https://doi.org/10.2352/EI.2024.36.8.IMAGE-242

Generalizing Handwriting and Scene-Text Detection in Images*

***Taewook Kim*†, *Gaurav Patel*†, *Qian Lin*‡, *Jan P. Allebach*†, *and Qiang Qiu*†**
†***Purdue University***
‡***HP Inc***

Abstract

In this paper, we present a deep-learning approach that unifies handwriting and scene-text detection in images. Specifically, we adopt adversarial domain generalization to improve text detection across different domains and extend the conventional dice loss to provide extra training guidance. Furthermore, we build a new benchmark dataset that comprehensively captures various handwritten and scene text scenarios in images. Our extensive experimental results demonstrate the effectiveness of our approach in generalizing detection across both handwriting and scene text.

Introduction

Image-based text detection is one essential task with a wide range of applications. Its importance lies in its ability to extract and locate textual information within images and videos. Many critical downstream tasks can then be enabled, such as text recognition [20], document understanding [1], and text image generation [2].

Building upon the significance of text detection, our objective is to develop a robust model capable of accurately localizing both handwritten and printed text within real-world contexts. Text detection has been intensively studied in the scene text domain, while it remains to be explored further in the handwritten domain, and even less attention has been provided to detect both scene text and handwriting simultaneously. Hence, we develop a unified text detection model that is generalizable across both handwriting and scene text domains.

To improve cross-domain generalization, we adopt an adversarial learning scheme [5]. This learning process involves a discriminator and a backbone network, where the discriminator classifies the domain of features extracted from the backbone network by minimizing the domain classification loss. The backbone network, on the other hand, is trained to maximize the domain classification loss. In this way, the backbone network is learned to extract domain-invariant features as it is encouraged to extract features that are not distinguishable by the domain classifier. We further improve the model's detection capabilities by devising additional loss functions and synthesizing handwriting images to augment training data. We collect a new benchmark dataset that reflects real-world scenarios. By improving model cross-domain generalization, we observe a significant increase in detection accuracy on both handwriting and scene text domains.

Related Works

There has been significant recent exploration of deep convolutional network-based techniques within the field of text detection. These approaches can be broadly categorized into two groups: 1) methods reliant on anchors, and 2) methods that operate without anchors. In the following section, we provide a brief overview of the previous methods within each of these categories.

Anchor-based text detectors are the approaches that are extensions of standard region-based object detectors [21, 15]. Based on Faster R-CNN [21], [17] proposed RRPN that is capable of generating inclined proposals with angle information that fits more accurately to the ground truth text regions. TextBoxes [12] is a fully-convolutional text detection model that extends SSD [15] with different designs of default box and different convolutional kernel sizes to be more applicable to the word aspect ratios that are different from the general objects. The methods mentioned above have demonstrated impressive improvement from the existing methods, however, a significant drawback is that the majority of these methods depend on predefined anchor designs, and are inherently limited in detecting texts in arbitrary shapes.

Anchor-free text detectors formulate text detection as a text segmentation problem. EAST [25] proposed a pipeline that adopts fully convolutional networks (FCN) [16] to directly predict text regions to eliminate redundant steps. [11] proposes PSENet which adopts the FPN [13] structure to generate fused feature maps of various scales, then expand predicted kernels to obtain final text detection results via progressive scale expansion algorithm.

Unlike anchor-based detectors, anchor-free detectors are more capable of detecting texts of arbitrary shapes, but most of the approaches suffer from heavy overhead. To overcome this issue, [23] proposes an efficient segmentation-based framework that can detect texts in arbitrary shapes, with little decrease in performance. Considering the importance of the model being efficient to be deployed in real-world scenarios, we chose [23] as our baseline model.

Proposed Approach

Our detection model is built from Pixel Aggregation Network (PAN) [23], which is an efficient and accurate arbitrary-shaped text detector equipped with a low computational-cost segmentation head and learnable post-processing. To generalize the detection capability across both scene text and handwriting images, we deploy an adversarial learning scheme [5] to encourage the backbone network to extract domain-invariant features from images. Furthermore, we devise a new loss function, called *inverted dice loss*, which further boosts the performance of the detection network. We first briefly summarize our adopted baseline model [23].

