Medical Vision Seminar

——Wei Lou

（MICCAI2021） TransFuse: Fusing Transformers and CNNs forMedical Image Segmentation

—— Yundong Zhang, Huiye Liu and Qiang Hu1

1. Motivation

Deep CNNs' drawbacks:

1. Low-level features are washed out by consecutive multiplications.
2. Local information is discarded, as the spatial resolution is reduced gradually.
3. Training parameter-heavy deep nets with small medical image datasets tends to be unstable and easily overfitting.

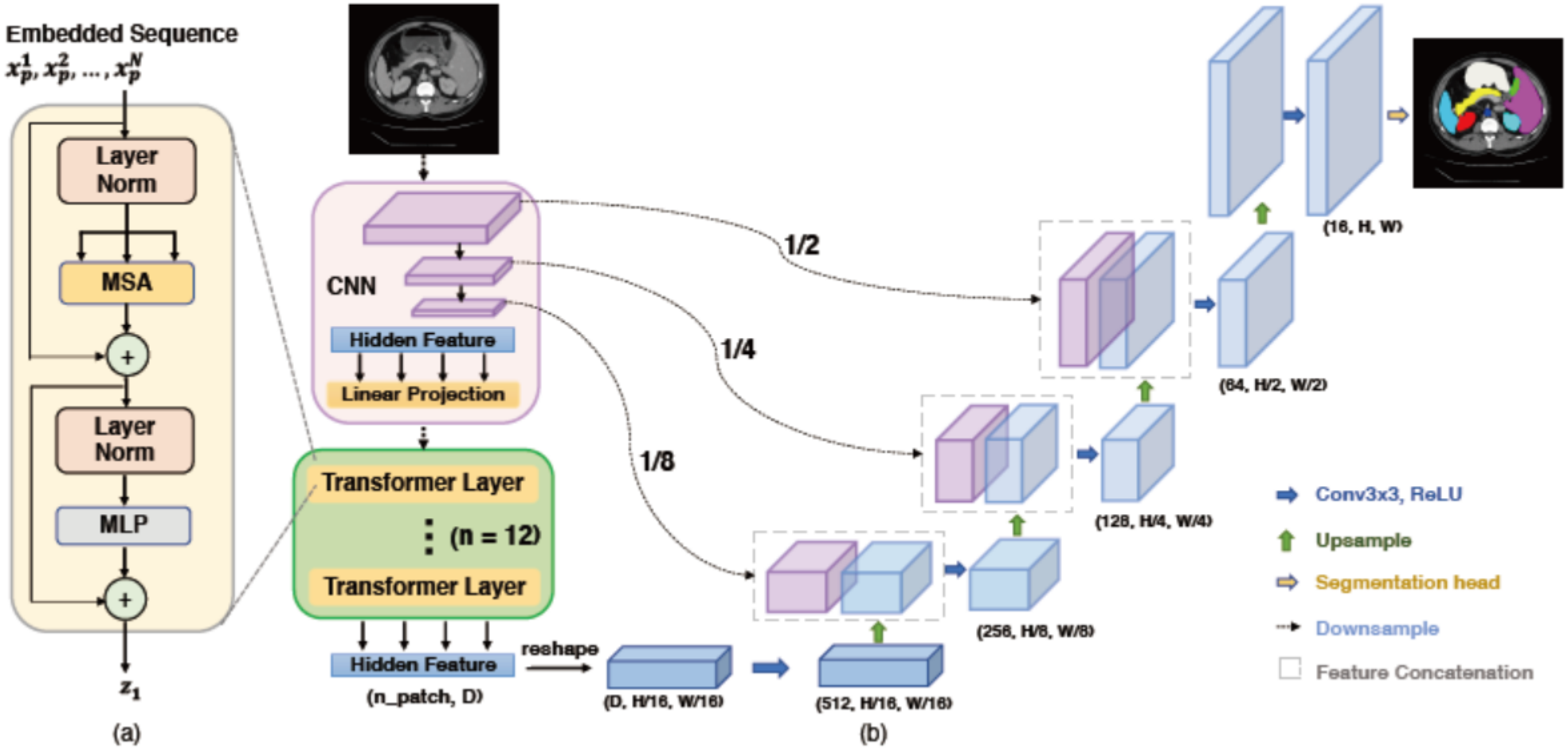
Typical transformers' drawbacks:

1. Limitations in capturing fine-grained details. Especially for medical images.

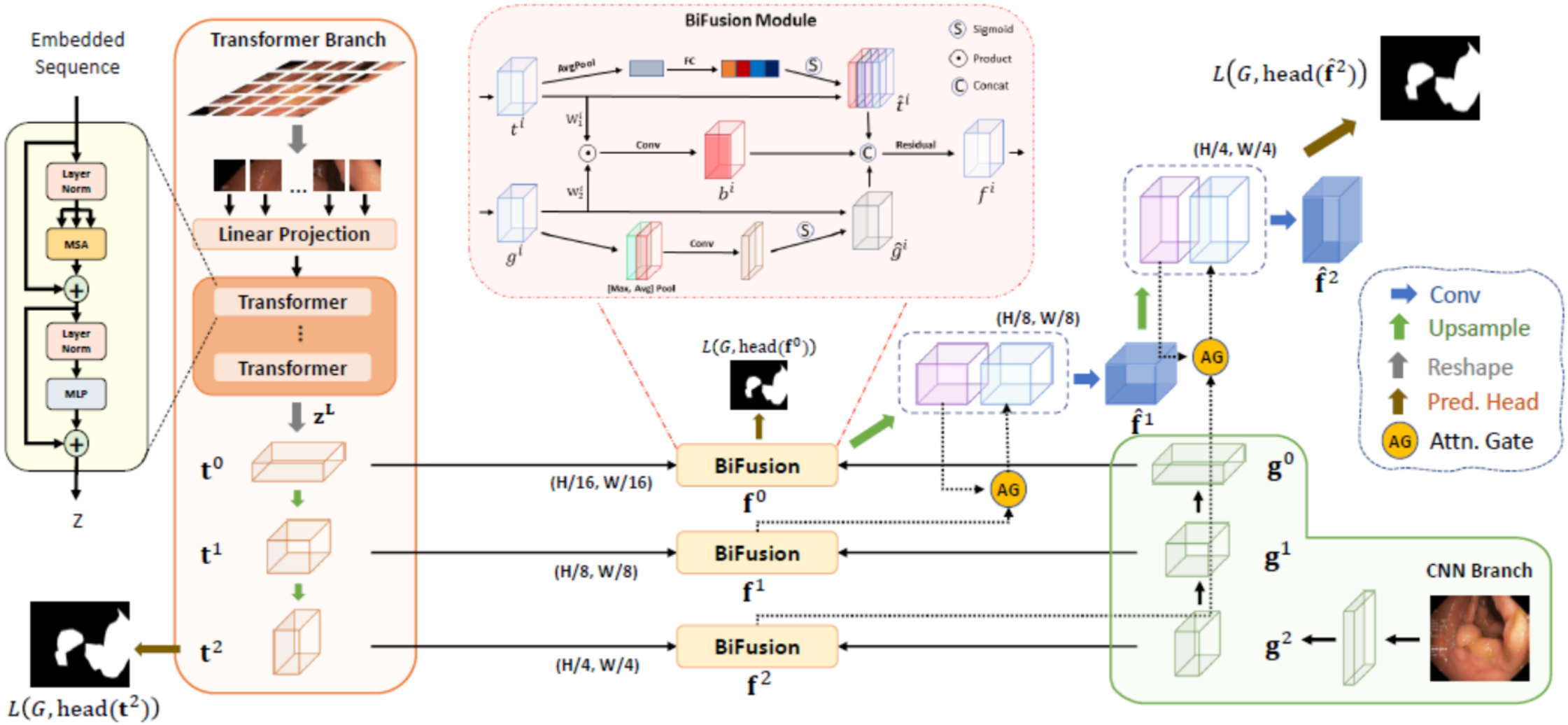
Aims:

A novel fusion module to combine these two operators.

