Supplementary Information

No gain from attention or graph modules in matched ResNet-50, ResNet-18, EfficientNet-B0 and MobileNetV2 baselines for nine-class chest radiographs

Kui Wang a,*, Rui Li b, Yinzhen Li c, Yazhou Zhang d, Licong Wang e, Xiaolei Tong a, Qianwen Wang a, Yong Feng a, Dongsheng Rui a, Shangzhi Xu a, Xianghui Zhang a,*

a School of Public Health, Shihezi University, Beier Road, Shihezi, Xinjiang 832003, China

b Shihezi People's Hospital, 45 Beisan Road, Shihezi, Xinjiang 832000, China

c First Affiliated Hospital of Shihezi University, 107 Beier Road, Shihezi, Xinjiang 832008, China

d Manas County People's Hospital, 26 Fenghuang East Road, Manas County, Changji, Xinjiang 832200, China

e Baiyin City Center for Disease Control and Prevention, No. 477 Chang'an Road, Fangzhi Road Sub-district, Baiyin, Gansu 730900, China

* Corresponding authors: Kui Wang (1136127654@qq.com; Tel 0086-13579745760); Xianghui Zhang (zxh818@shzu.edu.cn)

Table S1 | Per-class AUC as mean ± SD across the ten seeds (42–51)

Source: recomputed from the stored per-seed predictions of the 60 runs (code/tb9_single_label/output/predictions/, code/tb9_single_label_perblock/output/predictions/), released with the analysis scripts in the repository (Code availability).

Class	baseline	SE	CBAM	BAM	GCN	GAT
No finding	0.7276 ± 0.0059	0.7272 ± 0.0062	0.7243 ± 0.0054	0.7286 ± 0.0076	0.7247 ± 0.0049	0.7198 ± 0.0043
Atelectasis	0.8057 ± 0.0078	0.8062 ± 0.0069	0.8053 ± 0.0057	0.8056 ± 0.0058	0.8059 ± 0.0055	0.8062 ± 0.0063
Cardiomegaly	0.9382 ± 0.0045	0.9388 ± 0.0038	0.9374 ± 0.0055	0.9391 ± 0.0022	0.9384 ± 0.0030	0.9392 ± 0.0037
Effusion	0.8689 ± 0.0054	0.8696 ± 0.0037	0.8655 ± 0.0036	0.8683 ± 0.0050	0.8703 ± 0.0033	0.8701 ± 0.0030
Infiltration	0.7202 ± 0.0064	0.7244 ± 0.0113	0.7205 ± 0.0048	0.7272 ± 0.0044	0.7201 ± 0.0056	0.7210 ± 0.0063
Mass	0.7966 ± 0.0066	0.7927 ± 0.0083	0.7924 ± 0.0110	0.7969 ± 0.0072	0.7899 ± 0.0074	0.7853 ± 0.0107
Nodule	0.7620 ± 0.0077	0.7602 ± 0.0071	0.7570 ± 0.0108	0.7658 ± 0.0069	0.7593 ± 0.0057	0.7439 ± 0.0089
Pneumonia	0.7012 ±	0.6991 ±	0.6934 ±	0.6962 ±	0.6980 ±	0.6849 ±

	0.0153	0.0136	0.0165	0.0152	0.0130	0.0115
Pneumothorax	0.8702 ± 0.0044	0.8734 ± 0.0054	0.8696 ± 0.0062	0.8711 ± 0.0053	0.8726 ± 0.0043	0.8722 ± 0.0040

Note: SD is the sample standard deviation across seeds (ddof = 1). The ensemble (ten-seed probability average) per-class values reported in the main Table 3 are systematically 0.01–0.04 higher than the per-seed means, as expected for an ensemble; both are reproduced from the same prediction files.