Baseline Model

PAN [23] is a segmentation-based model that enables text detection in arbitrary shape. It consists of a lightweight backbone

*Research supported by HP Inc, Palo Alto, CA

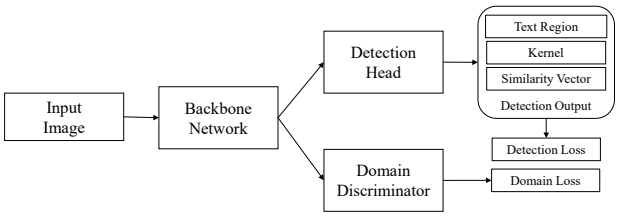

Figure 1. *Overview of the proposed model. In addition to the detection loss, our model is trained with a domain classification loss, which significantly improves generalization capability across different domains. Additionally, we incorporate the inverted dice loss, which further augments our training process by providing valuable guidance.*

network (ResNet-18 [7]), and a segmentation head comprised of a Feature Pyramid Enhancement Module (FPEM) and Feature Fusion Module (FFM). FPEM follows the backbone network and is a cascaded U-shaped module. It consists of upscale and downscale enhancement processes that improve the representation power by combining feature maps in different scales. FFM is a module following the FPEM, that combines the feature pyramids at different cascading stages. To avoid the text lying close to each other to be recognized as the same text instance, PAN [23] proposes a learnable pixel aggregation algorithm as post-processing. The detection head of PAN not only predicts the text region, but also the kernels that act as a cluster center of each text instance (Fig. 1). Starting from the predicted kernel region, it progressively expands up to the predicted text region. To ensure that the text region and the kernel of the same instances are aggregated, the following loss is deployed [23],

$$\mathscr{L}_{agg} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|T_i|}\sum_{p\in T_i}\log(Dist_{pk}(p,K_i)+1), \tag{1}$$

where T_i is the i^{th} text instance, N is the batch size, $Dist_{pk}$ is the distance between the text pixel and the kernel of the same text instance. This is defined as follows [23],

$$Dist_{pk}(p,K_i) = \max(||S_p(p)-S_k(K_i)||-\delta_{agg},0)^2, \tag{2}$$

where δ_{agg} is a constant set to 0.5, and S_p is the similarity vector of pixel p, and S_k is the similarity vector of kernel K_i. The similarity vector of kernel K_i is defined as [23],

$$S_k(K_i) = \sum_{p\in K_i} S_p(p)/|K_i|. \tag{3}$$

To keep the kernel of different text instances to maintain distance from each other, kernel discrimination loss $\mathscr{L}_{dis}$ is also defined [23],

$$\mathscr{L}_{dis} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\log(Dist_{kk}(K_i,K_j)+1), \tag{4}$$

$$Dist_{kk}(K_i,K_j) = \max(\delta_{dis}-||S_k(K_i,S_k(K_j)||,0)), \tag{5}$$

where $Dist_{kk}$ is the distance between kernels from different text instances, and δ_{dis} is a constant set to 3. Also, the text loss $\mathscr{L}_{text}$, and the kernel loss $\mathscr{L}_{ker}$ are formulated as [23],

$$\mathscr{L}_{text} = 1-\frac{2\sum_i \hat{Y}_t(i)Y_t(i)}{\sum_i \hat{Y}_t(i)^2+\sum_i Y_t(i)^2}, \tag{6}$$

$$\mathscr{L}_{ker} = 1-\frac{2\sum_i \hat{Y}_k(i)Y_k(i)}{\sum_i \hat{Y}_k(i)^2+\sum_i Y_k(i)^2}, \tag{7}$$

where $\hat{Y}_t$ and $\hat{Y}_k$ are the predicted text segmentation results and kernel segmentation results of i^{th} pixel, respectively, and Y_t and Y_k are the ground truth. The two losses are originally from [19]. Here, following [11], the ground truth of the kernels is generated by proportionally shrinking the polygon spanning the text region by the ratio of r.