2. Previous work——TransUnet



3. Methods (two parallel branches)



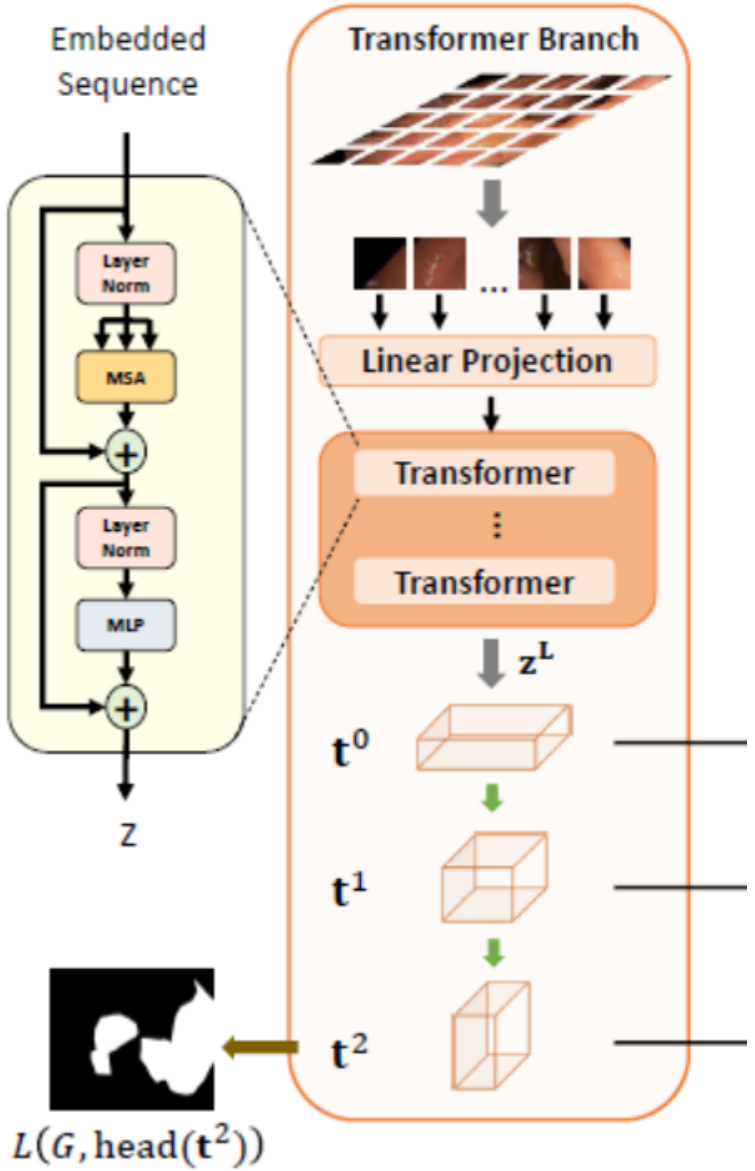
3.1 Transformer branch

- **Encoder**

1. The input images is evenly divided into N patches, size as $\frac{H}{16} * \frac{W}{16} * 3$, the patches are then flattened and passed into a linear embedding layer. Obtaining the raw embedding sequence e:
2. A learnable positional embedding is added to e.
3. The embeddings then are processed by several transformers.

- **Decoder**

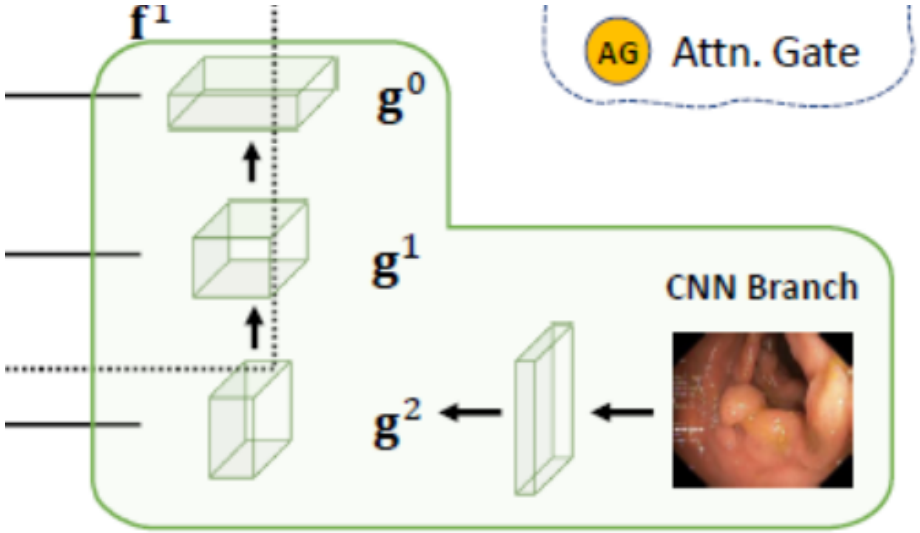
1. The output Z^L is reshaped back to $\frac{H}{16} * \frac{W}{16} * D_0$.
2. Using 2 upsampling convs to recover spatial resolution.



3.2 CNN branch (Reduce the deep layers)

Resnet50 but without the last block, results in 3 feature maps

$$(\frac{H}{16} * \frac{W}{16} * C_0, \frac{H}{8} * \frac{W}{8} * C_1, \frac{H}{4} * \frac{W}{4} * C_2)$$



3.3 BiFusion Module

$$\hat{\mathbf{t}}^i = \text{ChannelAttn}(\mathbf{t}^i) \qquad \hat{\mathbf{g}}^i = \text{SpatialAttn}(\mathbf{g}^i)$$

