Supplementary Material: Towards Ship License Plate Recognition in the Wild: A Large Benchmark and Strong Baseline

Baolong Liu[1,2,3], **Ruiqing Yang**[1], **Roukai Huang**[1], **Wenhao Xu**[1], **Xin Pan**[4]
Chuanhuang Li[1], **Bin Wang**[5], **Xun Wang**[1,3], **Jianfeng Dong**[1,3*]

[1]Zhejiang Gongshang University, [2]Key Laboratory of Public Security Information Application Based on Big-Data Architecture, Ministry of Public Security, [3]Zhejiang Key Laboratory of Big Data and Future E-Commerce Technology, [4]Zhejiang University and [5]Zhejiang Key Laboratory of Artificial Intelligence of Things (AIoT) Network and Data Security

Due to space limitations, we further report and discuss the annotation and diversity of the dataset, additional experimental results, and implementation details of the baseline model in the supplementary material:

- Further analysis of data annotation and diversity.
- More ablation and comparative experiments.
- Details on the implementation of the baseline.

Further Analysis of Data Annotation and Diversity

The annotation process of challenging samples. Due to the complex character layout of SLPs and the harsh natural conditions in outdoor environments, SLPs often face challenges such as character occlusion and blurring, leading to low-quality imaging. However, as shown in Figure 1, each ship is expected to have its SLP written in multiple locations on the ship's body, such as the bow, stern, and hull, and the characters on these plates are consistent. Therefore, we utilize multiple SLPs for mutual verification to infer annotations for those plates with unreadable individual characters. As shown in Figure 2, for SLPs that are difficult to read and have complex character layouts, we will select other SLPs from the same ship that are easier to read as reference SLPs to help us complete the annotation of such samples. Moreover, some ships have SLPs printed simultaneously in Chinese and English, with the English SLP being the Pinyin of the Chinese SLP. As a result, we can accurately infer the English SLP based on the Chinese SLP, even if some characters of the English SLP are unclear. Lastly, many SLPs exhibit layout disorder during the printing process. We refer to relevant regulations and uniformly annotated SLPs following the sequence of Chinese characters, numbers, and English letters.

Additionally, to ensure high-quality SLP annotation, we develop a data annotation platform. Figure 3 shows partial screenshots of our data annotation crowdsourcing platform. We establish a streamlined SLP data production and management platform that allows for SLP data distribution, data annotation, status viewing, quality inspection, and correction of incorrect annotations. In addition, our system can perform statistics and metrics viewing on the currently annotated data. Our system can track the number of SLPs from different locations and different ships, as well as the annotation volume and qualification rate for each annotator. These statistical metrics help us monitor the status and progress of SLP data annotation in real time and make necessary adjustments as quickly as possible to ensure the final creation of a representative and high-quality SLP dataset. Based on this platform, as shown in Figure 4, we also perform multi-attribute labeling for each SLP.

Figure 1: The diagram of SLP annotation pipeline

Although the regions of SLPs are small in the whole image (as shown in Figure 1), the image resolution is not low, as our images are collected using high-resolution surveillance cameras with a resolution of 3392×2008. The average resolution in our SLP34K dataset is 215×72, which is roughly equivalent to some public scene text recognition (STR) datasets (such as 218×84 in MLT19) or higher than them (such as 109×35 in LSVT).

Further exploration of the diversity of SLP34K. Our intelligent waterway monitoring system has installed cameras in eight different configurations across three rivers to capture images of ships. As a result, we collect images from a span of 42 months over the past 5 years (As illustrated in Figure 5). This ensures that our dataset includes a diverse and representative range of SLPs. We also analyze the distribution of samples with different attribute values in the training and testing sets. As shown in Figure 7a, it is evident

*Corresponding author, dongjf24@gmail.com

(a) We select clear and readable SLPs from the same boat to help complete the reasoning and annotation of hard SLPs.

