

Article

A Self-Mutual Learning Framework Based on Knowledge Distillation for Scene Text Detection

Weisheng Zheng ¹, Xiaofei Zhang ², Kefan Qu ², Ye Tao ², Juan Feng ^{2,*} and Wangpeng He ² 

¹ Dongfang Electric Digital Technology Co., Ltd., Chengdu 610218, China

² School of Aerospace Science and Technology, Xidian University, Xi'an 710071, China

* Correspondence: fengj@xidian.edu.cn

Abstract

Knowledge distillation serves as a prevalent model compression strategy within scene text detection, enabling the transfer of learned representations from a high-capacity teacher architecture to a streamlined student counterpart. Building upon this concept, deep mutual learning alleviates dependence on the teacher model through interactive learning among student models. However, existing deep mutual learning networks inadequately address the complex redundant backgrounds and text feature distributions in scene text images, failing to effectively balance the trade-off between model performance and lightweight design. To address this issue, this paper proposes an improved self-mutual learning framework based on deep mutual learning. By employing a design that incorporates parallel multi-detection heads and interactive learning, the proposed approach simplifies the model training process while significantly improving detection accuracy. Specifically, the framework introduces a pruning mechanism that enables different detection heads to capture input features with varying degrees of sparsity. This not only reduces interference from redundant backgrounds but also leads to a more lightweight implementation. Moreover, varying feature sparsity among detection heads promotes more diverse knowledge exchange throughout mutual learning. This substantially boosts the distilled model's resilience in intricate text environments. Comprehensive evaluations show that our approach achieves superior F-measure scores compared to leading knowledge distillation methods.

Keywords: knowledge distillation; mutual learning; scene text detection; lightweight

1. Introduction

Scene text detection [1–4] is crucial for applications such as scene understanding, information extraction, and knowledge storage. However, text in natural scenes often suffers from uneven lighting, cluttered backgrounds, and variations in font, size, orientation, shape, and aspect ratio, making detection highly challenging.

In recent years, data-driven deep learning techniques guided by causal factors [5] have provided new perspectives for text detection in natural scenes. Among them, segmentation-based scene text detection methods demonstrate excellent robustness to arbitrarily shaped or curved text in complex environments due to their pixel-level prediction capability. He proposes a physically constrained dual-branch feature fusion (PC-DFF) model for passive anti-collision [6]. Yet, their large model sizes and slow inference hinder deployment on resource-limited devices. To address this, knowledge-distillation-based text detection [7,8] has become a popular solution for lightweight compression. During distillation, student networks learn “dark knowledge” from softened teacher outputs, improving performance



Academic Editors: Arkaitz Zubiaga and Manohar Das

Received: 11 January 2026

Revised: 16 February 2026

Accepted: 28 February 2026

Published: 2 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

while balancing accuracy, speed, and robustness. However, traditional distillation [9,10] still heavily depends on a well-trained teacher, limiting the student's robustness in complex scenes. Deep Mutual Learning (DML) [11,12] alleviates this issue by enabling student models to learn collaboratively, reducing reliance on a teacher and further promoting model lightweighting.

Originally proposed by Hinton et al. [13], knowledge distillation functions by fostering a transfer of learned representations between a sophisticated teacher and a compact student model. It is typically classified into offline, online, and self-distillation, distinguished by the specific nature and role of the teacher architecture. Notably, FitNet [14] pioneered knowledge transfer by passing feature outputs from intermediate layers, allowing the student model to have a different network structure from the teacher. Attention Transfer (AT) [15] proposed a straightforward feature extraction strategy, computing attention maps of intermediate features to facilitate knowledge transfer. Building on this, the emergence of the DML framework freed knowledge distillation from its reliance on a pre-trained teacher model by establishing an interactive learning mechanism among student models, demonstrating superior performance compared to traditional offline distillation. However, due to the lack of guidance from a teacher model, student models often lack robustness in diversely representing complex sample features during knowledge transfer. To address this, Zhang et al. [11] proposed a strategy where student models learn from each other, surpassing traditional offline distillation frameworks that rely on softened targets. Chung et al. [16] combined the mutual learning framework with adversarial networks, training a discriminator to encourage models to learn different feature distributions. It is noteworthy that relatively few knowledge distillation algorithms are available in the text detection domain. Existing mainstream distillation methods have not fully considered the distinctive image characteristics in scene text images, resulting in limited performance improvement for lightweight detection models.

Traditional knowledge distillation methods are mainly developed for image classification [17,18] and rarely consider the diverse key features and redundant backgrounds in scene text images. When directly applied to text detection, these methods often lead to suboptimal performance. To alleviate this, recent works introduce adversarial learning into distillation [19,20] to enhance feature diversity by modeling scene feature distributions. Meanwhile, key features in distillation essentially originate from the model itself [21], yet mainstream methods largely overlook the quality of feature maps attended by the network. In scene text images, text pixels occupy only a small proportion, causing different detection heads to focus on highly similar feature maps, which leads to semantic redundancy and unnecessary computation.

