ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447346

VIEWING WRITING AS VIDEO: OPTICAL FLOW BASED MULTI-MODAL HANDWRITTEN MATHEMATICAL EXPRESSION RECOGNITION

Hanbo Cheng *Jun Du*⋆ *Pengfei Hu* *Jiefeng Ma* *Zhenrong Zhang* *Mobai Xue*

University of Science and Technology of China, Hefei, Anhui, China
jundu@ustc.edu.cn

ABSTRACT

Handwritten Mathematical Expression Recognition (HMER) forms a crucial task in the domain of document intelligence. It encompasses online and offline modalities, which utilize the trajectory sequence and static image as input, respectively. It is intuitive to utilize both online and offline modalities to build a more powerful recognition system. However, a formidable challenge arises as a result of the substantial heterogeneity between the online and offline modalities, which consequently leads to considerable obstacles in their alignment and fusion. In this work, we perceive the writing process as a video and introduce the Aggregated Optical Flow Map (AOFM) to represent the online modality, which is more compatible with the offline modality. Additionally, we propose the Optical Flow Aware Network (OFAN) in order to automatically extract, align, and fuse the features across online and offline modalities. Through experiment analysis, our method can be seamlessly applied to multiple existing offline HMER models, thereby yielding stable and substantial enhancements across CROHME 2014, 2016, and 2019 datasets. The code in this work is available at https://github.com/Hanbo-Cheng/OFAN.git.

***Index Terms*—** Handwritten Mathematical Expression Recognition, Aggregated Optical Flow Map, Multi-Modal, Attention

1. INTRODUCTION

Handwritten Mathematical Expression Recognition (HMER) is a significant branch of document intelligence, which is required by many applications such as education, technology document digitization, and office automation. Diverging from regular text line recognition tasks, HMER presents greater challenges due to the intricate 2D structures inherent in mathematical expressions. Despite the impressive accomplishments of large language models (LLM) in natural language processing and multi-modal tasks, they still fall short of conventional methods in multiple OCR-related tasks, especially the HMER. Therefore, it's still necessary to design specialized methods for the HMER [1].

From the perspective of input data, the HMER task comprises two distinct categories: offline HMER and online HMER [2]. The former utilizes static images, while the latter employs dynamic handwriting trajectories as input. The encoder-decoder structure is widely used in both the offline and online OCR-related tasks [3, 4, 5]. In the encoder stage, in offline HMER, WAP [6] proposes to use CNN-based encoder to process the static image. For online HMER, [7, 8] transform the sequence of trajectory points into 8-dimensional vectors and leverages an RNN-based encoder to extract the feature. In the decoder stage, most of the online and offline HMER methods apply the RNN-based or transformer-based decoder with attention mechanism [3, 9, 10]. Generally, the dynamic handwriting trajectory in online HMER provides more comprehensive motion information during the writing process, which proves to be significantly beneficial in addressing ambiguous handwriting [8]. However, due to the irregular writing orders among different writers and the lack of global spatial information, the online HMER method often encounters limitations of lower structure analyzing capability and incorrect prediction orders [2]. Compared to the trajectory sequence, the static image has lower sensitivity to the writing orders and possesses more global spatial information [8]. However, the offline HMER method usually encounters the challenge of ambiguous writing, such as "B" and "β", "s" and "5" [6].

To fully exploit the advantage of online and offline information, many works have explored the multi-modal HMER method. [2] initially proposes a multi-modal HMER architecture. However, a significant heterogeneity still exists between the features derived from the online and offline modalities, leading to substantial challenges in the alignment and fusion processes of multi-modal data [2]. Later, [11] suggests aligning and merging the features from different modalities through strokes. Although the use of stroke masks to align the online and offline features can alleviate the alignment and fusion problems, the irregular writing orders still impose a substantial passive impact on the process of recognition [11].

Optical flow is a technique to describe the motion of pixels in the video, which reflects the movement of objects or cameras in consecutive frames [12]. In this study, to bridge the gap between online and offline HMER as well as enhance their complementarity, we view the writing process as a video and introduce a novel technique, the Aggregated Optical Flow Map (AOFM) as illustrated in Fig. 1 to represent the online modality. Our AOFM is easy to obtain and incorporates both motion and global spatial information. Based on AOFM, we introduce an encoder-decoder architecture called the Optical Flow Aware Network (OFAN). To be more specific, the encoder incorporates a two-branch CNN structure to process the input from online and offline modalities respectively. The decoder encompasses a multi-modal attention module that can automatically align and fuse features extracted from disparate modalities. In the experiment section, we implement our proposed HMER framework on multiple extant offline methods and evaluate their performance on CROHME 2014, 2016, and 2019 datasets. Evidenced by consistent and significant improvement, our proposed approach exhibits substantial compatibility and prominent performance.

