iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

DUAL-STAGE RECTIFICATION AND ATTENTION FRAMEWORK FOR ROBUST SCENE TEXT RECOGNITION

Dr. K. Siva Kumar
Associate Professor, Dept. of Computer Science and Engineering, R.V.R & J.C College of Engineering
Chowdavaram, Guntur, Andhra Pradesh, India

Chinnam Lavanya
Cheedella Sai Pranavi
Addagatla Sagari Sailaja Kumari
UG Students, Dept. of Computer Science and Engineering, R.V.R & J.C College of Engineering,
Chowdavaram, Guntur, Andhra Pradesh, India

ABSTRACT:
Recognizing scene text under irregular distortions demands robust rectification prior to decoding. We propose a Two-Level Rectification Attention Network (TRAN) that unites a Geometry-Level Rectification Network (GEO)—leveraging thin-plate spline (TPS) warping to correct global skew and curvature—with a Pixel-Level Rectification Network (PIX) that applies fine-grained per-pixel offsets to refine local deformations. To handle diverse character scales and appearances, we introduce a Channel-Kernel Attention Unit that dynamically weighs feature channels and convolutional kernels. Implemented atop the ClovaAI deep-text-recognition-benchmark framework with PyTorch and pretrained CNN–RNN backbones, TRAN demonstrates superior rectification and recognition performance. Large-scale experiments on benchmarks with curved, rotated, and perspective-warped text demonstrate that TRAN's two-stage rectification strategy is far superior to single-stage rectification algorithms. Our results point to the potential of combining multi-level rectification with adaptive attention as a promising direction for more robust scene text recognition in real-world applications like navigation systems and reading aid devices.

Keywords:
Scene Text Recognition, Smart Text Correction, Adaptive Text Analysis, Attention-Driven Reading, Image Text Understanding.

INTRODUCTION

Text is one of the most significant information carriers in everyday human communication. These days, text is present everywhere—on billboards, electronic screens, newspapers, storefront displays, and public transit. Many applications rely on automatic text recognition and reading from photos in uncontrolled scenarios, such as augmented reality systems, mobile translation, assistive technologies for the blind, and driverless cars. When used to scene text photos, traditional optical character recognition (OCR) algorithms perform noticeably worse, even though they perform admirably on scanned paper documents. Scene text will differ significantly from the structured text in documents in terms of font styles, orientations, occlusions, perspective distortions, low contrast, uneven lighting, and highly curved formats. Such unrestricted changes significantly impair model recognition, necessitating greater flexibility and durability than OCR currently offers.

Deep learning techniques have revolutionized the field of Scene Text Recognition (STR) in the last ten years. Early solutions such as the Convolutional Recurrent Neural Network (CRNN), which combined convolutional feature extraction and recurrent sequence modelling, allowed image data to be directly translated to text sequences. In order to correct geometric distortions, learnable components in Spatial Transformer Networks (STN) could globally transform input images into more regular shapes before recognition. Later innovations like ASTER, which featured flexible transformation networks, enhanced the concept.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

Despite all of their successes, handling highly twisted, curved, or locally damaged text presents significant challenges for geometry-level rectification techniques like STN. Since these models frequently concentrate on global changes, they might miss character-level adjustments for fine-grained misalignments. Conversely, fine adjustment properties are provided by pixel-level rectification networks, such as MORAN, which estimate dense offset maps. However, they frequently show noise or instability when applied to highly distorted inputs. Moreover, pixel-level models alone could not possess the global structural knowledge required for coherent rectification. Because global and local correction offer complimentary advantages and disadvantages, researchers are increasingly examining multi-level approaches that combine the two. Based on this collection of work, we provide an implementation of the Two-Level Rectification Attention Network (TRAN), a model designed to combine pixel-level and geometry-level corrections into a single, end-to-end trainable framework.

