Research Report

视频场景文字检测与识别算法研究

2021年 10月

指导老师：刘绍辉 汇报人： 研一 舒言

目录

CONTENTS

第一部分

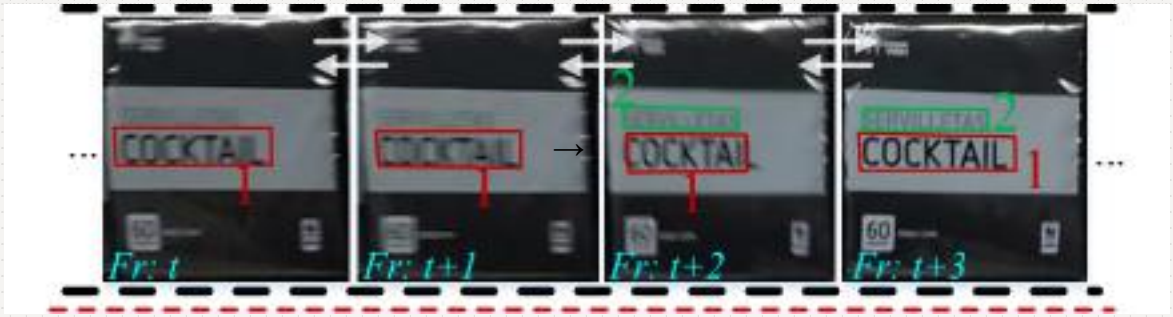
任务概述

1 评测任务

ICDAR 2021 Competition on Scene Video Text Spotting

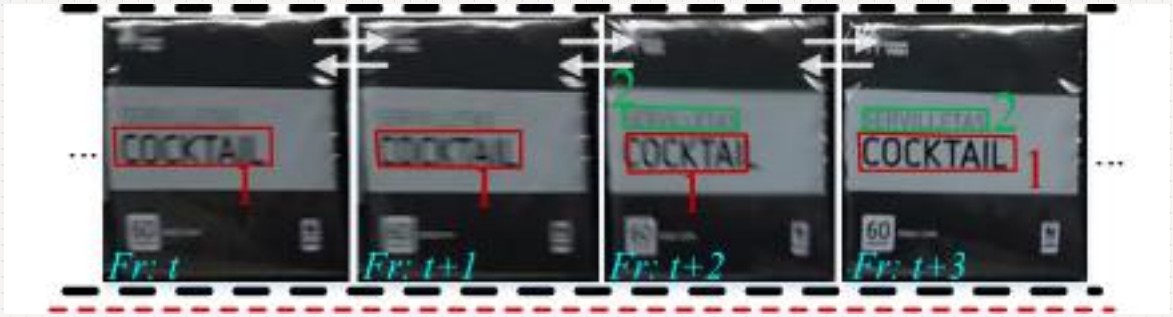
Zhanzhan Cheng[1,2⋆], Jing Lu[2⋆], Baorui Zou[3⋆], Shuigeng Zhou[3], and Fei Wu[1]

[1] Zhejiang University, Hangzhou, China
{11821104,wufei}@zju.edu.cn
[2] Hikvision Research Institute, Hangzhou, China
lujing6@hikvision.com
[3] Fudan University, Shanghai, China
{18210240270,sgzhou}@fudan.edu.cn



Text Detection → Text Tracking → Text Recognition

1 评测任务



Text Detection → Text Tracking → Text Recognition

(a)


```
{
    "video_19_4/67.jpg": {
        "content_ann": {
            "bboxes": [
                [
                    210,
                    646,
                    746,
                    670,
                    763,
                    752,
                    213,
                    704
                ],
                [
                    513,
                    355,
                    677,
                    275,
                    682,
                    327,
                    517,
                    397
                ]
            ]
        }
    },
    "video_13_6/405.jpg": {
        "content_ann": {
            "bboxes": [
                [
                    765,
                    618,
                    773,
                    634,
                    764,
                    658,
                    739,
                    622
                ]
            ]
        }
    }
}
```

