Structure-Aware Handwritten Text Recognition via Graph-Enhanced Cross-Modal Mutual Learning

Ji Gan[1,2,3], **Yupeng Zhou**[1,2], **Yanming Zhang**[4], **Jiaxu Leng**[1,2,3] and **Xinbo Gao**[*1,2,3]
[1] School of Computer Science and Technology, Chongqing University of Posts and Telecommunications
[2] Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications
[3]Chongqing Institute for Brain and Intelligence, Guangyang Bay Laboratory
[4]State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences
{ganji@, s230232052@stu.}cqupt.edu.cn, ymzhang@nlpr.ia.ac.cn, {lengjx, gaoxb}@cqupt.edu.cn

Abstract

Existing handwriting recognition methods only focus on learning visual patterns by modeling low-level relationships of adjacent pixels, while overlooking the intrinsic geometric structures of characters. In this paper, we propose a novel graph-enhanced cross-modal mutual learning network GCM to fully process handwritten text images alongside their corresponding geometric graphs, which consists of one shared cross-modal encoder and two parallel inverse decoders. Specifically, the encoder simultaneously extracts visual and geometric information from the cross-modal inputs, and the decoders fuse the multi-modal features for prediction under the guidance of cross-modal fusion. Moreover, two parallel decoders sequentially aggregate cross-modal features in inverse orders (V→G and G→V) but are enhanced through mutual distillation at each time-step, which involves one-to-one knowledge transfer and fully leverages complementary cross-modal information from both directions. Notably, only one branch of GCM is activated in inference, thus avoiding the increase of the model parameters and computation costs for testing. Experiments show that our method outperforms previous state-of-the-art methods on public benchmarks such as IAM, RIMES, and ICDAR-2013 when no extra training data is utilized.

1 Introduction

Handwritten text recognition (HTR) is considered a high-level human-computer interaction, which aims to convert humans' handwriting into characters or texts. HTR still remains challenging since humans' handwriting can be very arbitrary and has distinct writing styles. With the recent advances in deep learning techniques, it has witnessed significant progress in HTR [Yousef *et al.*, 2020; Kang *et al.*, 2022; Li *et al.*, 2023], in which the most widely used approaches are attention-based encoder-decoder models or networks equipped with the connectionist temporal classification (CTC) [Graves *et al.*, 2008] auxiliary objectives.

However, most existing methods for HTR simply operate on handwriting data from a single modality, i.e., either visual images or geometric graphs alone. Specifically, visual-based methods simply focus on exploiting visual patterns of handwritten texts by modelling the low-level relationships of adjacent pixels, while ignoring the intrinsic geometric structures of characters. Instead, computational methods based on structured, relational information can better demonstrate human-like learning. Particularly, recent advances [Gan *et al.*, 2023; Chen *et al.*, 2024] demonstrate that handwritten characters can also be considered as geometric graphs, which can explicitly learn the geometric structures of characters by utilizing graph-based networks [Yao *et al.*, 2019]. Although such graph-based methods are primarily proposed for recognizing isolated handwritten characters, they can be feasibly extended to HTR by integrating the encoder-decoder framework.

Nevertheless, it remains under-explored whether the combination of visual and geometric information can benefit handwriting analysis. Existing uni-modal approaches typically overlook the fact that both visual and geometric information of characters are advantageous for handwriting recognition, in which the two distinct modalities can complement each other. Therefore, it is highly probable to achieve better recognition performance by adequately exploiting and fusing both visual and geometric properties of handwritten texts through multi-model learning. In other words, a structure-aware handwriting recognition approach that explicitly incorporates both characters' geometric structures and visual appearances is promising to achieve better performance.

Furthermore, it also suffers from inadequate cross-modal fusion and conflicting optimization directions when directly adopting existing multi-modal fusion techniques for HTR. Recent advances [Huang *et al.*, 2022b; Peng *et al.*, 2022] demonstrate that multi-modal learning may be inferior to the uni-modal models under certain circumstances, since multi-modal models sometimes cannot jointly utilize all modalities well due to the conflicting optimization directions and uncoordinated modality convergence. Moreover, existing encoder-decoders typically adopt the two-dimensional coverage attention to focus on the current visual region of characters for

*Corresponding author: Xinbo Gao

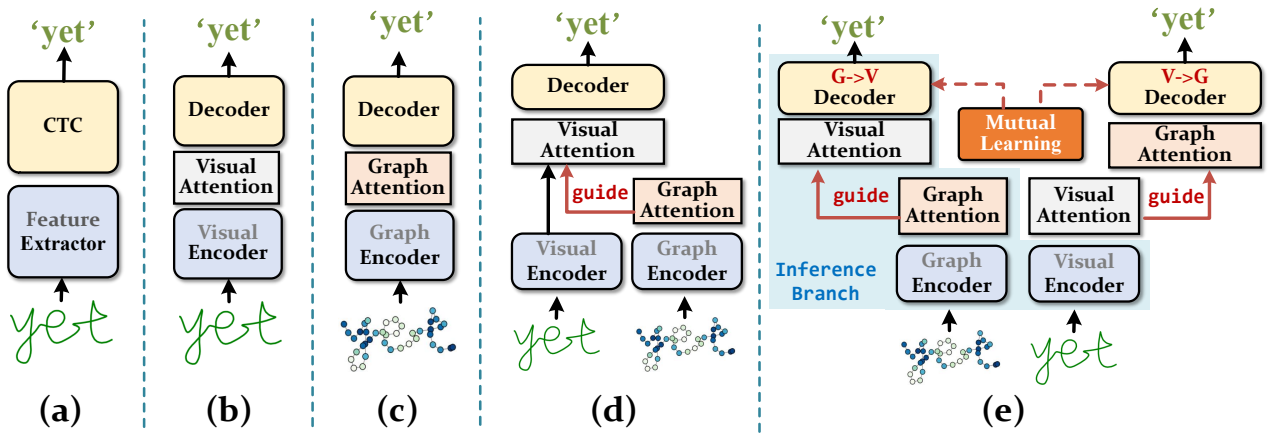

Figure 1: Typical architectures and our proposed GCM for HTR: (a) CTC-based methods; (b) Attention-based encoder-decoders on visual images; (c) Encoder-decoders on geometric graphs; (d) Graph-enhanced encoder-decoders (baseline); (e) The proposed GCM, which further introduces bidirectional mutual learning on cross-modality data. Notably, only one decoding branch of GCM will be activated in inference.

literal prediction, while ignoring the fact that geometric structures are more reasonable and intuitive for identifying characters, which therefore may cause the problem of attention drift. Nevertheless, it remains unclear how to appropriately and effectively integrate the geometric modality data to complement the visual modality for better performance.