Text Detection Improvement

Most of the existing text detection works have been focused on improving the performance of either scene-text or handwriting-only, while less attention has been provided to detecting both the scene-text and handwriting well. In this work, we design a way to generalize on both scene text and handwriting domains. Specifically, we adopt an adversarial domain generalization learning scheme [5] so that the model can extract domain-invariant features that are useful for both scene-text and handwriting detection. Specifically, we deploy a domain discriminator that tries to distinguish the domain of input images by minimizing the domain classification loss, while the encoder is adversarially optimized to extract domain-invariant features by maximizing the domain classification loss. We empirically observe significant model performance improvement after adopting the adversarial learning.

Figure 2. *The gradient reversal layer (GRL) of the proposed model. The GRL layer is inserted between the backbone network and the discriminator network. During backpropagation, we multiply the gradient from the discriminator by -1 to reverse the sign of the gradient. In this way, the backbone network is trained to maximize the loss of the domain loss, which results in domain invariant feature extraction of the backbone network, and improved generalization capability.*

Domain Generalization for Text Detection. Existing text detection methods typically assume that training and test data are sampled from similar distributions. When this assumption no longer holds, detection models may fail to produce reliable predictions. This is due to the *shift* in data distribution. In Domain Generalization (DG), several different but related domain(s) are given to be generalized to unseen target distribution by minimizing the shifts among given domains. To achieve this, we deploy a domain discriminator that is trained to minimize the domain classification error. To encourage the backbone network to extract domain-invariant features, a gradient reverse layer (GRL) [5] is inserted between the backbone network and the discriminator (Fig. 2), to reverse the sign of the gradient while backpropagating the gradient originated from the *domain classification loss*, which is formulated as,

$$\mathscr{L}_{domain} = -\mathbb{E}\left[\sum_{d\in\mathscr{D}}\mathbf{1}_{[d=k]}\log\sigma(Disc(Enc(I^k))\right], \tag{8}$$

where $\mathscr{D}$ is the domains, *Disc*, *Enc* denote discriminator and backbone network, respectively, and $\mathbf{1}$ is an indicator function which equals 1 if Image I^k is from domain $d = k$, and σ is the softmax function.

By reversing the gradient, the backbone network is trained to maximize this loss while the discriminator is trained to minimize it simultaneously; hence, the discriminator and the backbone network are adversarial to each other. As the training progresses, we empirically observe that the discriminator fails to minimize the domain loss, implying that the backbone network succeeded in extracting features that the domain of features could not be discerned by the discriminator.

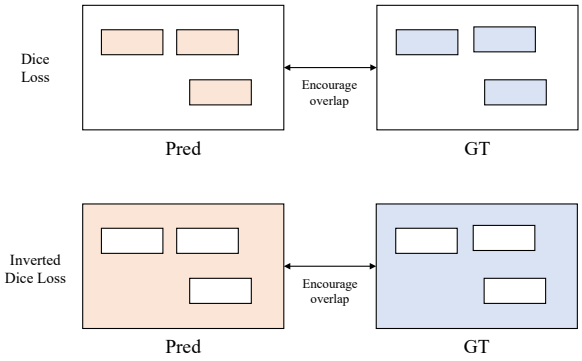

Figure 3. *Conceptual difference between the conventional dice loss [19] and the inverted dice loss.*

Inverted Dice Loss. The adopted dice loss (Eqn. 6) encourages a high overlap of the predicted foreground region with the ground truth. In this work, we introduce an additional training guidance by extending the dice loss. Specifically, we train the model to predict the ground truth background region as well. This can be done by simply inverting the ground truth foreground text region (Fig. 3). The inverted dice loss is formulated as follows,

$$\mathscr{L}_{inv} = 1 - \frac{2\sum_i \hat{Y}_b(i) Y_b(i)}{\sum_i \hat{Y}_b(i)^2 + \sum_i Y_b(i)^2}, \tag{9}$$

where $\hat{Y}_b$ and Y_b are the predicted segmentation of background and ground truth background.

