Bringing Suzhou Numerals into the Digital Age: A Dataset and Recognition Study on Ancient Chinese Trade Records

Ting-Lin Wu[♣] **Zih-Ching Chen**[♦] **Chen-Yuan Chen**[♥]
Pi-Jhong Chen[♠] **Li-Chiao Wang**[†]
[♣]National Yang Ming Chiao Tung University [♦]NVIDIA AI Technology Center [♥]National Chengchi University
[♠]National Central University [†]Academia Sinica
[♣]morris0401.cs11@nycu.edu.tw [♦]virginiac@nvidia.com [♥]111301049@nccu.edu.tw
[♠]peter20511@gmail.com [†]lcwang@gate.sinica.edu.tw

Abstract

Suzhou numerals, a specialized numerical notation system historically used in Chinese commerce and accounting, played a pivotal role in financial transactions from the Song Dynasty to the early 20th century. Despite their historical significance, they remain largely absent from modern OCR benchmarks, limiting computational access to archival trade documents. This paper presents a curated dataset of 773 expert-annotated Suzhou numeral samples extracted from late Qing-era trade ledgers. We provide a statistical analysis of character distributions, offering insights into their real-world usage in historical bookkeeping. Additionally, we evaluate baseline performance with handwritten text recognition (HTR) model, highlighting the challenges of recognizing low-resource brush-written numerals. By introducing this dataset and initial benchmark results, we aim to facilitate research in historical documentation in ancient Chinese characters, advancing the digitization of early Chinese financial records. The dataset is publicly available at our huggingface hub, and our codebase can be accessed at our github repository.

1 Introduction

Suzhou numerals, a traditional numerical notation system originating in ancient China, played a crucial role in trade, accounting, and daily transactions in East Asia (Yang and Zhang, 2019). Characterized by their unique brush-based calligraphic style and distinct structural patterns, Suzhou numerals differ significantly from modern numerical systems. Despite their historical and cultural importance, digitization and computational analysis remain underdeveloped (Liu et al., 2021), posing challenges to both preservation and automatic recognition.

Digital preservation of Suzhou numeral is imperative due to their profound historical and cultural significance, yet their survival is at risk. Ac-

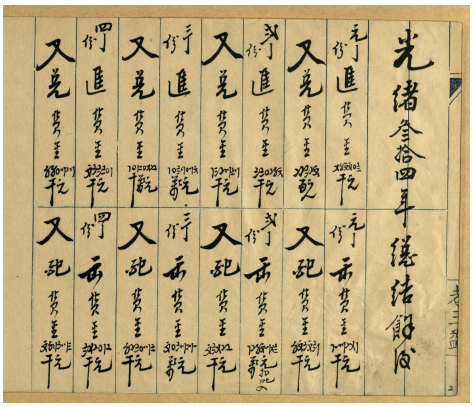

Figure 1: Excerpt from a late Qing-era accounting ledger (dated the 4th year of Emperor Guangxu's reign), preserved in the Hechang Firm in Nagasaki(長崎和昌號) archives.

cording to Jchi (2011) these numerals evolved from Song dynasty arithmetic rod methods and were widely disseminated in Ming China and Japan via educational. Li et al. (2022) further reveals that, before Arabic numerals prevailed, Suzhou numerals were integral to commerce and measurement. As Suzhou numerals fade from use—gradually replaced by Arabic digits—their preservation becomes increasingly urgent to prevent cultural loss and retain a unique part of China's mathematical heritage.

In the past decade, HTR has made significant progress through the use of deep neural networks Doermann and Tombre (2014). Unlike traditional HTR systems that employ hand-crafted features, these networks are data-hungry and require significant amounts of training data to learn, generalize, and be deployed in real-world scenarios. Suzhou numerals are rarely discussed in current studies, which focus primarily on standard Chinese characters or modern digits. This paper aims to address the gap by introducing a dataset of Suzhou numerals for HTR research.

We introduce the first dataset of Suzhou numer-

Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 83–89
May 4, 2025.

