IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 August 2022, accepted 16 August 2022, date of publication 29 August 2022, date of current version 6 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3202639

RESEARCH ARTICLE

FormulaNet: A Benchmark Dataset for Mathematical Formula Detection

FELIX M. SCHMITT-KOOPMANN[1,2], ELAINE M. HUANG[2], HANS-PETER HUTTER[1], (Member, IEEE), THILO STADELMANN[3,4], (Senior Member, IEEE), AND ALIREZA DARVISHY[1]

[1]Institute of Applied Information Technology, ZHAW, 8401 Winterthur, Switzerland
[2]People and Computing Laboratory, University of Zurich, 8050 Zurich, Switzerland
[3]Center for Artificial Intelligence, ZHAW, 8400 Winterthur, Switzerland
[4]European Centre for Living Technology (ECLT), 30123 Venice, Italy

Corresponding author: Felix M. Schmitt-Koopmann (scmx@zhaw.ch)

This work was supported by the Bridge Discovery program of the Swiss National Science Foundation under Grant 194677.

ABSTRACT One unsolved sub-task of document analysis is mathematical formula detection (MFD). Research by ourselves and others has shown that existing MFD datasets with inline and display formula labels are small and have insufficient labeling quality. There is therefore an urgent need for datasets with better quality labeling for future research in the MFD field, as they have a high impact on the performance of the models trained on them. We present an advanced labeling pipeline and a new dataset called FormulaNet in this paper. At over 45k pages, we believe that FormulaNet is the largest MFD dataset with inline formula labels. Our experiments demonstrate substantially improved labeling quality for inline and display formulae detection over existing datasets. Additionally, we provide a math formula detection baseline for FormulaNet with an mAP of 0.754. Our dataset is intended to help address the MFD task and may enable the development of new applications, such as making mathematical formulae accessible in PDFs for visually impaired screen reader users.

INDEX TERMS Automatic annotation, dataset, document analysis, deep learning, mathematical formula detection, page object detection.

I. INTRODUCTION

The 2008 United Nations Convention on the Rights of Persons with Disabilities [1] and the 2019 European Accessibility Act [2] require that everyday products and services be usable for people with disabilities. Nevertheless, many technologies remain inaccessible; PDFs are one such technology that frequently present a barrier for readers with visual impairments. This is especially true for scientific PDFs. For example, mathematical formulae in PDFs are usually not tagged with alternative text, making it impossible for screen reader software to read them out in a comprehensible way. Research has shown that most authors of scientific documents are unfamiliar with the concept of PDF accessibility, or lack the tools to support it [3].

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

Document analysis offers high potential for new applications, including applications for people with disabilities. One such application is the automated addition of accessibility tags to a PDF. Such accessibility tags allow a visually impaired person to read a PDF with a screen reader. Currently, tags must be added manually, which requires a great deal of time, expert knowledge, and awareness [3].

With effective document analysis, the tagging process could be automated or semi-automated, thus reducing the required time and expert knowledge necessary. This could help to increase the overall availability of tagged PDFs and as a result, give visually impaired people more complete access to information. However, the challenges of automated document analysis have not yet been solved. Searching for simple text in documents is currently possible [4]; however, the detection of more complex structures within a text, such as tables, graphs, or formulae remains problematic.

New data-driven approaches have enabled significant advancements in the document analysis field [5]. Most data-driven document analysis solutions work with images of document pages. This has the advantage that the approach can be applied regardless of the document format and version.

The first step planned for our document analysis pipeline is page object detection (POD). It aims to locate logical objects in document pages with a high semantic level, e.g., paragraphs, footnotes, tables, figures, or mathematical formulae. In the next step, these objects will be processed by formula recognition, figure classification, text analysis, and other means.

The POD task is often divided into subtasks of locating a single logical object at a time. Despite the progress of POD in recent years [4], [6], [7], some objects are still challenging to identify and need to be addressed further. One of these open problems is mathematical formula detection (MFD) [8]. MFD is especially important for scientific documents from STEM fileds (science, technology, engineering, and mathematics), because mathematical formulae are often important objects for the understanding of STEM articles. Automated processing of formulae could help to simplify and improve many tasks, such as searching for mathematical formulae in documents, extracting mathematical formulae, and making mathematical formulae accessible.

In recent years, many MFD models have been proposed [4], [6], [7], but one problem that the authors of this paper have identified is that the MFD datasets they have been evaluated on have been of limited size and quality.

A selection of the most popular POD datasets is presented in Table 1. Existing POD datasets [9], [10], [11], [12], [13], [14], [15] are of limited value for the MFD issues we are attempted to address because of three reasons. First, most POD dataset were not intended for the MFD task and hence, consider no mathematical formulae or only display formulae but not inline formulae. Second, existing datasets with inline formulae tend to be small for deep learning approaches with less than 10k pages. Third, the mathematical formulae labels have insufficient quality or are incorrect. In this, paper, we propose a new large-scale and high-quality dataset for the MFD task of scientific PDF documents. It is created from the LaTeXsource [16] of papers from arXiv.org [17].

The main contributions of this paper are as follows: (a) a novel large-scale, high-quality dataset for MFD with practical relevance for document accessibility and, in conjunction with the provided baselines, scientific use as a benchmark suite; (b) an advanced fully automated labeling pipeline for constructing similar high-quality datasets of POD of nearly any size.

Due to copyright issues, we can only provide the links to the papers used and the postprocessing scripts to reconstruct FormulaNet, but not the images of FormulaNet. The scripts are publicly available at https://github.com/felix-schmitt/FormulaNet. Due to the compiling of the LaTeXfiles, the resulting pixel values may differ. We observed that on average 0.1% of the binary pixel values and 10.4% of the color pixel values variate.

TABLE 1. Overview of a selection of the most popular POD datasets.

