Paper Survey and Some Thoughts for Scene Text Recognition

Tsai-Shien Chen (陳在賢)

Wednesday, October 13, 2021

Media IC and System Lab
Graduate Institute of Electronics Engineering
National Taiwan University

Outline

- Introduction: Scene Text Recognition
- Introduction: Contrastive Learning
- How can Contrastive Learning help?

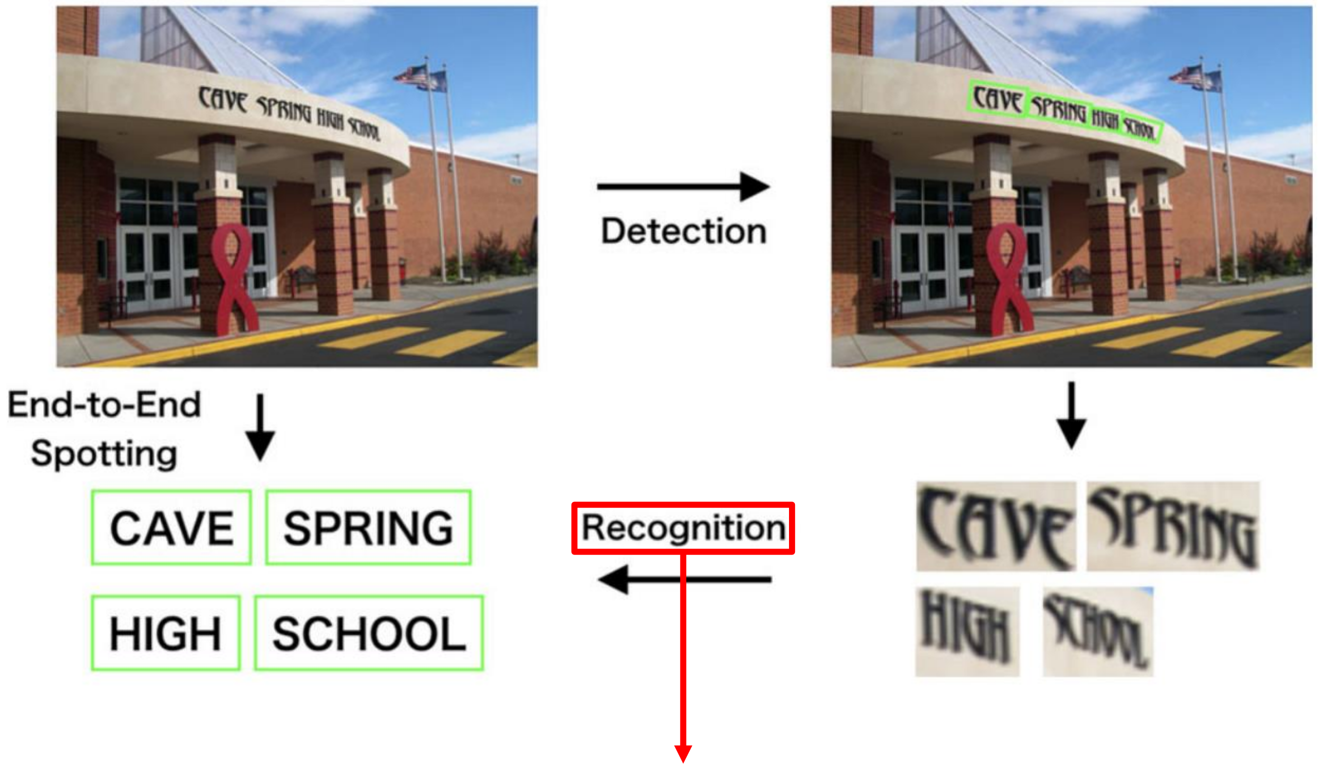
Outline

- Introduction: Scene Text Recognition
- Introduction: Contrastive Learning
- How can Contrastive Learning help?

What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis, ICCV 2019 Oral (citation: 182)
Scene text detection and recognition: The deep learning era, IJCV 2021 (citation: 145)