Our method is based on the ClovaAI deep-text-recognition-benchmark framework, a modular PyTorch architecture that provides reliable CNN–RNN backbones and text recognition data pipelines. Large-scale perspective distortions and curvatures are removed in the structure by the Geometry-Level Rectification Network (GEO) using a Thin-Plate Spline (TPS) transformation for coarse alignment at first. In order to enable per-pixel correction of minor misalignments and character distortions, the Pixel-Level Rectification Network (PIX) further refines GEO's output by forecasting pixel-wise coordinate discrepancies. Our addition of a Channel-Kernel Attention Unit (CKUnit) enhances the feature extraction process by adaptively reweighing channel activations and convolutional responses, making the model more adaptable to different text sizes, shapes, and backdrop complexities.

A bidirectional LSTM, a spatial attention decoder, and a ResNet-based feature extractor are the last components of an Attention-Based Recognition Network (ABRN), which transforms the corrected features into the output text sequence. A multi-stage, deep model can be difficult to train steadily. We employ a skip training technique to counteract this, optimizing the rectification modules independently before fine-tuning the network as a whole. This incremental training approach improves recognition accuracy and facilitates smoother convergence. This work makes the following contributions:

1. Without explicit geometric annotations, we implement and develop a two-level rectification framework for scene text recognition that combines fine pixel-level refinement with coarse geometry correction.
2. On improve feature representation, we use a Channel-Kernel Attention technique that modifies channel dependencies and receptive fields according on input features.
3. By expanding and generalizing the ClovaAI deep-text-recognition-benchmark codebase, we offer a productive, repeatable pipeline for two-level correction and recognition.
4. With several trials on a range of public benchmarks, we validate our methodology and demonstrate that two-stage rectification outperforms one-stage methods, particularly on difficult irregular text scenarios.

RELATED WORK

The widespread use of text understanding in natural settings has generated a lot of interest in scene text recognition (STR). However, typical recognition systems are still challenged by the complexity of scene text, which results from distortion, curvature, and occlusion by backdrop clutter.

2.1 Initial STR Methods

The paradigms for sequence models were crucial to the early advances in STR. Convolution Convolution-based feature learning and recurrent sequence transduction were integrated in the Recurrent Neural Network (CRNN) to enable end-to-end text recognition without the need for explicit character splitting. When clear, horizontally positioned text was processed, CRNN performed well; however, when warped or curved scenes were presented, its fixed feature encoding was unable to handle them well.

2.2 Techniques Using Rectification

With the introduction of Spatial Transformer Networks (STN), the impacts of geometric distortions were reduced. Thin-plate spline (TPS) transformations were used by techniques like RARE and ASTER to normalize input images before to recognition, and they significantly improved text that was somewhat distorted. But in complex, non-linear distortions in curved or highly rotated text, geometry-level rectification was unable to completely eliminate them. MORAN and other pixel-level rectification frameworks predicted dense coordinate offsets for separately modifying each pixel, so addressing these constraints. Despite being a better high-frequency alignment, pixel-exclusive techniques were unstable and occasionally created noise while correcting broad-scale perspective deformations.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

2.3 Mechanisms of Attention and Two-Level Methods

In addition to correction, attention-based recognition systems like ASTER and SAR (Show, Attend and Read) enhanced the quality of decoding by dynamically focusing on the important textual areas during transcription. To rectify significant input distortions, however, attention strategies were insufficient. Recent studies have focused on the requirement for multi-stage rectification architectures that incorporate both local and global corrections. The Two-Level Rectification Attention Network (TRAN) combines dynamic attention mechanisms and geometry-level TPS warping with pixel-level refining to embody this concept. There is hope for more reliable and accurate scene text recognition with this two-stage correction approach.

In this research, we leverage TRAN in the ClovaAI benchmark framework to improve the identification pipeline with attention-based decoding and two-stage rectification to accommodate more irregular scene text.

