

Received May 1, 2022, accepted May 19, 2022, date of publication May 27, 2022, date of current version June 3, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3178597

RSTC: A New Residual Swin Transformer for Offline Word-Level Writer Identification

PEIRONG ZHANG 

College of Electronic and Information Engineering, South China University of Technology, Guangzhou 510041, China

e-mail: eeprzhang@mail.scut.edu.cn


This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61936003, and in part by the Natural Science Foundation of Guangdong Province (GD-NSF) under Grant 2017A030312006.

ABSTRACT Writer identification has steadily progressed in recent decades owing to its widespread application. Scenarios with extensive handwriting data such as page-level or sentence-level have achieved satisfactory accuracy; however, word-level offline writer identification is still challenging owing to the difficulty of learning good feature representations with scant handwriting data. This paper proposes a new Residual Swin Transformer Classifier (RSTC), which comprehensively aggregates local and global handwriting styles and yields robust feature representations with single-word images. The local information is modeled by the Transformer Block through interacting strokes and global information is featurized by holistic encoding using the Identity Branch and Global Block. Moreover, the pre-training technique is exploited to transfer reusable knowledge learned from a task similar to writer identification, strengthening RSTC's representation of handwriting features. The proposed method is tested on the IAM and CVL benchmark datasets and achieves state-of-the-art performance, which demonstrates the superior modeling capability of RSTC for word-level writer identification.

INDEX TERMS Writer identification, handwriting analysis, vision transformer, deep learning.

I. INTRODUCTION

Handwriting data is a type of behavioral biometric analogous to iris and fingerprint, with distinct personal characteristics that allow persons to be identified. It has been profoundly deployed in many fields, such as security and law enforcement. Writer identification is defined as finding an individual writer in a large dataset, which stems from the broader domain of handwriting recognition. According to the data acquisition manner, this task can be categorized as online and offline [1], [2]. For online writer identification, it refers to recording temporal data produced in writing procedures, such as coordinates, angles, and pressures using specific devices. For offline writer identification, merely static handwriting images are recorded and analyzed. Consequently, identifying writers in offline scenarios is considered more challenging with fewer features input to the identification system. In addition, depending on the handwriting content, writer identification can be dichotomized as text-dependent and text-independent [2]. The former means that each writer must write the same content; the latter, which is considered

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



(a) Several examples words in IAM dataset of different writers.



(b) Several examples words in CVL dataset of different writers.

FIGURE 1. Word images that have a low amount of handwriting.

more difficult, denotes that the writer is allowed to write different content.

In most cases, we can only access a small amount of personal handwriting data, such as a few words or even letters, as depicted in Figure 1. It presents a tremendous challenge to identify a person as handwriting data is insufficient. Earlier studies widely exploited hand-crafted feature extractors, such as the vector of locally aggregated descriptors (VLAD) [3] and co-occurrence histogram [4], which capture local handwriting styles of the writer. On the other hand, convolutional neural networks (CNN) [5], [6] and recurrent neural networks (RNN) [7] were increasingly popular for deep local features

extraction. Furthermore, the combinations of hand-crafted feature extractors and deep neural networks [8]–[10] have also been well explored. As attention-based methods have recently become the mainstream, several studies [11]–[14] have focused on attention-based methods to enhance modeling in feature space, which are usually combined with CNN. The aforementioned methods have achieved promising performances in page-level or sentence-level scenarios. However, they did not deliver satisfactory results in offline word-level scenarios. The reasons are two-fold. (1) Most of the existing methods only concentrated on local features, but ignored the modeling of global features, which limited models' performance. (2) Owing to the limited handwriting data and sparse connections of neighboring strokes of words, hand-crafted feature extraction techniques and CNN/RNN were unable to encode enough local and global information to capture discriminative handwriting styles. Although attention-based methods slightly improve the accuracy, they only optimized the representation of local features without involving global dependency. Hence, existing methods failed to yield robust features for single-word images, resulting in poor performances.

Recently, the Vision Transformer (ViT) [15] has achieved considerable success in many fields of computer vision, such as object detection and image classification [15]. And various vision transformers [16]–[18] have been proposed. The transformer in vision fields reduces pixel information lost in long-range modeling using self-attention mechanism and captures shallow and deep features of images based on different architectures, which means that it could be a potential approach for improving the accuracy of offline word-level writer identification. Furthermore, the pre-training technique has been widely leveraged for its effectiveness in transferring knowledge, such as in the large pre-trained language models (LMs) [19], [20]. However, it has not been investigated in writer identification.

This paper proposes a new effective Residual Swin Transformer (RSTC) to explicitly handle the extraction of local and global features of single-word images in a comprehensive manner. To better model local fine-grained features such as letter strokes, the Transformer Block is introduced, which mines interconnections of strokes by the interaction of pixel tokens. To characterize global handwriting styles such as the spatial architecture of words, the Identity Branch and the Global Block are proposed. The former encodes the entire image through patch embedding to a feature vector and the latter adaptively incorporates all extracted local and global features, resulting in joint optimization of the Transformer Block and Identity Branch. In general, the proposed RSTC coordinates local and global feature modeling and forms robust final features although the handwriting data is limited. Additionally, the pre-training technique is exploited to facilitate RSTC's performance, where word recognition is adopted as the pre-training task to provide transferable knowledge.