Scene Text Recognition

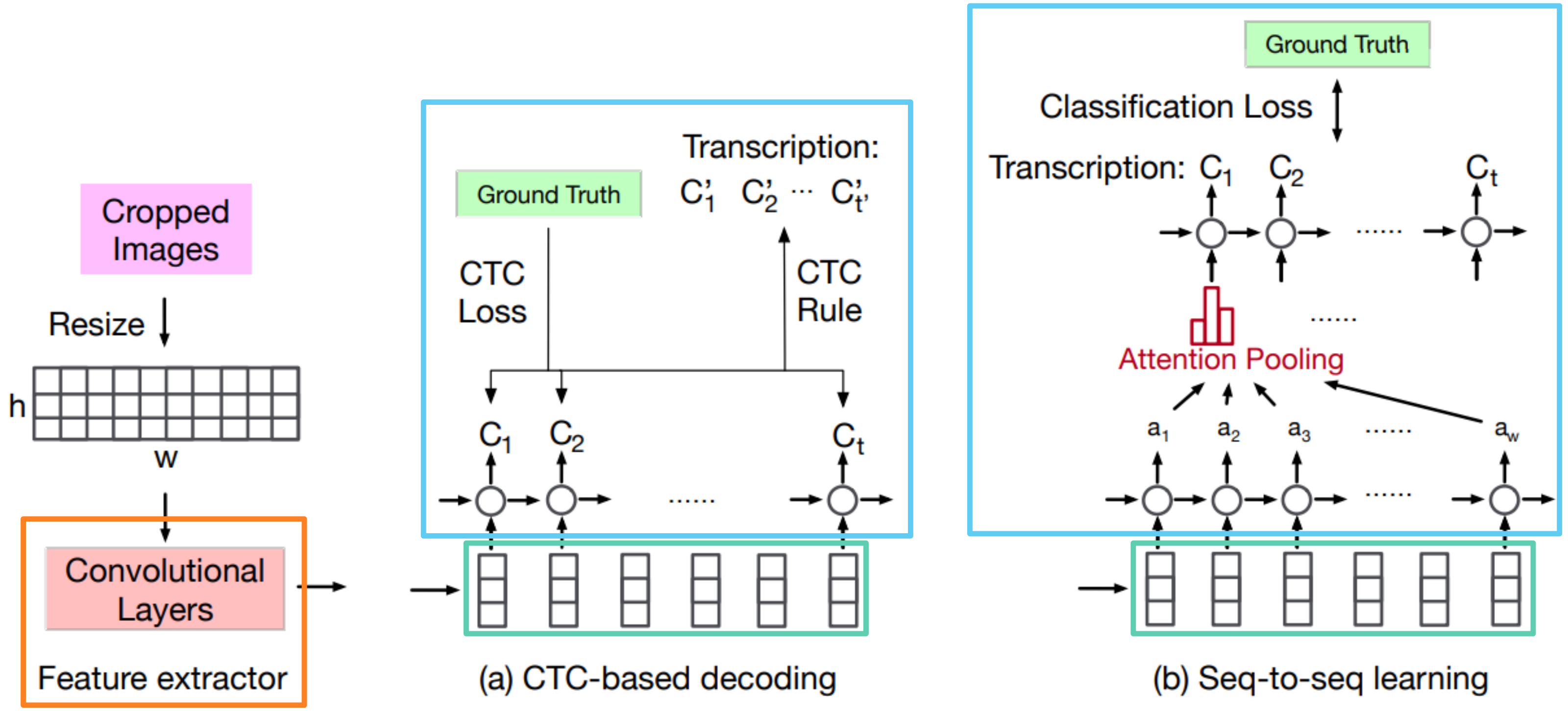
- Task definition

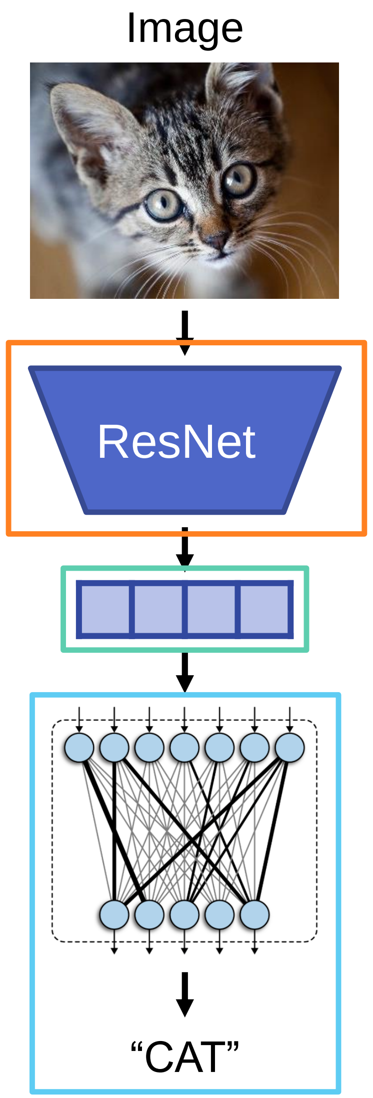


More related to representation learning!!
I will focus on Scene Text Recognition.

Scene Text Recognition

- Previous pipeline (2-stage)