While Deep Mutual Learning (DML) reduces reliance on pre-trained teachers, standard DML approaches typically require heterogeneous network architectures to ensure feature diversity among students. Applying DML directly to a single lightweight model with identical detection heads often leads to 'homogenization,' where heads collapse into identical representations, failing to generate informative supervisory signals. Simply adding noise or dropout is often insufficient for the pixel-level precision required in scene text detection. Therefore, the specific focus of this work lies in creating a mechanism that induces structural diversity within a unified network without increasing inference cost. Unlike prior works that treat pruning solely as a compression tool, we uniquely repurpose the pruning mechanism as a 'diversity generator' to drive effective self-mutual learning. Specifically, we draw inspiration from pruner-based feature regularization [22,23] and propose a Self-Mutual Learning (SML) framework built upon the interactive learning mechanism of DML. First, to overcome the heavy reliance on well-trained, large-scale teacher models, SML builds upon the interactive learning mechanism of DML, enabling a single lightweight

student model to learn collaboratively through multiple parallel detection heads without any external teacher guidance. Second, to address the lack of feature diversity and the homogenization problem inherent in vanilla mutual learning, we introduce a sparsity-driven feature regularization module. By applying channel-wise pruning to the input features of each head, we deliberately induce different sparsity patterns, forcing the parallel heads to capture orthogonal and complementary text representations. Third, to mitigate the semantic redundancy caused by standard distillation methods ignoring the sparse nature of text pixels in cluttered backgrounds, we employ a gradient split-and-reset strategy. This mechanism dynamically filters redundant background channels and allows the model to strictly focus on key, distinct semantic features during the mutual distillation process, where prediction consistency is maintained [24].

The remainder of this paper is organized as follows: Section 2 provides a detailed description of the proposed Self-Mutual Learning (SML) framework, including the baseline model architecture and the gradient split-and-reset strategy. Section 3 presents the experimental setup, followed by a comprehensive evaluation, ablation studies, and qualitative visual comparisons on three public benchmark datasets. Finally, Section 4 concludes the paper and discusses potential directions for future work.

In summary, our contributions are as follows:

(1) We propose a pruning-assisted Self-Mutual Learning (SML) framework for scene text detection. Unlike traditional knowledge distillation that relies on a heavy pre-trained teacher, our framework enables a single lightweight student model to boost its own performance through the parallel construction and interactive learning of multiple detection heads. This streamlines the training pipeline while maintaining high efficiency.

(2) We introduce a sparsity-driven feature diversification mechanism. By incorporating an innovative pruner with a gradient split-and-reset strategy, we deliberately induce controlled sparsity differences among features. This mechanism forces different detection heads to capture complementary text representations, effectively preventing the suboptimal equilibrium often found in vanilla mutual learning.

(3) Comparative experiments on three public datasets demonstrate that the proposed method outperforms other mainstream knowledge distillation approaches in terms of the F-measure metric.

2. Methods

In this section, we provide a detailed breakdown of the proposed approach. Section 2.1 presents the baseline text detection model, covering both its architectural design and the associated loss functions. Section 2.2 then describes the proposed SML framework, in which a pruner is trained to construct multiple detection heads, enabling collaborative learning among them to enhance overall performance. Finally, Section 2.3 explains how the pruner is effectively trained to produce diverse input features through gradient decomposition and gradient reset.

2.1. Baseline Model

We adopt DBNet [25] as our baseline, a segmentation-based text detection framework composed of three main components: (a) Backbone—extracts multi-level image features and contains most of the network parameters; (b) Neck—bridges the backbone and the detection head, using Feature Pyramid Network (FPN) [26] to fuse features from different layers and thereby improve robustness to scale variations; (c) Detection Head—applies differentiable binarization to the predicted probability and threshold maps to generate an approximate binary text map.

Referring to DBNet, the overall loss L_{DB} adopted in our detector consists of three parts: the probability map loss L_s , the binary map loss L_b , and the threshold map loss L_t . The total loss is formulated as:

$$L_{DB} = L_s + \alpha \times L_b + \beta \times L_t \quad (1)$$

where L_s is the binary cross-entropy loss, L_b and L_t are the Dice loss and L1 loss, respectively, and α and β are set to 1.0 and 10 respectively.

2.2. Self-Mutual Learning Framework

Based on the aforementioned observations, our overall framework is illustrated in Figure 1. The core philosophy of the proposed Self-Mutual Learning (SML) framework is to enable a single model to learn from multiple divergent perspectives without the guidance of an external teacher. By maximizing parameter sharing while simultaneously generating sufficiently diverse internal representations, the framework facilitates effective knowledge transfer.

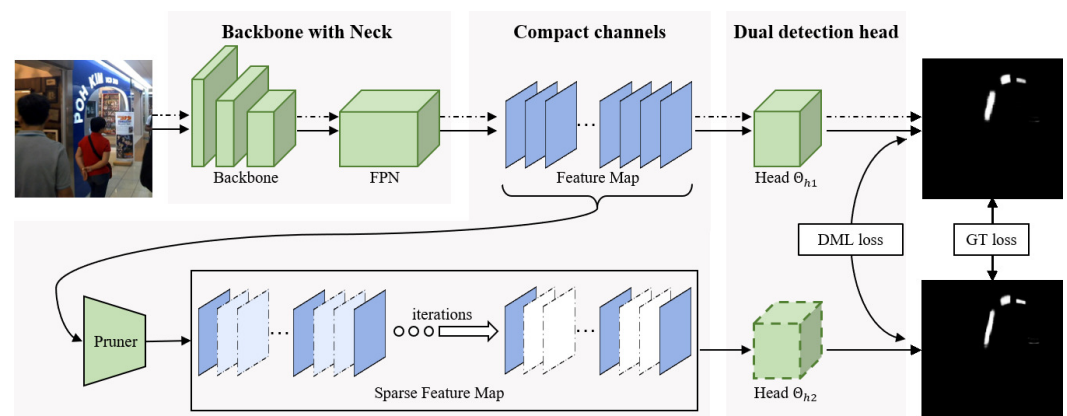


Figure 1. Overview of the Self Mutual Learning framework. Additional detection heads are added to provide supplementary knowledge signals.