In summary, the main contributions of this paper include:

- This paper proposes a new Residual Swin Transformer Classifier (RSTC) for offline word-level writer identification. To the best of our knowledge, this is the first paper that introduces the vision transformer-based model in the domain of writer identification.
- RSTC models fine-grained local features using the Transformer Block and global features using the proposed Identity Branch and Global Block, and then comprehensively aggregate them. Therefore, it effectively solves the difficulty of extracting robust features with a low amount of handwriting data.
- Extensive experiments on two widely used benchmarks demonstrate that the proposed method is superior to the existing approaches, indicating the effectiveness of RSTC.

The remainder of this paper is organized as follows. Section II presents a review of the related works of writer identification. Section III introduces the proposed methodologies. Section IV describes the implementation details and experimental results. Section V concludes the paper.

II. RELATED WORKS

Considerable research has been conducted in the field of writer identification, including online and offline methods. They generally involved three stages: data preprocessing, feature extraction, and writer decision. Based on different schemes of the latter two stages, the existing techniques can be grouped into hand-crafted feature-based and deep learning-based methods.

A. HAND-CRAFTED FEATURE-BASED METHODS

Before the advancement of deep learning, the conventional methods extracted hand-crafted features from handwriting images, and then combined them with decision systems to make classifications.

1) TEXTURAL FEATURES

Considering handwriting as texture [21], textural features have been widely exploited. Bertolini *et al.* [22] utilized the Local Binary Pattern (LBP) [23] and Local Phase Quantization (LPQ) [22] features of texture blocks and passed them to the SVM classifier. Newell and Griffin [24] applied the oriented Basic Image Feature (oBIF) Column scheme to encode handwriting images. They used a bank of six Derivative-of-Gaussian (DtG) filters to extract texture features as encoding representations. Said *et al.* [21] and Mridha *et al.* [10] leveraged the Gabor filter, which is a useful descriptor for texture extraction. Similar to Gabor filters, the XGabor filters were used in [25] with the feature relation graph (FRG).

2) ALLOGRAPHIC FEATURES

The allograph of handwritten characters contains features of their shapes. Bulacu *et al.* [26] used a segmentation method to extract allographic features by generating shape codebooks and computing probability distribution functions (PDF). Ruben *et al.* [27] captured the occurrence probability

of allographic elements of handwriting to distinguish writers. In [28], Niels *et al.* novelly proposed to extract allograph frequency vectors for each writer and matched writers by similarity.

3) CONTOUR-BASED FEATURES

In [29], K-adjacent segment features were used to model the character contours on the entire page. Brink *et al.* [30] extracted Quill-Hinge features from contours of the ink-trace to make representations. In [31], Lai *et al.* proposed Pathlet features, which extracted oriented fragments of handwriting contours using the Ramer-Douglas-Peucker algorithm and then described them with Path Signature or Log Path Signature features. In addition, they calculated the Scale Invariant Feature Transform (SIFT) features. The combination of Pathlet and SIFT features achieved excellent results on page-level writer identification.

B. DEEP LEARNING-BASED METHODS

Following the advancement of deep learning, neural network-based methods, such as CNN and RNN, have been widely employed for writer identification. They are not typically used alone, but combined with existing feature extraction techniques.

1) CNN-BASED

Christlein *et al.* [5] exploited CNN as the feature extractor and used Gaussian Mixture Models (GMM) Supervector encoding to aggregate the features. Yang *et al.* [8] developed a CNN architecture and used Path Signature to extract features. They also proposed a data-augmentation method called DropStroke and concluded that CNN needs a large amount of data to achieve reasonable performances. Xing and Qiao [6] proposed an end-to-end DCNN model. With multistream input and parameter sharing, their model leveraged the spatial features between patches and achieved good results when the input was a sentence or a Chinese character. Rehman *et al.* [32] investigated the effect of different frozen layers of CNN on the identification rate of writers on several datasets. In [10], Mridha *et al.* explored combining the thresholded-Gabor filter and CNN for Indic language writer identification in word-level scenarios. He and Schomaker [33] proposed a deep-adaptive learning model based on single-word images. They used a two-stream CNN, one for the main task and the other for the auxiliary task, which receives shared feature representations. Subsequently, they introduced FragNet [34] that also had two modeling paths. The first path used the feature pyramid to receive entire images as input and the second one predicted writers' identities through fragments of the images.

2) RNN-BASED

Liu *et al.* [7] employed LSTM and the Path Signature to capture sequential information of handwritings. He *et al.* [35] adopted the residual learning mechanism and introduced a residual RNN block that received the local feature map

computed by the CNN block. Additionally, they studied the effects of horizontal and vertical segmentation of feature maps.

3) ATTENTION-BASED

The attention mechanism was universally exploited in recent years. In writer identification, this technique was usually combined with CNN or RNN to enhance deep feature extractions. Shaikh *et al.* [14] employed cross attention and soft attention to concentrate on highly correlated pixel regions of handwriting pairs. Chen *et al.* [11] proposed the Letters and Styles Adapters (LSA) to encode different letters, which were inserted between CNN and LSTM. They also proposed Hierarchical Attention Pooling (HAP) to aggregate features. Ngo [13] proposed an attention key point filter to select key points in historical documents, which aimed to simulate the selection of the SIFT algorithm.

