It's About Time: Analog Clock Reading in the Wild

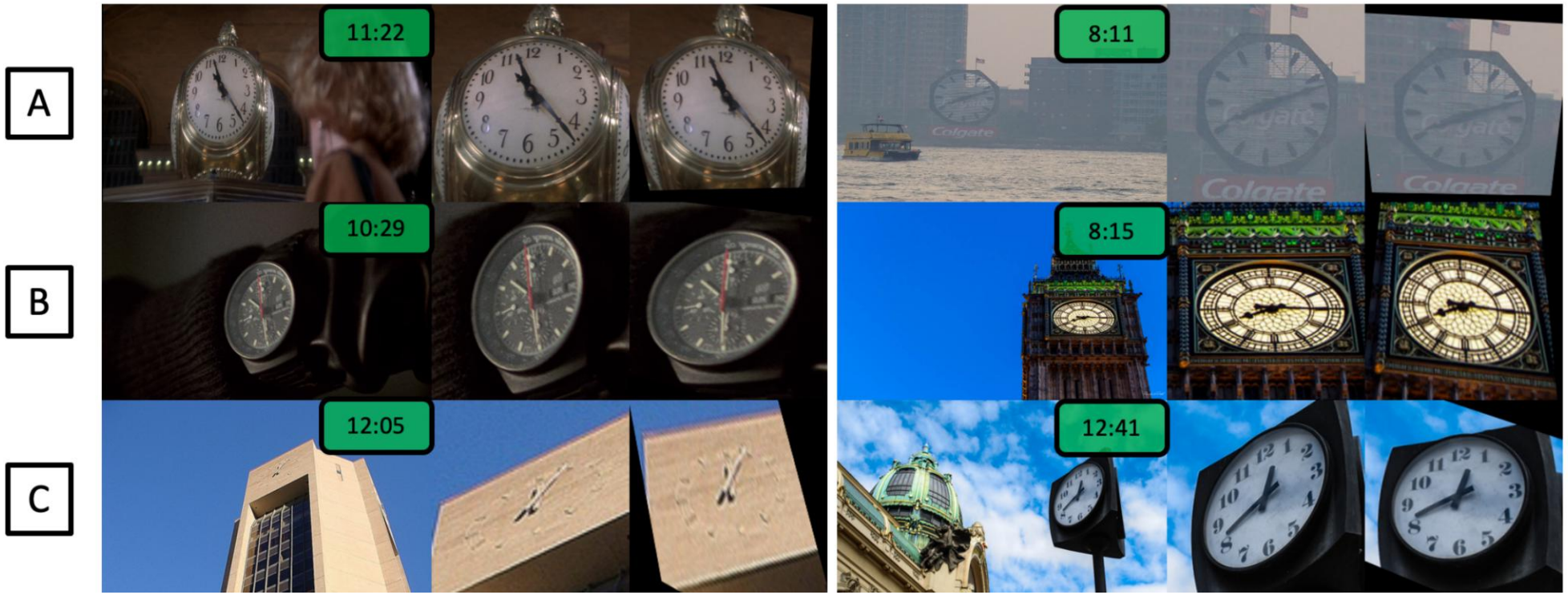
Charig Yang Weidi Xie Andrew Zisserman

VGG, Department of Engineering Science, University of Oxford

{charig,weidi,az}@robots.ox.ac.uk

https://charigyang.github.io/abouttime/

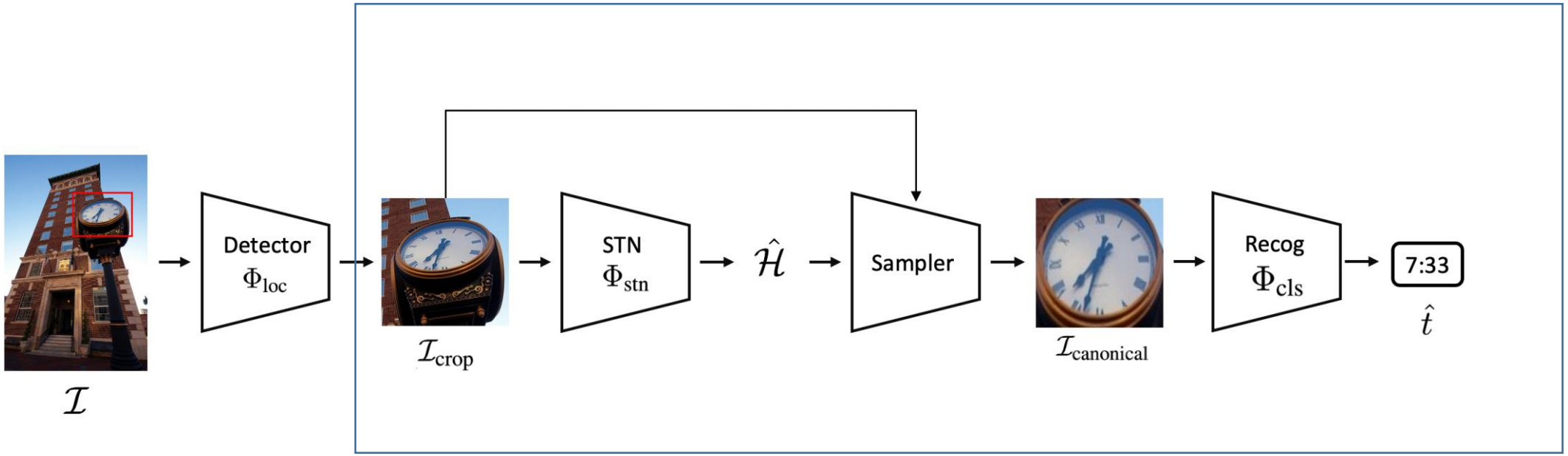
Introdcution



Ideas

- 1. proposing a synthetic generator to generate syn-clocks.
- 2. proposing a framework to solve problem.
- 3. proposing a specific weekly-supervised way.
- 4. releasing a new dataset.

