IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 March 2023, accepted 22 May 2023, date of publication 29 May 2023, date of current version 23 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3280547

APPLIED RESEARCH

RMFPN: End-to-End Scene Text Recognition Using Multi-Feature Pyramid Network

RUTURAJ MAHADSHETTI, GUEE-SANG LEE, (Member, IEEE), AND DEOK-JAI CHOI, (Member, IEEE)
Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, South Korea
Corresponding author: Guee-Sang Lee (gslee@jnu.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2018R1D1A3B05049058.

ABSTRACT Scene text recognition (STR) plays an important role in various computer vision activities. STR has been a desirable research topic in the computer community, and deep learning-based STR methods have gained tremendous outcomes over the past few years. Earlier state-of-the-art scene text recognition approaches even deliver a notable quantity of inaccurate yields when applied to images caught in real-world environments. Because these images lose precise text content information, previous methods generate less robust features and semantic information about text content. To address this issue, we propose a new approach called Residual Multi-Feature Pyramid Network(RMFPN), which integrates ResNet and Multi-Feature Pyramid Networks to grab multi-level relations, enrich the functionality, and generalization of the feature extractor. We build RMFPN with two convolutional pyramids as a feature extractor, which improves the robustness of features and semantic information to endure scene text recognition of various scales. Comprehensive experiments on diverse datasets demonstrate that our proposed method can acquire significant performance accuracy. The proposed RMFPN acquires a 0.61%, 1.2%, 1%, and 0.2% improvement on SVT, IC15, SVTP, and CUTE datasets.

INDEX TERMS Scene text recognition, deep learning, convolutional neural network, transformer, multi-feature pyramid network.

I. INTRODUCTION

In many languages, the text is used to communicate, record, and inherit culture. It is one of the more effective innovations of humans and has played a significant function in learning knowledge. The main goal of scene text recognition (STR) is to determine characters and text in natural scenes. Scene text recognition precisely accesses and utilizes textual information in the natural scene. It has several real-world applications, such as autonomous vehicles, partially sighted person assistance, robot navigation, instant translation, navigation, and document analysis. The challenging problem is recognizing text from natural scene images due to severe blur, perspective distortion, irregularity, and diversity of text shapes. Scene text recognition of irregular text from naturalistic images and recognizing text from low-resolution images have become seductive research subjects in the computer vision society after a resurgence of neural networks and improvement in publicly available vision datasets. The recent ICDAR robust reading competitions have shown the incipiency of advanced deep learning techniques.

Recently, several deep learning strategies have reached state-of-the-art achievement on scene text recognition, which performs well only with normal (regular) text that is often plane and frontal. However, irregular texts are arbitrarily oriented and curved. Most of the recent works [2], [3], [4], [12], [16], [19], [30], [31], [35] have endeavored to enhance the performance of scene text recognition using deep networks, such as super-resolution [2], [5], [7], [9], [12], the attention approach [4], [16], [23], elevating the backbone networks [4], [19], [30], and rectification modules [3], [21], [24], [30], [62]. Scene Text Recognition methods [31], [37], [39]

The associate editor coordinating the review of this manuscript and approving it for publication was Tallha Akram.

FIGURE 1. Sample images from the public datasets.

perform well, but their performance significantly falls with low-resolution and partially occluded images. As shown in Fig 1, the public datasets contain different kinds of images, such as normal images, low-resolution images, and partially occluded images. Low-resolution text images may be present in multiple cases, such as a photo taken with fewer focal cameras or an image squeezed to reduce disk usage. Proposed recognition methods usually use interpolation methods (bicubic, bilinear) when addressing low-resolution images. However, upsampled images are still blurred.

Existing super-resolution-based text recognition methods enrich the grade of low-resolution images and text recognition accurateness. Tran et al. [32] embraced LapSRN [7] to improve image quality and text content details. Wang et al. [12] and Bílkova and Hradiś [9] utilized a GAN-based approach with perceptual losses and CTC loss [36]. In these strategies, the bicubic and downsampling operations are used to generate low-resolution images from suitable quality images for the recognition process, while real-world images are more degraded and tricky. Recently, Wang et al. [33] presented a TSRN approach and a new dataset for low-resolution images. The gradient profile loss is utilized to capture sharp text boundaries. Ma et al. [45] introduce a TPGSR by orienting the categorical text prior details into the model training process.

The MORAN [62] trained using weak supervision, which allows it to be more flexible and adaptable than traditional methods. It can rectify images that have complicated distortions without being limited by geometric constraints. Qian et al. [35] proposed a deep learning-based framework using the upsampling approach as a preprocessing task to improve image resolution. Wang et al. [31] proposed a VisionLAN framework where linguistic knowledge is fused with a vision model and accelerates the speed. However, the framework needs an extra training process to achieve a linguistic ability or deep structure to ensure recognition accuracy, hindering its efficiency. Fang et al. [37] integrated language knowledge and used parallel prediction. Still, the existing framework's performance falls on low-quality images.