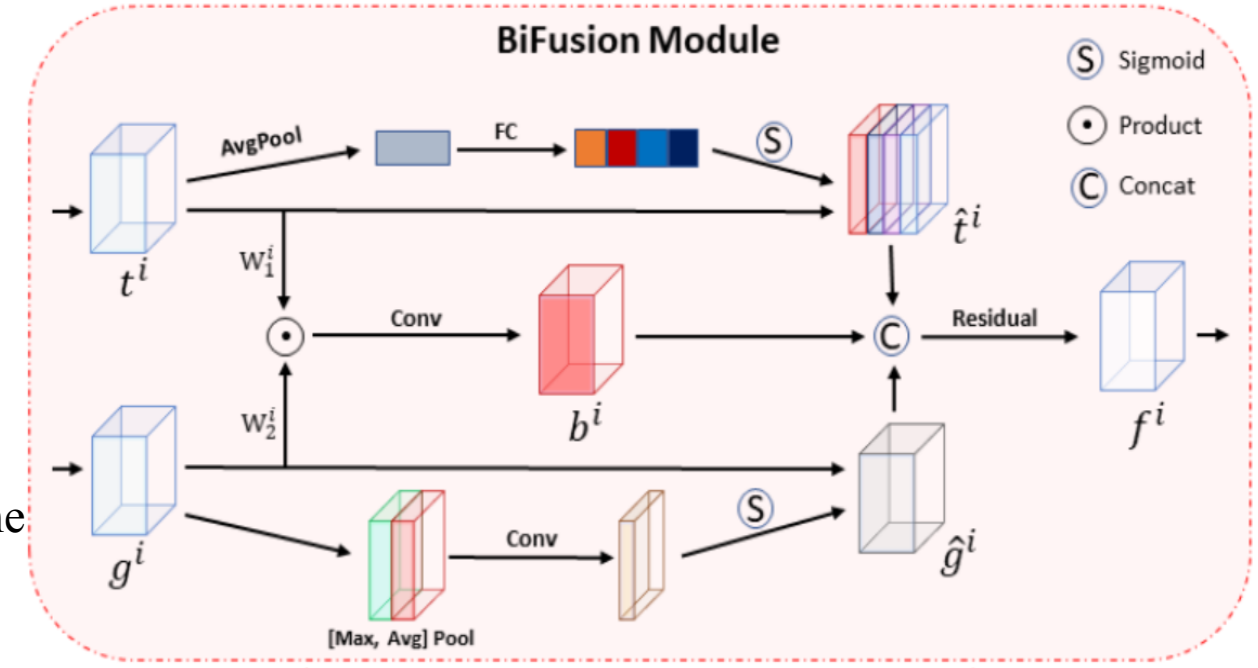
$$\hat{\mathbf{b}}^i = \text{Conv}(\mathbf{t}^i\mathbf{W}_1^i \odot \mathbf{g}^i\mathbf{W}_2^i) \qquad \mathbf{f}^i = \text{Residual}([\hat{\mathbf{b}}^i, \hat{\mathbf{t}}^i, \hat{\mathbf{g}}^i])$$

Channel attention:

$H * W * C_2$ -> $1 * 1 * C_2$ ->$1 * 1 * C_2$ corresponding to the attention value for each channel.

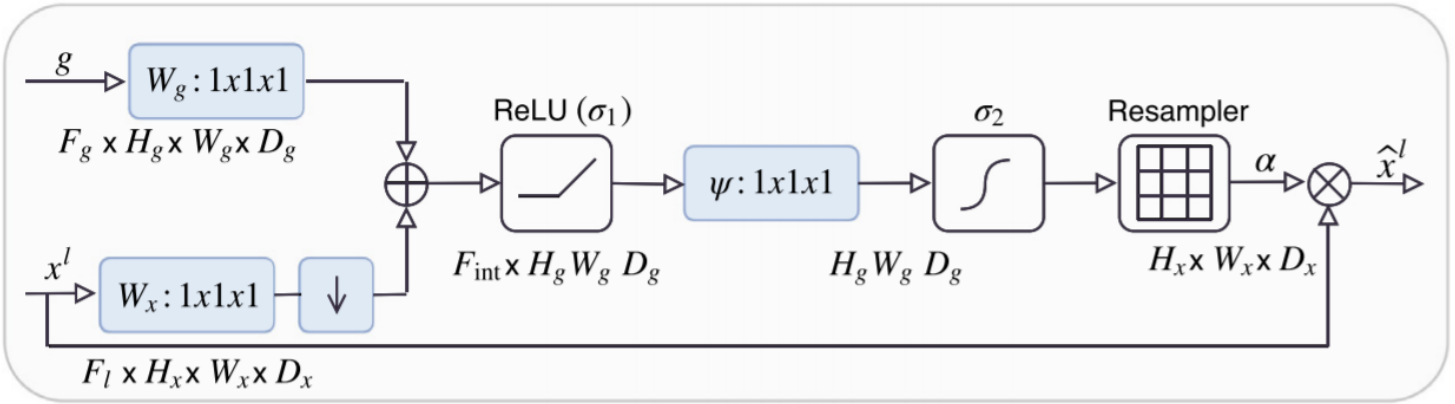
Spatial attention:

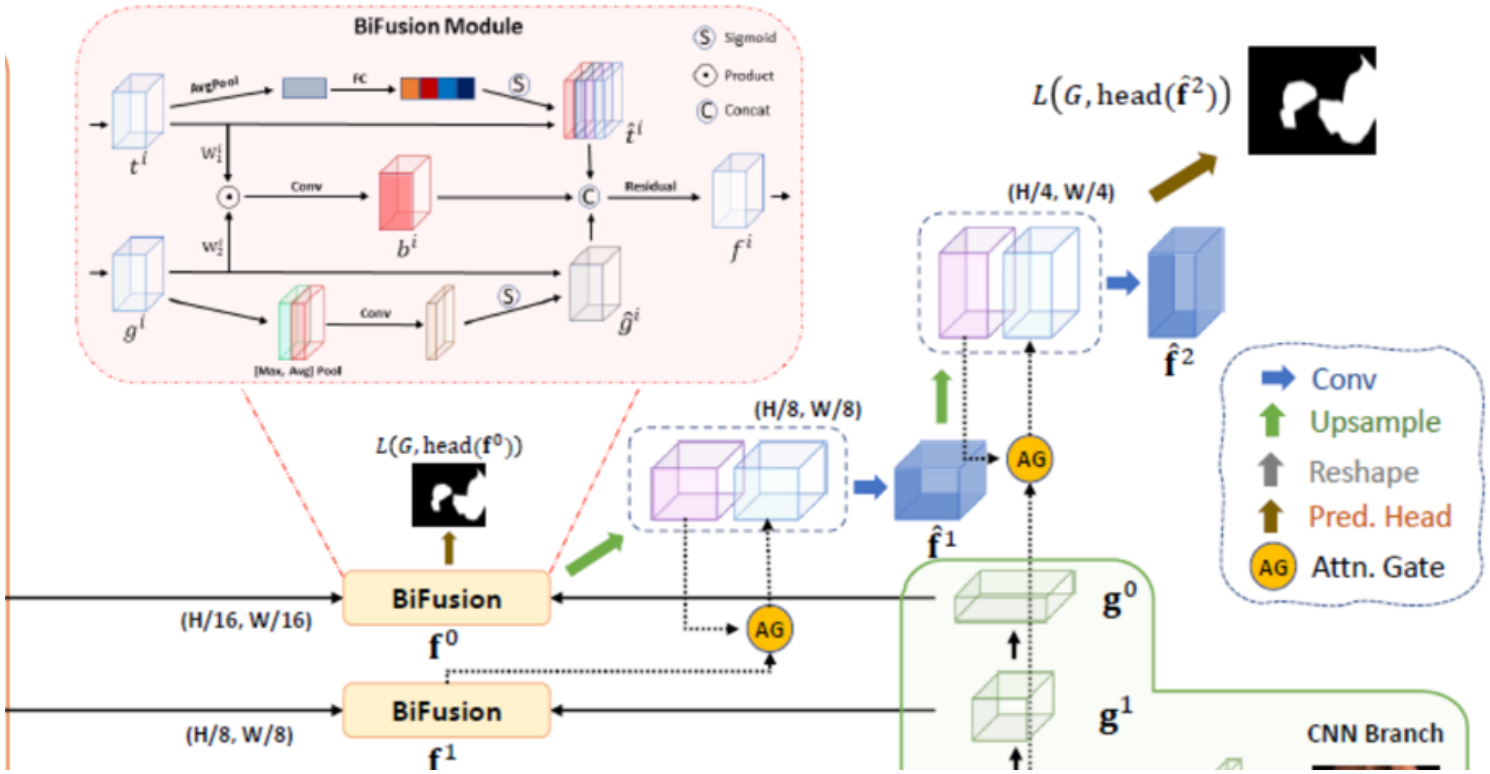
$H * W * C_2$ -> $2 * H * W$ ->$1 * H * W$ corresponding to the attention value for each pixel.



