Medical Vision Seminar

——Chenyu Liu

2022_01_19

(ICCV2021)CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification——

Chun-Fu (Richard) Chen, Quanfu Fan, Rameswar Panda

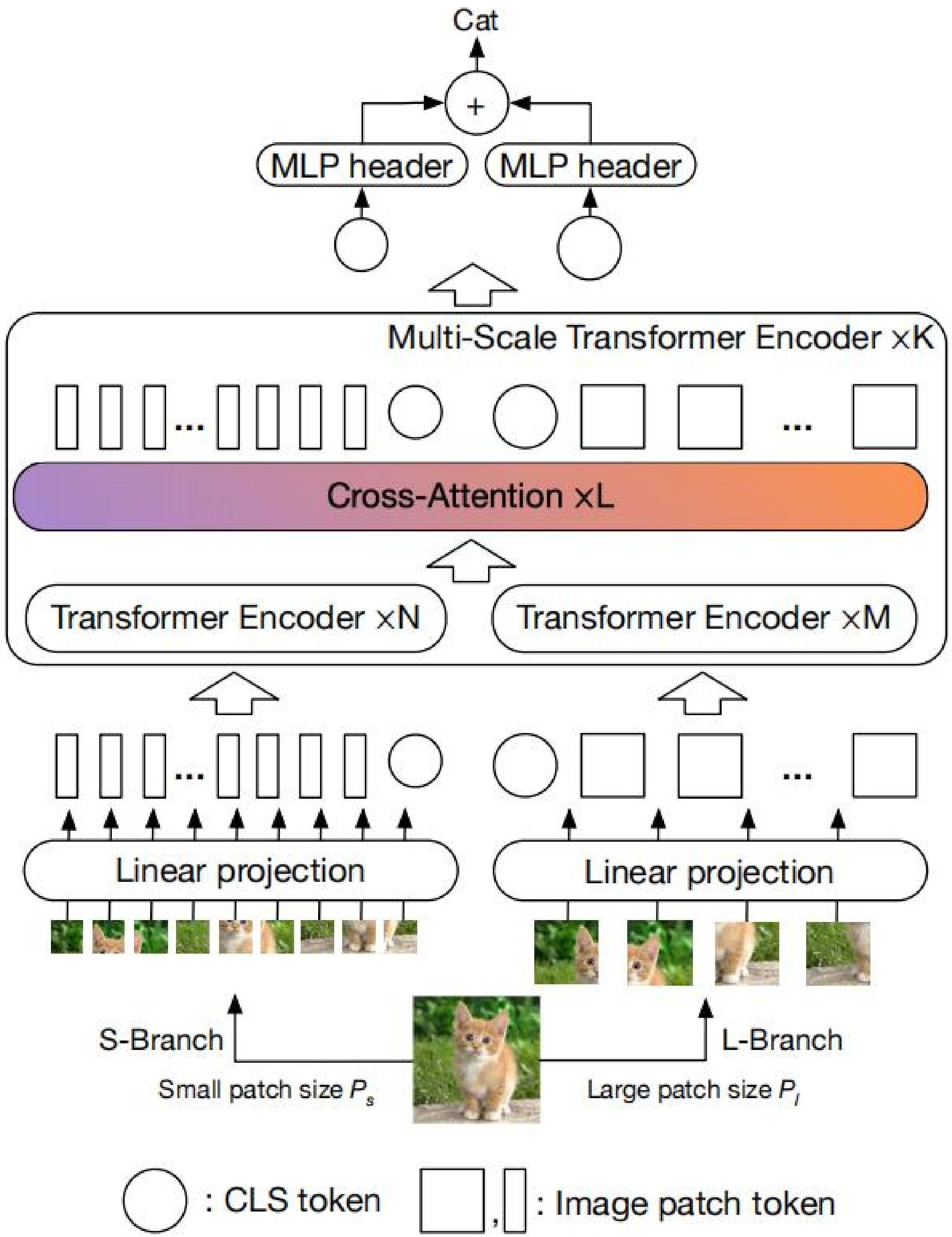
1. Motivation

1. The granularity of the patch size affects the accuracy and complexity of ViT; So they propose a novel dual-branch vision transformer to extract multi-scale feature representations for image classification.

