electronics

MDPI

Article

A Multi-Modal Retrieval Model for Mathematical Expressions Based on ConvNeXt and Hesitant Fuzzy Set

Ruxuan Li [1,2,3], Jingyi Wang [1,2,3] and Xuedong Tian [1,2,3,*]

[1] School of Cyber Security and Computer, Hebei University, Baoding 071002, China; lrx@stumail.hbu.edu.cn (R.L.); wangjingyi@stumail.hbu.edu.cn (J.W.)
[2] Institute of Intelligent Image and Document Information Processing, Hebei University, Baoding 071002, China
[3] Hebei Machine Vision Engineering Research Center, Hebei University, Baoding 071002, China
* Correspondence: xdtian@hbu.edu.cn

Abstract: Mathematical expression retrieval is an essential component of mathematical information retrieval. Current mathematical expression retrieval research primarily targets single modalities, particularly text, which can lead to the loss of structural information. On the other hand, multi-modal research has demonstrated promising outcomes across different domains, and mathematical expressions in image format are adept at preserving their structural characteristics. So we propose a multi-modal retrieval model for mathematical expressions based on ConvNeXt and HFS to address the limitations of single-modal retrieval. For the image modal, mathematical expression retrieval is based on the similarity of image features and symbol-level features of the expression, where image features of the expression image are extracted by ConvNeXt, while symbol-level features are obtained by the Symbol Level Features Extraction (SLFE) module. For the text modal, the Formula Description Structure (FDS) is employed to analyze expressions and extract their attributes. Additionally, the application of the Hesitant Fuzzy Set (HFS) theory facilitates the computation of hesitant fuzzy similarity between mathematical queries and candidate expressions. Finally, Reciprocal Rank Fusion (RRF) is employed to integrate rankings from image modal and text modal retrieval, yielding the ultimate retrieval list. The experiment was conducted on the publicly accessible ArXiv dataset (containing 592,345 mathematical expressions) and the NTCIR-mair-wikipedia-corpus (NTCIR) dataset.The MAP@10 values for the multimodal RRF fusion approach are recorded as 0.774. These substantiate the efficacy of the multi-modal mathematical expression retrieval approach based on ConvNeXt and HFS.

Keywords: mathematical expressions retrieval; ConvNeXt; FDS; HFS; multi-modal

Citation: Li, R.; Wang, J.; Tian, X. A Multi-Modal Retrieval Model for Mathematical Expressions Based on ConvNeXt and Hesitant Fuzzy Set. *Electronics* **2023**, *12*, 4363. https://doi.org/10.3390/electronics12204363

Academic Editor: George A. Tsihrintzis

Received: 21 September 2023
Revised: 14 October 2023
Accepted: 19 October 2023
Published: 20 October 2023

1. Introduction

Mathematical expressions play a significant role in scientific communication and calculations. Mathematical Information Retrieval (MIR) is a crucial component of Information Retrieval (IR) that deals with searching for specific mathematical expressions, concepts, or objects. Efficient indexing and retrieval of mathematical expressions have become the most challenging part of MIR due to the growing use of mathematical content in various scientific documents, educational materials, and web information in MathML or LaTeX format. Many academics have been working on mathematical information retrieval research recently and have seen some success [1–3].

Due to the two-dimensional structure of mathematical expressions and the complexity of mathematical symbol types and semantics, traditional search engines primarily designed for plain text retrieval often struggle to fulfill the required criteria. Mathematical expression retrieval encounters numerous challenges:

1. Mathematical Context: Understanding the context in which a mathematical expression is used is crucial, as the same notation can have different meanings in distinct mathematical subfields or domains.

2. Equivalent transformations of a mathematical expression: Mathematical expressions in different forms may convey the same meaning (e.g., $a \times (x + z)$ and $a \times x + a \times z$, $a^{\frac{1}{2}}$ and $\sqrt{a}$).
3. Multimodal Content: Multimodal content presents mathematical expressions in various forms, in addition to textual modes like MathML and LaTeX, these expressions can also be represented through images. Different modalities of mathematical expressions exhibit distinct characteristics, necessitating the effective handling of diverse data types.
4. Structure Complexity: Mathematical expressions can be symbolically complex, involving nested functions, subscripts, superscripts, and specialized symbols that require intricate parsing and interpretation.

Current research in multimodal studies is expanding rapidly, exploring the integration of various data types and modes to gain a more comprehensive understanding of complex phenomena [4,5]. In response to the challenges outlined above, specifically Challenge 3 and Challenge 4, this paper makes the following contributions:

1. Introducing the image modality in mathematical expression retrieval. Images can capture the visual aspects of mathematical expressions, providing a richer and more comprehensive representation compared to plain text. Combining image modal with text modal allows a more comprehensive understanding of mathematical content. This integration can enhance retrieval accuracy by considering multiple data types simultaneously.
2. Building upon extracting image features from mathematical expression images, we have devised a symbol-level feature extraction method to obtain a more comprehensive set of image feature information. This enhancement ensures that the ranking results produced by the image modality retrieval module are more rational and accurate.
3. We employ algorithmic analysis to extract attributes from the textual modality of mathematical expressions. Subsequently, we construct a table of attributes for textual modality expressions. By introducing hesitant fuzzy sets and leveraging their advantages in handling multi-attribute evaluation criteria, we calculate the similarity between expressions.
4. We opt for an appropriate fusion sorting method to combine and rank the retrieval results from both image and text modalities, resulting in the final ranking outcome. Furthermore, a subset of literature from the publicly accessible ArXiv dataset was extracted and utilized to construct a dataset encompassing 592,345 mathematical expressions.

2. Related Work

Regarding mathematical expression retrieval of the text modal, expression trees are widely used for storing and processing mathematical expressions and have been applied to mathematical expression retrieval by many scholars. Goel et al. [6] have undertaken studies on Math Word Problems, utilizing a tree-matching algorithm for the matching of mathematical expressions; they performed pair-wise matching on expression trees through post-order traversal. Pfahler et al. [7] incorporated unsupervised embedding learning and Graph Convolutional Neural Networks (GCNNs) for learning mathematical representations. In order to facilitate effective nearest-neighbor queries, mathematical operations represented in XML format were processed as graphical data and embedded into a low-dimensional vector space. Schellenberg et al. [8] employed substitution trees to index and retrieve mathematical expressions in LaTeX representation, but the insertion bias limits its performance. Hu et al. [9] proposed WikiMirs, using a method of generalization that is hierarchical to produce subtrees from the representation trees of mathematical expressions, which can support substructure and similarity matching of mathematical expressions. Zhong et al. [10] presented a dynamic pruning algorithm for inverted index, representing mathematical expressions as OPTs (Operator Trees), which improves retrieval efficiency for substructures of mathematical expressions.

