Signature detection is one of the crucial predecessor procedures to prove the authenticity of handwritten signatures. Despite the increasing demand for automatically verifying the signatures on contracts, bills, reports, etc., the research based on signature detection is very limited. One of the impediments to signature detection is the lack of public annotated datasets due to the interests of privacy and confidentiality. In this paper, we explore a data augmentation method, the Copy-Paste augmentation, to alleviate the scarcity of signed documents. Modeling signature detection as an object detection task, we experiment on Chinese technical documents using different object detection models. The experiment shows that YOLOv5 with Copy-Paste augmentation performs best. Based on the experiment, we propose a prototype system for signature detection.

**Index Terms**—signature detection, object detection, Copy-Paste augmentation
technical documents to explore the possibility of Chinese signature detection.

In this paper, we use Copy-Paste augmentation [1] and experiment with different signature detection models, including Faster R-CNN [2], and YOLOV5 [3] on scanned Chinese technical documents and our synthesized data to construct a pipeline signature detection system with high efficiency. To evaluate our research, we conduct ablation studies on different training schemes and different training data. In the experiment, we find that YOLOv5 outperforms Faster R-CNN, and through Copy-Paste augmentation, we can achieve a higher performance of 0.738 mAP@0.5:0.95.

The rest of this paper is organized as follows. Section II introduces the related work of Copy-Paste augmentation and signature detection. Section III demonstrates our prototype system for signature detection. Section IV describes the method we used. Section V demonstrate our experiments in details. Section VI presents the result of our ablation studies. Finally, section VII concludes our work.

II. RELATED WORK

A. Copy-Paste Augmentation

Copy-Paste augmentation is a simple way of augmenting data in the computer vision field. For object detection specifically, Copy-Paste augmentation is to paste the bounding boxes of various objects on new image backgrounds to synthesize data. In many object detection tasks, a specific class of objects may lack enough data, which can be alleviated by Copy-Paste augmentation. Georgakis et al. [4] and Gupta et al. [5] focus more on the realism of synthesized data by selecting the possible regions for objects to paste on. Dwibedi et al. [1] propose a pipeline of collecting images, pasting, and blending objects into scenes.

B. Signature Detection

With the fast development of deep learning and growing interest in the object detection field, signature detection is usually modeled as an object detection task. Thus, state-of-the-art methods in the object detection field can also be applied to signature detection. Sharma et al. [6] explores YOLOv2 [7] and Faster R-CNN on Tobacco 800 dataset [8]. Hauri [9] researches YOLOv5, Faster R-CNN, RetinaNet [10] on building applications, and the results indicate that YOLOv5 outperforms other models. Yan et al. [11] experiment Faster R-CNN, DETR [12], and YOLOv3 [13] model on their Chinese document offline signature forgery detection benchmark, ChiSig. They find that Faster R-CNN outperforms other detection methods.

III. PROTOTYPE SYSTEM

To meet the demands of authenticity verification for signatures, we build a prototype system as the signature detection module of the whole system. The prototype system proves the feasibility of our methods and further validates our methods.
Following the conclusions of our experiments and ablation studies, we decide to use Copy-Paste augmentation to build synthesized data to pretrain the detector and to use the YOLOv5 detector for our signature detection. The workflow of our prototype system is shown in Fig. 1.

To reduce labor costs and time consumption, we try our best to reduce processes requiring humans to participate. In our prototype system, two processes need humans, the annotating process and the signature collection process. The annotating process is unavoidable, but the amount of annotation can reduce because of our synthesized data. We can omit the signature collection process if the signatures on the documents are in a similar handwritten style to the signatures on the ChiSig dataset or any other public signature detection dataset.

After annotating data through CVAT [14], an open source annotation platform, we can automatically extract the background images with no signature annotations as synthesized backgrounds. And by collecting the signatures from the target group, we can alleviate the problem we may meet for different domains of signatures. Then we add signatures from the ChiSig dataset and the collected signatures on the backgrounds to build our synthesized data. Then we pretrain the YOLOv5 detector on our synthesized data and finetune the detector on annotated data afterward.