Scene text image includes semantic (linguistic) information and visual texture about the text content. The recent NLP-based STR method focuses on achieving linguistic information to encourage the recognition process. Many frameworks use visual features and semantic information in two distinct parts, like vision and language models. The vision model obtains the visual representations of texture details without assuming the semantic knowledge of words. The linguistic model illustrates the association between characters via the semantic learning structure. As shown in Fig 2, previous methods significantly fail to recognize text content from low-quality images. We introduce a novel framework using a multi-feature pyramid network to improve the robustness of the visual representations and semantic features.

The recent text and object detection approaches [54], [56], [58], [59], [61], [63] have shown the effectiveness of local features from different layers and multi-scale feature aggregation. Huang et al. [54] propose a framework using feature aggregation that focuses on overcoming the disadvantages of CTC and attention-based decoders. Yu et al. [56] have shown the effectiveness of feature aggregation from distinct layers to more suitably merge semantic and spatial information for recognition, localization, and detection. Tang et al. [55] presented a text detection network employing the FPN approach to improve the feature capability of texts of distinct scales and enrich detection accuracy. Dang and Lee [53] proposed a boundary feature-guided approach utilizing multi-task learning for text segmentation, which used both local feature information and transferred global features to acquire detailed structural feature information. The Quadbox framework, introduced by Keserwani et al. [58], uses quadrilateral geometry. The approach uses indirect regression, where all points in the quadrilateral are moved to the center, and vector regression is applied. Wu et al. [61] developed a feature fusion pyramid network that addresses the issue of combining low-level and high-level features by using two attention modules and a residual network based on FPN. Liu et al. [59] proposed FTPN, a network that combines FPN and Bi-LSTM to improve recall rates by leveraging multi-scale features. In contrast, Xie et al. [63] proposed SPCNET, which integrates FPN and instance segmentation and uses the semantic segmentation branch to capture context information and guide the detection branch.

Previous methods have achieved promising results on several benchmarks. However, there are some challenges, such as recognizing the text from low-quality images and different fonts. To address these challenges, we focus on these problems to capture robust visual and semantic information. We propose a framework using a multi-feature pyramid network approach, which mainly embraces a ladder network and produces multi-scale features with several layers to improve the robustness of semantic features and the ability of visible representation of different scales. The feature pyramid network (FPN) [34] has features from top to bottom and combines them. It gradually merges them with semantic features to obtain multi-scale features of the input image. The proposed RMFPN method enhances insufficient extracted features in the recognition process and improves accuracy.

FIGURE 2. Failure outcomes from the previous framework, such as 1) VisionLAN [31] and 2) ABINet [37]. The wrong prediction is presented in red, and accurate predictions are presented in black.

We utilize a feature pyramid for upsampling visual features and combine these features to capture the text content. The formula for the aggregation function Fn combines information from a sequence of layers $(y_1, \ldots, y_n)$ that become progressively deeper and more semantically significant. The function express in Eq. 1, where f is the integration node.

$$Fn(y_1, \ldots, y_n) = \begin{cases} y_1 & \text{if } n = 1 \\ Fn(f(y_1, y_2), \ldots, y_n) & \text{otherwise} \end{cases} \quad (1)$$

RMFPN framework incorporates multi-scale features to facilitate the network's ability to capture meaningful information from both shallow and deep layers. First, resize all extracted features as resizes as the initial layer of size $R^{h \times w \times d}$ where h is the height, w is the width, and d is the channel size. The convolution pyramid also enhances the features on various scales. Then the convolution pyramid is employed to enrich the features on various scales. The proposed framework increases the feature extractor's ability to generate robust visible features about the text content that rapidly boosts the recognition performance and overcomes false positive outcomes.

The primary contributions of this work are three-fold:

1) The proposed robust framework can effectively overcome false positives and improve recognition accuracy.
2) RMFPN approach significantly enriched the semantic information and visual feature ability of text content from indiscernible and degraded images.
3) The RMFPN framework outperforms state-of-the-art approaches on various standard benchmarks containing text image samples of distinct forms, such as horizontal and irregular text.

II. RELATED WORKS

A. SCENE TEXT RECOGNITION

Computer vision research has long focused on scene text recognition (STR). STR research has achieved considerable advancement over the last few years. Generally, previous Scene text recognition (STR) work is classified into two classes, the first Language-free approach, and the second language-aware approach.

The first category is **language-free approaches** [43], [47], [48], [51]. Language-free approaches consider text recognition as a classification task and especially depend on visible information to predict words. Patel et al. [25] have proposed a method that generates an extra lexicon for an input image to increase the performance of the text recognition system. Zhang et al. [51] proposed a framework that treats the text recognition process as a visual compare operation. To predict the text sequence, estimate the similitude between the input image feature and the predetermined alphabet characters. Liao et al. [19] introduce a text spotting method, which detects and recognizes text instances of arbitrary shapes. The spatial Attention mechanism is employed to predict the character order in words. Commonly, the language-free approaches disregard all semantic regulations during the text recognition process.