Neural network methods have achieved significant progress in natural language-

related tasks, but their performance on mathematical language-related tasks remains an active research area. Gao et al. [11] proposed a formula vector generation method based on "formula2vec" by analyzing feature differences between natural and mathematical languages. In pursuit of attaining heightened semantic information during embedding, Dadure et al. [12] proposed a contextual formula embedding method that retrieves syntactically and semantically similar formulas, sub-formulas, and parent formulas, highlighting the importance of formula context in mathematical information retrieval. Peng et al.'s MathBERT pre-training model [13] can capture the semantic structure information of formulas by concurrently training the formulas and the contexts that relate to them. Dai [14] proposed NTFEM, which extends N-ary tree representations of MathML formulas to one-dimensional linear sequences, uses a word embedding model to obtain the sub-structure vector, and applies a weighting function to obtain a weighted average embedding vector.

In the realm of image-based mathematical expression retrieval, Marinai et al. [15] proposed a mathematical symbol retrieval approach based on visual bag-of-words encoding, which employed self-organizing maps to cluster shape context into appropriate visual dictionaries, facilitating efficient retrieval of mathematical symbols. Zanibbi et al. [16] used content-based image retrieval to match binarized and decomposed query images with expression images.

In summary, current mathematical expression retrieval mainly focuses on unimodal approaches. For text modal retrieval, methods based on representation trees and neural networks embedding expressions into vectors predominate. However, these methods may overlook structural information. Image modal retrieval, on the other hand, tends to decompose mathematical expression images into symbol images or connected components for retrieval, often emphasizing symbol similarity at the expense of semantic information. To address these limitations, our study integrates image and text retrieval outcomes, achieving a multimodal retrieval model based on mathematical expression images and text. This approach considers both modalities' similarities, leading to more rational retrieval outcomes.

In the field of Mathematical Information Retrieval (MIR), precise assessment of expression similarity holds paramount importance. While prevalent models, such as those proposed in [13,14], rely on cosine distance, the intricate two-dimensional characteristics of mathematical expressions demand a more objective approach. The utilization of Hesitant Fuzzy Set (HFS) theory emerges as an apt choice for addressing uncertainty and multi-attribute evaluation, offering a versatile means of representing hesitant information [17–21]. The integration of HFS theory into mathematical expression retrieval enhances the holistic assessment of diverse attribute features.

3. The Proposed Model Overview

Figure 1 represents the framework diagram of the multi-modal retrieval model for mathematical expressions based on ConvNeXt and HFS. Users input images and text of query mathematical expressions and receive a ranked list after retrieval. This system retrieves mathematical expressions from both image and text modalities. In the Image-Modal Retrieval Module, ConvNeXt [22] is used for image feature extraction and Symbol Level Features Extraction (SLFE) for symbol-level features of the mathematical expression images. Initial retrieval is conducted based on image features, followed by ranking based on symbol-level similarity, yielding descending search outcomes. In the Text-Modal Retrieval Module, the Formula Description Structure (FDS) [23] is utilized for the analysis of textual expressions, resulting in multiple attribute features of mathematical symbols. Additionally, the Hesitant Fuzzy Set (HFS) [24] is used to compute hesitant fuzzy similarity between mathematical query and candidate expressions, with results returned in descending similarity order. For result enhancement, the Rank Fusion Module employs Reciprocal Rank Fusion (RRF) to combine rankings generated by distinct modules.

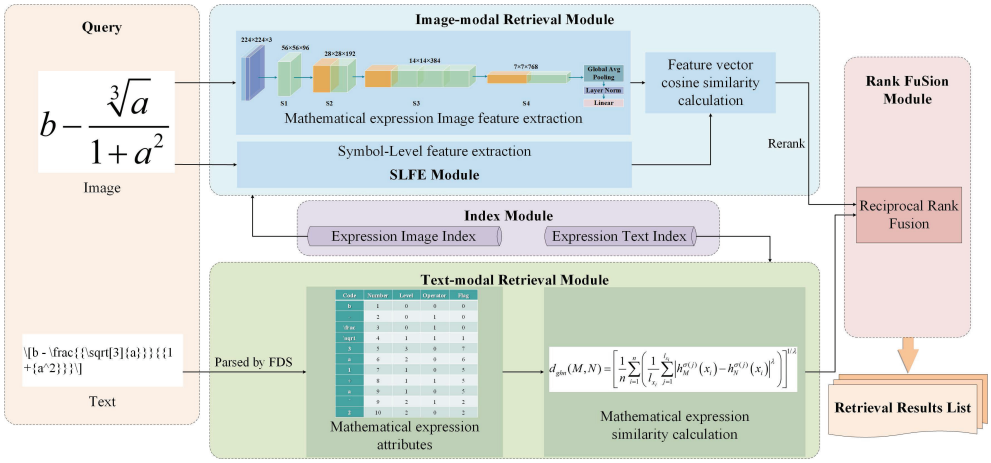

Figure 1. Framework diagram of mathematical expression multi-modal retrieval.

In our research, we broaden the modal of mathematical expression retrieval to include both image and text forms, allowing retrieval from LaTeX and MathML formats. Additionally, HFS enables a flexible representation of uncertain information, facilitating a comprehensive assessment of attribute influences on decision-making. The membership degrees of mathematical expression attributes obtained through FDS analysis are computed using hesitant fuzzy sets, and the similarity between the query and candidate expressions is determined based on the HFS similarity calculation formula.

It is essential to highlight that, as of now, the method proposed in this paper faces limitations in retrieving semantically equivalent formulae to the queried ones. This is attributed to the following factors: The transformation of semantically equivalent mathematical expressions necessitates a profound understanding of domain-specific knowledge, which currently poses challenges for automated processing. The inherent semantics of mathematical expressions can be notably ambiguous, even for human comprehension, particularly when presented in isolation without adequate context. For example, the expression "mn" can be interpreted either as a variable "mn" or as a multiplication operation between "m" and "n". This complexity further complicates machine-based semantic equivalence determination.

