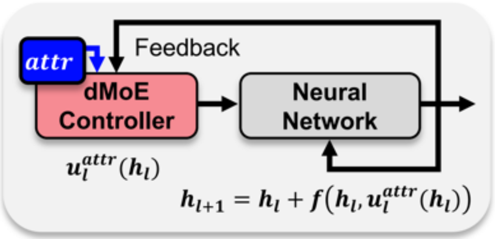
- Mode-switching Control

$$u_t(h_t) = \underline{\kappa_{s(attr)}(h_t)}.$$

Multiple sub-strategies governed by distributional attribute

discretization

Distribution-aware MoE

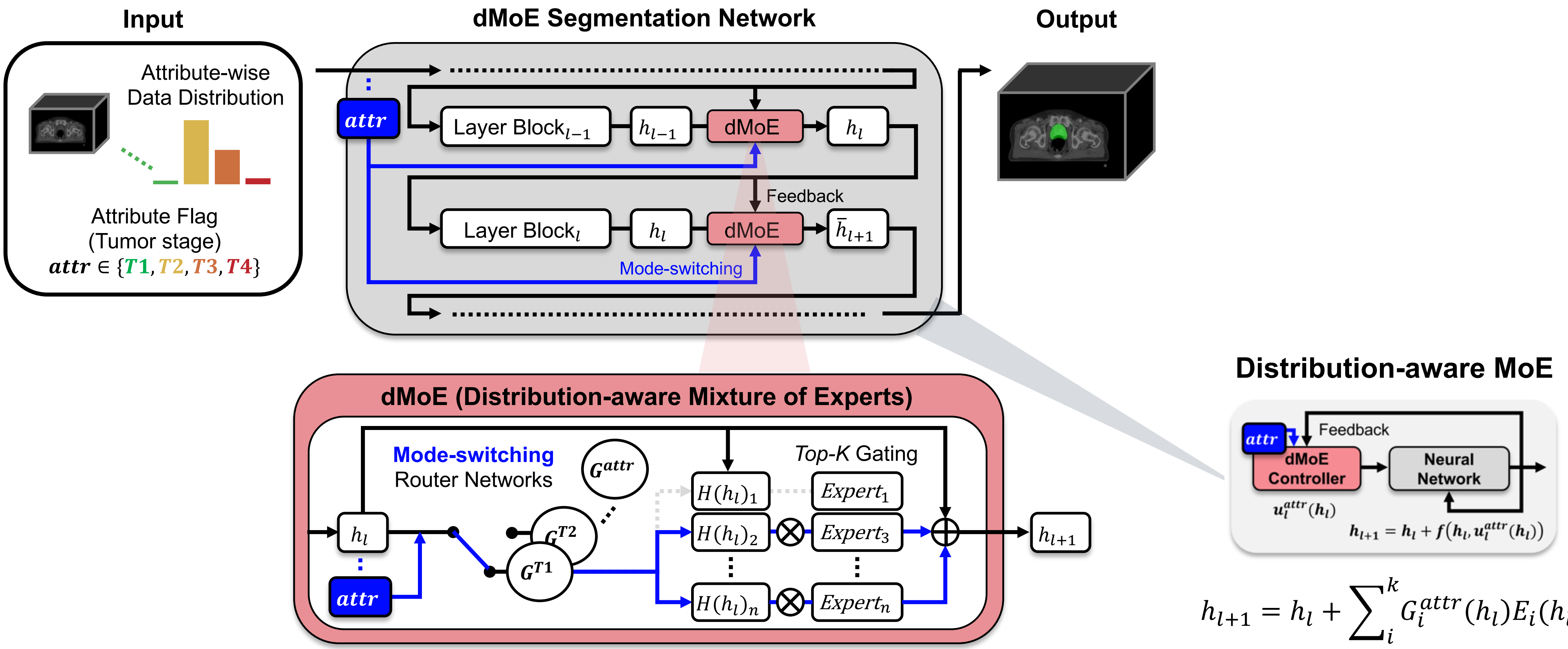


$$h_{l+1} = h_l + \sum_i^k G_i^{attr}(h_l) E_i(h_l)$$

MoE, N. Shazeer et al., *ICLR* 2017; J. C. Doyle et al., 2013; J. D. Boskovic et al., *IEEE* 2021

Distribution-aware Mixture of Expert (dMoE)