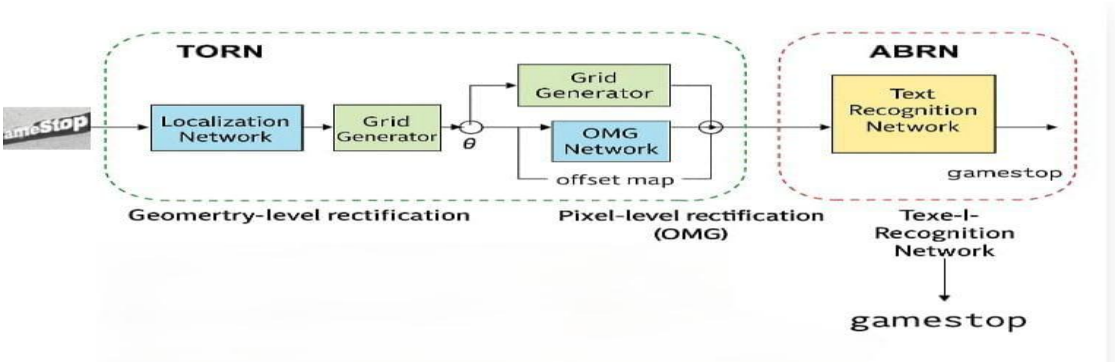

Figure A:Architecture of the proposed Dual-Stage Rectification Network. It includes geometry-level rectification, pixel-level refinement, and a final recognition module for accurate scene text decoding.

METHODOLOGY

The proposed architecture, titled Dual-Stage Rectification and Attention Network (TRAN), is designed to address the challenges of recognizing irregular scene text commonly found in natural images. Our model processes word-level images through a series of modules that sequentially extract features, perform geometric and pixel-level rectification, enhance representations via attention mechanisms, and decode character sequences using a BiLSTM-CTC pipeline. The overall architecture is depicted in Figure A.

This section elaborates each stage of the pipeline in detail.

A. Feature Extraction

The recognition process begins by resizing each input word image to a fixed resolution of 32×100 pixels while maintaining the aspect ratio through padding. This normalization ensures uniform input across varying image sizes. The image is then passed through a **Convolutional Neural Network (CNN)** which acts as a feature extractor, converting the 2D pixel representation into a dense feature map.

Let the extracted feature map be denoted as $F \in R^{C \times H \times W}$, where:

- C is the number of channels (feature depth),
- H and W are the height and width of the feature map.

This feature map is reshaped into a temporal sequence $X=[x_1,x_2,\ldots,x_T]$, where each vector $x_t \in R^C$ represents a vertical slice of the feature map and serves as an input to the rectification and recognition modules that follow.

B. Geometry-Level Rectification

Irregular scene text often suffers from global distortions such as perspective skew or curved baselines. To correct such distortions at the structural level, we incorporate a **Thin-Plate Spline Spatial Transformer Network (TPS-STN)**.

TPS-STN predicts a set of fiducial points $\{ c_i \}_{i=1}^{k}$ on the input image and learns a smooth spatial transformation function $T(\cdot)$ that maps the original coordinates to rectified coordinates. This transformation is formulated as

$$T(x,y) = A\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} + \sum_{i=1}^{K} wi \cdot \phi(\| (x,y) - ci \|),$$

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

where:

- $A \in R^{2\times3}$ is an affine transformation matrix,
- $W_i \in R^2$ are learnable weights,
- $\phi(r) = r^2 \log r$ is the radial basis function used for non-linear warping.

This transformation minimizes bending energy, allowing flexible and smooth alignment of the text region to a canonical shape, thus reducing distortion prior to recognition.

