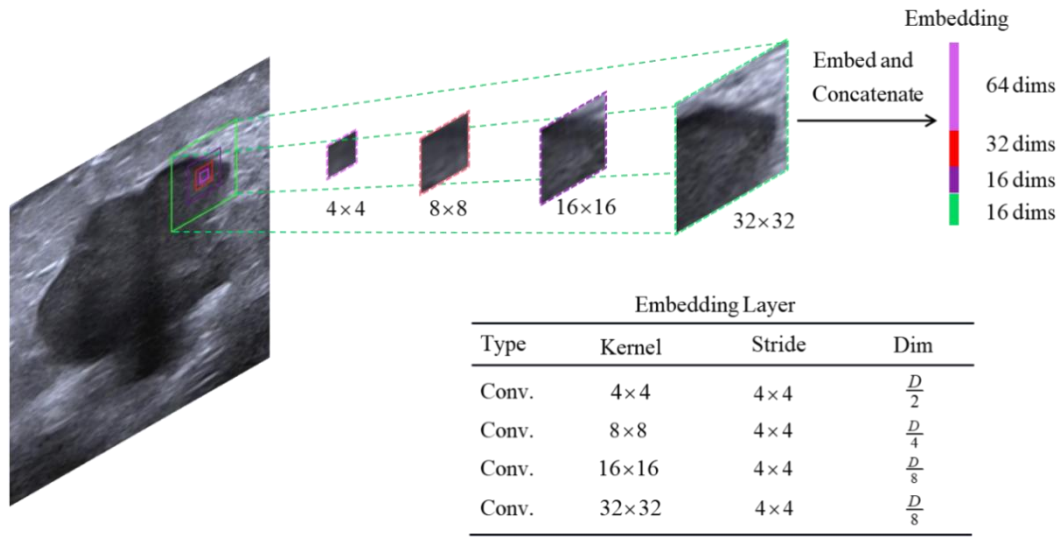
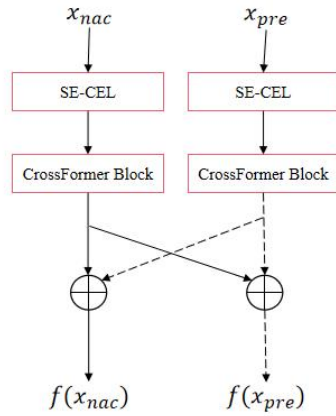


Supplementary Figure 1. Squeeze-and-Excitation block [1]. For the image, the features are compressed in the channel direction to obtain $1 \times 1 \times C$ dimensional features, and during the training process, weighting is computed for the features and the weights are applied to the extracted features to enrich the feature semantics.



Supplementary Figure 2. SE-CEL module, which selects different convolution operations to obtain different output dimensions, so that cross-scale features can be obtained, and in the output phase of the module, the different dimensions are combined to obtain a total of $64+32+16+16$ for a total of 128 dimensional scale features[2].



Supplementary Figure 3. Dou-Ult Net module, which extracts nonlinear parametric features, the features are weighted to provide consistent positional information.

Supplementary Table 1. Characteristics of patients in training, validation and Test cohorts of initial-baseline patients