4. Methods

4.1. Retrieval of Mathematical Expressions in Image Modal

4.1.1. Extraction of Mathematical Expression Image Features

While meeting the requirements, the configuration of ConvNeXt-Tiny is simpler and has fewer parameters compared to other versions. It exhibits faster feature extraction capabilities, facilitating the expedited construction of feature databases. Consequently, this study employs the ConvNeXt-Tiny neural network model [22] for feature extraction in the image modal of mathematical expressions. Figure 2 shows the network structure of ConvNeXt-Tiny. It consists of four stages with different numbers of ConvNeXt blocks, each producing a feature map with different dimensions. It has strong feature extraction ability, few parameters, and low hardware requirements during training.

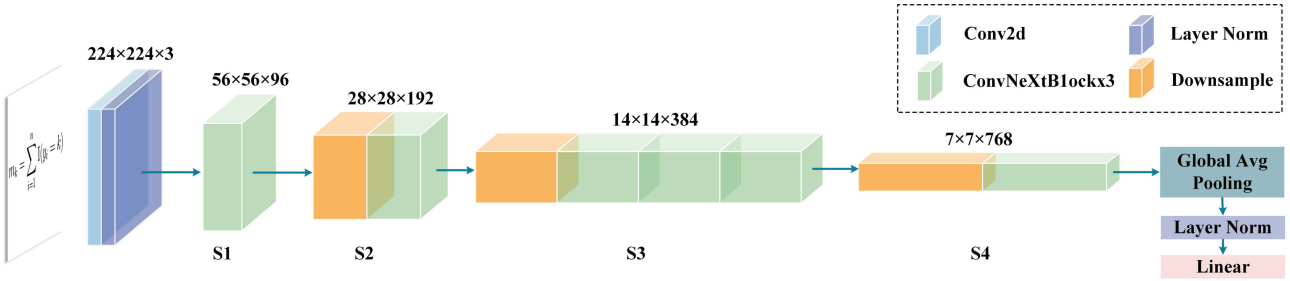

Figure 2. Network structure of ConvNeXt-Tiny.

The Visual Transformer (ViT) [25] has emerged as an effective alternative to convolutional neural networks for various computer vision tasks, exemplified by models such as the Swin Transformer [26]. Leveraging the layer structure, downsampling techniques, activation functions, data processing methods, anti-bottleneck architecture, and deep convolution inspired by the Swin Transformer, ConvNeXt [22] further enhances the image feature extraction performance.

After extracting image features using the Convnext network, an image feature index is constructed. During retrieval, cosine distance is employed to calculate the similarity between feature vectors.

4.1.2. Symbol-Level Feature Extraction Module

Compared with other types of images, mathematical expression images have unique features, and the extraction of symbol-level features is helpful to the retrieval of mathematical expressions. Mathematical expressions are distinct in their two-dimensional structure, complex symbols, and specific spatial arrangements. To accurately retrieve and evaluate mathematical expressions, it is essential to capture the symbol-level details, which are crucial for understanding their meaning and structure.

The Symbol Level Features Extraction (SLFE) module extracts symbol-level features from mathematical expression images by segmenting the input image into individual symbol blocks. Given a query image of a mathematical expression E_Q, consider $E_S = \{e_{s1}, e_{s2}, \cdots, e_{sn}\}$ as the set of symbol blocks representing expression elements and n as representing the total count of these symbol blocks ($e_{si} \in \mathrm{R}^{30\times30}$). Position vectors are calculated for element symbol blocks to retain spatial information. We use the connected component labeling algorithm to obtain element symbol blocks in mathematical expression images [27].

Figure 3a shows attached symbols "a" and "c", which are separated using the connected component labeling algorithm (Figure 3b). Each resulting component forms an element symbol block of 30 × 30 pixels. Position vectors $P_{si} = \{p_1, p_2, \cdots, p_n\}$ of the element symbol block e_{si} are computed for the symbol block set S, where $p_j = (t_j, d_j, l_j, r_j, ro_j)$ represents the position vector for the j-th element symbol block (Figure 4).

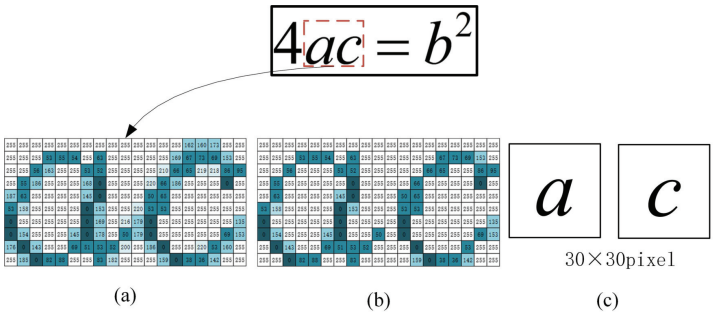

Figure 3. Example image of segmentation (**a**) adhered grayscale image (**b**) two components after segmentation (**c**) symbol block.

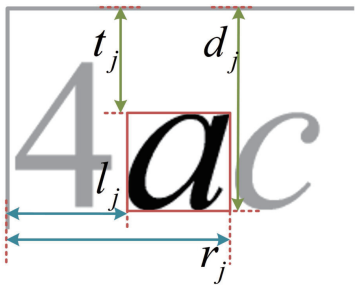

Figure 4. Position vector feature map of the element symbol block.

Here t_j, d_j, l_j, and r_j represent the distance between the top, down, left, and right edges of the j-th element symbol block and the top and left edges of the query image E_Q. The position vector elements are standardized to range between 0 and 1 using Equation (1), where $d_{\max} = \max\{d_1, \ldots, d_n\}$, $r_{\max} = \max\{r_1, \ldots, r_n\}$, $\frac{d_{\max}}{r_{\max}} = ro_i$ are the height-to-width ratio of E_Q, for recording the size proportion of the expression.

$$\mathbf{p}_i = \left(\frac{t_i}{d_{\max}}, \frac{d_i}{d_{\max}}, \frac{l_i}{r_{\max}}, \frac{r_i}{r_{\max}}, \frac{d_{\max}}{r_{\max}} \right) \tag{1}$$

In this study, we created a database of indexed mathematical expression images with their image features and symbol-level features. During the query process, we retrieve results based on image feature similarity, then sort by symbol-level feature similarity before outputting the final results.

