SSAN: A Symbol Spatial-Aware Network for Handwritten Mathematical Expression Recognition

Haoran Zhang[1,2,3], **Xiangdong Su**[1,2,3 *], **Xingxiang Zhou**[1,2,3], **Guanglai Gao**[1,2,3]

[1]College of Computer Science, Inner Mongolia University, China
[2]National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, China
[3]Inner Mongolia Key Laboratory of Multilingual Artificial Intelligence Technology, China
zhryidr962@gmail.com, cssxd@imu.edu.cn, zxx.w@outlook.com, csggl@imu.edu.cn

Abstract

The great challenge of handwritten mathematical expression recognition (HMER) is the complex structures of the expressions, which are directly related to the symbol spatial positions. Existing HMER methods typically employ attention mechanisms in the decoder of their models to implicitly perceive the symbol positions, or employ symbol counting and tree-based strategies to model the symbol spatial relation. However, these methods still cannot effectively capture the structural information of formulas, thus negatively impacting the symbol decoding in HMER. To deal with this problem and enhance the HMER performance, this paper proposes a novel auxiliary task, namely predicting the symbol spatial distribution map of handwritten expression images. On such basis, this paper designs a symbol spatial-aware network (SSAN) for this task, which is jointly optimized with the HMER model. Specifically, considering the similarity of the symbol spatial positions between the handwritten mathematical expression images and their corresponding printed templates, we obtain the symbol spatial distribution map by first generating printed templates from LaTeX ground-truth for handwritten formula images and then replacing the connected components of printed templates with 2D Gaussian distribution maps of the same size. Meanwhile, due to the loose alignment of the symbol spatial positions between handwritten and printed formula images, and misclassification of similar symbols, we further propose a coarse-to-fine alignment strategy and an attention-guided symbol masking strategy in SSAN to tackle these issues. Extensive experiments demonstrate that SSAN significantly improves the recognition performance of the HMER models, and the proposed auxiliary tasks are more effective in enhancing HMER performance than existing auxiliary tasks. Code is available at https://github.com/Howrunz/SSAN.

Introduction

Handwritten Mathematical Expression Recognition (HMER) aims to convert formula images into LaTeX sequences, which is widely used in answer sheet scoring, office automation, and document understanding. Unlike handwritten text line recognition, HMER faces not only diverse handwriting styles but also complex spatial relationships (Zhou et al. 2013; Li et al. 2019; He, Tan, and Bi 2020). These challenges complicate the accurate modeling of long-distance correlations and adherence to LaTeX grammatical specifications (Zhao et al. 2021), especially with nested structures, lengthy sequences, and other intricate configurations. As a result, the recognition accuracy often varies with the complexity of the mathematical expression (ME) structure, further reducing overall performance.

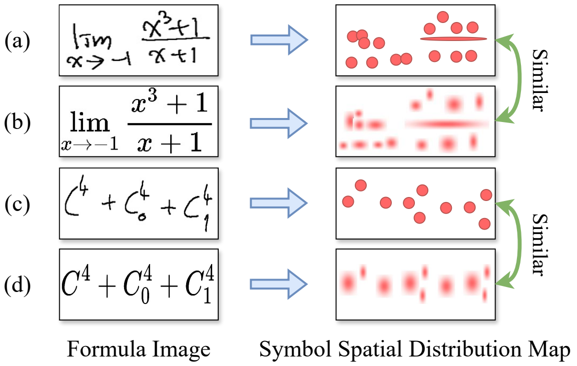

(a)
(b)
(c)
(d)
$\lim_{x \to -1} \frac{x^3+1}{x+1}$
$C^4 + C_0^4 + C_1^4$
Similar
Similar
Formula Image
Symbol Spatial Distribution Map

Figure 1: Illustration of symbol spatial distribution maps for handwritten mathematical expression images and their printed templates. For easy comparison, the symbol spatial distribution maps of (a) and (c) are manually annotated, while (b) and (d) are generated using the proposed method.

Prevalent approaches regard HMER as an image-to-sequence task and design attention-based encoder-decoder models for this task. These models depend on the attention mechanism to facilitate the symbol decoding in generating the target LaTeX sequences. For example, WAP (Zhang et al. 2017) incorporated a coverage attention mechanism to improve recognition accuracy. ABM (Bian et al. 2022) proposed a bidirectional mutual learning method to help the model understand the long-distance correlations in MEs. Some recent works focus on enhancing the capability of recognizers to understand the positional and hierarchical relationships among symbols in formula images. CAN (Li et al. 2022) introduced a symbol counting task with a weakly supervised counting module to ensure precise alignment between attention features and symbol regions. BPD (Li et al. 2024) proposed a tree-based model for explicit symbol recognition and relational prediction between sym-

[*]Corresponding author.

bols. Although these methods advance the HMER, they still cannot effectively capture the symbol spatial information and thus negatively impact the symbol decoding in HMER.

To better perceive the structural information of MEs, we propose a novel auxiliary task for HMER, namely predicting the symbol spatial distribution map of handwritten expression images. On such basis, we design a symbol spatial-aware network (SSAN) for this task, and jointly optimize it with the HMER model. The motivations are based on our findings that the spatial distribution of formula symbols closely relates to their semantic relationships. People usually write formulas following standard formats, like those in printed mathematical expressions in textbooks, which leads to an inherent similarity in the spatial distribution of symbols between handwritten MEs and their printed templates, as shown in Fig. 1. Therefore, predicting the symbol spatial distribution map can explicitly enforce the model to capture each symbol, thus benefiting the HMER task. Specifically, we first generate the symbol spatial distribution map by applying connected component analysis to printed templates produced from target LaTeX sequences and creating 2D Gaussian distribution maps based on the centers of these connected components. Next, we incorporate SSAN into an attention-based encoder-decoder HMER model and jointly optimize them within a multi-task framework to mitigate the attention drift problem of the HMER model.