The aforementioned methods that focused on word-level scenarios used CNN or RNN to learn handwriting styles and perform identification. The results are unsatisfactory because they generally consider local features but ignored global features. Besides, limited handwriting data restricted models' performances. To this end, this paper proposes the Residual Swin Transformer Classifier (RSTC) to tackle these problems by jointly capturing local and global writing styles.

III. METHODOLOGY

This section first describes the design of the model in detail. Then, the pre-training technique is discussed. Finally, the data augmentation techniques and the loss function are presented.

A. RSTC: RESIDUAL SWIN TRANSFORMER CLASSIFIER

The overall architecture of the proposed RSTC model is shown in Figure 2. The model consists of (1) the Transformer Block, (2) the Identity Branch, and (3) the Global Block. The former aims to minutely model the local handwriting styles, and the latter two are responsible for capturing global handwriting styles and performing feature integration.

1) TRANSFORMER BLOCK

Since writers have their unique style of handwriting, the strokes of handwritten words share strong similarities. Hence, modeling the relationship between different strokes is a key factor in learning local handwriting styles. Considering the local modeling ability, the tiny version of the advanced Swin-Transformer [16] (Swin-T) is adopted as the Transformer Block. The core modules of Swin-Transformer are window multi-head self-attention (W-MSA) and shifted window multi-head self-attention (SW-MSA). W-MSA computes self-attention in nonoverlapping windows, which covers each part of the word and finely captures the features of adjacent strokes. SW-MSA re-partitions the image and intersects windows, which unearths mutual connections between neighboring parts of the word, effectively encoding remote features.

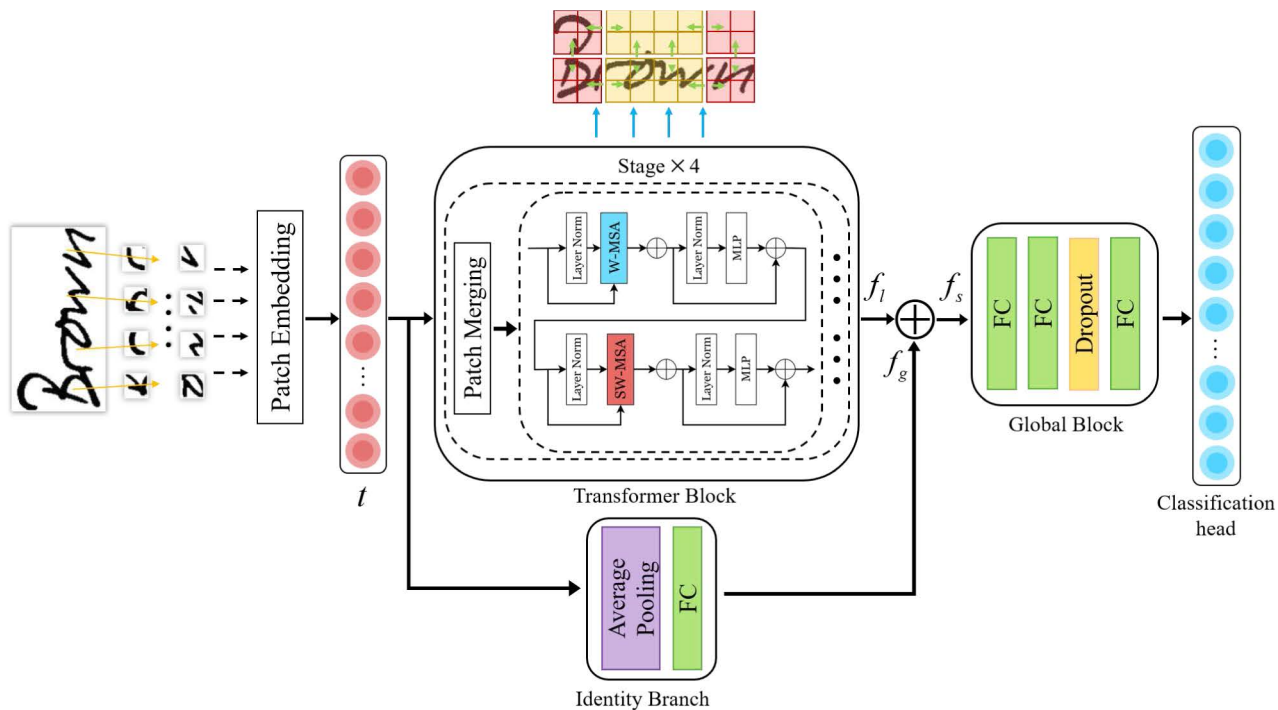


FIGURE 2. The overall architecture of the Residual Swin Transformer Classifier (RSTC). The Transformer Block captures local dependency of word images. The Identity branch adds an identity mapping of the original images to capture global dependency. All local and global features are aggregated by the Global Block.