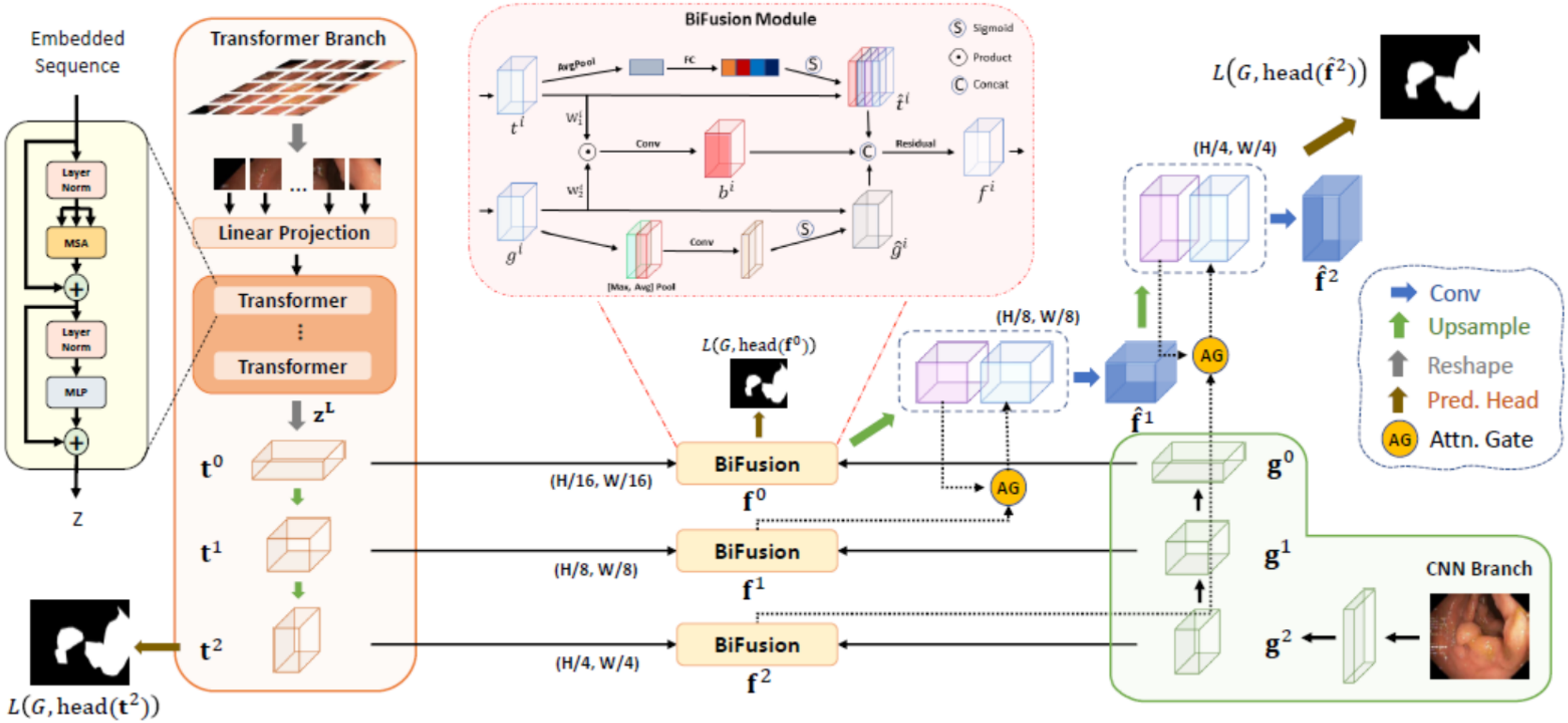
3.4 Additive attention GATE (AG)

$$\hat{\mathbf{f}}^{i+1} = \mathrm{Conv}([\mathrm{Up}(\hat{\mathbf{f}}^{i}), \mathrm{AG}(\mathbf{f}^{i+1}, \mathrm{Up}(\hat{\mathbf{f}}^{i}))])$$





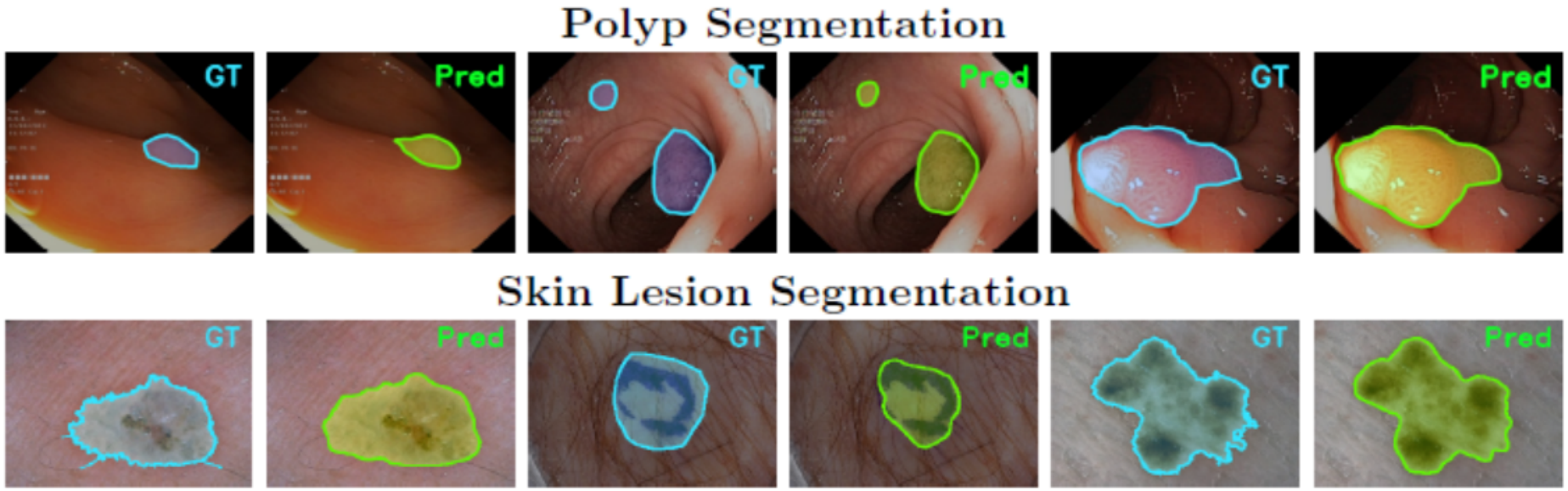
3.5 Three Losses



4. Experiments

4.1 Dataset

Four segmentation tasks with different imaging modalities, disease types, target objects, target sizes, etc. are considered:
1. Polyp Segmentation (息肉) 2. Skin Lesion Segmentation 3. Hip Segmentation (髋) 4. Prostate Segmentation (前列腺)



4.2 Results on polyp segmentation datasets

Table 1: Quantitative results on polyp segmentation datasets compared to previous SOTAs. The results of [4] is obtained by running the released code and we implement SETR-PUP. '-' means results not available.