However, SSAN still faces two challenges: (1) the symbol spatial distribution in handwritten formula images does not fully align with printed templates; (2) there are many misclassifications among similar symbols. The reason for the first challenge is that symbol variation and position shift are natural in handwritten MEs. The second challenge arises because the HMER model does not entirely learn the discriminative features of similar symbols. Therefore, we propose two strategies in SSAN to deal with the above challenges, including coarse-to-fine alignment strategy and attention-guided symbol masking strategy. For the coarse-to-fine alignment strategy, SSAN's prediction target transitions from the symbol spatial distribution map of printed images in the first training stage to the predicted spatial attention map of the well-trained HMER model in the second training stage. For the attention-guided symbol masking strategy, we randomly mask some parts of symbols based on the spatial attention maps to force the HMER model to focus on the discriminative features of similar symbols.

Experimental results show that integrating SSAN as well as the two training strategies into baseline models consistently improves the recognition accuracy. Significantly, DWAP+SSAN outperforms DWAP by 9.37%, 9.37%, 11.79%, and 6.74% on the CROHME 2014, CROHME 2016, CROHME 2019, and HME100K datasets, respectively. The main contributions of this paper are as follows:

- We propose an auxiliary task that predicts the symbol spatial distribution map to facilitate the HMER model to capture the formula structure better, and introduce a 2D Gaussian distribution map of symbols as symbol spatial distribution map. To the best of our knowledge, this is the first design of such a task to enhance HMER.
- We design the symbol spatial-aware network (SSAN) to predict symbol spatial distribution and jointly optimize the HMER model within a multi-task framework. Experimental results demonstrate that incorporating the SSAN into baseline models can obtain consistent improvement in recognition accuracy.
- We propose two strategies to further promote the SSAN in HMER. The coarse-to-fine training strategy addresses the loose alignment of symbol spatial distribution between handwritten formula images and printed templates, and the attention-guided masking strategy mitigates the misclassification of similar symbols.

Related Work

HMER Methods

With the success of deep neural networks, the encoder-decoder framework (Cho et al. 2014; Sutskever, Vinyals, and Le 2014) showed promise in recognizing handwritten mathematical formulas. (Deng et al. 2017) first applied this approach to HMER, and it gained popularity after Zhang et al. (Zhang et al. 2017) noted its similarities to machine translation. They enhanced it with a coverage attention mechanism (Tu et al. 2016), resulting in a more powerful WAP model. (Zhang, Du, and Dai 2018) further improved WAP by using DenseNet (Huang et al. 2017) as an encoder for better handling of multi-scale symbols and introduced a tree decoder in DWAP-TD (Zhang et al. 2020) for complex formulas. Building on this, SAN (Yuan et al. 2022) introduced grammatical rules into the encoder-decoder network, improving its ability to predict formula structures by translating LaTeX tokens into parsing trees.

Following the success of the Transformer (Vaswani et al. 2017), (Zhao et al. 2021) first adapted it for bidirectional mathematical formula modeling. (Ding, Chen, and Huo 2021) enhanced this with multi-head attention and a stacked decoder. (Zhao and Gao 2022) proposed CoMER, a Transformer-based model with an attention refinement module leveraging self-coverage and cross-coverage techniques. Other approaches focused on creative data augmentation. (Li et al. 2020) proposed a random scale enhancement strategy, and (Yang et al. 2022) proposed a tree-based multi-level augmentation strategy.

Multi-Task Learning

Multi-task learning leverages shared knowledge among different tasks to prevent overfitting, enhance feature representation, and improve performance and generalization. (Wu et al. 2018, 2020) were among the first to incorporate adversarial learning tasks in recognizing mathematical formulas. (Truong et al. 2020) introduced a weakly supervised symbol classification task, enhancing the encoder's feature representation. (Bian et al. 2022) proposed ABM, a bi-directional mutual learning network with a shared encoder and two decoders working in opposite directions, improved by attention aggregation. (Li et al. 2022) added a weakly-supervised counting module to HMER for predicting symbol quantities, thus correcting attention mechanism errors. (Fu et al. 2023) proposed SLAN, using relation-level counting and a

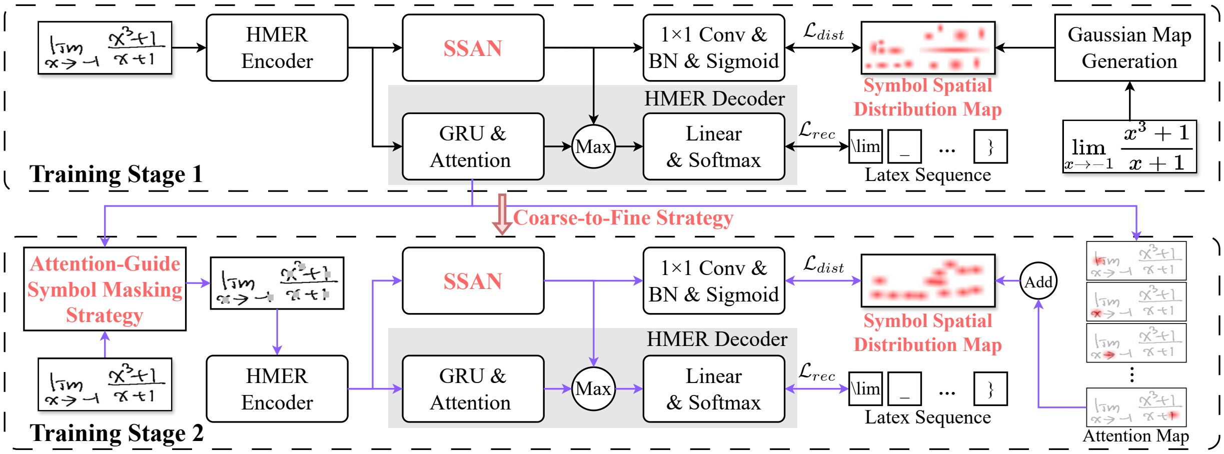


Figure 2: The overall architecture of the HMER model incorporates the symbol spatial-aware network (SSAN).

context-aware decoder to address attention issues. (Liu et al. 2023) introduced a semantic interaction learning method SAM with a semantic-aware module to enhance symbol interaction. (Zhang et al. 2023) proposed GCN for coarse-grained classification of mathematical symbols.

