Research Report

利用视觉语言预训练框架做OCR相关任务

2022年 5月

指导老师：刘绍辉 汇报人： 研一 舒言

1

Vision-Language Pre-Training for Boosting Scene Text Detectors

Sibo Song[1*] Jianqiang Wan[1*] Zhibo Yang[1] Jun Tang[1] Wenqing Cheng[2] Xiang Bai[2] Cong Yao[1]

[1]DAMO Academy, Alibaba Group

[2]Huazhong University of Science and Technology

{sibosongzju,hustwjq,yangzhibo450,yaocong2010}@gmail.com

xixing.tj@alibaba-inc.com {xbai,chengwq}@hust.edu.cn

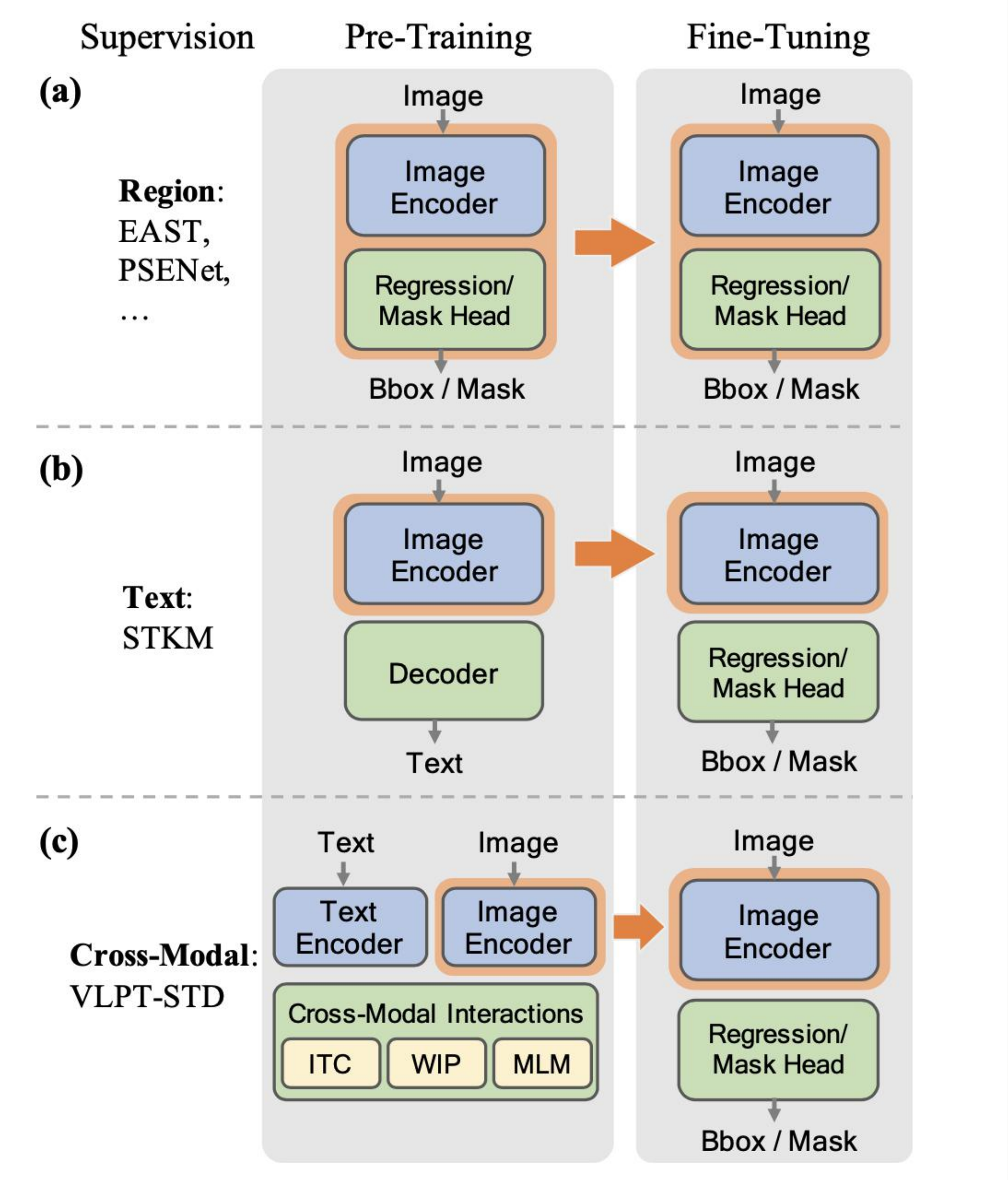
目录

CONTENTS

第一部分

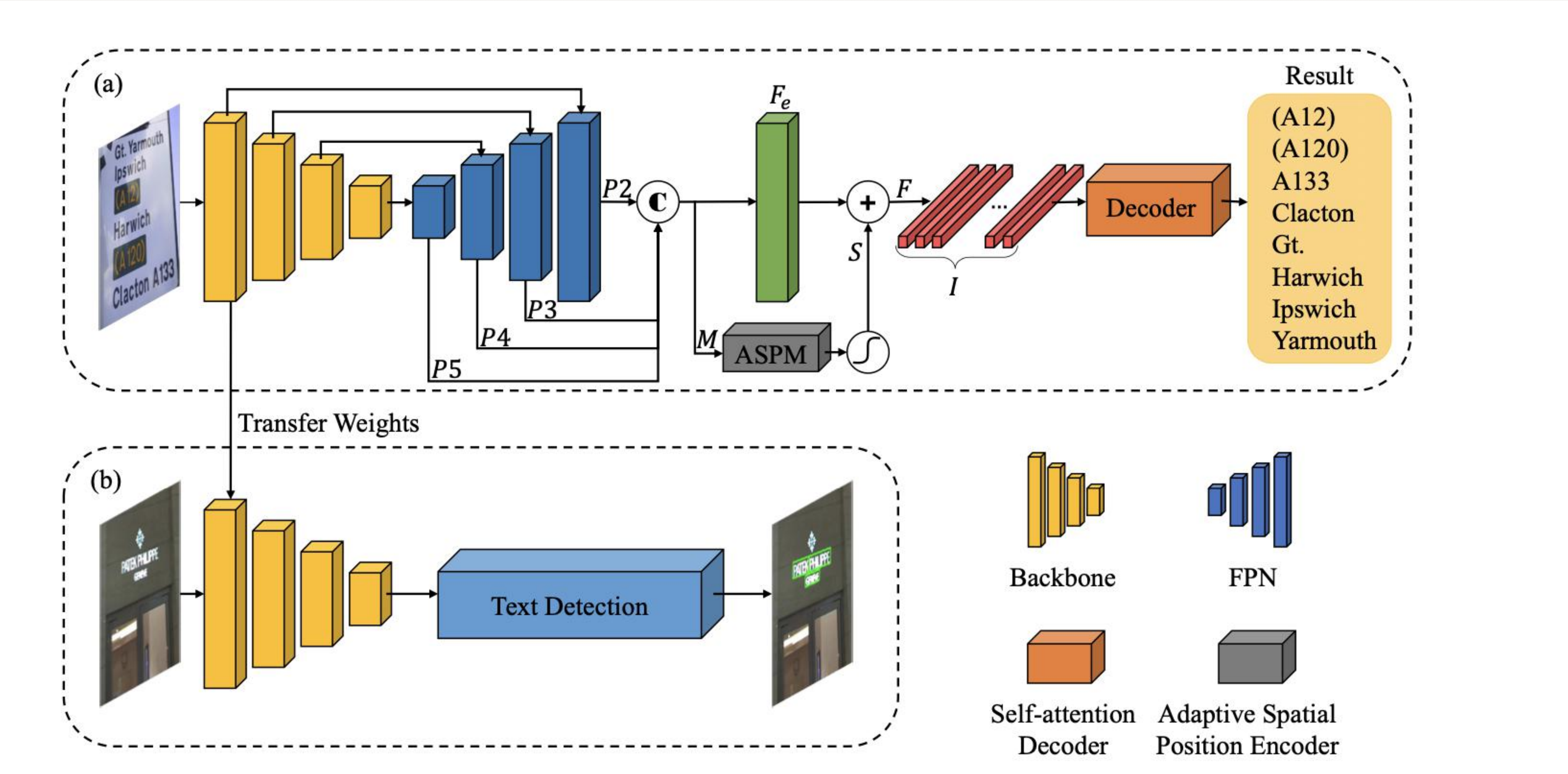
任务概述

1 任务



Supervision
Pre-Training
Fine-Tuning
(a)
Region: EAST, PSENet, …
Image
Image Encoder