Dataset	Pages	Inline Labels	Inline-Accuracy	Display-Accuracy
Marmot [9]	400	Yes	76.90%	88.72%
ICDAR 2017 POD (corrected) [10]	2,417	No	-	-
IBEM [11]	8,272	Yes	96.72%	83.38%
FormulaNet	**46,672**	**Yes**	**98.08%**	**97.86%**
GROTOAP2 [12]	119,334	No	-	-
PubLayNet [13]	364,232	No	-	-
TableBank [14]	417,234	No	-	-
DocBank [15]	500,000	No	-	-

The remainder of this paper is organized as follows: Chapter II presents related work and existing datasets. Chapter III presents our definition of inline and display formulae and introduces our dataset and labeling pipeline. Chapter IV presents the baseline model and experiments to demonstrate the improvement in labeling quality. Chapter V provides concluding remarks.

II. RELATED WORK AND EXISTING DATASETS

POD has been an active research area for several years [4], [6], [7]. The MFD subtask has been researched since at least 1968 [18] and efforts in this area have increased in recent years. Traditional MFD solutions are rule-based. However, object recognition using deep learning models has achieved good results and is replacing traditional rule-based approaches. Modern MFD models use convolutional neural networks (CNN) and build upon state-of-the-art object detections models, e.g., Faster-RCNN [19], Mask-RCNN [20], and FCOS [21]. The major challenge with MFD is the variation in complexity between small single mathematical elements and large mathematical formulae. Research [23] has shown that deformable CNNs [22], with their adaptive geometric transformation, have the ability to handle large variations in size. Furthermore, Generalized Focal Loss [24] reduces the imbalance issue of positive/negative sampling of large and small objects. As baseline model, we use the 1st place solution of the in ICDAR 2021 Competition on Mathematical Formula Detection [23] with small modifications. It is built upon FCOS and uses both modifications.

The competition [4] showed that MFD models can achieve excellent results in terms of F1 scores, but inline formulae are still challenging for these models and additional work is needed to address. One reason is that large existing POD datasets do not include labels for inline formulae (ref. Table 1) and the ones containing inline formulae are limited in size and labeling quality. We explain this lack of dataset with inline formulae by the fact that inline formulae are uncommon and often not crucial for the understanding of non-STEM documents. Furthermore, the separation between inline formulae and text is not clearly defined, as presented in Chapter III-A. However, STEM documents contain many inline formulae,

and their correct processing is important for many applications, such as accessible PDFs.

We are aware of only two publicly available MFD datasets with inline formulae based on not rearranged articles such as omitting content and changing layout. One is the Marmot dataset [9] with 400 pages. Due to its small size, it is not ideal for deep learning approaches. The largest dataset with inline formulae is the IBEM dataset [11] with 8,272 pages, which is 20 times larger than Marmot, but it is still small for deep learning approaches. In comparison, DeepScores [25], an object detection dataset for music scores, which is a comparable object detection task, contains 300,000 pages. The IBEM dataset was created for the ICDAR 2021 Competition on Mathematical Formula Detection [4] to run the latest performance competition of MFD models. It was created in a fashion similar to FormulaNet, by detecting specific formula patterns in the LaTeXcode. The patterns detected were then used to create the ground truth labels.

The large-scale POD datasets are not designed for the MFD task and hence, contain no inline formulae labels. With FormulaNet, we narrow the gap between MFD datasets and large-scale POD datasets.

III. FormulaNet

This section describes the construction details and characteristics of the FormulaNet dataset. FormulaNet uses papers about High Energy Physics on arXiv.org from the years 2000, 2002, and 2003. We used the High Energy Physics papers for the FormulaNet dataset not only because such PDFs comprise many formulae, but also to make it more comparable to the IBEM dataset, which also uses High Energy Physics papers from arXiv.org.

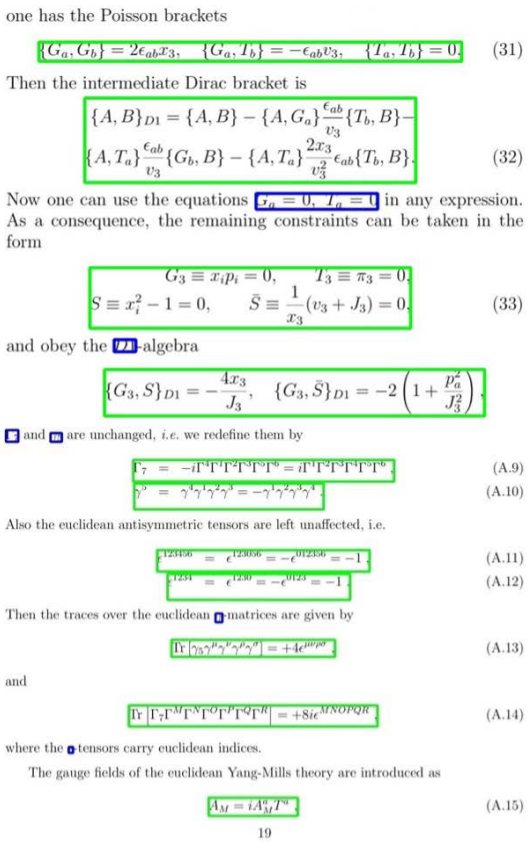
A. LABEL DEFINITIONS

There are no widely accepted standard definition for inline forumlae or display formulae. For the purposes of this reasearch, we provide working definitions of these terms based on the rules detected from the Marmot, ICDAR, and IBEM datasets:

1) INLINE FORMULAE

We define inline formulae as all math-typed elements embedded in a text, except plain numbers.