The multi-task learning methods previously mentioned direct the model's perception of structural information in handwritten mathematical formulas by considering semantic consistency, leveraging weak supervision, and considering the sequence generation order. Unlike these methods, we propose an auxiliary task designed to predict the symbol spatial distribution map within the mathematical formula images. This task aids the model in intuitively understanding the positional relationships between symbols, thereby effectively capturing the structural information of MEs.

Methodology

Overview Architecture

Given the crucial influence of symbol position on how the model understands the formula, we aim to strengthen the HMER model's ability of capturing the spatial structure of formula images and understanding the relations between formula images and their target LaTeX sequences. Since there is inherent similarity in the spatial positions of the symbols between handwritten mathematical expressions and their printed templates, this paper proposes an auxiliary task of predicting the symbol spatial distribution map of the handwritten formula image and specifically designs a symbol spatial-aware network (SSAN) for this task to improve the HMER performance. We incorporate SSAN into an attention-based encoder-decoder HMER model and jointly optimize the HMER model and SSAN within a multi-task framework. Meanwhile, we propose a coarse-to-fine alignment strategy and an attention-guided symbol masking strategy for HMER to tackle the loose alignment of the symbol spatial distribution and the misclassification of similar symbols, respectively. In real applications, the HMER encoder and HMER decoder can be replaced with any HMER model that uses spatial attention. In the following sections, we use DWAP (Zhang, Du, and Dai 2018) as the HMER model to illustrate our approach.

The overall training architecture is shown in Fig. 2. According to the proposed coarse-to-fine strategy, the training process is divided into two stages. In the first stage, we jointly optimize the HMER model with the SSAN. Here, the learning target of SSAN is the symbol spatial distribution map of the printed template corresponding to the input formula images. Through joint optimization with SSAN in the first stage, the decoder of the HMER model has been able to learn better attention maps than without SSAN. In spite of this, it is worth noting that there is loose alignment between the symbol spatial distribution maps of the handwritten mathematical expressions and that of their printed templates. To ensure the self-consistency of the HMER model (Farquhar et al. 2021; Bonatti and Mohr 2022; Wang et al. 2023), we adjust the learning target of SSAN in the second stage of training to the attention graph of the trained HMER decoder in the first stage. Through the coarse-to-fine strategy, the SSAN can enable the HMER model to better capture the structure information of formula images and achieve higher recognition performance.

To avoid the misclassification of similar symbols in HMER, we also design an attention-guided masking strategy and use it in the second training stage. According to this strategy, we mask a small part of the symbols in the input formula images base on the learned attention maps. The strategy enforces the HMER models to further learn the discriminative features of similar symbols, thus improving the HMER performance. The details of our methodology are described in the following sections.

Symbol Spatial Distribution Map

We generate printed formula images using word-level HMER labels and an open-source LaTeX rendering tool. We then conduct a connected component analysis on the printed images, yielding connected components $C = C_1, C_2, \ldots, C_N$, where N is the total number of components and $C_n = P_1, P_2, \ldots, P_M$ denotes the M pixel coordinates within a given component. The center coordinates

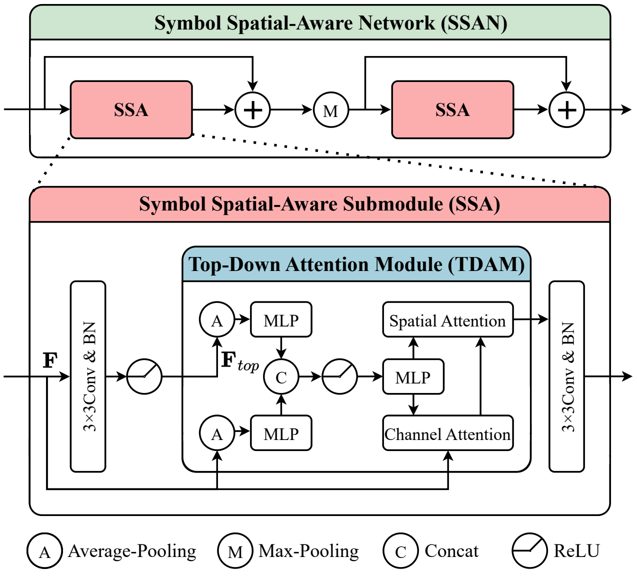


Figure 3: The architecture of the symbol spatial-aware network (SSAN), consisting of two SSA submodules.

of each connected component can be calculated using

$$x_c = \frac{x_1 + x_2 + ... + x_M}{M},\ y_c = \frac{y_1 + y_2 + ... + y_M}{M}, \quad (1)$$

where (x, y) is the pixel coordinate within the connected component, and (x_c, y_c) is its center point.

However, the symbol spatial distribution map generated using only the center points of connected components is sparse. This sparsity limits the model's ability to learn structural details in MEs. Sparse maps require more resources to handle zero values during training, which will lengthen the training time, lead to the risks of over-fitting on nonzero data, and lower the generalization. Therefore, we employ a 2D Gaussian map to better represent the spatial distribution of symbols.

We generate the corresponding 2D Gaussian map centered on the midpoint of the connected component, using

$$G(x, y) = \exp\left(-\frac{(x - x_c)^2}{2\sigma_x^2} - \frac{(y - y_c)^2}{2\sigma_y^2}\right), \quad (2)$$

where the σ_x and σ_y are set to $1/4$ of the width and height of the bounding box enclosing the connected component, respectively. This guarantees that the generated 2D Gaussian map corresponds closely to the size of the connected component.