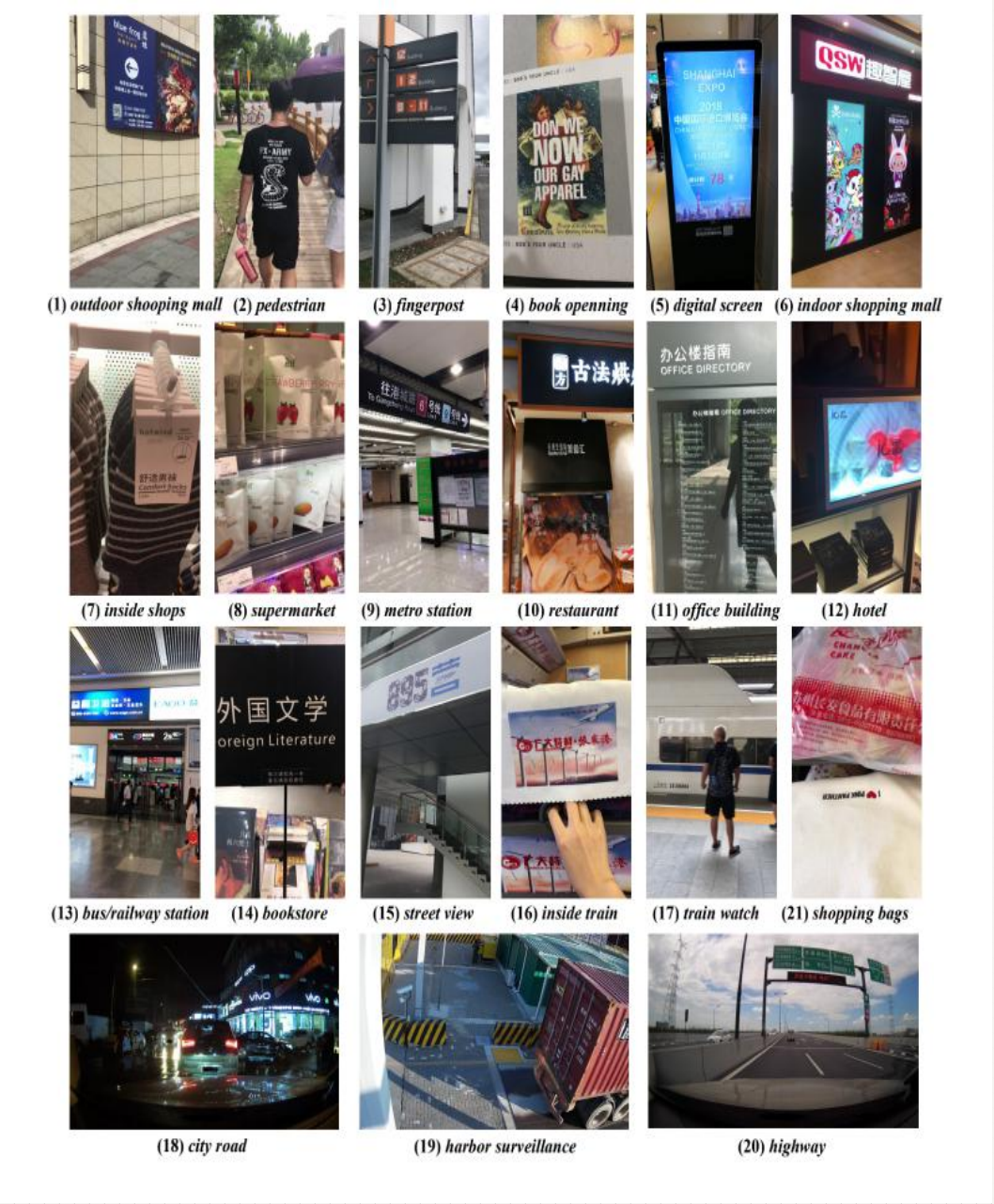

(b)


```
{
    "video_10_3": {
        "1304": {
            "tracks": [
                "158,938_569_962_569_962_589_938_589"
            ]
        },
        "1348": {
            "tracks": [
                "310,605_779_648_779_648_811_605_811",
                "313,586_799_632_799_632_827_586_827",
                "309,607_782_647_782_647_810_607_810",
                "311,604_783_639_783_639_808_604_808",
                "312,599_791_639_791_639_819_599_819"
            ]
        },
    },
    "video_9_1": {
        "1140": {
            "tracks": [
                "40,181_1102_520_1102_520_1181_181_1181",
                "30,526_1156_871_1156_871_1234_526_1234",
                "41,145_1099_484_1099_484_1183_145_1183",
                "36,397_1176_735_1176_735_1257_397_1257",
                "45,124_1172_446_1172_446_1259_124_1259",
                "44,104_1151_438_1151_438_1235_104_1235"
            ]
        },
    }
}
```


(c)


```
{
    "video_10_3": {
        "1304": {
            "tracks": [
                "158,938_569_962_569_962_589_938_589,DAVAR"
            ],
            "text": "DAVAR"
        },
        "1348": {
            "tracks": [
                "310,605_779_648_779_648_811_605_811,LAB",
                "313,586_799_632_799_632_827_586_827,LAB",
                "309,607_782_647_782_647_810_607_810,LAB",
                "311,604_783_639_783_639_808_604_808,LAB",
                "312,599_791_639_791_639_819_599_819,LAB"
            ],
            "text": "LAB"
        }
    },
    "video_9_1": {
        "1140": {
            "tracks": [
                "40,181_1102_520_1102_520_1181_181_1181,JUICE",
                "30,526_1156_871_1156_871_1234_526_1234,JUICE",
                "41,145_1099_484_1099_484_1183_145_1183,JUICE",
                "36,397_1176_735_1176_735_1257_397_1257,JUICE",
                "45,124_1172_446_1172_446_1259_124_1259,JUICE",
                "44,104_1151_438_1151_438_1235_104_1235,JUICE"
            ],
            "text": "JUICE"
        }
    }
}
```

1 难点



(1) *outdoor shooping mall* (2) *pedestrian* (3) *fingerpost* (4) *book openning* (5) *digital screen* (6) *indoor shopping mall*

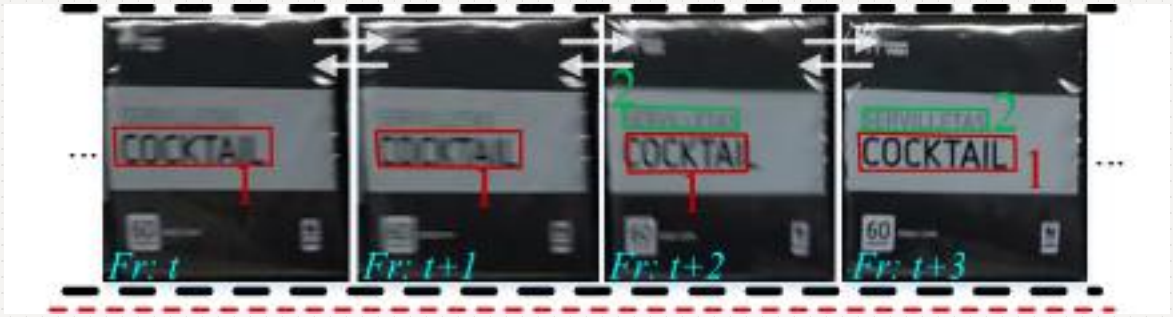
(7) *inside shops* (8) *supermarket* (9) *metro station* (10) *restaurant* (11) *office building* (12) *hotel*

(13) *bus/railway station* (14) *bookstore* (15) *street view* (16) *inside train* (17) *train watch* (21) *shopping bags*

(18) *city road* (19) *harbor surveillance* (20) *highway*

blurring，rotation，illumination.....

1 评测指标



Text Detection → **Text Tracking** → **Text Spotting**

Text Detection	Text Tracking	Text Spotting
IoU-based	Similar to multi-object tracking	Detection+Tracking +recognition
Precision	ATA	Precision
Recall	MOTA	Recall
F-measure	MOTP	F-measure

$$MOTA = 1 - \frac{\sum(FN + FP + IDSW)}{\sum GT} \in (-\infty, 1]$$

1 评测提交结果

User ID	Rank	F-score$_s$	Precision$_s$	Recall$_s$	ATA$_s$	MOTA$_s$	MOTP$_s$	Affiliations
tianqihenhao	1	0.5308	0.6655	0.4414	0.4549	0.5913	0.8421	TEG, Tencent
DXM-DI-AI -CV-TEAM	2	0.4755	0.6435	0.3770	0.4188	0.4960	0.8142	DuXiaoman Financial
panda12	3	0.4183	0.5243	0.3479	0.3579	0.5179	0.8427	IA, CAS
lzneu09	4	0.3007	0.3611	0.2576	0.2737	0.4255	0.8330	Northeastern University
yucheng3	5	0.2964	0.3506	0.2567	0.2711	0.4246	0.8332	University of Chinese Academy of Sciences
tangyejun	6	0.2284	0.2527	0.2084	0.2121	0.3676	0.8337	*
tiendv	7	0.0813	0.1402	0.0572	0.0802	0.0887	0.7976	University of Information Technology
enderloong	8	0.0307	0.0239	0.0429	0.0357	0.0159	0.7813	*
colorr	9	0.0158	0.0085	0.1225	0.0146	0.0765	0.8498	*
weijiawu3	10	0.0077	0.0041	0.0550	0.0088	-0.1530	0.7670	Zhejiang University
BOE_AIoT_CTO	11	0.0000	0.0000	0.0000	0.0000	-0.0003	0.0000	BOE