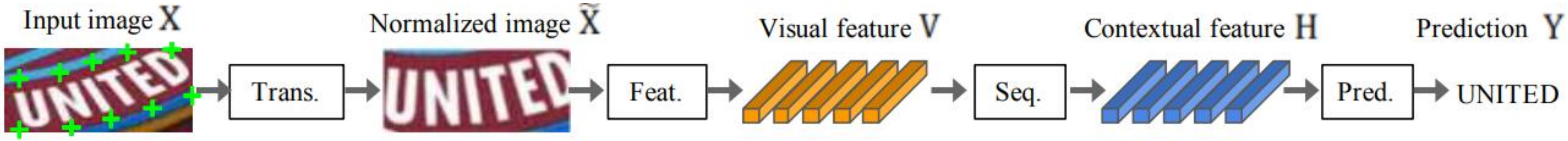




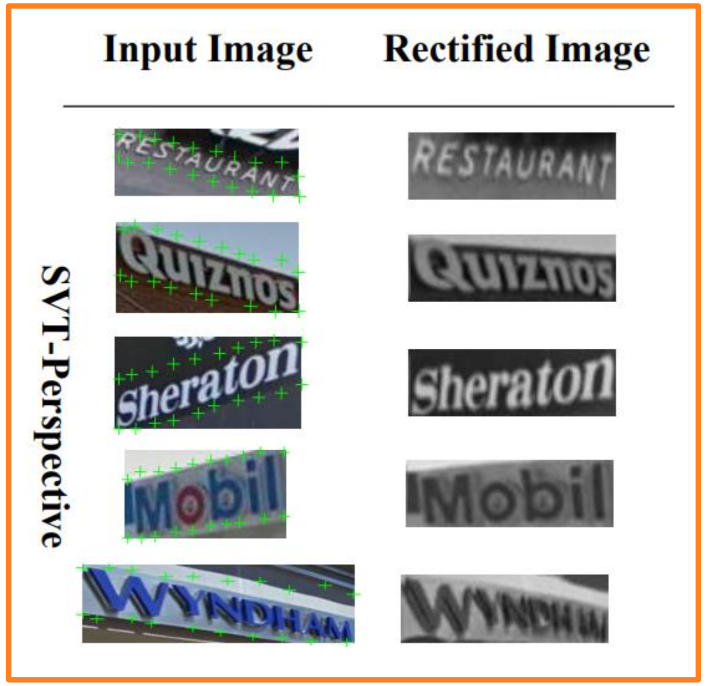
Encoder　Representation　Decoder

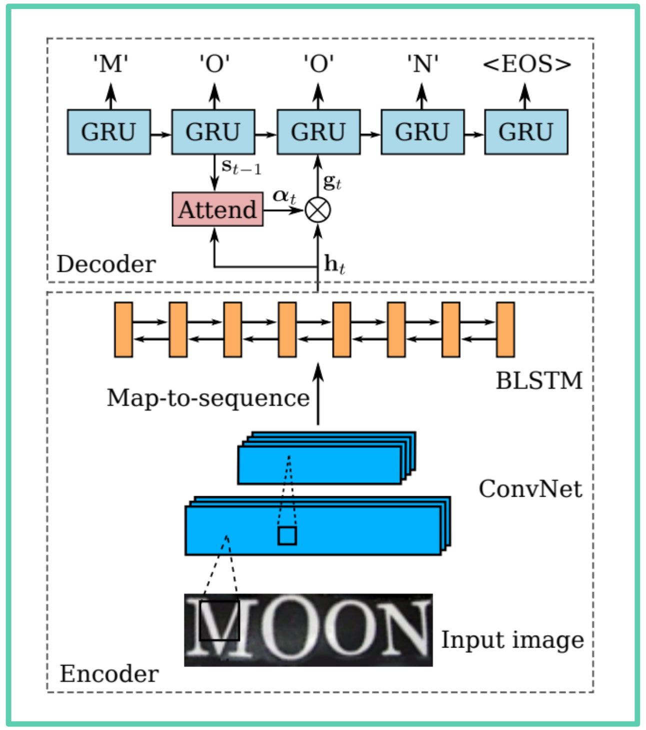
Scene Text Recognition

- State-of-the-art pipeline (4-stage)



  - Transformation
  - Feature extraction (encoder)
  - Sequence modeling
  - Prediction (decoder)





Robust Scene Text Recognition with Automatic Rectification, CVPR 2016 (citation: 401)

Scene Text Recognition

- Experiment environment

	Model	Year	Train data	IIIT 3000	SVT 647	IC03 860	IC03 867	IC13 857	IC13 1015	IC15 1811	IC15 2077	SP 645	CT 288	Time ms/image	params $\times 10^6$
Reported results	CRNN [23]	2015	MJ	78.2	80.8	89.4	–	–	86.7	–	–	–	–	160	8.3
	RARE [24]	2016	MJ	81.9	81.9	90.1	–	88.6	–	–	–	71.8	59.2	<2	–
	R2AM [15]	2016	MJ	78.4	80.7	88.7	–	–	90.0	–	–	–	–	2.2	–
	STAR-Net [17]	2016	MJ+PRI	83.3	83.6	89.9	–	–	89.1	–	–	73.5	–	–	–
	GRCNN [26]	2017	MJ	80.8	81.5	91.2	–	–	–	–	–	–	–	–	–
	ATR [28]	2017	PRI+C	–	–	–	–	–	–	–	–	**75.8**	69.3	–	–
	FAN [4]	2017	MJ+ST+C	87.4	85.9	–	94.2	–	93.3	70.6	–	–	–	–	–
	Char-Net [16]	2018	MJ	83.6	84.4	**91.5**	–	90.8	–	–	60.0	73.5	–	–	–
	AON [5]	2018	MJ+ST	87.0	82.8	–	91.5	–	–	–	**68.2**	73.0	**76.8**	–	–
	EP [2]	2018	MJ+ST	88.3	**87.5**	–	94.6	–	**94.4**	**73.9**	–	–	–	–	–
	Rosetta [3]	2018	PRI	–	–	–	–	–	–	–	–	–	–	–	–
	SSFL [18]	2018	MJ	**89.4**	87.1	–	**94.7**	**94.0**	–	–	–	73.9	62.5	–	–
Our experiment	CRNN [23]	2015	MJ+ST	82.9	81.6	93.1	92.6	91.1	89.2	69.4	64.2	70.0	65.5	4.4	8.3
	RARE [24]	2016	MJ+ST	86.2	85.8	93.9	93.7	92.6	91.1	74.5	68.9	76.2	70.4	23.6	10.8
	R2AM [15]	2016	MJ+ST	83.4	82.4	92.2	92.0	90.2	88.1	68.9	63.6	72.1	64.9	24.1	2.9
	STAR-Net [17]	2016	MJ+ST	87.0	86.9	94.4	94.0	92.8	91.5	76.1	70.3	77.5	71.7	10.9	48.7
	GRCNN [26]	2017	MJ+ST	84.2	83.7	93.5	93.0	90.9	88.8	71.4	65.8	73.6	68.1	10.7	4.6
	Rosetta [3]	2018	MJ+ST	84.3	84.7	93.4	92.9	90.9	89.0	71.2	66.0	73.8	69.2	4.7	44.3
	Our best model		MJ+ST	**87.9**	**87.5**	**94.9**	**94.4**	**93.6**	**92.3**	**77.6**	**71.8**	**79.2**	**74.0**	27.6	49.6