Another approach is **language-aware methods** [3], [12], [18], [50]. In this category, the proposed methods follow all semantic rules to guide the recognition task. Wang et al. [31] introduced the VisionLAN method. VisionLAN is a vision-language-based framework where linguistic knowledge is fused with a vision model and accelerates the computational cost. Lee and Osindero [49] employ the recurrent neural network (RNN) to capture the sequential dynamics in words. The MORAN [62] trained using weak supervision, which allows it to be more flexible and adaptable than traditional methods. It can rectify images that have complicated distortions without being limited by geometric constraints. Xue et al. [60] introduced the I2C2W framework, which breaks down text recognition into two tasks that are interconnected to reduce the impact of geometric and photometric distortions. The accuracy of the model may decline under certain conditions,

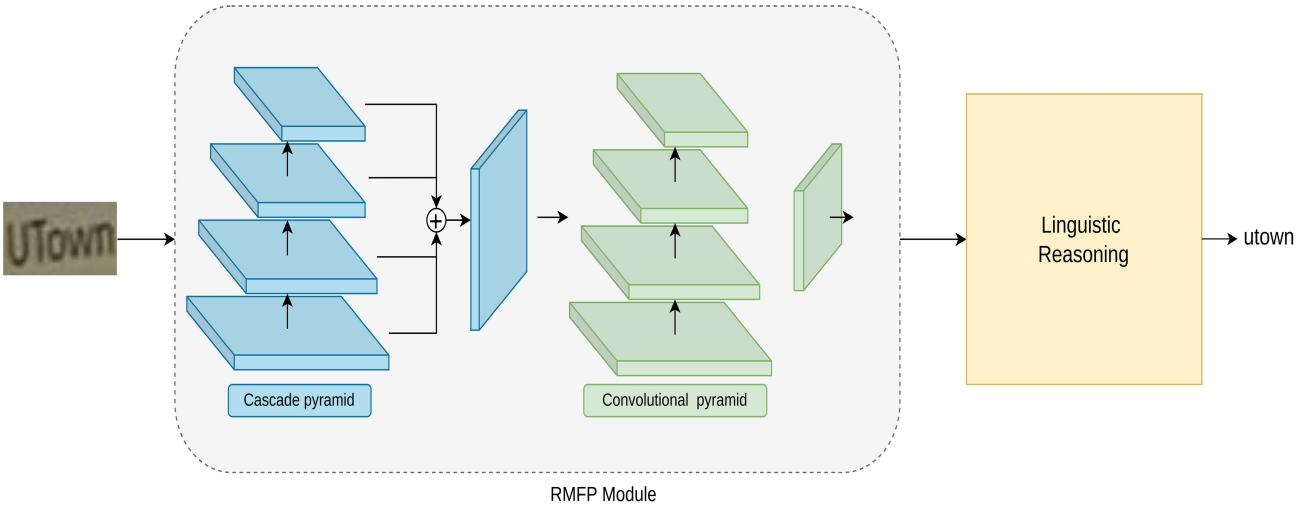


FIGURE 3. The architecture of the proposed RMFPN, RMFPN consists of two parts: RMFP, and Reasoning model.

such as an increase in character similarity or a blurred or low-resolution input image. ASTER [30] performs a rectification process first. Then use a Recurrent neural network to grab the semantic details by utilizing the outcome of the earlier step. Such a sequential process in RNN restricts the computational capability and the performance of semantic reasoning [13]. These approaches reach appreciable results in the scene text recognition technique. Previous methods focus on deep features instead of semantic information from shallow features. Different from the previous framework, we utilize shallow features from multiple layers. Thus, RMFPN is feasible to enrich visual and semantic information from degraded images.

B. LINGUISTIC REASONING PREDICTION

ViLBERT [21] strategy considers textual and visual inputs separately and trains their network employing two processes. **BERT** [11] introduces a task based on the transformer method to mask the tokens of input sentences, which is effective for long-range dependencies representation. Some researchers use this approach to address the vision-and-language task [21], [25], [31], [39]. Lu et al. [21] propose a visual-linguistic method, which carries both visible and semantic information as input.

III. PROPOSED METHOD

The detailed architecture of the proposed framework is shown in Fig 3. We integrate ResNet-50 and Multi-Feature Pyramid Networks. The backbone takes input as an image and extracts visual features. Then, the cascade pyramid and convolutional pyramid take extracted features as input to upscale features and enrich semantic information. Finally, Linguistic Reasoning parallelly predicts characters.

A. RESIDUAL MULTI-FEATURE PYRAMID MODULE (RMFP)

The RMFPN proposed to solve the problems of low-resolution, complex backgrounds. Our objective is to upgrade

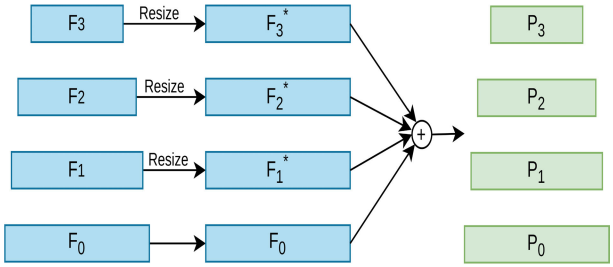


FIGURE 4. The architecture of residual multi-feature pyramid module.

the feature-capturing ability of the method and overcome false outcomes. As illustrated in Fig 4, a multi-feature pyramid network constructs multi-scale features with different layers to improve the representation capability of visible features of different scales and the robustness of semantic information. Multi-scale features are adopted in our framework to encourage the network can bind semantic details from shallow layers to deep ones. The cascaded pyramid combined with the ResNet-50 and a cascade approach utilizes to capture the features of distinct layers to emphasize the features. Finally, the second convolution pyramid upsamples features and improves semantic information.