After training, we visualize the results using FiftyOne [15], a platform integrating visualization, deployment, etc. Although there are some metrics for evaluating the models, the visualization platform can help us concentrate more on the specific cases directly through our eyes. We can improve our detector and choose the best model checkpoint for deployment after analyzing the failure cases through visualization.

IV. METHOD

In this section, we introduce Copy-Paste augmentation and two detectors, Faster R-CNN and YOLOv5 models, representing two-stage and one-stage detectors, respectively. Faster R-CNN and YOLOv5 models are both state-of-the-art models in the object detection field.

A. Copy-Paste augmentation

Considering the limited pages of signed documentation, we want to synthesize more signed documents to compose training data.

The steps of Copy-Paste augmentation include: collecting signatures and backgrounds and adding the signatures to the backgrounds. Besides finding the current available Chinese signature dataset, we collect a small number of signatures from our target group to improve the detectors’ performance. As for the collection of background images, we use the document images in our dataset with no signature annotations. To improve the ability for generalization, we randomize the following variables while pasting signatures into the backgrounds:

- The number of signatures in each document
- The positions where the signatures paste on
- The size of the signatures
- The angle of rotation of the signatures
- The sharpness of the signature images

B. Faster R-CNN

Faster R-CNN is a region-proposals based two-stage detector composed of two modules, including a deep, fully convolutional network that generates region proposals, namely Region Proposal Networks(RPN) and a Fast R-CNN detector to use these proposals for detection. The RPN module applies the concept of attention mechanisms, guiding the Fast R-CNN module to where to look. Although both RPN and Fast R-CNN detector are trained independently, Faster R-CNN manages to make the RPN share features with the Fast R-CNN by sharing the same convolutional layers, which solves the bottleneck problem of region proposal computation in the previous region-proposal based algorithm.

C. YOLOv5

YOLOv5 is one of the latest detection networks of the YOLO [16] series. YOLO models the object detection task as a regression problem. By dividing the image into grids and for each grid cell predicts bounding boxes, their confidences, and the accordingly class probabilities, they can construct a tensor to encode the predicted result. Therefore, YOLO models are able to connect the process of predicting class labels and bounding boxes together. YOLOv5 integrates other advanced skills and tricks in computer vision like Cross Stage Partial Networks(CSPNet) [17] to improve the previous YOLO models. YOLOv5 is a one-stage detector which makes it inference faster than previous two-stage detectors.

V. EXPERIMENT

In this section, we conduct experiments to evaluate the performances of different detectors, including Faster R-CNN and YOLOv5, which are state-of-the-art methods in the object detection field. We report the results in the metric of mAP, the most universally used evaluation metric in the object detection field.

A. Dataset

To train our models, we collect 7,732 scanned images in total, while only 1,782 images have annotated signatures on them. We split the annotated images into 1,642 images as the training set and 140 images as the validation set. We do not split the dataset randomly. Instead, we check every image in the validation set to ensure the annotations are accurate, and each image has several signatures to represent some of the most challenging detection samples in the dataset. For example, the signatures are signed very closely and may be hard to recognize.

B. Data Augmentation

Deep learning models require a mass amount of data to train, and the scarcity of annotated signatures may hurt our models’ performance. Thus we choose to use Copy-Paste augmentation to synthesize more data. We collect more signatures and randomly paste these signatures on scanned background images to simulate authentic images for detection. To illustrate
the process more clearly, we introduce the collection of background images and handwritten signatures and the process of adding signatures to backgrounds individually in the following.

**Background Images** To simulate real technical documents, we extract 5,765 document images with no annotation as background images for two reasons:

- These documents are in the same field as authentic documents. They are all technical documents and scanned in a similar environment.
- With no annotation, our bounding boxes will not be disturbed by other handwritten texts in images.