Scene Text Recognition

- Experiment environment (training)

	Model	Year	Train data	IIIT	SVT	IC03		IC13		IC15		SP	CT	Time	params
				3000	647	860	867	857	1015	1811	2077	645	288	ms/image	$\times 10^6$
Reported results	CRNN [23]	2015	MJ												
	RARE [24]	2016	MJ												
	R2AM [15]	2016	MJ												
	STAR-Net [17]	2016	MJ+PRI												
	GRCNN [26]	2017	MJ												
	ATR [28]	2017	PRI+C												
	FAN [4]	2017	MJ+ST+C												
	Char-Net [16]	2018	MJ												
	AON [5]	2018	MJ+ST												
	EP [2]	2018	MJ+ST												
	Rosetta [3]	2018	PRI												
	SSFL [18]	2018	MJ												
Our experiment	CRNN [23]	2015	MJ+ST												
	RARE [24]	2016	MJ+ST												
	R2AM [15]	2016	MJ+ST												
	STAR-Net [17]	2016	MJ+ST												
	GRCNN [26]	2017	MJ+ST												
	Rosetta [3]	2018	MJ+ST												
	Our best model		MJ+ST												

MJSynth (MJ): 8.9 M word boxes
SynthText (ST): 5.5 M word boxes

(a) MJSynth word boxes

(b) SynthText scene image

The large-scale training datasets are all synthetic…

Scene Text Recognition

- Testing environment

Real world scene text are relatively small-scale…
(Manual label annotations are too expensive…)

Benchmark (regular)

Benchmark	Description	# Train.	# Eval.
SVT	from Google Street View	100	250
IIIT5K	from Google image searches with querying “billboards” and “posters”	2000	3000
ICDAR 2013	for ICDAR 2013 competition	229	233

Benchmark (irregular)

Benchmark	Description	# Train.	# Eval.
ICDAR 2015	collected with Google Glass. contains perspective or blurry images	1000	500
SVT Perspective	collected from Google Street View contains perspective texts	-	639
CUTE80	captured by digital cameras or collected from the Internet.	-	80

Scene Text Recognition

- Some problems in scene text recognition…
  - Pre-training models of the encoder (ResNet) is based on ImageNet.
  - The training datasets are all synthetic.
  - The training outcome might be suboptimal to the real-world images.

- Potential Solution
  - A large-scale real-world unlabeled textual datasets
  - Self-supervised learning framework
    - SimCLR, MoCo, BYOL, …

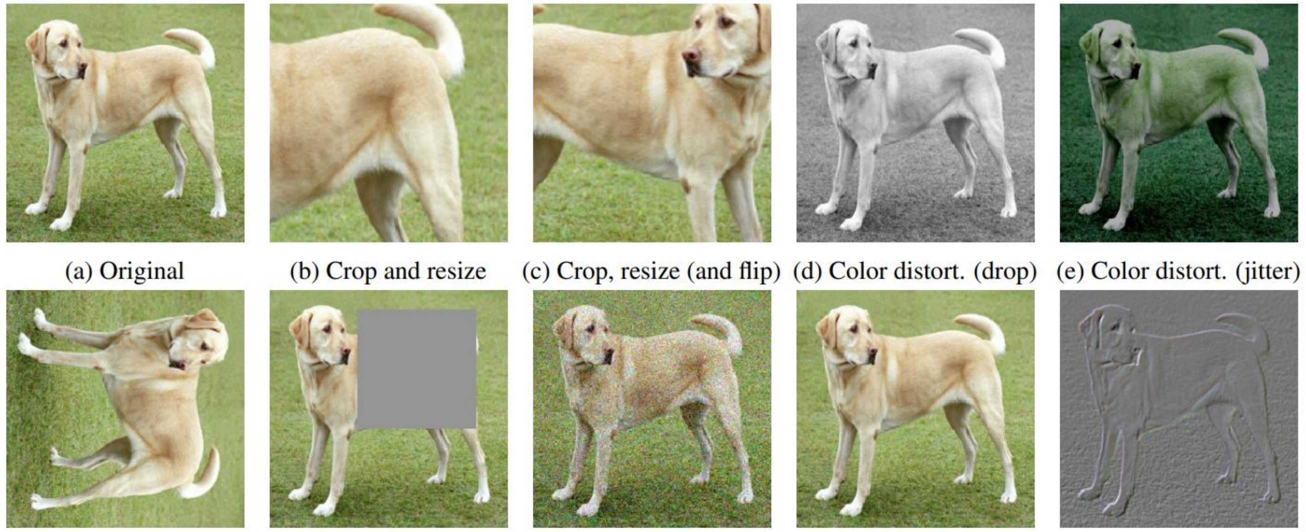
Outline

- Introduction: Scene Text Recognition
- Introduction: Contrastive Learning
- How can Contrastive Learning help?

A self-supervised learning framework which can train a good pre-training encoder without using labeled data!!

