ABSTRACT  A critical ideology of the existing Material Genome Project refers to the application of data and artificial intelligence to facilitate material innovation. The lack of data hinders the development of novel materials. The figures and captions in the material literature cover essential information regarding the entire document and have sufficient image sample data for research. Accordingly, how to extract figures and captions from the literature is critical to solve the lack of data. Though some PDF parsing tools are capable of extracting information from documents, they generally identify a document’s figures by parsing the document into a concrete structure. As impacted by the inconsistency of the form of different journals, they commonly achieve wrong recognition results. Thus, an efficient figure and caption extraction network FCENet is proposed in the present study. Inconsistent with other extraction tools, this study first attempts to adopt instance segmentation models to detect figures and their captions, and then extract them. FCENet developed in this study builds upon BlendMask and introduces a horizontal and vertical attention module. This study splits the BlendMask detection head into two branches, i.e., figure detection and caption detection, which increases final detection accuracy and speed. This study collects nearly 3000 material documents for model training and testing. As revealed from the last experiments and results, the performance of FCENet is significantly compared with that of other existing instance segmentation models. Its box and mask mAP (mean Average Precision) are 8.51% and 12.59% higher than those of BlendMask, respectively. This study hopes that considerable material image data can be acquired via FCENet and sufficiently support image data for machine learning and data mining in the material area.

INDEX TERMS  Document content extraction, instance segmentation, BlendMask.

I. INTRODUCTION  Scientific research literature represents the most cutting-edge research results in a range of research areas. Its research direction overall determines each area’s development and boosts research development in science and technology. When the mentioned research results flow from science and technology to industry, they are capable of boosting human society’s progress. The material area refers to one of the essential areas at present, and its development impacts other areas. On the whole, a wide variety of areas require materials in practice. The development and application of several critical materials lay the cornerstones for promoting social progress.

Research and development of novel materials originate from considerable test data. Such process can be accelerated by the analysis and handling of test data. However, it has also led to some difficulties currently facing research and development of novel materials: (1) Lack of data. For some material data, it cannot obtain large amounts of experimentally
achieved data as impacted by the enormous expense of test costs and time costs. (2) Influence of artificial subjective factors. Even though there is considerable material data, some test links are interfered with by artificial emotional factors under manual participation, thereby reducing the accuracy of tested results to a certain extent.

In the current data-driven era, however, the computer area has been of more apparent importance. As artificial intelligence is developed and applied continuously, it has gradually expanded to more areas. Many areas exploit artificial intelligence to allow computers to process data (e.g., biomedicine [1]–[3], agriculture [4], [5], remote-sensing [6], materials [7], and physics [8]), which have achieved effective results. Accordingly, this avoids the interference of artificial factors while lowering time costs.

To solve the lack of data, people began to explore how to extract figures and captions from material literatures. After all, figures and captions of scientific research documents involve essential information in the entire document. Collecting such information and fully exploiting it can enrich experimentally achieved data and has boosted the development of novel materials. As impacted by the complexity and diversity of journal layouts, the randomness of form and shape, as well as scale, figures and captions are difficult to extract directly from the document.

In the present study, a novel and more versatile method is proposed to extract figures and captions from material documents. Inconsistent with other methods, existing instance segmentation models on datasets (e.g., COCO) exhibits good performance, whereas good performance on the document image is not ensured. This study introduces instance segmentation to document image’s figures and captions extraction initially. Given the document image’s particularity, the instance segmentation model FCENet (Figure Caption Extract Net) proposed in this study achieves better results than other models. The work here is summarized below:

1) This study collects over 3,000 material documents, converts them into an image format page by page, and subsequently selects 10,000 from nearly 20,000 images to build an instance segmentation dataset. The dataset format complies with the COCO dataset. To be specific, there are 6000 images in the training set, 2000 images in the validation set, as well as 2000 images in the test set.

2) The present study develops a model FCENet. It builds upon BlendMask, which currently performs optimally in instance segmentation, and introduces a horizontal and vertical attention module while minimizing computational overhead. Box and mask mAP of FCENet in this study are 8.51% and 12.5% higher than those of BlendMask on the dataset here, respectively, and its speed has not decreased.

3) In this study, FCENet follows an allocation strategy of multi-scale targets, which is suitable for document images. Compared with the method adopted by BlendMask, the method here is more ideal for document images.

4) Initially, this study applies instance segmentation in figures and captions extraction of material documents. A novel idea is then presented for figures and captions extraction of the document.

II. RELATED WORK

Given the model’s structure and purpose, four perspectives are presented separately, i.e., figures and captions extraction, attention mechanism, multi-scale target allocation, as well as instance segmentation.