Overall Training Objective Function

Our model is trained in a multi-task learning fashion where the overall training objective is formulated as follows:

$$\min_{Enc,Head} \mathscr{L}_{text} + \alpha\mathscr{L}_{ker} + \beta(\mathscr{L}_{agg} + \mathscr{L}_{dis}) + \mathscr{L}_{inv}, \tag{10}$$

$$\min_{Disc} \max_{Enc} \gamma\mathscr{L}_{domain}, \tag{11}$$

where α, β, and γ are the hyperparameters, and Enc and $Head$ denote the backbone network and the detection head, respectively. Here the encoder and the discriminator are optimized adversarially, where the encoder tries to fool the discriminator by maximizing the domain loss. As the training progresses, we observe that the discriminator fails to minimize the domain loss, and this can be interpreted as the encoder succeeding in generating domain invariant loss.

Handwriting Data Synthesis

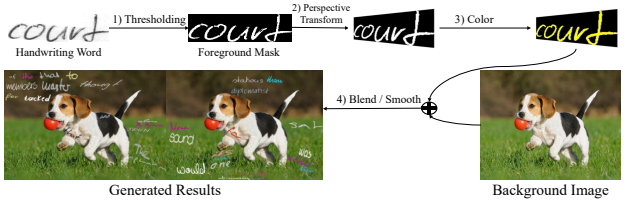

Figure 4. *The handwriting dataset synthesis process.*

The number of publicly available data on handwriting images is relatively scarce compared to the scene text images. In this regard, we augment our handwriting training data by synthesizing them (Fig. 4). The overall process is done by rendering the text on text-free background images [6]. Given a handwriting word image [18], we first threshold the word image to obtain the foreground mask image where the threshold is empirically determined here. After obtaining the mask, we color the foreground mask in random colors, then we perform the perspective transform on the foreground mask to be blended into text-free background images.

Specifically, we compute the transformation matrix between the coordinates of the rendered word in the synthetic scene-text image and the colored foreground mask, then apply the transformation matrix to blend the foreground mask into a text-free background image. Note that [6] provides the text-free background version of the text-rendered images and the coordinates of the rendered words. We re-use the coordinates of the rendered scene-text words and place the handwriting words by performing perspective transform. The overall process can be considered as a replacement of the scene text words with the handwriting words.

Experiments

Implementation Details

We mostly follow the same hyperparameter settings from [23]. We set α, β, and γ as 0.5, 0.25 and 0.05, respectively. ResNet [7] pre-trained on ImageNet [4] is adopted as the backbone. The dimension of the similarity vector is set to 4. We use Adam Optimizer [9] to train our model, with the learning rate of 10^{-3} for 50,000 iterations on a single GPU. The batch size is set to 13. We use GeForce RTX 2080 Ti for every experiment. We set the kernel shrinkage ratio r as 0.5. The input images are resized to a width of 640 pixels on the shorter side while maintaining their original aspect ratio by proportionally adjusting the longer side.

Datasets

Training Dataset. We use a merged set of the publicly available datasets, ST [6], GNHK [10], COCO [14], IC15 [8], and the synthesized handwriting datasets. We use the text-free background image provided from ST [6], and the foreground word images are obtained from IAM [18]. We consider each image set as an element of domain set $\mathscr{D} = \{ST, GNHK, COCO, IC15, SynthHand\}$. *SynthHand* here denotes the synthesized handwriting dataset. Thus, the domain classifier is a five-way classifier.

HP-Open Course Ware Dataset (HP-OCW). To evaluate the performance of our system, we manually collect and annotate

them using the open-source annotation tool, LabelMe [22]. We collected 433 videos from various institutes that provide YouTube video lectures. Among them, we extracted 10 consecutive frames from 50 video clips and manually annotated them which amounts to 500 images. In this dataset, both the handwriting and scene-text words are included.