An inline formula can consist of a single math element such as γ or a more complex formula consisting of multiple such elements. A single number is not considered as an inline formula for two reasons: First, in the existing datasets most numbers are not labeled as formulae. Second, numbers can already be processed through standard text optical character recognition (OCR). However, if a number comprises math structure elements like super-scripts or fractions, we consider it an inline formula because it is a mathematical construct, and text OCR will likely have problems interpreting it correctly. Mathematical elements within tables are not considered inline formulae because detecting a table structure is a challenging task, and detecting formulae within the table


one has the Poisson brackets

$$\{G_a, G_b\} = 2\epsilon_{ab}x_3, \quad \{G_a, T_b\} = -\epsilon_{ab}v_3, \quad \{T_a, T_b\} = 0, \tag{31}$$

Then the intermediate Dirac bracket is

$$\{A,B\}_{D1} = \{A,B\} - \{A,G_a\}\frac{\epsilon_{ab}}{v_3}\{T_b,B\} - \{A,T_a\}\frac{\epsilon_{ab}}{v_3}\{G_b,B\} - \{A,T_a\}\frac{2x_3}{v_3^2}\epsilon_{ab}\{T_b,B\}. \tag{32}$$

Now one can use the equations $G_a = 0,\ T_a = 0$ in any expression. As a consequence, the remaining constraints can be taken in the form

$$G_3 \equiv x_i p_i = 0, \quad T_3 \equiv \pi_3 = 0,$$
$$S \equiv x_i^2 - 1 = 0, \quad \bar{S} \equiv \frac{1}{x_3}(v_3 + J_3) = 0. \tag{33}$$

and obey the VA-algebra

$$\{G_3, S\}_{D1} = -\frac{4x_3}{J_3}, \quad \{G_3, \bar{S}\}_{D1} = -2\left(1 + \frac{p_a^2}{J_3^2}\right),$$

S and M are unchanged, *i.e.* we redefine them by

$$\Gamma_7 = -i\Gamma^1\Gamma^2\Gamma^3\Gamma^5\Gamma^6 = i\Gamma^1\Gamma^2\Gamma^3\Gamma^4\Gamma^5\Gamma^6, \tag{A.9}$$
$$\bar{\gamma}^5 = \gamma^4\gamma^1\gamma^2\gamma^3 = -\gamma^1\gamma^2\gamma^3\gamma^4. \tag{A.10}$$

Also the euclidean antisymmetric tensors are left unaffected, i.e.

$$\epsilon^{123456} = \epsilon^{123056} = -\epsilon^{012356} = -1, \tag{A.11}$$
$$\epsilon^{1234} = \epsilon^{1230} = -\epsilon^{0123} = -1, \tag{A.12}$$

Then the traces over the euclidean D matrices are given by

$$\mathrm{Tr}\left[\gamma_5\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma\right] = +4\epsilon^{\mu\nu\rho\sigma}, \tag{A.13}$$

and

$$\mathrm{Tr}\left[\Gamma_7\Gamma^M\Gamma^N\Gamma^O\Gamma^P\Gamma^Q\Gamma^R\right] = +8i\epsilon^{MNOPQR}, \tag{A.14}$$

where the D tensors carry euclidean indices.

The gauge fields of the euclidean Yang-Mills theory are introduced as

$$A_M = iA_M^a T^a \tag{A.15}$$

19

FIGURE 1. Examples of how multicolumn display labels are separated. Green shows the display formulae and blue the inline formulae.

is a subtask of this task. For the same reason, mathematical elements in figures are not labeled as inline formulae, because formulae within figures need to be considered separately, similar to formulae within tables.

2) DISPLAY FORMULAE

We define display formulae to be all-mathematical elements isolated from the running text. Multiline display formulae are separated depending on the formula references.

Formula references are not counted as part of a formula, because they are document structure elements and not part of the formula itself. This has the advantage that the bounding box size does not depend on the existence of a formula reference. Furthermore, we decided to only split up a multiline display formula into separate formulae if there is a formula reference on each line, as shown in Fig. 1. Splitting up a display formula line-by-line would have the effect of dividing a single formula into multiple parts, thus making it more complicated to process.

B. LABELING PROCESS

The labeling pipeline starts from the LaTeXsource files. It involves two labeling steps and one correction step as shown in Fig. 2. The first step is to modify the LaTeXcode to color each LaTeXobject. Depending on the object type, we use

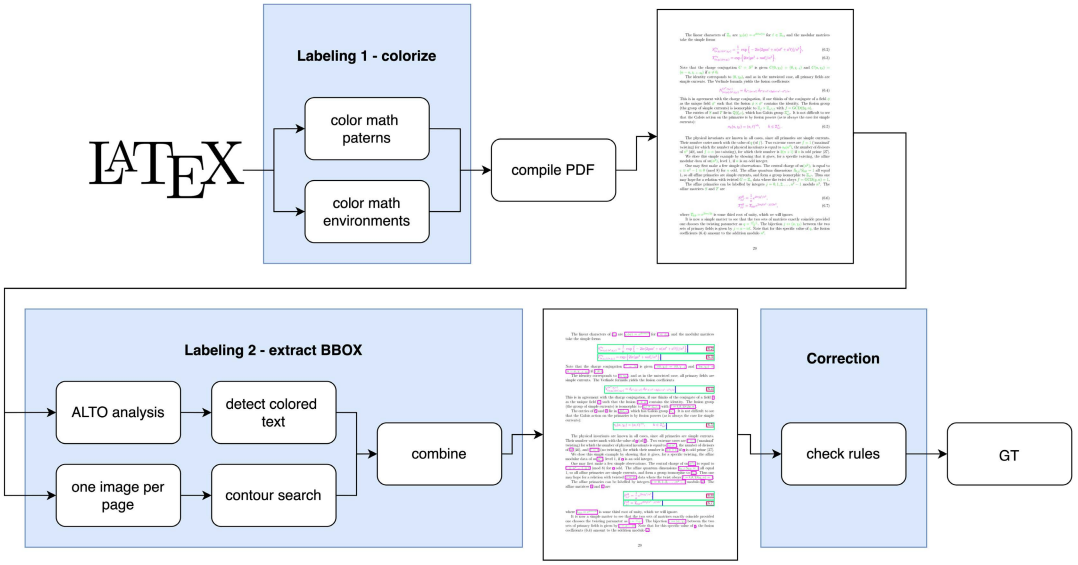


FIGURE 2. **Overview of the labeling pipeline.**

one or multiple colors to simplify the later separation. Two methods were combined to colorize the LaTeXcode. The first method uses regular expression search [26] to find predefined sequences in the LaTeXcode which are typical for a logical object class. Then, the sequences identified are colored with the xcolor package [27] and the following command:

```
\textcolor{l_color}{label}
```