The basic architecture of the RFPN divides into two sections. A cascade pyramid and a convolution pyramid are shown in Fig 4. F_0, F_1, F_2, and F_3 have extracted four source feature layers from ResNet-50. Max pooling operation is used to minimize the number of parameters and improves the global feature. Respectively, upsampled features from different layers (F_1^*, F_2^*, F_3^*) resize to the same size as F_0 to acquire the cascaded features. After that, all resized features are combined for recognition. This feature map includes both local and global features. The convolution pyramid also enhances the features on various scales. The convolution pyramid also enhances the features on various scales, such as P_0, P_1, P_2, and P_3. The configurations of the Residual

TABLE 1. Residual multi-feature pyramid module configuration.

Layer	Feature map size
Layer 1	192×32 × 32 × 128
Layer 2	192×64 × 16 × 64
Layer 3	192×128 × 8 × 32
Layer 4	192×512 × 8 × 32

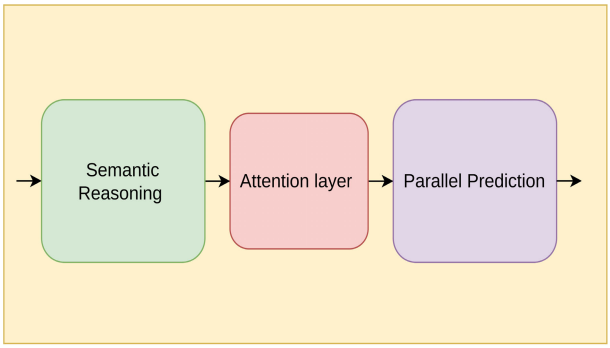


FIGURE 5. The architecture of the linguistic reasoning.

Multi-Feature Pyramid module are listed in Table 1. RMFPN comprises four layers, starting with the default stride value and gradually increasing it for the subsequent layers.

B. LINGUISTIC REASONING MODULE

We propose Reasoning Module (RM) consider visual and linguistic information concurrently in a suitable network. As a pure vision-based structure, The goal of the Reasoning Module is to use character-level information from the visual context to infer word-level predictions from features. The architecture of RM demonstrate in Fig. 5. It includes two parts, The Semantic Reasoning (SR) and the Parallel Prediction (PP) layer. SR layer holds N number of transformer units, which proves helpful for capturing the long-range dependencies in modern computer vision methods. Position encoding uses to perceive the pixel location information. The transformer units are employed in the RM to generate sequence order. Then, the Parallel Prediction layer proposes to predict the characters in parallel. To fulfill the lingustic modeling ability $w_i = f(w_N, \ldots\ldots, w_1)$, the reasoning procedure of the i^{th} character requires information about other characters. SR layer is a guide to signify the reliances between visual features to reason the semantics of characters.

C. LOSS FUNCTION

The comprehensive loss estimate is in two parts: RMFP loss and reasoning loss. We train our proposed approach using the following loss function:

$$L = \lambda_1 . L_{rmfp} + \lambda_2 . L_{lr} \tag{2}$$

where L_{rmfp} is a loss in RMFP module, and L_{lr} is a loss in LR module. We set $\lambda_1 = \lambda_2 = 0.5$ for L_{rmfp}, L_{lr}, and utilized cross-entropy loss function specify in Eq. 3 to calculate the loss. pt is the prediction of the model, gt is the ground truth, and N = 25.

$$L_* = \frac{1}{N} \sum_{t=1}^{N} \log(p_t | g_t) \tag{3}$$

IV. EXPERIMENT

In this section, We illustrate the persuasiveness of the RMFPN approach. The detailed discussion begins with the datasets utilized for training and analysis. Then present implementation details of the proposed model and evaluation facts. Next, we compare the proposed framework against SOTA approaches on standard public datasets consisting of both regular and irregular text images.

A. DATASET

The RMFPN used SynthText and SynthText90K datasets for training. Six datasets use to evaluate the framework (three regular (IC13, SVT, IIIT5k) and three irregular datasets (IC15, SVTP, CUTE80)).

SynthText [14] dataset has relatively 8-million synthetic word samples. **SynthText90K** [8] dataset contain of 9-million images covering 90k English word instances.

ICDAR 2013 (IC13) [26] includes 857 testing images. It covers data images from the IC03 dataset and grows the dataset with new images.

ICDAR 2015 (IC15) [22] contains 1811 testing images. The dataset has created with Google Glasses without considering accurate placement and focus.

IIIT 5K-Words (IIIT5k) [29] is gathered from the various websites and comprises 3000 testing image samples.

Street View Text (SVT) [17] has 647 cropped testing image examples from Google Street View. Noisy images are created by applying image processing operations.