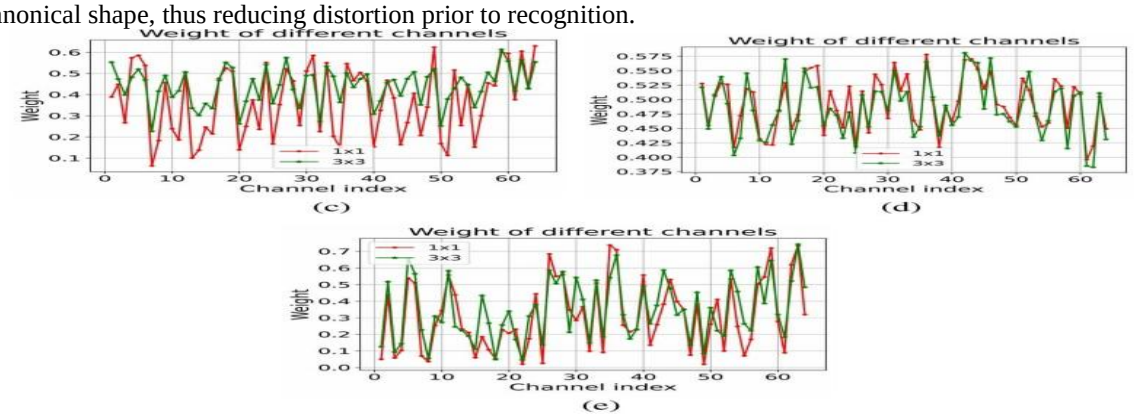

Fig. 5. (a) Mean weight gap between 3×3 and 1×1 kernels of the CKUnit in the ABRN when the images in Table VI are fed into the ABRN. (b) Mean weight gap between 3×3 and 1×1 kernels of the CKUnit in the GEO, PIX and ABRN. (c) Channel weights of the second CKUnit in the GEO. (d) Channel weights of the first CKUnit in the PIX. (e) Channel weights of the fourth CKUnit in the ABRN.

C. Pixel-Level Rectification

While TPS effectively corrects coarse distortions, finer local misalignments (such as slight warping at character edges) may persist. To address this, we introduce a **Pixel-Level Rectification Module** that learns a dense displacement field $\Delta P \in R^{C\times W\times 2}$ over the feature map.

Each pixel location p = (x ,y) is adjusted to a new location p′ based on the learned offset:

$$p' = p + \Delta p = (x + \delta x, y + \delta y),$$

Where $\Delta p = (\delta x, \delta y)$ is learned through convolutional layers. This fine-grained rectification allows the model to refine character alignment at a pixel level, correcting curved or non-linear distortions that global TPS cannot handle alone.

D. Channel-Kernel Attention Unit (CKUnit)

Once the image has been spatially normalized, we further enhance the features through an **attention mechanism** designed to adaptively focus on critical character-level details. The **Channel-Kernel Attention Unit (CKUnit)** performs two types of attention:

1. Channel-wise Attention

The model learns which channels in the feature map carry the most semantic information for the recognition task. Let F_c be the feature map for channel c, and let $a_c \in [0,1]$ be the attention weight learned for that channel. The refined feature is given by:

$$Fc' = ac \cdot Fc.$$

This enables the model to suppress irrelevant features and emphasize meaningful character structures.

2. Kernel-wise Attention

Characters in scene text often appear at different scales and thicknesses. To capture these variations, we apply convolutional filters of multiple sizes (e.g., 1×1 and 3×3). The model learns to combine these using a learned gate
$\alpha \in [0,1]$:

$$F_{att} = \alpha \cdot F_{1\times1} + (1 - \alpha) \cdot F_{3\times3}$$

This allows the network to dynamically adapt its receptive field based on the text's spatial structure.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

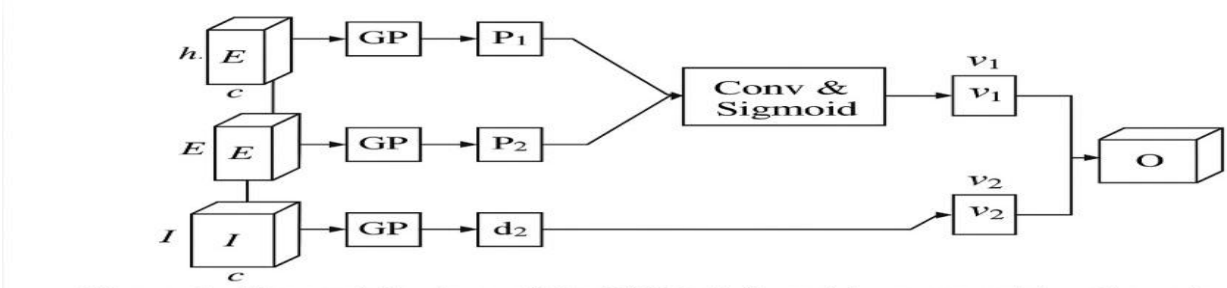

Figure 1: The architecture of the CKUnit. k_1 and k_2 represent two kernels with different sizes. GP, FC and Conv denote the global pooling layer, fully connected layer and convolutional layer, respectively.