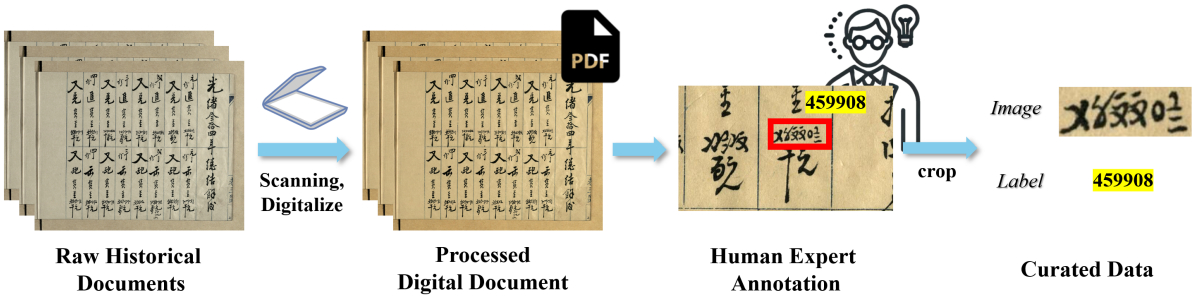

Figure 2: Flow chart of Suzhou Numerals dataset creation

als derived from historical trade records preserved by the Kinmen-based Liang family (1850–1930). These records, originating from the Hechang Firm in Nagasaki, illustrate real-world usage in financial ledgers, contracts, and transactions (Lin, 2020). Fig 1 shows the archievs of Hechang Firm in Nagasaki. Our dataset features:

- High-Quality Suzhou Numeral Samples: 773 manually annotated instances that span various brush styles and levels of degradation.
- Baseline for hand written Suzhou Numeral recognition: CRNN-based baseline model (Shi et al., 2015) for improved recognition of historical scripts .

By bridging cultural heritage preservation and AI-driven OCR, we advance handwritten character recognition for underrepresented scripts, enabling new lines of digital humanities research.

2 Related Work

2.1 Handwritten Text Recognition

HTR has been extensively studied in computer vision and pattern recognition. With the advent of deep learning, convolutional neural networks (CNN) (LeCun et al., 1998) (Krizhevsky et al., 2012) and recurrent neural networks (RNN) (Graves, 2013) have enabled end-to-end learning for HTR, achieving state-of-the-art results on MNIST (Deng, 2012b) and EMNIST (Cohen et al., 2017). Shi et al. (2015) proposed Convolutional Recurrent Neural Network (CRNN) which integrates CNN-based feature extraction with LSTM-based (Graves and Graves, 2012) sequence modeling, delivering powerful, end-to-end recognition performance for text recognition.

Although these approaches have shown high precision for Latin digits and standard Chinese characters, Suzhou numerals pose unique recognition challenges due to their stroke-based morphology, contextual variations, and historical degradation. Unlike modern printed numerals, their handwritten nature introduces stroke ambiguity, where numerals such as 〡 (2) and 〢 (3) differ by a single stroke, making them susceptible to misclassification. Additionally, numerals appear in both horizontal and vertical layouts, requiring flexible layout analysis for proper segmentation. The multiple variants of Suzhou Numerals, along with its mixed use with Chinese numerals and Manchu numbers, make its recognition extremely challenging (Saarela and Xue, 2023).

2.2 Datasets for Handwritten Text Recognition

Large-scale benchmarks have significantly advanced HTR research (Deng, 2012a), yet existing resources primarily target modern digits or standard Chinese text (Cohen et al., 2017; LeCun et al., 1998). Historical East Asian scripts, especially those used in commercial documents, remain notably underrepresented (Zhang et al., 2019). Saeed et al. (2024) introduces the Muharaf dataset which collecting over 1,600 historic handwritten Arabic manuscript images with CRNN-based baseline (Shi et al., 2015). Koch et al. (2023) introduces a tailored end-to-end handwritten text recognition system for Medieval Latin dictionary record cards. Moreover, due to limited archival access and the idiosyncrasies of brush-based writing, most publicly available Chinese OCR corpora do not isolate traditional numeric forms. In response, we introduce a new dataset of 773 annotated Suzhou numerals drawn from late Qing-era trade ledgers. Compared to previous Chinese OCR datasets that focus on general characters, ours specifically highlights the *numeric* brush strokes critical for historical accounting. To our

knowledge, this is the first publicly available corpus dedicated solely to Suzhou numerals, providing a foundation for future research in historical OCR and HTR.

