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

Abstract

The paper targets the challenging task of Ship License Plate (SLP) recognition. Existing methods for SLP recognition are hampered by the scarcity of large and publicly available datasets, leading to evaluations on small and non-representative datasets. To alleviate it, we have built a large dataset, called SLP34K, which consists of 34,385 images collected by an intelligent traffic surveillance system. The dataset is carefully manually annotated with text labels and attributes, and presents high data diversity by multiple installation locations and long capturing period of the cameras. Additionally, we propose a simple yet effective SLP recognition baseline method. The baseline is equipped with a strong visual encoder that benefits from initial pre-training via self-supervised learning, followed by further refinement through our devised semantic enhancement module. Extensive experiments on SLP34K verify the effectiveness of our proposed baseline. Moreover, while our baseline is designed for SLP recognition, it can also be used for common scene text recognition and achieve state-of-the-art performance on seven mainstream scene text recognition datasets.

Code, dataset, and supplementary materials — https://github.com/HuiGuanLab/SLP34K

(a) Our collected ship license plates

(b) Extreme imaging conditions

(c) Car license plates in CCPD dataset

Figure 1: Due to (a) complex character layouts and (b) extreme weather conditions, ship license plates face more intricate recognition challenges than (c) car license plates.

Introduction

Ship License Plate (SLP) is an important identity information that uniquely identifies ships. Hence, SLP recognition is crucial in the field of intelligent waterway traffic, which plays a significant role in a variety of ship-related applications, such as ship identity recognition, trajectory tracking, and navigation safety.

In recent years, there have been ongoing research efforts focusing on SLP recognition (Liu et al. 2017, 2022b; Zhou, Jiang, and Guo 2022; Wu et al. 2023), while the progress of this field has been limited. It is mainly due to the difficulty of capturing ship images, thus hindering the collection of large-scale SLP recognition datasets. Several datasets have been developed for ship license plate recognition (Liu et al. 2022b; Zhou, Jiang, and Guo 2022; Xu et al. 2024). For instance, Zhang *et al.* (Zhang et al. 2018) pioneered the creation of an SLP dataset comprising 6,000 images from 103 different ships. Afterwards, Liu *et al.* (Liu et al. 2022b) built a dataset of 2,350 SLP images, and Xu *et al.* (Xu et al. 2024) constructed a dataset that consists of 2,259 SLP images. However, all these datasets are small-scale, limiting their utility for training deep learning based models. In addition, the images of existing SLP datasets are usually captured in a specific location, which restricts the diversity of collected SLP images. This lack of diversity hampers the development of robust SLP recognition models that are capable of performing effectively in real-world scenarios.

Beyond the dataset limitations, SLP itself poses great challenges, primarily due to complex character spatial layout and visual degradation. This is mainly reflected in the

*Corresponding author, dongjf24@gmail.com

following aspects. *Complex character layout*: As illustrated in Figure 1a, SLPs are usually written in uncontrolled conditions, which leads to the characters of SLP being disordered in writing positions and formats. *Visual degradation*: As demonstrated in Figure 1b, ships frequently navigate in harsh environments where they encounter adverse weather conditions such as heavy fog and rain, or along with difficult imaging situations including back lighting and low lighting. The images captured under such conditions tend to suffer from visual degradation, such as blur, occlusion, and low illumination.

Targeting the above issues, in this work we first built a large dataset called SLP34K. The dataset was collected based on a real-world waterway intelligent traffic surveillance system. Images in the datasets were captured by eight surveillance cameras located at eight different locations over 42 months. Finally, by using more cameras located at more different locations and long-period capturing, we have obtained a dataset of 34,385 SLP images of high diversity. Furthermore, different from the existing dataset only provides the text label of license plates, we additionally perform annotation with two extra attributes, *i.e.,* text layout and recognition difficulty. Such attributes enable more finegrind experiment analysis, and also may inspire innovative approaches for SLP recognition in the future.

Additionally, as mentioned above that SLP images usually suffer from complex character spatial layout and visual degradation, we argue that a strong visual encoder is necessary for SLP recognition in the wild. To this end, we propose a simple yet effective SLP recognition baseline method, with a strong visual encoder. The strong visual encoder is achieved by two strategies. First, we employ self-supervised learning to pre-train the visual encoder, which enables the encoder to learn robust visual representations without labeled data. Second, we refine the encoder via a semantic enhancement module, which finetunes the encoder via a contrastive learning mechanism to enhance the semantic consistency between the SLP image and its text label.

It is worth noting that the most similar research direction to ours is car license plate recognition (Xu et al. 2018). Compared to ship license plates, car license plates are manufactured in relatively more controlled conditions, where characters are in a more standardized format, font styles, sizes, and colors (Figure 1c shows some examples of car license plates in the CCPD dataset (Xu et al. 2018)). Hence, the ship license plate recognition task in our work is more challenging.

To sum up, the main contributions of this paper are as follows:

- We have built to-date the largest dataset called SLP34K for the task of ship license plate recognition. The dataset is carefully manually annotated with text labels and attributes, and presents high data diversity by multiple installation locations and long capturing period of the cameras.
- We contribute a simple but effective baseline method for SLP recognition in the wild. The visual encoder of the baseline is enhanced through pre-training with self-supervised learning and further refined by semantic enhancement. The two modules are beneficial to SLP recognition.
- Experiments on SLP34K verify the effectiveness of our proposed baseline. Moreover, while our baseline is designed for SLP recognition, it can also be used for common scene text recognition and achieve state-of-the-art performance on seven mainstream datasets.