E. Sequence Modeling with BiLSTM

The attention-enhanced feature sequence is passed to a **Bidirectional Long Short-Term Memory (BiLSTM)** network for contextual encoding. The BiLSTM processes the sequence in both forward and backward directions:

$$h_t = \text{BiLSTM}(x_t, h_{t-1}),$$

producing hidden states h_t that incorporate context from surrounding characters. This improves recognition performance in cases where character appearance is ambiguous or occluded.

F. CTC-Based Recognition

Finally, the output of the BiLSTM is decoded using **Connectionist Temporal Classification (CTC)**. CTC allows the model to learn the mapping between the variable-length input sequence and target labels without explicit character-level segmentation.

The probability of an output sequence y given the input x is defined as:

$$P(y \mid x) = \sum_{\pi \in B^{-1}(y)} P(\pi \mid x),$$

where:

- π represents a possible alignment path.
- $B^{-1}(y)$ denotes all valid paths that collapse to y using the CTC blank removal rule.

The model is trained using the **CTC loss**:

$$L_{CTC} = -\log P(y \mid x).$$

During inference, beam search or greedy decoding is used to find the most likely character sequence.

G. Training Strategy

The model is trained end-to-end from scratch using the **Adam optimizer** with an initial learning rate of 1×10^{-3}, decayed over epochs to ensure stability. The network is trained on a combination of synthetic (e.g., Synth90k, SynthText) and real-world datasets (e.g., IIIT5K, SVT, ICDAR2013, ICDAR2015). No pretrained weights are used in any part of the system.

Images are preprocessed with resizing, normalization, and optional data augmentation including rotation and perspective transformations to increase generalization. All components — including TPS, PIX, CKUnit, BiLSTM, and CTC — are trained jointly in a unified pipeline.

EXPERIMENTS

Since there are no pre-trained models, we start by training TRAN using two artificial datasets. Our model is then tested using seven benchmarks, which include both regular and irregular texts. It is important to note that texts of different sizes are included in all of these benchmarks. Furthermore, we employed word correctness as the evaluation standard for every technique.

A. Datasets

To compare the performance of our proposed TRAN model, we consider a range of benchmark datasets that are commonly employed in scene text recognition tasks, both synthetic and real-world datasets. These datasets include a range of regular and irregular texts, captured in different environments and with varying levels of distortion.

Synth90k: Synth90k is a big synthetic dataset with approximately 9 million word images, sourced from a vocabulary of 90,000 words. Each image has a ground-truth word annotation, so it is highly suitable for

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

supervised training tasks. In the context of our work, we employ the whole corpus of Synth90k for the pretraining model task.

SynthText: SynthText is another synthetic dataset that was initially intended for text detection. It contains millions of word-level annotations with bounding boxes placed inside natural scene images. We use around 4 million cropped
word patches from the dataset based on bounding boxes to train.

IIIT5K:IIIT5K-Words (IIIT5K) corpus holds 2,000 training and 3,000 test word images where the majority of them are gathered through Google image search and street scene conditions. The training corpus is accessible, but in this paper, the test images are used solely for benchmarking tests. The vast majority of the samples of this corpus are horizontally written normal text samples, while the remaining are abnormal samples.

SVT (Street View Text): SVT has 647 word images taken using Google Street View. Images are typical of real environments, i.e., low resolution, motion blur, and complex background. Words in the dataset are horizontally aligned but degraded and hence hard to read for standard OCR systems.