**Handwritten Signatures** We collect handwritten signatures from two sources:

- 10,242 signatures from ChiSig [11] benchmark. Those signatures are collected from students, and their signatures are easier to recognize.
- 80 signatures collected from employees of China Qinshan Nuclear Power Co., Ltd. Their writing styles are more similar to the signatures on technical documents. We ask them to sign some unreal synthesized names to avoid privacy problems.

In order to balance the number of signatures with different handwritten styles, we reproduce 100 times of signatures to get 8,000 signatures of employees and merge these signatures with those from the ChiSig benchmark.

**Adding signatures to backgrounds** We randomly select signatures to paste on background images at random positions. Each background image has 3 or 4 signatures on it. All signatures are pasted with a random size of 5% – 10% of the background image size and a rotation within 10°. Also, we vary the sharpness of each signature from 0.8 to 1. All the random processes through synthesizing obey the uniform distribution. By repetitively synthesizing 3 times with the same background and different signatures, we get 17,295 images in total. We randomly split the whole dataset 9:1 as the training set and the validation set afterward.

The dataset we used for experiments is described in detail in Tab. I. A synthetic sample image is shown in Fig. 2.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>DETAILED INFORMATION ABOUT DATASETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>Validation set</td>
</tr>
<tr>
<td>Annotated data</td>
<td>1,642</td>
</tr>
<tr>
<td>Synthesized data</td>
<td>15,566</td>
</tr>
</tbody>
</table>

**C. Training Details**

To fully use our synthesized data, we first pretrain our detectors on the synthesized data, then initialize the detectors with the pretrained weights and finetune the detectors on the annotation data in the real situation.

We use MMDetection [18], an open-source object detection toolbox, to help train the Faster R-CNN detector. We use ResNet50 as the backbone of the detector and SGD as the optimizer. We train the Faster R-CNN detector for 12 epochs with a learning rate of 2.5e-3 and a batch size of 2. As for the YOLOv5 detector, we initialize the detector with the YOLOv5n6 pretrained model checkpoint, which has the fattest inference speed on both CPU and V100 with the pixels size of 1280 from the checkpoints Ultralytics provide. We use SGD as the optimizer. We train the YOLOv5 for 300 epochs with a learning rate of 1e-2 and a batch size of 16. We keep the same hyperparameter set through pretraining and finetuning.

**D. Evaluation Metrics**

To compare detection performance universally, we evaluate our methods through the most common metric – mean Average Precision (mAP). To compute mAP, first, we need to decide whether the bounding box is correctly detected by computing Intersection over Union (IoU). The definition of IoU is as the following equation (1), where A and B mean the area of the ground truth bounding box and predicted bounding box, respectively.

$$\text{IoU} = \frac{A \cap B}{A \cup B}$$  (1)
Then we need to calculate the precision and recall metrics, which are defined as:

\[
\text{Precision} = \frac{TP}{TP + FP} \\
\text{Recall} = \frac{TP}{TP + FN}
\]

where TP, FP, and FN represent true positive, false positive, and false negative, respectively.

Then AP is the area under the precision-recall curve, and mAP is the mean of APs of different categories. Since our experiment only involves one class-signatures, mAP is the same metric as AP.

According to the COCO format [19] of evaluation, we compute mAP when IoU is 0.5, which is the most used threshold for IoU in the object detection field, and IoU ranges from 0.5 to 0.95 with a step of 0.05, which can present the ability of detectors more comprehensively.

E. Results

We use the aforementioned methods to test on Chinese technical documents. The main results are shown in the following Tab. II.

<table>
<thead>
<tr>
<th>method</th>
<th>mAP@0.5</th>
<th>mAP@0.5:0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN</td>
<td>0.952</td>
<td>0.635</td>
</tr>
<tr>
<td>YOLOv5</td>
<td>0.988</td>
<td>0.738</td>
</tr>
</tbody>
</table>

The results show that YOLOv5 outperforms Faster R-CNN. We can also see that the mAP@0.5 metrics of both models are pretty satisfying, which indicates the efficiency of modeling signature detection as object detection. Compared with mAP@0.5, mAP@0.5:0.95 is much lower because the higher the IoU threshold, the number of missing boxes is higher.