Regression/ Mask Head
Bbox / Mask
(b)
Text: STKM
Decoder
Text
(c)
Cross-Modal: VLPT-STD
Text Encoder
Cross-Modal Interactions
ITC
WIP
MLM

1 STKM



(a)
Gt. Yarmouth
Ipswich
(A12)
Harwich
(A120)
Clacton A133
P2
P3
P4
P5
F_e
M
ASPM
S
F
I
...
Decoder
Result
(A12)
(A120)
A133
Clacton
Gt.
Harwich
Ipswich
Yarmouth
Transfer Weights
(b)
Text Detection
Backbone
FPN
Self-attention Decoder
Adaptive Spatial Position Encoder

1 STKM的缺陷

1. decoder是以字符级别输入学习的，并不能学习到词语的语义信息

2. 预训练侧重于从视觉到语言的单流向，缺少两种模态的交互

1

作者主要创新

1. 设计了视觉-语言交互的预训练框架，可以用于视觉和文本的特征对齐

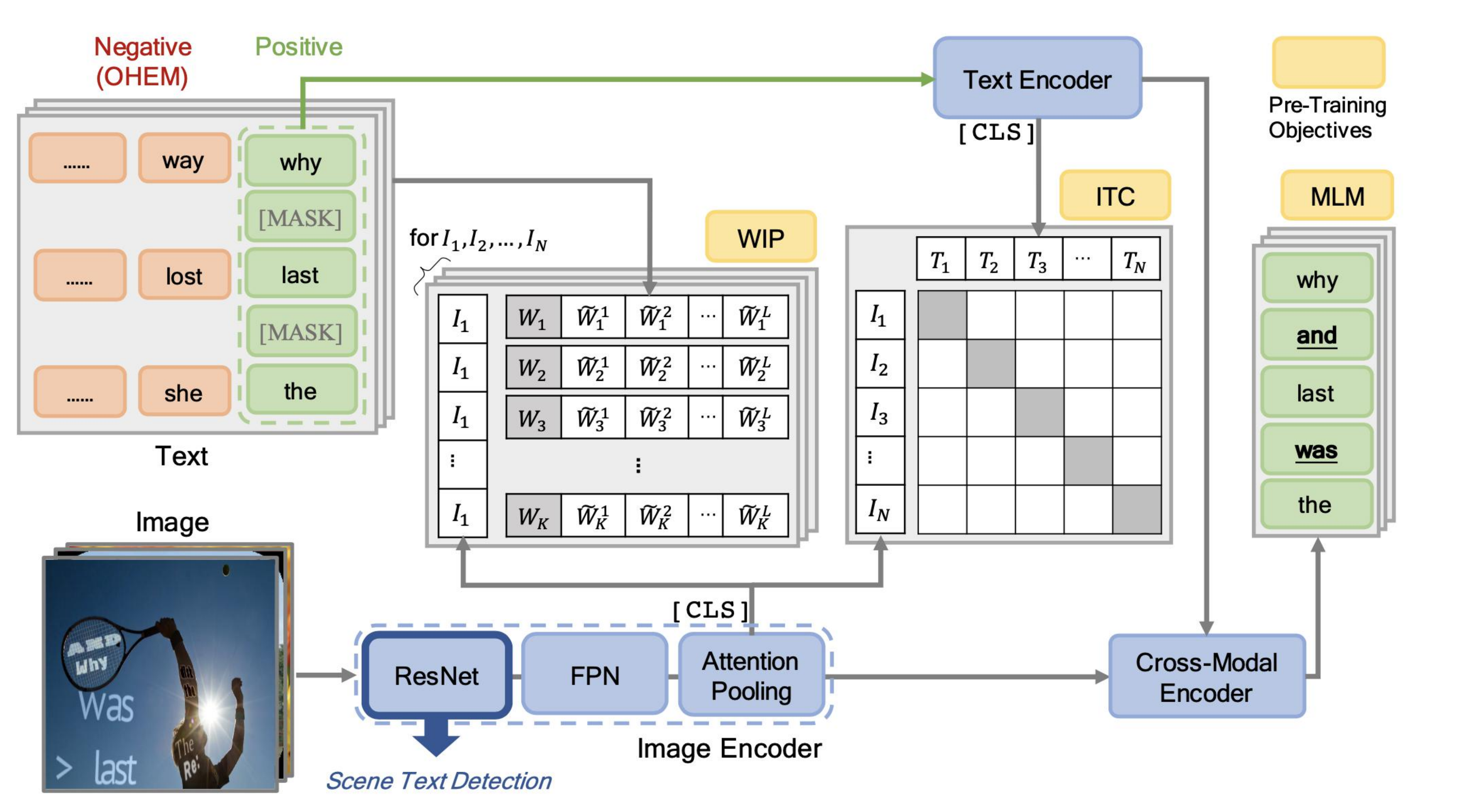
2. 设计了三种预训练代理任务，尤其是WIP（word-In-Image）任务，用于丰富视觉表示

