Computer Modeling in
Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2023.023229

ARTICLE

An Auto-Grading Oriented Approach for Off-Line Handwritten Organic Cyclic Compound Structure Formulas Recognition

Ting Zhang, Yifei Wang, Xinxin Jin, Zhiwen Gu, Xiaoliang Zhang and Bin He*

Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan, 430079, China

*Corresponding Author: Bin He. Email: hebin@ccnu.edu.cn

Received: 15 April 2022 Accepted: 22 August 2022

ABSTRACT

Auto-grading, as an instruction tool, could reduce teachers' workload, provide students with instant feedback and support highly personalized learning. Therefore, this topic attracts considerable attentions from researchers recently. To realize the automatic grading of handwritten chemistry assignments, the problem of chemical notations recognition should be solved first. The recent handwritten chemical notations recognition solutions belonging to the end-to-end trainable category suffered from the problem of lacking the accurate alignment information between the input and output. They serve the aim of reading notations into electrical devices to better prepare relevant e-documents instead of auto-grading handwritten assignments. To tackle this limitation to enable the auto-grading of handwritten chemistry assignments at a fine-grained level. In this work, we propose a component-detection-based approach for recognizing off-line handwritten Organic Cyclic Compound Structure Formulas (OCCSFs). Specifically, we define different components of OCCSFs as objects (including graphical objects and text objects), and adopt the deep learning detector to detect them. Then, regarding the detected text objects, we introduce an improved attention-based encoder-decoder model for text recognition. Finally, with these detection results and the geometric relationships of detected objects, this article designs a holistic algorithm for interpreting the spatial structure of handwritten OCCSFs. The proposed method is evaluated on a self-collected data set consisting of 3000 samples and achieves promising results.

KEYWORDS

Handwritten chemical structure formulas; structure interpretation; components detection; text recognition

1 Introduction

Preparation and grading of tests are the key activities in instruction, which could reflect the students' cognitive level and provide sources for teachers to improve their teaching. To realize auto-grading is meaningful as it has multiple functions such as reducing the workload of teachers, providing immediate feed-backs to students and supporting highly personalized learning. It has been widely used in subjects of English and Computer to grade English composition [1] and computer program [2,3]. This paper will be focused on automatic grading of handwritten chemistry assignments. Obviously, handwritten chemical notations recognition is the preliminary technique.

Handwriting recognition could be divided into online and offline two cases. In the online case, the input is a sequence of stokes while the input is an image for the offline. The published works regarding handwritten chemical notations recognition were mainly focused on the online case as more information (time) is available compared to the offline case. However, the offline case could support more application scenarios, such as auto-grading of chemistry paper tests. In daily study of chemistry, examinations for a long time to come will still be based on paper tests. Thus, in this work, we expect to solve the problem of offline handwritten OCCSFs recognition. This task is very challenging from two respects: (1) Large intra-class variance. Handwritten cases contain deformations in size, shape, and other variations. (2) Complex 2-dimensional structures. OCCSFs usually contain one or more rings accompanied by multiple text chains. These features make the task become a tricky problem.

With the development of deep learning techniques, many landmark achievements were released in the offline handwriting recognition field. Dominant methods belong to the end-to-end trainable class including ones based on connectionist temporal classification (CTC) [4] and others based on attention mechanism [5]. These methods achieved significant success on 1D text recognition [6] or 2D handwritten math expressions recognition tasks [7,8] owing to the strong capabilities of learning robust feature representation and accessing the global contextual information. They usually take as input an image of handwritten texts and output the sequence of labels directly. However, the accurate alignment between the output sequence of labels and the input image is difficult to determine with the existing approaches, which is important for auto-grading to diagnose errors or generate fine-grained feed-backs.

Currently, the published solutions to handwritten chemical notations recognition serve only the aim of reading notations into electrical devices to better prepare relevant e-documents instead of auto-grading students' assignments. The recent handwritten chemical notations recognition solutions [9] belonging to the end-to-end trainable category suffered from the problem of lacking the accurate alignment information between the input and the output. However, this accurate alignment information is required in grading assignments to diagnose errors and generate feed-backs at a fine-grained level. To tackle this limitation, we propose an auto-grading oriented approach for off-line handwritten OCCSFs Recognition. OCCSFs, as a typical two dimensional graphics language, have a complex spatial structure. Hand-drawn OCCSFs recognition is an appealing task as it exhibits big challenges for the complex spatial structure and variable writing style. Fig. 1 presents some examples of offline handwritten OCCSFs. In this work, we focus on OCCSFs with one or two ring structures. These types cover almost all the OCCSFs appearing in K12 education. To obtain the accurate alignment between the input and the output, we propose a components-detection-based approach for offline handwritten OCCSFs recognition.

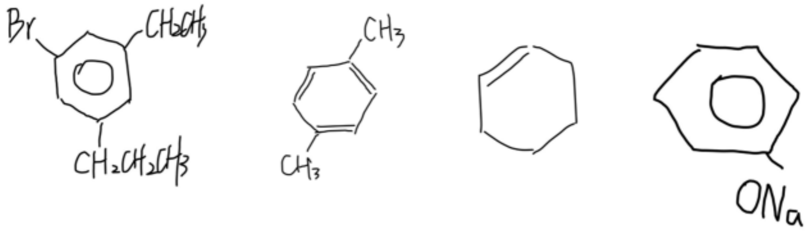


Figure 1: Samples of offline handwritten OCCSFs

Deep convolutional neural networks (DCNNs) [10] demonstrated excellent performance on image classification tasks. It also boosted the developments in the object detection field, bringing on dramatic improvements on accuracy. Currently, CNN-based object detection algorithms generally could be divided into two categories, being two-stage detectors and one-stage detectors. The two-stage methods such as Faster-RCNN [11] firstly generate region proposals and then address detection as a classification problem over region proposals. The one-stage methods like SSD [12,13], and YOLO [14] skip the region proposal generation step and predict bounding boxes and confidences for multiple categories directly. These one-stage methods have comparable performance with two-stage methods and yet are faster [12].

OCCSFs contain both graphical components and text components. The idea proposed in this work for handwritten OCCSFs recognition is to regard these components as different objects and use object detection algorithms to detect them. Then with detection results and spatial relationships between detected components, we analyse the structure. As the problem involves both graphical objects and text objects detection, the common detectors SSD and YOLO are adopted both to detect the predefined objects in OCCSFs. Next, the detected text components need to be recognized further. As reported in [15], attention-based methods can achieve higher recognition accuracy than CTC-based methods on isolated word recognition tasks, but perform worse on sentence recognition tasks. Apparently, text components in chemical OCCSFs appearing in K12 are similar to words instead of sentences. Thus we use an improved attention-based model for text components recognition by mitigating the existing error accumulation problem. Finally, an algorithm is proposed for interpreting the formula structure, which takes the detection results as input and outputs the interpretation results. This work is an extended version of the paper [16] published in 2021 International Conference on Engineering, Technology, and Education. Compared with [16], the work is extended from several aspects. Firstly, component detection is improved by introducing YOLOv5. Secondly, text components recognition is performed via integrating scheduled sampling into Decoupled Attention Network. Thirdly, OCCSFs with multiple rings are considered in this paper.