Table S2 | Per-seed macro-AUC of all eight trained configurations

Configuration	Seeds 42–51 (macro-AUC)	Mean ± SD
baseline	0.7958, 0.8021, 0.7985, 0.7972, 0.7963, 0.7990, 0.7990, 0.7987, 0.8029, 0.8002	0.7990 ± 0.0023
SE (post-backbone)	0.7973, 0.7969, 0.7936, 0.7993, 0.7987, 0.7991, 0.7972, 0.8010, 0.8068, 0.8007	0.7991 ± 0.0035
CBAM (post-backbone)	0.7953, 0.7960, 0.7898, 0.7948, 0.7986, 0.7990, 0.7952, 0.7973, 0.7981, 0.7974	0.7962 ± 0.0027
BAM (post-backbone)	0.8056, 0.7972, 0.8040, 0.7998, 0.8006, 0.8024, 0.8010, 0.7978, 0.7918, 0.7986	0.7999 ± 0.0039
GCN (post-backbone)	0.8019, 0.8012, 0.7955, 0.7980, 0.8005, 0.7928, 0.7993, 0.7997, 0.7945, 0.7933	0.7977 ± 0.0034
GAT (post-backbone)	0.7947, 0.7934, 0.7946, 0.7896, 0.7917, 0.7928, 0.7954, 0.7976, 0.7939, 0.7927	0.7936 ± 0.0022
SE (per-block)	0.7726, 0.7721, 0.7767, 0.7781, 0.7761, 0.7495, 0.7725, 0.7722, 0.7756, 0.7701	0.7716 ± 0.0082
CBAM (per-block)	0.7017, 0.7170, 0.7243, 0.7357, 0.7206, 0.7368, 0.7320, 0.7607, 0.7070, 0.7504	0.7286 ± 0.0184

Note: macro-AUC is the unweighted mean of the nine one-versus-rest AUCs. The baseline's secondary metrics, which accompany the macro-AUC quoted in the main text, are a

weighted AUC of 0.8039 ± 0.0024 (ensemble 0.8184) and an accuracy of 0.4324 ± 0.0046 (ensemble 0.4570) across the ten seeds (42–51), recomputed from the same stored predictions.

Table S3 | Training-set size: per-class cap raised to the limit of availability

ResNet-50, cap_train = 4,800 (Normal 4,800, Pneumonia 258, Cardiomegaly 874), giving 22,898 training and 2,383 validation images; validation is unchanged from the main benchmark and identical across all ten seeds. Five configurations by ten seeds (42–51). Source: recomputed by the authors from the 50 stored per-seed prediction files (code/tb9_multibackbone/output/scale/predictions/).

Configuration	macro-AUC (mean ± SD)	Δ vs baseline	95% CI of Δ	Seeds negative	Wilcoxon p	TOST (±0.01)	MDE
baseline	0.8112 ± 0.0019	n/a	n/a	n/a	n/a	n/a	n/a
SE, post-backbone	0.8125 ± 0.0019	+0.0012	[−0.0002, +0.0025]	3/10	0.1309	equivalent	0.0021
CBAM, post-backbone	0.8095 ± 0.0016	−0.0018	[−0.0035, −0.0005]	8/10	0.0195	equivalent	0.0024
SE, per-block	0.7795 ± 0.0114	−0.0317	[−0.0388, −0.0250]	10/10	0.0020	worse	0.0105
CBAM, per-block	0.7476 ± 0.0138	−0.0636	[−0.0713, −0.0559]	10/10	0.0020	worse	0.0117

Note: Δ is the paired per-seed difference against the baseline; the 95% CI is a 5,000-resample bootstrap of those paired differences; TOST uses the pre-specified ±0.01 margin; MDE is the minimum detectable effect at 80% power for n = 10. Position contrasts (per-block minus post-backbone) were −0.0330 (95% CI [−0.0401, −0.0264]) for SE and −0.0619 (95% CI [−0.0706, −0.0529]) for CBAM, both negative in 10 of 10 seeds (p = 0.0020). Benjamini–Hochberg control across the four comparisons against the baseline rejects CBAM, SE per-block and CBAM per-block, with thresholds 0.0125, 0.025, 0.0375 and 0.050 for the ordered p values; the largest rejected comparison has p = 0.0195 against its threshold of 0.0375. CBAM is equivalent within the margin and negative by the paired test, and both readings are reported.

Table S4 | Multi-backbone: ResNet-18, EfficientNet-B0 and MobileNetV2

Eight configurations by ten seeds (42–51) per backbone on the same data split and schedule as the main benchmark; the first half of the feature hierarchy is frozen for these backbones,

mirroring the ResNet-50 freeze proportionally rather than matching it on parameter count. Source: recomputed from the 240 stored per-seed prediction files (code/tb9_multibackbone/output/backbone/predictions/).