Related Work

Ship License Plate Recognition Datasets

There are several existing datasets for ship license plate recognition. Zhang *et al.* (Zhang et al. 2018) were the first to construct an SLP dataset, where 6000 SLP images were collected from 103 different ships. Images were captured by surveillance cameras located in one position over one month. The SLPR dataset (Liu et al. 2022b) contains 2,350 SLP images, which were captured using handheld smartphones or digital single-lens cameras. The annotations for this dataset were conducted on a per-line basis. It means that an SLP of multiple-line was split into multiple instances, which damaged the context information of the SLP. Xu *et al.* (Xu et al. 2024) constructed a dataset containing 2,259 SLP images, where images were also captured by surveillance cameras located in one position. Although there are a few SLP recognition datasets, it has the following issues: (1) The existing datasets are small-scale, which limits the use for training deep learning based models; (2) Images are captured by one or two cameras or even mobile phones in a specific location, which makes it challenging to guarantee a diverse collection of SLP samples. (3) All existing datasets have not been publicly released, making them difficult to access. To alleviate these issues, in this work we collect to-date the largest SLP recognition dataset, where images are captured by eight surveillance cameras located in multiple positions. We will provide comparisons and statistics for all these datasets later.

Ship License Plate Recognition Methods

SLP recognition (Liu et al. 2017, 2022b) is one of the research contents with important application value in the intelligent analysis of waterway ships, and its research interest has been increasing in recent years. Early work explored problems such as tilt correction and multi-line distribution in the SLP recognition process (Liu et al. 2018b,a). Since the low-quality blur phenomenon of SLP causes greater difficulties in recognition, an SLP super-resolution method is proposed in (Wu et al. 2023). In recent years, SLP recognition methods based on deep learning have gradually become mainstream (Liu et al. 2022b, 2018a; Zhou, Jiang, and Guo 2022; Xu et al. 2024). Such methods generally consist of deep convolutional neural networks, deep recurrent neural networks and spatial transformation neural networks. They can complete SLP recognition without segmenting SLP characters. However, despite these advancements, the number of SLP recognition methods remains limited due to the lack of publicly available datasets. Existing methods have only been tested for performance on private

Figure 2: SLP instances with different attributes.

SLP datasets in limited scenarios. To this end, this paper is the first to propose and open source an SLP recognition dataset with sufficient quantity and diverse data distribution based on a real-life waterway intelligent monitoring system and professional-level surveillance cameras.

The SLP34K Dataset

In this section, we describe how we collect representative images, annotate data, and split the dataset. Finally, we would like to summarize the existing SLP recognition datasets together and make a comparison.

Image Acquisition

We collect SLP images from an inland waterway intelligent traffic surveillance system that has been operating stably for many years. This system is equipped with eight Hikvision iDS-2CD9371-KS high-definition cameras, strategically positioned at different locations along various river sections. This strategic placement facilitates the capture of a wider range and more detailed images of SLPs. The cameras are able to automatically capture images of passing ships, and the resolution of the images is 3392×2008. As images captured in too short time intervals are similar, we only keep one and filter out other similar images to allow collected images to be diverse. Such a simple strategy helps prevent redundancy in our dataset by avoiding the inclusion of near-duplicate images. Moreover, we collect images from a span of 42 months over the past 5 years. It allows us to collect images of various seasons, weather, and lighting, thus covering a variety of real scenes in the wild. Finally, we have collected 16,000 images containing ship license plates.

Data Annotation

In order to obtain annotations, we construct a local crowdsourcing service. Unlike most text recognition-related datasets that only provide character annotation, we perform annotation with two extra attributes, *i.e.,* text layout and recognition difficulty. Such attributes enable more fine-grind experiment analysis, and also may inspire innovative approaches for SLP recognition in the future. To boost the annotation efficiency, we have developed a web-based platform specifically for data annotating, facilitating the SLP cropping, attribute and label annotations.

SLP cropping. As SLPs usually make up a small part of images, we first crop the SLP region which is then used for subsequent annotation. Similar to the common object detection annotation, we manually annotate the SLP region with a bounding box, which is the minimum bounding rectangle covering the whole SLP.

Text label. As a ship license's characters in China are typically comprised of Chinese, number, and English, we label the text content in the order of Chinese, number, and English characters to meet the official writing requirements and achieve a unified annotation style. Moreover, the dataset includes hard SLP samples with unreadable characters, we annotate these according to easy SLP samples cropped from the same ship. Concretely, as a ship often has SLP written in multiple locations on its body, we can easily infer the unclear characters by referring to the clear characters in the easy SLP of the same ship.

Attribute annotation. Some examples annotated with attributes are illustrated in Figure 2.

Text layout. We notice that SLP images collected in the wild exhibit diverse text layouts, prompting us to annotate them accordingly. To categorize the distinct characteristics of the text layout in SLPs, we use three attribute values: single-line, multiple-line, and vertical. Specifically, the three attribute values refer to the characters of an SLP being written in a single horizontal line, spread across multiple lines, and written vertically, respectively.

Recognition difficulty. The recognition difficulty is defined as how difficulty the SLP can be recognized, and has two attribute values: easy and hard.