3. 实验证明该方法超越了之前的预训练方法。（STKM）

第二部分

方法

2 方法介绍



Negative (OHEM)
Positive
...... way why
[MASK]
...... lost last
[MASK]
...... she the
Text
Image
Text Encoder
[CLS]
Pre-Training Objectives
for $I_1, I_2, \ldots, I_N$
WIP
ITC
MLM
I_1 W_1 $\tilde{W}_1^1$ $\tilde{W}_1^2$ ··· $\tilde{W}_1^L$
I_1 W_2 $\tilde{W}_2^1$ $\tilde{W}_2^2$ ··· $\tilde{W}_2^L$
I_1 W_3 $\tilde{W}_3^1$ $\tilde{W}_3^2$ ··· $\tilde{W}_3^L$
I_1 W_K $\tilde{W}_K^1$ $\tilde{W}_K^2$ ··· $\tilde{W}_K^L$
T_1 T_2 T_3 ··· T_N
I_1 I_2 I_3 I_N
why
and
last
was
the
[CLS]
ResNet
FPN
Attention Pooling
Cross-Modal Encoder
Image Encoder
Scene Text Detection

2 Image Encoder

1. ResNet+FPN+Attention_pooling

FPN: $\mathcal{F}_c = \mathrm{Conv}_{1\times 1, s2}([\mathrm{DS}_{\times 2}(P_2); P_3; \mathrm{US}_{\times 2}(P_4); \mathrm{US}_{\times 4}(P_5)])$

Attention_pooling: Multi-head attention layer

$x^I \longrightarrow \mathbf{V} = \{V_{[\mathtt{CLS}]}, V_1, ..., V_S\} \in \mathbb{R}_d$