Street View Text-Perspective (SVTP) [10] is snipped from Google Street View. The SVTP dataset includes 645 test images. The dataset contains perspectively distorted images.

CUTE80 (CUTE) is introduced in [6] for irregular and crooked text recognition. The dataset has 288 testing images, which crop from full images using annotated words.

B. IMPLEMENTATION DETAILS

The proposed RMFPN method trained end-to-end utilizing Adam optimizer with learning rate 1e-4. We employ ResNet-50 as a feature extractor. Initially, we initialize the default stride value then it assigns 2 for stages 3 and 4 to pare the feature map dimension. We utilize default weights. The recognition process covers 37 characters, including a-z alphabets, 0-9 numbers, and an end token symbol. The maximum length of the outcome order (N) is assigned to 25. All input images rescale into 64 × 256. The data augmentation process consists of color jittering, random rotation, and perspective distortion. The proposed framework train from scratch without finetuning on diverse datasets and experiments using Label smoothing and warming up. We perform the experiment on 2 NVIDIA GTX 2080ti GPUs with batch

TABLE 2. Scene text recognition accuracy compared with other STR methods on six standard benchmarks.

Method	Language Model	IC13	SVT	IIIT5K	IC15	SVTP	CUTE
GTC [64]	✗	94.30	92.90	95.50	82.50	86.20	92.30
TextScanner [52]	✗	92.90	90.10	93.90	79.40	84.30	83.30
CAFAN [48]	✗	91.40	82.10	92.00	-	-	79.90
ACE [47]	✗	89.70	82.60	82.30	68.9	70.10	82.60
Mask TextSpotter [19]	✗	95.30	91.80	95.30	78.20	83.60	88.50
SE-ASTER [15]	✓	93.80	89.60	92.80	80.00	81.40	83.60
Cheng et al. [1]	✓	93.30	85.90	87.40	70.60	-	-
VisionLAN [31]	✓	95.80	95.70	91.70	83.70	86.00	88.50
MORAN [62]	✓	93.20	88.30	91.20	77.80	79.70	81.90
ABINet [37]	✓	97.30	93.50	96.20	86.00	89.30	89.20
ASTER [30]	✓	91.80	89.50	93.40	76.10	78.50	79.50
RPI [57]	✓	92.90	91.70	95.10	78.10	84.80	91.70
I2C2W [60]	✓	95.00	91.70	94.40	82.80	83.10	93.10
SynthTIGER [25]	✓	87.90	84.50	89.80	69.50	74.60	74.00
S-GTR [40]	✓	95.80	94.10	96.80	87.90	84.60	92.30
CornerTransformer [38]	✓	96.40	94.60	95.90	86.30	91.50	92.00
MVLT [28]	✓	97.30	94.70	**96.80**	87.20	90.90	91.30
LevOCR [42]	✓	96.85	92.89	96.63	86.42	88.06	91.67
MGP-STRF [41]	✓	97.32	94.74	96.40	87.24	91.01	90.28
SVTR-L [46]	✓	97.20	91.70	96.30	86.60	88.40	95.10
Zhang et al. [27]	✓	**97.70**	94.30	96.50	85.40	89.30	91.30
Baseline+FPN	✓	96.30	95.00	96.10	87.80	89.00	93.90
RMFPN	✓	97.38	**96.31**	96.79	**89.10**	**92.50**	**95.30**

size 192 and 8 epochs. PyTorch is used to implement the proposed network.

C. DATA AUGMENTATION

Data augmentation is a necessary operation in many computer vision applications, such as object detection, text classification, semantic segmentation, and image classification. Specifically, the data augmentation process is utilized in the training phase. We experiment and prefer common operations such as random rotation, color jittering, and perspective distortion. As described in Table 4, applying the data augmentation approach, the average performance accuracy improves by 1.7%, which specifies that data augmentation is vital to scene text recognition like different strategies.

D. ABLATION STUDY

1) THE EFFECTIVENESS OF FPN

To demonstrate the significance of shallow features and robust semantic information, we apply Feature Pyramid Network to capture semantic information from the various layers. Table 3 depicts the average accuracy performance of the baseline and other tasks. The average accuracy of FPN increases by 0.9% as compared to the baseline, which shows its benefits of shallow features, the superiority of the upsampled features, and the capability to capture detailed feature information.

2) THE EFFECTIVENESS OF RMFP

In this section, we explore the significance of the proposed approach to capturing the robustness of semantic information and shallow layer feature. Ablation studies are conducted on IC13, IC15, SVT, IIT5k, SVTP, and CUTE dataset, which is accountable for evaluating performance

TABLE 3. The comparisons between baseline, FPN, and RMFPN. Six standard benchmarks utilize to estimate the average accuracy. The performances are compared during the identical training phase.

Methods	Average accuracy(%)
Baseline	92.25
Baseline+FPN	93.69
RMPN	**94.69**

TABLE 4. Ablation study of data augmentation. "DA" means data augmentation.