In summary, the contribution of the paper is three-fold:

(1) We view the writing process as a video and propose the **A**ggregated **O**ptical **F**low **M**ap (AOFM) to represent the online modality in the HMER task. The AOFM not only narrows the gap between online and offline data but also preserves extensive spatial information.

⋆ corresponding author

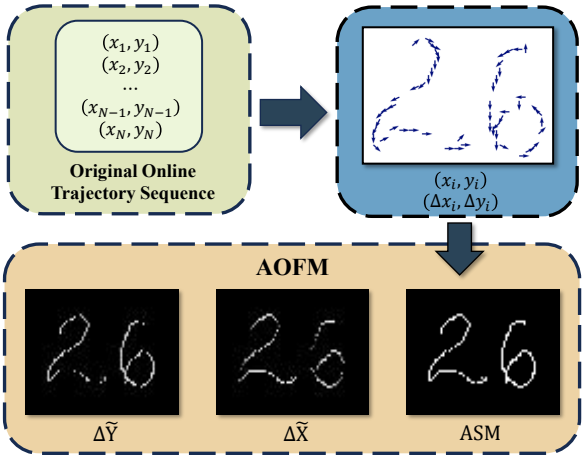

Fig. 1. The generation of the AOFM.

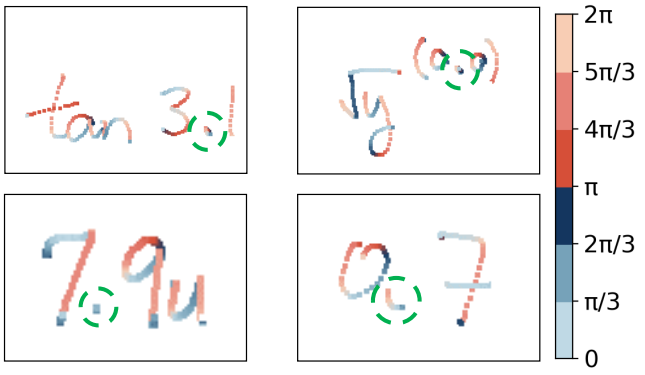

Fig. 2. The visualization of writing direction using gradient map.

(2) Based on the AOFM, we introduce the **O**tical **F**low **A**ware **N**etwork (OFAN) to accomplish the complementary integration of offline and online data.

(3) Our OFAN can be generalized to various existing models and achieve new state-of-the-art results. The experiment results demonstrate the superiority of our method over the previous best method.

2. METHOD

In this section, we introduce the Aggregated Optical Flow Map (AOFM) and the Optical Flow Aware Network (OFAN) based on encoder-decoder architecture for the HMER task.

2.1. Aggregated Optical Flow Map

The inherent nature of the writing process aligns harmoniously with the concept of a video. In the online trajectory sequence, each individual sampling point can be perceived as a distinct frame. Optical flow is a technique which is commonly applied in video-related tasks, such as video understanding and action recognition [12, 13]. The optical flow map records the movement direction and velocity of every pixel in each frame, which encompasses both the global spatial and motion information. Meanwhile, the optical flow map usually expresses a considerable complementarity with image [13]. However, the generation of conventional optical flow map is usually time and space consuming [14], which severely limits its application in HMER. Fortunately, We observe the online HMER exhibits unique traits distinguishing it from conventional video inputs. These distinctive traits encompass:

- Single dynamic point throughout the writing process
- Few intersections in the writing trace
- Direct and simplified acquisition of optical flow from the trajectory sequence.

Due to these traits, in contrast to the conventional optical flow map, we are able to streamline the tracking process to a single point and aggregate all the motion information pertaining to the writing process within a single frame, namely the Aggregated Optical Flow Map (AOFM). The generation process is illustrated in Fig.1.