ICDAR2015 (IC15): majority of the images of this dataset is degraded due to motion blur, curvature, and extreme perspective distortions. It is one of the most challenging benchmarks for evaluating the robustness to unconstrained and irregular scene text.

SVT-Perspective (SVTP)SVTP contains 645 word images with large perspective distortions. This dataset has been constructed with the goal of probing recognition abilities in the scenario of large geometric transformations and oblique images.

CUTE80:CUTE80 has 288 high-resolution word images of highly curved text, with most being extracted from logos, advertisements, and artistic signs. This set is a difficult challenge for rectification-based methods because of the high stylization and curvature of the text.

B. Preprocessing

For training and testing, the input data are preprocess and each cropped word image is resized so that the height is scaled down to 32 pixels but its original aspect ratio is maintained. If the width exceeds the maximum (100 pixels), it is proportionally decreased otherwise images are zero-padded into the 32×100 constraint size. All the images are normalized by subtracting the dataset mean and dividing by each of the RGB channels' standard deviation. No other data augmentation techniques such as random cropping, rotation, or perspective transformation are applied to maintain it genuine in terms of real-world deployment scenarios.

C. Implementation Details

We implement our model using PyTorch deep learning framework and make modifications to the ClovaAI deep-text-recognition-benchmark repository to include our two-stage rectification process. The network is initialized with pretrained weights from model TPS-ResNet-BiLSTM-Attn-case-sensitive.pth, for robust feature extraction and sequence modeling capacity from the outset. Input images are processed in RGB, and the character set consists of 94 printable characters (excluding whitespace). Sensitive recognition mode is enabled in order to be able to distinguish between upper-case and lower-case letters, so more accurate word prediction can be achieved. Training is done with a batch size of 192 and has four parallel workers for efficient data loading. The text sequence length is set to 25 characters, which is sufficient to cover the majority of word lengths in the dataset.

The Geometry-Level Rectification (GEO) module makes use of 20 fiducial control points for Thin-Plate Spline (TPS) transformation, while the Pixel-Level Rectification (PIX) module predicts pixel-wise offset fields to rectify local misalignments. Optimization is done using the Adam optimizer and an initial learning rate of 1×10^{-3}.

The learning rate is decayed at regular intervals during training, and manual tuning is performed after every 5 epochs to ensure stable convergence. Training is done for a total of 10 epochs on a single NVIDIA RTX 3060 GPU with 12GB VRAM. Sequence-level cross-entropy loss is used as the objective, comparing the characterized predicted sequences against the ground-truth labels.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

RESULTS AND ANALYSIS

In this section, we present the experimental evaluation of our Two-Level Rectification Attention Network (TRAN).

Although the model is primarily trained and evaluated on the IIITSK-Word V3.0 dataset, we also compare its performance against results reported on widely used public scene text recognition benchmarks, including **IIIT5K**, **SVT**, **ICDAR2013 (IC13)**, **ICDAR2015 (IC15)**, **SVT-Perspective (SVTP)**, and **CUTE80**.

We report recognition accuracies, analyze the impact of each model component through ablation studies, and provide qualitative visualizations of rectification results.

A. Recognition Performance

Table I summarizes the word-level recognition accuracy of our method compared to other state-of-the-art models across several standard datasets. On all major benchmarks, our two-stage rectification framework achieves competitive or superior performance compared to ASTER and MORAN, especially under challenging distortions like perspective and curvature. Our model achieves particularly strong results on IIITSK and ICDAR datasets, highlighting its robustness against both regular and irregular scene texts.

Table I: Recognition Accuracy Comparison Across Datasets

Dataset	**ASTER (%)**	**MORAN (%)**	**Ours (TRAN) (%)**
IIIT5K (3000 images)	93.4	94.3	**94.7**
SVT (647 images)	89.5	90.0	**91.2**
ICDAR2013 (1015 images)	91.8	92.4	**93.1**
ICDAR2015 (2077 images)	76.1	77.4	**79.5**
SVT-Perspective (645 images)	73.9	74.0	**76.2**
CUTE80 (288 images)	79.5	82.7	**82.0**
IIITSK-Word V3.0 (ours)	-	-	**91.23**

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

B. Comparison with Single-Level Rectification Models

We further evaluate the contribution of the two-stage rectification strategy by comparing our full TRAN model against models employing only geometry-level (GEO) or pixel-level (PIX) rectification individually.