4.2. Retrieval of Mathematical Expressions in Text Modal

4.2.1. Parsing and Extraction of Textual Attributes in Mathematical Expressions

The FDS [23] is used to analyze mathematical expressions and extract their properties. Each symbol in the expression has four attribute values: Number, Level, Operator, and Flag. As shown in Equation (2), the attributes obtained by parsing and extraction using FDS are represented as a quadruple array.

$$A(S_i) = (N_i, L_i, O_i, F_i) \tag{2}$$

The meaning of each attribute value is as follows: (1) N_i denotes the sequence number of S_i in the expression. (2) L_i represents the horizontal baseline position of the symbol S_i in the expression, as shown in Figure 5. (3) O_i is the function code of S_i , indicating whether the current symbol is an operator ($O_i = 1$) or an operand ($O_i = 0$). (4) F_i is the spatial flag information of S_i , which shows how the present symbol and the prior symbol are related. F_i can take values from 1 to 8, representing "above", "superscript", "right", "subscript", "below", "contains", "left-superscript", and "left-subscript". The main baseline's symbols have $F = 0$.

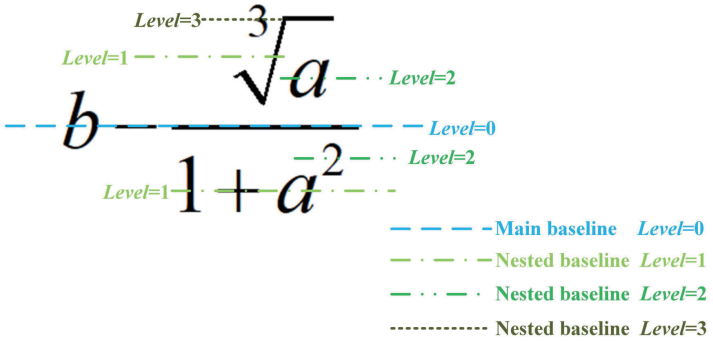

Figure 5. Sample plot of the baseline in the mathematical expression.

4.2.2. Text Similarity Calculation of Mathematical Expressions

FDS analysis on mathematical expressions yields four attribute values. A single distance calculation may introduce errors in similarity measurement. A fuzzy set [28] is a group of items with a variety of membership grades. To deal with multi-attribute evaluation indicators, we used hesitant fuzzy sets proposed by Torra [24] to calculate the similarity.

Definition 1. *Let X be a non-empty set. A hesitant fuzzy set [29] H is defined as follows:*

$$H = \left\{ < x, h_{H(x)} > | \ x \in X \right\} \tag{3}$$

In this context, $h_{H(x)}$ is the hesitant fuzzy element, which is a set of several possible degrees of membership of an element x in the set H of the elements in $X = \{x_1, x_2, \cdots, x_n\}$ [29]. Its value range is [0,1].

Let M and N be hesitant fuzzy sets on the non-empty set $X = \{x_1, x_2, \cdots, x_n\}$. The generalized hesitant fuzzy standard distance and similarity [29] between M and N are respectively defined as:

$$d_{ghn}(M,N) = \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{l_{x_i}} \sum_{j=1}^{l_{x_i}} \left| h_M^{\sigma(j)}(x_i) - h_N^{\sigma(j)}(x_i) \right|^{\lambda} \right) \right]^{1/\lambda} \tag{4}$$

$$sim(M,N) = 1 - d_{ghn}(M,N) \tag{5}$$

In this equation, $d_{ghn}(M,N)$ denotes the generalized hesitant fuzzy standard distance between hesitant fuzzy sets M and N, while $sim(M,N)$ represents their corresponding similarity [29]. $\lambda > 0$, when $\lambda = 1$, $d_{ghn}(M,N)$ is the hesitant fuzzy Hamming distance, and when $\lambda = 2$, $d_{ghn}(M,N)$ is the hesitant fuzzy Euclidean distance; $h_M^{\sigma(j)}(x_i)$ and $h_N^{\sigma(j)}(x_i)$ refer to the j-th largest element values in $h_M(x_i)$ and $h_N(x_i)$, respectively. Additionally, $l_M(x_i)$ and $l_N(x_i)$ represent the number of elements in $h_M(x_i)$ and $h_N(x_i)$, respectively. $l_{x_i} = \max(l_M(x_i), l_N(x_i))$.

The mathematical expression of the four-tuple attribute (N_i, L_i, O_i, F_i) obtained from FDS analysis is used to construct a hesitant fuzzy element set $h_{A(x)} = \{\mu_{h_N}, \mu_{h_L}, \mu_{h_O}, \mu_{h_F}\}$ as an evaluation attribute. $\mu_{h_N}, \mu_{h_L}, \mu_{h_O}, \mu_{h_F}$ represent the hesitant fuzzy membership functions corresponding to each evaluation attribute, as shown in Table 1.

Table 1. Formulas of hesitant fuzzy membership functions for evaluation attributes.

Membership Function Formula	Function Description
$\mu_{h_N}(ME_{\mathrm{DB}i}, ME_{\mathrm{Q}}) = e^{-\left(\frac{\mathrm{Number}_{\mathrm{DB}i} - \mathrm{Number}_{\mathrm{Q}}}{\sigma}\right)^2}$	σ is a balancing factor that ensures the value of μ_{h_N} remains in the range of [0,1].
$\mu_{h_L}(ME_{\mathrm{DB}i}, ME_{\mathrm{Q}}) = e^{-\alpha \lvert \mathrm{Level}_{\mathrm{DB}i} - \mathrm{Level}_{\mathrm{Q}} \rvert}$	α is a balancing factor that ensures the value of μ_{h_L} remains in the range of [0,1].
$\mu_{h_O}(ME_{\mathrm{DB}i}, ME_{\mathrm{Q}}) = \{(s_o, Operator(x))\}$	$Operator(x) = 1$ indicates that the current symbol is an operator, and $s_o = 1$. Otherwise, the current symbol is an operand, $s_o = 0.5$.
$\mu_{h_F}(ME_{\mathrm{DB}i}, ME_{\mathrm{Q}}) = \left\{ \left(f_o, Flag_{(\mathrm{DB}i,\mathrm{Q})} \right) \right\}$	$Flag_{(\mathrm{DB}i,\mathrm{Q})}$ is used to determine the spatial positional relationship of identical terms in two mathematical expressions. If the positional relationship between the two is the same, $f_o = 1$; otherwise, $f_o = 0$.