To address those challenges, we introduce a graph-enhanced cross-modal mutual learning network (GCM) to fully process handwriting images alongside their corresponding geometric graphs for HTR. Firstly, we introduce a graph-enhanced cross-modal decoder to aggregate the geometric contexts from graphs and then employ such graph contexts to guide the visual coverage attention on visual regions in the G→V direction, thus achieving structure-aware handwriting recognition. Secondly, considering that geometric contexts can guide the decoding process on visual features (from G→V), visual context can also guide the decoding process on geometric features in the inverse direction (from V→G). Intuitively, the cross-modal features from two inverse directions should be decoded into identical predictions, and thus, it is possible to enhance each other by minimizing their distances in the shared latent space. Therefore, we introduce two parallel decoders that sequentially aggregate cross-modal features in inverse orders (V→G and G→V) but are enhanced through mutual distillation at each time step. This eventually achieves one-to-one knowledge transfer and fully leverages complementary cross-modal information from both directions. Fig. 1 highlights the differences between our GCM and other prevalent approaches, and our contributions are listed as follows:

- We achieve structure-aware handwritten text recognition via graph-enhanced cross-modal mutual learning (GCM). Notably, cross-modal methods that fuse both geometric graphs and visual appearances remain under-explored, and we are also among the first to introduce mutual learning for graph-enhanced cross-modal fusion in HTR.
- We propose a novel bidirectional cross-modal mutual learning strategy that sequentially aggregates cross-modal features in inverse orders (V→G and G→V) and further enhances each other through mutual distillation. Notably, only one decoding branch is activated in inference, thus avoiding the increase of model parameters and computation costs for testing and deployment.
- Extensive experiments demonstrate that the proposed GCM surpasses previous state-of-the-art (SOTA) methods for HTR on benchmarks such as IAM, RIMES, and ICDAR-2013 when no extra training data is utilized.

2 Related Work

2.1 Handwritten Text Recognition

It has witnessed significant progress in HTR in recent years, which can be briefly divided into CTC-based and attention-based methods. For CTC-based methods, [Graves *et al.*, 2008; Graves and Schmidhuber, 2008] proposed to utilize recurrent neural networks (RNNs) [Lipton *et al.*, 2015] for segmentation-free HTR. Furthermore, [Coquenet *et al.*, 2020; Ingle *et al.*, 2019] adopted convolutional neural networks (CNNs) to model the sequential handwritten texts in parallel. Recently, attention-based methods have emerged for HTR with promising performance [Yousef *et al.*, 2020]. [Kang *et al.*, 2022; Li *et al.*, 2023] further introduced Transformers for HTR due to their better ability to model global context relationships. Moreover, [Gan *et al.*, 2023] proposed to view handwriting characters as geometric graphs and further employed graph Transformers to recognize those graphs. Nevertheless, most existing methods for HTR simply operate a single modality alone, while leaving the multi-modal methods that fuse visual and geometric properties under-explored.

2.2 Mutual Learning for Handwriting Analysis

Mutual learning [Zhang *et al.*, 2018; Zhu *et al.*, 2018; Guo *et al.*, 2020] refers to a learning process in which an ensemble of networks learn collaboratively and teach each other via knowledge transfer. However, very few methods have attempted to utilize mutual learning for handwriting analysis. Nevertheless, we are among the first to adopt mutual learning into multi-modal methods for HTR with promising results.