The second method colors complete LaTeXenvironments with the following LaTeXcommand:

```
\AtBeginEnvironment{l_env}{l_color}
```

The modified LaTeXfile is used to render a PDF of the paper with the colored logical objects. In the second part, the colored objects of the modified PDF are detected and combined into one bounding box by heuristic rules. A combination of two methods is used to enhance the labeling quality. One method converts the PDF into the ALTO format [28] with pdfalto [29]. The resulting XML files contain information about the elements detected and it allows the identification of all colored elements. Since pdfalto is an OCR engine mainly for text it does not detect all symbols correctly. We therefore apply the second method to find the missing symbols.

For the second step, a PNG image of each page is rendered using a modified version of pdf2image [30] without anti-aliasing. This modification allows us to create images with clear contours which simplifies the contour search (OpenCV implementation [31]). This enables the detection of all missing colored pixels such as bars, heads, and other special math symbols. All BBOXs of the pdfalto and contour search are then combined with heuristic rules. Using only contour search would make it complicated or even impossible to get the correct combination of contours to a BBOX.

The last step is the correction step. It detects labeling errors, and depending on the errors detected, deletes entire pages or even the whole document. The rules applied are based on our observations during developing the pipeline, e.g.:

- These rules indicate an error in the coloring step:
  - If the paper has 3 or fewer pages, the document is discarded.
  - If the paper has no inline or display formulae, the document is discarded.
  - If there exist black pixels in a 30-pixel border of the document, the document is discarded.
- These rules indicate and error in the extracting BBOX step:
  - If there are more than 3 small display formulae, the page is discarded.
  - If there are not enough black pixels in an image, the page is discarded.
  - If the sum of all label areas is less than 10% of the page, the page is discarded.

After the correction step, a txt-file of each page is created with the detected BBOXs and a corresponding JPG image of the page with a resolution of 1447 $\times$ 2048 is saved. If the ratio of the document does not match the image ratio, a white border is added.

C. FormulaNet CHARACTERISTIC

FormulaNet consists of 46,672 pages with 175,685 display labels and 825,838 inline labels. Besides formula labels, FormulaNet contains 11 other labels (display reference, display both, header, table, figure, paragraph, caption, footnote, footnote reference, list, bibliography). We have randomly split the dataset into training (95% of the pages) and test (5% of

TABLE 2. Distribution of the labels of the FormulaNet dataset.

Label	Train (44,338 pages)		Test (2,334 pages)	
	Total	Per Page	Total	Per Page
Inline Formulae	784,978	17.71±13.2	40,860	17.51±13
Display Formulae	166,759	3.76±2.9	8,936	3.83±2.8
Bibliography	1,086	0.02±0.2	56	0.02±0.2
Caption	3,671	0.08±0.3	203	0.09±0.3
Display Reference	144,800	3.27±2.9	7,799	3.34±2.8
Display Formulae + Reference	144,800	3.27±2.9	7,799	3.34±2.8
Footnote	8,109	0.18±0.4	438	0.19±0.4
Footnote Reference	11,576	0.26±0.7	632	0.06±0.3
Header	20,818	0.47±0.6	1,082	0.46±0.6
List	2,539	0.06±0.3	136	0.06±0.3
Paragraph	283,933	6.4±2.7	15,008	6.43±2.6
Table	1,145	0.03±0.2	49	0.02±0.2

the pages) sets. The distribution of the labels can be found in Table 2.

IV. COMPARISON WITH OTHER MFD DATASETS

To present the advantages of the proposed dataset, we used the currently best available FCOS model, I.e. [21] with selected modifications from Zhong [23]. We identified two main benefits of this model: First, the FCOS model is an object detection model without anchor boxes. The main advantage of an anchor-free object detection model is that it avoids the complicated calculations related to anchor boxes and has no anchor box hyper-parameters. Second, it uses the Generalized Focal Loss [24]. This allows the model to handle the large size differences between inline formulae and display formulae. Furthermore, these modifications have shown to be successful in competition [4]. The model is built upon Zhong's implementation [32], which uses the MMDetection toolbox [33]. Since we trained the models with one NVIDIA Tesla-V100, we used the ResNetSt-50 model and not the suggested ResNetSt-101. We trained the model with the training datapoints of the FormulaNet dataset and, for comparison, with the Tr00, Tr01, Tr10, Va00, Va01, Ts00, and Ts01 datapoints of the IBEM dataset. As we used one GPU for training, we increased the batch size from 3 to 5, decreased the learning rate from 10^{-3} to 10^{-4}, and trained it for 24 epochs. The model config files are publicly available on https://github.com/felix-schmitt/FormulaNet and the results can be reproduced by using the framework from Zhong [32].

A. EXPERIMENTS

We demonstrate the high quality of our labels and the resulting advantage for the model training with three experiments. The first experiment, which we call "Labeling Quality", investigates the quality of the labels. The second experiment is named "Dataset Comparison"; it analyses the prediction errors on existing datasets of the model trained with FormulaNet. The third experiment, "Out-of-Sample", investigates the generalization capability of models trained with FormulaNet. All results of the experiments should be interpreted with some caution, as only a randomized sample of the test PDFs was examined, and the evaluation was carried out manually.

FIGURE 3. Example image from Marmot dataset. Red shows the GT of Marmot and blue the predicted bounding box. Due to our definition of display formulae, this was counted as correct.