- An optimal control-inspired approach to achieve distribution-aware adaptation of network
- Specific focus on radiotherapy target volume contouring in Radiation Oncology



$$h_{l+1} = h_l + \sum_i^k G_i^{attr}(h_l)E_i(h_l)$$

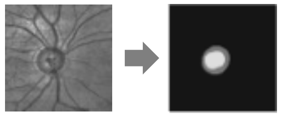
Experimental Results

Experimental Settings

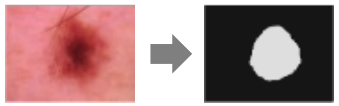
- Diverse medical image segmentation datasets

Table 6. Detailed distribution of data across attribute subgroups.

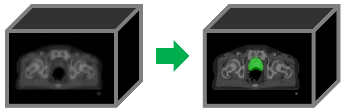
	Harvard-FairSeg				HAM10000						Radiotherapy Target Dataset				
Dataset	Total	Attribute (Race)			Total	Attribute (Age)					Total	Attribute (T-stage)			
		Asian	Black	White		≥ 80	≥ 60	≥ 40	≥ 20	< 20		T1	T2	T3	T4
Trainset (%)	7945 (100)	750 (9)	1174 (15)	6021 (76)	8137 (100)	191 (2)	1324 (16)	3693 (45)	2356 (31)	573 (7)	721 (100)	26 (4)	227 (31)	425 (59)	43 (6)
Testset	2000	169	299	1532	1061	121	469	328	120	24	275	11	129	114	21



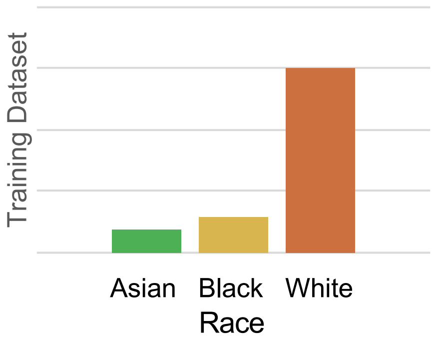
Scanning Laser Ophthalmoscopy (SLO) Fundus for Optic Cup & Neuroretinal Rim

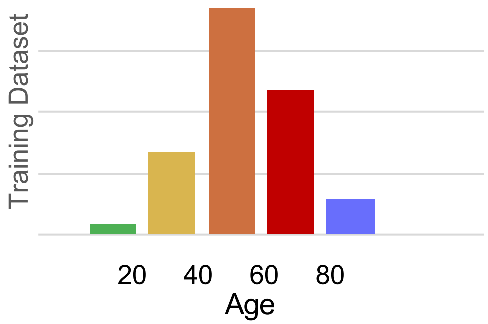


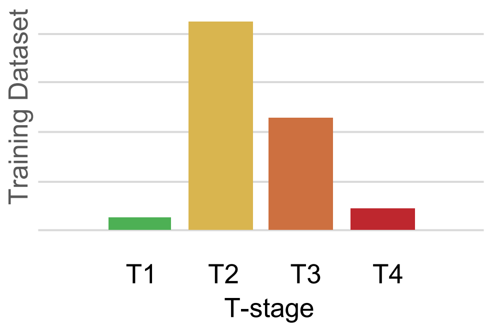
Dermatology Skin for Lesion



Pelvic CT for Radiotherapy Target







Harvard-FairSeg, Y. Tian et al., *ICLR* 2024; HAM10000, P. Tschandl et al., *Scientific Data* 2018; Radiotherapy Target Dataset, Severance Hospital, South Korea (*IRB numbers 4-2023-0179, 9-2023-0161, and 3-2023-0396*)

Experimental Settings

- 2D Transformer architectures

Table 4. dMoE within Transformer (TransUNet).