As shown in Figure 2, the preprocessed single-word images are first converted into pixel tokens t through the patch embedding layer. Subsequently, t is input to the Transformer Block. There are in total four inside Stages, denoted as Stage-1, 2, 3, and 4, which contain different numbers of consecutive internal blocks where W-MSA and SW-MSA are implemented in. In addition, all stages excluding Stage-1, down-sample the feature map at a rate of two. Thus, the Transformer Block progressively expands feature maps' receptive fields and hierarchically extracts deep local features. It finally yields local feature representations f_l of the word images.

2) IDENTITY BRANCH

Inspired by the residual-learning mechanism in [36], the Identity Branch is proposed, which contains a global average pooling (GAP) layer and a fully-connected (FC) layer. Because the embedded pixel tokens t are directly computed from the entire image, the Identity branch receives t as input and transforms them into one-dimensional vectors f_g through the compression of GAP and projection of FC. Hence, f_g denotes the extracted global features.

3) GLOBAL BLOCK

The Global Block is introduced to aggregate the local features f_l and global features f_g , which comprises three FC layers and one dropout layer. f_l are first added up with f_g to form f_s , and FC layers receive f_s as input to make further

integration, resulting in an aggregator of all writing style features. The dropout layer randomly drops several nodes of the previous layer, which helps prevent overfitting and makes the Global Block more adaptive.

Finally, features formed before the Classification Head cover the local and global handwriting styles of the writers, leading to more robust classification evidence and better model generalizability.

B. PRE-TRAINING

This paper chose handwritten word recognition as the pre-training task, whereas the main task is writer identification. Handwritten word recognition is a typical task in the handwriting field, and can be defined as a multi-classification task analogous to writer identification. The two tasks are similar but do not overlap, thus no leakage problem exists. In the pre-training stage, RSTC learns knowledge k_{wr} from word recognition on larger datasets. In the training stage, k_{wr} is transferred and combined with the knowledge k_{wid} learned from writer identification, as illustrated in Equation 1:

$$\begin{aligned}
 k_{final} &= k_{wid} \oplus k_{wr} \\
 f_{final} &= RSTC(data, k_{final}) \tag{1}
 \end{aligned}$$

where $data$ is the handwritten word images fed into RSTC in the training stage. Based on more reasonable parameters, k_{wr} optimized the final feature representations f_{final} .

C. DATA AUGMENTATION

Perspective transformation is used as the data augmentation method, whereby a multi-point transformation is performed on the image without changing the writing style of each user. This process is described as:

$$[x', y', z'] = [x, y, z] \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \quad (2)$$

where $x, y,$ and z denote the original pixel coordinates and $x', y',$ and z' denote the transformed coordinates. Some examples are shown in Figure 3.

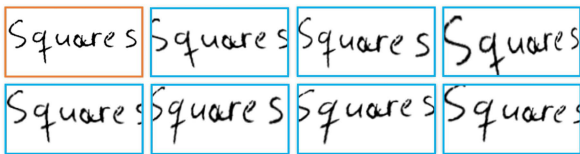


FIGURE 3. Examples of data augmentation using the perspective transformation. The upper left one is the original image. Others are the results of applying the algorithm to the original image.

D. LOSS FUNCTION

Label smoothing cross entropy loss [37] is adopted as the multi-classification loss. The original one hot label y_i becomes \hat{y}_i :

$$\hat{y}_i = y_i(1 - \varepsilon) + \frac{\varepsilon}{K} = \begin{cases} 1 - \varepsilon, & i = target \\ \frac{\varepsilon}{K}, & i \neq target \end{cases} \quad (3)$$

where K denotes the number of classes, and ε denotes the label smoothing factor. Label smoothing performs regularization to prevent the overfitting of network. Hence, the label smoothing cross-entropy loss can be interpreted as follows:

$$H(q', p) = (1 - \varepsilon)H(q, p) + \varepsilon H(\frac{1}{K}, p) \quad (4)$$

where q denotes the original ground-truth distribution, and q' denotes the ground-truth distribution after label smoothing. In addition, p is the prediction distribution computed by the model.

IV. EXPERIMENTS

A. DATASETS

Three publicly available benchmark datasets are used in this paper as follows:

- **The IAM dataset** [38] contains handwritten English text of 657 users in unconstrained scenarios, with at least one page of text per user. Each page was scanned at a resolution of 300dpi and saved as PNG images with 256 gray levels. These texts were automatically divided into isolated words, with writer labels provided.
- **The CVL dataset** [39] contains 310 users. 283 users contributed five pages (one German and four English) and the other 27 users contributed seven pages

(one German and six English). Similar to the IAM datasets, it also provided labeled isolated word images.

- **The GNHK dataset** [40] is a dataset that contains 688 pages of English handwriting in the wild. The images were captured by cameras in different regions around the world. Each page had a corresponding JSON file with the same name, which recorded the positions of the English words or letters on the corresponding image in the form of key-value pairs, where the key was the exact English word and the value was the position represented by a bounding box.

In the pre-training stage, IAM, CVL, and GNHK datasets are used. In the training stage, IAM and CVL datasets are split into the training set and testing set with a ratio of 4:1. Details are illustrated in Table 1 and 2.

TABLE 1. Pre-training dataset details.

Dataset	Images	Word Classes
IAM [38]	68072	2313
CVL [39]	97349	300
GNHK [40]	20621	1836
Total	186042	2518

TABLE 2. Training dataset details.