(b) We select SLPs with regular character distribution to assist in annotating the SLPs with complex character layout.

Figure 2: The annotation process of (a) SLPs that are difficult to read and (b) SLPs with complex character layout.

that there are sufficient samples with five different attribute values—single-line, multi-line, vertical, easy, and hard—in both the training and testing sets. This demonstrates the rationale behind the partitioning strategy for the SLP34K dataset. In addition, we conduct statistics on the top 30 characters (Figure 7b), the aspect ratio distribution of SLPs (Figure 7c), and the distribution of SLP character lengths (Figure 7d). These statistics reveal the diversity of samples in terms of character variety, character length, and aspect ratio within the SLP34K dataset. Moreover, from Figure 8, we can find that, apart from the manually labeled samples with the five attribute values, we discover that the dataset also contains some distinct samples, such as those with low lighting conditions, characters written in reverse order, and characters segmented into sections.

More Ablation and Comparative Experiments

Ablation experiments on baseline method using different semantic enhancement approaches. There are some other works proposing multimodal learning approaches based on visual-language integration for text recognition, e.g., ABINet. As shown in Figure 6a, existing visual-language multimodal text recognition methods often involve using separately trained language models to extract semantic features of the recognized text after visual features encoder and text recognition decoder. Since SLPs suffer from severe visual degradation, we believe a robust visual encoder is necessary. As shown in Figure 6b, architecturally, unlike typical methods, our visual-language multimodal learning is performed at the encoder level rather than after the decoder. Moreover, our semantic enhancement module is only used during the training phase and not required during inference, thus avoiding additional computational overhead. Finally, as shown in Table 1, based on our baseline method, we conduct ablation experiments on the proposed semantic enhancement module and the iterative correction using a language model. The experimental results demonstrate that with the inclusion of the semantic enhancement module, our baseline achieves an accuracy of 83.53% on the SLP34K dataset, surpassing the 81.80% obtained without using the semantic enhancement module. However, neither the model parameters nor the FLOPs have increased. In contrast, using only ABINet's language model fusion (LM Fusion) semantic enhancement method also improves accuracy but introduces additional computational overhead. The highest recognition accuracy is achieved when both semantic enhancement modules are used simultaneously. Overall, our proposed semantic enhancement module enhances SLP recognition accuracy without increasing inference computational overhead, presenting a more balanced and efficient semantic enhancement approach.

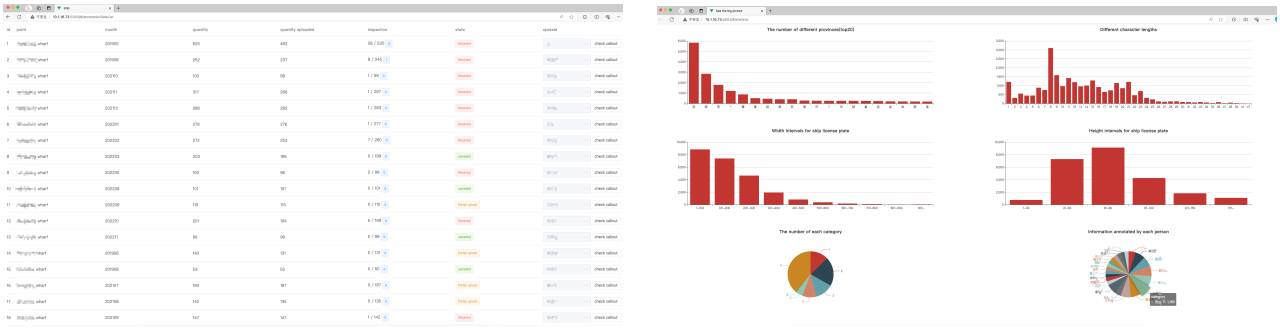
Ours	LM Fusion	Accuracy	Parameters	FLOPs
-	-	81.80	**115M**	**0.45G**
✓	-	83.53	**115M**	**0.45G**
-	✓	83.79	146M	4.46G
✓	✓	**83.96**	146M	4.46G

Table 1: Performance comparison of our baseline with different semantic enhancement modules.