Annotation quality control. To ensure the annotation quality, we recruit annotators with related background knowledge, and conduct cross-annotation and inspection mechanisms (Guo et al. 2024). Specifically, the annotation was performed by 23 volunteers (3 staff and 20 graduate students in our lab) who are engaged in computer vision and artificial intelligence research, with basic background knowledge of the STR task and data annotation. An SLP image was presented with two annotators for labeling, and they were asked to perform the annotation task independently. If SLP images were annotated with consistent labels, they would be directly included in the dataset. Otherwise, an inspection mechanism would be activated, where our staff volunteers would be asked to verify the results and give final annotation results.

Dataset Partition

To partition the dataset, we employ a stratified random sampling strategy to make samples with different attributes more balanced on the training and test sets (Tian et al. 2022, 2023). Specifically, we first group the dataset into three groups (*i.e.,* single-line, multiple-line, and vertical) according to their layout attributes. For each group, we further split samples into *easy* and *hard* subgroups, and randomly split the samples into the subgroups. For each subgroup, 20% samples are used for testing, and the remaining 80% are used for training. Finally, the training set contains a total of 27,501 images, and the test set has 6,884 images.

Dataset	#Image	Location	Period	Attribute	Public
Liu *et al.*	2350	-	-	No	No
Xu *et al.*	2259	1	-	No	No
Zhang *et al.*	6000	1	1 month	No	No
Ours	**34385**	**8**	**42 months**	**Yes**	**Yes**

Table 1: Comparison of ship license plate recognition datasets.

Dataset		#SLP Images			
		Train	**Test**	**Total**	**Ratio**
All		27501	6884	34385	-
Layout	single	8112	2042	10154	29.53%
	multi	15384	3847	19231	55.93%
	vertical	4005	995	5000	14.54%
Difficulty	easy	6233	1572	7805	22.70%
	hard	21268	5312	26580	77.30%

Table 2: Statistics of the attributes on the proposed dataset.

Dataset Statistics

Table 1 summarizes the statistics and comparisons among different SLP recognition datasets. Our dataset is the largest dataset in terms of SLP images, which has 34,385 images with carefully annotated labels. It is about 15 times the size of exiting datasets contracted by Liu *et al.* (Liu et al. 2022b) and Xu *et al.* (Xu et al. 2024). A major limitation of existing datasets is the constrained capturing location. For instance, Xu *et al.* (Xu et al. 2024) and Zhang *et al.* (Zhang et al. 2018) captured images from a specific location, limiting the diversity of the collected SLP samples. By contrast, we utilize eight surveillance cameras positioned in eight different locations. Moreover, our collected images were captured over 42 months, further enhancing the diversity of the dataset.

Also, the existing datasets only provide character annotation of SLP, we perform annotation with two extra attributes. Table 2 lists the statistics of attributes on our dataset. SLP34k exhibits considerable diversity in character spatial layouts. Specifically, the distribution of layouts in a single line, multiple lines, and vertically are 29.53%, 55.93%, and 14.54%, respectively. It can be observed that more than half of the dataset consists of structurally complex SLPs where characters are distributed across multiple lines but belong to the same SLP. Moreover, our dataset contains a high proportion of hard samples, with 77.30% of SLP images manually labeled as hard. Compared to the car license plates, SLPs in our dataset captured in the wild generally exhibit noticeable visual degradation, which poses greater challenges for SLP recognition.

Last but not least, all existing SLP recognition datasets have not been publicly released, making them difficult to access. Our dataset will be made publicly available for research purposes.

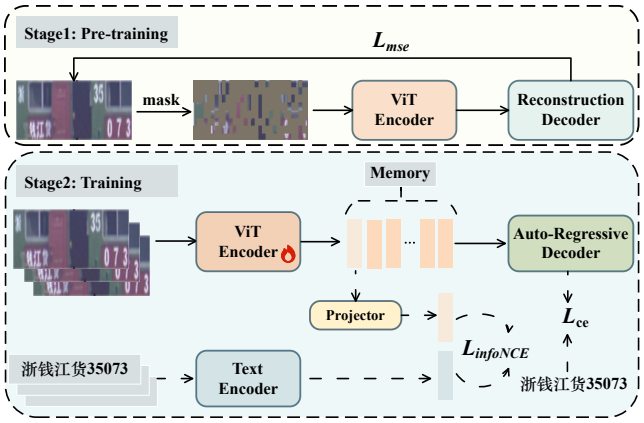

Figure 3: The framework of our proposed baseline, which adopts a classic encoder-decoder architecture.

The Proposed Baseline

As demonstrated in the above section, SLP recognition is challenging as images typically show diverse layouts (spatial distribution of characters in some SLP images is disordered) and suffer from visual degradation, such as blurring, and low lighting. Therefore, we argue that training a strong visual encoder is beneficial for SLP recognition. To this end, we devise a simple but effective baseline. The framework of our proposed baseline is illustrated in Figure 3, which adopts a classic encoder-decoder architecture. In order to boost the visual encoder, we employ self-supervised learning to pre-train the visual encoder, and propose a semantic enhancement module to guide the training of the visual encoder. For the decoder in our baseline, we directly adopt the auto-regressive decoder of the PAESeq method (Bautista and Atienza 2022), considering its good performance on scene text recognition.