Framework



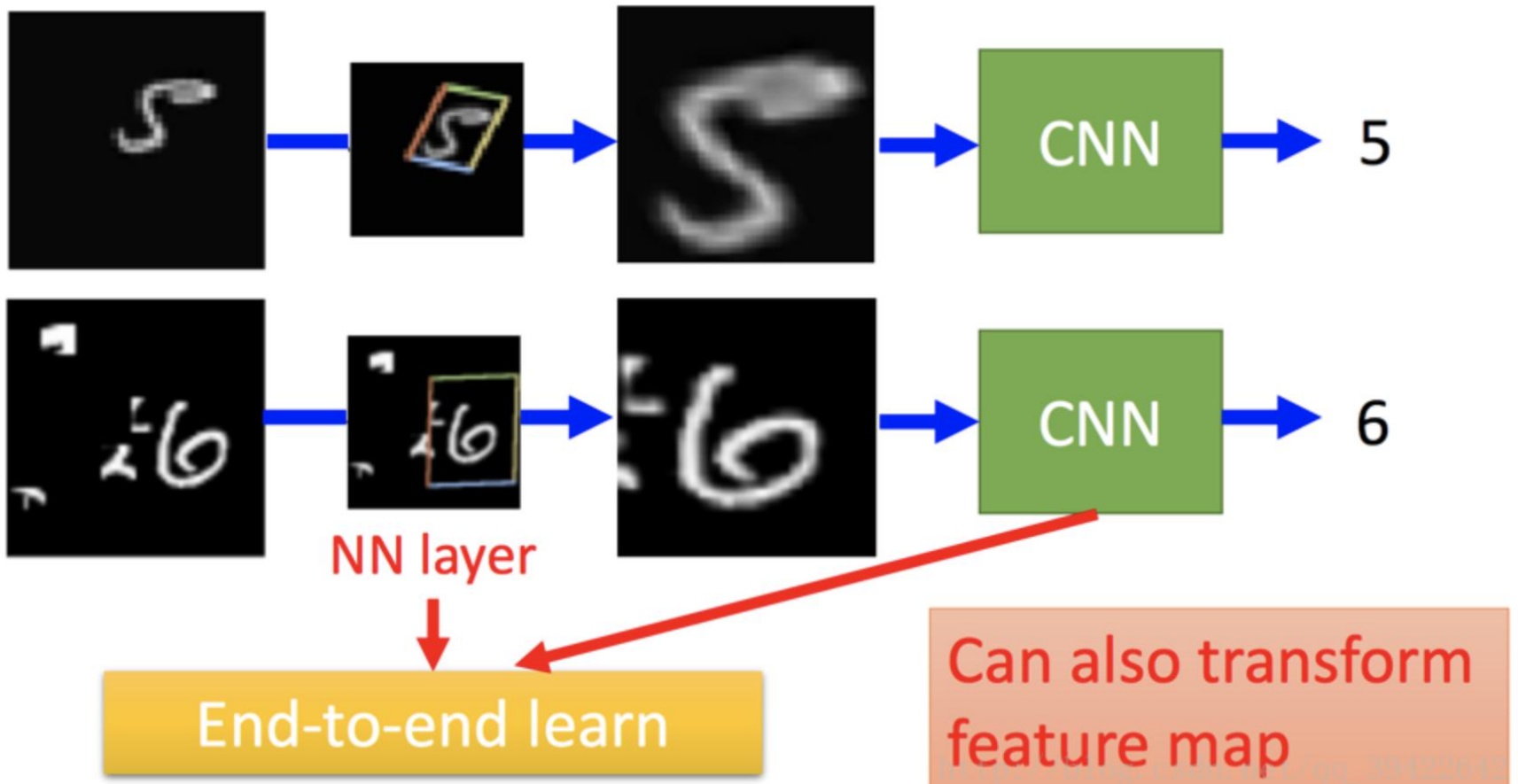
$$\hat{\mathcal{H}} = \Phi_{\text{stn}}(\mathcal{I}_{\text{crop}}) \in \mathcal{R}^{3\times 3}$$

$$\mathcal{I}_{\text{canonical}} = \text{SAMPLER}(\mathcal{I}_{\text{crop}}, \hat{\mathcal{H}}) \in \mathcal{R}^{3\times h\times w}$$

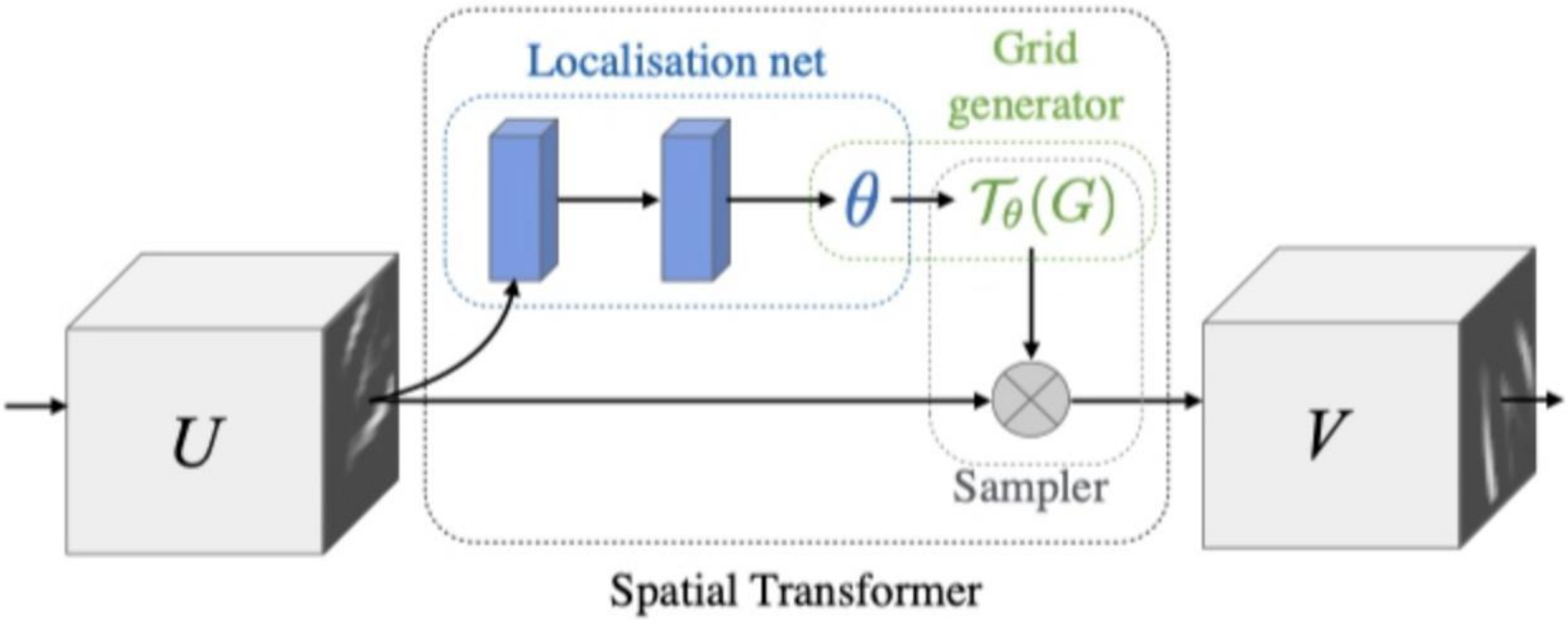
$$\mathcal{L} = \mathcal{L}_{\text{stn}} + \mathcal{L}_{\text{cls}} = \sum |\hat{\mathcal{H}} - \mathcal{H}| + \sum \hat{t}\log(t)$$