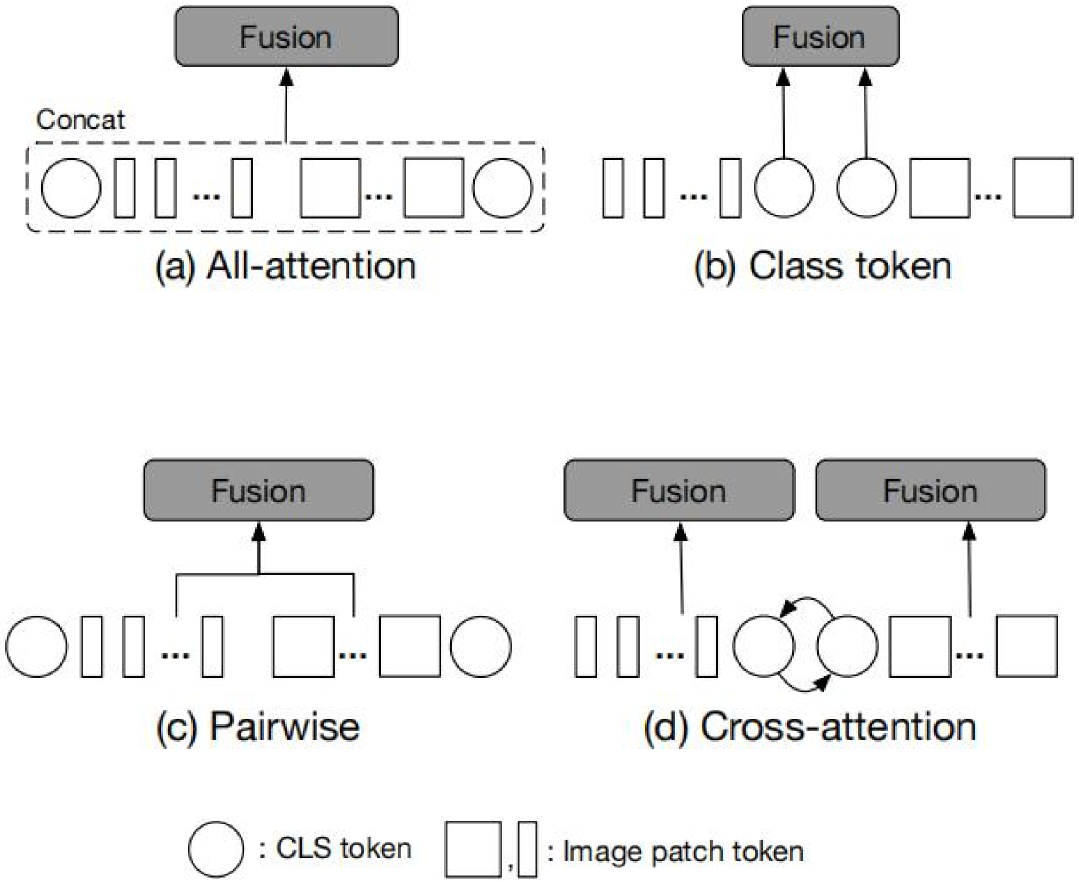
2. Effective feature fusion is the key for learning multiscale feature representations. So they develop a simple yet effective token fusion scheme based on cross-attention

2. Multi-Scale Vision Transformer

each encoder consists of
two branches:
(1) **L-Branch**: a large (primary) branch that utilizes coarse-grained patch size (Pl) with more transformer encoders and wider embedding dimensions,
(2) **S-Branch**: a small (complementary) branch that operates at fine-grained patch size (Ps) with fewer encoders and smaller embedding dimensions.