2 Text Encoder

$$\mathbf{W} = \{W_{[\mathrm{CLS}]}, W_1, W_2, \cdots, W_K\} \in \mathbb{R}_d$$

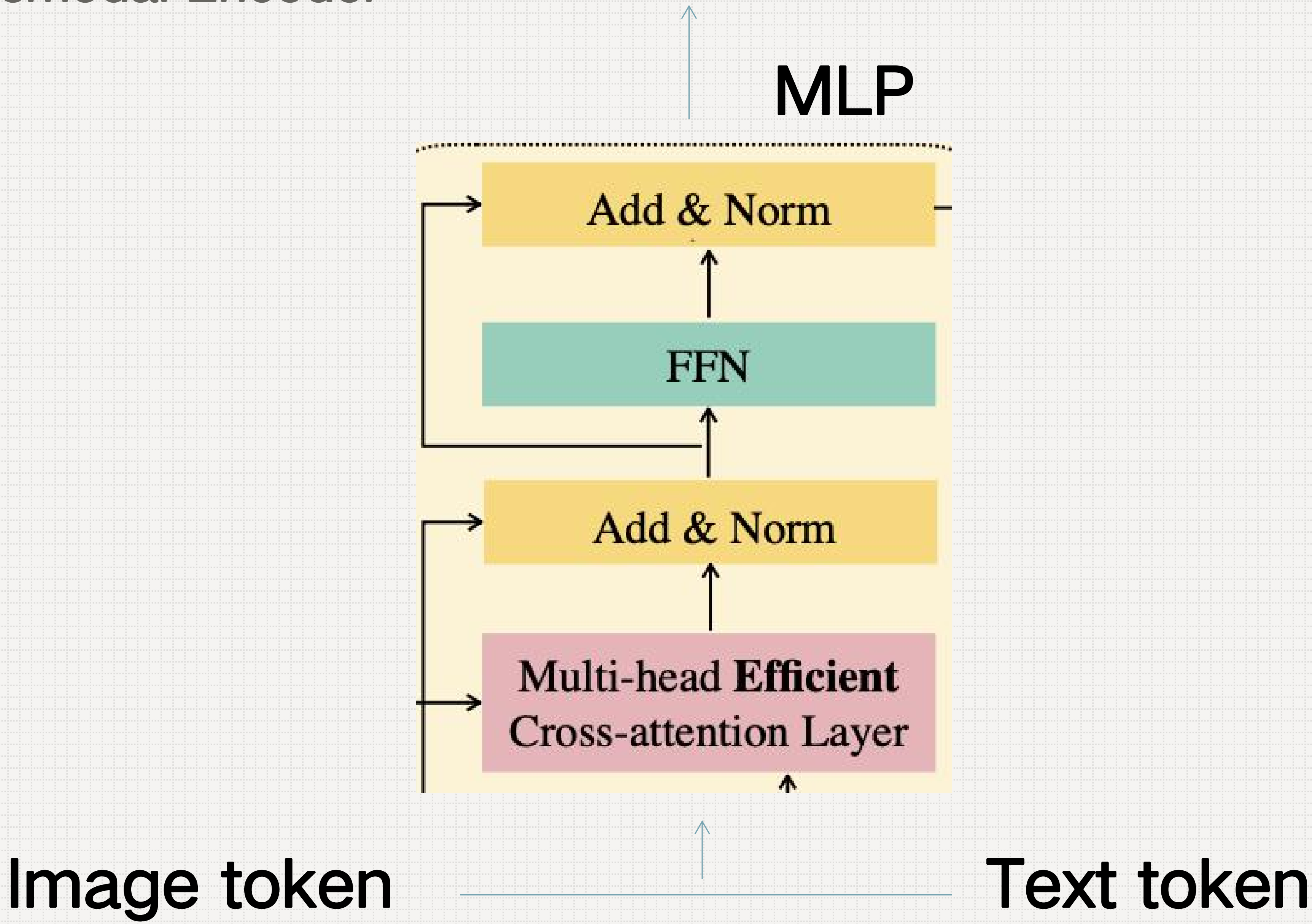
$\downarrow$

Bert

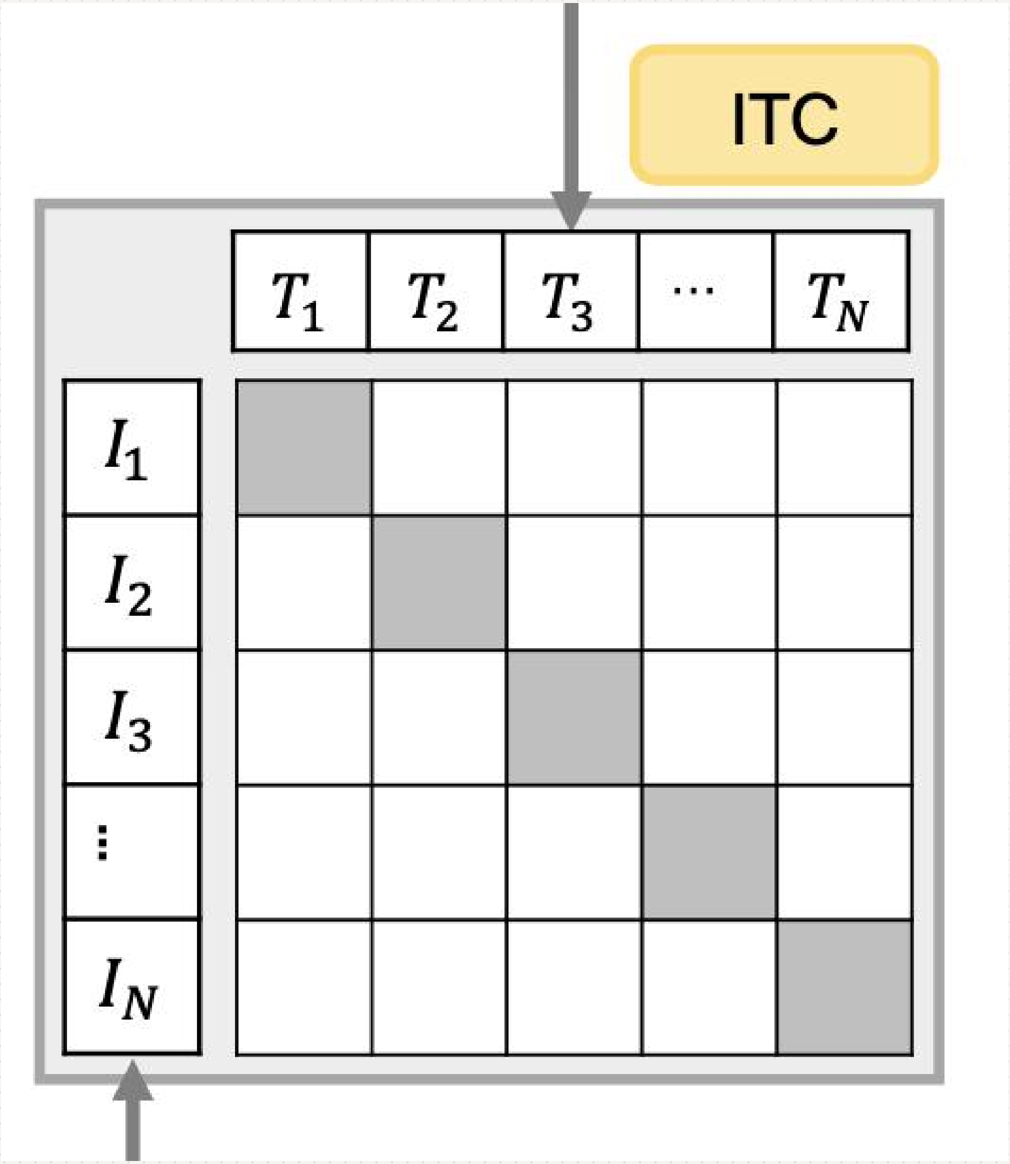
$\downarrow$

$$W_{[\mathrm{CLS}]}$$

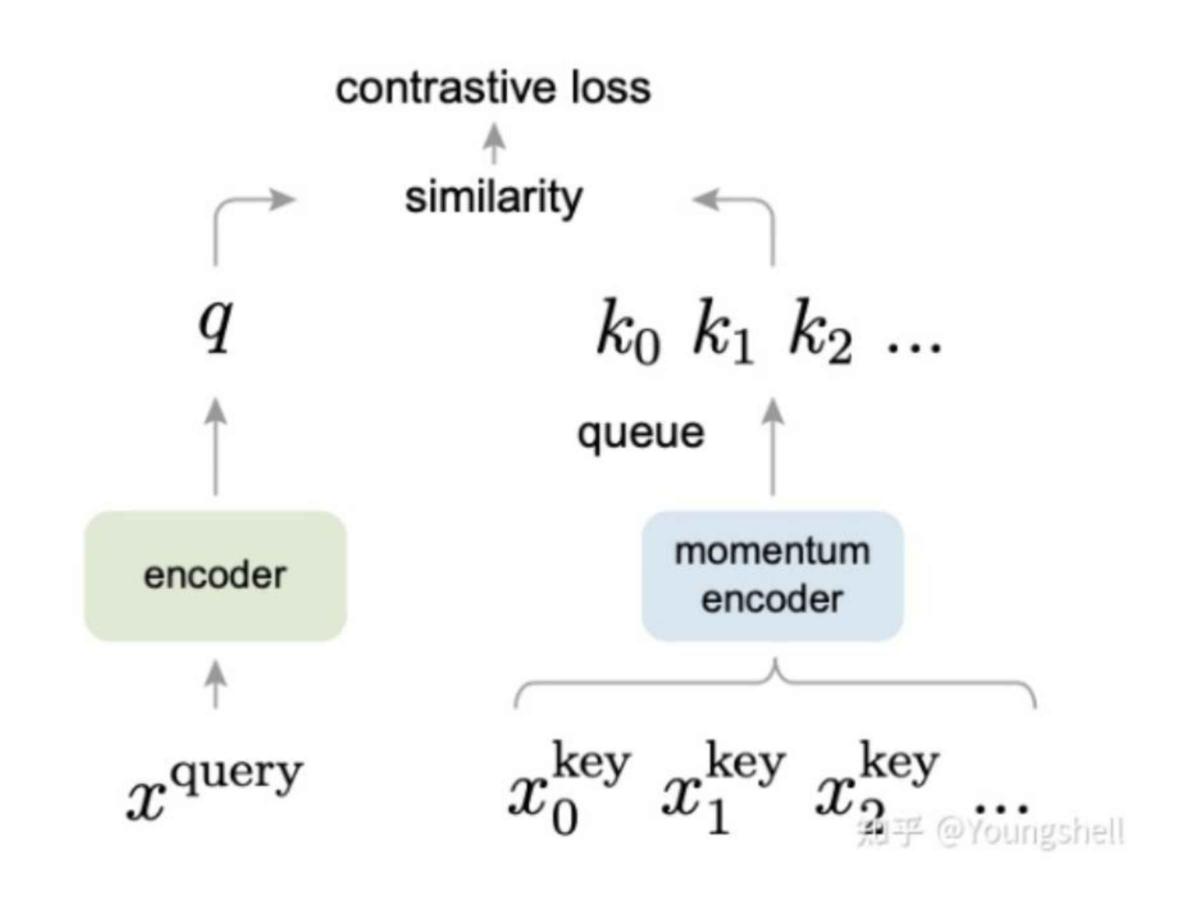
2 Crossmodal Encoder



2 Pretext task1 ITC (Image-Text Contrastive learning)



2 InfoNCE loss



$$L_q = -log \frac{exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{k} exp(q \cdot k_i / \tau))}$$