The original data format of the online modality is a variable length point sequence, denoted as:

$$[x_1, y_1, s_1], [x_2, y_2, s_2], \cdots, [x_N, y_N, s_N] \tag{1}$$

Where the x_i, y_i represent the coordination of the i^{th} sampling point and s_i is the stroke index. To alleviate the reliance on stroke information, we refrained from using s_i.

Our approach entails extracting both the coordination information and the movement direction, denoted as (x_i, y_i) and $(\Delta x_i, \Delta y_i)$.

$$\Delta x_i = x_{i+1} - x_i \quad \Delta y_i = y_{i+1} - y_i \tag{2}$$

Specifically, To alleviate the impact of variable writing velocity, we normalize the direction vector $(\Delta x_i, \Delta y_i)$ by converting it into a unit vector and obtain $(\Delta\tilde{x}_i, \Delta\tilde{y}_i)$:

$$\Delta\tilde{x}_i = \frac{\Delta x_i}{\sqrt{{\Delta x_i}^2 + {\Delta y_i}^2}} \quad \Delta\tilde{y}_i = \frac{\Delta y_i}{\sqrt{{\Delta x_i}^2 + {\Delta y_i}^2}} \tag{3}$$

Eventually, we put the direction $(\Delta\tilde{x}_i, \Delta\tilde{y}_i)$ in the position of (x_i, y_i), where $\Delta\tilde{\mathrm{X}}, \Delta\tilde{\mathrm{Y}} \in \mathbb{R}^{H\times W}$:

$$\Delta\tilde{\mathrm{X}}_{x_i,y_i} = \Delta\tilde{x}_i \quad \Delta\tilde{\mathrm{Y}}_{x_i,y_i} = \Delta\tilde{y}_i \tag{4}$$

Additionally, some characters such as "!", "÷", "i" and ".", comprise small-scale strokes that inherently lack a fixed writing pattern. In Fig. 2, we manifest such a phenomenon by utilizing the hue to represent the writing direction, denoting the angle within the range of 0 to 2π. The figure suggests the same stroke "." has variable writing patterns in different cases. Such a phenomenon leads to the consequence that only using the optical flow map fails to establish a stable feature, which eventually causes ambiguity and confusion. To ameliorate this issue, we introduce the method called Auxiliary Static Map (ASM). The ASM, denoted as $\mathrm{S} \in \mathbb{R}^{H\times W}$, is a single-channel Boolean map, where the elements represent the presence of optical flow information in the corresponding pixel. Eventually, we form the AOFM by concatenating the $\Delta\tilde{\mathrm{X}}$, $\Delta\tilde{\mathrm{Y}}$, and ASM:

$$\mathrm{AOFM} = [\Delta\tilde{\mathrm{X}}; \Delta\tilde{\mathrm{Y}}; \mathrm{S}] \tag{5}$$

2.2. Optical Flow Aware Network

As illustrated in Fig. 3, our architecture leverages the AOFM as the online modality input and static image as the offline counterpart. The online and offline data parallelly pass through a two-branch symmetric CNN-based encoder to extract the high-level features. Then a multi-modal decoder aligns and fuses the features from different modalities to generate the target sequence. Although we select the DWAP [15] to demonstrate our proposed method, the proposed structure can also be applied to the transformer based [10] or tree decoder based [16] method. In the encoder stage, following [15], we use a pair of DenseNet [17] to extract the feature from the AOFM and static image. We define the output feature map processed by two DenseNet as $\boldsymbol{A}_{\text{off}}, \boldsymbol{A}_{\text{on}} \in \mathbb{R}^{D \times H \times W}$ respectively.

In the decoder stage, the model aims to generate a target sequence $\boldsymbol{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]$. The target sequence can be the tree structure label [16] or the LaTeX string. Our decoder incorporates two layers of Gated Recurrent Unit (GRU) [18] and a multi-modal attention module. The first GRU layer aims to establish a shared query vector for attention operation. The second GRU layer aims to generate the target sequence step by step. The multi-modal attention module employs the shared query vector to retrieve, align and fuse the significant feature from online and offline modalities. The overall process can be denoted as:

$$\hat{\boldsymbol{h}}_t = \text{GRU}_1(\mathbf{y}_{t-1}, \boldsymbol{h}_{t-1}) \tag{6}$$

$$\boldsymbol{c}_t = f_{\text{MA}}(Q = \hat{\boldsymbol{h}}_t, K = \boldsymbol{A}_{\text{off}}|\boldsymbol{A}_{\text{on}}, V = \boldsymbol{A}_{\text{off}}|\boldsymbol{A}_{\text{on}}) \tag{7}$$

$$\boldsymbol{h}_t = \text{GRU}_2(\boldsymbol{c}_t, \hat{\boldsymbol{h}}_t) \tag{8}$$

where GRU_1 and GRU_2 represent the first and second GRU layer, f_{MA} indicates the multi-modal attention module, $\boldsymbol{c}_t$ is the context vector, $\hat{\boldsymbol{h}}_t$ and $\boldsymbol{h}_t$ represent the hidden state of the first and second layer of the GRU cell in t^{th} decoding step, respectively.