Characteristics	Training cohort n= 391(80%)	Validation cohort n= 49(10%)	<i>p</i> value	Test cohort 1 n= 51(10%)	<i>p</i> value	Multimodal Test cohort n=105 (100%)	<i>p</i> value
Age			0.694		0.778		0.620
Mean ± SD	53.08±9.33	53.63±9.25		52.69±8.99		52.57±8.996	
BMI			0.655		0.501		0.559
Mean ± SD	24.89±3.35	24.68±3.36		24.55±3.77		25.11±3.43	
Menopausal status			0.252		0.919		0.917
Pre/peri-Menopausal	110(28.1%)	10(20.4%)		14(27.5%)		29(27.6%)	
Post-Menopausal	281(71.9%)	39(79.6%)		37(72.5%)		76(72.4%)	
Tumor Size			0.609		0.905		0.801
Mean ± SD	29.67±8.81	28.98±8.85		29.51±9.13		29.60±18.671	
T stage			0.937		0.865		0.009
T 1	57(14.6%)	7(14.3%)		9(17.6%)		14(13.3%)	
T2	321(82.1%)	40(81.6%)		40(78.4%)		85(81.0%)	
T3	4(1.0%)	1(2.0%)		1(2.0%)		6(5.7%)	
T4	9(2.3%)	1(2.0%)		1(2.0%)		0(0%)	
Phenotypes			0.509		0.139		0.423
HR+/HER2 -	211(54.0%)	25(51.0%)		24(47.1%)		51(48.6%)	
Triple negative	53(13.6%)	10(20.4%)		4(7.8%)		11(10.5%)	
HR-/HER2 +	67(17.1%)	9(18.4%)		9(17.6%)		23(21.9%)	
HR+/HER2 +	60(15.3%)	5(10.2%)		14(27.5%)		20(19.0%)	
Receptor status							
ER			0.323		0.687		0.278
Negative	126(32.2%)	21(42.9%)		14(27.5%)		38(36.2%)	
Low positive	12(3.1%)	1(2.0%)		1(2.0%)		6(5.7%)	
Positive	253(64.7%)	27(55.1%)		36(70.6%)		61(58.1%)	
PR			0.166		0.545		0.819
Negative	159(40.7%)	25(51.0%)		23(45.1%)		44(41.9%)	
Positive	232(59.3%)	24(49.0%)		28(54.9%)		61(58.1%)	
HER2			0.488		0.190		0.092
Negative	104(26.6%)	17(34.7%)		10(19.6%)		18(17.1%)	
low expression	160(40.9%)	18(36.7%)		18(35.2%)		44(41.9%)	
Positive	127(32.5%)	14(28.6%)		23(45.1%)		43(41.0%)	
Ki-67(%)			0.600		0.245		0.392
Mean ± SD	31.55±21.17	29.86±21.81		28.69±15.67		29.60±18.671	
clinical TNM stage			0.919		0.834		0.004
I-II	333(85.2%)	42(85.7%)		44(86.3%)		77(73.3%)	
III	58(14.8%)	7(14.3%)		7(13.7%)		28(26.7%)	
Nodal status			0.565		0.527		0.583
Node Positive	367(93.9%)	47(95.9%)		49(96.1%)		97(92.4%)	
Node negative	24(6.1%)	2(4.1%)		2(3.9%)		8(7.6%)	
Chemotherapy regimen			0.076		0.276		0.418
Use of anti-HER2 agents in HER2-positive cancer	102(26.1%)	12(24.5%)		19(37.3%)		33(31.4%)	
Taxane - and platinium-based	12(3.1%)	5(%)		2(3.9%)		4(3.8%)	
Taxane-based	8(2.0%)	0(0.0%)		0(0.0%)		4(3.8%)	
Anthracycline combined with taxane	269(68.8%)	32(65.3%)		30(58.8%)		64(61.0%)	

BI-RADS			0.103		0.287		0.354
2, 3, 4a	3(0.8%)	2(4.1%)		1(2.0%)		2(1.9%)	
4b	22(5.6%)	2(4.1%)		4(7.8%)		5(4.8%)	
4c	141 (36.1%)	22(44.9%)		12(23.5%)		30(28.6%)	
5	225(57.5%)	23(46.9%)		34(66.7%)		68(64.8%)	
pCR status			0.962		0.928		0.303
Non-pCR	286(73.1%)	36(73.5%)		37(72.5%)		82(78.1%)	
pCR	105(26.9%)	13(26.5%)		14(27.5%)		23(21.9%)	

Note: pCR pathologic complete response; ER estrogen receptor; PR progesterone receptor; HER2, human epidermal growth factor receptor 2

Supplementary Table 2. Performance of six DL Models based on WSIs according to validation and test cohort 1.

	VGG16	ResNet50	DenseNet121	ViT	Swin-transformer	SE-CrossT
Validation cohort (n=49)						
ACC	0.705(0.670-0.726)	0.796(0.763-0.816)	0.796(0.726-0.811)	0.813(0.789-0.819)	0.784(0.738-0.809)	0.857(0.783-0.890)
AUC	0.511(0.470-0.527)	0.675(0.627-0.684)	0.728(0.698-0.769)	0.711(0.659-0.719)	0.734(0.609-0.791)	0.810(0.777-0.819)
SENS	0.709(0.630-0.713)	0.706(0.649-0.739)	0.710(0.653-0.741)	0.706(0.679-0.745)	0.710(0.682-0.727)	0.730(0.678-0.750)
SPEC	0.752(0.721-0.769)	0.755(0.701-0.773)	0.741(0.717-0.773)	0.761(0.740-0.800)	0.769(0.711-0.807)	0.788(0.726-0.818)
PPV	0.711(0.686-0.711)	0.751(0.672-0.788)	0.758(0.706-0.794)	0.751(0.677-0.757)	0.743(0.712-0.777)	0.751(0.680-0.779)
NPV	0.719(0.689-0.742)	0.732(0.690-0.754)	0.723(0.648-0.736)	0.746(0.690-0.782)	0.761(0.719-0.797)	0.751(0.712-0.781)
Test cohort 1(n= 51)						
ACC	0.723(0.684-0.733)	0.647(0.622-0.675)	0.745(0.689-0.749)	0.687(0.613-0.706)	0.725(0.665-0.731)	0.776(0.746-0.792)
AUC	0.583(0.621-0.689)	0.664(0.618-0.697)	0.637(0.578-0.655)	0.597(0.552-0.630)	0.651(0.606-0.685)	0.703(0.649-0.745)
SENS	0.660(0.600-0.688)	0.693(0.661-0.693)	0.658(0.599-0.660)	0.693(0.664-0.719)	0.727(0.678-0.728)	0.727(0.647-0.763)
SPEC	0.706(0.680-0.742)	0.704(0.646-0.710)	0.717(0.664-0.720)	0.733(0.684-0.748)	0.738(0.718-0.768)	0.738(0.689-0.759)
PPV	0.692(0.665-0.694)	0.732(0.705-0.749)	0.732(0.705-0.749)	0.723(0.687-0.726)	0.754(0.705-0.780)	0.724(0.680-0.725)
NPV	0.704(0.632-0.708)	0.718(0.684-0.743)	0.714(0.669-0.750)	0.722(0.682-0.732)	0.731(0.684-0.759)	0.729(0.680-0.737)