A. DOCUMENT FIGURE AND CAPTION EXTRACTION

According to document types, documents are split into structured documents (documents stored in PDF format) and unstructured documents (early scanned documents). Considerable works have established two research directions around document types. For unstructured documents, Bhownik et al. [9]–[12] all used conventional image processing methods (e.g., region classification or connected components) to identify document structure. Still, the document image is not identical to the natural scene image, so the final effect is generally unsatisfactory. For structured documents, more tools are adopted to extract figures. The mentioned tools parse documents into XML/HTML format and afterward extracted them. For instance, Praczyk et al. [13], in the area of high-energy physics, proposed PDFPlo-tExtractor, and Clark et al. [14] developed PDFigures2 in the area of computer science. The mentioned tools exploit clustering and classification to separate images, while in the biomedical area, LI et al. [15] submitted PDFigCapX, which first separates text and image regions, and conducts the connected component analysis to analyze the image. Image content is identified, and then the layout information of PDF is adopted to obtain the final figures and captions. The mentioned methods are capable of achieving effective results in their respective areas. However, as impacted by the diversity of structured documents in various areas, results are overall ineffective. Especially for extracting figures and captions from journals with considerable differences in contents and layouts, the error will be more prominent, and most methods do not associate figures with their respective captions. Accordingly, numerous shortcomings remain in the extraction of figures and their captions from documents.

B. ATTENTION MECHANISM

In brief, attention mechanism arouses the attention from the neural network on specific information in the image. SENet [16](Squeeze Excitation Network) took the attention mechanism in the channel dimension. It enhances features and suppresses some unimportant features by multiplicative weighting. CBAM [17](Convolutional Block Attention Module) proposed channel attention and spatial attention to weight features, while introducing the attention mechanism into the channel and spatial dimensions. In addition, DANet
tion based on the Anchor-free idea, this situation has been consuming. Still, with the detection framework’s introduction, the two-stage detection framework are significantly time-consuming. Hence, the mentioned instance segmentation models based on the respective instance. Subsequent MaskScoringRCNN object detection framework FasterRCNN, thereby segmenting the multi-scale problem effectively. Small targets correspond to large-scale feature maps, and large targets represent small-scale feature maps. ROI Pooling of FasterRCNN corresponding all target regions to the feature maps of a specific scale. CenterMask, and ROIAlign of MaskRCNN employs formulas to calculate feature maps corresponding to each target region directly. Besides, BlendMask’s leading network uses FCOS (Fully Convolutional One-Stage). Since FCOS refers to an anchor-free detection framework, it is not identical to those mentioned anchor-based frameworks. It was limiting the target’s height and width to solve the problem of multi-scale target allocation. Each feature map represents a scale range. However, this method does not apply to identify targets with a massive difference in height and width. For instance, it will assign a caption with a larger length to a feature map with a smaller scale, but caption itself is a target with a smaller scale. This method will cause problem of missed inspection. Thus, FCENet in this study adopts a distribution strategy more suitable for document images.

C. MULTI-SCALE TARGET ALLOCATION
As impacted by the different scales of targets in images, FPN [22] (Feature Pyramid Networks) is introduced to address the multi-scale problem effectively. Small targets correspond to large-scale feature maps, and large targets represent small-scale feature maps. ROI Pooling of FasterRCNN corresponding all target regions to the feature maps of a specific scale. CenterMask, and ROIAlign of MaskRCNN employs formulas to calculate feature maps corresponding to each target region directly. Besides, BlendMask’s leading network uses FCOS (Fully Convolutional One-Stage). Since FCOS refers to an anchor-free detection framework, it is not identical to those mentioned anchor-based frameworks. It was limiting the target’s height and width to solve the problem of multi-scale target allocation. Each feature map represents a scale range. However, this method does not apply to identify targets with a massive difference in height and width. For instance, it will assign a caption with a larger length to a feature map with a smaller scale, but caption itself is a target with a smaller scale. This method will cause problem of missed inspection. Thus, FCENet in this study adopts a distribution strategy more suitable for document images.

D. INSTANCE SEGMENTATION
Combining target detection and semantic segmentation is a challenging task. There have long been two directions, i.e., a bottom-up semantic segmentation-based method and a top-down detection-based method. Existing detection-based methods still dominate, and the most classic one is MaskRCNN, adding a semantic segmentation branch to object detection framework FasterRCNN, thereby segmenting the respective instance. Subsequent MaskScoringRCNN [26] based on MaskRCNN, a scoring mask quality unit, is introduced to improve the segmentation accuracy. However, the mentioned instance segmentation models based on the two-stage detection framework are significantly time-consuming. Still, with the detection framework’s introduction based on the Anchor-free idea, this situation has been dramatically optimized. Like CenterMask, by combining position attention to segment each instance, PolarMask [27] abandoned the detection boxes by using the Cartesian coordinate system. Still, it returned the contour of the instance based on the polar coordinate system. DeepSnake [28] is based on the detection boxes, and it uses a circular convolution structure to regress the instance contour, then obtain each instance mask. Unlike the mentioned two directions, Yolact [29] (You Only Look At Coefficients) proposed a novel idea, i.e., to process the detection task and the semantic segmentation task in parallel and then linearly combine high-level semantic information with underlying spatial information to generate the final mask. The subsequent Yolact++ [30] also introduced a mask score branch to improve segmentation accuracy. BlendMask introduces an attention mechanism to better integrate the detection branch’s semantic information with the spatial information of the semantic segmentation branch. BlendMask surpasses MaskRCNN for accuracy and speed. Accordingly, a more robust model BlendMask is taken here, as the basic structure of FCENet.