The main contributions of this work are as follows:

- We propose an auto-grading oriented approach for off-line handwritten OCCSFs recognition, which could output the final recognition results, as well as the accurate alignment between the input and output. This accurate alignment information is indispensable for auto-grading to diagnose errors or generate fine-grained feed-backs. The approach addresses the problem by defining the different types of components as objects, then adopting object detection algorithms to detect different components, next recognizing the text objects if exist and finally using the detection results to interpret the structure.
- An improved attention-based model for text components recognition is proposed via mitigating the existing error accumulation problem.
- Several metrics at the object level are defined to better analyze the effects of components detection results on later structure interpretation.
- An off-line handwritten OCCSFs dataset which consists of 3000 samples is built and later will be released freely for the research aim.

This article will explain the related works in Section 2. Section 3 introduces the built dataset and Section 4 details the proposed approach. Section 5 gives the experimental details and results. Finally, Section 6 concludes this work and puts forward some ideas for future work.

2 Related Works

In this section, we review the literatures from three aspects, being chemical structure formulas recognition, object detection and text recognition which are closely related to our work.

2.1 Chemical Structure Formulas Recognition

The data to be handled could be in print format or handwritten format. Further, handwriting data could be again divided into on-line and off-line two cases. In the on-line case, the input is a sequence of stokes while the input is an image for the off-line. Even though time information is not available compared to the on-line case, the off-line case could support more application scenarios, such as the auto-grading of chemistry paper tests.

Regarding chemical structural formulas recognition, the published works mainly focused on the on-line case. In 2007, Ouyang et al. designed an on-line recognition system for hand-drawn chemical diagrams [17]. The system used a trained classifier to locate and recognize chemical symbols, and then generated the initial structure considering the spatial context. Finally, the system used chemical knowledge to check the legitimacy of the interpreted structure and modified it if necessary. To promote the work in [17], they proposed "ChemInk" [18] a real-time recognition system which used a jointly trained conditional random field to combine multiple levels of visual features. The framework accessed different levels of details to enhance the system robustness to noise and drawing variations, thus improving the performance. Sadawi et al. [19] proposed a rule-based method where they used rules to identify atoms and bonds and deal with possible ambiguities. Sun et al. proposed a free-sketch recognition method [20] for chemical structural formulas. A dual-mode method was used to distinguish character input and non-character input first. Then they adopted an attribute graph to model sketched chemical structural formula and utilized domain knowledge to rectify the relationships among elements.

The research on *off-line* chemical structure formulas recognition goes into two branches—the rule-based category [21] and the end-to-end trainable category [9]. Bukhari et al. [21] proposed a system to automatically analyse the printed 2-D chemical structures in document images using traditional image processing techniques. The proposed recognition process consisted of a series of operations (totally 9) based on open-source libraries such as OpenCV-3.3. However, with embedded algorithms like Line Segment Detector and Hough Circle, it is difficult to deal with handwritten inputs which have multi-variations. Literature [9] published an attention-mechanism-based method which translated a bitmap image of a molecule directly into a SMILES—a machine-readable chemical format. This deep-learning-based method has a stronger generalization capability compared to the rule-based one. However, the accurate alignment between the output sequence of labels and the input image is difficult to determine. Unfortunately, the alignment information is indispensable for auto-grading to diagnose students' errors and generate fine-grained feedback. Thus, to break these limitations to enable auto-grading of handwritten chemical assignments at a fine-grained level, we propose a component-detection-based approach for off-line handwritten OCCSFs recognition.

2.2 Object Detection

Before the era of CNNs, traditional object detection methods mainly consisted of three steps, being region selection, feature extraction and classification. Deformable Part Model (DPM) [22] and Selective Search [23] were two state-of-the-art methods.

In 2014, Girshick proposed the R-CNN [24] detection algorithm which is the first CNN-based object detection algorithm achieving impressive results. This method combined selective search region proposals and CNN-based post-classification together. The subsequent research [11,25,26] improved the quality of region proposal generation or post-classification stage alternatively. Post-classification is costly and time-consuming as it needs to process thousands of image regions. SPPnet [25] speeded up the R-CNN detection algorithm notably by introducing a spatial pyramid pooling layer between the convolutional layer and fully connected layer to improve feature extraction. However, the training of SPPnet was still a multi-stage pipeline, far away from being end-to-end. Fast R-CNN [26] strengthened SPPnet again by proposing a single-stage training algorithm that jointly learned to classify object proposals and refine their spatial locations. Faster R-CNN [11] improved the quality of proposal generation by replacing selective search method with a region proposal network (RPN). Furthermore, it proposed an alternative training method to integrate RPN with Fast R-CNN. In general, these methods consist of region proposal generation and proposal classification and are named as two-stage detectors.

Different from the above-mentioned two-stage detectors, YOLO [14] and SSD [12] discarded region proposal generation step and predicted bounding boxes and confidences for multiple categories directly, therefore possessing high speed. They were named as one-stage detectors. The common understanding is that YOLO performs better on smaller objects and SSD performs better on larger objects. Overall, one-stage detectors have comparable performance with two-stage methods and yet possess high speed. Thus we propose a method for component detection of handwritten OCCSFs based on the one-stage detectors. Both YOLO and SSD will be tested in our task to compare the performances.

2.3 Text Recognition

The proposed methods for text recognition can be roughly divided into 2 classes, being segmentation-based and segmentation-free. The segmentation-based methods [27–29] usually involve segmenting characters, recognizing characters and combining the recognition results into the final outputs. However, accurate character segmentation is very difficult, especially for the handwriting input.

The segmentation-free methods encode the text input as a whole and decode the encoded features directly into the sequence of labels in an end-to-end trainable manner. With the strong ability of visiting global context information, this type of method achieved promising performance on a series of text recognition tasks. Two representative methods were CTC-based and attention-based. In [6], a novel neural network architecture was proposed, which integrated CNN (for feature extraction), RNN (for sequence modelling) and CTC (for transcription) into a unified framework. The attention mechanism was originally proposed in neural machine translation [5] and later was introduced into text recognition [30,31]. Different neural network models were proposed to work as the encoder to encode an input text image into a one-dimensional feature sequence [30,31] or a two-dimensional feature map [7]. The latter one retained the vertical spatial information. The attention model learned to focus on a specific region of the feature sequence or feature map at each time step. The conventional soft-attention mechanism proposed in [5] was developed in later works [7,32] to achieve better alignments. The decoder outputted the sequence of labels in an auto-regressive way. For the decoder module, the recurrent neural network was the most widely used model.