In f_{MA}, we use a shared query vector to retrieve and align the significant feature from online and offline respectively, which can be expressed as:

$$\boldsymbol{c}_t^{\text{off}} = f_{\text{attn}}(Q = \hat{\boldsymbol{h}}_t, K = \boldsymbol{A}_{\text{off}}, V = \boldsymbol{A}_{\text{off}}) \tag{9}$$

$$\boldsymbol{c}_t^{\text{on}} = f_{\text{attn}}(Q = \hat{\boldsymbol{h}}_t, K = \boldsymbol{A}_{\text{on}}, V = \boldsymbol{A}_{\text{on}}) \tag{10}$$

where $\boldsymbol{c}_t^{\text{off}}$ and $\boldsymbol{c}_t^{\text{on}}$ are the context vector for offline and online modality, the f_{attn} is the additive attention mechanism [19].

After generating the $\boldsymbol{c}_t^{\text{off}}$ and $\boldsymbol{c}_t^{\text{on}}$, the multi-modal attention module fuses them and provides a shared context vector:

$$\boldsymbol{c}_t = \boldsymbol{W}_f([\boldsymbol{c}_t^{\text{off}}; \boldsymbol{c}_t^{\text{on}}]^\top) + \boldsymbol{b} \tag{11}$$

where $[\boldsymbol{c}_t^{\text{off}}; \boldsymbol{c}_t^{\text{on}}]$ is the concatenation operation, $\boldsymbol{W}_f \in \mathbb{R}^{D \times 2D}$ and $b \in \mathbb{R}^D$ are trainable parameters.

Then we utilize the hidden state of the second layer of GRU $\boldsymbol{h}_t$, the context vector $\boldsymbol{c}_t$, and embedding of y_{t-1} to estimate the probability of y_t:

$$\text{p}(\mathbf{y}_t|\boldsymbol{I}_{\text{off}}, \boldsymbol{I}_{\text{on}}, \mathbf{y}_{t-1}) = \sigma(\boldsymbol{W}_o\phi(\boldsymbol{W}_h\boldsymbol{h}_t + \boldsymbol{W}_c\boldsymbol{c}_t + \boldsymbol{W}_{\text{y}}\boldsymbol{E}(\mathbf{y}_{t-1}))) \tag{12}$$

where the σ and ϕ are the softmax and maxout activation function, $\boldsymbol{I}_{\text{off}}, \boldsymbol{I}_{\text{on}}$ are offline and online input data, and $\boldsymbol{W}_o \in \mathbb{R}^{K \times m}, \boldsymbol{W}_h \in \mathbb{R}^{m \times n}, \boldsymbol{W}_c \in \mathbb{R}^{m \times D}, \boldsymbol{W}_{\text{y}} \in \mathbb{R}^{m \times n}, \boldsymbol{E} \in \mathbb{R}^{K \times m}$ are trainable parameters.

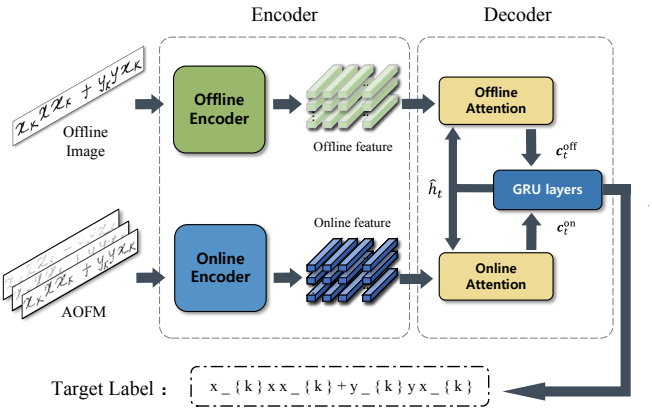

Fig. 3. Architecture of Optical Flow Aware Network adapted from DWAP (OFAN-DWAP).