III. FCENet (FIGURE CAPTION EXTRACT NET)
Though BlendMask has achieved effective results on the COCO dataset, BlendMask remains insufficient for document images. After all, document images are not identical to natural scene images. For document images, FCENet has made several changes based on BlendMask. The present section elucidates FCENet.

A. OVERALL NETWORK STRUCTURE
FCENet is an instance segmentation model consisting of a backbone network, an attention module, a detection network, as well as a mask branch. The backbone network covers ResNet and FPN. The attention module is composed of horizontal attention and vertical attention. Horizontal attention strengthens the horizontal direction’s features, and vertical attention enhances the vertical direction’s features. The detection network comprises a figure detection network and a caption detection network, capable of detecting figures and captions of the document. The mask branch combines low-level spatial information with high-level semantic information to obtain the final instance mask.

The overall framework of FCENet builds upon BlendMask with three modifications, i.e., (1) added horizontal and vertical attention, (2) divided detection tower into a figure and a caption detection tower, and (3) split box attention module into a figure and a caption box attention in top module for prediction top attention. Lastly, figures and captions are extracted from the document through the box and mask. The entire network framework is illustrated in Fig. 1.

The attention module refers to a vital part of the model FCENet in this study. It weights feature maps to improve the network’s overall performance. The detail of this module will be discussed in Subsection B.

The detection network of FCENet is similar to FCOS. The difference is that FCENet divided the detection network into
Algorithm 1 Blender Module Fusion Process

Input: bottom module base $B \in \mathbb{R}^{C \times H \times W}$, top T prediction boxes $P = \{p_t \in \mathbb{R}^{d \times 4} | d = 1 \ldots T\}$, top attention $A_{fig} = \{a_{fig} \in \mathbb{R}^{|T| \times 7 \times 7}\}$, $A_{caption} = \{a_{caption} \in \mathbb{R}^{|T| \times 7 \times 7}\}$
Output: Instance Mask

Begin
Step 1: Fusion top attention
$A = \text{CONCAT}(A_{fig}, A_{caption})$

Step 2: Generate target area
$r_t = \text{RoIPool}_{56 \times 56}(B', p_t). \ \forall t \in \{1 \ldots T\}$

Step 3: Linear interpolation
$a_t^i = \text{interpolate}_{7 \times 7 \rightarrow 56 \times 56}(a_t), \ \forall t \in \{1 \ldots T\}$

Step 4: Generate attention score map
$s_t = \text{softmax}(a_t^i), \ \forall t \in \{1 \ldots T\}$

Step 5: Generate mask $m_t$
$m_t = \sum_{i=1}^{T} r_t^i \ast r_t^i, \ \forall t \in \{1 \ldots T\}$

End

map $V'$. Then getting new feature maps by concat these three feature maps. The overall process is below:

Where $f_1^{1 \times 1}$ represents a $1 \times 1$ convolution, $f_2^{5 \times 5}$ represents two $5 \times 5$ convolutions, $f_3^{7 \times 1}$ represents a $7 \times 1$ convolution, $f_4^{1 \times 7}$ means a $1 \times 7$ convolution, $f_{HAM}^{1 \times 1}$ means horizontal attention module, $f_{VAM}^{1 \times 1}$ means vertical attention module, and $f_{CONCAT}$ means concat.

1) HORIZONTAL ATTENTION MODULE
The horizontal attention module is illustrated in Fig. 3.

It processes horizontal feature map $H$ at channel and position levels, respectively, obtain $H_c = \{h_{cj} \in \mathbb{R}^{C \times H \times W} | j = [1, 2 \ldots C']\}$ and $H_s = \{h_{sj} \in \mathbb{R}^{C \times H \times W} | j = [1, 2 \ldots C']\}$, then add $H = \{h_j \in \mathbb{R}^{C \times H \times W} | j = [1, 2 \ldots C']\}$ to obtain the final horizontal attention-weighted feature map

FIGURE 1. The overall framework of FCENet. FPN as Feature Pyramid Network, Attns as Attentions, fig as Figure.

B. ATTENTION MODULES
The attention module consists of the horizontal attention module and the vertical attention module. The overall structure is illustrated in Fig. 2.