As reported in [15], attention-based methods can achieve higher recognition accuracy than CTC-based methods on isolated word recognition tasks, but perform worse on sentence recognition tasks. The text components in chemical organic structure formulas appearing in K12 could be regarded as words instead of sentences, which do not contain vertical structures. Thus, in this paper, we adopt the attention-based encoder-decoder model for text components recognition.

The typical attention-based method generates the target sequence of tokens one by one. It predicts the current token depending on both the previous token and the feature currently focused. Furthermore, the computation of the feature currently focused also depends on the previous token. That means once there is a mistake happening in the historical decoding, the error could be propagated quickly along the sequence via these dependencies. To tackle the limitations, Decoupled Attention Network (DAN) was proposed in [33] which decoupled the dependency of the computation of the feature currently focused on the previous token. It used a convolutional alignment module that computed the focused weights of each time step based on visual features from the encoder only. However, DAN solved this error accumulation problem of the attention-based model partially not totally since the prediction of the current token still related to the previous token. At the training stage, the ground-truth previous tokens are available but not available at inference. This discrepancy between training and inference could lead to errors that propagate quickly along the sequence. Scheduled sampling [34] was proposed by Google to mitigate this discrepancy. In this work, we combine DAN and scheduled sampling to improve the attention-based model and then apply it for text components recognition.

3 The Off-Line Handwritten OCCSFs Dataset

In this work, we focus on OCCSFs with one or two-ring structures. These types cover almost all the OCCSFs appearing in K12 education. To our knowledge, there is no public offline handwritten OCCSFs data set available yet. Thus a data set should be created first. We collected common handwritten OCCSF images such as aromatic hydrocarbon, halogenated hydrocarbon, cyclohexane, cyclohexene, xylene, trinitrotoluene and other derivatives. As this work aims for auto-grading eventually, some samples collected may not follow the chemical grammar rules. Totally, a data set was built consisting of 2000 one-ring structure images and 1000 two-ring structure images. For the data annotation, we labelled the collected samples by 4 predefined objects (*benzene, ring, doublebond, textchain*. Details can be found in Section 4.1) and save the annotation information as VOC2007 format which is supported by both SSD and YOLO. Table 1 provides the statistics of the built data set. According to these statistics, it is known that each sample contains 3.485 objects in average. Fig. 2 illustrates some collected samples (two-ring structures only).

Table 1: The statistics of the off-line handwritten OCCSFs dataset

Items	No.
Objects	10455
benzene	3105
ring	897
doublebond	1689
textchain	4764

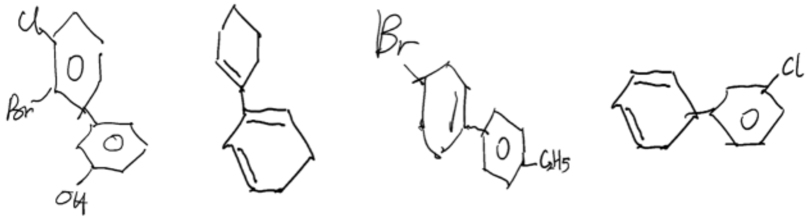


Figure 2: Samples of two-ring structure

4 The Proposed Method

In this paper, we consider the idea of using deep learning object detector to locate and classify the components in hand-drawn OCCSFs and interpreting the structure with the detection results. Specifically, a components-detection-based method (as shown in Fig. 3) was proposed which includes mainly 3 steps, being components detection, text components recognition and structure interpretation respectively. In the coming paragraphs, we will introduce these 3 steps in detail.

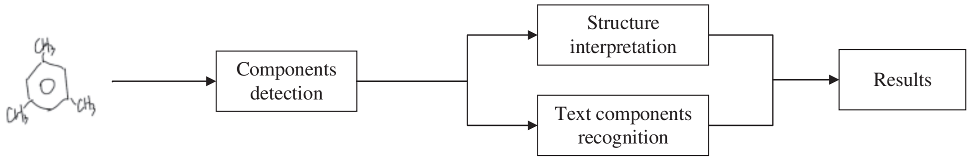


Figure 3: Illustration for the framework of the proposed method

4.1 *Components Detection*

We define 4 types of components as objects, namely *benzene*, *ring*, *doublebond* (*bond*2), *textchain* shown in Fig. 4, with which the structure could be interpreted unambiguously. These 4 types of components, not only have distinctive visual features considering the outlook but correspond to the minimum knowledge units in chemistry domain. The former point could ease the burden of object detector; the latter one could support auto-grading and personalized feedback generation at fine-grained level which is meaningful for intelligent education. One point that needs to be noted is that only a regular hexagon with a circle inside is annotated as *benzene* and the other cases are labelled as *ring* no matter how many bonds are inside. As explained previously, we use one-stage detectors to locate and classify the predefined graphical and text objects in OCCSFs. As verified by the experimental results, YOLO is more friendly to our task than SSD. Thus, we introduce the main procedures of components detection based on YOLOv5 which was initially released in 2020. As illustrated in Fig. 5, YOLOv5 consists of three parts: backbone, neck and prediction. To improve the robustness of the model, different data augmentation techniques are used first such as mosaic, random affine (scale and translation), augment HSV, random horizontal flip. Then these images are adjusted into a fixed size (such as 512 × 512) to be fed into YOLOv5's backbone for feature extraction. The backbone of YOLOv5 is mainly composed of CBS, CSP and SPPF, generating three feature maps of different scales. Then, the neck which includes CBS, Upsample, Concat and CSP is adopted to fuse these feature maps. Finally, the fused features will be sent to the prediction part to produce a diverse set of predictions.