Module	Layer Block	Resample	dMoE	Data dimension (C × H × W)
In	- Conv	- -	- -	$Ch_{in} \times 224 \times 224$ $1 \times 14 \times 14$
Encoder	AttentionBlock_1 AttentionBlock_2 ⋮ $\text{AttentionBlock}_{11}$ $\text{AttentionBlock}_{12}$	- - - - -	dMoE	$768 \times (14 \times 14)$ $768 \times (14 \times 14)$ ⋮ $768 \times (14 \times 14)$ $768 \times (14 \times 14)$
Decoder	UpResBlock_4 UpResBlock_3 UpResBlock_2 UpResBlock_1	Up Up Up Up	- - - -	$256 \times 28 \times 28$ $128 \times 56 \times 56$ $64 \times 112 \times 112$ $16 \times 224 \times 224$
Out	Conv	-	-	$Ch_{out} \times 224 \times 224$

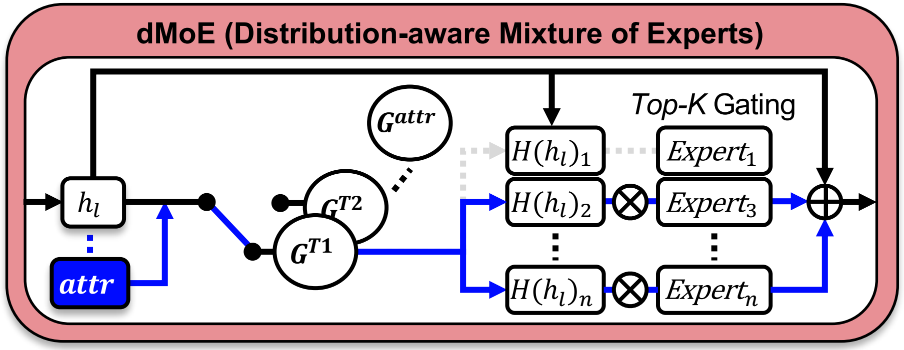
- 3D Residual U-Net architectures

Table 5. dMoE within 3D CNN (3D ResUNet).

Module	Layer Block	Resample	dMoE	Skip-Connection	Data dimension (C × H × W × D)
In	Conv	-	-	-	$Ch_{in} \times 384 \times 384 \times 128$
Encoder	ResBlock_1 ResBlock_2 ResBlock_3 ResBlock_4 ResBlock_5	Down Down Down Down Down	dMoE_1 dMoE_2 dMoE_3 dMoE_4 dMoE_5	↴ ↴ ↴ ↴	$48 \times 192 \times 192 \times 64$ $48 \times 96 \times 96 \times 32$ $96 \times 48 \times 48 \times 16$ $192 \times 24 \times 24 \times 8$ $384 \times 12 \times 12 \times 4$
Decoder	UpResBlock_4 UpResBlock_3 UpResBlock_2 UpResBlock_1	Up Up Up Up	- - - -	↵ ↵ ↵ ↵	$192 \times 24 \times 24 \times 8$ $96 \times 48 \times 48 \times 16$ $48 \times 96 \times 96 \times 32$ $48 \times 192 \times 192 \times 64$
Out	TransposeConv	Up	-	-	$Ch_{out} \times 384 \times 384 \times 128$

- dMoE training

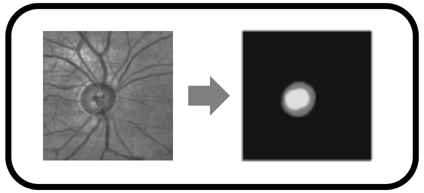
  - ✓ Top-K : 2
  - ✓ #n of Expert : 8
  - ✓ Expert : MLP (Linear – ReLU – Linear – Dropout)
  - ✓ Training : Single NVIDIA A100 80 GB memory GPU

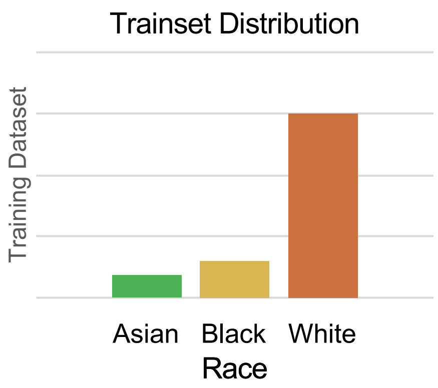


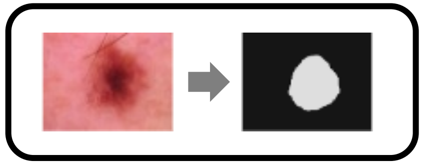
TransUNet, J. Chen et al., *arXiv* 2021; 3D U-Net, Ö. Çiçek et al., *MICCAI* 2016

Improving Fairness in 2D Medical Image Segmentation

Data / Attribute







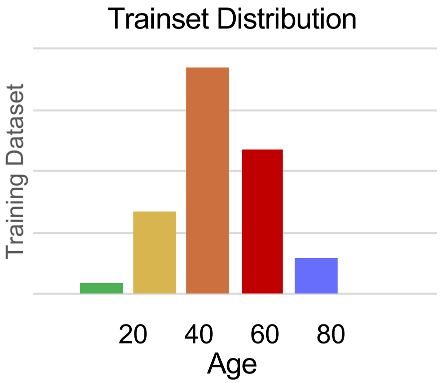


Table 1. Comparison on 2D Harvard-FairSeg dataset with **race** as the distribution attribute.