Contrastive Learning

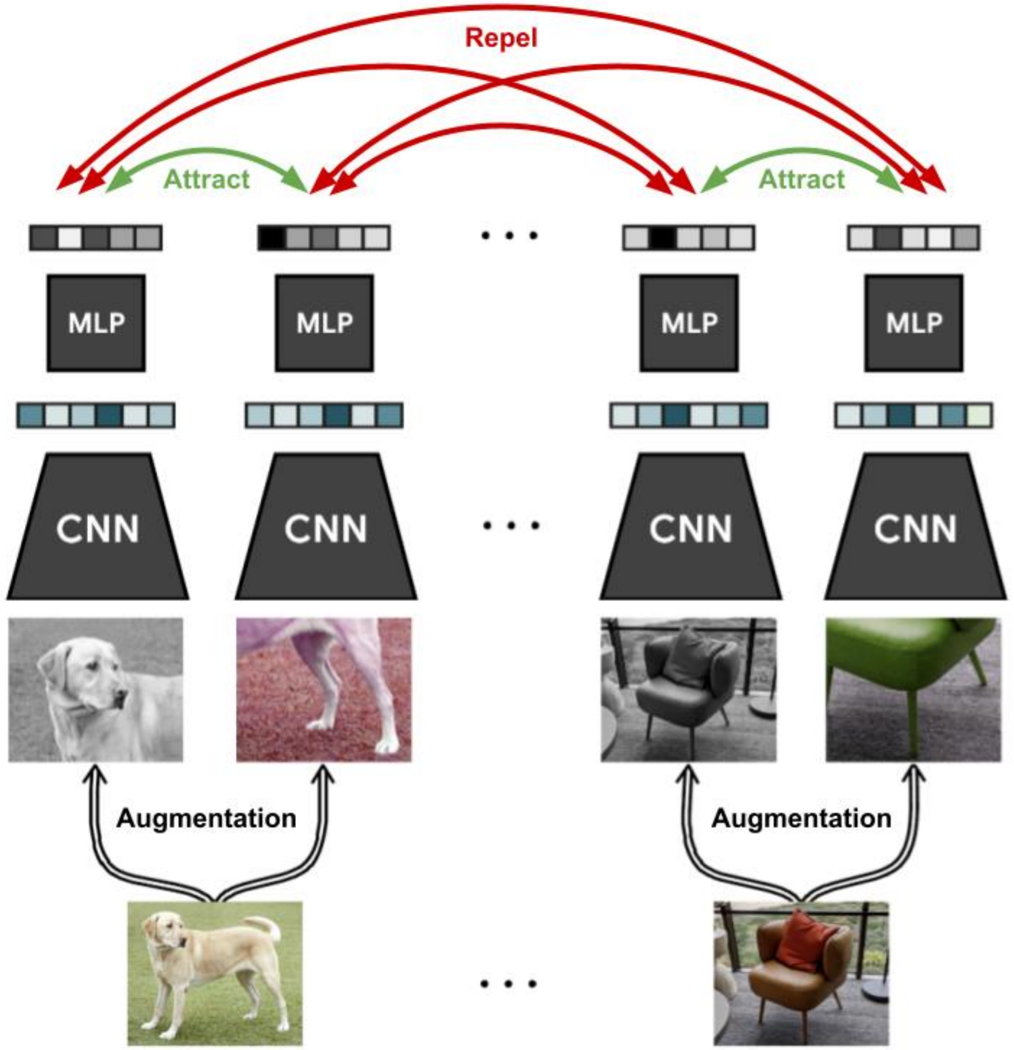
- A representative framework: SimCLR
  - A set of image augmentation



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)
(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

  - InfoNCE loss

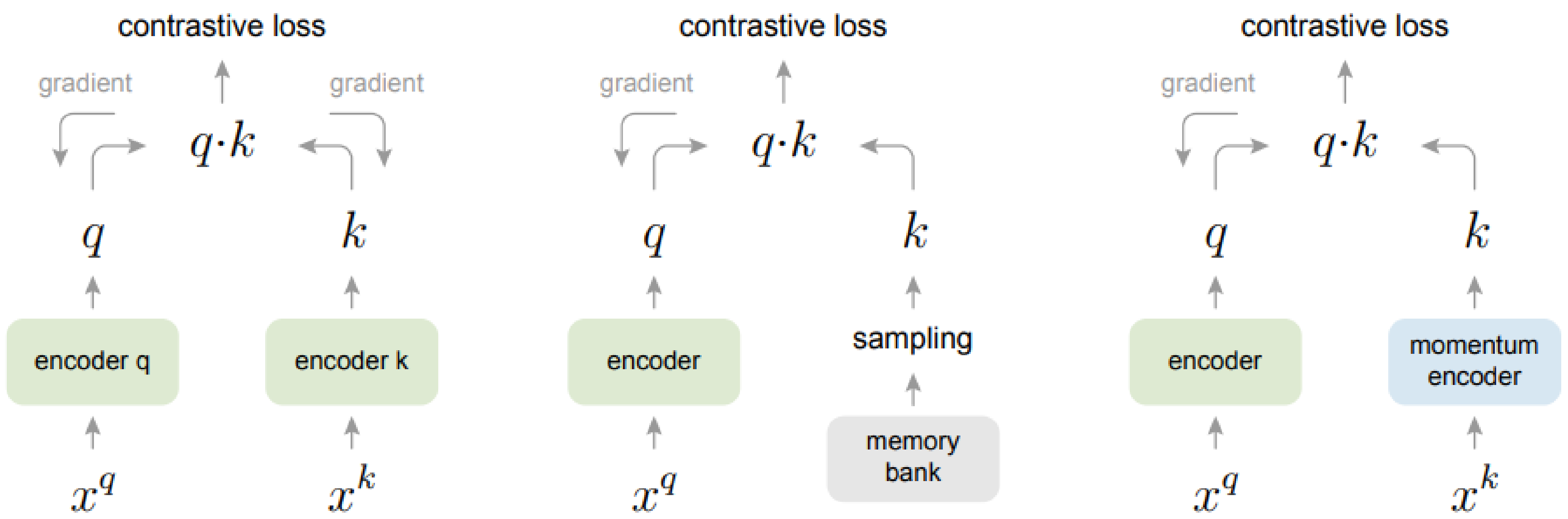
$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$