Definition 2. *Let ME_{Q} be a query expression, $ME_{\mathrm{DB}i}(i = 1, 2, \cdots, n)$ be a dataset containing n mathematical expressions.*

Definition 3. *The formula for calculating the similarity between two mathematical expressions is as follows:*

$$sim(ME_{\mathrm{Q}}, ME_{\mathrm{DB}i}) = 1 - \left[\frac{1}{4} \sum \left(\frac{1}{l_{P_p}} \sum_{j=1}^{l_{P_p}} \left| h^{\sigma(j)}_{ME_{\mathrm{Q}}}(P_p) - h^{\sigma(j)}_{ME_{\mathrm{DB}i}}(P_p) \right|^{\lambda} \right) \right]^{1/\lambda} \tag{6}$$

Here, $h_{ME_{\mathrm{Q}}}(P_p)$ and $h_{ME_{\mathrm{DB}i}}(P_p)$ represent the sets of hesitant fuzzy elements corresponding to ME_{Q} and $ME_{\mathrm{DB}i}$, respectively. The elements in $h_{ME_{\mathrm{Q}}}(P_p)$ and $h_{ME_{\mathrm{DB}i}}(P_p)$ represent the membership degrees of each attribute value of expressions ME_{Q} and $ME_{\mathrm{DB}i}$, respectively. l_{P_p} represents the number of evaluated attribute values, $h^{\sigma(j)}_{ME_{\mathrm{Q}}}(P_p)$ and $h^{\sigma(j)}_{ME_{\mathrm{DB}i}}(P_p)$ represent the degree of membership values of the $\sigma(j)$-th factor in $h_{ME_{\mathrm{Q}}}(P_p)$ and $h_{ME_{\mathrm{DB}i}}(P_p)$, respectively.

4.3. Ranking Fusion of Multi-Modal Retrieval Results

The Reciprocal Rank Fusion (RRF) [30] combines the image and text modal retrieval ranking. We use Equation (7) to calculate the fusion score S_{RRF} of a mathematical expression e, based on the retrieval ranking $r(e)$ and a constant k to reduce the influence of highly scored documents.

$$S_{\mathrm{RRF}}(e \in ME_{\mathrm{DB}i}) = \sum_{r \in R} \frac{1}{k + r(e)} \tag{7}$$

To determine the optimal value for k, we conducted experiments with various values. The experimental results are presented in Table 2. As seen in Table 2, with the increase in the value of k, the MAP obtained from the fusion results initially increases and then decreases. The maximum MAP value is achieved at k = 60. Therefore, in our experiments, we choose a value of k equal to 60.

Table 2. The influence of k on the final MAP value of RRF fusion results.

k	**20**	**40**	**60**	**80**	**100**
MAP	0.763	0.767	0.774	0.771	0.769

Illustratively, taking the query expression "ME" as an example, the top 5 mathematical expression IDs, ranks, and fusion scores for multimodal retrieval results are presented in Table 3. This culminates in a retrieval ranking of {3,4,2,1,6,5}. In this case, the top five results in the image-modal retrieval did not include the mathematical expression with ID = 5. However, the top five results in the text-modal retrieval contained the expression with ID = 5. This discrepancy might be due to the image modality retrieval method ranking this candidate expression lower (beyond the 5th position), resulting in a score of 0 in the image modality retrieval calculation. Consequently, $S_{\mathrm{RRF}}(\mathrm{id}_5) = \frac{1}{60+5} = 0.01538$,

Table 3. Illustration of fusion score calculation for multimodal retrieval results.

ID of Result	**Rank of Image**	**Rank of Text**	**Fusion Score**
1	5	4	$S_{\mathrm{RRF}}(\mathrm{id}_1) = \frac{1}{60+5} + \frac{1}{60+4} = 0.03100$
2	4	2	$S_{\mathrm{RRF}}(\mathrm{id}_2) = \frac{1}{60+4} + \frac{1}{60+2} = 0.03175$
3	3	1	$S_{\mathrm{RRF}}(\mathrm{id}_3) = \frac{1}{60+3} + \frac{1}{60+1} = 0.03226$
4	2	3	$S_{\mathrm{RRF}}(\mathrm{id}_4) = \frac{1}{60+2} + \frac{1}{60+3} = 0.03200$
5	Null	5	$S_{\mathrm{RRF}}(\mathrm{id}_5) = \frac{1}{60+5} = 0.01538$
6	1	Null	$S_{\mathrm{RRF}}(\mathrm{id}_6) = \frac{1}{60+1} = 0.01639$

This method is not affected by similarity scores and only depends on the ranking of retrieval results. It gives higher rankings to result items that a ranking model strongly prefers and ranks result items that are weakly preferred by multiple models less highly.

5. Experimental Results and Discussion

5.1. Experimental Dataset and Environment

We obtained 592,345 mathematical expressions from 14,274 articles in the ArXiv, a free and open dataset created by Cornell University researchers, and extracted 250,045 expressions from 11,770 articles in the NTCIR-mair-Wikipedia-corpus (NTCIR). We employ the method proposed by Xu et al. [31] to extract mathematical expressions images from scientific literature. The expressions are stored in both image and LaTeX text formats. The experimental environment is shown in Table 4.

Table 4. Experimental environment.

Experimental Environment	Configuration
Processor	AMD EPYC
RAM	16 GB
Operating system	Windows 10
Graphics card	RTX A5000
Video memory	24 GB
Python version	3.8
Pytorch version	1.11.0

5.2. Evaluation Protocol and Metrics

We randomly select ten representative mathematical expressions from the dataset for our experiments, as shown in Table 5. We employ Mean Average Precision (MAP) and Discounted Cumulative Gain (DCG) [32] to evaluate the effectiveness of our method in this paper. Three graduate students majoring in computer science have assigned a relevance judgment to the top 20 retrieval results for query expressions. Relevance judgments range from 1 (irrelevant) to 5 (highly relevant) based on the top-k retrieval results for all queries. This approach simulates user assessment of retrieval results and allows us to calculate Average Precision (AP), MAP, and nDCG to evaluate the effectiveness of our model.

Table 5. 10 query expressions in system experiment.

No.	Query Expressions	No.	Query Expressions
1	$\mathcal{M}(x,y) = \frac{1}{2}\left(1 - \mathrm{erf}\left[\frac{\sqrt{\pi}d(x,y)}{\delta}\right]\right)$	6	$\cos(\theta)$
2	$\frac{a}{b}$	7	$\mathcal{L} = \prod_{i=1}^{N} p(\eta_i, E_i, \phi_i, \Pi, \eta_0)$
3	$K = \begin{bmatrix} f \cdot \mu_x & \gamma & u_x \\ 0 & f \cdot \mu_y & u_y \\ 0 & 0 & 1 \end{bmatrix}$	8	$\Omega(t) = \sum_{k=0}^{\infty} H_k(t)$
4	$\mathrm{x} + \mathrm{y}$	9	$\hat{\Omega}_s = \int_0^{\tau} \hat{P}(S \mid T = t, H)dt$
5	$\sigma(x) = \frac{1}{1+\exp(-x)}$	10	$\dot{r}(0) \triangleq \lim_{\tau \to 0} \frac{dr(\tau)}{d\tau}$

5.3. System Experiment

The average precision and MAP were calculated for ten query expressions, as shown in Table 6.