Method	All (n=2000) ES-Dice (CIs)	All (n=2000) Dice (CIs)	All (n=2000) ES-IoU (CIs)	All (n=2000) IoU (CIs)	Asian (n=169) Dice	Asian (n=169) IoU	Black (n=299) Dice	Black (n=299) IoU	White (n=1532) Dice	White (n=1532) IoU
Rim Segmentation										
TransUNet† (Chen et al., 2021)	0.703	0.793	0.585	0.671	0.746	0.616	0.731	0.599	0.811	0.691
+ ADV† (Madras et al., 2018)	0.700	0.791	0.583	0.668	0.741	0.612	0.729	0.598	0.809	0.689
+ DRO† (Sagawa et al., 2019)	0.700	0.790	0.581	0.667	0.747	0.618	0.723	0.590	0.808	0.689
+ FEBS† (Tian et al., 2024)	0.705	0.795	0.587	0.673	0.748	0.619	0.733	0.602	0.813	0.694
+ FairDiff‡ (Li et al., 2024)	0.716	0.800	0.596	0.680	0.757	0.628	0.743	0.611	0.817	0.699
+ MoE	0.733 (0.713-0.752)	0.804 (0.799-0.809)	0.614 (0.596-0.633)	0.685 (0.680-0.691)	0.760	0.635	0.763	0.635	0.817	0.701
+ dMoE	**0.743 (0.723-0.763)**	**0.813 (0.808-0.818)**	**0.627 (0.608-0.645)**	**0.698 (0.692-0.704)**	**0.769**	**0.645**	**0.776**	**0.652**	**0.825**	**0.713**
Cup Segmentation										
TransUNet† (Chen et al., 2021)	0.828	0.848	0.730	0.753	0.827	0.728	0.849	0.758	0.850	0.755
+ ADV† (Madras et al., 2018)	0.826	0.841	0.727	0.743	0.825	0.726	0.842	0.748	0.843	0.744
+ DRO† (Sagawa et al., 2019)	0.820	0.844	0.725	0.748	0.820	0.723	0.847	0.753	0.846	0.750
+ FEBS† (Tian et al., 2024)	0.825	0.846	0.727	0.750	0.825	0.725	0.848	0.755	0.848	0.751
+ FairDiff‡ (Li et al., 2024)	0.825	0.848	0.736	0.753	0.832	0.735	0.848	0.757	0.850	0.754
+ MoE	0.830 (0.809-0.847)	0.854 (0.849-0.860)	0.739 (0.720-0.754)	0.762 (0.755-0.768)	**0.845**	**0.757**	0.842	0.748	0.857	0.765
+ dMoE	**0.832 (0.810-0.853)**	**0.862 (0.856-0.867)**	**0.745 (0.722-0.765)**	**0.773 (0.766-0.779)**	0.844	0.755	**0.851**	**0.761**	**0.866**	**0.777**

† Metric reported from (Tian et al., 2024). ‡ ES-metrics are recalculated using Eq. (19), based on metrics reported in the original paper (Li et al., 2024), for a fair comparison.

Table 2. Comparison on 2D HAM10000 dataset for skin lesion segmentation with **age** as the distribution attribute.