2.1 Multi-scale fusion Module



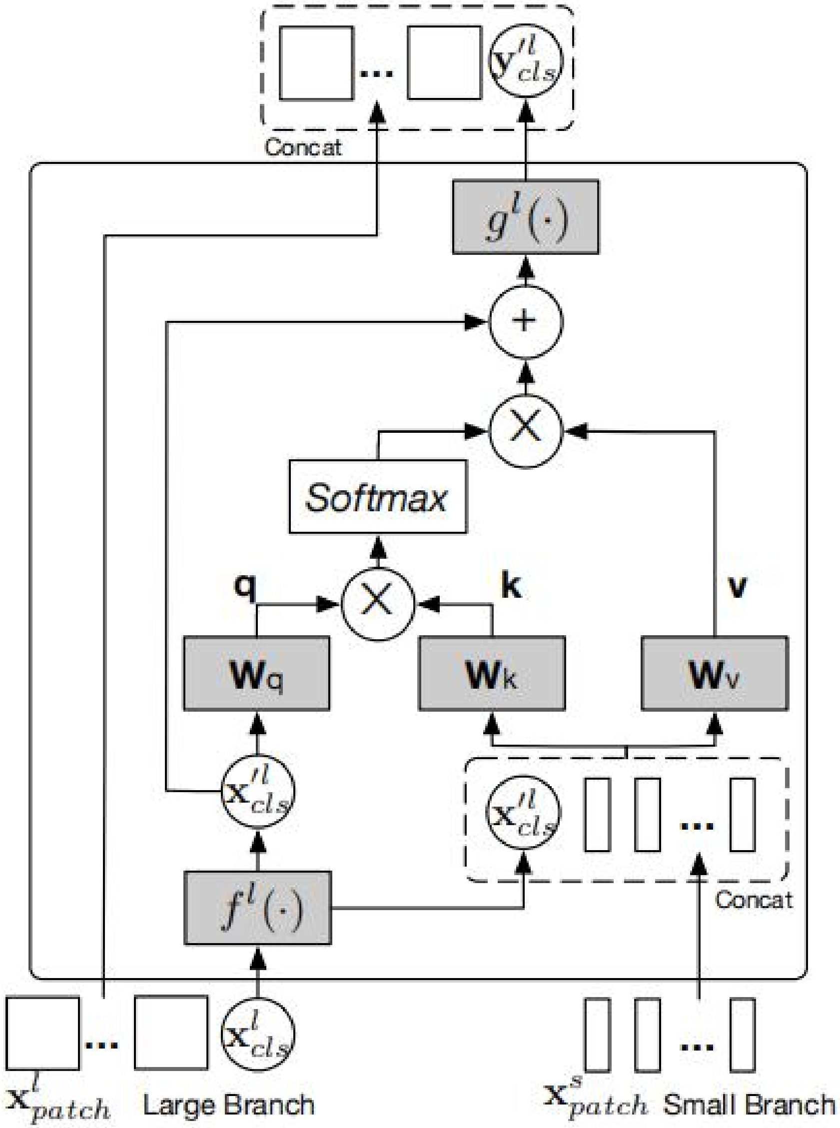


Figure 4: **Cross-attention module for Large branch**.

4.1 Experiment
-Comparisons with DeiT.

Model	Patch embedding	Patch size Small	Patch size Large	Dimension Small	Dimension Large	# of heads	M	r
CrossViT-Ti	Linear	12	16	96	192	3	4	4
CrossViT-S	Linear	12	16	192	384	6	4	4
CrossViT-B	Linear	12	16	384	768	12	4	4
CrossViT-9	Linear	12	16	128	256	4	3	3
CrossViT-15	Linear	12	16	192	384	6	5	3
CrossViT-18	Linear	12	16	224	448	7	6	3
CrossViT-9†	3 Conv.	12	16	128	256	4	3	3
CrossViT-15†	3 Conv.	12	16	192	384	6	5	3
CrossViT-18†	3 Conv.	12	16	224	448	7	6	3

Model	Top-1 Acc. (%)	FLOPs (G)	Throughput (images/s)	Params (M)
DeiT-Ti	72.2	1.3	2557	5.7
CrossViT-Ti	73.4 (+1.2)	1.6	1668	6.9
CrossViT-9	73.9 (+0.5)	1.8	1530	8.6
CrossViT-9†	**77.1** (+3.2)	2.0	1463	8.8
DeiT-S	79.8	4.6	966	22.1
CrossViT-S	81.0 (+1.2)	5.6	690	26.7
CrossViT-15	81.5 (+0.5)	5.8	640	27.4
CrossViT-15†	**82.3** (+0.8)	6.1	626	28.2
DeiT-B	81.8	17.6	314	86.6
CrossViT-B	82.2 (+0.4)	21.2	239	104.7
CrossViT-18	82.5 (+0.3)	9.0	430	43.3
CrossViT-18†	**82.8** (+0.3)	**9.5**	418	44.3