A fundamental challenge in self-mutual learning is preventing the dual detection heads, Θ_{h1} and Θ_{h2} , from collapsing into identical states. Unlike existing DML approaches that rely solely on random initialization to induce diversity, our proposed SML method explicitly introduces and enforces structural diversity through channel-level pruning. Specifically, we introduce a sparse pruner P to generate complementary feature representations. In contrast to conventional pruning methods that are primarily designed for model compression, the pruner in our framework serves as a mechanism for generating structural stochasticity. By masking different subsets of channels for each detection head, we compel the two heads to rely on distinct feature patterns when solving the same detection task. This tangible advantage ensures that even if one head predominantly focuses on texture information, the other is forced to attend to structural cues such as boundaries or shapes, thereby effectively preventing feature homogenization. By applying channel-wise pruning to the FPN output feature map F , we derive a compact feature representation $F' = P(F)$. This explicitly reduces the covariance between the feature spaces of the two heads. Subsequently, the original and compact features are fed into two distinct detection heads for joint training. Let P_1 and P_2 denote the outputs of Θ_{h1} and Θ_{h2} , respectively. This forced information bottleneck ensures that the two detection heads observe the scene from different levels of abstraction and capture text boundaries as well as regional textures, thereby significantly enriching the shared backbone when distilled back during the backward pass.

Driven by these diverse inputs, the dual heads generate complementary probability and threshold maps. To facilitate knowledge transfer, we employ a mutual distillation mechanism. Specifically, the symmetric Kullback-Leibler (KL) divergence is utilized as the distillation L_{DML} :

$$L_{\text{DML}} = \frac{L_{\text{KL}}(P_1 \parallel P_2) + L_{\text{KL}}(P_2 \parallel P_1)}{2} \quad (2)$$

This objective functions as a consistency constraint, forcing each head to align its predictions with the high-confidence regions of the other. By minimizing this discrepancy, the model effectively distills the collective knowledge from both heads back into the shared backbone during the backward pass. This design ensures that Θ_{h1} benefits from the sparse, focus-driven signals of Θ_{h2} , while Θ_{h2} is guided by the comprehensive global features of Θ_{h1} . During training, the framework is optimized via a joint loss that balances individual detection accuracy and mutual alignment:

$$L_{\text{total}} = L_{\text{DB}}(P_1) + L_{\text{DB}}(P_2) + L_{\text{DML}}(P_1, P_2) \quad (3)$$

where $L_{\text{DB}}(P_1)$ and $L_{\text{DB}}(P_2)$ represent the ground-truth losses for the two detection heads, respectively.

In summary, the proposed SML framework enhances model performance in two aspects. First, by constraining the auxiliary detection head to utilize only a subset of channels, the model is encouraged to focus on the most informative features, which is particularly beneficial in complex scene text scenarios. Second, providing diverse input feature maps enriches the knowledge available to each head, thereby improving the effectiveness of mutual learning between the dual detection branches.

2.3. Gradient Split-and-Reset Strategy

Following the methodology in Section 2.2, a sparse pruner is optimized to produce heterogeneous feature maps. A prevalent and successful technique for developing sparse architectures involves incorporating a regularization term. In practice, when calculating the gradients for pruner P , this additional penalty is integrated with the primary objective L_{total} . Thus, the gradient $G(P)$ can be formulated as:

$$G(P) = \frac{\partial L_{\text{total}}}{\partial P} + \lambda * \frac{P}{\|P\|_2} \quad (4)$$

Here, λ is a hyperparameter that controls the strength of the penalty. However, conventional pruning approaches typically require fine-tuning a pre-trained network to induce sparsity, which is challenging for untrained models. This difficulty arises because the loss gradients and the sparsity-inducing penalty gradients often conflict with each other [27].

To tackle this issue, we introduce a simple yet effective technique for optimizing the sparse pruner via gradient decoupling during the training phase. As depicted in Figure 2, our process begins by aggregating the weights along the output channels; intuitively, lower weight sums reflect reduced network contribution and responsiveness. Based on this metric, the output channels of the pruner are partitioned into retained and pruned subsets. By independently routing the loss-derived gradients and the penalty gradients to these respective groups, we effectively circumvent optimization conflicts. Consequently, the ultimate gradient $G(P)$ for the pruner is defined as follows:

$$G(P) = (1 - M) * \frac{\partial L_{\text{total}}}{\partial P} + M * \lambda * \frac{P}{\|P\|_2} \quad (5)$$

where the hyperparameter λ is set to 100 in the experiments.

Ultimately, under the sole influence of the penalty gradient, redundant channels smoothly decay toward zero without interfering with the optimization of the task loss. Meanwhile, the remaining channels are completely isolated from the sparsity penalty. By learning to represent linear combinations of other channels, they form a more compact set of output channels. This split-and-reset strategy ensures that the training process remains highly stable, enabling the model to avoid catastrophic forgetting or underfitting of critical features, and preventing the gradient collapse that typically occurs when large penalties are directly applied to untrained networks.