Method	All (n=1061) ES-Dice (CIs)	All (n=1061) Dice (CIs)	All (n=1061) ES-IoU (CIs)	All (n=1061) IoU (CIs)	Age ≥ 80 (n=121) Dice	Age ≥ 80 (n=121) IoU	Age ≥ 60 (n=469) Dice	Age ≥ 60 (n=469) IoU	Age ≥ 40 (n=328) Dice	Age ≥ 40 (n=328) IoU	Age ≥ 20 (n=120) Dice	Age ≥ 20 (n=120) IoU	Age < 20 (n=24) Dice	Age < 20 (n=24) IoU
TransUNet (Chen et al., 2021)	0.792 (0.737-0.841)	0.876 (0.863-0.889)	0.714 (0.664-0.766)	0.824 (0.809-0.838)	0.862	0.787	0.868	0.809	0.888	0.846	0.895	0.857	0.875	0.839
+ FEBS (Tian et al., 2024)	0.757 (0.704-0.807)	0.858 (0.845-0.872)	0.667 (0.613-0.719)	0.798 (0.783-0.812)	0.831	0.747	0.844	0.774	0.884	0.837	0.871	0.827	0.869	0.830
+ MoE	0.796 (0.741-0.844)	0.882 (0.868-0.895)	0.721 (0.671-0.770)	0.833 (0.818-0.846)	**0.864**	**0.794**	0.875	0.820	0.889	**0.851**	**0.904**	**0.869**	**0.882**	**0.850**
+ dMoE	**0.801 (0.745-0.847)**	**0.884 (0.870-0.896)**	**0.725 (0.673-0.776)**	**0.834 (0.820-0.847)**	**0.864**	0.791	**0.881**	**0.824**	**0.890**	0.850	0.901	0.866	0.880	0.846

Improving Fairness in 3D Radiotherapy Target Contouring

Data / Attribute

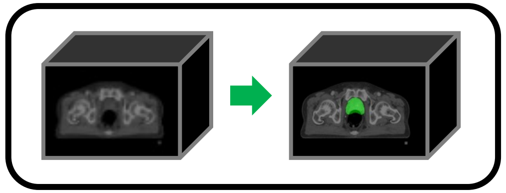
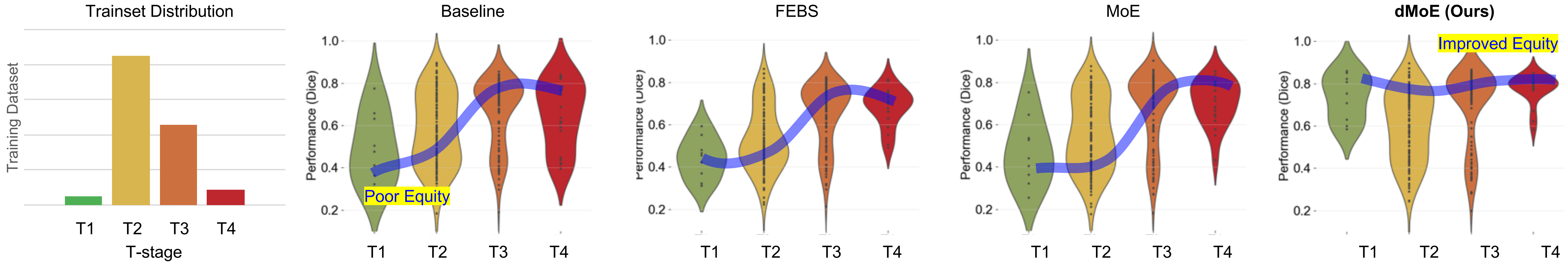
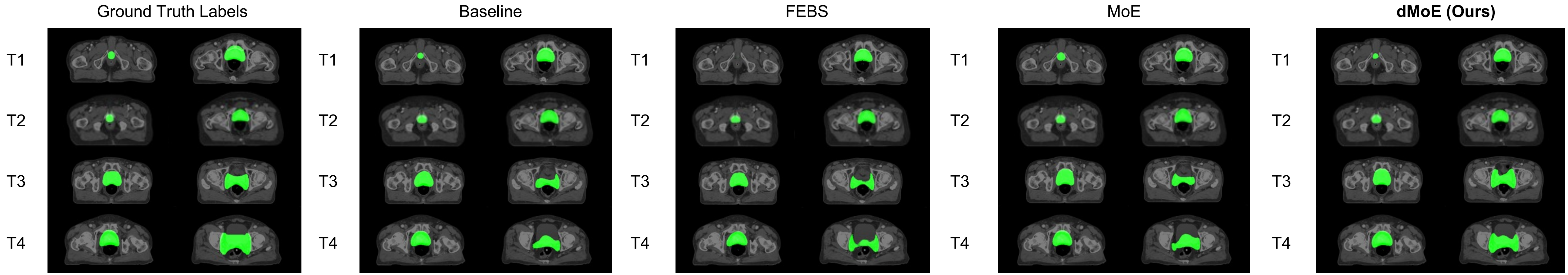


Table 3. Comparison on 3D radiotherapy target segmentation with **tumor stage** as the distribution attribute.