Table 1: Suzhou Numerals and Their Unicode Representations

Arabic	Suzhou numerals	Unicode
0	〇	U+3007
1	〡	U+3021
2	〢	U+3022
3	〣	U+3023
4	〤	U+3024
5	〥	U+3025
6	〦	U+3026
7	〧	U+3027
8	〨	U+3028
9	〩	U+3029

3 Data Collection

3.1 Source Material and Archival Records

Our dataset is derived from the **Hechang Firm in Nagasaki** archive (Hec) (Zhu, 2016) (Ichikawa, 1983) (Xu, 1988), which documents trade activities between China, Japan, and Southeast Asia from 1880 to 1930. The collection includes accounting ledgers, trade contracts, and commercial correspondence, where Suzhou numerals appear in transaction records, itemized cost lists, and within handwritten Chinese text (Fig 1). These materials provide a rich historical context, capturing variations in notation style and document formatting over time.

3.2 Digitization and Annotation

As illustrated in the flow chart in Figure 2. First, all documents have been scanned and digitized into high-resolution PDF files. Then, the portions containing the Suzhou numerals (0-9) in these documents were manually annotated by human experts. Finally, every portion was individually cropped into an image with an annotated label. Ambiguous cases, particularly those affected by fading or overlapping strokes, were cross-verified by multiple annotators for consistency.

3.3 Dataset Statistics

The final dataset comprises **773 annotated instances** of Suzhou numerals. We divide the dataset into training, testing, and evaluation sets with a ratio of 7:1.5:1.5, resulting in 541 samples for training, 116 for testing, and 116 for evaluation. The dataset captures natural variations in stroke thickness, numeral alignment, and stylistic nuances, providing a comprehensive representation of real-world Suzhou numeral usage.

Figure 3 illustrates the frequency distribution of individual digits (0-9) appearing in filename labels, providing insights into numerical biases or inconsistencies within the dataset.

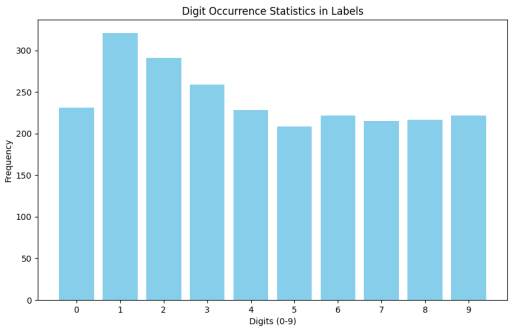

Figure 3: Histogram displaying the frequency of individual digits (0-9) appearing in the filename labels.

4 Experiments and Baseline HTR Results

We evaluate our proposed approach using with CRNN (Shi et al., 2015) to recognize brush-based Suzhou numerals. This section details the CRNN pipeline, training procedures, and the effects of rotation, padding, and pretrained checkpoints, for a baseline of our dataset.

CRNN Pipeline CRNN architecture introduced by Shi et al. (2015) is adapted as a baseline to address the recognition of handwritten Suzhou numeral sequences. The input to our system is a grayscale image $I \in \mathbb{R}^{H \times W \times 1}$ that contains a series of handwritten Suzhou numerals, where $H = 32$ and $W = 128$.

The brief introduction pipeline is as follows. For more details, please refer to Appendix B. First, the grayscale image is fed into a CNN (LeCun et al., 1998) (Krizhevsky et al., 2012) which extracts high-level feature maps. These feature maps are then reshaped into a sequential feature representation. Subsequently, these features are input to bidirectional LSTM (Graves and Graves, 2012) layers. Finally, the sequence features are fed into MLP and the output sequence is decoded into predicted Arabic digit sequence.

Table 2: Baselines of OCR models on Suzhou Numerals recognition. The table presents results for classic OCR and CRNN-based models under different training configs. The lowest Character Error Rate (CER) is achieved with a pretrained CRNN model incorporating both padding and rotation.

Model	Pretrained	Padding	Rotation	CER (%)
Tesseract	Yes	-	-	100.00
Tesseract	Yes	No	No	23.22
CRNN	No	No	No	5.450
CRNN	No	No	Yes	5.205
CRNN	No	Yes	Yes	3.645
CRNN	No	Yes	No	5.115
CRNN	Yes	No	Yes	5.150
CRNN	Yes	Yes	Yes	**3.570**

For the transcription of the sequential output, we adopt the CTC loss (Jaderberg et al., 2014). It proves essential for accommodating irregular spacing and partial strokes.

$$\mathcal{L}_{\text{CTC}} = -\ln p(y \mid \mathrm{x}),$$

which sums over all valid alignments between the input numeral sequence (x) and the ground truth numeral sequence (y).