Methods	Average accuracy(%)
RMFPN(without DA)	92.99
RMFPN+DA	**94.69**

on regular and irregular images. Table 3 displays the average recognition accuracy of RMFP, the baseline, and the baseline with a Feature Pyramid Network. We apply Multi-Feature Pyramid Network to capture robust features. Here, we evaluate the STR average accuracy. As demonstrated in Table 3, the proposed RMFPN remarkably enhances the average recognition accuracy by 2.44% on the baseline and 1.0% on the baseline with FPN. The proposed method significantly considers shallow features and deep features together, which precisely recognize text from degraded and low-resolution images. Comparison of outcomes on degraded images from the previous framework, such as 1) VisionLAN [31] and 2) ABINet [37] and the proposed framework shown in Fig. 6. The RMFPN method effectively recognizes text from degraded images. The results presented in Figure 7, showcasing different epochs, suggest that employing the proposed method can help achieve higher accuracy.

FIGURE 6. Comparison of outcomes on degraded images from the previous framework, such as 1) VisionLAN [31] and 2) ABINet [37] and the proposed framework. The recognition result strings of RMFPN are depicted beside each image. The first row and second row represent the output of ABINet and VisionLAN, respectively. Green characters are precisely predicted by our method.

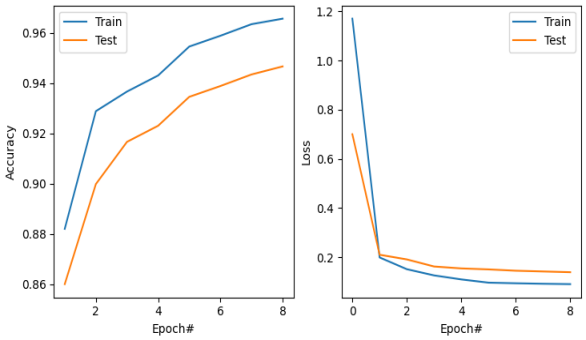


FIGURE 7. Accuracy and loss at different epochs are depicted during the training and testing stages.

E. RESULTS AND ANALYSIS

We equate the RMFPN framework with earlier state-of-the-art approaches on Six standard datasets in Table 2. Typically, the language-aware approaches achieve more promising than language-free approaches. While compared to language-free and language-aware frameworks, The proposed RMFPN delivers state-of-the-art accuracy across the six standard public benchmarks by adaptively considering the shallow feature and semantic information for feature enhancement. Our method can be easily used on various scene text recognition datasets without requiring any specific adjustments. The experimental outcomes demonstrate that our approach outperforms the existing methods. Compared to language-based methods, the proposed RMFPN method shows improvements of 0.61% on regular datasets such as SVT. For irregular datasets, the performance improves by 1.2%, 1%, and 0.2% on IC15, SVTP, and CUTE respectively. Our method provides a more intuitive way to utilize the shallow features with deep ones for scene text recognition. The RMFPN method is flexible in considering both robust visual and semantic information in the visual space.

We have evaluated the performance of RMFPN against the current state-of-the-art methods on three curved datasets (CUTE, SVTP, and CUTE) that have scene text images with noticeable geometric distortions. The results are presented in Table 2, which clearly shows that our RMFPN approach outperforms the existing state-of-the-art methods. The three curved datasets contain many images with severe geometric distortions, diverse shapes, and complicated and noisy backgrounds. Such challenges often cause misalignments of visual features at noisy time steps in previous methods, whereas our RMFPN method successfully avoids these difficulties.

The RMFPN framework can acquire more promising results than RPI [57], ASTER [30], SCRN [3], and MORAN [62] on an irregular benchmark, which embraces the rectification methodology used to correct perspective distortions and other forms of irregularities before recognition. As depicted in Tab. 1, the PMFPN framework gained 10.9%, 7.7%, and 3.6% for [57] and 11.30%, 12.80%, and 13.4% for [62] on IC15, SVTP, and CUTE datasets, respectively. For [3], the network achieved 10.4%, 11.7%, and 7.8% on irregular datasets. Our method effectively works on regular and irregular datasets and significantly achieves promising results compared to state-of-art methods.

V. CONCLUSION

In this paper, we presented a novel, efficient, and end-to-end scene text recognition framework based on blending shallow features with deep ones, enriching the semantic and visual information, which can enhance the recognition performance of heavily degraded images. We claim the importance of the relationship between shallow features and recognition. Robust, accurate visual features and semantic knowledge plays an important part in scene text recognition. The RMFPN framework effectively improves the ability of the feature extractor, which generates robust visual features and semantic information. Compared with the earlier model, RMFPN demonstrates a stronger feature capability. The proposed method archives pledging results on a different dataset. We will explore the potential of STR approaches in the future.

REFERENCES

[1] X. Chen, L. Jin, Y. Zhu, C. Luo, and T. Wang, "Text recognition in the wild: A survey," *ACM Comput. Surv. (CSUR)*, vol. 54, no. 2, pp. 1–35, 2021.

[2] J. Xu, Y. Chae, B. Stenger, and A. Datta, "Dense ByNet: Residual dense network for image super resolution," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2472–2481.

[3] M. Yang, Y. Guan, M. Liao, X. He, K. Bian, S. Bai, C. Yao, and X. Bai, "Symmetry-constrained rectification network for scene text recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct./Nov. 2019, pp. 9146–9155.