Symbol Spatial-Aware Network

This paper designs a symbol spatial-aware network (SSAN) to predict the symbol spatial distribution map, which is illustrated in Fig. 3. The SSAN consists of two SSA submodules, allowing SSAN to extract structural information at multiple scales. Residual connection and max-pooling efficiently combine the outputs from previous modules, improving feature integration and minimizing the loss of key information.

Given an encoder feature map $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$, we first use a convolutional layer with a 3×3 kernel to extract high-level semantic information, denoted as $\mathbf{F}_{top}$, formulated as:

$$\mathbf{F}_{top} = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}(\mathbf{F}))). \quad (3)$$

Subsequently, we introduce the Top-Down Attention Module (TDAM) (Jaiswal, Fernando, and Tan 2022) to enhance the model focus on symbol spatial positions within the lower-level local features, thus amplifying the critical region features. The TDAM jointly models high-level semantic and lower-level features to generate a guidance map $\mathbf{G}$, indicating key channels and spatial positions within the lower-level features:

$$\mathbf{G} = \mathbf{W}_G(\mathrm{ReLU}([\mathbf{W}_b \mathrm{Pool}(\mathbf{F}); \mathbf{W}_t \mathrm{Pool}(\mathbf{F}_{top})])), \quad (4)$$

where $\mathbf{W}_G$, $\mathbf{W}_b$ and $\mathbf{W}_t$ denote the weights of MLP, and Pool denotes the spatial average pooling operation. Next, channel attention is computed based on the attention guidance map, emphasizing channels containing key features:

$$\mathbf{F}_{channel} = \sigma(\mathbf{G}) \odot \mathbf{F}, \quad (5)$$

where σ denotes the sigmoid activation function, and $\odot$ denotes element-wise matrix multiplication. We further enhance the spatial position representations of the relevant symbol regions within channels that are recognized to contain crucial features through spatial attention:

$$\mathbf{F}_{spatial} = \mathbf{F}_{channel} \odot \sigma(\mathbf{G} * \mathbf{F}_{channel}), \quad (6)$$

where $*$ denotes pointwise convolution, it focuses attention on specific spatial positions, thereby enhancing the feature representation at these positions. We then use a 3×3 kernel convolutional layer to refine the features extracted from TDAM, enhancing their suitability for subsequent processing. The output symbol spatial distribution feature $\mathbf{L}$ of SSA submodule written as:

$$\mathbf{L} = \mathrm{BN}(\mathrm{Conv}(\mathbf{F}_{spatial})). \quad (7)$$

Joint Optimization of SSAN and HMER Model

To enhance the HMER model, we jointly optimize the SSAN and HMER model within a multi-task framework. The learning target of SSAN is the symbol spatial distribution map, while the target of the HMER model is the ground-truth LaTeX sequence. The symbol spatial distribution map from the SSAN branch provides complementary information to the HMER branch and alleviates attention drift. We use the cross-entropy function to compute the loss for the HMER task:

$$L_{rec} = -\sum_i \sum_c Y_{i,c} \log(P_{i,c}), \quad (8)$$

where i and c denote the index of the sample and category.

Since the target Gaussian distribution map is a single-channel map with values in the range $[0, 1]$, we employ a mapping head constructed with a 1×1 convolutional layer, batch normalization and a sigmoid function to transform the symbol spatial distribution feature into a single-channel map. We then use the smooth L1 loss (Ren et al. 2015) to compute the distance between the predicted symbol spatial distribution map S_{pred} and the ground-truth distribution map S:

$$L_{dist} = \begin{cases} 0.5(S_{pred} - S)^2, & |S_{pred} - S| < 1 \\ |S_{pred} - S| - 0.5, & otherwise. \end{cases} \tag{9}$$

We empirically combine the losses of the two branches as the total loss function for joint optimization:

$$L = L_{rec} + L_{dist}. \tag{10}$$

Comprehensive Strategies to Refine SSAN

Coarse-to-Fine Alignment Strategy. As mentioned, the proposed auxiliary task is to explicitly constrain the spatial attention of HMER. Concerning the similarity of the symbol spatial position between the handwritten MEs and their corresponding printed templates, we use the symbol spatial distribution map of the printed template corresponding to the input formula images as the learning target of SSAN. Since there is loose alignment between the symbol spatial position maps of the handwritten MEs and their corresponding printed templates, we propose coarse-to-fine alignment strategy. According to this strategy, we divide the training process into two stages. In the first stage, the learning target of SSAN is the symbol spatial distribution map of the printed template, while in the second stage, the learning target of SSAN is the attention map of the trained HMER decoder in the first stage. Through the coarse-to-fine strategy of adjusting the training target of SSAN, we can ensure the self-consistency of the HMER model and make the HMER model better capture the structure information of formula images, thereby improving overall recognition accuracy and robustness.

Attention-Guided Symbol Masking Strategy. According to our analysis, the misclassification of similar symbols arises from the unsatisfactory attention maps. For example, if both the attention maps focus on the top part of '9' and 'q', it may lead to errors. Therefore, we propose an attention-guided symbol masking strategy to enforce the HMER model to learn the discriminative features of similar symbols. Specifically, in the second training stage, we randomly mask a small part of the symbols according to the learned attention map from the first stage. Through this strategy, the HMER model tends to learn useful features beyond the attention areas for discriminating the similar symbols, such as the lower parts of '9' and 'q'.

Experiments

Datasets and Metrics

To evaluate the effectiveness of the proposed method, we conduct the experiment on the benchmark datasets named CROHME (Mouchère et al. 2014) and HME100K (Yuan et al. 2022). We use the Expression Recognition Rate (ExpRate), ≤ 1 Error, and ≤ 2 Error as the metrics to evaluate the performance of different methods in HMER.

Implement Details

In this paper, we jointly optimize SSAN and the baseline HMER models, using the same optimizer and learning rate adjustment strategy in the baseline models. The output feature channels for the two SSA submodules are 600 and 432. The kernel size and stride of the max-pooling layer are both set to 2. All models are trained on two NVIDIA V100 32GB GPUs, and the batch size is 20. For a fair comparison, we provide results with and without data augmentation. We employ scale augmentation (Li et al. 2020) with a scaling factor from 0.7 to 1.4 in the experiment with data augmentation.