Methods	Kvasir		ClinicDB		ColonDB		EndoScene		ETIS	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net [18]	0.818	0.746	0.823	0.750	0.512	0.444	0.710	0.627	0.398	0.335
U-Net++ [33]	0.821	0.743	0.794	0.729	0.483	0.410	0.707	0.624	0.401	0.344
ResUNet++ [13]	0.813	0.793	0.796	0.796	-	-	-	-	-	-
PraNet [8]	0.898	0.840	0.899	0.849	0.709	0.640	0.871	0.797	0.628	0.567
HarDNet-MSEG [11]	0.912	0.857	0.932	0.882	0.731	0.660	0.887	0.821	0.677	0.613
TransFuse-S	**0.918**	**0.868**	0.918	0.868	**0.773**	**0.696**	0.902	0.833	0.733	0.659
TransFuse-L	**0.918**	**0.868**	**0.934**	**0.886**	0.744	0.676	**0.904**	**0.838**	**0.737**	**0.661**
SETR-PUP [32]	0.911	0.854	0.934	0.885	0.773	0.690	0.889	0.814	0.726	0.646
TransUnet [4]	0.913	0.857	0.935	0.887	0.781	0.699	0.893	0.824	0.731	0.660
*TransFuse-L**	**0.920**	**0.870**	**0.942**	**0.897**	**0.781**	**0.706**	**0.894**	**0.826**	**0.737**	**0.663**

Table 2: Quantitative results on ISIC 2017 test set. Results with backbones use weights pretrained on ImageNet.

Methods	Backbones	Epochs	Jaccard	Dice	Accuracy
CDNN [31]	-	-	0.765	0.849	0.934
DDN [15]	ResNet-18	600	0.765	0.866	0.939
FrCN [1]	VGG16	200	0.771	0.871	0.940
DCL-PSI [3]	ResNet-101	150	0.777	0.857	0.941
SLSDeep [19]	ResNet-50	100	0.782	**0.878**	0.936
Unet++ [33]	ResNet-34	30	0.775	0.858	0.938
TransFuse-S	R34+DeiT-S	30	**0.795**	0.872	**0.944**

Table 3: Results on in-house hip dataset. All models use pretrained backbones from ImageNet and are of similar size ($\sim$ 26M). HD and ASD are measured in mm.

Methods	Pelvis		L-Femur		R-Femur	
	HD	ASD	HD	ASD	HD	ASD
Unet++ [33]	14.4	1.21	9.33	0.932	5.04	0.813
HRNetV2 [28]	14.2	1.13	6.36	0.769	5.98	0.762
TransFuse-S	**9.81**	**1.09**	**4.44**	**0.767**	**4.19**	**0.676**

Table 4: Quantitative results on prostate MRI segmentation. PZ, TZ stand for the two labeled classes (peripheral and transition zone) and performance (PZ, TZ and mean) is measure by dice score.

Methods	PZ	TZ	Mean	Params	Throughput
nnUnet-2d [12]	0.6285	0.8380	0.7333	29.97M	0.209s/vol
nnUnet-3d_full[12]	0.6663	0.8410	0.7537	44.80M	0.381s/vol
TransFuse-S	**0.6738**	**0.8539**	**0.7639**	**26.30M**	**0.192s/vol**

Table 5: Ablation study on parallel-in-branch design. Res: Residual.

Index	Backbones	Composition	Fusion	Kvasir	ColonDB
E.1	R34	Sequential	-	0.890	0.645
E.2	DeiT-S	Sequential	-	0.889	0.727
E.3	R34+DeiT-S	Sequential	-	0.908	0.749
E.4	R34+VGG16	Parallel	BiFusion	0.896	0.651
E.5	R34+DeiT-S	Parallel	Concat+Res	0.912	0.764
E.6	R34+DeiT-S	Parallel	BiFusion	0.918	0.773

Table 6: Ablation study on BiFusion module. Res: Residual; TFM: Transformer; Attn: Attention.

Fusion	Jaccard	Dice	Accuracy
Concat+Res	0.778	0.857	0.939
+CNN Spatial Attn	0.782	0.861	0.941
+TFM Channel Attn	0.787	0.865	0.942
+Dot Product	0.795	0.872	0.944

（ICCV2021） Fast Convergence of DETR with Spatially Modulated Co-Attention

—— Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, Hongsheng Li

1. Motivation

DETR's key drawbacks: Slow convergence.

The aim of this paper is to accelerate the convergence of DETR by using a Spatially Modulated Co-Attention (SMCA) mechanism. SMCA only replace the co-attention mechanism in the decoder, while keep other operations unchanged. By integrating multi-head and scale-selection attention designs into SMCA, our fully-fledged SMCA can achieve better performance compared to DETR