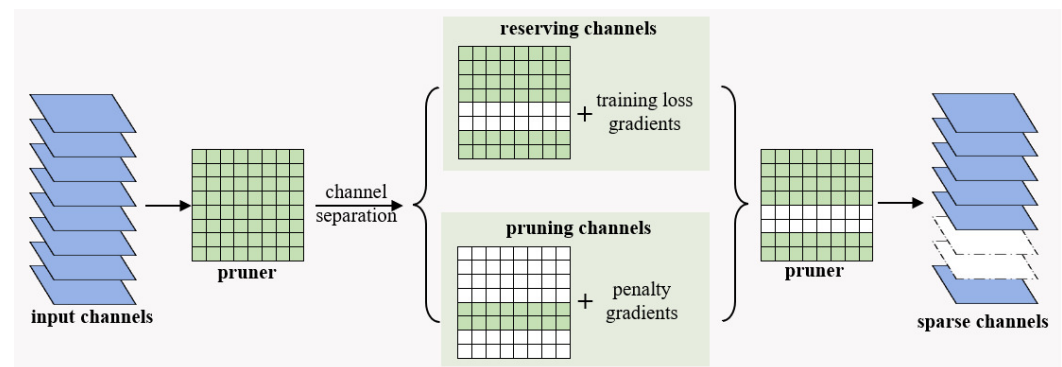


Figure 2. The pruner is trained to introduce feature diversity, where pruning is employed to enhance the distinction between different detection heads.

3. Experiments

3.1. Datasets

SynthText [28]: To provide a robust initialization for our models, we employ SynthText for pre-training. This comprehensive synthetic collection comprises more than 800,000 images tailored for detecting text within natural environments.

MSRA-TD500 [29]: Designed for multi-oriented text spotting, this dataset features a mix of English and Chinese scripts across various indoor and outdoor settings, such as malls and streets. It provides a total of 500 annotated images, partitioned into 300 for training and 200 for evaluation, with resolutions varying between 1296×864 and 1920×1280 . To ensure consistency with prior literature [25], the training set is augmented with 400 additional samples from HUST-TR400.

ICDAR 2015 [30]: This dataset, originating from the 2015 ICDAR challenge, features 1500 street-level images (1000 for training and 500 for testing) at a resolution of 720×1280 . Captured via wearable cameras, the images frequently exhibit motion blur and occlusions, presenting a rigorous test for text detection. All English text instances are localized using quadrilateral bounding boxes defined by four vertices.

Total-Text [31]: Specialized in curved and arbitrary-shaped text detection, this benchmark predominantly features street-level signage and billboards. Unlike traditional datasets, text regions here are delineated using polygons with varying vertex counts. The collection is partitioned into 1255 training and 300 testing samples, primarily encompassing English scripts with occasional Chinese characters.

3.2. Implementation Details

3.2.1. Model Structure

In this work, we adopt the DB algorithm and adapt it to a lightweight student model. Specifically, the backbone is replaced with the efficient MobileNetV3-large network, and the number of channels in the FPN is reduced from 256 to 96 to further decrease the model size. During training, additional detection heads are introduced to provide complementary

differential knowledge. These dual detection heads share the same architecture but maintain independent parameters. At inference time, the extra detection heads are removed, ensuring that the model remains lightweight and efficient.

3.2.2. Evaluation Metrics

Since scene text spotting fundamentally aligns with general object detection, its effectiveness is quantified using three standard metrics: Precision (P), Recall (R), and the F-measure ($F1$). To establish these, we first define TP as accurately localized text regions, FP as false alarms (incorrect predictions), and FN as missed ground-truth texts. Based on these fundamental counts, P evaluates the reliability of the model's positive outputs, while R assesses its capability to discover all existing text instances. The $F1$ -score subsequently unifies them by computing their harmonic mean. For all three metrics, an elevated value signifies a more robust model. The mathematical formulations are expressed in Equations (6)–(8):

$$P = \frac{TP}{TP + FP} \quad (6)$$

$$R = \frac{TP}{TP + FN} \quad (7)$$

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (8)$$

In these formulations, TP corresponds to the count of text instances accurately localized by the algorithm, FP refers to the frequency of false alarms or erroneous predictions, and FN signifies the volume of ground-truth texts that the model failed to detect.

3.2.3. Training Detail

Adopting the training protocols of the DB framework, our baseline model first undergoes a 100,000-iteration pre-training phase on SynthText to establish robust initial weights. This is followed by a 1200-epoch fine-tuning stage across three target datasets. To explicitly address text that is heavily blurred or occluded, we implement a dynamic masking strategy for highly distorted regions during optimization. The network is trained via Stochastic Gradient Descent (SGD) using a weight decay of 0.0001, a starting learning rate of 0.007, and a polynomial decay schedule. To promote model generalization, the input data is augmented using random cropping, flipping, and rotations spanning -10° to 10° , with all training samples uniformly scaled to 640×640 pixels. During evaluation, original aspect ratios are maintained while image heights are dynamically adjusted to suit each specific dataset.

3.3. Results

3.3.1. Compared with Previous Methods

The quantitative outcomes of our study are detailed in Table 1, utilizing Precision (P), Recall (R), and their harmonic average, the F-measure (F), as the primary criteria. To rigorously validate our proposed framework, we conduct evaluations across three prominent benchmarks: ICDAR2015, MSRA-TD500, and Total-Text. Our performance is benchmarked against a spectrum of established methodologies. These encompass fundamental Knowledge Distillation (KD) paradigms—namely ST [13], FitNet [14], and AT [15]—as well as DML [11] representing online distillation, and KDEM [32], a cutting-edge KD technique tailored for text spotting.