Training and Data Augmentation Given the script's variability in stroke density and ink clarity, we apply rotations (up to $\pm 20°$), random scaling (5–15%), and brightness alterations ([0.1, 0.2]). When performing data augmentation, we expand the training data by 3 times (1x original training data and 2x augmented data). We train using Adam (LR=0.0001), a batch size of 4, for 100 epochs. More training details can be found in Appendix C).

Rotation and Padding Effects We also investigate how rotation degrees and input padding influence recognition (Table 2). In cases without padding, a moderate rotation (10°) enhances accuracy, but larger angles (20°) start to degrade performance, presumably due to excessive numeral distortion. With padding, the results remain more stable as rotation increases; however, improvements largely plateau beyond 10°. These observations suggest that retaining contextual spacing around numerals helps mitigate augmentation artifacts, especially in heavily degraded scans.

Baseline Comparisons We report the baseline of our Suzhou numerals in different models. As shown in Table 2, Tesseract performs poorly

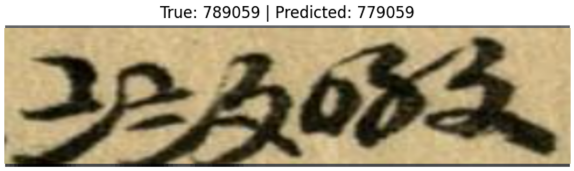

Figure 4: An example of misprediction. The lowest stroke of second character '〢' and the top left stroke of the third character '〤' is almost connected. Therefore, our model identifies these two strokes as a single stroke, and mistakenly recognized the second character as '〢'. See Fig 5 in Appendix for more examples.

(100% CER) on Suzhou numerals, reflecting the script's brush-based style and close integration with Chinese text. Finetuning CRNN attains significantly lower error rates. Once we incorporate padding and rotations, the CER decreases further to 3.645%.

Pretrained Checkpoints and Final Results Leveraging a CRNN checkpoint trained on Synth90k dataset (Jaderberg et al., 2014) yields the best outcome. After fine-tuning on Suzhou numerals, we achieve a CER of 3.57% (Table 2), illustrating that transfer learning is particularly effective in this under-resourced domain. Qualitative inspections reveal that most errors are due to faint strokes, especially confusing 〣 **(2)** with 〣 **(3)** or merging 〤 **(4)** and 〥 **(5)** in severely degraded regions (Figure 4). Despite these issues, the results confirm the viability of specialized neural architectures, even with limited training data, and highlight the importance of careful augmentation strategies when tackling historical scripts.

5 Conclusion and Future Work

We introduce the first dataset of Suzhou numerals, providing a critical resource for historical OCR and HTR. Our CRNN baseline, achieving a CER of 3.57%. Future work includes expanding the dataset with additional scribes and degraded samples, integrating attention-based models like Transformers for improved feature extraction, recognizing Suzhou numerals in multilingual documents, and enhancing reproducibility through code and dataset sharing. By bridging cultural preservation with machine learning, this work establishes a foundation for advancing OCR on underrepresented scripts, inviting further research and applications.

References

《長崎和昌號文書》 (t0856). 中研院臺史所檔案館數位典藏. https://tais.ith.sinica.edu.tw/sinicafrsFront/search/search_detail.jsp?xmlId=0000456122. Accessed: 2025-02-09.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. 2017. Emnist: Extending mnist to handwritten letters. *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926.

Li Deng. 2012a. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

Li Deng. 2012b. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142.

David Doermann and Karl Tombre. 2014. *Handbook of Document Image Processing and Recognition*. Springer.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Alex Graves and Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.

Shinai Ichikawa. 1983. 長崎華商「泰益號」🄵🄵文書簡介 昭和 57~ 58 年度科研成果報告の一部. 東南アジア研究年報, (24/25):71–106.

Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep features for text spotting. In *European Conference on Computer Vision*.

Shigeru Jchi. 2011. Mathematical books of the song, yuan, and ming dynasties and the "oranda fuch": The japanese transmission of suzhou numerals (studies in the history of mathematics). *RIMS Kkyroku (Proceedings of the Research Institute for Mathematical Sciences)*, 1739:128–137.