Table 6. The average precision and MAP@k of the multi-modal retrieval method on the two datasets.

Dataset	P@3	P@5	P@10	P@20	MAP@3	MAP@5	MAP@10
ArXiv	0.833	0.800	0.690	0.500	0.858	0.833	0.787
NTCIR	0.800	0.720	0.630	0.405	0.825	0.818	0.760

Because there were more similar expressions in the ArXiv dataset, the results were consistently better than those from the NTCIR dataset. The low P@20 values are be-

cause there are fewer than 20 expressions in the dataset which are highly similar to some complex queries.

5.4. Ablation Experiments

Through experiments, we observed that the SLFE module prioritizes results that exhibit greater structural and length similarity to the query expression. The relevance scores of the retrieved expressions were used to calculate P@k, MAP@10, and nDCG@10 values. We evaluated the impact of SLFE re-ranking on image-modal retrieval by comparing P@k, MAP@10, and nDCG@10 values. The results in Figure 6 show that SLFE significantly improves both retrieval accuracy and ranking results by attending to symbol-level features.

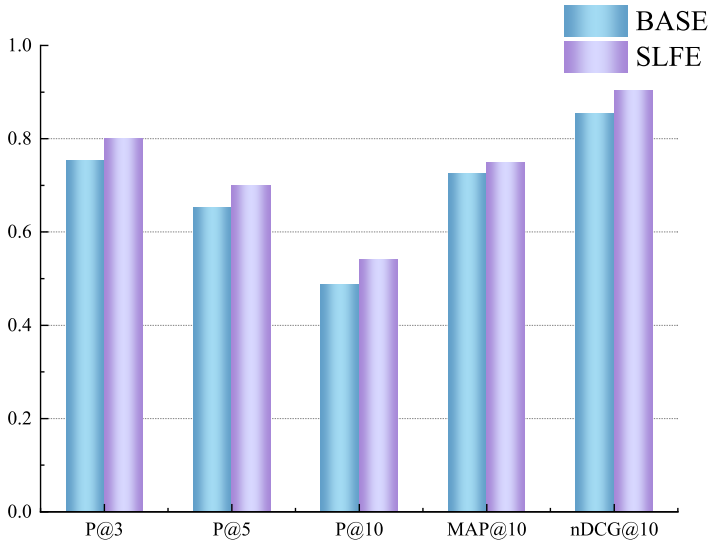

Figure 6. Ablation experimental results of the SLEF module.

The image-modal retrieval method is Method 1, the text-modal retrieval method is Method 2, and the image-text modal RRF method is Method 3. For the query $O(\log \frac{1}{\gamma})$, the top 5 results from Method 1 and Method 2 are shown in Table 7. Both methods retrieve exact matches as the first formula. Method 1 places a stronger emphasis on the structural information of mathematical expressions. For instance, because the candidate expression $\tilde{O}(\log \frac{1}{\gamma})$ closely aligns with the query expression $O(\log \frac{1}{\gamma})$ in terms of structure, it is ranked higher. Additionally, method 1 can retrieve formulas that share the same structure but have different symbolic representations, such as expression $O(\log \frac{\varepsilon}{\gamma})$. On the other hand, method 2 prioritizes the symbols within the expressions, such as candidate expressions that all contain $\log \frac{1}{\gamma}$.

Table 7. Method 1 vs. Method 2 results for query: $O(\log \frac{1}{\gamma})$.

Rank	Method 1	Method 2
1	$O(\log \frac{1}{\gamma})$	$O(\log \frac{1}{\gamma})$
2	$\tilde{O}(\log \frac{1}{\gamma})$	$d \cdot O(\log \frac{1}{\gamma})$
3	$O(\log \frac{\varepsilon}{\gamma})$	$\frac{1}{\gamma^2} O(\log \frac{1}{\gamma})$
4	$O(C_{\text{link}} \cdot d \log \frac{1}{\gamma})$	$T = O(\log \frac{1}{\gamma})$
5	$\tilde{O}(\frac{1}{\gamma^2} \log^2(\frac{\theta}{\varepsilon}))$	$O(C_{\text{link}} \cdot d \log \frac{1}{\gamma})$

The average P@k and MAP@k of the three methods are summarized in Figure 7. Text-modal retrieval outperformed image-modal retrieval because FDS is more concerned with the type of the symbol itself (e.g., whether the symbol is an operator, operand, constant, or variable). However, we found that within the results of image modality retrieval,

expressions more structurally similar to the query image are ranked higher. This suggests that image modality retrieval methods are more adept at capturing structural information inherent in expressions. The integration of both modalities can lead to more accurate and effective retrieval results. The image-text modal RRF method was effective, as the fusion ranking of two modal retrieval results was better than the single modal retrieval ranking, demonstrating the complementary results of the two modal retrievals.

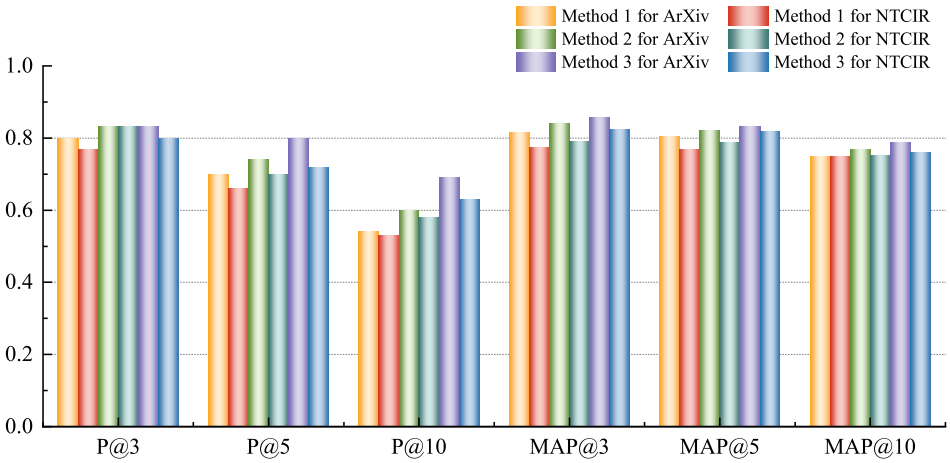

Figure 7. The average P@k and MAP@k values of the three methods.