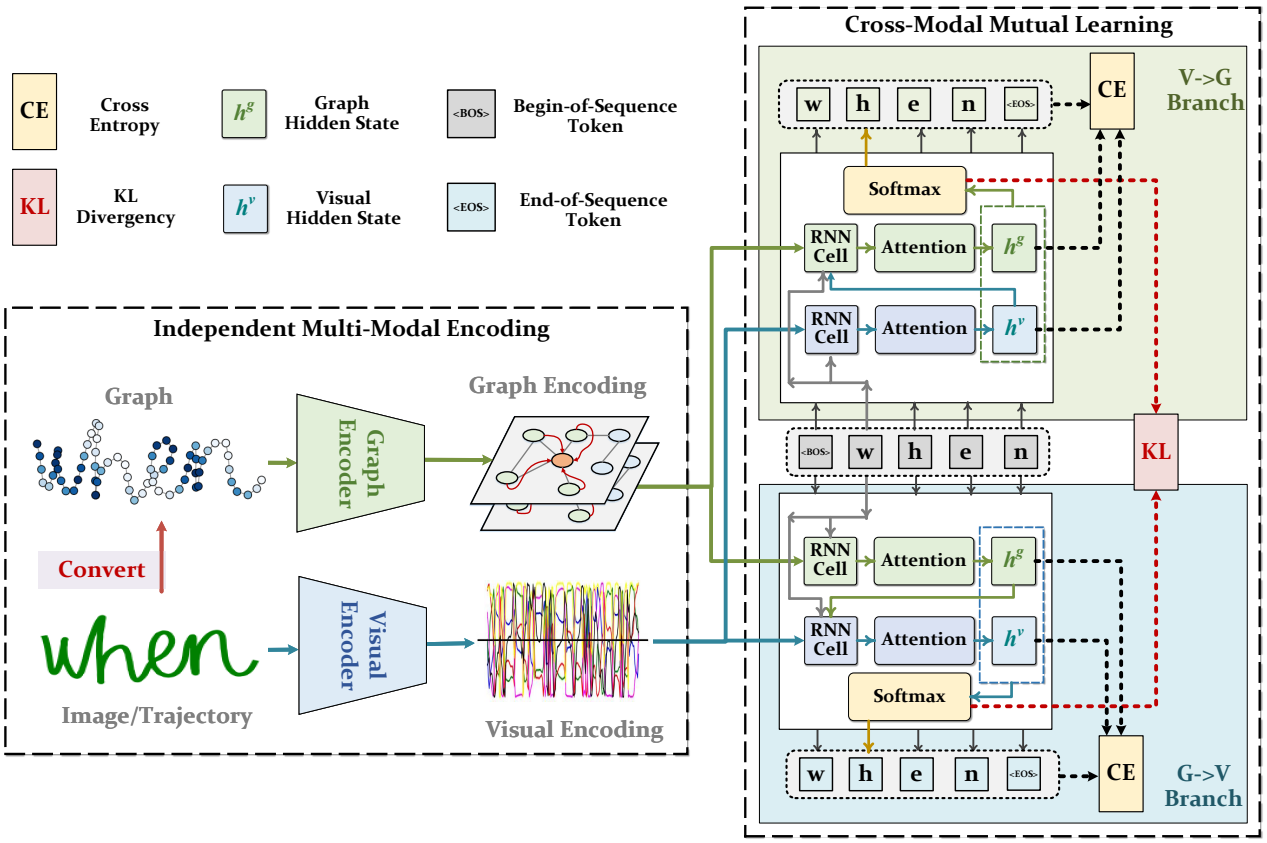

Figure 2: Overview of GCM for HTR. The proposed GCM consists of a shared cross-modal encoder and two parallel inverse decoders. Specifically, 1) the multi-modal encoder first extracts cross-modal features from both images and graphs respectively, and 2) the cross-modal decoders fuse multi-modal features under the guidance of cross-modal contexts through the hybrid multi-modal fusion and cascaded attention aggregation, 3) finally, two parallel decoders sequentially aggregate cross-modal features in inverse orders (V→G and G→V) but are enhanced through mutual distillation at each time step, which involves one-to-one knowledge transfer and adequate cross-modal fusion.

3 Methodology

As shown in Fig. 2, we propose a graph-enhanced cross-modal mutual learning network (GCM) to achieve structure-aware handwritten text recognition, which explicitly exploits both visual and geometric characteristics of handwritten texts. Given a handwritten text image $\mathbf{v}$, we first convert it into a skeleton-based graph $\mathbf{g}$ and then adopt the recognition model to produce the target character sequence $\mathbf{y} = (y_1, \cdots, y_l, \cdots, y_L)$ with the maximal probability, i.e.,

$$\mathbf{y}^* = \text{argmax}_{\mathbf{y}} P(\mathbf{y}|\mathbf{v}, \mathbf{g}), \tag{1}$$

where each character $y_l \in \mathcal{A}$ is in the alphabet set $\mathcal{A}$.

3.1 Graph Construction from Images

Following [Gan *et al.*, 2023], we first extract skeletons from handwriting images and then resample the key points in identical intervals while keeping their original connections as shown in Fig. 3. Eventually, we obtain geometric graphs from visual images without introducing any extra annotation costs.

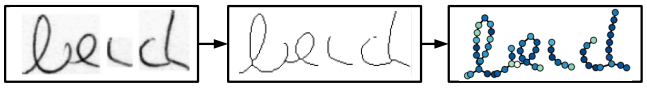

Figure 3: Graph construction from visual images.

3.2 Independent Multi-Modal Encoding

To exploit characters' visual and geometric properties, it is important to first encode the multi-modal data into a shared latent representation space for effectively capturing complementary information from each other. Since geometric graphs and visual images of characters have distinct data organization, a more appropriate way is to utilize two independent multi-modal encoders (i.e., the graph encoder and visual encoder) for extracting the cross-modal features separately.

Visual Encoder We adopt the most prevalent convolutional recurrent network (CRN) [Shi *et al.*, 2016] to directly extract the visual information from handwriting images, which is a hybrid network that fully integrates the advantages of both CNNs and RNNs. Given a handwriting image $\mathbf{v}$, a visual encoder $\mathcal{V}$ is utilized to extract visual features as $\tilde{\mathbf{v}} = \mathcal{V}(\mathbf{v})$.

Graph Encoder To fully exploit the geometric structures of handwritten texts, we utilize the pyramid graph Transformer (PyGT) [Gan *et al.*, 2023] to process character graphs, which is stacked by multiple graph-enhanced atten-

tion blocks. Given a character graph $\mathbf{g}$, the graph encoder $\mathcal{G}$ is adopted to exploit the local topology structures as $\tilde{\mathbf{g}} = \mathcal{G}(\mathbf{g})$.

3.3 Graph-Enhanced Cross-Modal Decoding

Considering that geometric and visual features are essentially complementary, the graph context can guide the decoding process on the visual regions of characters (from G→V), and similarly, visual context can also guide the decoding process on geometric features in the inverse direction (from V→G). To fully integrate the cross-modal features of characters, we introduce novel graph-enhanced cross-modal decoders in two inverse directions through hybrid multi-modal fusion and cascaded attention aggregation.

Graph → Visual Graph-Enhanced Decoder

G→V decoding branch aggregates geometric context from character graphs and further employs it to guide the coverage attention over visual images.

Graph Guidance The geometric context is captured as the graph guidance for the later decoding on visual images. Specifically, a graph recurrent cell $\mathcal{R}_g^{G\to V}$ first calculates the decoding context $\mathbf{c}_l^G$ of l-th character based on its previous graph hidden state $\mathbf{h}_{l-1}^G$ and the previous character y_{l-1} as

$$\mathbf{c}_l^G = \mathcal{R}_g^{G\to V}(\mathbf{h}_{l-1}^G, y_{l-1}). \tag{2}$$

The graph hidden state $\mathbf{h}_l^G$ is then updated based on the geometric feature $\tilde{\mathbf{g}}$ and decode context $\mathbf{c}_l^G$ via attention-based aggregation as

$$\mathbf{h}_l^G = \text{Softmax}(\tilde{\mathbf{g}} \cdot \mathbf{c}_l^G)^\top \cdot \mathbf{c}_l^G, \tag{3}$$

where $\top$ is the transpose operation. The obtained graph hidden state $\mathbf{h}_l^G$ can be considered as the graph guidance.

Graph-Guided Cross-Modal Decoding The graph guidance $\mathbf{h}_l^G$ is integrated to guide the decoding process on visual regions of characters based on the hybrid multi-modal fusion and cascaded attention aggregation. Specifically, the visual recurrent cell $\mathcal{R}_v^{G\to V}$ computes the current decoding state $\mathbf{c}_l^{G\to V}$ depending on its previous visual hidden state $\mathbf{h}_{l-1}^{G\to V}$ as well as the previous character y_{l-1}, and also fuses the graph guidance $\mathbf{h}_l^G$ at the early stage, i.e.,

$$\mathbf{c}_l^{G\to V} = \mathcal{R}_v^{G\to V}(\mathbf{h}_{l-1}^{G\to V} + \alpha_g \mathbf{h}_l^G, y_{l-1}), \tag{4}$$

where α_g is a learnable fusion parameter. Then, the visual hidden state $\mathbf{h}_l^{G\to V}$ is updated through attention-based aggregation as