Note: data in parentheses are 95% confidence intervals. ACC accuracy; AUC area under the receiver; SENS sensitivity; SPEC specificity; PPV positive predictive value; NPV negative predictive value; ViT vision transformer.

Supplementary Table 3. Performance of six DL Models based on ultrasound imaging according to validation and test cohort 1.

	VGG16	ResNet50	DenseNet121	ViT	Swin-transformer	SE-CrossT
Validation cohort (n=49)						
ACC	0.837(0.793-0.858)	0.837(0.806-0.847)	0.796(0.772-0.820)	0.837(0.758-0.856)	0.830(0.754-0.845)	0.878(0.811-0.880)
AUC	0.700(0.606-0.702)	0.794(0.763-0.804)	0.790(0.743-0.827)	0.799(0.669-0.847)	0.737(0.663-0.739)	0.837(0.806-0.855)
SENS	0.864(0.814-0.866)	0.876(0.807-0.897)	0.883(0.820-0.890)	0.881(0.825-0.916)	0.866(0.822-0.897)	0.883(0.819-0.901)
SPEC	0.846(0.769-0.857)	0.859(0.799-0.887)	0.876(0.819-0.900)	0.859(0.803-0.884)	0.869(0.822-0.902)	0.882(0.815-0.890)
PPV	0.826(0.770-0.864)	0.824(0.760-0.848)	0.840(0.795-0.864)	0.853(0.796-0.881)	0.853(0.817-0.886)	0.859(0.803-0.887)
NPV	0.823(0.788-0.834)	0.860(0.814-0.861)	0.857(0.820-0.879)	0.857(0.820-0.859)	0.861(0.821-0.898)	0.879(0.816-0.884)
Test cohort 1(n=51)						
ACC	0.745(0.712-0.776)	0.764(0.707-0.797)	0.725(0.665-0.751)	0.764(0.716-0.787)	0.773(0.745-0.804)	0.823(0.751-0.829)
AUC	0.571(0.558-0.640)	0.712(0.690-0.735)	0.695(0.671-0.716)	0.678(0.664-0.710)	0.701(0.682-0.719)	0.817(0.775-0.836)
SENS	0.831(0.773-0.844)	0.835(0.798-0.865)	0.846(0.794-0.856)	0.859(0.821-0.881)	0.835(0.798-0.846)	0.859(0.805-0.869)
SPEC	0.811(0.736-0.816)	0.843(0.795-0.867)	0.816(0.759-0.838)	0.820(0.775-0.854)	0.847(0.783-0.884)	0.867(0.821-0.882)
PPV	0.808(0.729-0.822)	0.805(0.747-0.812)	0.821(0.769-0.839)	0.841(0.810-0.855)	0.853(0.783-0.877)	0.833(0.783-0.870)
NPV	0.808(0.760-0.845)	0.812(0.787-0.841)	0.832(0.761-0.871)	0.837(0.788-0.865)	0.837(0.778-0.838)	0.855(0.810-0.865)

Note: data in parentheses are 95% confidence intervals. ACC accuracy; AUC area under the receiver; SENS sensitivity; SPEC specificity; PPV positive predictive value; NPV negative predictive value; ViT vision transformer.