第二部分

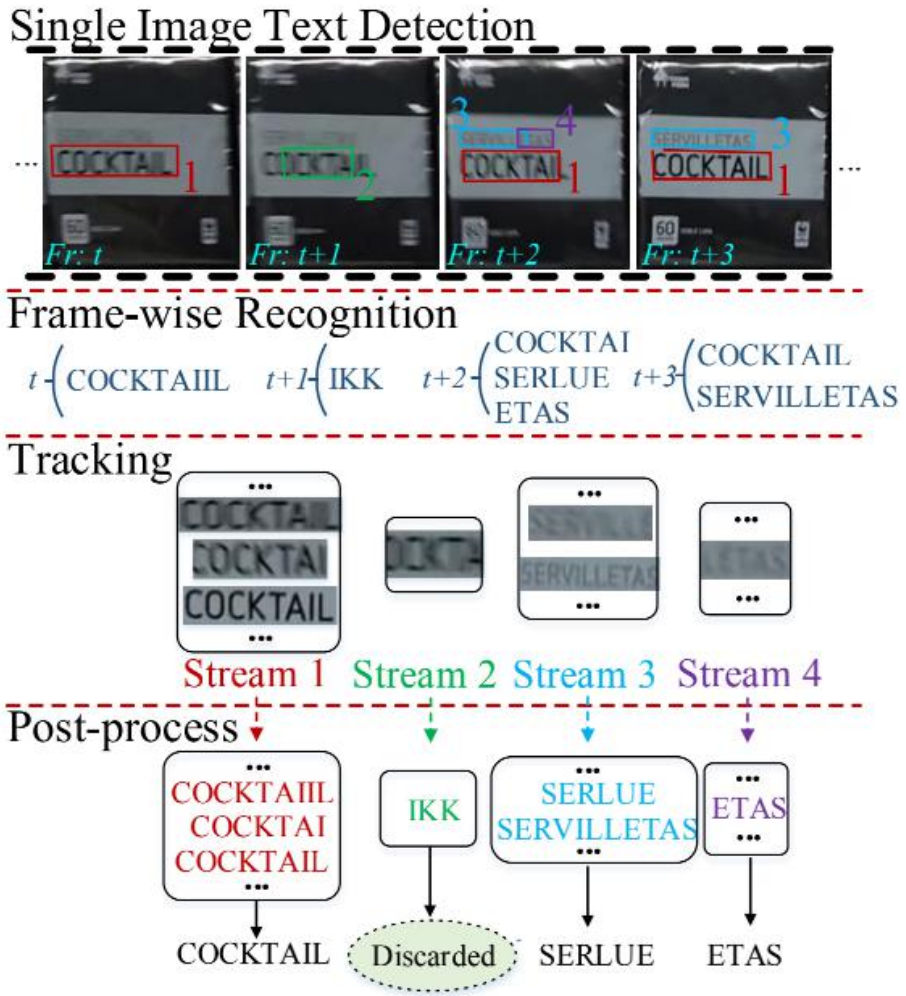
主流方法

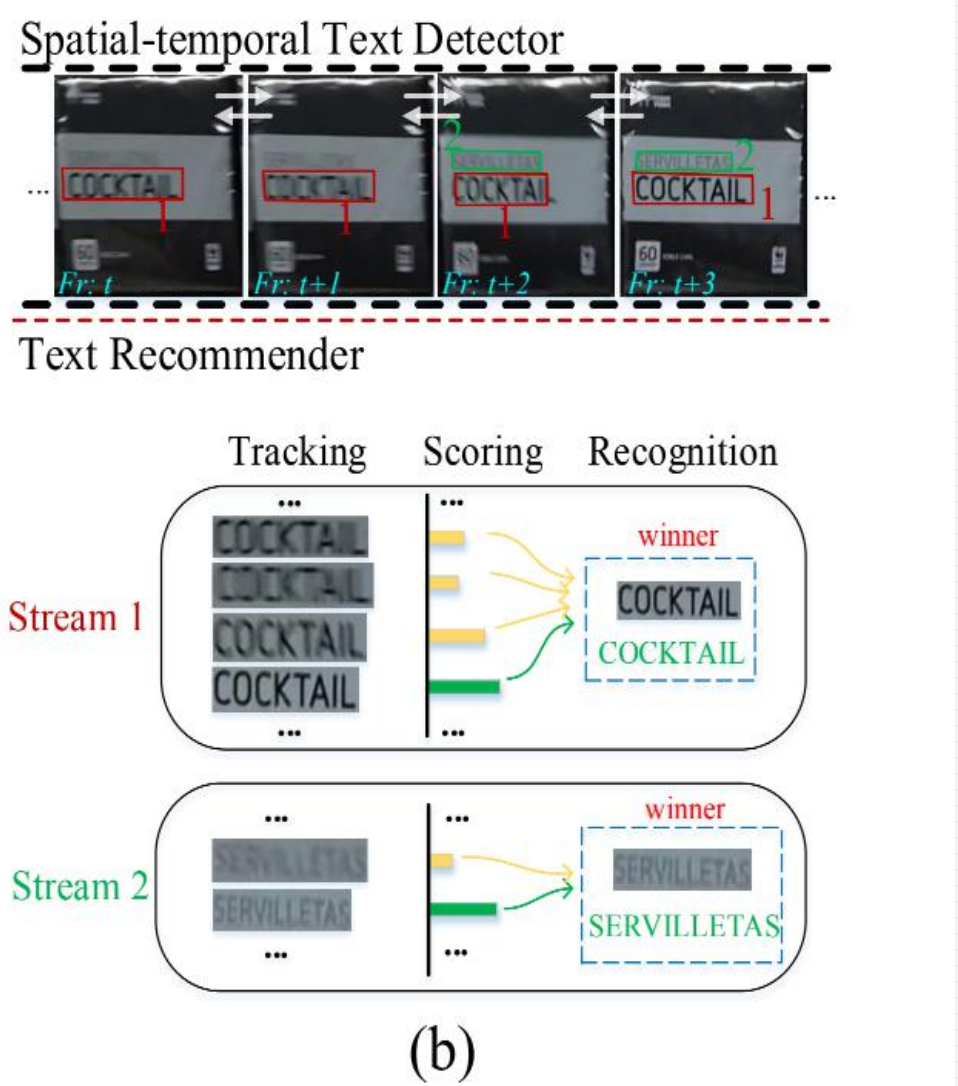
2 方法介绍

You Only Recognize Once: Towards Fast Video Text Spotting

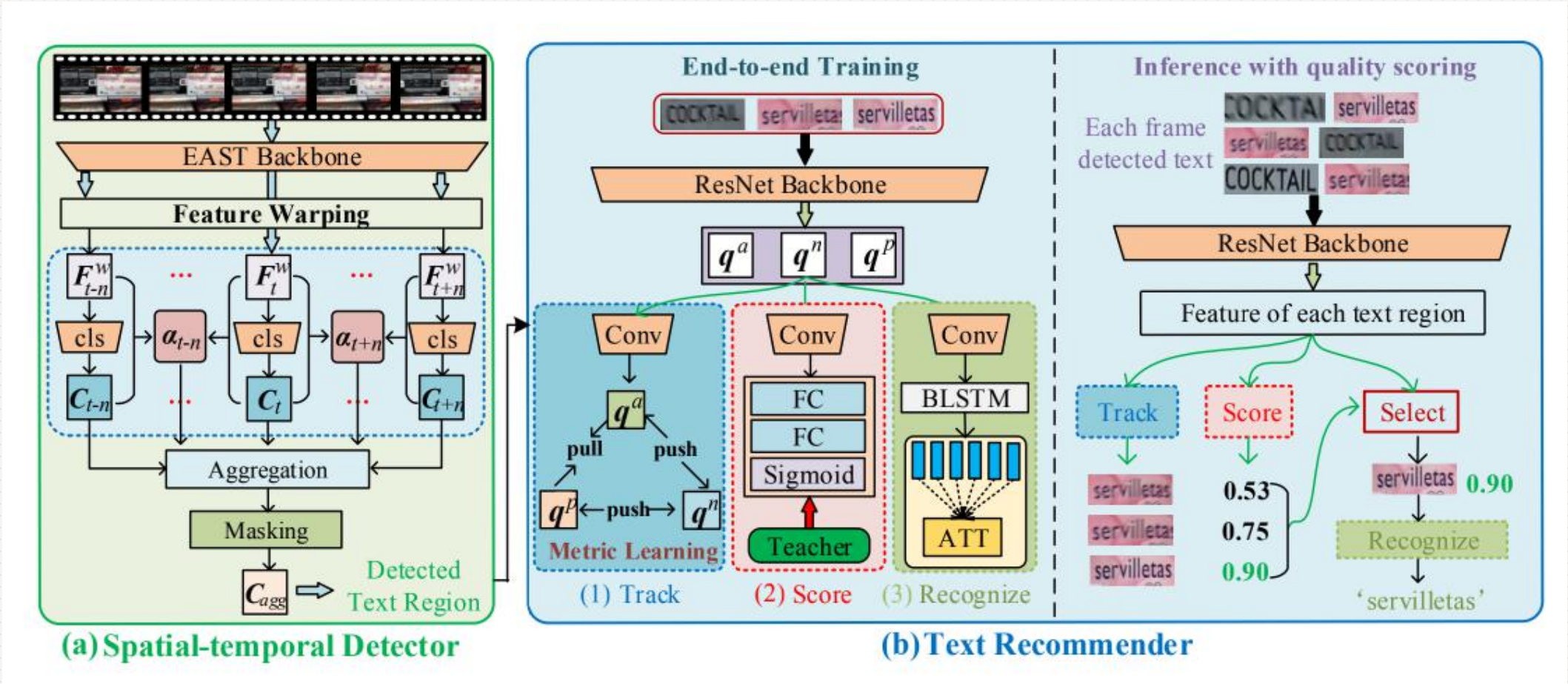
Zhanzhan Cheng[12*], Jing Lu[2*], Yi Niu[2], Shiliang Pu[2], Fei Wu[1+], Shuigeng Zhou[3]
[1]Zhejiang University, Hangzhou, China
chengzhanzhan@hikvision.com,wufei@cs.zju.edu.cn
[2]Hikvision Research Institute, China