[4] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, "Focusing attention: Towards accurate text recognition in natural images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5086–5094.

[5] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[6] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," *Exp. Syst. Appl.*, vol. 41, no. 18, pp. 8027–8048, Dec. 2014.

[7] W. Lai, J. Huang, N. Ahuja, and M. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5835–5843.

[8] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014. [Online]. Available: https://arxiv.org/abs/1406.2227

[9] Z. Bílková and M. Hradiš, "Perceptual license plate super-resolution with CTC loss," *Electron. Imag.*, vol. 32, no. 6, pp. 52-1–52-5, Jan. 2020.

[10] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan, "Recognizing text with perspective distortion in natural scenes," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 569–576.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[12] W. Wang, E. Xie, P. Sun, W. Wang, L. Tian, C. Shen, and P. Luo, "TextSR: Content-aware text super-resolution guided by recognition," 2019, *arXiv:1909.07113*.

[13] D. Yu, X. Li, C. Zhang, T. Liu, J. Han, J. Liu, and E. Ding, "Towards accurate scene text recognition with semantic reasoning networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12110–12119.

[14] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2315–2324.

[15] Z. Qiao, Y. Zhou, D. Yang, Y. Zhou, and W. Wang, "SEED: Semantics enhanced encoder–decoder framework for scene text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13525–13534.

[16] Z. Wojna, A. N. Gorban, D.-S. Lee, K. Murphy, Q. Yu, Y. Li, and J. Ibarz, "Attention-based extraction of structured information from street view imagery," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Nov. 2017, pp. 844–850.

[17] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1457–1464.

[18] H. Li, P. Wang, C. Shen, and G. Zhang, "Show, attend and read: A simple and strong baseline for irregular text recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8610–8617.

[19] M. Liao, P. Lyu, M. He, C. Yao, W. Wu, and X. Bai, "Mask TextSpotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 532–548, Feb. 2021.

[20] R. Litman, O. Anschel, S. Tsiper, R. Litman, S. Mazor, and R. Manmatha, "SCATTER: Selective context attentional scene text recognizer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11959–11969.

[21] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. 33rd Adv. Neural Inf. Process. Syst.*, 2019, pp. 13–23.

[22] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, "ICDAR 2015 competition on robust reading," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1156–1160.

[23] X. Yang, D. He, Z. Zhou, D. Kifer, and C. L. Giles, "Learning to read irregular text with attention mechanisms," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, vol. 1, Aug. 2017, pp. 3280–3286.

[24] H. Jiang, Y. Xu, Z. Cheng, S. Pu, Y. Niu, W. Ren, F. Wu, and W. Tan, "Reciprocal feature learning via explicit and implicit tasks in scene text recognition," in *Proc. Int. Conf. Document Anal. Recognit.* Cham, Switzerland: Springer, Sep. 2021, pp. 287–303.

[25] Y. Patel, L. Gomez, M. Rusinol, and D. Karatzas, "Dynamic lexicon generation for natural scene images," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 395–410.

[26] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazàn, and L. P. de las Heras, "ICDAR 2013 robust reading competition," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 1484–1493.

[27] X. Zhang, B. Zhu, X. Yao, Q. Sun, R. Li, and B. Yu, "Context-based contrastive learning for scene text recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 3353–3361.

[28] J. Wu, Y. Peng, S. Zhang, W. Qi, and J. Zhang, "Masked vision-language transformers for scene text recognition," 2022, *arXiv:2211.04785*.

[29] A. Mishra, K. Alahari, and C. Jawahar, "Scene text recognition using higher order language priors," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 127.1–127.11.

[30] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: An attentional scene text recognizer with flexible rectification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2035–2048, Sep. 2019.

[31] Y. Wang, H. Xie, S. Fang, J. Wang, S. Zhu, and Y. Zhang, "From two to one: A new scene text recognizer with visual language modeling network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14174–14183.

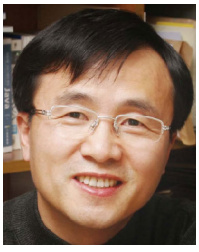
[32] H. T. M. Tran and T. Ho-Phuoc, "Deep Laplacian pyramid network for text images super-resolution," in *Proc. IEEE-RIVF Int. Conf. Comput. Commun. Technol. (RIVF)*, Mar. 2019, pp. 1–6.

[33] W. Wang, E. Xie, X. Liu, W. Wang, D. Liang, C. Shen, and X. Bai, "Scene text image super-resolution in the wild," 2020, *arXiv:2005.03341*.