Pre-training the Visual Encoder via Self-supervised Learning

Pre-training the visual encoder by supervised image classification on labeled images of ImageNet is the *de facto* choice for most multimedia and computer vision tasks (Tian et al. 2024). However, we argue that it is not an optimal choice for the ship license plate recognition task. As the visual encoder obtained by this has a strong association with the pre-defined categories, which limits the generalizability for some specific domains, such as SLP images, where images have weak relevance with these categories. Therefore, we opt to pre-train the visual encoder by self-supervised learning to diminish its association with the categories. For self-supervised learning, we employ the way of training masked autoencoders (MAE) (He et al. 2022). MAE first masks a large portion of the input image and feeds the remaining visible parts with an encoder. Besides, a decoder is employed to reconstruct the full image. Its training focuses on minimizing the reconstruction loss between the original image and the reconstructed output, effectively enabling the encoder to learn robust visual representations without labeled data. It is worth pointing out that we choose it as the pre-training way

as the masked operation in MAE somewhat resembles the visual degradation commonly observed in the SLP images. This similarity potentially makes the learned visual representations particularly effective for SLP recognition. Specifically, we utilize the structure and training way proposed in (He et al. 2022). A Vision Transformer is used as the encoder, and the mean squared error loss is used as the reconstruction loss. After pre-training, only the visual encoder is kept for second-stage training.

Enhancing the Visual Encoder via Semantic Enhancement

Although we pre-train the visual encoder through self-supervised learning, the pre-training process is task-agnostic. We further refine the visual encoder via semantic enhancement, which trains the visual encoder under the guidance of semantic labels in the scenario of SLP recognition. Specifically, on the basis of the visual encoder, we additionally introduce a textual encoder that takes the text label as input. As the SLP images are semantically relevant to their corresponding text labels, we add constraints to ensure that the encoded features from both the visual and textual encoders are closely aligned. To this end, a contrastive learning mechanism is employed across the visual and textual encoders. This mechanism aims to maximize the cosine similarity for pairs of visual and textual features corresponding to the same SLP, while minimizing it for mismatched pairs. More formally, given an SLP image annotated with a text label, we first obtain the global visual feature I and textual feature T via visual encoder and textual encoder. The visual feature I is obtained as the output of cls token, and a project head is further employed to align the feature dimensionality with the text feature. The textual feature is extracted by CLIP's textual encoder. Note that here we freeze the textual encoder as it empirically achieves better performance than the fine-tuned one. For the i^{th} sample in the same batch containing N samples, the loss of contrastive learning is implemented using infoNCE loss (Dong et al. 2022, 2023a; Zheng et al. 2023; Liu et al. 2022a), that is

$$\mathcal{L}_{infoNCE} = -\log \frac{\exp\left(I_i \cdot T_i^+ / \tau\right)}{\sum_{j=0}^{N} \exp\left(I_i \cdot T_j / \tau\right)} \tag{1}$$

where T_i^+ is the textual feature of the i^{th} sample's text label. The parameter τ, known as the temperature coefficient, is utilized to modulate the distribution of similarity scores across samples. It is worth noting that the textual encoder is not required during the inference phase, ensuring that the semantic enhancement introduces no additional computational overhead during the inference.

Training and Inference

After the model has been pre-trained by self-supervised learning, we train the whole model by jointly minimizing the cross-entropy loss $\mathcal{L}_{ce}$ for text label prediction (Liu et al. 2023) and infoNCE loss (Dong et al. 2023b) for semantic enhancement, and the whole loss is defined as:

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha \mathcal{L}_{infoNCE} \tag{2}$$

where α is a hyper-parameter to balance the importance of two losses, and $\mathcal{L}_{ce}$ is derived from (Bautista and Atienza 2022). During the inference, only the visual encoder and the auto-regressive decoder are employed for text prediction.

Evaluations

Experiment Settings

Dataset. Since our proposed dataset is the only SLP recognition dataset available to us, we conduct comprehensive experiments on it to verify the effectiveness of our baseline.

Performance metrics. Following the previous scene text recognition works (Karatzas et al. 2015; Bautista and Atienza 2022; Zhao et al. 2023), we adopt accuracy as the performance metric. The accuracy is defined as the percentage of correct test images where all characters in an image are correctly predicted. Besides, on our dataset, we also report the grouped accuracy according to the attribute values.

Comparison between Different Visual Encoders

As the visual encoder plays an important role in SLP recognition, we explore the influence of their architecture, pre-training manner, and fine-tuning in this section.

The results on our SLP34K dataset are summarized in Table 3. Note that we do not employ the proposed semantic enhancement in all models to remove its influence on the visual encoder. For each model, the fine-tuning one consistently outperforms the corresponding counterpart without the fine-tuning with a clear margin. This result is as expected. As images from the training dataset, such as ImageNet (Russakovsky et al. 2015) and WebImageText that consists of 400 million image-text pairs built by OpenAI (Radford et al. 2021), are much different from ship license plate images. For model structure, whether or not pre-training is used, the Transfomer-based ViT (row#2, row#6) is consistently better than the corresponding CNN-based ResNet152(He et al. 2016) (row#1, row#4), showing the Transfomer-based visual encoder is more suitable for SLP recognition.