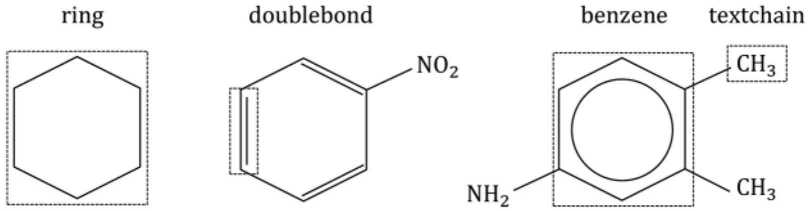


Figure 4: Illustration for the predefined objects

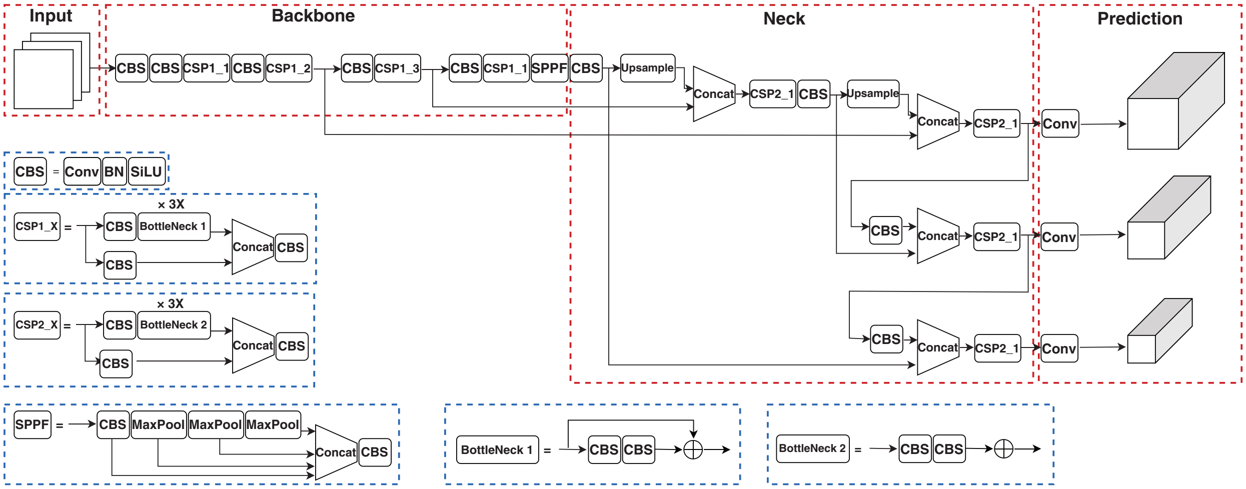


Figure 5: The overall network structure of YOLOv5

4.2 *Text Components Recognition*

Different from the other 3 types of components, *textchain* need to be recognized further to obtain the sequence of labels. In [35], an end-to-end trainable system was proposed for recognizing handwritten chemical formulas. The proposed system adopted the CNN + RNN + CTC framework. With the strategy of introducing additional labels, this framework could interpret the 'subscript' and 'superscript' existing in chemical formula. As stated in work [15], the attention-based method can achieve a higher recognition accuracy than the CTC-based method on isolated word recognition tasks and the text components in OCCSFs appearing in K12 are very similar to words. Therefore, we will adopt the attention-based encoder-decoder model for text components recognition. However, the classical attention-based encoder-decoder model suffers from the error accumulation problem. In [33], a decoupled attention network was proposed which solves the error accumulation problem partially. In this work, we combine DAN and scheduled sampling [34] to further mitigate the problem and then apply the improved attention-based model for text components recognition.

The improved attention-based encoder-decoder model is composed of three major parts: encoder, convolutional alignment module and text decoder with scheduled sampling. The details will be given in the following paragraphs.

4.2.1 Encoder

Resnet is adopted as the encoder to extract features from *textchain* components image regions. As Fig. 6 shows, it stacks 23 residual blocks. The size and number of kernels we use are embedded in each block.

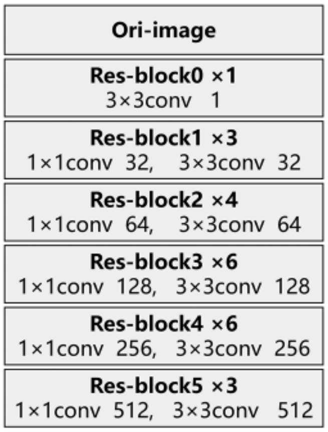


Figure 6: The structure of encoder

4.2.2 Convolutional Alignment Module

This module takes a FCN-like architecture to compute the attention map directly, which is different from the traditional attention mechanism where the computation of the features currently attended depends on the previous token generated by the decoder. First, the features at each scale extracted by the encoder are fed into the convolution stage which contains several down-sampling convolutional layers; Then the deconvolution stage, by adding the feature in the corresponding convolution layer, makes dense predictions per-pixel channel-wise. The number of channels equals to the number of decoding steps.

4.2.3 Decoder with Scheduled Sampling

The decoder takes the feature map (from the encoder) and the attention map (from the convolutional alignment module) as input and outputs the sequence of labels. During the procedure of decoding, the decoder uses a GRU (gated recurrent unit) layer to model the contextual information. As illustrated in Fig. 7, The current hidden state of GRU is determined by three sources: the previous hidden state h_{t-1}, the previous token y_{t-1} or g_{t-1} and the current context vector c_t. The current hidden state h_t is computed as the following formula:

$$h_t = GRU((e_{t-1}, c_t), h_{t-1})) \tag{1}$$

where e_{t-1} is the embedding vector of the previous token y_{t-1} or g_{t-1}. The computation of c_t depends on the encoded features and the attention map.

The previous token adopted being the ground truth token g_{t-1} or the decoded output y_{t-1} is decided by scheduled sampling. We use ϵ_i to represent the probability of taking the ground truth token in the i^{th} mini-batch of the training phase. By intuition, ϵ_i should favor the ground truth token more at the beginning as the model is not well trained yet and pay more attention to the decoded token since it is the real case in the inference phase. To decrease ϵ_i from 1 to 0, different decay functions can be chose, being linear decay, exponential decay and inverse sigmoid decay. Here we provide the linear decay function in detail. For more information, please refer to literature [34].

$$\epsilon_i = \max(\epsilon_{min}, k - c * i) \quad (2)$$

where $\epsilon_{min} \in (0, 1)$ represents the minimum probability to use the ground truth label, and k, c are the offset and slope of the decay, respectively.

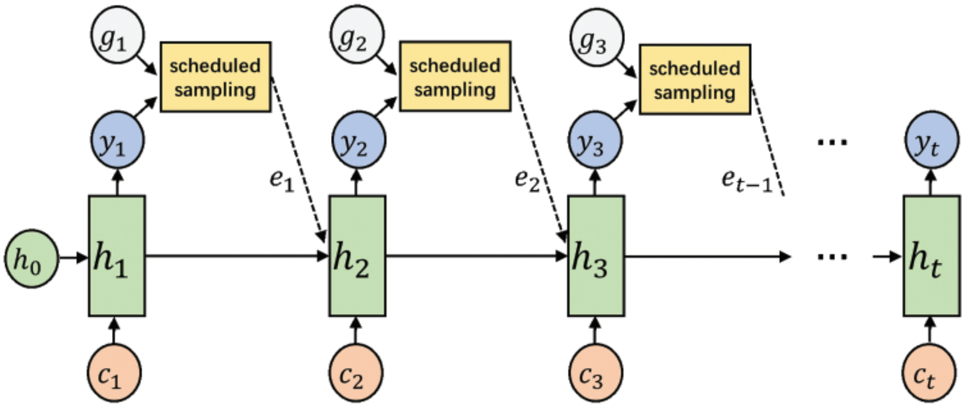


Figure 7: Detailed structure of the decoder with scheduled sampling. g_i represents the ground truth label

4.3 Structure Interpretation

In this work, we focus on the task of multi-ring structure interpretation. Ideally, we can obtain the accurate bounding boxes and categories of the predefined components in hand-drawn OCCSFs. With these bounding boxes and categories information, the structure can be interpreted unambiguously via analysing the spatial relationships between them. We introduce the proposed interpretation algorithm in detail as follows.

