An End-to-End Approach for Handwriting Recognition: From Handwritten Text Lines to Complete Pages

Supplementary Material

This is supplementary material for the paper "An End-to-End Approach for Handwriting Recognition: From Handwritten Text Lines to Complete Pages". We will further describe hyper-parameter details and additional experimental analysis.

A. Encoder Hyperparameters

Table SI 1 describes the main hyper-parameters of DANCER's encoder module according to each convolutional block (vanilla, octave, or depth-wise separable convolution).

Blocks	Config.	Values
Vanilla	Conv. filters	16 - 16 - 16
	Stride	1x1 - 1x1 - 1x1
Octave	Conv. filters	32 - 32 - 64 - 64 - 128 - 128 - 128 - 128 - 128 - 128
	Octave α	0.500 - 0.500 - 0.375 - 0.375 - 0.250 - 0.250 - 0.125 - 0.125 - 0.125 - 0.00
	Stride	1x1 - 2x2 - 1x1 - 2x2 - 1x1 - 2x2 - 1x1 - 1x1 - 1x1 - 1x1
	Max Pool	0x0 - 0x0 - 0x0 - 0x0 - 0x0 - 0x0 - 0x0 - 2x1 - 0x0 - 2x1
Separable Depth-Wise Block	Conv. filters	256 - 256 - 256
	Gated conv.	No - No - Yes
	Stride	1x1 - 1x1 - 1x1

Table SI 1. Hyper-parameter details of the DANCER's encoder. The encoder of DANCER can be seen as a three-part architecture. The initial section is composed of a standard convolutional block featuring three layers. Next, the middle section comprises 16-octave convolutional layers, integrating max pooling in certain layers. The final segment is characterized by a separable depth-wise convolutional block consisting of three layers, where the last layer uniquely incorporates a gated convolution within its depth-wise convolution operation. The values are ordered according to each respective layer.

B. Data Augmentation

Table SI 2 details the data augmentation techniques utilized during the training of DANCER. The probability of triggering data augmentation is set at 0.9. Each augmentation technique has an individual probability of being applied set at 0.1. Table S.2 comprehensively lists each augmentation method along with its specific parameters. These techniques were systematically applied to enrich the training dataset, improving the model's generalization ability on diverse handwritten documents.

Augmentation	Parameters
Elastic Distortion	$\alpha_{min} = 0.50$ $\alpha_{max} = 1.00$ $\sigma_{min} = 1.00$ $\sigma_{max} = 10.00$ $Min.\ Kernel\ Size = 3$ $Max.\ Kernel\ Size = 9$
Scale	$Min.\ Factor = 0.75$ $Max.\ Factor = 1.00$
Perspective	$Min.\ Factor = 0.00$ $Max.\ Factor = 0.40$
Dilation and Erosion	$Min.\ Kernel\ Size = 1.00$ $Max.\ Kernel\ Size = 3.00$
Color Jittering	$Hue = 0.20$ $Brightness = 0.40$ $Contrast = 0.40$ $Saturation = 0.40$
Gaussian Noise	$Standard\ Deviation = 0.50$
Gaussian Blurr	$Min.\ Kernel\ Size = 3.00$ $Max.\ Kernel\ Size = 5.00$ $\sigma_{min} = 3.00$ $\sigma_{max} = 5.00$
Sharpen	$\sigma_{min} = 0.00$ $\sigma_{max} = 1.00$ $Min.\ Strength = 0.00$ $Max.\ Strength = 1.00$

Table SI 2. List of data augmentation techniques applied during DANCER's training. When applying dilation and erosion augmentations, we randomly select one of the two morphological operations with equal probability.

C. Correlation between mAP_{CER} and CER

The training results of the DANCER model allow us to visualize the natural correlation between the Character Error Rate (CER) and the mean Average Precision CER (mAP_{CER}). According to Figure SI 1, we can notice that the CER has consistently decreased throughout the training, in-

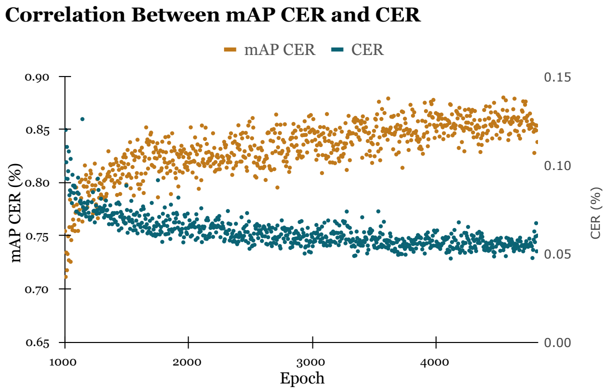

Figure SI 1. mAP$_{\text{CER}}$ and CER computed on READ's single-page validation set as a function of the training epochs.

dicative of the model's increasing proficiency in character recognition. Concurrently, the mAP$_{\text{CER}}$ follows an upward trajectory, suggesting the model's capability to accurately classify and place text regions within the document's layout. This inverse relationship between CER and mapCER as a function of training epochs underscores DANCER's effectiveness in transcription accuracy and spatial understanding of text by DANCER.