lujing6,niuyi,pushiliang@hikvision.com
[3]Fudan University, Shanghai, China
sgzhou@fudan.edu.cn

2 主要贡献

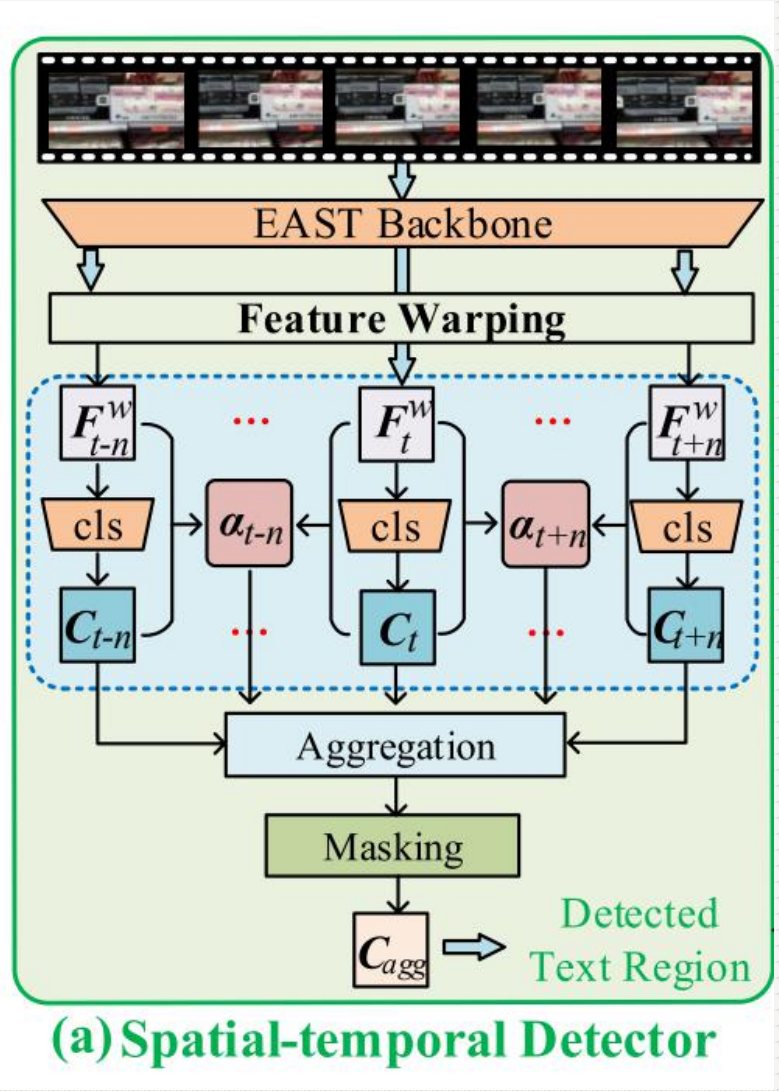

Single Image Text Detection
Fr: t
Fr: t+1
Fr: t+2
Fr: t+3
Frame-wise Recognition
t
COCKTAIIL
t+1
IKK
t+2
COCKTAI
SERLUE
ETAS
t+3
COCKTAIL
SERVILLETAS
Tracking
Stream 1
Stream 2
Stream 3
Stream 4
Post-process
COCKTAIIL
COCKTAI
COCKTAIL
IKK
SERLUE
SERVILLETAS
ETAS
COCKTAIL
Discarded
SERLUE
ETAS
(a)