4.3.1 Geometric Property of OCCSFs

The standard benzene structure is a regular hexagon with a circle or 3 bonds inside. There also exist other cases where a regular hexagon is with 0, 1, 2 bonds inside [36]. No matter circle or bonds are inside, no matter where the bonds are, a regular hexagon (ring) could be drawn in 2 formats, being horizontal and vertical as shown in Fig. 8. Obviously, each hexagon (ring) has 6 vertices and 6 edges. In some cases, the edge and the internal bond form a double bond together. If we link each vertex and edge of the hexagon to the origin, 12 axes (6 vertex axes and 6 edge axes) could be obtained and the angle between the adjacent axes is 30°. Being aware of chemical knowledge, we can conclude that *textchain* is linked to the vertex axis and *doublebond* is on the edge axis.

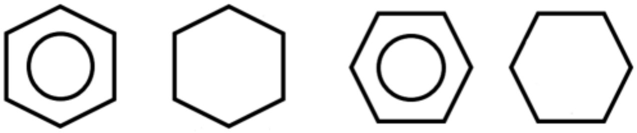


Figure 8: The vertical (left) and horizontal (right) formats of the ring structures which have a circle and 0 bond inside, respectively

4.3.2 Interpretation Algorithm

Given the geometric property of OCCSFs, it is intuitive to consider using the angle offset information between detected objects and standard formats to analyse the hand-drawn structure. Motivated by this idea, we design a bottom-up hand-drawn OCCSFs interpretation algorithm which analyzes the single-ring structure first, then connects two interpreted rings.

With respect to the single-ring structure interpretation, we first take the geometric center of detected *benzene* or *ring* object box as the origin to establish a rectangular coordinate system. Then draw an axis every 30° generating 12 axes totally. As shown in Fig. 9, the printed formula is presented with 12 axes. It can be seen that for the standard (printed) benzene structure, all *doublebond* objects are on the short dashed lines representing edge axes of the vertical format, and all *textchain* objects are more close to the long dashed lines representing vertex axes of the vertical format. We use the angle offset between detected objects (*doublebond, textchain*) and 12 axes to identify the format of the single-ring structure being horizontal or vertical. To achieve this, we propose a concept of angle offset index (AOI). The definition of AOI is as follows:

$$AOI = S\left(\sum_i \Delta_i^H - \sum_i \Delta_i^V\right) \tag{3}$$

where $\sum_i \Delta_i^H$ is the angle offset between the hand-drawn structure and the standard horizontal format. $\sum_i \Delta_i^V$ is the angle offset between the hand-drawn structure and the standard vertical format. $S(x)$ is the sigmoid function. When the first offset is greater than the second offset, the output is >0.5, which means that there is a high probability the structure is written in vertical format. Otherwise, the output is <0.5 representing that the structure is more likely written in horizontal format.

$$S(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

Next, we introduce how to compute the angle offset between the hand-drawn ring structure and the standard format. Δ_i^H denotes the angle offset between the i_{th} object (*doublebond* or *textchain*) and the standard horizontal format.

- If it is a *textchain* object, we compute the angle differences with 6 **vertex** axes of standard horizontal format and take the minimum one as $\Delta_i^H = min(\boldsymbol{\theta}_1^v, \boldsymbol{\theta}_2^v, \boldsymbol{\theta}_3^v, \boldsymbol{\theta}_4^v, \boldsymbol{\theta}_5^v, \boldsymbol{\theta}_6^v)$;
- If it is a *doublebond* object, we compute the angle differences with 6 **edge** axes of standard horizontal format and take the minimum one as $\Delta_i^H = min(\boldsymbol{\theta}_1^e, \boldsymbol{\theta}_2^e, \boldsymbol{\theta}_3^e, \boldsymbol{\theta}_4^e, \boldsymbol{\theta}_5^e, \boldsymbol{\theta}_6^e)$.

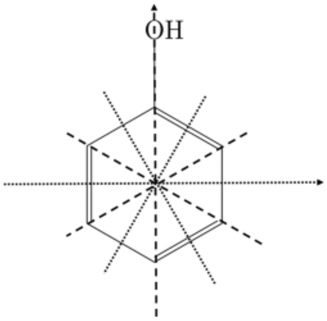


Figure 9: Printed formula is presented with 12 axes

Each angle difference is calculated by measuring the angle between the line connecting two origins (the origin of the detected *benzene* or *ring* box and the origin of the detected *textchain* or *doublebond* box) and the corresponding axis. In the similar way, we can compute Δ_i^V. In Fig. 10, the computation process of Δ_i^H and Δ_i^V for a *textchain* object is illustrated, respectively. Among 6 angles, $\boldsymbol{\theta}_1$ is the smallest one then we assign Δ_i^H and Δ_i^V with $\boldsymbol{\theta}_1$. Then the angle offsets of all the objects (*textchain* and *doublebond*) are combined to compute angle offset index (AOI). With the resulting AOI, we can determine the writing format being horizontal or vertical. Next step is to locate the *textchain* and *doublebond*. To address this problem, we compute the angle differences between the component to be located and 6 corresponding axes and find the minimum one to locate the component at the found axis.

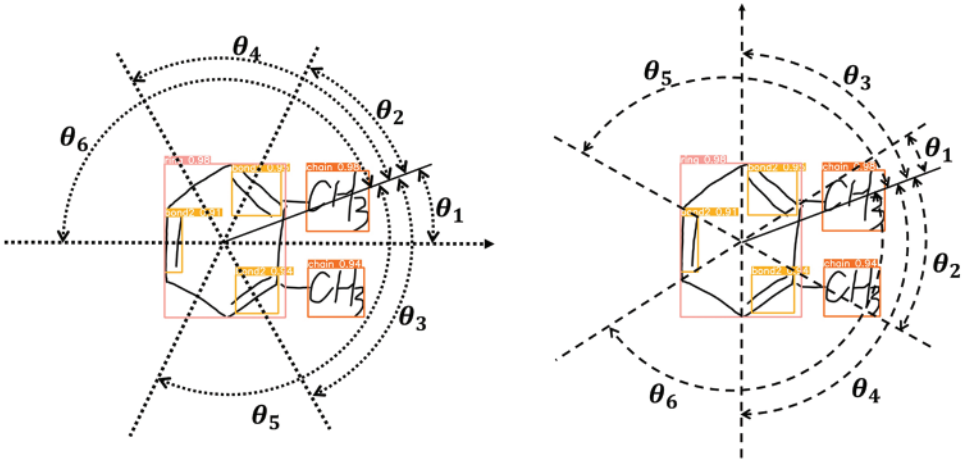


Figure 10: Illustration for the computation process of Δ_i^H (left) and Δ_i^V (right) for a *textchain* object

When the single-ring structure is interpreted, the rest task is to connect two rings if exist. The strategy we take is regarding one ring as *textchain* of the other ring and locating it with the same method aforementioned. Algorithm 1 provides the whole process for structure interpretation.