(a) Baselines

Backbone	macro-AUC (mean ± SD)
ResNet-50 (main benchmark)	0.7990 ± 0.0023
ResNet-18	0.7866 ± 0.0026
EfficientNet-B0	0.7890 ± 0.0022
MobileNetV2	0.7872 ± 0.0028

(b) Δ against each backbone's own baseline

Configuration	ResNet-18	EfficientNet-B0	MobileNetV2
SE, post-backbone	+0.0021 (1/10, p = 0.0039)	+0.0022 (3/10, p = 0.0371)	+0.0027 (4/10, p = 0.2754)
CBAM, post-backbone	+0.0003 (5/10, p = 0.8457)	+0.0020 (2/10, p = 0.0371)	+0.0001 (4/10, p = 0.7695)
BAM, post-backbone	−0.0005 (7/10, p = 0.4316)	+0.0007 (5/10, p = 0.6250)	−0.0014 (9/10, p = 0.0840)
GCN, post-backbone	−0.0020 (7/10, p = 0.0840)	−0.0014 (6/10, p = 0.1934)	−0.0014 (7/10, p = 0.1309)
GAT, post-backbone	−0.0005 (5/10, p = 0.6953)	+0.0012 (4/10, p = 0.4316)	−0.0003 (6/10, p = 0.9219)
SE, per-block	+0.0010 (4/10, p = 0.3223)	+0.0019 (4/10, p = 0.0840)	−0.0058 (8/10, p = 0.0098)
CBAM, per-block	−0.0018 (7/10, p = 0.1055)	−0.0168 (10/10, p = 0.0020)	−0.0169 (10/10, p = 0.0020)

Note: parentheses give the number of seeds with a negative difference and the paired Wilcoxon p. Of the 21 module-versus-baseline comparisons, 19 fall inside the ±0.01 equivalence margin and 2 are worse (CBAM per-block on EfficientNet-B0 and MobileNetV2); none is better. Benjamini–Hochberg control across the seven comparisons of each backbone rejects SE on ResNet-18 in the positive direction, CBAM per-block on EfficientNet-B0, and SE per-block together with CBAM per-block on MobileNetV2. Two features of this comparison are noted in the main text: EfficientNet-B0 already contains squeeze-and-excitation inside its MBConv blocks, so its baseline includes channel attention, and the freeze is proportional rather than parameter-matched.

(c) Position contrasts (per-block minus post-backbone)

Backbone	SE	CBAM
ResNet-50 (main benchmark)	−0.0330 (10/10, p = 0.0020)	−0.0619 (10/10, p = 0.0020)
ResNet-18	−0.0011 (6/10, p = 0.3223)	−0.0020 (8/10, p = 0.0488)
EfficientNet-B0	−0.0003 (4/10, p = 0.6953)	−0.0188 (10/10, p =

		0.0020)
MobileNetV2	−0.0085 (9/10, p = 0.0039)	−0.0171 (10/10, p = 0.0020)

(b) Ensemble prediction of the ten seeds

Class	baseline	SE	CBAM	BAM	GCN	GAT
No finding	0.7403	0.7410	0.7360	0.7407	0.7372	0.7328
Atelectasis	0.8203	0.8206	0.8213	0.8217	0.8231	0.8207
Cardiomegaly	0.9480	0.9492	0.9481	0.9494	0.9479	0.9476
Effusion	0.8817	0.8803	0.8779	0.8787	0.8835	0.8797
Infiltration	0.7347	0.7410	0.7379	0.7420	0.7346	0.7358
Mass	0.8150	0.8141	0.8124	0.8155	0.8092	0.8044
Nodule	0.7753	0.7748	0.7731	0.7793	0.7740	0.7592
Pneumonia	0.7223	0.7136	0.7101	0.7135	0.7177	0.6992
Pneumothorax	0.8870	0.8863	0.8847	0.8871	0.8891	0.8859

Ensemble per-class AUC (one-versus-rest). Per-seed mean ± SD values are given in panel (a); the per-class standard deviation ranges from 0.002 to 0.017 depending on class size. The best-to-worst spread across the six models is at most 0.0082 AUC for six of the nine classes, but reaches 0.0111 for mass, 0.0201 for nodule and 0.0231 for pneumonia. These three are the classes with the fewest images and the largest seed-to-seed variability, so the residual spread sits where the estimate is least stable rather than being concentrated in any one module.