Spatial-temporal Text Detector
Fr: t
Fr: t+1
Fr: t+2
Fr: t+3
Text Recommender
Tracking
Scoring
Recognition
Stream 1
winner
COCKTAIL
COCKTAIL
Stream 2
winner
SERVILLETAS
SERVILLETAS
(b)

2 Pipeline


EAST Backbone
Feature Warping
cls
Aggregation
Masking
Detected Text Region
(a) Spatial-temporal Detector
End-to-end Training
ResNet Backbone
Conv
FC
Sigmoid
Teacher
BLSTM
ATT
pull
push
Metric Learning
(1) Track
(2) Score
(3) Recognize
Inference with quality scoring
Each frame detected text
Track
Score
Select
0.53
0.75
0.90
0.90
Recognize
'servilletas'
(b) Text Recommender

2 Detection Module



(a) Spatial-temporal Detector

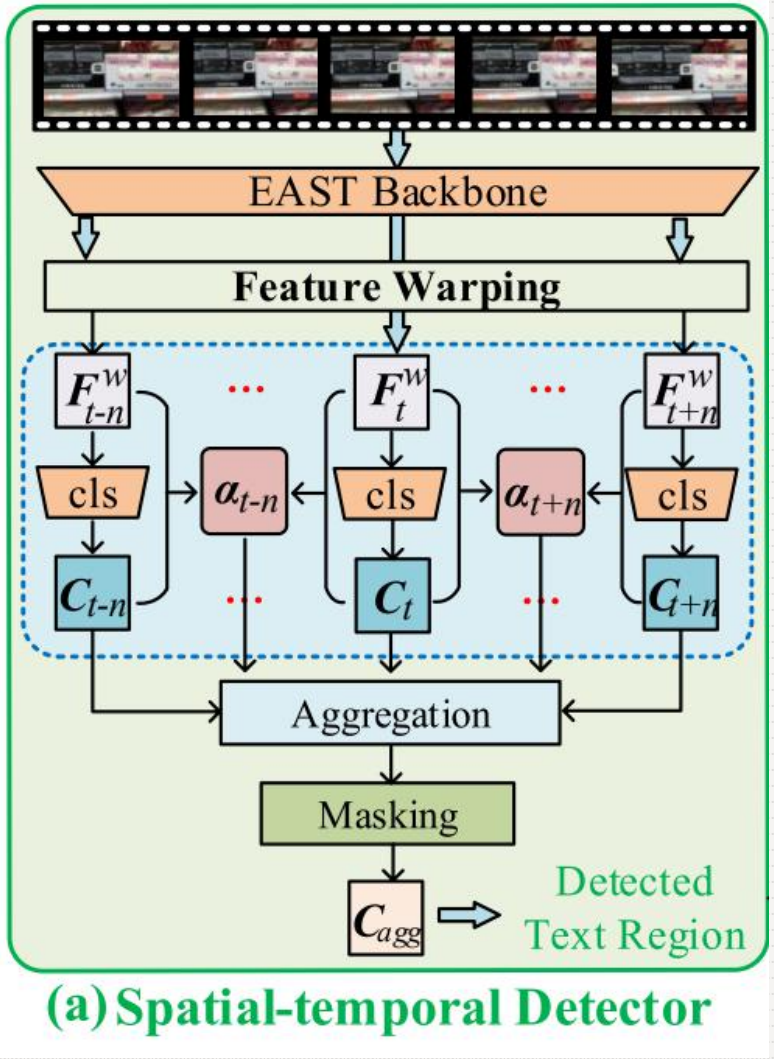
Input $I_{t-n} \cdots\cdots I_t \cdots\cdots I_{t+n}$