Contrary to our definition of display formulae, the Marmot dataset includes the reference number to the display formula bounding box as shown in Fig. 3. Through the different display formula definition, we did not count this as an error in the experiment "Labeling Quality" and we did not count it as an error if the model predicted the display formula without the reference number in the experiment "Dataset Comparison". Detailed experiment results are publicly available on https://github.com/felix-schmitt/FormulaNet/.

1) LABELING QUALITY

To investigate the labeling quality of the different datasets, we checked 100 randomly sampled pages of each dataset by hand. We counted the correct labels (CL), wrong labels (WL), wrong dimensions (WD), and missed labels (ML). CL BBOXs cover all pixels from the desired formula and no pixels from non-formula elements, while WD BBOXs contain pixels from non-formula elements or cover only parts of the desired formula. WL BBOXs cover no pixels from the corresponding formula or overlap with another BBOX. MLs are formulae that failed to be labeled as such. To make the results comparable, we put them in relation to the correct number of ground truth (CGT) labels, which is the sum of CL, WD, and ML. The pages without any labeling error (PWE) are the percentage of pages without any WL, WD, and ML of inline or display labels. This corresponds to the approximate amount of work required to clean up all errors manually. The results are shown in Table 3. The results for inline labels show that IBEM and FormulaNet have 7 times fewer labeling errors than Marmot, and furthermore, FormulaNet has 41% fewer labeling errors than IBEM. Marmot has the lowest ratio of WL, but the highest ratio of ML. The analysis of the errors revealed that the inline labels of Marmot are very accurate, but are missing many inline formulae compared to the other two datasets. Compared to IBEM, FormulaNet decreases the ratios of all three error types (WL, WD, ML) by 30-80%. One reason is FormularNet's consistent definition of inline formulae, in comparison with IBEM's inconsistent labeling of formulae in figures as inline formulae, as shown in Fig. 4.

The results for display formulae shows that the labeling errors of FormulaNet are 5–8 times less frequent than those of IBEM and Marmot. The lower labeling quality of IBEM and Marmot is primarily caused by not properly splitting and merging the display formulae as shown in Fig. 5.

Additionally, the PWE of FormulaNet shows that fewer than 16% of the pages have any labeling error, which is 3 and

TABLE 3. Results of "Labeling Quality" with the three datasets IBEM, Marmot, and FormulaNet. The table shows the ratios of correct labeled labels (CL) over the correct number of GT labels (CGT), wrong labels (WL) over CGT, wrong dimension of the BBOX (WD) over CGT, and missed labels (ML) over CGT for the two label types Inline and Display. Further, it shows the percentage of pages without a labeling error (PWE).

Label	Inline Formulae				Display Formulae				Pages
Dataset	CL/CGT (CL)	WL/CGT (WL)	WD/CGT (WD)	ML/CGT (ML)	CL/CGT (CL)	WL/CGT (WL)	WD/CGT (WD)	ML/CGT (ML)	PWE (PWE)
IBEM	96.72% (1533)	2.08% (33)	1.01% (16)	2.27% (36)	83.38% (316)	2.9% (11)	7.92% (30)	8.71% (33)	59% (41)
Marmot	76.9% (1808)	**0.38%** (9)	2.81% (66)	20.29% (477)	88.72% (354)	12.28% (49)	9.27% (37)	2.01% (8)	11% (89)
FormulaNet	**98.08%** (1529)	0.45% (7)	**0.38%** (6)	**1.54%** (24)	**97.86%** (365)	**0.27%** (1)	**1.61%** (6)	**0.54%** (2)	**84%** (16)

TABLE 4. Results of "Dataset Comparison" experiment with the datasets IBEM Ts10, IBEM Ts11, Marmot, and FormulaNet (test). The table shows the recall, precision for an IoU threshold of 0.5 and an NMS value of 0.4. The non-predicted GT BBOXs (NPs) and the wrongly predicted BBOXs (WPs) are manually checked if an NP should be not a GT (NGT) and if a WP should be a GT (SGT).

Label	Inline Formulae						Display Formulae					
Dataset	Recall	Precision	SGT/WP (WP)	SGT/CGT (SGT)	NGT/NP (NP)	NGT/GT (NGT)	Recall	Precision	SGT/WP (WP)	SGT/CGT (SGT)	NGT/NP (NP)	NGT/GT (NGT)
IBEM Ts10	94.73%	94.52%	36% (50)	1.98% (18)	39.58% (48)	2.09% (19)	94.64%	92.98%	58.33% (12)	4.14% (7)	66.67% (9)	**3.57%** (6)
IBEM Ts11	**95.16%**	**97.8%**	40% (15)	**0.87%** (6)	61.76% (34)	2.99% (21)	89.23%	82.08%	94.74% (38)	16.98% (36)	90.48% (21)	9.74% (19)
Marmot	82.93%	66.13%	84.63% (423)	27.39% (358)	27.65% (170)	4.72% (47)	75.47%	94.49%	28.57% (7)	1.46% (2)	61.54% (39)	15.09% (24)
FormulaNet (test)	94.91%	94.38%	**35.29%** (51)	1.98% (18)	**23.91%** (46)	**1.22%** (11)	**98.96%**	**95.5%**	**0%** (9)	**0%** (0)	**0%** (2)	**0%** (0)

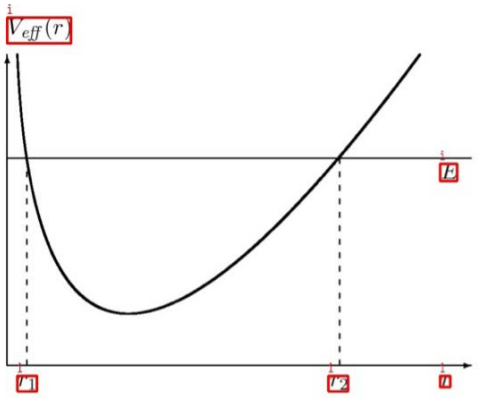


FIG. 1: The shape of our potential $V_{eff}(r)$ with $\eta > 0$.