Philipp Koch, Gilary Vera Nuñez, Esteban Garces Arias, Christian Heumann, Matthias Schöffel, Alexander Häberlin, and Matthias Assenmacher. 2023. A tailored handwritten-text-recognition system for medieval Latin. In *Proceedings of the Ancient Language Processing Workshop*, pages 103–110, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Wenhua Li et al. 2022. The application and cultural connotation of suzhou numerals in hainan during the late qing and republican periods. *Library Journal*, 41(12):112.

Man-houng Lin. 2020. Maritime trade networks in east asia: The kinmen merchant archives. *Journal of Chinese Historical Studies*, 30(2):78–96.

Chang Liu, Xu-Yao Zhang, and Qiu-Feng Wang. 2021. A survey of traditional chinese character recognition methods. *Pattern Recognition*, 112:107750.

Mårten Söderblom Saarela and Zhang Xue. 2023. A study on the quantification and reform of chinese characters in a late qing bannerman manuscript. *Bulletin of the Institute of Modern History, Academia Sinica*, (119):1–37.

Mehreen Saeed, Adrian Chan, Anupam Mijar, joseph Moukarzel, Gerges Habchi, Carlos Younes, amin elias, Chau-Wai Wong, and Akram Khater. 2024. Muharaf: Manuscripts of handwritten arabic dataset for cursive text recognition. In *Advances in Neural Information Processing Systems*, volume 37, pages 58525–58538. Curran Associates, Inc.

Baoguang Shi, Xiang Bai, and Cong Yao. 2015. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *Preprint*, arXiv:1507.05717.

Zifen Xu. 1988. 🄵支簿記法としての🄵門華僑簿記の事例研究-長崎在留の「泰益号」の簿記 (一九〇七-一九三四)-. 🄵🄵史学, 23(3):29–47.

Hongwei Yang and Jianguo Zhang. 2019. Chinese historical numbers: Morphology, evolution and usage. *Journal of Chinese Writing Systems*, 3(1):12–31.

X.Y. Zhang, F. Yin, Y.M. Zhang, C.L. Liu, and Y. Bengio. 2019. A survey of deep learning algorithms for optical character recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Delan Zhu. 2016. 公私領域之間 長崎僑領陳世望 (1901-1940). 國史館館刊, (50):1–46.

Appendix

A Examples of Mispredictions

We list some examples (Figure 5) which our model can't correctly predict the true labels.

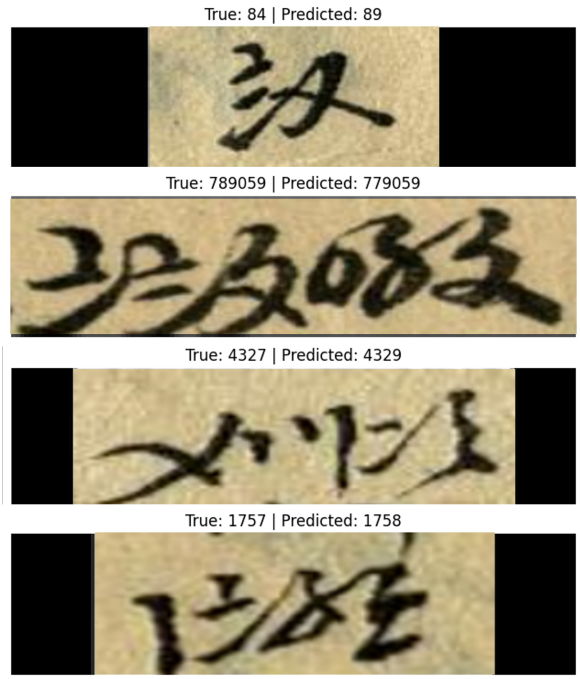

Figure 5: Some examples of mispredictions

B CRNN Model Pipline

We have the same model architecture as Convolutional Recurrent Neural Network (CRNN) (Shi et al., 2015) because we want to obtain a baseline for our Suzhou Numerals recognition task. The following is the detailed model architecture.