5.5. Contrast Experiments

SearchOnMath [33] is a retrieval tool for scientific literature and Wikipedia pages based on mathematical expressions. Tangent-CFT [34] is a mathematical expression embedding model that represents mathematical expressions using Operator trees (OPTs) and Symbol Layout Trees (SLT) and ultimately generates formula embeddings using fastText. Experimental trials were conducted utilizing the mathematical expressions from Table 5. A comparative analysis was performed between our proposed approach and other methods. The results for MAP@k and average P@k are presented in Table 8. The proposed image-text modal RRF method achieves an accuracy of 0.660 and an average precision of 0.774 for the top 10 results. The nDCG@10 results for each expression retrieval outcome under various methods are depicted in Figure 8.

Table 8. Comparison of P@k and MAP@k between other methods and the proposed method.

Method	P@3	P@5	P@10	MAP@3	MAP@5	MAP@10
Image modal retrieval method	0.784	0.680	0.535	0.796	0.787	0.749
Text modal retrieval method	0.833	0.720	0.590	0.816	0.805	0.760
Image-text RRF method	0.817	0.760	0.660	0.842	0.826	0.774
Search on Math	0.600	0.620	0.520	0.717	0.709	0.664
Tangent-CFT	0.820	0.683	0.583	0.809	0.785	0.739

The method proposed in this paper, which combines the attribute features of mathematical expression text and images, is able to simultaneously consider both the holistic characteristics and symbol-level features of the image modality, as well as the symbol types characteristic of the text modality. This results in ranking outcomes that better align with user requirements. Therefore, the NDCG@10 value of the method proposed in this paper is generally higher compared to the contrastive methods. In comparison to the SearchOnMath and Tangent-CFT methods, our approach offers user convenience, avoiding the cumbersome manual input of mathematical expressions. Additionally, it effectively balances the structural information of the image modality and the symbolic attributes of the text modality, optimizing the output ranking. Consequently, it efficiently circumvents the issue of mathematical expression text losing structural features.

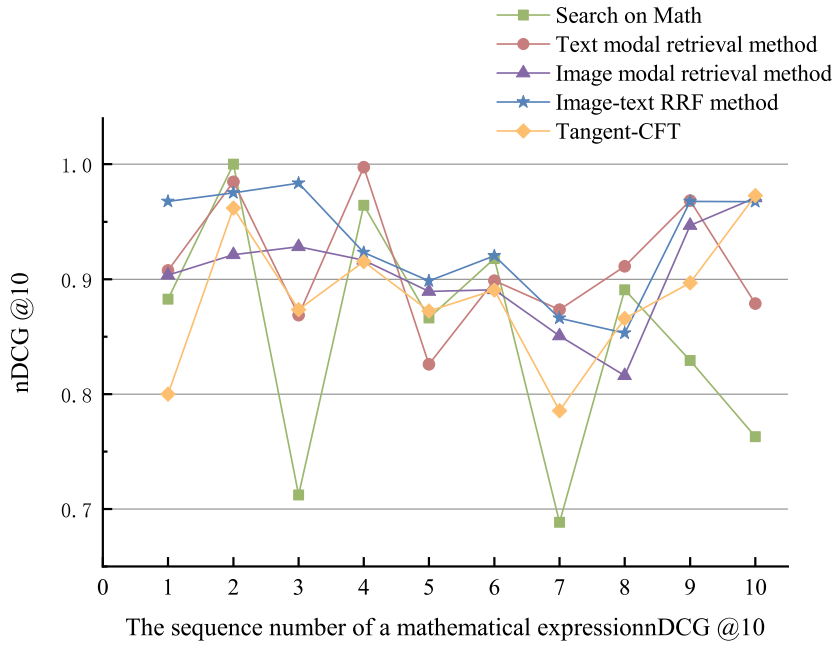

Figure 8. Comparison of nDCG@10 of the proposed method and other methods.

6. Conclusions

We have constructed a multimodal mathematical expression retrieval model by leveraging the RRF method to fuse the retrieval results from both image and text modalities. In this model, the image-modal retrieval module employs the ConvNext and SLDE modules to extract mathematical expression image features and symbol-level features, respectively. Based on the similarity of these features, the results are then outputted in descending order. Meanwhile, the text-modal retrieval module analyzes and extracts attribute values of textual expressions using FDS. To determine the hesitant fuzzy similarity between the candidate expression and the mathematical query expression, the hesitant fuzzy set theory is introduced in the meantime. The RRF amalgamates image and text modality retrieval outcomes, culminating in the ultimate combined result. Concurrently, a subset of literature from the publicly accessible ArXiv dataset was extracted and utilized to construct a dataset encompassing 592,345 mathematical expressions.

The multimodal retrieval model eliminates constraints on the user's input modalities for mathematical query expressions, thereby ensuring a comprehensive evaluation of both modality similarities. This approach enhances the rationality of retrieval outcomes.

In the future, our research will further explore multimodal mathematical expression retrieval in the following areas:

1. Enhancing image modality feature extraction by incorporating contextual parsing of mathematical expressions, thereby preserving a richer semantic understanding.
2. Extending the application of the proposed method to scientific literature retrieval. This involves integrating mathematical expressions with intrinsic attributes of scientific literature, including keywords, to enhance retrieval accuracy during scientific literature searches.