The empirical data reveals a consistent upgrade in the student network's F-measure across all evaluated benchmarks when applying our method. Detailed metrics show an absolute increase from 76.015% to 77.235% on the ICDAR2015 dataset, coupled with en-

hancements of 1.56% and 1.17% on MSRA-TD500 and Total-Text, respectively. Furthermore, our framework distinctly outperforms traditional baselines like ST, FitNet, and AT. The fact that ST and FitNet occasionally cause performance regressions on certain datasets highlights a key limitation. Conventional knowledge transfer mechanisms are misaligned with the unique requirements of segmentation-based text spotters. The simple logit-based or feature-based imitation is insufficient for DBNet, where the critical information lies in the sharp transition between text kernels and boundaries. Our SML, by contrast, focuses on internal consistency between diverse heads, which preserves these high-frequency boundary details.

Table 1. Comparison with other distillation methods.

Method	Param (M)	ICDAR2015			MSRA-TD500			Total-Text			FPS
		P	R	F	P	R	F	P	R	F	
Student_DB	25.9	82.16	70.73	76.02	85.40	72.34	78.33	85.82	71.94	78.27	27
ST (Hinton)	32.4	82.67	67.98	74.61	82.97	71.13	76.60	85.60	70.85	77.53	23
FitNet	27.2	82.32	70.15	75.75	85.89	73.20	79.04	85.26	73.70	79.06	25
AT	35.2	81.37	71.50	76.11	84.59	73.54	78.68	85.87	72.48	78.61	21
DML	28.5	81.71	71.64	76.35	84.95	73.71	78.93	84.62	73.07	78.42	21
KDEM	36.3	81.37	73.81	77.40	86.09	73.37	79.22	83.76	74.11	78.64	18
Ours	18.8	82.81	72.36	77.24	85.35	75.09	79.89	87.03	73.07	79.44	32

While our method slightly underperforms KDEM on ICDAR2015, this can be attributed to the complexity of street scenes in this dataset, which feature sparse target pixels and abundant background interference. Furthermore, our SML framework does not incorporate explicit supervision from a dedicated teacher model during training, which inevitably limits the upper bound of its performance. Although the self-mutual learning mechanism facilitates knowledge interaction and feature consistency among multiple detection heads, the supervisory signals are still mainly derived from the model's own predictions, and thus essentially belong to a form of weak or self-guided supervision. In the absence of high-level semantic constraints and structural priors provided by a stronger external teacher, the guidance signals remain limited in terms of discriminability, diversity, and stability. Specifically, in complex real-world scenarios where text foreground features are severely degraded, the detection heads—being constrained by the limited capacity and representational power of the lightweight model—tend to produce highly correlated predictions. Such correlation leads to considerable redundancy during the knowledge exchange process, thereby weakening the potential benefits of self-mutual learning. Furthermore, without high-quality soft labels or cross-level semantic constraints, the model exhibits limited capability in modeling hard samples and ambiguous boundary regions, which may result in inaccurate localization or missed detections. Consequently, the absence of a more powerful and expressive teacher model as a reference makes it difficult for the SML framework to obtain more discriminative and stable supervisory signals when facing highly challenging scenarios, thereby restricting further performance improvements on complex scene text detection tasks. Future work may explore transferring geometric-semantic joint representations from a pre-trained teacher to the lightweight student through multi-scale feature alignment and degeneration-aware knowledge distillation, potentially enhancing recognition performance.

Qualitative comparisons between the baseline and our SML method are shown in Figure 3. In Figure 3a, the input image is presented. Compared to the baseline, our method exhibits stronger focus on text regions, while attention in background regions is

more dispersed, indicating that the model has learned richer and more comprehensive feature representations.

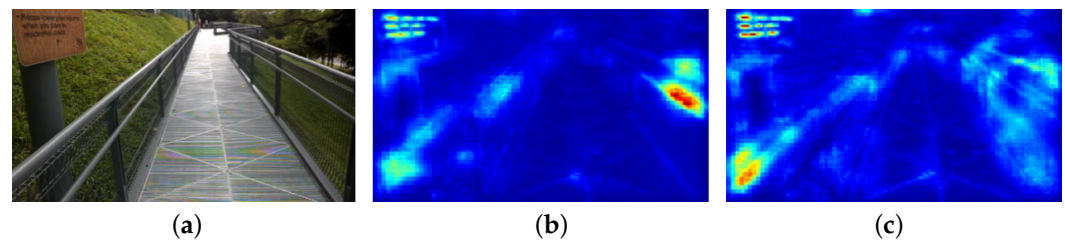


Figure 3. Comparison of attention maps generated by the baseline and SML. (a) Image, (b) Baseline, (c) SML.

3.3.2. Ablation Study

Table 2 details the ablation experiments for our proposed distillation strategy, executed entirely on the MSRA-TD500 benchmark. We initially assess the dual-head architecture by embedding a pruner within it. This pruning mechanism yields a 1.379% boost in the F1-score, suggesting that condensed feature channels enable the network to prioritize essential representations. Furthermore, by applying a Deep Mutual Learning (DML) loss to the segmentation outputs, we foster a collaborative learning environment between the parallel detection heads, which drives an additional F1-score surge of 1.566%.

Table 2. Ablation study on MSRA-TD500. The ✓ and × denote the inclusion and exclusion of the corresponding module.