4.2 Experiment
-Comparisons with SOTA.

Model	Top-1 Acc. (%)	FLOPs (G)	Params (M)
Peceiver [19] (arXiv, 2021-03)	76.4	–	43.9
DeiT-S [35] (arXiv, 2020-12)	79.8	4.6	22.1
CentroidViT-S [42] (arXiv, 2021-02)	80.9	4.7	22.3
PVT-S [38] (arXiv, 2021-02)	79.8	3.8	24.5
PVT-M [38] (arXiv, 2021-02)	81.2	6.7	44.2
T2T-ViT-14 [45] (arXiv, 2021-01)	80.7	6.1*	21.5
TNT-S [14] (arXiv, 2021-02)	81.3	5.2	23.8
CrossViT-15 (Ours)	81.5	5.8	27.4
CrossViT-15† (Ours)	**82.3**	6.1	28.2
ViT-B@384 [11] (ICLR, 2021)	77.9	17.6	86.6
DeiT-B [35] (arXiv, 2020-12)	81.8	17.6	86.6
PVT-L [38] (arXiv, 2021-02)	81.7	9.8	61.4
T2T-ViT-19 [45] (arXiv, 2021-01)	81.4	9.8*	39.0
T2T-ViT-24 [45] (arXiv, 2021-01)	82.2	15.0*	64.1
TNT-B [14] (arXiv, 2021-02)	**82.8**	14.1	65.6
CrossViT-18 (Ours)	82.5	9.0	43.3
CrossViT-18† (Ours)	**82.8**	9.5	44.3

*: We recompute the flops by using our tools.

Table 3: **Comparisons with recent transformer-based models on ImageNet1K.** All models are trained using only ImageNet1K dataset. Numbers are referenced from their recent version as of the submission date.

Model	Top-1 Acc. (%)	FLOPs (G)	Throughput (images/s)	Params (M)
ResNet-101 [15]	76.7	7.80	678	44.6
ResNet-152 [15]	77.0	11.5	445	60.2
ResNeXt-101-32×4d [43]	78.8	8.0	477	44.2
ResNeXt-101-64×4d [43]	79.6	15.5	289	83.5
SEResNet-101 [18]	77.6	7.8	564	49.3
SEResNet-152 [18]	78.4	11.5	392	66.8
SENet-154 [18]	81.3	20.7	201	115.1
ECA-Net101 [37]	78.7	7.4	591	42.5
ECA-Net152 [37]	78.9	10.9	428	59.1
RegNetY-8GF [30]	79.9	8.0	557	39.2
RegNetY-12GF [30]	80.3	12.1	439	51.8
RegNetY-16GF [30]	80.4	15.9	336	83.6
RegNetY-32GF [30]	81.0	32.3	208	145.0
EfficienetNet-B4@380 [34]	82.9	4.2	356	19
EfficienetNet-B5@456 [34]	83.7	9.9	169	30
EfficienetNet-B6@528 [34]	84.0	19.0	100	43
EfficienetNet-B7@600 [34]	84.3	37.0	55	66
CrossViT-15	81.5	5.8	640	27.4
CrossViT-15†	82.3	6.1	626	28.2
CrossViT-15†@384	83.5	21.4	158	28.5
CrossViT-18	82.5	9.03	430	43.3
CrossViT-18†	82.8	9.5	418	44.3
CrossViT-18†@384	83.9	32.4	112	44.6
CrossViT-18†@480	84.1	56.6	57	44.9

Table 4: **Comparisons with CNN models on ImageNet1K.** Models are evaluated under 224×224 if not spec-

4.2 Ablation Studies

Comparison of Different Fusion Schemes.

Effect of Patch Sizes.

Channel Width and Depth in S-branch.

Depth of Cross-Attention and Number of Multi-Scale Transformer Encoders.

Importance of CLS Tokens.

Fusion	Top-1 Acc. (%)	FLOPs (G)	Params (M)	Single Branch Acc. (%) L-Branch	S-Branch
None	80.2	5.3	23.7	80.2	0.1
All-Attention	80.0	7.6	27.7	79.9	0.5
Class Token	80.3	5.4	24.2	80.6	7.6
Pairwise	80.3	5.5	24.2	80.3	7.3
Cross-Attention	81.0	5.6	26.7	68.1	47.2

Table 6: **Ablation study with different fusions on ImageNet1K.** All models are based on CrossViT-S. Single branch Acc. is computed using `CLS` from one branch only.