Comparison with State-of-the-Art Methods

We integrate SSAN with the most representative DWAP (Zhang, Du, and Dai 2018), SAM (Liu et al. 2023) and CAN (Li et al. 2022), and compare them with previous state-of-the-art methods, including WAP (Zhang et al. 2017), DWAP-MSA (Zhang, Du, and Dai 2018), WS-WAP (Truong et al. 2020), PAL-v2 (Wu et al. 2020), DWAP-TD (Zhang et al. 2020), BTTR (Zhao et al. 2021), TDv2 (Wu et al. 2022), TSDNet (Zhong et al. 2022), ABM (Bian et al. 2022), SAN (Yuan et al. 2022), CoMER (Zhao and Gao 2022), SLAN (Fu et al. 2023), SAM (Liu et al. 2023), GCN (Zhang et al. 2023), and BPD-Coverage (Li et al. 2024).

Table 1 shows the results of baselines and the models with SSAN. From Table 1, we can see that SSAN significantly improves the performance of SAM-DWAP, DWAP, and CAN-DWAP, demonstrating its effectiveness and generalizability. Specifically, DWAP+SSAN demonstrates the most notable performance improvements on the CROHME 2014, 2016, 2019, and HME100K, with increases of 9.37%, 9.37%, 11.79%, and 6.74%, respectively. Additionally, DWAP+SSAN achieves state-of-the-art performance on the CROHME 2014, while CAN-DWAP+SSAN leads on the other three datasets. Moreover, on the three CROHME benchmarks, the baseline models combined with SSAN outperform the recent state-of-the-art method of BPD-Coverage. These results prove that the task of predicting the symbol spatial distribution benefits the HMER decoding, and the proposed SSAN can promote the HMER model to better capture the spatial structure of MES. It also suggests that the coarse-to-fine alignment and attention-guided symbol masking strategies can deal with the loose alignment of distribution maps between handwritten MEs and their printed templates and the misclassification of similar symbols.

Comparison with State-of-the-Art Methods with Data Augmentation

Table 2 presents the results after data augmentation under the same settings. Since most previous methods did not employ data augmentation, we neglect them and focus on the recent baselines with data augmentation. The results still exhibit the effectiveness of SSAN in enhancing the HMER models. The significant performance improvement from the proposed method underscores the great potential of combining symbol spatial distribution maps with multi-task learning in the HMER task.

Methods	From	CROHME 2014			CROHME 2016			CROHME 2019			HME100K		
		ExpRate↑	$\leq 1 \uparrow$	$\leq 2 \uparrow$	ExpRate↑	$\leq 1 \uparrow$	$\leq 2 \uparrow$	ExpRate↑	$\leq 1 \uparrow$	$\leq 2 \uparrow$	ExpRate↑	$\leq 1 \uparrow$	$\leq 2 \uparrow$
DWAP-MSA	PR 2018	52.80	68.10	72.00	50.10	63.80	67.40	–	–	–	–	–	–
WS-WAP	ICFHR 2020	53.65	–	–	51.96	64.34	70.10	–	–	–	–	–	–
PAL-v2	IJCV 2020	48.88	64.50	69.78	49.61	64.08	70.27	–	–	–	–	–	–
DWAP-TD	ICML 2020	49.10	64.20	67.80	48.50	62.30	65.30	51.40	66.10	69.10	62.60	79.05	85.67
BTTR	ICDAR 2021	53.96	66.02	70.28	52.31	63.90	68.61	52.96	65.97	69.14	64.10	–	–
TDv2	AAAI 2022	53.62	–	–	55.18	–	–	58.72	–	–	–	–	–
TSDNet	MM 2022	54.70	68.85	75.58	52.48	68.26	73.41	56.34	72.97	77.84	–	–	–
ABM	AAAI 2022	56.85	73.73	81.24	52.92	69.66	78.73	53.96	71.06	78.65	65.93	81.16	87.86
SAN	CVPR 2022	56.20	72.60	79.20	53.60	69.60	76.80	53.50	69.30	70.10	67.10	–	–
CAN-ABM	ECCV 2022	57.26	74.52	82.03	56.15	72.71	80.30	55.96	72.73	80.57	68.09	83.22	89.91
SLAN	ICMR 2023	56.18	73.02	81.36	54.83	72.98	80.04	54.46	73.19	80.53	–	–	–
SAM-CAN	ICDAR 2023	58.01	–	–	56.67	–	–	57.96	–	–	68.81	–	–
GCN*	ICASSP 2023	60.00	–	–	58.94	–	–	61.63	–	–	–	–	–
BPD-Coverage	PR 2024	60.65	–	–	58.50	–	–	61.47	–	–	–	–	–
DWAP	PR 2018	51.48	67.01	73.30	50.65	63.30	70.88	50.04	65.39	69.39	61.85	70.63	77.14
DWAP + SSAN	Ours	**60.85**	**75.56**	**82.25**	**60.02**	**76.22**	**83.28**	**61.83**	**79.08**	**86.08**	**68.74**	**84.42**	**90.10**
SAM-DWAP	ICDAR 2023	56.90	73.33	80.73	55.80	73.23	81.17	57.71	75.73	84.07	68.14	84.07	89.67
SAM-DWAP + SSAN	Ours	**60.04**	**76.06**	**82.76**	**59.37**	**74.89**	**83.35**	**60.63**	**77.73**	**84.49**	**68.88**	**84.58**	**90.32**
CAN-DWAP	ECCV 2022	57.00	74.21	80.61	56.06	71.49	79.51	54.88	71.98	79.40	67.31	82.93	89.17
CAN-DWAP + SSAN	Ours	**59.03**	**75.46**	**82.05**	**60.07**	**74.89**	**82.04**	**62.14**	**80.32**	**86.41**	**69.28**	**85.23**	**90.62**

Table 1: Comparison results with the previous SOTA methods. * represents the use of additional information.