Dataset	Images	Training	Testing	Writer Classes
IAM [38]	90644	71990	18654	657
CVL [39]	91467	73076	18391	310
Total	182111	145066	37045	967

B. PRE-PROCESSING

Figure 4 illustrates the entire process. First, all word images are transformed into gray-level images and Gaussian filter algorithm are applied to filter out noise. Second, they are binarized using the OTSU [41] algorithm. Subsequently, images are resized to a fixed size (96,192), which maintains the aspect ratio with white pixel padding until reaching the preset size to avoid distortions. Before input to the network, all images are augmented using the perspective transformation.



FIGURE 4. Preprocess procedure: Converted to gray-level image -> Gaussian Filtering -> OTSU Binarization -> Resizing and Padding.

C. METRICS

This paper uses the popular Top-1 and Top-5 accuracies [3], [8], [33], [34] as metrics. Top-k refers to calculating the

percentage of queries, where the k highest-ranked words are from the same writer. Top-1 denotes the probability that the first ranked word stems from the match writer. Top-5 denotes the probability that words stem from the correct writer rank in the top five.

D. IMPLEMENTATION DETAILS

The overall network architecture is implemented using Pytorch. The size of per batch data is set to 128. The model is optimized using AdamW [42] with the initial learning rate and weight decay rate of 0.0001 and 0.00001 respectively. The learning rate decreases according to the cosine annealing algorithm proposed in [43]. The number of epochs is 100 in the pre-training stage and 200 in the training stage. The label smoothing factor ε in Equation 3 and 4 is set to 0.1.

E. EXPERIMENTAL RESULTS

1) ABLATION STUDY

a: IDENTITY BRANCH AND GLOBAL BLOCK

To evaluate the effectiveness of the Identity Branch I and Global Block G in RSTC, the ablation studies on the IAM and CVL datasets are conducted separately and jointly. As shown in Table 3, when I is used, the accuracy on joint datasets Acc_{joint} is improved by 0.3%, whereas the accuracies on single datasets, a.k.a Acc_{IAM} and Acc_{CVL} , gain minor improvements. Surprisingly, in the case of adding G only, although Acc_{joint} is still better than the one of adding I , Acc_{IAM} and Acc_{CVL} suffer drops. This indicates G does better in describing features in the mixed-dataset scenarios, whereas I is more friendly to single-dataset. With two datasets in use, the characteristics of CVL and IAM might affect each other because of G , leading to more adaptive feature representations and boosting performance. Moreover, I brings global information, which is more valuable than extracting only local features. When applying I and G together, RSTC achieves the best performance as shown in the sixth row, owing to the full utilization of local and global features.

TABLE 3. Ablation studies with different components of RSTC. Note that the datasets are split into the training set and testing set with a ratio of 4:1.

RSTC Settings		IAM		CVL		CVL and IAM	
Identity Branch	Global Block	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
×	×	90.8	96.8	93.1	98.1	92.7	97.6
✓	×	90.9	96.9	93.1	98.1	93.0	98.0
×	✓	90.7	96.6	92.7	97.7	93.3	97.8
✓	✓	91.4	97.0	93.3	98.1	93.7	97.9

b: RSTC AND OTHER BACKBONES

Table 4 respectively lists the results between different backbones. The tested models include ResNet-34 [36], RSTC without pre-training weights, and RSTC with pre-training weights. It is clear that whether using the datasets separately or jointly, RSTC trained from scratch yields higher accuracies

TABLE 4. Comparisons between different backbones. Note that the datasets are split into the training set and testing set with a ratio of 4:1.

model	pre-training weights	IAM		CVL		CVL and IAM	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Resnet-34 [36]	-	82.7	93.7	83.8	95.2	83.5	94.8
RSTC	×	87.5	95.6	90.3	97.2	89.0	96.6
RSTC	✓	91.4	97.0	93.3	98.1	93.7	97.9

than Resnet-34, which indicates that the well-designed RSTC captures more valid handwriting details in word images than the CNN-based model. When pre-training weights are leveraged, RSTC gains a significant increase of 4.7% in the Top-1 Acc_{joint} compared to when randomly initialized. This proves that knowledge of writing style learned from word recognition benefits the extraction of more robust features in the writer identification task to achieve higher accuracy.

There is another point that should be discussed. From Tables 3 and 4, Acc_{CVL} is higher than Acc_{joint} and Acc_{IAM} in some cases. Without pre-training weights and using single datasets, the performance is expected to be worse because of less data; however, experimental results show the opposite. This is an explanation for this observation. The CVL dataset is text-dependent as writers wrote the same words, whereas the IAM dataset is text-independent and considered more complex. Therefore, with less data, identifying writers on CVL is easier and the accuracy should be correspondingly higher than on IAM or both of them together. However, with I , G , and pre-training weights, the Top-1 Acc_{joint} surpasses Top-1 Acc_{IAM} and Acc_{CVL} . It reveals that RSTC achieves the best performance when the data is sufficient and can handle more complex conditions.

2) COMPARISON WITH STATE-OF-THE-ART METHODS

Tables 5 and 6 illustrate comparisons of different methods. The two datasets are re-divided with the ratio of 7:3 as mentioned in [33] for fair comparisons.