Feature map outputs by FPN as the input, it contains three branches: (1) Calculate basic feature map $P_i$, (2) Calculate horizontal attention map $H'$, (3) Calculate vertical attention module, (3) Calculate basic feature map $P_i$, (3) Calculate horizontal attention map $H'$, (3) Calculate vertical attention module.
$H' \in \mathbb{R}^{C' \times H \times W}$. The overall operation is defined below:

$$H' = \sum_{j=1}^{C'} h_j + h_{s_j} + h_{c_j}$$

(3)

where $j$ represents the number of channels in the feature map, $j = [1, 2 \ldots C']$.

$$h_{s_j} = \alpha \times S \odot h_j$$

(4)

$$h_{c_j} = \beta \times (a_{jl} \odot a'_{j}) \otimes h_j$$

(5)

where $\alpha$ and $\beta$ denote two learnable parameters adopted to determine the weight of channel and position attention, $\odot$ which means element-wise multiplication, $\otimes$ means matrix multiplication. Attention map $S \in \mathbb{R}^{1 \times H \times W}$:

$$S = \frac{1}{1 - e^{-((f^3 \times \lambda \cdot \text{GAP}(H)) \cdot \text{GMP}(H))}}$$

(6)

where GAP and GMP represent global average pooling and global maximum pooling, respectively, $f^c$ represents full
Algorithm 2: The Overall Calculation Process of the Attention Module

**Input:** Feature map \( P = \left \{ p_i \in \mathbb{R}^{N \times 256 \times H \times W} \mid i = 2 \ldots 6 \right \} \)

**Output:** Feature map weighted by attention \( P_a = \left \{ p_a \in \mathbb{R}^{N \times 6 \times M \times W} \mid i = 2 \ldots 6 \right \} \)

**Begin**

**Branch 1:** Calculate basic feature map \( P_b = \left \{ p_b \in \mathbb{R}^{N \times C \times H \times W} \mid i = 2 \ldots 6 \right \} \)

\[
P_b = f^{1 \times 1}(P), \quad P_b = \left \{ p_b \in \mathbb{R}^{N \times C \times H \times W} \mid i = 2 \ldots 6 \right \}
\]

\[
p_b = f^{1 \times 1}(P_b)
\]

**Branch 2:** Calculate horizontal attention map \( H = \left \{ h_i \in \mathbb{R}^{N \times C \times H \times W} \mid i = 2 \ldots 6 \right \} \)

\[
P_M = f^{1 \times 1}(P), \quad P_H = \left \{ p_H \in \mathbb{R}^{N \times C \times H \times W} \mid i = 2 \ldots 6 \right \}
\]

\[
H = f^{1 \times 1}(P_H), \quad H = \left \{ h_i \in \mathbb{R}^{N \times C \times H \times W} \mid i = 2 \ldots 6 \right \}
\]

**Branch 3:** Calculate vertical attention map \( V = \left \{ v_i \in \mathbb{R}^{N \times C \times H \times W} \mid i = 2 \ldots 6 \right \} \)

\[
P_V = f^{1 \times 1}(P), \quad P_V = \left \{ p_V \in \mathbb{R}^{N \times C \times H \times W} \mid i = 2 \ldots 6 \right \}
\]

\[
V = f^{1 \times 1}(P_V), \quad V = \left \{ v_i \in \mathbb{R}^{N \times C \times H \times W} \mid i = 2 \ldots 6 \right \}
\]

\[
V = f_{VAM}(H)
\]

\[
P_a = f_{CONTACT}(P_b, (H^T + V)), \quad \dim = 1
\]

**End**

The overall structure of the vertical attention module is illustrated in Fig. 4.

2) VERTICAL ATTENTION MODULE

The overall structure of the vertical attention module is illustrated in Fig. 4.

Its basic structure complies with that of the horizontal attention module. Among them,

\[
V_c = \left \{ v_{cj} \in \mathbb{R}^{C \times H \times W} \mid j = 1, \ldots C^c \right \},
\]

\[
V_s = \left \{ v_{sj} \in \mathbb{R}^{C \times H \times W} \mid j = 1, \ldots C^s \right \}
\]

and the final feature map \( V' \in \mathbb{R}^{C \times H \times W} \) weighted by vertical attention. The overall operation is defined below:

\[
V' = \sum_{j=1}^{C} v_j + v_{cj} + v_{sj}
\]

where

\[
v_{cj} = \alpha' \times S' \odot v_j
\]

\[
v_{sj} = \beta' \times (y_{jl} \odot y') \odot v_j
\]

where \( \alpha' \) and \( \beta' \) respectively denote also two learnable parameters, the attention map \( S' \in \mathbb{R}^{1 \times H \times W} \):

\[
S' = \frac{1}{1 - e^{-\left(f^{1 \times 3}(H_{\text{AM}}(H))\right)}}
\]

Attention map \( Y = \left \{ y_{jl} \in \mathbb{R}^{C \times C'} \right \} \) (feature map \( Y_{\text{reshape}} \in \mathbb{R}^{C \times (H \times W)} \)) is multiplied by \( Y_{\text{reshape}} \in \mathbb{R}^{(H \times W) \times C} \) and then obtained after softmax function:

\[
y_{jl} = \frac{\exp (v_j \odot v_j)}{\sum_{i=1}^{C} \exp (v_j \odot v_j)}
\]