Method	All (n=275)				T1 (n=11)		T2 (n=129)		T3 (n=114)		T4 (n=21)	
	ES-Dice (CIs)	Dice (CIs)	ES-IoU (CIs)	IoU (CIs)	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
3D ResUNet (Çiçek et al., 2016)	0.487 (0.447-0.529)	0.610 (0.589-0.630)	0.367 (0.336-0.399)	0.462 (0.440-0.482)	0.493	0.341	0.569	0.420	0.659	0.511	0.656	0.506
+ FEBS (Tian et al., 2024)	0.434 (0.406-0.467)	0.586 (0.567-0.604)	0.322 (0.302-0.346)	0.433 (0.414-0.452)	0.442	0.288	0.528	0.374	0.652	0.501	0.685	0.527
+ MoE	0.452 (0.415-0.492)	0.608 (0.586-0.628)	0.342 (0.314-0.372)	0.461 (0.439-0.482)	0.492	0.345	0.542	0.393	0.674	0.532	0.708	0.557
+ dMoE	**0.499 (0.469-0.531)**	**0.650 (0.628-0.671)**	**0.384 (0.358-0.410)**	**0.506 (0.483-0.528)**	**0.718**	**0.571**	**0.585**	**0.435**	**0.693**	**0.556**	**0.778**	**0.641**

Note. The underlined value indicates the worst-group accuracy among distribution attributes for each method.





Computationally Efficient with Optimal Performance

	TransUNet	+MoE	+dMoE	**3D ResUNet**	+MoE	+dMoE
Input	224 W × 224 H			384 W × 384 H × 128 D		
GFlops	45.84	90.28	90.28	1542.36	1761.30	1761.30
Params (M)	91.67	129.46	129.51	13.28	26	26.05

Table 7. Computational complexity comparison.

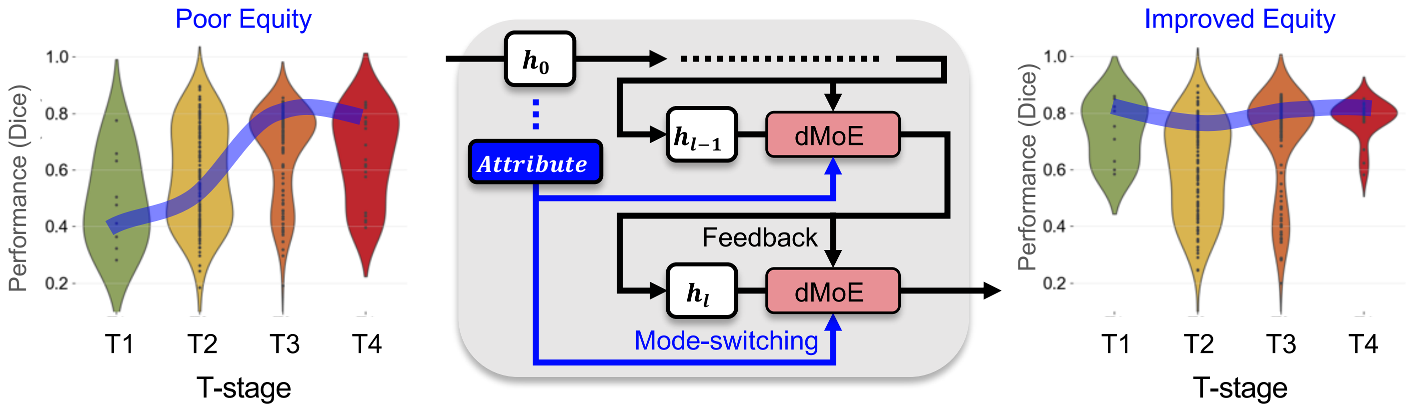
Method	GFlops ↓	All (n=275)		T1 (n=11)	T2 (n=129)	T3 (n=114)	T4 (n=21)
		ES-Dice(D)	Dice	Dice	Dice	Dice	Dice
dMoE (Ours)	**1761.30**	**0.499**	**0.650**	**0.718**	**0.585**	**0.693**	**0.778**
Multiple networks for each attribute	5729.44	0.457	0.606	0.599	0.515	0.681	0.760

Table 8. Comparison to multiple networks for each attribute.