FIGURE 4. Example page from IBEM Ts11. Red shows the GT inline labels, that are not inline labels with our inline definition.

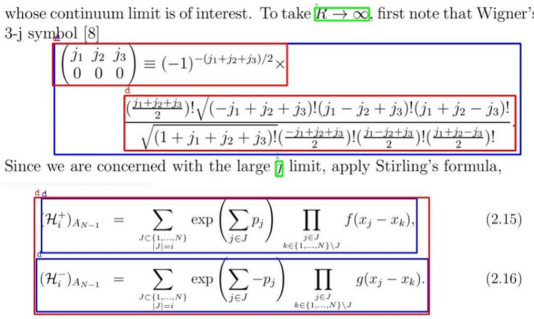


FIGURE 5. Examples from IBEM Ts11 of split and merge errors. Red shows the GT of IBEM and blue shows possible BBOX with our display definition.

6 times less than IEBM and Marmot, respectively. This also clearly indicates the better labeling quality of FormulaNet compared to IBEM and Marmot.

2) DATASET COMPARISON

The "Dataset Comparison" experiment investigates whether a model benefits from the high labeling quality of the FormulaNet dataset, and whether a model trained with FormulaNet can detect errors in existing datasets.

For the experiment, the model was trained with the FormulaNet dataset. We used the trained model to test the predictions on the IBEM Ts10 and IBEM Ts11 and Marmot datasets, and randomly selected 50 pages from each dataset. We used an Intersection of Union (IoU) threshold of 0.5 and an Non-maximum Suppression (NMS) value of 0.4 for the evaluation. Any non-predicted GT BBOXs (NPs) (with IoU smaller than 0.5 or no overlap) were manually checked to determine whether they are a correct GT or should not be a GT (NGT). Moreover, any incorrectly predicted BBOXs (WP) are manually checked for whether they should be a GT (SGT). For comparison, we have added the FormulaNet test set results. The results are shown in Table 4.

The high recall and precision values of the two IBEM test datasets indicate a similar labeling strategy of IBEM and FormulaNet. The model trained on the FormulaNet training set reached a combined F1 score (inline formulae and display formulae) of 94.49% for the 50 pages of IBEM Ts10, 93.97% for IBEM Ts11, and 94.26% for IBEM Ts10 + IBEM Ts11. Since the challenge [4] used an IoU threshold of 0.7, the

TABLE 5. Results of the "Out-of-Sample" experiment with 50 random pages of 1000 arXiv 2021 papers. The table shows the resulting recall, precision, WL over CGT, and WD over CGT for the two label types Inline Formulae and Display Formulae.

Label	Inline Formulae				Display Formulae			
Training Dataset	Recall	Precision	WL/CGT (WL)	WD/CGT (WD)	Recall	Precision	WL/CGT (WL)	WD/CGT (WD)
IBEM	68.5%	68.32%	14.31% (164)	17.63% (202)	73.27%	77.08%	1.98% (2)	19.8% (20)
FormulaNet	**76.53%**	**84.73%**	**5.32%** (61)	**8.29%** (95)	**82.18%**	**84.69%**	**0.99%** (1)	**13.86%** (14)

TABLE 6. Results of the two baseline models (FCOS-50 and FCOS-101). The COCO metric is used for the evaluation.

Model	mAP Inline	mAP Display	mAP	mAP@50	mAP@75
FCOS-50 [$\mu \pm \sigma$]	0.752±0.02	**0.755±0.02**	0.754±0.03	**0.921±0.02**	0.84±0.02
FCOS-101 [$\mu \pm \sigma$]	**0.756±0.02**	0.749±0.03	**0.755±0.03**	0.920±0.02	**0.841±0.02**

values are not fully comparable. With an IoU threshold of 0.7 and all pages of Ts10 and Ts11, the model reaches an F1 score of 84.58%, which is only 2% lower than the results in the challenge [4] without using the training data.

The lower precision and recall values on IBEM Ts11 for display formulae are a result of the small number of pages, along with an excessive number of split and merge errors of display formulae (shown in Fig. 5). Additionally, the high SGT and NGT ratios indicate that many of these errors are errors in the ground truth of IBEM Ts11. These results verify that the model trained with FormulaNet can detect labeling errors in the IBEM dataset.

The recall and precision values for our model tested with the Marmot test dataset are lower compared to the results on the two IBEM datasets. The corresponding accuracy of 88.02% for inline formulae and 76.51% for display formulae (86.81% combined) is slightly lower than the best models trained on Marmot [34]. However, the low NGT ratio and high SGT ratio for inline formulae of the Marmot dataset show that the Marmot inline labels are accurate, but not all inline formulae are in the GT, as the "Labeling Quality" experiment showed as well. The high NGT ratio of display formulas is primarily due to split and merge errors.

The precision and recall values with the FormulaNet test set show that the model accurately predicts inline and display formulae. The four display formulae indicators (SGT/WP, SGT/CGT, NGT/NP, and NGT/GT) are rather low with 0. We explain these zero values due to the small page set of 50 pages and hence few display formulae. However, the zero values indicate that the are only few labeling errors in the dataset and the model has learned very accurately to predict display formulae.

3) OUT-OF-SAMPLE

For the "Out-of-Sample" experiment, we randomly selected 50 pages from over 1000 arXiv papers from all fields from 2021. We trained our model once with the IBEM dataset and once with the FormulaNet dataset. The trained models predicted the labels of the 50 pages. Since there are no annotations for these pages, we manually checked each BBOX to see if it was correct, incorrect, and if BBOXs were missing from the page. The definitions of CL, WD, and WL are the same as for the experiment "Labeling Quality". The recall is calculated as the ratio of CL over CGT and the precision as the ratio of CL over the sum of CL, WL, and WD. The results are shown in Table 5.

Even on papers from other fields, the model makes better prediction if it is trained with the FormulaNet dataset compared to when it is trained on the IBEM dataset. The model trained with FormulaNet reaches an 11.72% higher recall and a 24.02% better precision for inline labels, and a 12.16% higher recall and a 9.87% better precision for display formulae.