\( y_{jl} \) represents the effect of the \( j^{th} \) vertical feature channel on the \( l^{th} \) horizontal feature channel, the attention map \( Y' = \left \{ y_j' \in \mathbb{R}^{C \times C'} \mid j = 1, 2 \ldots C' \right \} \):

\[
Y' = \frac{1}{1 - e^{-\left(f_c(GAP(H)) + f_c(GMP(H))\right)}}
\]

IV. EXPERIMENTS AND RESULTS

Dataset and assessment indicators. In this study, 3000 material documents are collected to generate an instance segmentation dataset complying with the COCO dataset format. It consists of 10,000 images, i.e., 6000 images in the training set, 2000 images in the validation set, as well as 2000 images in the test set. There are 1000 images from other areas (e.g., biomedicine, agriculture, and remote sensing) documents. There are 3 types of instance tags, except for background categories. The assessment indicator of this experiment is identical to that of the COCO dataset, which is the average precision (AP) of the box and mask, AP (AP50) at IOU 0.5, as well as AP at 0.5 (AP85). And four more assessment indicators (Precision, Recall, F1score, and Accuracy).

Training details. FCENet uses RestNet-50 and FPN as the backbone network. The number of channels in the bottom module is 128, and that of channels in the detection network reaches 64. The model here trains on rtx2060 GPU for 180K iterations, the batch size is 1, the maximum learning rate is 3e-5, and the learning rate at the end reaches 1e-6. Before 21.6K, the learning rate is raised linearly, and then the learning rate decreases slowly like a Cos function. Lr curve is illustrated in Fig. 5(left). Input images are resized to a maximum side length of 832 as well as a minimum size length of 672. All hyperparameters are set to be identical to those of BlendMask. Loss curve is illustrated in Fig. 5(right).

Verify details. The unit of inference time is ‘s’ in all tables. The present study verifies the performance of the model FCENet by a series of ablation experiments.
A. ABLATION EXPERIMENTS

In this study, ablation experiments are performed to verify FCENet’s effectiveness with attention module, aspect ratio allocation strategy, as well as detection network modules.

1) DISTRIBUTION STRATEGY

FCENet vs. BlendMask. Allocation strategy used by BlendMask, same as FCOS, limiting a scope box on the feature map of each scale to solve the allocation problem of multi-scale targets. In the present study, FCENet limits each scale feature map’s detection range of the target aspect ratio and splits the detection network into a figure detection network and a caption detection network. The methods of BlendMask and FCOS primarily aim at detecting conventional objects in natural scene images. As opposed to the mentioned, this study’s method aims to detect document images and items with noticeable differences in length and width.

The results of the mentioned methods are listed in Table 1. The method in the present study performs better on document images. Because the scope box limits the long side of the target, it will assign targets with noticeable differences in length and width to the feature map that does not match the scale. In contrast, this study uses the aspect ratio to solve this problem and transfer it to the appropriate scale feature map, thereby avoiding missed detection. The histogram of aspect ratio distribution is illustrated in Fig. 6.

2) DETECTION TOWER

FCENet vs. BlendMask. The detection tower of BlendMask, like FCOS, takes the feature map of the FPN network (Fig. 1)
TABLE 1. Comparison of different distribution methods of detection network.

| Distribution method | $AP_{box}$ | $AP_{mask}$ | $AP_{mask}^{50}$ | $AP_{mask}^{55}$ |
|---------------------|------------|-------------|------------------|------------------|
| Scope box           | 71.65      | 63.8        | 75.55            | 50.56            |
| Aspect ratio        | 73.28      | 67.28       | 81.46            | 51.86            |

The model used here does not add horizontal and vertical attention modules, and the number of channels in the detection tower is 256. The scope box is a comparison of the allocation methods used by FCENet, BlendMask, and FCOS.

Figure 6. Histogram of aspect ratio distribution. The abscissa is the inverse of aspect ratio, and the ordinate is the number of targets in the dataset.

TABLE 2. Comparison of detection tower.

| Detection tower       | $AP_{box}$ | $AP_{mask}$ | $AP_{mask}^{50}$ | $AP_{mask}^{55}$ |
|-----------------------|------------|-------------|------------------|------------------|
| Shared detection tower| 71.65      | 63.8        | 75.55            | 50.56            |
| Figure caption detection tower | 72.06  | 66.89       | 79.83            | 52.82            |

Compared with BlendMask, the model used here only differs from the detection tower. The shared inspection tower is a comparison between FCENet and BlendMask.