3. EXPERIMENTS

3.1. Dataset and Implement Details

For training and evaluation, our method utilizes the CROHME dataset [20], which currently stands as the most widely employed public dataset in the HMER task. In the CROHME training set, there are 8836 handwritten mathematical formulas. The CROHME encompasses three test sets: CROHME 2014, 2016, and 2019. These test sets incorporate 986, 1147, and 1199 mathematical expressions, respectively. In the original form, the CROHME dataset uses the trajectory sequence to represent the mathematical formula. We transform the original data into the static image and the AOFM, serving as the input of offline and online modality.

We employ our multi-modal HMER architecture in a plug-in mode on multiple existing offline HMER models. In comparison to the original offline approach, we introduce the following adaptations: (1) We incorporate an extra DenseNet encoder for online modality. (2) We replace the single-modal attention module with our proposed multi-modal attention module. As for the remaining components, we meticulously adhere to the configuration of the original offline model. The loss function is the cross entropy loss, and the optimization algorithm is the Adadelta [21], with a learning rate set to 1. All experiments were performed on a single NVIDIA 3090 24GB GPU.

3.2. Comparison with State-Of-The-Art Methods

In this section, we validate the effectiveness and compatibility of our method by applying it to five established offline HMER models: WAP [15], ABM [4], BTTR [10], TDv2 [9], and CoMER [22]. These models encompass diverse decoder structures such as RNN-based and transformer-based, and label categories including tree structure-based and LaTeX string-based. This careful selection ensures that they represent a comprehensive range of existing methods in the HMER task. The "-OFAN" indicates that we incorporate the proposed architecture into the existing offline model. To evaluate the model performance, we utilize the ExpRate measure, which represents the proportion of correctly predicted expressions.

Based on the results presented in Table 1, our approach demonstrates excellent compatibility with extant offline HMER models.

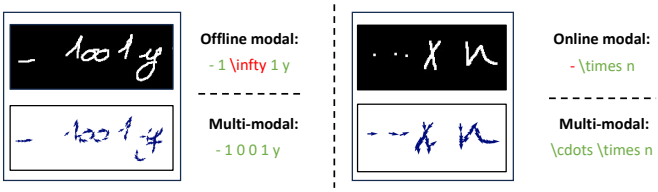

Fig. 4. The OFAN rectifies errors in the single-modal method. We use the arrow to represent the writing direction in online modality.

The experiment results convincingly reveal that our method consistently improves the performance of these models. Notably, the "CoMER-OFAN" configuration achieves considerable ExpRate of 60.24%, 61.63%, and 61.96% on CROHME 2014, 2016, and 2019 respectively. These results exhibit a significant advantage over the majority of existing single-modal and multi-modal HMER models.

Table 1. Results on the CROHME dataset, "off" and "on" indicates the offline and online modality, † represents our reproduced result.

Model	Modality		CROHME(ExpRate)		
	off	on	2014	2016	2019
WAP [6]	✓		46.55	44.55	-
DWAP-TD [16]	✓		49.10	48.50	51.40
SAN [23]	✓		56.2	53.6	53.5
CAN-DWAP [24]	✓		57.00	56.65	54.88
CAN-ABM [24]	✓		57.26	56.15	55.96
BTTR [10]	✓		53.96	52.31	52.96
TDv2 [9]	✓		53.56	55.18	58.72
GCN [25]	✓		60.00	58.94	61.63
TAP [8]		✓	48.47	44.81	-
G2G [26]		✓	54.46	52.05	-
MDR [3]		✓	55.8	52.5	53.6
MAN [2]	✓	✓	54.05	50.56	52.21
MMSCAN-D [11]	✓	✓	55.38	52.22	53.88
MMSCAN-E [11]	✓	✓	57.20	53.97	56.21
path signature [27]	✓	✓	58.92	59.46	**63.22**
OFAN-based multi-modal method					
DWAP† [15]	✓		50.51	49.34	48.70
DWAP-OFAN	✓	✓	55.78	54.40	53.38
ABM† [4]	✓		55.58	54.05	54.23
ABM-OFAN	✓	✓	57.71	55.01	56.30
BTTR† [10]	✓		54.05	55.01	57.38
BTTR-OFAN	✓	✓	58.27	57.45	58.46
TDv2† [9]	✓		54.87	54.58	57.88
TDv2-OFAN	✓	✓	59.73	58.41	60.13
CoMER† [22]	✓		58.92	57.89	59.21
CoMER-OFAN	✓	✓	**60.34**	**61.63**	61.96