Dual Head	Pruner	DML Loss	P	R	F
×	×	×	85.395	72.336	78.325
✓	✓	×	86.055	74.226	79.704
✓	✓	✓	84.482	75.773	79.891

As illustrated in Table 3, we compare the experimental results under various hyperparameter λ settings. When the penalty term is too small, the pruner fails to effectively induce sufficient sparsity discrepancies between channels. In this scenario, the features received by the dual detection heads remain largely identical, causing the self-mutual learning process to degenerate into conventional redundant training. At $\lambda = 100$, the penalty gradient is robust enough to suppress redundant channels toward zero, while the masking mechanism ensures that these intense penalty signals do not interfere with the retained channels. However, a further increase in λ imposes excessive constraints that force the model into over-simplification, leading to the loss of critical information for ambiguous text regions during the sparsification process.

Table 3. Sensitivity analysis of the penalty hyperparameter λ on the MSRA-TD500 dataset.

λ	P	R	F
0	83.922	73.851	78.565
10	84.014	74.212	78.810
100	84.482	75.773	79.891
300	84.318	73.914	78.774

As illustrated in Figure 4, the horizontal and vertical axes denote the indices of the input and output feature channels, respectively, while the color intensity reflects the degree of attention assigned to the model weights. Dark regions correspond to pruned connections,

whereas bright regions indicate parameters that play a critical role in feature modeling. Taking the model trained on the MSRA-TD500 dataset as an example, a pronounced sparsity pattern can be observed along the output channel dimension, suggesting that different detection heads learn representations with distinct sparsity characteristics. This phenomenon indicates that, by incorporating the pruning mechanism, the model can effectively suppress redundant responses caused by complex background patterns and remove unnecessary feature extractors without incurring noticeable degradation in detection performance, thereby achieving a more compact and efficient feature representation. Moreover, the diversity of channel-wise sparsity patterns across different detection heads provides complementary information for the self-mutual learning process, facilitating more effective knowledge interaction and ultimately enhancing the robustness of the model in complex scene text detection scenarios. Overall, our experiments confirm that the proposed SML framework effectively boosts model performance.

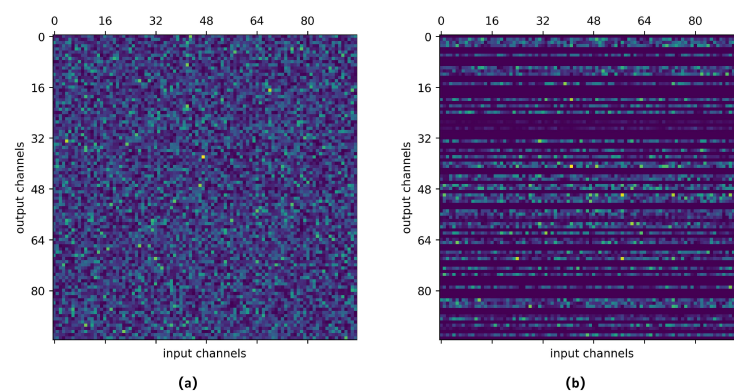


Figure 4. Visualization of the pruner throughout training. (a) Before Pruning (b) After Pruning.

3.3.3. Visual Comparison

To intuitively demonstrate the effectiveness of the proposed method, we present a qualitative comparison of detection results in Figure 5. This visualization covers representative challenging samples from the ICDAR2015, MSRA-TD500, and Total-Text datasets, where the rows correspond to the Ground Truth, DML, KDEM, and our proposed SML method. As illustrated by the green bounding boxes, the comparison highlights the distinct advantages of our approach in handling complex scenarios.

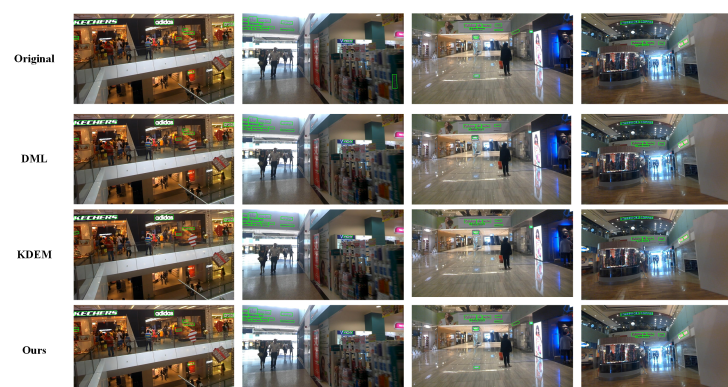


Figure 5. Detection results of different methods.

Observations from the first three columns demonstrate the model's superior capability in mitigating background interference and resolving dense text distributions. Specifically, in the first two columns, standard methods such as DML often struggle with “text-line adhesion,” where closely arranged text lines are erroneously merged into a single detection

box. In contrast, our SML method (Row 4) successfully separates these dense instances. This visual improvement directly supports the mechanism discussed in Section 3.3.4, suggesting that the pruner effectively suppresses background noise channels that typically bridge the gap between text lines, thereby generating sharper decision boundaries. This filtering capability is further evidenced in the third column, where baseline methods tend to generate False Positives on texture-rich backgrounds. SML, however, maintains high precision, aligning with the quantitative results in Table 1 and confirming that sparsity-driven training enables the model to focus on discriminative features rather than environmental noise.