[34] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[35] Y. Qian, W. Yuyang, and F. Su, "Robust scene text recognition through adaptive image enhancement," in *Proc. Brit. Mach. Vis. Conf.*, 2020.
[36] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, Jun. 2006, pp. 369–376.
[37] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7094–7103.
[38] X. Xie, L. Fu, Z. Zhang, Z. Wang, and X. Bai, "Toward understanding WordArt: Corner-guided transformer for scene text recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 303–321.
[39] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. Lee, "What is wrong with scene text recognition model comparisons? Dataset and model analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4714–4722.
[40] Y. He, C. Chen, J. Zhang, J. Liu, F. He, C. Wang, and B. Du, "Visual semantics allow for textual reasoning better in scene text recognition," in *Proc. 36th AAAI Conf. Artif. Intell.*, vol. 36, no. 1, Jun. 2022, pp. 888–896.
[41] P. Wang, C. Da, and C. Yao, "Multi-granularity prediction for scene text recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 339–355.
[42] C. Da, P. Wang, and C. Yao, "Levenshtein OCR," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 322–338.
[43] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017.
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[45] J. Ma, S. Guo, and L. Zhang, "Text prior guided scene text image super-resolution," 2021, *arXiv:2106.15368*.
[46] Y. Du, Z. Chen, C. Jia, X. Yin, T. Zheng, C. Li, Y. Du, and Y.-G. Jiang, "SVTR: Scene text recognition with a single visual model," 2022, *arXiv:2205.00159*.
[47] Z. Xie, Y. Huang, Y. Zhu, L. Jin, Y. Liu, and L. Xie, "Aggregation cross-entropy for sequence recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6531–6540.
[48] M. Liao, J. Zhang, Z. Wan, F. Xie, J. Liang, P. Lyu, C. Yao, and X. Bai, "Scene text recognition from two-dimensional perspective," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8714–8721.
[49] C. Lee and S. Osindero, "Recursive recurrent nets with attention modeling for OCR in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2231–2239.
[50] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "AON: Towards arbitrarily-oriented text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5571–5579.
[51] C. Zhang, A. Gupta, and A. Zisserman, "Adaptive text recognition through visual matching," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 51–67.
[52] Z. Wan, M. He, H. Chen, X. Bai, and C. Yao, "TextScanner: Reading characters in order for robust scene text recognition," 2019, *arXiv:1912.12422*.
[53] Q. Dang and G. Lee, "Document image binarization with stroke boundary feature guided network," *IEEE Access*, vol. 9, pp. 36924–36936, 2021.
[54] Y. Huang, C. Gu, S. Wang, Z. Huang, and K. Chen, "Spatial aggregation for scene text recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2021.
[55] Q. Tang, Q. Li, R. Wang, and Y. Lai, "Scene text detection with feature aggregation and receptive field enhancement," in *Proc. IEEE 5th Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, vol. 5, Mar. 2021, pp. 710–714.
[56] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.
[57] C. Xu, Y. Wang, F. Bai, J. Guan, and S. Zhou, "Robustly recognizing irregular scene text by rectifying principle irregularities," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1332–1339.
[58] P. Keserwani, A. Dhankhar, R. Saini, and P. P. Roy, "Quadbox: Quadrilateral bounding box based scene text detection using vector regression," *IEEE Access*, vol. 9, pp. 36802–36818, 2021.
[59] F. Liu, C. Chen, D. Gu, and J. Zheng, "FTPN: Scene text detection with feature pyramid based text proposal network," *IEEE Access*, vol. 7, pp. 44219–44228, 2019.
[60] C. Xue, J. Huang, W. Zhang, S. Lu, C. Wang, and S. Bai, "Image-to-character-to-word transformers for accurate scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–14, 2023.
[61] Y. Wu, L. Zhang, H. Li, Y. Zhang, and S. Wan, "Feature fusion pyramid network for end-to-end scene text detection," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, Jan. 2023, doi: 10.1145/3582003.
[62] C. Luo, L. Jin, and Z. Sun, "MORAN: A multi-object rectified attention network for scene text recognition," *Pattern Recognit.*, vol. 90, pp. 109–118, Jun. 2019.
[63] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, and G. Li, "Scene text detection with supervised pyramid context network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 9038–9045.
[64] W. Hu, X. Cai, J. Hou, S. Yi, and Z. Lin, "GTC: Guided training of CTC towards efficient and accurate scene text recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11005–11012.

RUTURAJ MAHADSHETTI received the B.E. degree in computer science and engineering from Shivaji University, India, in 2020. He is currently pursuing the M.S. degree in artificial intelligence convergence with Chonnam National University, South Korea. His research interests include computer vision, image processing, and deep learning.

GUEE-SANG LEE (Member, IEEE) received the B.S. degree in electrical engineering and the M.S. degree in computer engineering from Seoul National University, Republic of Korea, in 1980 and 1982, respectively, and the Ph.D. degree in computer science from The Pennsylvania State University, in 1991. He is currently a Professor with the Department of Electronics and Computer Engineering, Chonnam National University, Republic of Korea. His research interests include image processing, computer vision, and video technology.



DEOK-JAI CHOI (Member, IEEE) received the B.S. degree from the Department of Computer Engineering, Seoul National University, in 1982, the M.S. degree from the Department of Computer Science, KAIST, South Korea, in 1984, and the Ph.D. degree from the Department of Computer Science and Telecommunications, University of Missouri-Kansas City, Kansas City, MO, USA, in 1995. He is currently a Full Professor with the Department of Artificial Intelligence Convergence, Chonnam National University, South Korea. His research interests include human activity recognition, biometric authentication, and network security.