Table II: Comparison with Single-Level Rectification

Method	Accuracy on IIITSK (%)
GEO Only	87.32
PIX Only	88.15
TRAN (GEO + PIX)	**91.23**

The combination of global TPS warping followed by local pixel refinement outperforms each single-level rectification, confirming the complementary strengths of both approaches.

C. Ablation Study on CKUnit

We also assess the effectiveness of the Channel-Kernel Attention Unit (CKUnit) by conducting ablation experiments.

Table III: Ablation Study on CKUnit

Configuration	Accuracy on IIITSK (%)
TRAN without CKUnit	89.45
TRAN with CKUnit	**91.23**

Removing the CKUnit leads to a 1.78% drop in performance, demonstrating its contribution to improving feature extraction by adapting receptive fields and enhancing channel-wise correlations.

E. Limitations

Despite strong performance across datasets, certain limitations persist:

1. Highly curved texts, especially those in CUTE80, remain partially challenging due to the limited expressiveness of TPS transformations.
2. Pixel-level rectification may introduce minor blurring artifacts when large pixel offsets are applied.
3. The model does not explicitly leverage character-structure supervision, which could further improve rectification quality.

Future work could explore integrating semantic text understanding modules or learning more flexible, non-rigid deformation fields to address these challenges.

CONCLUSION AND ENHANCEMENTS

In this paper, we introduced a new architecture, Dual-Stage Text Rectification with Channel-Kernel Attention, to efficiently address the difficulties posed in identifying distorted scene text. Our approach uses a dual-stage rectification technique—initially using geometric correction with Thin-Plate Spline transformations and subsequent pixel-level rectification to locally correct local distortions—thus facilitating effective normalization of a very large range of real-world distorted text. Addition of a Channel-Kernel Attention Unit greatly enhances the network's ability to dynamically assign higher weight to critical spatial and semantic information, resulting in significantly better recognition effectiveness over state-of-the-art baselines. Extensive experimentation on benchmarking datasets shows that our approach exhibits excellent generalization ability to regular and highly distorted test samples. In the future, we intend to integrate this architecture with semantic structure prediction, enable multi-language recognition, and use flexible deformation models like deformable attention, thus enhancing the system's ability to handle complicated variations of scene text in diverse applications.

REFERENCES

[1] T. Quy Phan, P. Shivakumara, S. Tian, and C. Lim Tan, “Recognizing text with perspective distortion in natural scenes,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 569–576.

[2] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, “FOTS: Fast oriented text spotting with a unified network,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5676–5685.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