To explore the necessity of the pre-training, we compare the model using the pre-training to the corresponding one using the parameters randomly initialized (without the pre-training). We explore three pre-training manners *i.e.,* classification, contrastive learning, and self-supervised learning. For classification, it pre-trains the visual encoder using the labeled images of ImageNet. For contrastive learning, we adopt image-text contrastive learning used in CLIP (Radford et al. 2021). For self-supervised learning, the training manner of MAE is utilized. As shown in Table 3, we found that the model without the pre-training (row#2) outperforms the one with classification (row#6) or contrastive learning (row#8) for pre-training. But it is worse than the model using self-supervised learning for pr-training. The results allow us to conclude the following interesting findings: (1) The pre-training of the visual encoder is not always beneficial for SLP recognition. (2) For better performance, the pre-training using self-supervised learning based

Row ID	Visual Encoder	Manner	Dataset	Fine-tuning	Acc	Layout			Difficulty	
						single-line	multi-line	vertical	easy	hard
#1	ResNet152	-	-	✓	75.60	73.69	72.31	92.46	88.19	71.93
#2	ViT-B/16	-	-	✓	79.28	77.66	76.11	94.77	92.02	75.51
#3	ResNet152	CLS	ImageNet	×	0.68	0.1	0.31	3.32	0.7	0.68
#4				✓	74.08	68.23	65.83	87.34	86.68	64.63
#5	ViT-B/16	CLS	ImageNet	×	17.44	12.98	15.46	63.22	30.95	18.88
#6				✓	77.10	76.02	73.20	94.47	91.27	72.93
#7	ViT-B/16	CTS	WebImageText	×	11.72	7.69	6.71	39.4	17.42	10.05
#8				✓	57.98	54.43	54.75	77.89	74.79	53.05
#9	ViT-B/16	SFS	ImageNet	×	44.26	35.58	39.14	81.71	59.21	39.82
#10				✓	80.88	80.52	77.47	94.97	93.24	77.26
#11			Our proposed dataset	×	61.51	49.78	60.27	90.65	76.55	57.12
#12				✓	**81.80**	**82.09**	**77.93**	**95.98**	**94.26**	**78.09**

Table 3: Performance comparison between different encoders with various architectures, pre-training manners, pre-training dataset, and fine-tuning on the SLP34K dataset. Note that '-' denotes that the model is trained from scratch without pre-training. CLS, CTS, and SFS denote pre-training through classification, contrastive, and self-supervised manners, respectively.

Manner	SEM	Acc	Layout			Difficulty	
			single	multi	vertical	easy	hard
CLS	×	77.10	76.02	73.20	94.47	91.27	72.93
	✓	**79.17**	**76.99**	**76.19**	**95.18**	**91.91**	**75.41**
CTS	×	57.98	54.43	54.75	77.89	74.79	53.05
	✓	**64.22**	**61.83**	**60.32**	**84.32**	**79.58**	**59.71**
SFS	×	81.80	82.09	77.93	95.98	94.26	78.09
	✓	**83.57**	**82.37**	**80.95**	**96.18**	**94.27**	**80.40**

Table 4: Effectiveness of our proposed semantic enhancement (SEM) on the SLP34K dataset.

Method	Arch	Acc	Layout			Difficulty	
			single	multi	vertical	easy	hard
MATRN	Res34	56.60	57.76	50.25	78.69	73.61	51.56
SemiMTR	Res34	58.68	59.67	53.76	75.78	73.87	54.22
ABINet	Res34	66.52	65.44	62.30	84.92	81.01	62.22
Ours	Res34	**76.55**	**74.53**	**73.43**	**92.76**	**88.74**	**72.95**
ViTSTR	ViT	46.81	52.52	35.15	80.30	66.86	40.93
PARSeq	ViT	70.53	69.41	66.38	89.65	84.38	66.59
MAEREC	ViT	79.84	79.09	76.14	92.86	91.09	75.98
Ours	ViT	**83.57**	**82.37**	**80.95**	**96.18**	**94.27**	**80.40**

Table 5: Performance comparison of different methods on the SLP34K dataset with the same backbone network architecture (Arch). Res34 represents ResNet34.

on MAE is recommended. Additionally, when replacing the ImangeNet dataset (row#10) with our constructed SLP34K dataset (row#12) for pre-training, self-supervised learning with MAE obtains further performance gain.

Effectiveness of Semantic Enhancement

Table 4 shows the results of our model without or with the proposed semantic enhancement. Here all the models utilize the ViT as the visual encoder, and we also try three different pre-training manners. No matter what kind of pre-training manners are employed, using our proposed semantic enhancement consistently boosts performance. The results not only demonstrate the effectiveness of the semantic enhancement but also present its flexibility to various pre-training manners. Further, we observe that the performance gain of semantic enhancement is more obvious on the more challenging *Hard* samples. Take the results of the unsupervised pre-training as an example, incorporating extra semantic enhancement obtains about 2.3% absolute accuracy increase, which is much larger than that of *Easy* samples. We attribute it to the fact that *Hard* samples typically suffer from visual information degeneration. Semantic enhancement, to some extent, alleviates its issue by training the visual encoder under the guidance of the text semantic.