Supplementary Table 4. Ablation study result for the Squeeze-and-Excitation (SE) embedding block

	ACC	AUC	SENS	SPEC	PPV	NPV
Pathology						
Validation cohort (n=49)						
SE-CrossT	0.857(0.783-0.890)	0.810(0.777-0.819)	0.730(0.678-0.750)	0.788(0.726-0.818)	0.751(0.680-0.779)	0.751(0.712-0.781)
CrossT	0.813(0.761-0.818)	0.751(0.685-0.762)	0.741(0.711-0.753)	0.763(0.721-0.801)	0.755(0.686-0.778)	0.748(0.677-0.764)
Test cohort 1(n=51)						
SE-CrossT	0.776(0.746-0.792)	0.703(0.649-0.745)	0.727(0.647-0.763)	0.738(0.689-0.759)	0.724(0.680-0.725)	0.729(0.680-0.737)
CrossT	0.790(0.752-0.823)	0.720(0.694-0.733)	0.691(0.625-0.730)	0.721(0.694-0.759)	0.748(0.690-0.751)	0.718(0.692-0.745)
US						
Validation cohort (n=49)						
SE-CrossT	0.878(0.811-0.880)	0.837(0.806-0.855)	0.883(0.819-0.901)	0.882(0.815-0.890)	0.859(0.803-0.887)	0.879(0.816-0.884)
CrossT	0.837(0.792-0.872)	0.815(0.785-0.821)	0.898(0.840-0.913)	0.881(0.833-0.891)	0.859(0.821-0.874)	0.883(0.830-0.911)
Test cohort 1(n=51)						
SE-CrossT	0.823(0.751-0.829)	0.817(0.775-0.836)	0.859(0.805-0.869)	0.867(0.821-0.882)	0.833(0.783-0.870)	0.855(0.810-0.865)
CrossT	0.823(0.785-0.850)	0.795(0.756-0.815)	0.840(0.770-0.851)	0.847(0.808-0.864)	0.838(0.803-0.867)	0.827(0.763-0.828)
Multimodal						
Validation cohort(n=49)						
SE-CrossT	0.918(0.838-0.924)	0.851(0.806-0.855)	0.896(0.833-0.918)	0.888(0.861-0.895)	0.866(0.819-0.891)	0.885(0.830-0.896)
CrossT	0.857(0.820-0.887)	0.841(0.768-0.852)	0.879(0.842-0.913)	0.896(0.820-0.896)	0.871(0.810-0.872)	0.883(0.820-0.913)
Test cohort 1(n=51)						
SE-CrossT	0.882(0.854-0.920)	0.851(0.791-0.856)	0.878(0.828-0.893)	0.855(0.826-0.868)	0.840(0.775-0.864)	0.864(0.789-0.884)
CrossT	0.843(0.823-0.855)	0.809(0.773-0.840)	0.859(0.813-0.866)	0.841(0.808-0.879)	0.840(0.790-0.864)	0.848(0.827-0.862)

Note: data in parentheses are 95% confidence intervals. ACC accuracy; AUC area under the receiver; SENS sensitivity; SPEC specificity; PPV positive predictive value; NPV negative predictive value; US ultrasound.

Supplementary Table 5. Performance of six DL Models based on Multimodal according to multifocal lesions test cohort. Comparison with state-of-the-art networks.

multifocal lesions Test cohort(n=105)	VGG16	ResNet50	DenseNet121	ViT	Swin-transformer	SE-CrossT
ACC	0.762(0.738-0.806)	0.791(0.743-0.798)	0.791(0.751-0.824)	0.800(0.778-0.821)	0.848(0.834-0.879)	0.848(0.808-0.858)
AUC	0.487(0.464-0.503)	0.535(0.513-0.566)	0.543(0.496-0.553)	0.600(0.533-0.630)	0.713(0.668-0.714)	0.713(0.649-0.714)
SENS	0.786(0.760-0.791)	0.789(0.747-0.792)	0.794(0.726-0.825)	0.793(0.749-0.807)	0.793(0.778-0.818)	0.794(0.727-0.815)
SPEC	0.761(0.717-0.764)	0.811(0.758-0.853)	0.781(0.730-0.814)	0.802(0.769-0.804)	0.816(0.772-0.842)	0.803(0.789-0.826)
PPV	0.761(0.735-0.778)	0.793(0.755-0.840)	0.786(0.753-0.791)	0.793(0.751-0.836)	0.786(0.772-0.792)	0.815(0.780-0.820)
NPV	0.730(0.686-0.757)	0.779(0.755-0.780)	0.785(0.752-0.825)	0.772(0.736-0.821)	0.795(0.728-0.832)	0.777(0.759-0.805)

Note: data in parentheses are 95% confidence intervals. ACC accuracy; AUC area under the receiver; SENS sensitivity; SPEC specificity; PPV positive predictive value; NPV negative predictive value; ViT vision transformer.

Supplementary References

1. Hu J, Shen L, Albanie S, Sun G,Wu E. *Squeeze-and-excitation networks*. IEEE Trans Pattern Anal Mach Intell, 2020. **42**(8). 2011-2023.
2. Wang W, Yao L, Chen L, Lin B, Cai D, He X,Liu W. *Crossformer:A versatile vision transformer hinging on cross-scale attention*. arXiv preprint arXiv. **2108.00154**.