5 Experiments

5.1 Components Detection

5.1.1 Experiment Settings

The off-line handwritten OCCSFs dataset is divided into the training set, validation set and test set with the ratio of 6:2:2.

- GPU: Nvidia GeForce RTX 2080 Ti with 11G memory
- Library: Pytorch 1.9; CUDA 11.1
- Batch size: 16
- Learning rate: 0.01
- Learning rate decay: Cosine annealing

5.1.2 Evaluation Metrics

In addition to the common evaluation metrics such as precision, recall and mAP in the field of object detection, we define new metrics at the object level to better analyse the effects of components detection results on later structure interpretation.

$$cor. = \frac{n_c}{n} \quad wro. = \frac{n_w}{n} \quad mis. = \frac{n_m}{n} \tag{5}$$

where n denotes the number of objects of one category in the ground truth file, n_c is the number of objects correctly detected, n_w is the number of objects wrongly detected (two types of errors: the object is detected with the wrong label or detected by multiple times) and n_m is the number of objects undetected. Obviously, $n = n_c + n_w + n_m$.

5.1.3 Results

In this section, we detail the components detection results using one-stage detectors, SSD and YOLOv5. We first compare the performances of two detectors on our task with the metric of mAP@0.5. As shown in Table 2, both SSD and YOLO v5 have nearly perfect performances on *benzene*, *ring* and *textchain* detection. But YOLOv5 has achieved a higher mAP@0.5 on *bond2* than SSD which is consistent with the common conclusion that YOLO performs better on detecting smaller objects.

Algorithm 1: Structure Interpretation Algorithm

Input: detected boxes and their categories
Output: the interpretation results

```
Correlate textchain and double2 with the target benzene or ring via computing the minimum Euclidean distance
for all benzene or ring $\in$ objects do
    Find benzene or ring object box and take the geometric center as the origin $O(O_x, O_y)$.
    Calculate the angle offset index (AOI)
    for all textchain or double2 adhere to the selected benzene or ring do
        if $AOI > 0.5$ then
            if object.label == textchain then
                $\Delta = min$ (the angle differences with 6 vertex axes of standard vertical format) locate the object
            else if object.label == bond2 then
                $\Delta = min$ (the angle differences with 6 edge axes of standard vertical format) locate the object
            end if
        else if $AOI < 0.5$ then
            if object.label == textchain then
                $\Delta = min$ (the angle differences with 6 vertex axes of standard horizontal format) locate the object
            else if object.label == bond2 then
                $\Delta = min$ (the angle differences with 6 edge axes of standard horizontal format) locate the object
            end if
        end if
    end for
```

(Continued)

```
Algorithm 1: (Continued)
20:   end for
21:   Connect two rings if required
22:   Return the interpretation results
```

Table 2: Performance comparison of SSD and YOLOv5 on components detection task with the metrics of precision (P), recall (R) and mAP@0.5

	Detectors					
Components	SSD			YOLOv5		
	P	R	mAP@0.5	P	R	mAP@0.5
benzene	0.995	1	0.994	0.998	0.998	0.995
ring	0.995	1	0.995	0.994	1	0.995
textchain	0.996	0.997	0.995	0.995	0.997	0.993
bond2	0.963	0.963	0.977	0.997	1	0.994
all	0.987	0.990	0.990	0.996	0.999	0.994

To deeply analyse the detection results from the point of view of structure interpretation, we also give the detailed information with the proposed metrics in this work. Table 3 provides these interesting data. From the statistics we can see that SSD misses around 4% *bond2* components which will majorly affect later structure interpretation step. Consequently, YOLOv5 is chosen as the components detector in our work. Figs. 11 and 12 present some detection results using SSD and YOLOv5, respectively.

Table 3: Performance comparison of SSD and YOLOv5 on components detection task with the merics of *cor. wro. mis.*

	Object											
Method	*benzene*			*ring*			*textchain*			*bond2*		
	cor.	wro.	mis.	cor.	wro.	mis.	cor.	wro.	mis.	cor.	wro.	mis.
YOLOv5	0.998	0.002	0	1	0	0	1	0	0	1	0	0
SSD	1	0	0	1	0	0	0.995	0	0.005	0.957	0.003	0.040

5.2 *Text Chain Recognition*

5.2.1 *Data set*

There are 4764 *textchain* components in the off-line handwritten OCCSFs dataset which is in fact very limited in terms of quantity, scope and sequence length. Text components of chemical organic structure formulas look visually as same as chemical formulas, both composed of a sequence of chemical symbols. In order to better evaluate the proposed improved attention-based encoder-decoder model for text recognition, we use a larger handwritten chemical formulas dataset published in [35] which consists of 12,224 samples covering 97 chemical formulas. The data is divided into the training and test subset with the ratio of 8:2.

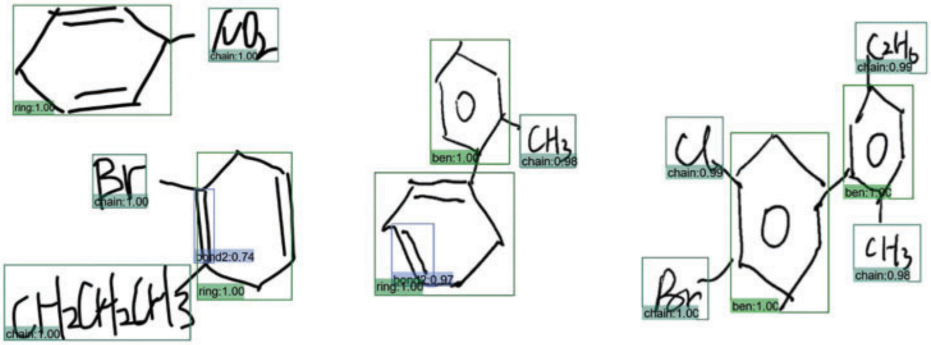


Figure 11: Detection results using SSD

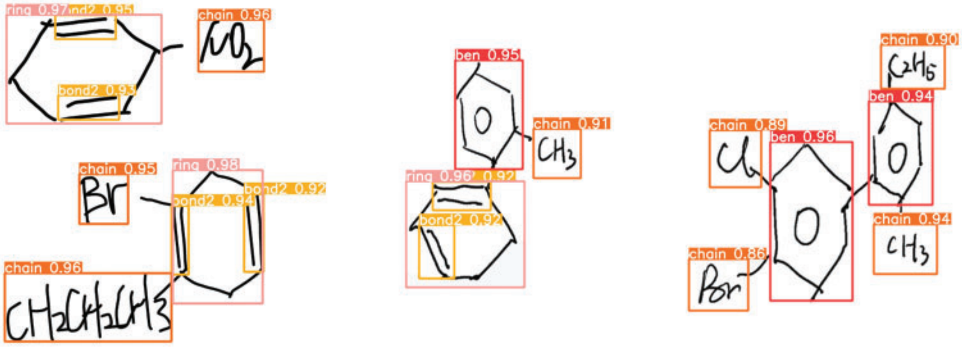


Figure 12: Detection results using YOLOv5

5.2.2 Experiment Settings

- GPU: Nvidia GeForce RTX 2080 Ti with 11 G memory
- Library: Pytorch 1.9; CUDA 11.1
- Batch size: 24
- Learning rate: 0.1
- Learning rate decay: 0.3162

