where X and Y are the binary image of the network output and that of the ground truth. $X \cap Y$ denotes the overlap between X and Y, and $|X|$ is the L_1-norm of image X.

The proposed method does not use any information from either mathematical grammar or the character recognition results. The image conversion module in the proposed method is requested to obtain information that is crucial to determine the components that should remain as MEs only from the appearance of documents in the surrounding image area. A limited size of small regions may cause difficulty regarding making a decision even for humans in this scenario.

C. POSTPROCESSING

Through the image conversion process, we obtain sub-block images that contain CCs that correspond to MEs. In the postprocessing stage, we reconstruct the page image and extract CCs that correspond to mathematical symbols and characters.

To reconstruct the entire page image, each sub-block image is rearranged in the equivalent position and the maximum pixel value among the overlapping pixels is assigned to the corresponding pixel in the page image.

Additionally, pixel-wise multiplication between the resized reconstructed image and the original image is performed to eliminate dilated pixels caused by the morphology operation in preprocessing, and artifacts and noise injected during the image conversion process.

IV. PERFORMANCE EVALUATION

A. DATASETS

For a quantitative evaluation of the performance of mathematical OCR, a number of datasets have been proposed in the literature. InftyCDB datasets [39], [40] are large collections of mathematical symbols and notation from actual mathematical documents. UW databases [41] contain 100 typeset MEs from 25 document pages. However, these datasets are not applicable for evaluating ME detection performance because the content in the dataset is rearranged not to keep the original articles because of copyright reasons.

We collected two large datasets to train U-Net and evaluate the performance of the proposed ME detection method. The datasets, called GTDB-1 and GTDB-2, consist of document page images collected from scientific journals and textbooks. The GTDB-1 dataset, which was used to train the U-Net model, contains 31 English articles on mathematics. The GTDB-2 dataset, which was used for quantitative and qualitative evaluations of the performance of the proposed method, contains 16 articles. Diverse font faces and mathematical notation styles are included in these articles. A list of the articles in both datasets is provided in the appendix.

The statistics of each dataset are shown in Table 1. The article pages were originally scanned at 600 dpi. The ground truth annotations for each math symbol and ordinary character were attached manually. [2]

TABLE 1. Statistics of the datasets: Two datasets collected from scientific journals and textbooks.

	GTDB-1	GTDB-2
# articles	31	16
# pages	544	343
# math symbols	162,406	115,433
# of ordinary text characters	646,714	507,412

B. EVALUATION EXPERIMENTS

To train the U-Net model, we extracted 1,000 pairs of sub-blocks from each document page and the corresponding ground truth image from the GTDB-1 dataset. The locations of sub-blocks on each page image were determined randomly. The dataset consisted of 544 images, so the total number of sub-block images for training was 544,000.

We mainly used mathematical symbol (character) recall R_s, precision P_s and F-measure F_s as the performance measures. Each measure is defined as follows:

$$R_s = \frac{n_{TP}}{n_{TP} + n_{FN}}, \tag{2}$$

$$P_s = \frac{n_{TP}}{n_{TP} + n_{FP}}, \tag{3}$$

$$F_s = \frac{2\,P_s R_s}{P_s + R_s}, \tag{4}$$

where n_{TP}, n_{FP} and n_{FN} are the numbers of correctly detected mathematical symbols, falsely detected symbols or ordinary text, and undetected mathematical symbols, respectively. Pixel-level majority voting is used for the symbol-level evaluation. If the majority of pixels in a candidate symbol were detected as a mathematical symbol, the candidate symbol is determined as a mathematical symbol.

We also used ME-based recall (R_e), precision (P_e) and F-measure (F_e) as supplemental performance measures. Their definitions are similar to (2)–(4), but the numbers in the equations are counted for regions.

To determine MEs over the detected mathematical symbols, mathematical layout analysis is requested to obtain the spatial relationship between the symbols. We do not intend to implement layout analysis in the present study. Therefore, note that the evaluations using ME-based measures are based on the assumption that the candidates of mathematical regions are obtained using some layout analysis method. In this study, we extracted candidate regions using the ground truth so that the candidate regions contain in-line and displayed MEs, and

[2] The ground truth annotation has been released to the public to benchmark OCR performance for scientific documents. Although we could not include the original document images of articles for copyright reasons, we provide hyperlinks on our website to the web pages of the original documents, where the readers can obtain the document images: `https://github.com/uchidalab/GTDB-Dataset/tree/master`