$$\mathbf{h}_l^{G\to V} = \text{Softmax}(\tilde{\mathbf{v}} \cdot \mathbf{c}_l^{G\to V})^\top \cdot \mathbf{c}_l^{G\to V}, \tag{5}$$

where $\tilde{\mathbf{v}}$ denotes visual features. Finally, we fuse the cross-modal contexts at the late stage as

$$\mathcal{O}_l^{G\to V} = \mathcal{F}([\mathbf{h}_l^{G\to V}, \mathbf{h}_l^G]), \tag{6}$$

where $\mathcal{O}_l^{G\to V}$ is the discriminative pattern in the G→V branch, and $\mathcal{F}$ is a fully connected layer. Finally, the probability of producing the l-th character in the G→V branch is calculated as

$$P^{G\to V}(y_l|\mathbf{y}_{<l}, \tilde{\mathbf{v}}, \tilde{\mathbf{g}}) = \text{Softmax}(\mathcal{O}_l^{G\to V}), \tag{7}$$

where $\mathbf{y}_{<l}$ denotes the previous $l-1$ characters.

Visual → Graph Visual-Enhanced Decoder

Similarly, we can also aggregate the visual context to guide the coverage attention over geometric structures of characters with the V→G decoding branch.

Visual Guidance Similarly, the visual context is computed through a visual recurrent cell $\mathcal{R}^{V\to G}$ based on the visual features $\tilde{\mathbf{v}}$, previous visual hidden state $\mathbf{h}_{l-1}^V$, and the previous character y_{l-1} as

$$\mathbf{c}_l^V = \mathcal{R}_v^{V\to G}(\mathbf{h}_{l-1}^V, y_{l-1}), \tag{8}$$

$$\mathbf{h}_l^V = \text{Softmax}(\tilde{\mathbf{v}} \cdot \mathbf{c}_l^V)^\top \cdot \mathbf{c}_l^V, \tag{9}$$

where the visual context $\mathbf{h}_l^V$ is treated as the visual guidance.

Visual-Guided Cross-Modal Decoding We similarly compute the discriminative pattern in the V→G branch as

$$\mathbf{c}_l^{V\to G} = \mathcal{R}_g^{V\to G}(\mathbf{h}_{l-1}^{V\to G} + \alpha_v \mathbf{h}_l^V, y_{l-1}), \tag{10}$$

$$\mathbf{h}_l^{V\to G} = \text{Softmax}(\tilde{\mathbf{v}} \cdot \mathbf{c}_l^{V\to G})^\top \cdot \mathbf{c}_l^{V\to G}, \tag{11}$$

$$\mathcal{O}_l^{V\to G} = \mathcal{F}([\mathbf{h}_l^V, \mathbf{h}_l^{V\to G}]). \tag{12}$$

Lastly, the probability of producing the l-th character in the V→G branch is calculated as

$$P^{V\to G}(y_l|\mathbf{y}_{<l}, \tilde{\mathbf{v}}, \tilde{\mathbf{g}}) = \text{Softmax}(\mathcal{O}_l^{V\to G}). \tag{13}$$

Eventually, we obtain the probabilities of the l-th character y_l in inverse directions at each time step.

3.4 Bidirectional Cross-Modal Mutual Learning

Since the cross-modal representations are complementary to each other, the cross-modal discriminative patterns from two inverse directions (i.e. $\mathcal{O}_l^{V\to G}$ and $\mathcal{O}_l^{G\to V}$) should produce the identical character y_l at each decoding step. Therefore, the cross-modal patterns in inverse fusion directions can enhance each other via one-to-one knowledge transfer, which is achieved by minimizing their distances in the shared latent space with better multi-modal alignment. Specifically, we introduce Kullback-Leibler (KL) divergency [Zhang *et al.*, 2018] to quantify the difference in prediction distributions between $\mathcal{O}_l^{V\to G}$ and $\mathcal{O}_l^{G\to V}$. The objective is to minimize the distance between the probability distributions of two branches at each decoding step through mutual distillation as

$$\mathcal{L}_{KL} = \sum_{l=1}^{L} \epsilon(\mathcal{O}_l^{V\to G}, T) log \frac{\epsilon(\mathcal{O}_l^{V\to G}, T)}{\epsilon(\mathcal{O}_l^{G\to V}, T)}, \tag{14}$$

where $\epsilon(\mathcal{O}_l^*, T) = \text{Softmax}(\mathcal{O}_l^*/T)$ is the soft probabilities with the temperature T. This eventually achieves one-to-one knowledge transfer between inverse branches and fully leverages the complementary cross-modal information from both directions.

3.5 End-to-End Optimization Objectives

Given the target sequence $\tilde{\mathbf{y}} = (\tilde{y_1}, \cdots, \tilde{y_l}, \cdots, \tilde{y_L})$, the model first minimizes the cross-entropy losses over multi-modal features in two inverse branches as

$$\mathcal{L}_M^{G\to V} = \sum_{l=1}^{L} -\tilde{y_l} \log \frac{\exp(\mathcal{O}_{l,\tilde{y_l}}^{G\to V})}{\sum_{y_l} \exp(\mathcal{O}_{l,y_l}^{G\to V})}, \tag{15}$$

$$\mathcal{L}_M^{V\to G} = \sum_{l=1}^{L} -\tilde{y_l} \log \frac{\exp(\mathcal{O}_{l,\tilde{y_l}}^{V\to G})}{\sum_{y_l} \exp(\mathcal{O}_{l,y_l}^{V\to G})}, \tag{16}$$

where $\mathcal{O}_{l,y_l}^{V\to G}$ and $\mathcal{O}_{l,y_l}^{G\to V}$ are the logits of y_l-th class. Moreover, we further introduce the uni-modal learning objectives besides multi-modal joint learning to alleviate the uncoordinated modality convergence issue [Huang *et al.*, 2022b] as

$$\mathcal{L}_U^{G\to V} = \mathcal{L}_{CE}^{V|G\to V} + \mathcal{L}_{CE}^{G|G\to V}, \tag{17}$$

$$= \sum\nolimits_{*\in\{V,G\}} \sum_{l=1}^{L} -\tilde{y}_l \log \frac{\exp(\mathcal{O}_{l,\tilde{y}_l}^{*|G\to V})}{\sum_{y_l}\exp(\mathcal{O}_{l,y_l}^{*|G\to V})}, \tag{18}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss, and $\mathcal{O}_{l,\tilde{y}_l}^{*|G\to V}$ denotes the uni-modal logits of the G→V branch, i.e., either the visual modal logit $\mathcal{O}_{l,\tilde{y}_l}^{V|G\to V}$ or the graph modal logit $\mathcal{O}_{l,\tilde{y}_l}^{G|G\to V}$ as

$$\mathcal{O}_{l,\tilde{y}_l}^{V|G\to V} = \mathcal{F}(\mathbf{h}_l^{G\to V}), \quad \mathcal{O}_{l,\tilde{y}_l}^{G|G\to V} = \mathcal{F}(\mathbf{h}_l^{G}). \tag{19}$$