Method	CROHME		
	2014	2016	2019
DWAP	57.91	55.88	56.79
CoMER	59.33	59.81	62.97
CAN-DWAP	60.65	59.11	62.80
DWAP + SSAN	**62.58**	**62.51**	**65.30**

Table 2: Comparison results with data augmentation. **All methods use the same augmentation setting.**

Comparison with Other Auxiliary Task

In fact, CAN (Li et al. 2022) and SAM (Liu et al. 2023) also design auxiliary tasks and modules to improve the HMER model DWAP. We conduct experiments to compare these two auxiliary tasks with our proposed task of predicting the symbol spatial distribution map by introducing their corresponding modules to DWAP. The results are shown in Table 3.

Table 3 shows that these three auxiliary tasks improve the performance of DWAP across all CROHME datasets, respectively. Among these three methods, SSAN performs the best, demonstrating its effectiveness in enhancing DWAP through symbol spatial distribution prediction. From Table 3, we can find that combining SSAN with CAN or SAM (No. 6, No. 7) brings little performance improvement, or even lowers the performance on some datasets when compared with SSAN (No. 4). This is because the learning target of these three tasks are not consistent and the optimization complexity is increased when more auxiliary task introduced. The performance of combining the three methods (No. 8) does not surpass that of SSAN individually (No. 4). In summary, our proposed task and SSAN can effectively improve the HMER model.

No.	Method	CROHME		
		2014	2016	2019
1	DWAP	51.48	50.65	50.04
2	+ CAN	57.00	56.06	54.88
3	+ SAM	56.80	55.62	56.21
4	+ SSAN	**60.85**	**60.02**	**61.83**
5	+ CAN & SAM	58.01	56.67	57.96
6	+ SSAN & CAN	**59.03**	**60.07**	**62.14**
7	+ SSAN & SAM	**60.04**	**59.37**	**60.63**
8	+ SSAN & CAN & SAM	57.61	**59.02**	**61.05**

Table 3: Effectiveness of Multi-task on DWAP. The bolded results indicate that the HMER model with SSAN outperforms its own performance.

Ablation Study

The proposed method jointly optimizes the HMER model, SSAN, the coarse-to-fine strategy and the attention-guided symbol masking strategy within a multi-task framework. The ablation study is shown in Table 4. All these blocks and strategies are sequentially added to the backbone model. The ”+ Joint optimization” represents the scenario in which we only used the SSAN block in the backbone without the distribution loss. The ”+ Distribution Loss” means that we used the distribution loss. The ”+ Coarse-to-fine alignment” and ”+ Attention-guided symbol masking” means using corresponding strategies in the training process.

Table 4 shows that the proposed SSAN and the two strategies are effective in enhancing the HMER model DWAP. Predicting the distribution map (”+ Distribution loss”) significantly improves the performance of DWAP. The two strategies further bring obvious improvement to DWAP on the above basis, indicating their effectiveness in HMER.

Method	CROHME			HME100K
	2014	2016	2019	
DWAP	51.48	50.65	50.04	61.85
+ Joint optimization	56.69	57.02	58.55	67.06
+ Distribution loss	58.72	58.24	59.97	68.37
+ Coarse-to-fine alignment	59.53	58.94	60.13	68.59
+ Attention-guided symbol masking	**60.85**	**60.02**	**61.83**	**68.74**

Table 4: Ablation study.

Method	Simple	Fraction	Radical	Sup./Sub.
DWAP	59.98	51.70	48.10	36.50
CAN-DWAP	61.56	57.67	53.33	43.83
DWAP + SSAN	**73.13**	**63.92**	**60.48**	**49.50**

Table 5: Recognition accuracy on different ME structures.

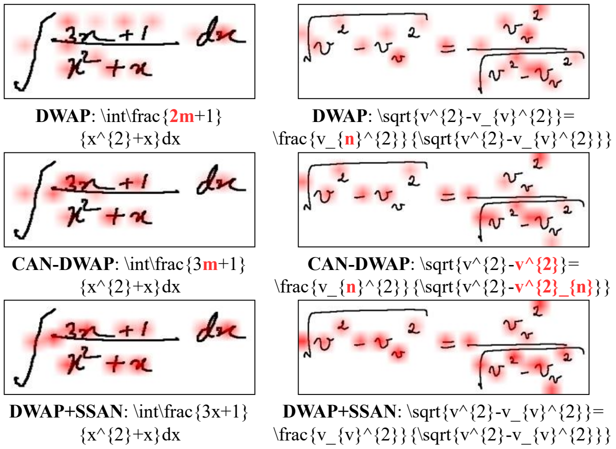
In-depth Analysis

In this section, we conduct an in-depth analysis of recognition accuracies on different formula structures, as well as the inference speed and parameters for DWAP, CAN-DWAP, and DWAP+SSAN. We also carry out the case study.

Recognition Accuracy on Different Formula Structures. We categorize the MEs in CROHME 2019 into four types: simple ME, fraction, radical, and ME with superscript or subscript structure. The simple ME is defined as a sequence where all symbols align on a single horizontal line. Table 5 shows that DWAP+SSAN achieves the best recognition accuracy across all four types, significantly outperforming DWAP. This indicates that the formula structure information learned through SSAN can effectively enhance HMER performance. The most notable improvement is seen in the subscript and superscript categories, with a 13% increase, demonstrating that the main advantage of SSAN is capturing the complex structure in MEs. This finding emphasizes the importance of the spatial distribution map introduced by SSAN, especially in recognizing nested and detail-rich formulas.

Inference Speed and Parameters. Table 6 shows the inference speed and parameters of the above mentioned methods. The FLOPs and FPS were calculated on the HME100K test dataset using an NVIDIA Tesla V100 GPU. DWAP+SSAN has 1.4 times the FLOPs and 3.5 times the parameters of DWAP, but is only 6.5% slower in FPS, indicating that SSAN can compute efficiently with minimal added computational cost. DWAP+SSAN explicitly has a similar inference speed and parameters as CAN-DWAP.