Model	Patch size Small	Large	Dimension Small	Large	K	N	M	L	Top-1 Acc. (%)	FLOPs (G)	Params (M)
CrossViT-S	12	16	192	384	3	1	4	1	81.0	5.6	26.7
A	**8**	16	192	384	3	1	4	1	80.8	6.7	26.7
B	12	16	**384**	384	3	1	4	1	80.1	7.7	31.4
C	12	16	192	384	3	**2**	4	1	80.7	6.3	28.0
D	12	16	192	384	3	1	4	**2**	81.0	5.6	28.9
E	12	16	192	384	**6**	1	**2**	1	80.9	6.6	31.1

Table 7: **Ablation study with different architecture parameters on ImageNet1K.** The **blue** color indicates changes from CrossViT-S.

（MICCAI 2021） Combining 3D Image and Tabular Data via the Dynamic Affine Feature Map Transform

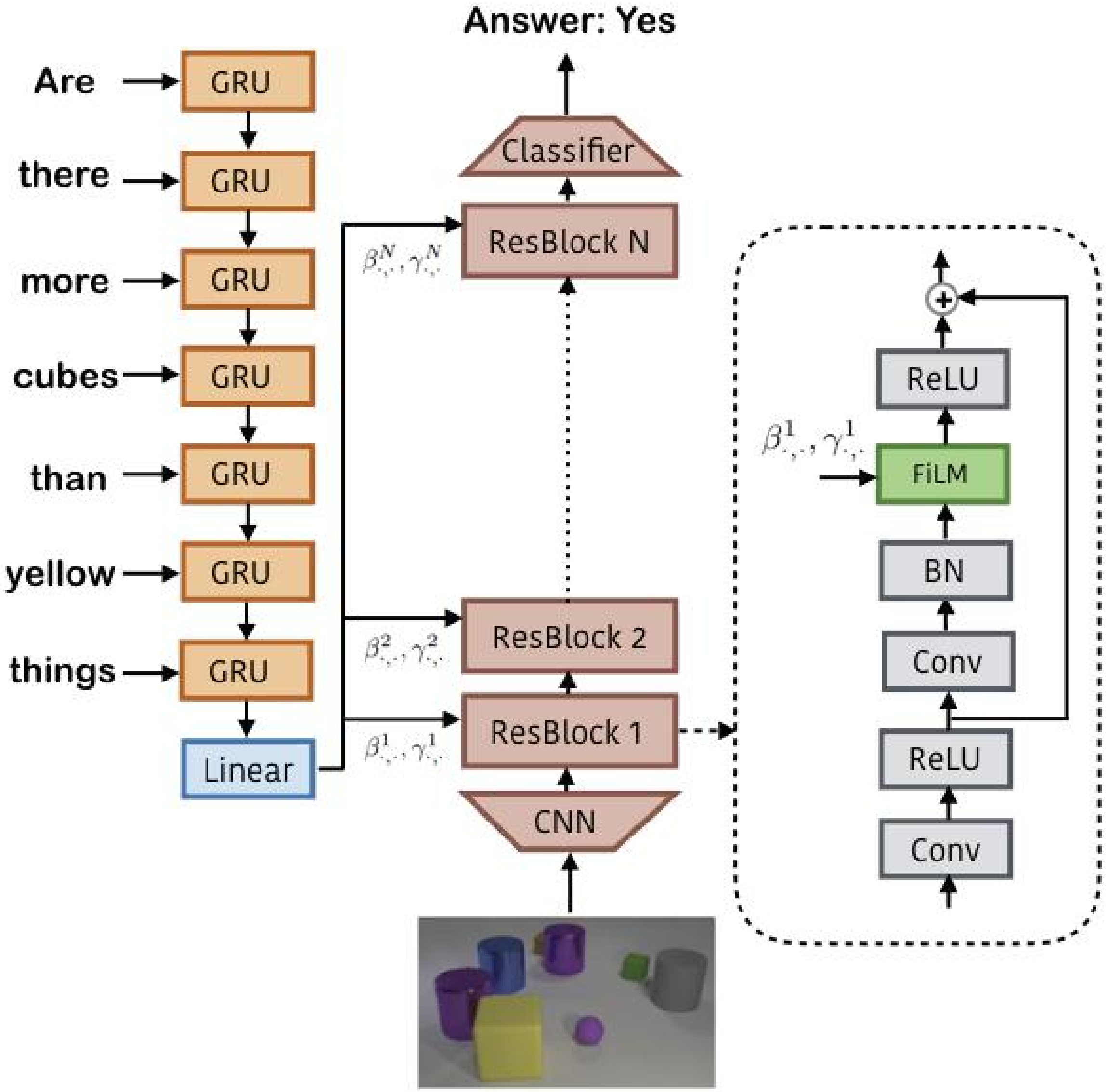
—— Sebastian Pölsterl(B) , Tom Nuno Wolf, and Christian Wachinger