5.2.3 Results

Table 4 provides the detailed results at character level and formula level of handwritten chemical formulas recognition with different methods proposed in different literature [33,35,37]. As can be seen, for the task of handwritten chemical formulas recognition, the vanilla attention-based method [37] performs better than the CTC-based method [35] which is consistent with the conclusion in [15] that attention-based methods can achieve higher recognition accuracy than CTC-based methods on isolated word recognition tasks. When the decoupled attention [33] is introduced, the accuracies increase again (96.70% → 98.61% at formula level) as it decouples the dependency of the computation of the feature at current time step on the previous token in this way easing the problem of error accumulation. Since the error propagation problem is not solved completely, we further integrate scheduled sampling [34] into DAN to improve the model. From the statistics, we can tell that the improved model performs better than DAN [33] which verifies the effectiveness of our method. Three decay functions are tested in our task, among which linear decay performs better than the other two functions. The best results are **99.62%** at the character level and **98.92%** at the formula level which are quite fine.

Table 4: Performance comparison of the proposed methods for handwritten chemical formulas recognition

	Metrics	
Methods	Character level accuracy	Formula level accuracy
[35][1]	—	95.30%
[37][1]	98.95%	96.70%
[33][2]	99.40%	98.61%
Our method (Linear decay)	**99.62%**	**98.92%**
Our method (Exponential decay)	99.45%	98.51%
Our method (Inverse sigmoid decay)	99.56%	98.80%

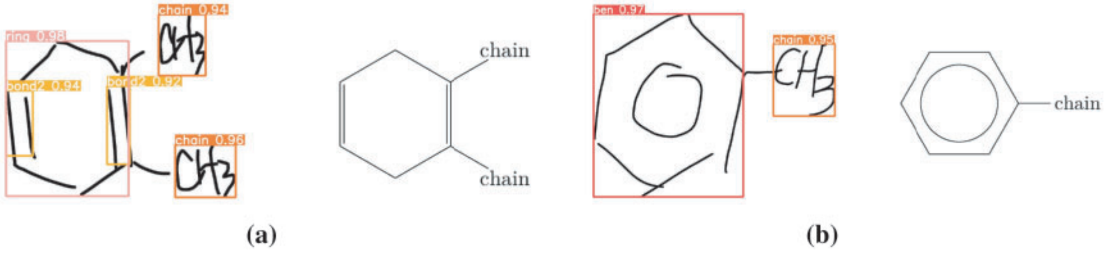
Notes: [1] The results are extracted directly from the published literature. [2] The results are reproduced by us.

We use the proposed model to recognize the detected *textchain* components.

5.3 Structure Interpretation

As introduced in Algorithm 1, the interpretation algorithm takes detected boxes and their categories as input and outputs the corresponding interpretation results. We respectively evaluate the proposed algorithm on the single-ring and multi-ring samples from the test set. An overall accuracy of 74.32% is achieved considering the errors from the components detection step, where 87.62% of single-ring samples and 45.12% of multi-ring samples are correctly interpreted. It can be seen from the statistics that the proposed algorithm performs well on single-ring samples. However, the result on multi-ring samples is not ideal which could be caused by the conflict between the complex structures of multi-rings and the limited representation capability of the proposed angle offset feature. In Fig. 13, we present some interpreted samples including the correct cases, as well as some error cases to indicate the directions to improve the structure interpretation algorithm in future works.

Overall, the proposed approach achieves a total accuracy of 73.52% for off-line handwritten OCCSFs recognition on a self-collected data set.



(a)

(b)

Figure 13: (Continued)

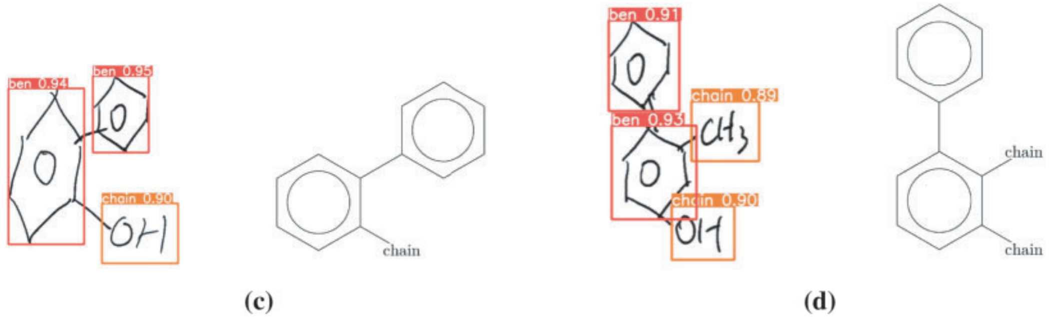


Figure 13: Illustration of some interpreted samples. (a) A correctly interpreted single-ring sample; (b) A wrongly interpreted single-ring sample; (c) A correctly interpreted multi-ring sample; (d) A wrongly interpreted multi-ring sample

6 Conclusion

In this work, we propose an auto-grading oriented approach for off-line handwritten OCCSFs recognition. The proposed method firstly defines different components of OCCSFs as objects and adopts the deep learning detector YOLOv5 to detect them. Then, for the detected text objects, we introduce an improved attention-based encoder-decoder model for text recognition. Finally, a holistic algorithm is designed for interpreting the single-ring structures. With the proposed method, the accurate alignment information between the input and output is available which makes the auto-grading of handwritten chemistry assignments at a fine-grained level possible.

At present, the structure interpretation algorithm works well for single-ring structures but has limited performance on multi-ring samples. It could be caused by the conflict between the complex structures of multi-rings and the limited representation capability of the proposed angle offset feature. No doubt, low robustness, low generalization capability and limited representation capability are the common problems of manually designed features. This weakness will be tackled by learning feature representations automatically in future.