Comparison with current multimodal large models. We evaluate the multimodal large models GPT-4o and Qwen-VL-Plus on our SLP34K dataset. Specifically, we input the SLP image along with the prompt of "You are now a high-performance text recognition model. I will input some images containing Chinese, English, and numeric text. Please accurately recognize and extract the text in a strict left-to-right, top-to-bottom reading order. Then, return the recognized text to me in the following sequence: Chinese characters, Arabic numerals, and English letters. Ensure that the returned text is arranged according to this specified order and that the information is complete and accurate." to the large model, instructing it to identify characters on the given image. The GPT-4o and Qwen-VL-Plus achieve accuracy of 17.05% and 6.43% on SLP34K, respectively. We attribute their poor performance to the limited exposure of these models to SLP data during training, which makes it difficult to achieve high recognition accuracy in the zero-shot testing.

Analysis of vertical text recognition performance. We notice that our method performs better on vertical text compared to horizontal text in the SLP34K dataset, and we conduct an analysis of this phenomenon. Based on our statistical analysis, we find that our dataset includes 122 unique characters for vertical text and 458 unique characters for horizontal text. The smaller number of unique characters in vertical text reduces the recognition challenge, which enables our method to achieve better performance on vertical text. As shown in Table 5 of the main paper, other methods also perform better on vertical text than on horizontal text, which is consistent with our baseline.

Qualitative analysis and failures. Figure 9 demonstrates some qualitative comparison results of challenging examples suffering from blur, low lighting, and so on, our baseline demonstrates superior recognition performance in many extreme cases. We attribute it to the fact that our model



(a) Our streamlined SLP data production and management platform

(b) Statistics and analysis of the current annotation status

Figure 3: The snapshot of our established annotation platform. It allows for (a) SLP data distribution, data annotation, status viewing, quality inspection, and correction of incorrect annotations. At the same time, (b) our annotation platform can also provide real-time statistics and analysis of the current annotation status.

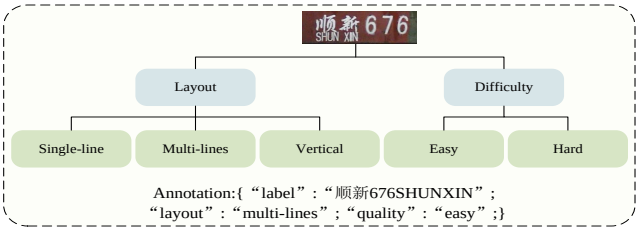


Figure 4: Multi-level multi-attribute SLP labeling.

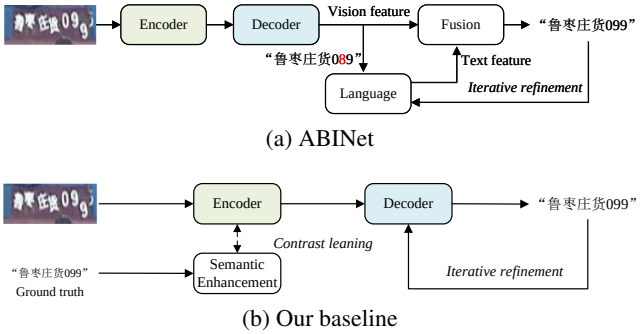
Camera ID	Proportion	Capturing years				
		2018	2019	2020	2021	2022
Camera 1	9.9%		←		→	
Camera 2	18.0%		←		→	
Camera 3	8.8%					←
Camera 4	15.0%		←		→	
Camera 5	9.0%		←		→	
Camera 6	12.5%		←		→	
Camera 7	16.2%	←			→	
Camera 8	10.5%	←		→		

Figure 5: Image proportion and capturing period of each camera

employs pre-training by MAE and semantic enhancement, while the compared methods are trained from scratch without semantic enhancement. Nonetheless, our method also encounters some cases of recognition failure. We analyze the bad samples and find that our method struggles to handle cases with severe visual degradation, such as low-light SLP images and heavily occluded characters (as shown in Figure 9).