$$\hat{t} = \Phi_{\text{cls}}(I_{\text{canonical}}) \in \mathcal{R}^{720}$$

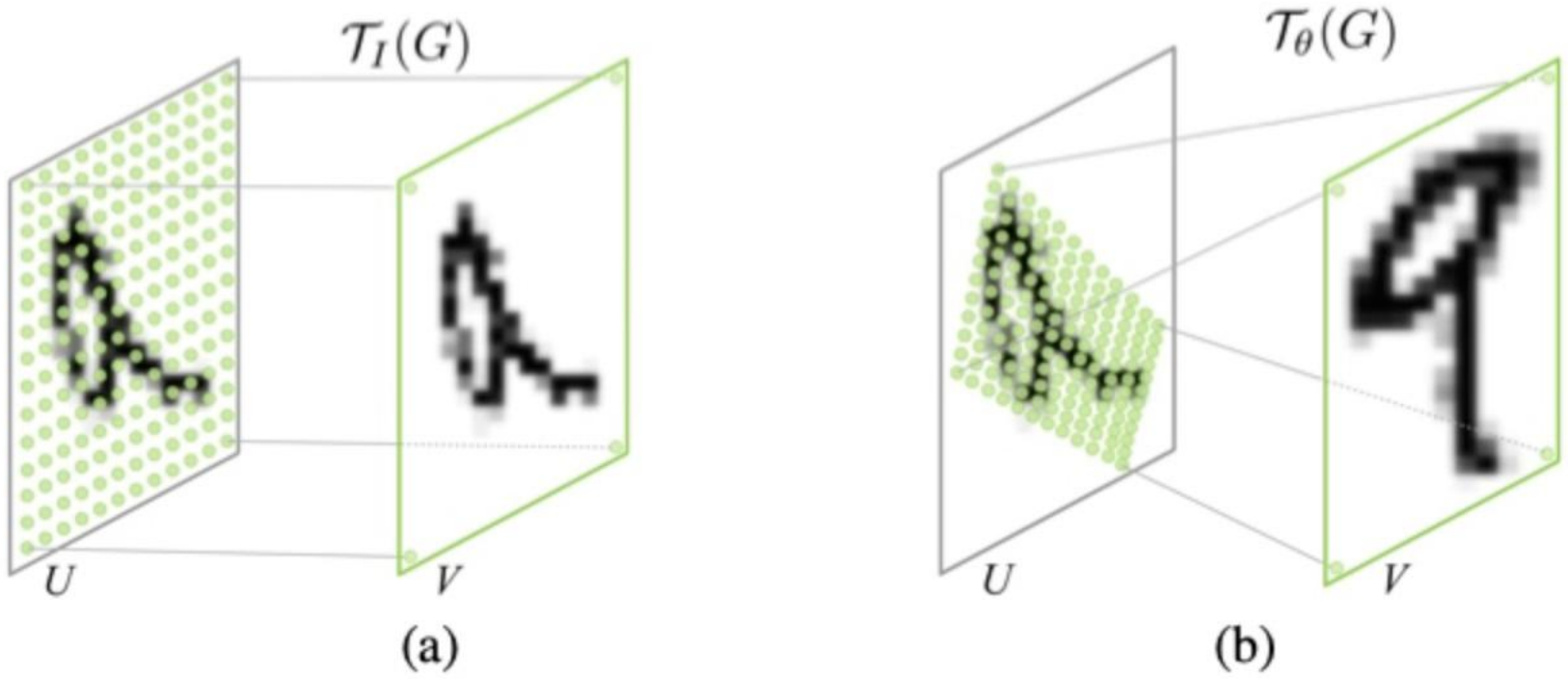
STN(1)


CNN
5
CNN
6
NN layer
End-to-end learn
Can also transform feature map

STN(2)

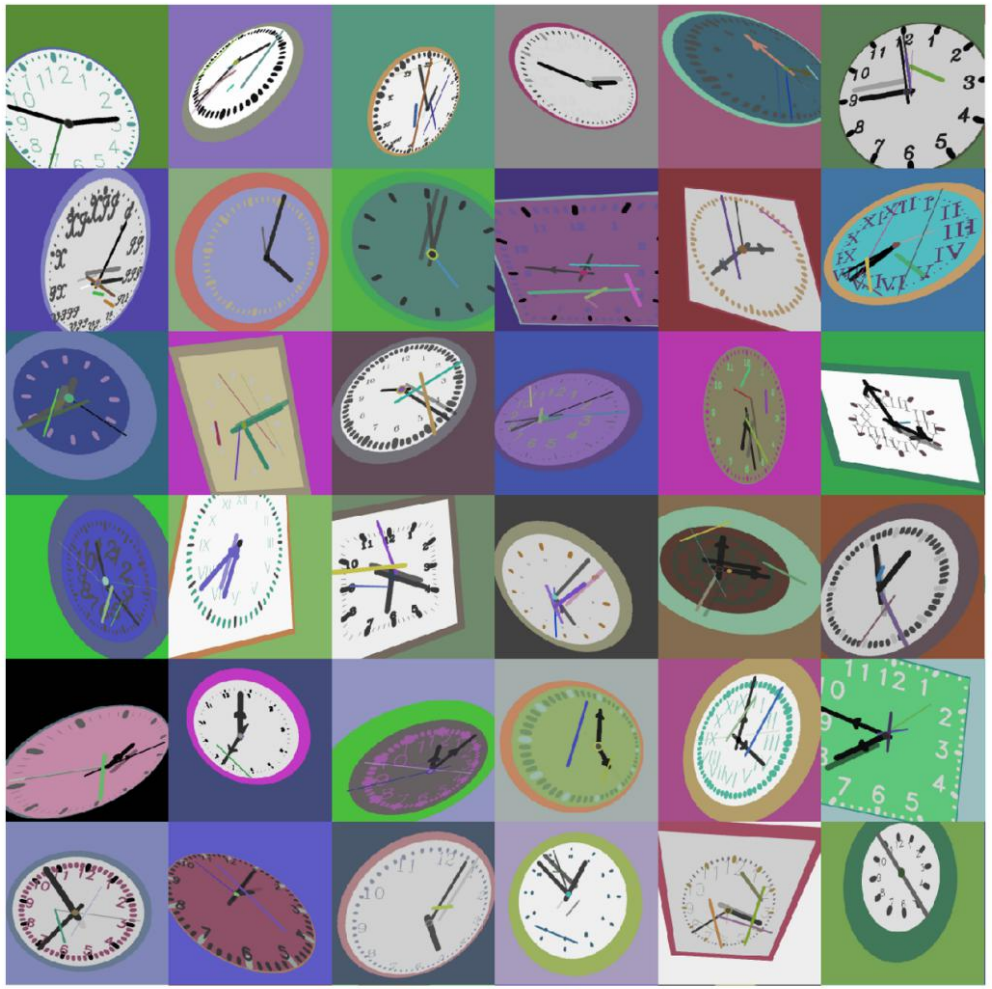

Localisation net
Grid generator
θ
$\mathcal{T}_\theta(G)$
U
V
Sampler
Spatial Transformer