Author Contributions: Data curation, R.L. and J.W.; Formal analysis, R.L., J.W. and X.T.; Funding acquisition, X.T.; Investigation, R.L.; Methodology, R.L., J.W. and X.T.; Software, R.L.; Supervision, X.T.; Validation, R.L. and X.T.; Visualization, R.L. and X.T.; Writing—original draft, R.L.; Writing—review & editing, R.L., J.W. and X.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Natural Science Foundation of Hebei Province of China (Grant No. F2019201329).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, X.; Zhu, Y.; Liu, S.; Ju, J.; Qu, Y.; Cheng, G. Dyrren: A dynamic retriever-reranker-generator model for numerical reasoning over tabular and textual data. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 13139–13147.
2. Satpute, A.; Greiner-Petter, A.; Schubotz, M.; Meuschke, N.; Aizawa, A.; Gipp, B. TEIMMA: The First Content Reuse Annotator for Text, Images, and Math. *arXiv* **2023**, arXiv:2305.13193.
3. Gipp, B.; Greiner-Petter, A.; Schubotz, M.; Meuschke, N. Methods and Tools to Advance the Retrieval of Mathematical Knowledge from Digital Libraries for Search-, Recommendation-, and Assistance-Systems. *arXiv* **2023**, arXiv:2305.07335.
4. Xu, P.; Zhu, X.; Clifton, D.A. Multimodal learning with transformers: A survey. *arXiv* **2023**, arXiv:2206.06488.
5. Suzuki, M.; Matsuo, Y. A survey of multimodal deep generative models. *Adv. Robot.* **2022**, *36*, 261–278. [CrossRef]
6. Goel, M.; Goyal, V.; Venktesh, V. MWPRanker: An Expression Similarity Based Math Word Problem Retriever. *arXiv* **2023**, arXiv:2307.01240.
7. Pfahler, L.; Morik, K. Self-Supervised Pretraining of Graph Neural Network for the Retrieval of Related Mathematical Expressions in Scientific Articles. *arXiv* **2022**, arXiv:2209.00446.
8. Schellenberg, T.; Yuan, B.; Zanibbi, R. Layout-based substitution tree indexing and retrieval for mathematical expressions. In Proceedings of the Document Recognition and Retrieval XIX, SPIE, Burlingame, CA, USA, 25–26 January 2012; Volume 8297, pp. 126–133.
9. Hu, X.; Gao, L.; Lin, X.; Tang, Z.; Lin, X.; Baker, J.B. Wikimirs: A mathematical information retrieval system for wikipedia. In Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, Indianapolis, IN, USA, 22–26 July 2013; pp. 11–20.
10. Zhong, W.; Rohatgi, S.; Wu, J.; Giles, C.L.; Zanibbi, R. Accelerating substructure similarity search for formula retrieval. In Proceedings of the Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, 14–7 April 2020; pp. 714–727.
11. Gao, L.; Jiang, Z.; Yin, Y.; Yuan, K.; Yan, Z.; Tang, Z. Preliminary Exploration of Formula Embedding for Mathematical Information Retrieval: can mathematical formulae be embedded like a natural language? *arXiv* **2017**, arXiv:1707.05154.
12. Dadure, P.; Pakray, P.; Bandyopadhyay, S. Embedding and generalization of formula with context in the retrieval of mathematical information. *J. King Saud-Univ.-Comput. Inf. Sci.* **2022**, *34*, 6624–6634. [CrossRef]
13. Peng, S.; Yuan, K.; Gao, L.; Tang, Z. Mathbert: A pre-trained model for mathematical formula understanding. *arXiv* **2021**, arXiv:2105.00377.
14. Dai, Y.; Chen, L.; Zhang, Z. An N-ary tree-based model for similarity evaluation on mathematical formulae. In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 11–14 October 2020; pp. 2578–2584.
15. Marinai, S.; Miotti, B.; Soda, G. Mathematical symbol indexing using topologically ordered clusters of shape contexts. In Proceedings of the 2009 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 26–29 July 2009; pp. 1041–1045.
16. Zanibbi, R.; Yuan, B. Keyword and image-based retrieval of mathematical expressions. In Proceedings of the Document Recognition and Retrieval XVIII, SPIE, San Francisco, CA, USA, 26–27 January 2011; Volume 7874, pp. 141–149.
17. Torra, V.; Narukawa, Y. On hesitant fuzzy sets and decision. In Proceedings of the 2009 IEEE International Conference on Fuzzy Systems, Jeju Island, Republic of Korea, 20–24 August 2009; pp. 1378–1382.
18. Farhadinia, B.; Aickelin, U.; Khorshidi, H.A. Uncertainty measures for probabilistic hesitant fuzzy sets in multiple criteria decision making. *Int. J. Intell. Syst.* **2020**, *35*, 1646–1679. [CrossRef]
19. Liu, P.; Zhang, X. A new hesitant fuzzy linguistic approach for multiple attribute decision making based on Dempster–Shafer evidence theory. *Appl. Soft Comput.* **2020**, *86*, 105897. [CrossRef]
20. Farhadinia, B.; Aickelin, U.; Khorshidi, H.A. Higher order hesitant fuzzy Choquet integral operator and its application to multiple criteria decision making. *arXiv* **2020**, arXiv:2011.08183.
21. Cai, L. Interval-Valued Hesitant Fuzzy Sets and Its Application to Decision Making. Ph.D. Thesis, Zhengzhou University, Zhengzhou, China, 2013.
22. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
23. Tian, X. A mathematical indexing method based on the hierarchical features of operators in formulae. In Proceedings of the 2nd International Conference on Automatic Control and Information Engineering (ICACIE 2017), Hong Kong, China, 26–28 August 2017; pp. 49–52.
24. Torra, V. Hesitant fuzzy sets. *Int. J. Intell. Syst.* **2010**, *25*, 529–539. [CrossRef]

25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
27. Samet, H.; Tamminen, M. Efficient component labeling of images of arbitrary dimension represented by linear bintrees. *IEEE Trans. Pattern Anal. Mach. Intell.* **1988**, *10*, 579–586. [CrossRef]
28. Zadeh, L.A. Fuzzy sets. *Inf. Control.* **1965**, *8*, 338–353. [CrossRef]
29. Xu, Z.; Xia, M. Distance and similarity measures for hesitant fuzzy sets. *Inf. Sci.* **2011**, *181*, 2128–2138. [CrossRef]
30. Cormack, G.V.; Clarke, C.L.; Buettcher, S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, USA, 19–23 July 2009; pp. 758–759.
31. Xu, X.; Tian, X.; Yang, F. A retrieval and ranking method of mathematical documents based on CA-YOLOv5 and HFS. *Math. Biosci. Eng.* **2022**, *19*, 4976–4990. [CrossRef] [PubMed]
32. Jarvelin, K.; Kekalainen, J. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the ACM SIGIR Forum*; ACM: New York, NY, USA, 2017; Volume 51, pp. 243–250.
33. Oliveira, R.M.; Gonzaga, F.B.; Barbosa, V.C.; Xexéo, G.B. A distributed system for SearchOnMath based on the Microsoft BizSpark program. *arXiv* **2017**, arXiv:1711.04189.
34. Mansouri, B.; Rohatgi, S.; Oard, D.W.; Wu, J.; Giles, C.L.; Zanibbi, R. Tangent-CFT: An embedding model for mathematical formulas. In Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, Santa Clara, CA, USA, 2–5 October 2019; pp. 11–18.