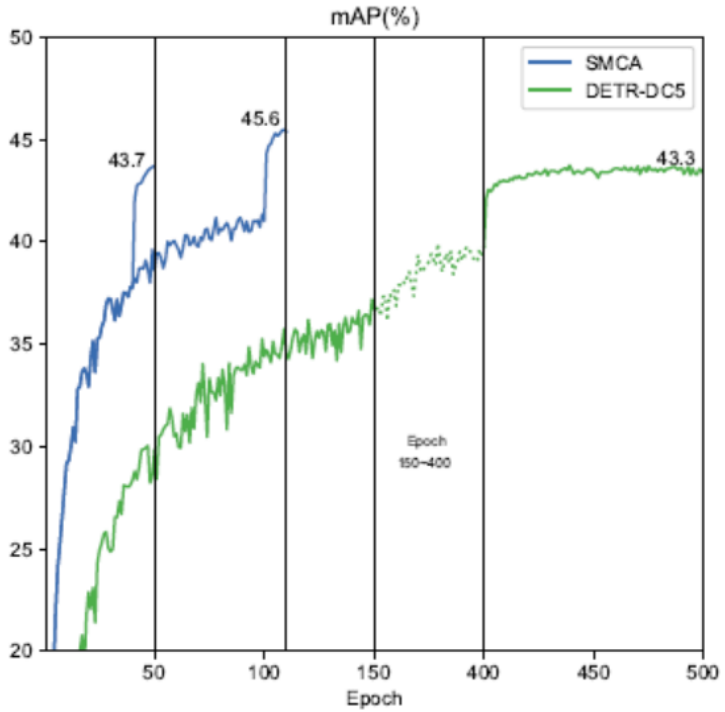
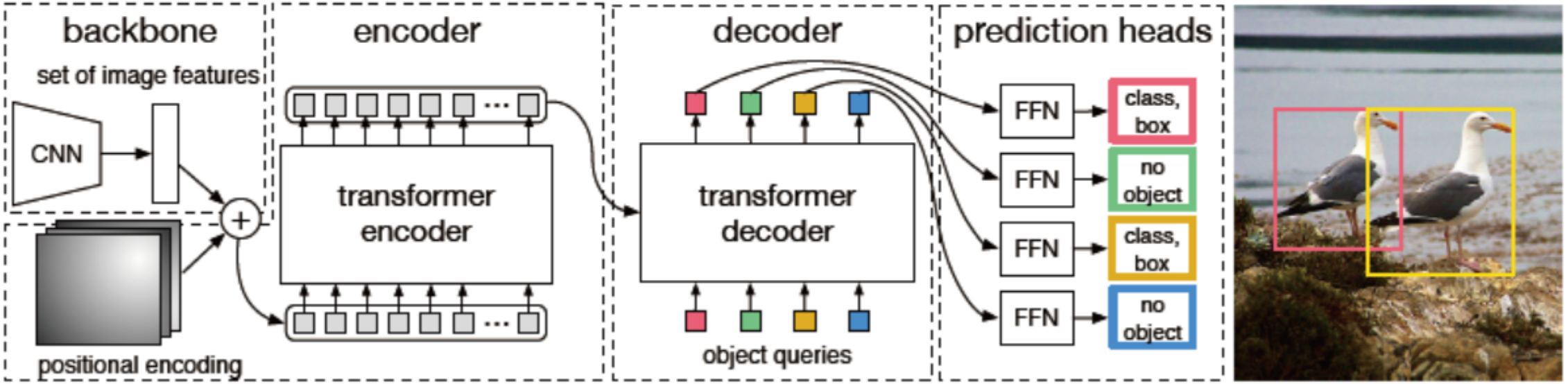
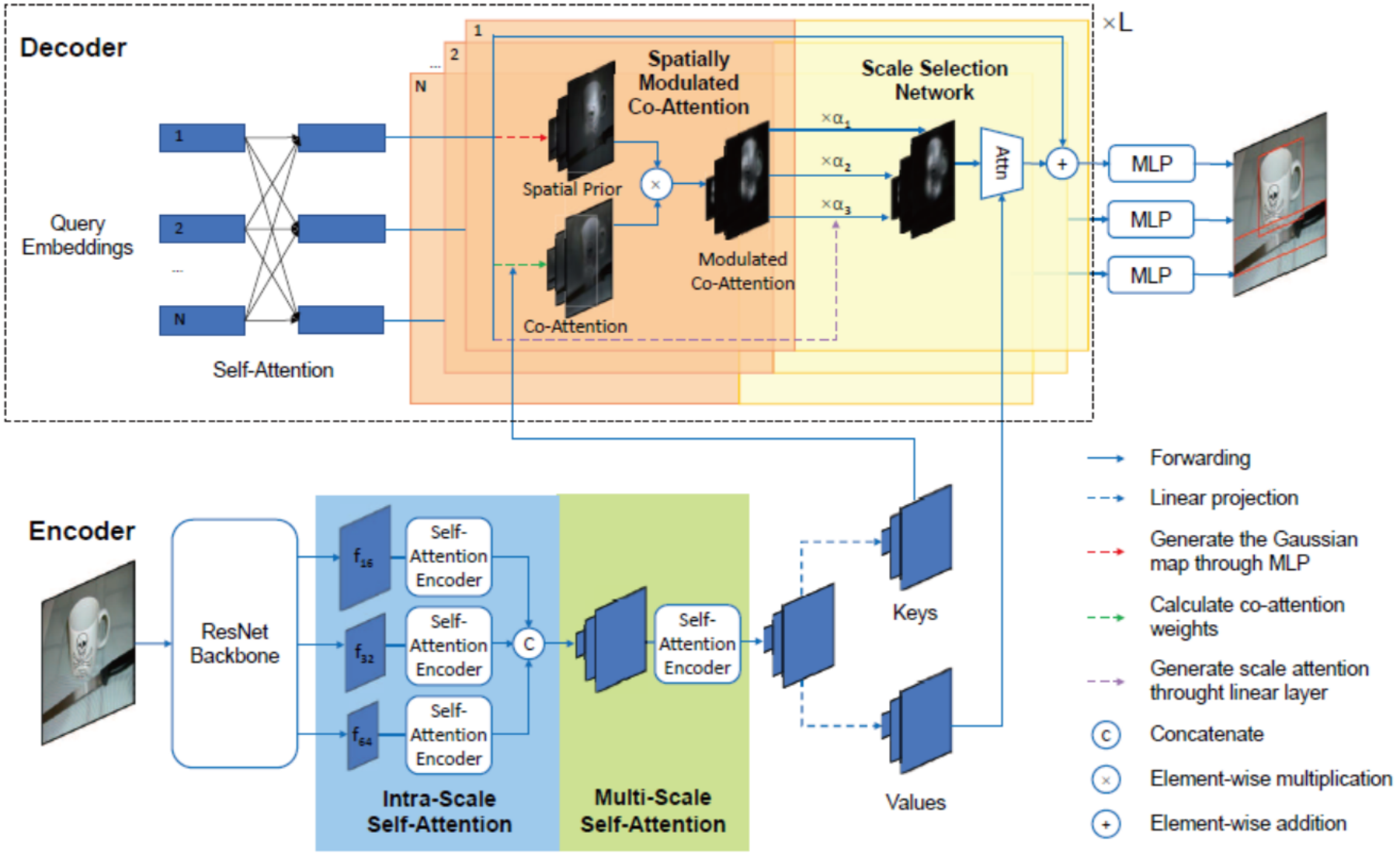


Figure 1. Comparison of convergence of DETR-DC5 trained for 500 epochs, and our proposed SMCA trained for 50 epochs and 108 epochs. The convergence speed of the proposed SMCA is much faster than the original DETR.