HP-Notepad Whiteboard Dataset (HP-NW) . We use randomly sampled 50 handwriting images from the test set of GNHK [10], and we collect an additional 51 images from the web and YouTube videos, where the handwritten words are written on either the whiteboard or notepad (Fig. 5).

Table 1. Comparison of the baseline with the existing methods.

Method	Precision	Recall	F-measure	FPS
CTPN	60.4	53.8	56.9	7.14
SegLink	42.3	40.0	40.8	10.7
EAST	78.7	49.1	60.4	21.2
CTD+TLOC	77.4	69.8	73.4	13.3
PSENet-1s	80.6	75.6	78.8	3.9
PCR	79.8	85.3	82.4	11.4
PAN (baseline)	84.6	77.7	81.0	39.8

Experiement Results of the Proposed Model

Comparison of baseline model with the existing methods. We first compare the performance of the baseline model on a recent challenging dataset for curve text detection dataset, CTW-1500 [24] with the existing state-of-the-art models that were originally reported in the paper [23] (Table 1). Compared to the best model in the table, PCR [3], our model achieved a slightly lower F1 score by 1 percentage point (p.p). However, it's important to note that our model excels in terms of efficiency, running approximately 3.5 times faster than PCR 1.

Evaluation Results on HP-OCW. We provide the evaluation results of our model on the manually collected HP-OCW dataset (Table 2). By adopting the domain generalization on our model, we could achieve a 4.77 p.p increase in the F1 score from the baseline.

Evaluation Results on HP-NW. We provide the evaluation results of our model on the manually collected HP-NW dataset (Table 3). By adopting the domain generalization on our model, we could achieve a 2.82 p.p increase in F1 score, and by adopting inverted dice loss we could achieve an additional increase of 0.73 p.p.

Table 2. Evaluation results on HP-OCW Dataset. DG denotes domain generalization.

Method	Precision	Recall	F1
Baseline	84.59	56.68	67.87
Baseline+DG	82.88	64.65	72.64

Table 3. Evaluation results on HP-NW dataset. DG denotes domain generalization, and ID denotes the inverted dice loss.

Method	Precision	Recall	F1
Baseline	90.70	85.29	87.91
Baseline+DG	92.71	88.83	90.73
Baseline+DG+ID	92.86	90.20	91.51

Table 4. Evaluation results on IC15 dataset. DG denotes domain generalization, and ID denotes the inverted dice loss.

Method	Precision	Recall	F1
Baseline	75.43	64.76	69.69
Baseline+DG	83.93	66.15	73.99
Baseline+DG+ID	84.68	68.13	75.51

Evaluation Results on IC15. We provide the evaluation results of our model on IC15 (Table 4). By adopting the domain generalization on our model, the model achieves a 4.30 p.p increase in F1 score, and by adopting both the domain generalization and the inverted dice loss, an additional improvement of 1.52 p.p could be achieved.

Advantage of Adversarial Domain Generalization. In general, the model achieved the improvements when domain generalization is applied to the baseline: achieving 4.77 p.p, 2.82 p.p, and 4.30 p.p improvements of F1 score from the baseline when evaluated on the HP-OCW dataset, HP-NW dataset, and IC15 dataset, respectively. Notably, the HP-NW dataset consists mostly of handwritten words, whereas the IC15 dataset consists mostly of scene-text images, and improvements were achieved in both domains.

Advantage of Inverted Dice Loss. The proposed inverted dice loss is advantageous as it further improves detection performance. When tested on the HP-NW dataset, we could achieve a 0.73 p.p improvement in F1 score when inverted dice loss is adopted on top of the model with the domain generalization (Table 3). When tested on the IC15 dataset, we could achieve 1.23 p.p improvement in F1 score when inverted dice loss is adopted on top of the domain generalization model (Table 4). This general trend in improvement validates the effectiveness of our proposed inverted dice loss.

Visualization Results

We provide the qualitative detection results of the proposed model on the HP-OCW dataset HP-NW, and IC15 [8] dataset (Fig. 5). The results demonstrate that our model is capable of detecting both the handwriting and scene (printed) text in real-world scenarios.