Furthermore, the fourth column (Total-Text) highlights the model's geometric adaptability. While competing methods frequently miss the tail ends of highly curved text (leading to False Negatives), SML achieves more complete coverage. This indicates that the mutual consistency constraint forces the student model to learn robust geometric representations, effectively reducing the "incomplete prediction" errors analyzed in the error analysis section. In summary, Figure 5 visually corroborates our experimental findings, confirming that the SML framework not only improves numerical metrics (F-measure) but also qualitatively enhances robustness against scale variations, background clutter, and geometric distortions.

3.3.4. Error Analysis

While quantitative results validate the effectiveness of the proposed SML method, a comprehensive error analysis remains crucial for understanding the behavior of the pruner and the limitations of the mutual learning mechanism. In this section, taking the ICDAR2015 and Total-Text datasets as examples, we conduct an in-depth analysis of potential detection errors. As shown in Table 1, although our method achieves improvements in F-measure, the gains in Recall are relatively limited compared to Precision. This phenomenon can be explained by the internal feature evolution visualized in Figure 4. As observed in the transition from Figure 4a (Before Pruning) to Figure 4b (After Pruning), the pruner enforces a highly sparse representation where a significant portion of feature channels are suppressed (indicated by dark regions). While this sparsity effectively removes background redundancy, it introduces a risk: if the feature activation for a small-scale or blurred text instance is weak and relies on these suppressed channels, the model fails to capture it. Consequently, the gradient split-and-reset strategy, which prioritizes dominant features, may misclassify these weak but valid signals as "redundant," leading to false negatives as seen in the missed detections in Figure 5. Given the substantial number of curved text instances in the Total-Text dataset, although the proposed dual-head mutual learning framework ensures prediction consistency, the fact that both branches share the same backbone combined with the lack of explicit geometric supervision may cause the model to converge to an incomplete prediction state. This results in insufficient coverage of the tail-end regions of highly curved text, where the shared features may not be discriminative enough to guide the regression of extreme boundary points. In summary, the error analysis indicates that while the pruner enhances model compactness and robustness against background interference, it introduces an inevitable trade-off, wherein certain weak yet valid feature signals may be sacrificed.

4. Conclusions

4.1. Summary

In this paper, we investigate knowledge distillation for scene text detection and propose a pruner-based Self-Mutual Learning (SML) framework to improve model efficiency without relying on an external teacher network. Different from conventional offline distillation methods, our approach enhances the representation capability of the model by

introducing multiple detection heads during training, enabling mutual knowledge exchange within a single network. This design simplifies the distillation process while avoiding the dependency on large pretrained teacher models. Crucially, the additional detection heads, the pruner, and the auxiliary loss functions are strictly confined to the training phase. After training, only a single, compact student model is retained for inference. Therefore, the SML framework introduces absolutely zero architectural overhead during deployment. After training, only a compact student model is retained for inference, resulting in a strictly computational cost. The experimental results demonstrate that the proposed SML framework effectively improves the performance of lightweight student models. The core driver of this improvement lies in the “denoising” effect of the sparsity-driven pruning mechanism. As discussed in Section 3.3.4, the pruner does not merely reduce the number of active channels; it acts as a semantic filter. By suppressing channels with low responsiveness (often corresponding to background noise), the model forces the remaining active channels to focus on highly discriminative text features. Furthermore, unlike traditional mutual learning where diversity arises from heterogeneous network architectures, our method induces diversity through differential feature sparsity. This forces the dual detection heads to learn consistent predictions from structurally different feature subsets, thereby enhancing robustness.

4.2. Limitations and Future Work

Nevertheless, since the proposed framework does not leverage explicit supervision from a stronger teacher model, the performance improvements are relatively moderate. The method still encounters difficulties in challenging scenarios, such as curved or heavily blurred text instances. In addition, although the proposed framework eliminates the need for a separate teacher network, the introduction of multiple detection heads during training leads to additional computational overhead in the training phase.

In future work, we plan to explore more effective strategies for providing stronger and more accurate supervisory signals, potentially by combining self-mutual learning with adaptive or hybrid distillation schemes. We also aim to further improve training efficiency and design a more lightweight and scalable distillation framework to better handle complex text patterns and meet the practical requirements of real-world deployment. Furthermore, although our current experiments validate the SML framework using the DBNet architecture with a MobileNetV3 backbone, the core mechanisms of SML—namely, sparsity-driven feature diversification and the mutual distillation loss—are inherently architecture-agnostic. Therefore, the proposed SML framework has strong potential to be seamlessly integrated into other popular segmentation-based or hybrid text detectors. Exploring the generalizability of SML across a wider range of text detection architectures remains a promising direction for our future research.