2. Related work - DETR



3. Methods

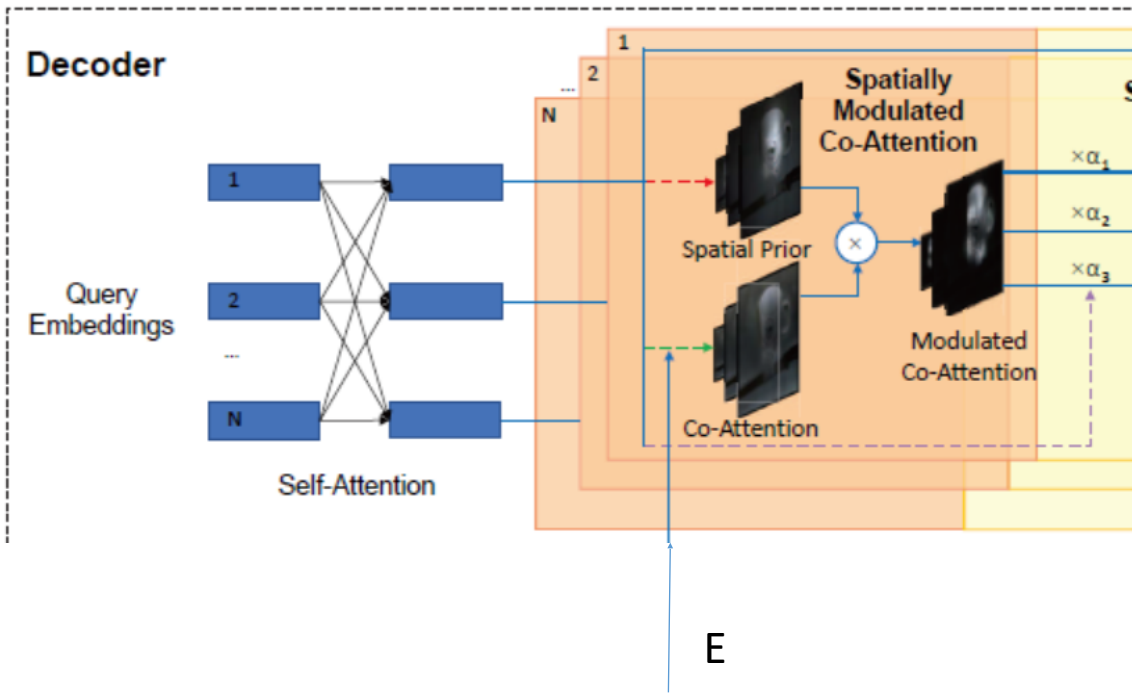

Decoder
×L
Query Embeddings
Self-Attention
Spatially Modulated Co-Attention
Spatial Prior
Co-Attention
Modulated Co-Attention
Scale Selection Network
Attn
MLP
MLP
MLP
Encoder
ResNet Backbone
Self-Attention Encoder
Self-Attention Encoder
Self-Attention Encoder
Intra-Scale Self-Attention
Self-Attention Encoder
Multi-Scale Self-Attention
Keys
Values
Forwarding
Linear projection
Generate the Gaussian map through MLP
Calculate co-attention weights
Generate scale attention throught linear layer
Concatenate
Element-wise multiplication
Element-wise addition

3.1 Spatially Modulated CoAttention

Core idea: Handcrafted query spatial priors

3.1.1 Dynamic spatial weight maps

- Each object query first dynamically predicts the center and scale of its responsible object.
- Create Gaussian-like distribution map by using these centers and scale: i,j is the spatial indices [0,W],[0,H]



$$c_h^{\text{norm}}, c_w^{\text{norm}} = \text{sigmoid}(\text{MLP}(O_q)),$$
$$s_h, s_w = \text{FC}(O_q),$$

$$G(i,j) = \exp\left(-\frac{(i-c_w)^2}{\beta s_w^2} - \frac{(j-c_h)^2}{\beta s_h^2}\right),$$

3.2.2 Spatially-modulated co-attention

Modulate the co-attention feature C_i (object query O_q / self-attention encoded feature E) with the spatial prior G.

$$C_i = \text{softmax}(K_i^T Q_i / \sqrt{d} + \log G) V_i.$$

3.2.3 Multi-head Spatially-modulated co-attention

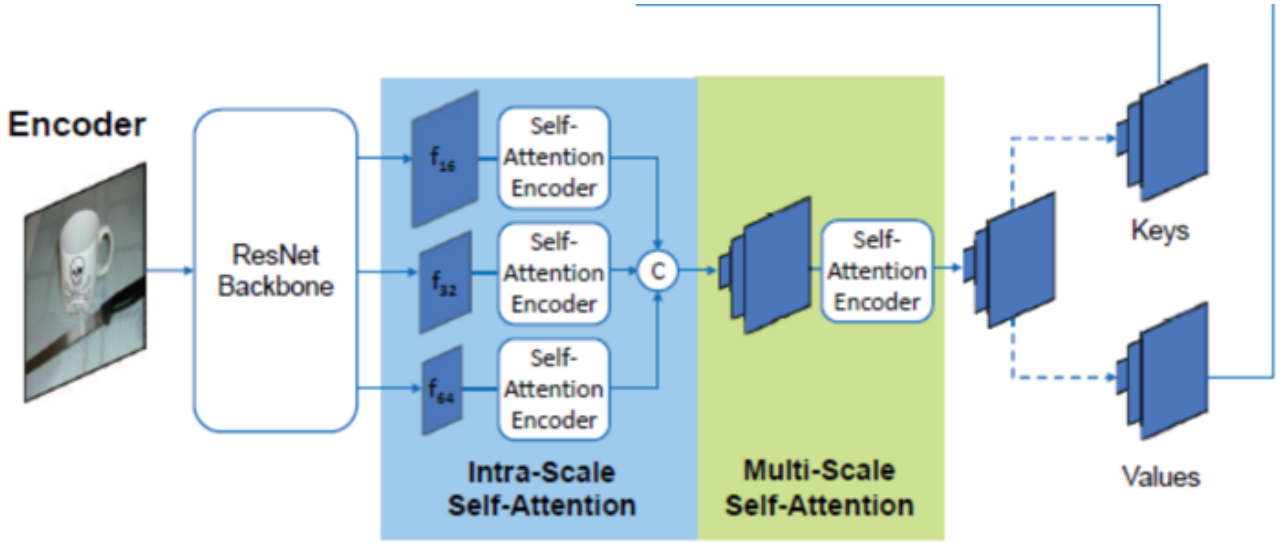
The prior map for each head is different. And it's calculated by head-specific centers and scales.

$$C_i = \mathrm{softmax}(K_i^T Q_i / \sqrt{d} + \log G_i) V_i \quad \text{for } i = 1, \ldots, H.$$

3.2.4 Multi-scale visual features

- Intra-scale self-attention
- The weights of the Transformer block (with self-attention and feedforward sub-networks) are shared across different scales.

(Experimental results show more generalization)



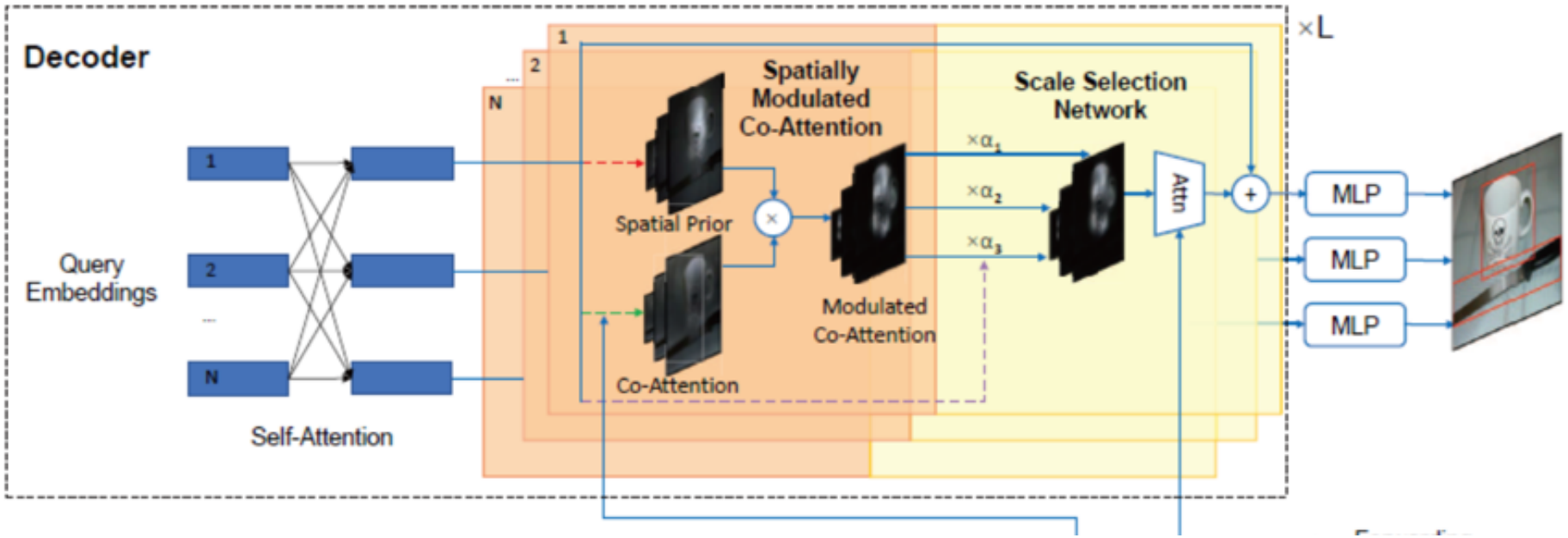
3.2.4 Scale-attention

- The outputs of encoder are E_{16}, E_{32}, E_{64} corresponding to downsampling rate.
- Observation: Some queries may only need information from certain scale. For example: small object in E_{64}
- Design to automatically select scale for each box using learnable scale-attention.
- K,V are generated by E