Conclusion

In this paper, we proposed a text detection method that can generalize to both handwriting and scene text images. Our model adopts PAN as a baseline, and we further improve this by deploying a domain discriminator. The deployed discriminator discriminates the domain of an image originated from. We train the backbone network to maximize the discriminator loss that results in domain-invariant learning. We evaluate the performance of the

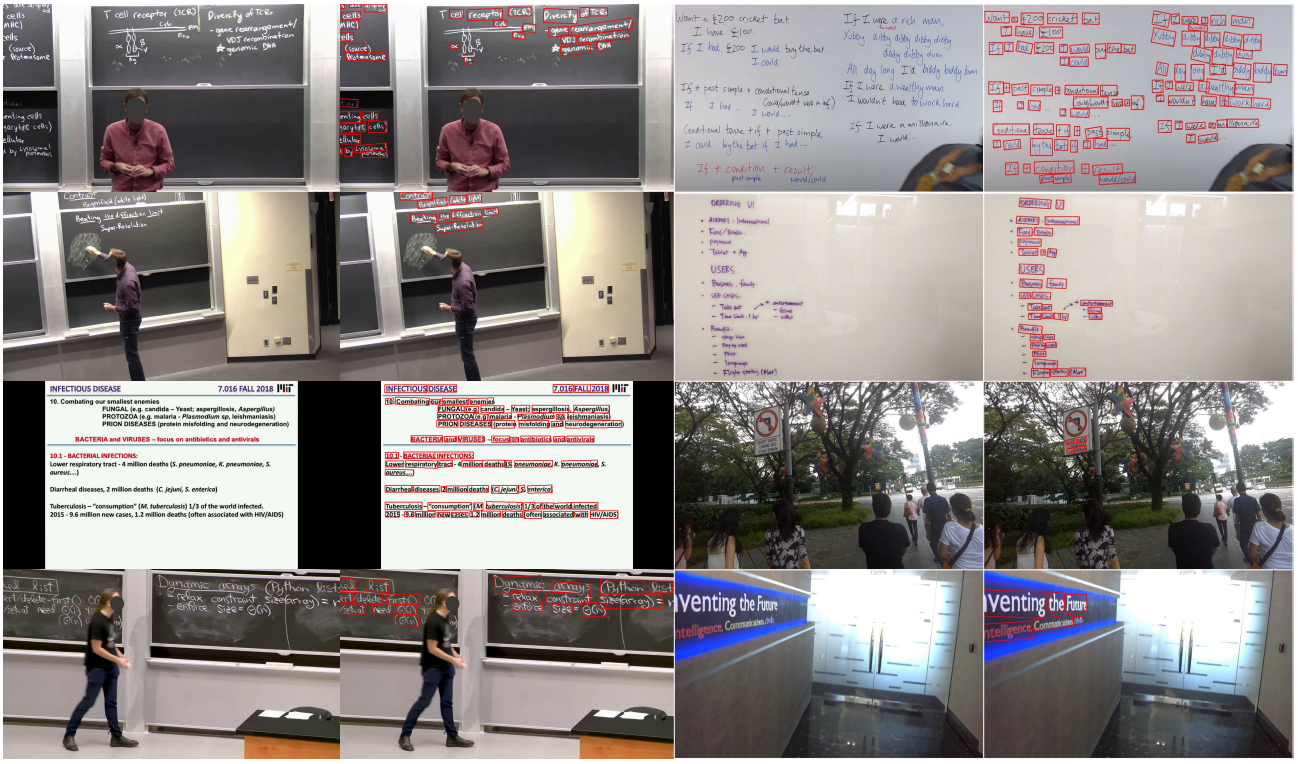

Figure 5. *The detection results on the HP-OCW, HP-NW, and IC15 datasets. The left two columns are from HP-OCW, and the right two columns are images of HP-NW and IC15. The odd columns are input images and the even columns are the detected results. The top two rows of the right two columns are images from HP-NW and the bottom two rows are from the IC15 dataset.*

proposed method in diverse settings, and the results validate the effectiveness of our method.