Implementation Details

Our baseline model's encoder utilizes ViT as the backbone network, consisting of 12 Transformer layers, with each Transformer layer having 12 attention heads. During



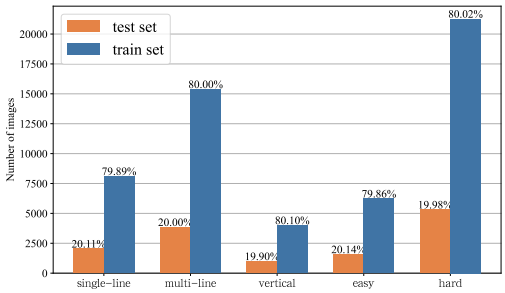
(a) ABINet

(b) Our baseline

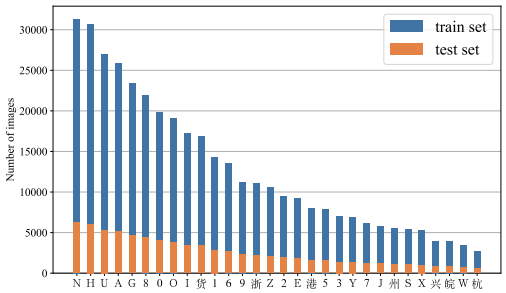
Figure 6: Different from the current representative (a) ABINet method, which add an additional language model on the decoder side for recognition result correction, (b) our proposed baseline method only performs semantic enhancement on the encoder side during the training phase.

the pre-training phase, the pixel reconstruction decoder also adopts the ViT network, comprising 8 Transformer layers, with each Transformer having 16 attention heads. The reconstruction loss is computed using the mean squared error loss function to measure the reconstruction error between the reconstructed pixels and the original pixels. Pre-training is conducted on the SLP34K dataset, we perform horizontal flipping and cropping for data augmentation. After completing the pre-training phase, semantic enhancement fine-tuning training is conducted based on the optimized encoder. During fine-tuning, the dimension of the global visual features outputted by the encoder is projected from 768 to 512 to match the dimension of the text features outputted by the CLIP text encoder, thus enabling contrastive learning. Besides visual-language contrastive learning. The SLP recognition decoder consists of 3 Transformer layers. Fine-tuning training incorporates random rotation, cropping, motion blur, and noise addition for data augmentation.

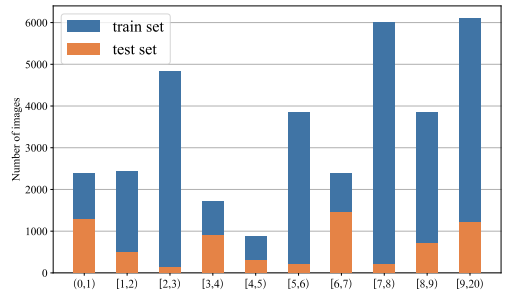
For the input processing, we resize the image size to 224×224, and split an image into 16×16 patches. The max-



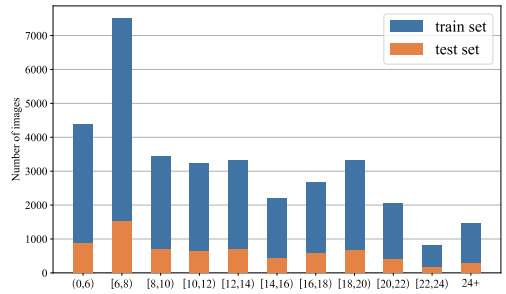
(a) Proportion of different attribute samples in training and test sets