Author Contributions: W.Z. and X.Z. wrote the original draft. K.Q. and Y.T. performed literature research. Writing—review and editing, J.F. and W.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Natural Science Basic Research Program of Shaanxi under Grant 2023JCYB289 and the Innovation Capability Support Program of Shaanxi under Grant 2022TD-37.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: Author Weisheng Zheng was employed by the company Dongfang Electric Digital Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Long, S.; He, X.; Yao, C. Scene text detection and recognition: The deep learning era. *Int. J. Comput. Vis.* **2021**, *129*, 161–184. [\[CrossRef\]](#)
- Zhu, Y.; Yao, C.; Bai, X. Scene text detection and recognition: Recent advances and future trends. *Front. Comput. Sci.* **2016**, *10*, 19–36. [\[CrossRef\]](#)
- Duan, C.; Fu, P.; Guo, S.; Jiang, Q.; Wei, X. Odm: A text-image further alignment pre-training approach for scene text detection and spotting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 15587–15597.
- Xin, Q.; Zhang, C.; Wang, Y.; Fan, C.; Yang, H.; Lang, Q.; Qi, H. DRA-Net: Dynamic Feature Fusion Upsampling and Text-Region Focus for Ancient Chinese Scene Text Detection. *Electronics* **2025**, *14*, 3324. [\[CrossRef\]](#)
- He, W.; Li, M.; Feng, C.; Wen, Z. A multi-stage domain generalisation method via exploiting causal disentanglement for machinery fault diagnosis. *Nondestruct. Test. Eval.* **2025**, 1–17. [\[CrossRef\]](#)
- He, W.; Fan, P.; Wen, Z.; Tang, X.; Zhang, S.; Gao, Y. Physically Constrained Dual-Branch Feature Fusion Model for Passive Anti-Collision While Drilling. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 6515611. [\[CrossRef\]](#)
- Bhunja, A.K.; Sain, A.; Chowdhury, P.N.; Song, Y.Z. Text is text, no matter what: Unifying text recognition using knowledge distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 983–992.
- Liu, Z.; Zhou, G.; Jia, Z.; Shi, F. Scene Text Detection in Foggy Weather Utilizing Knowledge Distillation of Diffusion Models. *IEEE Signal Process. Lett.* **2025**, *32*, 996–1000. [\[CrossRef\]](#)
- Kim, Y.; Rush, A.M. Sequence-level knowledge distillation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 1317–1327.
- Chen, G.; Choi, W.; Yu, X.; Han, T.; Chandraker, M. Learning efficient object detection models with knowledge distillation. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 742–751.
- Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H. Deep mutual learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4320–4328.
- Zhao, H.; Yang, G.; Wang, D.; Lu, H. Deep mutual learning for visual object tracking. *Pattern Recognit.* **2021**, *112*, 107796. [\[CrossRef\]](#)
- Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531. [\[CrossRef\]](#)
- Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. FitNets: Hints for Thin Deep Nets. *arXiv* **2015**, arXiv:1412.6550. [\[CrossRef\]](#)
- Zagoruyko, S.; Komodakis, N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. *arXiv* **2017**, arXiv:1612.03928. [\[CrossRef\]](#)
- Chung, I.; Park, S.; Kim, J.; Kwak, N. Feature-map-level Online Adversarial Knowledge Distillation. *arXiv* **2020**, arXiv:2002.01775.
- Luo, L.; Xue, D.; Feng, X. Automatic Diabetic Retinopathy Grading via Self-Knowledge Distillation. *Electronics* **2020**, *9*, 1337. [\[CrossRef\]](#)
- Belinga, A.G.; Tekouabou Koumetio, C.S.; El Haziti, M.; El Hassouni, M. Knowledge Distillation in Image Classification: The Impact of Datasets. *Computers* **2024**, *13*, 184. [\[CrossRef\]](#)
- Suh, S.; Rey, V.F.; Lukowicz, P. Tasked: Transformer-based adversarial learning for human activity recognition using wearable sensors via self-knowledge distillation. *Knowl.-Based Syst.* **2023**, *260*, 110143. [\[CrossRef\]](#)
- Ienco, D.; Dantas, C.F. Discom-kd: Cross-modal knowledge distillation via disentanglement representation and adversarial learning. *arXiv* **2024**, arXiv:2408.07080.
- Park, Y.J.; Cho, H.S. Subset-Aware Dual-Teacher Knowledge Distillation with Hybrid Scoring for Human Activity Recognition. *Electronics* **2025**, *14*, 130. [\[CrossRef\]](#)
- Yu, L.; Xiang, W. X-pruner: Explainable pruning for vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 24355–24363.
- Yao, G.; Wang, Y.; Xu, H.; Zhang, L.; Qi, M. Global-pruner: A stable and efficient pruner for retraining-free pruning of encoder-based language models. In Proceedings of the 28th Conference on Computational Natural Language Learning, Miami, FL, USA, 15–16 November 2024; pp. 46–55.
- Zhang, X.; Xiang, W. Diffusion Preference Alignment via Attenuated Kullback–Leibler Regularization. *Electronics* **2025**, *14*, 2939. [\[CrossRef\]](#)
- Liao, M.; Wan, Z.; Yao, C.; Chen, K.; Bai, X. Real-time scene text detection with differentiable binarization. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11474–11481.
- Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.

27. Ding, X.; Hao, T.; Tan, J.; Liu, J.; Han, J.; Guo, Y.; Ding, G. Resrep: Lossless cnn pruning via decoupling remembering and forgetting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 4510–4520.
28. Gupta, A.; Vedaldi, A.; Zisserman, A. Synthetic data for text localisation in natural images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2315–2324.
29. Yao, C.; Bai, X.; Liu, W.; Ma, Y.; Tu, Z. Detecting texts of arbitrary orientations in natural images. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2012; pp. 1083–1090.
30. Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V.R.; Lu, S.; et al. ICDAR 2015 competition on robust reading. In *Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR)*; IEEE: New York, NY, USA, 2015; pp. 1156–1160.
31. Ch'ng, C.K.; Chan, C.S. Total-text: A comprehensive dataset for scene text detection and recognition. In *Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*; IEEE: New York, NY, USA, 2017; Volume 1, pp. 935–942.
32. Chen, J.; Lai, Y.; Zeng, Y.; Yang, F. Knowledge distillation via entropy map for scene text detection. In *Proceedings of the 2021 16th International Conference on Computer Science & Education (ICCSE)*; IEEE: New York, NY, USA, 2021; pp. 506–511.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.