2 交叉熵 ->NCE loss -> InfoNCE loss

$$\hat{y}_+ = softmax(z+) = \frac{exp(z+)}{\sum_{i=0}^{k} exp(z_i)}$$

$$L(\hat{y}) = -\sum_{i \in K} y_i log(\hat{y}_i)$$

$$-log\frac{exp(z+)}{\sum_{i=0}^{k} exp(z_i)}$$

NCE(noise constrative estimation)

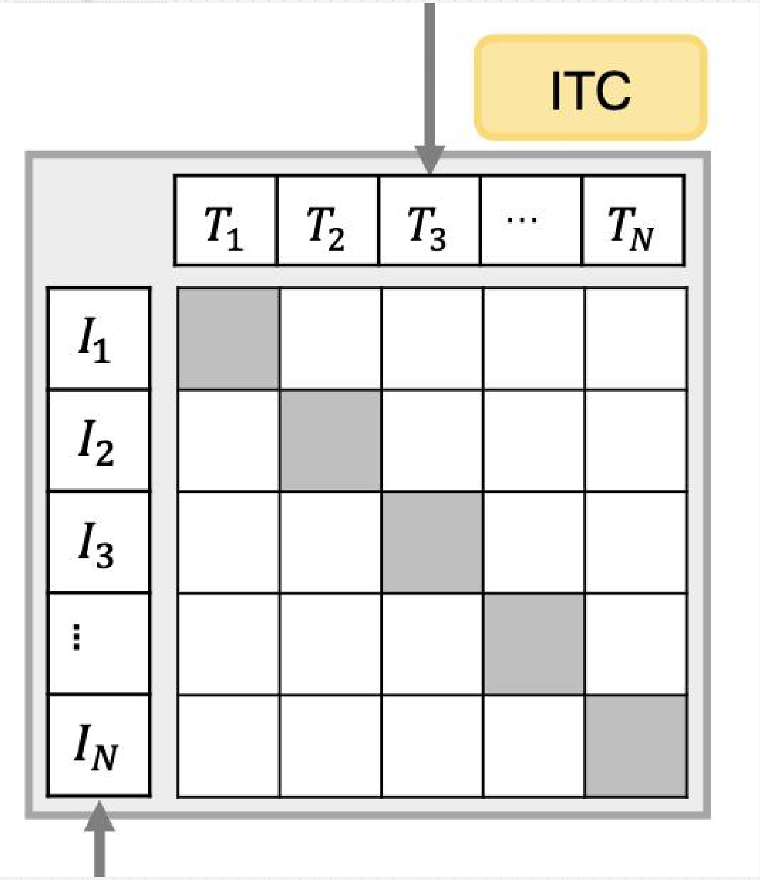
多分类变为二分类 + 负样本采样

InfoNCE

多分类，k为负样本的数量

$$L_q = -log\frac{exp(q \cdot k_+/\tau)}{\sum_{i=0}^{k} exp(q \cdot k_i/\tau))}$$