Comparison with Baseline Methods

Baseline selection. As models specifically designed for SLP recognition are not source-released, we can not compare them on our dataset. Instead, we compare with models targeted at conventional scene text recognition, as the SLP recognition can be regarded as a particular case of scene text recognition in the context of ship license. Given the rich literature, we have to be selective, choosing open-source models for fair and reproducible comparison. Besides, considering ResNet34 and ViT are more dominant in recent scene text recognition methods as backbones, we select methods using ResNet34 (MATRN (Na, Kim, and Park 2022), SemiMTR (Aberdam et al. 2022), ABINet (Fang et al. 2021)) and ViT (ViTSTR (Atienza 2021), PARSeq (Bautista and Atienza 2022), MAEREC (Jiang et al. 2023)) as backbones for comparison. Following the model structure and training process consistent with the original papers, we train

Method	test of Union14M-L							Avg.	Regular			Irregular			Avg.
	curve	multi oriented	artistic	contextless	salient	multi words	general		IIIT	IC13	SVT	IC15	SVTP	CUTE	
SVTR	72.4	68.2	54.1	68.0	71.4	67.7	77.0	68.4	95.9	95.5	92.4	83.9	85.7	93.1	91.1
MORAN	43.8	12.8	47.3	55.1	45.7	54.6	44.7	43.4	94.7	94.3	89.0	78.8	83.4	87.2	87.9
ASTER	38.4	13.0	41.8	52.9	31.9	49.8	66.7	42.1	94.3	92.6	88.9	77.7	80.5	86.5	86.7
NRTR	49.3	40.6	54.3	69.6	42.9	75.5	75.2	58.2	96.2	96.9	94.0	80.9	84.8	92.0	90.8
SAR	68.9	56.9	60.6	73.3	60.1	74.6	76.0	67.2	96.6	96.0	92.4	82.0	85.7	92.7	90.9
DAN	46.0	22.8	49.3	61.6	44.6	61.2	67.0	50.4	95.5	95.2	88.6	78.3	79.9	86.1	87.3
SATRN	74.8	64.7	67.1	76.1	72.2	74.1	75.8	72.1	97.0	97.9	95.2	87.1	91.0	96.2	93.9
RobustScanner	66.2	54.2	61.4	72.7	60.1	74.2	75.7	66.4	96.8	95.7	92.4	86.4	83.9	93.8	91.2
SRN	49.7	20.0	50.7	61.0	43.9	51.5	62.7	48.5	95.5	94.7	89.5	79.1	83.9	91.3	89.0
ABINet	75.0	61.5	65.3	71.1	72.9	59.1	79.4	69.2	97.2	97.2	95.7	87.6	92.1	94.4	94.0
VisionLAN	70.7	57.2	56.7	63.8	67.6	47.3	74.2	62.5	96.3	95.1	91.3	83.6	85.4	92.4	91.3
MATRN	80.5	64.7	71.1	74.8	79.4	67.6	77.9	74.6	98.2	97.9	96.9	88.2	94.1	97.9	95.5
MAEREC	88.8	83.9	80.0	85.5	84.9	87.5	85.8	85.2	98.5	98.1	**97.8**	89.5	94.4	98.6	96.2
Ours	**91.9**	**90.4**	**81.5**	**86.1**	**88.9**	**90.2**	**86.3**	**87.9**	**99.4**	**98.8**	97.7	**89.8**	**96.4**	**99.3**	**96.9**

Table 6: Performance comparison on the largest Union14M-L and six commonly used scene text recognition datasets.

their model on the training set of our SLP34K dataset, and evaluate the performance on the corresponding test set.

Experimental results. The results on our SLP34K dataset are summarized in Table 5. Please note that our proposed baseline is compatible with both traditional CNNs and the recent ViTs as the visual backbone. Using either ResNet34 or ViT as the visual backbone, our baseline consistently outperforms the compared methods with the same backbone by a clear margin. The results demonstrate that directly adapting the scene text recognition method for ship license plate recognition is suboptimal, as ship license plate images in the wild are typically in complex layouts and suffer from visual degradation. We attribute it to the fact that our model employs pre-training by MAE and semantic enhancement, while the compared methods are trained from scratch without semantic enhancement.

Comparison in Scene Text Recognition

Although our proposed baseline is designed for ship license plate recognition, it can also be used for common scene text recognition. To further verify its effectiveness, we conduct performance comparisons on common scene text recognition datasets.

Setup. Following the previous works (Bautista and Atienza 2022; Jiang et al. 2023), we conduct experiments on Union14M-L dataset (Jiang et al. 2023), and six commonly used scene text recognition benchmarks, namely the three regular text datasets IIIT5K (Mishra, Alahari, and Jawahar 2012), ICDAR2013 (Karatzas et al. 2013), and SVT (Wang, Babenko, and Belongie 2011), and the three irregular text datasets ICDAR2015 (Karatzas et al. 2015), SVTP (Phan et al. 2013), and CUTE80 (Risnumawan et al. 2014). Union14M-L is the largest scene text recognition dataset to this date. It consists of 14 million training images and over 0.4 million real-world test images. In 14 million training images, 10 million are unlabeled and 4 million are labeled. Hence, we utilize 10 million unlabeled images for the first pre-training stage by self-supervised learning, and use labeled images for the second training stage. For the other six benchmarks, we only utilize their test set for performance evaluation, corresponding to 3,000, 1,015, 647, 2,077, 645, and 288 images, respectively. All models are trained on the training set of Union14M-L. For the performance metric, we use word accuracy as done in previous work (Bautista and Atienza 2022).