1. Motivation

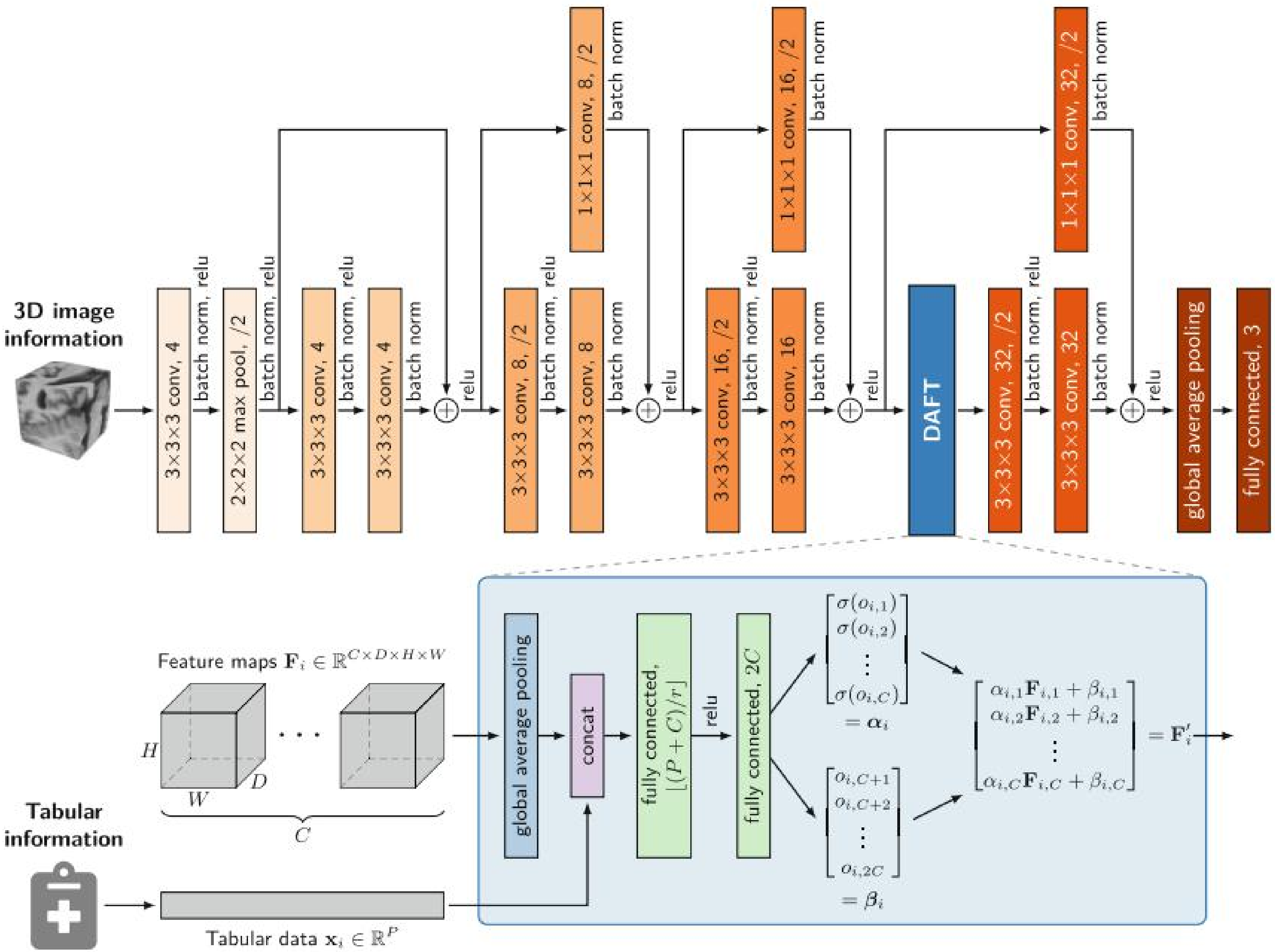
Prior work on diagnosing Alzheimer’s disease from magnetic resonance images of the brain established that convolutional neural networks (CNNs) can leverage the high-dimensional image information for classifying patients. However, little research focused on how these models can utilize the usually low-dimensional tabular information, such as patient demographics or laboratory measurements.