Model

Output

Text region mask

2 Detection Module



(a) Spatial-temporal Detector

$I_{t-n} \ldots\ldots I_t \ldots\ldots I_{t+n}$

$F_{t-n} \ldots\ldots F_t \ldots\ldots F_{t+n}$

$$F^w_{t+i} = Warp(F_{t+i}, flow_{(t+i,t)}),$$

MLP-based Classifier

$C_{t-n} \ldots\ldots C_t \ldots\ldots C_{t+n}$

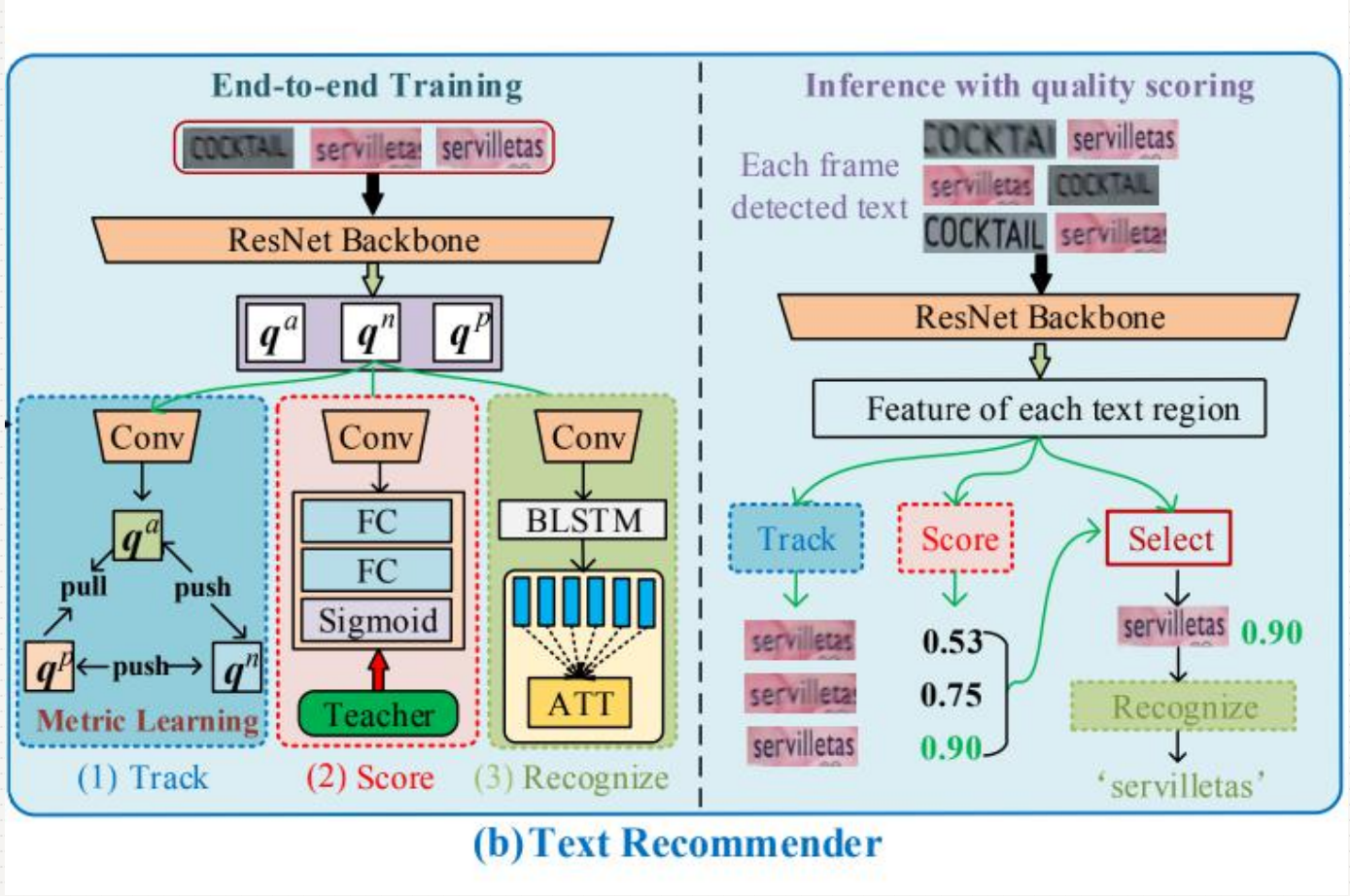
$$F^{trans}_{t+i} = ReLU(BN(WF^w_{t+i} + b)),$$

$$Sim_{t+i,t} = F^{trans}_{t+i} \odot F^{trans}_{t}$$

$$a_{t+i} = \frac{exp(Sim_{t+i,t} \odot C_{t+i})}{\sum_{i'=-n}^{n} exp(Sim_{t+i',t} \odot C_{t+i'})}.$$