References

[1] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[2] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. In *Advances in Neural Information Processing Systems*, 2023.

[3] Pengwen Dai, Sanyi Zhang, Hua Zhang, and Xiaochun Cao. Progressive contour regression for arbitrary-shape scene text detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.

[5] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 2015.

[6] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[8] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. Icdar 2015 competition on robust reading. In *International Conference on Document Analysis and Recognition*, 2015.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] Alex W. C. Lee, Jonathan Chung, and Marco Lee. Gnhk: A dataset for english handwriting in the wild. In *International Conference of Document Analysis and Recognition*, 2021.

[11] Xiang Li, Wenhai Wang, Wenbo Hou, Ruo-Ze Liu, Tong Lu, and Jian Yang. Shape robust text detection with progressive scale expansion network. *arXiv preprint arXiv:1806.02559*, 2018.

[12] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *AAAI Conference on Artificial Intelligence*, 2017.

[13] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 2014.

[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vi-*

sion, 2016.

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

[17] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *Transactions on Multimedia*, 2018.

[18] Urs-Viktor Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5:39–46, 2002.

[19] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International conference on 3D vision*, 2016.

[20] Gaurav Patel, Taewook Kim, Qian Lin, Jan P. Allebach, and Qiang Qiu. Self-attention enhanced recognition: A unified model for handwriting and scene-text recognition with improved inference. In *Electronic Imaging*, 2024.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 2015.

[22] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173, 2008.

[23] Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wenjia Wang, Tong Lu, Gang Yu, and Chunhua Shen. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *International Conference on Computer Vision*, 2019.

[24] Liu Yuliang, Jin Lianwen, Zhang Shuaitao, and Zhang Sheng. Detecting curve text in the wild: New dataset and new solution. *arXiv preprint arXiv:1712.02170*, 2017.

[25] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

Author Biography

Taewook Kim *is a Ph.D. student in Electrical and Computer Engineering at Purdue University, where he focuses on research in the fields of computer vision and machine learning. He earned his Bachelor's degree in Computer Engineering from Hongik University in South Korea, and his Master's degree in Computer Science and Engineering from Pohang University of Science and Technology (POSTECH) in South Korea.*

Gaurav Patel *received his Bachelor of Technology (B.Tech.) degree in Electronics and Communication Engineering from National Institute of Technology Raipur, India in 2020. He is currently pursuing his Ph.D. in the Elmore Family School of Electrical and Computer Engineering at Purdue University. His research interest includes Computer Vision, Machine Learning (Deep Learning), and Artificial Intelligence.*

Qian Lin *is an HP Fellow working on computer vision and deep learning research. She is also an adjunct professor at Purdue University. She joined Hewlett-Packard Company in 1992. She received her BS from Xi'an Jiaotong University in China, her MSEE from Purdue University, and her Ph.D. in Electrical Engineering from Stanford University. She is an inventor/co-inventor for 45 issued patents. She was awarded Fellowship by the Society of Imaging Science and Technology (IS&T) in 2012, Outstanding Electrical Engineer by the School of Electrical and Computer Engineering of Purdue University in 2013, and the Society of Women Engineers Achievement Award in 2021.*

Jan P. Allebach *is Hewlett-Packard Distinguished Professor of Electrical and Computer Engineering at Purdue University. He was named Electronic Imaging Scientist of the Year by IS&T and SPIE, and was named Honorary Member of IS&T, the highest award that IS&T bestows. He has received the IEEE Daniel E. Noble Award, the IS&T/OSA Edwin Land Medal, the IS&T Johann Gutenberg Prize, is a Fellow of the National Academy of Inventors, and is a member of the National Academy of Engineering.*

Qiang Qiu *is a researcher with expertise in computer vision and machine learning, focusing on deep learning, image understanding, and representation learning. He holds a Ph.D. in Computer Science from the University of Maryland, College Park. He currently serves as an Assistant Professor at Purdue University and has previously worked as an Assistant Research Professor at Duke University.*