STN(3)

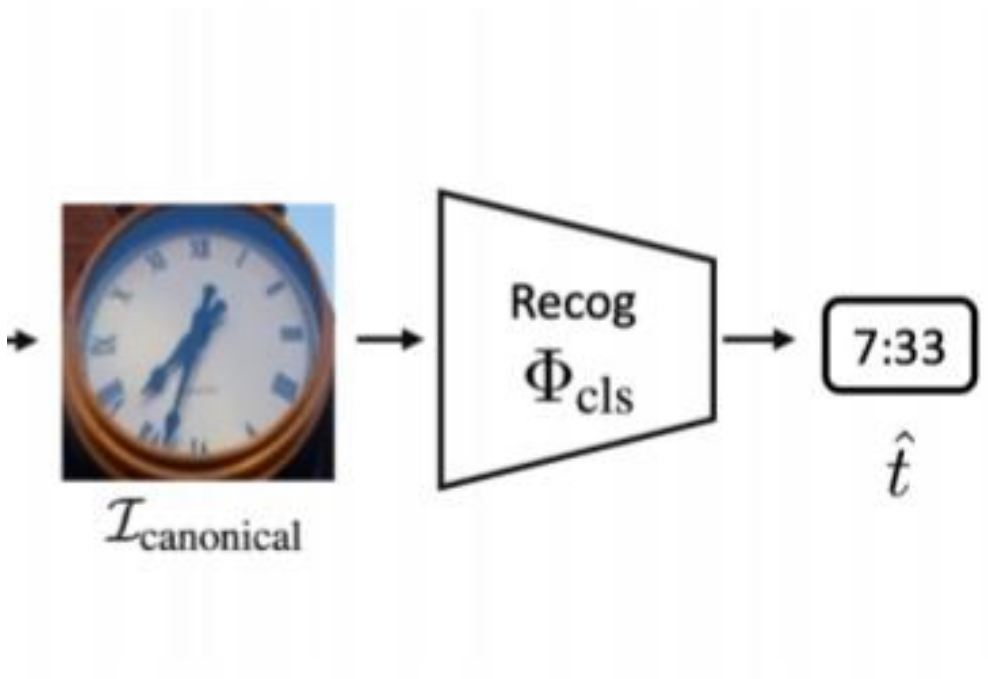


$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \mathcal{T}_\theta(G_i) = \mathtt{A}_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$$

Synthetic clock image generator



specific weekly-supervised way

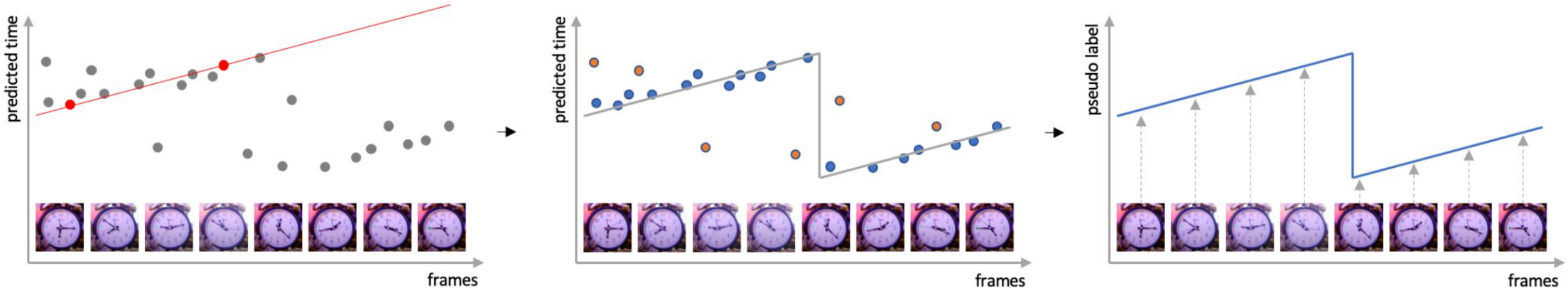


Classification without labels

Training process

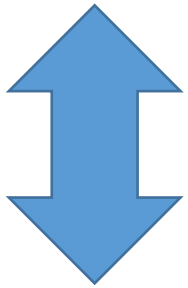
- 1. pre-training in synthetic datasets(with labels)
- 2. use pre-trained model to generate pseudo-labels.
- 3.real-training
- (How to eliminate domain gaps?)

Select pseudo-labels



Iterative Training

use pre-trained model to generate pseudo-labels.



real-training

Iteration	Videos	Frames
Pseudo-label Round 1	1490	1.0M
Pseudo-label Round 2	1881	1.2M
Pseudo-label Round 3	1987	1.3M