Similarly, the uni-modal learning objective is calculated as

$$\mathcal{L}_U^{V\to G} = \mathcal{L}_{CE}^{V|V\to G} + \mathcal{L}_{CE}^{G|V\to G} \tag{20}$$

with the uni-modal logits of the V→G branch as

$$\mathcal{O}_{l,\tilde{y}_l}^{V|V\to G} = \mathcal{F}(\mathbf{h}_l^{V}), \quad \mathcal{O}_{l,\tilde{y}_l}^{G|V\to G} = \mathcal{F}(\mathbf{h}_l^{V\to G}). \tag{21}$$

Finally, the overall training objective is summarized as

$$\mathcal{L}_{Total} = \mathcal{L}_M^{G\to V} + \mathcal{L}_M^{G\to V} + \mathcal{L}_U^{V\to G} + \mathcal{L}_U^{V\to G} + \lambda\mathcal{L}_{KL}, \tag{22}$$

where λ is the hyperparameter that controls the significance of the cross-modal mutual learning.

4 Experiments

4.1 Experiment Settings

Datasets All used datasets are publicly available and have the official dataset splits or proportions, which includes:

+ **IAM** [Marti and Bunke, 2002] is the most widely used public handwritten English text dataset, which contains 1539 handwritten pages comprising 115,320 words.
+ **RIMES** [Grosicki *et al.*, 2009] is a public French handwriting dataset, which is contributed by over 1300 people with 12,723 pages corresponding to 5605 mails.
+ **IAHEW-UCAS2016** [Gan and Wang, 2019] is a public in-air handwriting English word dataset, which contains 150,480 samples covering 2280 English words.
+ **ICDAR2013** [Yin *et al.*, 2013] is the most widely used Chinese handwriting dataset, which contains 3755 classes of Chinese characters with 224,419 handwriting samples.

Evaluation Metrics We use the most widely used *Character Error Rate* (CER) and *Word Error Rate* (WER) to evaluate the performance of handwriting recognition models.

Implementation Details The whole architecture is implemented with the PyTorch [Paszke *et al.*, 2017] deep learning framework. The model is optimized via the Adam [Kingma and Ba, 2015] algorithm with a batch size of 64. We set the initial learning rate to 0.001 and the hyperparameter of mutual learning λ to 0.75 by default, and the training process is terminated when the model reaches convergence. We set the beam width to 64 during the decoding stage. All experiments are conducted on a workstation with an Intel(R) Core(TM) i9-11900K CPU, 64GB RAM, and an RTX-4090 24GB GPU.

Method	Input	IAM (%)		RIMES (%)	
		CER	WER	CER	WER
[Espana *et al.*, 2010]	V	9.8	22.4	-	-
[Luong *et al.*, 2015]	V	10.8	35.1	6.8	28.5
[Bluche, 2016]	V	7.9	24.6	2.9	12.6
[Sueiras *et al.*, 2018]	V	8.8	23.8	4.8	15.9
[Bhunia *et al.*, 2019]	V	8.4	17.2	6.4	10.5
[Zhang *et al.*, 2019]	V	8.5	22.2	-	-
[Wang *et al.*, 2020]	V	6.4	19.6	2.7	8.9
[Cascianelli *et al.*, 2022a]	V	6.8	24.7	4.0	13.7
[Kang *et al.*, 2022]*	V	7.6	24.5	-	-
[Cascianelli *et al.*, 2022b]*	V	7.3	37.5	-	-
PyGT + ATTN (*baseline*)	G	10.2	21.7	5.7	11.7
CRN + ATTN (*baseline*)	V	7.1	17.1	3.4	8.1
+ GCM (ours)	G&V	**5.8**	**14.9**	**2.4**	**6.5**

Table 1: Results for offline HTR on IAM and RIMES. Notably, only the methods without using the extra training data are compared. In the table, V and G denote the visual and graph inputs respectively, and * denotes the Transformer-based models for HTR.

4.2 Comparison with Prior Works

To demonstrate the effectiveness of our method, we compare the proposed GCM with previous SOTA methods for HTR on public benchmarks. It is worth noting that our GCM can recognize both online and offline handwritten texts.

Results for Offline Handwritten Text Recognition As listed in Table 1, we compare our GCM with previous SOTA methods for HTR on two public offline handwritten text datasets IAM and RIMES. For a fair comparison, we only compare the methods without utilizing the extra synthetic or training data in Table 1. It can be observed that most previous methods are only based on the image modality, while graph-enhanced multi-modal methods are rarely explored. Moreover, our method also outperforms previous SOTA methods for offline HTR when no extra training data is utilized. Experiments demonstrate that it can achieve better performance for HTR by exploiting both visual and geometric information of characters, especially for scenarios with limited data.

Results for Online Handwritten Text Recognition It is worth noting that the proposed GCM can be feasibly extended to online HTR. As listed in Table 2, we extended GCM to

Method	CER (%)	WER(%)	
		Lexicon	None
RNN-CTC [Graves *et al.*, 2008]	2.87	3.79	11.47
RNN-ATTN [Gan and Wang, 2019]	3.14	3.39	11.37
CNN-ATTN [Gan *et al.*, 2019]	2.55	3.28	11.58
GCN-CTC [Gan *et al.*, 2023]	2.67	2.77	11.08
GCN-ATTN [Gan *et al.*, 2023]	1.61	2.72	5.13
GCN&CRN-ATTN [Chen *et al.*, 2024]	1.29	2.58	3.76
GCM (ours)	**1.04**	**1.89**	**3.14**

Table 2: Results for online HTR on IAHEW-UCAS2016.