From Table 5 and 6, we can see that the hand-crafted features [3], [30], [44]–[47] have significantly low accuracy in word-level scenarios. This is because the hand-designed features fail to provide stable feature representations owing to scant handwriting data. Performances of neural network-based methods [33]–[35] are better than those of hand-crafted features; however, they are still not satisfactory. The proposed RSTC outperforms them by a large margin in the word-level scenario, especially on the IAM dataset with an improvement of 4.3%. RSTC yields a better performance even when compared with the method which used sentences as input as shown in the seventh row of Table 5.

3) PERFORMANCE DISTRIBUTION

The Top-1 accuracy distributions of each writer between RSTC and Vertical GR-RNN [35] with different datasets are illustrated in Figure 5. The blue line represents the baseline of the distribution and each red point corresponds to

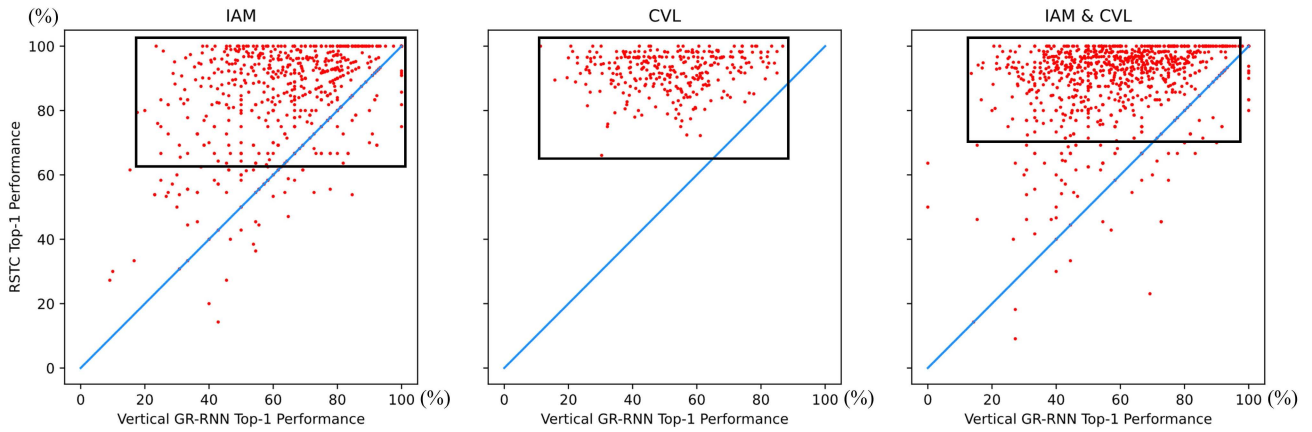


FIGURE 5. Top-1 accuracy distribution between RSTC and Vertical GR-RNN [35] on three scenarios. Note that the number of red points in each figure equals the number of writers of the corresponding dataset, which are 657, 310, and 967, respectively. The horizontal and vertical coordinates of all figures range from 0% to 100%.

TABLE 5. Comparison with Existing Methods on IAM Dataset. Note that the datasets are split into the training set and testing set with a ratio of 7:3.

method	scenario	Top-1	Top-5
curvature-free features [44]	word	15.7	32.1
textural and allographic features [45]	word	26.7	45.4
contour-based orientation [46]	word	30.5	49.8
co-occurrence [47]	word	37.2	57.8
ink-trace [30]	word	35.9	57.8
deep and hand-crafted descriptors [3]	sentence	86.3	96.1
deep-adaptive learning [33]	word	69.5	86.1
FragNet-64 [34]	word	85.1	95.0
Vertical GR-RNN (FGRR) [35]	word	85.9	95.2
Horizontal GR-RNN (FGRR) [35]	word	86.1	95.0
RSTC (This paper)	word	90.7	96.6

TABLE 6. Comparison with Existing Methods on the CVL Dataset. Note that the datasets are split into the training set and testing set with a ratio of 7:3.

method	scenario	Top-1	Top-5
curvature-free features [44]	word	12.8	29.6
textural and allographic features [45]	word	25.8	48.0
contour-based orientation [46]	word	28.8	51.4
co-occurrence [47]	word	30.0	52.4
ink-trace [30]	word	29.4	52.6
deep-adaptive learning [33]	word	78.6	93.2
FragNet-64 [34]	word	90.2	97.5
Vertical GR-RNN (FGRR) [35]	word	92.6	97.9
Horizontal GR-RNN (FGRR) [35]	word	92.4	97.8
RSTC (This paper)	word	92.7	97.9

one writer. Points above this line indicate that RSTC performs better, whereas points below indicate that Vertical GR-RNN performs better. Since most points are clustered in the upper left (in the black boxes), it suggests that Vertical GR-RNN yields poor performance on these writers and RSTC can improve the identification accuracies of these writers. On the CVL dataset, points are more aggregated than in other scenarios, which indicates that RSTC is more adaptive to writer-dependent data than to writer-independent data.