3.3. Ablation Study

To verify the effectiveness of AOFM and our multi-modal HMER method, we conduct ablation experiments on CROHME. The result is presented in Table 2. The "on" indicates AOFM serving as online input, while the "off" signifies image as offline input. The "ASM" implies whether to use the auxiliary static map in AOFM. The results reveal that the performance of the single-modal model using AOFM with ASM slightly surpasses the offline model, and the multi-modal method significantly outperforms the single-modal approach. As illustrated in Fig. 4, in the offline method, the grouping of characters "00" is recognized as "∞" due to the occlusion. In the online counterpart, the discrete sampling nature causes the recognition of "$\cdots$" as "$-$". Through the utilization of the OFAN, the deficiencies in the single-modal approach are significantly alleviated. Additionally, in AOFM, we append the proposed ASM to alleviate the random writing pattern of small-scale strokes. To explore the impact of ASM, we compared the method with and without ASM. The result in Table 2 suggests that the ASM improves the performance considerably, both in the single-modal and multi-modal methods.

In order to assess the efficacy of the multi-modal attention module, we compare the performance of different feature-aligning strategies. The result is illustrated in Table 3, where the "align method" denotes the multi-modal feature aligning strategy. The "attention" represents our proposed multi-modal attention module. The "concat" denotes the alignment of online and offline features through direct concatenation. The results clearly demonstrate that our proposed multi-modal attention module significantly enhances performance.

Table 2. Ablation study of offline and online modalities.

Model	off	on	ASM	CROHME (ExpRate)		
				2014	2016	2019
DWAP	✓			50.51	49.34	48.70
		✓		51.52	47.78	46.04
		✓	✓	51.82	51.53	47.37
	✓	✓		54.76	53.36	52.54
	✓	✓	✓	55.78	54.40	53.38

Table 3. Ablation study of of the alignment strategy.

Model	Align Method	CROHME (ExpRate)		
		2014	2016	2019
DWAP	concat	53.45	53.36	51.29
	attention	55.78	54.40	53.38

4. CONCLUSION

In this paper, we adopt a unique perspective by considering the writing process as a video and propose the Aggregated Optical Flow Maps (AOFM) to serve as the input of online HMER. Then we introduce the Optical Flow Aware Network (OFAN) to solve the online and offline HMER in a unified manner. Our proposed architecture incorporates a symmetric two-branch CNN encoder and a decoder with the multi-modal attention module, which can work in plug-in mode with most of the existing offline models. The experiment substantiates the considerable performance improvements yielded by our method. This framework not only promotes better fusion between the two modalities but also exhibits substantial adaptability and scalability.

5. ACKNOWLEDGMENT

This work was supported by Alibaba Group.

6. REFERENCES

[1] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al., "On the hidden mystery of ocr in large multimodal models," *arXiv preprint arXiv:2305.07895*, 2023.

[2] Jiaming Wang, Jun Du, Jianshu Zhang, and Zi-Rui Wang, "Multi-modal attention network for handwritten mathematical expression recognition," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 1181–1186.

[3] Jiaming Wang, Qing Wang, Jun Du, Jianshu Zhang, Bin Wang, and Bo Ren, "Mrd: A memory relation decoder for online handwritten mathematical expression recognition," in *Document Analysis and Recognition – ICDAR 2021*, Josep Lladós, Daniel Lopresti, and Seiichi Uchida, Eds., Cham, 2021, pp. 39–54, Springer International Publishing.

[4] Xiaohang Bian, Bo Qin, Xiaozhe Xin, Jianwu Li, Xuefeng Su, and Yanfeng Wang, "Handwritten mathematical expression recognition via attention aggregation based bi-directional mutual learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, pp. 113–121, Jun. 2022.

[5] Pengfei Hu, Jiefeng Ma, Zhenrong Zhang, Jun Du, and Jianshu Zhang, "Count, decode and fetch: A new approach to handwritten chinese character error correction," 2023.