Table S5 | Patient-level split (no patient contributes images to both sets)

ResNet-50, patient-level split (StratifiedGroupKFold on Patient ID, fold 0, data seed 42): 9,548 training and 2,367 validation images, with no patient overlap in either the capped or the uncapped split. Five configurations by ten seeds (42–51). Source: recomputed by the authors from the 50 stored per-seed prediction files (code/tb9_multibackbone/output/patientsplit/predictions/); the split assignment is released as Source Data.

Configuration	macro-AUC (mean ± SD)	Δ vs baseline	95% CI of Δ	Seeds negative	Wilcoxon p	TOST (±0.01)	MDE
baseline	0.7893 ± 0.0033	n/a	n/a	n/a	n/a	n/a	n/a
SE, post-backbone	0.7905 ± 0.0016	+0.0012	[−0.0014, +0.0039]	3/10	0.3223	equivalent	0.0040
CBAM, post-backbone	0.7885 ±	−0.0007	[−0.0028, +0.0011]	7/10	0.5566	equivalent	0.0030

	0.0030						
SE, per-block	0.7653 ± 0.0050	−0.0240	[−0.0286, −0.0193]	10/10	0.0020	worse	0.0071
CBAM, per-block	0.6960 ± 0.0466	−0.0933	[−0.1226, −0.0658]	10/10	0.0020	worse	0.0432

Note: conventions as in Table S3. Position contrasts (per-block minus post-backbone) were −0.0252 (95% CI [−0.0280, −0.0225]) for SE and −0.0925 (95% CI [−0.1220, −0.0649]) for CBAM, both negative in 10 of 10 seeds (p = 0.0020). Benjamini–Hochberg control across the four comparisons against the baseline rejects SE per-block and CBAM per-block only, with the smallest p-value sitting at the floor imposed by n = 10. Under the image-level split of the main benchmark the baseline was 0.7990 ± 0.0023 and the small negative effect of post-backbone CBAM reached significance (Δ −0.0028, p = 0.0195); under this patient-level split it did not (Δ −0.0007, p = 0.5566), which is consistent with the expectation that patient overlap attenuates differences between models. Because the patient-level split necessarily changes the validation set, its baseline is not directly comparable with the main benchmark's, and that comparison is between designs rather than paired; comparisons within each split remain paired seed by seed. Every CBAM per-block seed (lowest 0.6196) fell below the lowest baseline seed (0.7847), so the mean difference is not produced by an outlying seed.

Table S6 | Position contrasts between the three placements (ten seeds, identical data, schedule and seed set)

Contrast	Δ macro-AUC	95% CI of Δ	Seeds negative	Wilcoxon p	Significant under the position family?
SE in-branch vs SE post-backbone	+0.0012	[−0.0015, +0.0032]	2/10	0.1055	No
CBAM in-branch vs CBAM post-backbone	+0.0044	[+0.0022, +0.0065]	1/10	0.0137	Yes
SE in-branch vs SE per-block	+0.0287	[+0.0242, +0.0347]	0/10	0.0020	Yes
CBAM in-branch vs CBAM per-block	+0.0720	[+0.0602, +0.0835]	0/10	0.0020	Yes

The four placement contrasts form a secondary family analysed with Benjamini–Hochberg control at a false-discovery rate of 0.05 (thresholds 0.0125, 0.025, 0.0375 and 0.050); the three smallest p-values are rejected, and the 95% confidence interval of Δ is a 5,000-resample bootstrap of the paired per-seed differences, the same convention used in Table 4. The CBAM in-branch versus CBAM post-backbone difference is small (+0.0044) and has no counterpart for SE ($p = 0.1055$), and we therefore read the two non-degenerate placements as indistinguishable rather than as evidence that the canonical placement is better. The two contrasts against the per-block placement show that output gating, not position, accounts for the degradation.