A Simple Framework for Contrastive Learning of Visual Representations, ICML 2020 (citation: 2263)

Contrastive Learning

- Disadvantage of SimCLR
  - batchsize should be large (4096) to get enough negative samples…
- Solution: momentum contrastive learning (MoCo)

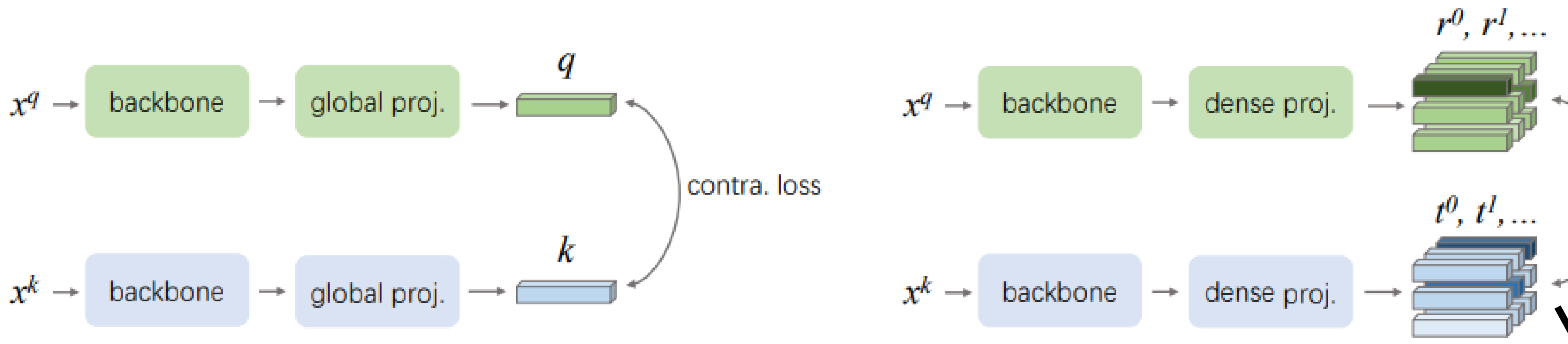


(a) end-to-end (b) memory bank (c) MoCo

Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020 (citation: 1805)

Contrastive Learning

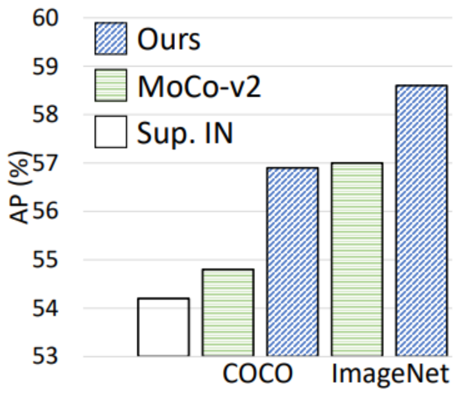
- Disadvantage of vanilla contrastive learning
  - No spatial information which is suboptimal for dense prediction task (e.g., semantic segmentation, object detection)
  - Features for scene text recognition also contain spatial information!!
- Solution: dense contrastive learning



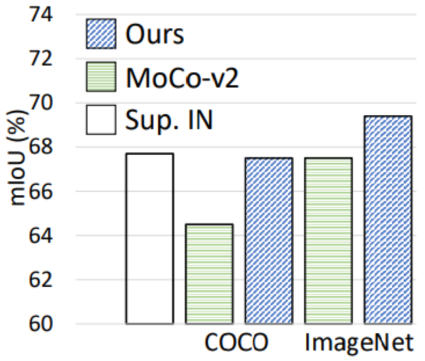
(a) Global Contrastive Learning **(b)** Dense Contrastive Learning

$$\mathcal{L} = (1-\lambda)\mathcal{L}_q + \lambda\mathcal{L}_r$$

Dense Correspondence across Views: Alignment the feature within different views



(a) Object Detection



(b) Semantic Segmentation

Dense Contrastive Learning for Self-Supervised Visual Pre-Training, CVPR 2021 Oral (citation: 35)

Outline

- Introduction: Scene Text Recognition
- Introduction: Contrastive Learning
- How can Contrastive Learning help?

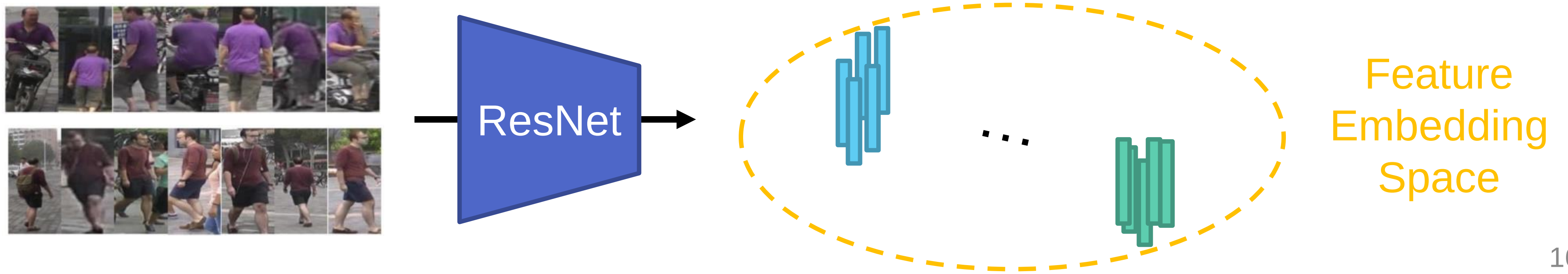
How can Contrastive Learning help?