Method	Input	Seq.	Acc(%)
Human Performance [Yin *et al.*, 2013]	V	×	96.13
DFE-DLQDF [Liu *et al.*, 2013]	V	×	92.72
HCCR-GoogleNet [Zhong *et al.*, 2015]	V	×	96.26
DirectMap-CNN [Zhang *et al.*, 2017]	V	×	96.95
M-RBC + IR [Yang *et al.*, 2017]	V	×	97.37
PyGT [Gan *et al.*, 2023]	G	×	96.49
DenseRAN [Wang *et al.*, 2018]	V	✓	96.66
FewshotRAN [Wang *et al.*, 2019]	V	✓	96.97
HDE-Net [Cao *et al.*, 2020]	V	✓	97.14
HCRN [Huang *et al.*, 2022a]	V	✓	96.70
CUE [Luo *et al.*, 2023]	V	✓	96.96
PyGT + RAN $(baseline)$	G	✓	95.20
DenseRAN* $(baseline)$	V	✓	96.74
+ GCM (ours)	G&V	✓	**97.12**

Table 3: Results for handwritten Chinese character recognition on ICDAR2013. In the table, "Seq." denotes whether the recognition method is radical-based, and * denotes our re-implementation.

online HTR on IAHEW-UCAS2016, in which each sample is written in mid-air with a single stroke. Additionally, we mainly compared GCM with prior prevalent approaches, including RNN-CTC, RNN-ATTN, etc. It can be observed that our multi-modal GCM outperforms the uni-modal models (such as RNN-Decoder & GCN-Decoder) and has achieved the new SOTA performance on IAHEW-UCAS2016, demonstrating its effectiveness for online HTR.

Results for Handwritten Chinese Character Recognition We demonstrate that GCM is also effective for radical-based handwritten Chinese character recognition (HCCR), in which handwritten Chinese characters are recognized as radical sequences. As shown in Table 3, we compare the GCM with previous SOTA methods for HCCR on ICDAR-2013. It is worth noting that radical-based methods generally perform inferior to conventional class-based methods, since it is more challenging to simultaneously recognize both individual structures and radicals of Chinese characters. Experimental results demonstrate that our GCM has achieved the SOTA performance among radical-based approaches for HCCR.

4.3 Ablation Study of Proposed Method

To investigate the contributions of individual parts of GCM, we have conducted extensive ablation studies of different modules or learning strategies. For a fair comparison, we applied greedy decoding for testing in all ablation studies.

Effectiveness of Cross-Modal Fusion As shown in Table 4, we compare the performance of different network configurations by integrating different modalities of data. The experimental results reveal several key insights: (i) directly processing images achieves high accuracy, as spatial information plays a crucial role in handwriting recognition; similarly, decoding geometric graphs also yields comparable accuracy, suggesting that geometric structures can also benefit handwriting recognition; (ii) the cross-modal encoder-decoder frameworks can achieve better performance than uni-modal ones, since the former explicitly leverages both the geometric structures and spatial information of handwritten characters; (iii) it can achieve better performance by enhancing the cross-modal fusion and multi-modal learning.

Method	Modality		IAM (%)		RIMES (%)	
	Image	Graph	CER	WER	CER	WER
Image Only	✓	×	7.06	17.05	3.35	8.11
Graph Only	×	✓	10.20	21.68	5.68	11.68
Uni-G→V	✓	✓	7.01	16.91	3.40	7.43
Uni-V→G	✓	✓	6.61	15.66	2.79	6.71
Bi-G→V	✓	✓	6.18	15.26	2.63	6.84
Bi-V→G	✓	✓	6.17	15.34	2.65	6.80

Table 4: Ablation study on the cross-modal fusion. In the table, "Uni-" denotes the uni-decoder models and "Bi-" denotes the bi-decoder models.

Method	Fusion Direct.		IAM (%)		RIMES (%)	
	G→V	V→G	CER	WER	CER	WER
Uni-G→V	✓	×	7.01	16.91	3.40	7.43
Uni-V→G	×	✓	6.61	15.66	2.79	6.71
Twin-G→V	✓	×	6.84	16.07	2.96	7.32
Twin-V→G	×	✓	6.32	15.57	2.74	6.96
Bi-G→V	✓	✓	6.18	15.26	2.63	6.84
Bi-V→G	✓	✓	6.17	15.34	2.65	6.80

Table 5: Ablation study on the bidirectional mutual learning. In the table, "Twin-" denotes the bi-decoders with identical directions, "Bi-" denotes the bi-decoders with inverse directions, and "Fusion Direct." denotes the fusion direction of cross-modal features.

Effectiveness of Bidirectional Mutual Learning As shown in Table 5, we compare the performance of cross-modal models with different fusion directions. It can be observed that: (i) the single-direction fusion (either G→V or V→G) can achieve satisfactory performance, since both geometric and visual features can benefit handwriting recognition; (ii) moreover, fusing cross-modal data with two identical directions can also improve the recognition performance via one-to-one knowledge transfer, but its improvement may become marginal since the branches with identical directions attain mirror differences; (iii) instead, the branches with the inverse directions can learn more distinguishing patterns and thus can be more complementary via mutual learning.

Necessity of Uni-Modal Learning Objectives As shown in Table 6, we compare the performance of cross-modal decoders with or without utilizing $\mathcal{L}_U$. It is worth noting that existing multi-modal fusion generally suffers from imbalanced multi-modal learning problems due to the conflicts of multi-modal optimization directions. However, such a problem can be effectively alleviated by adding uni-modal learning objectives besides multi-modal joint learning. As shown in Table 6, the models jointly optimized with both uni-modal and multi-modal objectives can achieve better performance.

Method	Uni-Modal $\mathcal{L}_U$	IAM (%)		RIMES (%)	
		CER	WER	CER	WER
Uni-G→V	×	10.14	21.56	5.44	10.46
	✓	7.01	16.91	3.40	7.43
Uni-V→G	×	10.46	22.40	5.61	10.64
	✓	6.61	15.66	2.79	6.71
Bi-G→V	×	9.70	21.71	4.57	10.61
	✓	6.18	15.26	2.63	6.84
Bi-V→G	×	9.57	21.31	4.58	10.43
	✓	6.17	15.34	2.65	6.80

Table 6: Effectiveness of uni-modal learning objective.

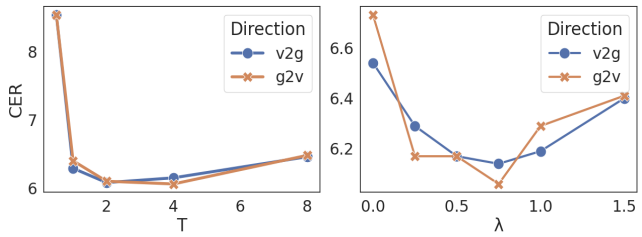

Figure 4: Sensitivity analysis of hyperparameters T and λ in $\mathcal{L}_{KL}$.