Experimental results. Table 6 summarizes the results on the test of Union14M-L, and the other six benchmarks. On Union14M-L, our proposed baseline consistently outperforms the previous methods (i.e., SVTR (Du et al. 2022), MORAN (Luo, Jin, and Sun 2019), ASTER (Shi et al. 2018), NRTR (Sheng, Chen, and Xu 2019), SAR (Li et al. 2019), DAN (Coquenet, Chatelain, and Paquet 2023), SATRN (Lee et al. 2020), RobustScanner (Yue et al. 2020), SRN (Yu et al. 2020), ABINet (Fang et al. 2021), VisionLAN (Wang et al. 2021), MATRN (Na, Kim, and Park 2022), MAEREC (Jiang et al. 2023)) that were specifically designed for scene text recognition. The results demonstrate the potential of our baseline for the common scene text recognition task. On the other six benchmarks, our baseline model performs the best on five datasets, and ranked second position on one dataset. It is worth noting the performance gain of our baseline over the second-best method on three irregular datasets (3% absolute improvement) is more significant than that on three regular datasets (1.5% absolute improvement). The result shows the advantage of our proposed method for challenging text, such as curved, blurry, rotated or even occluded samples contained in regular datasets.

Conclusion

In this paper, we contribute a large dataset, namely SLP34K, and a simple but strong baseline for SLP recognition. Extensive experiments on SLP34K showcase the effectiveness of our proposed method, and our baseline also achieves state-of-the-art performance on SLP recognition and mainstream scene text recognition datasets.

Acknowledgments

This research was supported by the Zhejiang Provincial Natural Science Foundation of China (LQ24F020005, No. LZ23F020004), National Natural Science Foundation of China (No. 62402438), Pioneer and Leading Goose R&D Program of Zhejiang (No. 2023C01212, No. 2023C01042), Young Elite Scientists Sponsorship Program by China Association for Science and Technology (No. 2022QNRC001), Fundamental Research Funds for the Provincial Universities of Zhejiang (No. FR2402ZD), Zhejiang Gongshang University "Digital+" Disciplinary Construction Management Project (No. SZJ2022C012, No. SZJ2022A002), and Open Research Project of the Key Laboratory of Public Security Information Application Based on Big-Data Architecture, Ministry of Public Security (No. 2021DSJSYS001). Additionally, we sincerely thank all the friends, colleagues, and students who assisted us with data collection and annotation throughout this work.

References

Aberdam, A.; Ganz, R.; Mazor, S.; and Litman, R. 2022. Multimodal semi-supervised learning for text recognition. *arXiv preprint arXiv:2205.03873*.

Atienza, R. 2021. Vision transformer for fast and efficient scene text recognition. In *International conference on document analysis and recognition*, 319–334. Springer.

Bautista, D.; and Atienza, R. 2022. Scene text recognition with permuted autoregressive sequence models. In *European conference on computer vision*, 178–196. Springer.

Coquenet, D.; Chatelain, C.; and Paquet, T. 2023. Dan: a segmentation-free document attention network for handwritten document recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(7): 8227–8243.

Dong, J.; Chen, X.; Zhang, M.; Yang, X.; Chen, S.; Li, X.; and Wang, X. 2022. Partially relevant video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 246–257.

Dong, J.; Peng, X.; Ma, Z.; Liu, D.; Qu, X.; Yang, X.; Zhu, J.; and Liu, B. 2023a. From Region to Patch: Attribute-Aware Foreground-Background Contrastive Learning for Fine-Grained Fashion Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1273–1282.

Dong, J.; Sun, S.; Liu, Z.; Chen, S.; Liu, B.; and Wang, X. 2023b. Hierarchical contrast for unsupervised skeleton-based action representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 525–533.

Du, Y.; Chen, Z.; Jia, C.; Yin, X.; Zheng, T.; Li, C.; Du, Y.; and Jiang, Y.-G. 2022. Svtr: Scene text recognition with a single visual model. *arXiv preprint arXiv:2205.00159*.

Fang, S.; Xie, H.; Wang, Y.; Mao, Z.; and Zhang, Y. 2021. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7098–7107.

Guo, D.; Li, K.; Hu, B.; Zhang, Y.; and Wang, M. 2024. Benchmarking Micro-action Recognition: Dataset, Method, and Application. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7): 6238–6252.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jiang, Q.; Wang, J.; Peng, D.; Liu, C.; and Jin, L. 2023. Revisiting scene text recognition: A data perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*, 20543–20554.

Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V. R.; Lu, S.; et al. 2015. ICDAR 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, 1156–1160. IEEE.

Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Bigorda, L. G.; Mestre, S. R.; Mas, J.; Mota, D. F.; Almazan, J. A.; and De Las Heras, L. P. 2013. ICDAR 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, 1484–1493. IEEE.

Lee, J.; Park, S.; Baek, J.; Oh, S. J.; Kim, S.; and Lee, H. 2020. On recognizing texts of arbitrary shapes with 2D self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 546–547.

Li, H.; Wang, P.; Shen, C.; and Zhang, G. 2019. Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8610–8617.

Liu, B.; Sheng, J.; Dun, J.; Zhang, S.; Hong, Z.; and Ye, X. 2017. Locating various ship license numbers in the wild: An effective approach. *IEEE Intelligent Transportation Systems Magazine*, 9(4): 102–117.

Liu, B.; Wu, S.; Zhang, S.; Hong, Z.; and Ye, X. 2018a. Ship license numbers recognition using deep neural networks. In *Journal of Physics: Conference Series*, volume 1060, 012064. IOP Publishing.