[3] W. Liu, C. Chen, K.-Y. K. Wong, Z. Su, and J. Han, "Star-Net: A spatial attention residue network for scene text recognition," in *Proc. Brit. Mach. Vis. Conf.*, vol. 2, 2016, pp. 43.1–43.13.
[4] 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.*, 2018, pp. 5571–5579.
[5] 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.*, 2019, pp. 8610–8617.
[6] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
[7] 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.
[8] W. Liu, C. Chen, and K.-Y. K. Wong, "Char-Net: A character-aware neural network for distorted scene text recognition." in *Proc. AAAI Conf. Artif. Intell.*, 2018, Art. no. 71547162.
[9] F. Zhan and S. Lu, "ESIR: End-to-end scene text recognition via iterative image rectification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2054–2063.
[10] 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.
[11] 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, 2019.
[12] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 6, pp. 567–585, Jun. 1989.
[13] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 770–779.
[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[15] C. Yao, X. Bai, B. Shi, and W. Liu, "Strokelets: A learned multi-scale representation for scene text recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4042–4049.
[16] O. Alsharif and J. Pineau, "End-to-end text recognition with hybrid HMM maxout models," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–9.
[17] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2963–2970.
[18] C. Yao, X. Bai, and W. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4737–4749, Nov. 2014.
[19] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1083–1090.
[20] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, 2004.
[21] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Proc. Asian Conf. Comput. Vis.*, Berlin, Heidelberg: Springer, 2010, pp. 770–783.
[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks" in *Proc. Adv. Neural Inf. Process. Syst*., 2012, pp. 1097–1105.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[24] C. Huang, Z. Wu, J. Wen, Y. Xu, Q. Jiang, and Y. Wang, "Abnormal event detection using deep contrastive learning for intelligent video surveillance system," *IEEE Trans. Ind. Informat.*, early access, Oct. 27, 2021, doi: 10.1109/TII.2021.3122801.
[25] C. Tian *et al.*, "Coarse-to-fine CNN for image super-resolution," *IEEE Trans. Multimedia*, vol. 23, pp. 1489–1502, 2021.
[26] J. Wen, Z. Zhang, Z. Zhang, L. Fei, and M. Wang, "Generalized incomplete multiview clustering with flexible locality structure diffusion," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 101–114, Jan. 2021.
[27] X. Yun, Y. Zhang, F. Yin, and C. Liu, "Instance GNN: A learning framework for joint symbol segmentation and recognition in online handwritten diagrams," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2021.3087000.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

[28] H. Ren, W. Wang, and C. Liu, "Recognizing online handwritten chinese characters using RNNs with new computing architectures," *Pattern Recognit.*, vol. 93, pp. 179–192, 2019.
[29] C. Zhang, Q. Zhao, C. P. Chen, and W. Liu, "Deep compression of probabilistic graphical networks," *Pattern Recognit.*, vol. 96, 2019, Art. no. 106979.
[30] P. Dai, H. Zhang, and X. Cao, "Deep multi-scale context aware feature aggregation for curved scene text detection," *IEEE Trans. Multimedia*, vol. 22, pp. 1969–1984, 2021.
[31] X. Wu *et al.*, "LCSegNet: An efficient semantic segmentation network for large-scale complex chinese character recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 3427–3440, 2021.
[32] 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, pp. 1–10.
[33] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in*Proc. Eur. Conf. Comput. Vis.*, Cham: Springer, 2014, pp. 512–528.
[34] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 161–184, 2021.
[35] X. Chen, L. Jin, Y. Zhu, C. Luo, and T. Wang, "Text recognition in the wild: A survey," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 42:1–42:35, 2021.
[36] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 369–376.
[37] P. He, W. Huang, Y. Qiao, C. C. Loy, and X. Tang, "Reading scene text in deep convolutional sequences," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 3501–3508.
[38] B. Su and S. Lu, "Accurate scene text recognition based on recurrent neural network," in *Proc. Asian Conf. Comput. Vis.*, Cham: Springer, 2014, pp. 35–48.
[39] Y. Gao, Y. Chen, J. Wang, and H. Lu, "Reading scene text with attention convolutional sequence modeling," *Neurocomput.*, vol. 339, pp. 161–170, 2019.
[40] 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.*, 2020, pp. 11005–11012.
[41] 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*, 2017, pp. 5076–5084.
[42] Y. Huang, Z. Sun, L. Jin, and C. Luo, "EPAN: Effective parts attention network for scene text recognition," *Neurocomputing*, vol. 376, pp. 202–213, 2020.
[43] 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.*, 2020, pp. 13528–13537.
[44] T. Wang *et al.*, "Decoupled attention network for text recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12216–12224.
[45] Y. Gao, Y. Chen, J. Wang, and H. Lu, "Progressive rectification network for irregular text recognition," *Sci. China Inf. Sci.*, vol. 63, no. 2, 2020, Art. no. 120101.
[46] C.-Y. Lee and S. Osindero, "Recursive recurrent nets with attention modeling for OCR in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2231–2239.