Hyperparameter Sensitivity Analysis of $\mathcal{L}_{KL}$ We also perform the sensitivity analysis on the hyperparameters λ and T in $\mathcal{L}_{KL}$, which control the impact of the mutual learning loss between two inverse branches. As shown in Fig. 4, $\mathcal{L}_{KL}$ with $\lambda = 0.75$ and $T = 2$ achieved the best performance.

4.4 Qualitative Analysis

Visualization of Cross-Modal Attention To demonstrate the effectiveness of cross-modal fusion, we compare the attention matrices and visualization results between cross-modal and uni-modal methods as shown in Fig. 5 & 6. It can be observed that uni-modal methods suffer from more served attention drift problems than cross-modal ones. Moreover, the cross-modal guidance can help the attention module more precisely locate the corresponding character regions, thus demonstrating the effectiveness of cross-modal fusion.

Visualization of Feature Distribution To demonstrate that GCM can learn more discriminative features, we further visualize feature distributions of lowercase characters between the cross-modal and uni-modal methods as shown in Fig. 7.

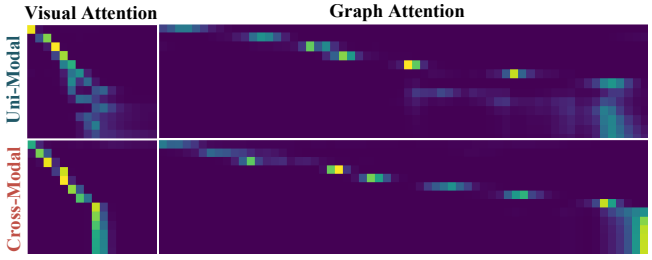

Figure 5: Comparison of cross-modal attention matrices.

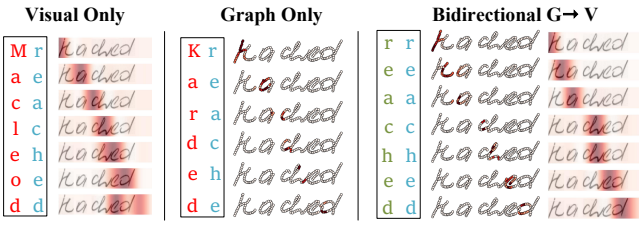

Figure 6: Comparison of cross-modal attention visualization.

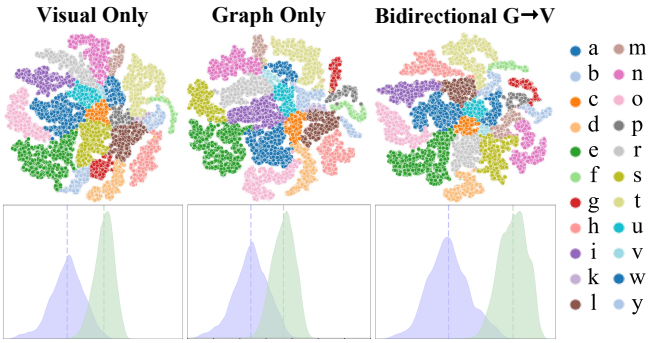

Figure 7: Comparison of feature distribution visualization.

We can observe that cross-modal methods can learn more discriminative feature distributions.

5 Conclusion

Most existing approaches for HTR typically operate on the visual modality alone, while ignoring the importance of the intrinsic geometric structures of characters. Instead, this paper aims to achieve a structure-aware handwriting recognition method that explicitly incorporates both geometric structures and visual appearances of characters. Particularly, we propose a novel graph-enhanced cross-modal mutual learning network GCM for handwritten text recognition, which can fully process handwriting images alongside their corresponding geometric graphs. Specifically, GCM first extracts both visual and geometric information from the cross-modal inputs, and then it utilizes two parallel decoders to sequentially aggregate cross-modal features in inverse orders (V→G and G→V). Furthermore, the two inverse branches can further enhance each other via bidirectional mutual learning, which involves one-to-one knowledge transfer and fully leverages complementary cross-modal information. Notably, only one decoding branch of GCM will be activated in inference, thus avoiding the increase of the model parameters and computation costs for testing and deployment. Extensive experiments demonstrate the effectiveness of the proposed GCM for HTR.

6 Limitation

The proposed GCM is unsuitable for recognizing complex scene texts, since it is challenging to obtain geometric graphs from scene text images when compared to handwriting images. Nevertheless, we leave this problem as our future work.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62206035, 62472060, and 62276258, the Science and Technology Innovation Key R&D Program of Chongqing under Grant No. CSTB2024TIAD-STX0023 and CSTB2023TIAD-STX0016, the Surface Project of Natural Science Foundation of Chongqing under Grant No. CSTB2022NSCQ-MSX0547, the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant No. KJQN202200612 and KJQN202400648, China Postdoctoral Science Foundation under Grant No. GZC20233362 and 2024MD754043, and the Open Projects Program of State Key Laboratory of Multimodal Artificial Intelligence Systems under Grant No. MAZS-2023-11.

References

[Bhunia *et al.*, 2019] Ayan Kumar Bhunia, Abhirup Das, Ankan Kumar Bhunia, Perla Sai Raj Kishore, and Partha Pratim Roy. Handwriting recognition in low-resource scripts using adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4767–4776, 2019.

[Bluche, 2016] Théodore Bluche. Joint line segmentation and transcription for end-to-end handwritten paragraph recognition. *Advances in Neural Information Processing Systems*, 29, 2016.

[Cao *et al.*, 2020] Zhong Cao, Jiang Lu, Sen Cui, and Changshui Zhang. Zero-shot handwritten Chinese character recognition with hierarchical decomposition embedding. *Pattern Recognition*, 107:107488, 2020.

[Cascianelli *et al.*, 2022a] Silvia Cascianelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Boosting modern and historical handwritten text recognition with deformable convolutions. *International Journal on Document Analysis and Recognition (IJDAR)*, 25(3):207–217, 2022.

[Cascianelli *et al.*, 2022b] Silvia Cascianelli, Vittorio Pippi, Martin Maarand, Marcella Cornia, Lorenzo Baraldi, Christopher Kermorvant, and Rita Cucchiara. The LAM dataset: a novel benchmark for line-level handwritten text recognition. In *International Conference on Pattern Recognition (ICPR)*, pages 1506–1513. IEEE, 2022.

[Chen *et al.*, 2024] Yuyan Chen, Xing Zhao, Ji Gan, Jiaxu Leng, Yan Zhang, and Xinbo Gao. Structure-aware in-air handwritten text recognition with graph-guided cross-modality translator. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5020–5024. IEEE, 2024.