- **Input:** Grayscale image of size $32 \times W$ (Height $\times$ Width), where the height is fixed and W is variable. We set $W = 128$ here.
- **Conv1:**
  - Kernel: 3×3, Filters: 64, Stride: 1, Padding: 1.
  - Output: $32 \times W \times 64$.
- **Pool1:**
  - Max Pooling: 2×2 with stride 2.
  - Output: $16 \times \frac{W}{2} \times 64$.
- **Conv2:**
  - Kernel: 3×3, Filters: 128, Stride: 1, Padding: 1.
  - Output: $16 \times \frac{W}{2} \times 128$.
- **Pool2:**
  - Max Pooling: 2×2 with stride 2.
  - Output: $8 \times \frac{W}{4} \times 128$.
- **Conv3:**
  - Kernel: 3×3, Filters: 256, Stride: 1, Padding: 1.
  - Output: $8 \times \frac{W}{4} \times 256$.
- **Conv4:**
  - Kernel: 3×3, Filters: 256, Stride: 1, Padding: 1.
  - Output: $8 \times \frac{W}{4} \times 256$.
- **Pool3:**
  - Max Pooling with kernel 2×1 (vertical pooling only).
  - Output: $4 \times \frac{W}{4} \times 256$.
- **Conv5:**
  - Kernel: 3×3, Filters: 512, Stride: 1, Padding: 1.
  - Output: $4 \times \frac{W}{4} \times 512$.
- **Conv6:**
  - Kernel: 3×3, Filters: 512, Stride: 1, Padding: 1.
  - Output: $4 \times \frac{W}{4} \times 512$.
- **Pool4:**
  - Max Pooling with kernel 2×1 (horizontal pooling only).
  - Output: $2 \times \frac{W}{4} \times 512$.
- **Conv7:**
  - Kernel: 2×2, Filters: 512, Stride: 1, No padding.
  - Output: $1 \times \frac{W}{4} \times 512$.

The final feature map is reshaped into a sequence:

$$\mathrm{x} = \{x_1, x_2, \ldots, x_T\}, \quad T = \frac{W}{4}, \quad x_i \in \mathbb{R}^{512}.$$

RNN Sequence Modeling

The sequential features are modeled by two layers of Bidirectional Long Short-Term Memory (BiLSTM):

- **BiLSTM Layer 1:**
  - Hidden Units: 256 in each direction.
  - Output per time step: 512-dimensional feature vector.
- **BiLSTM Layer 2:**
  - Hidden Units: 256 in each direction.
  - Output per time step: 512-dimensional feature vector.

Transcription Layer and CTC Loss

A fully connected layer with softmax activation is applied to the output of the final BiLSTM to obtain a probability distribution over the target character set augmented by a blank label for CTC. Formally, for each time step:

$$\text{Output dimension} = |\mathcal{A}| + 1,$$

where $\mathcal{A}$ denotes the set of target characters. In our case, $\mathcal{A}$ means a set of numerals from 0-9.

The network is trained using the Connectionist Temporal Classification (CTC) loss:

$$\mathcal{L}_{\text{CTC}} = -\ln p(y \mid \mathrm{x}),$$

which sums over all valid alignments between the input sequence (x) and the ground truth sequence (y).

C Training Hyperparameters

Table 3: Hyperparameters

Parameter	Value
img_height, img_width	32, 128
epochs	100
batch size	4
learning rate	1×10^{-4}
augmentation ratio	2x
rotation degree	20
brightness	0.1

D Example of Training and Evaluation loss

Figure 6 shows an basic example of training and evaluation loss graph. Hyperparameters are set same as Appendix C

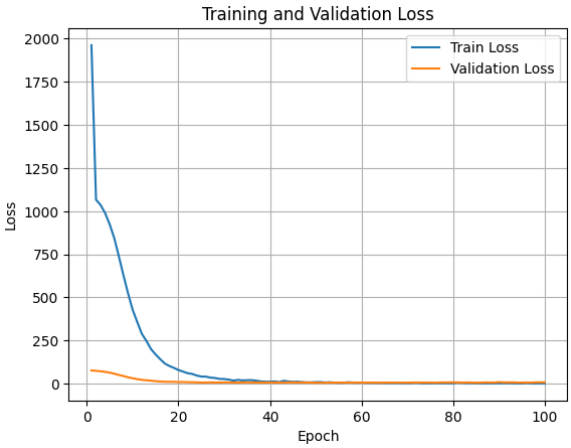

Figure 6: Basic training and evaluation loss graph