2 Pretext task1 ITC (Image-Text Contrastive learning)

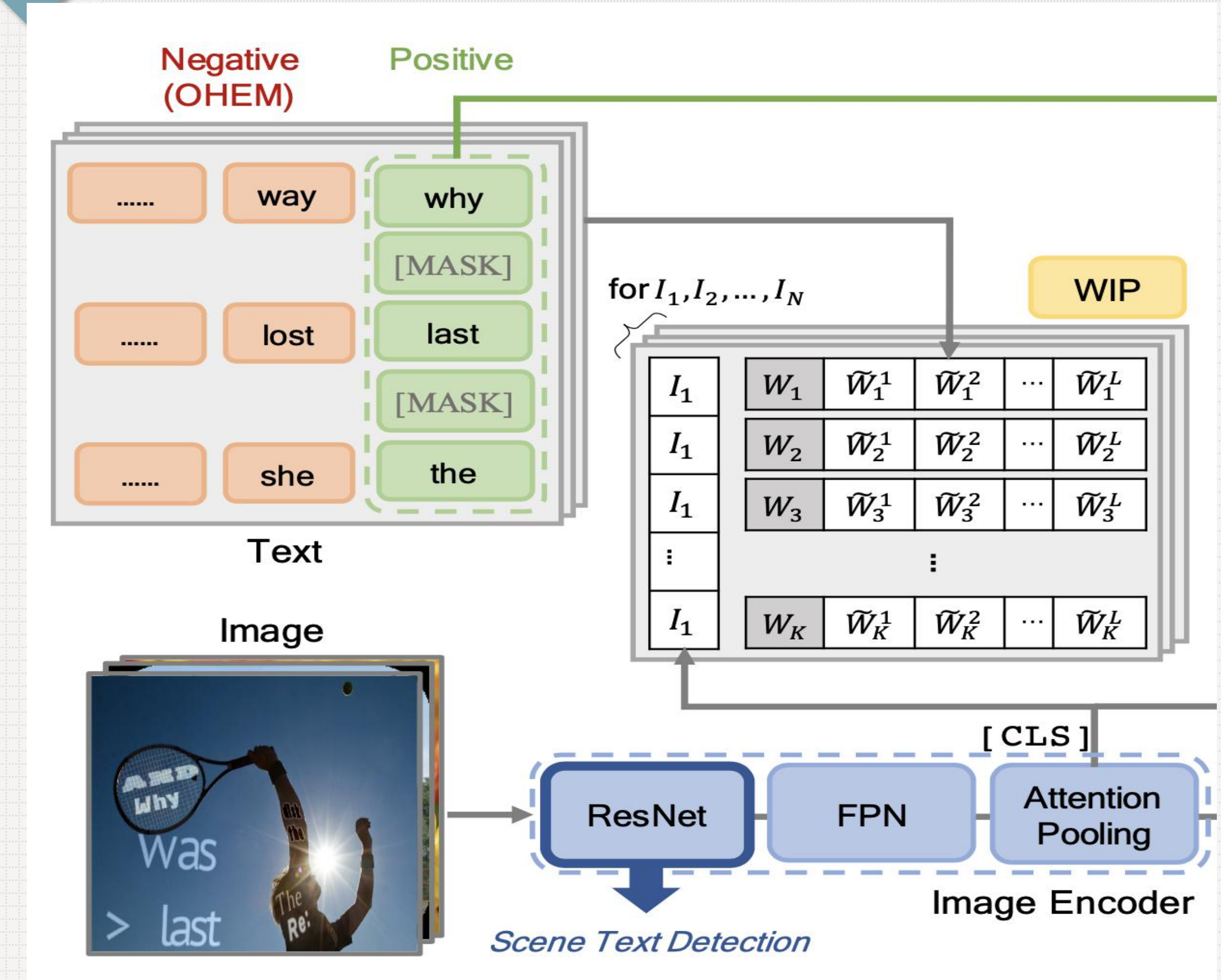


$$\mathcal{L}_{\mathrm{I2T}} = -\sum_{j} \log \frac{\exp\left(I_j \cdot T_j / \tau\right)}{\sum_{k=1}^{N} \exp\left(I_j \cdot T_k / \tau\right)}$$

$$\mathcal{L}_{\mathrm{T2I}} = -\sum_{j} \log \frac{\exp\left(T_j \cdot I_j / \tau\right)}{\sum_{k=1}^{N} \exp\left(T_j \cdot I_k / \tau\right)}$$

$$\mathcal{L}_{\mathrm{ITC}} = \lambda_1 \mathcal{L}_{\mathrm{I2T}} + \lambda_2 \mathcal{L}_{\mathrm{T2I}}$$