[Coquenet *et al.*, 2020] Denis Coquenet, Clément Chatelain, and Thierry Paquet. Recurrence-free unconstrained handwritten text recognition using gated fully convolutional network. In *International Conference on Frontiers in Handwriting Recognition*, pages 19–24. IEEE, 2020.

[Espana *et al.*, 2010] Salvador Espana, Boquera, Maria Jose Castro Bleda, Jorge Gorbe Moya, and Francisco Zamora Martinez. Improving offline handwritten text recognition with hybrid HMM/ANN models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):767–779, 2010.

[Gan and Wang, 2019] Ji Gan and Weiqiang Wang. In-air handwritten English word recognition using attention recurrent translator. *Neural Computing and Applications*, 31:3155–3172, 2019.

[Gan *et al.*, 2019] Ji Gan, Weiqiang Wang, and Ke Lu. A new perspective: Recognizing online handwritten Chinese characters via 1-dimensional CNN. *Information Sciences*, 478:375–390, 2019.

[Gan *et al.*, 2023] Ji Gan, Yuyan Chen, Bo Hu, Jiaxu Leng, Weiqiang Wang, and Xinbo Gao. Characters as graphs: Interpretable handwritten Chinese character recognition via pyramid graph transformer. *Pattern Recognition*, 137:109317, 2023.

[Graves and Schmidhuber, 2008] Alex Graves and Jürgen Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. *Advances in Neural Information Processing Systems*, 21, 2008.

[Graves *et al.*, 2008] Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868, 2008.

[Grosicki *et al.*, 2009] Emmanuèle Grosicki, Matthieu Carré, Jean-Marie Brodin, and Edouard Geoffrois. Results of the rimes evaluation campaign for handwritten mail processing. In *International Conference on Document Analysis and Recognition*, pages 941–945. IEEE, 2009.

[Guo *et al.*, 2020] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11020–11029, 2020.

[Huang *et al.*, 2022a] Guanjie Huang, Xiangyu Luo, Shaowei Wang, Tianlong Gu, and Kaile Su. Hippocampus-heuristic character recognition network for zero-shot learning in Chinese character recognition. *Pattern Recognition*, 130:108818, 2022.

[Huang *et al.*, 2022b] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably). In *International Conference on Machine Learning*, pages 9226–9259. PMLR, 2022.

[Ingle *et al.*, 2019] R. Reeve Ingle, Yasuhisa Fujii, Thomas Deselaers, Jonathan Baccash, and Ashok C. Popat. A scalable handwritten text recognition system. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 17–24. IEEE, 2019.

[Kang *et al.*, 2022] Lei Kang, Pau Riba, Marçal Rusiñol, Alicia Fornés, and Mauricio Villegas. Pay attention to what you read: non-recurrent handwritten text-line recognition. *Pattern Recognition*, 129:108766, 2022.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2015.

[Li *et al.*, 2023] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102, 2023.

[Lipton *et al.*, 2015] Zachary C. Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

[Liu *et al.*, 2013] Chenglin Liu, Fei Yin, Dahan Wang, and Qiufeng Wang. Online and offline handwritten Chinese character recognition: benchmarking on new databases. *Pattern Recognition*, 46(1):155–162, 2013.

[Luo *et al.*, 2023] Guofeng Luo, Dahan Wang, Xia Du, Huayi Yin, Xuyao Zhang, and Shunzhi Zhu. Self-information of radicals: A new clue for zero-shot Chinese character recognition. *Pattern Recognition*, 140:109598, 2023.

[Luong *et al.*, 2015] Minh Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

[Marti and Bunke, 2002] U.-V. 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.

[Paszke *et al.*, 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[Peng *et al.*, 2022] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022.

[Shi *et al.*, 2016] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304, 2016.

[Sueiras *et al.*, 2018] Jorge Sueiras, Victoria Ruiz, Angel Sanchez, and Jose F Velez. Offline continuous handwriting recognition using sequence to sequence neural networks. *Neurocomputing*, 289:119–128, 2018.

[Wang *et al.*, 2018] Wenchao Wang, Jianshu Zhang, Jun Du, Zi-Rui Wang, and Yixing Zhu. DenseRAN for offline handwritten Chinese character recognition. In *International Conference on Frontiers in Handwriting Recognition*, 2018.

[Wang *et al.*, 2019] Tianwei Wang, Zecheng Xie, Zhe Li, Lianwen Jin, and Xiangle Chen. Radical aggregation network for few-shot offline handwritten Chinese character recognition. *Pattern Recognition Letters*, 125:821–827, 2019.

[Wang *et al.*, 2020] Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and Mingxiang Cai. Decoupled attention network for text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12216–12224, 2020.

[Yang *et al.*, 2017] Xiao Yang, Dafang He, Zihan Zhou, Daniel Kifer, and C. Lee Giles. Improving offline handwritten Chinese character recognition by iterative refinement. In *International Conference on Document Analysis and Recognition*, volume 1, pages 5–10. IEEE, 2017.

[Yao *et al.*, 2019] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 33, pages 7370–7377, 2019.

[Yin *et al.*, 2013] Fei Yin, Qiufeng Wang, Xuyao Zhang, and Chenglin Liu. ICDAR 2013 Chinese handwriting recognition competition. In *International Conference on Document Analysis and Recognition*, pages 1464–1470. IEEE, 2013.

[Yousef *et al.*, 2020] Mohamed Yousef, Khaled F Hussain, and Usama S Mohammed. Accurate, data-efficient, unconstrained text recognition with convolutional neural networks. *Pattern Recognition*, 108:107482, 2020.

[Zhang *et al.*, 2017] Xuyao Zhang, Yoshua Bengio, and Chenglin Liu. Online and offline handwritten Chinese character recognition: A comprehensive study and new benchmark. *Pattern Recognition*, 61:348–360, 2017.

[Zhang *et al.*, 2018] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[Zhang *et al.*, 2019] Yaping Zhang, Shuai Nie, Wenju Liu, Xing Xu, Dongxiang Zhang, and Heng Tao Shen. Sequence-to-sequence domain adaptation network for robust text image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 2740–2749, 2019.

[Zhong *et al.*, 2015] Zhuoyao Zhong, Lianwen Jin, and Zecheng Xie. High performance offline handwritten Chinese character recognition using googlenet and directional feature maps. In *International Conference on Document Analysis and Recognition*, pages 846–850. IEEE, 2015.

[Zhu *et al.*, 2018] Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. *Advances in Neural Information Processing Systems*, 31, 2018.