- Contrastive learning has been applied to person re-identification
  - Maybe we can refer to their methods!! Unsupervised Pre-training for Person Re-identification, CVPR 2021
- But, first, what is re-identification?
  - Re-identification aims to give a single ID to the images of a same target.

ID #1 ID #3

ID #2 ID #4

  - Straightforward Solution:

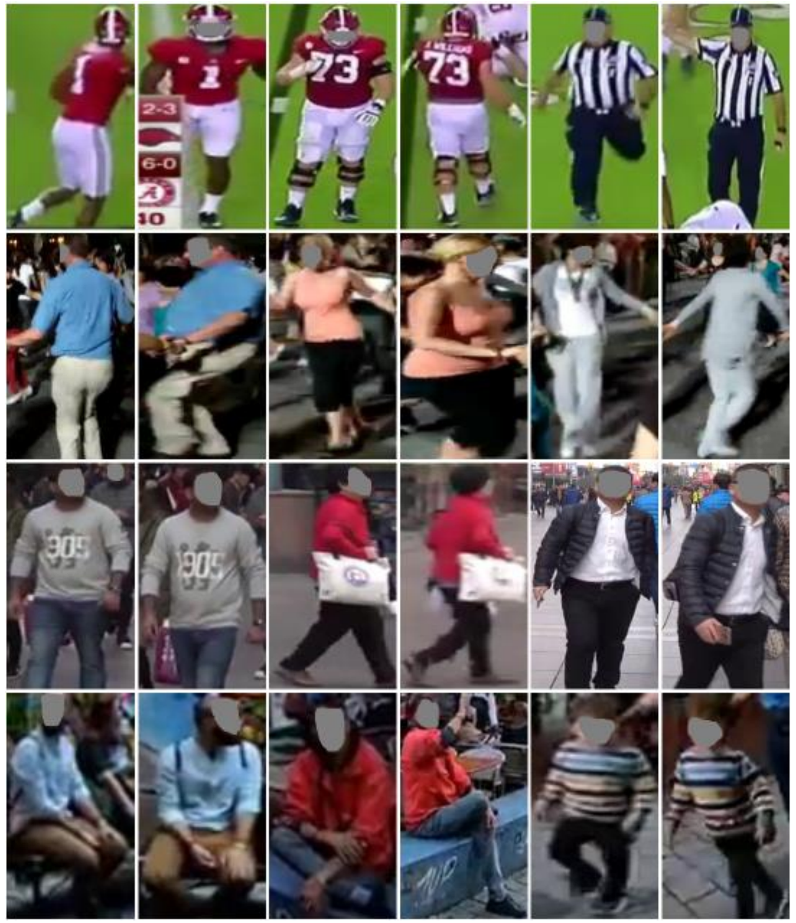


How can Contrastive Learning help?

- Re-identification has similar problems with scene text recognition
  - Existing datasets are in limited scales due to difficult data annotations
  - ImageNet pre-training models are not optimal for target task
    - Especially, the target tasks all use person or textual images
- Solution: large-scale unlabeled dataset + contrastive learning
- We will respectively discuss it in the following!!

How can Contrastive Learning help?

- Large-scale unlabeled dataset
  - Crawl YouTube video (query word: “cityname + streetview/scene”) and use YOLO-v5 to crop the person images
  - Advantage:
    - Large-scale (73k videos, 4.2M images)
    - Diverse places, lighting, ethnic, pose, resolution, etc.



Datasets	#images	#scene	#persons	environment	camera view	resolution	detector	crop size
VIPeR[16]	1,264	2	632	-	fixed	fixed	hand	128×48
GRID[28]	1,275	8	1,025	subway	fixed	fixed	hand	vary
CUHK03[26]	14,096	2	1,467	campus	fixed	fixed	DPM[12]+hand	vary
Market[44]	32,668	6	1,501	campus	fixed	fixed	DPM[12]+hand	128×64
Airport[25]	39,902	6	9,651	airport	fixed	fixed	ACF[11]	128×64
DukeMTMC[47]	36,411	8	1,852	campus	fixed	fixed	Hand	vary
MSMT17[39]	126,441	15	4,101	campus	fixed	fixed	FasterRCNN[32]	vary
LUPerson	4,180,243	46,260	$> 200k$	vary	dynamic	dynamic	YOLOv5	vary

How can Contrastive Learning help?

- Contrastive learning framework for re-identification
  - Similar to MoCo but they verify and change several augmentations:
  - Remove: color distortion related augmentation (channel drop, color jitter)
  - Add: RandomErasing (task-specific augmentation)
- Contrastive learning framework for scene text recognition
  - Use dense contrastive learning to preserve spatial information of features
  - Add some “task-specific augmentations” to make the model more robust
    - Random affine/TPS transformation, simulated over-explosion, … etc.

How can Contrastive Learning help?