My project

CNN

ResNet

Feature map

Predict Pointer/dail

Text detection

Detection Result

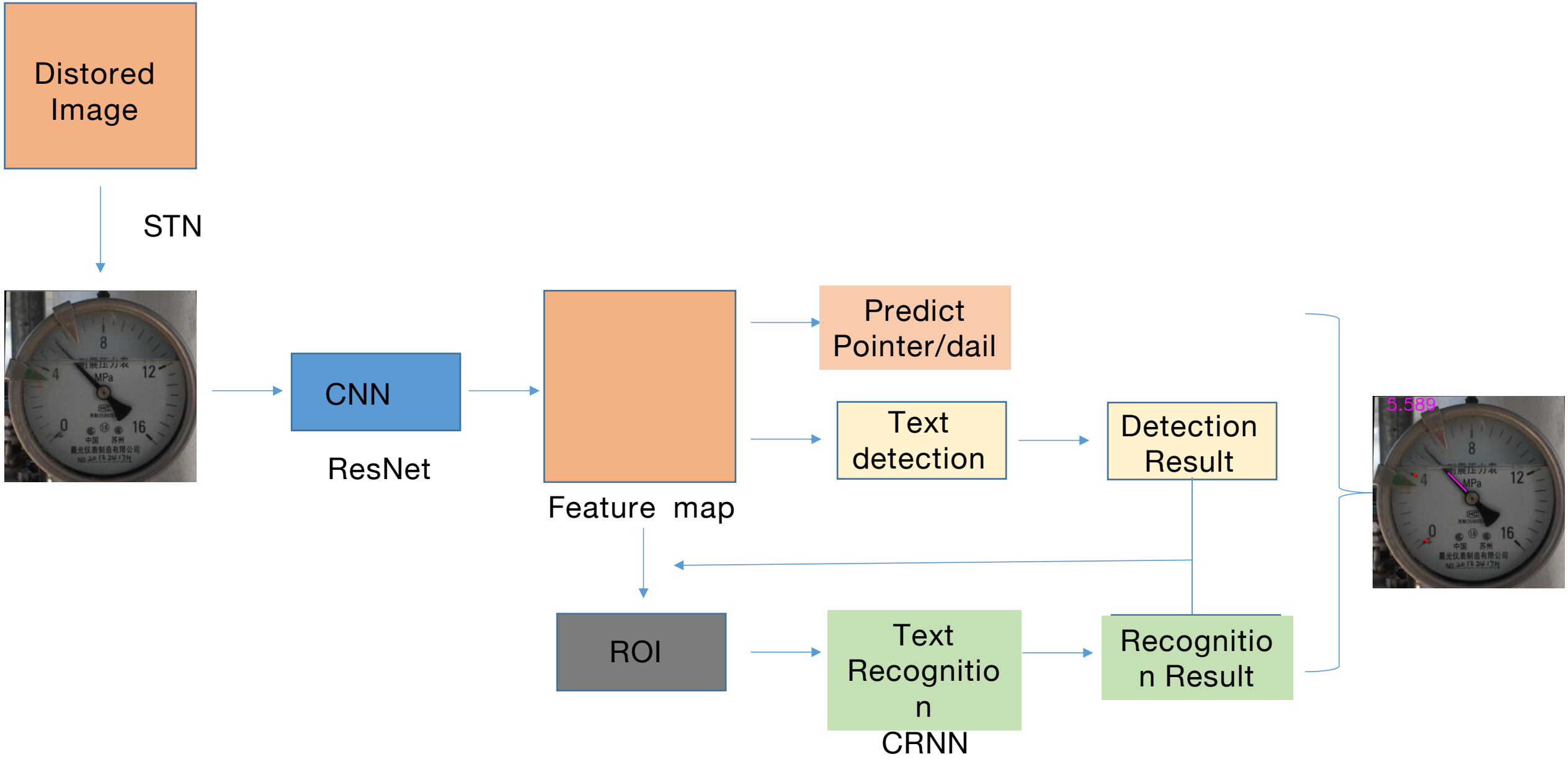
ROI

Text Recognition

CRNN

Recognition Result

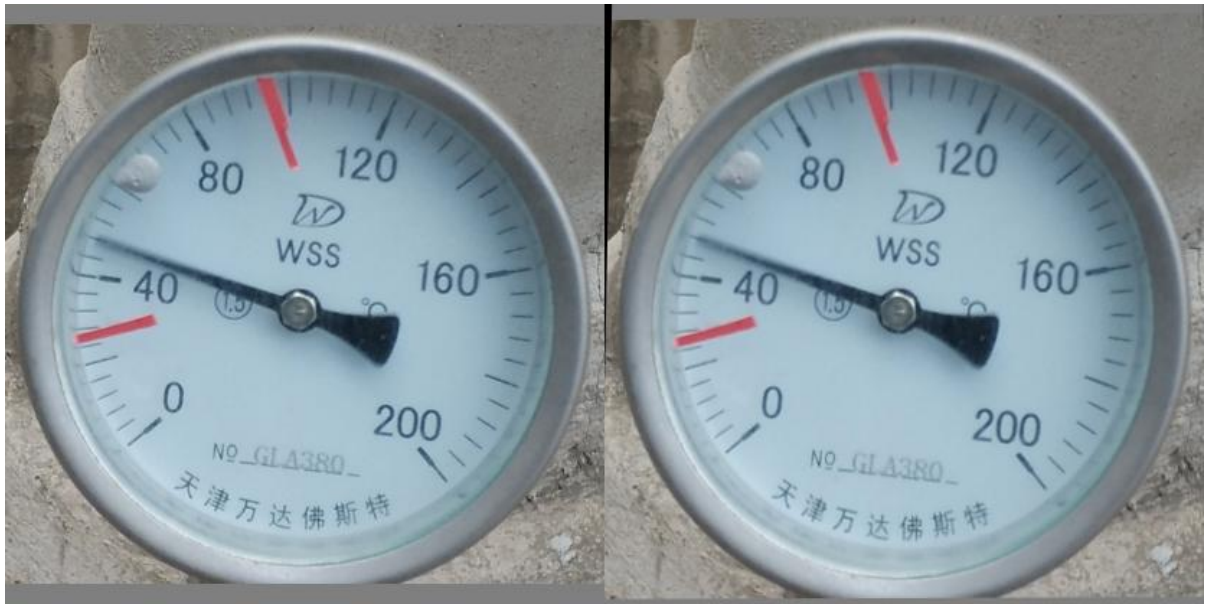
Future work


Distored Image
STN
CNN
ResNet
Feature map
Predict Pointer/dail
Text detection
Detection Result
ROI
Text Recognitio n
CRNN
Recognitio n Result
5.589

Current visulization result

- Number recognition result: 88.6%

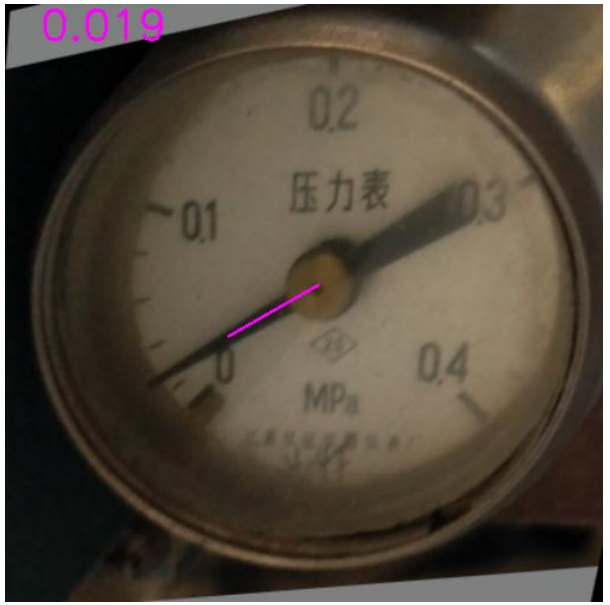
- Mean meter relative error: (pred-label)/label
- <3%