2 Pretext task2 WIP (Word-in-Image Prediction)

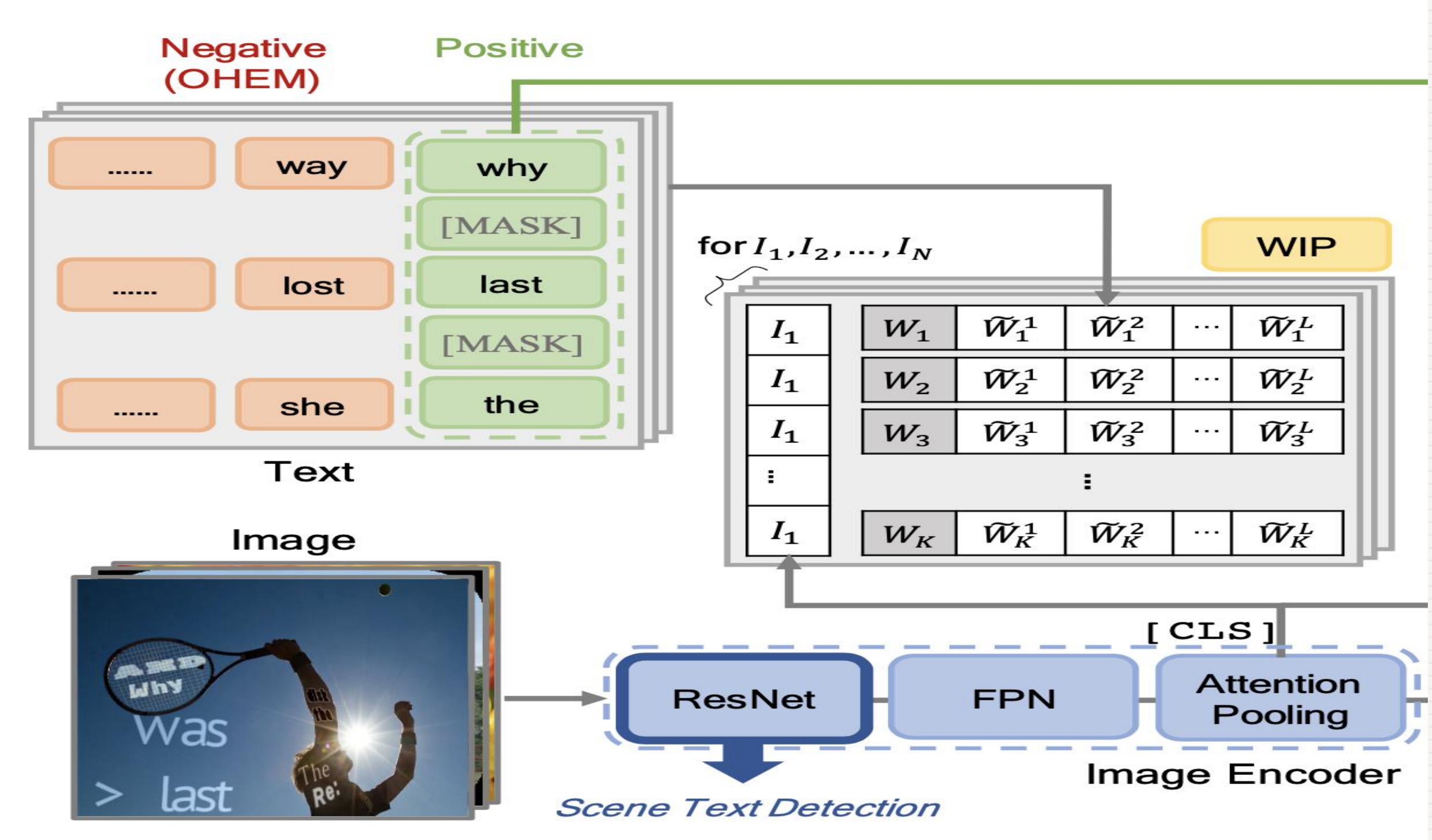


2 generate negative samples

OHEM + text embeddings` similarites

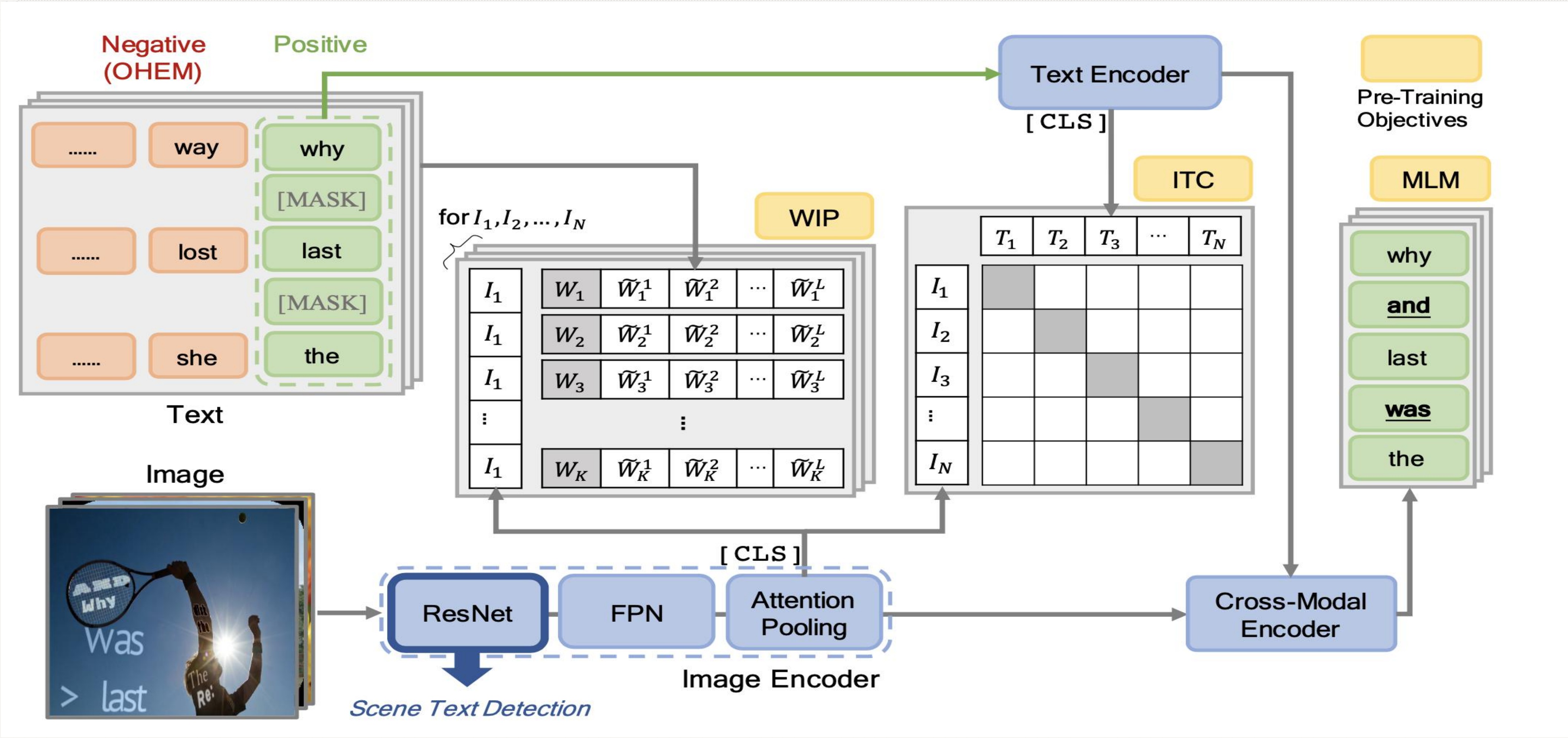
Query	Top-5 nearest neighbors from VLPT-STD				
eco	850	800	630	rca	600
vote	note	voice	work	role	write
sale	safe	scale	said	able	sake
north	worth	keith	math	norton	both
river	liver	layer	viper	driver	meter
right	light	night	rights	might	higher
special	specific	typical	serial	social	optical
affected	attached	selected	attacked	scattered	affiliated

2 Pretext task2 WIP (Word-in-Image Prediction)



$$\mathcal{L}_{\mathrm{WIP}} = -\sum_{k=1}^{K} \log \frac{\exp(I \cdot W_k/\tau)}{\exp(I \cdot W_k/\tau) + \sum_{l=1}^{L} \exp\left(I \cdot \widetilde{W}_k^l/\tau\right)}$$

2 Pretext task3 MLM (Masked Language Modeling)



$$\mathcal{L}_{\mathrm{MLM}} = -\mathbb{E}_{(W,V)} \log P_\theta(W_{\mathrm{masked}} | W_{\mathrm{unmasked}}, \mathbf{V})$$

2 comparison with SOTA

Methods	ICDAR2015			Total-Text			CTW1500		
	P	R	F	P	R	F	P	R	F
SegLink [46]	73.1	76.8	75.0	30.3	23.8	26.7	42.3	40.0	40.8
TextSnake [33]	84.9	80.4	82.6	82.7	74.5	78.4	67.9	85.3	75.6
TextDragon [10]	84.8	81.8	83.1	79.5	81.0	80.2	84.5	74.2	79.0
SAE [50]	84.5	85.1	84.8	-	-	-	82.7	77.8	80.1
PSENet + ST [1]	84.3	78.4	81.3	89.2	79.2	83.9	83.6	79.7	81.6
PSENet + STKM [1]	85.7	81.8	83.7	89.2	79.9	84.3	85.3	80.6	82.9
PSENet + Ours	86.0	82.8	**84.3**	90.8	82.0	**86.1**	86.3	80.7	**83.3**
Δ			**3.0↑**, **0.6↑**			**2.2↑**, **1.8↑**			**1.7↑**, **0.4↑**

2 comparison with SOTA

Methods	ICDAR2015			ICDAR2017			MSRA-TD500		
	P	R	F	P	R	F	P	R	F
SegLink [46]	73.1	76.8	75.0	-	-	-	86	70	77
TextField [55]	84.3	80.1	82.4	-	-	-	87.4	75.9	81.3
CRAFT [1]	89.8	84.3	86.9	80.6	68.2	73.9	88.2	78.2	82.9
GNNets [53]	90.4	86.7	88.5	79.6	70.1	74.5	-	-	-
EAST + ST [1]	89.6	81.5	85.3	75.1	61.9	67.9	86.9	77.6	82.0
EAST + STKM [1]	90.2	84.6	87.3	76.9	64.3	70.0	85.2	75.3	80.0
EAST + Ours	91.5	85.4	**88.3**	77.7	64.6	**70.5**	88.5	76.7	**82.2**
Δ			**3.0↑, 1.0↑**			**2.6↑, 0.5↑**			**0.2↑, 2.2↑**

[1] We report results using our reimplementation.