(b) The histogram of top-30 characters with the highest occurrence counts



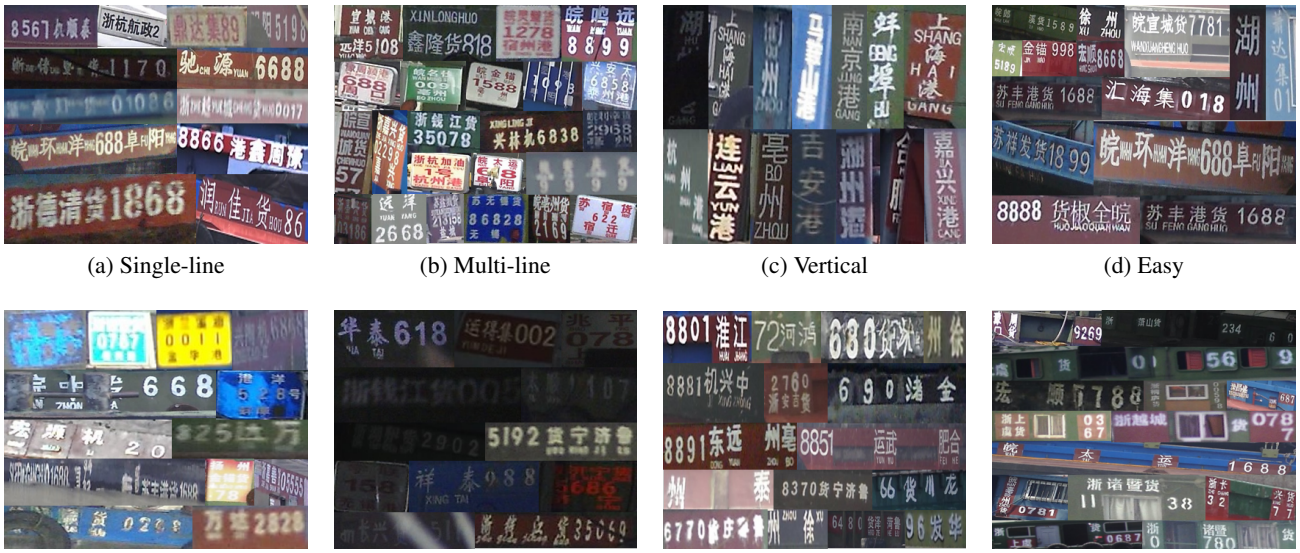
(c) Aspect ratio distribution statistics of samples in the SLP34K dataset



(d) Statistics of sample with different character lengths

Figure 7: Statistics on (a) the distribution ratios of samples with different attributes in the training and testing sets, (b) the top 30 most frequent characters, (c) sample aspect ratios, and (d) sample character lengths. These results demonstrate the good diversity of our proposed SLP34K dataset.

imum length of the character sequence is set to 50. For the first-stage model pre-training, We adopt the same learning rate as used in the original MAE paper, and empirically set the mask ratio to be 0.75. The pre-training epochs are 1,500. For the second-stage model training, we use the initial learning rate of the visual encoder as $5.25e{-}4$, and the other module as $1e{-}3$. Besides, we utilize warm up strategy and cosine learning rate decay policy as done in CLIP4STR. AdamW optimizer is adopted with a decoupled weight decay value of 0.2 and the batch set is set to 32. The training epochs are 100.



(a) Single-line (b) Multi-line (c) Vertical (d) Easy

(e) Hard (f) Low-lighting (g) Reverse (h) Discontinuous

Figure 8: Samples with different characteristics in the SLP34K dataset. In addition to (a)-(e) the five explicitly labeled attribute values, the dataset also contains various other diverse samples, e.g., (f) low-lighting SLPs, (g) SLPs with text written in reverse order, and (h) SLPs with multiple discontiguous positions.

(a) Single-line (b) Multi-line (c) Vertical

(d) Blurred (e) Low-lighting (f) Occluded

Figure 9: Examples of recognition results obtained by different methods. GT indicates the growth-truth.