$$\alpha_{16}, \alpha_{32}, \alpha_{64} = \text{Softmax}(\text{FC}(O_q)),$$

$$C_{i,j} = \text{Softmax}(K_{i,j}^T Q_i / \sqrt{d} + \log G_i) V_{i,j} \odot \alpha_j,$$

$$C_i = \sum_{\text{all } j} C_{i,j}, \quad \text{for } j \in \{16, 32, 64\},$$



3.2 SMCA box prediction

In original DETR, output feature D (N*C), a 3-layerMLP and a linear layer are used to predict the bounding box and classification confidence.

$$\begin{aligned}\text{Box} &= \text{Sigmoid}(\text{MLP}(D)),\\ \text{Score} &= \text{FC}(D),\end{aligned}$$

In SMCA, co-attention is constrained to be around the initially predicted object center

$$[c_h^{\text{norm}}, c_w^{\text{norm}}].$$

We then use the initial center as a prior for constraining bounding box prediction, which is denoted as:

$$\begin{aligned}\widehat{\text{Box}} &= \text{MLP}(D),\\ \widehat{\text{Box}}[:2] &= \widehat{\text{Box}}[:2] + [\widehat{c_h^{\text{norm}}}, \widehat{c_w^{\text{norm}}}],\\ \text{Box} &= \text{Sigmoid}(\widehat{\text{Box}}),\end{aligned}$$

This ensure that the prediction is highly related to the highlighted co-attention regions.

4. Experiments

Datasets: COCO 2017 dataset.

Method	Epochs	time(s)	GFLOPs	mAP	AP_S	AP_M	AP_L
DETR	500	0.038	86	42.0	20.5	45.8	61.1
DETR-DC5	500	0.079	187	43.3	22.5	47.3	61.1
SMCA w/o multi-scale	50	0.043	86	41.0	21.9	44.3	59.1
SMCA w/o multi-scale	108	0.043	86	42.7	22.8	46.1	60.0
SMCA	50	0.100	152	43.7	24.2	47.0	60.4
SMCA	108	0.100	152	45.6	25.9	49.3	62.6

Table 1. Comparison with DETR model over training epochs, mAP, inference time and GFLOPs.

Method		mAP	AP50	AP75
Baseline	DETR-R50	34.8	56.2	36.9
Head-shared Spatial Modulation	+Indep. (bs8)	40.2	61.4	42.7
	+Indep. (bs16)	40.2	61.3	42.9
	+Indep. (bs32)	39.9	61.0	42.4
Multi-head Spatial Modulation	+Fixed	38.5	60.7	40.2
	+Single	40.4	61.8	43.3
	+Indep.	41.0	62.2	43.6

Table 2. Ablation study on the importance of spatial modulation, multi-head mechanism. mAP, AP50, and AP75 are reported on COCO 2017 validation set.

Fixed: Only predict center with fixed scale
Single: Single scale for height and width
Indep: Separate scale for height and width

Method		mAP	Params (M)
SMCA		41.0	41.0
SMCA (2Intra-Multi-2Intra)		43.7	39.5
SMCA w/o SSA (2Intra-Multi-2Intra)		42.6	39.5
3Intra		42.9	37.9
3Multi		43.3	37.9
5Intra		43.3	39.5
Weight Share	Shared FFN	43.0	42.2
	Shared SA	42.8	44.7
	No Share	42.3	47.3

Table 3. Ablation study on the importance of combining intra-scale and multi-scale propagation, and the weight sharing for intra-scale self-attention. "Shared FFN" stands for only sharing weights of the feedfoward network of intra-scale self-attention. "Shared SA" stands for sharing the weights of the self-attention network. "No share" stands for no weight sharing in intra-scale self attention.

Weight sharing is better in intra-scale self-attention

Model	Epochs	GFLOPs	Params (M)	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DETR-R50 [4]	500	86	41	42.0	62.4	44.2	20.5	45.8	61.1
DETR-DC5-R50 [4]	500	187	41	43.3	63.1	45.9	22.5	47.3	61.1
Faster RCNN-FPN-R50 [4]	36	180	42	40.2	61.0	43.8	24.2	43.5	52.0
Faster RCNN-FPN-R50++ [4]	108	180	42	42.0	62.1	45.5	26.6	45.4	53.4
Deformable DETR-R50 (Single-scale) [46]	50	78	34	39.7	60.1	42.4	21.2	44.3	56.0
Deformable DETR-R50 (50 epochs) [46]	50	173	40	43.8	62.6	47.7	26.4	47.1	58.0
Deformable DETR-R50 (150 epochs) [46]	150	173	40	45.3	*	*	*	*	*
UP-DETR-R50 [5]	150	86	41	40.5	60.8	42.6	19.0	44.4	60.0
UP-DETR-R50+ [5]	300	86	41	42.8	63.0	45.3	20.8	47.1	61.7
TSP-FCOS-R50 [38]	36	189	*	43.1	62.3	47.0	26.6	46.8	55.9
TSP-RCNN-R50 [38]	36	188	*	43.8	63.3	48.3	28.6	46.9	55.7
TSP-RCNN+-R50 [38]	96	188	*	45.0	64.5	**49.6**	**29.7**	47.7	58.0
SMCA-R50	50	152	40	43.7	63.6	47.2	24.2	47.0	60.4
SMCA-R50	108	152	40	**45.6**	**65.5**	49.1	25.9	**49.3**	**62.6**
DETR-R101 [4]	500	152	60	43.5	63.8	46.4	21.9	48.0	61.8
DETR-DC5-R101 [4]	500	253	60	44.9	64.7	47.7	23.7	49.5	**62.3**
Faster RCNN-FPN-R101 [4]	36	256	60	42.0	62.1	45.5	26.6	45.4	53.4
Faster RCNN-FPN-R101+ [4]	108	246	60	44.0	63.9	47.8	27.2	48.1	56.0
TSP-FCOS-R101 [38]	36	255	*	44.4	63.8	48.2	27.7	48.6	57.3
TSP-RCNN-R101 [38]	36	254	*	44.8	63.8	49.2	29.0	47.9	57.1
TSP-RCNN+-R101 [38]	96	254	*	**46.5**	**66.0**	**51.2**	**29.9**	**49.7**	59.2
SMCA-R101	50	218	58	44.4	65.2	48.0	24.3	48.5	61.0

Table 4. Comparison with DETR-like object detectors on COCO 2017 validation set.