TABLE 3. Comparison of the number of channels in the detection tower.

| Model    | Att module | Channels | Time(s) | $AP_{box}$ | $AP_{mask}$ | $AP_{mask}^{50}$ | $AP_{mask}^{55}$ |
|----------|------------|----------|---------|------------|-------------|------------------|------------------|
| BlendMask| NO         | 256      | 0.54    | 71.85      | 65.8        | 75.55            | 50.56            |
| FCENet   | YES        | 128      | 0.47    | 71.9       | 67.55       | 79.65            | 56.75            |
|          |            | 256      | 0.5     | 76.43      | 71.76       | 87.54            | 55.81            |

3) THE NUMBER OF CHANNELS IN THE DETECTION TOWER
Time-cost increases as impacted by the attention module’s addition. The present study chooses to reduce the time-cost by compressing the number of channels in the detection tower. In order to find a balance point between time and performance, this study uses different channel numbers to measure the model’s performance. It tried 64, 128, and 256 channel numbers on FCENet and compared them with BlendMask. The results are listed in Table 3. The performance of FCENet decreases with the increase in channels, and time also increases a lot. Even FCENet with 64 channels is much better than BlendMask in performance. Considering performance and time cost comprehensively, unless otherwise specified, the number of detection tower channels used in the experimental model is 64.

4) ATTENTION MODULE
FCENet’s horizontal and vertical attention modules enhance horizontal and vertical features. In addition, horizontal attention, vertical attention, horizontal and vertical attention are compared to verify the model’s effectiveness. The results are listed in Table 4. Both horizontal attention and vertical attention dramatically improved the performance of this study’s model. Since the neatness of the document image’s content, the horizontal and vertical attention module of FCENet enhances the model’s performance by weighting the horizontal and vertical features. FCENet exhibits high performance in the document image.

5) TOP ATTENTION
In the top module of FCENet, to better distinguish between figures and captions, this study expands the box attention of BlendMask into two, i.e., fig-box attention and caption-box attention. Compared with single top attention, the results are listed in Table 5. The two-box attention methods are better.
As impacted by the massive difference between the figures and captions, single box attention cannot predict figures and captions’ top information well.

6) FUSION METHOD
Addition vs. concat. Attention module of FCENet, to fuse the feature maps $H'$, $V'$ and $P'_b$, the horizontal attention and vertical attention have weighted that. The present study tried element-wise addition and concat. The results are listed in Table 6. In contrast, the concat method outperforms the element-wise addition method. It is capable of more effectively integrating the horizontal feature map, the vertical feature map, as well as the base feature map. It retains features after attention weighting, while ensuring the model’s performance. As opposed to the mentioned, the element-wise addition method may cause overlapping effects and lose several features, which adversely affects model performance.

| Model          | $AP_{box}$ | $AP_{mask}$ | $AP_{mask_{50}}$ | $AP_{mask_{80}}$ |
|----------------|------------|-------------|------------------|-----------------|
| FCENet + a     | 79.34      | 73.75       | 88.7             | 58.67           |
| FCENet + b     | **80.16**  | **76.39**   | **92.63**        | **60.55**       |

Among them, a is single box attention, and b is two kinds of box attention.

| Fusion method       | $AP_{box}$ | $AP_{mask}$ | $AP_{mask_{50}}$ | $AP_{mask_{80}}$ |
|---------------------|------------|-------------|------------------|-----------------|
| Element-wise addition| 75.69      | 70.53       | 84.34            | 57.57           |
| Concat              | **80.16**  | **76.39**   | **92.63**        | **60.55**       |
TABLE 7. Comparison of results.

| Method   | Time/s | AP<sub>box</sub> | AP<sub>mask</sub> | AP<sub>IoU</sub> | PR<sub>box</sub> | PR<sub>mask</sub> | PR<sub>IoU</sub> | Acc<sub>box</sub> | Acc<sub>mask</sub> | Acc<sub>IoU</sub> |
|----------|--------|------------------|------------------|------------------|-----------------|-----------------|-----------------|----------------|-----------------|----------------|
| MaskRCNN | 1.62   | 67.46            | 63.21            | 74.2             | 56.36           | 62.66           | 59.33           | 60.95          | 60.67           |
| Yolact   | 0.35   | 60.5             | 61.47            | 74.05            | 51.94           | 57.97           | 55.09           | 56.26          | 51.05           |
| BlendMask| 0.54   | 71.65            | 69.87            | 75.55            | 50.96           | 81.52           | 58.35           | 64.48          | 75.21           |
| FCENet   | 0.43   | **80.16**        | **76.39**        | **92.63**        | **60.55**       | **75.51**       | **71.22**       | **73.21**      | **77.37**       |

<sup>AP<sub>box</sub> as Precision, AP<sub>mask</sub> as Recall, AP<sub>IoU</sub> as F1-score, Acc<sub>box</sub> as Accuracy (IoU at 0.85)</sup>

FIGURE 10. More comparison results of FCENet and BlendMask.

B. MAIN RESULTS

On the validation set in this study, we compare FCENet with BlendMask, MaskRCNN, and Yolact. As listed in Table 7, the all-around performance of FCENet in this study surpasses BlendMask, MaskRCNN, and Yolact. On RTX2060 GPU, the running speed of FCENet reaches 0.43s/image, the running speed of BlendMask is 0.54s/image, the running speed of MaskRCNN is 1s+/image, and the running speed of Yolact is 0.35s/image.