4) VISUALIZATION

The t-distributed Stochastic Neighbor Embedding (t-SNE) visualization is conducted to intuitively demonstrate the identification performance of RSTC, as shown in Figure 6. 50 writers with 16 word images for each are randomly selected in the testing set as the input of RSTC. Based on the final feature vectors computed by the last layer of the Global Block, the visualization results exhibit the distribution of

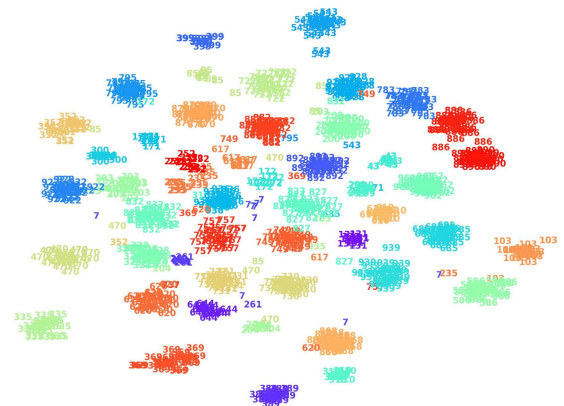


FIGURE 6. t-SNE visualization, each cluster represents the words written by one user.

feature vectors in a two-dimensional space after dimensionality reduction. For most users, their handwriting is gathered, and the clusters of each user are mostly isolated.

This indicates that the proposed RSTC can distinguish writers correctly so that they do not interfere with each other, highlighting its high identification ability in the word-level scenario.

V. CONCLUSION

This paper proposes a novel Residual Swin Transformer Classifier (RSTC) for offline word-level writer identification. To the best of our knowledge, this is the first attempt that successfully employed the vision transformer in this field. The Transformer Block in RSTC finely models the local features, and the introduced Identity Branch and Global Block deeply characterize global handwriting styles that have usually been ignored. Therefore, RSTC comprehensively extracts and tightly couples local and global features with limited handwriting data and yields robust feature representations for different writers. In addition, the pre-training technique that has not been popularly leveraged is exploited to boost RSTC's performance. Experimental results demonstrate the superiority of the proposed RSTC, surpassing the existing methods by a significant margin.

One limitation of RSTC is that when computing self-attention on white regions of word images, it may yield less useful information. The model should be guided to concentrate on the valid pixels of word images rather than the noise pixels. To this end, the future work is oriented to drawing on the technologies of deformable convolution and deformable ROI pooling [48] and combining them with the shifted-window attention mechanism to further focus on the word regions. The model is expected to capture more handwriting information using deformation. In this case, more supervision information such as the mask annotations for segmentation of words may be required to perform deformable attention.