VI. ABLATION STUDY

In this section, we conduct an ablation study to verify the necessity of Copy-Paste augmentation and discuss the results of the different training schemes.

A. Copy-Paste Augmentation

In order to test the effect of adding synthesized data, we run experiments only on annotated data. The results are shown in Tab. III:

<table>
<thead>
<tr>
<th>method</th>
<th>mAP@0.5</th>
<th>mAP@0.5:0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN</td>
<td>0.973</td>
<td>0.631</td>
</tr>
<tr>
<td>YOLOv5</td>
<td>0.990</td>
<td>0.720</td>
</tr>
</tbody>
</table>

We can compare the above results with the main results in section V. The results suggest that training only on annotated data can acquire quite well results. The mAP@0.5 of Faster R-CNN and YOLOv5 is even better than the detectors trained after Copy-Paste augmentation. This suggests that the original amount of annotated data is enough to produce good results. However, in terms of mAP@0.5:0.95, the detectors get better after Copy-Paste augmentation. As mAP@0.5:0.95 can present the performance of models more comprehensively, suggesting that Copy-Paste augmentation indeed makes a difference.

To demonstrate the effect of Copy-Paste augmentation more clearly, we can look at the loss curve during training, shown in Fig. 3. The initial loss of the YOLOv5 model is much lower after pretraining on synthesized data, suggesting that the detector learns useful knowledge from synthesized data. Although the detector without augmentation quickly converges, Fig. 3 shows that we can get better results even when we do not have enough annotated data by Copy-Paste augmentation.

![Fig. 3. Speed of convergence of YOLOv5](image)

B. Training Scheme

During the experiment, we find that different training schemes can significantly affect the final results. We try to train the detectors by two different approaches:

- We merge the synthesized data and annotated data to train detectors.
- We first pretrain the detectors using synthesized data and then finetune the model using annotated data.

We demonstrate our findings using Faster R-CNN as an example detector, and the results are shown in the following Tab. IV:

<table>
<thead>
<tr>
<th>Training Scheme</th>
<th>mAP@0.5</th>
<th>mAP@0.5:0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>merge dataset</td>
<td>0.834</td>
<td>0.593</td>
</tr>
<tr>
<td>pretrain and finetune</td>
<td>0.952</td>
<td>0.635</td>
</tr>
</tbody>
</table>
From the results above, we find that pretrain the model using synthesized data can produce better results than merging synthesized and annotated data. We have more epochs to learn from data by pretraining and finetuning the model. By pretraining on synthesized data, the detectors learn prior knowledge and, finetuning on annotated data, give the model the right direction to imitate. However, simply merging the synthesized and annotated data can confuse the model to learn unexpected features, thus harming the performance.

VII. CONCLUSION

In this paper, we explore different detectors on scanned Chinese technical documents and use Copy-Paste augmentation to enlarge our dataset. With respect to different detectors, we find YOLOv5 outperforms Faster R-CNN on signature detection. We conduct ablation studies on Copy-Paste augmentation and different training schemes to prove that Copy-Paste augmentation is valid and that the pretrained and finetuned scheme can improve the results remarkably. Based on the experiment results and ablation studies, we choose to build our prototype system using the YOLOv5 detector and Copy-Paste augmentation.

In future work, we hope to improve our approach by making suitable changes to Copy-Paste augmentation, like trying various blending skills. Also, we would like to experiment with more object detection models and other data augmentation methods to push the current limits of signature detection.

VIII. ACKNOWLEDGMENT

This work was funded by the National Key Research and Development Program of China (2020YFB1707803). We would like to thank China Qinshan Nuclear Power Co., Ltd., which provides scanned technical reports for us to research on. This research was led by Principal Supervisor Prof. Hongwei Wang and Prof. Gaoang Wang.

REFERENCES