2 comparison with SOTA

Methods	ICDAR2015			Total-Text			MSRA-TD500		
	P	R	F	P	R	F	P	R	F
DB + ST [1]	88.2	82.7	85.4	87.1	82.5	84.7	91.5	79.2	84.9
DB + STKM [1]	91.4	81.4	86.1	87.7	83.4	85.5	90.2	82.0	85.9
DB + Ours	92.0	81.6	**86.5**	88.7	84.0	**86.3**	92.3	84.9	**88.5**
Δ			**1.1↑, 0.4↑**			**1.6↑, 0.8↑**			**3.6↑, 2.6↑**

[1] We report results using our reimplementation.

2 Ablation studies

Image Encoder		Cross-Modal Encoder		PSENet	EAST
FPN†	Our FPN	w/o MHCA	w/ MHCA	CTW	IC15
√			√	82.7	88.0
	√	√		83.1	87.4
	√		√	83.3	88.3

			PSENet			EAST		
ITC	MLM	WIP	IC15	TT	CTW	IC15	IC17	TD500
			81.3	83.9	81.6	85.3	67.9	82.0
√			82.2	84.3	82.2	86.0	69.4	79.3
	√		84.5	85.9	83.1	87.7	70.2	81.5
		√	83.1	85.3	82.2	86.9	70.2	82.1
√	√		84.3	85.6	83.2	87.5	70.3	81.7
√		√	83.3	85.3	82.5	87.3	70.2	81.5
	√	√	84.7	85.8	82.9	87.6	70.4	81.9
√	√	√	84.3	86.1	83.3	88.3	70.5	82.2

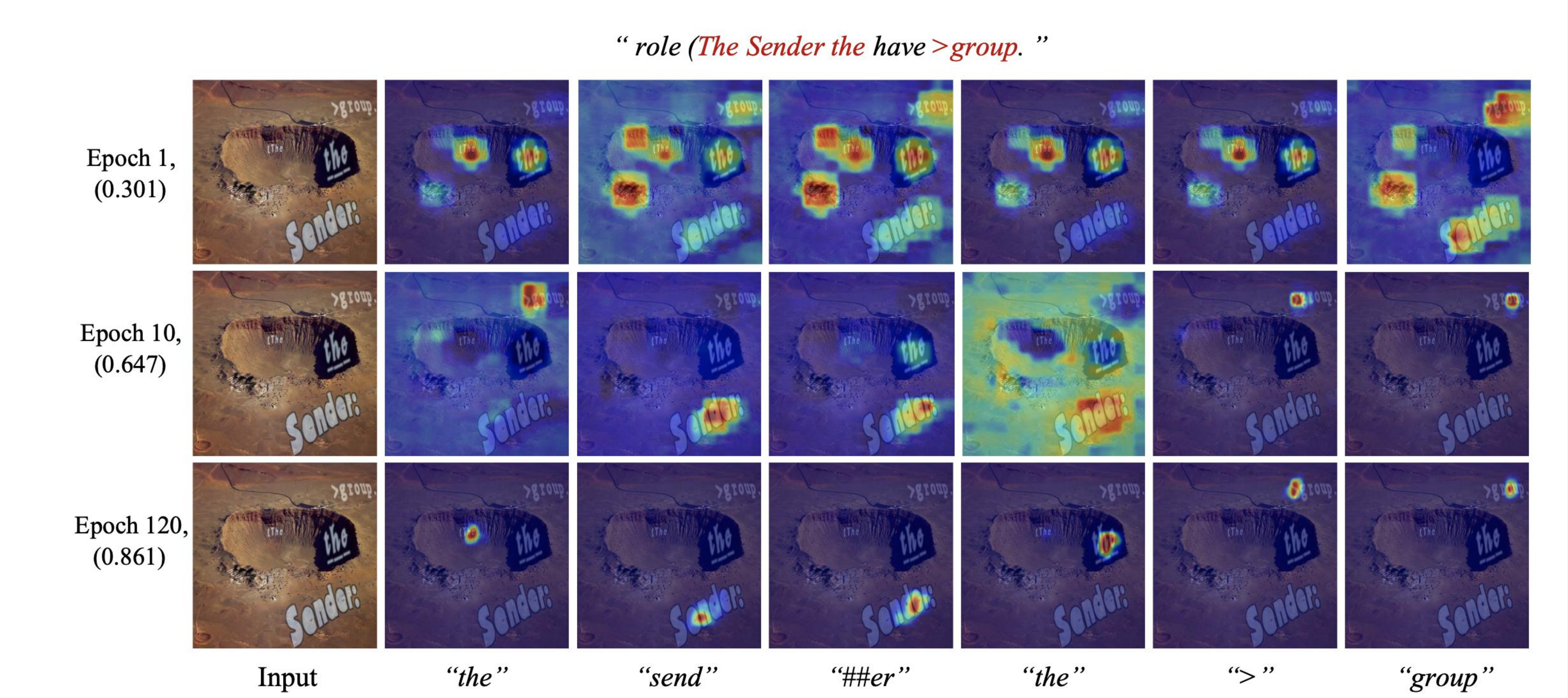
2 Ablation studies

Table 6. Ablation study on pre-training datasets with PSENet. **ST** denotes SynthText and **TO** denotes TextOCR. Only F-measure is presented.

Supervision	Pre-training Datasets	IC15	TT	CTW
Region	ST+TO (1 epoch)	82.24	84.52	81.75
Text (Ours)	ST (1 epoch)	84.33	86.14	83.30
Text (Ours)	ST+TO (1 epoch)	85.07	86.18	83.50

2 results

2 results



“ role (The Sender the have >group. ”
Epoch 1,
(0.301)
Epoch 10,
(0.647)
Epoch 120,
(0.861)
Input
“the”
“send”
“##er”
“the”
“>”
“group”

第三部分

未来展望

感谢聆听！

THANK YOU FOR WATCHING!