80
120
WSS
40
160
0
200
天津万达佛斯特


48.804
80
120
WSS
40
160
0
200
天津万达佛斯特


0.2
压力表
0.1
0.3
0
MPa
0.4


0.2
压力表
0.1
0.3
0
MPa
0.4


0.019
0.2
压力表
0.1
0.3
0
MPa
0.4


0.005


9.926
WSS


0.4
0.6
不锈钢压力表
MPa
0.2
0.8
1
大庆龙化新实业总公司


0.0
0.4
0.6
不锈钢压力表
MPa
0.2
0.8
1
大庆龙化新实业总公司


0.04
0.12
0
0.16
MPa
△P
MPa
合 格 证

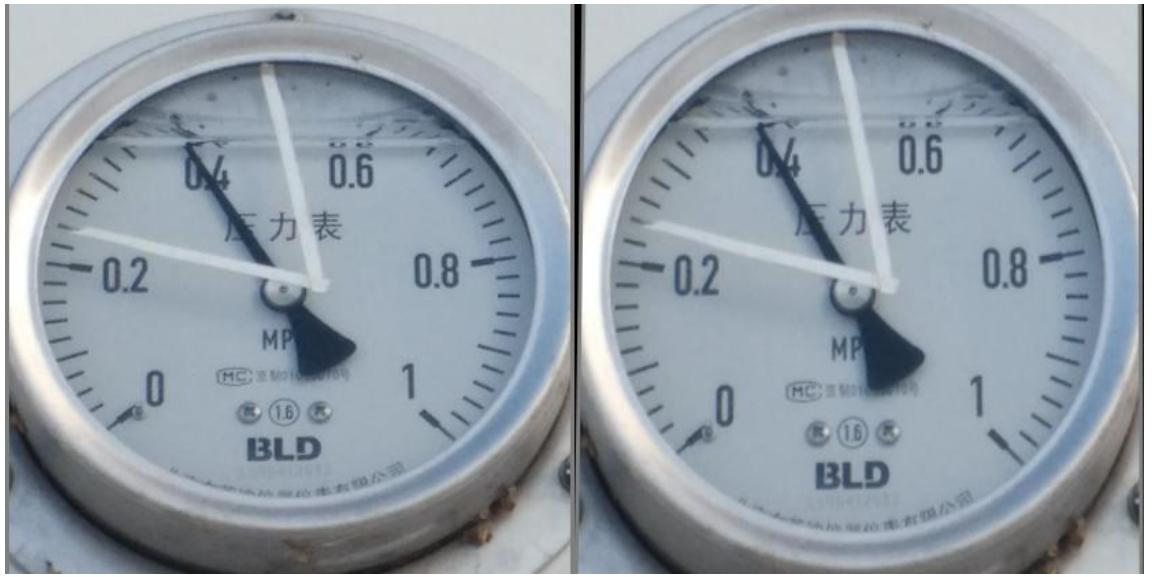

0.04
0.12
0
0.16
MPa
△P
MPa
合 格 证


0.056
0.04
0.12
0
0.16
MPa
△P
MPa
合 格 证


40
20
60
WSS
0
80
100


10.311
40
20
60
WSS
0
80
100


排气压力
供油压力
BLD


0.4
0.6
压力表
0.2
0.8
MPa
0
1
BLD


0.436
0.4
0.6
0.2
0.8
MPa
0
1
BLD

Text-DIAE: Degradation Invariant Autoencoders for Text Recognition and Document Enhancement

Mohamed Ali Souibgui[†1], Sanket Biswas[†1], Andres Mafla [†1], Ali Furkan Biten [†1], Alicia Fornés[1], Yousri Kessentini[2], Josep Lladós[1], Lluis Gomez[1], and Dimosthenis Karatzas[1]

[1] Computer Vision Center, Universitat Autònoma de Barcelona, Barcelona, Spain
{msouibgui,sbiswas,amafla,abiten,afornes,josep,lgomez,dimos}@cvc.uab.es

[2] Digital Research Center of Sfax, SM@RTS Laboratory, Sfax, Tunisia
yousri.kessentini@crns.rnrt.tn

Introduction

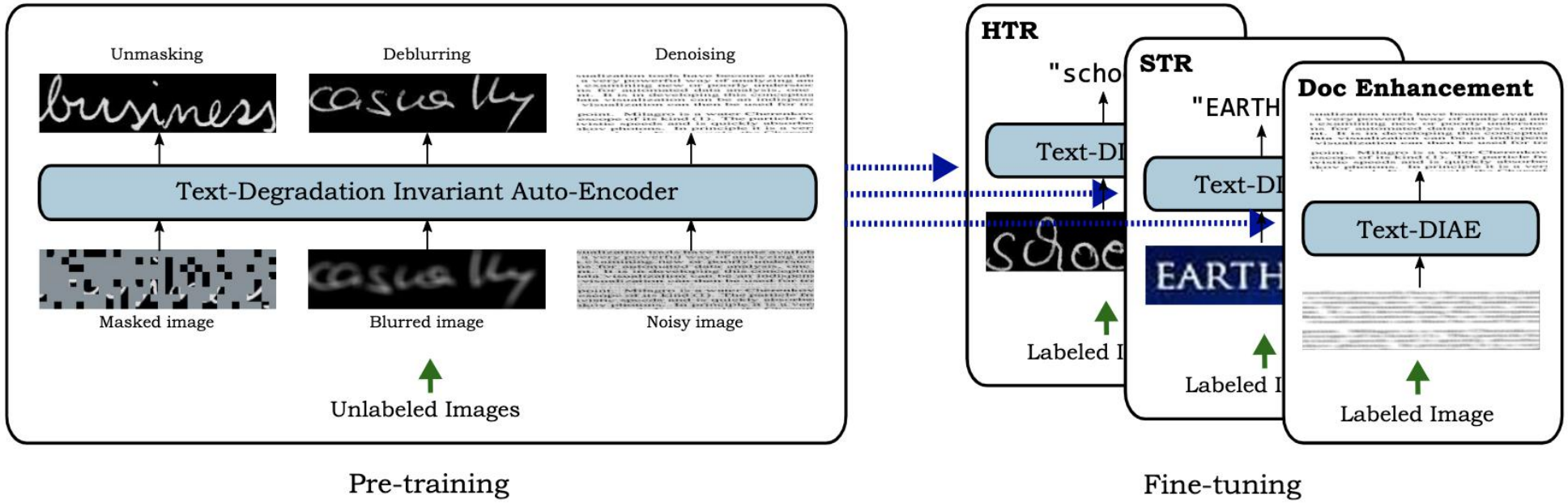


Fig. 1. Text-Degradation Invariant Auto-Encoder(Text-DIAE), we employ image reconstruction pretext tasks at pre-training. Masking, blurring and adding noise is employed to learn richer representations that outperform previous approaches.