Introduce the Dynamic Affine Feature Map Transform (DAFT), a general-purpose module for CNNs that dynamically rescales and shifts the feature maps of a convolutional layer, conditional on a patient’s tabular clinical information

2. Related work – FILM



3. Methods (task-level regularization)



4. Experiments

Table 1. Dataset statistics.

Task	Subjects	Age	Male	Diagnosis
Diagnosis	1341	73.9 ± 7.2	51.8%	Dementia (19.6%), MCI (40.1%), CN (40.3%)
Progression	755	73.5 ± 7.3	60.4%	Progressor (37.4%), median follow-up time 2.01 years

4.1 Compare with other methods

Table 2. Predictive performance for the diagnosis task (columns 4–5) and time-to-dementia task (columns 6–7). Values are mean and standard deviation across 5 folds. Higher values are better. I indicates the use of image data, T of tabular data, with L/NL denoting a linear/non-linear model.

	I	T	Balanced accuracy		Concordance index	
			Validation	Testing	Validation	Testing
Linear model	✗	L	0.571 ± 0.024	0.552 ± 0.020	0.726 ± 0.040	0.719 ± 0.077
ResNet	✓	–	0.568 ± 0.015	0.504 ± 0.016	0.669 ± 0.032	0.599 ± 0.054
Linear model /w ResNet features	✓	L	0.585 ± 0.050	0.559 ± 0.053	0.743 ± 0.026	0.693 ± 0.044
Concat-1FC	✓	L	0.630 ± 0.043	0.587 ± 0.045	0.755 ± 0.025	0.729 ± 0.086
Concat-2FC	✓	NL	0.633 ± 0.036	0.576 ± 0.036	0.769 ± 0.026	0.725 ± 0.039
1FC-Concat-1FC	✓	NL	0.632 ± 0.020	0.591 ± 0.024	0.759 ± 0.035	0.723 ± 0.056
Duanmu et al. [3]	✓	NL	0.634 ± 0.015	0.578 ± 0.019	0.733 ± 0.031	0.706 ± 0.086
FiLM [25]	✓	NL	0.652 ± 0.033	0.601 ± 0.036	0.750 ± 0.025	0.712 ± 0.060
DAFT	✓	NL	0.642 ± 0.012	$\mathbf{0.622 \pm 0.044}$	0.753 ± 0.024	$\mathbf{0.748 \pm 0.045}$

4.2 Ablation study

Configuration	Balanced accuracy	Concordance index
Before Last ResBlock	0.598 ± 0.038	0.749 ± 0.052
Before Identity-Conv	0.616 ± 0.018	0.745 ± 0.036
Before 1st ReLU	0.622 ± 0.024	0.713 ± 0.085
Before 2nd Conv	0.612 ± 0.034	0.759 ± 0.052
$\boldsymbol{\alpha}_i = \mathbf{1}$	0.581 ± 0.053	0.743 ± 0.015
$\boldsymbol{\beta}_i = \mathbf{0}$	0.609 ± 0.024	0.746 ± 0.057
$\sigma(x) = \text{sigmoid}(x)$	0.600 ± 0.025	0.756 ± 0.064
$\sigma(x) = \tanh(x)$	0.600 ± 0.025	0.770 ± 0.047
Proposed	0.622 ± 0.044	0.748 ± 0.045