As expected, the performance of both models is substantially lower compared to the performance in the "Dataset Comparison" experiment with the IBEM dataset. There are two reasons for the lower performance. First, we used our CL definition and not an IoU of 0.5 because of the manual evaluation of the results. Second, the papers in this test are not from the same research field as the papers during training (IBEM uses papers from the same research field as FormulaNet).

B. BASELINE RESULTS ON FormulaNet DATASET

For a baseline performance on FormulaNet, we present here the results of two of the models trained with the FormulaNet dataset. The smaller model (FCOS-50) uses the ResNetSt-50 as backbone, as used for the experiments, and the larger model (FCOS-101) is based on the ResNetSt-101 backbone. The evaluation was conducted on the FormulaNet test set with the COCO metric [35]. The models are trained on the training set of the FormulaNet dataset and evaluated on the test set of the FormulaNet dataset after 24 epochs. Table 6 presents the results of 5 runs of the two baseline models. The results show that the larger backbone ResNetSt-101 does not significantly improve the model performance and the dataset is challenging for MFD models. The baseline model configs are publicly available on https://github.com/felix-schmitt/FormulaNet and can be reproduced using the framework of [32].

V. CONCLUSION

In this paper, we presented the FormulaNet dataset, a new dataset to train and benchmark MFD. FormulaNet is the largest dataset comprising labeled display and inline formulae and achieves an unprecedented labeling quality for this problem. FormulaNet was created by an automated labeling pipeline which will make it possible to create large high-quality datasets for future MFD research and benchmarking. Due to our automated labeling process and our proposed definition of inline and display formulae, the labels are very consistent compared with existing datasets. In addition to the FormulaNet dataset, we provide a strong baseline with one of the current best MFD models.

Through the design of the labeling pipeline, the dataset is limited to LaTeXpapers. Furthermore, FormulaNet is based only on High Energy Physics papers from arXiv.org. However, the "Out-of-Sample" experiment showed that the dataset still generalizes well to out-of-sample datapoints.

Given the promising results of our experiments, we are optimistic that FormulaNet can serve as a new Benchmark dataset for MFD to help to advance research in this area, which may finally result in new applications with high impact regarding accessible scientific PDFs.

REFERENCES

[1] (2008). *Convention on the Rights of Persons with Disabilities—Articles | United Nations Enable*. Accessed: May 24, 2022. [Online]. Available: https://www.un.org/development/desa/disabilities/convention-on-therights-of-persons-with-disabilities/convention-on-the-rights-of-personswith-disabilities-2.html

[2] (2019). *Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services (Text with EEA relevance)*. vol. 151. Accessed: May 24, 2022. [Online]. Available: http://data.europa.eu/eli/dir/2019/882/oj/eng

[3] A. J. Rajkumar, J. Lazar, J. B. Jordan, A. Darvishy, and H.-P. Hutter, "PDF accessibility of research papers: What tools are needed for assessment and remediation?" in *Proc. Hawaii Int. Conf. Syst. Sci.*, Jan. 2020, pp. 4185–4194, doi: 10.24251/HICSS.2020.512.

[4] D. Anitei, J. A. Sánchez, J. M. Fuentes, R. Paredes, and J. M. Benedí, "ICDAR 2021 competition on mathematical formula detection," in *Document Analysis and Recognition—ICDAR*. Cham, Switzerland: Springe, 2021, pp. 783–795, doi: 10.1007/978-3-030-86337-1_52.

[5] T. Stadelmann, M. Amirian, I. Arabaci, M. Arnold, G. F. Duivesteijn, I. Elezi, M. Geiger, S. Lörwald, B. B. Meier, K. Rombach, and L. Tuggener, "Deep learning in the wild," in *Artificial Neural Networks in Pattern Recognition*, Cham, Switzerland: Springer, 2018, pp. 17–38, doi: 10.1007/978-3-319-99978-4_2.

[6] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "ICDAR2017 competition on recognition of documents with complex layouts—RDCL2017," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Nov. 2017, pp. 1404–1410, doi: 10.1109/ICDAR.2017.229.

[7] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "ICDAR2019 competition on recognition of documents with complex layouts–RDCL2019," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 1521–1526, doi: 10.1109/ICDAR.2019.00245.

[8] R. Zanibbi and D. Blostein, "Recognition and retrieval of mathematical expressions," *Int. J. Document Anal. Recognit.*, vol. 15, no. 4, pp. 331–357, Dec. 2012, doi: 10.1007/s10032-011-0174-4.

[9] *Marmot Dataset*. Accessed: Feb. 23, 2022. [Online]. Available: https://www.icst.pku.edu.cn/cpdp/sjzy/

[10] DFKI Cloud. *ICDAR-2017 POD(Corrected)*. Accessed: Feb. 23, 2022. [Online]. Available: https://cloud.dfki.de/owncloud/index.php/s/jrK3f9KEFSkwmgJ

[11] D. Anitei, J. A. Sánchez, and J. M. Benedí. *IBEM Mathematical Formula Detection Dataset*. Accessed: May 13, 2021, doi: 10.5281/zenodo.4757865.

[12] D. Tkaczyk, P. Szostek, and L. Bolikowski. (Sep. 29, 2015). *GROTOAP2*. RepOD. [Online]. Available: http://dx.doi.org/10.18150/8527338

[13] X. Zhong, J. Tang, and A. Jimeno Yepes, "PubLayNet: Largest dataset ever for document layout analysis," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 1015–1022, doi: 10.1109/ICDAR.2019.00166.

[14] M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, and Z. Li, "TableBank: Table benchmark for image-based table detection and recognition," in *Proc. 12th Lang. Resour. Eval. Conf.*, May 2022, pp. 1918–1925.