REFERENCES

- [1] C. Adak, B. B. Chaudhuri, and M. Blumenstein, "An empirical study on writer identification and verification from intra-variable individual handwriting," *IEEE Access*, vol. 7, pp. 24738–24758, 2019.
- [2] A. Rehman, S. Naz, and M. I. Razzak, "Writer identification using machine learning approaches: A comprehensive review," *Multimedia Tools Appl.*, vol. 78, no. 8, pp. 10889–10931, Apr. 2019.
- [3] A. Sulaiman, K. Omar, M. F. Nasrudin, and A. Arram, "Length independent writer identification based on the fusion of deep and hand-crafted descriptors," *IEEE Access*, vol. 7, pp. 91772–91784, 2019.
- [4] Z. Xia, T. Shi, N. N. Xiong, X. Sun, and B. Jeon, "A privacy-preserving handwritten signature verification method using combinational features and secure kNN," *IEEE Access*, vol. 6, pp. 46695–46705, 2018.
- [5] V. Christlein, D. Bernecker, A. Maier, and E. Angelopoulou, "Offline writer identification using convolutional neural network activation features," in *Pattern Recognition*. Cham, Switzerland: Springer, 2015, pp. 540–552.
- [6] L. Xing and Y. Qiao, "DeepWriter: A multi-stream deep CNN for text-independent writer identification," in *Proc. 15th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Oct. 2016, pp. 584–589.
- [7] M. Liu, L. Jin, and Z. Xie, "PS-LSTM: Capturing essential sequential online information with path signature and LSTM for writer identification," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Nov. 2017, pp. 664–669.
- [8] W. Yang, L. Jin, and M. Liu, "Chinese character-level writer identification using path signature feature, DropStroke and deep CNN," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 546–550.
- [9] V. Christlein and A. Maier, "Encoding CNN activations for writer recognition," in *Proc. 13th IAPR Int. Workshop Document Anal. Syst. (DAS)*, Apr. 2018, pp. 169–174.
- [10] M. F. Mridha, A. Q. Ohi, J. Shin, M. M. Kabir, M. M. Monowar, and M. A. Hamid, "A thresholded Gabor-CNN based writer identification system for indic scripts," *IEEE Access*, vol. 9, pp. 132329–132341, 2021.
- [11] Z. Chen, H.-X. Yu, A. Wu, and W.-S. Zheng, "Letter-level online writer identification," *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1394–1409, May 2021.
- [12] A. Srivastava, S. Chanda, and U. Pal, "Exploiting multi-scale fusion, spatial attention and patch interaction techniques for text-independent writer identification," *CoRR*, vol. abs/2111.10605, pp. 1–14, Nov. 2021.
- [13] T. T. Ngo, H. T. Nguyen, and M. Nakagawa, "A-VLAD: An end-to-end attention-based neural network for writer identification in historical documents," in *Proc. 16th Int. Conf. Document Anal. Recognit. (ICDAR)*. Cham, Switzerland: Springer, 2021, pp. 396–409.
- [14] M. A. Shaikh, T. Duan, M. Chauhan, and S. N. Srihari, "Attention based writer independent verification," in *Proc. 17th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Sep. 2020, pp. 373–379.
- [15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021.
- [16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
- [17] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–16.
- [18] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.
- [19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACLHLT*, 2019, pp. 4171–4186.
- [20] T. Brown *et al.*, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.
- [21] H. E. S. Said, T. N. Tan, and K. D. Baker, "Personal identification based on handwriting," *Pattern Recognit.*, vol. 33, no. 1, pp. 149–160, 2000.
- [22] D. Bertolini, L. S. Oliveira, E. Justino, and R. Sabourin, "Texture-based descriptors for writer identification and verification," *Expert Syst. Appl.*, vol. 40, no. 6, pp. 2069–2080, 2013.
- [23] A. Nicolaou, A. D. Bagdanov, M. Liwicki, and D. Karatzas, "Sparse radial sampling LBP for writer identification," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 716–720.
- [24] A. J. Newell and L. D. Griffin, "Writer identification using oriented basic image features and the delta encoding," *Pattern Recognit.*, vol. 47, no. 6, pp. 2255–2265, 2014.
- [25] B. Helli and M. E. Moghaddam, "A text-independent Persian writer identification based on feature relation graph (FRG)," *Pattern Recognit.*, vol. 43, no. 6, pp. 2199–2209, Jun. 2010.
- [26] M. Bulacu, L. Schomaker, and A. Brink, "Text-independent writer identification and verification on offline Arabic handwriting," in *Proc. 9th Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 2, Sep. 2007, pp. 769–773.
- [27] R. Fernandez-de-Sevilla, F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia, "Forensic writer identification using allographic features," in *Proc. 12th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Nov. 2010, pp. 308–313.
- [28] R. Niels, F. Grootjen, and L. Vuurpijl, "Writer identification through information retrieval: The allograph weight vector," in *Proc. 11th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*. Montreal, QC, Canada: Concordia Univ., 2008, pp. 481–486.
- [29] R. Jain and D. Doermann, "Offline writer identification using k -adjacent segments," in *Proc. 11th Int. Conf. Document Anal. Recognit. (ICDAR)*, 2011, pp. 769–773.
- [30] A. A. Brink, J. Smit, M. L. Bulacu, and L. R. B. Schomaker, "Writer identification using directional ink-trace width measurements," *Pattern Recognit.*, vol. 45, no. 1, pp. 162–171, 2012.
- [31] S. Lai, Y. Zhu, and L. Jin, "Encoding pathlet and SIFT features with bagged VLAD for historical writer identification," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3553–3566, 2020.

- [32] A. Rehman, S. Naz, M. I. Razak, and I. A. Hameed, "Automatic visual features for writer identification: A deep learning approach," *IEEE Access*, vol. 7, pp. 17149–17157, 2019.
- [33] S. He and L. Schomaker, "Deep adaptive learning for writer identification based on single handwritten word images," *Pattern Recognit.*, vol. 88, no. 1, pp. 64–74, 2019.
- [34] S. He and L. Schomaker, "FragNet: Writer identification using deep fragment networks," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3013–3022, 2020.
- [35] S. He and L. Schomaker, "GR-RNN: Global-context residual recurrent neural networks for writer identification," *Pattern Recognit.*, vol. 117, Sep. 2021, Art. no. 107975.
- [36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
- [38] U.-V. Marti and H. Bunke, "The IAM-database: An English sentence database for offline handwriting recognition," *Int. J. Document Anal. Recognit.*, vol. 5, no. 1, pp. 39–46, 2002.
- [39] F. Kleber, S. Fiel, M. Diem, and R. Sablatnig, "CVL-DataBase: An offline database for writer retrieval, writer identification and word spotting," in *Proc. 12th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2013, pp. 560–564.
- [40] A. W. C. Lee, J. Chung, and M. Lee, "GNHK: A dataset for English handwriting in the wild," in *Proc. 16th Int. Conf. Document Anal. Recognit. (ICDAR)*. Cham, Switzerland: Springer, 2021, pp. 399–412.
- [41] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.
- [42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–19.
- [43] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.
- [44] S. He and L. Schomaker, "Writer identification using curvature-free features," *Pattern Recognit.*, vol. 63, pp. 451–464, Mar. 2017.
- [45] M. Bulacu and L. Schomaker, "Text-independent writer identification and verification using textural and allographic features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 701–717, Apr. 2007.
- [46] I. Siddiqi and N. Vincent, "Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features," *Pattern Recognit.*, vol. 43, no. 11, pp. 3853–3865, 2010.
- [47] S. He and L. Schomaker, "Co-occurrence features for writer identification," in *Proc. 15th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Oct. 2016, pp. 78–83.
- [48] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.



PEIRONG ZHANG is currently pursuing the B.S. degree with the College of Electronics and Information Engineering, South China University of Technology. His research interests include deep learning, handwriting analysis and recognition, and computer vision.

...