- Experiments

ImageNet supervised pre-training
+ state-of-the-art re-id framework

LUP unsupervised pre-training
+ strong re-id baseline

Method	CUHK03	Market1501	DukeMTMC	MSMT17
PCB [36] (2018)	57.5/63.7	81.6/93.8	69.2/83.3	-
MGN [38] (2018)	67.4/68.0	86.9/95.7	78.4/88.7	-
MGN*	70.5/71.2	87.5/95.1	79.4/89.0	63.7/85.1
BOT [29] (2019)	-	85.9/94.5	76.4/86.4	-
DGNet [46] (2019)	-	86.0/94.8	74.8/86.6	52.3/77.2
IANet [23] (2019)	-	83.1/94.4	73.4/87.1	46.8/75.5
DSA [43] (2019)	75.2/78.9	87.6/95.7	74.3/86.2	-
Auto [31] (2019)	73.0/77.9	85.1/94.5	-	52.5/78.2
ABDNet [5] (2019)	-	88.3/95.6	78.6/89.0	60.8/82.3
OSNet [50] (2019)	67.8/72.3	84.9/94.8	73.5/88.6	52.9/78.7
SCAL [4] (2019)	72.3/74.8	89.3/95.8	79.6/89.0	-
P2Net [18] (2019)	73.6/78.3	85.6/95.2	73.1/86.5	-
MHN [2] (2019)	72.4/77.2	85.0/95.1	77.2/89.1	-
BDB [10] (2019)	76.7/79.4	86.7/95.3	76.0/89.0	-
SONA [41] (2019)	79.2/81.8	88.8/95.6	78.3/89.4	-
GCP [30] (2020)	75.6/77.9	88.9/95.2	78.6/87.9	-
SAN [24] (2020)	76.4/80.1	88.0/96.1	75.5/87.9	55.7/79.2
ISP [51] (2020)	74.1/76.5	88.6/95.3	80.0/89.6	-
GASM [21] (2020)	-	84.7/95.3	74.4/88.3	52.5/79.5
Ours(R50)+BDB	**79.6/81.9**	88.1/95.3	77.4/88.7	52.5/79.1
Ours(R50)+MGN	74.7/75.4	**91.0/96.4**	**82.1/91.0**	**65.7/85.5**
MGN(R101)	73.5/74.6	89.0/95.8	80.9/89.8	66.0/85.7
Ours(R101)+MGN	76.9/77.6	**92.0/97.0**	**84.1/91.9**	**68.8/86.6**

How can Contrastive Learning help?

- Experiments
  - Ablation study for different augmentations

Setting	Default	+RE	-GS	-GB	-CJ	-CJ+RE
mAP	73.4	74.2	73.2	73.3	74.0	74.7
$cmc1$	74.0	74.8	73.9	74.1	74.6	75.4

RE: Random Erasing
GS: Grayscale
GB: Gaussian Blur
CJ: Color Jitter

Max area	0.0	0.2	0.4	0.6	0.8
mAP	73.2	74.1	74.4	74.7	73.3
$cmc1$	73.8	74.3	75.3	75.4	73.7

Max erasing area for Random Erasing
0.4: commonly used in supervised re-ID
0.6: best for unsupervised pre-training

How can Contrastive Learning help?

- Summary
  - “Dense Contrastive Pre-training on Large-scale Unlabeled Dataset for Scene Text Recognition”
  - Large-scale Unlabeled Dataset
    - Diverse languages, environments (day/night/indoor/outdoor), …
  - Contrastive Learning Framework
    - Dense contrastive learning framework
    - Random spatial distortion? Random over-explosion?

How can Contrastive Learning help?

- Difference with previous approach
  - Large-scale Unlabeled Dataset

Real unlabeled datasets (Real-U)					
Book32 [14]	arXiv	2016	3.9M	3.7M	(88.9%)
TextVQA [44]	CVPR	2019	551K	463K	
~~ST-VQA [3]~~	~~ICCV~~	~~2019~~	~~79K~~	~~69K~~	
Total	–	–	4.6M	4.2M	

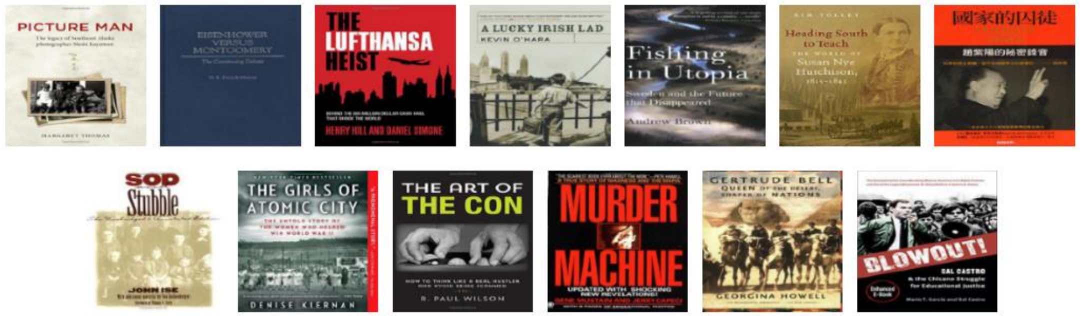


Fig. 2: The "Biographies & Memoirs" book covers that were classified by AlexNet as "History." While misclassified, many of these books also can relate to "History" despite the ground truth.

    - Still has a gap with real "scene text" dataset
  - Contrastive learning framework
    - Neglect spatial information
    - No task-specific augmentation

```
""" for self supervised learning on Feature extractor (CNN part) """
if SelfSL_layer == 'CNN':
    visual_feature = visual_feature.permute(0, 2, 1)          # [b, w, c] -> [b, c, w]
    visual_feature = self.AdaptiveAvgPool_2(visual_feature)   # [b, c, w] -> [b, c, 1]
```

Toward Scene Text Recognition With Fewer Labels, CVPR 2021