Case Study. In Fig. 4, we selected two typical examples of complex formula structures to illustrate how SSAN alleviates issues encountered by DWAP. In the first example, DWAP misrecognizes '$3x$' as '$2m$' due to attention drift. CAN-DWAP reduces this issue but still misclassifies 'x' as 'm'. SSAN alleviates attention drift by accurately perceiving symbol spatial positions and establishing correct long-distance correlations. The second example involves multiple nested structures and many position-sensitive elements. Both DWAP and CAN-DWAP misrecognize the subscript 'v' as 'n', suggesting a limitation in handling small and position-sensitive symbols. By introducing SSAN, which thoroughly learns symbol spatial distribution, the model improves recognition of complex structures and position-sensitive elements, effectively identifying multiple nested formulas. These examples demonstrate the effectiveness of SSAN.


DWAP: \int\frac{2m+1}{x^{2}+x}dx

DWAP: \sqrt{v^{2}-v_{v}^{2}}=\frac{v_{n}^{2}}{\sqrt{v^{2}-v_{v}^{2}}}

CAN-DWAP: \int\frac{3m+1}{x^{2}+x}dx

CAN-DWAP: \sqrt{v^{2}-v^{2}}=\frac{v_{n}^{2}}{\sqrt{v^{2}-v^{2}_{n}}}

DWAP+SSAN: \int\frac{3x+1}{x^{2}+x}dx

DWAP+SSAN: \sqrt{v^{2}-v_{v}^{2}}=\frac{v_{v}^{2}}{\sqrt{v^{2}-v_{v}^{2}}}

Figure 4: Attention map visualization for the DWAP, CAN-DWAP and DWAP+SSAN models.

Method	FLOPs	FPS	Parameters
DWAP	7.59G	20.52	4.72M
CAN-DWAP	11.24G	17.92	16.03M
DWAP + SSAN	10.34G	19.18	16.67M

Table 6: Comparison of inference speed (FLOPs and FPS) and the parameters of DWAP, CAN-DWAP, and DWAP+SSAN.

Conclusion

In this paper, we propose an auxiliary task to predict the symbol spatial distribution map to aid the HMER model in developing spatial-awareness of the formula images. We design a symbol spatial-aware network (SSAN) for this task and jointly optimized it with the HMER model. We also propose the coarse-to-fine alignment strategy and the attention-guided symbol masking strategy to enhance SSAN. Extensive experiments demonstrate that our method can significantly improve the HMER model's recognition performance, particularly for complex structural MEs. The experiments also prove the generalization of the proposed method. Furthermore, ablation studies confirm the relevance between symbol spatial distribution prediction and HMER, contributing to better recognition. Additionally, the coarse-to-fine alignment strategy overcomes performance limitations caused by the loose alignment of symbol distribution, while the attention-guided symbol masking strategy helps reduce similar symbol misclassification.

Acknowledgements

This work was funded by National Natural Science Foundation of China (Grant No. 62366036), National Education Science Planning Project (Grant No. BIX230343), Key R&D and Achievement Transformation Program of Inner Mongolia Autonomous Region (Grant No. 2022YFHH0077), The Central Government Fund for Promoting Local Scientific and Technological Development (Grant No. 2022ZY0198), Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (Grant No. NJYT24033), Inner Mongolia Autonomous Region Science and Technology Planning Project (Grant No. 2023YFSH0017), Joint Fund of Scientific Research for the Public Hospitals of Inner Mongolia Academy of Medical Sciences (Grant No. 2023GLLH0035).

References

Bian, X.; Qin, B.; Xin, X.; Li, J.; Su, X.; and Wang, Y. 2022. Handwritten Mathematical Expression Recognition via Attention Aggregation Based Bi-directional Mutual Learning. In *AAAI*, 113–121. AAAI Press.

Bonatti, C.; and Mohr, D. 2022. On the importance of self-consistency in recurrent neural network models representing elasto-plastic solids. *Journal of the Mechanics and Physics of Solids*, 158: 104697.

Cho, K.; van Merrienboer, B.; Gülçehre, Ç.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *CoRR*, abs/1406.1078.

Deng, Y.; Kanervisto, A.; Ling, J.; and Rush, A. M. 2017. Image-to-Markup Generation with Coarse-to-Fine Attention. In Precup, D.; and Teh, Y. W., eds., *ICML*, volume 70 of *Proceedings of Machine Learning Research*, 980–989. PMLR.

Ding, H.; Chen, K.; and Huo, Q. 2021. An Encoder-Decoder Approach to Handwritten Mathematical Expression Recognition with Multi-head Attention and Stacked Decoder. In Lladós, J.; Lopresti, D.; and Uchida, S., eds., *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, 602–616. Springer.

Farquhar, G.; Baumli, K.; Marinho, Z.; Filos, A.; Hessel, M.; van Hasselt, H. P.; and Silver, D. 2021. Self-consistent models and values. *Advances in Neural Information Processing Systems*, 34: 1111–1125.

Fu, Y.; Cai, W.; Gao, M.; and Zhou, A. 2023. Symbol Location-Aware Network for Improving Handwritten Mathematical Expression Recognition. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 516–524.

He, F.; Tan, J.; and Bi, N. 2020. Handwritten Mathematical Expression Recognition: A Survey. In Lu, Y.; Vincent, N.; Yuen, P. C.; Zheng, W.; Cheriet, F.; and Suen, C. Y., eds., *Pattern Recognition and Artificial Intelligence - International Conference, ICPRAI 2020, Zhongshan, China, October 19-23, 2020, Proceedings*, volume 12068 of *Lecture Notes in Computer Science*, 55–66. Springer.

Huang, G.; Liu, Z.; van der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *CVPR*, 2261–2269. IEEE Computer Society.

Jaiswal, S.; Fernando, B.; and Tan, C. 2022. TDAM: Top-Down Attention Module for Contextually Guided Feature Selection in CNNs. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *ECCV*, volume 13685 of *Lecture Notes in Computer Science*, 259–276. Springer.