Funding Statement: This work is supported by National Natural Science Foundation of China (Nos. 62007014 and 62177024), the Humanities and Social Sciences Youth Fund of the Ministry of Education (No. 20YJC880024), China Post Doctoral Science Foundation (No. 2019M652678) and the Fundamental Research Funds for the Central Universities (No. CCNU20ZT019).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Southavilay, V., Yacef, K., Reimann, P., Calvo, R. A. (2013). Analysis of collaborative writing processes using revision maps and probabilistic topic models. *The Third International Conference on Learning Analytics and Knowledge*, pp. 38–47. Leuven Belgium.
2. Singh, R., Gulwani, S., Solar-Lezama, A. (2013). Automated feedback generation for introductory programming assignments. *The 34th ACM Sigplan Conference on Programming Language Design and Implementation*, pp. 15–26. Seattle, Washington DC, USA.

3. Gulwani, S., Radiek, I., Zuleger, F. (2014). Feedback generation for performance problems in introductory programming assignments. *22nd ACM SIGSOFT International Symposium on the Foundations of Software Engineering, 45(11),* 41–51.
4. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning*, pp. 369–376. Pittsburgh Pennsylvania, USA.
5. Bahdanau, D., Cho, K., Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.
6. Shi, B., Bai, X., Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(11),* 2298–2304. DOI 10.1109/TPAMI.2016.2646371.
7. Zhang, J., Du, J., Zhang, S., Liu, D., Hu, Y. et al. (2017). Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition. *Pattern Recognition, 71,* 196–206. DOI 10.1016/j.patcog.2017.06.017.
8. Sakshi, S., Kukreja, V. (2021). A retrospective study on handwritten mathematical symbols and expressions: Classification and recognition. *Engineering Applications of Artificial Intelligence, 103,* 104292. DOI 10.1016/j.engappai.2021.104292.
9. Rajan, K., Zielesny, A., Steinbeck, C. (2020). DECIMER-towards deep learning for chemical image recognition. *Journal of Cheminformatics, 12(1),* 65. DOI 10.1186/s13321-020-00469-w.
10. Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 60(6),* 1097–1105.
11. Ren, S., He, K., Girshick, R., Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems, 39(6),* 91–99.
12. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. et al. (2016). Ssd: Single shot multibox detector. *European Conference on Computer Vision*, pp. 21–37. Amsterdam, The Netherlands, Springer.
13. Liao, M., Shi, B., Bai, X., Wang, X., Liu, W. (2017). TextBoxes: A fast text detector with a single deep neural network. *Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, California, USA.
14. Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788. Las Vegas, NV, USA.
15. Cong, F., Hu, W., Huo, Q., Guo, L. (2019). A comparative study of attention-based encoder-decoder approaches to natural scene text recognition. *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 916–921. Sydney, NSW, Australia.
16. Wang, Y., Zhang, T., Yu, X. (2021). A component-detection-based approach for interpreting off-line handwritten chemical cyclic compound structures. *IEEE International Conference on Engineering, Technology, and Education (TALE)*, Wuhan, Hubei Province, China.
17. Ouyang, T. Y., Davis, R. (2007). Recognition of hand drawn chemical diagrams. *AAAI, 7,* 846–851.
18. Ouyang, T. Y., Davis, R. (2011). Chemink: A natural real-time recognition system for chemical drawings. *Proceedings of the 16th International Conference on Intelligent User Interfaces*, pp. 267–276.
19. Sadawi, N. M., Sexton, A. P., Sorge, V. (2012). Chemical structure recognition: A rule-based approach. *Proceedings of SPIE 8297, Document Recognition and Retrieval XIX*. International Society for Optics and Photonics, Burlingame, California, USA.
20. Sun, P., Chen, Y., Lyu, X., Wang, B., Qu, J. et al. (2018). A free-sketch recognition method for chemical structural formula. *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pp. 157–162. Vienna, Austria: IEEE.

21. Bukhari, S. S., Iftikhar, Z., Dengel, A. (2019). Chemical structure recognition (CSR) system: Automatic analysis of 2D chemical structures in document images. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, NSW, Australia: IEEE.
22. Felzenszwalb, P. F., Girshick, R. B., McAllester, D., Ramanan, D. (2009). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9),* 1627–1645. DOI 10.1109/TPAMI.2009.167.
23. Uijlings, J. R., van de Sande, K. E., Gevers, T., Smeulders, A. W. (2013). Selective search for object recognition. *International Journal of Computer Vision, 104(2),* 154–171. DOI 10.1007/s11263-013-0620-5.
24. Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587. Columbus, OH, USA.
25. He, K., Zhang, X., Ren, S., Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9),* 1904–1916. DOI 10.1109/TPAMI.2015.2389824.
26. Girshick, R. (2015). Fast R-CNN. *Proceedings of International Conference on Computer Vision*, Santiago, Chile.
27. Bissacco, A., Cummins, M., Netzer, Y., Neven, H. (2013). Photoocr: Reading text in uncontrolled conditions. *IEEE International Conference on Computer Vision*, pp. 785–792. Sydney, NSW, Australia.
28. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A. (2014). Synthetic data and artificial neural networks for natural scene text recognition. *Annual Conference on Neural Information Processing Systems Deep Learning Workshop*, Montreal, Quebec, Canada.
29. Wang, T., Wu, D. J., Coates, A., Ng, A. Y. (2012). End-to-end text recognition with convolutional neural networks. *International Conference on Pattern Recognition*, pp. 3304–3308. Tsukuba, Japan.
30. Lee, C. Y., Osindero, S. (2016). Recursive recurrent nets with attention modeling for ocr in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2231–2239. Las Vegas, NV, USA.
31. Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X. (2016). Robust scene text recognition with automatic rectification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4168–4176. Las Vegas, NV, USA.
32. Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S. et al. (2017). Focusing attention: Towards accurate text recognition in natural images. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5086–5094. Venice, Italy.
33. Wang, T., Zhu, Y., Jin, L., Luo, C., Cai, M. (2020). Decoupled attention network for text recognition. *Proceedings of the AAAI Conference on Artificial Intelligence, 34(7),* 12216–12224.
34. Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems, 1,* 1171–1179.
35. Liu, X., Zhang, T., Yu, X. (2019). An end-to-end trainable system for offline handwritten chemical formulae recognition. *15th IEEE International Conference on Document Analysis and Recognition (ICDAR)*, pp. 577–582. Sydney, NSW, Australia.
36. Zheng, L., Zhang, T., Yu, X. (2019). Recognition of handwritten chemical organic ring structure symbols using convolutional neural networks. *15th IEEE International Conference on Document Analysis and Recognition, Workshop on Machine Learning (ICDAR-WML)*, pp. 165–168. Sydney, NSW, Australia.
37. Liu, X. (2019). *Comparison between CTC-based and attention-based methods for offline handwritten chemical formulae recognition (Master Thesis).*