[6] Jianshu Zhang, Jun Du, Shiliang Zhang, Dan Liu, Yulong Hu, Jinshui Hu, Si Wei, and Lirong Dai, "Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition," *Pattern Recognition*, vol. 71, pp. 196–206, 2017.

[7] Jianshu Zhang, Jun Du, and Lirong Dai, "A gru-based encoder-decoder approach with attention for online handwritten mathematical expression recognition," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, vol. 01, pp. 902–907.

[8] Jianshu Zhang, Jun Du, and Lirong Dai, "Track, attend, and parse (tap): An end-to-end framework for online handwritten mathematical expression recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 221–233, Jan 2019.

[9] Changjie Wu, Jun Du, Yunqing Li, Jianshu Zhang, Chen Yang, Bo Ren, and Yiqing Hu, "Tdv2: A novel tree-structured decoder for offline mathematical expression recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 2694–2702.

[10] Wenqi Zhao, Liangcai Gao, Zuoyu Yan, Shuai Peng, Lin Du, and Ziyin Zhang, "Handwritten mathematical expression recognition with bidirectionally trained transformer," in *Document Analysis and Recognition – ICDAR 2021*, Josep Lladós, Daniel Lopresti, and Seiichi Uchida, Eds., Cham, 2021, pp. 570–584, Springer International Publishing.

[11] Jiaming Wang, Jun Du, Jianshu Zhang, Bin Wang, and Bo Ren, "Stroke constrained attention network for online handwritten mathematical expression recognition," *Pattern Recognition*, vol. 119, pp. 108047, 2021.

[12] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[13] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, Cambridge, MA, USA, 2014, NIPS'14, p. 568–576, MIT Press.

[14] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu, "Self-supervised video representation learning by pace prediction," *CoRR*, vol. abs/2008.05861, 2020.

[15] Jianshu Zhang, Jun Du, and Lirong Dai, "Multi-scale attention with dense encoder for handwritten mathematical expression recognition," in *2018 24th international conference on pattern recognition (ICPR)*. IEEE, 2018, pp. 2245–2250.

[16] Jianshu Zhang, Jun Du, Yongxin Yang, Yi-Zhe Song, Si Wei, and Lirong Dai, "A tree-structured decoder for image-to-markup generation," in *Proceedings of the 37th International Conference on Machine Learning*, Hal Daumé III and Aarti Singh, Eds. 13–18 Jul 2020, vol. 119 of *Proceedings of Machine Learning Research*, pp. 11076–11085, PMLR.

[17] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016.

[18] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[19] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[20] Mahshad Mahdavi, Richard Zanibbi, Harold Mouchere, Christian Viard-Gaudin, and Utpal Garain, "Icdar 2019 crohme + tfd: Competition on recognition of handwritten mathematical expressions and typeset formula detection," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, pp. 1533–1538.

[21] MatthewD. Zeiler, "Adadelta: An adaptive learning rate method," *Cornell University - arXiv,Cornell University - arXiv*, Dec 2012.

[22] Wenqi Zhao and Liangcai Gao, "Comer: Modeling coverage for transformer-based handwritten mathematical expression recognition," in *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, Eds., Cham, 2022, pp. 392–408, Springer Nature Switzerland.

[23] Ye Yuan, Xiao Liu, Wondimu Dikubab, Hui Liu, Zhilong Ji, Zhongqin Wu, and Xiang Bai, "Syntax-aware network for handwritten mathematical expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4553–4562.

[24] Bohan Li, Ye Yuan, Dingkang Liang, Xiao Liu, Zhilong Ji, Jinfeng Bai, Wenyu Liu, and Xiang Bai, "When counting meets hmer: counting-aware network for handwritten mathematical expression recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 197–214.

[25] Xinyu Zhang, Han Ying, Ye Tao, Youlu Xing, and Guihuan Feng, "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)*, June 2023, pp. 1–5.

[26] Jin-Wen Wu, Fei Yin, Yan-Ming Zhang, Xu-Yao Zhang, and Cheng-Lin Liu, "Graph-to-graph: Towards accurate and interpretable online handwritten mathematical expression recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, pp. 2925–2933, May 2021.

[27] Zhe Li, Xinyu Wang, Yuliang Liu, Lianwen Jin, Yichao Huang, and Kai Ding, "Improving handwritten mathematical expression recognition via similar symbol distinguishing," *IEEE Transactions on Multimedia*, pp. 1–13, 2023.