On the test set, this study compares the results of FCENet with BlendMask, MaskRCNN, and Yolact. The visualized mask is illustrated in Fig. 9. FCENet in this study produces a higher-quality mask than BlendMask, MaskRCNN, and Yolact, and some models omit target (e.g., caption). Compared with other models, FCENet in this study is more suitable for processing document images.

After the final mask and box are generated, it is capable of extracting figures and captions of the documents. The overall process is presented in Fig. 7. The extracted results are illustrated in Fig. 8, where Fig. 8(a) is the document’s original image. The extracted results from documents show in Fig. 8(b).

Moreover, to verify the versatility of FCENet, this study collects documents in other areas to test FCENet. The results are illustrated in Fig. 11. FCENet of the present study is not merely prominent in material documents, and it also performed well in other areas.

C. DISCUSSIONS

Compared with BlendMask, FCENet in this study adds horizontal and vertical attention modules to enhance the horizontal and vertical features, which can more effectively process document images. It adopts aspect ratio to limit the target’s distribution on multi-scale feature maps and avoids missed detection caused by scale mismatch. Moreover, FCENet splits the detection tower into a figure detection tower and a caption detection tower, which is more flexible for differentiated targets (e.g., figures and captions).
The mentioned methods have enhanced the performance of FCENet.

Another advantage of FCENet is that it is capable of generating higher-quality masks and boxes of document images because the model in this study is more aimed at document images. Accordingly, with high-quality masks and boxes, it is allowed to better extract figures and captions of material documents. In comparison with other methods, FCENet in
FIGURE 12. More extraction results of figures and captions.
this study has ensured both quality, speed and versatility. Even on low-resolution images, its mAP has only dropped by 1.34%.

Lastly, this study visualizes the comparison results of FCENet(top) and BlendMask(bottom) on the test set (Fig. 10). Two sets of images are listed in rows. Moreover, figure and caption extraction results are visualized (Fig. 12), where the left side is the original image, and the right side is the extracted result. The present study’s model can also segment the figure’s sub-figures, but other extraction methods currently cannot do this.

V. CONCLUSION
Given the lack of image data in some areas of materials, this study chooses to collect image data from material documents. Instance segmentation is adopted here to extract figures and captions of material documents for the first time. It develops an instance segmentation model FCENet, based on BlendMask and suitable for processing document images. Second, this study collects and produces an instance segmentation dataset with COCO format and selected box and mask mAP as assessment indicators consistent with the COCO dataset. Besides, this study verifies the performance of the model by relevant ablation experiments. Compared with BlendMask, (1) FCENet adds horizontal and vertical attention modules, it enhances the features in the horizontal and vertical directions. Its box and mask mAP increase by 7.69% and 9.95%, respectively. (2) FCENet adopts an aspect ratio distribution strategy. It avoids the missed detection of targets with significant differences in length and width (e.g., captions). Its box and mask mAP increase by 1.63% and 3.84%, respectively. (3) FCENet splits the shared detection tower into figure caption detection towers and splits the top box attention into figure caption box attention. It is more flexible for recognizing targets with noticeable differences (e.g., figures and captions). Its box and mask mAP increase by more than 8%. In addition, the present study is also compared with other instance segmentation models. FCENet outperforms BlendMask and MaskRCNN on accuracy and speed, and it exceeds Yolact on accuracy too. The box and mask mAP of FCENet are respectively 8.51% and 12.59 % higher than BlendMask, and speed is 0.11s faster. For MaskRCNN and Yolact increased by over 10%. As revealed from experimental results, the mask mAP of FCENet reaches 76.39%, which performs better on document images. This instance segmentation framework is considered a novel way to collect material image data.