$$C_{t,agg} = \sum_{i=-n}^{n} a_{t+i} * C_{t+i}$$

2 Text Recommender



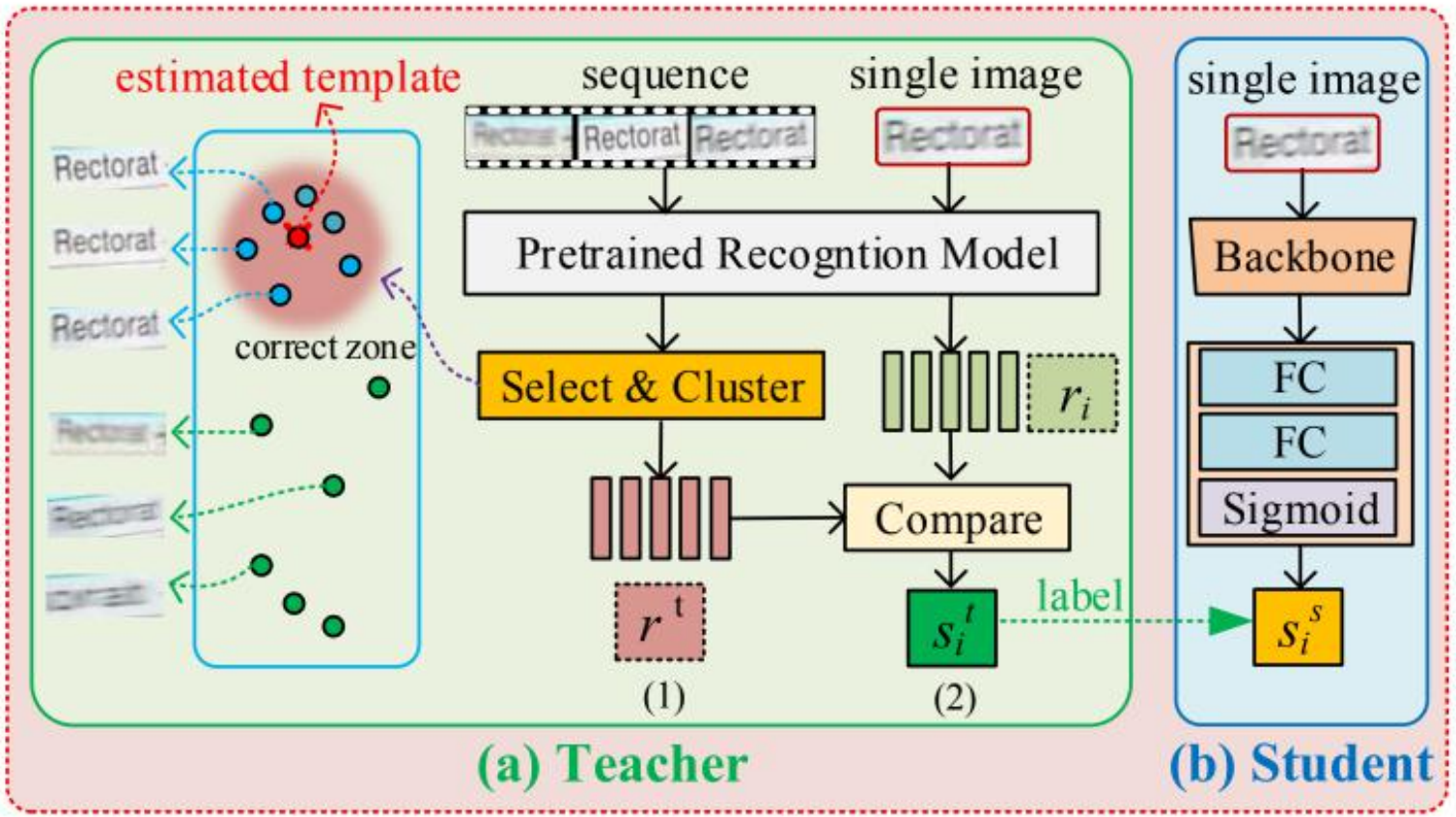
(b) Text Recommender

Quality Score

Tracking

Recognizing

2 Text Recommender--Quality score

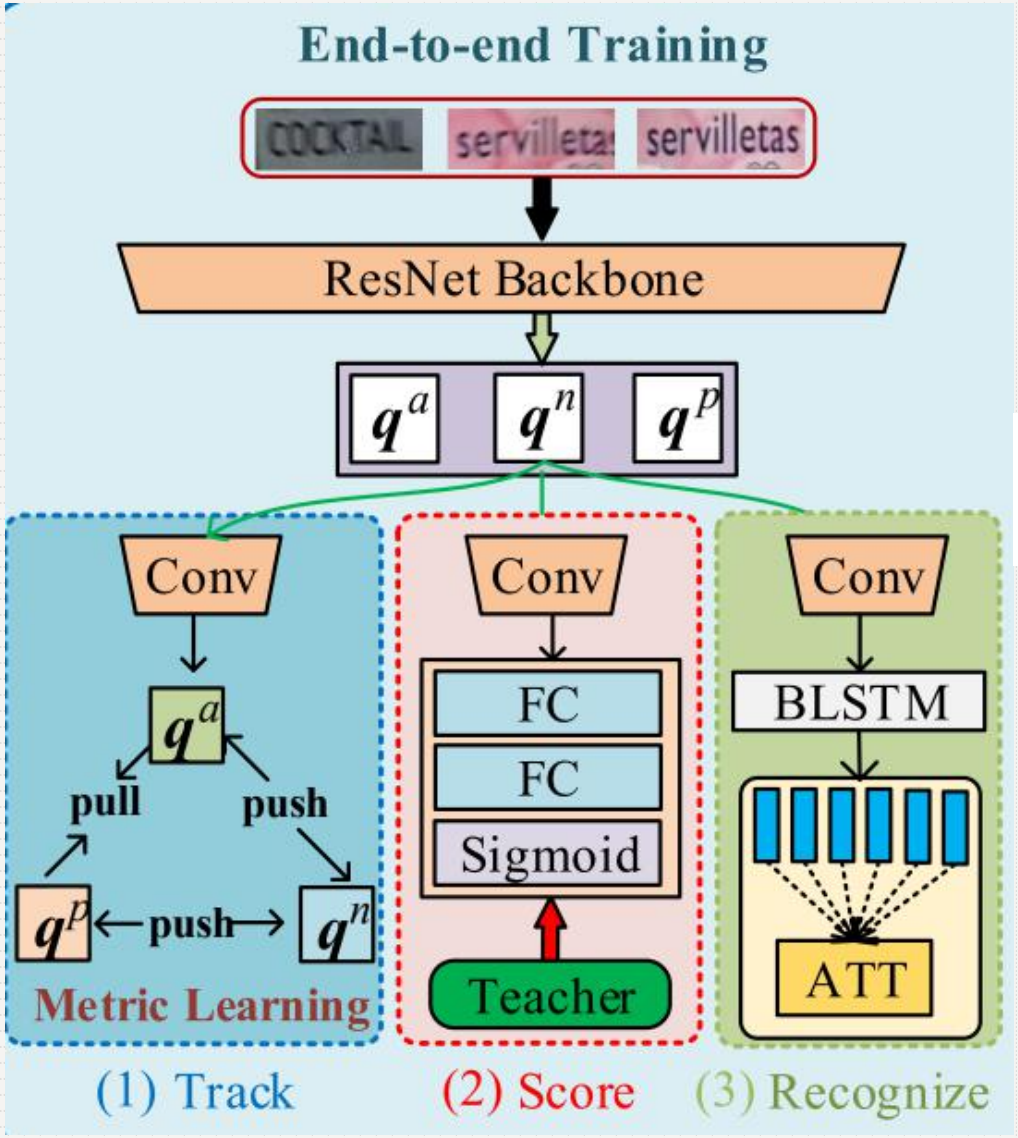


$$r^t = kmeans(r_1^{cor}, r_1^{cor}, ..., r_k^{cor}),$$

$$s_i^t = \frac{r^t \odot r_i}{||r^t|| * ||r_i||},$$

$$\mathcal{L}_S = \frac{1}{N} \sum_{i=0}^{N} ||s_i^t - s_i^s||$$

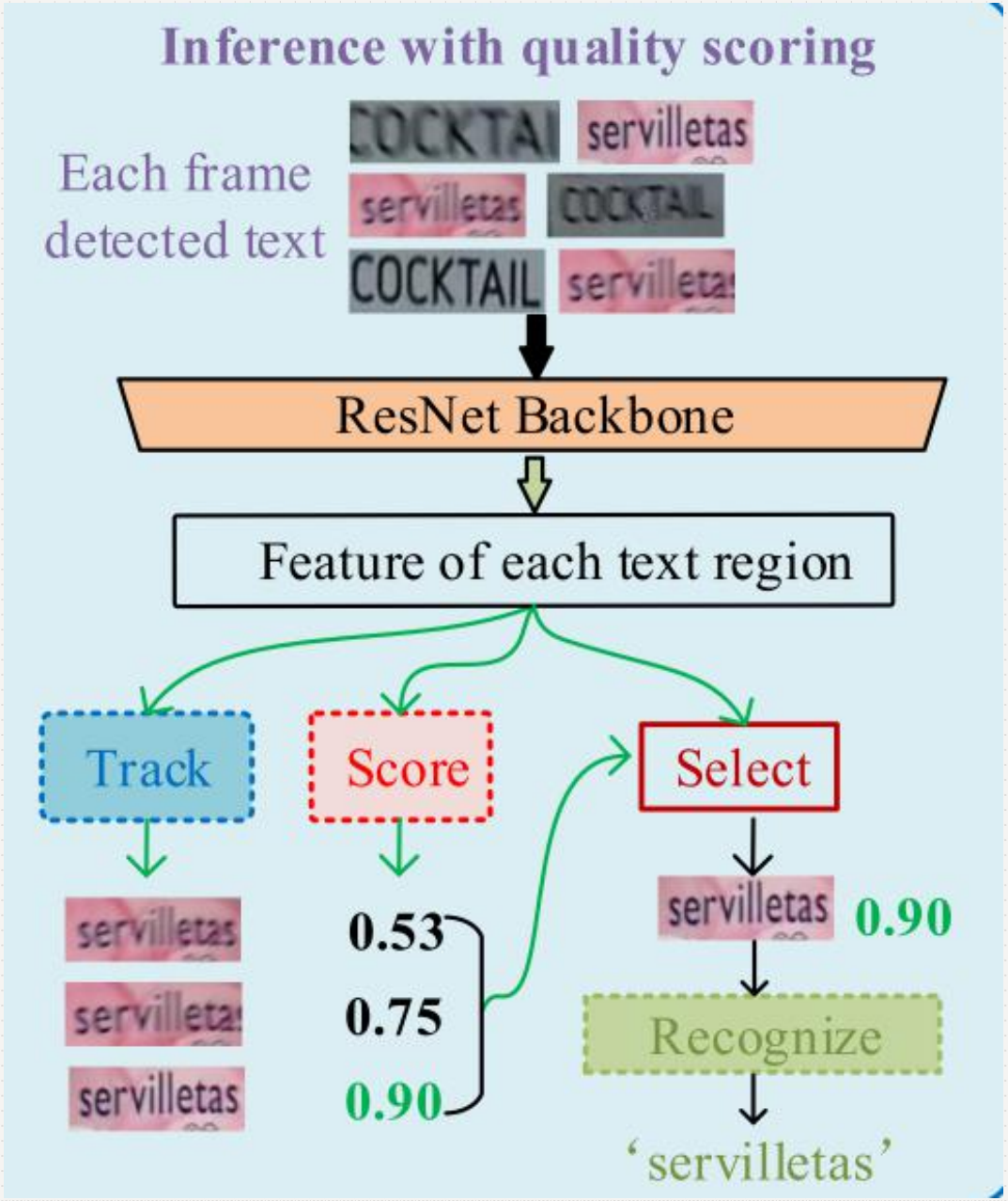
2 Text Recommender--Tracking+Recognition