Method

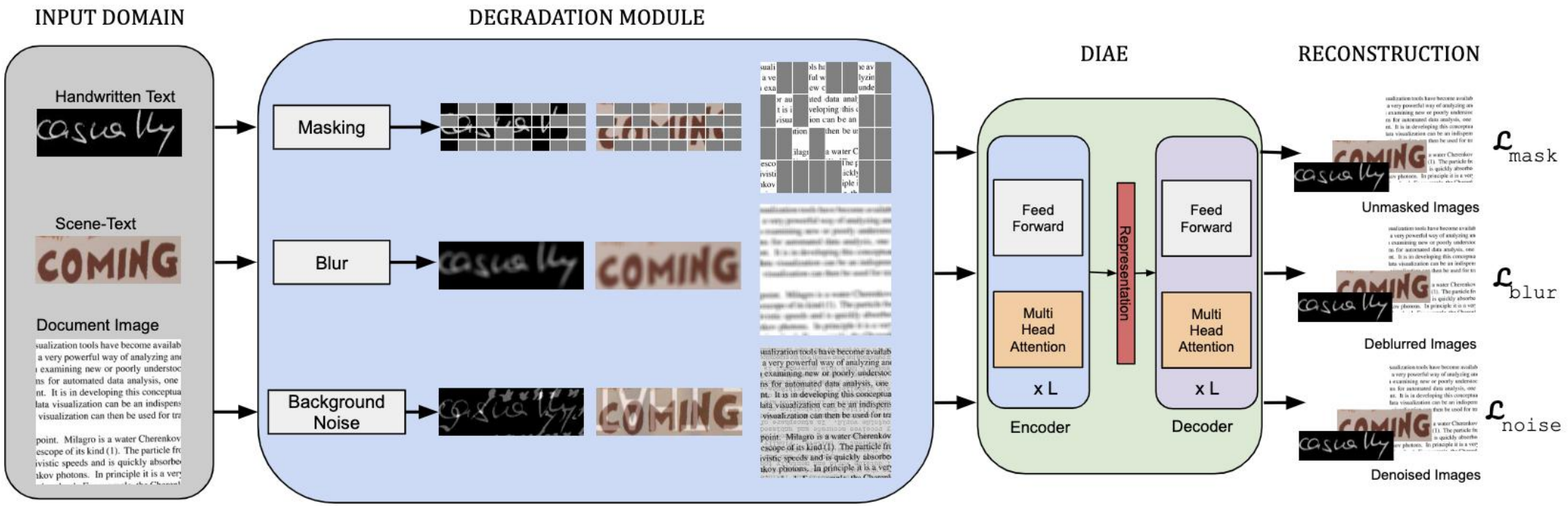


Fig. 2. Pre-training pipeline. Text-DIAE aims to learn degradation invariant representations. These are later used to reconstruct the input image with a specific learning objective for each degradation type.

$$z_{\mathcal{T}} = \mathcal{E}(\phi(I, \mathcal{T}); \theta_E)$$
$$I_r = \mathcal{D}(z_{\mathcal{T}}; \theta_D)$$

$$z_0 = E(I_d^p) + E_{pos}$$
$$z'_l = \mathrm{MSA}(\mathrm{LN}(z_{l-1})) + z_{l-1}, \;\; l = 1, \; \ldots \mathrm{L}$$
$$z_l = \mathrm{MLP}(\mathrm{LN}(z'_l)) + z'_l, \;\; l = 1, \; \ldots \mathrm{L}$$
$$z_{\mathcal{T}} = LN(z_L)$$

$$z'_l = \mathrm{MSA}(\mathrm{LN}(z_{l-1})) + z_{l-1} \;, \;\; l = 1, \; \ldots \mathrm{L}$$
$$z_l = \mathrm{MLP}(\mathrm{LN}(z'_l)) + z'_l \;, \;\; l = 1, \; \ldots \mathrm{L}$$
$$I_r = \mathrm{Linear}(z_L)$$

Fine-tune

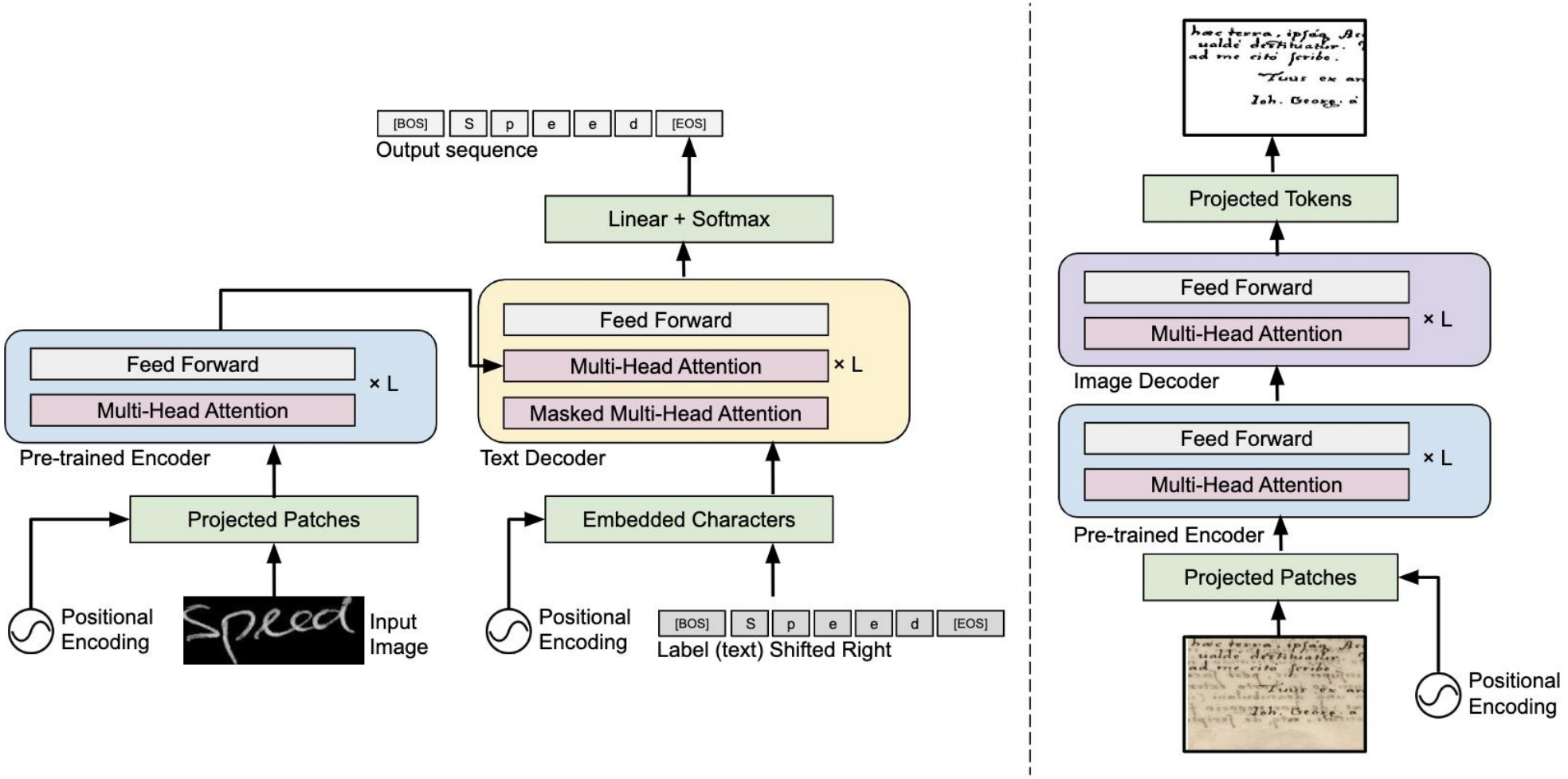

[BOS] S p e e d [EOS]
Output sequence
Linear + Softmax
Feed Forward
Multi-Head Attention
× L
Pre-trained Encoder
Feed Forward
Multi-Head Attention
Masked Multi-Head Attention
× L
Text Decoder
Projected Patches
Embedded Characters
Positional Encoding
Input Image
Positional Encoding
[BOS] S p e e d [EOS]
Label (text) Shifted Right
Projected Tokens
Feed Forward
Multi-Head Attention
× L
Image Decoder
Feed Forward
Multi-Head Attention
× L
Pre-trained Encoder
Projected Patches
Positional Encoding

Table 1. Representation quality. We evaluate the encoder capability of learning visual representations. This scenario is analogous as the linear probing in self-supervised models. We train a decoder with labelled data on top of a frozen encoder pre-trained on the proposed degradation. The column *Seen* refers to the number of samples in millions seen during pre-training. Word prediction in terms of Accuracy (Acc) and single edit distance (ED1) in handwritten and text recognition.