REFERENCES
[1] T. Ajiaparaoon, N. Rakratchatakul, M. Chantadisai, U. Vutrapongwutana, K. Kingphet, S. Sirisalapho, Y. Rakvongthai, T. Chaiwatanarat, and E. Chuangsawanich, “MaligNet: Semi-supervised learning for bone lesion instance segmentation using bone scintigraphy,” IEEE Access, vol. 8, pp. 27047–27066, 2020.
[2] M. Chung, M. Lee, J. Hong, S. Park, J. Lee, J. Lee, I.-H. Yang, J. Lee, and Y.-G. Shin, “Pose-aware instance segmentation framework from cone beam CT images for tooth segmentation,” Comput. Biol. Med., vol. 120, May 2020, Art. no. 103720.
[3] N. Kumar, R. Verma, D. Anand, Y. Zhou, O. F. Onder, E. Tsougkou, H. Chen, P. A. Heng, J. Li, Z. Hu, and Y. Wang, “A multi-organ nucleus segmentation challenge,” IEEE Trans. Med. Imag., vol. 39, no. 5, pp. 1380–1391, May 2020.
[4] Y. Tian, G. Yang, Z. Wang, E. Li, and Z. Liang, “Instance segmentation of apple flowers using the improved mask R-CNN model,” Biosystems Eng., vol. 193, pp. 264–278, May 2020.
[5] C. Chen, B. Li, J. Lui, T. Bao, and N. Ren, “Monocular positioning of sweet peppers: An instance segmentation approach for harvest robots,” Biosyst. Eng., vol. 196, pp. 15–28, Aug. 2020.
[6] H. Su, S. Wei, S. Liu, J. Liang, C. Wang, J. Shi, and X. Zhang, “HQ-ISNet: High-quality instance segmentation for remote sensing imagery,” Remote Sens., vol. 12, no. 6, p. 998, Mar. 2020.
[7] D. Chen, D. Guo, S. Liu, and F. Liu, “Microstructure instance segmentation from aluminum alloy metallographic image using different loss functions,” Symmetry, vol. 12, no. 4, p. 639, Apr. 2020.
[8] Z. Yang, R. Dong, H. Xu, and J. Gu, “Instance segmentation method based on improved mask R-CNN for the stacked electronic components,” Electronics, vol. 9, no. 6, p. 886, May 2020.
[9] S. Bhowmik, R. Sarkar, M. Nasipuri, and D. Doermann, “Text and non-text separation in offline document images: A survey,” Int. J. Document Anal. Recognit. (IJDAR), vol. 21, nos. 1–2, pp. 1–20, Jun. 2018.
[10] M. Mehri, P. Héroux, P. Gomez-Krámer, and R. Mullot, “Texture feature benchmarking and evaluation for historical document image analysis,” Int. J. Document Anal. Recognit., vol. 20, no. 1, pp. 1–35, Mar. 2017.
[11] L. O’Gorman, “The document spectrum for page layout analysis,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 11, pp. 1162–1173, Nov. 1993.
[12] F. Shafait, D. Keysers, and T. M. Breuel, “Performance evaluation and benchmarking of six-page segmentation algorithms,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 6, pp. 941–954, Jun. 2008.
[13] P. A. Praczyk, J. Nogueiras-Isso, and S. Mele, “Automatic extraction of figures from scientific publications in high-energy physics,” Inf. Technol. Libraries, vol. 32, no. 4, pp. 25–52, 2013.
[14] C. Clark and S. Divvala, “PDFFigures 2.0: Mining figures from research papers,” in Proc. 16th ACM/IEEE-CS Joint Conf. Digit. Libraries (JCDL), Jun. 2016, pp. 143–152.
[15] P. Li, X. Jiang, and H. Shatkay, “Figure and caption extraction from biomedical documents,” Bioinformatics, vol. 35, no. 21, pp. 4381–4388, Nov. 2019.
[16] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1310–1319.
[17] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Computer Vision—ECCV, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Springer, pp. 3–19, 2018.
[18] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3146–3154.
[19] R. Niu, X. Sun, W. Diao, K. Chen, and K. Fu, “Hmanet: Hybrid multiple attention network for semantic segmentation in aerial images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jan. 2020, pp. 1–15.
[20] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, “BlendMask: Top-down meets bottom-up for instance segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 8573–8581.
[21] K. M. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2980–2988.
[22] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” Tech. Rep., 2016.
[23] 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.
[24] Y. Wang, Z. Xu, H. Li, S. Cheng, and L. Yang, “CenterMask: Single shot instance segmentation with point representation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9313–9321.
[25] 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.
[26] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, “Mask scoring R-CNN,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6409–6418.
[27] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo, “PolarMask: Single shot instance segmentation with polar representation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 12193–12202.

[28] S. Peng, W. Jiang, H. Pi, X. Li, H. Bao, and X. Zhou, “Deep snake for real-time instance segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 8533–8542.

[29] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, “YOLACT: Real-time instance segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2019, pp. 9156–9165.

[30] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, “YOLACT++: Better real-time instance segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., to be published.

YINGLI LIU received the Ph.D. degree in materials science from the Kunming University of Science and Technology, in 2017. Since 2005, she has been a Teacher with the Computer Science Department of Kunming University of Science and Technology. Her research areas include machine learning, natural language processing, and materials genome projects.

CHANGKAI SI is currently pursuing the master’s degree with the Kunming University of Science and Technology. His research areas include material image processing, semantic segmentation, instance segmentation, and materials genome projects.

KAI JIN is currently pursuing the master’s degree with the Kunming University of Science and Technology. His research areas include material image processing, semantic segmentation, and materials genome projects.

Y. Liu et al.: FCENet: An Instance Segmentation Model for Extracting Figures and Captions From Material Documents

TAO SHEN (Member, IEEE) received Ph.D. degree from the Illinois Institute of Technology, in 2013. He is currently a Professor with the College of Information Engineering and Automation, Kunming University of Science and Technology. His research interests include blockchain technology, smart contract, and Energy of Things (IoT).

MENG HU is currently pursuing the master’s degree with the Kunming University of Science and Technology. His research areas include blockchain, peer-to-peer transactions, and the power Internet of Things.

* * *