$$\mathcal{L}_T = \mathcal{L}_{contra} + \lambda_t \mathcal{L}_{triplet},$$

$$\mathcal{L} = \lambda_1 \mathcal{L}_T + \lambda_2 \mathcal{L}_S + \lambda_3 \mathcal{L}_R,$$

2 Inference


Inference with quality scoring
Each frame detected text
ResNet Backbone
Feature of each text region
Track
Score
Select
0.53
0.75
0.90
servilletas
0.90
Recognize
'servilletas'

2 Ablation Experiments

Methods	*QSHR* (IC13/IC15)	*RCR* (IC13/IC15)	*FPS*
PCW	74.55/75.83	66.06/66.32	4.52
HFP	75.32/76.34	68.30/68.56	
TR ($\mathcal{L}_S$)	77.89/79.69	68.89/69.41	324.58
TR ($\mathcal{L}_S$+$\mathcal{L}_T$)	78.64/80.36	69.12/69.82	
TR ($\mathcal{L}_S$+$\mathcal{L}_R$)	81.23/83.03	69.92/70.69	
TR ($\mathcal{L}$)	**81.74/83.29**	**70.18/70.95**	

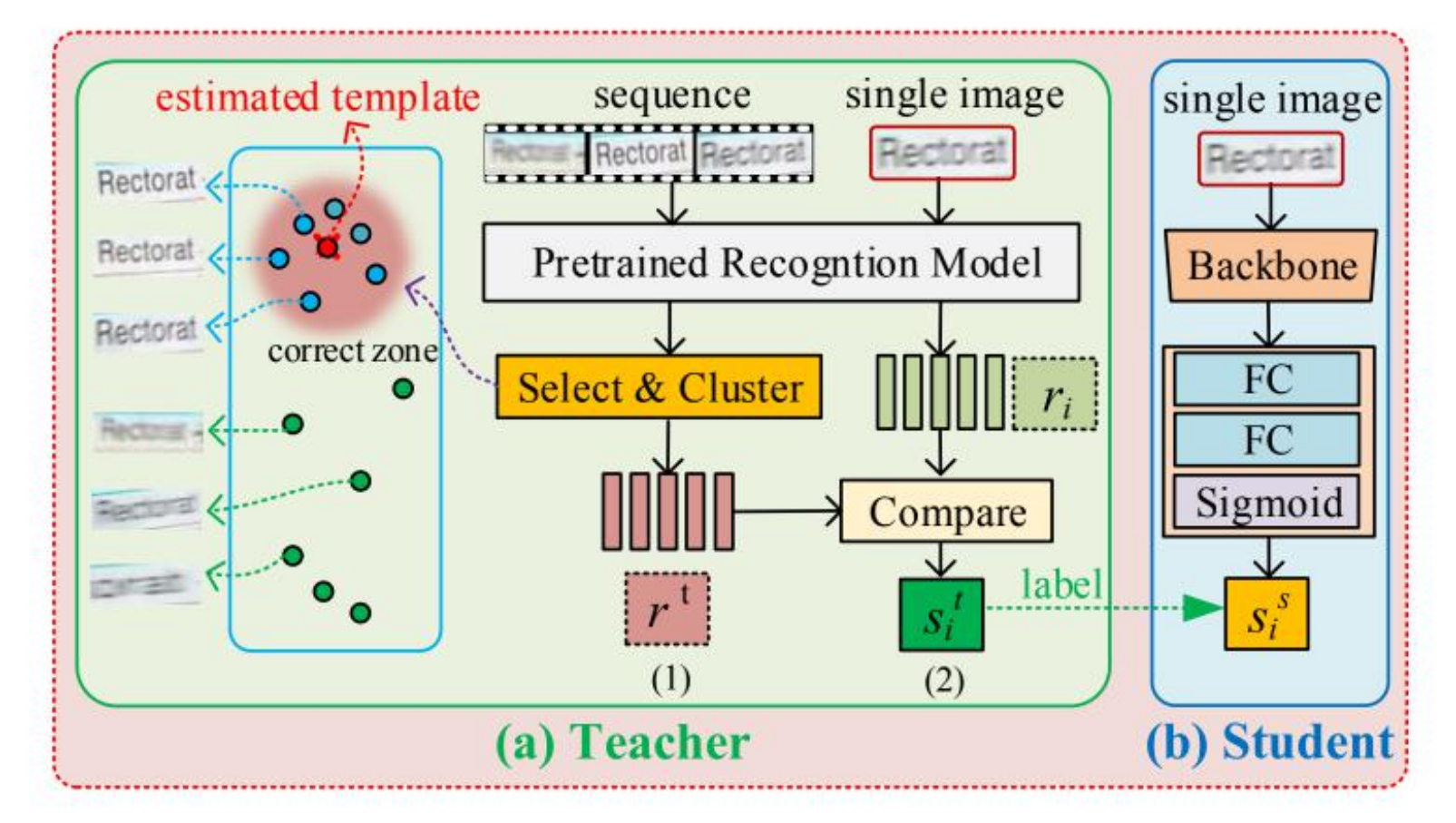
Methods	*QSHR* (IC13/IC15)	*RCR* (IC13/IC15)
PCW	41.73/45.66	59.78/60.62
HFP	39.37/41.73	58.96/60.06
TR	**51.18/54.33**	**66.14/67.71**

第三部分

未来展望

3 展望1--具体任务中视频场景识别技术的运用

3 展望2--关于模糊图像识别的新思路



感谢聆听!

THANK YOU FOR WATCHING!

2021年9月