Method	Encoder	Decoder	Handwritten Text						Scene-Text					
			IAM			CVL			IIIT5K			IC13		
			Acc	ED1	Seen	Acc	ED1	Seen	Acc	ED1	Seen	Acc	ED1	Seen
simCLR [14]	CNN	CTC	4.0	16.0	205.8	1.8	11.1	205.8	0.3	3.1	409.6	0.3	5.0	409.6
seqCLR [1]			39.7	63.3	205.8	66.7	77.0	205.8	35.7	62.0	409.6	43.5	67.9	409.6
simCLR [14]	CNN	Attn.	16.0	21.2	205.8	26.7	30.6	205.8	2.4	3.6	409.6	3.1	4.9	409.6
seqCLR [1]			51.9	65.0	205.8	74.5	77.1	205.8	49.2	68.6	409.6	59.3	77.1	409.6
Ours	ViT	Transf.	**71.0**	**82.1**	**4.7**	**78.1**	**81.5**	**1.2**	**77.1**	**87.8**	**9.1**	**92.6**	**95.6**	**18.2**

Table 2. Semi-supervised results. Accuracy obtained by fine-tuning a pre-trained model with varying percentages of the labeled dataset. Under this setting, we back-propagate the gradients through the specific decoder and the pre-trained encoder.

Method	Encoder	Decoder	Handwritten Text						Scene-Text	
			IAM			CVL			IIIT5K	IC13
			5%	10%	100%	5%	10%	100%	100%	100%
Supervised [1]	CNN	CTC	21.4	33.6	75.2	48.7	63.6	75.6	76.1	84.3
simCLR [14]			15.4	21.8	65.0	52.1	62.0	74.1	69.1	79.4
seqCLR [1]			31.2	44.9	76.7	66.0	71.0	77.0	80.9	86.3
Supervised [1]	CNN	Attention	25.7	42.5	77.8	64.0	72.1	77.2	83.8	88.1
simCLR [14]			22.7	32.2	70.7	59.0	65.6	75.7	77.8	84.9
seqCLR [1]			40.3	52.3	79.9	**73.1**	**74.8**	77.8	82.9	87.9
Supervised	ViT	Transformer	22.8	25.3	71.7	17.9	19.8	71.9	75.7	91.9
Ours			**49.6**	**58.7**	**80.0**	47.9	68.5	**87.3**	**86.1**	**92.0**

Table 4. Ablations of the degradations as pre-training objectives. Results in handwritten and scene-text recognition obtained by each pretext task. The performance is measured in terms of Word and Character error rates (WER and CER).

$\mathcal{L}_{mask}$	$\mathcal{L}_{blur}$	$\mathcal{L}_{noise}$	IAM			IC13		
			CER↓	WER↓	Avg.	CER↓	WER↓	Avg.
✓	✗	✗	**9.3**	**20.0**	**14.65**	4.5	8.0	6.25
✓	✓	✗	12.3	24.8	18.5	**4.2**	**8.0**	**6.10**
✓	✗	✓	11.1	23.3	17.2	4.8	8.6	6.70
✓	✓	✓	11.4	23.8	17.6	5.1	9.3	7.20

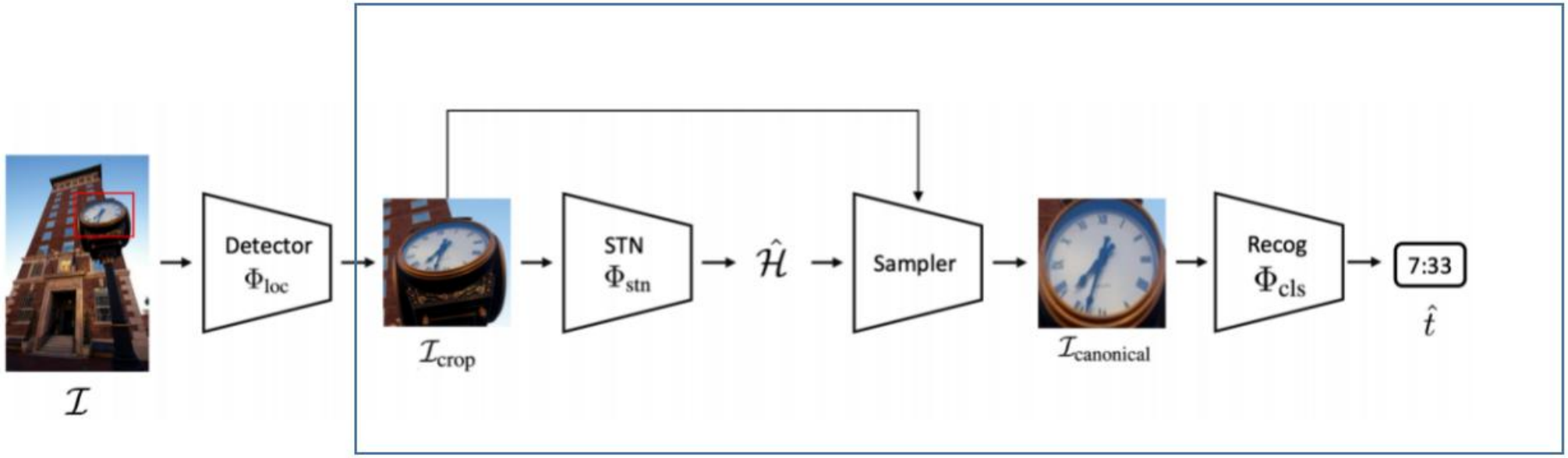
Fig. 4. Qualitative results of pre-training samples. The scenario on the left refers to handwritten text, while scene-text is depicted on the right. On each scenario, from left to right, the original, masked and reconstructed images are depicted.

Improvements

如何学习语义信息？

项目

- 之前两周主要工作：改进仪表纠正模块



Homography

delta

- Train： 已知椭圆仪表区域，然后用圆形拟合得到新的仪表区域，已知了H，就可以把（0,0）,(0,h),(w,0),(w,h)根据H映射到新图，求delta

- Train object： 学习delta

- Infer： 输入图片，输出delta，根据图像四个点求H，然后映射过去得到矫正图片