[15] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, and M. Zhou, "DocBank: A benchmark dataset for document layout analysis," in *Proc. 28th Int. Conf. Comput. Linguistics*, Dec. 2020, pp. 949–960, doi: 10.18653/v1/2020.coling-main.82.

[16] D. Knuth. (2021). *TeX Live*. LATEXVersion TeX Live. Accessed: May 31, 2022. [Online]. Available: https://tug.org/texlive/

[17] *Arxiv.org e-Print Archive*. Accessed: Feb. 23, 2022. [Online]. Available: https://arxiv.org/

[18] R. H. Anderson, "Syntax-directed recognition of hand-printed two-dimensional mathematics," in *Proc. Symp. Interact. Syst. Experim. Appl. Math. Proc. Assoc. Comput. Machinery Inc. Symp.*, New York, NY, USA, Aug. 1967, pp. 436–459, doi: 10.1145/2402536.2402585.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[20] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.

[21] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635, doi: 10.1109/ICCV.2019.00972.

[22] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[23] Y. Zhong, X. Qi, S. Li, D. Gu, Y. Chen, P. Ning, and R. Xiao, "1st place solution for ICDAR 2021 competition on mathematical formula detection," 2021, *arXiv:2107.05534*.

[24] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. 34th Int. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA, Dec. 2020, pp. 21002–21012.

[25] L. Tuggener, I. Elezi, J. Schmidhuber, M. Pelillo, and T. Stadelmann, "DeepScores—A dataset for segmentation, detection and classification of tiny objects," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3704–3709, doi: 10.1109/ICPR.2018.8545307.

[26] *Regex—Regular Expression Operations*. Regex Version 2022.1.18. Accessed: May 3, 2022. [Online]. Available: https://docs.python.org/3/library/re.html

[27] The LATEXProject. Xcolor. Xcolor Version 2.13. Accessed: May 3, 2022. [Online]. Available: https://ctan.org/pkg/xcolor

[28] *Analyzed Layout and Text Object (ALTO) XML Schema*. ALTO XML Version 4.2. Accessed: May 3, 2022. [Online]. Available: https://www.loc.gov/standards/alto/

[29] P. Lopez. Pdfalto. Pdfalto Version 0.5. Accessed: May 3, 2022. [Online]. Available: https://github.com/kermitt2/pdfalto

[30] E. Belvale. Pdf2image. Pdf2image Version 1.16.0. Accessed: May 3, 2022. [Online]. Available: https://pypi.org/project/pdf2image/

[31] *OpenCV*. Py-OpenCV Version 4.5.5. Accessed: May 3, 2022. [Online]. Available: https://opencv.org/

[32] Zhong. *1st Solution for ICDAR 2021 Competition on Mathematical Formula Detection*. Accessed: May 23, 2022. [Online]. Available: https://github.com/Yuxiang1995/ICDAR2021_MFD

[33] K. Chen *et al.*, "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.

[34] M. Z. Afzal, K. A. Hashmi, A. Pagani, M. Liwicki, and D. Stricker. (Jan. 2022). *DeHyFoNet: Deformable Hybrid Network for Formula Detection in Scanned Document Images*. [Online]. Available: http://dx.doi.org/10.20944/preprints202201.0090.v1

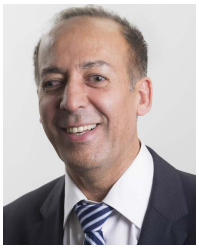
[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

FELIX M. SCHMITT-KOOPMANN received the B.Sc. degree in mechanical engineering and the M.Sc. degree in robotics, systems, and control from ETH Zürich, in 2019 and 2021, respectively. He is currently pursuing the Ph.D. degree with the People and Computing Laboratory, University of Zurich. He is a member of the Institute of Applied Informatics, ZHAW. His research interests include accessibility, AI, and document analysis.

ELAINE M. HUANG received the Ph.D. degree from the College of Computing, Georgia Institute of Technology, in 2006. She is currently a Professor in human–computer interaction with the Department of Informatics, University of Zurich (UZH), where she leads the People and Computing Research Group. Prior to joining UZH, in 2010, she was a Researcher at Motorola Laboratories and a Professor with the Department of Computer Science, University of Calgary. Her research interests include the use of technology to address issues of inequality and other societal challenges.

HANS-PETER HUTTER (Member, IEEE) received the Doctor of Technical Science degree in electrical engineering from ETH Zürich, in 1997. In 1997, he worked on hybrid HMM/ANN approaches to speech recognition over telephone lines. He joined the UBS Ubilaboratory as a Postdoctoral Researcher, where he worked on a European project for HMM-based speaker identification over the telephone. At the same time, he was a Co-Lecturer at ETHZ in two speech processing modules. In 1997, he joined the ZHAW Zurich University of Applied Sciences, Winterthur, where he worked as a Professor in computer science on various projects in the area of speech recognition and user centered design of graphical and voice user interfaces. In 2005, he founded the ZHAW School of Engineering, Institute of Applied Information Technology (InIT), together with his colleagues and was the Head of the Institute, until 2010. At the same time, he was also the Head of the Human-Information Interaction Group, InIT, which he is still leading today.

THILO STADELMANN (Senior Member, IEEE) received the Doctor of Science degree from Marburg University, Germany, in 2010, for his work on multimedia analysis and voice recognition. He worked in engineering and leadership roles in the automotive industry. He is currently a Professor of AI/ML with the ZHAW School of Engineering, Winterthur, Switzerland, the Director of the ZHAW Centre for Artificial Intelligence, and the Head of the Computer Vision, Perception and Cognition Group. He is a fellow of the European Centre for Living Technology, Venice, Italy.



ALIREZA DARVISHY is currently a Professor in ICT Accessibility and the Head of the ICT Accessibility Laboratory, Zurich University of Applied Sciences, Switzerland. He serves an Independent Reviewer for European research projects, such as the Active Assisted Living (AAL) program, and he is a Principle Investigator of the ''Accessible Scientific PDFs for All'' Project, funded by the Swiss National Science Foundation.