Conclusion

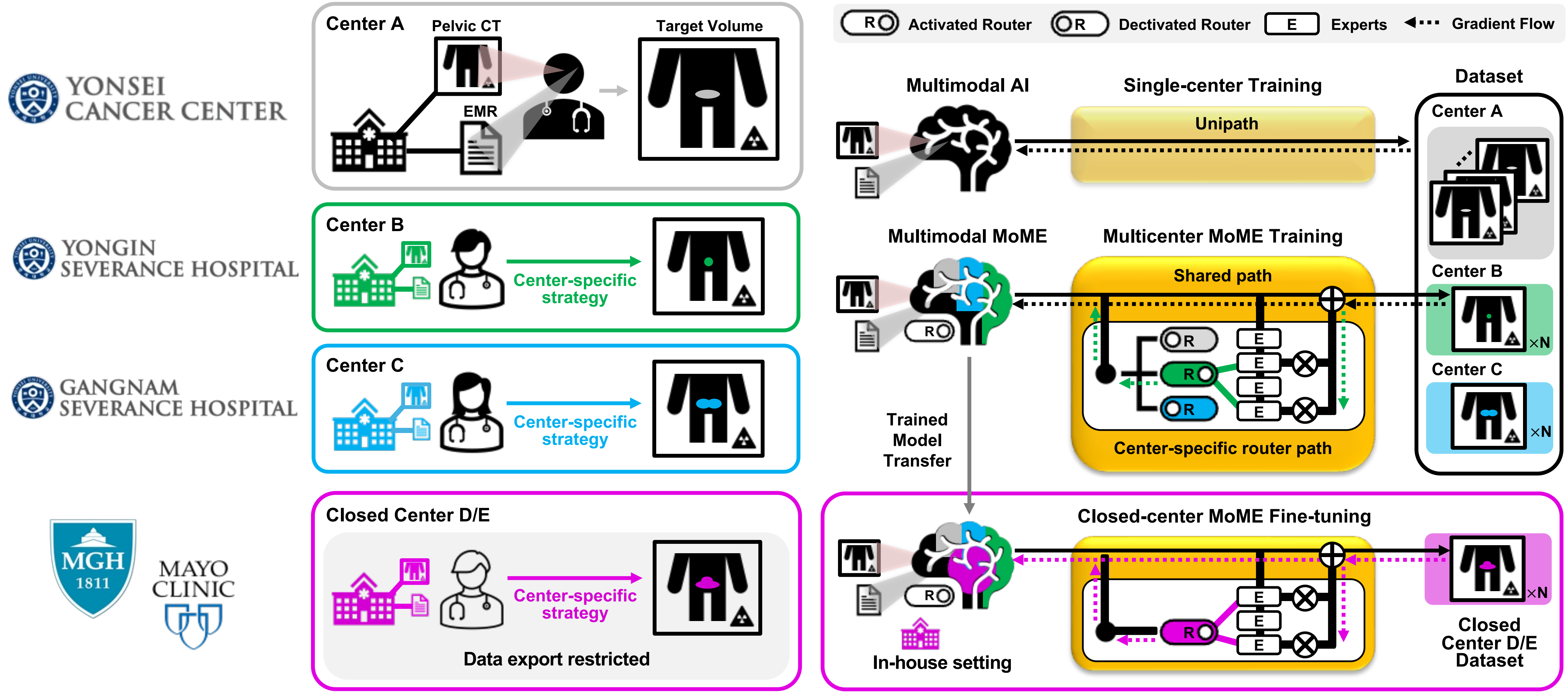
Summary

- We introduce **Distribution-aware** Mixture of Experts (dMoE).
- We enhance MoE gating mechanism to incorporate **distributional information as a mode-switching control** for adaptive parameter selection.



- dMoE advances **equitable and reliable AI-driven medical image analysis**.
- dMoE holds promise in **adapting trained models to unknown distributions**, thereby improving the success of **clinical AI integration across diverse hospitals**.

Future Work: Mixture of Multicenter Experts (MoME)



Y. Oh et al., In Revision for *IEEE TPAMI*

ICML
International Conference
On Machine Learning

Distribution-aware Fairness Learning in Medical Image Segmentation From A Control-Theoretic Perspective

International Conference on Machine Learning (ICML) 2025, (Top-2.6% Spotlight Paper)

Thank you!

https://github.com/tvseg/dMoE

ArXiv