Liu, B.; Zhang, S.; Hong, Z.; and Ye, X. 2018b. A horizontal tilt correction method for ship license numbers recognition. In *Journal of Physics: Conference Series*, volume 976, 012013. IOP Publishing.

Liu, B.; Zheng, Q.; Wang, Y.; Zhang, M.; Dong, J.; and Wang, X. 2022a. FeatInter: exploring fine-grained object features for video-text retrieval. *Neurocomputing*, 496: 178–191.

Liu, B.; Zheng, T.; Zheng, P.; Liu, D.; Qu, X.; Gao, J.; Dong, J.; and Wang, X. 2023. Lite-MKD: A Multi-modal Knowledge Distillation Framework for Lightweight Few-shot Action Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7283–7294.

Liu, D.; Cao, J.; Wang, T.; Wu, H.; Wang, J.; Tian, J.; and Xu, F. 2022b. SLPR: A deep learning based Chinese ship license plate recognition framework. *IEEE Transactions on Intelligent Transportation Systems*, 23(12): 23831–23843.

Luo, C.; Jin, L.; and Sun, Z. 2019. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90: 109–118.

Mishra, A.; Alahari, K.; and Jawahar, C. 2012. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*, 1–11. BMVA.

Na, B.; Kim, Y.; and Park, S. 2022. Multi-modal text recognition networks: Interactive enhancements between visual and semantic features. In *European Conference on Computer Vision*, 446–463. Springer.

Phan, T. Q.; Shivakumara, P.; Tian, S.; and Tan, C. L. 2013. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE international conference on computer vision*, 569–576.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Risnumawan, A.; Shivakumara, P.; Chan, C. S.; and Tan, C. L. 2014. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18): 8027–8048.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Sheng, F.; Chen, Z.; and Xu, B. 2019. NRTR: A no-recurrence sequence-to-sequence model for scene text recognition. In *2019 International conference on document analysis and recognition*, 781–786. IEEE.

Shi, B.; Yang, M.; Wang, X.; Lyu, P.; Yao, C.; and Bai, X. 2018. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9): 2035–2048.

Tian, Y.; Fu, H.; Wang, H.; Liu, Y.; Xu, Z.; Chen, H.; Li, J.; and Wang, R. 2024. RGB oralscan video-based orthodontic treatment monitoring. *Science China Information Sciences*, 67(1): 112107.

Tian, Y.; Jian, G.; Wang, J.; Chen, H.; Pan, L.; Xu, Z.; Li, J.; and Wang, R. 2023. A revised approach to orthodontic treatment monitoring from oralscan video. *IEEE Journal of Biomedical and Health Informatics*, 5827–5836.

Tian, Y.; Zhang, Y.; Chen, W.-G.; Liu, D.; Wang, H.; Xu, H.; Han, J.; and Ge, Y. 2022. 3D tooth instance segmentation learning objectness and affinity in point cloud. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(4): 1–16.

Wang, K.; Babenko, B.; and Belongie, S. 2011. End-to-end scene text recognition. In *2011 International conference on computer vision*, 1457–1464. IEEE.

Wang, Y.; Xie, H.; Fang, S.; Wang, J.; Zhu, S.; and Zhang, Y. 2021. From two to one: A new scene text recognizer with visual language modeling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14194–14203.

Wu, H.; Chen, J.; Wang, T.; Lai, X.; and Cao, J. 2023. Ship License Plate Super-Resolution in the Wild. *IEEE Signal Processing Letters*, 30: 394–398.

Xu, F.; Chen, C.; Shang, Z.; Peng, Y.; and Li, X. 2024. A CRNN-based method for Chinese ship license plate recognition. *IET Image Processing*, 18(2): 298–311.

Xu, Z.; Yang, W.; Meng, A.; Lu, N.; Huang, H.; Ying, C.; and Huang, L. 2018. Towards end-to-end license plate detection and recognition: A large dataset and baseline. In *Proceedings of the European conference on computer vision (ECCV)*, 255–271.

Yu, D.; Li, X.; Zhang, C.; Liu, T.; Han, J.; Liu, J.; and Ding, E. 2020. Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12113–12122.

Yue, X.; Kuang, Z.; Lin, C.; Sun, H.; and Zhang, W. 2020. Robustscanner: Dynamically enhancing positional clues for robust text recognition. In *European Conference on Computer Vision*, 135–151. Springer.

Zhang, W.; Sun, H.; Zhou, J.; Liu, X.; Zhang, Z.; and Min, G. 2018. DCNN Based Real-Time Adaptive Ship License Plate Recognition (DRASLPR). In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, 1829–1834. IEEE.

Zhao, S.; Quan, R.; Zhu, L.; and Yang, Y. 2023. CLIP4STR: A simple baseline for scene text recognition with pre-trained vision-language model. *arXiv preprint arXiv:2305.14014*.

Zheng, Q.; Dong, J.; Qu, X.; Yang, X.; Wang, Y.; Zhou, P.; Liu, B.; and Wang, X. 2023. Progressive localization networks for language-based moment localization. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2): 1–21.

Zhou, X.; Jiang, L.; and Guo, A. 2022. A Feature-enhanced Ship License Plate Recognition Algorithm Based on Matched Filter. In *Proceedings of the 2022 5th International Conference on Signal Processing and Machine Learning*, 172–178.