Li, B.; Yuan, Y.; Liang, D.; Liu, X.; Ji, Z.; Bai, J.; Liu, W.; and Bai, X. 2022. When Counting Meets HMER: Counting-Aware Network for Handwritten Mathematical Expression Recognition. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *ECCV*, volume 13688 of *Lecture Notes in Computer Science*, 197–214. Springer.

Li, H.; Wang, P.; Shen, C.; and Zhang, G. 2019. Show, Attend and Read: A Simple and Strong Baseline for Irregular Text Recognition. In *AAAI*, 8610–8617. AAAI Press.

Li, Z.; Jin, L.; Lai, S.; and Zhu, Y. 2020. Improving Attention-Based Handwritten Mathematical Expression Recognition with Scale Augmentation and Drop Attention. In *17th International Conference on Frontiers in Handwriting Recognition, ICFHR 2020, Dortmund, Germany, September 8-10, 2020*, 175–180. IEEE.

Li, Z.; Yang, W.; Qi, H.; Jin, L.; Huang, Y.; and Ding, K. 2024. A tree-based model with branch parallel decoding for handwritten mathematical expression recognition. *Pattern Recognition*, 149: 110220.

Liu, Z.; Yuan, Y.; Ji, Z.; Bai, J.; and Bai, X. 2023. Semantic graph representation learning for handwritten mathematical expression recognition. In *International Conference on Document Analysis and Recognition*, 152–166. Springer.

Mouchère, H.; Viard-Gaudin, C.; Zanibbi, R.; and Garain, U. 2014. ICFHR 2014 Competition on Recognition of On-Line Handwritten Mathematical Expressions (CROHME 2014). In *14th International Conference on Frontiers in Handwriting Recognition, ICFHR 2014, Crete, Greece, September 1-4, 2014*, 791–796. IEEE Computer Society.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. 91–99.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. 3104–3112.

Truong, T.; Nguyen, C. T.; Phan, K. M.; and Nakagawa, M. 2020. Improvement of End-to-End Offline Handwritten Mathematical Expression Recognition by Weakly Supervised Learning. In *17th International Conference on Frontiers in Handwriting Recognition, ICFHR 2020, Dortmund, Germany, September 8-10, 2020*, 181–186. IEEE.

Tu, Z.; Lu, Z.; Liu, Y.; Liu, X.; and Li, H. 2016. Modeling Coverage for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. 30.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-consistency improves chain of thought reasoning in language models. arXiv. *arXiv preprint arXiv:2203.11171*.

Wu, C.; Du, J.; Li, Y.; Zhang, J.; Yang, C.; Ren, B.; and Hu, Y. 2022. TDv2: A Novel Tree-Structured Decoder for Offline Mathematical Expression Recognition. In *AAAI*, 2694–2702. AAAI Press.

Wu, J.; Yin, F.; Zhang, Y.; Zhang, X.; and Liu, C. 2018. Image-to-Markup Generation via Paired Adversarial Learning. In Berlingerio, M.; Bonchi, F.; Gärtner, T.; Hurley, N.; and Ifrim, G., eds., *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part I*, volume 11051 of *Lecture Notes in Computer Science*, 18–34. Springer.

Wu, J.; Yin, F.; Zhang, Y.; Zhang, X.; and Liu, C. 2020. Handwritten Mathematical Expression Recognition via Paired Adversarial Learning. *IJCV*, 128(10): 2386–2401.

Yang, C.; Du, J.; Zhang, J.; Wu, C.; Chen, M.; and Wu, J. 2022. Tree-based data augmentation and mutual learning for offline handwritten mathematical expression recognition. *Pattern Recognition*, 132: 108910.

Yuan, Y.; Liu, X.; Dikubab, W.; Liu, H.; Ji, Z.; Wu, Z.; and Bai, X. 2022. Syntax-Aware Network for Handwritten Mathematical Expression Recognition. In *CVPR*, 4543–4552. IEEE.

Zhang, J.; Du, J.; and Dai, L. 2018. Multi-Scale Attention with Dense Encoder for Handwritten Mathematical Expression Recognition. In *ICPR*, 2245–2250. IEEE Computer Society.

Zhang, J.; Du, J.; Yang, Y.; Song, Y.; Wei, S.; and Dai, L. 2020. A Tree-Structured Decoder for Image-to-Markup Generation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 11076–11085. PMLR.

Zhang, J.; Du, J.; Zhang, S.; Liu, D.; Hu, Y.; Hu, J.; Wei, S.; and Dai, L. 2017. Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition. *PR*, 71: 196–206.

Zhang, X.; Ying, H.; Tao, Y.; Xing, Y.; and Feng, G. 2023. General category network: Handwritten mathematical expression recognition with coarse-grained recognition task. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Zhao, W.; and Gao, L. 2022. Comer: Modeling coverage for transformer-based handwritten mathematical expression recognition. In *ECCV*, 392–408. Springer.

Zhao, W.; Gao, L.; Yan, Z.; Peng, S.; Du, L.; and Zhang, Z. 2021. Handwritten Mathematical Expression Recognition with Bidirectionally Trained Transformer. In Lladós, J.; Lopresti, D.; and Uchida, S., eds., *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, 570–584. Springer.

Zhong, S.; Song, S.; Li, G.; and Chan, S. G. 2022. A Tree-Based Structure-Aware Transformer Decoder for Image-To-Markup Generation. In Magalhães, J.; Bimbo, A. D.; Satoh, S.; Sebe, N.; Alameda-Pineda, X.; Jin, Q.; Oria, V.; and Toni, L., eds., *ACM MM*, 5751–5760. ACM.

Zhou, X.; Wang, D.; Tian, F.; Liu, C.; and Nakagawa, M. 2013. Handwritten Chinese/Japanese Text Recognition Using Semi-Markov Conditional Random Fields. *IEEE TPAMI*, 35(10): 2413–2426.